DH-DETR: A Dynamic Hypergraph Detection Transformer for Enhanced Tiny Object Detection in UAV Imagery

Shengbo Wang[1], Wenzhu Yang[1, 2, *], Jinming Li[1]
[1]School of Cyber Security and Computer, Hebei University, Baoding, China
[2]Machine Vision Engineering Research Center, Hebei University, Baoding, China
[*]Corresponding Author. Email: wenzhuyang@hbu.edu.cn

***Abstract*—Tiny object detection in UAV imagery is inherently challenging due to extreme scale variations, limited pixel representations, and densely cluttered scenes. To address these challenges, we propose DH-DETR, a Dynamic Hypergraph Detection Transformer, specifically designed for tiny object detection in UAV imagery. DH-DETR incorporates three key components: a Hypergraph High-order Correlation Aggregator (HHCA) that captures cross-level and high-order dependencies to strengthen global structural reasoning; a lightweight Frequency–spatial Unified SElective (FUSE) module that enriches fine-grained details by injecting frequency-domain cues into shallow features while suppressing background noise; and a Dilated Depthwise Multi-Head Self-Attention (D2MHSA) block that efficiently enlarges the receptive field for long-range context modeling without quadratic complexity. Together, these components enable DH-DETR to preserve critical details, focus on discriminative regions, and achieve superior accuracy on challenging tiny object detection tasks. Extensive experiments on the VisDrone, UAVDT, and AI-TOD benchmarks demonstrate that DH-DETR achieves state-of-the-art performance, obtaining 28.8% AP on VisDrone, 19.1% AP on UAVDT, and 28.1% AP on AI-TOD, while maintaining competitive computational efficiency. These results highlight the effectiveness of integrating hypergraph-based high-order reasoning, frequency-domain enhancement, and efficient attention mechanisms for advancing high-resolution small-object detection in real-world UAV scenarios.**

***Index Terms*—Tiny object detection, Dynamic hypergraph, Detection Transformer, Unmanned aerial vehicles**

I. Introduction

Object detection is a fundamental task in computer vision, aiming to identify and localize objects in images. Despite remarkable progress driven by deep learning, detecting tiny objects remains a persistent challenge, especially in UAV imagery. In such scenarios, target objects often occupy only a few pixels, appear in densely packed distributions, and are heavily interfered by cluttered backgrounds and complex scene structures [1]. These factors severely weaken local discriminative cues and make it difficult to achieve robust detection under practical computational constraints.

To address these challenges, object detectors have evolved into CNN-based and Transformer-based paradigms. CNN-based methods such as YOLO [2] offer fast inference, but depend on handcrafted designs (e.g., anchors and NMS [3]), which limit flexibility and scalability in complex UAV scenarios with large variations in object scale and density. Detection Transformer (DETR) [4] introduced a new end-to-end detection paradigm by eliminating handcrafted components and formulating object detection as a set prediction problem. However, the original DETR suffers from slow convergence and limited performance on small objects, motivating extensive research on improved variants. Representative works include DN-DETR [5] for accelerating training via denoising, Sparse-DETR [6] for reducing computational overhead through sparse attention, UFA-DETR [7] for enhancing small-object representations, and Lite-DETR [8] for lightweight deployment. Deformable-DETR [9] and DINO [10] further improve detection accuracy by incorporating multi-scale and deformable attention mechanisms, while RT-DETR [11] balances efficiency and accuracy by decoupling the encoder. Despite these advances, achieving robust tiny-object detection in dense UAV scenes remains challenging, suggesting that existing improvements have not fully addressed the underlying representation bottlenecks.

In particular, existing detectors typically aggregate multi-scale information through feature pyramids or encoders based on pairwise interactions across scales or between queries and features. While effective for basic scale alignment, such formulations are inherently limited in modeling high-order structural relationships among multiple feature levels, spatial regions, and object instances. This limitation becomes critical in dense tiny-object scenarios, where objects share highly similar appearances and reliable detection requires joint reasoning over object groups and contextual structures, which cannot be adequately captured by pairwise interactions alone.

To overcome these limitations, we propose DH-DETR, a Dynamic Hypergraph Detection Transformer, tailored for tiny object detection in UAV imagery. Unlike conventional DETR variants that emphasize deformable attention or hierarchical fusion, DH-DETR is centered on dynamic hypergraph-based high-order reasoning, explicitly modeling many-to-many relationships across feature levels and spatial regions. At the core of DH-DETR is a dynamic Hypergraph High-order Correlation Aggregator (HHCA), which constructs a differentiable hypergraph during feature pyramid fusion to inject global structural semantics while preserving fine-grained local cues. Building upon this representation, a lightweight Frequency-spatial Unified SElective (FUSE) module enhances shallow

features with frequency-domain cues, and a Dilated Depthwise Multi-Head Self-Attention (D2MHSA) block efficiently enlarges the receptive field. Together, these components form a unified framework that balances local detail preservation and global contextual reasoning for dense tiny-object detection.

Our main contributions are summarized as follows:

- We propose DH-DETR, a novel Detection Transformer framework that integrates hypergraph-based reasoning with efficient feature pyramid, enabling effective tiny object detection in UAV imagery.
- We design a dynamic hypergraph high-order correlation aggregator that dynamically constructs differentiable hypergraphs to model high-order correlations across feature levels and spatial regions, facilitating cross-level semantic interaction in dense scenes.
- We introduce a lightweight frequency-spatial unified selective module that selectively injects frequency-domain cues into shallow features to enhance spatial details while suppressing background noise.
- We incorporate a dilated depthwise multi-head self-attention module to efficiently expand the receptive field and capture long-range context without quadratic complexity.

