Hierarchical Task-Decoupled Representation Learning for Efficient UAV Object Detection

1st Yaqi Liu
School of Cyber Security and Computer
Hebei University
Baoding, China
kiki6423@163.com

2nd Wenzhu Yang*
School of Cyber Security and Computer
Machine Vision Engineering Research Center
Hebei University
Baoding, China
wenzhuyang@hbu.edu.cn

***Abstract*—Object detection in UAV scenarios remains a significant challenge due to small objects, extreme scale variations, and complex backgrounds. Existing research often adopts uniform architectural designs, overlooking the distinct representational bottlenecks across hierarchical layers. Deep layers handle global semantics but suffer from severe feature homogenization and semantic degradation, while shallow layers preserve fine-grained details but are constrained by limited receptive fields and background interference. Moreover, complex designs increase computational overhead, hindering real-time deployment. To address these issues, we propose a lightweight Hierarchical Task-Decoupled representation learning network (HTDNet), consisting of three core modules: the Diverse Representation Reconstruction (DRR) module reconstructs structured semantics through feature decomposition and heterogeneous recomposition, mitigating feature homogenization in deep layers; the Cross-Layer Prior Modulation (CLPM) module leverages both semantic and spatial priors to guide targeted feature recalibration for small objects, suppressing background interference in shallow layers; and the Saliency-Guided Purification (SGP) module eliminates representational conflicts during multi-scale fusion through nonlinear modulation, enhancing small object saliency. Experimental results on VisDrone show that HTDNet outperforms the baseline by 6.3% mAP_{50} with only 45% of the parameters, achieving an optimal accuracy-efficiency trade-off, offering distinct advantages for real-time edge deployment. The generalization of HTDNet is further validated across AI-TOD and TinyPerson datasets.**

***Index Terms*—UAV images, Object detection, Small object, Lightweight, Task decoupling**

I. Introduction

In recent years, Unmanned Aerial Vehicle (UAV) technology has demonstrated immense potential in fields such as urban surveillance [1], disaster response [2], and precision agriculture [3]. Leveraging their flexible observation perspectives, UAVs can rapidly acquire high-resolution visual data from various altitudes, as illustrated in Fig. 1. Object detection [4] aims to automatically locate and classify objects within complex backgrounds, serving as a critical technology for UAVs to parse visual information in large-scale aerial scenes. Despite the rapid advancement of general object detection frameworks, accurately identifying small objects in UAV scenarios remains a major challenge.

Deep learning-based object detectors comprise two primary categories: two-stage and single-stage methods. Two-stage frameworks, such as Faster R-CNN [5] and Cascade R-CNN [6], achieve high localization accuracy through proposal-based refinement. In contrast, single-stage methods including the YOLO [7] series and SSD [8], enhance inference speed by directly regressing detections from global features. To address challenges in aerial imagery such as small objects, complex backgrounds, and high scale variability, existing research typically adopts two strategies: feature enhancement to improve fine-grained perception, such as attention mechanisms [9], and multi-scale fusion to enrich spatio-semantic representations via feature pyramids [10] and spatial alignment [11].

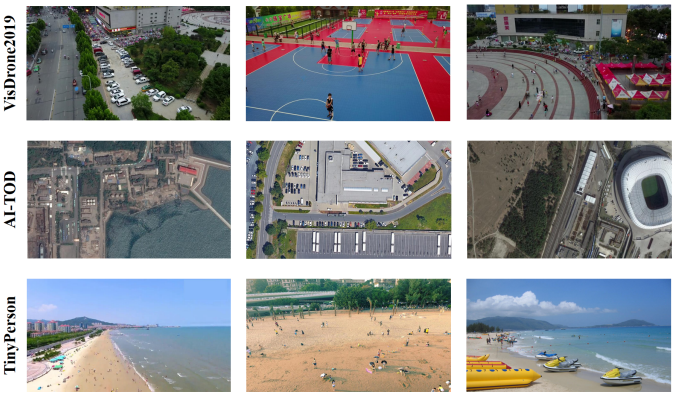

Fig. 1. Representative samples from the aerial benchmarks, illustrating diverse observation perspectives and altitudes in UAV-related scenarios.

However, contemporary research primarily adopts uniform architectural designs, overlooking the representational bottlenecks inherent in different levels of the feature pyramid. While deep layers capture global semantics, continuous downsampling and convolutional stacking often lead to feature homogenization, causing severe channel redundancy and semantic degradation, thereby limiting representational capacity in complex scenes. Shallow layers, though preserving spatial details, are highly susceptible to background noise due to limited receptive fields, resulting in false positives. Moreover, sophisticated designs incur prohibitive computational costs. Achieving high-precision real-time detection on resource-constrained UAV platforms remains a significant challenge.

