Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3071

TFE-NET: A TINY-AWARE FEATURE ENHANCEMENT NETWORK FOR COLLABORATIVE OPTIMIZATION IN SMALL OBJECT DETECTION

Qiang Zeng*, YiDan Chen
School of Computer Science and Artificial Intelligence, Beijing Technology and Business University, Beijing 102488, China.
Corresponding Author: Qiang Zeng, Email: 18178945645@163.com

Abstract: In object detection tasks, the sparse distribution, weak saliency and context-dependent nature of small objects pose three major challenges for perception systems. Although end-to-end detection architectures like the YOLO series have achieved a good balance between and speed accuracy in recent years, their utilization of shallow-layer features is low, resulting in performance bottlenecks in micro-object recognition. To address this, this paper proposes a small-object perception-enhanced detection framework, TFE-Net (Tiny-aware Feature Enhancement Network). By constructing a shallow high-resolution feature pathway and a multi-scale fine-grained semantic interaction module, it achieves a lightweight improvement of the YOLOv8s model structure. While maintaining the original model's computational complexity, this method significantly enhances the spatial perception and discrimination accuracy for extremely small objects. Experiments were conducted on the VisDrone dataset. Results show that the improved model boosts the mAP@0.5 from 0.386 to 0.421, with noticeable improvements in PR curves across all categories. This confirms the proposed strategy's ability to perceptually reconstruct and detect small objects in complex scenarios.
Keywords: Small object detection; Feature enhancement network; YOLOv8; Multi-scale fusion; Weak saliency awareness

1 INTRODUCTION

In today's rapidly advancing digital and intelligent era, small object detection (SOD), a key branch of computer vision, is growing in importance. It's widely used in intelligent transportation systems for real-time vehicle and pedestrian monitoring, in public safety for threat identification, in precision medicine for detecting cells and minor lesions, and in military surveillance for tracking distant small targets. Yet, SOD faces tough technical challenges. Small targets, occupying minimal pixel areas in images or videos, are hard to capture and recognize. Their lack of distinctiveness makes them hard to spot against complex backgrounds, causing mainstream detection models to suffer from severe perceptual degradation and fail to accurately locate and identify these targets. Moreover, the weak feature representation of small targets further increases detection difficulty.

Nowadays, the field of small object detection (SOD) faces numerous technical challenges, primarily caused by the following key factors. First, data sparsity is a significant issue. Limited samples of small objects and high annotation costs lead to class-imbalanced long-tailed distributions in training datasets, which restricts the effective learning ability of models. Second, during the process of multi-layer downsampling in convolutional neural networks, the fine-grained features unique to small objects are prone to being lost. This results in missing representations in high-level feature expressions, thereby affecting the final detection accuracy. Furthermore, the insufficiency of existing models in spatial context modeling makes it difficult to effectively capture the key discriminative information of small objects in complex backgrounds, thus weakening the ability to distinguish targets from backgrounds. These factors collectively result in "perception blind spots" in specific scales and regions, which seriously limits the performance and scope of application of SOD. The three core challenges in the SOD field are specifically as follows: Inadequate feature representation. Due to their limited size and sparse spatial distribution, the features of small objects gradually degenerate in deep networks, which impacts the extraction and recognition of detailed information; Perceptual scale imbalance. Traditional object detection frameworks focus on the detection performance of medium and large objects and lack dedicated path designs for small objects. This leads to unreasonable allocation of computing resources and attention mechanisms, causing the detection performance of small objects to be far below expectations; Missing contextual information. Accurate recognition of small objects is highly dependent on local contextual clues. However, most existing models lose a significant amount of spatial detail information during high-level feature abstraction. Due to the lack of effective context retention and fusion mechanisms, the detection difficulty is further increased, particularly in complex background scenarios. To address the above challenges, researchers have proposed a variety of innovative models and technical solutions aimed at improving the accuracy and reliability of SOD.

