Enhancing DETRs for Small Object Detection via Multi-Scale Refinement and Query-Aided Mining

Sisi Fu FUSISI@M.SCNU.EDU.CN
Zhiming Chen ZHIMINGCHEN@M.SCNU.EDU.CN
Xiaocheng Fang FANGXIAOCHENG@M.SCNU.EDU.CN
Jieyi Cai CAIJIEYI@M.SCNU.EDU.CN
Huanyu Liu LIUHUANYU@M.SCNU.EDU.CN
Huosheng Wen WENHUOSHENG@M.SCNU.EDU.CN
Bingzhi Chen* CHENBINGZHI@M.SCNU.EDU.CN
School of Artificial Intelligence, South China Normal University, Foshan, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Small object detection (SOD) aims to precisely localize and accurately classify objects from limited spatial extent and discernible features. Despite significant advancements in object detection driven by CNN-based and Transformer-based methods, SOD remains a significant challenge. This is primarily due to their minimal spatial dimensions and distinct features which pose difficulties in both computational efficiency and effective supervision. Particularly, Transformer-based detectors suffer from the high computational cost caused by the introduction of a feature pyramid network (FPN) and the sparse supervision for the encoder output due to insufficient positive queries. Current approaches attempt to mitigate these issues through sparse attention mechanisms and auxiliary one-to-many label assignment strategies. However, these approaches often still suffer from inefficiencies in processing multi-scale information and a deficiency in generating adequate positive queries for small objects. To address this issue, we propose a novel small object detector MRQM, which integrates **M**ulti-scale **R**efinement and **Q**uery-aided **M**ining. The scale-aware encoder strategically refines features across multiple scales from a bi-directional feature pyramid network (BiFPN) through iterative updates. This process not only reduces redundant computations but also significantly enhances the representation of features at various scales. Furthermore, the IoU-aware head integrates the dynamic anchors mining strategy and one-to-many label assignments to fully mine potential high-quality auxiliary positive queries for small instances, and mitigate issues related to sparse supervision for the encoder. Extensive experiments on the SODA-D and VisDrone datasets consistently demonstrate the superiority and effectiveness of our MRQM method.

Keywords: small object detection, Transformer-based models, multi-scale information, auxiliary positive queries

1. Introduction

As a fundamental task in computer vision, object detection requires localizing and classifying the instances. In recent years, remarkable progress in object detection has been achieved, primarily driven by two main research directions: CNN-based Ren et al. (2015) and Transformer-based Zhu et al. (2020) methods. Compared to objects of general scales,

small objects pose unique difficulties due to their limited spatial extent and lower distinguishable features, which can lead to significant drops in detection performance. This is particularly crucial in applications such as autonomous driving, aerial surveillance, and medical imaging where small objects often carry critical information. In the task of SOD, Transformer-based methods encounter more challenges compared to the extensively studied CNN-based methods. This difficulty is primarily due to two factors: 1) **substantial computational cost limits the effective use of multi-scale features from FPNs in Transformer-based methods,** 2) **insufficient number and quality of positive queries for small instances lead to sparse supervision for the encoder output.**

Earlier studies Lin et al. (2017a); Tan et al. (2020) have demonstrated that FPNs efficiently integrate features across diverse scales and resolutions. This capability enhances their ability to comprehend both fine-grained details and the global contextual information of small target objects. However, FPNs often entail prohibitive computational costs, particularly when used with Transformer-based detectors. Simply applying multi-scale features from FPNs to Transformer encoder layers is impractical because the computational complexity scales quadratically with the number of feature tokens. Concretely, handling a feature map with a spatial size of $H \times W$ requires a computational cost of $O(HW)$ for CNN-based detectors, while the complexity of the attention mechanism in Transformer-based detectors is $O(H^2W^2)$. To address this problem, Deformable DETR Zhu et al. (2020) and Sparse DETR Roh et al. (2021) substitute the original global dense attention mechanism with sparse attention. SMCA-DETR Gao et al. (2021) confines most Transformer encoder layers to be scale-specific and allows only one layer to integrate multi-scale features. However, as the number of tokens scales quadratically with the feature map size from FPNs, these methods still face significant computational and memory demands Zhang et al. (2023a).

In DETRs, limited positive queries result in sparse supervision for the encoder output. This limitation significantly impairs the encoder's ability to learn discriminative features and also adversely affects attention learning in the decoder. To address this issue, Group-DETR Chen et al. (2023) constructs group-wise one-to-many label assignments to exploit multiple positive object queries. Co-DETR Zong et al. (2023) presents a novel collaborative hybrid assignments training scheme. However, these methods face challenges in effectively handling small objects, as their label assignment relies heavily on overlap or distance metrics. Small instances typically occupy a minimal area, resulting in sparse overlaps between anchors and ground truth that fall below conventional positive IoU thresholds. Consequently, the current one-to-many label assignment methods struggle to generate sufficient high-quality positive samples for small instances, resulting in an insufficient number and quality of positive queries for small instances optimization in the decoder.

