Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

FHS-DEIM: Frequency-Aware Hybrid Encoding with Spiral Mamba for Infrared UAV Search and Rescue

Liqin Tian

tianliqin@ncist.edu.cn

Qinghai Normal University

Junyi Wu

Qinghai Normal University

Weizheng Wang

Adelaide University

Jiang Yi

University of Emergency Management

Yong Niu

Qinghai Normal University

Alsinglawi Belal S.

Zayed University

Research Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11500998/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

FHS-DEIM: Frequency-Aware Hybrid Encoding with Spiral Mamba for Infrared UAV Search and Rescue

Abstract

Infrared remote sensing from uncrewed aerial vehicles (UAVs) requires accurate person detection under low contrast, and vegetation occlusion to support visual inspection in search and rescue (SAR), is difficult because targets occupy few pixels, thermal contrast varies with the scene, and clutter edges can dominate the response. We present Frequency-Aware Hybrid Encoding with Spiral Mamba (FHS-DEIM), an end-to-end detection transformer built around a frequency-aware Mamba encoder and size-adaptive supervision. Frequency Interaction Reinforcement with Mamba (FIRE-Mamba) adapts Mamba to two complementary Haar representations: an approximation band and aggregated directional detail. Shared frequency attention and four Spiral Mamba branches model their context, and inverse Haar synthesis reconstructs a modulation at the input resolution. Haar Wavelet Downsampling (HWDS) processes four decimated subbands and restores three detail outputs for bottom-up pyramid reduction. Size-Adaptive Normalized Wasserstein Distance-Guided Matching-Aligned Loss (SA-NWD-MAL) interpolates between intersection-over-union and Wasserstein similarity according to matched-target size. On the person-focused Partially Occluded Person (POP) dataset and the five-category high-altitude infrared thermal UAV dataset (HIT-UAV), FHS-DEIM-M achieved average precision (AP) of 62.9% and 58.9%, respectively; FHS-DEIM-X reached 64.2% and 60.1%. In the M-scale component study, the complete configuration improved the baseline from 60.6% to 62.9% AP.

Keywords: Infrared person detection, UAV search and rescue, Mamba, discrete wavelet transform, normalized Wasserstein distance.

1 Introduction

Uncrewed aerial vehicles (UAVs) support search and rescue (SAR) by surveying terrain and supplying remote sensing imagery for person detection. As a UAV follows its search route, changes in viewing distance, perspective, and background produce a continuing stream of observations for inspection (Kubo et al. 2026; Woesle et al. 2026). The perception task is to turn this imagery into spatially identifiable candidates that can guide attention to possible persons (Liu et al. 2026a). In visual aerial SAR, Gotovac et al. (2020) showed how machine-generated candidate regions and AI-driven visual perception can assist human observers in reviewing search images. Thermal sensing supplies complementary evidence when visible appearance is weak: infrared thermography has supported personnel detection in robotic mine-rescue experiments (Szrek et al. 2021), while UAV infrared benchmarks document the effects of altitude and vegetation occlusion on person detection (Song et al. 2025; Suo et al. 2023). These applications connect aerial image acquisition to a concrete detection requirement: retaining enough evidence to detect a person and estimate a bounding box under changing observation conditions.

For UAV infrared person detection, that evidence is often spatially limited and difficult to separate from the surroundings. A distant person occupies few pixels, vegetation can obscure parts of the body, and local thermal contrast varies across the scene. Figure 1 illustrates the associated ambiguity: compact gray-level peaks coexist with extended background variation, so brightness alone does not determine an object instance or its extent. Research on pixel-level infrared small targets addresses weak responses through contextual modulation, shape information, and structured background separation (Dai et al. 2021; Wu et al. 2024; Zhang et al. 2022). Infrared box detection also explores multiscale enhancement and geometry-aware losses for dim maritime targets (Liu et al. 2026b). Person detection requires the remaining thermal fragments to support a bounding box. Lightweight infrared extraction and edge-aware multiscale fusion strengthen the feature evidence available to box detectors (Guan et al. 2026; Hu et al. 2026), but selecting useful detail requires considering the surrounding structure: an enhanced edge can belong to vegetation or a built object as well as to a person. This connects the observation problem to representation design. Fine spatial evidence must survive feature reduction, and contextual information must help distinguish a partially visible person from competing background responses. The resulting requirement concerns both what information the encoder preserves and how it relates that information across locations and scales.

Hybrid encoders in detection transformers (DETRs) provide a framework for meeting these requirements while controlling the cost of multiscale interaction. The real-time detection transformer (RT-DETR) separates attention-based interaction within a feature scale from convolutional fusion between scales (Zhao et al. 2024). Lite DETR further studies how interleaved updates of high- and low-resolution features and key-aware deformable attention distribute encoder computation (Li et al. 2023). Remote-sensing methods extend this design space through foreground selection, adaptive sampling, and hierarchical relations, including sparse attention and multiscale enhancement in DFE-DETR and multiscale hypergraph fusion in RSRT-DETR (Wu et al. 2025; Zhao et al. 2026). Related one-stage aerial detectors address the same scale-transfer problem: LMSFA-YOLO combines lightweight multiscale extraction with a high-resolution shallow feature layer, and ARF-YOLO combines dual-resolution attention with content-adaptive cross-level fusion (Chu et al. 2025; Zhang 2026a). Frequency-aware approaches add another means of organizing the information being fused: SAFF-DETR connects shallow frequency representations with cross-layer interaction, and PWL-RTDETR introduces multilevel wavelet convolution into a thermal detection neck (Guo et al. 2025; Zhi et al. 2025). In photovoltaic electroluminescence (EL) defect detection, YOLO-PVELAD similarly combines a DCT high-frequency pyramid with directional attention and a geometry-aligned dual-branch head to represent faint, anisotropic defects (Li et al. 2026). These studies motivate treating within-scale context and cross-scale transfer as coupled parts of the encoder. A context operator must return features aligned with the current grid, whereas a reduction operator must carry compact evidence to a coarser grid. For infrared SAR, their coordination matters because a useful contextual response can only assist person detection if the target detail remains available through subsequent fusion. DETR with improved matching (DEIM) and DEIMv2 provide the matching-aligned training and scalable detection framework in which we investigate this coordination (Huang et al. 2025a,b).

Within this framework, Mamba offers input-conditioned contextual propagation through selective state-space models (SSMs), with complexity linear in sequence length (Gu and Dao 2024). Visual Mamba organizes image features into scan sequences, making both the input representation and the route consequential to the context accumulated at each location (Guo et al. 2024; Liu et al. 2024b). LIR-Mamba demonstrates global and local Mamba scans in infrared box detection under laser interference, while FIRE-Net combines Fourier magnitude and phase processing with spatial Mamba for infrared small-target detection (Chen et al. 2026b; Li et al. 2025). These approaches motivate adapting Mamba to the structure of infrared evidence. Haar analysis offers

local approximation and directional-detail representations (Mallat 1989); the design question is how to propagate their context through Mamba and return it to the spatial feature pyramid. This representation problem is coupled to supervision of classification confidence and box geometry. For a tiny box, a small center or extent error can sharply reduce intersection-over-union (IoU), and nearby nonoverlapping boxes receive zero overlap quality. Matching-aligned objectives connect classification confidence to localization quality (Cai et al. 2024; Huang et al. 2025b), while normalized Wasserstein distance (NWD) provides a continuous measure of box displacement (Wang et al. 2021; Xu et al. 2022). The combined research question is therefore how to coordinate frequency-aware Mamba context, detail-preserving scale reduction, and size-dependent supervision so that compact infrared evidence can support both recognition and precise box localization.

We propose Frequency-Aware Hybrid Encoding with Spiral Mamba (FHS-DEIM) to address this question within the DEIMv2 encoder and training objective. Its central component, Frequency Interaction Reinforcement with Mamba (FIRE-Mamba), applies shared frequency attention (FA) and four Spiral Mamba routes to a Haar approximation band and aggregated directional detail, then reconstructs contextual modulation at the input resolution. FHS denotes frequency-aware, hybrid, and Spiral; hybrid encoding combines local feature processing, Mamba context, and multiscale fusion. Haar Wavelet Downsampling (HWDS) processes all four Haar subbands and restores three directional outputs during bottom-up reduction, connecting frequency enhancement to scale transfer. Size-Adaptive Normalized Wasserstein Distance–Guided Matching-Aligned Loss (SA-NWD-MAL) adjusts the quality target and regression penalty for already matched predictions according to target size. The design preserves the three-scale decoder interface and allows the roles of contextual propagation, reduction, and supervision to be examined separately. Evaluation on the Partially Occluded Person (POP) dataset and the high-altitude infrared thermal UAV dataset (HIT-UAV) relates the method to vegetation-occluded persons and broader aerial thermal scenes containing persons and other categories (Song et al. 2025; Suo et al. 2023). Separate training and testing on each benchmark examine the method's generalization across these detection settings.

The contributions are threefold:

- We develop FIRE-Mamba by coupling two-band Haar representation, shared FA, four-route Spiral Mamba propagation, and inverse reconstruction. This adapts Mamba's input representation and spatial routing to same-resolution infrared context enhancement.
- We integrate FIRE-Mamba with HWDS, which enhances four Haar subbands and restores three detail groups from a shared high-frequency bottleneck before pyramid reduction. Their distinct placements connect Mamba context modeling with cross-scale detail transfer in the hybrid encoder.
- We introduce SA-NWD-MAL to interpolate between IoU and NWD according to matched-target size and add a size-weighted localization term. It provides continuous geometric supervision for nearby tiny boxes during training.

We evaluate accuracy and computational cost on two UAV infrared benchmarks and measure throughput on two hardware platforms. Component and operator ablations, precision–recall (PR) curves, detections, and aligned response maps examine how the proposed components affect the resulting tradeoffs.

2 Related Work

Table 1 summarizes representative approaches by task, mechanism, and relation to FHS-DEIM. The following subsections examine person detection in UAV infrared SAR, efficient hybrid encoding and multiscale fusion, and the roles of Mamba, frequency representation, and supervision.

Table 1: Representative related work and the design questions addressed by FHS-DEIM, the comparison concerns mechanisms and task settings.

Method	Task	Main mechanism	Relation to FHS-DEIM
Lite DETR (Li et al. 2023)	General box detection	Interleaved multiscale updates and key-aware deformable attention	Efficient distribution of encoder computation across scales.
SAFF-DETR (Zhi et al. 2025)	Remote-sensing box detection	Shallow frequency fusion and cross-layer spatial/channel frequency attention	Prior frequency-aware detection; emphasizes joint multiscale interaction.
DFE-DETR (Wu et al. 2025)	Remote-sensing small objects	Sparse attention, multiscale enhancement, and deformable large kernels	Spatial selection and adaptive feature extraction.
PWL-RTDETR (Guo et al. 2025)	Thermal box detection	Reparameterization and multilevel wavelet convolution in the neck	Direct infrared wavelet precedent; convolutional frequency processing.
MMAF-YOLO (Yang et al. 2026)	UAV box detection	Multiscale aggregation, modulation, and high-resolution prediction	Preserves small-object detail through fusion and prediction resolution.
LIR-Mamba (Chen et al. 2026b)	Infrared box detection	YOLO with selective SSMs and global/local scans	Infrared Mamba detection with global/local routing and a different detection architecture.
CSM-DETR (Zhang 2026b)	UAV construction monitoring	MobileMamba and multiscale fusion	Mamba context in an application-specific DETR.
FIRE-Net (Li et al. 2025)	Infrared small-target detection	Fourier magnitude/phase and spatial Mamba interaction	Prior frequency–spatial Mamba design using a different transform.
Align-DETR (Cai et al. 2024)	General box detection	Aligned soft targets and ranked intermediate-layer supervision	Connects classification and localization quality.
FHS-DEIM	UAV infrared person detection	Two-band Spiral Mamba, four-subband reduction, and SA-NWD-MAL	Coordinates Mamba context, scale transfer, and size-dependent supervision.

2.1 Person Detection in UAV Infrared Search and Rescue

The perception problem in aerial SAR is to detect a person within a changing search footprint and provide a candidate whose position can be inspected. As the UAV advances along a route, the same search objective encounters different projected target sizes, viewpoints, and backgrounds. Gotovac et al. (2020) studied the connection between candidate detection and human inspection in visible-light UAV imagery. Thermal sensing has also supported ground-robot mine rescue through combined infrared and visible detections (Szrek et al. 2021). These applications establish the relevance of person detection under difficult observation conditions. In UAV infrared imagery, viewing altitude reduces the spatial evidence available for an individual person, while vegetation occlusion can leave only portions of the thermal body shape visible (Song et al. 2025; Suo et al. 2023). A useful representation must therefore support recognition from limited evidence and preserve a spatially meaningful estimate of object extent. Temporal evidence supports a further SAR cue: Zhang et al. (2021) distinguish breathing from background motion through frequency variability in drone and camera videos. This approach extracts periodic life signs across frames; the present task instead estimates person instances and box extents from each infrared image.

Figure 1 illustrates the visual evidence available to the detector. The source scene contains vegetation, parked vehicles, and built structures. The three selected regions of interest (ROIs) show compact elevated gray-level regions against nonuniform surroundings; ROI 3 also contains an extended elevated band next to the compact peak. The local blocks show where contrast is concentrated, while the surfaces make its spatial distribution explicit. For a detector, the relevant distinction is whether a compact

response belongs to a person and how it relates to adjacent structure. This motivates preserving local evidence while incorporating a wider spatial context to distinguish the candidate from surrounding clutter.

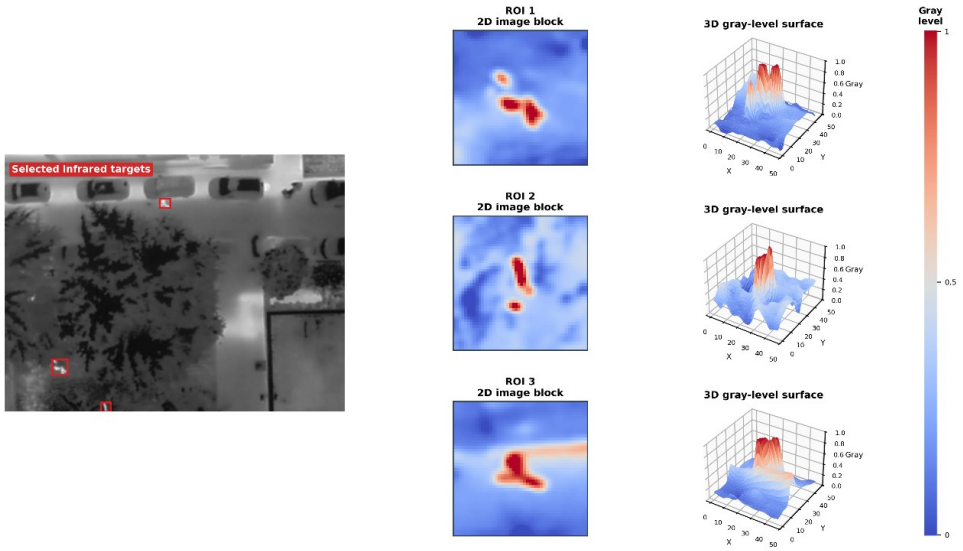


Fig. 1: Local contrast in an infrared search scene. The source image marks three selected regions; the corresponding two-dimensional blocks and three-dimensional surfaces show their gray-level structure. Compact elevated regions coexist with spatially extended background variation. Colors and surface heights use the displayed 0–1 gray-level scale.

Partial occlusion makes this distinction more demanding because the visible body shape may be incomplete (Song et al. 2025). Pixel-level infrared small-target methods provide useful ways to organize weak foreground evidence: ACMNet uses asymmetric contextual modulation, ISNet emphasizes target shape, and RPCANet unfolds low-rank and sparse separation (Dai et al. 2021; Wu et al. 2024; Zhang et al. 2022). Their common relevance is the use of surrounding or structural information to strengthen a target response that is difficult to identify locally. Bounding-box person detection must additionally associate the available evidence with an object instance and estimate its extent. A small visible fragment can contribute to that estimate even when it does not describe the complete body contour. This makes the relation between local detail and spatial context central to the detection problem. The POP and HIT-UAV benchmarks provide complementary settings for examining it through vegetation-occluded persons and high-altitude thermal observations (Song et al. 2025; Suo et al. 2023).

Infrared box detectors address the preservation of this evidence at several architectural points. LIRS-YOLO develops lightweight feature extraction for small infrared objects, and IR-DETR combines edge-aware processing with multiscale fusion (Guan et al. 2026; Hu et al. 2026). These designs emphasize the quality of the features supplied to prediction. PWL-RTDETR uses wavelet convolution, partial convolution, and reparameterization within thermal detection (Guo et al. 2025), while MMAF-YOLO combines multiscale aggregation, feature modulation, and a high-resolution prediction branch in experiments that include thermal imagery (Yang et al. 2026). They connect small-object evidence to the representation and resolution maintained by the detector. Complementary modalities provide another source of discrimination: Wang et al. (2025) fuse infrared radiation and visible texture for small-object detection in urban traffic scenes. FHS-DEIM performs person detection from a single infrared image, making context and detail within that modality the resources available to its encoder.