To address these issues, we propose a Hierarchical Task-Decoupled Representation Learning Network (HTDNet),

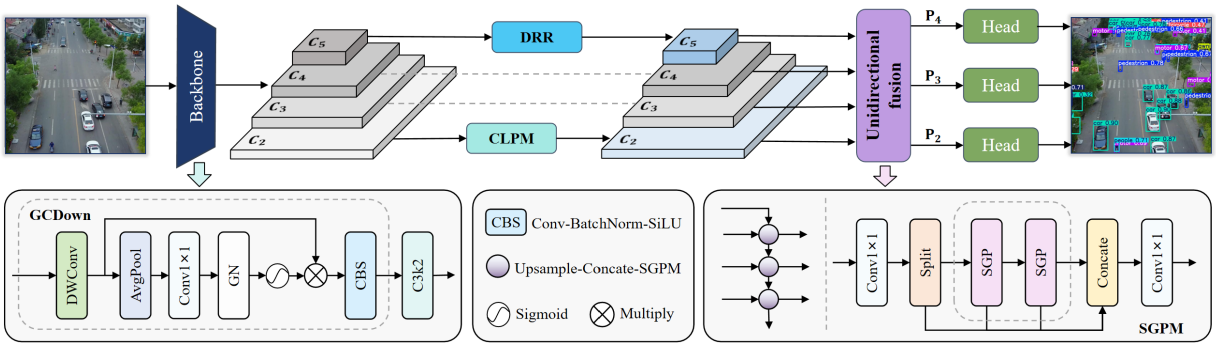

Fig. 2. Overview of our proposed HTDNet architecture.

adopting differentiated enhancement strategies for inherent representation flaws across deep, shallow, and neck layers. First, the Diverse Representation Reconstruction (DRR) module achieves efficient global modeling and reshapes semantic diversity via feature decomposition and heterogeneous restructuring, mitigating downsampling-induced homogenization. Second, the Cross-Layer Prior Modulation (CLPM) module leverages semantic and spatial priors for denoising and representation recalibration, enabling targeted enhancement of small objects. Finally, the Saliency-Guided Purification (SGP) module employs soft-gating for nonlinear filtering, alleviating cross-scale feature aliasing and enhancing discriminative ability for small objects.

The main contributions of this paper are summarized as follows:

- We propose HTDNet, which overcomes inherent multi-level representation bottlenecks through task decoupling and differentiated enhancement strategies. Leveraging a lightweight architecture, HTDNet significantly boosts detection performance with minimal computational cost.
- To enhance small-object detection, we design three core components: DRR reconstructs structured semantics via channel decomposition and restructuring, mitigating semantic degradation and feature homogenization in deep layers. CLPM uses dual priors to separate foreground and background, enabling selective refinement of shallow fine-grained details. SGP employs nonlinear soft-gating to alleviate cross-scale representation conflicts in feature fusion, enhancing the saliency response for small objects.
- Extensive experiments on multiple challenging aerial benchmarks demonstrate that HTDNet significantly improves small-object detection performance, achieving a superior balance between accuracy and inference efficiency on resource-constrained platforms.

II. Related Work

A. Feature Reconstruction

Feature reconstruction aims to reshape or recover high-quality representations from degraded or redundant feature maps. EFR-ACENet [12] utilizes an explicit feature reconstruction module to model inter-channel dependencies, restoring lost spatial information via context enhancement. TELOD [13] reconstructs discriminative features from low-resolution inputs through progressive pixel-level refinement. To address the information loss in UAV imagery, GAN-FPN [14] employs Generative Adversarial Networks to reconstruct high-resolution feature pyramids, recovering the structural details of small objects from low-level features. However, these methods focus on recovering spatial details, while overlooking the semantic homogenization of deep features.

B. Attention Mechanisms

Attention mechanisms assign weights to feature map regions or channels, enabling the extraction of key information from complex backgrounds. To enhance spatial perception of small targets, EPAF-Net [15] introduces a positional attention module that captures geometric information by orthogonally compressing features across vertical, horizontal, and channel dimensions. LUD-YOLO [16] integrates multi-scale attention units within residual structures for adaptive selection of discriminative features. VSTDet [17] simulates biological visual pathways via a reverse attention module, enhancing saliency perception at low resolutions. To improve global context modeling, MI-DETR [18] employs an orthogonal attention mechanism, constructing weight distributions to reduce feature redundancy. However, complex designs incur heavy computational costs, hindering real-time edge deployment.

C. Cross-Layer Fusion

Cross-layer fusion aims to integrate multi-scale information to handle dramatic scale variations in UAV images. To enhance information flow, Gold-YOLO [19] achieves global fusion by aggregating and redistributing features across different levels. FSENet [20] utilizes cross-layer redundancy suppression to filter background noise while strengthening the discriminative representation of small targets. Furthermore, MFS-YOLO [21] implements dynamic cross-layer feature flow association via parallel branches to address feature sparsity. However, cross-layer fusion inevitably induces representational conflicts, interfering with precise object localization.

