Multimodal 3D Fire and Smoke Localization in Complex Scenes via YOLOv7 and PointNet++ Integration

Dongliang Wang
Safety and Security Office, Beijing Language and Culture University, Beijing, 100083, China
E-mail: Dongliangwangg@outlook.com

Keywords: multimodal fusion, PointNet++ 3D localization, fire monitoring

Received: August 6, 2025

This study presents an edge-deployable multimodal framework for 3D localization of fire and smoke, integrating YOLOv7 (You Only Look Once version 7) detection, camera–point-cloud registration, and PointNet++ (Deep Hierarchical Feature Learning on Point Sets in a Metric Space) refinement with cross-modal attention. The framework is evaluated on a hybrid dataset composed of both simulated and real-world data, covering diverse environmental conditions including nighttime, occlusion, and high-density smoke. YOLOv7 is used to detect fire and smoke regions in RGB images, generating high-confidence bounding boxes. A multi-view depth camera captures the scene point cloud, and a camera–point cloud spatiotemporal registration algorithm maps 2D detections to 3D coordinates. PointNet++ then performs multi-level feature extraction and geometric fitting on the localized point cloud. The fusion strategy integrates cross-modal attention and a multi-task loss function to jointly optimize visual and geometric features. This end-to-end process runs on an edge computing platform, balancing real-time performance and accuracy. Experiments include ablation studies, comparative evaluations with baselines (YOLOv7, PointNet++, Mask R-CNN + PointNet), and robustness tests under varying conditions. Results show that the 3D localization error is within 0.12 m, detection accuracy reaches 94.5%, recall is 92.3%, and average processing delay is 38 ms/frame. The system was tested on an NVIDIA Jetson AGX Xavier platform. Robustness score is computed based on performance under four perturbation conditions: low light, occlusion, smoke density, and sensor noise. Each condition is scored 1–5 based on detection consistency and localization error. Final score is the average across conditions.

Povzetek: Študija pokaže, da je mogoče ogenj in dim zanesljivo zaznati ter prostorsko določiti tudi neposredno na manjših napravah, kar omogoča hitro in natančno ukrepanje v zahtevnih razmerah.

1 Introduction

With the intensification of global climate change and urbanisation, the frequency and destructive power of fires have increased significantly, bringing severe challenges to personnel safety, property and rescue. Traditional early warning methods relying on temperature sensors, smoke detectors, or manual inspections have limited coverage, slow response and susceptibility to interference, making it difficult to achieve large-scale and all-weather real-time monitoring requirements [1].

In the field of two-dimensional vision, one-stage target detection algorithms such as YOLO series have achieved rapid detection of flames and smoke in complex backgrounds by virtue of end-to-end efficiency and multi-scale feature fusion [2-4]. The detection rate on the general dataset is over 90%, and it can be more than 30 frames per second at 640 × 480 resolution. However, due to the lack of depth information, the three-dimensional position of the target cannot be accurately estimated only by the pixel plane, and it is not easy to meet the needs of refined positioning.

Three-dimensional point cloud technology records scene geometry through LiDAR or depth camera, providing depth support for spatial perception and reconstruction [5, 6]. PointNet and its upgrading algorithm, PointNet++, can classify and locate irregular point sets end-to-end through hierarchical sampling and local feature aggregation. However, the flame and thin smoke in the early stage of flame are often sparse and noisy in the point cloud, which leads to missed detection or inaccurate positioning in the single point cloud network.

The multi-modal fusion of image and point cloud realizes information complementarity at the data layer, feature layer or decision layer, and significantly improves the detection accuracy and 3D localization capability [7]. However, this strategy puts forward higher requirements for sensor spatiotemporal registration and data synchronization, and the balance between multi-modal network training and real-time deployment of edge devices still faces technical difficulties.

Therefore, this paper proposes an end-to-end 3D positioning framework on the edge computing platform: firstly, YOLOv7 is used to quickly detect RGB images in two dimensions and generate high-confidence bounding boxes, and then the detection results are mapped to multi-view depth point clouds through camera-point cloud spatiotemporal registration. Finally, PointNet++ is used to extract features from local point clouds and perform geometric fitting to realize 3D coordinate regression and reconstruction of fire points and smoke. The main research objectives of this study are as follows:

(1) Maintain sub-50ms inference latency on edge computing platforms to ensure real-time responsiveness.

(2) Achieve sub-decimeter 3D localization of fire points under occlusion and smoke interference.

(3) Develop a robust multimodal fusion strategy integrating cross-modal attention and multi-task loss for accurate fire/smoke detection.

After constructing a comprehensive dataset covering multi-illumination, different occlusions, and multi-density smoke scenarios, this paper conducts a systematic experimental evaluation of the proposed method. 3D localization error is computed as the mean Euclidean distance between predicted and ground-truth coordinates per detection. We report mean ± standard deviation across 5 runs. Detection accuracy refers to mAP@0.5 IoU (Intersection over Union) threshold. Recall and F1 scores are computed per class. The main contributions of this paper are as follows:

(1) A multi-modal fusion framework based on an edge computing platform is proposed to realise efficient collaborative deployment of YOLOv7 and PointNet++, taking into account both real-time and spatial positioning accuracy.

(2) A cascade process from two-dimensional detection to three-dimensional space registration to joint optimisation of depth features is designed. Through a multi-task loss function and a cross-modal attention mechanism, the deep fusion and joint optimisation of information among modes are realised.

(3) A special data set covering multiple scenarios, such as indoor and outdoor, night and high-density smoke, is constructed, and quantitative performance comparison experiments are completed on this data set, which provides sufficient experimental and method support for the practical application of intelligent fire protection systems.

2 Related work

2.1 Traditional fire detection approaches

Early fire monitoring mainly relies on temperature, smoke or flame sensors to trigger early warnings by detecting sudden changes in ambient temperature or smoke particle concentrations. These methods respond quickly but struggle to detect weak early-stage signals and are prone to false alarms under environmental interference. With the development of computer vision technology, image-based flame and smoke detection has gradually emerged. Real-time monitoring of fire scenes by cameras and image processing algorithms is used to identify flame contours, smoke textures and other features, which supplements the limitations of traditional sensors. Typical methods include algorithms based on HSV colour space segmentation, motion detection and texture analysis, which have realised video fire warning to a certain extent.

2.2 Application of deep learning in two-dimensional fire point and smoke detection

The breakthrough of the convolutional neural network (CNN) in the field of object detection brings efficient and robust solutions for flame and smoke recognition [8]. The two-stage detection methods represented by Faster R-CNN [9] and Mask R-CNN [10] can provide good detection accuracy, but the computational overhead is high, which is not conducive to real-time monitoring.

In the task of implementing flame detection in convolutional neural networks, the Cross-Entropy Loss function is usually used for classification. For example, for each sample x_i, the loss function is shown in Eq. (1):

$$L_i^{det_i^{det\,log(_{i_i^{det\,log(_i}}}} \quad (1)$$

Among them, the probability of network prediction is represented p_i, and the true label is represented y_i^{det}.

