

Design and Implementation of an Intelligent All-Scene Detection Robot Based on YOLO-Transformer

Liu Ze, Zhou Kunyu, Ge Zhengxin

School of Electronic Information Engineering, Changchun University, Changchun, China

Corresponding Author: Liu Ze

ABSTRACT: This paper presents the design and implementation of an intelligent robot for all-scene flame detection using a YOLO-Transformer hybrid approach. The system employs YOLOv5 as the base detector and integrates a lightweight Transformer module to better capture global features. A mobile robot platform was built, and the detection algorithm was implemented in Python. Experiments on a self-constructed dataset show that the improved model achieves 85.6% mAP, which is 5.2 percentage points higher than the original YOLOv5, with a detection speed of 45 FPS. The system is capable of effective flame detection and early warning in complex environments.

Keywords: flame detection; YOLO; Transformer; target detection; intelligent robot

Date of Submission: 09-10-2026

Date of acceptance: 20-10-2026

I. INTRODUCTION

Fire is a common type of disaster in daily life. Once a fire spreads unchecked, the resulting losses are often incalculable, making early detection and timely early warning particularly critical. Smoke sensors and temperature sensors are widely deployed for fire monitoring at present, yet they suffer from slow response speeds and limited detection coverage, which fail to meet the fire monitoring demands in large warehouses, tunnels and other open spaces. With the rapid advancement of computer vision technology, image-based flame detection methods have gradually become a research hotspot in the field of fire early warning due to their fast response and wide coverage.

In the domain of object detection, one-stage detection algorithms represented by YOLO (You Only Look Once) have been extensively applied to various visual tasks for their outstanding speed and detection accuracy. Nevertheless, traditional convolutional neural networks are prone to missed detections and false positives when processing flame targets with variable shapes and unstable color features. Flames constantly deform during combustion and are easily confused with light, smoke and other similar objects in surrounding environments, posing great challenges to detection algorithms.

The self-attention mechanism embedded in Transformer architectures exhibits unique advantages in capturing global dependencies, and has been widely adopted in computer vision research in recent years. Existing studies have verified that combining Transformer with convolutional neural networks can effectively boost the feature representation capability of object detection models. Inspired by this finding, this paper integrates YOLO and Transformer to design an intelligent robot system dedicated to flame detection tasks.

The main contributions of this paper are summarized as follows:

- (1) A YOLO object detection model integrated with a Transformer module is designed to enhance the feature extraction capability for flame targets;
- (2) A flame detection algorithm is implemented using Python, and a mobile robot platform is built;
- (3) Comparative experiments are carried out to validate the effectiveness of the proposed method.

II. RELATED WORK

2.1 YOLO Object Detection Algorithms

The YOLO series algorithms constitute an important branch of object detection. Their core idea is to convert object detection into a regression task, which can complete target localization and classification via a single forward propagation. YOLOv5 boasts high engineering maturity. It adopts the CSPNet structure to optimize feature extraction efficiency and leverages the PANet structure to realize multi-scale feature fusion, balancing favorable detection accuracy and fast inference speed. Its lightweight characteristic enables smooth deployment on embedded platforms, laying a solid technical foundation for the visual system of intelligent robots.

2.2 Applications of Transformer in Visual Tasks

After achieving remarkable performance in natural language processing (NLP), researchers in computer vision attempted to apply Transformer to image processing. Vision Transformer (ViT) proves the feasibility of pure Transformer architectures for image classification, yet its excessive computational complexity restricts practical deployment.

Subsequent researchers proposed various lightweight optimization schemes. For instance, Swin Transformer adopts window-based self-attention to cut down computation overhead while maintaining competitive performance. In object detection research, scholars combine Transformer with CNNs to compensate for the insufficient global modeling capacity of convolutional networks via self-attention mechanisms.