III. Proposed Method

The overall framework of HTDNet is illustrated in Fig. 2. The network comprises three core components: the CSP-

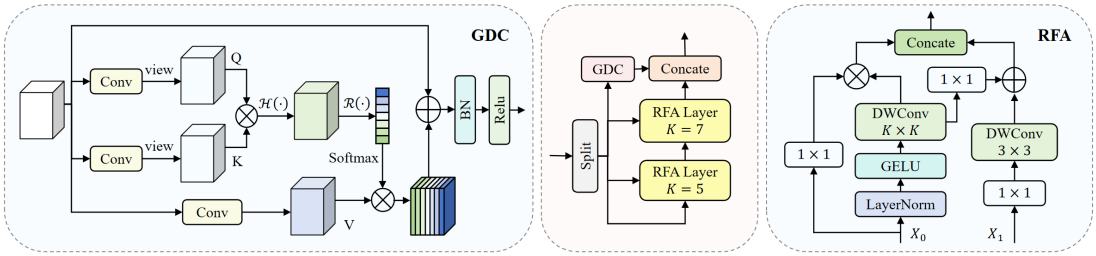

Fig. 3. Overview of the proposed DRR module.

DarkNet backbone extracts multi-scale features and performs hierarchically decoupled representation learning via DRR and CLPM, mitigating deep-layer semantic homogenization and shallow-layer noise interference. The neck employs unidirectional feature fusion, with the SGP module resolving representation conflicts during cross-scale fusion to enhance small object saliency. Furthermore, the detection heads are shifted from the $\{P_3, P_4, P_5\}$ layers to the $\{P_2, P_3, P_4\}$ layers to capture higher-resolution spatial details. In addition, a lightweight Context-Guided Downsampling (CGDown) module is introduced, replacing redundant strided convolutions with depthwise convolutions and global context computation. Through task-relevant feature enhancement, HTDNet addresses inherent representation bottlenecks across network layers, achieving a balance between computational efficiency and accuracy.

A. Diverse Representation Reconstruction (DRR)

To overcome the issues of feature uniformity and semantic degeneration in traditional feature extraction methods, we propose the Diverse Representation Reconstruction Module (DRR), as shown in Fig. 3. DRR generates informative feature representations that capture local details, global semantics, and channel discriminability through channel decomposition and feature reconstruction. Specifically, DRR first divides the input features into four parts: $\{X_0, X_1, X_2, X_3\}$. Among these, $\{X_0, X_1, X_2\}$ are passed into the Receptive Field Aggregation Unit (RFA), while $\{X_3\}$ is processed by the Global Feature Refinement Unit (GFR) for efficient context modeling. The structure is represented as follows:

$$X_{\text{out}} = \text{Concat}\left(\text{GFR}(X_3),\ \text{RFA}_2\left(\text{RFA}_1(X_0, X_1),\ X_2\right)\right) \quad (1)$$

RFA adopts a dual-path design for differentiated processing. After layer normalization, the modulation path generates modulated signals via 1×1 convolution, GELU, and a large kernel convolution, which are then element-wise multiplied with the feature projected by 1×1 convolution:

$$G = \text{DWConv}_{K\times K}\left(\text{GELU}\left(\text{Conv}_{1\times1}(X_i)\right)\right) \quad (2)$$

$$X_{\text{gate}} = G \odot \text{Conv}_{1\times1}(X_i) \quad (3)$$

where $\text{Conv}_{1\times1}(\cdot)$ denotes the 1×1 convolution operation, $\text{GELU}(\cdot)$ represents the nonlinear activation function, $\text{DWConv}_{K\times K}(\cdot)$ indicates the depthwise separable convolution with a kernel size of $K \in [5, 7]$, $\odot$ denotes the element-wise multiplication, and X_i represents either X_0 or the output from the preceding RFA module.

The local path extracts fine-grained local information through a 1×1 convolution and 3×3 depthwise convolution, and the information is fused through a residual connection with the modulation signal:

$$X_{\text{local}} = \text{DWConv}_{3\times3}\left(\text{Conv}_{1\times1}(X_j)\right) + \text{Conv}_{1\times1}(G) \quad (4)$$

where $X_j \in \{X_1, X_2\}$. Finally, the features from the two paths are aggregated across channels. We stack two RFA units serially to achieve hierarchical and progressive feature enhancement, maintaining computational efficiency while gradually expanding the receptive field from 5×5 to 7×7, completing dynamic fusion of feature channels.

In parallel, the X_3 feature enters the GFR and is first passed through a 3×3 convolution layer to generate Q, K, and V mappings. After dimensional adjustment, the dot-product similarity between Q and K is calculated, and the result $F \in \mathbb{R}^{C\times C\times H\times W}$ serves as the attention score matrix. Then, F is summed along the channel dimension and globally pooled along the remaining spatial dimensions to obtain the vector $F' \in \mathbb{R}^{C\times1}$. Each element of F' represents the average global relevance strength of the corresponding channel across the entire feature map, used for subsequent feature recalibration. After performing a softmax operation for normalization, dynamic weighting is applied to V, and the weighted result is residual-connected with the original input. Finally, BN and ReLU are applied for normalization and nonlinear activation. The process is formulated as follows:

$$A = \text{Softmax}\left(\mathcal{R}\left(\mathcal{H}\left(QK^{\top}/\sqrt{d}\right)\right)\right) \quad (5)$$

$$X_{\text{final}} = \text{ReLU}\left(\text{BN}\left(A \cdot V + X_3\right)\right) \quad (6)$$

where d is the feature dimension scaling factor; $\mathcal{R}(\cdot)$ and $\mathcal{H}(\cdot)$ are the summation operations over different dimensions; $\text{Softmax}(\cdot)$ represents the normalization operation; $\cdot$ denotes the dot product operation; $\text{BN}(\cdot)$ and $\text{ReLU}(\cdot)$ refer to Batch Normalization and ReLU activation functions, respectively.