To address these challenges, we propose a small object detector MRQM based on the scale-aware encoder and IoU-aware head. Concretely, motivated by the spatial redundancy of high-resolution features, we introduce a straightforward yet effective encoder block. This block partitions the multi-scale features from FPNs into high-level features (e.g., $\mathcal{F}_3, \mathcal{F}_4, \mathcal{F}_5$) and low-level features (e.g., $\mathcal{F}_2$). These high-level and low-level features are updated in an interleaved manner, ensuring efficient utilization of the multi-scale information. In addition, enlightened by the auxiliary head in Co-DETR Zong et al. (2023), we introduce an IoU-aware mechanism to dynamically optimize the anchors IoU threshold for small objects in the auxiliary head, which can mine more high-quality small object positive queries for the

decoder, thereby improving the small object detection performance. In summary, we make three main contributions:

- We introduce an efficient scale-aware encoder that iteratively refines multi-scale features from FPNs. This method enhances the representation of both high-level and low-level features in an interleaved manner, reducing redundant computations and thereby optimizing the utilization of multi-scale feature information.
- We introduce an auxiliary IoU-aware head that integrates the dynamic anchors mining strategy and one-to-many label assignments to generate sufficient high-quality positive queries for small instances. This approach mitigates the sparse supervision for the encoder and inadequate optimization in the decoder.
- The experiment results on the SODA-D and VisDrone datasets exhibit the superiority of our MRQM to detect these instances with extremely limited sizes. Specifically, MRQM significantly enhances the Co-DINO baseline, achieving an improvement of 2.3% AP on the SODA-D dataset and an increase of 3.4% AP on the VisDrone dataset.

2. Related Works

2.1. Object Detection

In general, existing object detection studies mainly encompass two streams: CNN-based object detectors and Transformer-based object detectors. Intuitively, CNN-based detectors, such as Faster R-CNN Ren et al. (2015), RetinaNet Lin et al. (2017b), and FCOS Tian et al. (2019), explicitly define surrogate regression and classification tasks, which require manual design of many components, such as anchor points, ROI thresholds, and non-maximum suppression (NMS). The intricate detection pipeline, hyper-parameter intensive, coupled with not fully end-to-end, results in difficulties in training and the sub-optimal performance of the CNN-based methods. Inspired by the Transformer paradigm, the recently proposed Transformer-based object detectors, such as DETR Carion et al. (2020), Deformable DETR Zhu et al. (2020), and DINO Zhang et al. (2022), aim to simplify the detection pipeline by leveraging self-attention mechanism. These models eliminate the need for hand-crafted components like anchors and NMS, enabling a more end-to-end approach. Transformer-based detectors offer improved global context modeling and achieve better detection performance. However, due to the computational complexity of the self-attention mechanism, challenges in computational efficiency and training convergence still remain.

2.2. Small Object Detection

Small object detection aims at the challenging task of accurately detecting small objects under the constraint of low-quality and low-resolution representation. Existing studies on small object detection primarily focus on enhancing feature representation, scale invariance, and robustness to occlusion and background clutter. CNN-based methods like Faster R-CNN Ren et al. (2015) and YOLO Redmon and Farhadi (2017) often struggle with small objects due to the limited resolution of feature maps and the inherent challenges in distinguishing fine details. To address these issues, techniques such as Feature Pyramid

Networks Lin et al. (2017a) have been introduced, enhancing feature hierarchies to capture the representation of small targets. Super-resolution-based methods Rabbi et al. (2020); Zhang et al. (2023b) aims at restoring the distorted structures of small objects instead of simply amplifying their ambiguous appearance, which empowers the model to mine the intrinsic correlations between small-scale objects and large-scale ones, thereby enhancing the semantic representation of small objects. Despite these advancements, SOD remains a challenging problem, necessitating further research to develop more deliberate designs for powerful paradigms working better in small object detection.

2.3. Spatial Redundancy

Inspired by the utilization of the spatial redundancy of CNNs, several works Sun et al. (2021); Zhang et al. (2023a) Zhang et al. (2023c) perform sparse operations over feature maps to avoid computation at less informative locations and focus on the most informative parts of the feature maps. Technically, PerforatedCNN Figurnov et al. (2016) and Dynamic Convolution Verelst and Tuytelaars (2020) generate pixel masks through different deterministic sampling and small gating networks, respectively. Xie et al.Xie et al. (2020) employ stochastic sampling and interpolation networks with Gumbel-Softmax distribution and sparsity loss to enhance the training of sparse masks. Specifically, PnP-DETR Wang et al. (2021a) dynamically allocates encoding operations to more informative feature tokens. QueryDet Yang et al. (2022), operating on feature pyramids, leverages sparse high-resolution feature computations guided by coarse predictions to enhance small object detection efficiency. IMFA Zhang et al. (2023a) exploits sparse multi-scale features from just a few crucial locations to improve refined detection. Sparse Semi-DETR Shehzadi et al. (2024) advances semi-supervised object detection by introducing a query refinement module and pseudo-label filtering, enhancing the detection of small objects in complex scenes.

2.4. Auxiliary Techniques

Auxiliary techniques have emerged as pivotal strategies to enhance the performance of Transformer-based models in SOD. Methods implemented in ViDT Song et al. (2021) and MDef-DETR Maaz et al. (2022) such as auxiliary decoding/encoding loss, improve training by introducing scale-specific objectives. Iterative box refinement Song et al. (2021) progressively enhances detection accuracy, while top-down supervision leverages semantic guidance for better object identification. Pre-training on extensive datasets followed by task-specific fine-tuning, as demonstrated by models like FP-DETR Wang et al. (2021b) and CBNet Cai et al. (2022), improves feature representation. Data augmentation Oksuz et al. (2020) addresses various imbalance problems. Techniques for improving the decoder such as one-to-many label assignments Zong et al. (2023) and denoising training Zhang et al. (2022) further refine model performance. Collectively, the existing auxiliary techniques provide a multifaceted approach to bolster the capabilities of Transformer models in SOD, addressing challenges such as class imbalance, localization accuracy, and convergence stability through a variety of innovative strategies.

