Adaptive Anchor Clustering and Lightweight Bidirectional Fusion for UAV-Based Fire Detection

Xuepeng Huang, Zhiqiang Liu*, Xu Zhang, Wenjing Li
[1,3]Inner Mongolia University of Technology, Hohhot, China
[2]Inner Mongolia Technical University of Construction, Hohhot, China
Email: 18088630924@163.com, liuzq@imut.edu.cn

***Abstract*—Early fire detection is critical for reducing casualties and property losses; however, existing vision-based methods often exhibit limited robustness in complex scenarios involving small targets and heterogeneous data distributions. To address these challenges, this paper proposes an adaptive fire detection framework based on structural optimization of YOLOv11, which enhances scale representation and multi-scale feature fusion. Specifically, a K-means–based clustering strategy is introduced to better capture the scale distribution of flame and smoke targets, providing prior information to improve the matching process, while a lightweight bidirectional feature pyramid network (BiFPN) incorporating depthwise separable convolutions is employed to strengthen fine-grained feature representation for small and sparse fire targets with limited computational overhead. Extensive experiments conducted on a multi-source fire dataset demonstrate that the proposed method achieves 78.443% mAP_{50}, 56.327% $mAP_{50–95}$, and 70.019% Recall, outperforming the baseline YOLOv11 by 4.131%, 3.646%, and 5.402%, respectively. Further ablation and cross-dataset generalization experiments verify the effectiveness and complementarity of the proposed components, particularly in small-target and multi-scale scenarios. These results indicate that the proposed approach provides a robust and efficient solution for real-world fire detection applications.**

***Index Terms*—Fire–smoke recognition,UAV,Lightweight architecture,Small object detection**

I. Introduction

Fire disasters are sudden and highly destructive public safety incidents, and their early detection is critical for reducing casualties and economic losses. Timely identification of smoke and flame sources at the early stage can effectively suppress fire spread. However, visual fire detection remains challenging due to limited training samples, complex backgrounds, illumination variations, and airflow disturbances, which impose stringent requirements on detection accuracy and robustness [1].

Conventional fire detection methods mainly rely on physical sensors, such as smoke and temperature detectors, which typically respond only after threshold exceedance, resulting in delayed alarms and high false-positive rates in complex environments. Manual inspection is similarly inadequate for large-scale and diverse scenarios [2]. Although deep learning–based object detection has achieved notable progress, UAV-based fire detection still suffers from several limitations: small and highly diverse datasets that hinder generalization [3], poor performance on small-scale flames and sparse smoke targets, and difficulties in balancing detection accuracy with real-time efficiency.

To address these issues, this study proposes an improved fire detection framework based on YOLOv11. The main contributions are summarized as follows:

- An adaptive K-means–based anchor generation strategy is introduced to better match the scale characteristics of fire and smoke targets.
- A P2-enhanced bidirectional feature pyramid network is designed to improve shallow feature propagation and enhance small-target detection.
- Comprehensive experiments are conducted on multiple datasets with diverse scene characteristics to evaluate accuracy and generalization performance.

Overall, the proposed method aims to improve detection accuracy and reduce miss rates in fire monitoring tasks through targeted architectural and training strategies, providing a more robust and practical solution for UAV-based fire detection.

II. Related Work

A. YOLOv11 Architecture

YOLOv11 [4], released by Ultralytics in September 2024, represents the latest generation of real-time object detectors, achieving an improved balance between detection accuracy and computational efficiency. In this work, an optimized YOLOv11 architecture is adopted, consisting of three main components: backbone, neck, and head, which respectively perform feature extraction, multi-scale feature fusion, and object prediction [5].

In the backbone, the conventional C2f module is replaced with the C3K2 module, which employs a dual-branch convolutional design to enhance feature representation while reducing parameter count and computational cost. The backbone also utilizes the CBS block, composed of Convolution, Batch Normalization, and SiLU activation, which is widely used in YOLO-based networks to stabilize training and improve early-stage feature extraction.

In the neck, a Bidirectional Feature Pyramid Network (BiFPN) is introduced to enable efficient multi-scale feature fusion. By establishing both top-down and bottom-up information paths with learnable fusion weights, BiFPN enhances small-target perception while maintaining high inference speed. In the head, YOLOv11 employs a lightweight

C2PSA module to explicitly model spatial positional dependencies, improving localization accuracy and boundary discrimination in complex scenes [6]. The overall architecture is illustrated in Fig. 1.

