A Frequency-Aware and Entropy-Guided Transformer for Real-Time Tiny Object Detection in Remote Sensing

Dongdong An[1], Jiezhi Bao[1], Shuo Zhang[2*], Wenbing Tang[3], Qin Zhao[1]

[1]College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, China

[2]School of Electronic Information Engineering, Shanghai DianJi University, Shanghai, China

[3]Northwest A&F University, Yangling, China

Email: {andongdong, 1000569275}@shnu.edu.cn, szhang@sdju.edu.cn

***Abstract*—Real-time aerial object detection remains challenging due to tiny targets and cluttered backgrounds. In remote-sensing imagery, small objects often occupy only a few pixels, and their weak high-frequency cues are easily overwhelmed by dominant low-frequency background components during feature extraction and fusion. Moreover, common feature-pyramid fusion relies on upsampling, which can introduce spatial misalignment and degrade geometric localization. In densely populated scenes, classification confidence and box regression quality may jointly deteriorate, causing valid targets to be suppressed without an explicit spatial prior. We propose FRE-Det (Frequency-Residual and Entropy Detector), a task-oriented architecture that integrates time–frequency analysis and information-theoretic guidance. First, we design a Dual Frequency Enhanced Hybrid Encoder that performs frequency filtering via the Fast Fourier Transform (FFT) to isolate and enhance weak target responses against background interference. Second, we introduce a Wavelet Reconstruction Block (WRB) to replace naive upsampling with wavelet-guided reconstruction, improving spatial consistency during multi-scale interaction. Third, to mitigate missed detections in dense regions, we propose Local Spatial Entropy-guided Uncertainty Minimization (LSE-UM), which uses local information density as a spatial prior to recover true targets based on signal presence. Extensive experiments on VisDrone2019 demonstrate state-of-the-art performance, achieving $AP_{50} = 49.3\%$ and $AP_S = 19.9\%$ while running at 87 FPS.**

***Index Terms*—Remote Sensing, Tiny Object Detection, Real-Time Detection, Frequency Domain Learning**

I. INTRODUCTION

Object detection constitutes a fundamental pillar of computer vision, serving as the perceptual engine for modern earth observation systems [1]. In critical applications ranging from maritime surveillance to disaster response, the imperative to rapidly and accurately localize targets is paramount [2], [3]. While generic deep learning detectors have achieved remarkable success, their direct transfer to remote sensing remains suboptimal due to the unique challenges of aerial imagery: expansive fields of view, extreme background clutter, and the predominance of tiny objects—often fewer than 32×32 pixels—possessing inherently weak feature representations [4], [5]. Beyond resolution constraints, satisfying the stringent latency requirements of embedded UAV platforms has led to the rise of Real-Time Object Detection architectures [6], such as the YOLO series and the recent RT-DETR [7]. However, directly deploying these models for aerial tiny object detection entails significant challenges primarily due to the severe attenuation of small target features within standard architectures.

The primary limitation of current real-time models lies in the inefficacy of multi-scale feature interaction [8]. Tiny targets inherently possess faint feature responses that are easily overwhelmed by the high-spectral-energy noise of complex backgrounds—a phenomenon we define as *Feature Submersion*. Furthermore, standard upsampling operations in feature fusion pyramids inevitably introduce aliasing and sub-pixel deviations. While negligible for large objects, this *Spatial Misalignment* destroys the geometric integrity required for precise tiny object localization. In addition to feature-level deficiencies, Transformer-based query initialization presents a secondary bottleneck in dense scenarios [9]. Standard strategies select object queries by solely evaluating task-specific predictions; however, in dense clusters, intrinsic feature ambiguity triggers a simultaneous degradation of classification and localization confidence. Lacking an explicit spatial prior, these mechanisms discard valid targets as background, ignoring the physical reality of their high local information density.