RetinaNet, a refined model based on Feature Pyramid Network (FPN), is widely used in small object detection. Its essence lies in constructing a top-down feature pyramid to enhance the perception of small objects by fusing multi-scale features. For instance, Tian et al. proposed a small object detection algorithm based on an improved RetinaNet model[1]. This algorithm addressed the low accuracy of traditional object detection algorithms when dealing with objects in horizontal and aerial images. Ahmad et al. introduced a detector integrated with RetinaNet to enhance low-level

semantic information and high-level spatial resolution[2], thereby effectively improving the superiority of small object detection in aerial images. Ahmad et al. adopted an anchor optimization method to improve the baseline framework's accuracy, thereby enhancing the detection of extremely small objects[3]. Nevertheless, RetinaNet has limitations, such as over-reliance on high-level features causing detail loss, and a linear feature fusion method lacking non-linear feature reshaping, which affect the detection of tiny objects.

DETR, based on the Transformer architecture, has pioneered a new paradigm for object detection. Its core is using self-attention mechanisms to model global spatial dependencies and enhance contextual information perception. For example, Dubey et al. proposed a normalized inductive bias for object detection with data fusion[4], improving DETR's accuracy in detecting small objects. Dai et al. replaced the cross-attention module with a dynamic attention module based on ROI[5], achieving faster convergence with fewer training epochs. Cao et al. introduced a new decoder layer to improve localization accuracy, especially for small objects[6]. However, DETR has limitations in handling local micro features. Specifically, the self-attention mechanism tends to focus on global features and overlook local details of small objects, leading to suboptimal performance in extremely small object detection tasks.

The YOLO series holds a significant position in object detection, with YOLOv5 and YOLOv8 being representative. YOLOv5 offers more efficient feature extraction and a flexible network structure, improving detection accuracy while maintaining real-time performance. YOLOv8 further introduces technologies like Decoupled Head and Task-Aligned Assigner to enhance training efficiency and detection accuracy. For example, Wang et al. introduced a query-based model with a new pipeline to address remote detection challenges in driving scenarios[7]. Sun et al. combined optical flow with background suppression images as auxiliary inputs[8], significantly improving the detection of infrared moving small objects. Shen et al. incorporated deformable convolution modules and a dynamic non-monotonic focusing mechanism into the backbone network[9], addressing object detection challenges in complex remote sensing image tasks. Despite the YOLO series' excellent real-time performance and detection accuracy, its lack of a dedicated perception path for shallow high-frequency information limits the model's spatial resolution for tiny objects, making it difficult to adapt to high-density small object distributions in complex scenes.

Currently, traditional perception models, constrained by the single paradigm of global representation of deep features, fail to leverage the descriptive advantages of shallow features for local fine-grained information, leading to significant capability gaps in small target perception. Most models also lack effective feature interaction mechanisms, making it hard to achieve fine perception and precise localization of small targets under controllable computational complexity, which further increases the difficulty of SOD. Against this backdrop, SOD research is showing significant paradigm evolution. On the one hand, the hierarchical semantic fusion paradigm is gaining attention. By enhancing shallow feature representation and introducing cross-scale feature interaction, it significantly improves the model's perception granularity of small targets. On the other hand, the cross-domain feature reconstruction direction is emerging. By leveraging the complementarity of multimodal feature spaces to fill representation gaps, it provides a new technical path to maximize perceptual integrity.

The SOD problem is essentially a representation deficiency caused by the imbalanced distribution of small-sample categories in high-dimensional feature spaces. Its core lies in constructing effective high-resolution descriptors within a limited feature perception domain. Starting from the dual perspectives of computational graph spatial modeling and information flow optimization, this paper proposes an architectural adjustment strategy with weakly-supervised feature enhancement capabilities to maximize information integrity in perceptual scenarios. In summary, this paper makes the following technical contributions to the field of SOD: (1) Proposes TFE-Net (Tiny-aware Feature Enhancement Network), an optimized network model based on the YOLOv8s framework. TFE-Net incorporates a dedicated small target perception branch, integrating shallow feature enhancement, multi-scale fine-grained perception fusion, and local non-linear feature reshaping strategies. This innovative architecture effectively expands the feature representation space while maintaining the model's original computational efficiency, significantly improving the model's ability to identify extremely small targets in complex scenes. By enhancing shallow features to capture local fine-grained information of small targets, fusing multi-scale features to enhance the model's perception of different-sized small targets, and reshaping local non-linear features to further improve feature representation, the model can more accurately identify extremely small targets that occupy minimal pixel areas in images. (2) In architectural design, TFE-Net balances practicality and computational efficiency. Through feature path reparameterization and perceptual redundancy compression, TFE-Net not only enhances SOD performance but also ensures efficient inference in practical deployment. Feature path reparameterization optimizes the feature extraction process for more efficient computational resource utilization, while perceptual redundancy compression reduces unnecessary computational overhead and speeds up inference. This balance enables TFE-Net to quickly and accurately process complex visual scenes in practical applications, meeting the demands of real-time applications.