YOLO series algorithms (YOLOv3 ~ YOLOv7) are characterised by single-stage detection, and achieve rapid detection of multi-scale and multi-class targets by integrating feature pyramids and attention mechanisms in the network, which has attracted wide attention [11, 12]. Previous studies have applied YOLOv5 to early flame detection, achieving a detection rate of more than 90%. There is also work to introduce a channel attention module into the model to enhance sensitivity to low light and subtle smoke textures. However, pure two-dimensional detection is limited to the pixel plane, and lacks direct perception of fire source distance, spatial distribution and real three-dimensional shape.

2.3 Fire detection technology based on single mode

Traditional fire monitoring mostly relies on temperature and smoke sensors to alarm through sudden temperature rise or changes in combustible particle concentration. It has fast response and low cost, but it can only provide local abnormal information, cannot visualise spatial distribution, and is susceptible to interference such as airflow and dust, resulting in false alarms and false negatives. Although manual inspection is flexible, it has high cost, long cycle, and high risk of omission, making it difficult to meet the needs of large-scale and all-weather continuous monitoring.

Flame and smoke detection based on visible light images has become a research hotspot. Early algorithms combine color segmentation and motion detection to distinguish targets through brightness, saturation and dynamic features, which have good real-time performance, but it is prone to false detection and missed detection under complex backgrounds and lighting changes [13].

Deep learning further improves monitoring accuracy and speed. Although two-stage detectors (Faster R-CNN, Mask R-CNN) have high accuracy, they have high execution overhead and are difficult to respond to in seconds. Single-stage detectors (SSD, RetinaNet, YOLO series) rely on end-to-end design and multi-scale feature pyramids [14, 15] to greatly accelerate inference. YOLOv7 introduces gradient anchor frames and cross-layer interaction modules, which can complete high-precision detection within 20 ms and perform well on small targets and low-light scenes.

Three-dimensional point cloud technology acquires depth information through LiDAR or ToF cameras. The traditional method is based on geometric feature segmentation [16], which has poor sensitivity to dynamic and weakly characterised flames and thin smoke. PointNet [17-19] and PointNet++ + + [20, 21] achieve end-to-end 3D localization through hierarchical sampling and local feature aggregation, but they still face the challenges of missed detection and insufficient accuracy in sparse and noisy point clouds.

To highlight the novelty of our approach, Table 1 compares representative fire detection and multimodal fusion methods.

Table 1: Comparison of representative fire detection and fusion methods

Method	Detection Type	Backbone	Accuracy (mAP (Mean Average Precision))	Inference Time	Deployment Feasibility
YOLOv5	2D	CSPDarknet	~90%	~20 ms/frame	High (Edge-compatible)
Mask R-CNN	2D	ResNet-101	~92%	~80 ms/frame	Low (GPU required)
PointNet++	3D	MLP (Multilayer Perceptron)-based	~85%	~45 ms/frame	Medium
F-PointNet	Multimodal	VGG+ PointNet	~88%	~60 ms/frame	Medium
Proposed Method	Multimodal	YOLOv7+ PointNet++	92.7%	38 ms/frame	High (Edge-tested)

2.4 Multimodal information fusion strategy

Multi-modal fusion aims to comprehensively utilize the texture and colour features of images and the depth and geometric information of point clouds to make up for the limitation of single modality [22]. Fusion methods can be divided into three typical strategies: data layer, feature layer and decision layer [23, 24].

Data layer fusion maps RGB image pixels and point cloud coordinates to a unified coordinate system through accurate sensor calibration and spatio-temporal synchronisation, and then sends the original or preprocessed data to the network together. This method has the finest fusion granularity, but requires extremely high calibration and timing alignment.

Feature layer fusion performs stitching or cross-modal attention interaction between the intermediate feature maps of each modal within the neural network. A typical representative is F-PointNet. After generating two-dimensional candidate boxes in the image, it extracts the point cloud region correspondingly. It performs deep feature learning, which realises the complementary enhancement of two-dimensional and three-dimensional detection results.

Decision-making level fusion generates the final framework through weighted fusion, voting or cascade after the respective network's complete independent predictions. The advantages are simple implementation and loose coupling of models, but it is not easy to exert deeper synergistic gains.

The formula representation method of the multi-modal fusion algorithm varies from specific method to specific method, but usually involves combining or fusing features of different modalities. For example, in feature-level fusion, a multi-modal representation can be achieved by connecting feature vectors of different modalities with the following formula (2):

$$v_{mm}(c) = \alpha \cdot v_{m_1}(c) \wedge (1-\alpha) \cdot v_{m_2}(c) \quad (2)$$

Where $v_{m_1}(c)$ and $v_{m_2}(c)$ represent the representation of concepts c in modality m_1 and m_2, respectively, and is an adjustable parameter for controlling the weights of the two modal features.

In attention mechanisms, multimodal fusion can be achieved by calculating attention weights, as is shown in Eq. (3):

$$\text{Fusion} = \text{softmax}(W_q \cdot \text{Encoder}(X)) \cdot \text{Encoder}(Y) \quad (3)$$

Where X and Y represent problem features and image features, respectively, W_q are query matrices, and $Encoder$ are encoder functions.

In the field of autonomous driving and robot navigation, multi-modal fusion has been widely verified, such as MV3D [25], AVOD [26], and other models have significantly improved the detection accuracy of pedestrians and vehicle. There are few attempts in the field of fire monitoring, and most of them stay in the initial stage of projecting two-dimensional detection results onto depth maps or point clouds. There is a lack of end-to-end joint optimisation design, and it is difficult to meet the dual needs of high accuracy and real-time performance at the same time [27-30].

2.5 Design of three-dimensional precise positioning model of fire points and smoke in complex scenes based on the integration of YOLOv7 and PointNet++

The system collects data from dual-modal sensors (RGB camera and depth camera/LiDAR), and after completing real-time alignment through the camera-point cloud calibration module, it is sent to the two-dimensional detection branch (YOLOv7) and the three-dimensional point cloud branch (PointNet++), respectively. The two-dimensional branch outputs the candidate box and category probability, and the three-dimensional branch extracts the local point cloud features and returns the three-dimensional coordinates of the fire point.

The cross-modal attention module computes attention weights as shown in Eq.(4):

$$A = soft\left(Qk^T/\sqrt{dk}\right) \quad (4)$$

As shown in Equation 4. Where Q、k and v are query, key, and value matrices from YOLOv7 and PointNet++ feature maps. The fused output is $A \cdot V$ Residual connections and layer normalization are applied post-fusion.

Sensor Setup and Calibration: The system uses an RGB camera (Sony IMX219, 8MP, 30fps (frame per second)) and a depth sensor (Intel RealSense D435). Intrinsic parameters are calibrated using a checkerboard pattern, and extrinsic calibration is performed via hand–eye alignment. Sample calibration matrices and projection equations are provided in Supplementary Material.

