forests

MDPI

Article

MCDet: Target-Aware Fusion for RGB-T Fire Detection

Yuezhu Xu, He Wang *, Yuan Bi, Guohao Nie and Xingmei Wang

College of Computer Science and Technology, Harbin Engineering University, 145 Nantong Street, Harbin 150001, China; xuyuezhu@hrbeu.edu.cn (Y.X.); bcxby123@hrbeu.edu.cn (Y.B.); nieguohao@hrbeu.edu.cn (G.N.); wangxingmei@hrbeu.edu.cn (X.W.)
* Correspondence: wanghe991104@hrbeu.edu.cn

Abstract

Forest fire detection is vital for ecological conservation and disaster management. Existing visual detection methods exhibit instability in smoke-obscured or illumination-variable environments. Although multimodal fusion has demonstrated potential, effectively resolving inconsistencies in smoke features across diverse modalities remains a significant challenge. This issue stems from the inherent ambiguity between regions characterized by high temperatures in infrared imagery and those with elevated brightness levels in visible-light imaging systems. In this paper, we propose MCDet, an RGB-T forest fire detection framework incorporating target-aware fusion. To alleviate feature cross-modal ambiguity, we design a Multidimensional Representation Collaborative Fusion module (MRCF), which constructs global feature interactions via a state-space model and enhances local detail perception through deformable convolution. Then, a content-guided attention network (CGAN) is introduced to aggregate multidimensional features by dynamic gating mechanism. Building upon this foundation, the integration of WIoU further suppresses vegetation occlusion and illumination interference on a holistic level, thereby reducing the false detection rate. Evaluated on three forest fire datasets and one pedestrian dataset, MCDet achieves a mean detection accuracy of 77.5%, surpassing advanced methods. This performance makes MCDet a practical solution to enhance early warning system reliability.

Keywords: object detection; forest fire detection; multimodal fusion; multidimensional representation synergy

Academic Editor: José Aranha

Received: 26 May 2025
Revised: 25 June 2025
Accepted: 26 June 2025
Published: 30 June 2025

Citation: Xu, Y.; Wang, H.; Bi, Y.; Nie, G.; Wang, X. MCDet: Target-Aware Fusion for RGB-T Fire Detection. *Forests* **2025**, *16*, 1088. https://doi.org/10.3390/f16071088

1. Introduction

Forest fire detection is a critical challenge in ecological conservation and disaster prevention. It plays a vital role in preserving biodiversity, mitigating carbon emissions, and safeguarding public safety. In recent years, Unmanned Aerial Vehicles (UAVs) equipped with visible-light sensors, integrated with deep learning models, have emerged as a mainstream solution for real-time fire monitoring [1,2]. However, smoke obscures the thermal core features of flames, high-reflectivity regions are frequently misclassified as fire, and nighttime conditions result in diminished target visibility. The perceptual capacity of visible-light systems is severely constrained in scenarios involving smoke occlusion, strong light reflection, and low-light environments.

Multimodal forest fire detection has emerged as a promising solution to enhance detection performance in complex environments, gaining increasing traction in recent research and applications [3,4]. By integrating complementary data from RGB and thermal imaging modalities, visible-light sensors capture visual features such as smoke texture and flame color, while infrared sensors utilize thermal radiation signals to penetrate obscurants

like smoke and identify heat sources under low-visibility conditions. This synergistic fusion enables robust environmental perception, overcoming limitations of single-sensor systems and ensuring reliable all-weather detection capabilities [5,6].

Recent studies indicate that fusing multimodal features can substantially enhance the accuracy of forest fire detection [7]. The degree of enhancement is influenced by both the fusion stage and the specific strategies employed. Studies demonstrate that features extracted independently from single-modal data, such as visible light and infrared, can generate independent detection results. These results are subsequently combined through decision-level fusion strategies [8], including weighted averaging and voting mechanisms, thereby integrating complementary information from multiple modalities [3,5,6,9]. Recent advances in multimodal fusion methodologies have introduced sophisticated strategies to address key challenges such as heterogeneous modality alignment [10,11], modality imbalance [12,13], and representation learning [14]. These approaches leverage attention mechanisms and feature pyramids to improve inter-modal interactions. However, persistent limitations in cross-modal feature interaction efficiency persist: (1) Feature confusion arises between the high-temperature infrared radiation signatures of flame cores and high-brightness regions in visible light domains, compounded by disparities between the semi-transparent visual characteristics of smoke and its weak infrared signatures. (2) Challenging forest environments, characterized by dense vegetation, complex terrain, and dynamic lighting variations, hinder model performance by diverting attention from essential fire-related features, which leads to both missed detections and an increased rate of false alarms.

To this end, this paper proposes a target-aware multimodal forest fire detection framework, MCDet. We first introduce Multidimensional Representation Collaborative Fusion (MRCF) to address feature interference arising from the visual confusion between high-temperature infrared radiation cores of flames and high-brightness regions in visible light, as well as the disparity between the semi-transparent appearance of smoke in visible spectra and its weak infrared signatures. The MRCF employs a discrete State Space Model (SSM) to extract spatial features from sequential images and incorporates a selective scanning mechanism to establish global cross-modal interactions. Integrated with deformable convolutions for enhanced local detail perception, this architecture forms a global–local collaborative fusion framework. Then, to resolve challenges in feature focusing due to vegetation occlusion and terrain interference, we design a content-guided attention network (CGAN). The CGAN module dynamically generates dual-dimensional spatial and channel attention weights through a gating mechanism, enabling adaptive feature aggregation. In addition, we adopt the WIoU loss function, which adjusts the gradient weights of hard samples via a non-monotonic focusing mechanism. It suppresses misleading gradient propagation in high-frequency noisy regions, thereby mitigating localization errors caused by blurred smoke boundaries and flame occlusion. Extensive experiments demonstrate that our MCDet achieves advanced performance across four challenging benchmark datasets. The main contributions of this paper are as follows:

- We propose a novel MRCF, which integrates discrete state-space modeling and global–local cross-modal interactions, effectively resolving visual confusion between flame infrared radiation and visible light interference.
- We propose CGAN, a model that captures fire targets' spatial distribution and high-response channel features using content-guided attention. Its dynamic gating mechanism improves feature discrimination amid vegetation occlusion and terrain interference.
- The WIoU loss function is introduced to significantly improve target focusing in complex scenarios. In addition, two large-scale multimodal forest fire datasets, D-Fire and Fire-dataset, are constructed, enriching the data resources in this field.

The paper is organized as follows: Section 2 reviews object detection and multimodal fire detection methods. Section 3 details the proposed MCDet. Section 4 presents experimental validation through comparative tests, ablation studies, and cross-scenario evaluations. Finally, Section 5 concludes with future research directions.

2. Related Works

2.1. Object Detection

Object detection is a fundamental research area in computer vision, focusing on identifying regions of interest (RoIs) within images and predicting their corresponding categories and spatial locations [15]. Contemporary detection frameworks are broadly divided into two primary categories: single-stage [16–18] and two-stage [19–21] approaches. Two-stage detectors, exemplified by models such as Fast R-CNN [19] and Faster R-CNN [20], employ a sequential process involving region proposal generation followed by detection. In contrast, single-stage detectors unify localization and classification into a single step, eliminating the need for explicit region proposals. The prominent examples include the YOLO series [16,22–24] and EfficientDet [18]. Carion et al. [25] introduced DETR, an end-to-end Transformer-based architecture [26] that significantly advanced object detection performance in both accuracy and real-time processing. DETR disrupted the dominance of conventional single- and two-stage paradigms by pioneering the integration of Transformer architectures into object detection. Transformers enable global modeling of remote feature dependencies but significantly increase the computational pressure of CNN-based methods. However, under challenging conditions, such as complex lighting, occlusions, or limited data availability, unimodal feature representations may prove inadequate [12,27], frequently resulting in missed detections or false positives.

2.2. Multimodal Fusion

Multimodal fusion techniques integrate data from diverse sources to address the perceptual constraints of single-modal systems in complex environments, including visible light, infrared, and thermal imaging. Existing approaches are broadly classified into three categories: convolution-based fusion [28], Transformer-based fusion [29–31], diffusion model [32], and state-space models [33]. Convolution-based methods typically employ channel concatenation or weighted fusion strategies to integrate multimodal information at input or intermediate feature layers. Dual-branch convolutional networks, for instance, extract features separately from visible and infrared modalities while utilizing spatial attention maps to dynamically modulate fusion weights. However, their reliance on local receptive fields limits their ability to model global contextual relationships, increasing susceptibility to cross-modal feature misalignment. Transformer-based approaches leverage self-attention and cross-attention mechanisms to establish global inter-modal interactions. Certain methods [34–36] serialize visible and infrared features to compute cross-modal correlation matrices, enhancing semantic consistency through multi-head attention. Despite their effectiveness, the quadratic computational complexity of Transformers presents efficiency challenges when processing high-resolution wildfire imagery. Recent advances in visual state-space models (VSSMs) utilize state-space equations to model spatial dependencies in image data. By implementing a four-directional scanning strategy and recursive hidden state updates, these models achieve linear computational complexity [37]. Although VSSMs maintain a global receptive field, the effective absence of a global–local collaborative mechanism leads to cross-modal confusion in specific forest fire detection tasks, thereby impeding comprehensive extraction of complementary information.

2.3. Forest Fire Detection

Early unimodal forest fire detection methods primarily relied on handcrafted feature extraction and traditional machine learning techniques, which identified flames and smoke based on their physical properties and image features [38]. For instance, support vector machine (SVM) classifiers were employed to recognize forest fire imagery [39]. Subsequent advances introduced deep learning-based unimodal approaches [40–44], including Faster R-CNN for synthetic smoke detection and attention-enhanced algorithms that learn abstract, enriched feature representations, thereby improving adaptability and generalization. However, unimodal methods rely on single-sensor data, rendering them vulnerable to environmental interference such as lighting variations, smoke occlusion, and terrain obstruction, which limit comprehensive feature acquisition.