DRR enhances channel diversity and feature quality in high-level feature maps through structured feature reconstruction, ultimately achieving an efficient synergistic representation of global context and local details, thereby alleviating the information loss of small objects.

B. Cross-Layer Prior Modulation (CLPM)

To address the challenges of contextual weakness and noise susceptibility in shallow feature maps for small objects, we

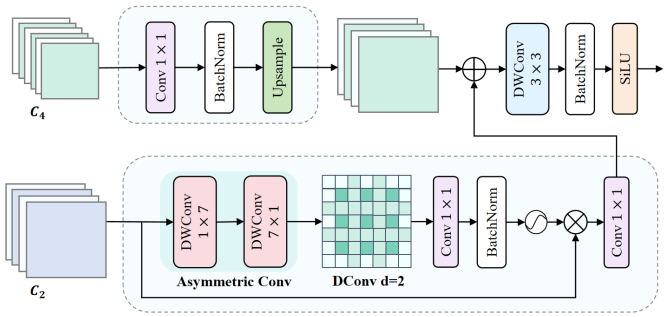

Fig. 4. Detailed architecture of the proposed CLPM module.

propose a Cross-Layer Prior Modulation (CLPM) module, as shown in Fig. 4. This module effectively suppresses background interference and highlights target saliency by leveraging priors from high-level semantics and spatial regions.

The CLPM module takes shallow features X_{shallow} and deep features X_{high} as inputs. The high-level features are first aligned across channels using a 1×1 convolution and Batch Normalization (BN), and then upsampled via bilinear interpolation to match the spatial resolution of the shallow features, generating the semantic prior map P_{sem}:

$$P_{\text{sem}} = \text{Upsample}(\text{BN}(\text{Conv}_{1\times1}(X_{\text{high}}))) \tag{7}$$

where $\text{Conv}_{1\times1}(\cdot)$ denotes the 1×1 convolution operation, $\text{BN}(\cdot)$ represents Batch Normalization, and $\text{Upsample}(\cdot)$ refers to bilinear upsampling. To minimize spatial alignment bias during cross-layer sampling, the C_4 layer is specifically selected as the input for X_{high}.

The shallow features are then passed through a spatial region branch to extract structural cues. Asymmetric and dilated convolutions are employed to construct the region prior P_{reg}. Asymmetric convolutions, utilizing 1×7 and 7×1 axial decompositions, capture directional contextual dependencies to provide local positional priors for small objects while maintaining low computational cost. The dilated convolution, with a dilation rate of 2, uses a depthwise separable convolution to enhance the saliency of small objects by expanding the local contrast range. Subsequently, a 1×1 convolution and BN are applied to weight the high-resolution features with an attention mask generated by the HardSigmoid function. The computation process is as follows:

$$X_{\text{reg}} = \text{Conv}_{3\times3,d=2}(\text{Conv}_{7\times1}(\text{Conv}_{1\times7}(X_{\text{shallow}}))) \tag{8}$$

$$M = \sigma(\text{BN}(\text{Conv}_{1\times1}(X_{\text{reg}}))) \tag{9}$$

$$P_{\text{reg}} = \text{Conv}_{1\times1}(X_{\text{shallow}} \odot M) \tag{10}$$

where $\text{Conv}_{1\times7}(\cdot)$ and $\text{Conv}_{7\times1}(\cdot)$ are asymmetric depthwise separable convolutions, $\text{Conv}_{3\times3,d=2}(\cdot)$ denotes the dilated convolution, $\sigma(\cdot)$ is the HardSigmoid function, and $\odot$ represents element-wise multiplication.

The enhanced low-level details and aligned high-level semantics are adaptively fused to achieve top-down semantic injection. Finally, further nonlinear transformations are performed through 3×3 depthwise separable convolutions and activation functions:

$$X_{\text{out}} = \text{SiLU}(\text{BN}(\text{Conv}_{3\times3}(P_{\text{sem}} + P_{\text{reg}}))) \tag{11}$$

By integrating region masks and cross-layer information, the CLPM module suppresses background noise, achieves targeted foreground enhancement, and compensates for semantic gaps, significantly improving detection accuracy of small objects.

C. Saliency-Guided Purification (SGP)

In multi-scale feature fusion, representational conflicts are inevitably introduced due to the receptive field mismatch and feature aliasing caused by upsampling and concatenation, which severely obscure the subtle signals of small objects. To address this issue, we propose the Saliency-Guided Purification Module (SGP) illustrated in Fig. 5.

