forests

MDPI

Article

FireRGBTNet: A Lightweight Forest Fire Detection Model Based on Efficient RGB–Thermal Fusion

Yifan Ma [1], Weifeng Shan [1,*], Maofa Wang [2], Yanwei Sui [1] and Mengyu Wang [1]

[1] School of Emergency Technology and Management, University of Emergency Management, Sanhe 065201, China; 24661712@st.cidp.edu.cn (Y.M.); 23661621@st.cidp.edu.cn (Y.S.); 23661710@st.cidp.edu.cn (M.W.)
[2] School of Applied Science, Beijing Information Science & Technology University, Beijing 100192, China; wangmaofa@bistu.edu.cn
* Correspondence: shanweifeng@cidp.edu.cn

Abstract

In recent years, UAV-based visible and thermal (RGB-T) multimodal object detection has demonstrated significant potential for monitoring forest fires in complex environments. However, restricted by the intrinsic discrepancies between modalities, existing RGB-T models still struggle to achieve optimal detection accuracy. To address the aforementioned issues, this paper proposes FireRGBTNet, a lightweight and efficient RGB-T fusion model for UAV-based forest fire detection. First, a heterogeneous dual-stream backbone is designed to precisely capture modality-specific information via customized feature extraction modules, utilizing Target-Enhanced Downsampling to adaptively preserve the fragile features of small targets. Second, a multi-scale semantic alignment enhancement branch is proposed to explicitly bridge the semantic gap through the dual constraints of statistical distribution and semantic direction. Finally, a multimodal spatial gated fusion module is constructed, which utilizes spatial gated interactions to effectively suppress multimodal noise and incorporates Mish Gated Linear Units to compensate for global semantics, thereby achieving high-quality heterogeneous feature fusion. Extensive experiments on the RGBT-3M dataset demonstrate that with a mere 4.13M parameters, FireRGBTNet achieves accuracies of 95.9% and 62.8% in terms of mAP@0.5 and mAP@0.5:0.95, respectively. The proposed model achieves an optimal trade-off between detection accuracy and computational efficiency, providing a highly effective solution for UAV-based forest fire monitoring in complex environments.

Keywords: object detection; UAV; forest fire; sensor fusion; multimodality

Check for updates

Academic Editors: Change Zheng, Fengjun Zhao, Xingdong Li and Yunfei Liu

Received: 13 July 2026
Revised: 4 August 2026
Accepted: 10 August 2026
Published: 12 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

As one of the most vital ecosystems on Earth, forests play an irreplaceable role in maintaining the global carbon cycle, conserving biodiversity, and regulating the climate. In recent years, influenced by climate change and human activities, forest fires have occurred with increasing frequency, posing severe threats to human lives and property while causing tremendous damage to the natural ecological environment [1]. Consequently, achieving early warning and real-time monitoring of forest fires is of paramount significance for mitigating disaster losses and protecting ecosystems. Traditional forest fire monitoring approaches primarily rely on manual patrols and fixed lookout towers, ground-based sensors, and satellite remote sensing. However, manual patrols and lookout towers suffer from major blind spots and limited coverage area [2]. Although ground-based sensors

and near-ground optical monitoring devices offer superior real-time performance, their deployment density remains constrained by high maintenance costs [3]. Meanwhile, satellite remote sensing technologies are broadly hindered by inadequate spatial resolution, poor timeliness, and long response latency [4]. Thus, these traditional methods struggle to concurrently meet the stringent requirements of precise target localization and rapid response for fire detection. Driven by recent advancements in computer vision and deep learning, combined with the advantages of unmanned aerial vehicles (UAVs)—such as high mobility, broad field-of-view coverage, and independence from complex terrain—UAV-based forest fire monitoring has emerged as a crucial research direction [5]. Nevertheless, complex forest environments place higher demands on the robustness and computational efficiency of detection algorithms. On the one hand, abrupt illumination variations and dense smoke occlusion frequently result in feature loss and false detections [6], while complicated background clutter and severe meteorological interference further degrade small-target detection performance [7]. On the other hand, constrained by the resource limitations of edge computing platforms onboard UAVs, models with high computational complexity are difficult to deploy for real-time edge inference [8].