The usefulness of feature enhancement also depends on background discrimination. COCO-FP shows in general object detection that substantial background false positives can remain despite improvements on conventional average precision (AP) benchmarks (Liu et al. 2024a). This finding motivates examining the spatial selectivity of a detector alongside its final box accuracy. In the infrared scene of Fig. 1, both compact regions and extended surrounding structure contribute image contrast. Enhancing all high-frequency structure can strengthen clutter, while broad aggregation can mix the evidence of nearby or partially visible targets. The scientific requirement is consequently selective preservation: compact evidence should remain available through the encoder, with contextual relationships supporting its interpretation. This requirement leads directly to the question of how within-scale processing and multiscale fusion are organized.

2.2 Efficient Hybrid Encoders and Multiscale Fusion

A hybrid encoder determines where contextual interaction occurs and how its output is exchanged between feature levels. RT-DETR separates attention-based within-scale interaction from convolutional cross-scale fusion (Zhao et al. 2024). This organization distinguishes the cost of contextual modeling from the cost of constructing a multiscale pyramid. Lite DETR studies a different allocation of encoder computation through interleaved high- and low-resolution feature updates and key-aware deformable attention (Li et al. 2023). Both make the treatment of resolution an explicit design choice: fine grids retain detailed spatial evidence, while the encoder must make that evidence available to features carrying broader context. DEIM and DEIMv2 provide the matching-aligned training and scalable detector framework used here (Huang et al. 2025a,b); our encoder question concerns how infrared information is processed and transferred within that framework.

Efficient feature use also depends on how representations reach the queries and on the deployment budget. Cross-feature interaction is explored in real-time end-to-end detection by Ge et al. (2025), while multi-time inquiries DETR (MI-DETR) uses parallel inquiry heads to obtain complementary information for decoder queries (Nan et al. 2025). For gas-pipeline defects, lite distance-content-aware DETR (L-DC-DETR) combines alternating multiscale updates and selective high-resolution processing with distance-content-aware decoding (Chen et al. 2026a). NeXt-DETR connects scalable architecture to efficient inference on resource-constrained systems (Choi and Lee 2026). These methods place efficiency choices across the encoder, decoder, and deployment interface. At the system level, the review by Khadam et al. (2025) identifies both edge and cloud deployment as common settings for artificial intelligence in Internet of Things systems. This broader deployment context motivates considering where inference executes alongside the accuracy of the detector, while the present encoder design focuses on the information exchanged before query decoding.

Remote-sensing encoder research provides several ways to improve this interaction for small and densely distributed objects. Position-DETR introduces position guidance, and SED-DETR develops scale-enhanced deformable attention (Yin et al. 2025; Zhao et al. 2025). These methods bring location and scale into the selection of informative features. DFE-DETR combines sparse attention, multiscale strip-convolution enhancement, and deformable large-kernel sampling (Wu et al. 2025), linking selective contextual interaction with adaptable spatial feature extraction. RSRT-DETR uses polar attention, multiscale hypergraph relations, and hierarchical fusion to model dense small-object scenes (Zhao et al. 2026). Across these approaches, the encoder is improved through changes to foreground selection, sampling geometry, or relationships among feature locations. For UAV infrared person detection, these choices motivate examining whether a context response remains connected to the compact evidence needed at the subsequent fusion stage.

Frequency-aware fusion introduces an additional choice about the representation being exchanged. SAFF-DETR combines shallow frequency representations with cross-layer spatial and channel frequency attention, enabling global multiscale interaction without repeated upsampling and downsampling (Zhi et al. 2025). PWL-RTDETR uses multilevel wavelet convolution and inverse reconstruction in a thermal detection neck (Guo et al. 2025). The former emphasizes frequency-aware interaction across feature levels, while the latter makes wavelet processing part of the convolutional representation. These precedents show that frequency information can participate in detection at different architectural locations and through different operators. Their relevance to FHS-DEIM is the joint consideration of frequency representation, contextual processing, and the spatial resolution at which information is retained or reconstructed.

This comparison identifies two connected roles for the proposed hybrid encoder. Same-resolution context enhancement must return its output to the feature grid used by lateral fusion, whereas bottom-up reduction must transfer approximation and detail to the next coarser grid. FHS-DEIM assigns frequency-aware Mamba propagation to the first role and wavelet subband enhancement to the second. FIRE-Mamba uses a two-band representation for contextual interaction and reconstructs the response at the input resolution. HWDS begins with four subbands and retains their decimated grid during reduction. Coordinating these placements allows the paper to examine Mamba context and cross-scale detail transfer as related components of the same infrared representation problem.

2.3 Mamba, Frequency Representation, and Supervision

Mamba uses selective SSMs to implement input-conditioned propagation with complexity linear in sequence length (Gu and Dao 2024). VMamba brings Mamba to visual features through directional scanning, and MambaIR applies it to image restoration (Guo et al. 2024; Liu et al. 2024b). In these adaptations, a two-dimensional feature map becomes a sequence along a prescribed route. The route determines the order in which locations contribute to the evolving state, and inverse routing places the resulting responses back on the image grid. Thus, even when two scans cover the same positions, they can organize contextual accumulation differently. For an infrared detector, the scan must be considered together with the representation supplied to Mamba and the spatial alignment required by subsequent pyramid fusion.

Application-specific detectors demonstrate how Mamba can supply this context within different architectures. LIR-Mamba combines YOLO and selective SSMs with global and local scans for infrared small-target box detection under laser interference (Chen et al. 2026b). CSM-DETR incorporates MobileMamba and multiscale feature fusion for construction-site monitoring from UAV imagery (Zhang 2026b). The former directly studies infrared detection, while the latter connects Mamba features to a transformer detector in a different aerial application. Together, they motivate examining the route and representation through which Mamba receives weak object evidence. FIRE-Mamba concentrates on this interface: frequency bands provide the sequence content, four Spiral routes propagate context, and inverse routing aligns their responses before the feature returns to the detection pyramid.

Frequency–spatial interaction supplies the basis for the representation choice. FIRE-Net processes Fourier magnitude and phase alongside a spatial Mamba branch for infrared small-target detection (Li et al. 2025). Haar analysis instead organizes each local neighborhood into one approximation component and three directional-detail components (Mallat 1989). Wavelet convolution further demonstrates how multiresolution processing can enlarge receptive fields (Finder et al. 2024). These approaches connect frequency content to contextual processing through different transforms. In FIRE-Mamba, the approximation band and mean directional detail form two inputs to shared FA and Mamba propagation. The four Spiral outputs are restored to a common grid, and inverse Haar synthesis turns the processed bands into spatial residual

modulation. HWDS uses the same four-subband starting representation for a different operation: it enhances approximation and detail while retaining the coarser grid needed by the bottom-up path. The contribution is the resulting Mamba adaptation and its coordination with scale transfer in the hybrid encoder.

The final part of this coordination concerns supervision of classification confidence and box geometry. Align-DETR connects classification targets to joint classification–localization quality and uses ranked weighting for intermediate-layer many-to-one supervision (Cai et al. 2024). DEIM develops matching-aligned loss (MAL) within dense-supervision training (Huang et al. 2025b). Quality-aligned supervision makes localization relevant to the confidence learned by a detector, which is consequential for tiny targets: a small geometric displacement can sharply reduce IoU or leave it at zero. NWD supplies a continuous similarity based on the displacement of box centers and extents (Wang et al. 2021; Xu et al. 2022). SA-NWD-MAL uses target size to interpolate between IoU and NWD for matched-positive quality targets and to weight an additional localization term. Its role is to make the supervision of compact objects sensitive to geometric proximity after query assignment. It complements the representation changes by addressing how the detector learns confidence and box geometry from the evidence retained by the encoder.

The local-contrast ambiguity in Fig. 1 connects the three lines of work: the encoder must preserve spatial evidence, Mamba must organize context that helps interpret it, and supervision must guide both the classification confidence and box geometry of the resulting candidate. FHS-DEIM brings these operations together through frequency-aware Mamba context, wavelet reduction, and size-dependent matched supervision.

3 Method

To address the loss of compact person evidence during scale reduction and its ambiguity with thermal clutter, FHS-DEIM combines frequency-aware Mamba context, subband-preserving reduction, and size-adaptive supervision. The architecture in Fig. 2 maps an infrared image to person detections through backbone extraction, hybrid encoding, and query-based prediction. The backbone supplies features at three resolutions: the finer grid retains spatial evidence for compact persons, while the deeper representations provide scene context. Channel projections align the three features to the common width used by the encoder. FIRE-Mamba processes the two deeper scales before pyramid fusion so that their contextual responses can enter the subsequent cross-scale exchange. The fine-scale backbone feature enters this exchange through the top-down path, where it is combined with information propagated from the deeper levels.

Within each FIRE-Mamba block, Haar analysis separates approximation and directional detail. Shared FA and four Spiral Mamba routes process the resulting frequency representation, and inverse synthesis returns a contextual modulation to the original feature grid. The aligned coarse features enter the top-down merges: information first moves from the coarsest level to the intermediate level and then to the finest level. The bottom-up path subsequently uses HWDS for two successive reductions. At each transition, HWDS enhances the Haar subbands and combines them on the decimated grid before fusion with the corresponding lateral feature. These directions connect deeper context with fine-grid detail, producing three encoder outputs for detection.

These outputs supply the DEIMv2 query decoder. IoU-aware query selection initializes object queries and reference boxes from the encoded representation, after which successive decoder layers refine class predictions and normalized bounding boxes. The training objective connects these predictions to target categories and geometry through the inherited detection and distribution-refinement losses. SA-NWD-MAL adjusts the quality and geometric supervision of already matched predictions according to target size, complementing the feature changes made by FIRE-Mamba and HWDS. Thus, the architecture links frequency-aware Mamba context and bidirectional

scale transfer to query-based detection, with the size-adaptive objective acting during training.

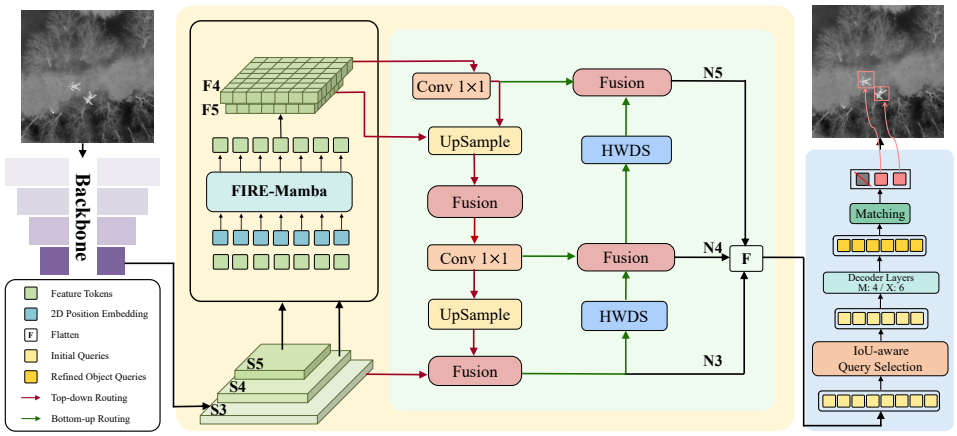


Fig. 2: FHS-DEIM architecture. FIRE-Mamba enriches two deeper feature scales, HWDS performs two bottom-up reductions, and the fused feature pyramid enters the query decoder. SA-NWD-MAL supervises matched predictions during training.

3.1 Hybrid Encoder

Fine spatial evidence and deeper contextual features must interact within the detection pyramid to support both weak-target recognition and box estimation. The hybrid encoder in Fig. 2 organizes within-scale processing and cross-scale exchange around this requirement. Its design divides this task between frequency-aware Mamba processing within a scale and wavelet downsampling at the transitions between scales. The detector's three-scale feature pyramid network (FPN) and path aggregation network (PAN) provide the paths through which the two operations interact.

To describe this flow, let $\mathbf{X}_l \in \mathbb{R}^{B\times C_e\times H_l\times W_l}$ denote the projected feature at level $l \in \{3,4,5\}$, ordered from the finest to the coarsest grid. The dimensions specify batch size B, shared channel width C_e, and spatial size H_l, W_l at level l. FHS-DEIM-M uses $C_e = 256$ and four decoder layers; FHS-DEIM-X uses $C_e = 384$ and six decoder layers. The common width makes the projected features compatible with the additions along the fusion paths.

The feature flow begins with FIRE-Mamba on levels 4 and 5. Within each block, the approximation band and mean directional band pass through shared FA and four Spiral Mamba routes. Reconstruction converts their contextual responses into modulation of the original-resolution feature, preserving its shape for lateral fusion. The top-down FPN then processes the pyramid from coarse to fine. At each transition, a lateral projection of the current coarse feature is upsampled, added to the next finer feature, and processed by the fusion block. The finest projected feature consequently combines its spatial detail with context propagated through the two deeper levels.

The bottom-up PAN starts from this fused fine-scale representation. Each of its two transitions applies HWDS to the current feature, using Haar analysis, approximation/detail enhancement, and channel reduction before addition to the saved lateral feature at the next coarser level. HWDS retains the decimated grid required for this transfer, while FIRE-Mamba has already returned its context response to the grid on which it entered the top-down path. The distinction determines the placement of the two modules: Mamba enriches within-scale context before fusion, and HWDS carries enhanced approximation and detail through the subsequent reductions. The resulting three encoder features retain width C_e and the pyramid resolutions expected by the decoder.

Both encoder components start from the same Haar subband analysis, which supports either same-resolution reconstruction or cross-scale reduction. For a generic input $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$, the fixed analysis kernels are

$$\begin{aligned} \mathbf{K}_{\mathrm{LL}} &= \tfrac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \qquad \mathbf{K}_{\mathrm{LH}} = \tfrac{1}{2}\begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}, \\ \mathbf{K}_{\mathrm{HL}} &= \tfrac{1}{2}\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, \quad \mathbf{K}_{\mathrm{HH}} = \tfrac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \end{aligned} \tag{1}$$

The Haar discrete wavelet transform (DWT) uses channelwise filtering and stride-two sampling to produce the approximation and three directional subbands:

$$\mathbf{D}_q = \left(\mathbf{K}_q *_{\mathrm{dw}} \mathbf{X}^{+}\right)\downarrow_2, \qquad q \in \{\mathrm{LL}, \mathrm{LH}, \mathrm{HL}, \mathrm{HH}\}, \tag{2}$$

where $*_{\mathrm{dw}}$ and $\downarrow_2$ denote depthwise filtering and decimation. For odd H or W, $\mathbf{X}^{+}$ replicates the final row or column before analysis. Each subband has shape $B \times C \times h \times w$, with $h = \lceil H/2 \rceil$ and $w = \lceil W/2 \rceil$. The matrices in Eq. (1) specify the directional signs and axes: local 2×2 information is redistributed across four bands on the coarser grid.

3.2 FIRE-Mamba

FIRE-Mamba couples FA, four-route Spiral Mamba propagation, and inverse Haar synthesis. Frequency separation supplies approximation and detail representations for context modeling, while the reconstructed response modulates spatial features at their original resolution. For an input $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$, channelwise layer normalization (LN) and the sigmoid linear unit (SiLU) form the local branch:

$$\begin{cases} \mathbf{F} = \mathrm{LN}(\mathbf{X}), \\ \mathbf{U} = \mathrm{SiLU}(\mathbf{F}). \end{cases} \tag{3}$$

Applying the analysis operator in Eq. (2) to the normalized feature $\mathbf{F}$ in Eq. (3) forms the two frequency inputs:

$$\begin{cases} (\mathbf{D}_{\mathrm{LL}}, \mathbf{D}_{\mathrm{LH}}, \mathbf{D}_{\mathrm{HL}}, \mathbf{D}_{\mathrm{HH}}) = \mathrm{DWT}_{\mathrm{Haar}}(\mathbf{F}), \\ \mathbf{D}_{\mathrm{HF}} = \tfrac{1}{3}(\mathbf{D}_{\mathrm{LH}} + \mathbf{D}_{\mathrm{HL}} + \mathbf{D}_{\mathrm{HH}}), \\ \mathbf{D} = \mathrm{Fold}_B(\mathbf{D}_{\mathrm{LL}}, \mathbf{D}_{\mathrm{HF}}). \end{cases} \tag{4}$$

The Haar kernels in Eq. (1) fix the directional convention. Each band retains shape $B \times C \times h \times w$ on the decimated grid defined by $h = \lceil H/2 \rceil$ and $w = \lceil W/2 \rceil$; odd dimensions are padded by replicating the final row or column. Folding in Eq. (4) first stacks the two bands as $B\times 2\times C\times h\times w$ and then merges the batch and band axes. The folded tensor $\mathbf{D} \in \mathbb{R}^{2B\times C\times h\times w}$ therefore contains each sample's LL and HF features in adjacent entries. Both bands use the same attention and Mamba parameters while retaining channel width C.