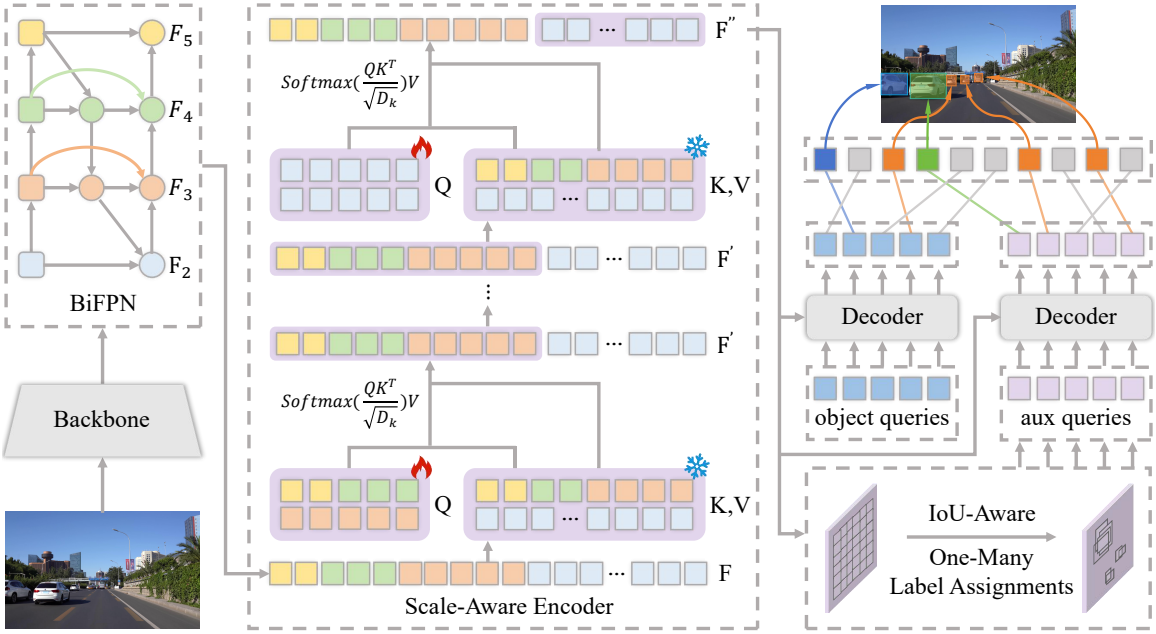

Figure 1: Architecture of the proposed Multi-Scale Refinement and Query-Aided Mining framework for SOD. **Left:** We utilize BiFPN to efficiently integrate features across diverse scales and resolutions, and use F2 ˜F5 to indicate the features from different BiFPN levels. **Middle:** The Scale-aware Encoder updates features from different levels at varying frequencies, thereby efficiently leveraging multi-scale features. **Right:** The IoU-aware head generates auxiliary queries via one-to-many label assignments and dynamic anchors mining. Queries from both the original and auxiliary branches jointly supervise the SAE output.

3. Methodology

In this section, we describe our MRQM for accurate small object detection. The main core of the proposed MRQM framework is the strategical refinement of multi-scale features from BiFPN through iterative updates while alleviating the sparse supervision for the encoder output by mining potential high-quality auxiliary positive queries for small instances. Fig. 1 illustrates the detection pipeline of the proposed MRQM, encompassing two essential modules: 1) Scale-aware encoder (SAE), and 2) IoU-aware head (IAH).

3.1. FPNs with Scale-Aware Encoder

The multi-scale features generated by FPNs possess an inherent structure: high-level features capture rich semantic information, and low-level features mainly capture local details related to small objects. However, the substantial redundancy in low-level features leads to significant computational costs when these multi-scale features are applied to Transformer encoder layers. Therefore, the iterative update of encoded image features is fundamental for MRQM to efficiently exploit multi-scale features. Specifically, we divide the multi-scale features $\mathcal{F}$ from BiFPN into low-level features $\mathcal{F}_L$ and high-level features $\mathcal{F}_H$. The

high-level features, considered the primary features, are updated more frequently, while the low-level features are updated less often. The limited number of frequently updated high-level features and the abundance of infrequently updated low-level features significantly reduce computational cost. As illustrated in Fig. 1, we effectively integrated features of different scales and resolutions using BiFPN, resulting in multi-scale features $\mathcal{F}$. The high-level features undergo multiple updates and are merged with the un-updated low-level features to form the multi-scale features $\mathcal{F}^{'}$ in SAE. Subsequently, the low-level features are updated once at the end of the encoder block and merged with the previously updated high-level features to produce the final refined multi-scale features $\mathcal{F}^{''}$, which serve as the encoder's output. This approach allows for the maintenance of a full-scale feature pyramid while substantially lowering computational costs.