In contrast, multimodal forest fire detection methods [45–48] integrate heterogeneous data from multiple sources to exploit complementary intermodal information. For example, Sun et al. [49] achieved precise flame detection and spatial localization through joint modeling of visible and infrared modalities. Similarly, Kim [50] addressed unimodal limitations in representing dynamic flame features by preprocessing infrared and visible images with channel number matching to enable spatial alignment. METAFusion [51] achieves feature fusion by mapping infrared and visible images into a unified meta-feature space, harnessing their distinct texture information by reconstructing fused features into a fusion image. Nonetheless, due to its reliance on convolutional networks, it may be sensitive to image misalignment because of limited local-range feature interaction. Conversely, ICAFusion [52], which is based on Transformer architecture, employs global cross-attention to enhance the semantic discriminability of fusion features. Despite the promise of multimodal approaches, feature fusion frameworks founded solely on either global semantic understanding or local texture characteristics encounter challenges in effectively identifying and localizing flame and smoke targets under occlusion or complex environmental conditions. To address these challenges, this paper introduces MRCF, a method integrating a global–local collaborative fusion framework that thoroughly exploits globally discriminative features and local texture details. This is achieved through global cross-modal interaction via SSM and nuanced local detail perception via deformable convolution.

3. Method

In this section, we first present an overview of the architectural framework of MCDet and then elaborate on its core component, MRCF. Subsequently, we delineate the multimodal detection framework implemented by the CGAN.

3.1. Overall Architecture of MCDet

As shown in Figure 1, MCDet enhances YOLOv5s [22] by integrating multimodal capabilities and comprises four core components: the image input module, the multimodal fusion backbone network, the neck network, and the head network.

The image input module employs Mosaic augmentation to diversify training data and standardize image dimensions, with Letterbox augmentation further addressing size inconsistencies [17].

The backbone network integrates five key elements: the Focus module, CBS module, C3 module, SPPF module, and the novel Multidimensional Representation Collaborative Fusion (MRCF) module. The Focus module reduces computational load and memory usage by merging different areas of the input image into the channel dimension, while preserving information integrity. To optimize GPU computational efficiency, the Focus module is replaced with a 6×6 convolutional layer of equivalent complexity. The embedded features are extracted individually through backbone blocks. Each block is composed of CBS, C3,

and SPPf, as detailed in [22]. The CBS module combines a convolutional layer, batch normalization, and a SiLU activation function. The C3 module incorporates residual-connected bottleneck blocks for efficient hierarchical feature extraction, while the SPPF module enhances multi-scale target detection by aggregating flame and smoke features through sequential max-pooling operations.

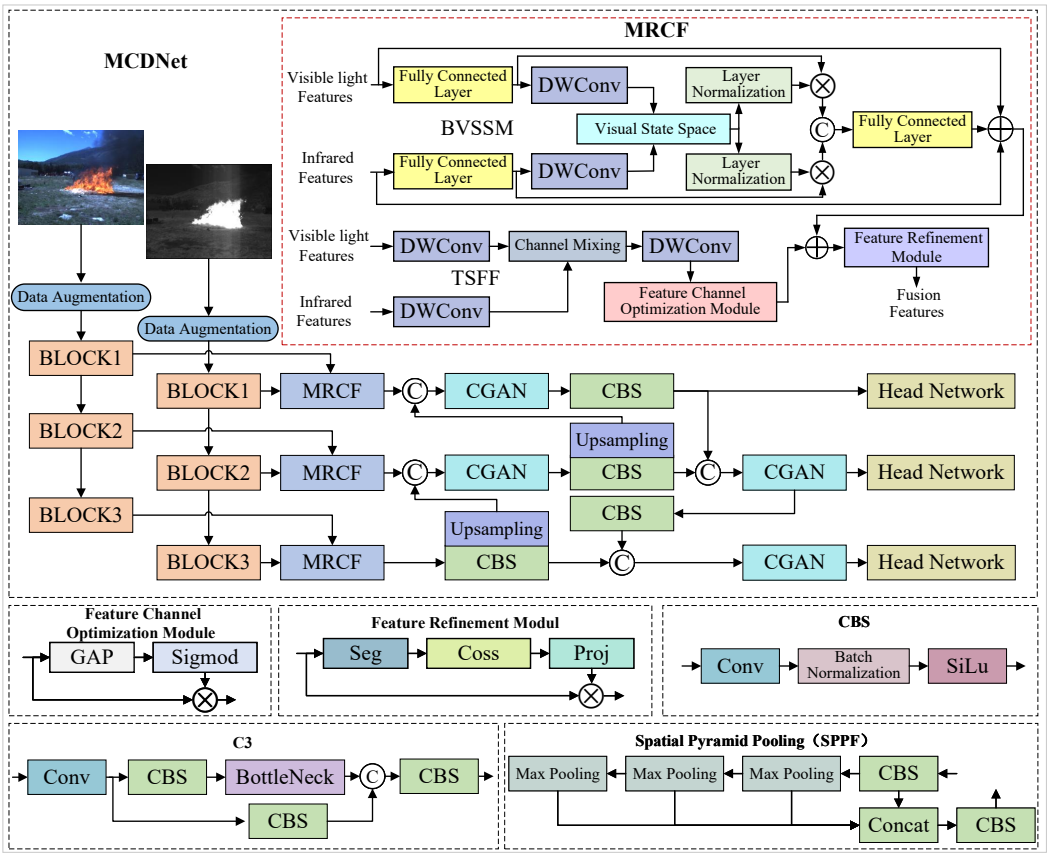

Figure 1. Overall Architecture of MCDet.

The neck network leverages Feature Pyramid Network (FPN) [53] and Path Aggregation Network (PAN) [54] architectures, augmented with the proposed MRCF and CGAN to fuse cross-modal semantics and multi-scale spatial features during downsampling and lateral connections.

The head network utilizes a decoupled architecture with shared feature inputs [22], where classification and regression branches operate independently under a multi-scale anchor-based framework, enabling adaptive matching and task-specific optimization.

3.2. Multidimensional Representation Collaborative Fusion

To investigate potential connections between visible light and infrared modalities, and to develop complementarily enhanced fusion features, the MRCF methodology approaches the fusion task from both global and local perspectives. It employs a bidirectional visual state-space module (BVSSM) to capture cross-modal long-range dependencies, and a Two-property Spectral Feature Fusion Module (TSFF) to improve local feature interactions between modalities. This approach aims to construct cross-modal joint representations that incorporate semantic information and local textures. The structural overview of the MRCF module is depicted in the upper right part of Figure 1.

3.2.1. Bidirectional Visual State Space Module

In the BVSSM module, the visible-light feature F_{vis} and infrared feature F_{ir} extracted from the backbone network are first processed via a Fully Connected Layer and a Depthwise Separable Convolution layer before being fed into the VSSMs [37]. To capture comprehensive information from both modalities, the VSSM employs a cross-scanning module that establishes spatial associations between pixels through a four-directional scanning strategy. However, this strategy introduces significant latency when processing high-resolution feature maps, rendering it unsuitable for real-time forest fire detection. To mitigate this, the VSSM adopts a dual-sequence processing approach. The features F_{vis} and F_{ir} (with dimensions $H \times W \times C$) are flattened along their spatial dimensions and concatenated into a one-dimensional sequence S_{Concat} of size $(2 \times H \times W) \times C$, expressed as

$$S_{\text{Concat}} = \text{Concat}(\text{Flatten}(DWConv(Liner(F_{vis}))), \text{Flatten}(DWConv(Liner(F_{ir})))) \quad (1)$$

An additional sequence S_{Inverse} is generated by reversing the order of the feature sequence S_{Concat}, called reversely scanning).

Then, discrete modeling captures the associations between spatial locations in a sequence to develop a globally aware feature representation. Given that different modalities contribute unequally to fire-related features, directly using VSSM-processed features risks diluting critical attributes such as flame thermal radiation and smoke morphology during fusion, thereby compromising detection capability. Consequently, after VSSM processing, a weighting operation emphasizes these key features, producing refined representations F_{vis} and F_{ir}. These refined features are concatenated along the channel dimension, and through a linear transformation with residual connections, the cross-modal joint representation F_{y_1} is obtained:

$$H_{vis}, H_{ir} = VSSM(S_{Concat}, S_{Inverse}), \quad (2)$$

$$F_{y_1} = Liner(Concat(Norm(H_{vis})\otimes)W_{vis}, (Norm(H_{ir})\otimes)W_{ir}) + F_{vis} + F_{ir}, \quad (3)$$

where W_{vis} and W_{ir} are learnable weights, and $Norm(\cdot)$ denotes layer normalization. The VSSM structure is illustrated in Figure 2.

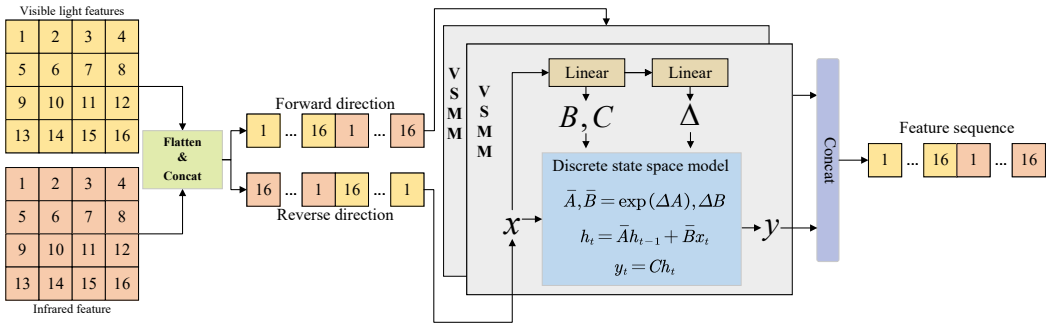

Figure 2. VSSM module structure.