FA applies global max pooling (GMP) and global average pooling (GAP) in parallel to the same normalized bands. Channel concatenation and projection combine these complementary descriptors into context that is added to the input:

$$\begin{cases} \mathbf{P} = \mathcal{C}_{\mathrm{p}}\left(\begin{bmatrix} \mathrm{GMP}(\mathrm{LN}(\mathbf{D})) \\ \mathrm{GAP}(\mathrm{LN}(\mathbf{D})) \end{bmatrix}\right), \\ \mathbf{F}_{\mathrm{b1}} = \mathbf{D} + \mathbf{P}. \end{cases} \tag{5}$$

The bracketed GMP and GAP responses are concatenated along channels. Each pooled descriptor has shape $2B \times C \times 1 \times 1$. The projection $\mathcal{C}_{\mathrm{p}}$ combines $2C$ descriptor channels into a C-channel context, and $\mathbf{P}$ is broadcast over the $h \times w$ grid during addition. A second residual transform refines the pooled context while retaining the first residual feature:

$$\mathbf{F}_{\mathrm{b2}} = \mathcal{C}_2(\mathrm{SiLU}(\mathcal{C}_1(\mathrm{LN}(\mathbf{F}_{\mathrm{b1}})))) + \mathbf{F}_{\mathrm{b1}}. \tag{6}$$

The context feature from Eq. (6) is projected to a modulation feature $\mathbf{G}$, while a separate projection of the original frequency input produces the value feature $\mathbf{V}$:

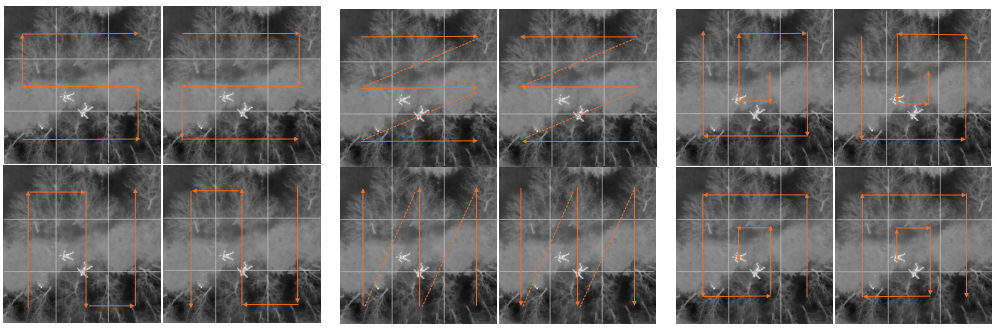
$$\begin{cases} \mathbf{G} = \mathcal{C}_{\mathrm{g}}(\mathbf{F}_{\mathrm{b2}}), \\ \mathbf{V} = \mathcal{C}_{\mathrm{v2}}(\mathrm{SiLU}(\mathcal{C}_{\mathrm{v1}}(\mathbf{D}))), \\ \mathrm{FA}(\mathbf{D}) = \mathbf{G} \odot \mathbf{V}. \end{cases} \tag{7}$$

All channel projections are learned, bias-free 1×1 convolutions; apart from $\mathcal{C}_{\mathrm{p}}$, they map C channels to C. LN uses $\epsilon = 10^{-6}$ and learned affine parameters. The projected modulation $\mathbf{G}$ in Eq. (7) is unconstrained, allowing signed rescaling of the value feature. The right-hand detail in Fig. 5 follows this FA computation from parallel pooling through residual refinement to $\mathbf{G} \odot \mathbf{V}$.

Mamba serializes the same $N = hw$ positions along four Spiral permutations:

$$\Pi_{\mathrm{sp}} = \{\pi_{\mathrm{cw,in}}, \pi_{\mathrm{ccw,in}}, \pi_{\mathrm{cw,out}}, \pi_{\mathrm{ccw,out}}\}. \tag{8}$$

In the route set in Eq. (8), the clockwise-inward route starts at the lower-left corner, and the counter-clockwise-inward route starts at the lower-right corner. Both traverse nested contours toward the center. The clockwise-outward route reverses the counter-clockwise-inward sequence, and the counter-clockwise-outward route reverses the clockwise-inward sequence. Each route visits every grid position once, with adjacent tokens remaining four-neighbor connected. Figure 3 compares these routes with corner-raster, cross, and continuous two-dimensional organizations.



(a) Four-corner raster (b) Cross and continuous routes (c) Four-route Spiral serialization

Fig. 3: Serialization designs. FHS-DEIM uses the four Spiral routes; corner-raster, cross, and continuous two-dimensional organizations provide the route comparisons.

Within each Mamba branch, a C-dimensional token is projected to inner width $2C$. For one inner channel, let $x_t, y_t \in \mathbb{R}$ denote the drive and output at position t, and let $\mathbf{h}_t \in \mathbb{R}^{d_{\mathrm{s}}}$ be its hidden state. The selective recurrence is

$$\begin{cases} \mathbf{h}_t = \overline{\mathbf{A}}_t \mathbf{h}_{t-1} + \overline{\mathbf{B}}_t x_t, \\ y_t = \mathbf{C}_t \mathbf{h}_t + D x_t, \end{cases} \tag{9}$$

where the discretized coefficients $\overline{\mathbf{A}}_t \in \mathbb{R}^{d_{\mathrm{s}} \times d_{\mathrm{s}}}$ and $\overline{\mathbf{B}}_t \in \mathbb{R}^{d_{\mathrm{s}}}$ govern state retention and input injection; $\mathbf{C}_t \in \mathbb{R}^{1 \times d_{\mathrm{s}}}$ projects the state to the output, and D is the skip coefficient. Input-dependent discretization and projections regulate propagation as in Mamba (Gu and Dao 2024). Each inner channel uses $d_{\mathrm{s}} = 16$ states. The branch uses sequence-convolution width 4 and expansion factor 2, then projects back to width C. Figure 4 illustrates Mamba recurrence and inverse routing.

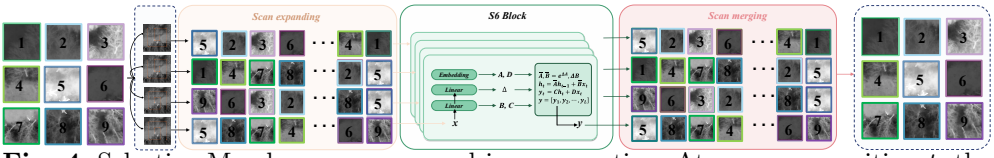


Fig. 4: Selective Mamba recurrence and inverse routing. At sequence position t, the hidden state $\mathbf{h}_t$ integrates the drive x_t through $\mathbf{h}_t = \overline{\mathbf{A}}_t \mathbf{h}_{t-1} + \overline{\mathbf{B}}_t x_t$, and the readout is $y_t = \mathbf{C}_t \mathbf{h}_t + D x_t$, as in Eq. (9). The inverse permutation restores each route to the image grid.

Channel normalization and local depthwise convolution (DWConv) prepare the FA response for sequential propagation. Each route has an independently parameterized Mamba branch. The inverse permutation for each route restores image coordinates before the four responses are averaged:

$$\begin{cases} \mathbf{Z} = \mathrm{LN}_C(\mathrm{FA}(\mathbf{D})), \\ \mathbf{Z}_r = \mathrm{Scan}_{\pi_r}(\mathrm{SiLU}(\mathrm{DWConv}_{3\times3}(\mathbf{Z}))), \\ \mathbf{Q} = \mathrm{LN}\left[\frac{1}{4}\sum_{r=1}^{4} \mathrm{Restore}_{\pi_r^{-1}}(\mathrm{Mamba}_r(\mathbf{Z}_r))\right] \\ \quad = \Phi(\mathrm{FA}(\mathbf{D})), \end{cases} \tag{10}$$

where LN_C denotes affine-free channel normalization, and Scan_{π_r} serializes the locally processed features into a sequence $\mathbf{Z}_r \in \mathbb{R}^{2B\times N\times C}$. Inward/outward traversals reverse accumulation relative to the center, while clockwise/counter-clockwise traversals reverse the contour order. Restoring the grid before averaging yields $\mathbf{Q} \in \mathbb{R}^{2B\times C\times h\times w}$. The operator Φ denotes the frequency-context transformation following FA, comprising normalization, DWConv, SiLU, the four Mamba branches, inverse routing, averaging, and final LN. The composed operator provides the frequency Mamba transformation in Fig. 5.

After unfolding the processed bands, the inverse discrete wavelet transform (IDWT) assigns the same fused-detail feature to all three Haar directional slots:

$$\begin{cases} (\mathbf{Q}_{\mathrm{LL}}, \mathbf{Q}_{\mathrm{HF}}) = \mathrm{Unfold}_B(\mathbf{Q}), \\ \mathbf{R} = \mathrm{IDWT}_{\mathrm{Haar}}(\mathbf{Q}_{\mathrm{LL}}, \mathbf{Q}_{\mathrm{HF}}, \mathbf{Q}_{\mathrm{HF}}, \mathbf{Q}_{\mathrm{HF}}), \\ \mathbf{S} = \mathbf{X} + \mathbf{U}, \\ \mathbf{Y} = \mathbf{S} + \mathbf{R} \odot \mathbf{U}. \end{cases} \tag{11}$$

Unfolding restores the $B \times 2 \times C \times h \times w$ organization before extracting the two $B \times C \times h \times w$ tensors. Synthesis returns $\mathbf{R}$ to the input grid, with any replicated boundary cropped. The residual structure separates direct local enhancement from frequency-conditioned modulation: $\mathbf{S} = \mathbf{X} + \mathbf{U}$ first combines the identity and activated normalized features, and the frequency-conditioned product is then added to the local residual feature. The same $\mathbf{U}$ participates in direct addition and multiplicative modulation, connecting the local response to the reconstructed Mamba context. The resulting relation $\mathbf{Y} = \mathbf{X} + \mathbf{U} + \mathbf{R} \odot \mathbf{U}$ preserves the input shape $B \times C \times H \times W$.

Interaction between frequency context and spatial features is motivated by FIRE-Net (Li et al. 2025). The present block implements this interaction through Haar analysis, shared two-band attention and Mamba processing, and residual modulation, as summarized in Fig. 5.

The selective recurrence grows linearly with route length N. The complete block also performs four-route scanning, wavelet analysis and synthesis, and precision conversion. The frequency branch is evaluated in 32-bit floating-point precision (FP32) during mixed-precision training, and the output enters the top-down FPN at its original feature resolution.

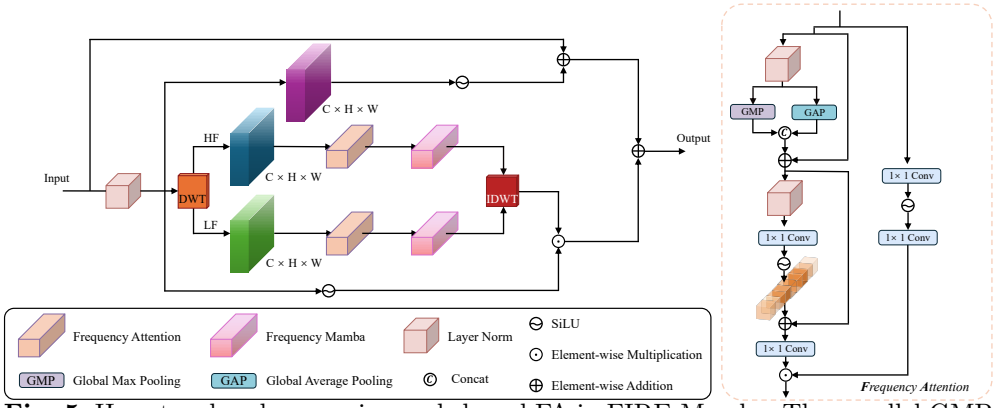


Fig. 5: Haar two-band processing and shared FA in FIRE-Mamba. The parallel GMP and GAP descriptors in Eq. (5) form $\mathbf{P}$; residual refinement produces $\mathbf{F}_{\mathrm{b2}}$, and the two projections yield $\mathrm{FA}(\mathbf{D}) = \mathbf{G} \odot \mathbf{V}$. The frequency Mamba transformation is $\mathbf{Q} = \Phi(\mathrm{FA}(\mathbf{D}))$ in Eq. (10). Inverse Haar synthesis supplies $\mathbf{R}$, giving $\mathbf{Y} = \mathbf{X} + \mathbf{U} + \mathbf{R} \odot \mathbf{U}$ in Eq. (11).

3.3 Haar Wavelet Downsampling

HWDS combines frequency-dependent enhancement with spatial reduction at the two PAN transitions. It applies the Haar analysis in Eq. (2) to an input $\mathbf{X}$ to obtain the approximation and three directional subbands, with $C = C_{\mathrm{e}}$.

Both enhancement branches use global-response channel attention (GRCA) to recalibrate channels according to their global response strength. For a feature $\mathbf{Z} \in \mathbb{R}^{B \times C_{\mathrm{a}} \times h \times w}$ with attention width C_{a},

$$\begin{cases} g_{bc} = \left(\sum_{u=1}^{h} \sum_{v=1}^{w} Z_{bcuv}^2 \right)^{1/2}, \\ n_{bc} = \dfrac{g_{bc}}{C_{\mathrm{a}}^{-1} \sum_{c'=1}^{C_{\mathrm{a}}} g_{bc'} + \epsilon}, \\ \mathrm{GRCA}(\mathbf{Z}) = \mathbf{Z} + \boldsymbol{\gamma} \odot (\mathbf{Z} \odot \mathbf{n}) + \boldsymbol{\beta}. \end{cases} \tag{12}$$

The stabilizer $\epsilon = 10^{-6}$ keeps the normalization denominator nonzero, and the indices b, c, u, v refer to batch, channel, row, and column. The channel response g_{bc} measures spatial response strength through the L2 norm, and n_{bc} expresses that norm relative to the mean response across channels. Both descriptors have shape $B \times C_{\mathrm{a}} \times 1 \times 1$ and are broadcast spatially. The learnable scale and shift $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^{1 \times C_{\mathrm{a}} \times 1 \times 1}$ are initialized to zero, so GRCA initially preserves its input. GRCA recalibrates the compressed features in both enhancement branches.

The low-frequency branch smooths the approximation with a fixed depthwise Gaussian filter $\mathcal{G}_{\mathrm{dw}}$, then learns a compact representation through depthwise and pointwise (PW) convolution followed by GRCA in Eq. (12):

$$\begin{cases} C_{\mathrm{b}} = \max(1, \lfloor 0.125C \rfloor), \\ \mathbf{L}_{\mathrm{b}} = \mathrm{GRCA}\big(\mathrm{PW}^{\mathrm{L}}_{C \to C_{\mathrm{b}}}(\mathrm{DWConv}_{3\times3}(\mathcal{G}_{\mathrm{dw}}(\mathbf{D}_{\mathrm{LL}})))\big), \\ \mathbf{L} = \mathrm{Restore}^{\mathrm{L}}_{C_{\mathrm{b}} \to C}(\mathbf{L}_{\mathrm{b}}). \end{cases} \tag{13}$$

DWConv and PW denote learned depthwise and pointwise convolutions. The low- and high-frequency branches both use attention width $C_{\mathrm{a}} = C_{\mathrm{b}}$. The bottleneck $\mathbf{L}_{\mathrm{b}}$ in Eq. (13) has C_{b} channels, and $\mathrm{Restore}^{\mathrm{L}}$ restores C channels. The fixed Gaussian filter uses the kernel $\frac{1}{16}[1,2,1]^{\mathsf{T}}[1,2,1]$ to smooth the approximation before learned mixing. The DWConv and PW operations are followed by batch normalization and SiLU

activation; the restore operation uses a 1×1 convolution with the same normalization and activation.

The high-frequency branch first applies a fixed depthwise Laplacian $\mathcal{L}_{\mathrm{dw}}$ with kernel $\left[\begin{smallmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{smallmatrix}\right]$ to each directional subband. It aggregates the filtered responses, enhances them in a shared bottleneck, and applies learned pixel attention (PA) followed by sigmoid modulation:

$$\begin{cases} \mathbf{H}_{\mathrm{cat}} = \mathrm{Concat}[\mathcal{L}_{\mathrm{dw}}(\mathbf{D}_{\mathrm{LH}}), \mathcal{L}_{\mathrm{dw}}(\mathbf{D}_{\mathrm{HL}}), \mathcal{L}_{\mathrm{dw}}(\mathbf{D}_{\mathrm{HH}})], \\ \mathbf{H}_{\mathrm{b}} = \mathrm{Aggregate}_{3C \to C_{\mathrm{b}}}(\mathbf{H}_{\mathrm{cat}}), \\ \mathbf{E}_{\mathrm{H}} = \mathrm{GRCA}\big(\mathrm{PW}^{\mathrm{H}}_{C_{\mathrm{b}} \to C_{\mathrm{b}}}(\mathrm{DWConv}_{3\times 3}(\mathbf{H}_{\mathrm{b}}))\big), \\ \widehat{\mathbf{H}}_b = \mathbf{E}_{\mathrm{H}} \odot \mathrm{sigmoid}(\mathrm{PA}_{1\times 1}(\mathbf{E}_{\mathrm{H}})), \\ [\widehat{\mathbf{H}}_{\mathrm{LH}}; \widehat{\mathbf{H}}_{\mathrm{HL}}; \widehat{\mathbf{H}}_{\mathrm{HH}}] = \mathrm{Restore}^{\mathrm{H}}_{C_{\mathrm{b}} \to 3C}(\widehat{\mathbf{H}}_b). \end{cases} \tag{14}$$

Concat and semicolons denote channel concatenation, sigmoid is the logistic activation, and $\odot$ is elementwise multiplication. The aggregate and restore operators use 1×1 convolution, batch normalization, and SiLU activation. $\mathbf{H}_{\mathrm{cat}}$ has $3C$ channels, the bottleneck tensors have C_{b} channels, and restoration produces three C-channel directional tensors. The concatenated, bottleneck, and restored tensors all occupy the $h \times w$ grid. The PA projection $\mathrm{PA}_{1\times 1}$ is a learned $C_{\mathrm{b}} \to C_{\mathrm{b}}$ convolution with bias. The resulting sigmoid gate has the same shape as $\mathbf{E}_{\mathrm{H}}$ and modulates each enhanced response. The Gaussian and Laplacian filters use unit stride and zero padding to preserve their subband grids.

