Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.048351

ARTICLE

# Enhancing Dense Small Object Detection in UAV Images Based on Hybrid Transformer

**Changfeng Feng[1], Chunping Wang[2], Dongdong Zhang[1], Renke Kou[1] and Qiang Fu[1,*]**

[1]Shijiazhuang Campus, People Liberation Army Engineering University, Shijiazhuang, 050003, China

[2]School of Information and Intelligent Engineering, Sanya University, Sanya, 572000, China

*Corresponding Author: Qiang Fu. Email: Fu_Qiang@aeu.edu.cn

Received: 05 December 2023 Accepted: 22 January 2024 Published: 26 March 2024

**ABSTRACT**

Transformer-based models have facilitated significant advances in object detection. However, their extensive computational consumption and suboptimal detection of dense small objects curtail their applicability in unmanned aerial vehicle (UAV) imagery. Addressing these limitations, we propose a hybrid transformer-based detector, H-DETR, and enhance it for dense small objects, leading to an accurate and efficient model. Firstly, we introduce a hybrid transformer encoder, which integrates a convolutional neural network-based cross-scale fusion module with the original encoder to handle multi-scale feature sequences more efficiently. Furthermore, we propose two novel strategies to enhance detection performance without incurring additional inference computation. Query filter is designed to cope with the dense clustering inherent in drone-captured images by counteracting similar queries with a training-aware non-maximum suppression. Adversarial denoising learning is a novel enhancement method inspired by adversarial learning, which improves the detection of numerous small targets by counteracting the effects of artificial spatial and semantic noise. Extensive experiments on the VisDrone and UAVDT datasets substantiate the effectiveness of our approach, achieving a significant improvement in accuracy with a reduction in computational complexity. Our method achieves 31.9% and 21.1% AP on the VisDrone and UAVDT datasets, respectively, and has a faster inference speed, making it a competitive model in UAV image object detection.

**KEYWORDS**

UAV images; transformer; dense small object detection

## 1 Introduction

Unmanned aerial vehicles (UAVs) are increasingly used for various purposes, such as disaster relief, urban monitoring, and land protection, owing to their compact dimensions, cost-effectiveness, and operational simplicity [1–4]. Fig. 1 shows that the object features in UAV vision are more complex than those in normal vision. This complexity lies in the dense clutter and numerous small objects, presenting challenges for object detection in UAV images [5].

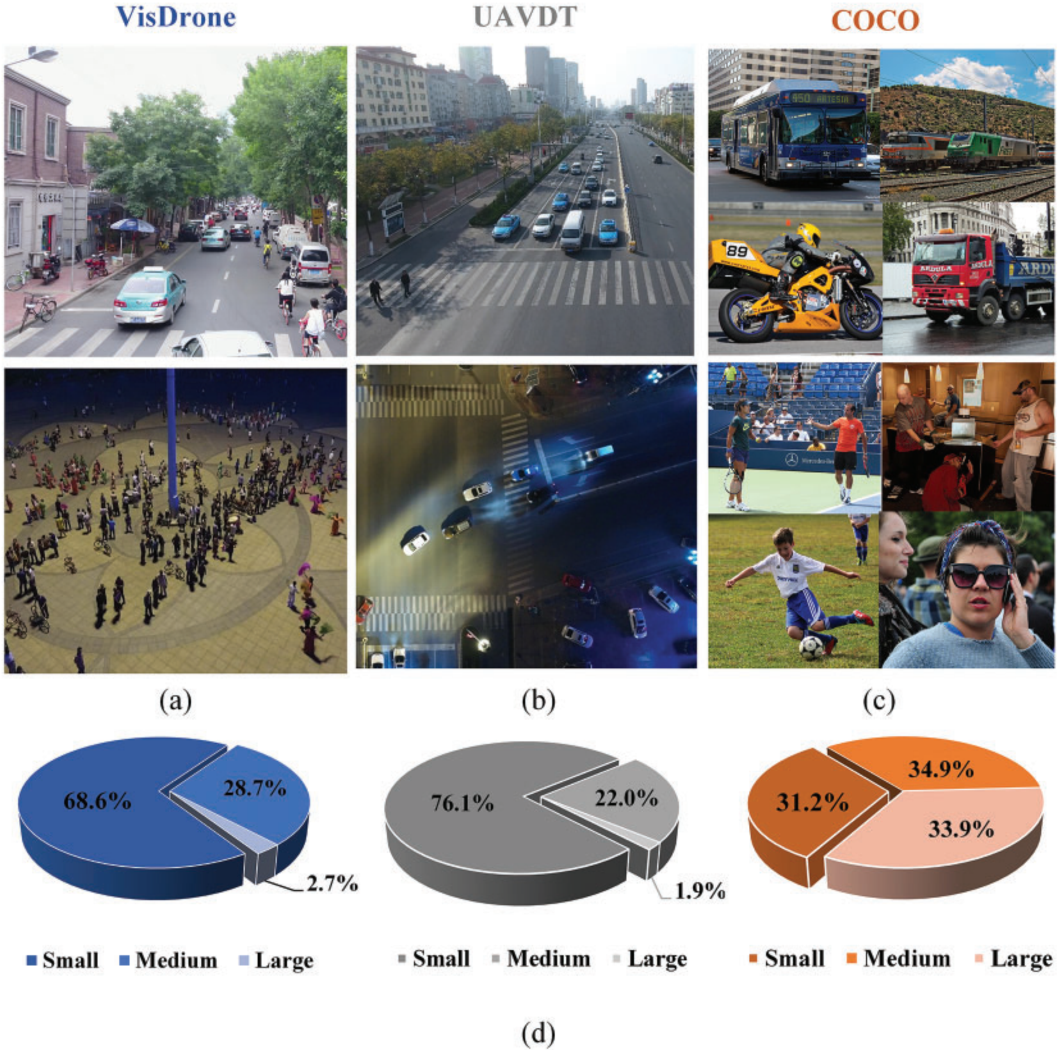


**Figure 1:** Comparative analysis of object characteristics in UAV and standard imagery. Panels (a) to (c) depict representative samples from the VisDrone, UAVDT, and COCO datasets. Panel (d) presents a statistical distribution chart of object sizes

The emergence of large-scale benchmark datasets has led to an innovative paradigm in object detection using convolutional neural networks (CNNs) [6–8]. Detectors designed for UAV imagery are usually modified on these CNN-based detectors and are broadly attributed to two kinds of pipelines: multi-inference and single-inference [9]. Multi-inference detectors tend to use a coarse-to-fine process to achieve higher accuracy, where coarse detectors are launched to localize dense subregions, and subsequently, fine detectors are used to detect more minor instances [10–11]. However, these multi-inference detectors result in significant latency, making them unsuitable for practical application. Single-inference detectors are generally more efficient, requiring only one pass over the image to generate predictions, although this efficiency often comes with trade-offs in accuracy [12–13]. Overall, existing CNN-based approaches exhibit limitations in reconciling efficiency with precision, and designing an accurate and efficient UAV image detector remains an outstanding challenge.

The advent of the transformer architecture, initially conceived for the natural language processing domain, has subsequently been adapted for visual tasks, with notable success for its global context awareness. In the object detection field, the DEtection TRansformer (DETR) leverages the transformer architecture to create an end-to-end detector [14]. This approach has revolutionized the standard workflow by dispensing with components such as predefined anchors and complex post-processing, which has led to remarkable performance enhancements across many downstream tasks [15–16]. However, it is paramount to acknowledge that existing DETRs are predominantly tailored for natural imagery, which presents pronounced challenges when repurposed for UAV image analysis.

- The computational complexity of the transformer's attention mechanism exhibits heightened sensitivity to the length of the input sequences, making it face an intolerable computational burden for processing high-resolution UAV images.

- Unlike CNNs, transformers inherently lack spatial inductive biases. They lack specific priors, making it difficult to learn features in exceptional cases (e.g., densely distributed and numerous small instances in the UAV view) without a large amount of data.

This work introduces a novel end-to-end transformer-based detector called the Hybrid DEtection TRansformer (H-DETR). To mitigate the computational intensity when handling high-resolution drone imagery, we present an improved transformer encoder, which optimizes the original attention-only encoder by hybridizing an efficient CNN-based feature fusion network. To address the challenge of dense object distributions, we embed a query filter process before query initialization to suppress similar queries. To improve the detection accuracy of small objects, we include an adversarial learning concept in the training scheme. With the above practices, we optimize the transformer-based model to make it more suitable for UAV images and balance accuracy and efficiency, which the previous model cannot achieve.

The primary contributions to this paper are as follows:

- We systematically analyze the transformer architecture's functionality and propose the Hybrid Transformer Encoder (HTE), designed to reduce the computational intensity of high-resolution UAV image processing without sacrificing the accuracy benefits.

- We design a Query Filter (QF) that addresses the specific issue of prediction redundancy that arises from dense clusters of similar objects in drone view.

- We introduce Adversarial Denoising Learning (ADL), an innovative enhancement technique for small object detection. ADL leverages adversarial noise to improve the robustness of the model against spatial and semantic perturbations commonly encountered in UAV image object detection.

- We conduct comprehensive evaluation experiments on the VisDrone and UAVDT datasets, confirming the proposed method significantly improves accuracy while reducing computational complexity. Our proposed H-DETR can balance detection accuracy and efficiency well, outperforming previous CNN-based models.

