Enhancing Asset Visibility in Smart Energy Systems: A Computational Intelligence Framework for Heterogeneous Infrastructure Inspection

Li Tan[1*], Yibo Li[1], Dongjiao Ge[2], Xiaofeng Lian[1], Yixin Peng[1]

[1] School of Computer and AI, Beijing Technology and Business University, Beijing, China

[2] Faculty of Data Science, City University of Macau, Macau (SAR), China

*tanli@th.btbu.edu.cn, 2431062079@st.btbu.edu.cn, djge@cityu.edu.mo, lianxf@th.btbu.edu.cn, 1749853240@qq.com

Abstract*—The fragmentation of energy infrastructure creates an Asset Visibility Gap, constraining digital twin construction and smart grid scheduling. While UAV imagery bridges this gap, precise perception faces challenges from feature dissipation of tiny assets and boundary aliasing in dense arrays. To address this, we propose HFA-YOLO, a lightweight detection framework based on YOLOv11 tailored for heterogeneous energy assets. The model reconstructs the feature pathway via three mechanisms: First, Space-to-Depth Convolution (SPD-Conv) establishes a lossless down-sampling channel to mitigate feature dissipation. Second, a Detail-Enhanced Unit (DEU) leverages gradient priors to sharpen boundaries and decouple dense components. Finally, a Hybrid Feature Aggregation (HFA) module employs frequency-spatial filtering to enhance weak features and suppress environmental noise. Experiments across photovoltaic, transmission tower, wind turbine, and VisDrone2019 scenarios demonstrate superior generalization. With only 3.06M parameters, the method achieves an optimal accuracy-efficiency trade-off, realizing mAP50 gains of 2.12%, 4.91%, 1.40%, and 4.18% in photovoltaic, transmission tower, wind turbine, and VisDrone2019 scenarios, respectively, compared to the baseline. Ablation studies verify the effectiveness of each component. HFA-YOLO offers a robust solution for the visibility bottleneck of energy assets. The source code is publicly available at https://github.com/HermitKin/HFA-YOLO.

***Index Terms*—Tiny Object Detection, YOLOv11, Feature Aggregation, UAV Inspection, Energy Infrastructure**

I. Introduction

The rapid proliferation of heterogeneous energy infrastructure, including distributed photovoltaics (PV), transmission towers, and wind turbines, introduces critical perception blind spots for grid operators due to asset dispersion and environmental complexity. Accurate auditing and defect detection of these assets are pivotal for downstream tasks such as power generation forecasting and maintenance scheduling. While Unmanned Aerial Vehicles (UAVs) have become the pivot for automated inspection, enabling remote and high-frequency monitoring, the transition from raw aerial imagery to actionable data is severely constrained by three intrinsic bottlenecks in aerial visual signals.

Feature Dissipation degrades high-frequency details of minute targets during CNN down-sampling, causing small components (e.g., insulators) to vanish in deep layers. *Boundary Blurring* in dense scenarios (e.g., PV arrays) creates semantic adhesion, leading to counting errors and localization drift due to insufficient sub-pixel gradient sensitivity. *SNR Imbalance* in complex industrial backgrounds causes weakly textured targets (e.g., wind turbine blades) to be overwhelmed by interferences like specular reflections, inducing high false alarm rates.

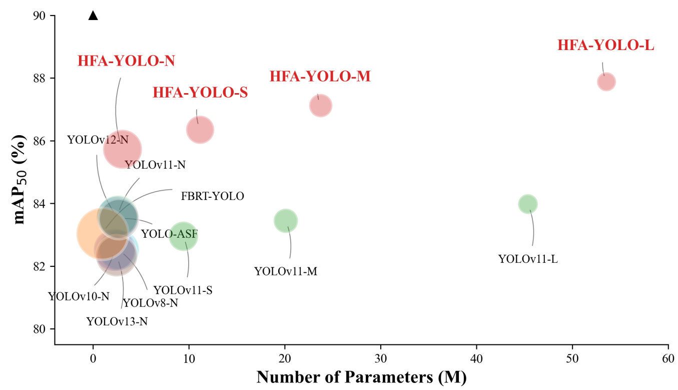

Fig. 1. Accuracy-Efficiency Comparison. Our HFA-YOLO series (red) dominates the Pareto frontier on the PV dataset, significantly outperforming existing mainstream and lightweight detectors.