Naively incorporating multi-scale features into the encoder leads to prohibitive computational complexity, as the sheer number of feature tokens across all scales is too large to be efficiently processed by the attention mechanism. This challenge motivates us to perform self-attention interaction updates at varying frequencies for hierarchical features with different token counts. In SAE, the high-level features $\mathcal{F}_H$ serve as queries Q to extract features from all scales, encompassing both low-level and high-level feature tokens. Formally, the update process can be described as

$$\mathcal{F}_H = Concat(\mathcal{F}_i), i \in [3,5];\ \mathcal{F}_L = \mathcal{F}_2, \tag{1}$$

$$Q = \mathcal{F}_H;\ K = V = \mathcal{F}, \tag{2}$$

$$\mathcal{F}_H^{'} = DeformAttn(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{3}$$

$$\mathcal{F}^{'} = Concat(\mathcal{F}_H^{'}, \mathcal{F}_L), \tag{4}$$

where $Concat$ refers to the concatenation of high-level features $\mathcal{F}_H$ into multi-scale features. Here, the query Q represents the initial high-level features $\mathcal{F}_H$, while the key K and value V correspond to the initial features $\mathcal{F}$ from all levels. After Q, K, and V are processed through the deformable attention layer, Q is updated. The updated high-level features $\mathcal{F}_H^{'}$ are then combined with the original low-level features $\mathcal{F}_L$ to obtain $\mathcal{F}^{'}$.

Low-level features $\mathcal{F}_L$ contain an excessive number of tokens, which is a critical factor in inefficient computation. To address this, SAE updates these low-level features at a lower frequency following a sequence of high-level feature fusion. Specifically, we use the initial low-level features, denoted as $\mathcal{F}_L$, as queries to interact with $\mathcal{F}^{'}$ to update their representation. Therefore, the update process can be formulated as

$$Q = \mathcal{F}_L;\ K = V = \mathcal{F}^{'}, \tag{5}$$

$$\mathcal{F}_L^{'} = DeformAttn(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{6}$$

$$\mathcal{F}^{''} = Concat(\mathcal{F}_H^{'}, \mathcal{F}_L^{'}). \tag{7}$$

Here, the query Q represents the initial low-level features $\mathcal{F}_L$, while the key K and value V correspond to $\mathcal{F}^{'}$. After Q, K, and V are processed through the deformable attention

layer, Q is updated. The updated low-level features $\mathcal{F}_L^{'}$ are then combined with the updated high-level features $\mathcal{F}_H^{'}$ to produce the output $\mathcal{F}^{''}$ of the SAE. SAE iteratively updates the multi-scale features from BiFPN, significantly reducing computational costs. This allows Transformer-based models to effectively integrate with FPNs.

3.2. IoU-Aware Head

In the one-to-one set matching paradigm of DETRs Carion et al. (2020), each ground truth will only be assigned to one specific query as the supervised target. Too few positive queries lead to inefficient cross-attention learning in the Transformer decoder. To alleviate this, Co-DETR Zong et al. (2023) employs an auxiliary one-to-many label assignment strategy to enrich the supervision of the encoder output, compelling it to be sufficiently discriminative to support the training convergence of these heads. However, this approach is insufficient for small objects. While the universal auxiliary head performs well on objects of general scales, it struggles with extremely small objects due to its inherent limitations. Specifically, the universal auxiliary head employs either overlap-based Lin et al. (2017b) or distance-based Tian et al. (2022) strategies to select positive anchors for training. These methods cannot ensure sufficient potential anchors for small objects, which have significantly smaller center regions. In other words, the positive sample criteria of the universal auxiliary head are overly stringent when applied to small or tiny objects, resulting in a limited number of samples available for optimization. Consequently, the universal auxiliary head fails to generate enough positive queries for small instances in the decoder.

To remedy the aforementioned issues of the universal auxiliary head in handling small instances, we introduce an IoU-aware mechanism. This mechanism incorporates a dynamic anchors mining strategy that adapts to the size variations of instances, thereby facilitating enhanced anchors mining for small objects. The IoU-aware mechanism dynamically assigns variable anchors IoU thresholds based on the object's area, ensuring that instances of varying sizes have access to a sufficient number of potential anchors. Concretely, for an object box with width w and height h, any anchors with an IoU exceeding threshold T_I will be considered positive samples. The threshold T_I is defined as follows:

$$T_I = \max(T_b, T_b + k \cdot (\frac{\sqrt{w \cdot h}}{A_s})^p), \tag{8}$$

where T_b represents the base IoU threshold. The constant k and the exponent p are used to control the growth rate of the threshold T_I and we often set k to 0.15 in experiments. Furthermore, A_s denotes the area of the minimal instance in the dataset, which ensures an adequate sampling of extreme-size objects and can be adjusted to suit different datasets. In essence, the IoU-aware mechanism is designed to assign lower thresholds to smaller targets but ensure these thresholds do not fall below T_b. This approach prevents an excessive number of low-quality samples from adversely affecting the optimization process. Meanwhile, the model employs a smoothly varying continuous threshold for determining positive samples, which is a more nuanced method.

In recent advancements in object detection, the introduction of a one-to-many label assignment strategy significantly enhances the labeling mechanism. Unlike the traditional one-to-one matching approach, where each ground-truth box is associated with a single

positive sample, the one-to-many strategy enables multiple candidates to correspond with the same ground-truth box. This paradigm shift is particularly advantageous for small object detection, as it expands the pool of positive samples, effectively addressing the sparse supervision challenge often faced with these instances.

Utilizing the dynamic anchors mining strategy and one-to-many label assignments of the IoU-aware auxiliary head, we meticulously generate an ample number of customized positive queries for the decoder. The mined customized queries Q_a can be generated using the following formula:

$$Q_a = Linear(PE(C_{pos})) + Linear(E(F''_{pos})), \tag{9}$$

where $PE(\cdot)$ stands for positional encodings and C_{pos} denotes the set of positive coordinates identified by the IoU-aware auxiliary head. The set of features, F''_{pos}, is extracted from the output of the Scale-Aware Encoder, specifically tailored according to the positive sample coordinate set C_{pos}. To facilitate the selection, we extract the corresponding features from $E(\cdot)$ based on the established index pair. The notation $Linear(\cdot)$ refers to a linear transformation applied to the feature set.