To address these limitations, we propose **FRE-Det** (Frequency-Residual and Entropy Detector), a specialized real-time architecture that transcends the spatial-only paradigm by incorporating Frequency Domain Analysis and Information Theory into the detection pipeline. First, we design a Dual-Frequency Enhanced Hybrid Encoder to resolve the interaction dilemma. This module utilizes global filtering based on the Fast Fourier Transform (FFT) to suppress background noise in the frequency domain, coupled with generative upsampling via the Inverse Discrete Wavelet Transform (IDWT) to robustly maintain high-frequency structural details of tiny objects. Second, to rectify selection failures in dense clusters, we introduce a Local Spatial Entropy-based Uncertainty Minimization (LSE-UM) strategy. By leveraging local information

*Corresponding author. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62302308.

density as a spatial prior, we align query selection with the physical presence of targets, effectively compensating for the semantic ambiguity inherent in weak-feature scenarios. The main contributions of this work are summarized as follows:

- We identify and analyze the specific mechanisms of feature submersion and spatial misalignment, providing new insights into small object feature degradation in real-time aerial detection.
- We propose a frequency-domain enhanced hybrid encoder that leverages FFT and IDWT to reconstruct high-frequency signatures and preserve tiny object details against complex background clutter.
- We develop an entropy-based query selection strategy (LSE-UM) to minimize uncertainty and improve localization priors in dense, spatially-ambiguous object scenarios.

II. Related Work

A. Real-Time Architectures

Real-time object detection is pivotal for latency-sensitive applications such as autonomous UAV surveillance. The field has been long dominated by the YOLO series, which evolved from anchor-based to anchor-free paradigms to achieve a commendable balance between speed and accuracy [10]. However, these CNN-based detectors rely on Non-Maximum Suppression (NMS), which introduces latency instability in dense scenarios [11]. Recently, RT-DETR [7] leveraged Transformer architectures to eliminate NMS via bipartite matching, yet its query selection remains heavily dependent on classification and regression quality.

This dependency is particularly problematic for tiny object detection, where extreme pixel scarcity and feature ambiguity lead to simultaneous degradation of semantic and localization metrics [8], [12]. While research into probabilistic metrics like KLDet [13] and GA strategies has stabilized training objectives, a fundamental bottleneck persists in feature representation. Standard encoders, including Feature Pyramid Networks (FPN), aggregate context based on activation magnitude, causing the weak signatures of tiny targets to be overwhelmed by high-energy background noise—a phenomenon we term feature submersion. Unlike generic detectors that discard uncertain predictions, our FRE-Det incorporates spatial information density to recover these valid targets based on physical signal existence.

B. Frequency-Domain Methods

Frequency analysis, a cornerstone of signal processing [14], has recently been integrated into deep learning to complement traditional spatial operations. Advanced mechanisms now utilize frequency-dynamic modulation for attention and adaptive filters for efficient token mixing. In low-level vision, Wavelet Transforms are widely adopted for image restoration [15] and domain adaptation [16] due to their unique ability to decouple style from content and preserve fine-grained textures.

Despite these advancements, frequency analysis is rarely optimized for the specific constraints of aerial tiny object detection. A critical oversight in standard architectures is that conventional upsampling acts as a lossy low-pass filter [17], smoothing out the sharp edges and high-frequency transitions that characterize tiny targets. To address this, FRE-Det incorporates the invertible Discrete Wavelet Transform (DWT) into the encoder. Unlike standard pooling or interpolation that discards information, DWT decomposes feature maps into low-frequency approximations and high-frequency details. By explicitly utilizing these high-frequency components, we ensure that faint object signatures are preserved and propagated through the network rather than being eroded by dominant background clutter.

III. Method

This section delineates the FRE-Det (Frequency-Residual and Entropy Detector) architecture. As illustrated in Fig. 1, FRE-Det adopts an end-to-end paradigm comprising a backbone, a *Dual-Frequency Enhanced Hybrid Encoder*, and a Transformer decoder. To counteract the *submersion* of weak target signatures, our encoder applies differentiated spectral processing to multi-scale features $\{C_3, C_4, C_5\}$: a Fourier-filtering branch preserves shallow high-frequency details, while an inverse wavelet-based branch reconstructs deep semantic features typically lost during upsampling (Sec. III-A). Furthermore, the *Local Spatial Entropy-guided Uncertainty Minimization* (LSE-UM) strategy rectifies localization failures in dense scenarios by initializing queries based on local *information density* rather than purely semantic confidence (Sec. III-B). Finally, the decoder iteratively refines these queries into the final predictions.