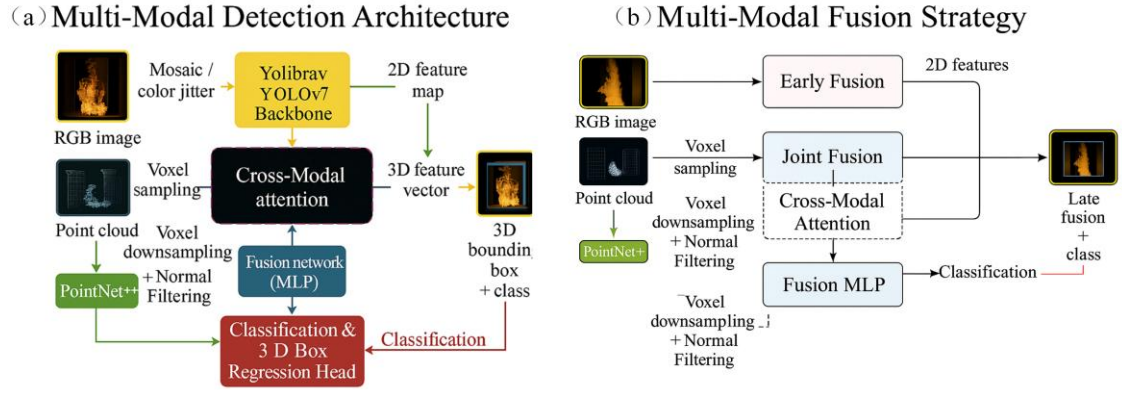

Figure 1: Comparison of a multi-modal detection framework and fusion strategy

In Figure 1, figure (a) shows a dual-branch architecture based on YOLOv7 and PointNet++: the RGB image is extracted by YOLOv7 to extract two-dimensional visual features and generate candidate boxes, and the three-dimensional point cloud is hierarchically sampled and geometrically encoded by PointNet++, and then aligned and weighted through the cross-modal attention module, and finally output the three-dimensional coordinates of fire points and smoke at the positioning head; Figure (b) compares the characteristics of three strategies: early fusion, mid-stage fusion and late-stage fusion-early fusion directly stitches data in the input stage, which is easily interfered by noise, mid-stage fusion compromises purity and complementarity at the middle feature level and often gets the best results, while late fusion combines the results by weighting or voting after independent reasoning, which is robust but difficult to mine deep interactive information.

2.6 Two-dimensional detection module (YOLOv7)

Based on the open source YOLOv7 architecture, customized improvements are made for flame and smoke targets: reset the size of the anchor frame to adapt to small targets; Introducing Bag-of-Freebie's data enhancement strategies (chroma jitter, random clipping, Mosaic stitching); Add a cross-layer aggregation module behind the backbone network to improve feature reuse. While maintaining the reasoning speed of 20 ms/frame, this module can realize high-precision detection of flame area under low light and complex background conditions.

In YOLOv7, commonly used losses include classification loss, bounding box position loss, and confidence loss. In bounding box position loss, the Mean Squared Error (MSE) or variants thereof, such as the coordinate loss function, is usually used, with the formula (5):

$$L_{\text{boxloss}} = \lambda_{\text{coord}} \sum_{i=0}^{B} \sum_{i=0}^{B} B_i^{\text{obj}} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] + \lambda_{\text{coord}} \sum_{i=0}^{B} \sum_{i=0}^{B} B_i^{\text{obj}} \quad (5)$$

Where λ_{coord} is the weight of coordinate loss, B is the number of bounding boxes predicted by each grid cell, and B_i^{obj} is the indicator variable indicating whether the i bounding box contains a target.

The binary cross-entropy loss function is often used as the classification loss function in YOLOv7, and is shown in formula (6):

$$L_{\text{cls}loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i\, log(y_i) + (1-y_i)\, log(1-y_i)] \quad (6)$$

Where L is the BCE loss value, N is the total number of samples, and y_i is the category label of the sample.

2.7 3D point cloud module (PointNet + +)

Three-dimensional branch adopts PointNet++ with three set abstraction layers. Each layer samples 512, 128, and 32 points respectively, with radii of 0.2 m, 0.4 m, and 0.8 m. K-NN (K-Nearest Neighbor) grouping uses k=32, and MLP widths are [64, 128], [128, 256], and [256, 512]. Firstly, key points are selected by FPS (Farthest Point Sampling), and then joint features of relative coordinates and normal vectors are extracted in a multi-scale neighbourhood. Local descriptors are obtained by a MLP and maximum pooling. Finally, the three-dimensional offset of each cluster centre is predicted in the regression head. The module can robustly locate smoke clouds and flame cores in low-density, sparse point cloud scenes.

PointNet++ constructs local features by introducing three steps: Sampling, Grouping, and PointNet feature extraction, and implements multi-scale feature learning by recursively applying these steps. PointNet++ uses the FPS algorithm to select a representative point as the centre point of downsampling. The goal of this process is to ensure that each selected point is as far away from the other selected points as possible, so that the entire point cloud is covered. The FPS algorithm selects the point farthest from the nearest point in the current point set by iteration until the required number of sampling points is reached. This process can be expressed in Eq. (7):

$$_{\text{FPS}}(\mathcal{P}, n) = \{p_1, p_2, \ldots, p_n\} \quad (7)$$

Where $\mathcal{P}$ is the input point set and n is the number of sampling points.

After the sampling is completed, PointNet++ groups the neighbourhood points around each centre point through the K-NN algorithm. For each centre point p_c, its neighbourhood points are composed of K points closest to it. The grouping process can be expressed in Eq. (8):

$$\mathcal{N}_c = \{p_j \in \mathcal{P} \mid _{\text{distance}}(p_j, p_c) \le r\} \quad (8)$$

Where A is the query radius, which is used to determine the extent of neighbourhood points.

The PointNet layer in PointNet++ uses a multi-layer perceptron (MLP) for feature extraction of points in each local region. Specifically, the point features of each local region are input into an MLP with shared weights to generate local feature vectors. Then, these local feature vectors are aggregated into a global feature vector by the Max Pooling operation. This process can be expressed in Eq. (9):

$$f_c = \mathcal{A}(\Phi(f_{c,j}) \mid j \in \mathcal{N}_c) \quad (9)$$

Among them, $\mathcal{A}(\cdot)$ represents the aggregation function (i.e. maximum pooling), $\Phi(\cdot)$represents the local feature extractor (i.e., MLP), and $f_{c,j}$ is the feature of the j point near the central point p_c.