In addressing the complex challenges of object detection in forest fire scenarios, early research predominantly focused on visible-light images. Early detection algorithms primarily relied on handcrafted operators such as color space transformations, dynamic contours, and textures. For instance, Prema et al. [9] proposed a multi-feature analysis-based smoke detection method grounded in the YUV color space. By processing through three distinct stages to reduce false alarms, this approach achieved an accuracy of 96.29% and a false alarm rate of 1.47%, thereby enhancing both the accuracy and response speed of early forest fire detection. However, the generalization capability of such methods is severely limited when confronted with complex and variable natural backgrounds. With the advancement of deep convolutional neural networks, detection paradigms based on the Faster R-CNN series have been extensively applied to fire scenarios, significantly improving feature extraction capabilities for flames and smoke. For example, Pan et al. [10] introduced a forest fire smoke region detection framework leveraging weakly supervised fine segmentation and a lightweight Faster R-CNN. By aggregating image-level information rather than relying on costly pixel-level annotations, the proposed framework utilizes its weakly supervised fine segmentation method and decision network to localize and segment forest fire smoke. On a self-constructed fire dataset, it attained a mean Intersection over Union (mIoU) of 70.2% and an Average Precision (AP) of 99.6%. Furthermore, considering that flames and smoke exhibit characteristics of dynamic diffusion and morphological variability, Majid et al. [11] attempted to incorporate attention mechanisms into fire object detection. Leveraging the powerful global context modeling capacity of attention mechanisms, the model's detection accuracy was further enhanced, achieving a test accuracy of 95.40% and a recall of 97.61% on a real-world fire image dataset. Nevertheless, the imaging quality of visible-light images is highly dependent on favorable lighting conditions and clear visibility. In nighttime environments or during the mid-to-late stages of a fire when dense smoke prevails, visual features of targets undergo severe degradation. Consequently, detectors based on RGB modalities frequently encounter critical missed-detection issues [12].

To overcome the susceptibility of visible-light imaging to environmental interference, some researchers have begun exploring the application of thermal technology in fire scenarios [13]. Thermal sensors are highly sensitive to temperature gradients, enabling them to operate independently of lighting variations and effectively penetrate smoke, thereby circumventing occlusion issues. However, due to the lack of rich textural details and clear edge contours in infrared images, thermal infrared detectors are highly susceptible to background thermal noise from heated rocks, reflective objects, or high-temperature wild

animals in complex forest environments, consequently inducing a substantial number of false positives. Meanwhile, the accurate identification of small infrared targets also poses significant challenges [14], rendering it highly prone to false positive alarms. Therefore, compared to employing thermal sensors alone for fire detection, leveraging the complementary advantages of the fine spatial texture features provided by visible-light imaging and the salient thermal radiation features offered by thermal infrared has become an inevitable choice for breaking through the performance bottlenecks of fire object detection in complex forest environments.

Existing multimodal object detection primarily relies on two stages—feature extraction and feature fusion—to achieve cross-modal information interaction. In the feature extraction stage, adopting a dual-stream backbone network for parallel feature extraction has become the mainstream architecture in this field. For instance, Wang et al. [15] employed a dual-stream ResNet backbone to independently extract RGB and infrared features. This symmetric architecture, leveraging ResNet's excellent feature representation capability, substantially improved object detection accuracy, achieving mean Average Precisions (mAP) of 78.42% and 83.63% on the DroneVehicle and VEDAI datasets, respectively. However, the doubled model parameter count also incurs high computational complexity and deployment costs. Meng et al. [16] extended YOLO, a generic lightweight model based on convolutional neural networks, into a dual-stream structure. By constructing two parallel backbone branches to independently extract and fuse features from visible-light and infrared images, it not only significantly improved multimodal detection accuracy but also maintained favorable inference efficiency, attaining a mAP@0.5 of 88.8% and a mAP@0.5:0.95 of 59.9% on the M3FD dataset. Although the aforementioned dual-stream architectures have achieved certain successes, their backbone networks mostly adopt identical network designs, overlooking the inherent feature discrepancies between modalities. Specifically, visible-light images are rich in fine-grained, high-frequency textural details, whereas infrared images primarily characterize thermal radiation distributions, where the background often manifests as large-area, low-frequency sparse regions, with effective information highly concentrated in a few high-temperature fire spots. Applying identical designs to both feature extraction branches—utilizing the same parameter scale for Thermal feature extraction as used in the RGB branch—not only leads to severe computational redundancy and parameter waste in the infrared branch when processing large areas of invalid dark backgrounds, but also constrains the effective extraction of fine-grained textural features from visible-light images. Furthermore, early-stage fire targets typically manifest as extremely small-pixel, high-brightness anomalous regions. To expand the receptive field, traditional CNNs generally implement downsampling via strided convolutions; such operations inevitably cause the feature signals of tiny targets to be diluted or even lost amidst background noise during layer-by-layer propagation. Most existing multimodal detection networks directly reuse generic backbone structures, lacking explicit feature preservation and enhancement mechanisms tailored for tiny fire targets. Consequently, they are highly prone to missed detections when confronting early-stage fires against complex forest backgrounds.