II. Related Work

A. Tiny Object Detection

Tiny object detection remains a fundamental challenge in UAV imagery, where targets occupy very limited pixels, exhibit irregular orientations, and suffer from severe semantic sparsity. Recent studies have explored multi-scale feature fusion and context-aware designs [12], attention-based enhancement [13], and Transformer-based architectures [11], achieving notable improvements on UAV benchmarks. However, many existing methods incur substantial computational overhead, introduce redundant or noisy features in densely populated scenes, or still rely on low-order, pairwise interactions, limiting their ability to capture high-order structural relationships and contextual dependencies among dense tiny objects.

B. Real-Time End-to-End Object Detection

Transformer-based detectors have advanced end-to-end object detection by formulating it as a set prediction problem. DETR [4] and its variants improve accuracy and efficiency through deformable attention, efficient fusion, and enhanced matching strategies. To adapt Transformers to UAV imagery, recent works introduce density-aware queries and task-specific designs [14], while lightweight variants focus on reducing computational cost [15]. Nevertheless, most existing methods rely on low-order or sparse interactions, limiting their ability to model dense structural relationships among tiny objects in complex UAV scenes.

C. High-order Feature Fusion for Object Detection

Recent advances in graph and hypergraph neural networks [16] have demonstrated the effectiveness of high-order relational modeling in capturing complex dependencies beyond pairwise interactions. By allowing multiple entities to be jointly connected through hyperedges, these methods provide a principled way to model structured relationships among features, spatial regions, and semantic concepts. However, their application to object detection remains limited. Most existing approaches either focus on classification or segmentation tasks, or rely on static and offline hypergraph construction based on fixed feature similarity, which prevents adaptation to image-specific content and restricts end-to-end optimization.

III. Method

We present DH-DETR, a Dynamic Hypergraph Detection Transformer tailored for tiny object detection in UAV imagery. The overall architecture is illustrated in Fig. 1. Specifically, the framework consists of three main parts. First, a hypergraph-centric feature fusion module, namely the HHCA, is embedded into the feature pyramid to explicitly model high-order many-to-many relationships across feature levels and spatial regions, enabling holistic structural reasoning beyond pairwise interactions. Second, a lightweight FUSE module is applied to shallow pyramid levels to selectively inject frequency-domain cues, enhancing fine-grained spatial details while suppressing background noise in cluttered UAV scenes. Third, a D2MHSA block is incorporated into the fusion stage to efficiently enlarge the receptive field and capture long-range contextual dependencies with manageable computational complexity.

A. Dynamic Hypergraph High-Order Correlation Aggregator

Most object detectors rely on convolutions or first-order attention with conventional feature pyramid fusion, which limits their ability to model high-order structural relationships and often disrupts contextual continuity, resulting in inaccurate localization in dense small-object scenarios. To overcome this limitation, we introduce a dynamic HHCA module into the feature pyramid, which explicitly models global high-order correlations via differentiable hypergraph reasoning, while preserving fine-grained local cues through parallel local refinement and shortcut pathways Fig. 2.

Given multi-scale backbone features $\{P_5, P_4, P_3\}$, P_5 and P_3 are resized to match the spatial resolution of P_4, concatenated along the channel dimension, and projected by a 1×1 convolution to form $X \in \mathbb{R}^{B\times 3C\times H\times W}$. The fused representation is evenly split along the channel dimension as:

$$[X_1,\, X_2,\, X_3] = \mathrm{Split}_{\text{channel}}(X), \quad X_i \in \mathbb{R}^{B\times C\times H\times W} \tag{1}$$

corresponding to a low-order local branch, a high-order hypergraph reasoning branch, and a shortcut branch, respectively.

To explicitly model high-order correlations, the middle branch X_2 is processed by two parallel DyHGC3 blocks, producing hypergraph-enhanced features:

$$H_1 = \mathrm{DyHGC3}(X_2), \quad H_2 = \mathrm{DyHGC3}(X_2) \tag{2}$$

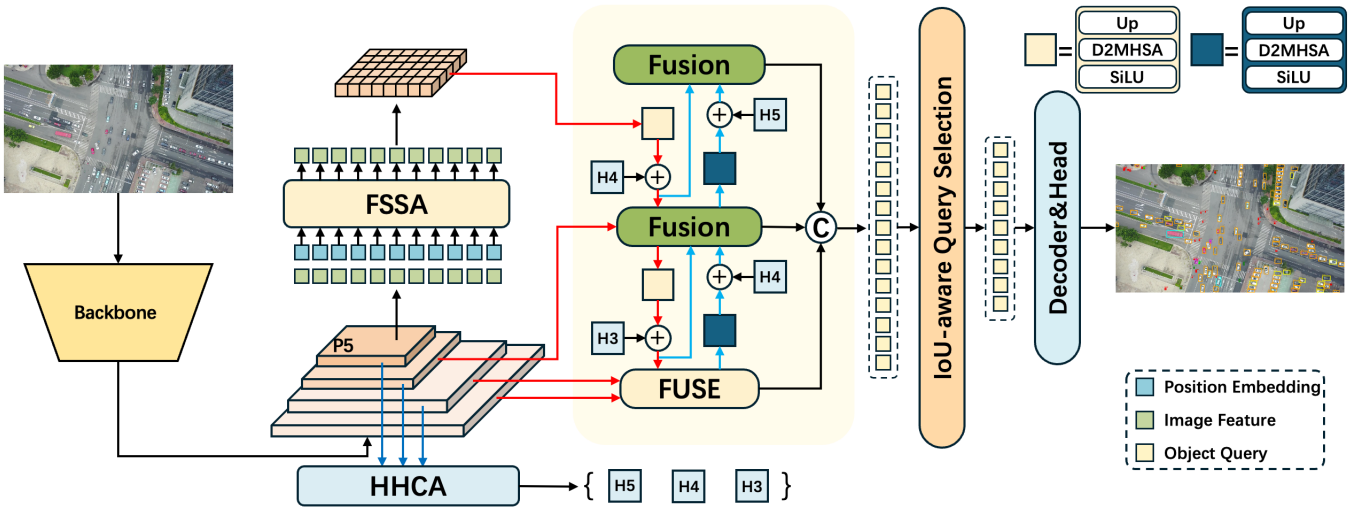

Fig. 1. Overall architecture of the proposed DH-DETR model.

which capture latent many-to-many relationships among spatial regions and feature channels. Their outputs are aggregated and fused with the input through a gated residual connection:

$$H' = X_2 + \tanh(g) \cdot \tfrac{1}{2}\,(H_1 + H_2) \tag{3}$$

where g is a learnable channel-wise gate that adaptively controls the contribution of high-order semantics.

In parallel, the low-order branch X_1 is refined by stacking n lightweight C3K2 modules, with each stage recursively defined as $L^{(i)} = \text{C3K2}(L^{(i-1)})$ and $L^{(0)} = X_1$, progressively enhancing local structures and boundary cues critical for small-object localization. Finally, the shortcut feature X_3, the fused high-order representation H', and the refined low-order features are concatenated and compressed by a 1×1 convolution to obtain the aggregated feature map:

$$H = \text{Conv}_{1\times1}\Big(\text{Concat}\big[X_3,\, H',\, L^{(1)}, \ldots, L^{(n)}\big]\Big) \tag{4}$$

The resulting feature is resized to generate multi-scale outputs $\{H_3, H_4, H_5\}$ aligned with the detection pyramid, enabling integration with subsequent refinement and decoding stages.

Each DyHGC3 block consists of dynamic hyperedge generation with location encoding and dynamic hypergraph convolution, as detailed in the following subsections.

1) Dynamic Hypergraph Generation with Location Encoding: Most hypergraph methods rely on offline and static incidence matrices built from feature similarity, which limits its adaptivity, gradient flow, and spatial awareness, resulting in unreliable associations in dense tiny-object scenarios. To address these limitations, we propose a location-aware hypergraph generation strategy that enables HHCA to dynamically construct a differentiable hypergraph during each forward pass by jointly modeling semantic similarity and geometric proximity.

Given an input feature map $X \in \mathbb{R}^{B\times C\times H\times W}$, we first reshape it into a token sequence $X' \in \mathbb{R}^{B\times N\times D}$ with $N = H \times W$, where each token corresponds to a spatial location. To capture global contextual information, we summarize the token set by concatenating its mean-pooled and max-pooled statistics:

$$Y = \left[\frac{1}{N}\sum_{i=1}^{N} X'_i,\ \max_i X'_i\right] \tag{5}$$

The resulting global descriptor is then augmented with 2-D sinusoidal positional encoding E to inject explicit spatial cues. A lightweight multi-layer perceptron $\phi(\cdot)$ transforms the position-enhanced descriptor into an offset term, which adapts a learnable prototype base P_0 into a batch-specific prototype set:

$$P = P_0 + \phi(Y + E), \quad P \in \mathbb{R}^{B\times M\times D} \tag{6}$$

This dynamic adaptation allows the prototype set to flexibly reflect scene-specific object distributions and spatial layouts.

Subsequently, we compute multi-head scaled dot-product similarities between tokens and prototypes, and normalize them via a row-wise softmax to obtain a soft incidence matrix:

$$A = \text{softmax}_j\left(\frac{(X'W^q)(PW^k)^\top}{\sqrt{d_k}}\right), \quad A \in \mathbb{R}^{B\times N\times M} \tag{7}$$

where W^q and W^k are learnable projections and d_k denotes the per-head feature dimension. The resulting matrix A encodes soft hyperedge memberships for each token, serving as a differentiable incidence matrix.

2) Dynamic Hypergraph Convolution: After the adaptive hyperedges are constructed, we perform Dynamic Hypergraph Convolution (DHGC) to propagate high-order information between vertices and hyperedges. Unlike conventional graph convolution that is restricted to pairwise message passing, DHGC enables many-to-many interactions by explicitly modeling vertex–hyperedge–vertex relations, which is particularly suitable for dense small-object scenarios.

Given the position-augmented token sequence $\tilde{Y} \in \mathbb{R}^{B\times N\times D}$ and the soft incidence matrix $A \in \mathbb{R}^{B\times N\times M}$, the DHGC operator proceeds in three stages. First, each hyperedge