A. Dual Frequency-domain Enhanced Hybrid Encoder

The proposed encoder adopts a frequency-aware topology to ensure signal fidelity during multi-scale feature interaction. Specifically, we decouple feature processing into two specialized spectral streams: a *filtering branch* utilizing Fourier transforms to suppress shallow-layer noise, and a *generation branch* employing wavelet reconstruction to recover deep semantic resolution. This dual-path mechanism preserves high-frequency structural signatures while mitigating the feature submersion typical of spatial-only architectures.

1) Shallow Feature Enhancement via HFE: To amplify faint target responses in noisy shallow layers (e.g., C_3), we introduce the High-Frequency Feature Enhancement (HFE) module. By transforming the input $F_{in} \in \mathbb{R}^{H \times W \times C}$ into the spectral domain via 2D Fast Fourier Transform (FFT), we obtain the spectrum $\mathcal{F}(u, v)$. To mitigate Gibbs artifacts, we design a learnable Gaussian Soft-Mask $M(u, v)$ that adaptively suppresses low-frequency background clutter while retaining high-frequency target impulses:

$$\begin{aligned} D(u,v) &= \sqrt{(u-u_0)^2 + (v-v_0)^2} \\ M(u,v) &= 1 - \exp\left(-\frac{D^2(u,v)}{2\sigma^2}\right) \end{aligned} \tag{1}$$

The filtered high-frequency response F_{high}, obtained through Inverse FFT (IFFT), is then modulated via dual-path context weighting. The *Channel Path* captures global spectral statistics

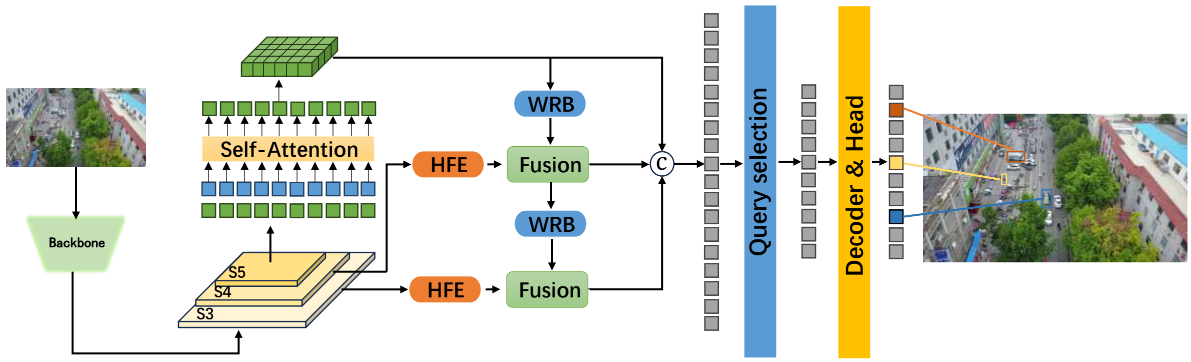

Fig. 1. The overall architecture of FRE-Det. The symbol ⓒ denotes concatenation.

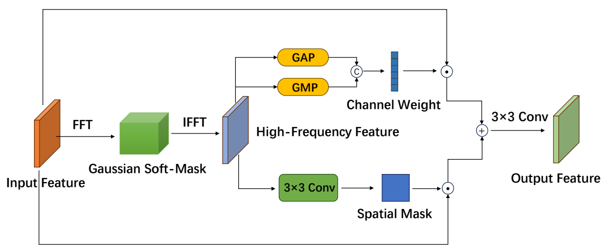

Fig. 2. Architecture of the High-Frequency Feature Enhancement (HFE) module. The symbols $\odot$, $\oplus$, and ⓒ denote the Hadamard product, element-wise addition, and concatenation operation, respectively.

through pooling and an MLP to yield W_{ch}, while the *Spatial Path* employs local receptive fields to distinguish valid target singularities from stochastic noise, generating W_{sp}. The enhanced output is:

$$F_{out} = \text{Conv}_{3\times3}\left((F_{in} \odot W_{ch}) \oplus (F_{in} \odot W_{sp})\right) \quad (2)$$

