An Enhanced Fire Perception Framework for Firefighting Robots: ECA-BiFPN Boosted YOLO-EB with Multi-Modal Fusion

Botao Ni, Lei Huang, Ying Xiang, Yan Zhu, Lin Li, Yunfei Zhou, Jingjing Yang, and Hao Tang

***Abstract*— To address the challenges posed by complex fire environments to flame perception and situational assessment in firefighting robots, this paper proposes a multimodal perception method integrating flame segmentation, spatial localization, and situational awareness. Based on an enhanced YOLO-EB segmentation network, this approach combines stereo vision and spectral information to achieve precise flame detection and localization. The YOLO-EB network incorporates an Efficient Channel Attention (ECA) mechanism and a Bidirectional Feature Pyramid Network (BiFPN) to enhance its representational capabilities, effectively balancing accuracy and real-time performance. Ablation experiments demonstrate that the synergistic effect of these modules significantly improves model performance: compared to the YOLOv8-seg baseline, the proposed model achieves a flame recall of 0.680 and a mean average precision (mAP) of 0.821. Leveraging high-quality segmentation masks, a stereo-spectral fusion perception framework is constructed, achieving precise flame localization at medium distances with an average error of 2–3%. Through dynamic fusion of spectral and visible-light features, the system attains an average situational awareness accuracy of 92.45%, significantly outperforming single-modal methods. Experimental results confirm that this integrated approach provides firefighting robots with stable and reliable capabilities for early fire detection, accurate localization, and dynamic assessment, demonstrating strong potential for practical deployment.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/10447

***Index Terms*—Firefighting robots, Flame segmentation, YOLO-EB, Binocular vision, Spectral fusion, Multi-modal perception**

The associate editor coordinating the review of this manuscript and approving it for publication was Luis Camarinha-Matos (*Corresponding author: Ying Xiang*).

This work was supported by the Joint Fund for Regional Innovation and De velopment of National Natural Science Foundation China (No. U24A2078), 20 24 Guangdong Provincial University Characteristic and Innovative Projects, (No. 2024KTSCX379).

B. Ni, L. Huang, and Ying Xiang are with the School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China (e-mails: nibotao@mails.gdut.edu.cn, huanglei@mails.gdut.edu.cn, and xiangy@gdut.edu.cn).

Y. Zhu and L. Li are with Beijing Qingniao Fire Protection Co., Ltd., Zhuolu County, Hebei Province 075600, China (e-mails: zoumin2@mails.gdut.edu.cn, and jiangbotao59@mails.gdut.edu.cn).

Y. Zhou is with Guangdong Transportation Vocational College, Guangzhou, 510650, China (e-mail: zengqingchuan@mails.gdut.edu.cn).

J. Yang is with the School of Physics and Electronic Information, Guangxi University for Nationalities, Nanning, 530006, China (e-mail: chenkai26@mails.gdut.edu.cn).

H. Tang is with the Collaborative Innovation Research Institute, Guangdong University of Technology, Heyuan, 517001, China (e-mail: fujiaming@mails.gdut.edu.cn).

I. Introduction

With the progression of urbanization and global climate change, extreme weather events—including intense heat, drought, and strong winds—are occurring with increasing frequency [1–3]. This trend elevates fire risks and underscores the urgent need for autonomous systems capable of operating in hazardous environments. Firefighting robots, which can perform detection and response tasks under high-temperature and smoke-obscured conditions, have thus emerged as a critical research focus. By integrating autonomous perception and execution, these robots enhance operational safety and endurance compared to traditional firefighting methods. A core requirement for such robots is the accurate assessment of fire scale and progression, which directly informs suppression strategies and personnel safety [4,5]. However, real-world fire scenarios present significant perceptual challenges: smoke occlusion, rapidly changing flame morphology, and incomplete sensor data often impede reliable fire-source localization and real-time situation awareness [6].

Current fire-detection approaches largely rely on conventional sensors for temperature, smoke, and gas monitoring. To improve robustness, multi-sensor fusion has gained attention. Systems such as RAS [7] and AIFD [8] combine physical sensors with visible or infrared imagery through feature- and decision-level fusion, achieving early fire detection with reduced false alarms. In parallel, spectral analysis has become a valuable tool for examining fire characteristics. By analyzing radiative signatures of combustion products (e.g., in the visible to mid-wave infrared bands), spectral methods enable flame-temperature estimation, fire-intensity assessment, and even fuel-type classification [9–

12]. Recent studies further show that integrating spectral data with visual or thermal imagery improves both detection reliability and combustion-state analysis [13,14].

In robotic vision, precise environmental perception is essential for navigation and interaction. For flame segmentation, encoder-decoder networks such as U-Net [15] are commonly used but often yield blurred edges and exhibit limited multi-scale fusion. Although improved variants like U³UNet [16] enhance edge delineation, they remain sensitive to complex backgrounds. Dilated-convolution-based methods [17] strengthen multi-scale representation at the cost of high computational overhead, hindering real-time deployment. By contrast, the YOLO series [18] offers a favorable balance between accuracy and inference speed, making it suitable for robotic platforms.

To address feature-selection and multi-scale challenges in fire-related vision tasks, attention mechanisms and feature-pyramid structures have been widely adopted. For instance, SE-Net and CBAM modules have been incorporated into custom networks to improve feature discrimination in fire and smoke detection [19,20], while designs like EFDNet employ channel attention with multi-scale supervision to trade off accuracy and efficiency [21]. Beyond these specialized architectures, lightweight general-purpose modules—notably Efficient Channel Attention (ECA) [22] and the Bidirectional Feature Pyramid Network (BiFPN) [23]—have shown strong performance across various vision applications. Nevertheless, their synergistic integration into a real-time, edge-oriented instance-segmentation framework for flame analysis remains underexplored. More importantly, existing fire-perception studies typically concentrate on isolated detection or segmentation tasks, lacking end-to-end integration of complementary sensing modalities such as stereo vision (for spatial localization) and spectral sensing (for intensity assessment)—a integration crucial for comprehensive robotic firefighting.

To bridge this gap, we propose an integrated fire-perception framework for firefighting robots. Our main contributions are threefold:

An Enhanced Segmentation Model (YOLO-EB): We optimize the YOLOv8-seg baseline by integrating the ECA mechanism and BiFPN architecture. This co-design enhances multi-scale feature representation and flame region recognition, improving segmentation accuracy and stability in complex, dynamic environments.

Mask-Constrained Stereo Ranging: High-quality segmentation masks are leveraged to constrain stereo matching. This strategy mitigates the adverse effects of unstable flame texture, amorphous morphology, and background clutter on ranging accuracy, enabling reliable spatial localization within the effective sensing range.

Multimodal Feature Fusion for Situational Assessment: Building upon precise segmentation and localization, we implement a fusion of visible-light and spectral features. This multimodal approach facilitates real-time assessment of fire intensity and combustion state, equipping firefighting robots with enhanced perceptual and decision-making capabilities.

II. Method

A. Overall Work and Framework Design

The overall workflow of the fire perception and assessment framework proposed in this study is illustrated in Fig. 1.