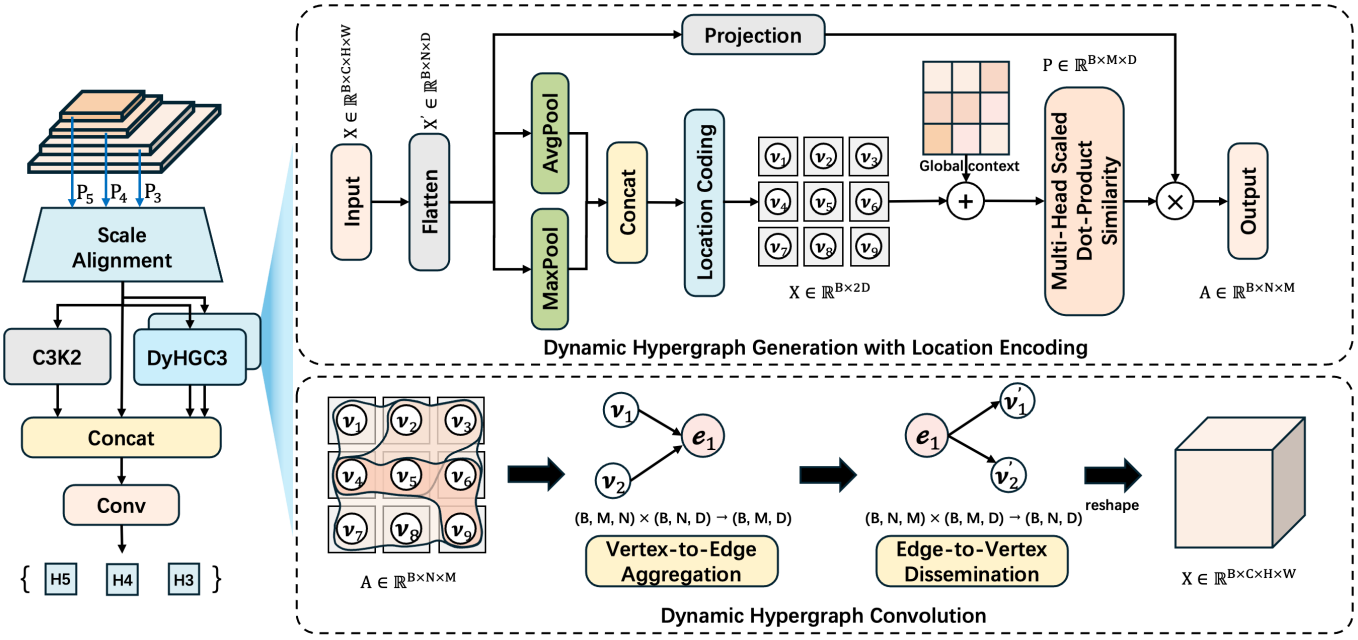

Fig. 2. Structure of the proposed HHCA module.

aggregates information from its incident vertices, forming hyperedge-level descriptors via weighted pooling:

$$H_e = A^\top \tilde{Y}, \quad H_e \in \mathbb{R}^{M \times D} \tag{8}$$

This operation summarizes the collective semantics of multiple spatial regions associated with the same hyperedge.

Next, the hyperedge descriptors are projected into the latent space through a linear transformation followed by a non-linear activation, and redistributed back to the vertex space to update vertex representations:

$$H'_e = \phi_e(H_e), \quad Z = \phi_v(AH'_e) \tag{9}$$

where $\phi_e(\cdot)$ and $\phi_v(\cdot)$ denote learnable projection layers. This bidirectional aggregation enables information exchange across distant spatial locations through shared hyperedges, effectively capturing high-order structural dependencies. To preserve original feature semantics and stabilize optimization, the updated vertex features are combined with the input tokens via a residual connection:

$$Z' = \tilde{Y} + Z \tag{10}$$

The fused representation is then reshaped back to the spatial domain, yielding a feature map $F \in \mathbb{R}^{B \times D \times H \times W}$.

Finally, a lightweight spatial refinement branch is applied to inject explicit coordinate priors. Specifically, coordinate attention followed by a depthwise convolution is used to enhance spatial sensitivity, and its contribution is adaptively regulated by a learnable gate:

$$F' = \tanh(g) \cdot \mathrm{DWConv}\big(\mathrm{CoordAtt}(Z')\big) \tag{11}$$

where g is initialized to zero, allowing the model to start from an identity mapping and progressively incorporate spatial refinement during training.

B. Frequency–spatial Unified Selective Module

Tiny-object detection relies heavily on fine-grained low-level features that preserve boundaries and textures. However, most existing detectors either discard shallow layers to reduce computational cost or directly fuse them into higher stages, which often amplifies noise and redundant details, thereby degrading localization accuracy in cluttered backgrounds. To overcome these limitations, we propose a lightweight FUSE module, which is applied to shallow pyramid levels (P_2 and P_3), as illustrated in Fig. 3.

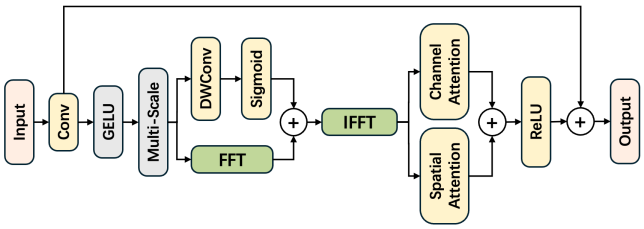

Fig. 3. Structure of the proposed FUSE module.

Given an input feature map $X \in \mathbb{R}^{B \times C \times H \times W}$, FUSE projects it into two parallel branches via 1×1 convolutions:

$$y = \mathrm{Conv}^{(y)}_{1\times1}(X), \quad x_{\mathrm{f}} = \mathrm{Conv}^{(f)}_{1\times1}(X) \tag{12}$$

where y preserves spatial details and x_{f} serves as a compact embedding for frequency modeling.