B. K-means Clustering

K-means clustering [8] is an unsupervised partitioning algorithm that groups data into K clusters by minimizing the sum of squared errors (SSE), defined as:

$$\min_{C_k k=1^K} \sum k = 1^K \sum_{\mathbf{x}_i \in C_k} |\mathbf{x}_i - \boldsymbol{\mu}_k|_2^2 . \tag{1}$$

Here, C_k and $\boldsymbol{\mu}_k$ denote the k-th cluster and its centroid, respectively. The algorithm iteratively alternates between cluster assignment and centroid update until convergence.

In object detection, each ground-truth bounding box can be represented as $\mathbf{b}_i = (w_i, h_i)$. Applying K-means to bounding box dimensions allows anchor boxes to better match the scale distribution of training data. However, flame and smoke targets often exhibit irregular shapes and large scale variations, leading to anchor–target mismatches when generic anchor settings are used. To address this issue, K-means clustering is employed to adaptively generate anchor boxes tailored to fire-related targets, thereby improving recall, particularly for small or weak objects.

C. Bidirectional Feature Pyramid Network(BiFPN)

Low-level features preserve fine spatial details, whereas high-level features encode richer semantic information [9]. To achieve more effective multi-scale fusion, a Bidirectional Feature Pyramid Network (BiFPN) [10] is integrated into the YOLOv11 neck. Unlike conventional FPNs with a single top-down pathway, BiFPN introduces bidirectional connections, enabling simultaneous top-down semantic propagation and bottom-up spatial enhancement, as shown in Fig. 2.

Let P_2, P_3, P_4, P_5 denote feature maps from different backbone stages. The bidirectional fusion process is formulated as:

$$\hat{P}l^{\uparrow} = \mathcal{F}\left(P_l, \mathrm{Up}(\hat{P}l+1)\right), \tag{2}$$

$$\hat{P}_l^{\downarrow} = \mathcal{F}\left(\hat{P}l^{\uparrow}, \mathrm{Down}(\hat{P}l-1^{\uparrow})\right). \tag{3}$$

To further enhance fusion flexibility, BiFPN adopts a weighted feature fusion strategy:

$$P_{\text{out}} = \mathrm{Conv}\left(\frac{\sum_i w_i P_i}{\sum_i w_i + \epsilon}\right), \tag{4}$$

where w_i are learnable non-negative weights normalized as:

$$w_i = \frac{\mathrm{ReLU}(\alpha_i)}{\sum_j \mathrm{ReLU}(\alpha_j) + \epsilon}. \tag{5}$$

By enabling adaptive bidirectional information flow, BiFPN significantly improves the detection of small, weak, and irregular fire targets in complex environments.

III. Methodology

A. UAV Fire Dataset Description

The UAV fire dataset used in this study consists of publicly available aerial fire images collected from multiple open-access sources. All images were captured by unmanned aerial vehicles at different altitudes and viewing angles, covering various scenarios such as forest fires, grassland fires, and wildland–urban interface areas.

Compared with ground-based datasets, UAV images generally exhibit wider fields of view and clearer target–background separation, with fire targets predominantly appearing at medium to large scales, while small and weak targets occur less frequently. All images were manually annotated with axis-aligned bounding boxes following a unified labeling protocol, and annotation quality was carefully verified.

Notably, this dataset is not used for training but serves as an independent benchmark to evaluate the cross-domain generalization capability of the proposed model under aerial observation conditions.

B. Fire-Specific Anchor Box Generation

Anchor box design plays a critical role in object detection performance, as mismatched anchors often lead to missed detections, especially for small targets [11]. To better align anchor configurations with the scale distribution of flame and smoke targets, a K-means–based clustering strategy is employed.

