Margin requirements for first page
Paper size this page US Letter
72 pt
1 in
25.4 mm

LH-DETR: A Lightweight Hybrid Architecture for End-to-End Object Detection in UAV Images

Feifei Xu[1], Lupeng Sun[1], Dongyang Li[1]*, Guoxiang Wu[1], and Chenchuan Lv[1]

***Abstract*— Object detection in unmanned aerial vehicles (UAVs) has become a research highlight at the intersection of computer vision and robotics technology, and its applications in security inspection, agricultural monitoring, disaster relief and others are becoming increasingly widespread. The key to achieving autonomous perception and decision-making of UAV lies in precise and real-time object detection. However, objects from the perspective of UAV often have characteristics such as small scale and dense distribution, coupled with limited onboard computing resources, which poses significant challenges to traditional detection algorithms. To address the trade-offs, this paper proposes LH-DETR, a lightweight hybrid architecture for end-to-end object detection, referring to three specialized innovations. We propose a Wavelet-Mamba Hybrid Block (WMHB), a novel backbone component that synergistically combines the linear-complexity of Mamba state-space model for capturing long-range dependencies with the multi-scale feature extraction capabilities of wavelet transforms. To better identify small objects, a Frequency-Aware Dynamic FFN (FAD-FFN) is designed to selectively amplify critical high-frequency components—like edges and textures—by analyzing features in the frequency domain. Additionally, AutoSliding Varifocal Loss (ASVLoss) is defined to stabilize the model's optimization, which is an adaptive loss function that dynamically shifts its focus from medium-quality to high-quality predictions as training progresses. Experiments on public aerial datasets demonstrate that LH-DETR achieves an outstanding balance between accuracy and speed, significantly improving detection performance for small objects while greatly reducing the computational complexity.**

I. INTRODUCTION

Object detection in Unmanned Aerial Vehicle (UAV) has emerged as a critical research focus, driven by applications in smart cities, precision agriculture, modern logistics, as well as military and national defense[1]. Although there has been significant progress in object detection for low-resolution images, these methods present considerable challenges in terms of efficiency and accuracy in high-resolution aerial images. Especially from a high-altitude perspective, the object size in the image is relatively small, densely distributed, and the background is complex and cluttered. Furthermore, the operational nature of UAVs necessitates real-time processing on resource-constrained onboard hardware, demanding model's high accuracy and efficient computing capability.

Traditional object detection has undergone a significant shift toward end-to-end architectures, which streamline the detection pipeline by eliminating hand-crafted components like anchor generation and non-maximum suppression (NMS). Detection Transformer (DETR)[2], [3]is the groundbreaking work, which applies a Transformer-based architecture to directly predict a set of objects. However, despite its design advantages, the original DETR suffers from extremely slow training convergence and large computational costs, making it difficult to adapt to real-time applications, especially in the context of UAVs.

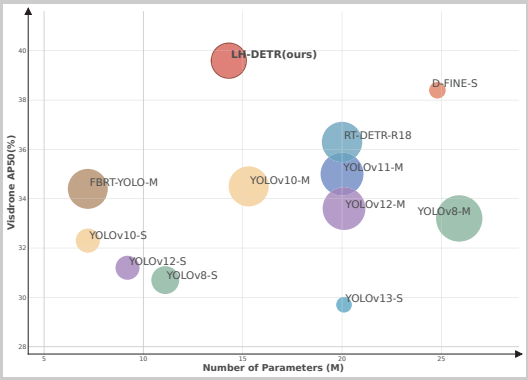

Fig. 1. LH-DETR is compared with other real-time detectors in terms of accuracy and efficiency on VisDrone dataset. The radius of the circle represents GFLOPs.

Based on the above work, more efficient variants have emerged. Deformable DETR[4] introduces the attention mechanism that focuses on a sparse set of key points, accelerating convergence. More recently, RT-DETR[5] achieves a major breakthrough by employing a hybrid encoder that effectively decouples intra-scale interaction and cross-scale fusion, making real-time, end-to-end object detection a feasible reality. Nevertheless, even state-of-the-art models like RT-DETR still face some ongoing challenges. As their reliance on high-resolution feature maps results in significant computational overhead, they still struggle to detect small, detail-rich objects prevalent in UAV images. Moreover, their internal Feedforward Networks (FFNs) deal with all spatial features uniformly, but high-frequency details such as edges and textures are crucial for identifying small objects.

To address these issues, we propose LH-DETR, a lightweight hybrid architecture for end-to-end object detection tailored specifically for UAV images. Our framework involves several specialized innovations designed to enhance both efficiency and accuracy in detecting multi-scale objects.

- We propose the Wavelet-Mamba Hybrid Block (WMHB), a novel backbone component that synergistically combines the linear-complexity, long-range dependency modeling of Mamba state-space models[6,7,8,9] with the multi-scale feature

*Corresponding author.
[1]The authors are with Shanghai University of Electric Power, Shanghai 201306, China (e-mail: dongyangli.ldy@shiep.edu.cn).

54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm

Margin requirements for the other pages
Paper size this page US Letter
54 pt
0.75 in
19.1 mm

extraction capabilities of wavelet transforms, allowing our model to efficiently capture both global context and fine-grained local details.
- To further improve small object detection, we design the Frequency-Aware Dynamic FFN (FAD-FFN), explicitly analyzing frequency-domain features to selectively enhance critical high-frequency components necessary for identifying small objects, while simultaneously reducing the number of parameters.
- We define AutoSliding Varifocal Loss (ASVLoss), which adaptively shifts the training focus from medium-quality predictions in early stages to high-quality predictions later.