General deep learning models struggle to balance noise suppression with fine-grained detail preservation, and their computational overhead exceeds aerial edge computing constraints. Consequently, the YOLO series remains the *de facto* choice for efficient perception. However, while YOLOv11 [1] provides a robust baseline, it lacks the architectural specialization needed for the specific signal degradation and frequency aliasing inherent to energy infrastructure.

To address the conflict between limited edge computing power and the need for high-precision detection of dense, tiny objects, we propose HFA-YOLO. By reconstructing the feature pathway via a frequency-aware perspective, it achieves an optimal parameter-accuracy trade-off. Extensive experiments across three core energy datasets demonstrate that HFA-YOLO consistently outperforms state-of-the-art YOLO variants (v8–

* Corresponding author.

v13) and lightweight models across all scales. As visualized in Fig. 1, our approach defines a new Pareto frontier of the performance-efficiency curve. Our main contributions are summarized as follows:

- We integrate **Space-to-Depth Convolution (SPD-Conv)** [2] to replace standard strided convolutions, ensuring lossless preservation and transmission of fine-grained details to mitigate feature vanishing for tiny targets.
- We propose a **Detail Enhancement Unit (DEU)** [3] that explicitly models gradient priors to sharpen local edge representations, effectively resolving boundary adhesion in dense object arrays.
- We design a **Hybrid Feature Aggregation (HFA)** module [4] for frequency-spatial collaborative filtering. It acts as a dynamic global filter to suppress complex industrial noise and enhance weak target structures.
- Extensive experiments on three energy datasets and the VisDrone2019 benchmark confirm that **HFA-YOLO** achieves an optimal balance between parameter efficiency and detection accuracy, proving its robustness for aerial edge deployment.

II. Related Work

A. Evolution of Object Detection and Aerial Adaptation Challenges

Object detection has evolved from high-precision two-stage algorithms (Faster R-CNN) to fast single-stage models (YOLO series [5]). Recent iterations (YOLOv8, YOLOv11) leverage anchor-free mechanisms and decoupled heads to optimize the latency-accuracy trade-off, maturing alongside the Feature Pyramid Network (FPN) paradigm [6]. Recently, spectral properties are also being explored to enhance complex texture modeling.

However, designed for natural scenes, these general-purpose models face a “scale gap” in aerial imagery, where tiny targets suffer from *Feature Vanishing*. Aerial-optimized methods address this via cropping strategies (ClusDet [7]), feature complementary mapping (FBRT-YOLO [8]), or dynamic convolutions [9], [10]. Yet, relying primarily on geometric alignment, they lack explicit mechanisms to suppress high-frequency environmental noise (e.g., shadows, reflections). Consequently, they fail the real-time robustness demands of UAV edge devices.

B. Specific Pain Points in Energy Infrastructure Perception

In intelligent energy systems, perception tasks have shifted to fine-grained quantitative asset mapping for PVs, transmission towers, and wind turbines. However, existing models overlook two core challenges: “dense adhesion” and “weak feature interference.”

In transmission inspections [11], minute fittings are frequently overwhelmed by complex backgrounds, causing missed detections [12]. For asset mapping, standard convolutions lack the sub-pixel gradient sensitivity needed to separate closely arranged PV modules or tower structures, resulting in severe counting errors [13], [14]. While segmentation approaches attempt to mitigate this, their computational costs prohibit edge deployment. Additionally, low-texture targets (e.g., reflective PV panels, white turbine blades) carry minimal semantics and suffer from illumination interferences [15], [16] that existing spatial attention mechanisms fail to filter.

A unified framework addressing tiny object recovery, dense array separation, and frequency-domain noise suppression remains absent. HFA-YOLO fills this gap, achieving high-precision perception of multi-source energy assets via a frequency-spatial collaborative recovery mechanism.

III. Method

A. Overall Network Architecture

Tailored to the constraints of energy infrastructure inspection, we propose HFA-YOLO (Fig. 2). Built upon YOLOv11, HFA-YOLO reconstructs the feature perception pathway through a tripartite architecture. SPD-Conv enables lossless feature down-sampling. DEU applies differential operators to sharpen boundaries in dense arrays. Concurrently, HFA leverages frequency-spatial synergistic filtering to suppress environmental noise.