To improve clustering stability and convergence quality, an optimized initialization strategy is adopted, and the distance metric is redefined using Intersection over Union (IoU) rather than Euclidean distance:

$$D(\text{box, centroid}) = 1 - IoU(\text{box, centroid}). \tag{6}$$

The clustering objective is formulated as:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} D(x, \mu_i). \tag{7}$$

Clustering is performed on the width–height distributions of all ground-truth bounding boxes in the training set. In this study, the number of anchors is set to $k = 9$, following common practices in anchor-based object detection frameworks, which provides a reasonable balance between representation capacity and model complexity. Compared with preset anchors, the clustered anchors improve the alignment between anchor boxes and ground-truth targets, leading to better matching quality during training and potentially increasing the number of effective positive samples. This contributes to improved detection performance, particularly for small fire targets.

C. Integration of BiFPN

In fire monitoring, early-stage flames and sparse smoke often manifest as high-frequency, dynamically changing boundaries in low-resolution images. Consequently, the design of the neck structure plays a pivotal role in determining the overall

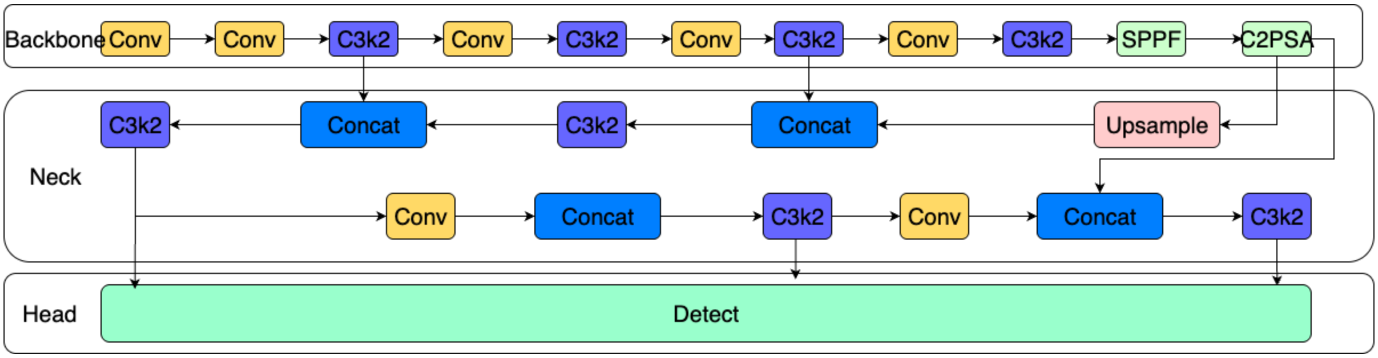

Fig. 1. Typical pipeline of fire and smoke detection based on computer vision and deep learning, including image acquisition, feature extraction, and decision-making stages. The illustration is adapted from representative fire detection frameworks reported in previous studies [7].

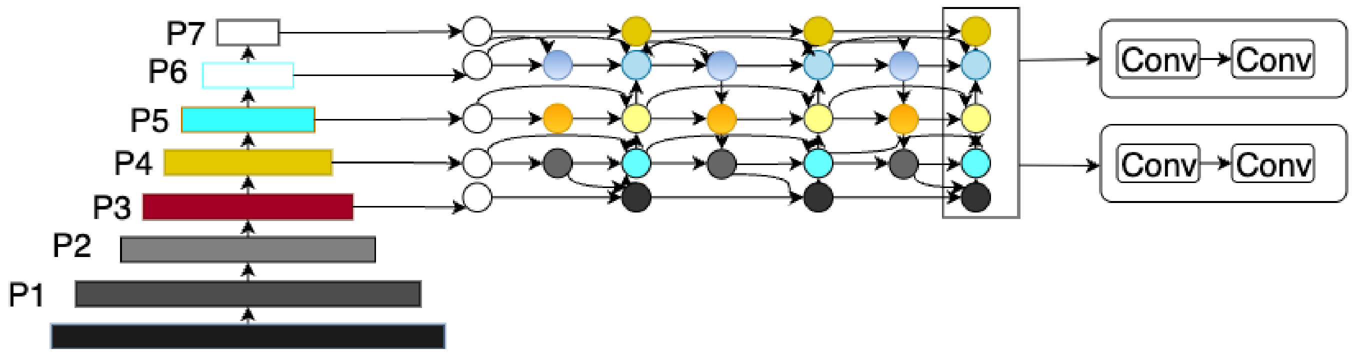

Fig. 2. Overall architecture of the BiFPN network.

detection performance. In this study, the original unidirectional Feature Pyramid Network (FPN) in YOLOv11 is replaced with a Bidirectional Feature Pyramid Network (BiFPN) to further enhance multi-scale feature aggregation.