2 METHOD

2.1 Overall Architecture Design

2.1.1 Core design philosophy

The core of TFE-Net's design is to boost feature perception and discrimination for tiny object detection. On the one hand, We introduce a Shallow Perception Enhancement Pathway (SPEP) to strengthen shallow feature extraction and highlight local details of small objects. On the other hand, A Cross-Scale Fine-Grained Aggregation Unit (CFAU) is

used to promote interaction and fusion of features at different scales, preserving crucial small object information. Also, Feature Flow Diversity Pathways are built to enrich feature propagation routes for better mining and utilization of small object features. Overall, TFE-Net inherits YOLOv8s' strong feature extraction ability and effectively alleviates feature degradation and insufficient representation of small objects. It offers a better feature foundation for tiny object detection and enhances detection performance in complex scenarios.

2.2 Core Module Design

TFE-Net has three key innovations: (1)Local Feature Perception Enhancement. TFE-Net strengthens expression of local fine-grained information by adding a small-object detection head and a dedicated shallow high-resolution feature pathway. The detection head is designed for small objects to capture their features precisely. The shallow high-resolution feature pathway preserves local details. This design improves the model's handling of local small-object features and boosts recognition accuracy. (2)Nonlinear Cross-Scale Feature Interaction. TFE-Net introduces a CFAU module for dynamic nonlinear interaction between features of different scales. Through nonlinear transformation and fusion, the CFAU module captures complex relationships between multi-scale features, enhancing context perception of small objects. (3) Feature Reconstruction and Alignment Optimization. TFE-Net uses feature space reparameterization via a selective feature enhancement module (SFEM) to compress redundant information. SFEM selectively enhances important features and suppresses redundant ones, optimizing representation for efficient object detection. This strategy also ensures feature consistency across levels and scales, improving overall model performance.

2.2.1 Shallow Perception Enhancement Pathway(SPSE)

The Shallow Perception Enhancement Pathway (SPEP) is a key component of TFE-Net for boosting small-object detection performance. Building on YOLOv8s' original three-layer detection heads, SPEP adds a new small-object detection head. This enables the model to better utilize the rich local detail information in shallow feature maps, which is crucial for small-object detection. Moreover, the Upsample-Concat-C2f module in SPEP plays a significant role in enhancing feature-map resolution. It strengthens spatial details through up-sampling and reconstructs high-dimensional features effectively via channel concatenation and convolution fusion. Which is,

$$F_{enhanced} = C2f(Concat(Upsample(F_{deep}), F_{shallow})) \quad (1)$$

Where, F_{deep} represents the deep feature map, $F_{shallow}$ represents the shallow feature map, and $C2f$ denotes the feature reconstruction module.

Specifically, in the neck network, deep-layer feature maps are first upsampled to achieve high-resolution feature maps. Then, the Concat module merges these upsampled features with early-stage features from the backbone network. Finally, the $C2f$ module performs multiple convolutions and skip connections to extract and fuse multi-level features, generating richer representations. This fusion strategy effectively combines multi-level feature information, enhancing feature expression. Consequently, the model's perception and detection accuracy for small objects are improved, as shown in Figure 1.

Figure 1 Shallow Perception Enhancement Pathway Diagram

Overall, the SPEP pathway effectively boosts the expressive density of the feature space. By enhancing local perception, it eases the issue of microscopic feature degradation and enables strong responsiveness reconstruction of small-object features.