Consequently, the auxiliary one-to-many label assignment branches share the same parameters with L decoder layers in the original main branch during training. Consequently, all queries in the auxiliary branch are treated as positive queries, eliminating the need for a matching process. Specifically, the loss for the l-th decoder layer in the auxiliary branch can be expressed as follows:

$$\mathcal{L}_l^{dec} = \hat{\mathcal{L}}(\hat{P}_l, P_{pos}). \tag{10}$$

$\hat{P}_l$ refers to the output predictions of the l-th decoder layer in the auxiliary branch, and P_{pos} refers to the positive proposals generated by the auxiliary branch. In IAH, we employ a one-to-many label assignments strategy and a dynamic anchors mining strategy to effectively mine a sufficient number of potential auxiliary positive queries for small instances.

4. Experiments

In this session, we compare our proposed MRQM on the SODA-D and VisDrone datasets for small object detection with several state-of-the-art methods. After that, we conduct ablation studies, parameter analysis, and visualization to evaluate the effectiveness of each component in our MRQM framework.

4.1. Dataset

SODA-D. Based on the traffic scenarios, SODA-D Cheng et al. (2023) consists of 24,828 high-quality images and 278,433 instances from nine categories, including people, rider, bicycles, motor, vehicle, traffic-sign, traffic light, traffic camera, and warning-cone. SODA-D exhibits significant diversity in time periods, geographical locations, weather conditions, shooting perspectives, and scenarios, benefitting the generalization of small object detection.
VisDrone. VisDrone Du et al. (2019) contains 10,209 high-resolution drone-captured images with a resolution of 2000 $\times$ 1500 and 542,000 instances covering 10 common object

Table 1: Comparison with detection approaches on the SODA-D: the subscripts eS, rS, gS, and N represent extremely small, relatively small, generally small and normal respectively.

Methods	**Ref.**	**AP**	**AP$_{50}$**	**AP$_{75}$**	**AP$_{eS}$**	**AP$_{rS}$**	**AP$_{gS}$**	**AP$_{N}$**	**Params**
CNN-based									
Faster RCNN Ren et al. (2015)	NIPS'15	28.9	59.4	24.1	13.8	25.7	34.5	43.0	**41M**
FCOS Tian et al. (2019)	ICCV'19	23.9	49.5	19.9	6.9	19.4	30.9	40.9	**32M**
ATSS Zhang et al. (2020)	CVPR'20	26.8	55.6	22.1	11.7	23.9	32.2	41.3	**32M**
YOLOX Ge et al. (2021)	CVPR'21	26.7	53.4	23.0	13.6	25.1	30.9	30.4	**99M**
Sparse RCNN Sun et al. (2021)	CVPR'21	24.2	50.3	20.3	8.8	20.4	30.2	39.4	**106M**
RFLA Xu et al. (2022)	ECCV'22	29.7	60.2	25.2	13.2	26.9	35.4	44.6	**45M**
CFINet Yuan et al. (2023)	ICCV'23	30.7	60.8	26.7	14.7	27.8	36.4	44.6	**49M**
Transformer-based									
Deformable-DETR Zhu et al. (2020)	ICLR'20	19.2	44.8	13.7	6.3	15.4	24.9	34.2	**40M**
Conditional-DETR Meng et al. (2021)	ICCV'21	25.7	52.8	15.0	7.9	20.3	28.0	36.5	**46M**
DAB-DETR Liu et al. (2021)	ICLR'22	27.2	55.1	20.6	10.3	22.5	31.9	37.2	**55M**
DINO Zhang et al. (2022)	ICLR'23	28.9	59.4	22.4	12.5	22.7	34.7	42.8	**56M**
Co-DINO Zong et al. (2023)	ICCV'23	32.2	61.1	28.9	15.3	28.4	38.9	48.4	**66M**
MRQM(Ours)	-	**34.5**	**65.1**	**31.4**	**16.5**	**30.6**	**41.4**	**52.3**	**60M**

categories in traffic scenarios. Its various environmental settings (urban and rural), scenes with different population densities, viewpoint variations, and heavy occlusions all pose severe challenges to small target detection.

4.2. Implementation Details

We implement our approach based on Pytorch and we mainly conduct experiments with Co-DINO–a state-of-the-art Transformer-based object detector with open-sourced implementation. All experiments in this study were conducted on a single RTX 3090 GPU. During the training phase, we employed AdamW as the parameter optimizer with a batch size of 2. Data augmentation was limited to random flipping. The model was trained using a 1× schedule (12 epochs), with an initial learning rate of 2e-4 and a weight decay of 0.0001.

4.3. Metrics

In our experiments, we assess performance using Average Precision(AP), calculated across multiple Intersections over Union (IoU) thresholds from 0.5 to 0.95 in increments of 0.05. Additionally, we report AP50 and AP75, corresponding to IoU thresholds of 0.5 and 0.75, respectively. The definition of AP varies across scales in the Visdrone and SODA-D datasets. In Visdrone, objects are classified into small (S), medium (M), and large (L) categories, representing sizes within (0, 32^2], (32^2, 96^2], and (96^2, ∞]. Hence, AP$_S$, AP$_M$, and AP$_L$ denote precision metrics for objects of different scales. Conversely, in the SODA-D dataset, objects are categorized as Small or Normal according to their areas. The category further

Table 2: Comparison with detection approaches on the Visdrone2019: the subscripts S, M, and L respectively indicate the size of the object as small, medium, and large.