In the feature fusion stage, some studies employ element-wise addition or channel concatenation strategies [17]. Although these methods incur low computational costs, they overlook the complementarity and conflicts between modalities; direct fusion tends to readily introduce noise. To better integrate features across different modalities, existing multimodal fusion schemes predominantly adopt cross-modal network designs based on attention mechanisms to achieve deep complementary information exchange between modalities. For example, Sun et al. [18] employed a channel attention mechanism to enhance features from different modalities along the channel dimension and achieved se-

mantic interaction through weight swapping, attaining an F1-score of 83% across multiple fire datasets. Furthermore, Shen et al. [19] utilized a powerful cross-attention module to perform feature modeling via a global perspective, aggregating bimodal semantic information across the entire range, which greatly improved the multimodal feature fusion effect, achieving a mAP@0.5 of 79.20%. Regarding cross-modal Transformer fusion, Zhang et al. [20] proposed the cross-modal fusion framework CMX, which leverages a cross-modal feature rectification module to rectify features of one modality using those of another, achieving state-of-the-art performance with an mIoU of 59.7% on the MFNet dataset; however, its Transformer structure based on global attention incurs substantial computational overhead. Nevertheless, despite the excellent performance of such fusion methods in conventional fusion scenarios, issues persist in UAV-based forest fire monitoring tasks. First, forest backgrounds are extremely vast and texturally complex, while early-stage fire targets typically occupy very few pixels. In global attention calculations, the fragile local features of tiny fire spots are highly susceptible to being overwhelmed by the vast and redundant useless background noise, causing the global attention mechanism to dilute the high local responses of small targets due to its excessive focus on global context when computing correlation matrices. Second, modal interference issues also exist in forest fire scenarios. The visible-light modality often contains backgrounds highly visually similar to flame characteristics, such as "red leaves" or "sunset glow," while edges of high-temperature targets in the infrared modality are frequently accompanied by significant thermal spillover. When these interfering features enter the interaction modules, they often induce erroneous semantic enhancement, leading to mutual amplification of both types of noise during cross-modal fusion and consequently degrading the model's detection performance. Finally, the excessive utilization of global attention computations brings massive parameters and computational loads, which also contradicts our original intention of lightweight deployment on UAVs for forest fire monitoring.

Furthermore, beyond the structural design of cross-modal interaction mechanisms, existing multimodal object detection methods also exhibit deficiencies during the early multimodal feature processing stage. Current mainstream multimodal detection approaches predominantly feed the extracted dual-stream features directly into a multimodal encoder for joint modeling to achieve feature fusion, neglecting the fact that RGB images and thermal images originate from different physical sensors and that their feature distributions reside in completely independent representation spaces, thereby exhibiting a substantial cross-modal semantic gap. Research in the field of multimodal representation learning has demonstrated that, without explicit feature alignment constraints, direct cross-modal fusion is highly prone to inducing severe semantic inequivalence and feature space misalignment issues [21]. Although some studies have attempted to introduce explicit semantic anchors [22] or contrastive learning [23] to narrow such heterogeneous representation discrepancies, traditional encoder-based solutions still rely on simply stacking deeper modules to enhance the model's fitting capacity. This strategy of relying on brute-force stacking not only increases the model parameter count but also introduces excessively complex fusion mechanisms and prohibitive computational overhead. Simultaneously, its practical effectiveness in bridging cross-modal representation disparities remains rather limited. More critically, this practice significantly elevates the risk of model overfitting, rendering the optimization process of cross-modal joint modeling exceptionally arduous.