## 2 Related Work

Object detection has been a cornerstone of computer vision research, leading to a significant body of literature. This section summarizes critical studies that have informed the development of our proposed methodology.

### *2.1 Visual Transformers*

Transformer was proposed by Vaswani et al. [17] as a new attention-based building block for machine translation. Attention mechanisms are neural network layers that aggregate information from the entire input sequence. Transformers introduced a self-attention layer to scan each element in the sequence and update it by aggregating global information. This ability was later shown to capture complex spatial dependencies in images, critically needed for computer vision tasks. Vision Transformer (ViT) [18] laid the groundwork by applying a pure transformer to sequences of image patches, fostering the model's ability to discern intricate patterns across the entire image. A notable derivative is the swing transformer [19], which introduces a hierarchical structure with shifted windows, enhancing the efficiency of the self-attention mechanism for dense prediction tasks. By their global receptive fields, these architectures have been instrumental in surpassing previous state-of-the-art methods across various vision tasks [20,21]. In our work, these seminal studies inform our architectural choices. As a bridge between the underlying transducer model and UAV imaging applications, we exploit the global modeling advantages of the transducer to improve object detection while mitigating the computational intensity typically associated with such models.

### *2.2 General Object Detection*

CNN-based object detection models can generally be classified as anchor-based or anchor-free detectors. Anchor-based detectors can be further divided into two-stage and one-stage. Two-stage detectors, such as Faster Region-based Convolutional Neural Network (Faster R-CNN) [6] and Cascade R-CNN [22], generate the proposed region and then classify and regress the objects within it. In contrast, one-stage detectors, such as You Only Look Once (YOLO) and Single Shot MultiBox Detector (SSD) [23], can directly perform this classification and regression of objects within their entire feature. Anchor-free detectors replace anchors, which can cause a significant computational burden, with more efficient alternatives such as centerness constraints or heat maps [24]. DETR is a significant innovation in object detection, utilizing a transformer architecture for end-to-end object detection with a cleaner process and better generalization. However, despite its advantages, DETR encounters optimization complexity and computational intensity. To alleviate these issues, conditional DETR [25] reduces query optimization complexity by explicitly identifying the extremity region of an object through a conditional space query. Sparse DETR [26] selectively updates the encoder tokens expected to be referenced by the decoder, reducing computational overhead. These improvements have enhanced the utility of DETR-like detectors, leading them to achieve excellent performance in several benchmarks. Real-Time DEtection TRansformer (RT-DETR) improves the efficiency of feature processing by decoupling the treatment of multi-scale features. Also, it proposes an IoU-aware query selection method to improve the inconsistency in the distribution of raw query classification scores and location confidence. However, the use of transformers in object detection is still a relatively new concept, and the above work has mainly focused on natural images without adaptation to the characteristics of drone view.

### *2.3 Object Detection in UAV Images*

Advancements in object detection have been substantial, yet their application to drone-captured imagery still presents unique challenges, such as small object instances. To address these challenges, some studies have used a coarse-to-fine pipeline. Global-local Self-Adaptive Network (GSANet) [27] filters out crowded sub-regions using a self-adaptive region selection algorithm. Then, it performs detailed detection after improving the resolution of the small sub-regions using a local super-resolution network. Unified Foreground Packing Multi-Proxy Detection (UFPMPDet) network [11] merges the unified foreground regions generated by a coarse detector into a mosaic image for detailed detection and employs a multi-proxy detection network to handle the significant confusion between inter-class similarities and intra-class variations of instances. The Clustered Detection (ClusDet) network [28] unifies object clustering and detection in an end-to-end framework that sequentially finds clustered regions and detects objects in those regions. Similarly, Zhang et al. [29] proposed an adaptive cropping method based on a difficult region estimation network to enhance the detection of challenging targets.

These dual-stage processes, although precise, incur obvious computational penalties [30,31]. Recently, some efforts have converged on developing optimized single-inference models to reconcile detection accuracy with operational practicality [32]. Hierarchical Shot Detector (HSD) [33] proposes a new Reg-Offset-Cls module and stacking strategy to achieve higher accuracy and speed by summarizing the drawbacks of single-stage detectors, such as the mismatch of bounding box classification and the insufficient accuracy of one-shot regression. Du et al. [34] presented a novel global Context-Enhanced Adaptive Sparse Convolutional (CEASC) network by optimizing sparse convolution, which reduces the computational effort while improving accuracy. However, the accuracy of all the above methods is not satisfactory. Capitalizing on their superior ability to capture global information, several works have attempted to integrate transformers into detectors for UAV images. Zhu et al. [35] merged a transformer-based prediction head with the YOLOv5 detection model, achieving remarkable performance improvements in large-scale variations and high-density contexts. To recognize high-level semantic information and enhance the perception of local geometric features, Multiple Attention Mechanism Enhanced YOLOX (MAME-YOLOX) [36] integrates the Swin transformer into the neck module of YOLOX. Due to the computationally intensive nature of the internal attention mechanisms in transformers, prior efforts have employed it solely as a localized enhancement module within CNN-based frameworks, thereby underutilizing its intrinsic global-aware capabilities. Instead, our approach uses a DETR-like architecture dominated by a complete transformer structure. It has been optimized for effectiveness and efficiency based on it, thus better unleashing the potential of the transformer in object detection in UAV Images.

## 3 The Proposed Method

In this section, we delineate the architecture of our proposed model, which is composed of a backbone, an encoder, and a decoder augmented with prediction heads. The overarching configuration is depicted in Fig. 2. Our design uses ResNet50 as the underlying backbone, with the outputs of its last three stages providing input features to the converter. The subsequent HTE (described in Section 3.1) is responsible for converting these multi-scale features into a sequence of features. The HTE is ingeniously structured and composed of cross-scale fusion and semantic enhancement modules. The latter, which processes solely the uppermost level features, is equipped with a self-attention layer and a subsequent feed-forward layer. Subsequently, the QF (described in Section 3.2) performs a selective refinement of the encoder feature sequence to extract a specified number of distinct queries, which are used as the initial set for the transformer decoder. Here, we use the classical six-layer structure for the transformer decoder. The filtered queries also pass through a self-attention layer before cross-attention operation with the encoder features and then through a feed-forward layer. Mirroring the semantic enhancement module, each progression of the filtered query is accompanied by layer normalization, ensuring stability and improved convergence. During training, these filtered queries are iteratively improved in the transformer decoder using the ADL method (described in Section 3.3). Finally, the prediction head maps the final query to the corresponding bounding box and confidence score.

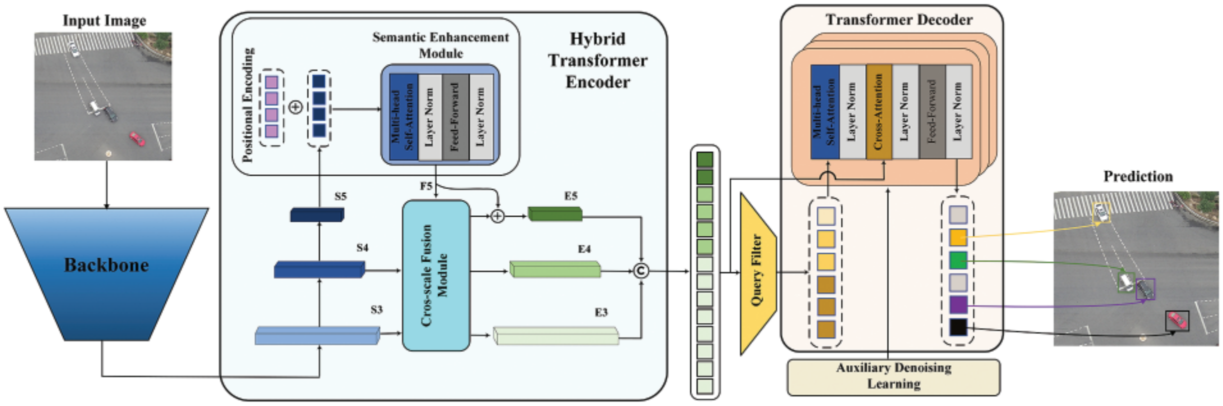
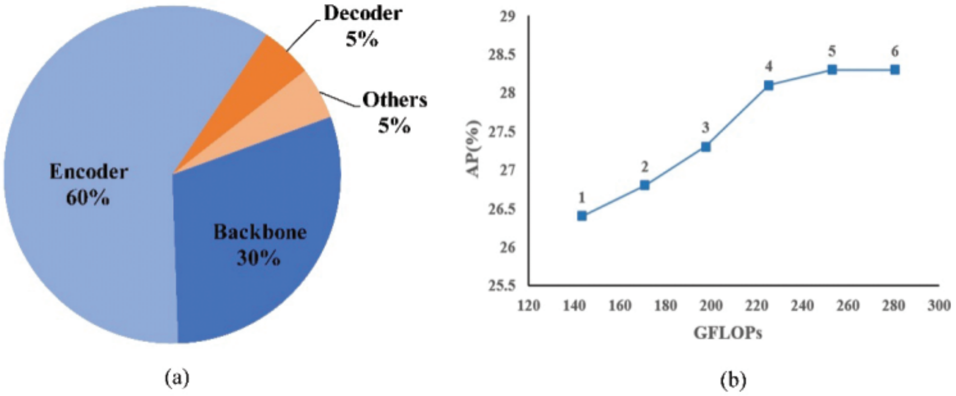