SGP consists of two primary components: feature mapping and feature purification. The feature mapping layer uses a 1×1 convolution for channel dimensionality reduction and nonlinear recombination, followed by a 3×3 convolution for feature extraction, as formulated below:

$$X_{map} = \text{Conv}_{1\times1}(\text{Conv}_{3\times3}(\text{Conv}_{1\times1}(X_{\text{in}}))) \tag{12}$$

To capture the spatial saliency distribution without adding extra parameters, the purification section employs a soft-gating mechanism, achieving feature de-noising through numerically stable exponential weighting. Specifically, the peak response intensity along the channel dimension is first extracted via maximum pooling, followed by a subtraction operation for numerical offset correction, defined as:

$$E = \exp(X_{\text{map}} - \max(X_{\text{map}})) \tag{13}$$

where $\max(\cdot)$ denotes the channel-wise maximum pooling, $-$ represents the subtraction operation, and $\exp(\cdot)$ is the exponential operation.

An exponential operation then nonlinearly amplifies salient signals while suppressing background noise, with the result being element-wise multiplied by the input features. To unify response scales across multi-scale features, the module aggregates global spatial information through average pooling and performs an inversion operation as a normalization factor. Finally, after Sigmoid activation, adaptive weighting is applied

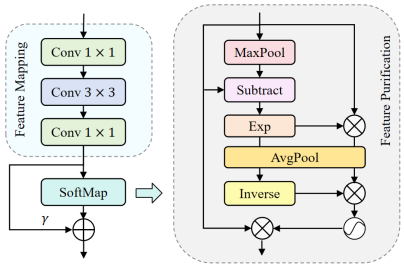

Fig. 5. Detailed components of the proposed SGP module.

TABLE I
COMPARISON OF DETECTION PERFORMANCE ON THE VISDRONE2019 TEST-DEV SET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Methods	Pedestrian	People	Bicycle	Car	Van	Truck	Tricycle	A.-tri.	Bus	Motor	mAP_{50}	Params	GFLOPs
Faster RCNN [5]	39.2	33.0	20.8	74.1	45.8	39.0	30.2	15.1	55.3	42.6	39.5	41.4	251.4
Cascade RCNN [6]	39.1	31.3	19.3	74.2	46.3	39.2	30.8	13.8	55.1	42.0	39.1	69.3	241.7
RT-DETR [22]	42.0	37.6	16.7	80.6	46.8	32.5	28.9	13.9	47.3	51.6	39.8	20.0	60.0
YOLOv5s [23]	40.5	31.9	12.5	78.8	44.8	36.1	25.8	15.9	54.0	43.1	38.3	8.7	24.0
YOLOv8s [24]	40.9	31.0	11.6	79.0	44.7	35.7	26.4	14.8	57.1	42.9	38.4	11.2	28.6
YOLOv9s [25]	41.3	32.2	12.4	79.6	45.9	38.9	29.4	15.7	56.8	44.1	39.6	7.1	26.7
YOLOv10s [26]	39.3	31.7	12.6	78.7	45.1	35.8	26.3	13.7	57.5	43.3	38.6	7.2	21.6
YOLOv10m [26]	43.6	35.0	15.4	80.5	48.0	40.1	29.9	18.1	58.3	46.6	41.5	15.4	59.1
YOLOv11s [27]	40.7	30.8	13.6	79.0	44.4	35.8	26.7	15.6	54.9	43.1	38.5	9.4	21.7
YOLOv11m [27]	48.0	36.2	18.2	82.0	48.6	40.0	32.8	19.3	60.4	49.3	43.5	20.1	68.0
LUDY [16]	44.8	34.3	14.5	80.9	48.4	39.4	29.8	16.9	62.2	46.2	41.7	10.3	–
LSOD-YOLO [28]	35.8	23.4	12.4	76.6	41.5	**40.1**	21.7	**22.1**	59.7	36.1	37.0	**3.8**	33.9
HIC-YOLO [29]	**50.5**	39.2	**21.4**	80.6	43.6	38.2	29.0	14.8	62.1	50.5	43.0	9.3	30.9
HTDNet (Ours)	**50.5**	**40.7**	20.9	**84.0**	**48.6**	38.9	**30.8**	19.3	**62.3**	**51.8**	**44.8**	4.3	**19.6**

to the original features, yielding the refined output through a residual structure:

$$S = \sigma\left(\frac{\sum(X_{\text{map}} \odot E)}{\sum E + \epsilon}\right) \tag{14}$$

$$X' = X_{\text{map}} + \gamma(X_{\text{map}} \odot S) \tag{15}$$

where $\odot$ denotes the element-wise multiplication, $\sum$ represents the summation operation along the channel dimension, σ is the Sigmoid activation function that constrains the output within the range of $(0, 1)$, ϵ is a small constant used for numerical stability, and γ denotes a learnable scaling factor.