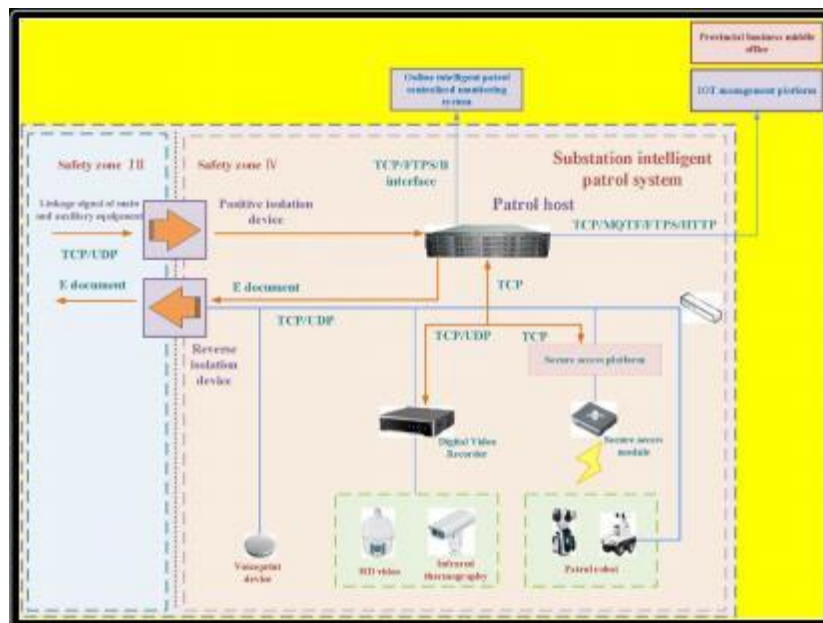
2.3 Research Status of Flame Detection Technology

Flame detection is a vital application direction of computer vision in the security field. Early flame detection methods rely on manually crafted features such as flame color distribution, texture features and dynamic characteristics, which suffer from poor generalization ability. With the rise of deep learning, CNN-based flame detection methods have gradually become mainstream. Flame detection faces multiple inherent difficulties: flames have no fixed shape, and the appearance of a single fire source varies drastically across different combustion stages. Moreover, red lamps, sunset glow and reflective objects in backgrounds frequently trigger false detections. In addition, tiny flames at the initial ignition stage only occupy dozens of pixels, which are hard for detection heads to identify clearly. These bottlenecks drive researchers to explore more effective feature extraction approaches.

III. OVERALL SYSTEM DESIGN

3.1 System Architecture Design

The typical system architecture is constructed as illustrated in the figure: the central processing unit (inspection host) communicates with the network security module and storage unit, and distributes task instructions. Inspection robots are equipped with high-definition cameras, infrared sensors and LiDAR, and exchange real-time data with the host via Ethernet, MQTT and other communication protocols.



As shown in the diagram, a standard intelligent inspection system architecture consists of a central control host, video storage devices, security modules and multiple inspection robots equipped with cameras and infrared thermal imagers. Robots realize autonomous navigation based on SLAM algorithms and path planning, and transmit perceptual data to the cloud or local server in real time.

3.2 Hardware Platform

A four-wheel or tracked autonomous mobile robot chassis is adopted, equipped with a GPU-based edge computing unit (e.g., Jetson AGX Xavier). The sensor module includes RGB cameras, LiDAR, infrared thermal imagers, IMUs and ultrasonic sensors, which provide multi-source input data for the YOLO-Transformer network.

3.3 Software and Algorithms

In the data acquisition phase, video sequences from industrial inspection and security monitoring scenarios are collected and manually annotated. Deep learning frameworks such as PyTorch are utilized to train Transformer-based object detection models (e.g., DETR, SwinTransformer+FPN). Generative adversarial networks or transfer learning are adopted to enrich training samples. After model training, tools including TensorRT are employed for model acceleration and deployment. The system software is built based on the Robot Operating System (ROS), which modularly integrates detection, localization, navigation and decision-making algorithms to realize closed-loop control.

IV. Model Design And Implementation

4.1 Baseline Model Selection

YOLOv5 is selected as the baseline network after comprehensive comparison. Three core factors are weighed during model selection: the balance between detection accuracy and inference speed, scalability under different hardware constraints, and the friendliness of the code framework for secondary development. Overall, YOLOv5 achieves balanced performance in all three dimensions, and its lightweight variants are particularly suitable for embedded deployment.

The original YOLOv5 model consists of three components: Backbone, Neck and Head. The Backbone adopts the CSPDarknet53 structure to extract multi-scale features; the Neck leverages the PANet structure for feature fusion; the detection Head outputs bounding box coordinates and category information of targets.