**Figure 2:** Overview of the H-DETR. Initially, multi-scale features are extracted using a backbone and then transformed into a sequence. Subsequently, a fixed number of queries are filtered out and fed into the transformer decoder. In the terminal phase, the decoder with prediction heads iteratively optimizes the queries and generates final predictions

### *3.1 Hybrid Transformer Encoder*

To reduce the computational consumption of the transformer, our study first critically evaluates the computational effort of each component of the DETR using Giga Floating Point Operations (GFLOPs), a measure of the computational complexity of a neural network model, and the results are shown in Fig. 3a. We find the transformer encoder to be the primary source of computational effort and further investigate the effect of the encoder on the accuracy of object detection in UAV images. Fig. 3b shows that the original encoder improves accuracy only slightly while almost doubling the GFLOPs. The above analysis suggests optimizing the vanilla transformer encoder is the key to mitigating the transformer's intensive computation without causing performance degradation in UAV image detection. We contend that the conventional encoder design incurs computational redundancy primarily due to two factors:



(a)

(b)

**Figure 3:** Analysis of the DETR-like model in the VisDrone dataset. (a) The distribution of the GFLOPs in the baseline model. (b) The AP and GFLOPs of baseline with various encoder layers

• The low-level features' intra-scale interactions, lacking in rich semantic content and high resolution, contribute to unnecessary computational load.

• The indiscriminate inter-scale feature interactions across different scales may elevate the risk of semantic ambiguity.

In response to these findings, we introduce the HTE, which hybridizes an attention-based semantic enhancement module with a CNN-based cross-scale fusion module, as shown in Fig. 2. The semantic enhancement module is modified based on the original encoder layer to process only high-level features with more semantic information, which avoids intra-scale interactions of low-level features. At the same time, we use the cross-scale fusion module to target the interactions between features at different levels to make the feature fusion more discriminative and less computationally intensive.

In particular, we unbend the highest-level feature $S_5$ as the initial *Query*, *Key*, and *Value* of the attention operation in the semantic enhancement module. It is worth noting that multi-head self-attention [17] replaces original deformable attention. We posit that applying self-attention to semantically enriched high-level features can discern the nexus between conceptual entities within the image. Moreover, multi-head attention divides tokens into multiple subspaces through distinct mapping matrices to elicit more nuanced feature representations. This process can be formulated as follows:

$$Q = K = V = Unbend\left(S_5\right) \tag{1}$$

$$Q_i = QW_i^Q, K_i = KW_i^K, V_i = VW_i^V \quad where\ W_i^Q, W_i^K, W_i^K \in \mathbb{R}^{d_{feat} \times d_k} \tag{2}$$

Here, $i$ denotes the $i$-th subspace, where $d_{feat} = d_{S_5} = 256$, $d_k = 32$. Self-attention operations are then performed within the subspaces to obtain the output $head_i$, which can be formulated as:

$$head_i = Attention\left(Q_i, K_i, V_i\right) = Softmax\left(\frac{Q_i\left(K_i\right)^T}{\sqrt{d_k}} V_i\right) \tag{3}$$

The results obtained from each head are concatenated and linearly mapped using the output matrix to get the final output $F_5$, as shown in Eq. (4). Here, I was set to 8, following the [14].

$$F_5 = Concat\left(head_1 \ldots head_i \ldots head_I\right) W^o \quad where\ W^O \in \mathbb{R}^{hd_v \times d_{feat}} \tag{4}$$

Subsequently, we perform a feature fusion operation on $\{S_3, S_4, F_5\}$ using a cross-scale fusion module. The structure of the cross-scale fusion module resembled that of the Path Aggregation Network (PANet) [37], as shown in Fig. 4, except that we introduced a Cross-Stage Partial Layer (CSPLayer) to enhance the gradient performance and reduce the computational operations [38]. Moreover, a simple residual connection is employed to prevent any decline in performance that the simultaneous operation of different mechanisms within the HTE might cause.

### *3.2 Query Filter*

In the transformer, each query comprises a series of embedding vectors that the decoder can optimize and subsequently map to classification scores and bounding boxes. Conceptually, a query represents a potential instance within an image, functioning as a hypothetical object template. When considering UAV imagery, the fact that objects are identical in densely distributed areas (e.g., vehicles in parking lots or crowds in public squares) introduces additional challenges for the model, as these conditions tend to result in queries consistent in spatial and semantic. This similarity between queries produces redundant predictions in some scenarios, affecting object detection accuracy.

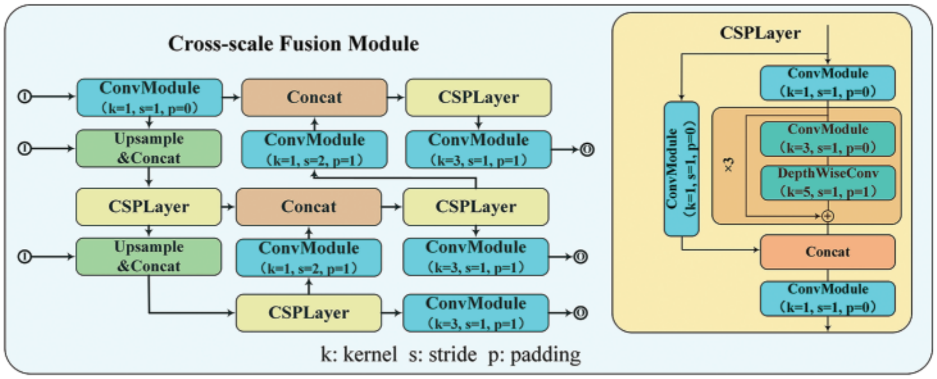

Cross-scale Fusion Module
ConvModule (k=1, s=1, p=0)
Concat
CSPLayer
Upsample &Concat
ConvModule (k=1, s=2, p=1)
ConvModule (k=3, s=1, p=1)
CSPLayer
Concat
CSPLayer
Upsample &Concat
ConvModule (k=1, s=2, p=1)
ConvModule (k=3, s=1, p=1)
CSPLayer
ConvModule (k=3, s=1, p=1)
CSPLayer
ConvModule (k=1, s=1, p=0)
ConvModule (k=1, s=1, p=0)
×3
ConvModule (k=3, s=1, p=0)
DepthWiseConv (k=5, s=1, p=1)
Concat
ConvModule (k=1, s=1, p=0)
k: kernel s: stride p: padding

**Figure 4:** Composition and data flow of cross-scale fusion module

To address the issue of query homogeneity, we design a QF, as shown in Fig. 5. Firstly, the encoder features are passed through an ancillary prediction head parallel to the decoder, encompassing classification and regression branches. After this, we employ class-agnostic non-maximum suppression (NMS) on the resulting proposals, prioritizing selecting the top N proposals with the highest scores. In contrast to the training-unknown NMS in conventional detectors, it does not directly impact the final detection results and is training-aware, so we set a positive intersection over union (IoU) threshold of 0.85. The coordinates of the top N proposals are then coded to produce positional embedding, while the encoder features undergo a linear layer activation to generate the corresponding content embedding. These content and positional embeddings are amalgamated to form the initial set of queries for the decoder.

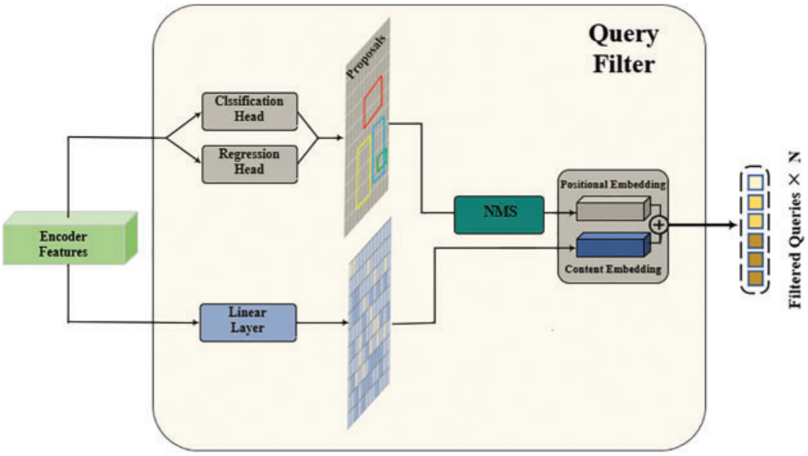


**Figure 5:** The structural details of the query filter

The function of QF is to map the query onto the practical proposals and perform pre-screening before initialization. This prevents the generation of redundant predictions during subsequent detection. Besides, our QF employs NMS not just as a post-processing step but integrates it into both the training and inference phases, thereby ensuring our model's strict adherence to end-to-end design principles.

### *3.3 Adversarial Denoising Learning*

As depicted in Fig. 6, the unique perspective afforded by drones can obscure the spatial localization and semantic differentiation of objects, particularly those of smaller scale. To enhance the model's capability to discern features of small objects, we have integrated an ADL method, drawing inspiration from the principles of adversarial training. Adversarial training is known to bolster network robustness through the incorporation of synthetic samples that contain deliberately introduced noise. In our methodology, we add spatial and semantic noise to instances in the image as synthetic samples.