B. Space-to-Depth Convolution (SPD-Conv)

To achieve lossless dimensionality reduction and preserve fine-grained spatial information, we deploy the Space-to-Depth Convolution [2] in the backbone and feature fusion paths, as illustrated in Fig. 3. This module couples a space-to-depth transformation with a non-strided convolution to establish a bijective mapping, losslessly projecting high-resolution spatial details into the channel domain. The SPD layer adopts a “space-to-channel” strategy. Given input $X \in \mathbb{R}^{H\times C\times W}$, it slices spatial pixels with a scale factor of 2, generating four sub-feature maps concatenated into X'. The formulation is defined as:

$$X'_{i,j,:} = \text{Concat}\left(X_{m,n}, X_{m+1,n}, X_{m,n+1}, X_{m+1,n+1}\right) \quad (1)$$

where $i \in [0, \frac{H}{2}-1]$ and $j \in [0, \frac{W}{2}-1]$ are spatial indices. We define base sampling indices as $m = 2i$ and $n = 2j$. This mathematically reversible transformation introduces no truncation, ensuring edges and high-frequency textures of minute targets are fully preserved in the expanded channel space.

Subsequently, to address channel redundancy and extract cross-channel semantics, a stride-1 standard convolution is appended:

$$Y = \mathcal{F}_{\text{Conv}}(X') = \sigma(\text{BN}(W * X' + b)) \quad (2)$$

where $W \in \mathbb{R}^{C_{out}\times 4C\times k\times k}$ represents weights. This layer serves a dual function: compressing channels from $4C$ to C_{out} for efficiency, and transforming pixel information into semantic features via linear combinations, allowing the network to trace back full-resolution details in deep layers.

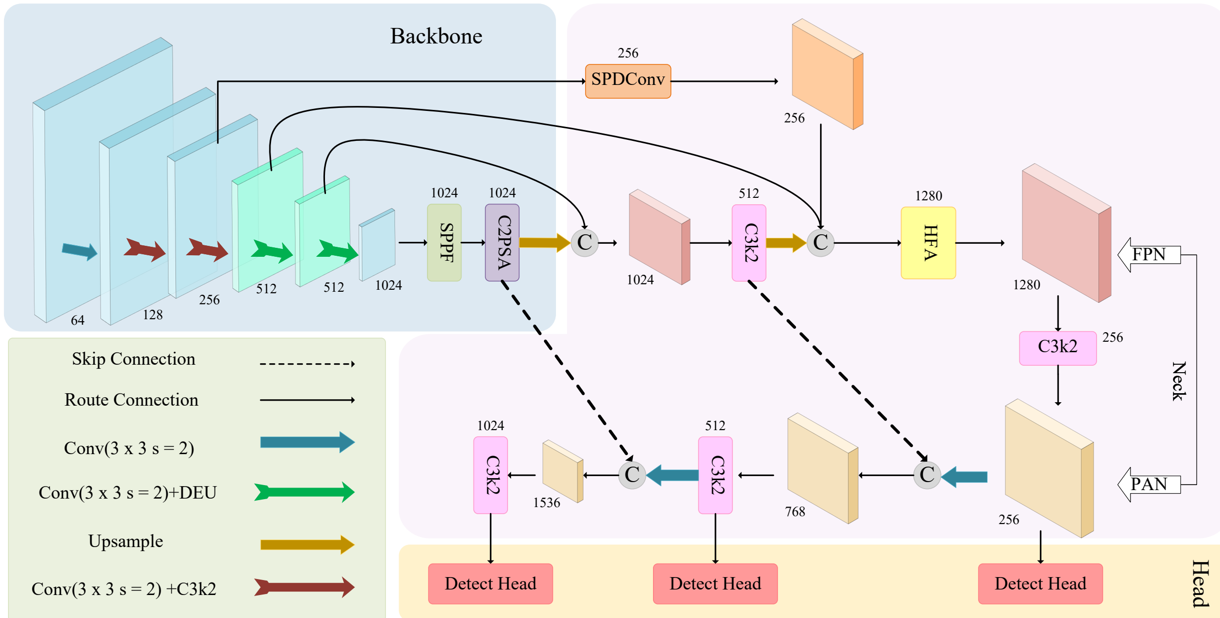

Fig. 2. Overview of the proposed HFA-YOLO network. It consists of three main improvements: SPD-Conv, DEU, and HFA module.

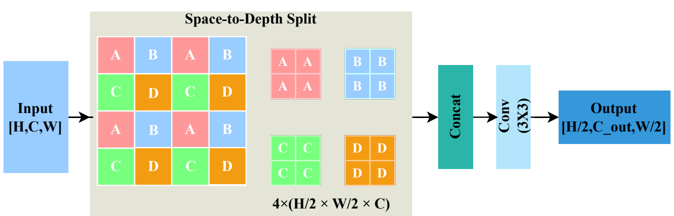

Fig. 3. The structure of Space-to-Depth Convolution. It illustrates the space-to-depth transformation that losslessly down-samples the feature map by remapping spatial pixels to the channel dimension, followed by a non-strided convolution.