Methods	Ref.	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params
CNN-based								
Faster RCNN Ren et al. (2015)	NIPS'15	20.1	33.4	21.1	12.1	30.4	40.0	**41M**
FCOS Tian et al. (2019)	ICCV'19	15.2	25.5	15.7	7.8	23.7	29.7	**32M**
ATSS Zhang et al. (2020)	CVPR'20	18.7	30.7	19.3	10.9	29.0	32.5	**32M**
YOLOX Ge et al. (2021)	CVPR'21	23.5	39.5	23.8	14.8	34.2	39.0	**99M**
Sparse RCNN Sun et al. (2021)	CVPR'21	8.5	15.5	8.1	4.9	12.5	18.6	**106M**
RFLA Xu et al. (2022)	ECCV'22	25.4	42.2	26.0	17.9	34.4	45.7	**45M**
CFINet Yuan et al. (2023)	ICCV'23	26.0	45.3	26.1	18.3	35.3	49.9	**49M**
Transformer-based								
Deformable-DETR Zhu et al. (2020)	ICLR'20	14.2	25.7	13.8	8.0	21.7	30.1	**40M**
Conditional-DETR Meng et al. (2021)	ICCV'21	26.4	37.7	27.4	16.2	36.2	35.6	**46M**
DAB-DETR Liu et al. (2021)	ICLR'22	27.8	40.8	26.9	16.4	35.5	39.8	**55M**
DINO Zhang et al. (2022)	ICLR'23	26.8	44.2	28.9	17.5	37.3	41.3	**56M**
Co-DINO Zong et al. (2023)	ICCV'23	28.5	46.7	29.9	20.5	38.2	46.3	**66M**
MRQM(Ours)	-	**31.9**	**51.4**	**33.6**	**24.1**	**41.7**	**53.4**	**60M**

Table 3: Ablation studies on two modules including BiFPN with scale-aware encoder (BiFPN with SAE) and IoU-aware head (IAH) on the VisDrone dataset.

Settings	BiFPN with SAE	IAH	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
I	✗	✗	28.5	46.7	29.9	20.5	38.2	46.3
II	✓	✗	30.7	49.1	30.7	23.7	39.5	50.8
III	✗	✓	30.1	48.5	31.6	23.1	38.9	51.0
IV	✓	✓	31.9	51.4	33.6	24.1	41.7	53.4

divides into extremely small (eS), relatively small (rS), generally small (gS), and normal(N), spanning size ranges of (0, 144], (144, 400], (400, 1024], and (1024, 2000], respectively.

4.4. Comparisons with State-of-The-Arts

To demonstrate the superiority of our MRQM approach, we conduct a comprehensive comparison with a wide range of state-of-the-art methods on the VisDrone and SODA-D. To ensure a fair comparison, we employ ResNet-50 as the underlying backbone architecture for all baseline models on two benchmark datasets.

Evaluation on SODA-D. As shown in Table 1, it can be observed that the proposed MRQM method greatly benefits from the SAE and the IAH outperforms all comparative baselines on benchmark datasets. Compared to the current best-performing method for

Table 4: Different definitions about the base IoU threshold T_b for IoU-aware head.

T_b	**AP**	**AP**$_{50}$	**AP**$_{75}$	**AP**$_S$	**AP**$_M$	**AP**$_L$
0.5	31.2	50.8	32.9	23.6	40.2	52.2
0.55	31.9	51.4	33.6	24.1	41.7	53.4
0.6	31.5	51.1	33.2	23.7	40.8	52.6
0.65	30.8	50.1	32.5	23.4	39.9	51.5

Table 5: Different definitions about the exponent p for IoU-aware head.

p	**AP**	**AP**$_{50}$	**AP**$_{75}$	**AP**$_S$	**AP**$_M$	**AP**$_L$
0.1	31.7	51.2	33.0	24.1	41.8	53.0
0.2	31.9	51.4	33.6	24.1	41.7	53.4
0.3	31.3	50.8	32.3	23.8	40.9	52.7

SOD, i.e., Co-DINO, our methods exhibit performance improvements in overall average precision(AP) score on the SODA-D dataset, achieving improvements of 2.3%(34.5% vs. 32.2%). In particular, for the three performance metrics of detecting small objects AP$_{eS}$, AP$_{rS}$ and AP$_{gS}$, our method also yields a substantial improvement of 1.2%, 2.2%, and 2.5%, respectively. These comparative results underscore the effectiveness of our approach.

Evaluation on VisDrone. According to Table 2, it can be observed that our MRQM framework also demonstrates impressive enhancements on the drone-captured dataset, even in challenging more complex environmental conditions, varying population densities, and severe occlusion scenarios. Intuitively, our model exhibits superior performance, surpassing all baseline models and achieving an overall AP score of 31.9%. Moreover, our method demonstrates notable superiority in small object detection, outperforming the second-best by 3.6% in AP$_S$. Additionally, AP$_M$ and AP$_L$ accuracies improve by 3.5% and 7.1%, respectively, highlighting the generalization and robustness of our approach. Notably, while enhancing the performance of detecting small objects, our method maintains a lower parameter count(**60M**), thereby improving detection efficiency. In contrast to the current state-of-the-art method Co-DINO, our framework not only achieves superior performance but also reduces parameter complexity. These compelling results underscore the effectiveness and dominance of our MRQM approach in advancing small object detection.