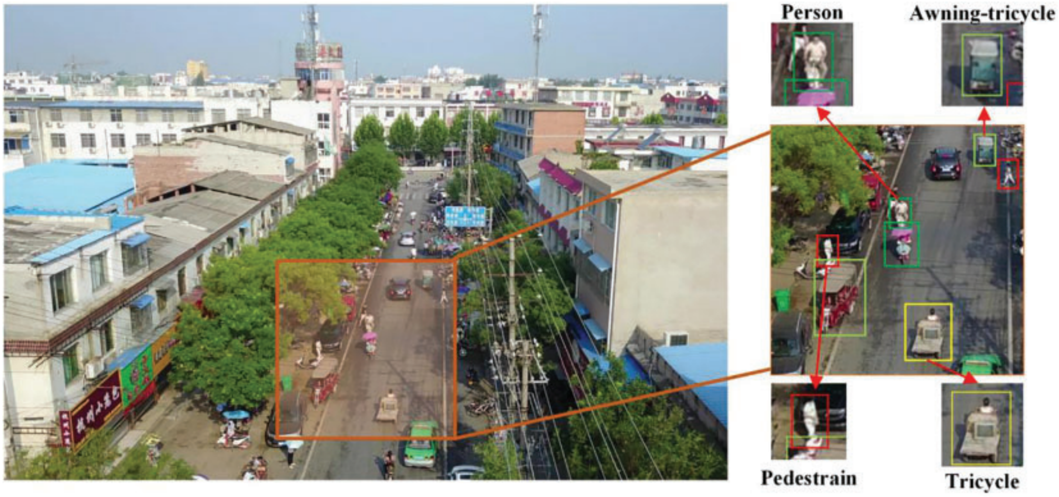


**Figure 6:** Challenging small instances in the drone's view

For a training image, the ground truth (GT) objects are denoted as $G = \{G_0 \ldots G_i \ldots G_{m-1}\}$, where $m$ represents the count of GT objects, and $G_i$ is defined by the tuple $(x_i, y_i, w_i, h_i, c_i)$, encapsulating the center's coordinates, width, height, and object category. To fabricate reconstructed objects with spatial noise, we apply counter-directional perturbations as described below:

$$R^P = \left\{R_0^P \ldots R_i^P \ldots R_{m-1}^P\right\} \qquad where\ R_i^P = \left(x_i, y_i, w_i + \alpha\frac{\lambda_1 w_i}{2}, h_i + \alpha\frac{\lambda_1 h_i}{2}, c_i\right) \tag{5}$$

$$R^N = \left\{R_0^N \ldots R_i^N \ldots R_{m-1}^N\right\} \qquad where\ R_i^N = \left(x_i, y_i, w_i + \alpha\frac{\lambda_2 w_i}{2}, h_i + \alpha\frac{\lambda_2 h_i}{2}, c_i\right) \tag{6}$$

Here, we set two different levels of noise to $w_i$ and $h_i$, which is to prevent the network from being sensitive to too small perturbations and thus causing performance degradation. $\alpha$ is a predetermined hyperparameter dictating the spatial noise's intensity. $\lambda_1$ and $\lambda_2$ are stochastic variables indicative of the noise magnitude, where $\lambda_1 \in (-0.25, 0.25)$ and $\lambda_2 \in (-0.5, -0.25] \cup [0.25, 0.5)$. This boundary condition ensures GT does not undergo a corner-point flip after adding the noise. Semantic noise is introduced through label flipping, randomly reassigning categories to $c_i$ within $R_i^P$ and $R_i^N$. The hyperparameter $\beta$ modulates the proportion of label flipping and the intensity of the semantic disturbance. These adversarial perturbed constructs, $R_i^P$ and $R_i^N$ are encoded into queries, subsequently processed by the transformer decoder for adversarial denoising, and then mapped by the detection head. The associated loss is computed by juxtaposing the denoised predictions with the GT, as expressed in the denoising loss function:

$$L_{DE} = \frac{1}{m} \sum\nolimits_{i=0}^{m} [L_{reg}\left(R_i^P, G_i\right) + L_{cls}\left(R_i^P, G_i; R_i^N, \varnothing\right)] \tag{7}$$

The loss function comprises two components: regression loss for the bounding boxes and classification loss. In this study, we adopted L1 and CIOU [39] losses for regression and focal loss [40] for classification loss.

The spatial noise in ADL allows the model to learn more detailed spatial information, while the semantic noise enables the model to learn label-boundary associations, driving the accuracy of the model's label predictions. The two synergies effectively improve the accuracy of small object prediction.

## 4 Experiments

### *4.1 Datasets and Evaluation Metrics*

We evaluated our method using two widely used benchmarks: VisDrone [41] and UAVDT [42].

The VisDrone dataset encompasses 8,599 high-resolution images from diverse UAV-mounted cameras across multiple environments, ranging from urban to rural scenes. The dataset includes 6,471 images in the training set, 548 in the validation set, and 1,580 in the test set. It spans ten object categories: pedestrian, people, bicycle, car, van, truck, tricycle, awning tricycle, bus, and motor. Following previous studies [11,13,43], we used 6,471 and 548 images for training and testing, respectively.

The UAVDT dataset consists of 23,258 images used for training and 15,069 for testing, following the previous work [11]. These images were captured at a resolution of 1,080 × 540 using a drone flying at low altitudes over urban areas. The tags include three types of vehicles: cars, buses, and trucks.

Accuracy is quantified using the AP along with AP at 0.5 ($\mathrm{AP}_{50}$) and 0.75 ($\mathrm{AP}_{75}$) IoU thresholds. Furthermore, we employed size-specific metrics—$\mathrm{AP}_{\mathrm{L}}$, $\mathrm{AP}_{\mathrm{M}}$, and $\mathrm{AP}_{\mathrm{S}}$—to evaluate detection efficacy across various object scales, where these metrics correspond to large, medium-sized, and small objects, respectively. Higher values in AP, $\mathrm{AP}_{50}$, $\mathrm{AP}_{75}$, $\mathrm{AP}_{\mathrm{L}}$, $\mathrm{AP}_{\mathrm{M}}$, and $\mathrm{AP}_{\mathrm{S}}$ indicate superior detection performance. For efficiency, we use GFLOPs and frames per second (FPS) as metrics to provide a holistic view of the model's performance.

### *4.2 Implementation Details*

Baseline model: DeNoising-DETR [44] with a pre-trained ResNet50 backbone is selected as our foundational model due to its rapid and reliable optimization, significantly reducing our training expenditures.

Training and testing strategies: The model is trained on the VisDrone dataset for 24 epochs with the AdamW optimizer, starting with a learning rate of 0.0002, which is reduced tenfold post-20 epochs. On the UAVDT dataset, the training lasts 12 epochs, with the learning rate following a similar reduction schedule after ten epochs. Consistent with [14], the learning rate for the backbone is set at $1 \times 10^{-5}$. In our ADL approach, the hyperparameters $\alpha$ and $\beta$ are set to 0.5 and 0.3, respectively. All training is performed on two NVIDIA A100 GPUs, with inference speed tests conducted on a single A100 GPU. The input resolutions we use in the VisDrone and UAVDT datasets are 1,333 × 800 and 800 × 800, while no test-time augmentation methods are used.

### *4.3 Quantitative Analysis*

This section quantitatively assesses our proposed methodology against recent state-of-the-art (SOTA) models utilizing two benchmark datasets for UAV-based object detection. To ensure a fair and comprehensive evaluation, we evaluate the models not only in terms of accuracy but also in terms of efficiency.

#### *4.3.1 Comparison with Other SOTA Models*

In our evaluation, we employed precision metrics (AP, $AP_{50}$, and $AP_{75}$) to assess the overall accuracy of our method on the VisDrone and UAVDT datasets. We benchmarked against other SOTA models, as depicted in Tables 1 and 2.

**Table 1:** Comparison with SOTAs on the validation set of the VisDrone dataset

| Model | Source | Backbone | AP | $AP_{50}$ | $AP_{75}$ |
|---|---|---|---|---|---|
| Cascade RCNN [22] | CVPR2018 | ResNet50 | 24.0 | 39.0 | 25.0 |
| YOLOv5-x [45] | GitHub2020 | CSPDarknet53 | 24.1 | 44.0 | 15.2 |
| GFL [46] | CVPR2021 | ResNet50 | 25.9 | 41.7 | 27.2 |
| ★ClusDet [28] | ICCV2019 | ResNet50 | 26.7 | 50.6 | 24.7 |
| QueryDet [13] | CVPR2022 | ResNet50 | 28.3 | 48.1 | 28.7 |
| CEASC [34] | CVPR2023 | Resnet18 | 28.7 | 50.7 | 28.4 |
| ★DMNet [47] | CVPRW2020 | ResNeXt101 | 29.4 | 49.3 | 30.6 |
| SDPNet [48] | TGRS2023 | ResNet50 | 30.2 | 52.5 | 30.6 |
| DSHNet [49] | WACV2021 | ResNet50 | 30.3 | 51.8 | 30.9 |
| ★GLSAN [27] | TIP2021 | ResNet50 | 30.7 | 55.4 | 30.0 |
| RT-DETR-X | Arxiv2023 | HGNetv2 | 31.0 | 52.0 | 30.9 |
| ★CZDet [10] | CVPRW2023 | ResNet50 | 33.2 | 58.3 | 33.2 |
| ★UFPMPDet [11] | AAAI2022 | ResNet50 | 36.6 | 62.4 | 36.7 |
| Our method | – | Resnet50 | 31.9 | 53.6 | 32.3 |

Note: The "★" indicates that the detector uses a coarse-to-fine pipeline.