The BiFPN establishes a closed-loop information flow through top-down and bottom-up pathways, allowing high-level semantic information to be propagated downward while low-level detailed features are simultaneously fed back upward. This bidirectional interaction significantly strengthens cross-scale feature fusion and ensures more comprehensive feature representation. Compared with the conventional element-wise summation used in FPN, BiFPN introduces a learnable weighted fusion mechanism, where adaptive weight parameters dynamically adjust the contribution of each feature level during the fusion process. The fusion can be mathematically expressed as:

$$P_{\text{out}} = \text{Conv}\left(\frac{\sum_i w_i \cdot P_i}{\sum_i w_i + \epsilon}\right). \tag{8}$$

Here,P_i denotes the input features from different levels, w_i represents the learnable fusion weights, and ϵ is a small constant to ensure numerical stability. This mechanism improves the flexibility and accuracy of cross-level feature fusion, thereby enhancing the perception of small fire targets.

To further improve small-object detection performance, the high-resolution P2 feature (160×160) from the backbone is incorporated into the BiFPN, allowing the network to better preserve boundary details of small flames. Meanwhile, to reduce computational cost, all convolution operations in BiFPN are replaced with Depthwise Separable Convolutions, which significantly decrease the number of parameters and FLOPs, thus improving computational efficiency while maintaining detection accuracy. The architecture of the improved model is illustrated in Fig. 3.

Experimental results demonstrate that the enhanced BiFPN module significantly improves the detection of small flames and smoke, achieving an increase of approximately 4.1% in mAP_{50} and 5.4% in Recall, while maintaining inference speed. These findings indicate that the proposed design effectively balances accuracy and efficiency in complex fire detection scenarios, providing strong support for early fire recognition.

The introduction of P2-level features is primarily motivated by the need to preserve fine-grained spatial information that is often lost in deeper layers. Such information is critical for accurately localizing small flames and sparse smoke regions, which are common in early-stage fire scenarios. By incorporating P2 into the BiFPN structure, the proposed network enhances multi-scale feature interaction while maintaining sensitivity to small-scale targets.

From an architectural perspective, shallow features typically contain richer spatial details, which are particularly critical for

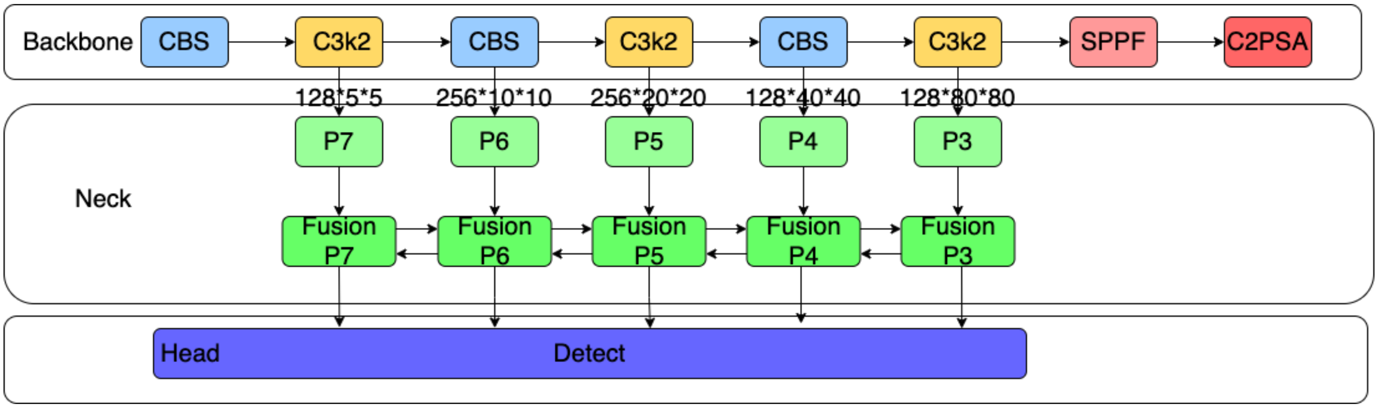

Fig. 3. Illustration of the improved model architecture.