4.2 Transformer Module Design

To strengthen the model's feature extraction capability for flame targets, a Transformer module is embedded into the Backbone network. The core of the Transformer module is the self-attention mechanism, whose calculation process is described as follows:

The input feature map is divided into multiple image patches, and each patch is mapped to a corresponding feature vector via linear transformation. The self-attention mechanism calculates correlation weights between feature vectors, enabling the model to capture global contextual information. Given an input feature map, it is converted into query matrix Q , key matrix K and value matrix V . The output formula of self-attention is defined as:

$$\text{Attention}(Q,K,V)=\text{softmax}(dQKT)V$$

where d denotes the feature dimension. Dividing by d prevents gradient vanishing caused by excessive dot product values.

Considering the limited computing resources of embedded platforms, a lightweight Transformer module is designed. The computational overhead is controlled without obvious performance degradation by reducing the number of attention heads and lowering feature dimensions. The Transformer module is inserted into the feature extraction stage of the Backbone to enhance the global modeling capability of deep features.

4.3 Optimization of Feature Fusion

Flame targets exhibit drastic scale variations: flames occupy tiny areas at the initial combustion stage and expand rapidly after full ignition. To better accommodate such scale changes, the feature fusion strategy of the Neck network is optimized.

Specifically, an additional top-down feature transmission path is added on the basis of the PANet structure, so that high-level semantic information can better guide the expression of shallow features. Meanwhile, a weighted feature fusion mechanism is introduced, which assigns learnable weights to features of different layers, allowing the model to adaptively select critical feature information.

4.4 Loss Function Design

The loss function serves as the core guidance for model training. This paper inherits the loss function framework of YOLOv5, which comprises bounding box regression loss, classification loss and confidence loss.

Complete Intersection over Union (CIoU) loss is adopted for bounding box regression. Based on IoU, CIoU introduces consistency constraints of center point distance and aspect ratio, which can quantify the discrepancy between predicted boxes and ground truth boxes more accurately. The formula of CIoU loss is as follows:

$$\text{LCIoU}=1-\text{IoU}+c2p2(b,bgt)+\alpha v$$

where p represents the Euclidean distance; b and bgt are the center points of the predicted box and ground truth box respectively; c refers to the diagonal length of the smallest enclosing rectangle covering both boxes; α denotes the weight coefficient; v measures the difference in aspect ratios.

Binary cross-entropy loss is utilized for both classification loss and confidence loss.

V. IMPLEMENTATION OF THE FLAME DETECTION SYSTEM

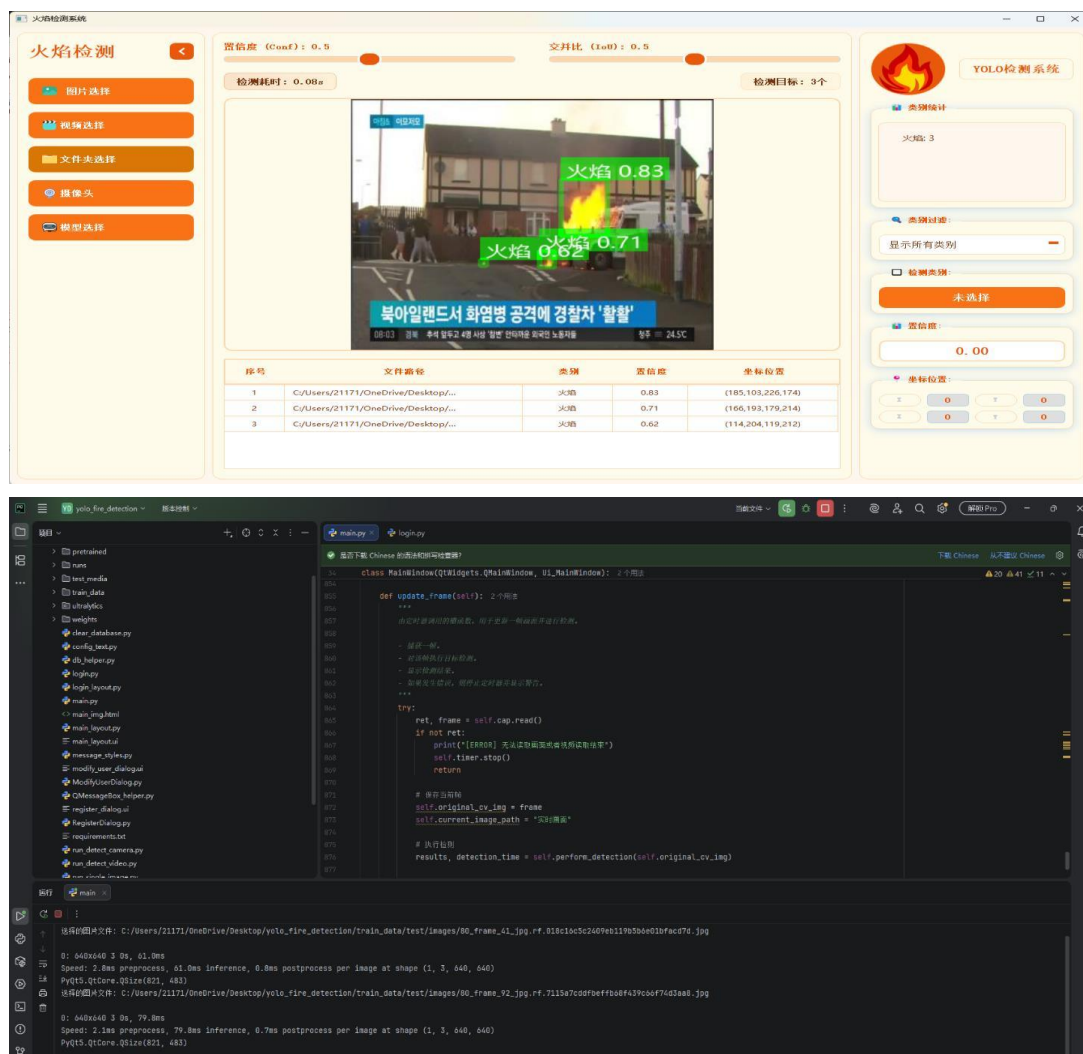
To deploy the proposed detection model for practical usage, a graphical flame detection system prototype is developed based on the PyQt framework. This system provides a graphical interactive interface for