In summary, existing multimodal forest fire detection methods still confront pressing challenges during the feature extraction and fusion stages. Firstly, traditional homogeneous dual-stream backbone networks neglect the intrinsic physical differences between visible and infrared image features, and conventional downsampling operations are highly susceptible to causing fragile fire features to attenuate or even vanish during propagation.

Secondly, due to the absence of explicit feature alignment constraints, direct interaction between heterogeneous features separated by inherent semantic gaps induces severe semantic disparity. Finally, conventional global attention fusion mechanisms not only incur massive computational overhead but are also highly prone to diluting faint target features by excessively focusing on redundant backgrounds, and may even amplify background interference noise across modalities. To systematically address the aforementioned issues, this paper proposes FireRGBTNet, a lightweight forest fire detection model based on efficient RGB–Thermal fusion. This model comprehensively exploits the feature advantages of each modality via customized heterogeneous backbone networks, utilizes an explicit semantic alignment mechanism prior to feature fusion to mitigate modal discrepancies, and achieves highly efficient cross-modal fusion with minimal parameters through a spatial gated attention mechanism, ultimately achieving high-precision, lightweight fire detection in complex forest environments. The main contributions of this paper are summarized as follows:

(1) A heterogeneous dual-stream backbone is designed. To address the inherent physical discrepancies between modalities, we propose the RGBBlock and ThermalBlock modules, which effectively mitigate the computational redundancy of the infrared branch while maximizing modal feature extraction. Concurrently, a Target-Enhanced Downsampling (TED) module is proposed. By employing an adaptive weighted fusion mechanism to dynamically regulate extremum features and local spatial features, TED prevents the attenuation and loss of feature intensity for diminutive, early-stage fire spots during network downsampling.

(2) A multi-scale semantic alignment enhancement (MSAE) branch is proposed. This branch projects heterogeneous features into a shared semantic space and achieves explicit alignment of cross-modal features across multiple scales by introducing the dual constraints of channel-level Wasserstein distribution alignment (CWA) and Cosine Loss.

(3) A multimodal spatial gated fusion (MSGF) module is constructed. To address the issue wherein global attention mechanisms tend to dilute small target features and amplify cross-modal noise, we utilize Spatial Gated Cross-Attention (SGCA) and Mish Gated Linear Units (MishGLU) within the MSGF. SGCA achieves the enhancement of diminutive fire spot features and the suppression of multimodal noise through multi-scale local receptive fields and cross-modal gating logic. Meanwhile, MishGLU provides global context compensation, ultimately synergizing to realize the high-quality fusion of RGB and Thermal features.

(4) An efficient end-to-end RGB-IR lightweight forest fire detection model is proposed. We construct a globally unified loss function incorporating auxiliary matching costs and other components, enabling end-to-end joint optimization of the entire lightweight detection model. Extensive experiments demonstrate that FireRGBTNet exhibits strong competitiveness compared with state-of-the-art object detection models, providing a promising solution for real-time UAV-based forest fire monitoring under resource-constrained scenarios.