SSMs, such as S4 [55] and Mamba [56], are mathematical frameworks utilized to describe and analyze the dynamics of systems over time. These models predict the subsequent state of the system based on input data, thereby facilitating advanced sequence modeling. Specifically, at time t, an SSM maps input x_t to output y_t via a hidden state h_t:

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t, \quad (4)$$

$$y_t = Ch_t, \quad (5)$$

where $\bar{A}$ is the evolution parameter, and $\bar{B}$ and C are projection parameters derived from input x_t. In deep learning algorithms, image patch sequences are discretized, and S4 employs the Zero-Order Hold (ZOH) technique to convert the continuous parameters into discrete counterparts:

$$\bar{A}, \bar{B} = \exp(\Delta A), \Delta B \tag{6}$$

These equations form a recursive sequence. At each iteration, S4 integrates the hidden state h_{t-1} from the preceding time step with the current input x_t to generate a new hidden state h_t. By processing feature sequences sequentially, VSSM facilitates feature extraction with long-range dependencies. Due to limitations imposed by input order, we employ bidirectional input strategies to extract various global features.

3.2.2. Two-Property Spectral Feature Fusion

In forest environments, the characteristics of flames and smoke vary significantly across different modalities, and their scale changes depending on the distance to the detection equipment. To adaptively capture local features at varying scales and modalities, deformable convolutions are employed to process the visible F_{vis} and infrared features F_{ir} from the backbone network. However, as deformable convolutions are originally designed for single-modality inputs, their inherent advantages may remain underutilized in multimodal scenarios. To address this limitation, a channel mixing strategy is subsequently adopted. Specifically, F_{vis} and F_{ir} are systematically interleaved and grouped along the channel dimension to form C groups, where each group comprises concatenated feature slices of size $H \times W \times 2$. This yields channel-mixed intermediate features G, formally defined as follows:

$$G_i = Concat(F_{vis}[:, :, i], F_{ir}[:, :, i]), i{=}1, \cdots, C. \tag{7}$$

Subsequently, group-wise convolution operations are applied to G_i, facilitating local multimodal interactions and producing the intermediate features $M_i = Conv(G_i)$. To further exploit the intricate interdependencies among multimodal feature channels, the Feature Channel Optimization Module (FCOM) processes M to explicitly model inter-channel dependencies through learnable weights, thereby generating refined features F_{y_2}.

$$F_{y_2} = M \otimes \text{Sigmoid}\Big(\text{GAP}(M)\Big), \tag{3-6}$$

The global average pooling (GAP) extracts global channel statistics, which are transformed into modulation weights W using a sigmoid-activated convolution. These weights are then multiplied element-wise with M to produce the optimized output F_{y_2}.

Initial features F_{y1} (from BVSSM) and F_{y2} (from TSFF) are fused via element-wise addition to generate F_y. To address spatial detail loss from direct fusion, a Feature Refinement Mechanism (FRM) applies dual-path refinement. F_y first enters a segmentation branch predicting a foreground–background mask m through convolutional processing:

$$m = \sigma\big(Conv\big(CBS\big(CBS\big(F_y\big)\big)\big)\big) \tag{8}$$

The channel-wise cosine similarity between m and F_y then computes a correlation vector v:

$$v = Coss\big(m, F_y\big) \tag{9}$$

A projection layer (two convolutional layers) processes v to generate channel weights, normalized via sigmoid σ:

$$s = \sigma(Conv(Conv(v))) \tag{10}$$

Finally, s modulates F_y through element-wise multiplication, yielding the following refined feature:

$$F = s \otimes F_y, \tag{11}$$

This suppresses conflicting information, reduces fusion redundancy, and preserves small-object details. The refined feature F from MRCF is then forwarded to the subsequent prediction module for forest fire detection.

3.3. Content-Guided Attention Network

The detection model may struggle to focus on essential fire features due to two challenges: high similarity in background interference and multi-scale feature conflicts in fused multi-modal data. This dual interference obstructs the effective extraction of critical fire event patterns. Therefore, we propose a multi-modal detection framework based on content-guided attention network. The framework employs content-guided attention to extract spatial distribution patterns and high-response channel features associated with fire targets. A dynamic gating mechanism generates adaptive weights, facilitating multi-dimensional feature aggregation and the construction of discriminative enhanced feature representations. Figure 3 illustrates the structure of the content-guided attention network.

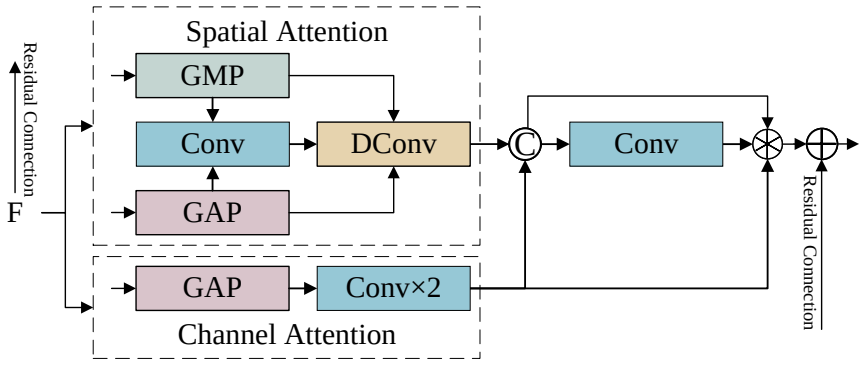

Figure 3. CGAN architecture.

For input feature F, attention is simultaneously computed across spatial and channel dimensions. In the spatial attention branch, the input undergoes global average pooling (GAP) and global maximum pooling (GMP) to extract spatial information. A 3×3 convolution layer then predicts the spatial offset Δp, capturing variations in location-specific importance:

$$\Delta p = Conv\left(F^S_{GMP}, F^S_{GAP}\right), \tag{12}$$

where F^S_{GMP} and F^S_{GAP} denote the GMP and GAP outputs, respectively. This offset Δp guides deformable convolution to produce spatial attention weights W_S:

$$W_S = DeformConv\left(F^S_{GMP}, F^S_{GAP}, \Delta p\right). \tag{13}$$

In the channel attention branch, global average pooling followed by convolutional operations generates channel attention weights W_C, quantifying channel-wise importance:

$$W_C = Conv\left(Max\left(0, Conv\left(F^C_{GAP}\right)\right)\right). \tag{14}$$

The spatial attention W_S and channel attention W_C are fused via a dual-path gating mechanism. Gating weights G are computed as

$$G = \sigma(Conv([W_C, W_S])). \tag{15}$$

In this context, the gating weights G can dynamically balance the important information between spatial and channel features across different channels. When processing fire event features, the model can reasonably allocate attention to spatial and channel information based on the input, effectively avoiding false detection issues caused by background interference.Subsequently, the coarse-grained fused feature W_{CF} is obtained by calculating G with the attention W_S and W_C. The formula is as follows:

$$W_{CF} = G \odot W_C + (1 - G) \odot W_S. \tag{16}$$

Finally, W_{CF} is combined with the original input F through residual connection to produce the output feature map F.

Multimodal Detection Method Based on CGAN

The workflow of the CGAN proceeds as follows. First, a dual-branch backbone network extracts features from visible light and infrared images, producing six distinct scale-specific feature maps: visible light modality features $p3_{vi}$, $p4_{vi}$, and $p5_{vi}$ and infrared modality features $p3_{ir}$, $p4_{ir}$, and $p5_{ir}$. These cross-modal features are fused hierarchically via MRCF modules to generate the joint modal features F_3, F_4, and F_5.

Next, the high-level feature F_5 is upsampled to align with the scale of F_4 and concatenated along the channel dimension. The concatenated features are then input into the CGAN module for processing, outputting the discriminative enhanced feature H_1. Then, H_1 is further concatenated with the low-level feature F_3, processed through the CGAN module, and results in the feature H_2. Feature H_2 is processed along two paths: the first path directly inputs into the prediction head for small-scale target localization and recognition. The second path concatenates H_2 with the convolution-processed features, inputs them into the CGAN module for processing, and generates feature H_3. H_3 is also processed along two branches: the first branch inputs into the prediction head for medium-scale target detection, while the second branch concatenates with the convolution-processed joint feature F_5, is processed through the CGAN module, and is then input into the prediction head for large-scale target detection.

Subsequently, the high-level feature F_5 is upsampled to spatially align with F_4 and concatenated along the channel axis. The combined features are fed into the CGAN module, yielding the discriminative-enhanced feature H_1. Next, H_1 is concatenated with the low-level feature F_3, processed by the CGAN module, and refined into H_2. The feature H_2 undergoes dual processing pathways: the first pathway directs H_2 to the prediction head for small-scale target localization and recognition, while the second pathway concatenates H_2 with convolution-processed intermediate features, processes them via the CGAN module, and generates H_3. Feature H_3 is similarly bifurcated: one branch feeds into the prediction head for medium-scale target detection, and the other merges with the convolution-processed joint feature F_5, undergoes CGAN refinement, and finally enters the prediction head for large-scale target detection.