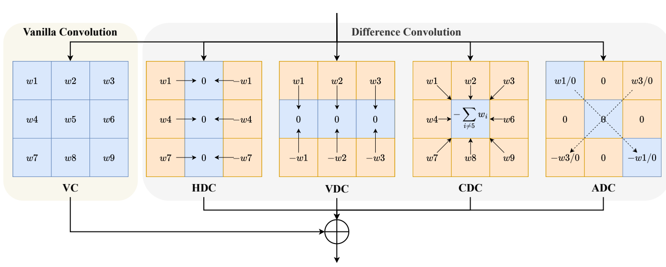

Fig. 4. Schematic illustration of the Detail-Enhanced Unit (DEU). It shows the parallel branches including the standard convolution and the four differential operators (central, horizontal, vertical, and diagonal), which are fused during inference via structural re-parameterization.

C. Detail-Enhanced Unit (DEU)

Inspired by DEANet [3], we design the Detail-Enhanced Unit, as shown in Fig. 4, to mitigate deep feature texture loss. It addresses feature smoothing in high-density scenarios, preventing object adhesion or missed detections caused by blurred boundaries. DEU reconstructs the feature interaction mechanism. To maintain gradient stability, DEU retains residual connections while introducing a kernel-level omni-dimensional gradient perception mechanism. It explicitly decouples the core convolution into five distinct weight components. An intensity perception term (W_{std}) captures the basic semantic topology. Meanwhile, four directional gradient terms (central W_{cd}, horizontal W_{hd}, vertical W_{vd}, and diagonal W_{ad} differences) act as learnable edge filters. The equivalent kernel is defined as:

$$W_{fused} = W_{std} + \sum_{k \in \{cd,hd,vd,ad\}} \theta_k \cdot W_k \tag{3}$$

where θ represents balancing coefficients. Physically, these operators act as learnable edge filters, capturing subtle variations of dense targets in all directions.

Furthermore, DEU employs Structural Re-parameterization. Via linear superposition, the multi-branch topology collapses into a standard 3×3 kernel W_{fused} during inference. This enhances boundary detection and reduces localization drift while maintaining inference speed, balancing feature enhancement with zero extra computational cost.

D. Hybrid Feature Aggregation (HFA)

Addressing background clutter and semantic ambiguity, we design the Hybrid Feature Aggregation module based on OmniKernel [4] architecture (depicted in Fig. 5). Functioning as a Global Semantic Filter, HFA extracts key fine-grained features by filtering noise from complex environments.

To balance high-order interactions with efficiency, HFA employs a Channel Partitioning Strategy. The input feature $X \in \mathbb{R}^{C\times H\times W}$ is partitioned along the channel dimension at a 1:3 ratio. Only 25% of the channels are routed into the evolution path to extract core semantics, while the remain-

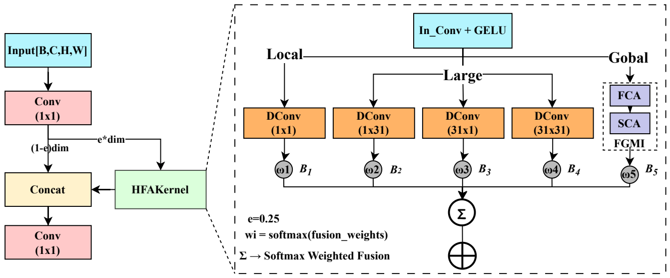

Fig. 5. The detailed schematic of the proposed Hybrid Feature Aggregation module. It illustrates the channel partitioning strategy, the parallel processing branches, and the dynamic gating fusion mechanism.

ing 75% bypass this transformation via identity mapping to preserve high-fidelity context.