2. Materials and Methods

2.1. Dataset

To comprehensively evaluate the multimodal object detection performance of the proposed model in complex forest environments, the experimental validation in this paper is based on the publicly available UAV aerial multi-modal forest fire dataset RGBT-3M [24]. This dataset was collected by the original authors using DJI Matrice 300 RTK and DJI Mavic 2 Enterprise drones across multiple representative forest regions in Anhui, Yunnan,

and Inner Mongolia, China, extensively covering authentic forest fire scenarios as well as outdoor simulated fire experiments. During the data construction phase, the original authors extracted frames at 5 frames per second from approximately 10 UAV aerial video sequences and adopted a frame-skipping strategy to reduce inter-frame similarity, ultimately obtaining 11,220 pairs of flame image pairs. Concurrently, the M-RIFT method was employed to achieve high-precision registration of visible-light and thermal infrared images captured by heterogeneous sensors, effectively eliminating spatial misalignment and geometric distortions between cross-modal images. This rigorous normalization process in temporal synchronization and spatial alignment highly aligns with the research requirements of this paper regarding cross-modal feature fusion. Specifically, this dataset encompasses fine-grained annotations for three categories of targets—Smoke, Fire, and Person—bridging the gap of lacking high-quality registered RGB-T datasets in complex forest fire scenarios. The original authors partitioned the dataset into a training set (7854 pairs) and a validation set (3366 pairs) at a ratio of 7:3. This paper directly adopts this standard split of the public dataset for model training and evaluation. In the specific experimental setup, to adapt to the network architecture of the proposed multimodal detection model, we performed preprocessing operations on the data at the model input end, uniformly padding all visible-light and thermal infrared images from the originally provided size of 640 × 480 to 640 × 640. Detailed statistical information of the RGBT-3M dataset is presented in Table 1, detailed information of the acquisition equipment is provided in Table 2, and partial representative image samples are displayed in Figure 1.

Table 1. Detailed information of the RGBT-3M dataset.

Dataset	Number of Image Pairs	Number of Labels	Person	Fire	Smoke
Train	7854	21,550	4148	7914	9488
Val	3366	9227	1740	3401	4086
Total	11,220	30,777	5888	11,315	13,574

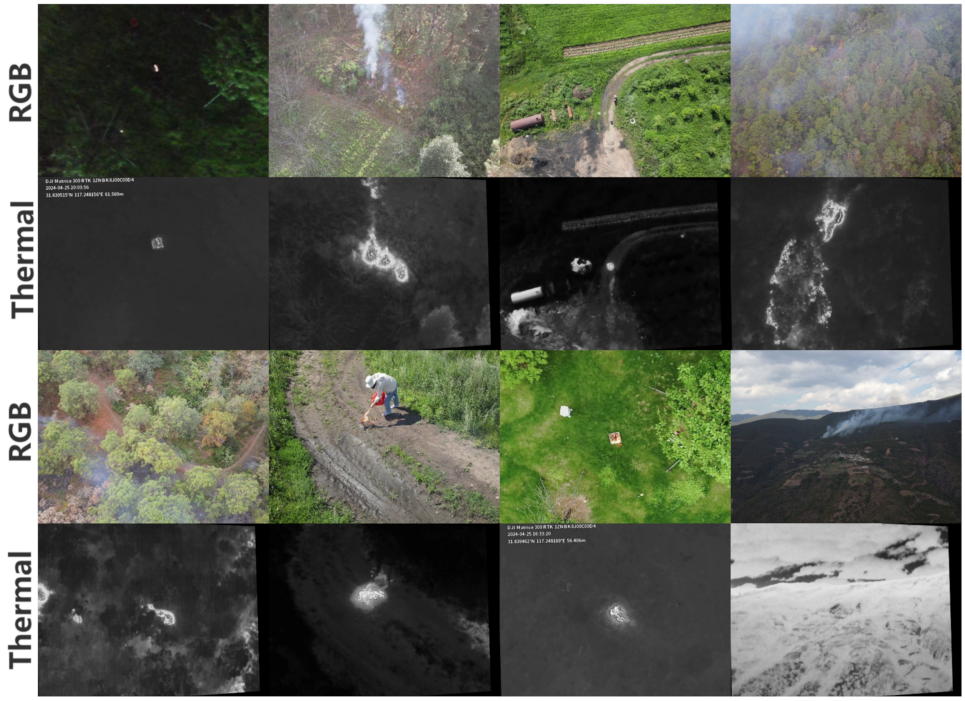


Figure 1. Partial display of the RGBT-3M dataset.

Table 2. Detailed information of the collection equipment.