IV. EXPERIMENTS AND ANALYSIS

A. Datasets

VisDrone2019 [30] is a leading UAV benchmark with 10,209 high-resolution images and 2.6 million annotated instances. Its primary challenge is small objects, with targets under 32×32 pixels comprising 44.7% of annotations. The dataset's complexity, including occlusion, scale imbalances, and diverse viewpoints, makes it a standard for evaluating multi-class recognition in aerial scenes. **AI-TOD [31]** focuses on tiny-object detection in aerial imagery, with 700,621 objects across 28,036 images and an average scale of just 12.8 pixels. The dense class distribution and background clutter increase localization difficulty, making it a platform for evaluating model robustness in small target detection. **TinyPerson [32]** targets long-distance pedestrian detection with 72,651 instances and most bounding boxes smaller than 20×20 pixels. Its high object density and complex textures make it ideal for evaluating performance in tiny-object settings.

B. Experimental Settings

All experiments were conducted on an NVIDIA RTX 4090 GPU using the PyTorch framework with CUDA 11.7. YOLO11s was used as the baseline, and the model was trained for 300 epochs with a 640×640 input size and batch size of 4. The SGD optimizer was used with a learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005, with a warm-up strategy. Training was done from scratch without pre-trained weights. MS COCO metrics were used to evaluate detection accuracy, while GFLOPS and Params were used to assess computational complexity and model efficiency.

TABLE II
COMPARISON OF DETECTION PERFORMANCE ON AI-TOD AND TINYPERSON DATASETS.

Method	AI-TOD (%)			TinyPerson (%)		
	mAP_{50}	Precision	Recall	mAP_{50}	Precision	Recall
YOLOv5s [23]	41.3	53.1	40.3	18.3	37.6	20.0
YOLOv8s [24]	42.3	54.1	41.7	18.5	39.1	19.3
YOLOv9s [25]	41.6	53.8	41.5	17.6	37.2	19.2
YOLOv10s [26]	41.1	53.0	41.0	18.2	37.8	19.6
YOLOv11s [27]	41.3	52.8	40.6	18.0	36.7	20.3
ARNet [33]	42.8	**61.4**	40.5	24.3	40.9	26.8
HTDNet (Ours)	**46.1**	59.7	**45.0**	**25.2**	**41.5**	**27.6**

C. Experimental Results

Experimental results on the VisDrone2019 dataset, shown in Table I, demonstrate that HTDNet achieves state-of-the-art performance with minimal computation. Reaching 44.8% mAP_{50} with only 19.6 GFLOPs, HTDNet outperforms competitors in 7 out of 10 categories, notably improving "people" and "motor" by 1.5% and 1.3% over second-best methods, respectively. This highlights HTDNet's significant advantage in small object detection. HTDNet exhibits slightly lower performance on "awning-tricycle" due to its high intra-class variability and scale differences. Compared to the YOLOv11s baseline, HTDNet boosts mAP_{50} by 6.3% while reducing parameters by 55%. Even compared to LSOD-YOLO, HTDNet yields a 7.8% mAP_{50} gain with only a 0.5M parameter increase. Furthermore, HTDNet demonstrates superior trade-offs in computation, parameter size, and accuracy compared to other lightweight models such as HIC-YOLO and LUDY, showing high efficiency in resource-constrained environments.

To further validate the effectiveness and generalizability of HTDNet, experiments on AI-TOD and TinyPerson, illustrated in Table II, show mAP_{50} gains of 4.8% and 7.2% over the baseline. Specifically, HTDNet reaches 25.2% mAP_{50} on

TinyPerson with SOTA precision and recall. On AI-TOD, it achieves the highest mAP_{50} of 46.1% and recall of 45.0%, despite slightly lower precision than the top method. These cross-dataset results confirm HTDNet's robust adaptability beyond UAV detection to diverse small object tasks.

TABLE III
ABLATION STUDY OF DIFFERENT COMPONENTS ON VISDRONE2019.

NET	CLPM	DRR	SGP	Params	GFLOPs	mAP_{50}	mAP
×	×	×	×	9.4M	21.7	38.5	23.4
✓	×	×	×	5.2M	19.5	42.8	25.8
✓	✓	×	×	5.2M	21.2	43.8	26.7
✓	✓	✓	×	4.4M	20.0	44.3	26.9
✓	✓	✓	✓	4.3M	19.6	44.8	27.3
✓	×	✓	×	4.4M	18.3	42.9	25.9

TABLE IV
COMPARISON OF DIFFERENT ATTENTION MECHANISMS ON VISDRONE.

Model	**mAP_{50} (%)**	**mAP (%)**	**Params**	**GFLOPs**
CBAM [34]	43.4	26.3	4.4M	18.8
CA [35]	42.9	25.9	4.3M	19.0
EMA [36]	43.5	26.4	4.3M	18.8
CLPM (Ours)	44.8	27.3	4.3M	19.6

D. Ablation Study

Table III presents the ablation results on the VisDrone2019 dataset, evaluating the effectiveness and generalization of each component in HTDNet. First, the structural refinements and the introduction of CGDown significantly enhance small object detection performance while reducing model complexity. We then performed a staged ablation study on the three core modules. Specifically, the DRR module alone reduces the parameters from 5.2M to 4.4M and GFLOPs to 18.3, while simultaneously improving mAP_{50}, demonstrating its efficient global modeling capability. Similarly, the CLPM module alone improves mAP_{50} from 42.8% to 43.8%, yielding a 0.9% mAP gain with nearly constant parameters. Their combination