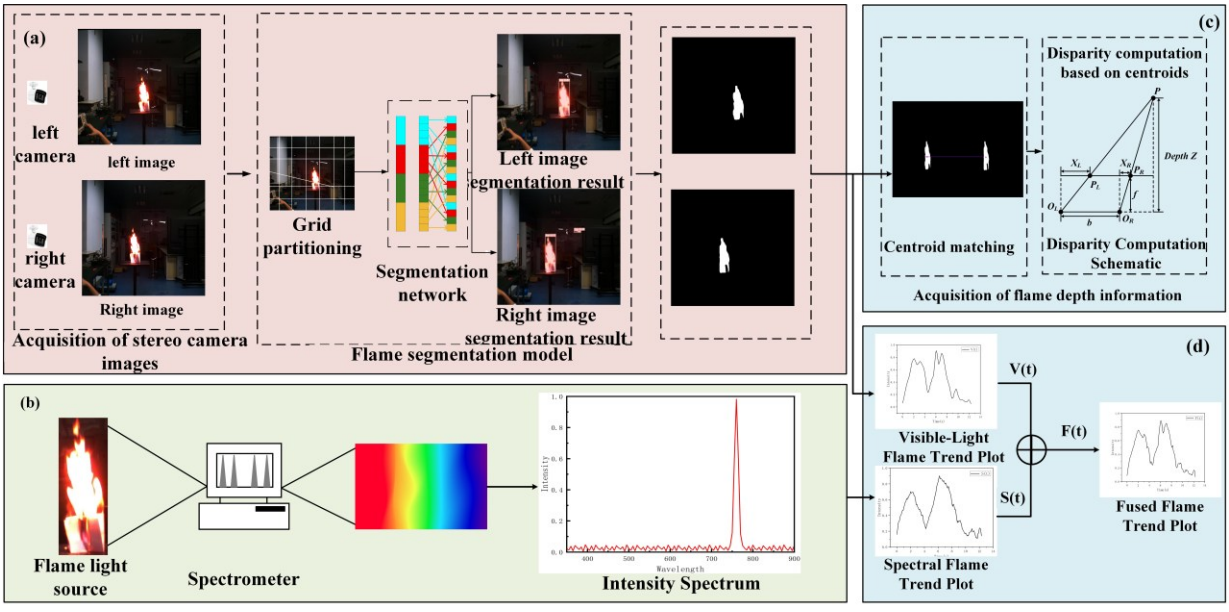

Fig. 1. (a) The binocular vision platform synchronously captures fire scenes, generating left and right view flame binary masks after processing by the flame segmentation network; (b) The spectrometer synchronously collects fire source radiation information and generates spectral intensity distributions; (c) The center of mass of the flame region is extracted based on the segmentation mask, and parallax is calculated through left-right view matching, combined with a depth model to derive the spatial depth of the fire source; (d) V(t) denotes the visible-light fire intensity time signal derived from statistical analysis of red channel pixels in segmented regions, while S(t) represents the spectral fire intensity time signal in characteristic wood combustion bands. Dynamic fusion of these signals yields the composite metric F(t) for fire intensity assessment.

Centered on high-precision synchronous acquisition and collaborative processing of multi-source information, this framework achieves precise fire source localization and combustion intensity assessment through the fusion of visual, spatial, and spectral data. The workflow comprises four key stages: stereo image segmentation, synchronous spectral data acquisition, depth calculation based on disparity, and multi-modal trend analysis.

During the visual perception stage, a binocular vision system is employed for synchronous acquisition and segmentation of fire scenes, as shown in Fig. 1(a). The system consists of two precisely calibrated industrial cameras arranged with a fixed baseline and synchronized via hardware triggering to ensure temporally consistent stereo image pairs—a critical requirement given the dynamic nature of flames. The acquired images are processed by an enhanced YOLO-EB flame segmentation network incorporating Efficient Channel Attention (ECA) and a Bidirectional Feature Pyramid Network (BiFPN). This architecture strengthens multi-scale feature representation while suppressing background interference, enabling accurate pixel-level segmentation of flames with diverse morphologies. The ECA mechanism enhances channel-wise feature discrimination, while BiFPN facilitates bidirectional cross-scale information flow, preserving both fine boundary details and global context. The resulting segmentation masks improve delineation of flame boundaries and small targets, providing reliable visual constraints for subsequent spatial ranging and multimodal analysis.

As illustrated in Fig. 1(b), spectral information acquisition is synchronously integrated with the binocular vision system using a unified trigger signal, enabling simultaneous sampling of visible images and spectral radiation data. This ensures strict temporal alignment between modalities. By analyzing intensity variations of characteristic emission lines associated with elements such as sodium (Na) and potassium (K)—which exhibit distinct radiative behaviors during combustion of common materials—the system captures critical combustion state information. These emission lines are particularly indicative of flame temperature and combustion phase. The spectral data serve as an essential physical basis for assessing fire intensity and combustion characteristics, forming a key modality in multimodal fire perception [25]. The high temporal resolution of the spectrometer enables detection of rapid changes in combustion dynamics that may not be apparent in visible imagery alone.

Fire source localization is achieved via a centroid-based binocular depth estimation method, as shown in Fig. 1(c). High-quality flame segmentation masks are utilized to constrain cross-view matching, effectively reducing the influence of flame texture variability on disparity estimation. Traditional stereo matching algorithms often fail in fire scenarios due to the absence of stable texture features. By leveraging the geometric centroids of segmented flame regions, the proposed method circumvents the need for dense pixel-wise matching, which is computationally expensive and error-prone under such conditions. The spatial depth of the fire source is robustly calculated from the parallax of flame centroids in stereo images, combined with camera intrinsic parameters and baseline length. This centroid-based approach improves computational efficiency and enhances localization stability, particularly at medium ranges where flame masks remain consistent across views. The resulting depth estimates serve as critical inputs for downstream tasks such as fire spread prediction and targeted robotic response.

Based on the triangulation principle, the depth Z of this point relative to the camera's image plane can be calculated using the following formula:

$$Z = \frac{b \times f}{x_L - x_R} \tag{1}$$

$$x_c = \frac{\sum_{x=1}^{W}\sum_{y=1}^{H} x \cdot M(x,y)}{\sum_{x=1}^{W}\sum_{y=1}^{H} M(x,y)} \tag{2}$$

$$y_c = \frac{\sum_{x=1}^{W}\sum_{y=1}^{H} y \cdot M(x,y)}{\sum_{x=1}^{W}\sum_{y=1}^{H} M(x,y)} \tag{3}$$

Where W and H denote the image width and height, respectively. After acquiring the centroid coordinates of the flame masks in the left and right images separately, a robust disparity estimation d can be obtained based on the difference in their horizontal coordinates. Subsequently, d is substituted into the depth formula to calculate the fire source distance Z.

By using the geometric centroid of the segmented flame region for stereo matching, the proposed method effectively alleviates the influence of flame texture instability and edge ambiguity, enabling robust and accurate fire source localization, particularly at medium and short distances. As shown in Fig. 1(d), multimodal flame trend analysis is performed by fusing temporal visible-light and spectral signals. The visible signal $V(t)$ is derived from segmentation results (e.g., flame area or red-channel intensity), while the spectral signal $S(t)$ is obtained from characteristic emission line intensities. Both signals are normalized and smoothed to produce standardized sequences $V_{\text{norm}}(t)$ and $S_{\text{norm}}(t)$ for subsequent fusion analysis.

The core fusion process employs a dynamic weighting mechanism based on local signal stability. For each time point i, a sliding window of length 2r+1 is used to compute the local mean μ and standard deviation σ. The quality score for each signal at time i is then calculated as:

$$\text{score}_i = \frac{|x_i - \mu|}{\sigma + \varepsilon} \tag{4}$$

where ε is a small constant to avoid division by zero. The scores are linearly normalized to the range [0, 1] to obtain $\text{score}_{\text{spec}}(t)$ and $\text{score}_{\text{vis}}(t)$ .The dynamic fusion weight $\alpha(t)$ is derived from the relative quality scores:

$$\alpha(t) = \frac{\text{score}_{\text{spec}}(t)}{\text{score}_{\text{spec}}(t) + \text{score}_{\text{vis}}(t) + \varepsilon} \tag{5}$$

Finally, the fused fire intensity signal is computed as:

$$F(t) = \alpha(t) \cdot S_{\text{norm}}(t) + \big(1 - \alpha(t)\big) \cdot V_{\text{norm}}(t) \tag{6}$$

The algorithm is implemented as follows:

Algorithm 1: Dynamic Multi-modal Fusion based on Sliding Window Confidence: Input: Normalized visible signal

sequence $V_{norm}[1..T]$, Normalized spectral signal sequence $S_{norm}[1..T]$, Sliding window radius r; Output: Fused fire intensity sequence $F[1..T]$.