UAV	Camera	FPS	Resolution
DJI Matrice 300 RTK	H20T	30	visible image: 1920 × 1080 infrared image: 640 × 512
DJI MAVIC 2 Enterprise	All-in-one camera	30	visible image: 1920 × 1080 infrared image: 900 × 720

2.2. Methods

2.2.1. FireRGBTNet Overall Architecture

To address critical challenges such as the inherent semantic gap between RGB and thermal multimodal images in complex forest fire scenarios, the susceptibility of diminutive fire spot features to attenuation, and the difficulty of efficiently fusing heterogeneous features, this paper innovatively proposes a multimodal object detection network, FireRGBTNet. The overall network architecture of FireRGBTNet is illustrated in Figure 2. We primarily adopt an end-to-end object detection paradigm consisting of a heterogeneous dual-stream backbone, a neck, and a detection head.

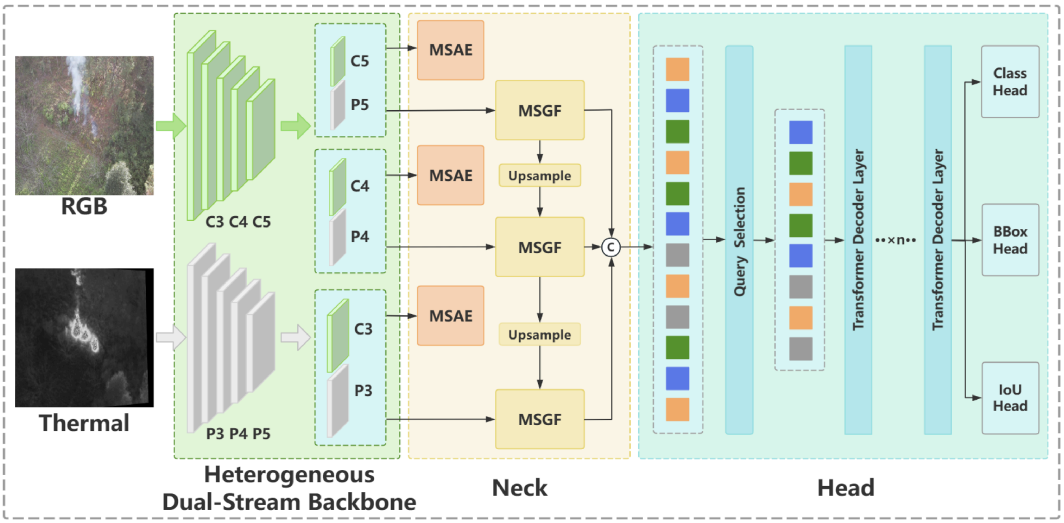


Figure 2. Overall architecture of FireRGBTNet.

Heterogeneous dual-stream backbone. As the foundation of feature extraction, we depart from the traditional homogeneous two-stream architecture and design a heterogeneous dual-stream backbone. This backbone is primarily responsible for independently receiving RGB and thermal images and extracting modal features that align with their respective physical imaging characteristics through customized modules. Simultaneously, to address the challenge wherein diminutive fire spots and smoke targets in the early stages of a forest fire are easily lost within deep networks, we introduce Target-Enhanced Downsampling (TED) at this stage. TED adaptively amplifies and preserves critical high-frequency weak signals during the dimensionality reduction process.

Neck. During the cross-modal feature interaction stage, the Neck component is primarily composed of the multi-scale semantic alignment enhancement (MSAE) and the multimodal spatial gated fusion (MSGF) modules. Specifically, MSAE is responsible for explicitly aligning the two-stream features through dual constraints of statistical distribution and semantic orientation prior to feature fusion, thereby bridging the semantic gap and significantly alleviating the optimization burden on subsequent fusion modules. Following feature alignment, a Feature Pyramid Network (FPN) structure [25] is constructed within the Neck. MSGF performs efficient feature fusion of multi-scale cross-modal features along

the top-down path. This design not only facilitates the thorough interaction of information across different receptive fields but also provides the detection head with high-quality features that possess both rich global semantics and local spatial details.