Extensive experiments conducted on public aerial imagery benchmarks, including VisDrone, UAVVaste, and UAV-PDD, demonstrate that our proposed LH-DETR substantially outperforms previous detectors in the trade-off between parameter size and accuracy (AP), as illustrated in Fig. 1.Together, these contributions enable LH-DETR to set a new standard for efficient and accurate real-time object detection on UAV platforms.

II. Related Work

A. End-to-End Object Detectors

End-to-end object detectors have recently attracted significant attention due to their simplified architectures, which remove traditional components such as anchor generation and non-maximum suppression (NMS). Among these, DETR (Detection Transformer)[2], proposed by Carion et al. , stands out as the first model to incorporate Transformer-based architecture in object detection. It directly predicts objects without the need for anchors or NMS, offering a streamlined approach to detection tasks. However, despite its advantages, DETR faces several challenges, particularly slow convergence and high computational costs. To address these limitations, several variations have been proposed. Deformable DETR[4], for instance, improves upon DETR by introducing deformable attention, which better handles high-resolution feature maps by sampling a sparse set of key points. Furthermore, RT-DETR[5] introduces a hybrid encoder-decoder architecture that improves multi-scale feature processing and optimizes real-time inference by decoupling intra-scale interactions and cross-scale fusion. These advancements make real-time, end-to-end object detection feasible in resource-constrained environments like UAVs, where fast inference is crucial for practical deployment.

B. State Space Models in Vision

In recent years, State Space Models (SSMs)[6] have generated growing interest in vision tasks, primarily due to their linear complexity in capturing long-range dependencies. Initially inspired by control systems, SSMs treat inputs and outputs as sequences, with hidden states evolving according to a predefined dynamic model. Mamba[7], as a state-of-the-art SSM architecture, introduces an efficient input selection mechanism, making it highly suitable for vision tasks. Mamba is optimized for GPUs through a hardware-aware algorithm that leverages kernel fusion, parallel scanning, and recomputation to enhance both training and inference efficiency. Several adaptations of Mamba have further advanced its performance in visual tasks. For instance, Vim[8]introduces a bidirectional Mamba block, which demonstrates superior speed and memory advantages over Vision Transformers (ViTs)[9], [10] in high-resolution settings. VMamba[11] and EfficientVMamba[12] have proposed innovative Cross-Scan and Efficient Scan techniques to further boost the model's efficiency in handling visual data.

Building upon these advancements, our work integrates the Mamba model with wavelet transforms to propose a lightweight visual network, the Wavelet-Mamba Hybrid Block (WMHB). This module concurrently models long-range dependencies with high efficiency while designing wavelet transforms for multi-scale feature extraction.This synergy significantly enhances both feature representation and computational efficiency for complex UAV images.

C. Real-Time Object Detection in UAV Images

Detecting objects in UAV images presents unique challenges, especially when dealing with small objects and occlusions, which are common in high-resolution aerial images. Although traditional object detection models like YOLO[13], [14], [15], [16], [17], [18] and FCOS[19] perform well in general settings, they struggle with the complexities of small object detection in UAV images, where computational efficiency and high accuracy are both required. Techniques such as feature pyramids and multi-scale processing have been proposed to enhance small object detection by improving feature representation at various scales. YOLOv4[20] and YOLOv5[21] are popular single-stage models for real-time object detection, but they still suffer from the need for NMS, which slows down inference and reduces the overall efficiency.

In contrast, RT-DETR[5] has shown great promise in real-time UAV object detection, surpassing traditional YOLO-based methods by eliminating NMS and using a Transformer-based architecture to model long-range dependencies. However, RT-DETR still faces challenges in detecting small objects due to its reliance on high-resolution feature maps, which are computationally expensive. Our work addresses these issues by incorporating multi-scale feature fusion that integrates both spatial and frequency-domain information, significantly improving small object detection in UAV images without the high computational cost.

D. Feature Fusion and Multi-Scale Information Extraction

Feature fusion and multi-scale information extraction are essential for object detection, particularly in scenarios involving small or densely packed objects. Feature Pyramid Networks (FPN) have been widely used for combining features from different layers of a neural network to enhance object detection performance. However, the semantic gaps between different layers' feature maps often lead to misalignment, especially when fusing high-level semantic features with low-level spatial information. Methods like PANet[22] introduce

54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm

Margin requirements for the other pages
Paper size this page US Letter
54 pt
0.75 in
19.1 mm