1. $Initialize\ fusion\ sequence\ F$
2. $for\ time\ point\ i = r+1\ to\ T-r\ \ step\ s\ do$:

// Extract current sliding window data (centered at i)

3. $V_{win} = V_{norm}[i-r:\ i+r]$
4. $S_{win} = S_{norm}[i-r:\ i+r]$

// Calculate statistics within the window (correspond to Eq. 4)

5. $\mu_V, \sigma_V = MeanAndStd(V_{win})$
6. $\mu_S, \sigma_S = MeanAndStd(S_{win})$

// Calculate Z-score and quality score for current time (Eq.4)

7. $\text{score}_\text{v} = |\text{V}_\text{norm}[\text{i}] - \mu\text{v}|/(\sigma\text{v} + \epsilon)$
8. $\text{score}_\text{s} = |\text{S}_\text{norm}[\text{i}] - \mu\text{s}|/(\sigma\text{s} + \epsilon)$

// Calculate dynamic fusion weight (Eq. 5)

9. $\alpha = \text{score}_\text{S}/(\text{score}_\text{S} + \text{score}_\text{V} + \epsilon)$

// Perform weighted fusion(corresponding to Eq.6)

10. $\text{F}[\text{i}] = \alpha \cdot \text{S}_\text{norm}[\text{i}] + (1-\alpha) \cdot \text{V}_\text{norm}[\text{i}]$
11. End for
12. Return F

The parameter r defines the temporal scale for assessing local signal consistency, while the sliding step size s (default s=1 sample) determines the granularity for generating multiple evaluation segments. Based on the prior knowledge that flame dynamics evolve continuously at a second-level scale and through a preliminary parameter sweep on the validation set (testing r=3,5,7), we set r=2 (corresponding to a window length of 2r+1=5 samples, approximately 0.5 seconds). This configuration was empirically determined to provide an optimal trade-off between smoothing short-term noise and effectively capturing genuine trend changes, as corroborated by comparative experiments on the validation set. The constant ϵ (set to 1×10^{-6}) is a negligible value introduced solely for numerical stability to prevent division by zero; its specific magnitude has no significant impact on the algorithm's performance, as it is orders of magnitude smaller than typical signal variations.

To quantify the multimodal fusion system's ability to track fire dynamics, we define the Trend Consistency Accuracy by comparing the instantaneous trend symbols of the fused output F(t) and the infrared ground truth R(t):

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} I\!I\Big(\text{sign}\big(F(i) - \text{F}(i-1)\big) = \text{sign}\big(R(i) - \text{R}(i-1)\big)\Big) \times 100\% \quad (7)$$

where $\text{sign}(\text{x})$ returns +1 for a rising trend ($x > \varepsilon$), −1 for declining ($x < -\varepsilon$), and 0 for stable, with ε being a small threshold. N is the total number of comparison points, and $I\!I(\cdot)$ is the indicator function.

Finally, the trend output unit performs linear fitting (e.g., based on the least squares method) on the fusion sequence $F(t)$ within a short time window, characterizing the instantaneous rate of change in fire intensity through the slope k of the fitted curve. By adaptively fusing and analyzing temporally synchronized and complementarily characterized visible and spectral information, this method enables stable quantification of fire development trends, providing effective support for fire situation assessment and early warning.

B. YOLO-EB

To address challenges in fire scenarios such as complex background interference, inadequate representation of multi-scale flame features, and limited segmentation accuracy for irregular fire patterns, this paper enhances the YOLOv8-seg segmentation framework through two targeted improvements: integration of an Efficient Channel Attention (ECA) mechanism and adoption of a Bidirectional Feature Pyramid Network (BiFPN). Specifically, the backbone network incorporates the ECA module to strengthen the network's response to discriminative flame feature channels, thereby improving feature focusing in prominent fire zones; the neck network adopts BiFPN for multi-scale feature fusion, enhancing the model's perception of flame regions at different scales through bidirectional information flow and weighted fusion mechanisms. The overall improved network architecture is illustrated in Fig. 2.

The overall architecture of the proposed improved YOLO-EB segmentation network incorporates the following specific enhancements:

(1) Integration of the Efficient Channel Attention (ECA) Mechanism: The ECA module is embedded in the backbone network to adaptively recalibrate channel-wise feature responses by capturing local cross-channel interactions. Unlike conventional attention mechanisms that rely on dimensionality reduction—which may lead to information loss—ECA employs a lightweight one-dimensional convolution with an adaptive kernel size determined by the channel dimension. This enables the network to model inter-channel dependencies more efficiently, directing its focus toward discriminative features associated with flame regions while suppressing irrelevant background noise. As a result, the network exhibits enhanced sensitivity to prominent fire zones and improved resistance to visual distractions commonly present in fire scenes, such as smoke, reflections, or cluttered environments.

(2) Adoption of the Bidirectional Feature Pyramid Network (BiFPN): The original feature fusion structure in the neck network is replaced by BiFPN, which introduces a learnable weighted fusion mechanism across multi-scale features. By integrating both top-down and bottom-up information flows, BiFPN facilitates effective interaction between high-level semantic cues and low-level spatial details. This bidirectional architecture ensures that flame instances of varying scales—from small ignition points to large-scale fire regions—are accurately represented and segmented. The weighted feature fusion further enhances the model's capacity to maintain spatial consistency in segmentation boundaries, which is particularly critical for irregular and dynamic flame morphologies.

In summary, the aforementioned architectural enhancements enable the YOLO-EB model to achieve robust and accurate flame segmentation in complex dynamic environments, providing a reliable foundation for subsequent spatial localization and multimodal fusion tasks.