2) Generative Deep Feature Reconstruction via WRB: To rectify spatial misalignment and aliasing in deep semantic features (e.g., C_5), the Wavelet Reconstruction Block (WRB) reformulates resolution recovery as a generative Inverse Discrete Wavelet Transform (IDWT) task. Unlike interpolation, WRB models the input F_{deep} as the low-frequency approximation (F_{LL}) and explicitly predicts the missing high-frequency sub-bands—Horizontal (F_{LH}), Vertical (F_{HL}), and Diagonal (F_{HH})—to reconstruct sharp boundaries. Specifically, a 3×3 grouped convolution expands the channels to $4C$ to represent these sub-bands. The high-resolution feature $F_{up} \in \mathbb{R}^{2H\times 2W\times C}$ is reconstructed for each 2×2 block (i, j) as:

$$\begin{aligned}
F_{up}(2i, 2j) &= (F_{LL} - F_{LH} - F_{HL} + F_{HH})_{i,j} \\
F_{up}(2i, 2j+1) &= (F_{LL} - F_{LH} + F_{HL} - F_{HH})_{i,j} \\
F_{up}(2i+1, 2j) &= (F_{LL} + F_{LH} - F_{HL} - F_{HH})_{i,j} \\
F_{up}(2i+1, 2j+1) &= (F_{LL} + F_{LH} + F_{HL} + F_{HH})_{i,j}
\end{aligned} \quad (3)$$

This precise reconstruction preserves geometric integrity for localization, bridging deep semantics and high-resolution spatial alignment.

B. Local Spatial Entropy-guided Uncertainty Minimization Query Selection

Standard query initialization in Transformer-based detectors predominantly relies on task-specific confidence, which often falters in dense aerial clusters due to feature ambiguity. This results in a "Dual Failure" state—simultaneous degradation of classification and localization scores—causing valid tiny targets to be erroneously discarded as background noise. To resolve this, we propose the Local Spatial Entropy-guided Uncertainty Minimization (LSE-UM) strategy. Grounded in information theory, our approach treats valid targets as *spatial singularities* characterized by high information density, providing a physical prior that remains robust even when semantic confidence is low.

The encoder feature map $P \in \mathbb{R}^{H\times W\times C}$ is first projected into a Signal Energy Map E via the ℓ_2-norm to decouple raw signal strength from prediction heads:

$$E(i, j) = ||P(i, j)||_2 = \sqrt{\sum_{k=1}^{C} (P(i, j, k))^2} \quad (4)$$

To quantify spatial uncertainty, we model the local energy distribution within a $k \times k$ neighborhood Ω. The normalized energy values form a local spatial probability distribution $p(x, y)$, reflecting the topology of the feature response:

$$p(x, y) = \frac{E(x, y)}{\sum_{(m,n)\in\Omega} E(m, n)}, \quad (x, y) \in \Omega \quad (5)$$

While background clutter manifests as dispersed energy, tiny objects appear as concentrated, pulse-like peaks. We utilize Local Spatial Entropy (LSE) to explicitly measure this dispersion:

$$H_{LSE}(i, j) = -\sum_{(x,y)\in\Omega} p(x, y) \log_2(p(x, y) + \epsilon) \quad (6)$$

where a lower H_{LSE} indicates higher spatial certainty [24]–[26]. To integrate this prior, the final query selection score S_{final} is formulated by modulating the joint task confidence S_{task} with the inverted entropy:

$$S_{final} = S_{task} \cdot (1 + \beta \cdot e^{-H_{LSE}}) \quad (7)$$

TABLE I
COMPARISON WITH ADVANCED DETECTORS

Model	**Backbone**	**Params (M)**	**FPS**	**AP$_{50}$**	**AP$_{75}$**	**AP$_S$**	**AP$_M$**	**AP$_L$**
Faster R-CNN [18]	R-50	41.2	28	42.4	24.4	15.6	36.5	43.0
RetinaNet [19]	R-50	39.3	-	38.7	23.0	11.5	36.2	42.3
ATSS [20]	R-50	31.9	-	40.2	24.6	14.3	35.7	40.9
TODD [21]	R-50	31.8	30	32.2	20.5	9.5	30.3	43.0
YOLOv5-M [22]	-	21.2	63	30.1	18.3	7.7	27.6	38.3
YOLOv5-L [22]	-	46.5	55	31.9	19.6	8.3	29.7	40.7
YOLOv8-M [23]	-	25.8	85	34.7	20.4	9.8	30.7	42.7
YOLOv8-L [23]	-	43.2	72	36.9	22.0	10.6	33.0	45.6
YOLOv10-L [10]	-	24.4	**126**	35.8	20.7	10.1	30.8	43.4
YOLOv10-X [10]	-	29.5	95	38.0	22.4	11.5	33.3	46.6
DINO-DETR [6]	R-50	47.5	-	43.4	24.7	14.6	37.5	50.8
RT-DETR [7]	R-50	42.3	98	38.5	22.4	12.7	32.5	43.4
HCTD [11]	R-50	45.4	94	43.7	24.6	14.7	37.8	50.6
FRE-Det (Ours)	R-50	38.8	87	**49.3**	**29.2**	**19.9**	**40.9**	**57.9**

where β regulates the sensitivity to spatial entropy. This mechanism acts as a Spatial Rectifier, boosting target rankings with high signal singularity. By anchoring query initialization to physical signal peaks, LSE-UM recovers tiny objects otherwise missed due to semantic ambiguity.

IV. EXPERIMENTS

A. Dataset Description

To rigorously evaluate the robustness of FRE-Det in complex aerial surveillance, we leverage the VisDrone2019 benchmark. Developed by the AISKYEYE team at Tianjin University, this large-scale dataset encapsulates diverse geographic environments across 14 Chinese cities, ranging from dense urban centers to rural highways. It presents significant challenges through extreme environmental variability, including dynamic lighting, severe occlusions, and fluctuating crowd densities. Following the official protocol, we utilize 6,471 images for training and 548 for validation across 10 common ground categories. Crucially, the preponderance of tiny objects with extremely low spatial resolution in VisDrone2019 underscores its suitability as a testbed to verify how our proposed frequency-domain enhancement and entropy-guided query selection strategies mitigate feature erosion.

B. Implementation Details

FRE-Det is implemented in PyTorch and trained on two NVIDIA Tesla V100 GPUs with a batch size of 8 and the AdamW optimizer. To ensure a fair comparison, the input resolution is standardized to 640×640 for FRE-Det and other real-time detectors (e.g., YOLO series, RT-DETR), while traditional Transformer-based models maintain their default 800×1333 size. We employ an ImageNet-pretrained ResNet-50 backbone without manual hyperparameter tuning for baseline models to maintain experimental integrity. Evaluation is conducted on the VisDrone2019 validation set across 10

TABLE II
PERFORMANCE RESULTS ON DIFFERENT CLASSES

Class	**Instances**	**AP$_{50}$**	**mAP**	**FPS**
all	38759	49.3	29.2	87
pedestrian	8844	54.4	25.5	–
people	5125	49.7	20.6	–
bicycle	1287	26.5	12.3	–
car	14064	84.3	58.9	–
van	1975	50.8	35.4	–
truck	750	44.0	29.4	–
tricycle	1045	38.2	21.4	–
awning-tricycle	532	20.4	12.1	–
bus	251	63.9	47.7	–
motor	4886	60.7	29.0	–

categories following the standard COCO protocol. We report AP_{50}, AP_{75}, and mAP (averaged over IoU $[0.50 : 0.95]$) as primary metrics. Given our focus on tiny object detection, AP_S (area 32×32 pixels) is emphasized, while AP_M and AP_L are reported to demonstrate multi-scale robustness.