3.4. Loss Function

We incorporate WIoU loss [57] to address the limitations of the CIoU loss in YOLOv5s [22] for forest fire detection. First, a two-layer distance-aware attention mechanism is employed to construct WIoUv1. This mechanism captures relative spatial relationships between targets using distance information, enabling preliminary boundary distinction for partially occluded fire sources and mitigating redundant detections. The formulation is as follows:

$$L_{WIoUv1} = R_{WIoU} \times L_{IoU} L_{WIoUv1} = R_{WIoU} \times L_{IoU}, \tag{17}$$

$$R_{WIoU} = \exp\left(\frac{(b_x^{g^t} - b_x)^2 + (b_y^{g^t} - b_y)^2}{c_w^2 + c_h^2}\right), \tag{18}$$

$$L_{IoU} = 1 - \text{IoU}, \tag{19}$$

where $b_x^{g^t}$ and $b_y^{g^t}$ denote the ground truth center coordinates, while b_x and b_y represent the predicted box coordinates. c_w and c_h represent the width and height of the minimum enclosing rectangle for the predicted box and the ground truth box, respectively. Subsequently, WIoUv1 is enhanced by integrating a dynamic non-monotonic gradient gain allocation strategy, which leverages the outlier degree β to adaptively prioritize targets of varying scales. By assigning higher gradient gains to small-scale fire targets, this strategy amplifies parameter updates during backpropagation, accelerating feature learning and improving detection performance for small-scale fires. The formulation is defined as

$$L_{WIoU} = r \times L_{WIoUv1}, \tag{20}$$

$$r = \frac{\beta}{\delta \alpha^{\beta - \delta}}, \tag{21}$$

$$\beta = \frac{L_{IoU}^*}{L_{IoU}}, \tag{22}$$

where α and δ are hyperparameters, β quantifies the outlier degree, and r denotes the gradient gain.

4. Experiments

This section begins by introducing the selected dataset and experimental configuration. Subsequently, the experimental setup is publicly released to ensure reproducibility. The methodology is then validated through comparative evaluations across four benchmark datasets, followed by an in-depth analysis of the results. Finally, ablation studies are conducted to isolate the contributions of individual components.

4.1. Datasets

(1) Corsican Fire dataset

The Corsican Fire dataset [58] consists of 635 pairs of real visible and near-infrared fire images acquired at the University of Corsica. All flames and smoke are annotated to facilitate detection accuracy assessment, with dense smoke clusters labeled and sparse smoke excluded. Flames originating from the same source are jointly annotated.

(2) D-Fire dataset

The D-Fire dataset [59] is a publicly accessible resource designed for flame and smoke detection via simulated forest fire scenarios, widely employed to evaluate fire detection models. It comprises 21,527 images categorized by fire event type: 1164 with flames only, 5867 with smoke only, 4658 with both flames and smoke, and 9838 unlabeled background samples. The dataset includes annotations for 14,692 flame instances and 11,865 smoke instances. After removing unlabeled or erroneously labeled images, 11,681 images were retained, allocated into training (9388), validation (1146), and test (1147) sets.

To generate high-quality cross-modal simulation data, this study trains a CUT (Contrastive Unpaired Translation) model [60] for visible-to-infrared image conversion based on the Corsican Fire dataset [58]. The model adopts an unsupervised image translation framework grounded in contrastive learning, with its core mechanism leveraging contrastive loss to capture semantic correspondences between deep features of input (visible) and

generated (infrared) images. Specifically, CUT employs an encoder-generator architecture with shared weights: the encoder extracts multi-level features from input visible images, while the generator reconstructs images in the target domain (infrared) style based on these features. The critical contrastive learning strategy forces the model to precisely preserve the original image's content structures (e.g., fire field morphology, object contours) while transforming styles (e.g., thermal radiation patterns), by maximizing the similarity between features of input image patches and their corresponding generated patches (positive pairs) while minimizing similarity to unrelated patches (negative pairs). This explicit feature correspondence learning not only enhances model interpretability but also fundamentally ensures strong semantic consistency between generated and original images, establishing a theoretical foundation for reliable simulated data. During training, input images were uniformly resized to 512 × 512 pixels, and the model underwent 200 epochs to ensure full convergence, with other hyperparameters following default settings from the original literature [60]. Using the trained CUT model, we performed cross-modal translation on visible images from the D-Fire and Fire-datasets, generating a corresponding infrared simulation dataset with high semantic consistency.

(3) Fire-dataset

The Fire-dataset [61] is a simulated dataset developed for fire detection and emergency response systems research. It encompasses five primary fire scenario categories—forest, urban, industrial, indoor, and urban-rural interface—under diverse environmental conditions, including variations in lighting, weather, and terrain. While the dataset exclusively provides visible light images with annotated flame objects, this work synthesizes infrared modality data using the Corsican Fire bimodal dataset.

(4) LLVIP dataset

Due to the scarcity of multimodal datasets for forest fire scenes, in order to comprehensively assess the model's detection performance, this paper further conducted experiments on the publicly available pedestrian detection dataset LLVIP. The LLVIP dataset [62] is a benchmark resource tailored for multimodal vision tasks in low-light conditions, primarily utilized for developing pedestrian detection methods through infrared and visible light fusion.

4.2. Experimental Details

All experiments are performed on a system running Ubuntu 22.04, utilizing an NVIDIA GTX 3090 GPU for training and testing. Unless otherwise stated, the training configuration included 120 iterations, a batch size of 12, and an initial learning rate of 0.01, optimized with Stochastic Gradient Descent (SGD). The hyperparameters for the YOLOv5s baseline network strictly adhered to the official configurations. To ensure comparability with prior research, evaluation metrics consistent with those employed in existing studies are adopted.

4.3. Comparative Experiment

4.3.1. Comparison on the D-Fire Dataset

We conducted experiments on the D-Fire simulation dataset and compared the proposed MCDet method with advanced approaches both quantitatively and qualitatively. Table 1 summarizes the performance comparison of various models on the dataset. As shown in Table 1, MCDet achieves 2.2% higher precision than YOLOv10s and a 14.7% improvement in mAP@0.5:0.95 over ICAFusion. This performance advantage stems from MCDet's content-guided attention network, which generates spatial importance maps that selectively modulate the feature pyramid to enhance target regions while suppressing background noise. Faster R-CNN's anchor-based selection mechanism demonstrates limitations

in handling small fire targets and occlusions, constrained by predefined anchor ratios and pooling-layer quantization errors. MCDet addresses these limitations with a 6.8% precision advantage. While YOLOv11s employs dynamic anchor matching to address class imbalance between flame and smoke categories, achieving strong performance in single-modality detection, its reliance on unimodal data limits feature representation in complex scenarios, resulting in inferior performance compared to MCDet. Among multimodal approaches, ICAFusion's iterative cross-attention mechanism introduces feature redundancy during multimodal interaction, explaining MCDet's superior performance with a 13.6% higher mAP@0.5. For other comparison algorithms, we use their default parameter settings and select the best results from multiple experiments. The results of MCDet are averaged over 10 experiments, with standard deviation values for precision, recall, mAP@0.5, and mAP@0.5:0.95 being 5.57 $\times$ 10^{-4}, 5.22 $\times$ 10^{-4}, 6.68 $\times$ 10^{-5}, and 2.90 $\times$ 10^{-4}, respectively.

Table 1. Results on the D-Fire dataset. The optimal value is indicated in boldface.

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95
FasterRCNN [20]	74.5	71.6	76.0	46.0
YOLOv5s [22]	78.3	71.5	77.9	45.9
YOLOv10s [23]	79.1	72.3	77.5	45.4
YOLOv11s [24]	80.4	**73.3**	79.3	47.2
IEGOD [63]	74.1	65.9	76.5	44.8
DETR [25]	72.3	68.7	75.2	43.5
DINO [64]	76.4	70.2	78.2	46.2
CalNet [65]	74.8	66.7	72.0	35.2
ICAFusion [52]	76.5	60.8	66.5	34.1
METAFusion [51]	80.1	73.0	76.5	44.9
ProbEn [14]	75.6	67.5	77.3	45.1
MCDet	**81.3**	72.7	**80.1**	**48.8**

Figure 4 illustrates a comparison of detection outcomes on the D-Fire simulation dataset. In the first and second columns, smoke and tree interference result in missed flame detections by the YOLOv5s model. The third column demonstrates that feature similarities between clouds and smoke lead to the failure of YOLOv5s in detecting smoke. In contrast, the MCDet method successfully identifies both flames and smoke undetected by YOLOv5s under complex interference conditions, showcasing its robustness.

Figure 4. Comparison of detection results of D-Fire dataset. The cyan arrows indicate the classification and confidence.

4.3.2. Comparison on the Corsican Fire Dataset

The Corsican Fire dataset comprises two target classes: smoke and flames. Primarily composed of low-light scenarios, the dataset features black smoke that is prone to blending

with background elements due to low contrast. Table 2 summarizes the benchmarking results of various models on this dataset. The low-contrast nature of black smoke in dim environments renders its edge features indistinct, while the high brightness of flame regions causes attention shifts during feature extraction, reducing model sensitivity to smoke targets. Consequently, smoke detection recall decreases significantly, resulting in a combined recall rate of just 51.8% for both fire and smoke. Furthermore, the dataset lacks occlusion cases, and flames' relatively larger size enhances their distinguishability from low-light backgrounds. Notably, the MCDet model demonstrates superior precision in flame detection, achieving 84.4% precision for both fire and smoke, outperforming other models substantially. The MDCNet results of standard deviation values for precision, recall, mAP@0.5, and mAP@0.5:0.95 being 6.57 $\times$ 10^{-4}, 6.99 $\times$ 10^{-4}, 4.25 $\times$ 10^{-2}, and 6.74 $\times$ 10^{-4}, respectively.

Table 2. Results on the Corsican Fire dataset. The optimal value is indicated in boldface.