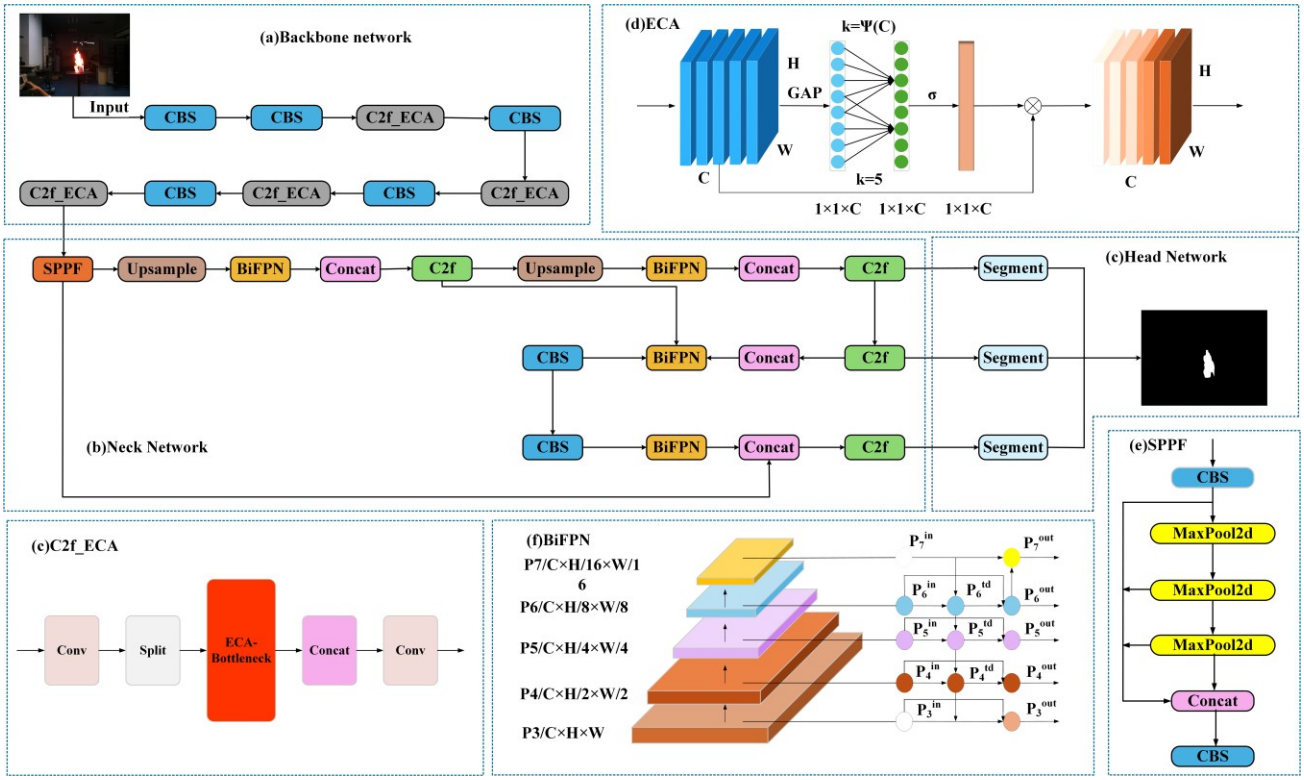

Fig. 2. Overall architecture of the proposed improved YOLO-EB segmentation network. The network consists of three core components: (a) backbone network, (b) neck network, (c) head network. Key functional modules include: (d) Efficient Channel Attention (ECA) module, (e) C2f_ECA module architecture, (f) Bidirectional Feature Pyramid Network (BiFPN) module architecture, and (g) Spatial Pyramid Pooling Fast (SPPF) module architecture.

To address the common issues of edge blurring, missed detections, and false positives in fire source segmentation tasks under complex backgrounds, existing research typically incorporates attention mechanisms to enhance the model's feature representation capabilities. Among these, the Squeeze-Excitation (SE) attention mechanism compresses spatial information through Global Average Pooling (GAP) and utilizes fully connected layers to model inter-channel dependencies. Convolutional Block Attention Modules (CBAM) further integrate channel and spatial attention, generating spatial weight maps through average and max pooling. While these methods effectively enhance target feature representation, their reliance on fully connected layers or multi-branch structures significantly increases model parameters and computational complexity. This limits real-time performance in instance segmentation tasks and fails to meet resource constraints in edge computing scenarios like firefighting robots.

In contrast, the Efficient Channel Attention (ECA) mechanism generates channel description vectors through global pooling and employs one-dimensional convolutions instead of fully connected layers to capture cross-channel interactions, thereby avoiding information loss from channel dimension reduction. Its parameter scale is reduced by more than an order of magnitude compared to the SE attention mechanism. Additionally, ECA introduces an adaptive kernel size adjustment strategy that dynamically balances local and global information modeling capabilities based on the number of channels, achieving a better trade-off between computational efficiency and feature representation. Leveraging these advantages, this paper integrates the ECA attention mechanism into the C2f module of the YOLOv8-seg backbone network, with its specific structure shown in Fig. 2(d).

The ECA module first aggregates spatial information via global average pooling to obtain a channel descriptor v_c :

$$v_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_c(i,j) \tag{8}$$

Where$F \epsilon R^{C \times H \times W}$ is the input feature map. A 1D convolution with an adaptive kernel size k (determined by channel dimension C) is then applied to capture cross-channel interactions, producing channel attention weights ω_c through a sigmoid activation. The final output feature map is obtained by channel-wise multiplication:

$$F_{out}^{(c,i,j)} = F^{(c,i,j)} \omega_c \tag{9}$$

This lightweight design enhances the network's focus on discriminative flame features while minimizing computational overhead.

To address the limitations of existing object detection and segmentation models in multi-scale fire feature fusion, particularly the constrained accuracy in locating and detecting small-scale fire sources, this paper replaces the original feature fusion structure in the YOLOv8-seg neck network with a Bidirectional Feature Pyramid Network (BiFPN, as shown in Fig. 2). BiFPN takes the multi-resolution feature maps output from the backbone network as input, sequentially labeled as P_3, P_4, P_5, P_6, and P_7, where P_3 denotes high-resolution low-level features and P_7 denotes low-resolution high-level features. In the top-down path, high-level features are fused with low-level

features through progressive upsampling to enhance the semantic information of low-level features. In the bottom-up path, low-level features are propagated to high-level features through progressive down sampling to supplement the spatial details of high-level features. Finally, the outputs from both fusion paths are integrated through weighted combination to form the final feature representation, calculated as follows:

$$P_i^{out} = \mathrm{Conv}\left(\frac{w_1 \cdot P_i^{in} + w_2 P_i^{td} + w_3 P_{i-1}^{out}}{w_1 + w_2 + w_3 + \epsilon}\right) \quad (10)$$

where: P_i^{in} is the intermediate feature from the top-down path, and w_j are learnable weights normalized via Softmax:

$$w_j' = \frac{\exp(w_j)}{\sum_j \exp(w_j)} \quad (11)$$

Finally, through multi-layer stacking and repeated bidirectional feature fusion, the network progressively optimizes its feature representation capabilities. Simultaneously, it retains only the connection paths between key layers to effectively integrate multi-scale features while reducing overall computational overhead.

This enables the network to adaptively emphasize critical information while suppressing redundant interference in complex backgrounds, thereby highlighting prominent flame regions and enhancing the stability of multi-scale segmentation. The overall structure of BiFPN is illustrated in Fig. 2(f).

III. Experimental Analysis and Discussion of Results

A. Dataset and Experiment Settings

The training dataset used in this experiment is sourced from the publicly available Fire Segmentation Dataset (FSD). The dataset is divided into training and testing sets at a 9:1 ratio, with an additional 10% of the training set allocated as a validation set. The FSD dataset encompasses a wide range of typical fire and non-fire scenarios, including grasslands, forests, buildings, road environments, and small-scale fire incidents, demonstrating strong scene diversity. The dataset comprises 3,203 annotated fire images and 1,797 non-fire images. The non-fire samples include complex visual scenarios prone to misclassification, such as sunsets and urban lighting, which help enhance the model's ability to distinguish background interference. Regarding segmentation annotations, the dataset contains approximately 6,000 fire source instance labels and 3,000 smoke instance labels. The overall annotations exhibit balanced distribution across both category counts and spatial patterns, ensuring robust diversity. The dataset is available through its official open-source repository [26]. The dataset is available via its official open-source repository [26]. The dataset is available at: https://github.com/suyixuan123s/Fire-Segmentation-Dataset.git.

Spectral data were collected by our research team using a spectrometer in an experimental setting. Meanwhile, real-time fire combustion data were synchronously acquired via an infrared camera as the ground truth to reflect the actual fire combustion status, serving as the basis for evaluating the quality of the fused data.

B. Implementation Details

All experiments were conducted under the configuration specified in Table I. Models were trained for up to 300 epochs with early stopping (patience=100), a batch size of 16, and an input size of 640×640. We used SGD (momentum=0.937, weight decay=5×10^{-4}) with an initial learning rate of 0.01, employing a linear warmup for 3 epochs followed by cosine annealing decay.

For data augmentation, we applied Mosaic (disabled in the last 10 epochs), random horizontal flipping (p=0.5), and HSV adjustments. To ensure reproducibility, a fixed random seed (0) was used, and all reported results are averaged over five independent runs.

Inference speed (FPS) was measured on an RTX A5000 GPU with batch size 1, following 100 warm-up iterations and averaged over 1000 consecutive runs.

TABLE I
Experimental Environment and Configuration Parameters

Schedule	Capacity
operating system	Window
runtime environment	Cuda 11.3
	Python3.8.10
	Pytorch 1.11.0
GPU	RTX A5000
CPU	Intel Xeon Silver 4120R

C. Dataset Evaluation

Model performance was assessed using standard metrics: mean Average Precision (mAP@0.5) for segmentation accuracy, Frames Per Second (FPS) for inference speed, and parameter count for model complexity. To statistically validate improvements, paired t-tests (α=0.05) were conducted, with significance denoted as *p<0.05, **p<0.01, and ***p<0.001. The reliability of measurements (e.g., ranging) is reported with 95% confidence intervals. For fire trend assessment, trend accuracy measures the alignment between the predicted and ground-truth intensity trend.