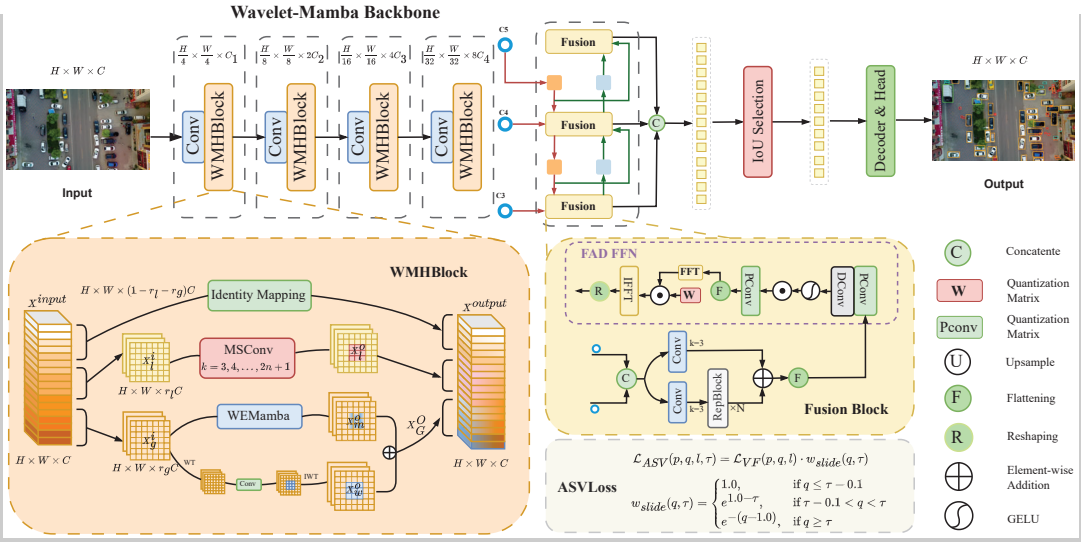

Fig. 2. Framework of LH-DETR.The model's backbone exploits the Wavelet-Mamba Hybrid Block (WMHB) for efficient multi-scale feature extraction, capturing both global and local details. A Frequency-Aware Dynamic FFN (FAD-FFN) is designed to enhance the perception of small objects by amplifying critical high-frequency details. Finally, the novel AutoSliding VarifocalLoss (ASVLoss) improves model accuracy through a more robust and precise optimization.

additional bottom-up paths to improve information flow, but these approaches often increase computational complexity, which is undesirable in real-time systems.

Our approach combines multi-scale feature fusion in both the spatial and frequency domain, leveraging the power of wavelet transforms to capture fine-grained details in high-frequency components while preserving global contextual information in low-frequency components. This strategy not only improves the accuracy of small object detection but also maintains efficiency, making it suitable for real-time applications in UAVs. Furthermore, by using selective scanning and dynamic frequency processing techniques, our model optimizes feature alignment and fusion, offering a significant advantage over traditional feature fusion methods.

III. Method

A. Wavelet-Mamba Hybrid Block

To address the challenges posed by dense multi-scale objects, high proportion of small objects, and complex backgrounds in UAV images, we propose the Wavelet-Mamba Hybrid Block (WMHB) as the main component of our backbone network. The WMHB module, depicted in Fig. 2, innovatively integrates wavelet analysis with the Mamba state-space model to form an efficient and expressive feature extraction unit. This module employs a channel-splitting strategy, partitioning the input feature tensor $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ into three mutually exclusive subspaces:

$$\mathbf{X} = [\mathbf{X}_{\text{global}}, \mathbf{X}_{\text{local}}, \mathbf{X}_{\text{identity}}]. \tag{1}$$

Here $\mathbf{X}_{\text{global}}$, $\mathbf{X}_{\text{local}}$, and $\mathbf{X}_{\text{identity}}$ are subspaces containing fractions r_g, r_l, and $1-r_g-r_l$ of the total channels, respectively. This partitioning allows for computationally independent processing paths while subsequent fusion preserves full feature expressiveness. To optimize GPU memory access, the number of global channels is always maintained as a multiple of 16. Each subspace is then handled by a specialized branch:

- **Wavelet-Enhanced Mamba (WEmamba):** Models long-range dependencies for the global feature channels.
- **Multi-Scale Convolution (MSConv):** Extracts local features at multiple receptive fields.
- **Identity Mapping:** Preserves information by directly passing the remaining channels.

Finally, the outputs from each branch are fused via channel concatenation and a projection layer, with a residual connection to the input:

$$\mathbf{X}^{\text{out}} = \mathscr{P}([\mathbf{X}^{\text{out}}_{\text{global}}, \mathbf{X}^{\text{out}}_{\text{local}}, \mathbf{X}^{\text{out}}_{\text{identity}}]) + \mathbf{X}. \tag{2}$$

This design effectively balances global modeling capability with computational complexity, making it particularly suitable for multi-scale feature extraction in UAV images.

1) Wavelet-Enhanced Mamba: The Wavelet-Enhanced Mamba (WEmamba) is designed to address the high computational complexity of traditional Transformers on high-resolution UAV images by synergistically modeling high-frequency details and global dependencies with a combination of the wavelet transform and the Mamba state-space model. First, the input feature tensor $\mathbf{X}^I \in \mathbb{R}^{h\times w\times c_g}$ is decomposed into multi-band features via a Discrete Wavelet Transform (DWT):

$$\mathbf{X}^I_{wt} = \text{DWT}\left(\mathbf{X}^I, [\phi_{LL}, \phi_{LH}, \phi_{HL}, \phi_{HH}]\right). \tag{3}$$

Here, $\phi_{LL}, \phi_{LH}, \phi_{HL}, \phi_{HH}$ are the low-pass and high-pass filters generated by the Daubechies 1 (db1) wavelet basis. The resulting high-frequency components $(\phi_{LH}, \phi_{HL}, \phi_{HH})$capture edge and texture information, while the low-frequency component (ϕ_{LL}) preserves the global structure. Subsequently, the high-frequency features are enhanced for local perception using depthwise separable

54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm

Margin requirements for the other pages
Paper size this page US Letter
54 pt
0.75 in
19.1 mm

convolutions, while the low-frequency features are fed into the Mamba module for long-range dependency modeling:

$$\mathbf{X}_m^O = \mathrm{SS2D}\left(\sigma\left(\mathrm{Linear}\left(\mathbf{X}_{LL}^I\right)\right)\right), \tag{4}$$

where σ denotes the GELU activation function. The SS2D component employs selective scanning to achieve global modeling with linear complexity. In contrast to the quadratic complexity $O(n^2)$ of traditional self-attention, Mamba's computational complexity is linear $O(n)$, making it highly suitable for large-scale UAV image inputs. Finally, the processed components are concatenated and reconstructed back to the original feature dimensions using an Inverse Discrete Wavelet Transform (IDWT):

$$\mathbf{X}_w^O = \mathrm{IDWT}\left(\mathrm{Concat}([\mathbf{X}_m^O, \mathbf{X}_{LH}^O, \mathbf{X}_{HL}^O, \mathbf{X}_{HH}^O])\right). \tag{5}$$

The key advantage of WEmamba lies in the synergy: the wavelet decomposition provides a rich, multi-scale feature representation, while Mamba efficiently establishes long-range dependencies within that representation. This allows the model to capture both global context and fine-grained local details simultaneously.