Head. In the final target prediction stage, we adopt the mainstream RT-DETR [26] detection methodology, treating object detection as a direct set prediction problem. This approach completely eliminates the reliance on prior anchors and Non-Maximum Suppression (NMS) [27] post-processing typical of traditional methods, thereby ensuring the efficiency and real-time performance of the model during practical deployment. (It should be noted that the RT-DETR decoder adopted in this paper has a hidden dimension of 128, 2 decoder layers, and 300 queries.)

2.2.2. Heterogeneous Dual-Stream Backbone

Existing multimodal object detection networks typically employ a symmetric dual-stream architecture, utilizing homogeneous feature extraction backbones for both RGB and thermal inputs. However, this homogeneous design overlooks the intrinsic discrepancies between modal characteristics: RGB images are characterized by dense, fine-grained, and high-frequency textures, whereas thermal images generally exhibit large-scale, low-frequency sparse backgrounds interspersed with localized salient thermal radiation. To address this issue, this paper proposes a heterogeneous dual-stream backbone. By accounting for these inter-modal feature differences, we design differentiated feature extraction modules—specifically, the RGBBlock and ThermalBlock—for the respective backbones. This approach maximizes the extraction of modality-specific features while effectively mitigating the computational redundancy incurred when the network processes extensive sparse infrared backgrounds. Furthermore, during the feature downsampling stage, conventional networks typically rely on convolution operations with a stride of 2. Nevertheless, in the nascent stages of a forest fire, fire targets often occupy only a nominal number of pixels, resulting in extremely fragile feature responses. The inherent local smoothing effect of standard convolutions can easily cause these weak yet critical local high-frequency signals (such as minute high-temperature anomalies or incipient smoke textures) to be diluted or entirely lost amidst massive background noise during feature propagation. To effectively overcome the problem of feature attenuation in small targets, we further design a Target-Enhanced Downsampling (TED) module, which adaptively focuses on and amplifies high-response regions during the downsampling process, thereby maximizing the preservation of fragile features for minuscule early-stage fire points. The specific architectural design is illustrated in Figure 3.

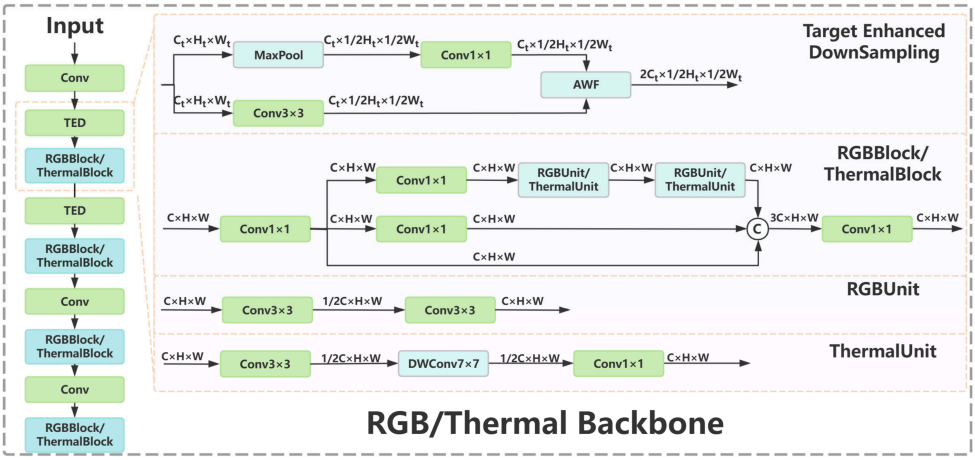


Figure 3. Heterogeneous dual-stream backbone architecture diagram, where $C = 2C_t$, $H = 1/2H_t$, $W = 1/2W_t$.