**Table 2:** Comparison with SOTA models on the test set of the UAVDT dataset

| Model | Source | Backbone | AP | $AP_{50}$ | $AP_{75}$ |
|---|---|---|---|---|---|
| ★ClusDet [28] | ICCV2019 | ResNet50 | 13.7 | 26.5 | 12.5 |
| ★DMNet [47] | CVPRW2020 | ResNet50 | 14.7 | 24.6 | 16.3 |
| YOLOv5-x [45] | GitHub2020 | CSPDarknet53 | 14.9 | 25.1 | 14.7 |
| GFL [46] | CVPR2021 | ResNet50 | 15.7 | 28.1 | 16.5 |
| Cascade R-CNN [22] | CVPR2018 | ResNet50 | 16.0 | 28.0 | 17.4 |
| ★GLSAN [27] | TIP2021 | ResNet50 | 17.0 | 28.1 | 18.8 |
| CEASC [34] | CVPR2023 | Resnet18 | 17.1 | 30.9 | 17.8 |

(Continued)

**Table 2 (continued)**

| Model | Source | Backbone | AP | $AP_{50}$ | $AP_{75}$ |
|---|---|---|---|---|---|
| DSHNet [49] | WACV2021 | ResNet50 | 17.8 | 30.4 | 19.7 |
| ★UFPMPDet [11] | AAAI2022 | ResNet50 | 24.6 | 38.7 | 28.0 |
| Our method | – | Resnet50 | 21.1 | 36.3 | 22.8 |

Note: The "★" indicates that the detector uses a coarse-to-fine pipeline.

**VisDrone.** For the VisDrone dataset, our method's detection performance is cataloged in Table 1, juxtaposed with other cutting-edge models. Relative to general detectors such as Cascade R-CNN, YOLOv5-x, and Generalized Focal Loss (GFL), our approach achieves superior AP enhancements of 7.9%, 7.8%, and 6.0%, respectively. These improvements underscore the inadequacy of general-purpose models in UAV-based image detection tasks.

Furthermore, we contrasted our method with UAV-specific detection models. Against transformer-based QueryDet, our proposed solution achieves advancements of 3.6% in AP, 4.6% in $AP_{50}$, and 5.5% in $AP_{75}$. With Dual Sampler and Head Network (DSHNet), designed to address the long-tail distribution in UAV imagery, our method surpasses it by 1.6% in AP, 1.8% in $AP_{50}$, and 1.4% in $AP_{75}$. Against the recent innovations CEASC and Scale Decoupled Pyramid Network (SDPNet), utilizing the same ResNet50 backbone, our method demonstrates higher performance by 3.2% in AP, 2.9% in $AP_{50}$, and 3.9% in $AP_{75}$; and 1.7% in AP, 1.1% in $AP_{50}$, and 1.7% in $AP_{75}$, respectively. Even with the Density-Map guided object detection Network (DMNet), which employs the extensive ResNeXt101 backbone, our method maintains a lead of 2.5% in AP, 4.3% in $AP_{50}$, and 1.7% in $AP_{75}$. Notably, for multi-inference detectors ClusDet, DMNet, and GLSAN, which adopt a coarse-to-fine strategy, our method prevails with accuracies exceeding 5.2%, 2.5%, and 1.2% in AP. Compared to the latest research RT-DETR, our approach is 0.9% higher in AP, even when comparing variants with the largest sizes.

Our method does not obscure its comparative disadvantage in accuracy against the latest multi-inference detectors, lagging behind UFPMPDet and Cascaded Zoom-in Detector (CZDet) by 4.5% and 1.3% in AP, respectively. However, this precision is achieved at the expense of computational efficiency. UFPMPDet necessitates foreground packing before subsequent fine detection, and CZDet employs a higher test-time resolution (1,500 × 1,500) with adaptive cropping. We eschewed these operations to avoid compromising detection efficiency, aligning with our initial objective to design a practical detector. This trade-off will be examined in greater detail in the subsequent efficiency analysis.

**UAVDT.** The results for the UAVDT dataset are shown in Table 2. Our approach achieves satisfactory detection performance despite the UAVDT dataset containing a wide variety of scene images, and the data could be more unbalanced, leading to the unstable performance of many algorithms. When positioned against generic models like YOLOv5-x, GFL, and Cascade R-CNN, our proposed approach registers an improvement of AP: 6.2%, AP: 5.4%, and AP: 5.1%, respectively. Compared to the UAV image detectors CEASC and DSHNe, the approach we proposed has a higher AP: 4.0%, $AP_{50}$: 5.4%, $AP_{75}$: 5.0% than CEASC; higher AP: 3.3%, $AP_{50}$: 5.9%, $AP_{75}$: 3.1% than DSHNet. Compared with the existing superior multi-inference detectors, including ClusDet, DMNet, and GLSAN, the AP of our method is improved by 7.4%, 6.4%, and 4.1%. However, our approach is still 3.5% inferior on AP compared to UFPMPDet, in line with the VisDrone dataset.

Compared to other methods, our model achieves stable and consistent performance on both datasets, demonstrating the generality of our approach across different distributions.

### 4.3.2 Comparison of Multiscale Objects

It is worth noting that the UAV images have a more pronounced multiscale problem due to altitude variations and viewpoint offsets. Thus, the model must be robust to detect targets at multiple scales. To better demonstrate the effectiveness of our approach to the scale challenge, as shown in Table 3, we compare it with other methods through quantitative experiments at multiple scales.

**Table 3:** Multiscale comparison with other SOTA models on the validation set of the VisDrone dataset

| Model | Backbone | Resolution | AP | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|
| Cascade RCNN [22] | ResNet50 | 1,333 × 800 | 24.0 | 16.6 | 37.0 | 39.5 |
| YOLOv5-x [45] | CSPDarknet53 | 1,333 × 800 | 24.1 | 15.3 | 35.6 | 38.4 |
| GFL [46] | ResNet50 | 1,333 × 800 | 25.9 | 16.9 | 37.3 | 41.4 |
| ★ClusDet [28] | ResNet50 | 1,000 × 600 | 26.7 | 19.1 | 40.8 | 54.4 |
| ★DMNet [47] | ResNeXt101 | 1,000 × 600 | 28.2 | 21.6 | 41.0 | 56.9 |
| QueryDet [13] | ResNet50 | 2400 × 2400 | 28.3 | 17.9 | 30.4 | 36.7 |
| SDPNet [48] | ResNet50 | 1,333 × 800 | 30.2 | 22.6 | 39.6 | 39.8 |
| ★CZDet [10] | ResNet50 | 1,500 × 1,500 | 33.2 | 26.1 | 42.6 | 43.4 |
| Our method | Resnet50 | 1,333 × 800 | 31.9 | 23.2 | 42.9 | 51.0 |

Note: The "★" indicates that the detector uses a coarse-to-fine pipeline.

For Cascade RCNN, YOLOv5-x, and GFL, which use the same resolution input as us, our method has a significant advantage in accuracy on multiple scales. For both ClusDet and DMNet addressing the object scale variable distribution problem in UAV images, our accuracy exceeds both detectors by AP: 5.2% and AP: 3.7%. Our method performs better for QueryDet and SDPNet, which address the small object problem. On the small object subset, our method outperforms these two models by $AP_S$: 5.3% and $AP_S$: 0.6%, respectively. Our method achieves higher accuracy on medium and large targets beyond CZDet but lags a bit on small targets because CZDet uses a larger resolution and image cropping. In summary, our approach achieves satisfactory detection results on multiple scales, which shows that our model can cope well with scale variation challenges.

### 4.3.3 Comparison of Computational Complexity

To judge the superior performance of our method more comprehensively, we evaluated its computational complexity. The evaluation metric of inference speed can accurately reflect the computational complexity of the model; therefore, we use the inference speed FPS on the same hardware as a metric to compare with other works in this section, and the results are shown in Fig. 7.