2) Multi-Scale Convolution: The Multi-Scale Convolution (MSConv) block is a critical component in the WMHB responsible for local feature extraction. It employs a depthwise separable convolution structure to efficiently capture features across multiple receptive fields. The input local feature tensor $\mathbf{X}_L^I \in \mathbb{R}^{h\times w\times c_l}$ is processed in parallel by depthwise convolutions with different kernel sizes. The resulting feature maps are then concatenated along the channel dimension:

$$\mathbf{X}_j^O = \mathrm{DConv}\left(\mathbf{X}_L^I, k_j\right), \quad j \in \{3,5,7\}, \tag{6}$$

$$\mathbf{X}_L^O = \mathrm{Concat}\left([\mathbf{X}_3^O, \mathbf{X}_5^O, \mathbf{X}_7^O], \mathrm{dim}=1\right), \tag{7}$$

where k_j represents convolution kernels with varying receptive fields ($3\times3, 5\times5, 7\times7$). This multi-kernel design significantly reduces the parameter count and computational load—achieving an 87% parameter reduction compared to standard convolutions—while simultaneously enhancing the model's adaptability to multi-scale targets.

Within the WMHB architecture, the proportion of feature channels allocated to the MSConv block is typically configured to be between 20% and 30%. This allocation strategy maximizes local feature extraction efficiency under constrained computational resources. For UAV images characterized by dense small objects, MSConv excels at capturing essential boundary and texture information, thereby complementing the global context modeling provided by the WEmamba block.

3) Identity Mapping and Channel Proportion Control: To balance computational efficiency with information integrity, a proportion of the input channels, defined by $1-\xi-\eta$, is preserved via an identity mapping. Here,ξ and η represent the channel proportions allocated to the global (WEmamba) and local (MSConv) branches, respectively. The processed features from all three branches are then fused to reconstruct the final output tensor:

$$\mathbf{X}^O = \mathrm{Proj}\left(\mathrm{Concat}\left[\mathbf{X}_w, \mathbf{X}_L, \mathbf{X}_{\mathrm{identity}}\right]\right), \tag{8}$$

where $\mathrm{Proj}(\cdot)$ is a 1×1 convolutional projection layer that restores the channel dimension, and $\mathbf{X}_{\mathrm{identity}}$ denotes the channels passed through the identity mapping. Our experiments on the VisDrone dataset show that an allocation of $\xi = 0.6$ and $\eta = 0.3$ yields an optimal trade-off, improving the AP_{50} by 2.4% with only a 3.2% increase in inference latency.

B. Frequency-Aware Dynamic FFN

The Axial Intra-Scale Feature Interaction (AIFI) module in RT-DETR[5] enables efficient global feature modeling via sparse attention. However, its traditional Feedforward Network (FFN) component handles all spatial positions uniformly, disregarding the varying importance of different frequency components within image features. This limitation is particularly detrimental for UAV-based object detection, where small objects are often characterized by high-frequency details such as edges and textures.

To address this issue, we propose the Frequency-Aware Dynamic FFN (FAD-FFN), a lightweight but powerful enhancement to the standard FFN that explicitly models and exploits multi-frequency information. Unlike methods confined to the spatial domain, FAD-FFN selectively amplifies critical frequency components identified through frequency-domain analysis, all while maintaining computational efficiency.

1) Architecture Design: The FAD-FFN module employs a three-stage pipeline that deals with features in the frequency domain while preserving spatial information. As illustrated in Fig. 2, the architecture consists of the following stages:

a) Local Frequency Transformation.: The input feature tensor $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ is first partitioned into non-overlapping 8×8 patches, yielding $\frac{H}{8}\times\frac{W}{8}$ distinct blocks. A 2D Real-valued Fast Fourier Transform (RFFT) is then applied to each block:

$$\mathbf{X}_{\mathrm{patch}} = \mathrm{Reshape}(\mathbf{X}) \xrightarrow{\mathrm{RFFT}} \mathbf{X}_{\mathrm{freq}} \in \mathbb{R}^{C\times\frac{H}{8}\times\frac{W}{8}\times8\times5}, \tag{9}$$

where the dimension of 5 corresponds to the unique frequency points $(8/2+1)$ obtained due to the Hermitian symmetry of the RFFT of a real-valued input.

b) Learnable Frequency Filtering.: A channel-specific learnable filter, $\mathbf{W}_{\mathrm{freq}} \in \mathbb{R}^{C\times1\times1\times8\times5}$, is applied to the frequency-domain representation via element-wise multiplication: $\mathbf{Y}_{\mathrm{freq}} = \mathbf{X}_{\mathrm{freq}} \odot \mathbf{W}_{\mathrm{freq}}$. This filter, initialized with all ones, is trained to selectively amplify frequency components most relevant to the object detection task. Unlike dynamic mechanisms that generate filters from input content, our approach uses a fixed set of learnable parameters, ensuring computational stability while providing frequency-aware enhancement.

c) Feature Reconstruction and Fusion.: The filtered frequency-domain features are transformed back into the spatial domain using an Inverse RFFT (IRFFT). This output is then fused with a parallel, lightweight spatial-processing branch:

$$\mathbf{Y} = \mathrm{IRFFT}(\mathbf{Y}_{\mathrm{freq}}) + \mathscr{F}(\mathbf{X}), \tag{10}$$

where $\mathscr{F}(\cdot)$ denotes a sequence of depthwise separable convolutions, a GELU activation, and a channel projection.

54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm

Margin requirements for the other pages
Paper size this page US Letter
54 pt
0.75 in
19.1 mm

This dual-branch design ensures that both the original spatial structure and the enhanced frequency characteristics are retained.