2.2.2 Cross-Scale Fine-Grained Aggregation Unit(CFAU)

The Cross-Scale Fine-Grained Aggregation Unit (CFAU) is a key module in TFE-Net for boosting feature fusion and small-object detection. It uses multi-path nonlinear interaction strategies to achieve dynamic weighting and context enhancement of features at different scales. Specifically, CFAU has three core mechanisms:

(1) Multi-Scale Adaptive Fusion: CFAU dynamically adjusts the weights of features at different scales, allowing effective fusion based on their importance. This adaptive weighting strategy improves the flexibility and adaptability of feature fusion.

(2) Spatial-Aware Enhancement: CFAU emphasizes enhancing spatial information in local regions of feature maps. It uses spatial attention mechanisms to highlight areas of small objects and reduce background noise interference.

(3) Channel Recalibration: CFAU dynamically adjusts the weights of feature channels. It enhances important channels and suppresses unimportant ones, further optimizing feature representation and improving discriminability.

In summary, CFAU combines these three core mechanisms through multi-path nonlinear interactions. Which is,

$$F_{\mathrm{CFAU}} = \sum_{i=1}^{N} \alpha_i \cdot \mathcal{A}(F_i) \quad (2)$$

Where, $\mathcal{A}$ denotes the adaptive attention module, α_i signifies dynamic weights, and F_i represents feature maps of different scales.

CFAU enhances the model's local context perception and expands the spatial coverage and diversity of feature representations through non-linear cross-scale feature fusion.

2.2.3 Selective Feature Enhancement Module(SFEM)

The Selective Feature Enhancement Module (SFEM) in TFE-Net optimizes feature representation and boosts model performance. It uses a Convolution-BatchNorm-Activation (CBS) pattern with a SiLU activation function to enhance non-linear feature expression. Which is

$$F_{improve} = CBS(Concat(C2f(F_{deep}), F_{enhanced})) \tag{3}$$

Here, F_{deep} represents the deep feature map, $F_{enhanced}$ denotes the shallow feature map obtained from the Shallow Perception Enhancement Pathway (SPEP), and CBS represents the feature re-extraction module, which is:

$$CBS = SiLu(BatchNorm(Conv(F_i)) \tag{4}$$

Here, F_i represents the input feature maps of different scales to the module.

Specifically, the feature map first undergoes a convolution operation to extract higher-level features. Then, BatchNorm and the SiLU activation function are applied to enhance feature expression through the CBS module. Next, the processed feature map is concatenated with the one from the first part to integrate features from different sources. Finally, the $C2f$ module is used again to further fuse and enhance the concatenated features, extracting more robust and discriminative representations. As shown in Figure 2, SFEM's feature-channel selective enhancement mechanism dynamically adjusts channel weights, suppressing irrelevant features and boosting key-feature expression.

Figure 2 Feature Selective Enhancement Module Diagram

SFEM enables dynamic selection and enhancement of local features in the feature space. Through nonlinear mapping and channel recalibration, it improves the discriminative power and information density of feature distributions. This design boosts feature discrimination and enhances the model's detection accuracy and robustness for small objects.

2.3 Improved Architecture Overall Process

The improved architecture presented in this paper enhances YOLOv8s in multiple ways to boost tiny-object detection. The Backbone retains YOLOv8s' original structure for feature extraction. The Neck adds a Shallow Perception Enhancement Pathway (SPEP), which, with the Upsample-Concat-C2f module, introduces shallow high-resolution features, crucial for small-object detection. It also incorporates a Cross-Scale Fine-Grained Aggregation Unit (CFAU) for adaptive weighting and dynamic nonlinear interaction of multi-scale features, thereby strengthening contextual perception of small objects. The Head retains the original Anchor-Free decoupled head and adds a new small-object detection head to enhance the model's ability to detect tiny objects. Moreover, the architecture uses a Selective Feature Enhancement Module (SFEM) to boost feature distinctiveness and refine predictions. In summary, this architecture, as shown in Figure 3, optimizes every component to fully leverage multi-level features, significantly improving tiny-object detection performance.