Finally, a grouped channel projection combines the four enhanced bands with an average-pooling shortcut:

$$\begin{cases} \mathbf{V} = \mathrm{Concat}[\mathbf{L}, \widehat{\mathbf{H}}_{\mathrm{LH}}, \widehat{\mathbf{H}}_{\mathrm{HL}}, \widehat{\mathbf{H}}_{\mathrm{HH}}], \\ \mathcal{H}_{\mathrm{HWDS}}(\mathbf{X}) = \mathrm{Reduce}_{4C \to C}(\mathbf{V}) + \mathrm{AvgPool}^{\mathrm{ceil}}_{2,2}(\mathbf{X}). \end{cases} \tag{15}$$

The final channel reduction combines the enhanced approximation and directional detail using 1×1 convolution with two channel groups, batch normalization, and SiLU activation. The pooling operator uses a 2×2 window, stride two, and ceil-sized output; any remaining boundary-size difference is aligned to the wavelet grid by bilinear interpolation. The HWDS output has shape $B \times C \times h \times w$ and can be added to the next lateral feature in the bottom-up PAN. Figure 6 summarizes how both frequency branches reach this decimated output without inverse wavelet synthesis.

HWDS supplies the approximation and directional inputs to separate enhancement branches. Its shared high-frequency bottleneck then restores three directional outputs before the final $4C \to C$ reduction, while the pooling branch provides a direct path from the input. Figure 7 compares standard downsampling, the fast Fourier transform (FFT), Haar-DWT, and HWDS with fixed scene content, target coordinates, and display scale.

In the selected POP scene, the four operators retain different portions of the compact thermal structures and surrounding vegetation. HWDS separates the principal foreground responses while preserving the half-resolution output required by PAN fusion. This qualitative comparison complements the common-protocol detection results in Table 12.

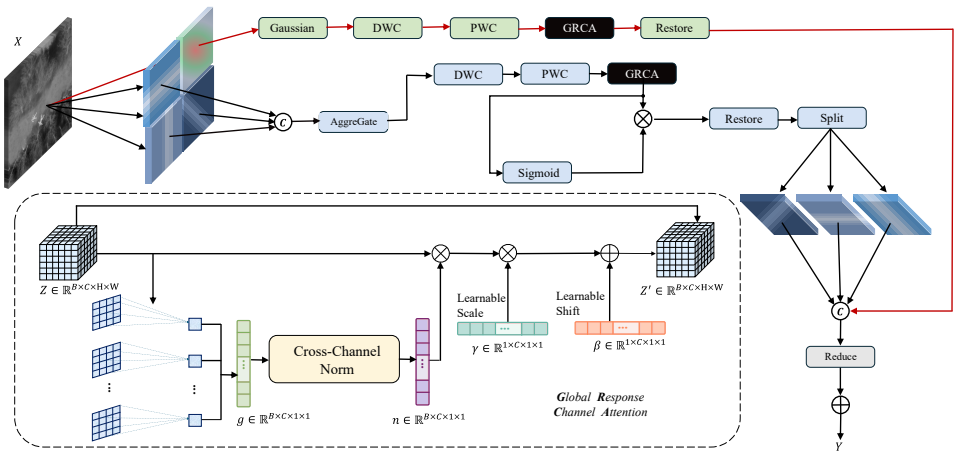
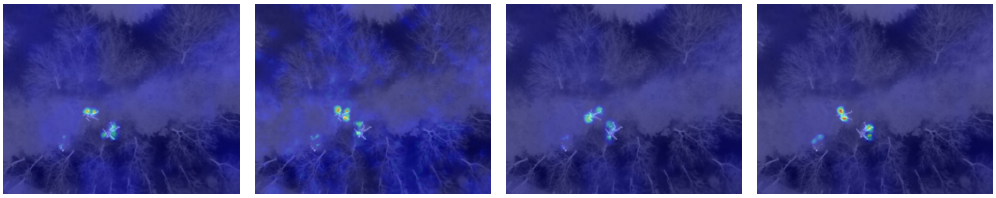


Fig. 6: HWDS architecture. The approximation output $\mathbf{L}$ and restored directional outputs $\widehat{\mathbf{H}}_{\mathrm{LH}}, \widehat{\mathbf{H}}_{\mathrm{HL}}, \widehat{\mathbf{H}}_{\mathrm{HH}}$ form the concatenated feature $\mathbf{V}$ in Eq. (15). Gaussian smoothing and directional Laplacian filtering precede learned enhancement. The high-frequency gate uses the enhanced feature: $\widehat{\mathbf{H}}_b = \mathbf{E}_{\mathrm{H}} \odot \mathrm{sigmoid}(\mathrm{PA}_{1\times1}(\mathbf{E}_{\mathrm{H}}))$ in Eq. (14). Grouped reduction of $\mathbf{V}$ is added to the average-pooling shortcut to form $\mathcal{H}_{\mathrm{HWDS}}(\mathbf{X})$.



Standard downsampling | FFT | Haar-DWT | HWDS

Fig. 7: Aligned responses of standard downsampling, FFT, Haar-DWT, and HWDS. Each column uses the same POP scene, target coordinates, and display scale to compare the spatial selectivity of the reduced representation.

3.4 SA-NWD-MAL

For a compact target, a small center displacement can sharply reduce overlap or produce zero IoU. SA-NWD-MAL couples the classification-quality target to a smoother geometric similarity in this regime and adds a corresponding localization term. The target dimensions control both adjustments, so their contribution decreases continuously as the target becomes larger. The detector retains its assignment procedure, auxiliary supervision, and distribution-refinement losses; the formulation below distinguishes their matched sets and normalizers before defining the complete objective.

Let $\mathcal{A}$ collect the ordinary prediction stages: the final and intermediate decoder outputs, the preliminary box predictions, and the encoder predictions. Each stage s has its own classification matching $\mathcal{M}_s$. For box and fine-grained localization (FGL) supervision, the stage-specific matchings are combined by retaining, for each query, its most frequently assigned target across stages:

$$\begin{cases} j_i^\star \in \arg\max\limits_{j} \sum\limits_{s\in\mathcal{A}} \mathbb{1}[(i,j)\in\mathcal{M}_s], \\ \mathcal{U} = \{(i, j_i^\star) : i \text{ occurs in at least one } \mathcal{M}_s\}. \end{cases} \tag{16}$$

In Eq. (16), query indices include image membership, so assignments are combined only within the same image. Classification uses $\mathcal{M}_s$; ordinary box regression and FGL use the query-unique set $\mathcal{U}$. Let N_{gt} and $N_{\mathcal{U}}$ be the per-device mean target count and

union-pair count, respectively, each lower-bounded by one. With distributed training, counts are summed across devices and divided by the number of devices before clipping. Let $\mathcal{D}$ collect the denoising stages, with predefined positive correspondences $\mathcal{M}_{\mathrm{dn},s}$ and G_s denoising groups. The classification and regression normalizers are

$$\begin{cases} N_{\mathrm{cls}}^{(s)} = \begin{cases} N_{\mathrm{gt}}, & s \in \mathcal{A}, \\ G_s N_{\mathrm{gt}}, & s \in \mathcal{D}, \end{cases} \\ N_{\mathrm{reg}}^{(s)} = \begin{cases} N_{\mathcal{U}}, & s \in \mathcal{A}, \\ G_s N_{\mathrm{gt}}, & s \in \mathcal{D}. \end{cases} \end{cases} \tag{17}$$

The normalizers in Eq. (17) preserve the distinction between classification matches and regression pairs. For either type of matched pair, write the normalized predicted and target boxes as $\mathbf{b}_i^{\mathrm{p}}, \mathbf{b}_i^{\mathrm{t}}$, with $\mathbf{b} = (c_x, c_y, w, h)$. Horizontal and vertical coordinates are normalized by their corresponding input dimensions. The Gaussian box representation and NWD quality are (Wang et al. 2021; Xu et al. 2022)

$$\begin{cases} g(\mathbf{b}) = [c_x, c_y, w/2, h/2]^{\mathsf{T}}, \\ q_{\mathrm{nwd},i} = \exp\left(-\dfrac{\| g(\mathbf{b}_i^{\mathrm{p}}) - g(\mathbf{b}_i^{\mathrm{t}}) \|_2}{C_{\mathrm{nwd}}} \right). \end{cases} \tag{18}$$

The two center coordinates specify position, and $w/2, h/2$ are diagonal standard deviations. Exponential decay in Eq. (18) converts center-and-extent displacement to $(0, 1]$, preserving a nonzero similarity for nearby nonoverlapping boxes. The constant $C_{\mathrm{nwd}} = 0.02$ sets the decay scale in normalized coordinates. Target size then controls the NWD contribution to the classification-quality target:

$$\begin{cases} s_i = \sqrt{w_i^{\mathrm{t}} h_i^{\mathrm{t}}}, \\ a_i = \mathrm{clip}\left(1 - \dfrac{s_i}{\tau_s}, 0, 1 \right), \\ q_{\mathrm{sa},i} = (1 - a_i) q_{\mathrm{iou},i} + a_i q_{\mathrm{nwd},i}. \end{cases} \tag{19}$$

The scale threshold in Eq. (19) is $\tau_s = 0.05$. For $s_i \geq \tau_s$, the weight a_i is zero and the classification-quality target reduces to IoU. Smaller targets receive progressively greater NWD weight. Continuous variation of the mixture weight avoids an abrupt change at the scale threshold. Because the weight depends on target dimensions, predicted box size cannot alter its supervision weight.

For query i and class c, let $t_{ic}^{(s)}$ indicate a matched positive class, and let $p_{ic}^{(s)}$ be the predicted probability. Denote stop-gradient by sg. MAL weights binary cross-entropy (BCE) using a detached soft target for positives and confidence-dependent weights for all remaining query–class pairs (Huang et al. 2025b):

$$\begin{cases} \tilde{q}_{ic}^{(s)} = t_{ic}^{(s)} \, \mathrm{sg}(q_{\mathrm{sa},i}^{(s)})^{\gamma}, \\ \omega_{ic}^{(s)} = t_{ic}^{(s)} + (1 - t_{ic}^{(s)}) \, \mathrm{sg}(p_{ic}^{(s)})^{\gamma}, \\ \mathcal{L}_{\mathrm{mal}}^{(s)} = \dfrac{1}{N_{\mathrm{cls}}^{(s)}} \displaystyle\sum_{i,c} \omega_{ic}^{(s)} \, \mathrm{BCE}(p_{ic}^{(s)}, \tilde{q}_{ic}^{(s)}). \end{cases} \tag{20}$$

The focusing exponent in Eq. (20) is $\gamma = 1.5$, and $\mathrm{BCE}(p, q) = -q \log p - (1 - q) \log(1 - p)$. Classification quality is evaluated using each ordinary stage's own matches and the predefined correspondences of each denoising stage. A matched positive has unit outer weight and target q_{sa}^{γ}; negative classes and unmatched queries have target zero and weight $\mathrm{sg}(p)^{\gamma}$. NWD therefore changes the desired confidence of a small matched target, while the inherited MAL weighting continues to emphasize confident negative responses. Detachment separates this classification-quality supervision from the gradients supplied by box regression.

Let $\mathcal{R}_s = \mathcal{U}$ for ordinary stages and $\mathcal{R}_s = \mathcal{M}_{\mathrm{dn},s}$ for denoising stages. The localization components combine coordinate distance, generalized intersection-over-union (GIoU), and size-weighted NWD:

$$\begin{cases} \mathcal{L}_1^{(s)} = \dfrac{1}{N_{\mathrm{reg}}^{(s)}} \displaystyle\sum_{(i,j)\in\mathcal{R}_s} \|\mathbf{b}_i^{\mathrm{p}} - \mathbf{b}_j^{\mathrm{t}}\|_1, \\ \mathcal{L}_{\mathrm{giou}}^{(s)} = \dfrac{1}{N_{\mathrm{reg}}^{(s)}} \displaystyle\sum_{(i,j)\in\mathcal{R}_s} \left[1 - \mathrm{GIoU}(\mathbf{b}_i^{\mathrm{p}}, \mathbf{b}_j^{\mathrm{t}})\right], \\ \mathcal{L}_{\mathrm{nwd}}^{(s)} = \dfrac{1}{N_{\mathrm{reg}}^{(s)}} \displaystyle\sum_{(i,j)\in\mathcal{R}_s} a_j(1 - q_{\mathrm{nwd},ij}), \\ \mathcal{L}_{\mathrm{box}}^{(s)} = 5\mathcal{L}_1^{(s)} + 2\mathcal{L}_{\mathrm{giou}}^{(s)} + 2\mathcal{L}_{\mathrm{nwd}}^{(s)}. \end{cases} \tag{21}$$

The pair (i, j) associates a predicted query with its assigned target, and $q_{\mathrm{nwd},ij}$ measures their NWD similarity. In Eq. (21), the L1 term supervises the four box coordinates, GIoU captures region geometry, and NWD adds a size-weighted distance on the Gaussian representation. Unlike the detached quality in MAL, q_{nwd} in the NWD regression term carries gradients to the predicted box. A target with $a_j = 0$ contributes no additional NWD penalty. With no positive pairs, the regression sums are zero. The four-coordinate NWD distance adds $O(|\mathcal{R}_s|)$ work per stage and is computed in FP32 for half- or bfloat16 inputs.

The fine-grained distribution refinement and self-distillation objectives are inherited from the D-FINE/DEIMv2 detector (Huang et al. 2025a; Peng et al. 2025a). For a stage with predicted box-side distributions, let $P_{id}^{(s)}(k)$ be the probability of bin k for side $d \in \{1, 2, 3, 4\}$. The encoded target distance lies between bins k_{id}^- and k_{id}^+, with interpolation weights $\eta_{id}^- + \eta_{id}^+ = 1$. FGL retains IoU weighting:

$$\begin{aligned} \mathcal{L}_{\mathrm{fgl}}^{(s)} = -\frac{1}{N_{\mathrm{reg}}^{(s)}} \sum_{(i,j)\in\mathcal{R}_s} \sum_{d=1}^{4} \mathrm{sg}(q_{\mathrm{iou},ij}) \\ \times \left[\eta_{id}^- \log P_{id}^{(s)}(k_{id}^-) + \eta_{id}^+ \log P_{id}^{(s)}(k_{id}^+)\right]. \end{aligned} \tag{22}$$

The target distances and interpolation weights follow the inherited distance-to-bin mapping. The FGL term in Eq. (22) encourages each predicted side distribution to place mass around the target location, complementing supervision on the decoded box. FGL retains IoU weighting; size adaptation affects the MAL quality targets and the additional NWD regression term.

For a student stage with a teacher distribution, let $\mathbf{z}_{id}^{(s)}$ and $\mathbf{z}_{id}^{\star}$ be student and teacher side logits. The distribution self-distillation term $\mathcal{L}_{\mathrm{ddf}}$ uses Kullback–Leibler (KL) divergence at temperature $T = 5$:

$$\begin{cases} \ell_{id}^{(s)} = T^2\,\mathrm{KL}\Big(\mathrm{sg}[\mathrm{softmax}(\mathbf{z}_{id}^{\star}/T)] \,\Big\|\, \mathrm{softmax}(\mathbf{z}_{id}^{(s)}/T)\Big), \\ v_i^{(s)} = \begin{cases} \mathrm{sg}(q_{\mathrm{iou},i}^{(s)}), & i \text{ is matched}, \\ \mathrm{sg}(\max_c p_{ic}^{\star}), & i \text{ is unmatched}, \end{cases} \\ \mathcal{L}_{\mathrm{ddf}}^{(s)} = \dfrac{\sqrt{n_+}\,\langle v_i^{(s)}\ell_{id}^{(s)}\rangle_+ + \sqrt{n_-}\,\langle v_i^{(s)}\ell_{id}^{(s)}\rangle_-}{\sqrt{n_+} + \sqrt{n_-}}, \end{cases} \tag{23}$$

where angle brackets denote separate means over matched and unmatched box sides, with an empty-group mean set to zero. Ordinary stages define the matched group using $\mathcal{U}$, while denoising stages use their predefined positive correspondences. The counts n_+, n_- enumerate the four box sides of matched and unmatched ordinary queries, respectively, and the same counts are reused for denoising distillation. The teacher is the final decoder output for the corresponding ordinary or denoising query group, and $p_{ic}^{\star}$ denotes its class probability. Teacher probabilities and quality weights

are detached. Computing matched and unmatched means separately in Eq. (23) prevents the larger unmatched group from determining the average solely through its size; the square-root counts retain a controlled dependence on group population. Distillation is applied only when the teacher output comes from a different prediction stage.