To capture multi-scale textures, the frequency branch aggregates local patterns using lightweight depthwise convolutions

with different receptive fields:

$$u = \mathrm{DW}_{1\times1}(x_\mathrm{f}) + \mathrm{DW}_{3\times3}(x_\mathrm{f}) + \mathrm{DW}_{5\times5}(x_\mathrm{f}) \tag{13}$$

The aggregated feature u is transformed into the frequency domain, modulated by a learnable complex gate conditioned on the residual branch, and mapped back to the spatial domain:

$$\hat{u} = \mathrm{IFFT}\big(\mathrm{FFT}(u) \odot G(y)\big) \tag{14}$$

which enables selective enhancement of informative low-frequency components, while suppressing background-dominated spectral responses. Channel attention and spatial gating are then applied to reweight the enhanced feature:

$$\widetilde{u} = \hat{u} \odot \mathrm{CA}(\hat{u}) \odot \mathrm{SG}(\hat{u}) \tag{15}$$

where CA and SG denote channel attention and spatial gating, respectively, which adaptively emphasize informative channels and spatial regions after spectral modulation. Finally, $\widetilde{u}$ is fused with the residual branch using learnable scalar weights:

$$z = \alpha \odot \widetilde{u} + \beta \odot y, \qquad \alpha = 0,\ \beta = 1 \tag{16}$$

allowing FUSE to start as an identity mapping and progressively inject spectral context during training.

C. Dilated Depthwise Multi-Head Self-Attention

Standard depthwise convolutions lack long-range modeling capability, while global self-attention incurs prohibitive computational cost. To address this trade-off, we introduce the Dilated Depthwise Multi-Head Self-Attention (D2MHSA) module, which enriches QKV features with dilated contextual cues at negligible extra overhead, enabling efficient long-range dependency modeling during multi-scale feature pyramid fusion, as illustrated in Fig. 4.

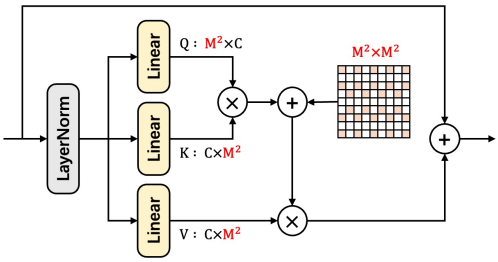

Fig. 4. Structure of the proposed D2MHSA module.

Given an input feature map $X \in \mathbb{R}^{B\times C\times H\times W}$, D2MHSA adopts a two-stage design that combines dilated local aggregation with lightweight self-attention.

In the first stage, the input features are projected and filtered using a 1×1 convolution followed by a dilated depthwise convolution, producing the query, key, and value embeddings:

$$[Q, K, V] = \mathrm{DWConv}^{3\times3,\,d=2}\big(\mathrm{Conv}^{1\times1}(X)\big) \tag{17}$$

where the dilation factor enlarges the effective receptive field to 5×5, allowing each token to encode broader local context prior to attention computation without introducing additional quadratic complexity.

In the second stage, the projected features are reshaped into multi-head representations and ℓ_2-normalised to enable cosine-based attention:

$$Q_h, K_h = \mathrm{Norm}(\mathrm{reshape}(Q, K)),\ V_h = \mathrm{reshape}(V) \tag{18}$$

followed by temperature-scaled multi-head attention,

$$A_h = \mathrm{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_k}} \cdot \tau_h\right), \qquad O_h = A_h V_h \tag{19}$$

where $d_k = C/H$ is the per-head feature dimension and τ_h is a learnable temperature that adaptively controls attention sharpness for each head.

The attended features from all heads are concatenated, reshaped back to the spatial domain, and fused through a 1×1 convolution:

$$Y = \mathrm{Conv}^{1\times1}\Big(\mathrm{reshape}^{-1}\big(\mathrm{Concat}_{h=1}^{H}(O_h)\big)\Big) \tag{20}$$

yielding the final output $Y \in \mathbb{R}^{B\times C\times H\times W}$. A residual connection enables the module to initialise as an identity mapping and progressively inject dilated contextual cues during training.

IV. Experiments

A. Datasets

We evaluate our method on three public UAV benchmarks: **VisDrone** [17] is a large-scale urban UAV benchmark with 10 categories, characterized by extreme scale variation, cluttered backgrounds, and frequent occlusions, and contains 6,471 training, 548 validation, and 1,610 test images. **UAVDT** [18] focuses on vehicle detection from aerial platforms and includes approximately 80,000 frames with 0.84 million annotations across three categories, posing challenges due to small object sizes, large viewpoint variations, and severe motion blur. **AI-TOD** [19] is specifically designed for tiny object detection, where the average object size is only 12.8 pixels, and features dense object distributions and large inter-class scale variations, with 11,214 training, 2,804 validation, and 14,018 test images.

B. Experiment configuration and evaluation metrics

1) Experiment configuration: Following D-FINE [20], our model adopts HGNetv2 as the CNN backbone, a single-layer transformer encoder, and a DETR-style transformer decoder. All experiments are conducted on an NVIDIA RTX 3090 GPU using PyTorch 2.3.0 and CUDA 12.4. The model is trained for 132 epochs with input images resized to 640×640 and a batch size of 4. Standard data augmentations from D-FINE, including random scale jittering, horizontal flipping, random cropping, and mild color perturbations, are applied to enhance robustness and generalization.

TABLE I
DETECTION PERFORMANCE COMPARISON ON THE VISDRONE DATASET.