Inside the evolution path, a "frequency-spatial" dual calibration strategy is applied. It sequentially employs Frequency Channel Attention (FCA) to capture global frequency context via FFT, and Spatial Channel Attention (SCA) to enhance local responses:

$$\begin{aligned} \text{FCA}: \quad & a_f = \text{Conv}_{1\times1}(\text{GAP}(X_0)), \\ & X_{fca} = \mathcal{F}^{-1}(a_f \odot \mathcal{F}(X_0)) \\ \text{SCA}: \quad & a_s = \text{Conv}_{1\times1}(\text{GAP}(X_{fca})), \\ & X_{sca} = a_s \odot X_{fca} \end{aligned} \tag{4}$$

Purified features X_{sca} enter a multi-branch system. Spatially, anisotropic kernels $\{B_1, \ldots, B_4\}$ with scales $\{1 \times 1, 1 \times 31, 31 \times 1, 31 \times 31\}$ capture long-range contexts. In the frequency domain, Frequency-domain Gated Modulation and Interaction (FGMI) serves as the fifth branch B_5. Following the Convolution Theorem, FGMI interacts in spectral space via FFT, capturing global low-frequency energy with linear complexity:

$$\begin{aligned} Y_{fgmi} = \alpha \odot \text{Re}\left(\mathcal{F}^{-1}(\mathcal{F}(X_1) \odot \mathcal{F}(X_2))\right) \\ + \beta \odot X_{sca} \end{aligned} \tag{5}$$

where $\mathcal{F}$ denotes 2D FFT, $\odot$ is the Hadamard Product, and α, β are learnable parameters.

Finally, branches are aggregated via Dynamic Gating Fusion. Softmax generates normalized weights γ to guide the fusion of diverse receptive fields with residual connections:

$$\gamma_i = \frac{e^{\omega_i}}{\sum_{j=1}^{5} e^{\omega_j}}, \quad Y_{\text{Kernel}} = \text{ReLU}\left(X_0 + \sum_{i=1}^{5} \gamma_i \odot B_i\right) \tag{6}$$

The aggregated feature Y_{Kernel} is concatenated with the reserved 75% high-fidelity channels, unifying local details and macroscopic views.

IV. Experiments

This section evaluates the proposed HFA-YOLO framework. We first detail the datasets (three energy scenarios and the VisDrone2019 benchmark), evaluation metrics, and implementation settings. Next, quantitative and qualitative comparisons with state-of-the-art detectors, alongside inference speed analysis, demonstrate our optimal accuracy-efficiency trade-off for real-time edge deployment. Finally, ablation and sensitivity analyses validate the synergistic contributions of the proposed modules and structural configurations.

A. Datasets and Evaluation Metrics

To evaluate performance across heterogeneous scenarios, we utilized three energy datasets (Photovoltaics, Transmission Tower, Wind Turbine) alongside the VisDrone2019 benchmark. These datasets serve as a robust testbed for challenges including tiny object loss, dense adhesion, and weak feature interference. To ensure rigorous evaluation and reproducibility, the precise image distributions for training, validation, and testing sets are detailed as follows: Photovoltaics (4,365/1,215/304), Transmission Tower (412/93/64), Wind Turbine (5,709/1,631/816), and VisDrone2019 (6,471/548/1,610). Following standard COCO protocols, we report mean Average Precision (mAP_{50} and $\text{mAP}_{50:95}$), Precision (P), Recall (R), and parameter count (Params) to comprehensively assess the accuracy-efficiency trade-off.

B. Implementation Details

Experiments were implemented in PyTorch on an NVIDIA RTX 4070S GPU using YOLOv11n as the baseline. The model was trained for 300 epochs via SGD with an initial learning rate of 0.01, linear learning rate decay, a weight decay of 0.0005, and a batch size of 32. Input images were resized to 640×640.

C. Comparative Experiments

As evidenced in Table I, HFA-YOLO establishes a new state-of-the-art across all model scales. Notably, HFA-YOLO-N surpasses the strong baseline YOLOv11-N by 2.12% in mAP_{50} and 2.42% in $mAP_{50:95}$ with a marginal parameter increase of 0.48M. This validates that frequency-domain enhancement captures the fine-grained semantics of dense PV components more effectively than general-purpose architectures. Furthermore, our method achieves a superior accuracy-efficiency trade-off compared to specialized lightweight detectors. While FBRT-YOLO achieves extreme compression (0.97M), it lags behind ours by 2.28% in $mAP_{50:95}$ due to compromised feature representation. Similarly, our model outperforms YOLO-ASF [17] by 2.22% in mAP_{50}, confirming that the frequency-spatial collaborative filtering in HFA handles specular reflections more robustly than pure spatial attention mechanisms.