Let $\mathcal{S} = \mathcal{A} \cup \mathcal{D}$, and let $\mathcal{S}_{\text{fgl}}$ and $\mathcal{S}_{\text{ddf}}$ contain the stages with side-distribution supervision and teacher distillation, respectively. The complete training objective is

$$\begin{aligned}\mathcal{L}_{\text{total}} = &\sum_{s\in\mathcal{S}} \left(\mathcal{L}_{\text{mal}}^{(s)} + \mathcal{L}_{\text{box}}^{(s)}\right) \\ &+ 0.15 \sum_{s\in\mathcal{S}_{\text{fgl}}} \mathcal{L}_{\text{fgl}}^{(s)} + 1.5 \sum_{s\in\mathcal{S}_{\text{ddf}}} \mathcal{L}_{\text{ddf}}^{(s)}.\end{aligned} \tag{24}$$

The complete objective in Eq. (24) assigns MAL, L1, GIoU, NWD, FGL, and distribution self-distillation the coefficients $1, 5, 2, 2, 0.15$, and 1.5. The same component coefficients apply to the supervised ordinary and denoising stages. Auxiliary decoder, preliminary, and encoder predictions make the objective act throughout prediction refinement; denoising predictions use their constructed correspondences and group-scaled normalization. Within this retained training framework, SA-NWD-MAL aligns small-target confidence with geometric proximity and supplies a compatible box gradient. During inference, the decoder produces class probabilities and normalized bounding boxes without evaluating the training losses.

4 Experiments and Results

4.1 Benchmarks and Evaluation Protocol

Desktop and embedded measurements use the hardware and software environment summarized in Table 2.

Table 2: Experimental hardware and software environment.

Component	Specification
Desktop GPU	NVIDIA GeForce RTX 4070 Ti
Edge device	NVIDIA Jetson Orin Nano 8GB
Edge power mode	MAXN_SUPER
Edge software	JetPack 6.2.1
Processor	Intel Core i5-12600K
System memory	48 GB DDR5
CUDA compilation tools	Release 11.8, V11.8.89
PyTorch	2.7.1
Python	3.12.12

Experiments use an NVIDIA RTX 4070 Ti graphics processing unit (GPU), an Intel Core i5-12600K, 48 GB DDR5 memory, CUDA compilation tools release 11.8 (V11.8.89), PyTorch 2.7.1, and Python 3.12.12. Embedded measurements use an NVIDIA Jetson Orin Nano 8GB in MAXN_SUPER mode with JetPack 6.2.1. Both FHS-DEIM variants use 640×640 input; M has encoder width 256 and four decoder layers, while X has encoder width 384 and six decoder layers. Tables 3 and 4 separately summarize the comparison methods and their training protocols.

Table 3: Characteristics of comparison methods.

Source	Year	Algorithm	Application scenario
IEEE TGRS	2026	LIRS-YOLO (Guan et al. 2026)	Lightweight infrared small-object detection from UAVs; evaluated on HIT-UAV and DroneVehicle.
IEEE JSTARS	2025	KDFF-DETR (Peng et al. 2025b)	Spaceborne infrared ship detection on ISDD, NUDT-SIRST-Sea, ISOD, and IRSDSS.
IEEE JSTARS	2026	IR-DETR (Hu et al. 2026)	Infrared small-object box detection on IRSTD-1K, SIRST-UAVB, SIRST-v1, and NUDT-SIRST.
IEEE/CVF CVPR	2024	RT-DETR-L (Zhao et al. 2024)	General real-time end-to-end detection with a ResNet-50-m (R50-m) backbone.
arXiv	2025	DEIMv2-HGNetv2-M (Huang et al. 2025a)	Multiscale real-time end-to-end detection and the direct architectural baseline used here.
ICCV	2021	Conditional DETR-R50 (Meng et al. 2021)	General end-to-end detection with accelerated convergence; originally evaluated on COCO.
ICLR	2021	Deformable DETR-R50 (Zhu et al. 2021)	General multiscale end-to-end detection with sparse deformable sampling; originally evaluated on COCO.

Table 4: Published reference protocols for third-party methods and the training settings of FHS-DEIM.

Method	Training setting
LIRS-YOLO (Guan et al. 2026)	200 epochs; SGD; LR 0.01; weight decay 0.01; batch 16; input 640^2; no pretraining
KDFF-DETR (Peng et al. 2025b)	72 epochs; AdamW; LR 10^{-4}; weight decay 10^{-4}; 2000-step warmup; batch 16; input 640^2
IR-DETR (Hu et al. 2026)	100 epochs; SGD; LR 0.01; momentum 0.937; weight decay 5×10^{-4}; batch 12; input 640
RT-DETR-L (Zhao et al. 2024)	ResNet-50-m (R50-m) backbone; 72 epochs; AdamW; base LR 10^{-4}; weight decay 10^{-4}; 2000-step linear warmup; batch 4 per GPU on four GPUs; training scales 480–800; evaluation input 640^2
DEIMv2-HGNetv2-M (Huang et al. 2025a)	102 epochs; AdamW; LR 4×10^{-4}; backbone LR 4×10^{-5}; total batch 32; input 640^2
Conditional DETR-R50 (Meng et al. 2021)	50 epochs; AdamW; transformer LR 10^{-4}; backbone LR 10^{-5}; total batch 16
Deformable DETR-R50 (Zhu et al. 2021)	50 epochs; AdamW; LR 2×10^{-4}; backbone LR 2×10^{-5}; total batch 32
FHS-DEIM-M	102 epochs; AdamW; LR 4×10^{-4}; backbone LR 4×10^{-5}; weight decay 10^{-4}; batch 8; input 640^2; encoder width 256; four decoder layers; two FIRE-Mamba stages; two HWDS reductions; SSM state dimension 16, local-convolution width 4, and expansion factor 2
FHS-DEIM-X	58 epochs; AdamW; LR 5×10^{-4}; backbone LR 5×10^{-6}; weight decay 1.25×10^{-4}; batch 32; input 640^2; encoder width 384; six decoder layers; two FIRE-Mamba stages; two HWDS reductions; SSM state dimension 16, local-convolution width 4, and expansion factor 2

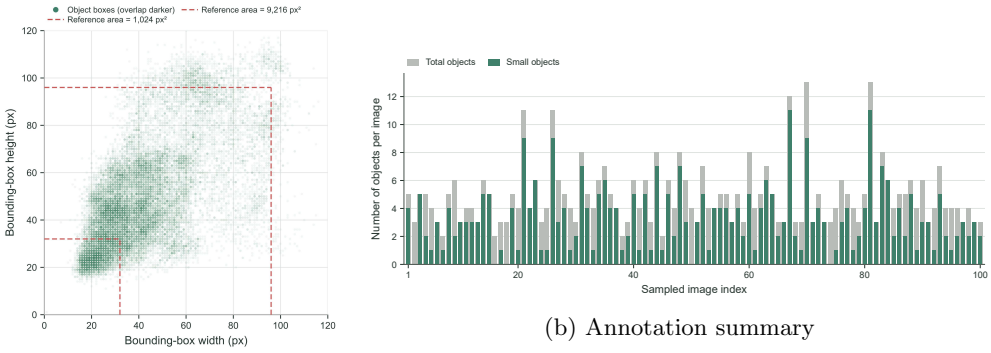
SGD: stochastic gradient descent; LR: learning rate.

The comparison spans Faster R-CNN (Ren et al. 2015); single-stage detectors including SSD, RTMDet, YOLO11, and YOLO26 (Jocher and Qiu 2024; Jocher et al. 2026; Liu et al. 2016; Lyu et al. 2022); and query-based architectures including Conditional DETR, Deformable DETR, and the infrared KDFF-DETR (Meng et al. 2021;

Peng et al. 2025b; Zhu et al. 2021). These families place the proposed encoder and supervision changes in the context of established detection architectures.

Accuracy is reported as AP under Common Objects in Context (COCO) evaluation. Epoch count ranges from 24 to 200 in the comparison rows, and optimizer, batch size, and input resolution also differ by method. Table 4 records the published reference protocol for each third-party method and the actual settings of FHS-DEIM. All reported accuracy tables use the corresponding test split. Component, scan, and downsampling variants use a common POP test protocol and support direct within-table contrasts. The M and X configurations change encoder width and decoder depth together; each DEIMv2 and FHS-DEIM pair therefore provides a configuration-level comparison under its corresponding training recipe. Parameter count measures model size, and billions of floating-point operations (GFLOPs) measure nominal arithmetic. Throughput in frames per second (FPS) is measured with batch size 1 and PyTorch 2.7.1 in FP32 over one complete pass through the POP test split; the timed pipeline includes image reading, preprocessing, host-to-device transfer, network inference, decoding, and post-processing on the indicated RTX 4070 Ti or Jetson Orin Nano 8GB device.

Datasets. The POP dataset was acquired with the long-wave infrared camera of a DJI Matrice 30T for wilderness SAR (Song et al. 2024, 2025).



(a) Box width and height

(b) Annotation summary

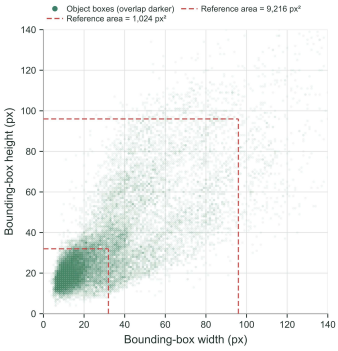
Fig. 8: Bounding-box distribution and annotation summary of POP. The scatter plot shows object width versus height, and the summary panels describe the corresponding annotation inventory.

It contains open and vegetation-occluded people recorded at approximately 30, 50, and 70 m under 19 acquisition conditions. Figure 8 summarizes the box distribution and annotation inventory. POP contains 5,548 training, 2,320 validation, and 900 test images, for 8,768 images in total. The POP statistics in Table 5 cover all three splits and include 26,811 person instances, with a mean box size of 42.30 × 49.24 pixels. Because POP contains only the person category (class 0), the corpus-level and class-specific counts and mean dimensions are identical. The released COCO annotation file stores the constant 1.1 in every instance's `area` field. We reconstruct the field from the bounding box as

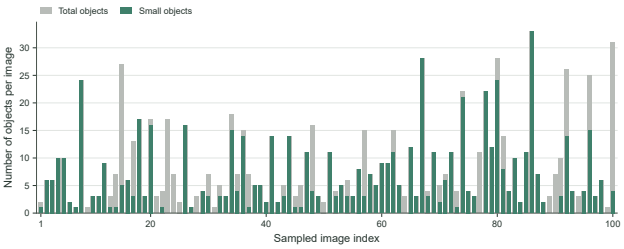
$$\widetilde{a}_i = w_i h_i, \tag{25}$$

while leaving the boxes, categories, and dataset split unchanged. The areas corrected by Eq. (25) are then used by the standard COCO small-, medium-, and large-object evaluators (Lin et al. 2014).

HIT-UAV was captured from a high-altitude UAV platform with thermal imagery and flight metadata covering 60–130 m altitude and 30°–90° viewing angles (Suo et al. 2023).



(a) Box width and height



(b) Annotation summary

Fig. 9: Bounding-box distribution and annotation summary of HIT-UAV. The scatter plot shows object width versus height, and the summary panels describe the corresponding annotation inventory.

The official HIT-UAV release contains 2,898 images, divided into 2,029 training, 579 test, and 290 validation images, with 24,899 objects in total. The annotated corpus contains 2,866 images; the remaining 32 images have no object annotations. The distribution statistics summarize this annotated corpus across the dataset, with a global mean box size of 28.74×31.88 pixels. The five class IDs are 0 (Person), 1 (Car), 2 (Bicycle), 3 (OtherVehicle), and 4 (DontCare). In our evaluation, DontCare is treated as the regular fifth category, Others, and contributes to AP together with the other four categories. Figure 9 shows the box distribution and annotation summary, while Table 5 gives the corpus-level and class-specific counts and mean dimensions. Models are trained and evaluated separately on each benchmark's corresponding splits. HIT-UAV therefore examines generalization to a broader multicategory aerial thermal setting that includes person detection.

Table 5: Dataset distribution and size characteristics. POP statistics cover all training, validation, and test images. HIT-UAV statistics cover annotated images across all splits, excluding 32 images without annotations. DontCare (ID 4) is evaluated as the fifth category, Others. Mean box dimensions are in pixels.

Dataset	Corpus				Class			
	Images	Objects	Mean w	Mean h	ID	Count	Mean w	Mean h
POP	8,768	26,811	42.30	49.24	0 (person)	26,811	42.30	49.24
HIT-UAV	2,866	24,899	28.74	31.88	0	12,312	12.95	19.96
					1	7,311	52.93	52.22
					2	4,980	30.74	30.03
					3	148	62.68	68.80
					4	148	45.45	43.90

Distribution structure. The five HIT-UAV class counts sum to 24,899. Class 0 contributes 12,312 objects, about 49.4% of the annotated corpus; classes 1 and 2 contribute 7,311 and 4,980, while classes 3 and 4 contain only 148 objects each. The smaller classes provide fewer examples for assessing their detection performance, while differences in box scale motivate reporting scale-stratified AP alongside aggregate AP. The class-0 mean box is smaller than the global mean, whereas classes 1, 3, and 4 have substantially larger mean dimensions.

Evaluation metrics. For confidence cutoff c, precision and recall are

$$\text{Precision}(c) = \frac{\text{TP}(c)}{\text{TP}(c) + \text{FP}(c)}, \tag{26}$$

$$\text{Recall}(c) = \frac{\text{TP}(c)}{\text{TP}(c) + \text{FN}(c)}, \tag{27}$$

where $\text{TP}(c)$, $\text{FP}(c)$, and $\text{FN}(c)$ count true positives, false positives, and false negatives at cutoff c, respectively.

Following the stated COCO-style evaluation protocol (Lin et al. 2014), let $p(r; \tau_{\text{IoU}})$ denote precision at recall r and IoU threshold τ_{IoU}. The precision envelope and the 101-point AP are

$$p_{\text{interp}}(r; \tau_{\text{IoU}}) = \max_{\widetilde{r} \geq r} p(\widetilde{r}; \tau_{\text{IoU}}), \tag{28}$$

$$\text{AP}_{\tau_{\text{IoU}}} = \frac{1}{101} \sum_{r \in \{0, 0.01, \ldots, 1\}} p_{\text{interp}}(r; \tau_{\text{IoU}}), \tag{29}$$

$$\text{AP} = \frac{1}{10} \sum_{\tau_{\text{IoU}} \in \{0.50, 0.55, \ldots, 0.95\}} \text{AP}_{\tau_{\text{IoU}}}\,. \tag{30}$$

AP_{50} and AP_{75} use individual IoU thresholds; headline AP averages thresholds from 0.50 to 0.95 and therefore evaluates tighter geometric agreement than AP_{50} alone. Category aggregation uses the person category on POP and all five categories, including DontCare evaluated as Others, on HIT-UAV. AP_S, AP_M, and AP_L condition evaluation on the COCO small, medium, and large object-area ranges, using the reconstructed POP areas defined above rather than the class-level mean widths in Table 5. AP_S is emphasized because both benchmarks contain small targets.

4.2 Comparative Experiments

The comparison includes two-stage and single-stage convolutional neural network (CNN) detectors and end-to-end Transformer detectors, covering general and infrared-specific approaches. Across the result tables, red boldface marks the best value and blue boldface the second-best distinct value in each metric column, following the indicated optimization direction; ties receive the same formatting.

POP accuracy. The POP comparison tests person detection across target sizes and overlap criteria using the complete FHS-DEIM configurations (Table 6).

Table 6: Comparative experimental results on the POP dataset.

Model	Epoch	AP (%) ↑	AP_{50} (%) ↑	AP_{75} (%) ↑	AP_S (%) ↑	AP_M (%) ↑	AP_L (%) ↑
Two-stage CNN detectors							
Faster R-CNN (Ren et al. 2015)	24	45.6	74.2	52.5	40.8	44.4	54.1
Single-stage CNN detectors							
SSD300 (Liu et al. 2016)	100	39.5	70.5	46.7	37.6	38.3	47.5
RTMDet-tiny (Lyu et al. 2022)	100	51.2	74.0	56.4	44.1	47.8	61.2
YOLO11m (Jocher and Qiu 2024)	200	59.7	84.4	68.5	56.7	57.2	68.3
YOLO11x (Jocher and Qiu 2024)	200	60.4	85.6	69.4	57.3	58.4	69.6
YOLO26m (Jocher et al. 2026)	200	60.6	85.2	69.5	57.2	58.1	71.7
YOLO26x (Jocher et al. 2026)	200	61.2	87.2	70.6	58.4	59.5	73.8
LIRS-YOLO (Guan et al. 2026)	200	60.7	85.8	71.0	58.7	59.8	70.4
End-to-end Transformer detectors							
DEIMv2-M (Huang et al. 2025a)	102	60.6	86.6	70.1	57.7	60.5	73.1
DEIMv2-X (Huang et al. 2025a)	58	62.4	**88.9**	**72.5**	58.6	61.3	74.3
RT-DETR-L (Zhao et al. 2024)	72	57.8	82.2	68.6	53.5	50.5	65.6
Conditional DETR (Meng et al. 2021)	50	49.5	75.2	56.4	45.6	49.8	62.3
IR-DETR (Hu et al. 2026)	100	59.5	86.1	69.7	57.3	59.5	69.9
KDFF-DETR (Peng et al. 2025b)	72	59.2	84.2	70.2	57.9	60.3	71.3
Ours							
FHS-DEIM-M	102	**62.9**	88.3	72.4	**59.8**	**62.1**	**75.3**
FHS-DEIM-X	58	**64.2**	**90.3**	**75.4**	**61.6**	**63.7**	**77.2**

FHS-DEIM improves detection across target sizes while retaining gains at the stricter overlap threshold (Table 6). Relative to DEIMv2-M, FHS-DEIM-M gains 2.3 AP points, with increases of 1.7 in AP_{50} and 2.3 in AP_{75}. The corresponding X-scale gains are 1.8, 1.4, and 2.9 points. The larger AP_{75} than AP_{50} improvement, especially for X, shows that the advantage persists when detections must satisfy tighter geometric agreement. Small-object AP also increases by 2.1 points for M and 3.0 for X, alongside positive medium- and large-object changes. Thus, the observed benefit addresses compact-person detection without sacrificing the other size strata in POP.