Method	Reference	AP	AP_{50}	AP_S	AP_M	AP_L	Params (M)	GFLOPs
CenterNet [21]	NeurIPS2019	23.9	45.5	14.4	36.3	44.9	32.0	206.1
YOLOv8-M [22]	–	24.0	40.2	14.8	36.0	44.6	25.9	78.9
YOLOv8-L [22]	–	26.1	42.7	14.9	36.6	45.0	43.7	165.2
YOLOv9-M [23]	ECCV2025	25.1	41.7	15.4	38.0	45.7	20.1	76.8
YOLOv10-M [24]	arXiv2024	24.3	40.4	14.8	36.7	42.8	15.4	59.1
YOLOv10-L [24]	arXiv2024	26.3	43.1	15.1	36.9	43.0	24.4	120.3
YOLOv11-M [25]	arXiv2024	25.2	41.7	16.3	37.0	46.5	20.0	67.7
YOLOv12-M [26]	arXiv2025	25.1	41.7	15.6	37.4	48.4	20.2	67.5
DETR [4]	ECCV2020	24.1	40.1	12.4	30.8	38.1	60.0	187.0
Deformable DETR [9]	ICLR2021	27.1	42.2	19.6	34.4	40.2	40.0	173.0
DINO [10]	ICLR2023	29.4	46.2	22.9	38.4	44.5	47.0	279.0
RT-DETR-R18 [11]	CVPR2024	26.7	44.6	18.9	38.0	44.9	20.0	60.0
RT-DETR-R50 [11]	CVPR2024	28.4	47.0	19.0	38.9	47.2	42.0	136.0
Sparse DETR [6]	ICLR2022	27.3	42.5	20.2	35.0	40.2	40.9	121.0
D-FINE-M [20]	ICLR2025	25.7	42.9	17.7	35.2	46.7	19.1	56.4
D-FINE-L [20]	ICLR2025	26.1	44.0	18.3	35.9	46.9	30.7	90.8
DH-DETR (Ours)	–	**28.8**	**47.7**	**21.0**	**38.8**	**49.8**	23.9	91.2

TABLE II
DETECTION PERFORMANCE ON UAVDT AND AI-TOD DATASETS.

Method	UAVDT (%)		AI-TOD (%)		Params (M)
	AP	AP_{50}	AP	AP_{50}	
DETR [4]	11.3	23.4	16.2	42.1	60.0
YOLOv8-M [22]	13.9	28.7	13.6	29.0	25.9
YOLOv10-M [24]	15.1	29.7	14.3	33.1	15.4
EfficientDet [27]	12.2	31.4	16.1	39.0	52.0
Deformable DETR [9]	18.6	31.9	17.0	45.9	40.0
RT-DETR-R18 [11]	11.1	20.7	20.8	54.4	20.0
SGF-YOLO [28]	15.1	28.3	25.1	48.9	34.7
SO-DETR-R18 [29]	13.1	22.9	19.3	49.1	20.5
D-FINE-M [20]	15.6	29.8	24.4	54.6	19.2
DH-DETR (Ours)	**19.1**	**32.3**	**28.1**	**57.8**	23.9

2) Evaluation Metrics: We adopt the standard COCO evaluation metrics, including AP averaged over IoU thresholds from 0.50 to 0.95 with a step of 0.05, and AP_{50} at an IoU threshold of 0.50. To analyze scale-sensitive performance, we further report AP_S, AP_M, and AP_L, corresponding to objects with areas smaller than 32^2, between 32^2 and 96^2, and larger than 96^2 pixels, respectively.

C. Comparison with state-of-the-art models

1) Results on the VisDrone Dataset: Table I summarizes the performance of representative state-of-the-art detectors on the VisDrone dataset. Existing methods achieve AP ranging from 23.9% to 29.4%, with many efficient models exhibiting limited accuracy in dense UAV scenarios, particularly for small objects. While recent high-capacity models such as DINO reach strong accuracy (29.4% AP), this improvement comes at the cost of substantially increased computation (279 GFLOPs), which limits practical deployment. In contrast, DH-DETR achieves 28.8% AP with a balanced computational cost of 91.2 GFLOPs, outperforming most competing approaches under comparable efficiency budgets. Notably, DH-DETR attains the highest AP_{50} (47.7%) and delivers strong small-object performance with an AP_S of 21.0%. Compared with the baseline D-FINE-M, DH-DETR yields a +3.1 AP improvement, demonstrating its effectiveness in enhancing fine-grained localization and representation for dense small-object detection in UAV imagery.

2) Results on the UAVDT and AI-TOD Dataset: To further evaluate the effectiveness and generalization ability of the proposed DH-DETR, we conduct comprehensive comparisons with representative state-of-the-art detectors on the UAVDT and AI-TOD datasets, as reported in Table II. These two benchmarks are particularly challenging due to severe scale variation, dense object distributions, and extremely small object sizes. On the UAVDT dataset, DH-DETR achieves the best overall performance, reaching 19.1% AP and 32.3% AP_{50}, outperforming all compared methods. Notably, DH-DETR surpasses strong baselines such as SGF-YOLO and D-FINE-M while maintaining a moderate computational cost, indicating its superior capability in handling dense vehicle detection under complex aerial viewpoints. On the more challenging AI-TOD dataset, which focuses on extremely tiny objects, DH-DETR consistently delivers the highest accuracy, achieving 28.1% AP and 57.8% AP_{50}. Compared with existing lightweight detectors, DH-DETR demonstrates a clear advantage in both localization accuracy and robustness to background clutter, highlighting the effectiveness of high-order relational modeling for tiny-object detection.

D. Ablation Study

In this section, we conduct ablation studies on the VisDrone dataset to systematically evaluate the contribution of each proposed module. The baseline corresponds to the D-FINE-M detector without any of the newly introduced components.