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv5s [22]	80.6	**55.4**	65.0	35.4
YOLOv11s [24]	81.3	54.8	66.1	36.7
CalNet [65]	72.8	46.7	61.2	31.0
ICAFusion [52]	74.0	45.8	55.6	28.7
METAFusion [51]	78.1	51.2	62.4	32.1
ProbEn [14]	73.3	50.7	64.3	33.5
MCDet	**84.4**	51.8	**69.1**	**42.7**

Figure 5 depicts a comparative analysis of detection performance on the Corsican Fire dataset. YOLOv5s exhibits notable limitations in detecting low-contrast targets. In the first column, the color characteristics of the black smoke closely resemble the background, causing the feature extraction network to struggle with distinguishing smoke features, resulting in missed detections. Examination of the second and third columns reveals that while YOLOv5s can partially identify low-contrast targets, its predicted bounding boxes display significant positional offsets. However, after integrating the content-guided attention network, MCDet effectively enhances edge and contour features of low-contrast smoke targets, thereby significantly improving both detection accuracy and localization precision under low-contrast conditions.

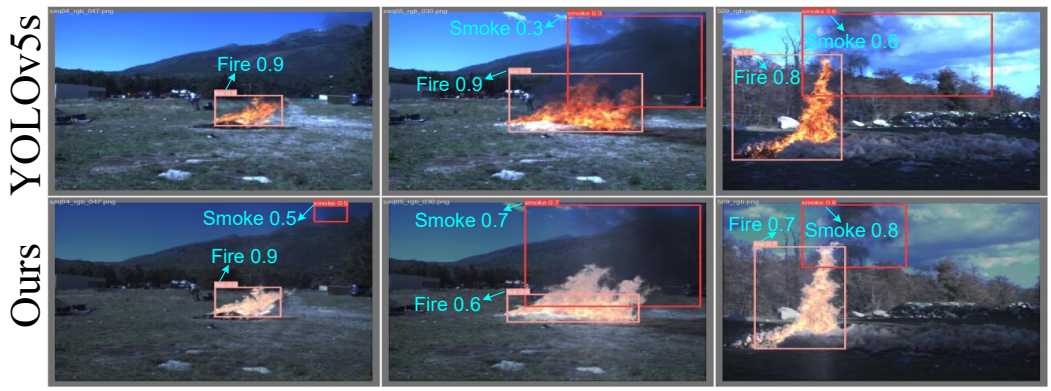

Figure 5. Comparison of detection results of Corsican Fire dataset. The cyan arrows indicate the classification and confidence.

4.3.3. Comparison on the Fire-Dataset

Table 3 summarizes the performance of various models on the Fire-dataset. The MCDet framework, employing a content-guided attention network, effectively reduces

smoke occlusion interference and achieves superior scores across multiple evaluation metrics compared to baseline models. YOLOv11s demonstrates a notable advantage in recall performance due to its optimized target anchor matching strategy; however, its robustness is compromised in fire scenarios with substantial smoke interference, resulting in lower mAP@0.5 and mAP@0.5:0.95 scores compared to MCDet. While the METAFusion model integrates a meta-learning-assisted fusion strategy to address modality discrepancies, it encounters delayed feature responses under significant flame scale variations. MCDet surpasses METAFusion by a margin of 10.6% in the mAP@0.5:0.95 metric. The MDCNet results of standard deviation values for precision, recall, mAP@0.5, and mAP@0.5:0.95 are 4.37×10^{-4}, 5.51×10^{-4}, 4.40×10^{-3}, and 6.58×10^{-4}, respectively.

Table 3. Results on the Fire-dataset. The optimal value is indicated in boldface.

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv5s [22]	62.9	44.4	48.3	21.2
YOLOv11s [24]	64.2	45.5	57.8	27.8
CalNet [65]	59.5	40.1	46.5	19.3
ICAFusion [52]	59.5	33.5	41.3	17.0
METAFusion [51]	64.0	45.2	53.5	23.8
ProbEn [14]	62.3	44.0	54.4	25.7
MCDet	**66.7**	**46.9**	**59.1**	**28.1**

Figure 6 displays a comparison of detection results on the Fire simulation dataset. In the first column, YOLOv5s exhibits evident missed-detection instances. The second and third columns demonstrate that while YOLOv5s successfully identifies the core flame region, the predicted bounding box centers are significantly offset, and the edge coverage exhibits substantial discrepancies compared to the ground truth. This limitation is primarily attributed to YOLOv5s' inability to effectively integrate shallow texture details with deep semantic information in its prediction head, resulting in reduced accuracy when localizing flame boundaries. By implementing a content-guided attention network, flame-edge localization capabilities are markedly improved, thereby enhancing the model's overall detection performance.

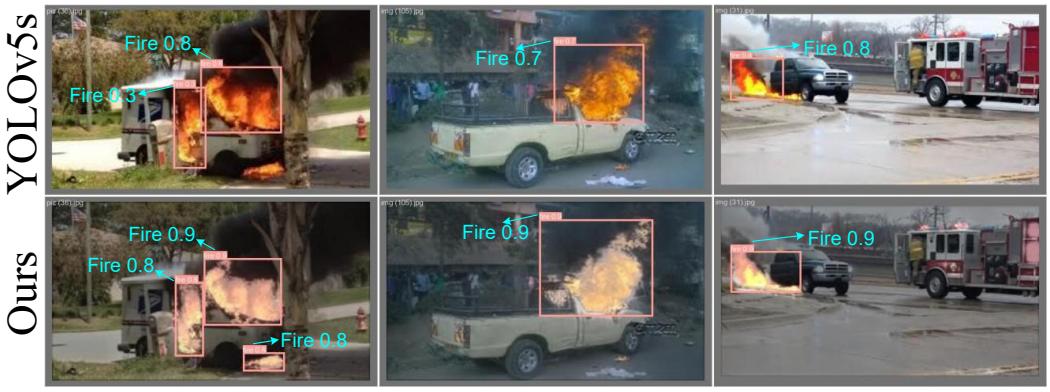

Figure 6. Comparison of detection results of the Fire dataset. The cyan arrows indicate the classification and confidence.

4.3.4. Comparison on the LLVIP Dataset

Table 4 compares the detection performance of representative models on the LLVIP dataset. MCDet demonstrates superior performance under the AP@0.5:0.95 metric. The integration of an attention mechanism enables MCDet to enhance target localization accuracy in complex scenarios, particularly in addressing common challenges of infrared

imagery, such as low contrast and occlusion. While MCDet is marginally outperformed by Fusion-Mamba on the AP@0.5 metric, it maintains superior performance in more stringent comprehensive evaluations, suggesting enhanced robustness and multi-scale adaptability. The MDCNet results of standard deviation values for mAP@0.5, and mAP@0.5:0.95 being 3.14×10^{-2}, and 6.87×10^{-4}, respectively.

Table 4. Results on the LLVIP dataset. The optimal value is indicated in boldface.

Model	mAP@0.5	mAP@0.5:0.95
DeformableDETR [66]	88.7	45.5
FasterRCNN [20]	90.1	49.2
DINO [64]	90.5	52.3
YOLOv5s [22]	90.8	52.7
TIRDet [67]	96.3	64.2
PoolFuser [68]	80.3	38.4
ProbEn [14]	93.4	51.5
MetaFusion [51]	91.0	56.9
Fusion-Mamba [69]	96.8	62.8
DM-Fusion [70]	88.1	53.1
CAMF [28]	89.0	55.6
CSAA [71]	94.3	59.2
Diff-IF [72]	93.3	59.5
DDFM [32]	91.5	58.0
GAFF [73]	94.0	55.8
DIVFusion [74]	89.8	52.0
MCDet	**96.9**	**69.8**

4.3.5. Ablation Experiments

In this section, we first conduct ablation studies on individual components of MCDNet. Then we investigate the impact of the dual branches in MRCF. Finally, critical parameters of the loss function are examined.

4.3.6. Ablation on MCDet

Table 5 illustrates the impact of the content-guided attention network (CGAN) and loss function selection on model performance. The introduction of CGAN enhances texture feature extraction at smoke spreading edges, yielding a 0.9% improvement in smoke detection performance (mAP@0.5:0.95). Analysis of loss functions reveals that incorporating the Efficient Intersection over Union (EIoU) loss introduces an aspect ratio constraint, mitigating flame bounding box deformation and improving flame mAP@0.5:0.95 by 1%. In contrast, the static angle penalty of the Structured Intersection over Union (SIoU) loss increases smoke precision by 1.7% but reduces flame mAP@0.5:0.95 by 0.8%, as its rigid shape constraints conflict with the irregular morphology of flames. Notably, the Wise Intersection over Union (WIoU) loss employs a dynamic weighting strategy that prioritizes low-brightness flame samples, enhancing flame precision by 0.6% and smoke mAP@0.5:0.95 by 2.5%. These performance variations arise from intrinsic differences between smoke and flame characteristics. Smoke detection relies on spatial continuity features and regular geometric spreading patterns, which MCDet's spatial attention module optimizes by enforcing boundary aspect ratio consistency via EIoU. Conversely, flames exhibit irregular edges, high brightness gradients, and localized highlights. WIoU's dynamic focus adaptively strengthens gradient responses in high-brightness flame regions, while MCDet's channel-spatial fusion enhances saliency detection in bright areas. Critically, SIoU's angle penalty underperforms WIoU in recall, mAP@0.5, and mAP@0.5:0.95 metrics due to its over-constraint on the stochastic shapes of both smoke and flames. Figure 7 illustrates the

variation curves of precision, recall, mAP@0.5, and mAP@0.5:0.95 for MRCF and MCDet models during training.

Table 5. The impact of components on MCDet performance. The optimal value under each category is indicated in boldface.