C. Comparison with Advanced Detectors

Table I summarizes the quantitative comparison between FRE-Det and state-of-the-art detectors. Utilizing a ResNet-50 backbone, FRE-Det achieves 49.3% AP_{50}, outperforming the strong baseline HCTD (43.7%) and RT-DETR (38.5%) by 5.6% and 10.8%, respectively. Notably, FRE-Det demonstrates superior localization precision, reaching 29.2% AP_{75} compared to 24.6% for HCTD. This improvement suggests that our Entropy-Guided Query Selection (LSE-UM) effectively suppresses background noise and refines query initialization for more accurate bounding box regression. Furthermore, FRE-Det significantly advances small object detection in remote sensing, attaining 19.9% AP_S—a 35.4% relative improvement over HCTD (14.7%). This underscores the efficacy of the

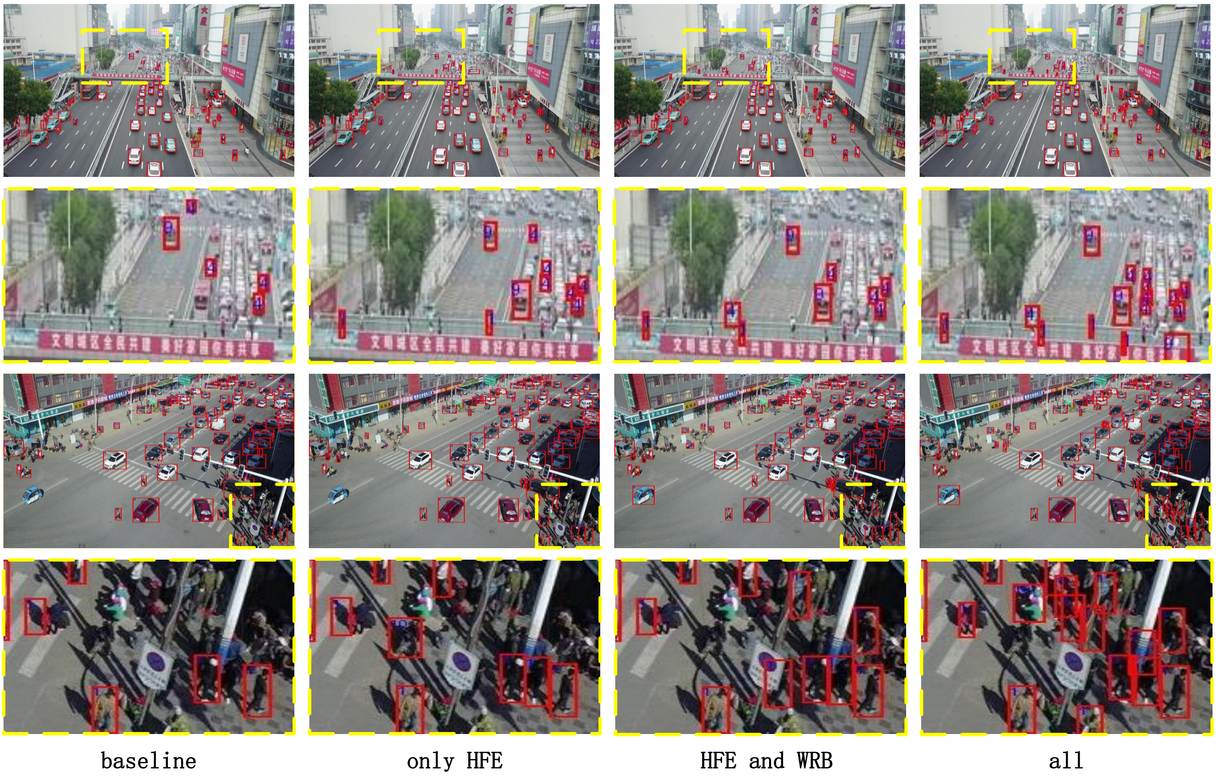

Fig. 3. Visual comparison of ablation study results.

Dual-Frequency Hybrid Encoder in recovering critical high-frequency details for tiny, dense targets. In terms of efficiency, FRE-Det maintains a superior accuracy-complexity trade-off; with 38.8M parameters, it is more lightweight than both RT-DETR (42.3M) and HCTD (45.4M). Although its inference speed (87 FPS) is lower than YOLOv10-L (126 FPS), the substantial accuracy gains render FRE-Det highly suitable for precision-critical surveillance applications.