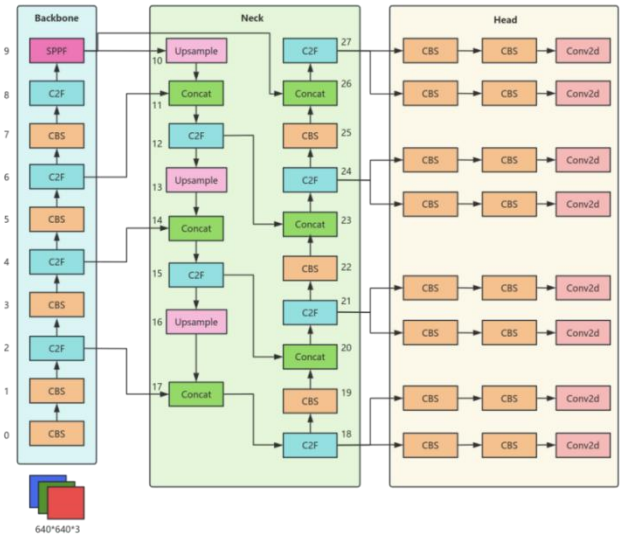

Figure 3 Improved Structure Diagram

3 EXPERIMENTAL DESIGN AND RESULT ANALYSIS

3.1 Experimental Environment and Settings

The experiments were conducted on a server with an NVIDIA RTX 4090 GPU, which excels at parallel computing and handles large-scale deep learning tasks efficiently. The software framework uses PyTorch for YOLOv8s and builds on the Ultralytics team's official implementation. Training hyperparameters are set as follows: SGD optimizer, initial learning rate (lr0) of 0.01, final learning rate (lrf) of 0.01, batch size of 32, image size of 640×640, and 300 training epochs. A linear annealing strategy was adopted to ensure stable model convergence and prevent loss of small-object features. The loss function combines YOLOv8's CIoU Loss, DFL, and Cross-Entropy Loss to enhance small-object localization accuracy.

3.2 Dataset Selection and Feature Analysis

3.2.1 Overview of the VisDrone dataset

The VisDrone dataset, collected by Tianjin University's Machine Learning and Data Mining Laboratory (AISKYEYE Team), is a benchmark for small object detection in low-altitude UAV scenarios[10]. It features complex real-world scenes with challenges like occlusion, multi-scale changes, complex backgrounds, and varying illumination, as shown in Figure 4. These characteristics make it a key benchmark for assessing and improving small object detection models. Additionally, its diversity in weather, lighting, and urban-rural backgrounds enables models trained on it to adapt better to various practical scenarios, thus enhancing their generalization and robustness.

Figure 4 Dataset Scenario Diagram

3.2.2 Categories and sample distribution

The VisDrone dataset comprises 10 common urban-scene object categories, such as pedestrian, car, bus, and bicycle. These categories, typical in urban environments, are applicable across various scenarios. The dataset exhibits a long-tail distribution, where some classes have many samples and others few. In VisDrone, common classes like vehicles and pedestrians have abundant samples, offering rich training data. This helps models learn their features well during training. But less common classes, such as special devices or specific animals, have few samples, causing the long-tail distribution. Due to this distribution, VisDrone is great for testing model robustness and generalization. It can effectively evaluate how well models handle class-imbalance issues.

3.2.3 Challenges

In the VisDrone dataset, objects are typically small, occupying less than 2% of the image on average, which is challenging for detection. The scenes often include multiple object categories that are densely packed, partially occluded, and frequently overlapping, further complicating recognition and differentiation. Additionally, the uneven background in the dataset significantly interferes with local feature perception, demanding greater robustness and adaptability from models during detection. These characteristics make the VisDrone dataset an ideal platform for evaluating object detection models.