2) Computational Efficiency: FAD-FFN is designed to achieve an optimal balance between feature enhancement and computational efficiency. Its advantages in this regard are detailed below.

a) Parameter Efficiency.: By redesigning the standard FFN, FAD-FFN reduces the total parameter count by 25%, from $8C^2$ down to $6C^2$. The additional frequency-domain filter only introduces $5C$ parameters (one for each of the 5 frequency points per channel), a negligible amount relative to the overall model size.

b) Computational Complexity.: A theoretical complexity analysis highlights the efficiency of our approach:Standard FFN (4x expansion): $8C^2HW$ FLOPs;FAD-FFN: $6C^2HW + 18CHW + 12HW$ FLOPs. The computational cost of FAD-FFN is lower than that of a standard FFN when $C > 10$, a condition that holds true in most practical scenarios where the channel count (C) is typically 256 or greater. The use of local 8×8 block processing ensures that the FFT operation's contribution to the overall computation is minimal.

c) Practical Performance.: On the VisDrone dataset (with an input size of 640×640), FAD-FFN increases inference latency by only 3.2% compared to a more streamlined 2x expansion FFN baseline, while, as shown in our experimental results, significantly improving small object detection performance. This makes it highly suitable for real-time object detection on UAV platforms.

Unlike prior frequency-domain methods, our approach is notable for its simplicity and efficiency. FAD-FFN avoids complex dynamic mechanisms or global frequency transformations, opting instead for local frequency processing with minimal overhead. This design makes it highly practical for resource-constrained environments while delivering substantial performance gains.

C. AutoSlidingVarifocalLoss: An Adaptive Threshold Quality-Aware Loss Function

The design of the loss function has a decisive impact on the performance of object detection models. Traditional cross-entropy loss, for instance, exhibits significant limitations when confronted with the severe class imbalance between positive and negative samples. Although Focal Loss mitigates this issue by re-weighting hard and easy samples, it does not sufficiently account for continuous gradients in prediction quality. More recently, Varifocal Loss[23] introduces an asymmetric weighting strategy using Intersection over Union (IoU) as a soft label, thereby improving the detector's focus on high-quality predictions.

Despite these advances, we observe that a model's attentional focus should change dynamically throughout training. In the early stages, learning should prioritize medium-quality predictions to establish a robust performance baseline quickly. Conversely, in the later stages, the focus must shift to high-quality predictions to refine and optimize the final accuracy. This observation motivates the need for a more adaptive loss function.

Based on this observation, we propose the AutoSlidingVarifocalLoss (ASVLoss), a loss function featuring an adaptive weight-adjustment mechanism. ASVLoss partitions the IoU axis into three continuous intervals, each assigned a weighting factor that evolves dynamically throughout the training. This is controlled by a dynamic reference point, τ, defined as $\tau = \max(\text{auto_iou}, 0.2)$, where auto_iou is the average IoU value of the current training batch. The loss function reweights predictions based on three key regions of quality, determined by the ground-truth IoU (q) and the dynamic reference point (τ):

$$w_{\text{slide}}(q,\tau) = \begin{cases} 1.0, & \text{if} \quad q \leq \tau - 0.1 \\ e^{1.0-\tau}, & \text{if} \quad \tau - 0.1 < q < \tau \\ e^{-(q-1.0)}, & \text{if} \quad q \geq \tau \end{cases}. \tag{11}$$

This weighting factor, $w_{\text{slide}}(q,\tau)$, is then applied to a standard Varifocal Loss formulation.

In these equations, p is the predicted score, q is the ground truth IoU, l is the class label, $\sigma(\cdot)$ represents the sigmoid function, and α and γ are hyperparameters.

- **Low-Quality Region ($q \leq \tau - 0.1$):** The weight is fixed at 1.0. This ensures that low-quality samples continue to provide valid gradients, which is particularly crucial during the early stages of training.
- **Transition Region ($\tau - 0.1 < q < \tau$):** The weight is set to $e^{1.0-\tau}$. This creates a smooth transition between the low- and high-quality regions, preventing optimization instability that may arise from abrupt changes in the loss landscape.
- **High-Quality Region ($q \geq \tau$):** The weight decays exponentially according to the factor $e^{-(q-1.0)}$. This forces the model to increasingly focus on high-quality samples in the later stages of training, thereby improving the ranking of precise predictions.

The final AutoSlidingVarifocalLoss is formulated by multiplying the standard Varifocal Loss($\mathscr{L}_{\text{VF}}$) with our dynamic weighting factor:

$$\mathscr{L}_{\text{ASV}}(p,q,l,\tau) = \mathscr{L}_{\text{VF}}(p,q,l) \cdot w_{\text{slide}}(q,\tau) \tag{12}$$

As training progresses, the reference point τ initialized at 0.2 gradually increases as the model's average prediction quality improves. This causes the boundaries of the three weighting regions to "slide" upwards along the IoU axis, enabling a smooth transition of the model's learning focus. Initially, when τ is small, the transition region receives a high weight, encouraging the model to rapidly establish a robust detection baseline by focusing on medium-quality predictions. In the later stages of training, as τ increases, the high-quality region becomes the primary optimization focus. This shift compels the model to refine its predictions on the most accurate bounding boxes, thereby driving the mean Average Precision (mAP) to higher levels.

54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm

Margin requirements for the other pages
Paper size this page US Letter
54 pt
0.75 in
19.1 mm

IV. Experiments

Implementation Details: All models are trained on an NVIDIA GeForce RTX 4090 GPU and tested for inference speed on a single NVIDIA GeForce RTX 3090 GPU to ensure a fair comparison. We train all models for 300 epochs using the Stochastic Gradient Descent (SGD) optimizer. We set the initial learning rate to 0.01, momentum to 0.937, weight decay to 0.0005, and batch size to 4. Unless otherwise specified, the input image resolution is resized to 640×640.