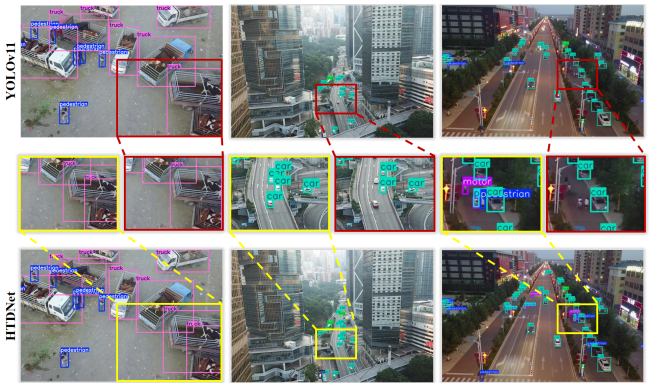

Fig. 6. Visual comparison of detection results.

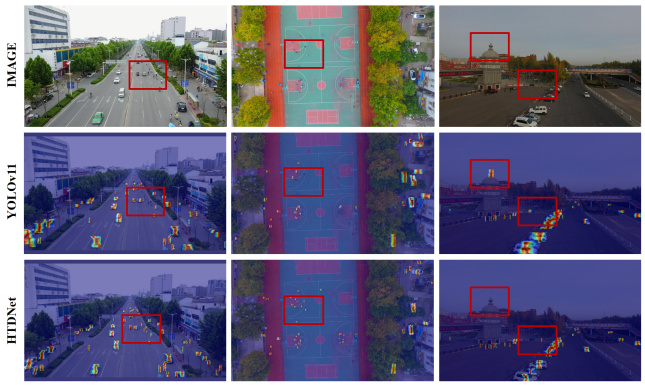

Fig. 7. Heatmap visualization comparison between the baseline and HTDNet.

further boosts mAP_{50} to 44.3%, highlighting their strong synergy. Finally, by incorporating the SGP module into the neck's unidirectional fusion, HTDNet reaches 44.8% mAP_{50}. Compared to the baseline, this represents a 6.3% improvement with a 55% reduction in parameters and a significant decrease in computational cost. Overall, HTDNet effectively addresses UAV-specific detection challenges through its hierarchical task-decoupled design, achieving a superior balance between performance and efficiency.

Furthermore, we conducted a component-level ablation study on CLPM by comparing it with several representative attention mechanisms, including CBAM, EMA and CA, to evaluate its feature guidance efficacy for small objects. By integrating semantic and spatial priors, CLPM facilitates the precise recalibration of foreground features. Experimental results indicate that CLPM achieves superior accuracy with only a small increase in computational cost and comparable parameter efficiency. Specifically, it surpasses the second-best method by 1.3% in mAP_{50} and 0.9% in mAP, demonstrating its robust ability to suppress complex background interference.

E. Visualization

To validate the performance of the proposed method, we provide a visual comparison on the VisDrone test-dev set. Fig. 6 shows detection results of YOLOv11 and HTDNet. In the first column, the proposed method achieves more accurate object localization. The second example highlights HTDNet's improved performance in detecting small objects, while the third demonstrates its effectiveness in complex backgrounds. Fig. 7 shows heatmaps, where YOLOv11 exhibits redundant saliency on large objects, while HTDNet's more balanced attention increases focus on small objects, reducing false negatives and false positives. These results underscore HTDNet's robustness in UAV scenarios.

V. CONCLUSION

This paper proposes HTDNet, an efficient object detection framework tailored for UAV scenarios, which effectively addresses the heterogeneous challenges across different network levels through task decoupling. Three core modules are introduced: **DRR** reconstructs structured semantics via feature

decomposition to mitigate semantic sparsity and homogenization in deeper layers; **CLPM** leverages semantic and spatial priors to guide the precise localization of small objects, effectively suppressing background interference in shallow layers; and **SGP** eliminates response bias during multi-scale fusion through nonlinear denoising, thereby enhancing the saliency response for small objects. Extensive experiments on three benchmarks demonstrate that HTDNet significantly improves detection performance while maintaining low model complexity, providing a robust solution for real-time deployment on resource-constrained platforms. Future work will explore knowledge distillation for model compression and algorithm optimization to maximize accuracy under extreme constraints.

ACKNOWLEDGMENT

The authors thank the Natural Science Foundation of Hebei Province, China Project (F2024201012) for their financial support and the support of the High-Performance Computing Center of Hebei University.

REFERENCES

[1] H. Zheng, Z. Chang, Y. Li, J. Zhu, W. Wang, Q. Yang, C. Xie, J. Zhang, and J. Liu, "An efficient and fast image mosaic approach for highway panoramic UAV images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 10 454–10 467, 2024.

[2] A. Bouguettaya, H. Zarzour, A. M. Taberkit, and A. Kechida, "A review on early wildfire detection from unmanned aerial vehicles using deep learning-based computer vision algorithms," *Signal Processing*, vol. 190, p. 108309, 2022.