3.3 Result Evaluation Metrics and Assessment Protocols

The key metrics for this experiment include mAP@0.5, mAP@0.5:0.95, Precision, and Recall. The mAP@0.5 measures average precision at an IoU threshold of 0.5, a critical indicator of object detection performance. mAP@0.5:0.95 evaluates overall model performance across multiple IoU thresholds. Precision assesses prediction accuracy, i.e., the proportion of correct predictions among all predicted objects. Recall measures the model's ability to recall objects, i.e., the proportion of correct predictions among all actual objects. These complementary metrics offer a comprehensive view of model performance.

Regarding evaluation protocols, all metrics are computed on the validation set to ensure reproducibility. Experimental settings are kept consistent, with only the addition of the small-object detection head being compared. This approach accurately reflects the performance contribution of TFE-Net.

3.4 Quantitative Result Analysis

The following is a quantitative analysis of the experimental results based on the aforementioned experimental setup.

Table 1 Basic Model Performance Table

Class	mAP@0.5	Class	MAP@0.5
pedestrian	0.408	truck	0.357
people	0.322	tricycle	0.267
bicycle	0.130	awning-tricycle	0.153
car	0.786	bus	0.553
van	0.446	motor	0.436
all-classes	0.386		

Table 1 shows the original model's varying performance across categories. Small objects like bicycles and tricycles have lower average precision, indicating challenges in detecting small targets due to their subtle features and smaller pixel presence in images. Additionally, the overall recall is low, suggesting frequent missed detections, especially for small objects. This implies the model's insufficient shallow-feature expression, as it fails to fully capture and utilize local details of small targets, affecting detection comprehensiveness.

Table 2 TFE-Net Model Performance Table

Class	mAP@0.5	Class	MAP@0.5
pedestrian	0.451↑ (+4.3%)	truck	0.394↑ (+3.7%)
people	0.355↑ (+3.3%)	tricycle	0.316↑ (+4.9%)
bicycle	0.158↑ (+2.8%)	awning-tricycle	0.168↑ (+1.5%)
car	0.807↑ (+5.1%)	bus	0.621↑ (+6.8%)
van	0.469↑ (+2.3%)	motor	0.471↑ (+3.5%)
all-classes	0.421 ↑ (+3.5%)		

Table 2 shows TFE-Net detection performance across categories, with significant improvements, especially for "hard-to-detect" objects like pedestrians and bicycles. Compared to the original model, the mAP@0.5 for pedestrian detection increased from 0.408 to 0.451, and for bicycle detection, it rose from 0.130 to 0.158. Overall, the mAP@0.5 improved by 3.5%, indicating enhanced accuracy in detecting small and confusing objects. The results demonstrate that TFE-Net enhances shallow perception paths, enabling better re-expression and re-localization of tiny targets. This constructs more layered and discriminative feature maps and highlights the model's robustness and reliability in small object detection.