4.5. Ablation Studies

In our ablation studies, we conduct a comprehensive analysis of the proposed MRQM method on the VisDrone to validate the effectiveness of the different components. The comparative results are presented in Table 3.

Effect of BiFPN with Scale-Aware Encoder. The results of "MRQM w/o BiFPN with SAE" clearly indicate that the BiFPN efficiently integrates features across diverse scales and resolutions to enhance the generalization ability of Transformer-based detectors for detecting instances of various sizes. While the scale-aware encoder refines multi-scale

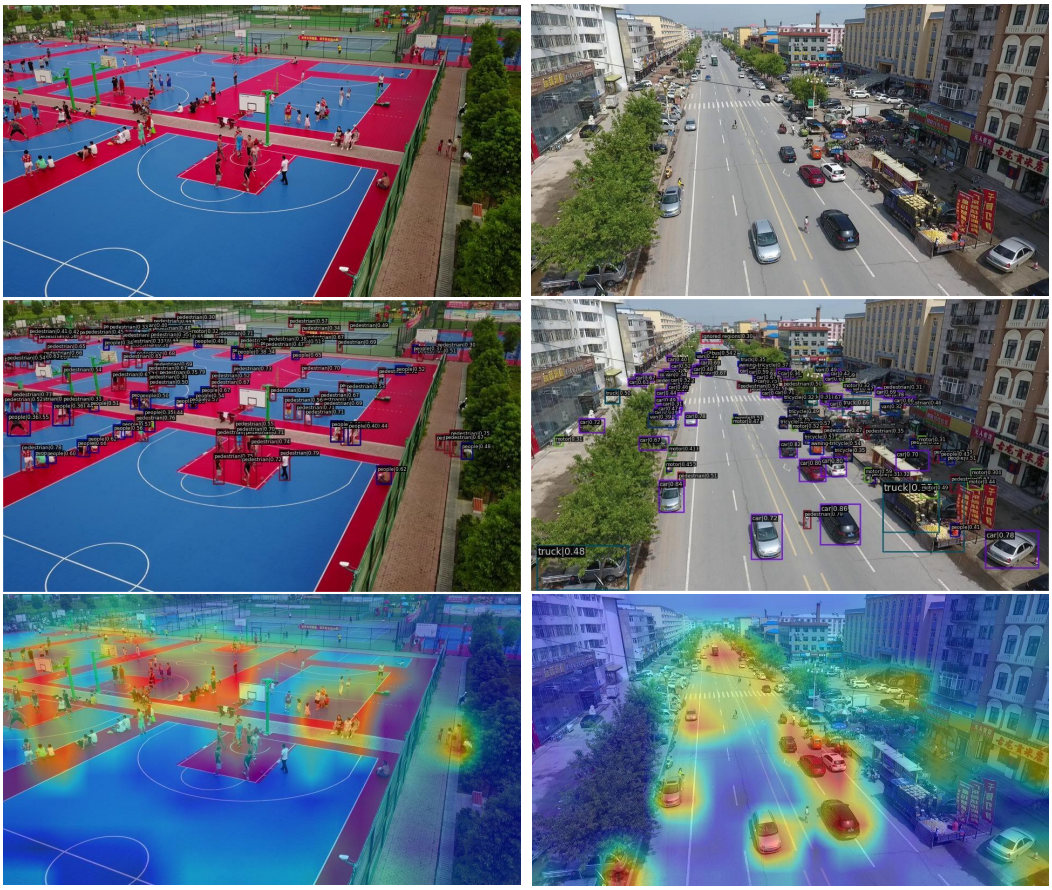

Figure 2: Visualization of the detection results and the query heatmap for small objects of our MRQM on the VisDrone dataset.

features from BiFPN, reducing redundant computations and thereby optimizing the utilization of multi-scale feature information.

Effect of IoU-Aware Head. Without the IoU-aware head, "MRQM w/o IAH" shows a notable drop in performance, highlighting the contribution of the IoU-aware head in fully mining sufficient potential auxiliary positive queries of small instances for the decoder to alleviate the sparse supervision for the encoder output.

4.6. Parameter Analysis

For parameter analysis, we examine the optimal performance of the auxiliary query mining strategy within the IAH. We conducted a comprehensive analysis under different configurations. As shown in Tables 4 and 5, the best results occur when T_b is 0.55 and p is 0.2. Introducing smoothly varying dynamic thresholds improved detection performance. However, performance declines when T_b drops to 0.5, likely due to low-quality samples dominating. The exponent p controls the growth rate of the dynamic IoU threshold T_I. At $p = 0.2$, T_I increases steadily, ensuring balanced detection performance.

4.7. Visualization Results

To further evaluate the performance of our MRQM approach, we visualize the detection results and the query heatmaps for small objects on VisDrone. As illustrated in Figure 2, by introducing SAE to leverage more discriminative features, our multi-scale refinement methods can more accurately detect small objects. Meanwhile, it can be clearly seen that our model can successfully recognize most small objects with high confidence scores and our IAH can locate the coarse positions of small objects by mining small target positive queries, enabling our model to detect them more effectively. In particular, We also show some limitations. In the second image of the query heatmap shown in Figure 2, some regions of large objects are falsely activated. Although these areas are not misdetected, this causes the detection head to process irrelevant locations, thereby impairing detection efficiency.