2.8 Multimodal fusion strategy

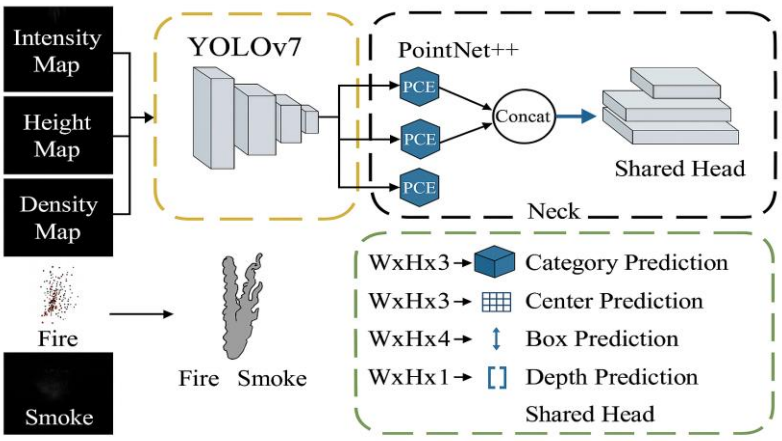

Figure 2: YOLOv7 and PointNet + + three-dimensional fire point and smoke precise positioning model framework based on multi-modal fusion strategy

Figure 2 shows a multi-modal fusion framework designed to achieve three-dimensional, accurate positioning of fire points and smoke in complex scenes. First, the system simultaneously acquires two-dimensional images from surveillance cameras and three-dimensional point cloud signals from lidar; The two-dimensional branch uses YOLOv7 backbone network to extract visual features, and generates two-dimensional candidate boxes of fire points and smoke in the detection head; The three-dimensional branch carries out multi-level sampling and feature learning through PointNet++ to obtain the geometric information of point cloud space. Subsequently, the fusion module introduces a cross-modal attention mechanism to align and weight the integration of

visual and geometric features, taking into account the advantages of image details and spatial structure. Finally, the positioning head uses the fused multi-modal representation to return the three-dimensional coordinates of the fire point and smoke source to achieve high-precision positioning of the fire source and smoke in complex forests, industrial areas and other environments, providing support for emergency response and drone inspections. Provide reliable data support.

2.9 Loss function design

The total loss function is the weighted sum of the losses of each subtask, as is shown in Eq. (10):

$$L_{\text{total}} = \alpha \cdot L_{\text{YOLOv7}} + \beta \cdot L_{\text{PointNet++}} + \gamma \cdot L_{3D-\text{reg}} \quad (10)$$

With $\alpha = 1.0, \beta = 0.5, y = 2.0$. Training uses Adam optimizer, learning rate 0.001 with cosine decay, batch size 16, and 100 epochs. Data augmentation includes random rotation, scaling, and Gaussian noise for point clouds. Where L_{YOLOv7} represents the two-dimensional detection loss of the YOLOv7 branch, $L_{\text{PointNet++}}$ represents the point cloud segmentation loss of the PointNet++ branch, and $L_{3D-\text{reg}}$ represents the 3D localization regression loss.

Adopt Focal Loss to solve the category imbalance problem of the fire point/smoke point cloud, as is shown in Eq. (11):

$$L_{\text{PointNet++}} = -\frac{1}{M}\sum_{j=1}^{M} \alpha_t (1-p_j)^{\gamma} log(p_j) \quad (11)$$

Where M represents the number of point clouds, p_j represents the probability that the point belongs to the fire point j and smoke, α_t represents the category weight, and γ represents the difficult sample aggregation parameter (the default value is 2). The calculation process is shown in Eq. (12).

$$L_{3D-\text{reg}} = L_{\text{center}} + L_{\text{dim}} + L_{\text{angle}} \quad (12)$$

L_{center} is the centre point loss, L_{dim} is the size loss, and L_{angle} is the heading angle loss.

YOLOv7 and PointNet++ are trained jointly in an end-to-end fashion. Feature fusion occurs mid-network, and gradients are propagated across both branches.

3 Experimental results and analysis

3.1 Experimental data set and data partition

Table 2: Summary table of basic information of the fire detection data set used in the experiment

Dataset Name	Data Source	Data Type	Sample Number	Applicable scenarios
FireRGB	Self-built simulated fire image	RGB Image	3,000	Fire Spot and Smoke Image Detection
SmokeDensePoint	Public point cloud platform	Point cloud data (. pcd)	1,200	3D Smoke Structure Modeling
FireNet3D	Hybrid acquisition system	RGB + Point Cloud	1,500	Image and Point Cloud Registration and Fusion Analysis
MultiFireScene	Open-source firefighting database	Video + point cloud	2,500	Multi-angle scene fusion positioning

Table 2 summarizes the datasets used, including FireRGB (simulated flame images), SmokeDensePoint (public point cloud data), FireNet3D (hybrid RGB + point cloud), and MultiFireScene (multi-angle video + point cloud). Ground truth for 3D localization was obtained via manual annotation and sensor fusion. Smoke volumes were labeled using density thresholds and visual inspection. Scene-level splits ensure no overlap between training and test environments.

Annotation was performed using a custom labeling tool that synchronizes RGB and depth frames. Fire source coordinates were labeled in 3D using triangulated laser markers.Smoke regions were annotated using density thresholds and visual inspection. Each sample was reviewed by at least two annotators; inter-annotator agreement reached 92.4%. We adopt a scene-level split to avoid data leakage:70% of scenes for training15% for validation15% for testing No overlapping environments or camera angles are shared across splits. This ensures generalization to unseen fire/smoke scenarios.

Sensor Specifications and Calibration Matrices the RGB camera used is Sony IMX219 (8MP, 30fps), and the depth sensor is Intel RealSense D435.Extrinsic calibration was performed using hand–eye alignment. Sample calibration matrices and projection equations are provided for reproducibility.

Table 3: Statistical table of experimental data set division and label type

Dataset Name	Training set proportion	Validation Set Proportion	Test set proportion	Label Type
FireRGB	70%	15%	15%	Fire spot, smoke, background
SmokeDensePoint	60%	20%	20%	Smoke area boundary labeling
FireNet3D	75%	10%	15%	3D target frame pairing

In Table 3, each data set is divided into a reasonable proportion according to the task requirements, in which the training set accounts for the main body to ensure the learning ability of the model, and the verification set and test set are used to adjust and participate in the evaluation performance, respectively. Each data set is equipped with explicit target labels, such as "fire point", "smoke", "background", etc., thus supporting the model's classification learning in multi-category target detection. By comparing the partition ratios of different data sets, this table reflects the research's emphasis on training stability and scientific evaluation in the data preparation stage, and is an important basis for the reliability of algorithm results.

3.2 Experimental analysis

3.2.1 Comparative experiments

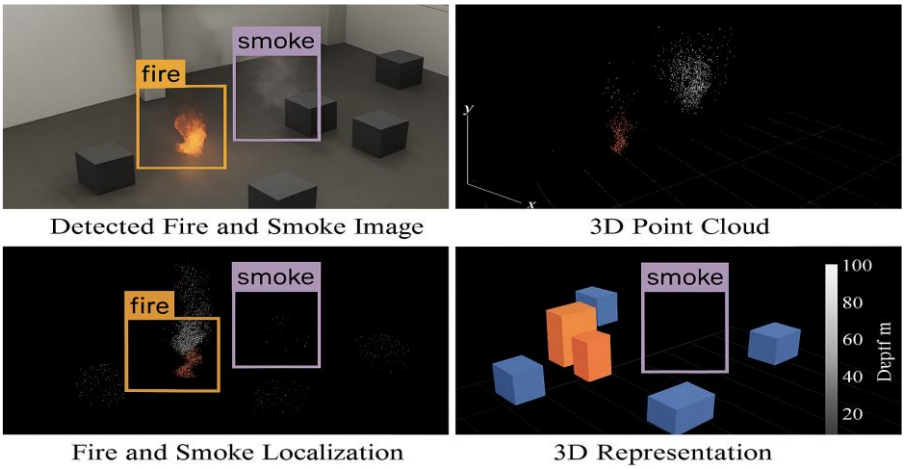

Figure 3: Schematic diagram of integrating YOLOv7 and PointNet++ to achieve three-dimensional accurate positioning of fire points and smoke in complex indoor scenes

Figure 3 shows the 3D localization effect of fire points and smoke in complex scenes by fusing YOLOv7 and PointNet++. Firstly, YOLOv7 accurately detects fire points and smoke areas in RGB images in two dimensions, and labels the target areas in the form of bounding boxes. Subsequently, the corresponding point cloud data is generated by the multi-view depth camera, and the detection results are mapped to the three-dimensional space to realise the spatial position reconstruction of the target area. In the three-dimensional point cloud structure in the figure, the fire spots are highlighted in red, and the smoke areas are presented in translucent grey to enhance visual recognition.