D. Comparative Experiment with Other Algorithms

To objectively evaluate the comprehensive performance of the improved YOLO-EB model in fire area segmentation tasks within complex dynamic environments, this paper conducts comparative experiments against multiple mainstream object detection and semantic segmentation models (encompassing lightweight models and high-precision segmentation models). The experimental results are summarized in Table II. The comparison reveals that existing mainstream models exhibit certain limitations in fire segmentation tasks: While traditional high-precision segmentation models (such as Deeplabv3+, Mask R-CNN, and YOLACT) achieve high segmentation accuracy, they generally suffer from high computational complexity, large parameter sizes, and limited inference speeds,

making them unsuitable for meeting the real-time and edge deployment requirements of firefighting robots. In contrast, lightweight models (e.g., YOLOv8(n), U-Net, and PSPNet) offer advantages in model size and inference speed. However, their segmentation accuracy and stability are insufficient, with noticeable false negatives, making it difficult to provide a reliable visual foundation for subsequent fire source ranging and multimodal feature extraction.

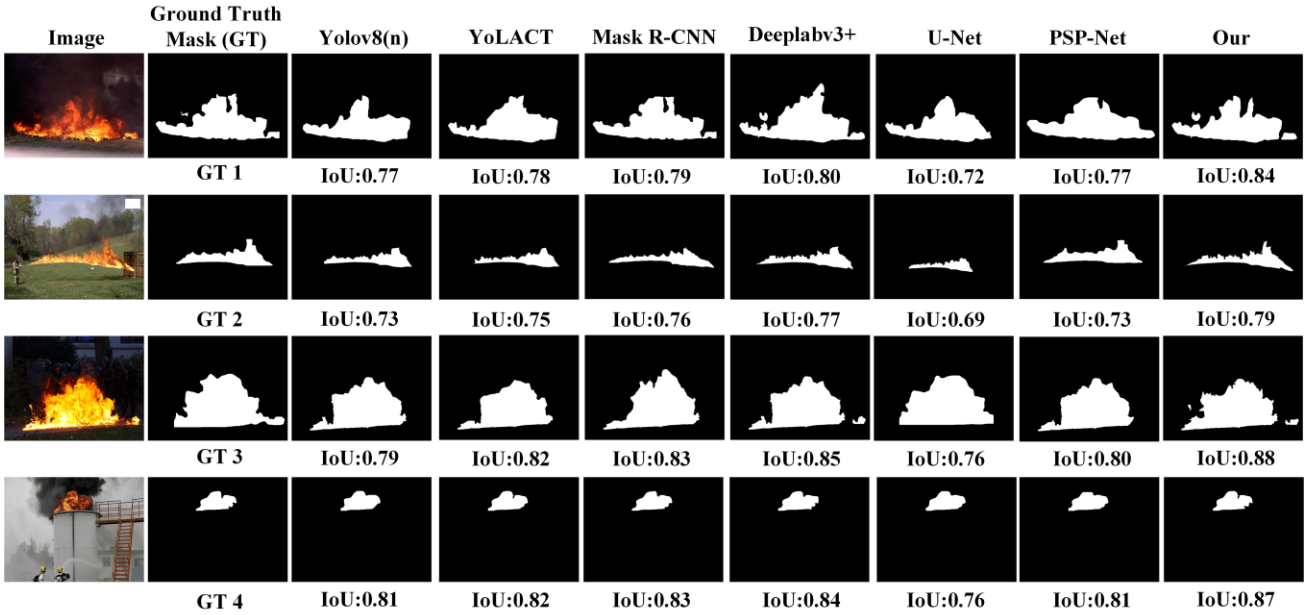

Fig. 3. Segmentation results of seven methods on four test samples are shown, with the IoU value of each prediction indicated below. The samples cover different scene conditions. Higher IoU values indicate higher segmentation accuracy (range 0-1).

TABLE II
COMPARATIVE EXPERIMENT OF DIFFERENT SEGMENTATION ALGORITHMS

Model	Recall(Mean ± SD)	mAP@0.5 (Mean ± SD)	FPS(Mean ± SD)	Params
Yolov8(n)	0.662± 0.011	0.777± 0.003	113.6± 2.1	12.43
YoLACT	0.651± 0.014	0.788± 0.005	31.2± 1.5	129.87
Mask R-CNN	0.667± 0.012	0.792± 0.004	33.7± 1.8	62.00
Deeplabv3+	0.657± 0.012	0.799± 0.004	47.1± 2.3	136.62
U-Net	0.640± 0.016	0.736± 0.006	97.1± 3.5	26.67
PSP- Net	0.638± 0.013	0.778± 0.005	52.2± 2.7	58.65
YOLO-EB	0.680± 0.008	0.821± 0.002	95.3± 1.9	17.10

As summarized in Table II, YOLO-EB achieves a mean recall of 0.680 ± 0.008 and a mean average precision (mAP) of 0.821 ± 0.002 over five independent runs, demonstrating statistically significant improvements over existing methods. This corresponds to a false negative rate (FNR) of 32.0%. Compared to the widely used YOLOv8(n) baseline (recall: 0.662 ± 0.011, FNR: 33.8%; mAP: 0.777 ± 0.003), YOLO-EB exhibits an absolute increase of 0.018 in recall (reducing FNR by 1.8 percentage points) and 0.044 in mAP (paired t-test, $p < 0.001$). Although the recall remains at 0.680, the notably higher mAP and lower standard deviation across runs indicate more consistent segmentation boundaries and better localization of detected flames.

In terms of inference efficiency, YOLO-EB maintains a throughput of 95.3 ± 1.9 frames per second (FPS), approximately 16% lower than YOLOv8(n) (113.6 ± 2.1 FPS) but more than twice as fast as high-accuracy models such asDeepLabv3+ (47.1 ± 2.3 FPS) and Mask R-CNN (33.7 ± 1.8 FPS), comfortably meeting the real-time requirement of 30 FPS for robotic firefighting platforms. With only 17.10 M parameters—a 38% increase over YOLOv8(n) yet merely 12–13% of those of DeepLabv3+ (136.62 M) and YOLACT (129.87 M)—YOLO-EB offers a favorable balance of accuracy, speed, and model complexity, making it suitable for deployment on resource-constrained edge devices.

To fully verify the effectiveness of each module in the proposed algorithm, module tests are conducted using the same dataset and training parameters. Taking the YOLOv8-seg model as the basic network framework, modules are gradually added for ablation experiments. Specific results are shown in Table III, where the checked items indicate that the corresponding module is incorporated into the basic YOLOv8-seg model.

TABLE III
ABLATION STUDY OF PROPOSED MODULES (ECA AND BIFPN) ON THE YOLOV8-SEG BASELINE

Group	ECA	BiFPN	Recall (Mean ± SD)	mAP@0.5 (Mean ± SD)	p-value (vs. Baseline)	FPS (Mean ± SD)	Params
1			0.659±0.011	0.777±0.003	-	113.6± 2.10	12.43
2	√		0.668±0.009	0.789±0.003	0.019*	109.8± 2.10	12.55
3		√	0.672±0.008	0.797±0.003	0.003**	98.5± 1.90	16.85
4	√	√	0.680±0.008	0.821±0.002	<0.001***	95.3± 1.90	17.1

Table III presents the results of our ablation study, which quantifies the individual and combined contributions of the ECA and BiFPN modules to YOLO-EB's performance. All comparisons are based on five independent runs, with statistical significance assessed using paired t-tests ($\alpha = 0.05$).

The ECA module alone (Group 2) yields a statistically significant improvement in mAP from 0.777 to 0.789 ($p = 0.023$), while increasing parameters by only 0.12 M and reducing FPS by 3.4%. The recall improves from 0.662 to 0.671, reducing the false negative rate from 33.8% to 32.9%. This demonstrates ECA's effectiveness in enhancing flame feature representation with minimal computational overhead.