Table II validates robustness across Transmission Tower, Wind Turbine, and VisDrone2019 datasets. Specifically, HFA-YOLO boosts mAP_{50} by 4.91% for transmission towers (suppressing false positives) and Recall by 2.48% for wind turbines (enhancing edge perception). Furthermore, a 4.18% mAP_{50} gain on the VisDrone2019 benchmark verifies its strong generalization capabilities for universal aerial objects, ruling out overfitting to specific energy data. These results

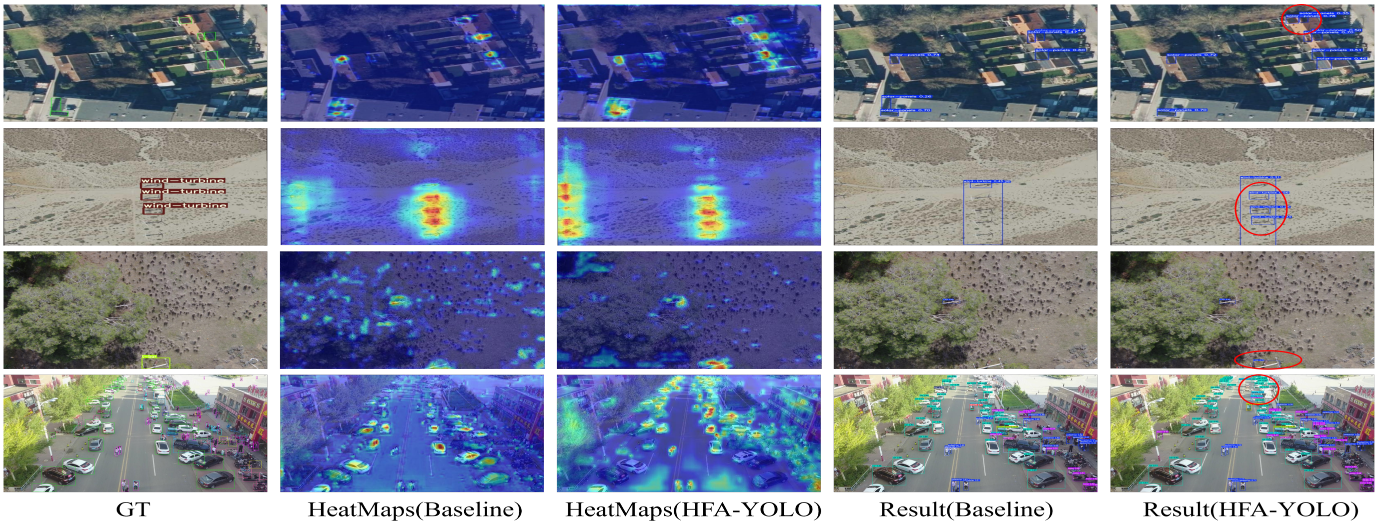

Fig. 6. Qualitative comparison on photovoltaic panels, wind turbines, transmission towers, and VisDrone2019 (top to bottom). Columns show Ground Truth, Grad-CAM heatmaps, and detection results. HFA-YOLO exhibits focused activation on tiny and dense targets compared to the baseline's diffuse response. Red circles indicate false negatives in the baseline successfully recovered by our method.

TABLE I
QUANTITATIVE COMPARISON ON THE PHOTOVOLTAIC PANEL VAL SET.

Model	**P (%)**	**R (%)**	**mAP$_{50}$ (%)**	**mAP$_{50:95}$ (%)**	**Params (M)**
YOLOv8-N	79.01	77.85	82.52	63.59	2.69
YOLOv10-N	77.34	77.01	82.52	63.31	2.27
YOLOv12-N	78.91	78.15	83.57	64.48	2.51
YOLOv13-N	78.77	75.26	82.33	63.29	2.45
FBRT-YOLO [8]	76.63	79.57	83.04	65.03	**0.97**
YOLO-ASF [17]	79.55	78.12	83.51	64.61	2.67
YOLOv11-N	80.48	76.41	83.61	64.89	2.58
YOLOv11-S	80.53	77.48	82.96	65.38	9.41
YOLOv11-M	80.95	78.42	83.45	65.92	20.11
YOLOv11-L	81.32	79.15	83.98	66.55	45.37
HFA-YOLO-N	83.05	82.09	85.73	67.31	3.06
HFA-YOLO-S	83.65	82.85	86.35	67.15	11.16
HFA-YOLO-M	84.28	83.52	87.12	68.05	23.73
HFA-YOLO-L	**84.95**	**84.15**	**87.88**	**68.92**	53.54