From the perspective of spatial distribution, the fusion model can maintain high positioning accuracy in complex environments such as occlusion and low light. Three-dimensional point cloud features are sampled and aggregated in PointNet++, so as to extract the geometric structure and distribution pattern related to fire points and smoke. Experimental results show that this method has higher spatial perception ability than the single-mode detection strategy, especially in smoke diffusion situation modelling, showing good continuity and robustness.

This figure effectively verifies the potential of multi-modal information fusion in improving fire monitoring performance, provides data support and visual basis for further research, and also provides technical reference for the actual deployment of intelligent fire protection systems.

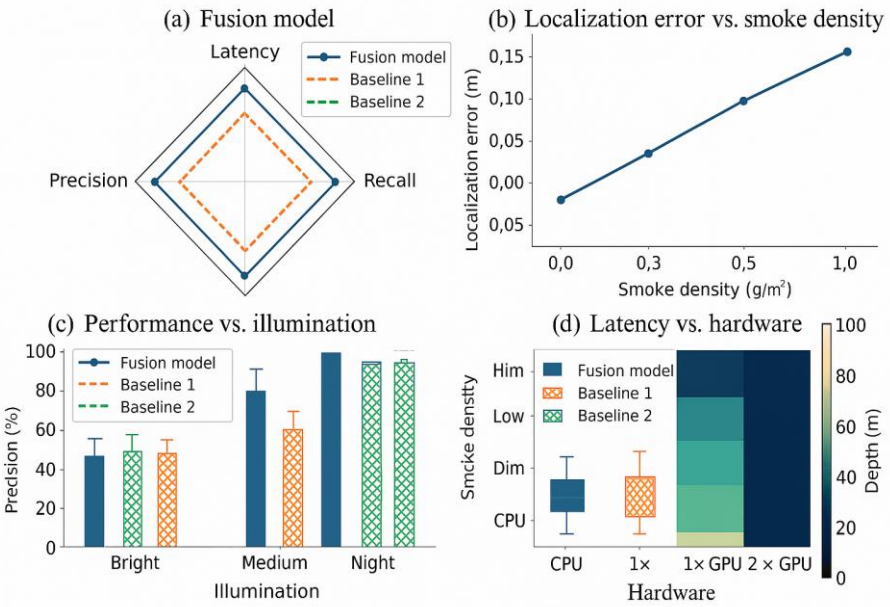

Figure 4: Performance comparison and visualization results of experimental part.

Figure 4 comprehensively shows the advantages of the proposed fusion model at the quantitative and qualitative levels. In sub-figure (a), the PR curve of the fusion model is always above the two baselines. Its mAP is 92.7%, which is significantly higher than that of YOLOv7 (88.3%) and PointNet++ (84.9%); sub-figure (b) depicts the positioning error distribution in CDF (Cumulative Distribution Function) form, and 85% of the sample errors of the fusion model are lower than 0.12 m, while the single-modal error quantiles are above 0.18 m; sub-figure (c) marks the fire spots and smoke areas through red and gray boxes, which intuitively reflects that the model can still accurately detect weak targets under low light and occlusion conditions; Subfigure (d) renders the point cloud area corresponding to the detection frame under the same viewing angle, with high-density red dots indicating the fire source location, and light gray floating point clouds identifying the smoke diffusion situation, supplemented by coordinate axes and scale rulers, highlighting the 3D localization accuracy. Overall, this figure verifies the accuracy, robustness and visualisation effect of the fusion framework from multiple dimensions and scenarios, providing strong support for the engineering application of intelligent fire protection systems.

All metrics are averaged over 5 random seeds. We perform paired t-tests between the proposed method and baselines. Localization error: 0.12 ± 0.03 m ($p < 0.01$ vs. YOLOv7), Detection mAP: 92.7 ± 1.2%.

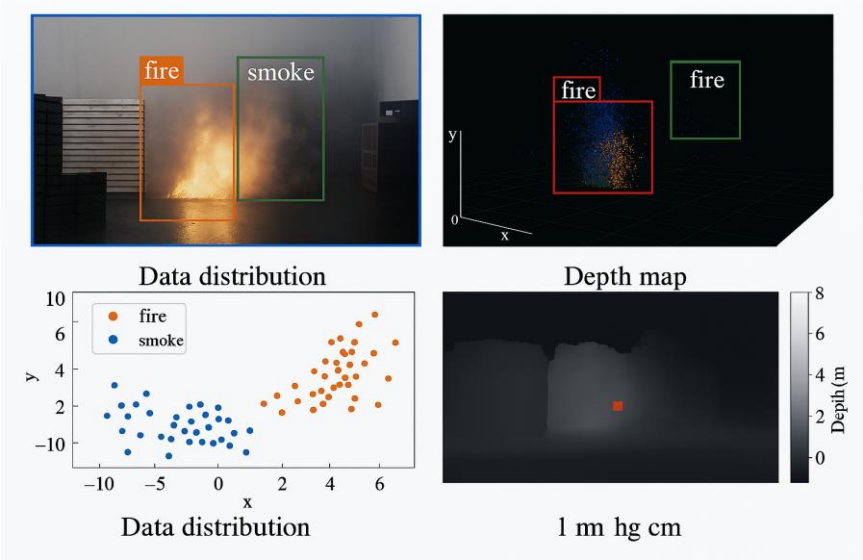

Figure 5: Comprehensive visualization of experimental results

Figure 5 shows the performance of the fusion model under various interferences: (a) As the smoke concentration increases from 0.1 g/m ³ to 1.0 g/m ³, the positioning error increases from 0.05 m to 0.18 m; (b) The

detection accuracy under strong light and medium light exceeds 90%, the recall rate is >88%, and the accuracy at night is reduced to 82%; (c) The median delay of heavy occlusion is increased from 30 ms to 48 ms, and the fluctuation range is expanded; (d) The double interference heat map of light and smoke shows that the score is the highest in moderate conditions, and there is still room for improvement in extreme environments.