Model	Class	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv5s + MRCF	All	80.9	73.1	78.9	47.5
	Smoke	87.3	79.5	84.7	54.6
	Fire	74.5	66.7	73.1	40.4
YOLOv5s + MRCF + CGAN	All	81.3	73.4	78.6	48.0
	Smoke	87.8	80.0	84.2	55.5
	Fire	74.8	66.8	73.0	40.5
YOLOv5s + MRCF + CGAN + EIOU	All	81.5	75.3	79.1	48.8
	Smoke	88.2	81.5	84.9	56.2
	Fire	74.8	**69.1**	**73.3**	**41.4**
YOLOv5s + MRCF + CGAN + SIOU	All	**82.1**	74.2	78.3	47.2
	Smoke	**89.0**	81.0	84.0	54.8
	Fire	**75.2**	67.4	72.6	39.6
YOLOv5s + MRCF + CGAN + WIOU(MCDet)	All	81.3	**76.6**	**80.1**	**48.8**
	Smoke	87.5	**84.1**	**87.5**	**57.1**
	Fire	75.1	69.0	73.1	40.6

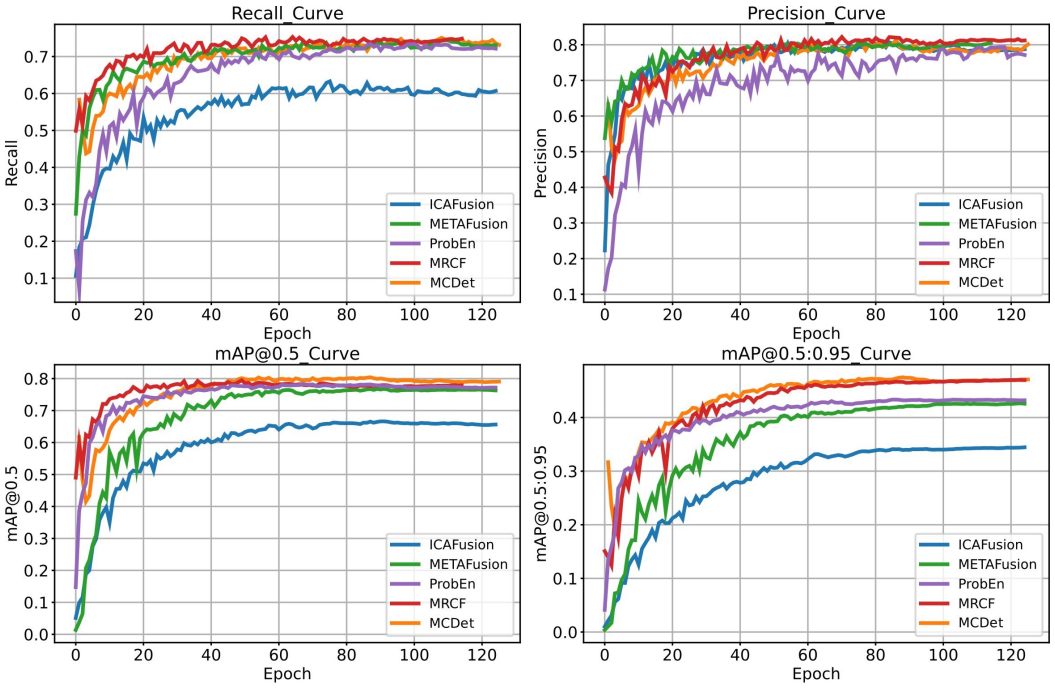

Figure 7. Performance comparison plot of the training process.

Table 6 details the efficiency of our method under various configurations. The basic bimodal detection approach using feature-level direct addition (Addition) exhibits the smallest parameter count of 14M and an inference speed of 7.6 ms. However, the introduction of the MRCF incurs the most significant inference latency, with the time surging to 20.8 ms. This is primarily attributed to the substantial increase in GFLOPs caused by its complex cross-modal feature interaction operations. Subsequently, adding the CGAN module only slightly increased the inference time to 22.2 ms. Following the introduction of the WIOU loss function, the online inference overhead remains manageable. The MRCF module is crucial for achieving the performance leap but also constitutes the primary

computational bottleneck; in contrast, CGAN and WIOU further enhance model accuracy or robustness with acceptable efficiency costs or zero overhead.

Table 6. The impact of components on detection efficiency.

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95	GFLOPs	Parameters	Time
YOLOv5s + Addition	76.8	69.8	74.3	39.5	28.2	14.0 M	7.6 ms
Baseline + MRCF	80.9	73.1	78.9	47.5	82.4	15.5 M	20.8 ms
Baseline + MRCF + CGAN	81.3	73.4	78.6	48.0	85.0	16.0 M	22.2 ms
Baseline + MRCF + CGAN + WIOU	81.3	76.6	80.1	48.8	85.0	16.0 M	22.2 ms

4.3.7. Ablation on MRCF

Table 7 demonstrates the performance impact of the MRCF module. To eliminate confounding effects from multimodal inputs, a bimodal YOLOv5s baseline was employed for ablation studies. In the modality fusion stage, a naive element-wise addition approach (denoted as "addition") was initially tested. As shown in the table, this method produced suboptimal results, as the inherent disparities between visible and infrared features caused additive noise amplification and redundant feature propagation, thereby degrading model performance. Subsequent implementation of the TSFF framework with deformable convolutions improved irregular flame shape representation, yielding measurable gains: +1.9% recall, +2.8% mAP@0.5, and +5.8% mAP@0.5:0.95, particularly for smoke detection. The BVSSM module further enhanced performance through long-range cross-modality dependency modeling, achieving comparable results to TSFF while demonstrating superior global context integration but limited local feature capture. Ultimately, the MRCF module delivered the most substantial improvements—4.1% precision, 4.6% mAP@0.5, and 9% mAP@0.5:0.95 gains—confirming its robustness in complex forest environments through adaptive feature fusion and enhanced generalization capacity.

Table 7. The impact of MRCF on model performance. The optimal value under each category is indicated in boldface.

Model	Class	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Multimodal YOLOv5s + Addition	All	76.8	69.8	74.3	39.5
	Smoke	83.9	76.1	79.0	46.2
	Fire	69.7	63.5	69.6	32.8
Multimodal YOLOv5s + TSFF	All	78.8	71.7	77.1	45.3
	Smoke	85.9	78.2	81.7	52.7
	Fire	71.7	65.2	72.5	37.9
Multimodal YOLOv5s + BVSSM	All	79.5	72.3	77.5	46.1
	Smoke	86.4	77.6	83.6	53.4
	Fire	72.6	**67.0**	71.4	38.8
Multimodal YOLOv5s + MRCF	All	**80.9**	**73.1**	**78.9**	**48.5**
	Smoke	**87.3**	**79.5**	**84.7**	**55.6**
	Fire	**74.5**	66.7	**73.1**	**41.4**

Visible light modalities have advantages in characterizing the texture details of flames and smoke, while infrared modalities are sensitive to thermal radiation, capable of penetrating environmental obstructions to capture the thermal characteristics of obscured fire sources. Different fusion strategies create variations in cross-modal joint feature representations. To verify the capability of existing fusion strategies in constructing cross-modal joint representations and to visually demonstrate the impact of different fusion mechanisms on multimodal information interaction, this section compares the results of different fusion methods (Figure 8).

The additive fusion retains the most in other noisy non-fire backgrounds. The TSFF method utilizes deformable convolution to deeply integrate local features, resulting in enhanced brightness and sharpened edges in the flame center of the fusion image, with some background effectively suppressed to a dark tone. The BVSSM method excels in processing global information, effectively presenting the details of large target areas. The images generated by the TSFF method appear brighter in small target flame areas and have clearer flame edges, while the fusion images produced by the BVSSM method are brighter and clearer in large flame areas. The MRCF fusion method combines advantages: it retains BVSSM's ability to handle global information while enhancing local flame details through TSFF. The image background is mostly dark-toned, with very low redundant noise interference, significantly highlighting the flame area. Therefore, the MRCF module can effectively leverage multimodal complementary information, improving the detection performance of fire targets in complex forest environments.

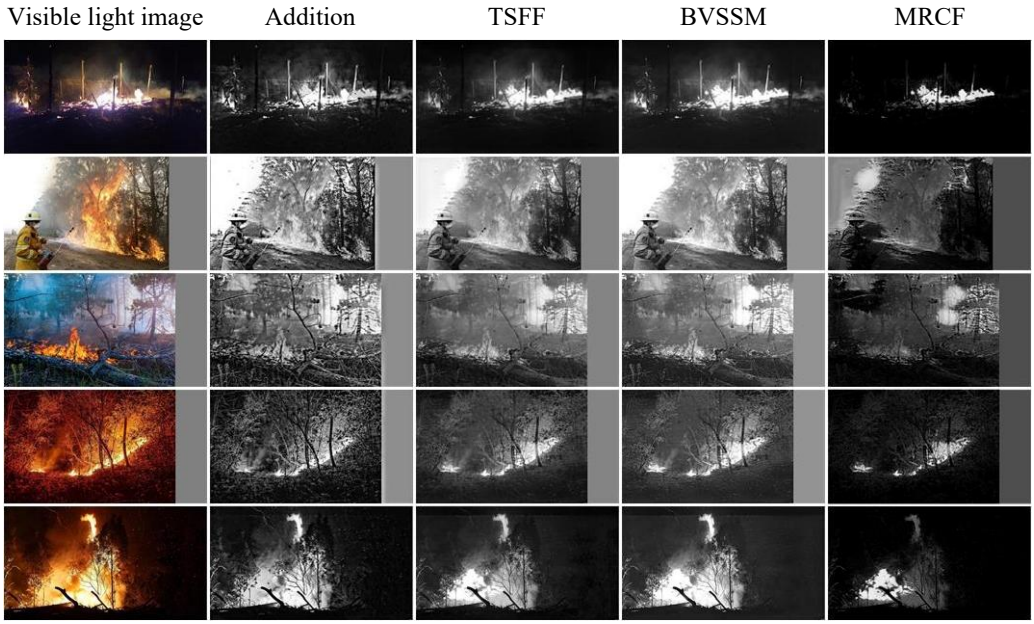

Figure 8. Visualization of the fusion results.