3.5 Visualization and Phenomenon Attribution

3.5.1 Analysis of small target enhancement effects

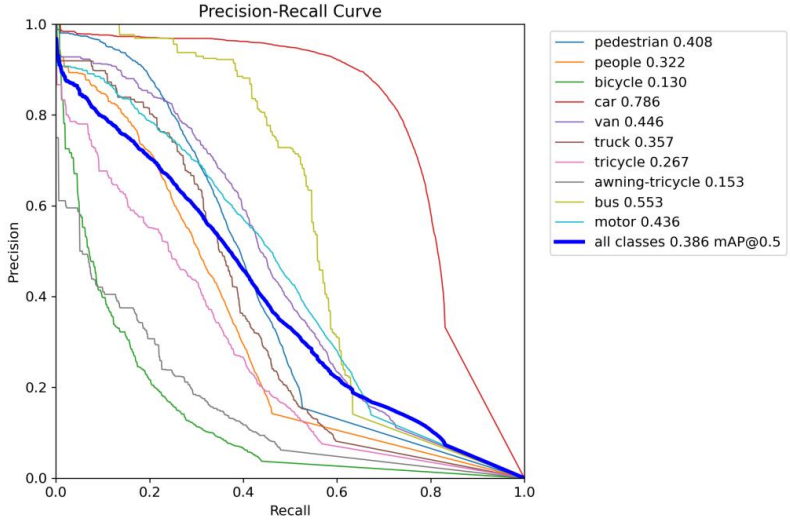

Figure 5 Basic Model PR Curve

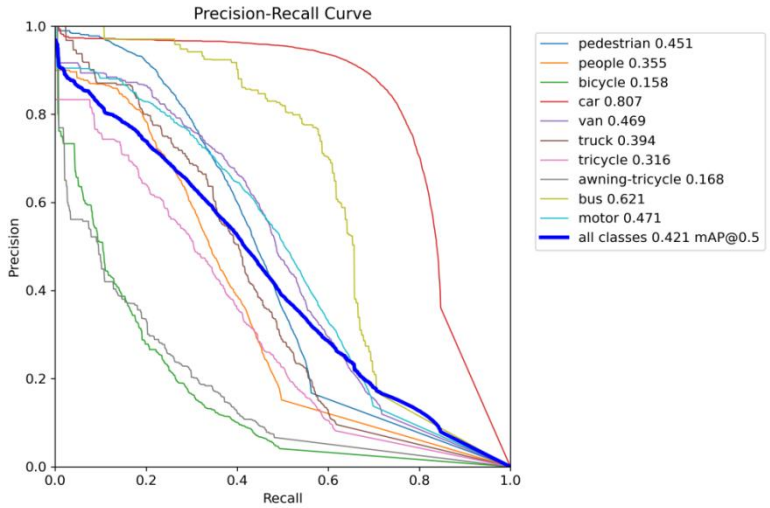

Figure 6 TFE-Net Model PR Curve

By comparing the PR curve trajectories in Figures 5 and 6, it can be observed that the latter achieves a Pareto frontier breakthrough in precision-recall co-optimization, indicating that the improved model has higher precision. For the pedestrian category, the PR curve is more stable in the high recall rate region, indicating more accurate detection of pedestrians. The curve for the bicycle category also shows a significant upward shift, indicating a more substantial enhancement of the model's ability to detect small objects. In terms of specific data, the TFE-Net model shows varying degrees of improvement in the mAP@0.5 values across all categories. For the pedestrian category, it improves from 0.408 to 0.451, and for the bicycle category, it increases from 0.130 to 0.158. This suggests a significant improvement in the model's performance when dealing with small and easily confusable objects. Moreover, the overall mAP@0.5 across all categories improves from 0.386 to 0.421, indicating an enhancement in the model's overall detection performance. Additionally, the regions where the original model suffered from severe missed detections are effectively captured by the improved model. This demonstrates that the improved model has enhanced its ability to express shallow features and can better capture the local details of small objects, thereby reducing missed detections. This is also consistent with the improvement in the PR curve, indicating a significant improvement in the model's ability to detect small objects in complex scenes.

3.5.2 Attribution analysis of improvement effects

The improved model in this study has achieved remarkable performance gains in small object detection, and the main reasons can be attributed to the implementation of the following key strategies. Firstly, a specially designed small object detection head (i.e., a shallow pathway) has been introduced, which effectively compensates for the feature compression and detail loss of small objects caused by the traditional downsampling process. This detection head directly utilizes high-resolution feature maps from shallow layers for processing, thereby significantly reducing feature loss and substantially enhancing the model's ability to represent the features of small objects. Secondly, the feature fusion structure of the model has been optimized, which significantly improves the cross-scale semantic consistency. The optimized feature fusion mechanism can more efficiently integrate feature information from different scales, enabling the model to perform more accurately when detecting objects of varying scales. Finally, high-resolution feature maps have been made to participate directly in the detection process, which has greatly improved the model's ability to recall tiny objects. High-resolution feature maps contain richer local detail information, which helps the model to more accurately locate and identify small objects, thereby effectively reducing the occurrence of missed detections. These design improvements work together in synergy to collectively drive the significant enhancement of the TFE-Net model's detection performance in small object detection tasks.