The BiFPN module alone (Group 3) provides a larger mAP gain to 0.797 ($p = 0.005$) but incurs greater computational cost, adding 4.42 M parameters and reducing FPS by 13.0%. It further improves recall to 0.676, corresponding to an FNR of 32.4%. This reflects BiFPN's stronger multi-scale fusion capabilities at the expense of increased complexity.

Crucially, combining both modules (Group 4) produces synergistic effects: mAP rises to 0.821, representing a 5.7% relative improvement over the baseline with highly significant statistical evidence ($p < 0.001$). The recall reaches 0.680, achieving the lowest false negative rate of 32.0% among all configurations. The model maintains real-time performance at 95.3 FPS while achieving the lowest mAP variance among all configurations (SD = 0.002), indicating enhanced stability.

These results validate that ECA and BiFPN provide complementary benefits: ECA enhances feature discrimination with minimal cost, while BiFPN improves multi-scale representation at higher computational expense. Their combination yields better performance for fire segmentation tasks, balancing accuracy, speed, and model efficiency for edge deployment.

E. Comparison of Ranging Results at Different Distances

The spatial localization accuracy of our stereo vision system was evaluated at six distances (50–1000 cm), with 20 independent measurements per point. Performance metrics include the mean measured distance, standard deviation, 95% confidence intervals (t-distribution, df=19), and average error rate (Table IV and Fig.4)

The system achieves high accuracy at medium distances (50–300 cm), with error rates of 2.06–3.15% and narrow confidence intervals (e.g., 50 cm: 51.08±2.06 cm, 95% CI [50.11, 52.05] cm). This demonstrates reliable flame localization for typical firefighting robot operational ranges.

Accuracy declines beyond 300 cm due to reduced parallax resolution, with error rates reaching 4.18% at 500 cm and 7.37% at 1000 cm, accompanied by significantly wider confidence intervals. These results confirm the system's suitability for medium-range fire source localization, while highlighting the need for sensor fusion in long-range applications.

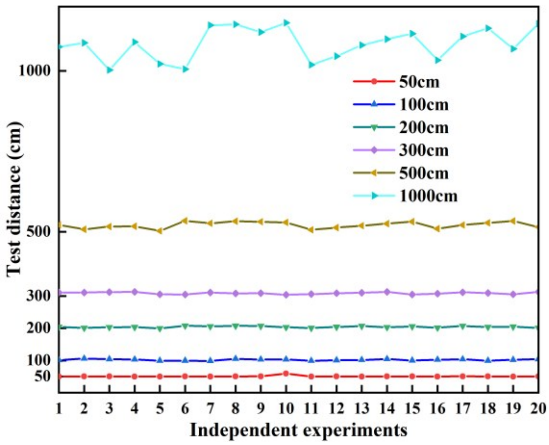

Fig. 4. Ranging experiment results under different target distances (50, 100, 200, 500, and 1000 cm). Each distance point was measured 20 times to evaluate the stability and consistency of the ranging results.

TABLE IV
STEREO VISION RANGING ACCURACY AT DIFFERENT TARGET DISTANCES

Actual distance (cm)	Average measurement distance (cm)	95% Confidence Interval (cm)	Average error rate (%)
50	51.08 ± 2.06	[50.11, 52.05]	2.16
100	102.06± 2.29	[100.99,103.13]	2.06
200	204.40± 2.47	[203.24,205.56]	2.20
300	309.05± 3.07	[307.61,310.49]	3.15
500	520.73 ± 9.96	[516.07,525.39]	4.18
1000	1078.7 ± 47.3	[1056.5,1100.9]	7.37

F. Comparative Experiments on Multimodal Data Fusion

This study conducted a statistical trend accuracy analysis on ten sets of flame monitoring data, with each set representing a continuous 12.5-second combustion process, to evaluate the performance of single features versus multimodal fusion features in reflecting actual flame evolution trends. To ensure statistical robustness, we employed a sliding window analysis based on the parameters defined in Section II.A (window radius r=2, step size s=1, window length 5 samples). This approach yielded multiple independent accuracy estimates per experimental group, enabling the calculation of the reported means and standard deviations (Table V and Fig. 5).

Experimental results indicate that the trend accuracy of visible-light features (V(t)) is relatively low, with a mean accuracy of 86.22 % and the highest variability (± 1.76 %), reflecting its susceptibility to environmental interference such as smoke and illumination changes. Spectral features (S(t)) demonstrate better mean accuracy (88.08 %) and moderate variability (± 1.52 %). In contrast, the proposed fused features (F(t)) not only achieve the highest mean accuracy (92.45 %) across all ten experimental sets but also exhibit the lowest variability (± 1.28 %). This outcome indicates that when single features exhibit limitations in stability, the dynamic weighted fusion method effectively leverages complementary information, thereby simultaneously enhancing both the accuracy and the reliability of trend determination. The fused features consistently outperform single features in every experimental group, validating the effectiveness and robustness of the proposed multimodal fusion approach for practical flame evolution trend analysis.

IV. CONCLUSION AND DISCUSSION

This paper presents an integrated flame localization and situational awareness system for robotic firefighters, which synergistically combines an enhanced YOLO-EB segmentation model with stereo vision and spectral information. The YOLO-EB model, optimized with an Efficient Channel Attention (ECA) mechanism and a Bidirectional Feature Pyramid Network (BiFPN), achieves a segmentation recall of 0.680 (False Negative Rate: 32.0%) and an mAP of 0.821, demonstrating an effective balance among accuracy, real-time inference speed (95.3 FPS), and parameter efficiency (17.10 M). For robotic navigation, the stereo-vision subsystem maintains an average ranging error of 2–3 % within 3 meters, while a novel spectral-visible fusion module achieves over 92 % accuracy in fire intensity trend assessment.

TABLE V
TREND ACCURACY OF MULTI-MODAL FUSION ACROSS EXPERIMENTAL GROUPS

Group	Spectrum (%) (Mean ± SD)	Visible (%) (Mean ± SD)	Fused (%) (Mean ± SD)
Group1	89.96±1.45	85.14±1.80	94.76±1.20
Group2	88.02±1.50	87.04±1.75	91.89 ±1.25
Group3	87.13±1.55	85.68±1.70	92.47±1.30
Group4	88.08±1.52	85.85±1.78	90.96±1.35
Group5	89.64±1.48	87.18±1.72	91.16±1.28
Group6	88.83±1.47	86.29±1.76	94.76±1.22
Group7	86.37±1.60	85.32±1.82	89.6±1.40
Group8	87.72±1.53	86.9±1.74	93.63±1.26
Group9	87.5±1.54	86.97±1.73	91.34±1.32
Group10	87.51 ±1.54	85.81 ±1.79	93.9±1.24
Average accuracy	88.08 ±1.52	86.22±1.76	92.45±1.28

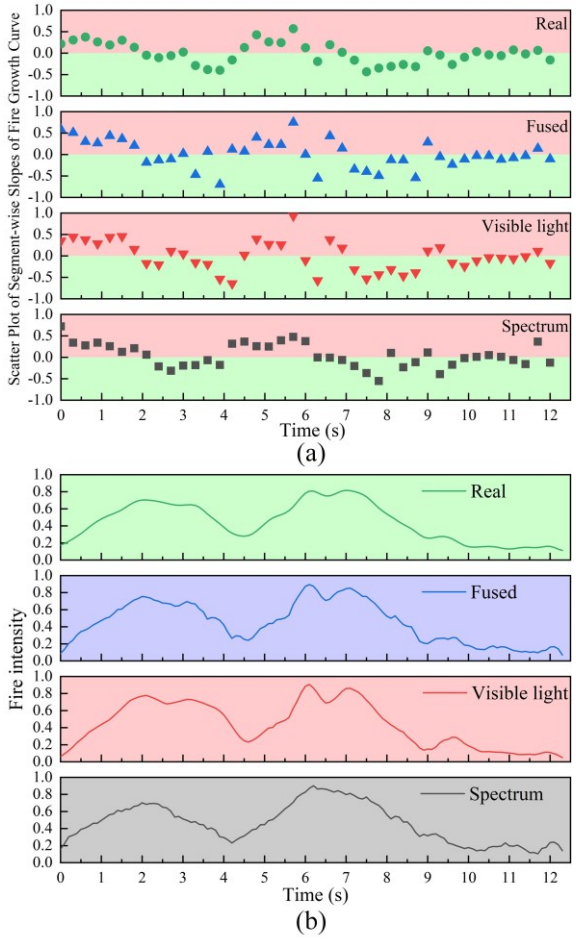

Fig. 5. (a) Flame intensity trends captured by the fire assessment system during a 12.5-second wood combustion process. “Spectrum” denotes combustion trends recorded by the spectrometer, “Visible Light” represents flame variations captured by the CCD camera, “Fused” indicates flame trends after multimodal information fusion, and “Real” corresponds to combustion intensity reference data captured by the infrared camera. (b) Scatter plot of fire intensity data. The red background area indicates the phase of increasing flame intensity, while the green background area indicates the phase of decreasing flame intensity.