capturing the edges and textures of tiny flames and sparse smoke. In this study, high-resolution P2 features from the backbone are incorporated into the BiFPN fusion pathway, enabling fine-grained information to be effectively preserved and reinforced during multi-scale feature integration. Meanwhile, the bidirectional information flow mechanism of BiFPN enhances semantic representation while mitigating feature degradation, thereby improving the discriminability of small targets in complex backgrounds.The experimental results demonstrate more pronounced performance gains on datasets with a higher proportion of small targets, further validating the effectiveness of the proposed design for recognizing weak and early-stage fire signals.

D. UAV Scenario-Specific Optimization

UAV-based fire detection is often affected by motion blur, occlusion, and illumination variations [12]. To enhance robustness, horizontal flipping (probability 0.2) and label smoothing (factor 0.05) are applied during training to improve generalization.

IV. Experiments

A. Experimental Setup

Due to the absence of a fully open and authoritative fire recognition dataset, this study constructs a large-scale multi-scene fire dataset by integrating images from ImageNet, VisFire, and BowFire. The dataset covers diverse fire scenarios, including forest, indoor, vehicle, industrial, and UAV-based aerial fires, providing comprehensive scene diversity for training and evaluation. The dataset statistics are summarized in Table I.

To ensure fair comparison, all models were trained under a unified setting. Training was conducted for 300 epochs with a fixed input resolution of 640×640. The Adam optimizer was used with an initial learning rate of 3×10^{-5} and gradual decay. Data augmentation included horizontal flipping (probability 0.2) and label smoothing (factor 0.05). The batch size was set to 48, and mixed-precision training with RAM caching was enabled. Model performance was evaluated using mAP_{50} and Recall. All experiments were implemented in PyTorch on an NVIDIA RTX 3090 GPU (CUDA 11.7).

TABLE I
Dataset Classification

Type	Number
Fire image	22407
Smoke image	17073
Fire bounding box	36463
Smoke bounding box	20975
Training set	16960
Validation set	5650
Test set	5650

B. Evaluation Metrics

The proposed improved model was evaluated on the test set through a series of experiments to verify its performance. The evaluation metrics used in this study include Precision, Recall, F1-score, and Average Precision (AP) [13].

For binary classification tasks [14], each prediction can be categorized into one of four outcomes based on the relationship between the ground-truth label and the predicted result: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). These relationships are summarized in the confusion matrix presented in Table II.

TABLE II
Confusion Matrix

Actual Values	Predicted	Confusion Matrix
Positive	Positive	TP
Positive	Negative	FN
Negative	Positive	FP
Negative	Negative	TN

Standard evaluation metrics, including Precision, Recall, and mAP, are adopted to evaluate the model performance.

C. Comparison of results

1) Comparison Between Traditional K-means and Improved K-means: To evaluate the effectiveness of the proposed IoU-based K-means clustering strategy, a comparison was conducted with traditional K-means under identical training settings. As shown in Table III, the improved method achieves consistent gains in mAP_{50}, $mAP_{50–95}$, and Recall, demonstrating that IoU-based distance effectively reduces anchor–target mismatch, particularly for small and elongated fire targets.

Notably, the number of parameters and FLOPs remains unchanged, since anchor clustering is performed offline and does not affect the network architecture or inference process.

2) Effect of BiFPN: To further validate the effectiveness of the proposed method, comparative experiments were conducted on representative baseline and detection models under a unified training and evaluation protocol. All models were trained on the combined fire dataset and evaluated using consistent metrics, including mAP_{50}, $mAP_{50–95}$, and Recall.

The results demonstrate that the proposed method consistently outperforms the baseline YOLOv11, particularly in scenarios involving small-scale flames and sparse smoke under complex backgrounds. Compared with existing approaches, the proposed framework shows improved sensitivity to weak fire targets while maintaining comparable computational complexity. The comparison results are presented in Table IV.