4.3.8. Ablation on Loss Function

Table 8 presents the ablation study for the dynamic factor α, with a fixed anomaly-degree benchmark $\delta = 3$. This experiment investigates how adjusting α influences the decay rate of gradient gain and its subsequent impact on detection performance. Results indicate that the model achieves optimal performance at $\alpha = 1.4$, where the gradient gain decay optimally balances the learning weights of high- and low-quality samples. When α exceeds 1.4, the accelerated decay rate excessively suppresses gradient contributions from certain samples, significantly degrading detection performance for smoke and flames. For instance, at $\alpha = 2.5$, precision decreases by 2.9% and mAP@0.5:0.95 declines by 4.4%.

In the ablation study of the anomaly-degree benchmark δ (with $\alpha = 1.4$ fixed), Table 9 demonstrates that $\delta = 3$ yields the best overall detection performance, achieving optimal results across all four evaluation metrics. As a dynamic threshold for anchor box anomaly degree, δ governs the model's attention to samples of varying quality by modulating the gradient gain allocation strategy. At $\delta = 3$, gradient gain peaks, directing the model's focus to medium-quality anchor boxes and significantly enhancing detection accuracy for translucent smoke regions. Deviations from $\delta = 3$ reduce performance: at $\delta = 2$, premature decay of gradient gain lowers precision by 1.8%, while at $\delta = 6$, the excessively high threshold causes neglect of medium-quality samples, further diminishing performance.

Table 8. The impact of dynamic factors of WIoU loss function on model performance. The optimal value is indicated in boldface.

Model	Difference	Precision	Recall	mAP@0.5	mAP@0.5:0.95
MCDet + WIoU-1	$\alpha = 1.0$	79.5	74.1	77.9	47.8
MCDet + WIoU-2	$\alpha = 1.4$	**81.3**	**76.6**	**80.1**	**48.8**
MCDet + WIoU-3	$\alpha = 1.6$	80.8	75.9	79.8	49.2
MCDet + WIoU-4	$\alpha = 1.9$	80.1	74.8	78.6	48.5
MCDet + WIoU-5	$\alpha = 2.5$	78.2	72.3	76.4	46.1

Table 9. Impact of WIoU loss function anomaly benchmark point on model performance. The optimal value is indicated in boldface.

Model	Difference	Precision	Recall	mAP@0.5	mAP@0.5:0.95
MCDet + WIoU-1	$\delta = \mathbf{2}$	79.8	73.2	77.6	46.3
MCDet + WIoU-2	$\delta = \mathbf{3}$	**81.3**	**76.6**	**80.1**	**48.8**
MCDet + WIoU-3	$\delta = \mathbf{4}$	80.1	75.4	78.9	47.5
MCDet + WIoU-4	$\delta = \mathbf{5}$	79.1	74.0	77.3	46.1
MCDet + WIoU-5	$\delta = \mathbf{6}$	78.5	72.7	77.1	45.2

4.4. Discussion

This paper proposes the MCDet framework to address core challenges in multimodal forest fire detection, including feature confusion between infrared radiation and visible light highlights, missed detections due to vegetation occlusion, and increased false alarm rates. The global–local cross-modal interaction mechanism established by the MRCF module significantly mitigates visual confusion between flame infrared radiation and visible light highlights. The dynamic gating weight generation capability of the CGAN successfully overcomes feature dispersion defects caused by vegetation occlusion and terrain interference. Meanwhile, the WIoU loss function optimizes localization accuracy in scenarios with blurred smoke boundaries and flame occlusion. Comprehensive superiority across four benchmark tests validates the framework's robustness: achieving a 9% improvement in mAP@0.5:0.95 on the D-Fire dataset, increasing smoke detection precision by 4.1% in low-contrast Corsican scenarios, reducing localization error by 2.5% in flame occlusion tests on the Fire-dataset, and notably demonstrating cross-scene generalization with 69.8% AP@0.5:0.95 on the LLVIP dataset.

Despite excellent performance in controlled forest areas, the model faces inherent limitations when transferred to wildfire and urban fire scenarios: extreme thermal convection and flying ember dispersion in wildfires may impair MRCF's thermal feature extraction capability, such as turbulence interference from rapidly spreading firelines. In urban environments, thermal reflections from glass curtain walls and high-temperature interference sources like vehicle engines exhibit feature distribution differences compared to forest highlights, leading to increase in false alarms for CGAN's attention mechanism during industrial fire tests. Future work will focus on two breakthroughs: developing dynamic feature decoupling modules to enhance generalization for wildfire embers and urban thermal interference sources, and integrating fire spread physical models to construct cross-scene prediction mechanisms. This will advance the detection framework's paradigm shift from confined forest areas to open disaster scenarios.

5. Conclusions

To address the challenges of cross-modal feature confusion and occlusion in multimodal fire detection, this paper proposes a multimodal fusion framework based on multidimensional representation synergy, named MCDet. First, the Multidimensional

Representation Collaborative Fusion Module (MRCF), which employs state-space modeling and deformable convolution to facilitate global–local cross-modal feature interactions, thereby enhancing complementary feature extraction. Then, the content-guided attention network (CGAN), designed to adaptively aggregate multidimensional features via dynamic gating mechanisms, suppressing background interference while enhancing discriminative representations of fire events. Finally, a WIoU loss function that improves localization of challenging samples by dynamically adjusting gradient weights, thereby mitigating misdirection from high-frequency noise. Experimental results demonstrate that our framework significantly enhances detection accuracy and robustness, enabling precise distinction between flames, smoke, and background interference in complex forest environments. Notably, the framework is compatible with both single-stage and two-stage detectors while maintaining lightweight inference capabilities. Building on this multidimensional representation synergy strategy, future work will explore its potential applications in multimodal disaster monitoring systems.

Author Contributions: Conceptualization, Y.X. and Y.B.; methodology, Y.B.; software, Y.B.; validation, X.W.; resources, Y.X.; data curation, Y.B.; writing—original draft preparation, H.W.; writing—review and editing, G.N.; visualization, H.W. and Y.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the following grants: Primary Research and Development Plan of Heilongjiang Province (Grant No. GA23A903), Aeronautical Science Foundation of China (Grant No. 202400550P6001), and Fundamental Research Funds for the Central Universities in China (Grant No. 3072024XX0602).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, Y.; Zheng, C.; Liu, X.; Tian, Y.; Zhang, J.; Cui, W. Forest fire monitoring method based on UAV visual and infrared image fusion. *Remote Sens.* **2023**, *15*, 3173. [CrossRef]
2. Abdusalomov, A.; Umirzakova, S.; Bakhtiyor Shukhratovich, M.; Mukhiddinov, M.; Kakhorov, A.; Buriboev, A.; Jeon, H.S. Drone-Based Wildfire Detection with Multi-Sensor Integration. *Remote Sens.* **2024**, *16*, 4651. [CrossRef]
3. Hare, J.; Tomsick, J.A.; Buisson, D.J.; Clavel, M.; Gandhi, P.; García, J.A.; Grefenstette, B.W.; Walton, D.J.; Xu, Y. NuSTAR observations of the transient galactic black hole binary candidate Swift J1858. 6–0814: A new sibling of V404 Cyg and V4641 Sgr? *Astrophys. J.* **2020**, *890*, 57. [CrossRef]
4. Jiao, Z.; Zhang, Y.; Xin, J.; Mu, L.; Yi, Y.; Liu, H.; Liu, D. A deep learning based forest fire detection approach using UAV and YOLOv3. In Proceedings of the 2019 1st International Conference on Industrial Artificial Intelligence (IAI), Shenyang, China, 23–27 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
5. Hu, Y.; Zhan, J.; Zhou, G.; Chen, A.; Cai, W.; Guo, K.; Hu, Y.; Li, L. Fast forest fire smoke detection using MVMNet. *Knowl.-Based Syst.* **2022**, *241*, 108219. [CrossRef]
6. Akhloufi, M.A.; Toulouse, T.; Rossi, L.; Maldague, X. Multimodal three-dimensional vision for wildland fires detection and analysis. In Proceedings of the 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), Montreal, QC, Canada, 28 November–1 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
7. Chen, X.; Hopkins, B.; Wang, H.; O'Neill, L.; Afghah, F.; Razi, A.; Fulé, P.; Coen, J.; Rowell, E.; Watts, A. Wildland fire detection and monitoring using a drone-collected rgb/ir image dataset. *IEEE Access* **2022**, *10*, 121301–121317. [CrossRef]
8. Meel, P.; Vishwakarma, D.K. Multi-modal fusion using Fine-tuned Self-attention and transfer learning for veracity analysis of web information. *Expert Syst. Appl.* **2023**, *229*, 120537. [CrossRef]
9. Kizilkaya, B.; Ever, E.; Yatbaz, H.Y.; Yazici, A. An effective forest fire detection framework using heterogeneous wireless multimedia sensor networks. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2022**, *18*, 1–21. [CrossRef]
10. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.