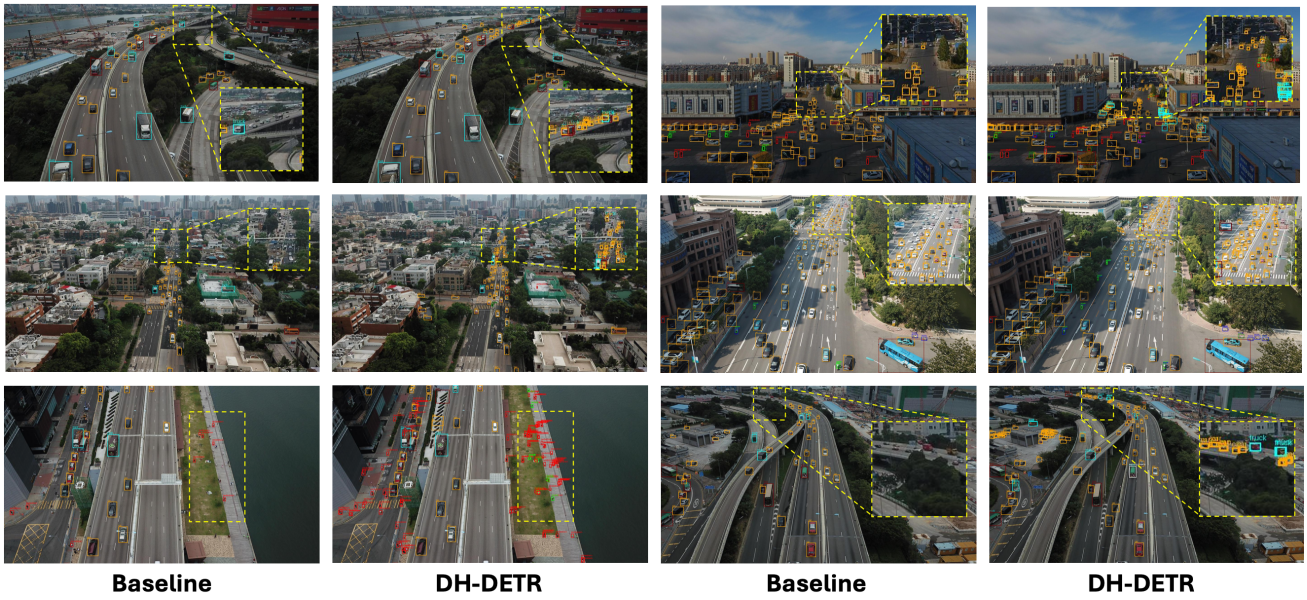

Fig. 5. Visual comparison results of DH-DETR and Baseline methods on the VisDrone dataset.

TABLE III
ABLATION RESULTS ON THE VISDRONE DATASET.

Method	AP (%)	AP_{50} (%)	Params (M)
Baseline	25.7	42.9	19.2
+ HHCA	27.5	46.2	21.9
+ FUSE	27.1	45.2	21.3
+ D2MHSA	26.8	43.9	19.9
+ HHCA + FUSE	28.1	47.0	23.2
+ HHCA + D2MHSA	27.9	46.8	22.4
+ HHCA + FUSE + D2MHSA	**28.8**	**47.7**	23.9

We first evaluate the impact of each individual component through controlled ablation experiments. Incorporating the dynamic HHCA module into the baseline yields a clear performance improvement, increasing AP from 25.7% to 27.5%. This gain demonstrates that explicitly modeling high-order structural relationships across feature levels and spatial regions effectively strengthens feature interactions, which is particularly beneficial in dense UAV scenes with complex object layouts. Introducing the FUSE module independently improves AP to 27.4%, demonstrating its ability to enhance fine-grained spatial details while suppressing background interference. Similarly, adding the D2MHSA module results in a consistent performance gain, validating its effectiveness in capturing long-range contextual dependencies with limited computational overhead.

We further investigate the complementary effects of different modules. Combining HHCA and FUSE leads to an AP of 28.1%, outperforming all single-module variants, which suggests that high-order relational modeling and frequency-aware feature enhancement are mutually beneficial. Finally, integrating all three components achieves the best overall performance, with AP reaching 28.8% and AP_{50} improving to 47.7%. Compared with the baseline, this corresponds to an absolute improvement of +3.1 AP. Although the full model incurs a modest increase in computational cost, the performance gain is substantial and well justified for dense small-object detection. These results demonstrate that each proposed component contributes positively to detection accuracy, and their joint integration effectively addresses the challenges of fine-grained localization and representation in UAV imagery.

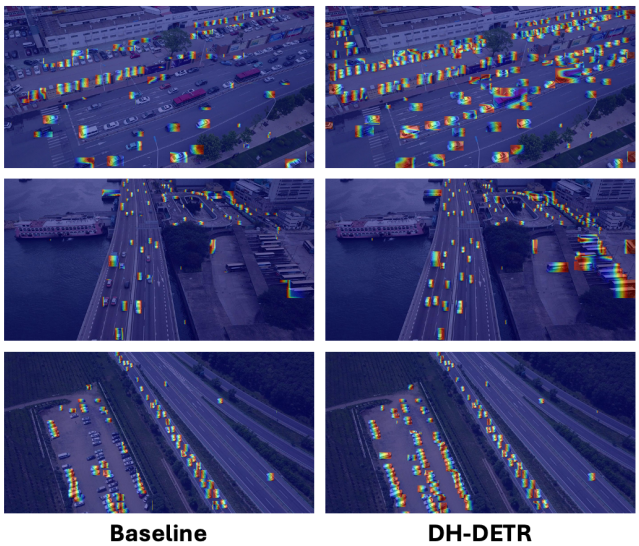

Fig. 6. Attention heatmap visualization of DH-DETR.