Technical Robustness and Design Rationale

The system's robustness in challenging conditions is a core design consideration. The multi-modal fusion employs a dynamic weighting strategy, not a fixed scheme, which continuously evaluates the short-term reliability of each sensor stream. This design directly addresses inherent weaknesses: when the visible-light channel suffers from missed detections (as quantified by its recall limit) due to issues like smoke obscuration, its influence is automatically attenuated, thereby leveraging the complementary spectral information (targeting Na/K emission lines) which is more penetrative to smoke and stable under varying illumination. This mechanism provides a principled approach to mitigate performance degradation in complex scenes.

Critical Limitations and Deployment Constraints

We explicitly acknowledge that the recall of 0.680 indicates a significant false-negative risk, primarily for flames under extreme occlusion, at a very small scale, or with severe motion blur. Therefore, for safety-critical deployment, this vision framework must be part of a multi-sensor redundant architecture, supplemented by thermal cameras and other sensing modalities, and operate under a human-in-the-loop supervision scheme. Practical deployment is further constrained by onboard compute resources, power budget, and the robot's thermal resilience in high-temperature environments.

The current validation has several boundaries that must be expanded for practical application:

Environmental Extremes: Systematic evaluation under dense smoke, lens contamination (water, soot), strong glare/reflections, and rapid illumination changes—common in real firefighting—was not conducted and remains a critical gap.

Dynamic & Multi-Target Scenarios: Performance under violent robot motion-induced blur and with multiple, densely clustered flames requires further assessment.

Fuel and Scenario Generality: Experiments focused on wood fires. Performance on fires from different fuel types (e.g., liquids, plastics, metals)—which exhibit distinct spectral and morphological signatures—is untested and constitutes a major limitation for generalizability.

Performance Across Flame Sizes: A standardized evaluation across flame sizes (small, medium, large) is currently absent, limiting insight into model performance across varied fire scenarios. This omission is due to the lack of size annotations in public datasets like FSD. More fundamentally, categorizing flames by pixel area alone is inadequate, as it confuses true fire scale with camera distance. A physically meaningful assessment requires combining segmentation with distance estimation—a core function of our system. Establishing a distance-aware size benchmark is therefore a vital open challenge for enabling finer-grained diagnosis of fire perception models.

The centroid-based stereo matching method, while computationally efficient, has specific limitations. First, significant shape asymmetry of the flame between the two views—caused by perspective or flickering—can introduce sub-pixel centroid errors, particularly at longer distances. Second, the current implementation ranges only the largest connected mask region, which may merge multiple separate flames or fail to localize individual sources in multi-fire scenarios. Third, partial occlusion (e.g., by obstacles or smoke) can yield incomplete masks, shifting the centroid away from the true flame center. These factors contribute to the rising ranging uncertainty beyond ~5 m shown in Table IV. Within the intended operational range ($\leqslant 3$ m), however, the short baseline reduces perspective disparity, and the high-quality segmentation from YOLO-EB keeps the centroid a stable and efficient proxy for dense matching—a deliberate trade-off suitable for real-time robotic deployment. Future work may integrate lightweight optical-flow checks or sparse feature matching within the mask to further mitigate such edge-case errors.

Pathway from Validation to Application

This work serves as a proof-of-concept validation of the "segmentation-stereo ranging-multi-modal fusion" pipeline under representative challenges. The limitations outlined above define the essential stress tests required to transition from a capable method to a field-ready system.

Future work will therefore prioritize:

(1) Data and Scenario Expansion: Building and benchmarking on datasets encompassing the adverse conditions listed above.

(2) Robustness Enhancements: Integrating auxiliary sensors (e.g., lidar for long-range depth, inertial measurement units for motion compensation) and exploring online image quality assessment to diagnose sensor degradation.

(3) Advanced Perception Modules: Developing fuel-type classification based on spectral-temporal signatures to enable adaptive perception, and creating more refined fire-spread prediction models that incorporate real-time environmental context.

(4) Granular Performance Benchmarking: Collaborating to establish and adopt standardized evaluation protocols for fire segmentation, including distance-aware flame size categorization. This involves defining size thresholds (small, medium, large) based on estimated physical dimensions rather than pixel area alone, utilizing the system's own ranging capabilities. Such benchmarks will enable more precise diagnosis of model strengths and weaknesses across different fire scales.

(5) Focus on sensor-fusion strategies: integrating a lightweight LiDAR or radar module provides absolute depth priors to rectify scale ambiguity; furthermore, deep fusion of thermal imagery with stereo geometry enables more robust 3D reconstruction of high-temperature regions — evolving the

system from pure ranging to comprehensive thermal-geometric situational awareness.

In summary, the proposed framework demonstrates superior potential over single-modal approaches for detecting medium-range wood fires amidst smoke and light variations, providing a foundational perception module for robotic firefighters. Its ultimate practicality hinges on rigorous validation and extension across the full spectrum of real-world firefighting challenges, which forms the immediate and critical roadmap for our subsequent research.

Conflict of Interest

Data Availability Statement: For reasonable requirements, t he corresponding data can be provided as requested.
Conflicts of Interest: No potential conflict of interest was repo rted by the author(s).

Acknowledgments

The author extends sincere gratitude to his advisor, Professor Ying Xiang, for his invaluable guidance and resources. Heartfelt thanks are also extended to Lei Huang for his assistance in experimental data collection and manuscript preparation. Appreciation is given to Yunfei Zhou, Yan Zhu, Lin Li, Jingjing Yang, and Hao Tang for their invaluable guidance throughout the writing process. Their contributions were crucial to the successful completion of this project.

References

[1] S. Yang, W. Ding, and Y. Wang, "Real-time flame detection and localization for firefighting robots," *Robot. Auton. Syst.*, vol. 194, Art. p. 105147, 2025. doi:10.1016/j.robot.2025.105147.

[2] J. Zhang, L. Liu, and H. Wang, "YOLOGX: An improved forest fire detection algorithm based on deep learning," *Front. Environ. Sci*., vol. 12, Art. p. 1486212, 2024. doi : 10.3389/fenvs.2024.1486212.

[3] H. Deng, D. Li, S. Cai et al., "Spatio-temporal dynamics of forest fire occurrence in Yunnan, China from 2001 to 2021 based on MODIS," *npj Nat. Hazards*, vol. 2, Art. no. 52, 2025. doi.org/10.1038/s44304-025-00102-6.

[4] Z. Dong, C. Zheng, F. Zhao, G. Wang, Y. Tian, and H. Li, "A deep learning framework: Predicting fire radiative power from the combination of polar-orbiting and geostationary satellite data during wildfire spread," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens*., vol. 17, pp. 10827–10841, May 2024. doi:10.1109/JSTARS.2024.3403146.

[5] G. Y. Wang, "Fire Source Range Localization Based on the Dynamic Optimization Method for Large-Space Buildings," *IEEE Sens.*, vol. 18, no. 6, Art.pp. 1954, 2018. doi:10.3390/s18061954.