The broader comparison places this gain against general and infrared-specific detectors. FHS-DEIM-X attains the highest value in every POP accuracy column, exceeding the strongest non-FHS aggregate result, DEIMv2-X, by 1.8 AP points. FHS-DEIM-M exceeds LIRS-YOLO, the strongest infrared-specific aggregate row, by 2.2 points. Moving from FHS-DEIM-M to X adds 1.3 AP points and 3.0 AP_{75} points, indicating that the larger configuration particularly benefits detection under tighter overlap.

HIT-UAV accuracy. With separate training and testing on HIT-UAV, the second benchmark examines generalization to a broader aerial thermal distribution containing persons and other categories. The reported metrics aggregate all five categories, including DontCare evaluated as Others (Table 7).

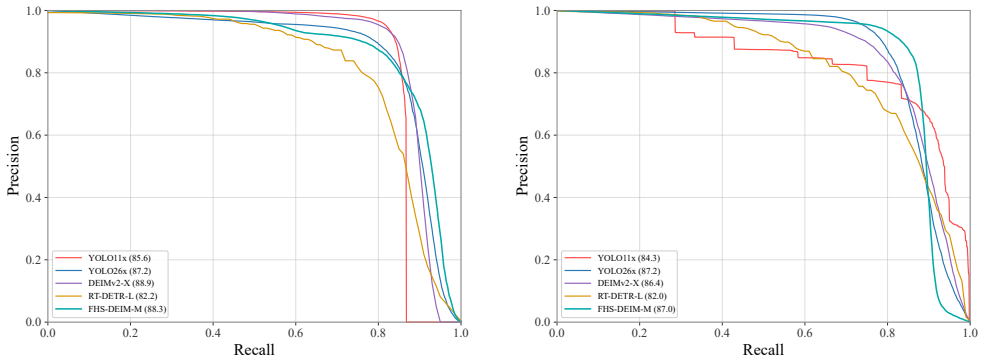
Table 7: Comparative experimental results on the HIT-UAV dataset. Metrics aggregate all five categories: Person, Car, Bicycle, OtherVehicle, and DontCare evaluated as Others.

Model	Epoch	AP (%) ↑	AP_{50} (%) ↑	AP_{75} (%) ↑	AP_S (%) ↑	AP_M (%) ↑	AP_L (%) ↑
Two-stage CNN detectors							
Faster R-CNN (Ren et al. 2015)	24	40.1	70.2	46.2	34.8	47.6	36.9
Single-stage CNN detectors							
SSD300 (Liu et al. 2016)	100	43.5	71.9	49.7	37.2	50.4	38.3
RTMDet-tiny (Lyu et al. 2022)	100	49.8	79.0	56.2	42.4	56.8	40.1
YOLO11m (Jocher and Qiu 2024)	200	54.8	83.7	62.1	47.8	61.3	42.3
YOLO11x (Jocher and Qiu 2024)	200	55.4	84.3	62.9	48.5	62.0	42.9
YOLO26m (Jocher et al. 2026)	200	55.7	84.9	63.3	49.0	62.4	43.2
YOLO26x (Jocher et al. 2026)	200	56.4	87.2	64.1	49.8	63.1	45.2
LIRS-YOLO (Guan et al. 2026)	200	56.5	**88.4**	64.4	50.1	63.4	45.0
End-to-end Transformer detectors							
DEIMv2-M (Huang et al. 2025a)	102	54.4	85.1	62.1	50.3	59.6	47.4
DEIMv2-X (Huang et al. 2025a)	58	56.1	86.4	63.4	51.2	61.2	**48.1**
RT-DETR-L (Zhao et al. 2024)	72	51.5	82.0	58.6	44.0	58.7	41.2
Conditional DETR (Meng et al. 2021)	50	46.5	77.2	52.3	39.4	53.1	36.0
IR-DETR (Hu et al. 2026)	100	57.5	86.5	65.8	50.8	63.8	44.4
KDFF-DETR (Peng et al. 2025b)	72	57.2	87.1	65.2	49.1	62.9	44.8
Ours							
FHS-DEIM-M	102	**58.9**	87.0	**67.2**	**52.2**	**64.5**	45.8
FHS-DEIM-X	58	**60.1**	**89.1**	**69.8**	**54.1**	**65.7**	**48.2**

The gains are particularly pronounced at stricter overlap (Table 7): FHS-DEIM-M improves AP_{75} by 5.1 points over DEIMv2-M, compared with 1.9 points in AP_{50}, while X gains 6.4 and 2.7 points, respectively. Aggregate AP increases by 4.5 points for M and 4.0 for X. This pattern supports improved detection with tighter box agreement across the evaluated categories. Its scale dependence is also informative: M gains 1.9 AP_S and 4.9 AP_M points, but AP_L falls from 47.4 to 45.8; X gains 2.9, 4.5, and 0.1 points in the three size strata. The improvement therefore concentrates on small and medium objects in this benchmark, with a different large-object tradeoff for the two configurations.

FHS-DEIM-X leads every HIT-UAV accuracy column and exceeds the strongest infrared-specific aggregate baseline, IR-DETR, by 2.6 AP points; FHS-DEIM-M exceeds it by 1.4 points. The ordering changes at the looser overlap threshold: LIRS-YOLO reaches 88.4 AP_{50}, above the 87.0 of FHS-DEIM-M, although its aggregate AP and AP_{75} are lower. This contrast explains why AP_{50} alone is insufficient to assess the desired box accuracy. The X configuration raises AP_{50} to 89.1 while retaining the strongest aggregate and scale-stratified results, providing a more balanced accuracy profile in this comparison.

PR curves. Confidence thresholding determines how many candidate detections are retained and how much false-positive burden accompanies higher recall. The curves in Fig. 10 expose this operating tradeoff at IoU = 0.50.

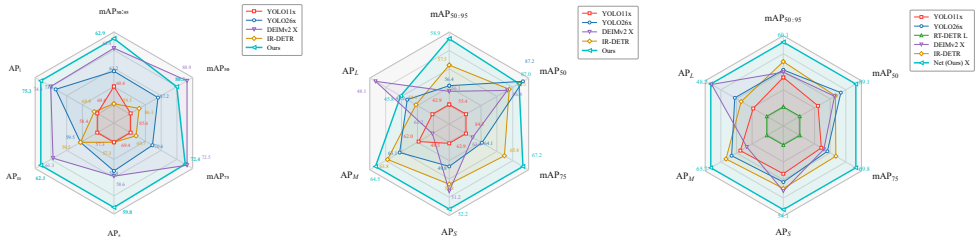


(a) POP (b) HIT-UAV

Fig. 10: Test-split PR curves at IoU = 0.50 on POP and HIT-UAV. Parenthesized legend values are AP_{50} (%); HIT-UAV AP_{50} averages all five evaluated categories, including Others.

On POP, FHS-DEIM-M reaches 88.3 AP_{50}, above YOLO11x, YOLO26x, and RT-DETR-L, but 0.6 points below DEIMv2-X. Its curve crosses the competing curves and retains comparatively higher precision in part of the high-recall tail; the aggregate ordering therefore combines different advantages across recall intervals. On HIT-UAV, FHS-DEIM-M reaches 87.0 AP_{50} and maintains higher precision than DEIMv2-X over much of the intermediate-to-high recall region, while YOLO26x has a slightly higher integrated value of 87.2. The steep final decline in both panels identifies the cost of admitting low-ranked candidates when recall approaches its maximum. For person search, the relevant choice is an operating region that retains weak targets with an acceptable inspection burden; the PR evidence describes that choice more directly than a single AP value.

Multidimensional accuracy profiles. The radar plots in Fig. 11 assess whether an aggregate improvement is accompanied by a weakness in overlap tolerance or target scale.



(a) POP (b) HIT-UAV, M comparison (c) HIT-UAV, X comparison

Fig. 11: Multidimensional performance projections for POP, the FHS-DEIM-M comparison on HIT-UAV, and the HIT-UAV X-scale comparison. Polygon radii encode the metrics listed on each chart.

On POP, FHS-DEIM-M leads the five plotted detectors in AP, AP_S, AP_M, and AP_L, while DEIMv2-X remains 0.6 points higher in AP_{50} and 0.1 points higher in AP_{75}. On HIT-UAV, FHS-DEIM-M improves five metrics over the plotted DEIMv2-X row, but its AP_L is 2.3 points lower, producing a visible contraction along the large-object axis. The X-scale panel has no corresponding accuracy deficit among its plotted methods and reaches the column maxima in Table 7. These profiles distinguish broadly distributed gains from a metric-specific tradeoff. The scientific interpretation rests on

the labeled axes and their normalization; the polygon serves as a visual summary of the tabulated metrics.

Detection across scene structures. LIRS-YOLO (Guan et al. 2026) and IR-DETR (Hu et al. 2026) are selected as infrared-specific CNN and Transformer comparators, emphasizing lightweight small-object feature extraction and edge-aware multiscale fusion, respectively. Together with the direct DEIMv2-M baseline, they provide complementary references for examining target localization and background discrimination in the visual comparisons. Scene-level examples connect the aggregate errors to the spatial structures encountered during person search. The nine POP frames cover woodland, landscaped vegetation and path boundaries, and building-adjacent backgrounds (Figs. 12–14). Each group emphasizes a different source of ambiguity: partial visibility and neighboring people, compact targets beside extended edges, or weak targets across heterogeneous surfaces. The same four detector columns and equal full-frame display scale permit inspection at corresponding coordinates. Target coverage, box extent, and confidence are considered separately because they describe different aspects of a detection.

Woodland scenes. Person-shaped thermal structures interact with branch clutter in the three rows of Fig. 12.

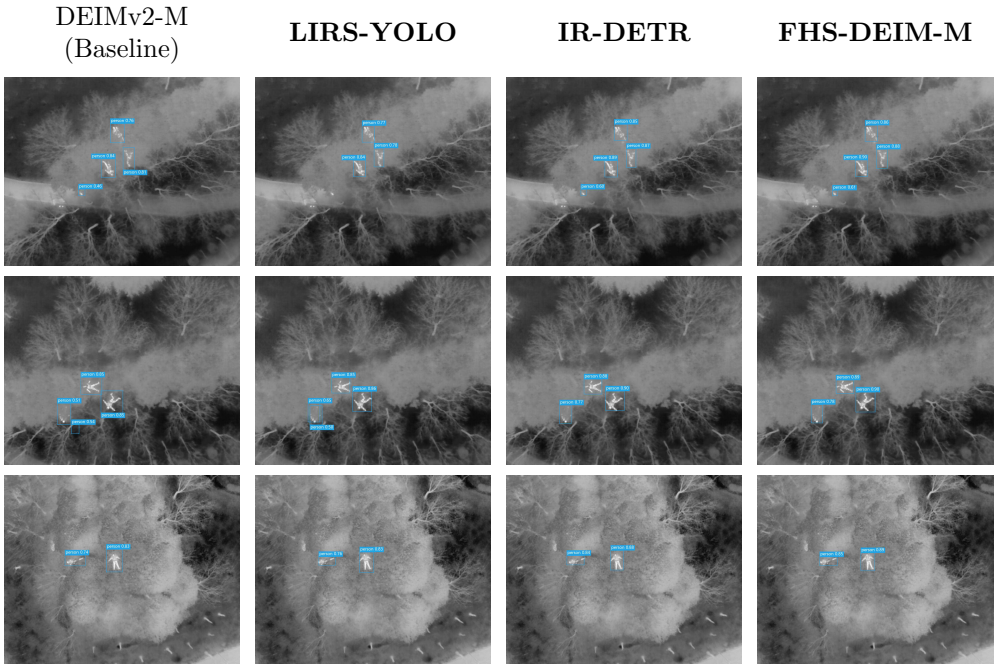


Fig. 12: Woodland scenes with branch occlusion and irregular vegetation texture. The three examples emphasize separated and closely spaced people within structurally cluttered natural backgrounds. Columns show DEIMv2-M (baseline), LIRS-YOLO, IR-DETR, and FHS-DEIM-M at the same full-frame display scale.

The irregular vegetation texture surrounds both separated and closely spaced people, making target coverage and box placement important to the comparison. In the top row of Fig. 12, FHS-DEIM-M localizes all four visible people, including the small lower-left instance absent from the LIRS-YOLO output. The middle row places adjacent detections close to the darker canopy boundary, whereas the bottom row separates a horizontally extended person-shaped region from a more vertically extended one. These arrangements expose two distinct localization demands: preserving coverage when vegetation intersects a person and avoiding a single oversized box spanning

neighboring thermal structures. The visual comparison therefore concerns the association of each box with an individual visible person, rather than simply the presence of a high-contrast region in the scene.

Landscaped vegetation and path boundaries. Compact people must remain distinguishable from the extended canopy and path edges in Fig. 13.

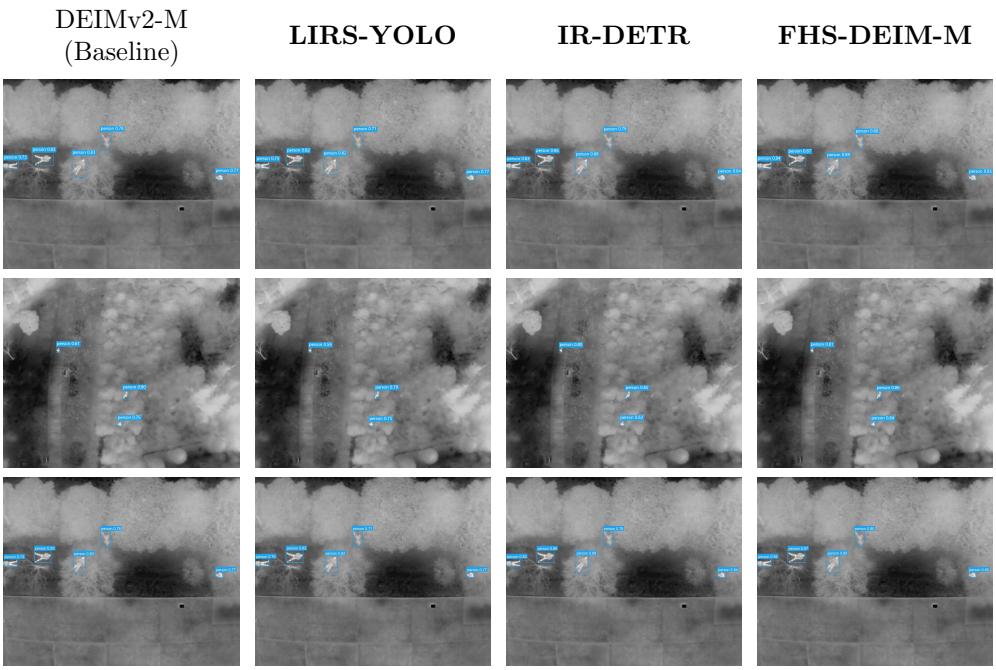


Fig. 13: Landscaped vegetation and path boundaries. Curved or straight artificial boundaries coexist with canopy texture, providing examples of small or dispersed people near competing background edges. Columns show DEIMv2-M (baseline), LIRS-YOLO, IR-DETR, and FHS-DEIM-M at the same full-frame display scale.

The four methods retain corresponding person detections in the displayed frames, so their main visible differences concern the scores assigned to weak structures. In the middle row of Fig. 13, the small upper-left person receives 0.81 from FHS-DEIM-M and 0.80 from IR-DETR, compared with 0.61 from DEIMv2-M and 0.59 from LIRS-YOLO. The nearby elongated boundary occupies a much larger region than this person. In the bottom row, the detections remain separated along the vegetation strip despite the broad smooth surface below it, while the top row tests the same separation beside curved paths and rounded canopy texture. These examples show how weak person evidence is scored amid competing background structure. The scores describe detector confidence; geometric agreement is assessed separately by the overlap-based results.

Building-adjacent scenes. Heterogeneous roof, pavement, and vegetation regions create a different confidence pattern in Fig. 14.