TABLE I

Comparison of detection performance (AP%, AP_{50}, AP_S, AP_M, AP_L) and model complexity (Params/FLOPs/FPS) on VisDrone. Detectors are grouped by category. Best results are in **bold** and second-best are underlined.

Model	Params	FLOPS	FPS	AP(%)	AP_{50}(%)	AP_S(%)	AP_M(%)	AP_L(%)
Real-time Object Detectors								
YOLOv8-S[24]	11.1M	28.5G	-	17.3	30.7	7.8	26.9	37.2
YOLOv8-M[24]	25.9M	78.9G	-	19.0	33.2	9.0	29.4	41.7
YOLOv10-S[14]	7.2M	21.4G	-	17.9	32.3	8.6	27.8	36.1
YOLOv10-M[14]	15.3M	58.9G	-	19.5	34.5	9.7	30.0	41.4
YOLOv11-M[15]	20.0M	67.7G	52.0	20.3	35.0	9.8	31.2	41.3
YOLOv12-S[16]	9.2M	21.2G	-	17.6	31.2	8.1	27.4	35.6
YOLOv12-M[16]	20.1M	67.2G	-	19.2	33.6	9.4	29.8	38.6
YOLOv13-N[17]	**6.2M**	**2.45G**	-	13.3	24.4	5.5	21.0	31.7
YOLOv13-S[17]	20.1M	9.0G	-	16.7	29.7	7.7	25.8	38.7
FBRT-YOLO-M[18]	7.2M	58.7G	-	19.6	34.4	9.4	30.9	42.1
End-to-end Object Detectors								
RT-DETR-R18[5]	20.0M	60.0G	55.6	20.8	36.3	11.3	30.5	41.3
RT-DETR-R50[5]	42.0M	136G	53.5	22.2	39.1	12.7	32.1	45.6
D-FINE-S[25]	24.8M	10.1G	-	21.9	38.4	12.2	32.1	39.7
D-FINE-M[25]	56.3M	19.1G	-	**24.2**	**41.7**	**13.9**	34.4	48.5
Real-time End-to-end Object Detectors for UAV Images(ours)								
LH-DETR(ours)	14.3M	44.9G	**57.0**	22.4	39.6	12.9	**41.8**	**55.7**

A. Results on Visdrone Dataset

We first compare LH-DETR with a wide range of SOTA detectors on the challenging VisDrone dataset. As shown in Table I, our model demonstrates a superior balance of accuracy and efficiency. Compared to the baseline RT-DETR-R18, LH-DETR achieves significant performance gains: Although reducing the parameter count by 28.5% (from 20.0M to 14.3M) and lowering FLOPs by 25.2% (from 60.0G to 44.9G), our model improves the overall AP by 1.6 percentage points (22.4% vs. 20.8%). Crucially, for small object detection (AP_S), the performance is boosted by 1.6 percentage points (12.9% vs. 11.3%).

Even when compared to the much larger RT-DETR-R50, LH-DETR surpasses it in overall AP by 0.2 percentage points while using only about one-third of the parameters and computational cost. Furthermore, our model shows remarkable improvements in detecting medium and large objects, increasing AP_M by 9.7 percentage points and AP_L by 10.1 percentage points, respectively. Although D-FINE-M achieves a higher overall AP of 24.2% and better small object detection (AP_S of 13.9%), it is nearly four times larger in terms of parameters (56.3M vs. our 14.3M), making it less suitable for deployment on resource-constrained UAV hardware. In contrast, our LH-DETR provides a more balanced profile, significantly outperforming D-FINE-M in detecting medium and large objects while maintaining a lightweight architecture ideal for real-world aerial applications.

In Table II, we further compare LH-DETR with methods specifically designed for UAV images. The results show

TABLE II

Comparison of AP(%)and Params/FPS with UAV Image detectors on VisDrone.The best results are in **bold**, and the second-best are underlined.

Model	Params	FLOPS	FPS	AP(%)	AP_{50}(%)
QueryDet[26]	33.9M	212.0G	18.2	20.1	39.6
ClusDet[27]	30.2M	207.0G	15.6	18.5	**42.3**
DCFL[28]	36.1M	157.8G	22.4	-	24.1
HIC-YOLOv5[29]	**9.4M**	**31.2G**	**61.0**	17.8	36.0
LH-DETR(ours)	14.3M	44.9G	57.0	**22.4**	39.6

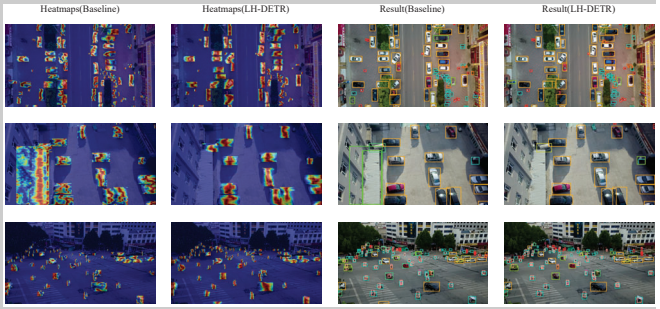

Fig. 3. Visualization results of the detection heatmaps on VisDrone. The highlighted areas represent the regions that the network is focusing on.

that, compared to QueryDet, our model increases AP by 2.3 percentage points (22.4% vs. 20.1%) while using 57.8% fewer parameters. Similarly, although ClusDet achieves a high AP_{50} of 42.3%, its overall AP is only 18.5%, falling significantly short of the performance of our model and requiring substantially more parameters and FLOPs. This demonstrates that our lightweight hybrid design is not only efficient but also more effective than existing specialized approaches for complex aerial scenes.