TABLE II
GENERALIZATION VALIDATION AND EFFICIENCY COMPARISON ON HETEROGENEOUS DATASETS

Dataset	**Model**	**P (%)**	**R (%)**	**mAP$_{50}$ (%)**	**mAP$_{50:95}$ (%)**	**FPS**
Photovoltaics	Baseline	80.48	76.41	83.61	64.89	**865.49**
	Ours	**83.05**	**82.09**	**85.73**	**67.31**	592.91
Transmission Tower	Baseline	76.90	73.24	73.82	42.17	105.07
	Ours	**82.88**	**75.20**	**78.73**	**43.23**	**105.92**
Wind Turbine	Baseline	82.33	82.28	87.39	66.46	**888.30**
	Ours	**84.10**	**84.76**	**88.79**	**67.08**	623.45
VisDrone 2019	Baseline	43.70	33.60	33.10	19.20	236.41
	Ours	**49.15**	**36.51**	**37.28**	**22.26**	**350.88**

demonstrate that the architecture achieves precise perception with a minimal parameter increase of 0.48M.

Furthermore, to address the strict latency constraints of UAV edge computing, we evaluated inference efficiency (Table II). Across all scenarios, HFA-YOLO maintains excellent real-time processing speeds. Notably, on the highly dense VisDrone benchmark, our method achieves 350.88 FPS end-to-end, vastly outperforming the YOLOv11n baseline (236.41 FPS). This occurs because our method effectively suppresses background noise and generates fewer redundant bounding boxes, significantly accelerating the Non-Maximum Suppression (NMS) stage. Although the baseline exhibits higher FPS on sparser datasets (e.g., Wind Turbine) with minimal NMS overhead, HFA-YOLO consistently delivers ultra-real-time efficiency suited for aerial deployment while securing substantial accuracy gains.

Qualitative results in Fig. 6 corroborate the quantitative advantages. HFA-YOLO generates focused heatmaps on PV arrays and turbines, enabling sub-pixel boundary decoupling and noise suppression amidst ground clutter. In summary, the proposed method resolves the perception bottleneck of heterogeneous targets, bridging the asset visibility gap for high-fidelity digital twins.

D. Ablation Studies

We conducted ablation studies on the VisDrone2019 validation set to verify the contribution of each proposed component: SPD-Conv, DEU, and HFA. Quantitative results are detailed in Table III. Integrating SPD-Conv yields a 0.80% mAP$_{50}$ improvement (Row 2), confirming its efficacy in mitigating fine-grained information loss during down-sampling. The HFA module delivers the most significant standalone gain of 2.41% (Row 3), validating that frequency-spatial synergy effectively suppresses noise while capturing global context. While DEU shows moderate standalone performance, its integration into the full architecture (Row 8 vs. Row 7) further elevates precision. Ultimately, the complete model achieves 37.28%

TABLE III
ABLATION STUDY RESULTS ON VISDRONE2019 (VAL)

SPD-Conv	DEU	HFA	P (%)	R (%)	mAP_{50} (%)	$mAP_{50:95}$ (%)
-	-	-	43.70	33.60	33.10	19.20
✓	-	-	44.09	34.14	33.90	19.88
-	-	✓	46.83	35.44	35.51	20.80
-	✓	-	45.01	33.74	33.42	19.39
✓	✓	-	44.65	34.41	34.26	20.01
-	✓	✓	46.95	34.96	35.03	20.54
✓	-	✓	47.14	36.63	36.54	21.62
✓	✓	✓	**49.15**	**36.51**	**37.28**	**22.26**

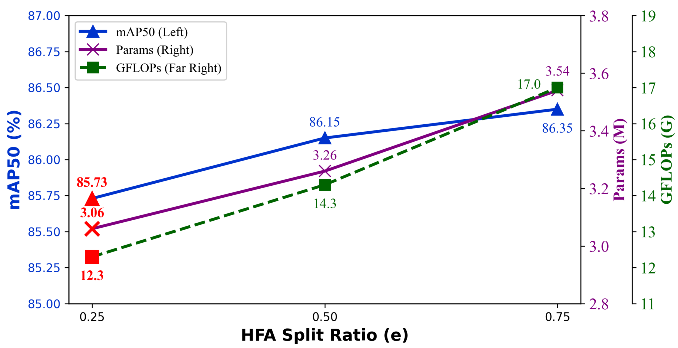

Fig. 7. Sensitivity analysis of the HFA partition ratio e. The red markers at $e = 0.25$ highlight the optimal accuracy-efficiency trade-off, avoiding the diminishing returns observed at higher ratios.

mAP_{50}, a 4.18% improvement over the baseline. This demonstrates the complementarity of our design: SPD preserves low-level spatial details, DEU sharpens local boundaries, and HFA aggregates global semantics.