In addition to the improvement under the IoU = 0.5 criterion, the proposed model also achieves a higher $mAP_{50–95}$ of 56.327%, representing a 3.646 percentage point increase over the baseline. This metric evaluates performance across multiple IoU thresholds (0.5–0.95), reflecting stronger localization capability under stricter conditions. The improvement indicates that the proposed feature fusion strategy enhances detection robustness across different object scales and boundary ambiguities.

On the fire dataset, the baseline YOLOv11 achieves an mAP_{50} of 74.312% and a Recall of 64.617%. With the proposed enhancements, the improved model reaches 78.443% in mAP_{50} and 70.019% in Recall, corresponding to gains of 4.131 and 5.402 percentage points, respectively. These improvements demonstrate enhanced detection performance, particularly in reducing missed detections of small and weak targets. The performance gain is achieved with a marginal increase of 2.9% in parameters and 1.3% in FLOPs.

D. Ablation Study

To analyze the contribution of each module, ablation experiments were conducted using YOLOv11 as the baseline. As shown in Table V, introducing K-means clustering and BiFPN individually improves performance, while their joint integration achieves the best results.

Specifically, adaptive anchor generation mainly improves anchor matching with minimal computational overhead, whereas BiFPN yields larger gains in Recall and $mAP_{50–95}$, indicating improved robustness under strict localization requirements. The combined configuration achieves optimal performance with only a slight increase in parameters and FLOPs, confirming the complementary effects of the two modules.

E. Generalization Experiment

To evaluate cross-dataset generalization, the model trained on the joint dataset was tested on VisFire, BowFire, and UAV datasets. As shown in Table VI, the improved model consistently outperforms the baseline across all datasets, demonstrating strong transferability despite differences in viewpoint, background complexity, and target scale distribution.

Performance gains are more pronounced on VisFire and BowFire, particularly in Recall, further indicating that the proposed adaptive anchor configuration and bidirectional feature fusion are especially effective for small-scale and weak fire targets. Overall, these results confirm the robustness and cross-scene applicability of the proposed method.

V. Conclusion and Future Work

This paper presents an enhanced YOLOv11-based framework for early-stage fire detection by improving scale representation and feature fusion. A K-means-based clustering strategy is employed to provide scale priors, and a P2-enhanced bidirectional feature pyramid network (BiFPN) is introduced to strengthen multi-scale feature aggregation, particularly for small and low-saliency targets.

Experimental results demonstrate that the proposed method consistently outperforms the baseline YOLOv11 in terms of mAP and Recall, with more significant improvements observed in scenarios dominated by small-scale flames and sparse smoke under complex backgrounds. These results validate the effectiveness of incorporating high-resolution features and bidirectional fusion for improving small-target detection performance.

Despite these improvements, the current framework relies solely on single-frame RGB inputs and does not explicitly model temporal dynamics, which may limit its robustness in challenging conditions such as low illumination or severe occlusion.

Future work will explore temporal modeling approaches, such as ConvLSTM and Transformer-based methods, as well as multi-modal data integration, to further enhance detection robustness and adaptability in real-world fire monitoring scenarios.

Acknowledgment

This work was supported in part by the Natural Science Foundation of Inner Mongolia Autonomous Region (Grant Nos. 2025MS06003, 2025MS06056), in part by the Science and Technology Program of Inner Mongolia Autonomous Region (Grant No. 2021GG0250), in part by the National Natural Science Foundation of China (Grant No. 61962044), and in part by the Fundamental Research Funds for the Universities Directly Under the Autonomous Region (Grant No. JY20220324)

TABLE III
COMPARISON OF BASELINE AND IMPROVED MODELS

Method	mAP_{50}	$mAP_{50–95}$	Recall	Params (M)	FLOPs (G)
Traditional K-means	74.312%	52.681%	64.617%	23.8	122.5
Improved K-means (IoU-based)	74.834%	53.105%	65.308%	23.8	122.5

TABLE IV
COMPARISON OF BASELINE AND IMPROVED MODELS

Model	mAP_{50}	$mAP_{50–95}$	Recall	Params (M)	FLOPs (G)
Base	74.312%	52.681%	64.617%	23.8	122.5
Improved	78.443%	56.327%	70.019%	24.5	124.1