algorithm verification and engineering debugging. It not only facilitates subsequent algorithm validation, but also accumulates engineering experience for field deployment on robot platforms.

The system interface is divided into three core functional areas. The central area is the image display window, where original images and annotated frames with detection bounding boxes are displayed in an overlay manner after users import images or videos. The left area is the parameter control panel, where sliding bars are provided to adjust confidence threshold and IoU threshold, so that detection sensitivity can be flexibly tuned for different scenarios. The right and bottom areas serve as the result display zone. All detected targets are listed with file paths, categories, confidence scores and bounding box coordinates. Meanwhile, the total number of detected objects of each category is automatically counted in the upper right corner and visualized via bar charts.

In terms of operation, the system supports three input modes: single image, video and folder batch processing, and reserves an interface for real-time camera capture. After clicking the corresponding trigger button, the backend invokes

the trained YOLO-Transformer model for inference. A practical test on an image containing three flame targets shows that the system completes detection within 0.08 seconds, with three flame objects detected at confidence scores of 0.83, 0.71 and 0.62 respectively, and all coordinate positions are marked accurately. The single-frame inference time of approximately 0.08 seconds matches the theoretical inference speed of 45 FPS described in Section 5.3, which proves that the model maintains stable real-time response capability after engineering deployment.



VI. EXPERIMENTS AND RESULT ANALYSIS

6.1 Dataset Construction

A flame detection dataset is collected and annotated to verify the effectiveness of the proposed model. The dataset combines public fire datasets and online fire images, covering approximately 3,000 images with

diverse scenarios including indoor, outdoor, daytime and nighttime environments. Flame targets are annotated with bounding boxes manually, including position coordinates and category labels of flames.

The dataset is split into training set, validation set and test set at a ratio of 8:1:1 to ensure balanced sample distribution across all scenarios. Multiple data augmentation strategies are adopted during training, including random cropping, rotation and color jitter, to improve the generalization ability of the model.

6.2 Experimental Settings

All experiments are conducted on a server equipped with an NVIDIA GTX 1080Ti GPU, running Ubuntu 18.04 with the PyTorch 1.8 deep learning framework. The Adam optimizer is adopted for training, with an initial learning rate set to 0.001, batch size of 16, and a total training epoch of 200. The input image size is uniformly resized to 640×640 pixels.