TABLE III
ABLATION STUDY RESULTS

Module	**HFE**	**WRB**	**LSE-UM**	AP_{50}	AP_S
Baseline				43.7	14.7
1	✓			46.1 (+2.4)	17.8 (+3.1)
2	✓	✓		47.4 (+3.7)	18.5 (+3.8)
3	✓	✓	✓	49.3 (+5.6)	19.9 (+5.2)

To systematically investigate the model's robustness across diverse categories, we detail class-wise performance in Table II. The model demonstrates clear dominance in detecting rigid and distinct objects, achieving AP_{50} scores of 84.3% for *car* and 63.9% for *bus*. Even for dense non-rigid categories like *pedestrian* and *motor*, it maintains competitive performance (54.4% and 60.7%). However, we observe performance drops in categories with complex articulations or extreme scale variations, such as *awning-tricycle* (20.4%) and *bicycle* (26.5%). This can be attributed to the inherent structural complexity and the "long-tail" distribution of these classes (e.g., only 532 instances for *awning-tricycle* vs. 14k for *car*), suggesting a direction for future optimization via data augmentation or specific structural modeling.

D. Ablation Study

To evaluate the individual contribution of each component in FRE-Det, we conduct a stepwise ablation study starting from the HCTD baseline. Quantitative results are summarized in **Table III**, with qualitative comparisons illustrated in **Fig. 3**.

The baseline HCTD struggles with tiny targets due to the inherent "low-pass filter" effect of deep architectures, which suppresses high-frequency edge details. Integrating the *High-Frequency Feature Enhancement* (HFE) module yields an immediate performance boost: AP_{50} increases by 2.4%, and notably, AP_S jumps by 3.1% (a 21.1% relative improvement). This gain validates our hypothesis that explicit spectral filtering revitalizes faint textures and prevents tiny targets from being submerged in complex background noise.

Building upon HFE, replacing standard bilinear upsampling with the *Wavelet Reconstruction Block* (WRB) further elevates AP_{50} to 47.4%. While standard upsampling introduces spatial aliasing that compromises sub-pixel geometric integrity, WRB leverages the invertibility of the IDWT to achieve precise reconstruction. This mechanism effectively mitigates spatial misalignment and preserves sharp boundaries, providing a robust structural foundation for the detection head.

Finally, incorporating the *Local Spatial Entropy-guided Uncertainty Minimization* (LSE-UM) strategy achieves peak performance, reaching 49.3% in AP_{50} and 19.9% in AP_S. The 1.9% incremental gain over the previous step highlights the limitations of purely semantic-based query selection. By utilizing spatial entropy as a prior for information density, LSE-UM forces the model to attend to regions of high signal complexity, effectively recovering valid targets previously discarded due to semantic ambiguity.

As illustrated in **Fig. 3**, the baseline suffers from significant omission errors in crowded scenes. Following the sequential integration of frequency-domain modules (HFE and WRB), object boundaries become sharper and the recall rate for distant targets improves. The full FRE-Det demonstrates the most robust capability, successfully localizing tiny instances overlooked by the baseline. This confirms the synergistic efficacy of coupling spectral feature enhancement with entropy-driven query refinement.

V. Conclusion

In this paper, we presented FRE-Det, an end-to-end real-time detector tailored for complex aerial surveillance. To address feature erosion and query omission, we introduced a Dual-Frequency Enhanced Encoder that leverages FFT and Wavelets (HFE and WRB) to preserve tiny object details. Furthermore, our proposed LSE-UM strategy utilizes local spatial entropy to recover valid targets in dense clusters, effectively bridging the confidence-existence gap.

Experiments on the VisDrone2019 benchmark demonstrate that FRE-Det achieves state-of-the-art performance with $AP_{50} = 49.3\%$ at 87 FPS, setting a new standard for tiny object detection ($AP_S = 19.9\%$). These results validate the efficacy of explicitly modeling spectral and spatial uncertainty. Future work will extend this frequency-aware paradigm to aerial instance segmentation and lightweight edge deployment.

References

[1] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, K. Wu, D. Hu *et al.*, “Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 672–27 683.