5. Conclusions

In this paper, we proposed a small object detector MRQM based on the scale-aware encoder and IoU-aware head, in which the former can iteratively refine multi-scale features within the encoder module, reducing redundant computations and thereby optimizing the utilization of multi-scale feature information. Then the IoU-aware head mines sufficient auxiliary queries for small objects for the decoder to alleviate the sparse supervision of the encoder output. Extensive experiments on the small object detection datasets SODA-D and VisDrone consistently demonstrate our approach outperforming state-of-the-art methods.

6. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62302172, 62176077, and 62176065), in part by the Guangdong International Science and Technology Co.operation Project (Grant No. 2023A0505050108), in part by the Shenzhen Key Technical Project (Grant Nos. JSGG20220831092805009, JSGG20201103153 802006), and in part by the Opening Project of Guangdong Province Key Laboratory of Information Security Technology (Grant No. 2023B1212060026).

References

Likun Cai, Zhi Zhang, Yi Zhu, Li Zhang, Mu Li, and Xiangyang Xue. Bigdetection: A large-scale benchmark for improved object detector pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 4777–4787, 2022.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 213–229. Springer, 2020.

Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group detr: Fast detr training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 6633–6642, 2023.

Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)*, 2023.

Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops(ICCVW)*, pages 0–0, 2019.

Mikhail Figurnov, Aizhan Ibraimova, Dmitry P Vetrov, and Pushmeet Kohli. Perforatedcnns: Acceleration through elimination of redundant convolutions. *Advances in Neural Information Processing Systems(NIPS)*, 29, 2016.

Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of detr with spatially modulated co-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, pages 3621–3630, 2021.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 2117–2125, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, pages 2980–2988, 2017b.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. In *International Conference on Learning Representations(ICLR)*, 2021.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *European Conference on Computer Vision(ECCV)*, pages 512–531, 2022.

Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, pages 3651–3660, 2021.

Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)*, 43(10):3388–3415, 2020.

Jakaria Rabbi, Nilanjan Ray, Matthias Schubert, Subir Chowdhury, and Dennis Chao. Small-object detection in remote sensing images with end-to-end edge-enhanced gan and object detector network. *Remote Sensing*, 12(9):1432, 2020.

Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 7263–7271, 2017.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pages 2164–2173. MIT Press, 2015.

Byungseok Roh, JaeWoong Shin, Wuhyun Shin, and Saehoon Kim. Sparse detr: Efficient end-to-end object detection with learnable sparsity. *arXiv preprint arXiv:2111.14330*, 2021.

Tahira Shehzadi, Khurram Azeem Hashmi, Didier Stricker, and Muhammad Zeshan Afzal. Sparse semi-detr: Sparse learnable queries for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 5840–5850, 2024.

Hwanjun Song, Deqing Sun, Sanghyuk Chun, Varun Jampani, Dongyoon Han, Byeongho Heo, Wonjae Kim, and Ming-Hsuan Yang. Vidt: An efficient and effective fully transformer-based object detector. In *International Conference on Learning Representations(ICLR)*, 2021.

Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 14454–14463, 2021.

Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10781–10790. IEEE, 2020.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019.

Zhi Tian, Xiangxiang Chu, Xiaoming Wang, Xiaolin Wei, and Chunhua Shen. Fully convolutional one-stage 3d object detection on lidar range images. *Advances in Neural Information Processing Systems(NIPS)*, 35:34899–34911, 2022.

Thomas Verelst and Tinne Tuytelaars. Dynamic convolutions: Exploiting spatial sparsity for faster inference. In *Proceedings of the ieee/cvf Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 2320–2329, 2020.

Tao Wang, Li Yuan, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. Pnp-detr: Towards efficient visual analysis with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, pages 4661–4670, 2021a.

Wen Wang, Yang Cao, Jing Zhang, and Dacheng Tao. Fp-detr: Detection transformer advanced by fully pre-training. In *International Conference on Learning Representations(ICLR)*, 2021b.

Zhenda Xie, Zheng Zhang, Xizhou Zhu, Gao Huang, and Stephen Lin. Spatially adaptive inference with stochastic feature sampling and interpolation. In *European Conference on Computer Vision(ECCV)*, pages 531–548. Springer, 2020.

Chang Xu, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. Rfla: Gaussian receptive field based label assignment for tiny object detection. In *European Conference on Computer Vision(ECCV)*, pages 526–543. Springer, 2022.

Chenhongyi Yang, Zehao Huang, and Naiyan Wang. Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 13668–13677, 2022.

Xiang Yuan, Gong Cheng, Kebing Yan, Qinghua Zeng, and Junwei Han. Small object detection via coarse-to-fine proposal generation and imitation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, pages 6317–6327, 2023.

Gongjie Zhang, Zhipeng Luo, Zichen Tian, Jingyi Zhang, Xiaoqin Zhang, and Shijian Lu. Towards efficient use of multi-scale features in transformer-based object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 6206–6216, 2023a.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

Jiaqing Zhang, Jie Lei, Weiying Xie, Zhenman Fang, Yunsong Li, and Qian Du. Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023b.

Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 9759–9768, 2020.

Shilong Zhang, Xinjiang Wang, Jiaqi Wang, Jiangmiao Pang, Chengqi Lyu, Wenwei Zhang, Ping Luo, and Kai Chen. Dense distinct query for end-to-end object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 7329–7338, 2023c.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations(ICLR)*, 2020.

Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, pages 6748–6758, 2023.