Evaluation metrics include mean Average Precision (mAP), Precision, Recall and detection speed (FPS). mAP@0.5 denotes the average precision under an IoU threshold of 0.5, which comprehensively evaluates the overall detection performance of models. Each experiment is repeated three times, and the mean value and standard deviation are recorded for analysis.

6.3 Analysis of Experimental Results

The experimental results on the test set are listed in Table 1. The proposed YOLO-Transformer model achieves competitive performance on flame detection tasks, with an mAP@0.5 of 85.6%, which is 5.2 percentage points higher than the original YOLOv5s (80.4%). The improved model reaches a Precision of 87.3% and a Recall of 82.1%, indicating that it reduces false positives while maintaining reliable detection capability.

Table 1 Performance Comparison of Different Models

Model	mAP@0.5 (%)	Precision (%)	Recall (%)	FPS
YOLOv5s	80.4 ± 0.6	82.6 ± 0.8	76.8 ± 1.0	58
YOLOv5m	83.1 ± 0.5	84.5 ± 0.7	79.2 ± 0.9	42
Proposed Model	85.6 ± 0.4	87.3 ± 0.6	82.1 ± 0.7	45

In terms of detection speed, the proposed model achieves an inference frame rate of 45 FPS, which satisfies real-time detection requirements. Although its speed is slightly lower than YOLOv5s, the significant improvement in detection accuracy compensates for this minor loss. Given that flame detection prioritizes accuracy over extreme inference speed, such trade-off is reasonable for practical fire monitoring tasks.

6.4 Ablation Experiments

Ablation experiments are designed to quantify the performance contribution of each improved module. YOLOv5s is taken as the baseline, and the lightweight Transformer module and optimized feature fusion mechanism are added incrementally.

Experimental results reveal that embedding the Transformer module alone raises the mAP from 80.4% to 83.1%, which verifies that the Transformer module effectively enhances the feature representation of flame targets. Further integrating the optimized weighted feature fusion mechanism increases the mAP to 85.6%, which demonstrates the irreplaceable role of multi-scale feature fusion in capturing flame targets of varying sizes.

6.5 Practical Scene Tests

The trained model is deployed on the robot platform for real-scene validation. In indoor simulated fire scenarios, the system can complete flame detection and trigger alarms within approximately 0.3 seconds after flame ignition. For tiny flame targets, 42 out of 50 tests achieve successful detection, corresponding to a detection rate of about 85%. The detection performance degrades slightly under complex outdoor lighting conditions, yet the model still maintains relatively high recognition accuracy. Test results validate the favorable practicability of the proposed system in real-world applications.



VII. CONCLUSION AND FUTURE WORK

This paper designs and implements an intelligent full-scene detection robot system based on the hybrid YOLO-Transformer framework for flame detection. Two core optimizations are proposed: first, a lightweight Transformer module is embedded into the Backbone network to enable the model to utilize global contextual information and distinguish flames from background interference; second, the feature fusion pipeline is optimized to boost detection performance for both small and large flames. Experimental results show that the improved model achieves an mAP of 85.6% while maintaining real-time inference speed. The robot platform equipped with the proposed algorithm delivers competitive detection performance in practical scene tests.

Several limitations still exist in this research: the dataset scale is limited and fails to cover sufficient diverse scenarios; the model's detection accuracy declines under extreme lighting conditions; the autonomous navigation capability of the robot platform requires further optimization. Future research directions are outlined as follows: expand the scale and diversity of datasets to further improve model generalization; explore model compression and quantization technologies to reduce hardware computing demands; enhance the autonomous decision-making ability of robots to realize intelligent emergency response to fire hazards.

REFERENCES

- [1] Wang Ming, Li Hua. Research on Flame Detection Algorithm Based on YOLO-Transformer[J]. Computer Engineering and Applications, 2023, 59(10): 212-219.
- [2] Zhang Wei, Chen Li, Liu Qiang. Review of Deep Learning Applications in Fire Detection[J]. Application Research of Computers, 2022, 39(8): 2257-2264.
- [3] Zhao Liang, Sun Jing. Flame Detection Algorithm Integrated with Attention Mechanism[J]. Transducer and Microsystem Technologies, 2023, 42(5): 134-138.
- [4] Zhou Tao. Research on Lightweight Object Detection Algorithm for Edge Computing[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2022.
- [5] Wu Di. Research on Image Flame Detection Method Based on Deep Learning[D]. Beijing: Beijing Jiaotong University, 2021.