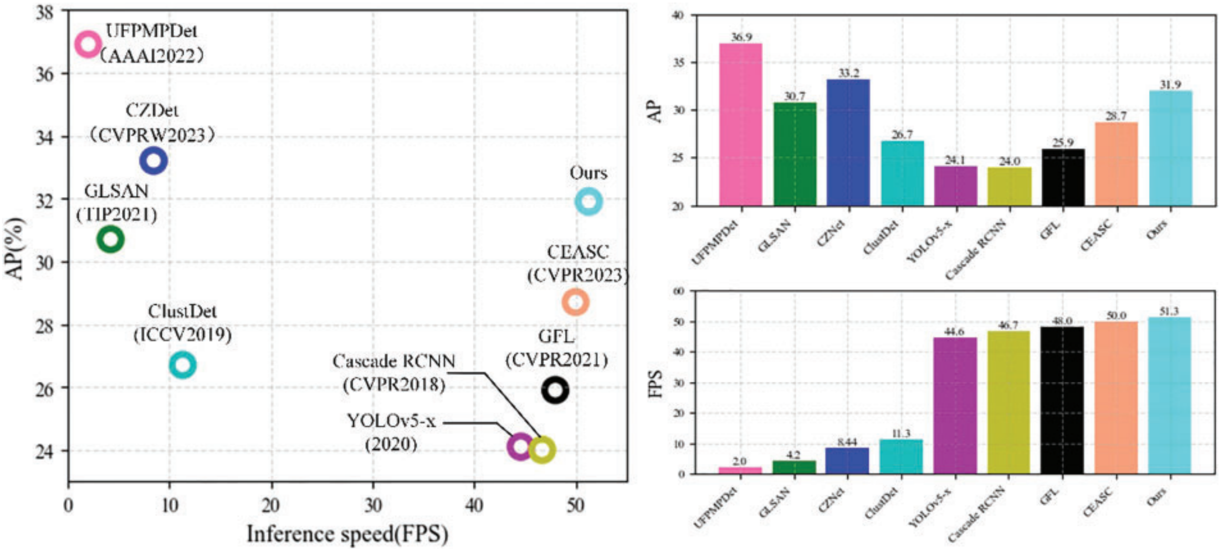


**Figure 7:** Comparison of accuracy and FPS of different models

Our method has higher accuracy and inference speed than the general detectors YOLOv5-x, Cascade RCNN, and GFL. Compared to the latest high-performance UAV image detector, CEASC, we achieve an FPS lead of 1.3, while the accuracy is 3.2% higher. At the same time, our method has a significant speed advantage over models that use a coarse-to-fine approach. The GLSAN requires an additional super-resolution procedure for densely populated regions, and ClusDet mandates multiple reviews of the segmented chips. Compared to these two models, our proposed method has higher accuracy and is about 12× and 5× faster in inference speed. Compared to UFPMNDet and CZDet, our method achieves about 25× and 6× speedups, respectively, thanks to our lightweight design and end-to-end pipeline. In contrast, the multi-inference step and large input resolution in UFPMNDet and CZDet increase the detection accuracy, making it impossible to achieve real-time detection (even on the high-performance A100 GPU).

The previous comparison highlights our contribution to UAV image detection. In UAV image detection, the balance between speed and accuracy is crucial, and our approach maximizes this balance, outperforming existing detectors.

### *4.4 Ablation Study*

A series of ablation studies are performed to elucidate the impact of individual modifications within the comprehensive model. Detailed findings from these studies are systematically compiled and presented in Table 4.

**Table 4:** Ablation studies about detection results on the validation set of VisDrone

| Method | N | AP | $AP_{50}$ | $AP_{75}$ | APs | $AP_M$ | $AP_L$ | GFLOPs | FPS |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 900 | 28.3 | 49.3 | 27.9 | 20.5 | 38.3 | 44.3 | 280.58 | 41.0 |
| Variant A | 900 | 26.4 | 46.3 | 25.7 | 19.4 | 36.7 | 44.4 | **143.36** | 49.5 |

(Continued)

**Table 4 (continued)**

| Method | N | AP | $AP_{50}$ | $AP_{75}$ | APs | $AP_M$ | $AP_L$ | GFLOPs | FPS |
|---|---|---|---|---|---|---|---|---|---|
| Variant B | 900 | 27.2 | 48.1 | 26.4 | 18.8 | 37.4 | 46.6 | 174.08 | 47.3 |
| Variant C | 900 | 28.5 | 50.6 | 28.0 | 19.8 | 39.1 | 46.4 | 149.50 | 48.2 |
| HTE | 900 | 28.9 | 50.8 | 28.5 | 20.4 | 39.6 | 45.3 | 149.52 | 50.5 |
| +QF | 900 | 29.9 | 52.0 | 29.5 | 21.0 | 40.0 | 52.3 | 149.41 | 51.9 |
| +ADL | 900 | 31.2 | 53.3 | 31.2 | 22.4 | 40.1 | 50.7 | 149.41 | **51.9** |
| +Queries | 1200 | **31.9** | **53.6** | **32.3** | **23.2** | **42.9** | 51.0 | 153.60 | 51.3 |
| +Queries | 1500 | 31.7 | 54.1 | 31.8 | 23.1 | 42.4 | **53.0** | 160.77 | 45.7 |

### *4.4.1 Effect of HTE*

To validate the superiority of our HTE, we experimented with a series of encoder designs, culminating in the HTE, as shown in Fig. 8. Compared to the original encoder, our Variant A slashes GFLOPs by 78% with a mere 1.9% drop in accuracy, revealing the baseline's inefficiency. Introducing a cross-scale fusion module, Variant B boosts accuracy by 0.8% over Variant A, confirming the merit of merging convolution with attention mechanisms. Variant C increases AP by 1.9% while cutting GFLOPs by 16.9%, supporting our design ideas in Section 3.1. Adding residual connectivity prevents possible performance degradation while improving accuracy by 0.4%. Compared to the baseline, our final HTE model cuts GFLOPs by an impressive 87% while raising accuracy by 0.6%, showing marked improvements in efficiency and performance.

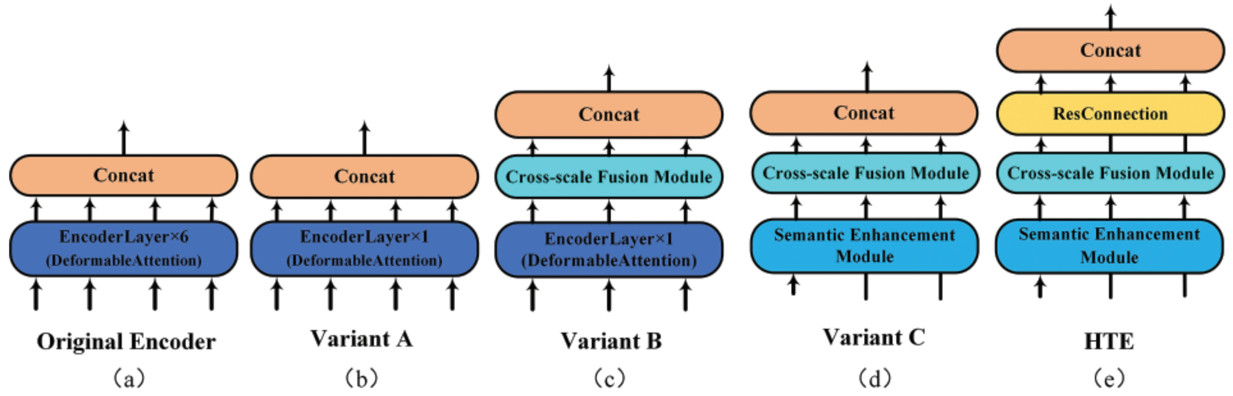
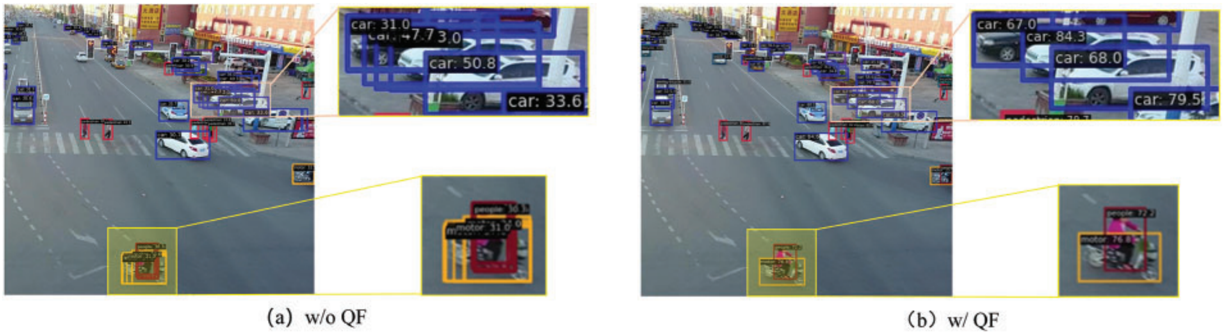


**Figure 8:** The evolution process of the HTE. (a) Baseline's encoder, (b)–(d) Encoder variants and (e) HTE

### *4.4.2 Effect of QF*

To visually encapsulate the impact of the QF, Fig. 9 contrasts the predictive outcomes before and after QF implementation, offering a tangible illustration of its capacity to minimize superfluous predictions of the model. The integration of the QF proved instrumental in excluding redundant predictions, thereby refining the precision of the model as quantified by enhancements in the mean AP, $AP_{50}$, and $AP_{75}$ metrics by 1.0%, 1.2%, and 1.0%, respectively. Additionally, the QF facilitates a more streamlined initialization of queries compared to the baseline methodology, which has the ancillary benefit of incrementally boosting the FPS.



(a) w/o QF

(b) w/ QF

**Figure 9:** Visualization of the effect of query filter

### *4.4.3 Effect of ADL*

Another approach we propose for improving small object detection in UAV images is ADL. As evidenced by the data in row 7 of Table 4, the integration of ADL furnishes a notable uplift in the model's performance metrics, with an increase in AP by 1.3%, $AP_{50}$ by 1.3%, and $AP_{75}$ by 1.7% across the entire dataset. More specifically, AP has a 1.4% enhancement for the subset of small objects. The ADL's deployment thus emerges as a pivotal modification, endowing the model with a heightened sensitivity to the nuanced characteristics of smaller instances and translating into a discernible leap in overall detection accuracy.