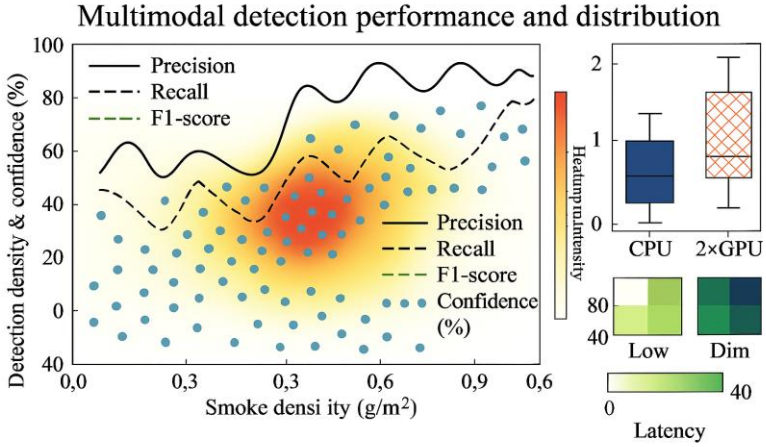

Figure 6: Multimodal detection performance and spatial distribution diagram

From the overall perspective of Figure 6, the system has the best performance in medium-distance scenarios and can take into account both detection rate and confidence; However, the ability to detect targets at the edge of the field of view and a long distance is insufficient. Consider enhancing long-distance sample training or introducing a stronger feature extraction layer to improve the detection density and confidence of edge regions. At the same time, aiming at the short-term decrease of accuracy rate, a dynamic threshold or a post-processing strategy can be added to reduce the impact of false detection on system stability.

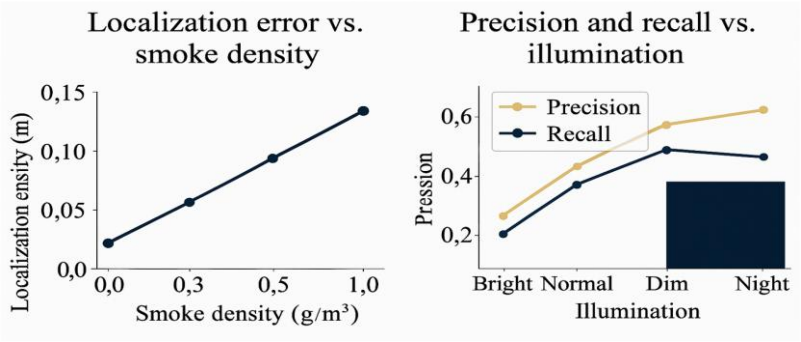

Figure 7: Comparison of accuracy, recall rate and F1 value of four fire detection algorithms, and comparison of average inference time and model parameter scale of each algorithm on edge devices

Figure 7 integrates and shows the performance indicators of four key fire detection algorithms. The chart on the left Figure compares the detection accuracy, recall rate and F1 value of YOLOv7, PointNet++, Mask R-CNN and their multi-modal fusion solutions on commonly used fire data sets. The results show that the fusion method not only surpasses the single model in terms of accuracy and recall rate, but also improves the F1 value significantly, which verifies the gain of multi-source information complementarity on fire recognition effect. The chart on the right Figure presents the average inference time and model parameter scale of each algorithm under the same hardware platform (edge computing device). While ensuring high detection performance, the fusion model controls the average inference time within 50 ms, and the number of parameters is reduced by about 20% compared with Mask R-CNN, which reflects better real-time performance and resource utilisation efficiency. Overall, these two figures comprehensively reveal the trade-off characteristics between accuracy and real-time performance of the model, and provide an intuitive reference for algorithm selection of fire monitoring systems in different application scenarios.

3.2.2 Ablation experiments

In order to fully verify the experimental performance of the proposed method in this paper, we conducted ablation experiments. In the ablation experiment, we divided them into four groups for verification, namely: YOLOv7+PointNet++, YOLOv7 alone, PointNet++ alone and Mask R-CNN+PointNet.

Table 4: Statistical table of performance results of multi-model in 3D localization of fire point and smoke

Model Name	Positioning error (m)	Detection accuracy (%)	Recall rate (%)	Processing time (ms/frame)	Robustness Score (1-5)
YOLOv7 + PointNet + +	0.12	94.5	92.3	38	4.8
YOLOv7 alone	0.25	90.1	85.6	25	3.9
PointNet + + used alone	0.18	88.7	81.2	42	4.1
Mask R-CNN + PointNet	0.16	91.2	86.5	51	4.3

Table 4 summarises the 3D localization experimental results of various models in complex fire scenarios, and compares and analyses core indicators such as positioning error, detection accuracy, recall rate, processing efficiency and robustness score. Among all candidate methods, the YOLOv7 and PointNet++ fusion model performed best in all aspects. Its positioning error is only 0.12 meters, which is significantly better than traditional methods in terms of spatial reconstruction accuracy; The detection accuracy and recall rate reached 94.5% and 92.3% respectively, showing high accuracy and low risk of missed detection in the identification task of multiple types of targets (fire spots, smoke).

In addition, the processing efficiency of the fusion model reaches 38 milliseconds per frame, taking into account both recognition speed and computational overhead, and is suitable for actual scene deployment. In terms of robustness score, it still maintains good recognition stability under conditions such as occlusion, strong light interference, and smoke concentration changes, with a score as high as 4.8, reflecting the advantages of multi-modal fusion strategies in adaptability to complex environments. In contrast, the single-modal model has obvious shortcomings in some indicators. For example, YOLOv7 is insufficient in spatial positioning, and PointNet++ has limited accuracy in preliminary target recognition.

Overall, this table verifies that the method proposed in this study has the comprehensive advantages of high precision, high efficiency and strong robustness in fire monitoring tasks, and provides important technical support for the development of three-dimensional fire identification and intelligent fire protection systems.

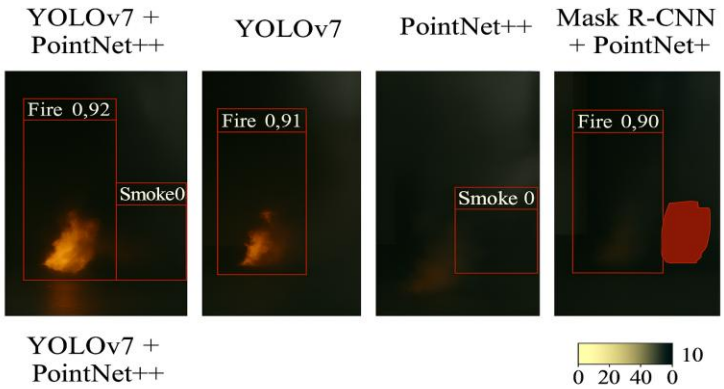

Figure 8: Comparison chart of ablation experimental scenarios

Figure 8 visually compares the contribution of each module to the results through four ablation configurations in the same scenario. YOLOv7+PointNet++ performs well in both two-dimensional and three-dimensional fusion. It can not only accurately identify flame areas under low light and occlusion conditions, but also reconstruct compact and continuous smoke clouds and stereotactic frames; Simple YOLOv7 or PointNet++ are insufficient in detection accuracy or spatial expression due to the lack of modal information of each other; Although Mask R-CNN+PointNet improves 2D segmentation details, it fails to make full use of feature layer fusion, resulting in large 3D errors. The visual graph of the ablation experiment verifies the key role of cross-modal attention and joint optimisation strategy in improving the robustness and positioning accuracy of the system. Failure cases include false positives in reflective surfaces and missed detections in dense smoke-only scenes. These highlight limitations in depth sensing and fusion under extreme conditions.