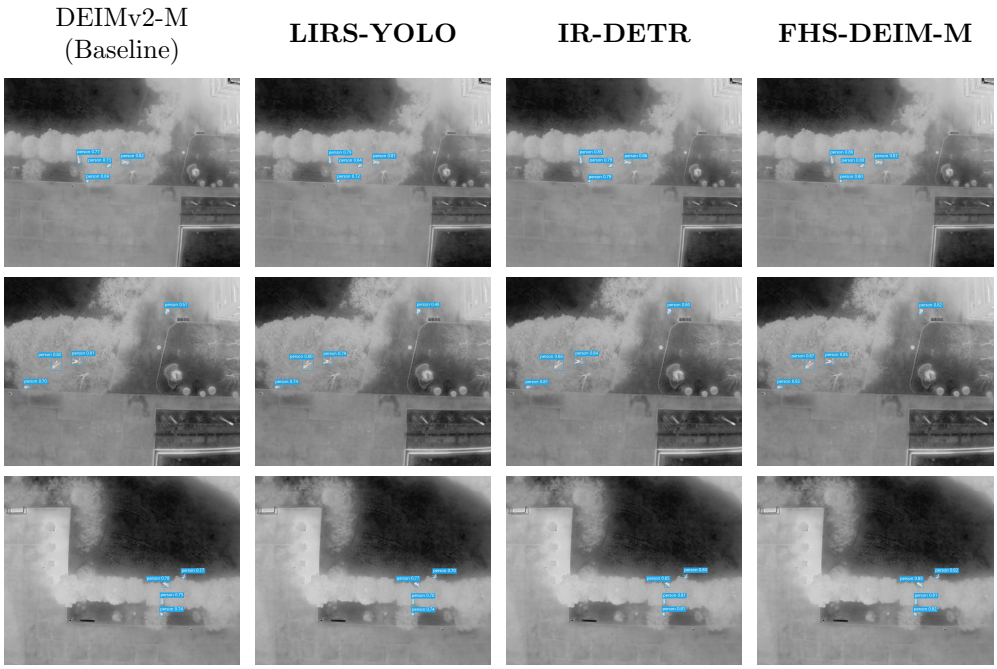


Fig. 14: Building-adjacent scenes with roof, pavement, and vegetation interfaces. The examples expose heterogeneous thermal backgrounds and different projected target sizes around built infrastructure. Columns show DEIMv2-M (baseline), LIRS-YOLO, IR-DETR, and FHS-DEIM-M at the same full-frame display scale.

In the middle row of Fig. 14, the upper-right person lies near the smooth-surface and vegetation interface. FHS-DEIM-M assigns this detection 0.82, close to the 0.80 of IR-DETR and above the 0.57 and 0.48 of DEIMv2-M and LIRS-YOLO. All four outputs contain a box there, making the example informative about weak-target scoring rather than an additional recovered person. The top and middle rows also vary the projected size and orientation of the people, so retaining a detection and fitting its full extent remain separate requirements. In the bottom row, the people occupy a narrow vegetation strip beside a roof whose boundary extends across a much larger spatial range. Together, these cases show why detection must associate local thermal structure with an individual person even when the surrounding built features dominate the image.

Detector-response comparison. To examine the spatial effects of feature enhancement, we visualize the output responses of the respective enhancement modules in each detector and the corresponding feature stage in the baseline. The response maps in Fig. 15 examine spatial selectivity before considering the extent encoded by a detection box.

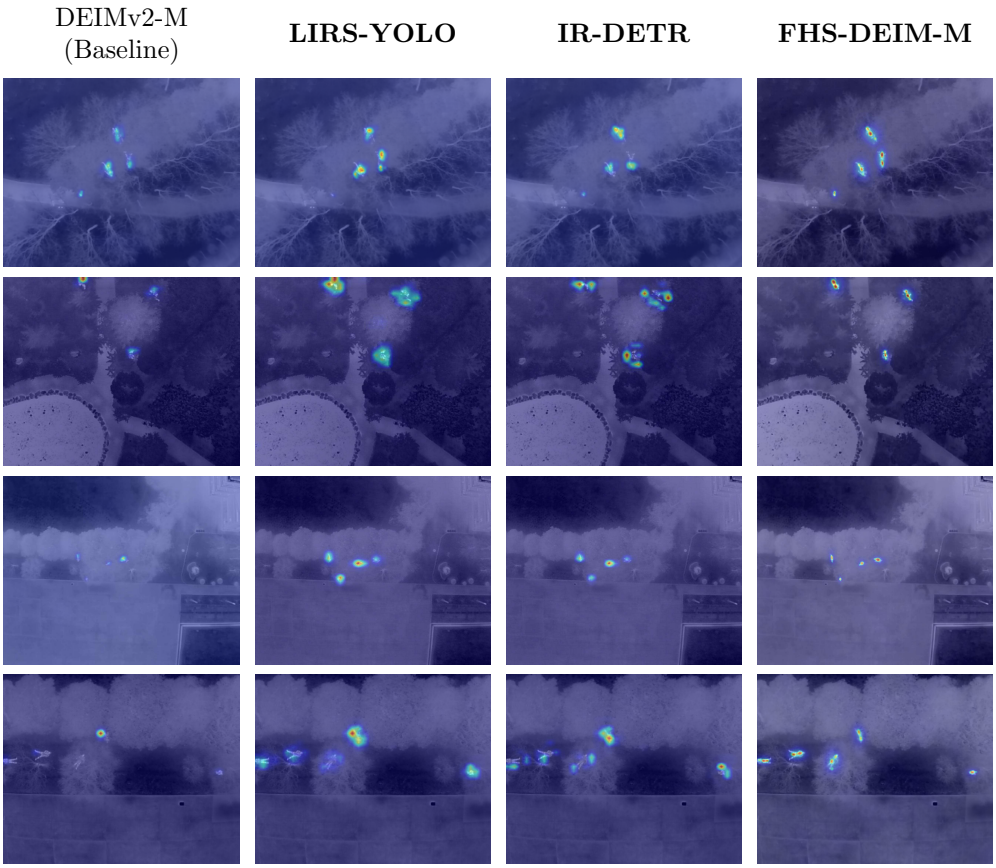


Fig. 15: Detector responses from DEIMv2-M (baseline), LIRS-YOLO, IR-DETR, and FHS-DEIM-M for four POP examples, shown from left to right. From top to bottom, the rows correspond to the top row of Fig. 12, the top row of Fig. 13, the top row of Fig. 14, and the bottom row of Fig. 13.

In the top row of Fig. 15, FHS-DEIM-M produces separated responses around the visible group, whereas DEIMv2-M and LIRS-YOLO concentrate their strongest activations at fewer locations. The FHS-DEIM-M maps in the second, third, and fourth rows likewise retain distinct responses around multiple person-shaped structures across vegetation and built backgrounds. Spatially separated responses are relevant when several weak targets coexist: a dominant activation at one person can otherwise leave the remaining locations poorly represented. The top row can be read together with the individual detections in the woodland example, while the third and fourth rows connect the response pattern to the heterogeneous interfaces discussed above. This evidence supports the spatial interpretation of the AP_S and AP_{75} gains in Table 6; the component ablation examines which design changes contribute to those gains.

Computational efficiency. Measured throughput on the desktop and embedded devices evaluates the processing cost associated with the accuracy results (Table 8).

Table 8: Computational efficiency under the batch-one FP32 POP test protocol. Throughput includes image reading, preprocessing, transfer, inference, decoding, and post-processing.

Model	Params (M) ↓	GFLOPs ↓	FPS (GPU) (frames/s) ↑	FPS (Single device) (frames/s) ↑
LIRS-YOLO (Guan et al. 2026)	**7.8**	**17.7**	**212.4**	**45.9**
KDFF-DETR (Peng et al. 2025b)	46.7	145.6	76.9	6.5
IR-DETR (Hu et al. 2026)	**9.5**	85.5	128.2	13.3
RT-DETR-L (Zhao et al. 2024)	36.0	100.0	137.6	15.7
DEIMv2-M (Huang et al. 2025a)	18.5	**47.1**	**197.3**	**31.3**
Conditional DETR-R50 (Meng et al. 2021)	44.0	90.0	119.5	16.9
Deformable DETR-R50 (Zhu et al. 2021)	40.0	173.0	79.1	9.1
FHS-DEIM-M	22.4	48.2	189.7	30.1

FHS-DEIM-M obtains its accuracy gains with a modest measured throughput reduction relative to DEIMv2-M (Table 8). It contains 22.4 M parameters and requires 48.2 GFLOPs, increases of 3.9 M and 1.1 GFLOPs. Under the complete-pipeline protocol, throughput changes from 197.3 to 189.7 FPS on the RTX 4070 Ti and from 31.3 to 30.1 FPS on the Jetson Orin Nano 8GB. The similar relative reduction on both devices places the feature and supervision gains in the context of practical processing cost. LIRS-YOLO remains lighter and faster, reaching 45.9 FPS on the embedded device, whereas FHS-DEIM-M provides higher aggregate and small-object AP on both benchmarks. The relevant outcome is therefore an accuracy–throughput tradeoff under a measured pipeline, with the appropriate choice depending on the required detection quality and frame-processing budget.

4.3 Ablation Studies

Single-module effects. The three-factor component experiment separates isolated effects from interactions in the M-scale host (Table 9). A checkmark indicates an enabled component; a blank cell indicates that it is disabled.

Table 9: Three-factor M-scale component ablation on the POP test split.

Module[1]			Params (M) ↓	GFLOPs ↓	AP (%) ↑	AP_{50} (%) ↑	AP_S (%) ↑
H	F	L					
			18.4647	**47.0798**	60.6	86.6	57.7
✓			**18.7374**	**46.8898**	61.1	86.6	58.3
	✓		22.1032	48.3675	**61.4**	87.2	**58.8**
		✓	**18.4647**	**47.0798**	60.8	86.8	58.2
✓	✓		22.3760	48.1775	61.2	87.7	58.4
✓		✓	**18.7374**	**46.8898**	60.5	**87.9**	58.4
	✓	✓	22.1032	48.3675	60.3	87.6	58.2
✓	✓	✓	22.3760	48.1775	**62.9**	**88.3**	**59.8**

[1]H: HWDS; F: FIRE-Mamba; L: SA-NWD-MAL.

Evaluated individually against the M-scale all-off ablation host, HWDS raises AP from 60.6 to 61.1 and AP_S from 57.7 to 58.3. FIRE-Mamba reaches 61.4 AP and 58.8

AP_S, giving the largest isolated gains in both metrics, at 0.8 and 1.1 points. SA-NWD-MAL reaches 60.8 AP, 86.8 AP_{50}, and 58.2 AP_S: AP, AP_{50}, and AP_S rise by 0.2, 0.2, and 0.5 points, respectively. It leaves parameters and GFLOPs unchanged because it operates only in the training objective.

Pair and full-model interaction. The component combinations show that the effect of one design depends on the representation and supervision with which it is trained. HWDS with FIRE-Mamba reaches 61.2 AP, HWDS with SA-NWD-MAL reaches 60.5 AP, and FIRE-Mamba with SA-NWD-MAL reaches 60.3 AP, changes of +0.6, -0.1, and -0.3 points from the all-off host. Relative to the sums of the corresponding isolated AP changes, the pairwise departures are -0.7, -0.8, and -1.3 points. Enabling all three yields 62.9 AP, 88.3 AP_{50}, and 59.8 AP_S, gains of 2.3, 1.7, and 2.1 points over the same host. The full-model AP gain is 0.8 points above the sum of the isolated changes. Adding SA-NWD-MAL to both encoder modules raises AP from 61.2 to 62.9, compared with its 0.2-point gain in isolation. The full-configuration gain exceeds the sum of the isolated gains, highlighting the interaction between feature enhancement and size-adaptive supervision.

The resource columns separate inference-graph changes from training-only supervision. HWDS changes the model to 18.7374 M parameters and 46.8898 GFLOPs, whereas FIRE-Mamba changes it to 22.1032 M and 48.3675 GFLOPs. Combining the two encoder modules produces 22.3760 M parameters and 48.1775 GFLOPs. Adding SA-NWD-MAL leaves these values unchanged, so the full model and the two-module encoder configuration have identical graph complexity.

Plug-and-play transfer. Inserting the same module bundle into RT-DETR-L and Conditional DETR tests its transfer across two additional end-to-end detector families (Table 10).

Table 10: Plug-and-play comparison on the POP test split. The same proposed FHS module bundle is inserted into two retained end-to-end detector families.

Configuration	AP (%) ↑	AP_{50} (%) ↑	AP_S (%) ↑
Original end-to-end detectors			
RT-DETR-L (Zhao et al. 2024)	**57.8**	**82.2**	**53.5**
Conditional DETR (Meng et al. 2021)	49.5	75.2	45.6
With the proposed FHS modules			
RT-DETR-L (Zhao et al. 2024)	**58.9**	**85.3**	**56.8**
Conditional DETR (Meng et al. 2021)	53.2	80.7	48.3

Inserting the proposed modules into RT-DETR-L raises AP from 57.8 to 58.9, AP_{50} from 82.2 to 85.3, and AP_S from 53.5 to 56.8 (Table 10). The largest change in this host is the 3.3-point increase for small targets. For Conditional DETR, AP increases from 49.5 to 53.2, AP_{50} from 75.2 to 80.7, and AP_S from 45.6 to 48.3. Its largest change is the 5.5-point increase at an IoU threshold of 0.50.

The positive results in both hosts show that the module bundle can improve two additional end-to-end detector families through their compatible multiscale features and matched normalized boxes. The different metric profiles indicate that the effect depends on the original detector's error profile. This experiment supports transfer of the combined design across the two evaluated hosts, complementing the component ablation that separates its contributions within the M-scale host.

Scan-routing comparison. The scan comparison uses the FIRE-Mamba-only configuration from Table 9, replacing Spiral Scan within FIRE-Mamba with Cross Scan or

Continuous 2D Scan while keeping all other settings unchanged (Table 11). The Spiral Scan row corresponds to the FIRE-Mamba-only result in the component ablation. All three variants have 22.1032 M parameters and 48.3675 GFLOPs.

Table 11: M-scale scan-policy mechanism comparison on the POP test split.

Variant	Params (M) ↓	FLOPs (G) ↓	AP (%) ↑	AP_{50} (%) ↑	AP_S (%) ↑
w/ Cross Scan	22.1032	48.3675	**60.8**	**86.6**	**58.7**
w/ Cont. 2D Scan			59.3	85.5	57.1
Spiral Scan			**61.4**	**87.2**	**58.8**

Spiral Scan reaches 61.4 AP, 87.2 AP_{50}, and 58.8 AP_S, exceeding Cross Scan by 0.6, 0.6, and 0.1 points and Continuous 2D Scan by 2.1, 1.7, and 1.7 points. It achieves the highest accuracy in all three metrics at the same parameter count and GFLOPs. Against Cross Scan, the small-object margin is only 0.1 point, so the advantage is more apparent in aggregate AP and AP_{50}. Against Continuous 2D Scan, the improvement extends more clearly across all three metrics. These results support Spiral routing within the evaluated FIRE-Mamba setting without an increase in model size or arithmetic cost.

Frequency-downsampling mechanism comparison. The reduction-operator experiment evaluates the balance between preserving detection evidence and reducing feature resolution (Table 12). The HWDownSample row corresponds to the HWDS-only configuration in Table 9, with 18.7374 M parameters and 46.8898 GFLOPs.

Table 12: Frequency-downsampling mechanism comparison on the POP test split.

Variant	Params (M) ↓	FLOPs (G) ↓	AP (%) ↑	AP_{50} (%) ↑	AP_S (%) ↑
w/ FFT DS.	**22.0986**	**47.5708**	**61.0**	85.5	57.5
w/ Haar-DWT.	22.4918	48.3736	60.9	**86.4**	**57.6**
HWDownSample	**18.7374**	**46.8898**	**61.1**	**86.6**	**58.3**

HWDownSample (HWDS) reaches 61.1 AP, 86.6 AP_{50}, and 58.3 AP_S. Compared with FFT downsampling, it gains 0.1 AP, 1.1 AP_{50}, and 0.8 AP_S points while reducing parameters by 3.3612 M and computation by 0.6810 GFLOPs. Compared with Haar-DWT downsampling, it gains 0.2, 0.2, and 0.7 points while reducing parameters by 3.7544 M and computation by 1.4838 GFLOPs. HWDS thus achieves the highest accuracy with the fewest parameters and GFLOPs among these variants. The larger margins in AP_S than in aggregate AP are consistent with a benefit for small targets in this comparison. The Haar-DWT comparison shows the benefit of the HWDS enhancement and gating design over direct wavelet downsampling. The aligned responses in Figure 7 provide complementary spatial observations.

5 Discussion

5.1 Cross-dataset behavior and component interaction

Across the two separately trained and evaluated benchmarks, the consistent gains in aggregate AP, AP_S, and AP_{75} connect the proposed design to two demands of aerial

thermal detection: retaining compact targets and predicting boxes that remain correct under stricter overlap. Their magnitude and scale distribution differ. POP improves in all three size strata, whereas HIT-UAV shows an M-scale large-object decrease and a nearly unchanged X-scale large-object result. This pattern makes target scale part of the interpretation of the improvement. HIT-UAV averages all five evaluated categories, including persons and Others, so its aggregate result broadens the evidence for aerial thermal detection beyond the person-centered POP setting.