4 CONCLUSION

This paper presents TFE-Net (Tiny-aware Feature Enhancement Network), a lightweight framework for enhancing shallow perception in YOLOv8s to address performance degradation in small object detection. By integrating shallow high-resolution feature pathways and local perception paths, TFE-Net significantly improves detection of tiny objects. Its architecture focuses on minor structural modifications to build a shallow perception framework with high feature sensitivity and discrimination, offering an optimized, cost-effective solution for small object detection. From a theoretical perspective, TFE-Net's design emphasizes weak feature reconstruction and diverse feature paths, providing a new framework for identifying small objects in complex settings. In practical terms, it efficiently enhances YOLOv8's architecture, ensuring compatibility with mainstream inference frameworks and showcasing strong transferability and deployability. This makes it suitable for various scenarios like traffic monitoring, UAV security, and industrial defect detection.

However, TFE-Net has certain limitations. It hasn't optimized loss function reconstruction or dynamic label allocation, and its local robustness can be further enhanced. It also mainly depends on RGB images and lacks multimodal data

fusion capabilities. Future research will focus on several key areas. We will integrate infrared, depth, and radar data for multimodal enhancement to enrich semantic representation. Unsupervised and self-supervised learning methods will be explored to reduce reliance on labeled data and improve adaptability in data-sparse areas. Architecture optimization via Neural Architecture Search (NAS) will be conducted to achieve high-performance, low-power detection. We will also apply end-to-end DETR structures to small object tasks, promoting the shift from "Anchor to Attention." Additionally, cross-scale dynamic perception and local refinement will be emphasized. By utilizing lightweight convolutional modules and high-resolution feature pathways, we can enhance shallow features and boost local fine-grained feature expression for better small object detail capture. Efficient perception path designs, such as incorporating GhostNet and MobileNet modules along with feature sparsity compression, will continue to be developed. These approaches balance efficiency and effectiveness by reducing computational complexity while maintaining perception performance. Overall, small object detection is moving towards multi-layer, cross-scale, non-linear feature reconstruction. This evolution offers new breakthroughs for micro-perception in complex environments. Feature space density reconstruction and perception path diversity expansion are becoming crucial in SOD research. These improvements will enhance the model's robustness, adaptability, and efficiency, unlocking more potential in complex scenarios. We believe TFE-Net will play a key role in advancing the field of small object detection.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Tian H, Zheng Y, Jin Z. Improved RetinaNet model for the application of small target detection in the aerial images//IOP Conference Series: Earth and Environmental Science. IOP Publishing, 2020, 585(1): 012142.

[2] Ahmed M, Wang Y, Maher A, et al. Fused RetinaNet for small target detection in aerial images. International Journal of Remote Sensing, 2022, 43(8): 2813-2836.

[3] Ahmad M, Abdullah M, Han D. Small object detection in aerial imagery using RetinaNet with anchor optimization//2020 International conference on electronics, information, and communication (ICEIC). IEEE, 2020: 1-3.

[4] Dubey S, Olimov F, Rafique M A, et al. Improving small objects detection using transformer. Journal of Visual Communication and Image Representation, 2022, 89: 103620.

[5] Dai X, Chen Y, Yang J, et al. Dynamic detr: End-to-end object detection with dynamic attention//Proceedings of the IEEE/CVF international conference on computer vision. IEEE, 2021, 10: 2988-2997.

[6] Cao X, Yuan P, Feng B, et al. Cf-detr: Coarse-to-fine transformers for end-to-end object detection//Proceedings of the AAAI conference on artificial intelligence. AAAI, 2022, 36(1): 185-193.

[7] Wang H, Liu C, Cai Y, et al. YOLOv8-QSD: An improved small object detection algorithm for autonomous vehicles based on YOLOv8. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 1-1.

[8] Sun S, Mo B, Xu J, et al. Multi-YOLOv8: An infrared moving small object detection model based on YOLOv8 for air vehicle. Neurocomputing, 2024, 588: 127685.

[9] Shen L, Lang B, Song Z. DS-YOLOv8-based object detection method for remote sensing images. IEEE Access, 2023, 11: 125122-125137.

[10] Du D, Zhu P, Wen L, et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results//Proceedings of the IEEE/CVF international conference on computer vision workshops. IEEE, 2019, 1: 0-0.