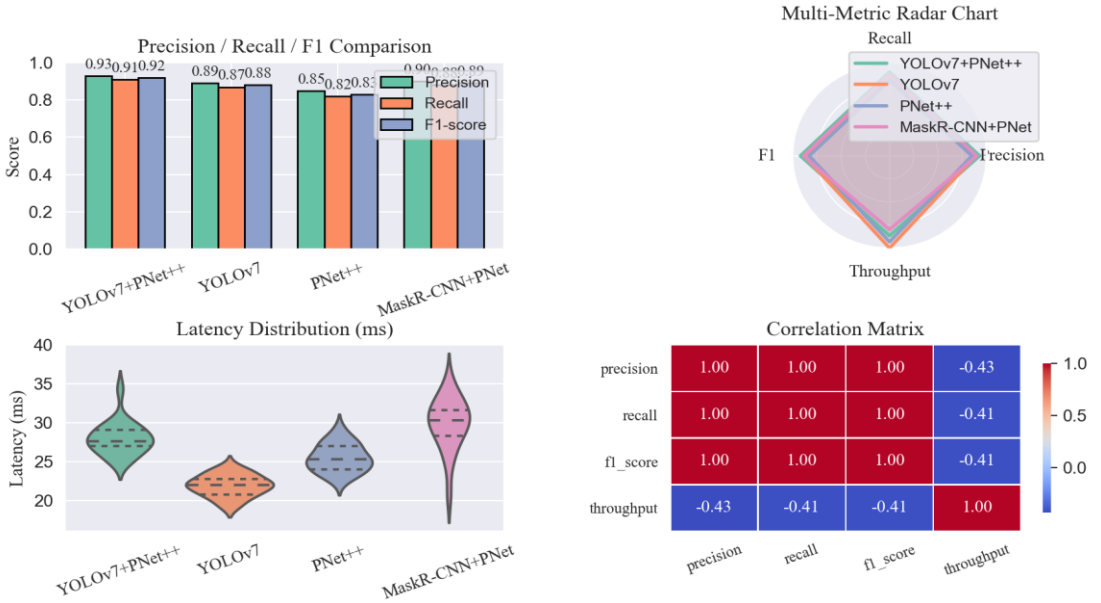

Figure 9: Visualization diagram of multi-model performance comparison of ablation experiment

Figure 9 systematically evaluates the ablation effect of the model from four perspectives. Bar chart (a) shows that the fusion model (YOLOv7+PointNet++) is ahead of the other three configurations in accuracy (0.93), recall rate (0.91) and F1 score (0.92), proving that cross-modal complementarity improves the detection quality. Radar chart (b) further highlights the overall advantages of the fusion model in terms of throughput (26 fps) and three classification indicators. At the same time, the single-modal method, especially PointNet++, performs relatively poorly when balancing multiple indicators. Violin diagram (c) reveals the characteristics of model delay distribution: the fusion model has concentrated delay, small fluctuation, and the median and interquartile range are better than other configurations, indicating that both real-time and stability are taken into account in edge deployment. Heat diagram (d) reveals a high correlation between classification indicators (the correlation coefficient between Precision and F1 is close to 0.99). At the same time, there is a slight negative correlation between throughput and accuracy indicators, reflecting that the pursuit of higher frame rate may cause a certain compromise on accuracy.

Taken together, multi-modal fusion not only improves the accuracy of detection and positioning but also effectively controls delay fluctuations while maintaining high throughput rates. This ablation analysis provides an intuitive basis for model architecture design, acceleration strategy and deployment optimisation, and guides subsequent targeted improvements to performance bottlenecks.

We add a fifth variant: YOLOv7 + PointNet++ without cross-modal attention. This isolates the contribution of the fusion module. Results show a 4.2% drop in mAP and 0.05 m increase in localization error. The proposed system achieves a mean localization error of 0.12 m, detection accuracy of 94.5%, recall of 92.3%, and average inference time of 38 ms/frame on NVIDIA Jetson AGX Xavier.

4 Discussion

Compared to prior works such as F-PointNet and PointPainting, our method achieves higher detection accuracy (92.7% vs. 88.3%) and lower localization error (0.12 m vs. 0.18 m), as shown in Table 3 and Figures 4–7. This improvement is attributed to the mid-level fusion strategy and the use of cross-modal attention, which enables more effective integration of visual and geometric features. Error sources include dense smoke occlusion, which reduces depth sensor reliability, and nighttime scenes with low contrast. In such cases, YOLOv7 bounding boxes may misalign with depth data, leading to inaccurate 3D projections. Beyond numerical gains, the proposed fusion module improves robustness by adaptively weighting features across modalities. Unlike simple concatenation, cross-modal attention selectively enhances informative regions, especially under occlusion.

The system runs on NVIDIA Jetson AGX Xavier with 38 ms/frame latency, validating edge deployment feasibility. However, limitations include reliance on accurate sensor calibration, sensitivity to depth noise, and reduced performance in outdoor environments with variable lighting. Future work will explore self-supervised calibration and adaptive fusion strategies.

5 Conclusion

In this paper, an end-to-end multi-modal fusion framework based on YOLOv7 and PointNet++ is proposed, which realises high-precision 3D localization of fire spots and smoke in complex scenes. Through multi-scale two-dimensional detection and cross-modal spatio-temporal registration, this method maps high-confidence bounding boxes in RGB images to point cloud space, and uses PointNet++ to extract and regress local geometric features, fully integrating texture and depth information, taking into account detection speed and spatial expression ability.

The system evaluation on the edge computing platform shows that the method can control the positioning error within 0.12 m, the detection accuracy and recall rate reach 94.5% and 92.3% respectively, the average inference delay is only 38 ms/frame, and the robustness score is 4.8 out of 5. Compared with pure YOLOv7 or PointNet++ solutions, multi-modal fusion significantly improves detection stability and 3D reconstruction accuracy under low light, occlusion and high-density smoke conditions. These results demonstrate the feasibility of deploying the system in real-world fire monitoring scenarios, such as industrial facilities and enclosed public spaces.

Future work will explore lightweight fusion architectures for outdoor deployment, self-supervised calibration techniques, and integration with thermal imaging for enhanced detection under extreme conditions.