[2] S. Zhang, J. Huang, W. Tang, Y. Wu, T. Hu, X. Xu, and J. Liu, “Dimsod: A diffusion-based framework for multi-modal salient object detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 10, 2025, pp. 10 103–10 111.

[3] R. Sapkota and M. Karkee, “Object detection with multimodal large vision-language models: An in-depth review,” *Information Fusion*, vol. 126, p. 103575, 2026.

[4] H. Hu, S. Chen, Z. You, and J. Tang, “Fsenet: Feature suppression and enhancement network for tiny object detection,” *Pattern Recognition*, vol. 162, p. 111425, 2025.

[5] S. Zhang, J. Huang, W. Tang, L. Tian, Y. Wei, and J. Liu, “Multi-modal salient object detection via a unified diffusion model,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[6] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, “Dn-detr: Accelerate detr training by introducing query denoising,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 619–13 627.

[7] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, “Detrs beat yolos on real-time object detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 16 965–16 974.

[8] Z. Shi, J. Hu, J. Ren, H. Ye, X. Yuan, Y. Ouyang, J. He, B. Ji, and J. Guo, “Hs-fpn: High frequency and spatial perception fpn for tiny object detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 7, 2025, pp. 6896–6904.

[9] N. Hoanh and T. V. Pham, “End-to-end transformer-based detection with density-guided query selection for small objects,” *Neurocomputing*, p. 131554, 2025.

[10] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han *et al.*, “Yolov10: Real-time end-to-end object detection,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 107 984–108 011, 2024.

[11] H. Xue, Z. Tang, Y. Xia, L. Wang, and L. Li, “Hctd: A cnn-transformer hybrid for precise object detection in uav aerial imagery,” *Computer Vision and Image Understanding*, p. 104409, 2025.

[12] S. Zhang, J. Huang, S. Chen, Y. Wu, T. Hu, and J. Liu, “Sod-diffusion: Salient object detection via diffusion-based image generators,” in *Computer Graphics Forum*, vol. 43, no. 7. Wiley Online Library, 2024, p. e15251.

[13] J. Jiang, Y. Zhang, and S. Zhang, “Deep vision transformer with association divergence for image anomaly detection and localization,” in *International Conference on Neural Information Processing*. Springer, 2024, pp. 1–15.

[14] S. Zhang, X. Hu, and J. Liu, “Tranbf: Deep transformer networks and bayesian filtering for time series anomalous signal detection in cyber-physical systems,” in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[15] S. Zhang, J. Huang, W. Tang, J. Liu, L. HAN, J. Li, H. Yuan, and Z. Fan, “Simplediffusion: A lightweight and efficient conditional diffusion model for multi-modal salient object detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 15, 2026, pp. 12 753–12 761.

[16] Q. Bi, J. Yi, H. Zheng, H. Zhan, Y. Huang, W. Ji, Y. Li, and Y. Zheng, “Learning frequency-adapted vision foundation model for domain generalized semantic segmentation,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 94 047–94 072, 2024.

[17] S. Zhang, X. Hu, and J. Liu, “Causal fusion of convolutional neural network and vision transformer for image anomaly detection and localization,” in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[18] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[19] T.-Y. Ross, G. Dollár *et al.*, “Focal loss for dense object detection,” in *proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2017, pp. 2980–2988.

[20] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, “Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9759–9768.

[21] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, “Tood: Task-aligned one-stage object detection,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2021, pp. 3490–3499.

[22] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, J. Fang, C. Wong, Z. Yifu, D. Montes *et al.*, “ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations,” *Zenodo*, 2022.

[23] H. Xue, X. Wang, Y. Xia, Z. Tang, L. Li, and L. Wang, “Enhanced yolov8 for small object detection in uav aerial photography: Yolo-uav,” in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[24] S. Zhang and J. Liu, “Anomalous signal detection for cyber-physical systems using interpretable causal neural network,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[25] S. Zhang, J. Huang, Y. Wu, T. Hu, W. Tang, and J. Liu, “Seg-diffusion: Text-to-image diffusion model for open-vocabulary semantic segmentation,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[26] S. Zhang and J. Liu, “Feature-constrained and attention-conditioned distillation learning for visual anomaly detection,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 2945–2949.