TABLE V
RESULTS OF THE ABLATION EXPERIMENTS

Version	mAP_{50}	Recall	$mAP_{50–95}$	Params (M)	FLOPs (G)
Base	74.312%	64.617%	52.681%	23.8	122.5
+ K-means	74.834%	65.308%	53.105%	23.9	122.8
+ BiFPN	76.219%	67.125%	54.327%	24.3	123.9
+ K-means + BiFPN	78.443%	70.019%	56.327%	24.5	124.1

TABLE VI
RESULTS OF THE GENERALIZATION EXPERIMENT

Dataset	Model	mAP_{50}	$mAP_{50–95}$	Recall	F1-score
Self-built UAV	YOLOv11	77.563%	55.842%	67.231%	0.715
	Improved	79.215%	57.103%	68.012%	0.732
VisFire	YOLOv11	75.342%	53.214%	65.783%	0.695
	Improved	77.891%	55.896%	69.234%	0.726
Bow Fire	YOLOv11	74.678%	52.487%	64.321%	0.681
	Improved	77.345%	54.963%	67.890%	0.716

REFERENCES

[1] Y. Dong, Z. Li, X. Li, and X. Li, “Using ontology and rules to retrieve the semantics of disaster remote sensing data,” Systems Engineering and Electronics (English Edition), vol. 35, no. 5, pp. 1211–1218, 2024.

[2] C. Jin, A. Zheng, Z. Wu, and C. Tong, “Real-time fire smoke detection method combining a self-attention mechanism and radial multi-scale feature connection,” Sensors, vol. 23, no. 6, 2023.

[3] Q. Zhu, Z. Xu, and L. Hu, “Semi-supervised image classification based on deep clustering and pre-trained models,” Proc. 7th Int. Conf. Computer Information Science and Artificial Intelligence, pp. 281–285, 2024.

[4] Z. He, K. Wang, T. Fang, L. Su, R. Chen, and X. Fei, “Comprehensive performance evaluation of YOLOv11, YOLOv10, YOLOv9, YOLOv8 and YOLOv5 on object detection of power equipment,” 2024.

[5] B. Jin, Z. Lu, L. He, and P. Yu, “FT-YOLO: A UAV-oriented small-object detection algorithm based on feature enhancement and task synchronization,” The Journal of Supercomputing, vol. 81, no. 16, 2025.

[6] T. Ghosh, A. Gangopadhyay, and L. Das, “Leveraging deep learning models for structural crack detection,” Innovative Infrastructure Solutions, vol. 10, no. 9, 2025.

[7] S. Zhuo, H. Bai, L. Jiang, X. Zhou, X. Duan, Y. Ma, and Z. Zhou, “SCL-YOLOv11: A lightweight object detection network for low-illumination environments,” IEEE Access, vol. 13, pp. 47653–47662, 2025.

[8] V. K. Dehariya, S. K. Shrivastava, and R. C. Jain, “Clustering of image data set using K-means and fuzzy K-means algorithms,” 2011.

[9] S. Xiong, W. Zuo, Z. Liu, J. Xin, and N. Zheng, “Dark object detection based on detail feature enhancement and cross-aligned attention,” Proc. 8th Int. Conf. Electrical, Mechanical and Computer Engineering (ICEMCE), pp. 2050–2056, 2024.

[10] N. S. Syazwany, J. H. Nam, and S. C. Lee, “MM-BiFPN: Multi-modality fusion network with Bi-FPN for MRI brain tumor segmentation,” IEEE Access, 2021.

[11] A. S. Geetha, “YOLOv4: A breakthrough in real-time object detection,” 2025.

[12] Y. Chang, Z. Wang, L. Qiu, S. Wang, and Y. Bai, “Reinforcement learning–based adaptive fault-tolerant antidisturbance control for UAVs subjected to external disturbances, input uncertainties, and structural uncertainties,” Journal of Aerospace Engineering, vol. 38, no. 1, 2025.

[13] J. Liu, S. Zhang, R. Ma, J. Sha, W. Kang, and M. Lin, “A method for analyzing and predicting the influencing factors of carbon emission intensity based on fuzzy comprehensive evaluation algorithm and data mining,” IOP Publishing Ltd, 2024.

[14] A. Goddard, K. Du, and Y. Xiang, “Mining invariance from nonlinear multi-environment data: Binary classification,” 2024.