E. Visualization Analysis

To further validate the effectiveness of DH-DETR, Fig. 5 presents qualitative comparisons with the baseline D-FINE

on the VisDrone dataset. DH-DETR consistently produces more accurate and robust detections, particularly for dense small objects in UAV imagery, by reducing false positives and generating tighter bounding boxes in crowded scenes. Moreover, the attention heatmaps in Fig. 6 reveal that DH-DETR focuses more precisely on dense small-object regions, clearly highlighting fine-grained object boundaries while suppressing irrelevant background responses. This behavior indicates that DH-DETR is able to exploit discriminative spatial cues and structured context, which is critical for reliable detection in complex urban UAV scenarios.

V. Conclusion

In this paper, we presented DH-DETR, a novel end-to-end detection framework for reliable tiny object detection in UAV imagery. DH-DETR integrates three core components: the HHCA module, the FUSE module, and the D2MHSA module, which jointly enhance early feature retention, model high-order structural dependencies, and enable efficient long-range contextual reasoning for dense aerial scenes. Extensive experiments on VisDrone, UAVDT, and AI-TOD datasets demonstrate that DH-DETR consistently outperforms DETR-based counterparts in terms of detection accuracy, particularly for small objects, while maintaining competitive computational efficiency. These results highlight the effectiveness of dynamically modeling high-order correlations and selectively enriching spatial detail with frequency cues in dense aerial scenarios.

Looking forward, future work may extend the proposed hypergraph-based design to instance or panoptic level detection, explore adaptive bounding box representations, or integrate foundation models to improve cross-domain generalization. We believe DH-DETR provides valuable insights into designing efficient yet expressive transformer-based detectors for real-world UAV applications.

Acknowledgments

The authors thank the Natural Science Foundation of Hebei Province, China (Project No. F2024201012) for financial support, and the High-Performance Computing Center of Hebei University for providing computational resources.

References

[1] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 91–99.

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. European Conf. Computer Vision (ECCV)*, 2020, pp. 213–229.

[5] F. Li, H. Zhang, S. Liu, J. Zhang, L. Ni, Y. Hou, T. Wang, Y. Qiao, H. Li, and J. Sun, "DN-DETR: Accelerate DETR training by introducing query denoising," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13619–13627.

[6] D. Meng, X. Chen, Z. Fan, G. Xu, Y. Zhang, X. Li, Z. Yuan, J. Sun, and J. Wang, "Sparse DETR: Efficient end-to-end object detection with learnable sparsity," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[7] X. Wang, X. Zhang, and X. Li, "UFA-DETR: Uncertainty-aware feature aggregation for tiny object detection," *arXiv preprint*, 2023.

[8] Y. Liu, H. Wang, and Q. Chen, "Lite-DETR: A lightweight transformer for object detection," in *Proc. AAAI Conf. Artificial Intelligence*, 2022.

[9] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[10] Y. Zhang, H. Fang, J. Wang, X. Bai, H. Lu, and H. Hu, "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2023.

[11] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, and Q. Dang, "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 16965–16974.

[12] Q. Li, W. Zhang, Y. Liu, and J. Wang, "ESOD: Efficient small object detection on high-resolution images," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023.

[13] F. C. Akyon, S. O. Altinuc, and A. Temizel, "Slicing aided hyper inference for small object detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 123–132.

[14] J. Li, H. Xu, W. Zhang, and L. Wang, "UAV-DETR: Uncertainty-aware vision transformer for small object detection in UAV imagery," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2023.

[15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.

[16] S. Bai, Z. Zhang, and P. H. S. Torr, "Hypergraph convolution and hypergraph attention for semi-supervised classification," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12033–12042.

[17] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," in *Proc. European Conf. Computer Vision Workshops (ECCVW)*, 2018.

[18] D. Du, Y. Qi, H. Yu, Y. Yang, K. Du, G. Zhang, Y. Wang, J. Zhang, and L. Wen, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. European Conf. Computer Vision (ECCV)*, 2018.

[19] Q. Li, W. Zhang, Y. Liu, and J. Wang, "AI-TOD: Tiny object detection in aerial images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2021.

[20] Y. Peng, H. Li, P. Wu, Y. Zhang, X. Sun, and F. Wu, "D-FINE: Redefine regression task of DETRs as fine-grained distribution refinement," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2025.

[21] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019, pp. 12897–12906.

[22] G. Jocher and Ultralytics, "YOLOv8: Ultralytics YOLO," *GitHub repository*, 2023.

[23] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv9: Learning what you want by learning where to look," *arXiv preprint arXiv:2402.13616*, 2024.

[24] W. Wang, J. Chen, X. Zhu, Y. Li, and H. Zhang, "YOLOv10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.

[25] Ultralytics, "YOLOv11: Next generation ultralytics YOLO detector," 2024.

[26] Ultralytics, "YOLOv12: Ultralytics YOLO for advanced object detection," 2025.

[27] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10781–10790.

[28] Y. Cheng, T. Wang, and W. Zhang, "Real-time UAV small object detection: An efficient approach using SGFNet and dynamic loss optimization," *Digital Signal Processing*, 2025, Art. no. 105543.

[29] C. Wang, Y. Liang, and H. Xu, "SO-DETR: Small object detection transformer with progressive resolution learning," *Pattern Recognition*, vol. 140, p. 109543, 2023.