Fig. 7 presents the sensitivity analysis of the channel partition ratio e. Increasing e induces a linear parameter surge (3.06M to 5.12M) but yields saturated performance gains (mAP_{50}: 85.73% to 86.35%). This phenomenon of diminishing returns confirms that processing 25% of frequency channels suffices for global context modeling, whereas higher ratios introduce computational redundancy. Consequently, we adopt $e = 0.25$ as the optimal configuration to balance detection accuracy and model lightness.

V. CONCLUSION

This paper proposes HFA-YOLO to address perceptual bottlenecks in aerial imagery. By synergizing SPD-Conv, DEU, and HFA, the model effectively resolves scale variations and background interference, achieving an optimal accuracy-efficiency balance (mAP_{50} of 85.73% with 3.06M parameters) for energy infrastructure inspection. However, the reliance on a single RGB modality remains a limitation for identifying internal defects, such as photovoltaic hotspots. Future work will focus on integrating infrared multi-modal fusion to enable all-weather thermal anomaly monitoring and applying model quantization and pruning to satisfy the rigorous power constraints of UAVs, thereby establishing a comprehensive perception loop for the smart grid lifecycle.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 62472010 and in part by the Chongqing Natural Science Foundation under Grant CSTB2024NSCQ-MSX0687.

REFERENCES

[1] G. Jocher and J. Qiu, "Ultralytics YOLO11," 2024, accessed: 2024-10-01. [Online]. Available: https://github.com/ultralytics/ultralytics

[2] R. Sunkara and T. Luo, "No more strided convolutions or pooling: a new CNN building block for low-resolution images and small objects," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2022, pp. 443–459.

[3] Z. Chen, Z. He, and Z.-M. Lu, "DEA-Net: single image dehazing based on detail-enhanced convolution and content-guided attention," *IEEE Transactions on Image Processing*, vol. 33, pp. 1002–1015, 2024.

[4] Y. Cui, W. Ren, and A. Knoll, "Omni-kernel network for image restoration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 1426–1434.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[6] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[7] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8311–8320.

[8] Y. Xiao, T. Xu, Y. Xin, and J. Li, "FBRT-YOLO: faster and better for real-time aerial image detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 8, 2025, pp. 8673–8681.

[9] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, "Large selective kernel network for remote sensing object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 794–16 805.

[10] L. Chen, Y. Fu, L. Gu, C. Yan, T. Harada, and G. Huang, "Frequency-aware feature fusion for dense image prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 763–10 780, 2024.

[11] X. Tao, D. Zhang, Z. Wang, X. Liu, H. Zhang, and D. Xu, "Detection of power line insulator defects using aerial images analyzed with convolutional neural networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 4, pp. 1486–1498, 2018.

[12] R. Jenssen, D. Roverso, M. Castro, S. Reinemo, A. Johansen, and J. H. Jørgensen, "Intelligent monitoring and inspection of power line components powered by UAVs and deep learning," *IEEE Power and Energy Technology Systems Journal*, vol. 6, no. 1, pp. 11–21, 2019.

[13] A. Greco, C. Pironti, A. Saggese, M. Vento, and V. Vigilante, "A deep learning based approach for detecting panels in photovoltaic plants," in *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*, 2020, pp. 1–7.

[14] K. He and L. Zhang, "Automatic detection and mapping of solar photovoltaic arrays with deep convolutional neural networks in high resolution satellite images," in *2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2)*. IEEE, 2020, pp. 3068–3073.

[15] J. J. Vega Díaz, M. Vlaminck, D. Lefkaditis, S. A. Orjuela Vargas, and H. Luong, "Solar panel detection within complex backgrounds using thermal images acquired by UAVs," *Sensors*, vol. 20, no. 21, p. 6219, 2020.

[16] L. Wang and Z. Zhang, "Automatic detection of wind turbine blade surface cracks based on UAV-taken images," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 9, pp. 7293–7303, 2017.

[17] M. Kang, C.-M. Ting, F. F. Ting, and R. C.-W. Phan, "ASF-YOLO: a novel YOLO model with attentional scale sequence fusion for cell instance segmentation," *Image and Vision Computing*, vol. 147, p. 105057, 2024.