The component evidence shows that the effects are configuration-dependent. FIRE-Mamba gives the largest isolated gains in both AP and AP_S, while HWDS produces smaller positive changes. Their two-module combination falls below FIRE-Mamba alone in both metrics. SA-NWD-MAL raises aggregate AP by 0.2 points in isolation. Adding it to HWDS alone or FIRE-Mamba alone reduces AP by 0.6 or 1.1 points, respectively, whereas adding it to both encoder modules raises AP by 1.7 points. The complete configuration exceeds the all-off host by 2.3 AP and 2.1 AP_S points. The strongest result comes from jointly combining contextual representation, cross-scale detail transfer, and size-adaptive supervision. The mechanism comparisons further identify Spiral Scan and HWDS as the most accurate choices among the evaluated scanning and downsampling operators.

5.2 Mamba representation and directional detail

FIRE-Mamba changes the information and routing supplied to Mamba while retaining its selective state update. Approximation and aggregated detail share the FA and Mamba parameters, and the four Spiral branches accumulate context in different spatial orders. Inverse routing aligns their outputs before averaging and reconstruction. At the same parameter count and GFLOPs, Spiral Scan improves all three reported accuracy metrics over Cross Scan and Continuous 2D Scan. This supports the relevance of spatial routing in the evaluated FIRE-Mamba setting. The 0.1-point AP_S margin over Cross Scan remains small, whereas the gains over Continuous 2D Scan are more pronounced across all three metrics.

The frequency representation also determines which distinctions are retained. Averaging directional detail reduces differences between horizontal, vertical, and diagonal structures; the processed detail tensor is subsequently reused in the three inverse-Haar synthesis slots. HWDS restores three detail outputs from a shared high-frequency bottleneck for scale transfer. Its Haar and Laplacian responses can also react to strong background edges, which motivates learned enhancement and gating before subband reduction. The scene-grouped detections and response maps make this competition visible across branches, path boundaries, and building-adjacent backgrounds. They complement the quantitative comparisons by showing where person-related responses occur within each scene.

Spatial alignment and scale-aware supervision complement these feature operations. Replicate padding and inverse-synthesis cropping preserve the original feature extent. NWD supplies a continuous distance for small boxes, with C_{nwd} and τ_s defined in the normalized box coordinate system.

5.3 Efficiency and application scope

The efficiency result places the frequency-aware Mamba design within a complete detection pipeline. Linear sequence-length complexity describes the selective recurrence, whereas measured throughput also includes the frequency operations, feature fusion, decoding, and data movement used by the detector. The modest throughput reduction relative to DEIMv2-M shows that the improved accuracy is retained at a similar processing rate on both tested devices. At the same time, the faster LIRS-YOLO result makes the operational choice explicit: applications with a tighter frame-processing budget may favor its lower cost, while FHS-DEIM-M offers higher aggregate and small-object accuracy under the reported protocols.

For UAV infrared SAR, the relevant output is a detected person with a bounding box that can support subsequent inspection. POP directly examines people in open and vegetation-occluded wilderness scenes, while HIT-UAV broadens the evaluation to a different aerial thermal distribution and multiple categories. The results support improved bounding-box detection under these benchmark conditions. The combination of localization accuracy, small-target performance, and measured embedded throughput supports the use of FHS-DEIM for locating people in UAV infrared imagery.

6 Conclusion

FHS-DEIM adapts Mamba to infrared UAV detection through a two-band Haar representation, shared FA, four-route Spiral propagation, and inverse reconstruction. FIRE-Mamba returns this contextual modulation to the original spatial grid. HWDS then supports cross-scale detail transfer by processing four decimated subbands and restoring three detail outputs from a shared high-frequency bottleneck before pyramid reduction. SA-NWD-MAL complements these encoder changes with size-dependent classification-quality targets and continuous geometric supervision for small boxes.

On POP, FHS-DEIM-M and FHS-DEIM-X reach 62.9 and 64.2 AP, improving on their corresponding DEIMv2 configurations by 2.3 and 1.8 points. On HIT-UAV, aggregate AP over all five evaluated categories reaches 58.9 and 60.1, with gains of 4.5 and 4.0 points. The component experiment identifies FIRE-Mamba as the strongest isolated contributor to AP and AP_S and shows a larger joint improvement when contextual propagation, frequency downsampling, and classification and box supervision are combined. The scan-policy and downsampling comparisons further connect performance to the particular spatial routes and scale-reduction operators used by the detector.

The M configuration achieves 189.7 FPS on the RTX 4070 Ti and 30.1 FPS on the Jetson Orin Nano 8GB under the stated complete-pipeline FP32 protocol. Together, the accuracy results, scene-based visual comparisons, and measured throughput establish FHS-DEIM as an effective Mamba-based detector for the evaluated UAV infrared detection tasks. The results also make the scale-dependent behavior and the cost of frequency-aware state propagation explicit, providing a basis for further study of reliable person detection under changing target size and background structure.

References

Cai Z, Liu S, Wang G, et al (2024) Align-DETR: Enhancing End-to-end Object Detection with Aligned Loss. In: 35th British Machine Vision Conference (BMVC). BMVA, URL https://bmvc2024.org/proceedings/211/, paper 211

Chen C, Zhang T, Ma C, et al (2026a) L-DC-DETR: A lite distance-content aware DETR for gas pipeline defect detection. Signal, Image and Video Processing 20:16. https://doi.org/10.1007/s11760-025-05070-y

Chen T, Hu X, Wang J (2026b) LIR-Mamba: consolidating YOLO and selective SSM with global-local scanning for robust infrared small target detection under laser interference. Scientific Reports 16:23713. https://doi.org/10.1038/s41598-026-64040-7

Choi CY, Lee SW (2026) NeXt-DETR: A scalable and efficient transformer-based detector for resource-constrained systems. Information Sciences 731:122913. https://doi.org/10.1016/j.ins.2025.122913

Chu Y, Wang J, Ma L, et al (2025) LMSFA-YOLO: A lightweight target detection network in Remote sensing images based on Multiscale feature fusion. Journal of King Saud University Computer and Information Sciences 37:63. https://doi.org/10.1007/s44443-025-00069-4

Dai Y, Wu Y, Zhou F, et al (2021) Asymmetric contextual modulation for infrared small target detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 950–959, https://doi.org/10.1109/wacv48630.2021.00099

Finder S, Amoyal R, Treister E, et al (2024) Wavelet convolutions for large receptive fields. In: Computer Vision – ECCV, pp 363–380, https://doi.org/10.1007/978-3-031-72949-2_21

Ge J, Huang B, Lv J, et al (2025) Improving real-time end-to-end object detection with cross-feature interaction. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, pp 341–346, https://doi.org/10.1109/SMC58881.2025.11343447

Gotovac S, Zelenika D, Marušić v, et al (2020) Visual-Based Person Detection for Search-and-Rescue with UAS: Humans vs. Machine Learning Algorithm. Remote Sensing 12(20):3295. https://doi.org/10.3390/rs12203295

Gu A, Dao T (2024) Mamba: Linear-time sequence modeling with selective state spaces. In: Conference on Language Modeling, URL https://openreview.net/forum?id=tEYskw1VY2

Guan J, Yao X, Zhao Q, et al (2026) LIRS-YOLO: A lightweight YOLO-based method for infrared features and small object detection. IEEE Transactions on Geoscience and Remote Sensing 64:4703914. https://doi.org/10.1109/TGRS.2026.3690625

Guo C, Luo P, Ma Z, et al (2025) An efficient reparameterized small object detection transformer for thermal infrared images. Scientific Reports 15:44599. https://doi.org/10.1038/s41598-025-28330-w

Guo H, Li J, Dai T, et al (2024) MambaIR: A simple baseline for image restoration with state-space model. In: Computer Vision – ECCV, pp 222–241, https://doi.org/10.1007/978-3-031-72649-1_13

Hu J, Zhao M, Liu B, et al (2026) IR-DETR: An infrared small object detector combining edge-aware mechanism and multiscale feature fusion. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 19:7264–7279. https://doi.org/10.1109/JSTARS.2026.3652368

Huang S, Hou Y, Liu L, et al (2025a) Real-time object detection meets DINOv3. arXiv preprint arXiv:2509.20787, https://doi.org/10.48550/arXiv.2509.20787

Huang S, Lu Z, Cun X, et al (2025b) DEIM: DETR with improved matching for fast convergence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 15162–15171, https://doi.org/10.1109/cvpr52734.2025.01412

Jocher G, Qiu J (2024) Ultralytics YOLO11. Software, version 11.0.0, URL https://github.com/ultralytics/ultralytics, accessed 7 Sep 2026

Jocher G, Qiu J, Liu M, et al (2026) Ultralytics YOLO26: Unified real-time end-to-end vision models. arXiv preprint arXiv:2606.03748, https://doi.org/10.48550/arXiv.2606.03748

Khadam U, Davidsson P, Spalazzese R (2025) A systematic literature review on AI in IoT systems: Tasks, applications, and deployment. Internet of Things 34:101779. https://doi.org/10.1016/j.iot.2025.101779

Kubo S, Yamada K, Yoshida H (2026) Automated Victim Detection from UAV Thermal Infrared Imagery for Nighttime Search and Rescue Using Multi-Pose Ground Camera Data. Remote Sensing 18(14):2279. https://doi.org/10.3390/rs18142279

Li F, Zeng A, Liu S, et al (2023) Lite DETR: An Interleaved Multi-Scale Encoder for Efficient DETR. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 18558–18567, https://doi.org/10.1109/CVPR52729.2023.01780

Li J, Sun F, Cheng L (2026) Geometry-aligned dual-branch detection with directional crack-aware attention for EL imaging of photovoltaic cells. Journal of King Saud University Computer and Information Sciences 38:182. https://doi.org/10.1007/s44443-026-00566-0

Li Y, Wang L, Chen S (2025) Frequency–spatial interaction reinforcement paradigm for infrared small target detection. IEEE Transactions on Geoscience and Remote Sensing 63:5008915. https://doi.org/10.1109/TGRS.2025.3624449

Lin TY, Maire M, Belongie S, et al (2014) Microsoft COCO: Common objects in context. In: Computer Vision – ECCV, pp 740–755, https://doi.org/10.1007/978-3-319-10602-1_48

Liu L, Guo W, Huang S, et al (2024a) From COCO to COCO-FP: A Deep Dive into Background False Positives for COCO Detectors. arXiv preprint arXiv:2409.07907, https://doi.org/10.48550/arXiv.2409.07907

Liu S, Wan Y, Li K, et al (2026a) A brain-inspired low-light and infrared multimodal remote sensing detection network for nighttime personnel search and rescue. Information Fusion 136:104518. https://doi.org/10.1016/j.inffus.2026.104518

Liu W, Anguelov D, Erhan D, et al (2016) SSD: Single shot multibox detector. In: Computer Vision – ECCV, pp 21–37, https://doi.org/10.1007/978-3-319-46448-0_2

Liu Y, Tian Y, Zhao Y, et al (2024b) VMamba: Visual state space model. In: Advances in Neural Information Processing Systems, pp 103031–103063, https://doi.org/10.52202/079017-3273

Liu Z, Dong L, Li C, et al (2026b) Maritime infrared dim and small targets detection via multi-scale enhancement network and semi-ellipsoid loss. Infrared Physics & Technology 155:106527. https://doi.org/10.1016/j.infrared.2026.106527

Lyu C, Zhang W, Huang H, et al (2022) RTMDet: An empirical study of designing real-time object detectors. arXiv preprint arXiv:2212.07784, https://doi.org/10.48550/arXiv.2212.07784

Mallat S (1989) A theory for multiresolution signal decomposition: The wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 11(7):674–693. https://doi.org/10.1109/34.192463

Meng D, Chen X, Fan Z, et al (2021) Conditional DETR for fast training convergence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 3651–3660, https://doi.org/10.1109/iccv48922.2021.00363

Nan Z, Li X, Dai J, et al (2025) MI-DETR: An object detection model with multi-time inquiries mechanism. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4703–4712, https://doi.org/10.1109/CVPR52734.2025.00443

Peng Y, Li H, Wu P, et al (2025a) D-FINE: Redefine Regression Task of DETRs as Fine-grained Distribution Refinement. In: International Conference on Learning Representations (ICLR), URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6cf58a87e3097e7d1f9be3e8693a93de-Abstract-Conference.html

Peng Y, Tao Z, Zhang J, et al (2025b) KDFF-DETR: Knowledge-driven feature fusion DETR for spaceborne infrared ship detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 18:17638–17653. https://doi.org/10.1109/JSTARS.2025.3586955

Ren S, He K, Girshick R, et al (2015) Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html

Song Z, Yan Y, Cao Y, et al (2024) An infrared dataset for partially occluded person detection in complex environment for search and rescue. OSF, https://doi.org/10.17605/OSF.IO/KMCVA

Song Z, Yan Y, Cao Y, et al (2025) An infrared dataset for partially occluded person detection in complex environment for search and rescue. Scientific Data 12:300. https://doi.org/10.1038/s41597-025-04600-0

Suo J, Wang T, Zhang X, et al (2023) HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection. Scientific Data 10:227. https://doi.org/10.1038/s41597-023-02066-6

Szrek J, Zimroz R, Wodecki J, et al (2021) Application of the Infrared Thermography and Unmanned Ground Vehicle for Rescue Action Support in Underground Mine—The AMICOS Project. Remote Sensing 13(1):69. https://doi.org/10.3390/rs13010069

Wang J, Xu C, Yang W, et al (2021) A normalized Gaussian Wasserstein distance for tiny object detection. arXiv preprint arXiv:2110.13389, https://doi.org/10.48550/arXiv.2110.13389

Wang J, Yang H, Zhang Z, et al (2025) Detection of moving small targets in infrared images for urban traffic monitoring. Internet of Things 33:101673. https://doi.org/10.1016/j.iot.2025.101673

Woesle C, Fischer-Brandies L, Buettner R (2026) A Robust Occlusion-Aware Deep Learning Architecture for Thermal Aerial Person Classification in Search-and-Rescue Missions. IEEE Access 14:10923–10938. https://doi.org/10.1109/ACCESS.2026.3655411

Wu F, Zhang T, Li L, et al (2024) RPCANet: Deep unfolding RPCA based infrared small target detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 4809–4818, https://doi.org/10.1109/wacv57701.2024.00474

Wu S, Yang H, Liao L, et al (2025) Dynamic small object feature enhancement and detection for remote sensing images. Scientific Reports 15:37225. https://doi.org/10.1038/s41598-025-21134-y

Xu C, Wang J, Yang W, et al (2022) Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark. ISPRS Journal of Photogrammetry and Remote Sensing 190:79–93. https://doi.org/10.1016/j.isprsjprs.2022.06.002

Yang L, Wu X, Wang X, et al (2026) MMAF-YOLO: A UAV object detection network based on multi-scale mutual aggregation and fine-grained feature modulation. Journal of King Saud University Computer and Information Sciences 38:354. https://doi.org/10.1007/s44443-026-00787-3

Yin H, Zhu Z, Wang H (2025) SED-DETR: A Scale-Enhanced Deformable Detection Transformer for Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing 63:1–12. https://doi.org/10.1109/TGRS.2025.3571033

Zhang L (2026a) ARF-YOLO: Attention-guided adaptive resolution-aware feature learning for UAV remote sensing object detection. Journal of King Saud University Computer and Information Sciences 38:652. https://doi.org/10.1007/s44443-026-00892-3

Zhang L (2026b) CSM-DETR: construction site monitoring via Mamba-Enhanced detection transformer for UAV aerial imagery. Journal of King Saud University Computer and Information Sciences 38:210. https://doi.org/10.1007/s44443-026-00628-3

Zhang L, Yang X, Shen J (2021) Frequency variability feature for life signs detection and localization in natural disasters. Remote Sensing 13(4):796. https://doi.org/10.3390/rs13040796

Zhang M, Zhang R, Yang Y, et al (2022) ISNet: Shape matters for infrared small target detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 877–886, https://doi.org/10.1109/cvpr52688.2022.00095

Zhao H, Chen K, Zhang Y, et al (2026) RSRT-DETR: Hierarchical Polar Attention and Multiscale Hypergraph Networks for Dense Remote Sensing Small-Object Detection. IEEE Transactions on Geoscience and Remote Sensing 64:1–21. https://doi.org/10.1109/TGRS.2026.3677535

Zhao W, Deng X, Gao F, et al (2025) Position-DETR: Step-by-Step Position-Guided Small Object Detection in Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing 63:1–14. https://doi.org/10.1109/TGRS.2025.3603636

Zhao Y, Lv W, Xu S, et al (2024) DETRs beat YOLOs on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 16965–16974, https://doi.org/10.1109/cvpr52733.2024.01605

Zhi Y, Zhao J, Song C, et al (2025) SAFF-DETR: An End-to-End Object Detection Network for Remote Sensing Images With Targets of Varying Sizes Based on Scale Adaptation and Frequency Fusion. IEEE Transactions on Geoscience and Remote Sensing 63:5646415. https://doi.org/10.1109/TGRS.2025.3618293

Zhu X, Su W, Lu L, et al (2021) Deformable DETR: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=gZ9hCDWe6ke