As a result, for more experiments, we mainly select RT-DETR as the primary baseline for our comparative analysis. RT-DETR represents a major breakthrough, establishing a new standard for real-time, end-to-end object detection and surpassing traditional YOLO-based methods in UAV contexts. And relatively speaking, its parameter count is closest to our method, so it is a formidable and highly relevant state-of-the-art benchmark.

Qualitative Results. Fig. 3 offers a visual comparison between the heatmaps and detection results generated by our LH-DETR and RT-DETR-R18, illustrating LH-DETR's superior performance in aerial image detection. The results reveal that our approach more effectively concentrates on small and densely arranged objects.

For instance, in the dense traffic scene presented in the first row, RT-DETR-R18's attention is diffuse and scattered, as shown by its broad heatmap. In contrast, the heatmap from LH-DETR shows highly concentrated activations focused on individual vehicles. This enhanced focus translates directly to the final output, where our model successfully detects vehicles in the cluttered center of the road that are either missed or poorly localized by RT-DETR-R18. The second row, depicting a parking lot, further highlights this advantage. RT-DETR-R18's heatmap incorrectly activates on

54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm

Margin requirements for the other pages
Paper size this page US Letter
54 pt
0.75 in
19.1 mm

non-vehicle areas, whereas our model's attention is cleanly focused on the cars, resulting in more precise bounding boxes. The third row demonstrates the challenge of detecting small, dense pedestrians in a crowded plaza. Although the baseline struggles to distinguish individual people, our LH-DETR demonstrates a remarkable ability to pinpoint each person with a focused activation. This fine-grained localization is a direct result of our model's enhanced ability to deal with high-frequency details, which is critical for small object detection. These examples highlight our framework in handling attribute-sensitive and spatially challenging queries that remain difficult for existing detectors.

TABLE III
COMPARISON OF AP(%) WITH OTHER DETECTORS ON UAVVASTE

Model	Params	FLOPS	FPS	AP(%)	AP_{50}(%)
YOLOv11-M[15]	20.0M	67.7G	52.0	30.6	64.0
RT-DETR-R18[5]	20.0M	60.0G	55.6	47.7	78.6
RT-DETR-R50[5]	42.0M	136G	53.5	48.6	79.4
HIC-YOLOv5[29]	**9.4M**	**31.2G**	**61.0**	30.5	65.1
LH-DETR(ours)	14.3M	44.9G	57.0	**49.0**	**79.7**

B. Results on UAVVaste Dataset

To verify the generalizability of our model, we conduct experiments on two additional specialized UAV datasets. As shown in Table III, on the UAVVaste dataset, our model improves the AP by 1.3 percentage points compared to the baseline RT-DETR-R18 (49.0% vs. 47.7%) while maintaining its efficiency advantages. It also outperforms the larger RT-DETR-R50 by 0.4 percentage points in AP. Furthermore, when compared to the similarly lightweight HIC-YOLOv5, LH-DETR achieves a massive 18.5 percentage point performance leap in AP (49.0% vs. 30.5%). This indicates our model's strong capability in identifying objects with diverse appearances in specialized contexts.

TABLE IV
COMPARISON OF AP(%) AND PARAMS/FPS ON UAV-PDD BY USING OUR METHODS WITH BASELINE.

Model	Params	FLOPS	FPS	AP(%)	AP_{50}(%)
RT-DETR-R18[5]	20.0M	60.0G	55.6	31.8	63.4
RT-DETR-R50[5]	42.0M	136G	53.5	32.2	64.0
LH-DETR(ours)	14.3M	44.9G	**57.0**	**33.0**	**64.3**

C. Results on UAV-PDD Dataset

The results on the UAV-PDD dataset, presented in Table IV, further underscore the real-time performance of our model. Compared to the baseline RT-DETR-R18, LH-DETR achieves a 1.2 percentage point increase in AP (33.0% vs. 31.8%) and a 2.5% increase in inference speed (57.0 vs. 55.6 FPS), all while reducing the parameter count by 28.5%. Even when compared against the heavyweight RT-DETR-R50, our model is higher in AP by 0.8 percentage points and is 2.5% faster. This is particularly significant as pavement distress detection requires identifying small, low-contrast details, proving that our frequency-aware enhancements and efficient backbone are highly effective for fine-grained analysis in real-time applications.

TABLE V
ABLATION STUDY: CUMULATIVE IMPACT OF LH-DETR COMPONENTS.

WMHB	FAD FFN	ASVLoss	Params	GFLOPs	AP(%)	AP_{50}(%)
✓			13.7M	43.4G	21.1	37.3
	✓		20.6M	61.5G	20.9	36.8
		✓	20.0M	60.0G	20.8	36.9
✓		✓	13.7M	43.4G	21.6	38.1
✓	✓		14.3M	44.9G	22.1	38.9
✓	✓	✓	14.3M	44.9G	**22.4**	**39.6**

D. Ablation Study on LH-DETR

To comprehensively evaluate the efficacy and efficiency of each proposed component, we conduct ablation studies on the VisDrone dataset using RT-DETR-R18 as the baseline. The quantitative results are summarized in Table V.

- **Wavelet-Mamba Hybrid Block (WMHB):** As the primary driver of efficiency, replacing the backbone with WMHB yields a 0.3% AP improvement (20.8→21.1). Crucially, this design significantly lowers model complexity, reducing the parameter count by approximately 31.5% (20.0M→13.7M) and GFLOPs by 27.7% (60.0G→43.4G). This demonstrates that coupling wavelet-based decomposition with Mamba's global modeling generates a feature representation that is both more robust and computationally efficient for aerial scenes compared to standard CNN backbones.
- **Frequency-Aware Dynamic FFN (FAD-FFN):** Incorporating FAD-FFN results in a consistent performance improvement, reaching 20.9 AP and 36.8 AP_{50}. Although this introduces a marginal increase in computational cost (+0.6M Params), the resulting performance gains justify the trade-off. This confirms that explicitly amplifying high-frequency cues effectively preserves the details of small, textured targets with minimal overhead.
- **AutoSliding Varifocal Loss (ASVLoss):** While the overall AP remains comparable to the baseline (20.8), ASVLoss notably boosts AP_{50} to 36.9 (+0.6% over baseline). Significantly, as an optimization of the training objective, ASVLoss improves localization quality without incurring any additional inference latency or parameter overhead.