[3] H. Wang, P. Du, S. Yang, Z. Zhang, and J. Ning, "A spatial arrangement preservation-based stitching method via geographic coordinates of UAV for farmland remote sensing image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[4] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023.

[5] 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, pp. 1137–1149, 2016.

[6] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1483–1498, 2021.

[7] 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 (CVPR)*, 2016, pp. 779–788.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 21–37.

[9] Z. Han, Z. Ma, and R. Wang, "PIAENet: Pyramid integration and attention enhanced network for object detection," *Information Sciences*, vol. 670, p. 120576, 2024.

[10] I. Ahmad, W. Lu, S. B. Chen, J. Tang, and B. Luo, "Lightweight oriented object detection with dynamic smooth feature fusion network," *Neurocomputing*, vol. 628, p. 129725, 2025.

[11] X. Fu, Z. Yuan, T. Yu, and Y. Ge, "DA-FPN: Deformable convolution and feature alignment for object detection," *Electronics*, vol. 12, no. 6, p. 1354, 2023.

[12] J. Ji, Y. Zhao, A. Li, X. Ma, C. Wang, and Z. Lin, "EFR-ACENet: Small object detection for remote sensing images based on explicit feature reconstruction and adaptive context enhancement," *Engineering Applications of Artificial Intelligence*, vol. 151, p. 110722, 2025.

[13] M. Li, Y. Zhao, G. Gui, F. Zhang, B. Luo, C. Yang, W. Gui, K. Chang, and H. Wang, "Object detection on low-resolution images with two-stage enhancement," *Knowledge-Based Systems*, vol. 299, p. 111985, 2024.

[14] U. Ahmad, J. Liang, T. Ma, K. Yu, F. Mehmood, and F. Banoori, "Small aerial object detection through GAN-integrated feature pyramid networks," *Applied Soft Computing*, vol. 171, p. 112834, 2025.

[15] X. Ding, R. Zhang, Q. Liu, and Y. Yang, "Real-time small object detection using adaptive weighted fusion of efficient positional features," *Pattern Recognition*, vol. 167, p. 111717, 2025.

[16] Q. Fan, Y. Li, M. Deveci, K. Zhong, and S. Kadry, "LUD-YOLO: A novel lightweight object detection network for unmanned aerial vehicle," *Information Sciences*, vol. 686, p. 121366, 2024.

[17] Y. Niu, C. Lin, X. Jiang, and Z. Qu, "VSTDet: A lightweight small object detection network inspired by the ventral visual pathway," *Applied Soft Computing*, vol. 171, p. 112775, 2025.

[18] B. Peng, S. Xiong, Y. Wang, T. S. Zhou, J. Li, H. Xiao, and R. Xiong, "MI-DETR: A small object detection model for mixed scenes," *Displays*, vol. 88, p. 103052, 2025.

[19] C. Wang, W. He, Y. Nie, J. Guo, C. Liu, K. Han, and Y. Wang, "Gold-YOLO: Efficient object detector via gather-and-distribute mechanism," in *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023. [Online]. Available: https://arxiv.org/abs/2309.11331

[20] 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.

[21] S. Jobaer, X. S. Tang, and Y. Zhang, "A deep neural network for small object detection in complex environments with unmanned aerial vehicle imagery," *Engineering Applications of Artificial Intelligence*, vol. 148, p. 110466, 2025.

[22] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 1–14.

[23] G. Jocher. (2022) Yolov5 release v6.2. [Online]. Available: https://github.com/ultralytics/yolov5/releases/tag/v6.2

[24] Jocher, Glenn. (2023) Yolov8. [Online]. Available: https://github.com/ultralytics/ultralytics

[25] C. Wang and H. Y. M. Liao, "Yolov9: Learning what you want to learn using programmable gradient information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 1–18.

[26] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 1–21.

[27] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.

[28] H. Wang, J. Liu, J. Zhao, J. Zhang, and D. Zhao, "Precision and speed: Lsod-yolo for lightweight small object detection," *Expert Syst. Appl.*, vol. 269, p. 126440, 2025.

[29] S. Tang, S. Zhang, and Y. Fang, "Hic-yolov5: Improved yolov5 for small object detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2024, pp. 6614–6619.

[30] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, and Y. Zhang, "VisDrone-DET2019: The vision meets drone object detection in image challenge results," [dataset] arXiv:2001.06303, 2019. [Online]. Available: http://arxiv.org/abs/2001.06303

[31] J. Wang, W. Yang, H. Guo, R. Zhang, and G. S. Xia, "Tiny object detection in aerial images," [dataset] IEEE, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9406139

[32] X. Yu, Y. Gong, N. Jiang, Q. Ye, and Z. Han, "Scale match for tiny person detection," [dataset] IEEE, 2020.

[33] X. Deng, Z. Xu, W. Yang, H. Li *et al.*, "Arnet: An aggregation and recalibration network for small object detection in aerial imagery," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2025.

[34] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[35] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 13 713–13 722.

[36] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Yun, "Efficient multi-scale attention module with cross-spatial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 5886–5895.