### *4.4.4 Effectiveness of Additional Queries*

Our investigation into the influence of query quantity reveals that an augmentation by 300 queries (to N =1200) yields a 0.7% ascension in accuracy, alongside a tolerable increase in computational load of 4.19 GFLOPs. Conversely, an escalation to 1500 queries results in a marginal 0.2% regression in accuracy and a 7.17 GFLOP hike in computational demand. This diminishment is ascribed to the confounding effects of excessive queries on the model's one-to-one matching capability. Consequently, the optimal model employs a query tally of N = 1200, striking a reasonable balance between precision and computational expenditure.

## ***4.5 Qualitative Analysis***

### *4.5.1 Visualization of Different Detectors*

To facilitate a more direct comparison of model performance, we visualize the predictions of some selected models, as shown in Fig. 10.

The first image, depicting a scene with minimal and conspicuous objects, reveals that the Cascade RCNN and CEASC models exhibit instances of missed detections. Conversely, models like YOLOv5-x, GFL, and QueryDet display a spectrum of false positives. Meanwhile, our model provides the most accurate predictions. The subsequent image presents a complex scenario with numerous small, densely clustered objects. All models, except ours, struggle with significant miss detection for small instances, including the latest iterations like CEASC. However, our method discerns the most challenging small objects at an equivalent input resolution.

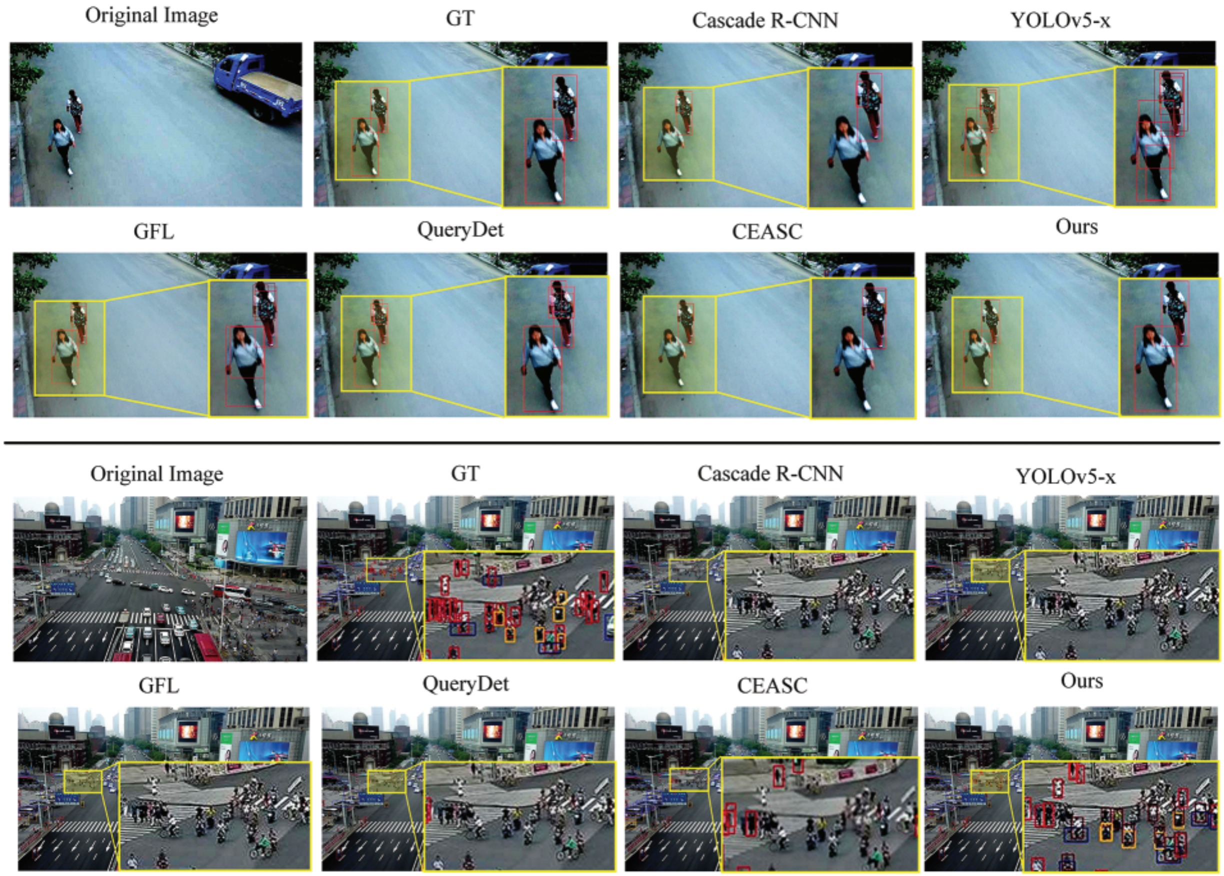


**Figure 10:** Visualization of comparative prediction outcomes among other models

In summary, our approach manifests superior detection proficiency in environments with sparse and dense object distributions, particularly distinguishing itself in accurately identifying small objects commonly elusive to conventional detection methods, which may be attributed to the global-aware attentional and the proposed modification.

### *4.5.2 Visualization in Challenging Scenes*

To more explicitly demonstrate the efficacy of our methodology, we present the detection results across various datasets in Fig. 11. It is observable that our detector consistently achieves high-performance detection across a multitude of intricate scenarios characterized by varying times of day (including day, dusk, and night), diverse perspectives (including side, frontal, and top views), and fluctuating lighting conditions (including low illumination and overexposure). Despite these complexities, our approach maintains commendable detection outcomes.

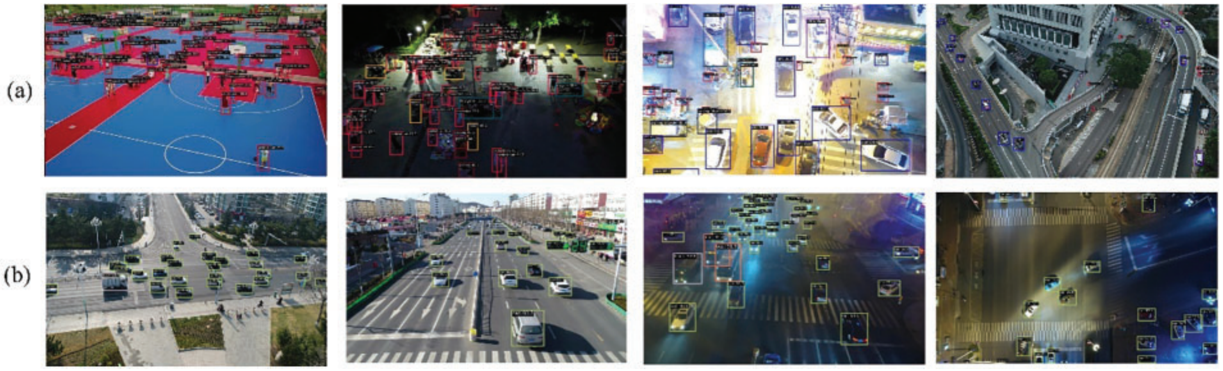


**Figure 11:** Detection results for different challenge scenarios. (a) VisDrone dataset. (b) UAVDT dataset

## 5 Conclusions

In this study, we have developed H-DETR, an end-to-end detector tailored for UAV image object detection. By hybridizing the attention-only encoder with a cross-scale fusion network, we designed HTE, which significantly simplifies the computation of the transformer. Supplementary to this, we propose the QF and the ADL to address the challenges posed by dense, small objects prevalent in UAV imagery. The QF leverages a training-aware non-maximum suppression to alleviate query consistency issues from density distribution in drone images. At the same time, the ADL augments the network's capacity to distinguish and extract discriminating features from small instances. Extensive experimental analyses on two distinct datasets have been conducted to ascertain the efficacy of our proposed approaches. Our model demonstrates superior detection speed on both datasets, achieving AP of 31.9% and 21.2%, respectively. Compared to more accurate models, our model attains speed improvements by 25× and 6×. This achievement underscores our model's unparalleled balance between precision and computational efficiency, surpassing existing CNN-based methods. Furthermore, our approach provides valuable insight for further integrating transformer architectures into UAV image analysis. However, it is essential to note that the currently available computational resources on UAV platforms are exceedingly limited, hindering our model's effective deployment on mobile platforms. As we look ahead, our research will aim to reduce computational requirements through strategies such as model pruning and knowledge distillation. Our ultimate objective is to realize the deployment of our model on UAV platforms in real-world scenarios, thereby facilitating further practical applications.

**Acknowledgement:** The authors would like to thank Xipeng Cao, author of CF-DETR, for his valuable suggestions during the experimental phase, and the authors would like to thank the editors and reviewers for their valuable work.

**Funding Statement:** This research was funded by the Natural Science Foundation of Hebei Province (F2021506004).

**Author Contributions:** The authors confirm contribution to the paper as follows: study conception and design: Changfeng Feng, Renke Kou; data collection: Changfeng Feng, Qiang Fu; analysis and interpretation of results: Changfeng Feng, Qiang Fu, Dongdong Zhang, Chunping Wang; draft manuscript preparation: Changfeng Feng. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** The data used in the study is publicly available. The download link has been noted in the manuscript.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

[1] X. Bai, L. Guo, H. Huo, J. Zhang, Y. Zhang and Z. L. Li, “Rse-net: Road-shape enhanced neural network for road extraction in high resolution remote sensing image,” *Int. J. Remote. Sens.*, May 2023. doi: 10.1080/01431161.2023.2214277.