Synergistic Effects: Combining these modules leads to cumulative performance gains. The integration of both WMHB and FAD-FFN boosts AP to 22.1%, highlighting the strong synergy between the efficient feature extractor and the frequency-aware enhancement module. Ultimately, the full LH-DETR model achieves peak performance (22.4% AP) while maintaining a highly lightweight profile (14.3M Params vs. Baseline 20.0M). This successfully validates our design goal of creating an efficient detector tailored for resource-constrained UAVs (Fig. 4).

54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm

Margin requirements for the other pages
Paper size this page US Letter
54 pt
0.75 in
19.1 mm

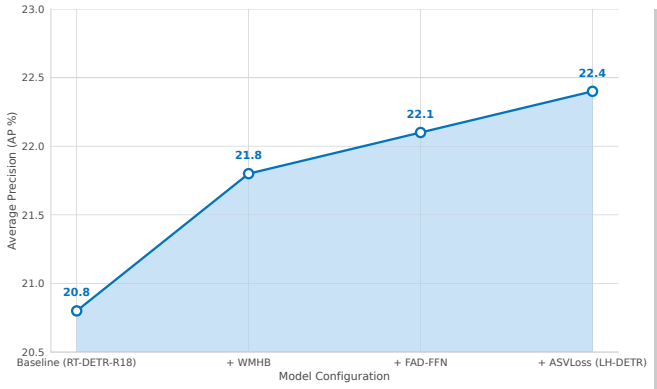

Fig. 4. Ablation Study: Cumulative Impact of LH-DETR Components.

V. CONCLUSIONS

In this paper, we propose a new real-time DETR detector for aerial image detection, named LH-DETR. Specifically, we devise a feature extractor based on wavelet transform and Mamba, which aims to deeply fuse local and global information and reduce computational complexity. Meanwhile, a Fourier transform-based FFN is designed to enhance multi-scale object perception and improve the network's ability to process features at different scales. For aerial image detection, our method mitigates the redundancies found in prior DETR-based detectors, which results in a significant improvement in the network's processing velocity.LH-DETR advances the technological methodologies of existing real-time detection models, offering a new perspective for the widespread industrial application of DETR-based architectures.

While our experiments are conducted on high-performance GPUs to validate algorithmic effectiveness, we acknowledge that deployment on embedded UAV platforms (e.g., NVIDIA Jetson) requires further optimization. Future work will focus on optimizing non-standard operators (like DWT) for TensorRT to bridge the gap between theoretical efficiency and edge deployment.

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)*. IEEE, 2024, pp. 455–462.
[2] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.
[3] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk, "Toward transformer-based object detection," *arXiv preprint arXiv:2012.09958*, 2020.
[4] 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.
[5] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 16 965–16 974.
[6] J. D. Hamilton, "State-space models," *Handbook of econometrics*, vol. 4, pp. 3039–3080, 1994.
[7] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.
[8] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.
[9] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.
[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
[11] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," *Advances in neural information processing systems*, vol. 37, pp. 103 031–103 063, 2024.
[12] X. Pei, T. Huang, and C. Xu, "Efficientvmamba: Atrous selective scan for light weight visual mamba," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 6, 2025, pp. 6443–6451.
[13] G. Wang, Y. Chen, P. An, H. Hong, J. Hu, and T. Huang, "Uav-yolov8: A small-object-detection model based on improved yolov8 for uav aerial photography scenarios," *Sensors*, vol. 23, no. 16, p. 7190, 2023.
[14] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, *et al.*, "Yolov10: Real-time end-to-end object detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 107 984–108 011, 2024.
[15] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.
[16] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025.
[17] M. Lei, S. Li, Y. Wu, H. Hu, Y. Zhou, X. Zheng, G. Ding, S. Du, Z. Wu, and Y. Gao, "Yolov13: Real-time object detection with hypergraph-enhanced adaptive visual perception," *arXiv preprint arXiv:2506.17733*, 2025.
[18] Y. Xiao, T. Xu, Y. Xin, and J. Li, "Fbrt-yolo: Faster and better for real-time aerial image detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 8, 2025, pp. 8673–8681.
[19] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.
[20] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
[21] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, J. Fang, C. Wong, Z. Yifu, D. Montes, *et al.*, "ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations," *Zenodo*, 2022.
[22] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9197–9206.
[23] H. Zhang, Y. Wang, F. Dayoub, and N. Sunderhauf, "Varifocalnet: An iou-aware dense object detector," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8514–8523.
[24] G. C. Jocher, "A. & qiu, j.(2023). ultralytics yolo."
[25] 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," *arXiv preprint arXiv:2410.13842*, 2024.
[26] C. Yang, Z. Huang, and N. Wang, "Querydet: Cascaded sparse query for accelerating high-resolution small object detection," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 13 668–13 677.
[27] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8311–8320.
[28] C. Xu, J. Ding, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "Dynamic coarse-to-fine learning for oriented tiny object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7318–7328.
[29] S. Tang, S. Zhang, and Y. Fang, "Hic-yolov5: Improved yolov5 for small object detection," in *2024 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2024, pp. 6614–6619.

54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm
54 pt
0.75 in
19.1 mm