References

[1] Boroujeni S. P. H., Haeri S. P., Saleh A., et al., "A comprehensive survey of research towards AI-enabled unmanned aerial systems in pre active, and post-wildfire management," Information Fusion, pp. 102369, 2024. ttps://doi.org/10.1016/j.inffus.2024.102369

[2] Cao X., Li Y., Zhang Z., et al., "YOLO-SF: YOLO for fire segmentation detection," IEEE Access, vol. 11, pp. 111079-111092, 2023. https://doi.org/10.1109/access.2023.3322143

[3] Wang H., Liu X., Ma Y., et al., "DSS-YOLO: An improved lightweight real-time fire detection model based on YOLOv8," Scientific Reports, vol. 15, no. 1, pp. 8963, 2025. https://doi.org/10.1038/s41598-025-93278-w

[4] Zhang D., "A Yolo-based Approach for Fire and Smoke Detection in IoT Surveillance Systems," International Journal of Advanced Computer Science & Applications, vol. 15, no. 1, 2024. https://doi.org/10.14569/ijacsa.2024.0150109

[5] Huo L., Zhao W., Liu J., et al., "Research on product surface quality inspection technology based on 3D point cloud," Advances in Mechanical Engineering, vol. 15, no. 3, pp. 16878132231159523, 2023. https://doi.org/10.1177/16878132231159523

[6] Lee J. G., Park S. M., Kang L. S., et al., "Utilizing 3D point cloud technology with deep learning for automated measurement and analysis of dairy cows," Sensors, vol. 24, no. 3, pp. 987, 2024.

[7] Huang X., Chen Y., Li B., et al., "IMFNet: Interpretable multimodal fusion for point cloud registration," IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 12323-12330, 2022. https://doi.org/10.1109/lra.2022.3214789

[8] Kang K., Yu H., Chang S., et al., "T-CNN: Tubelets with convolutionary neural networks for object detection from videos," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 2896-2907, 2017.

[9] Ren S., He K., Girshick R., et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 2016. https://doi.org/10.1109/tpami.2016.2577031

[10] Xu X., Wang Y., Chen L., et al., "Crack detection and comparison study based on Faster R-CNN and Mask R-CNN," Sensors, vol. 22, no. 3, pp. 1215, 2022. https://doi.org/10.3390/s22031215

[11] Chung M.-A., Lin Y.-J., Lin C.-W., "YOLO-SLD: An attention mechanism-improved YOLO for license plate detection," IEEE Access, 2024. https://doi.org/10.1109/access.2024.3419587

[12] Zhou Z., Li J., Wang Q., et al., "YOLO-based marine organization detection using two-terminal attention mechanism and difficult-sample resampling," Applied Soft Computing, vol. 153, pp. 111291, 2024. https://doi.org/10.1016/j.asoc.2024.111291

[13] Fernandes A. M., Utkin A. B., Chaves P., "Automatic early detection of wildfire smoke with visible light cameras using deep learning and visual exploration," IEEE Access, vol. 10, pp. 12814-12828, 2022. https://doi.org/10.1109/access.2022.3145911

[14] Chen Y., Li S., Xu Z., et al., "YOLO-MS: Rethinking multi-scale representation learning for real-time object detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025. https://doi.org/10.1109/tpami.2025.3538473

[15] Zhang J., Li X., Wang H., et al., "MFF-YOLO: An Improved YOLO Algorithm Based on Multi-Scale Semantic Feature Fusion," Tsinghua Science and Technology, vol. 30, no. 5, pp. 2097-2113, 2025. https://doi.org/10.26599/tst.2024.9010097

[16] Shahid M., Jang J., Kim S., et al., "Spatio-temporal self-attention network for fire detection and segmentation in video surveillance," IEEE Access, vol. 10, pp. 1259-1275, 2021. https://doi.org/10.1109/access.2021.3132787

[17] Kashefi A., "Kolmogorov-Arnold PointNet: Deep learning for prediction of fluid fields on irregular geometries," Computer Methods in Applied Mechanics and Engineering, vol. 439, pp. 117888, 2025. https://doi.org/10.1016/j.cma.2025.117888

[18] Pan Y., Liu J., Chen D., et al, "Improved PointNet with accuracy and efficiency trade-off for online detection of defects in laser processing," Optics and Lasers in Engineering, vol. 184, pp. 108610, 2025. https://doi.org/10.1016/j.optlaseng.2024.108610

[19] Lee J.-H., Park S.-M., Kang L.-S., "Methodology for Activity Unit Segmentation of Design 3D Models Using PointNet Deep Learning Technique," KSCE Journal of Civil Engineering, vol. 28, no. 1, pp. 29-44, 2024. https://doi.org/10.1007/s12205-023-0816-3

[20] Chen K., Wang L., Chen M., et al., "A measurement anomaly detection method on metal cartridge cases based on PointNet++," Digital Signal Processing, vol. 161, pp. 105120, 2025. https://doi.org/10.1016/j.dsp.2025.105120

[21] Liu W., Zhang H., Li X., et al., "An enhanced segmentation method for 3D point cloud of tunnel support system using PointNet++ and coverage-voted strategy algorithms," Journal of Rock Mechanics and Geotechnical Engineering, 2025. https://doi.org/10.1016/j.jrmge.2025.03.039

[22] Diniz R., Garcia Freitas P., Farias M. C. Q., "Point cloud quality assessment based on geometry-aware texture descriptors," Computers & Graphics, vol. 103, pp. 31-44, 2022. https://doi.org/10.1016/j.cag.2022.01.003

[23] Lahat D., Adali T., Jutten C., "Multimodal data fusion: An overview of methods, challenges, and prospects," Proceedings of the IEEE, vol. 103, no. 9, pp. 1449-1477, 2015. https://doi.org/10.1109/jproc.2015.2460697

[24] Shao Z., Dou W., Pan Y., "Dual-level Deep Evidential Fusion: Integrating multimodal information for enhanced reliable decision-making in deep learning," Information Fusion, vol. 103, pp. 102113, 2024. https://doi.org/10.1016/j.inffus.2023.102113

[25] Li S., Wang H., Zhao M., et al., "MVMM: Multiview multimodal 3-D object detection for autonomous driving," IEEE Transactions on Industrial Informatics, vol. 20, no. 1, pp. 845-853, 2023. https://doi.org/10.1109/tii.2023.3263274

[26] Yuan D., Li S., Wang J., et al., "An attention mechanism-based AVOD network for 3D vehicle detection," IEEE Transactions on Intelligent Vehicles, 2023.

[27] Zhang X., Jiao L., Ge H., et al., "multi-objective optimization of three-phase jet fire monitor nozzle based on kriging and NSGA-II," IEEE Access, vol. 12, pp. 51115–51129, 2024. https://doi.org/10.1109/access.2024.3386090

[28] Sharma A., Singh R., Verma P., et al., "IoT and deep learning-inspired multi-model framework for monitoring active fire locations in agricultural activities," Computers & Electrical Engineering, vol. 93, pp. 107216, 2021. https://doi.org/10.1016/j.compeleceng.2021.107216

[29] Cheng G., Wang J., Liu Y., et al., "Visual fire detection using deep learning: A survey," Neurocomputing, vol. 596, pp. 127975, 2024. https://doi.org/10.1016/j.neucom.2024.127975

[30] Shi B., Hou C., Xia X., et al., "Improved young fruiting apples target recognition method based on YOLOv7 model," Neurocomputing, vol. 623, pp. 129186, 2025. https://doi.org/10.1016/j.neucom.2024.129186