[2] V. Kamath and A. Renuka, “Deep learning based object detection for resource constrained devices: Systematic review, future trends and challenges ahead,” *Neurocomputing*, vol. 531, pp. 34–60, Apr. 2023.

[3] Z. Huang, G. Li, X. Sun, Y. Chen, J. Sun and Z. Ni, “Siamese dense pixel-level fusion network for real-time UAV tracking,” *Comput. Mater. Contin.*, vol. 76, no. 3, pp. 3219–3238, 2023. doi: 10.32604/cmc.2023.039489.

[4] S. Ali, A. Jalal, M. Alatiyyah, K. Alnowaiser, and P. Jeongmin, “Vehicle detection and tracking in UAV imagery via YOLOv3 and kalman filter,” *Comput. Mater. Contin.*, vol. 76, no. 1, pp. 1249–1265, 2023. doi: 10.32604/cmc.2023.038114.

[5] G. Tian, J. Liu, and W. Yang, “A dual neural network for object detection in UAV images,” *Neurocomputing*, vol. 443, pp. 292–301, Jul. 2021.

[6] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[7] Z. Tian, C. Shen, H. Chen, and T. He, “FCOS: Fully convolutional one-stage object detection,” in *Proc. 2019 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 9626–9635.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788.

[9] J. Xu, Y. L. Li, and S. Wang, “AdaZoom: Towards scale-aware large scene object detection,” *IEEE Trans. Multimedia*, vol. 25, pp. 4598–4609, 2023.

[10] A. Meethal, E. Granger, and M. Pedersoli, “Cascaded zoom-in detector for high resolution aerial images,” in *Proc. 2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Vancouver, BC, Canada, Jun. 2023, pp. 2046–2055.

[11] Y. Huang, J. Chen, and D. Huang, “UFPMP-Det: Toward accurate and efficient object detection on drone imagery,” in *36th AAAI Conf. Artif. Intell.*, Vancouver, BC, Canada, Jun. 2022, pp. 1026–1033.

[12] T. Wang, Z. Ma, T. Yang, and S. Zou, “PETNet: A YOLO-based prior enhanced transformer network for aerial image detection,” *Neurocomputing*, vol. 547, pp. 126384, Aug. 2023.

[13] C. Yang, Z. Huang, and N. Wang, “QueryDet: Cascaded sparse query for accelerating high-resolution small object detection,” in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 13658–13667.

[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, “End-to-end object detection with transformers,” in *Proc. Eur. Conf. Comput. Vis. (ECCV) 2020*, Glasgow, UK, Aug. 2020, pp. 213–229.

[15] A. Deng, G. Han, D. Chen, T. Ma, X. Wei and Z. Liu, “Interframe saliency transformer and lightweight multidimensional attention network for real-time unmanned aerial vehicle tracking,” *Remote. Sens.*, vol. 15, no. 17, pp. 4249, Jan. 2023.

[16] F. Li *et al.*, “Lite DETR: An interleaved multi-scale encoder for efficient DETR,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 18558–18567.

[17] A. Vaswani *et al.*, “Attention is all you need,” in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA, Curran Associates Inc., Dec. 2017, pp. 6000–6010.

[18] A. Dosovitskiy *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv.2010.11929, 2021.

[19] S. Lin and B. Guo, “Swin Transformer: Hierarchical vision transformer using shifted windows,” in *2021 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, Canada, Oct. 2021, pp. 9992–10002.

[20] N. Wang, W. Zhou, J. Wang, and H. Li, “Transformer meets tracker: Exploiting temporal context for robust visual tracking,” in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 1571–1580.

[21] A. A. Aleissaee *et al.*, “Transformers in remote sensing: A survey,” *Remote. Sens.*, vol. 15, no. 7, pp. 1860, Mar. 2023.

[22] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6154–6162.

[23] W. Liu *et al.*, “SSD: Single shot multibox detector,” in *2016 Eur. Conf. Comput. Vis.*, Amsterdam, Netherlands, 2016.

[24] H. Law and J. Deng, “CornerNet: Detecting objects as paired keypoints,” *Int. J. Comput. Vision*, vol. 128, no. 3, pp. 642–656, Mar. 2020. doi: 10.1007/s11263-019-01204-1.

[25] D. Meng *et al.*, “Conditional DETR for fast training convergence,” in *2021 IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, Canada, Aug. 2021, pp. 3631–3640.

[26] B. Roh, J. W. Shin, W. Shin, and S. Kim, “Sparse DETR: Efficient end-to-end object detection with learnable sparsity,” arXiv:2111.14330, 2022.

[27] S. Deng *et al.*, “A global-local self-adaptive network for drone-view object detection,” *IEEE Trans. Image Process.*, vol. 30, pp. 1556–1569, 2021. doi: 10.1109/TIP.2020.3045636.

[28] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, “Clustered object detection in aerial images,” in *2019 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 8310–8319.

[29] J. Zhang, J. Huang, X. Chen, and D. Zhang, “How to fully exploit the abilities of aerial image detectors,” in *2019 IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, Korea (South), Oct. 2019, pp. 1–8.

[30] J. Feng, J. Wang, and R. Qin, “Lightweight detection network for arbitrary-oriented vehicles in UAV imagery via precise positional information encoding and bidirectional feature fusion,” *Int. J. Remote. Sens.*, vol. 44, no. 15, pp. 4529–4558, Aug. 2023. doi: 10.1080/01431161.2023.2197129.

[31] H. Zhang, M. Sun, Q. Li, L. Liu, M. Liu and Y. Ji, “An empirical study of multi-scale object detection in high resolution UAV images,” *Neurocomputing*, vol. 421, no. 5, pp. 173–182, Jan. 2021. doi: 10.1016/j.neucom.2020.08.074.

[32] S. Cao, T. Wang, T. Li, and Z. Mao, “UAV small target detection algorithm based on an improved YOLOv5s model,” *J. Vis. Commun. Image R.*, vol. 97, no. 4, pp. 103936, Dec. 2023. doi: 10.1016/j.jvcir.2023.103936.

[33] J. Cao, Y. Pang, J. Han, and X. Li, “Hierarchical regression and classification for accurate object detection,” *IEEE Trans. Neur. Net. Lear.*, vol. 34, no. 5, pp. 2425–2439, May 2023. doi: 10.1109/TNNLS.2021.3106641.

[34] B. Du, Y. Huang, J. Chen, and D. Huang, “Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images,” in *2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, Canada, Jun. 2023, pp. 13435–13444.

[35] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, “TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, Canada, Aug. 2021, pp. 2778–2788.

[36] C. Shen, C. Ma, and W. Gao, “Multiple attention mechanism enhanced YOLOX for remote sensing object detection,” *Sensors*, vol. 23, no. 3, pp. 1261, Jan. 2023. doi: 10.3390/s23031261.

[37] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path aggregation network for instance segmentation,” in *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8759–8768.

[38] C. Y. Wang, H. Y. M. Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh and I. H. Yeh, “CSPNet: A new backbone that can enhance learning capability of CNN,” in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, Jun. 2020, pp. 1571–1580.

[39] Z. Zheng *et al.*, "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," *IEEE Trans. Cybernetics*, vol. 52, no. 8, pp. 8574–8586, Aug. 2022. doi: 10.1109/TCYB.2021.3095305.

[40] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020. doi: 10.1109/TPAMI.2018.2858826.

[41] P. Zhu *et al.*, "Detection and tracking meet drones challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7380–7399, Nov. 2022. doi: 10.1109/TPAMI.2021.3119563.

[42] D. Du *et al.*, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Cham, Springer, 2018, pp. 370–386.

[43] R. Zhang, Z. Shao, X. Huang, J. Wang, and D. Li, "Object detection in UAV images via global density fused convolutional network," *Remote. Sens.*, vol. 12, no. 19, pp. 3140, Jan. 2020. doi: 10.3390/rs12193140.

[44] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni and L. Zhang, "DN-DETR: Accelerate DETR training by introducing query denoising," in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 13609–13617.

[45] G. Jocher, "YOLOv5 by Ultralytics," 2020. Accessed: Nov. 23, 2023. [Online]. Available: https://github.com/ultralytics/yolov5

[46] X. Li, W. Wang, X. Hu, J. Li, J. Tang and J. Yang, "Generalized focal loss V2: Learning reliable localization quality estimation for dense object detection," in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11632–11641.

[47] C. Li, T. Yang, S. Zhu, C. Chen, and S. Guan, "Density map guided object detection in aerial images," in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, Jun. 2020, pp. 737–746.

[48] Y. Ma, L. Chai, and L. Jin, "Scale decoupled pyramid for object detection in aerial images," *IEEE Trans. Geosci. Remote.*, vol. 61, pp. 1–14, 2023. doi: 10.1109/TGRS.2023.3298852.

[49] W. Yu, T. Yang, and C. Chen, "Towards resolving the challenge of long-tail distribution in UAV images for object detection," in *2021 IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, IEEE, Jan. 2021, pp. 3257–3266.