11. Chen, Y.C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. Uniter: Learning universal image-text representations. *arXiv* **2019**, arXiv:1909.11740.
12. Zhou, K.; Chen, L.; Cao, X. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020, Proceedings, Part XVIII 16*; Springer: Cham, Switzerland, 2020; pp. 787–803.
13. Dasgupta, K.; Das, A.; Das, S.; Bhattacharya, U.; Yogamani, S. Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 15940–15950. [CrossRef]
14. Chen, Y.T.; Shi, J.; Ye, Z.; Mertz, C.; Ramanan, D.; Kong, S. Multimodal object detection via probabilistic ensembling. In *Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022*; Springer: Cham, Switzerland, 2022; pp. 139–158.
15. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning. PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
16. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
17. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
18. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
19. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13 December 2015; pp. 1440–1448.
20. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef] [PubMed]
21. Wang, S.; Li, Y.; Qiao, S. ALF-YOLO: Enhanced YOLOv8 based on multiscale attention feature fusion for ship detection. *Ocean Eng.* **2024**, *308*, 118233. [CrossRef]
22. Jocher, G.; Stoken, A.; Borovec, J.; Changyu, L.; Hogan, A.; Diaconu, L.; Poznanski, J.; Yu, L.; Rai, P.; Ferriday, R.; et al. ultralytics/yolov5: V3. 0. Zenodo 2020. Available online: https://zenodo.org/records/3983579 (accessed on 25 June 2025).
23. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
24. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725.
25. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Springer: Cham, Switzerland, 2020; pp. 213–229.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [CrossRef]
27. Xiao, Y.; Yang, M.; Li, C.; Liu, L.; Tang, J. Attribute-based progressive fusion network for rgbt tracking. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 2831–2838.
28. Tang, L.; Chen, Z.; Huang, J.; Ma, J. Camf: An interpretable infrared and visible image fusion network based on class activation mapping. *IEEE Trans. Multimed.* **2023**, *26*, 4776–4791. [CrossRef]
29. Wang, Z.; Chen, Y.; Shao, W.; Li, H.; Zhang, L. SwinFuse: A residual swin transformer fusion network for infrared and visible images. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5016412. [CrossRef]
30. Li, H.; Wu, X.J. CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Inf. Fusion* **2024**, *103*, 102147. [CrossRef]
31. Yang, Y.; Zhou, N.; Wan, W.; Huang, S. MACCNet: Multiscale Attention and Cross-Convolutional Network for Infrared and Visible Image Fusion. *IEEE Sens. J.* **2024**, *24*, 16587–16600. [CrossRef]
32. Zhao, Z.; Bai, H.; Zhu, Y.; Zhang, J.; Xu, S.; Zhang, Y.; Zhang, K.; Meng, D.; Timofte, R.; Van Gool, L. DDFM: Denoising diffusion model for multi-modality image fusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 8082–8093.
33. Zhang, T.; Zhu, Y.; Zhao, J.; Cui, G.; Zheng, Y. Exploring state space model in wavelet domain: An infrared and visible image fusion network via wavelet transform and state space model. *arXiv* **2025**, arXiv:2503.18378.
34. Liu, X.; Wang, Z.; Gao, H.; Li, X.; Wang, L.; Miao, Q. HATF: Multi-modal feature learning for infrared and visible image fusion via hybrid attention transformer. *Remote Sens.* **2024**, *16*, 803. [CrossRef]
35. Jiang, M.; Wang, Z.; Kong, J.; Zhuang, D. MCFusion: Infrared and visible image fusion based multiscale receptive field and cross-modal enhanced attention mechanism. *J. Electron. Imaging* **2024**, *33*, 013039. [CrossRef]
36. Liu, J.; Zhang, L.; Zeng, X.; Liu, W.; Zhang, J. MATCNN: Infrared and Visible Image Fusion Method Based on Multi-scale CNN with Attention Transformer. *arXiv* **2025**, arXiv:2502.01959. [CrossRef]
37. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual state space model. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 103031–103063.

38. Gaur, A.; Singh, A.; Kumar, A.; Kumar, A.; Kapoor, K. Video flame and smoke based fire detection algorithms: A literature review. *Fire Technol.* **2020**, *56*, 1943–1980. [CrossRef]
39. Mahmoud, M.A.I.; Ren, H. Forest fire detection and identification using image processing and SVM. *J. Inf. Process. Syst.* **2019**, *15*, 159–168.
40. Zheng, S.; Zou, X.; Gao, P.; Zhang, Q.; Hu, F.; Zhou, Y.; Wu, Z.; Wang, W.; Chen, S. A forest fire recognition method based on modified deep CNN model. *Forests* **2024**, *15*, 111. [CrossRef]
41. El-Madafri, I.; Peña, M.; Olmedo-Torre, N. Real-time forest fire detection with lightweight CNN using hierarchical multi-task knowledge distillation. *Fire* **2024**, *7*, 392. [CrossRef]
42. Jin, L.; Yu, Y.; Zhou, J.; Bai, D.; Lin, H.; Zhou, H. SWVR: A lightweight deep learning algorithm for forest fire detection and recognition. *Forests* **2024**, *15*, 204. [CrossRef]
43. Khan, S.; Khan, A. Ffirenet: Deep learning based forest fire classification and detection in smart cities. *Symmetry* **2022**, *14*, 2155. [CrossRef]
44. Ghosh, R.; Kumar, A. A hybrid deep learning model by combining convolutional neural network and recurrent neural network to detect forest fire. *Multimed. Tools Appl.* **2022**, *81*, 38643–38660. [CrossRef]
45. Bhamra, J.K.; Anantha Ramaprasad, S.; Baldota, S.; Luna, S.; Zen, E.; Ramachandra, R.; Kim, H.; Schmidt, C.; Arends, C.; Block, J.; et al. Multimodal wildland fire smoke detection. *Remote Sens.* **2023**, *15*, 2790. [CrossRef]
46. Fodor, G.; Conde, M.V. Rapid deforestation and burned area detection using deep multimodal learning on satellite imagery. *arXiv* **2023**, arXiv:2307.04916.
47. Al Duhayyim, M.; Eltahir, M.M.; Omer Ali, O.A.; Albraikan, A.A.; Al-Wesabi, F.N.; Hilal, A.M.; Hamza, M.A.; Rizwanullah, M. Fusion-Based Deep Learning Model for Automated Forest Fire Detection. *Comput. Mater. Contin.* **2023**, *77*, 1355–1371. [CrossRef]
48. Alipour, M.; La Puma, I.; Picotte, J.; Shamsaei, K.; Rowell, E.; Watts, A.; Kosovic, B.; Ebrahimian, H.; Taciroglu, E. A multimodal data fusion and deep learning framework for large-scale wildfire surface fuel mapping. *Fire* **2023**, *6*, 36. [CrossRef]
49. Sun, F.; Yang, Y.; Lin, C.; Liu, Z.; Chi, L. Forest fire compound feature monitoring technology based on infrared and visible binocular vision. *J. Phys. Conf. Ser. Iop Publ.* **2021**, *1792*, 012022. [CrossRef]
50. Kim, D.; Ruy, W. CNN-based fire detection method on autonomous ships using composite channels composed of RGB and IR data. *Int. J. Nav. Archit. Ocean. Eng.* **2022**, *14*, 100489. [CrossRef]
51. Zhao, W.; Xie, S.; Zhao, F.; He, Y.; Lu, H. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 13955–13965.
52. Shen, J.; Chen, Y.; Liu, Y.; Zuo, X.; Fan, H.; Yang, W. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognit.* **2024**, *145*, 109913. [CrossRef]
53. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
54. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
55. Gu, A.; Goel, K.; Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv* **2021**, arXiv:2111.00396.
56. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
57. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.
58. Toulouse, T.; Rossi, L.; Campana, A.; Celik, T.; Akhloufi, M.A. Computer vision for wildfire research: An evolving image dataset for processing and analysis. *Fire Saf. J.* **2017**, *92*, 188–194. [CrossRef]
59. de Venancio, P.V.A.; Lisboa, A.C.; Barbosa, A.V. An automatic fire detection system based on deep convolutional neural networks for low-power, resource-constrained devices. *Neural Comput. Appl.* **2022**, *34*, 15349–15368. [CrossRef]
60. Park, T.; Efros, A.A.; Zhang, R.; Zhu, J.Y. Contrastive learning for unpaired image-to-image translation. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part IX 16; Springer: Cham, Switzerland, 2020; pp. 319–345.
61. Olafenwa, M. FireNET 2019. Available online: https://github.com/OlafenwaMoses/FireNET (accessed on 25 June 2025).
62. Jia, X.; Zhu, C.; Li, M.; Tang, W.; Zhou, W. LLVIP: A visible-infrared paired dataset for low-light vision. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 3496–3504.
63. Liu, H.; Jin, F.; Zeng, H.; Pu, H.; Fan, B. Image enhancement guided object detection in visually degraded scenes. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 14164–14177. [CrossRef]
64. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv* **2022**, arXiv:2203.03605.
65. He, X.; Tang, C.; Zou, X.; Zhang, W. Multispectral object detection via cross-modal conflict-aware learning. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 1465–1474.

66. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
67. Wang, Z.; Colonnier, F.; Zheng, J.; Acharya, J.; Jiang, W.; Huang, K. TIRDet: Mono-modality thermal infrared object detection based on prior thermal-to-visible translation. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 2663–2672.
68. Cao, Y.; Fan, Y.; Bin, J.; Liu, Z. Lightweight transformer for multi-modal object detection (student abstract). *Proc. AAAI Conf. Artif. Intell.*, **2023**, *37*, 16172–16173. [CrossRef]
69. Dong, W.; Zhu, H.; Lin, S.; Luo, X.; Shen, Y.; Liu, X.; Zhang, J.; Guo, G.; Zhang, B. Fusion-mamba for cross-modality object detection. *arXiv* **2024**, arXiv:2404.09146.
70. Xu, G.; He, C.; Wang, H.; Zhu, H.; Ding, W. Dm-fusion: Deep model-driven network for heterogeneous image fusion. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 10071–10085. [CrossRef]
71. Cao, Y.; Bin, J.; Hamari, J.; Blasch, E.; Liu, Z. Multimodal object detection by channel switching and spatial attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 403–411.
72. Yi, X.; Tang, L.; Zhang, H.; Xu, H.; Ma, J. Diff-IF: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Inf. Fusion* **2024**, *110*, 102450. [CrossRef]
73. Zhang, H.; Fromont, E.; Lefèvre, S.; Avignon, B. Guided attentive feature fusion for multispectral pedestrian detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 72–80.
74. Tang, L.; Xiang, X.; Zhang, H.; Gong, M.; Ma, J. DIVFusion: Darkness-free infrared and visible image fusion. *Inf. Fusion* **2023**, *91*, 477–493. [CrossRef]