[6] P. Vorwerk, "Classification in Early Fire Detection Using Multi-Sensor Nodes—A Transfer Learning Approach," *IEEE Sens.*, vol. 24, no. 5, Art. pp. 1428, 2024.

[7] F. Khan, Z. Xu, J. Sun, F. M. Khan, A. Ahmed, and Y. Zhao, "Recent Advances in Sensors for Fire Detection," *Sens.*, vol. 22, no. 9, Art. no. 3310, 2022. doi:10.3390/s22093310.

[8] X. Deng, "An Indoor Fire Detection Method Based on Multi-Sensor Fusion and a Lightweight Convolutional Neural Network," *Sens*., vol. 23, no. 24, Art. pp. 9689, 2023. doi:10.3390/s23249689.

[9] T. Ren, M. F. Modest et al., "Machine learning applied to retrieval of temperature and concentration distributions from infrared emission measurements," *Appl. Energy*, vol. 252, Art. pp. 113448, 2019. doi:10.1016/j.apenergy.2019.113448.

[10] C. Magro, O. C. Goncalves et al., "Remote sensing of volatile organic compounds release during prescribed fires in pine forests using open-path Fourier transform infra-red spectroscopy," *Int. J. Wildland Fire,* vol. 33, no. 4, Art. no. WF23019, 2024. doi:10.1071/WF23019.

[11] K. L. Minatre, M. M. Arienzo et al., "Charcoal analysis for temperature reconstruction with infrared spectroscopy," *Front. Earth Sci*., vol. 135, Art. pp. 1354080, 2024. doi:10.3389/feart.2024.1354080.

[12] D. Sper et al., "Wildfire Detection Using Convolutional Neural Networks and PRISMA Hyperspectral Imagery: A Spatial-Spectral Analysis," *Remote Sens.*, vol. 15, no. 19, Art. pp. 4855, 2023. doi:10.3390/rs15194855.

[13] K. Thangavel et al., "Autonomous Satellite Wildfire Detection Using Hyperspectral Imagery and Neural Networks: A Case Study on Australian Wildfire," *Remote Sens*., vol. 15, no. 3, Art. p. 720, 2023. doi:10.3390/rs15030720.

[14] W. N. Sun et al., "A Study on Flame Detection Method Combining Visible Light and Thermal Infrared Multimodal Images," *Fire Technol*., vol. 61, pp. 2167–2188, 2025. doi:10.1007/s10694-024-01676-9.

[15] D. Spiller et al., "Wildfires Temperature Estimation by Complementary Use of Hyperspectral PRISMA and Thermal," *JGR: Biogeosciences*, vol. 127, no. 12, Art. p. e2022JG007055, 2022. doi:10.1029/2022JG007055.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. *IEEE Int. Conf. Med. Image Comput. Comput.-Assisted Interv. (MICCAI)*, 2015, pp. 234–241. doi:10.1007/978-3-662-54345-0_3.

[17] H. Feng, J. Qiu, L. Wen et al., "U3UNet: An accurate and reliable segmentation model for forest fire monitoring based on UAV vision," *Neural Networks*, vol. 175, pp. 262–275, 2025. doi: 10.1016/j.neunet.2025.107207.

[18] P. Gonçalves et al., "Fire segmentation using a DeepLabv3+ architecture," in Proc. 2025 *IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, 2025. doi:10.1117/12.2573902.

[19] Z. Deng, S. Hu, S. Yin, Y. Wang, A. Basu, and I. Cheng, "Multi-step implicit Adams predictor-corrector network for fire detection," *IET Image Process*., vol. 16, no. 9, pp. 2338–2350, 2022. doi: 10.1049/ipr2.12491.

[20] T. Li, H. Zhu, C. Hu, and J. Zhang, "An attention-based prototypical network for forest fire smoke few-shot detection," *J. Forestry Res*., vol. 33, no. 5, pp. 1493–1504, 2022. doi: 10.1007/s11676-022-01457-6.

[21] S. Li, Q. Yan, and P. Liu, "An efficient fire detection method based on multiscale feature extraction, implicit deep supervision and channel attention mechanism," *IEEE Trans. Image Process*., vol. 29, pp. 8467–8475, 2020. doi: 10.1109/TIP.2020.3016431.

[22] H. Du, Q. Li, Z. Guan, H. Zhang, and Y. Liu, "An improved lightweight YOLOv8 network for early small flame target detection," *Processes*, vol. 12, no. 9, Art. pp. 1978, 2024. doi: 10.3390/pr12091978.

[23] S. Li et al., "Efficient Channel Attention for Flame Feature Enhancement in Fire Detection," *IEEE Trans. Ind. Electron*., vol. 70, no. 12, pp. 12764–12774, 2023. doi:10.3390/fire8020038 /TIE.2023.3287654.

[24] D. Meng and X. Wu, "YOLOv5s-RBC: A Lightweight Fire Detection Algorithm with BiFPN for Multi-Scale Feature Fusion,"

SSRN Electron. J., vol. 34, pp. 1–12, 2025. doi: 10.18280/ts.380437

[25] F. A. Castillo, L. Arias, and J. Cifuentes, "Biomass flame spectroscopy technique to identify wood species through spectral emission during combustion processes," *Measurement*, vol. 202, Art. pp. 115581, 2025. doi:10.1016/j.measurement.2024.115581.

[26] X. Cao, Y. Su, X. Geng et al., "YOLO-SF: YOLO for Fire Segmentation Detection," *IEEE Access*, vol. 11, pp. 111079–111092, 2023. doi: 10.1109/access.2023.3322143.

Botao Ni is currently a Master's student at the School of Information Engineering, Guangdong University of Technology, Guangzhou, China. His research interests include computer vision-based environmental perception, deep learning models for flame segmentation and detection, and multimodal information fusion for intelligent robotic systems in complex scenarios.

Lei Huang is currently a Master's student at the School of Information Engineering, Guangdong University of Technology, Guangzhou, China. His research focuses on the analysis of spectral signatures during combustion processes, machine learning-based identification of combustible materials, and the integration of spectral data with visual perception for enhanced fire characterization.

Ying Xiang is currently a Professor at the School of Information Engineering, Guangdong University of Technology, Guangzhou, China. He has published over 100 research papers in international journals and conferences. His research interests include computer vision and pattern recognition, optical imaging and spectroscopy, machine learning algorithms for visual understanding, and their applications in intelligent perception systems for robotics and industrial automation.

Yan Zhu is currently a Manager at Beijing Qingniao Fire Protection Co., Ltd., Zhuolu County, Hebei, China. His work focuses on the development and deployment of advanced fire protection technologies and equipment, with expertise in fire suppression systems, firefighting robotics, and the practical implementation of intelligent safety solutions in industrial and urban environments.

Lin Li is currently a Manager at Beijing Qingniao Fire Protection Co., Ltd., Zhuolu County, Hebei, China. He has extensive experience in fire safety engineering and has been involved in the development of firefighting robotic systems, particularly in system integration, field testing, and the translation of research innovations into operational firefighting platforms.

Yunfei Zhou received the Ph.D. degree from Huazhong University of Science and Technology, Wuhan, China. He is currently a Lecturer at Guangdong Communication Polytechnic, Guangzhou, China. His research interests include intelligent transportation systems, multi-sensor fusion for environmental perception, and the application of computer vision in traffic monitoring and autonomous navigation.

Jingjing Yang is currently an Associate Professor at the School of Physics and Electronic Information, Guangxi University for Nationalities, Nanning, China. Her research interests include signal processing theory and applications, machine learning for remote sensing data analysis, and the development of intelligent algorithms for environmental monitoring and resource detection.

Hao Tang is currently a Researcher at the Collaborative Innovation Research Institute, Guangdong University of Technology, Heyuan, China. His research interests include robotic perception and control, computer vision for object detection and scene understanding, and multi-sensor integration for autonomous systems in complex environments.