(a) RGBBlock

In forest fire scenarios, RGB images contain extremely rich fine-grained visual information, such as the complex edges of smoke, dynamic color gradients of flames, and intricate textures of the surrounding environment. To fully extract these dense, high-frequency local features, conventional feature extraction networks typically stack a large number of convolutions, which inevitably incurs significant computational overhead. Furthermore, as network depth increases, complex architectures are highly prone to the vanishing gradient problem during backpropagation. This makes it difficult for the model to effectively update weights in shallow layers, subsequently leading to a degradation in the model's ability to capture and represent subtle texture features during the early stages of a fire. To effectively extract dense high-frequency texture features while maintaining stable gradient propagation, we design an RGBBlock module within the backbone network dedicated to processing RGB images, as depicted in Figure 3. Given that lightweight depthwise separable convolutions tend to decouple the feature interaction between spatial and channel dimensions, the RGBBlock retains the use of standard convolutions to maximize the capture of complex joint spatial-channel features within the RGB modality. Simultaneously, to address the gradient and efficiency bottlenecks encountered when stacking standard convolutions, we incorporate the design principles of the Efficient Layer Aggregation Network (ELAN) [28]. By constructing multiple parallel information flow paths, we achieve multi-level feature reuse and gradient path optimization without a substantial increase in computational complexity. Specifically, the forward propagation process of the RGBBlock can be described by the following mathematical expressions. Set the input feature map as $X \in \mathbb{R}^{C_{in} \times H \times W}$. First, a 1×1 convolution is applied for feature integration and dimensionality alignment, yielding the base feature tensor X_0:

$$X_0 = \mathrm{Conv}_{1\times1}(X) \tag{1}$$

Within this layer aggregation architecture, the information flow of X_0 is further decoupled into three parallel feature propagation paths to achieve multi-level feature reuse across various receptive field scales. To maximize the preservation of original fine-grained spatial features, the first path performs no non-linear transformations on X_0. F_1, acting as a shortcut to retain the original features, is transmitted directly to ensure the lossless propagation of high-frequency textures and provide the shortest gradient backpropagation channel:

$$F_1 = X_0 \tag{2}$$

The second path employs a 1×1 convolution to perform cross-channel linear mapping. By facilitating inter-channel feature interaction and preliminary refinement without altering the spatial receptive field, this path generates the feature representation F_2:

$$F_2 = \mathrm{Conv}_{1\times1}(X_0) \tag{3}$$

The third path extracts more complex fine-grained features through multi-level feature aggregation. Specifically, after undergoing dimensionality reduction via a 1×1 convolution, this branch passes through two RGBUnit feature extraction modules specifically tailored for the RGB modality to capture spatial textures and local contextual features F_3 within a broader receptive field:

$$F_3 = \Phi_{\mathrm{RGB}}(\Phi_{\mathrm{RGB}}(\mathrm{Conv}_{1\times1}(X_0))) \tag{4}$$

In this formulation, Φ_{RGB} denotes the feature extraction module of a single RGBUnit. To address the high-frequency and dense characteristics inherent in the RGB modality, this module employs standard convolutions to extract joint spatial and channel interaction

features. Simultaneously, to expand the receptive field while effectively managing computational redundancy, the module adopts a bottleneck architecture. Set $F \in \mathbb{R}^{C \times H' \times W'}$ represents an arbitrary intermediate feature tensor input to the extraction unit. The unit comprises two consecutive 3×3 convolutions, during which the feature channels undergo a compression and expansion sequence of $C \to \frac{1}{2}C \to C$ during propagation. Accordingly, the specific mathematical expression for Φ_{RGB} is as follows:

$$\Phi_{\mathrm{RGB}}(F) = \mathrm{Conv}_{3\times3}^{\frac{1}{2}C \to C}(\mathrm{Conv}_{3\times3}^{C \to \frac{1}{2}C}(F)) \tag{5}$$

Finally, the network concatenates these three features with varying receptive fields—namely, the original detailed features F_1, the channel interaction features F_2 and the large-receptive-field enhanced features F_3 —along the channel dimension, at which point the feature dimensionality expands to $3C$. The concatenated features then undergo a 1×1 convolution for cross-channel information fusion and dimensionality alignment, yielding the module's final enhanced feature representation Y_{out}:

$$Y_{\mathrm{out}} = \mathrm{Conv}_{1\times1}(\mathrm{Concat}([F_1, F_2, F_3])) \tag{6}$$

(b) ThermalBlock

Unlike RGB imagery, thermal images in forest fire scenarios typically exhibit large-scale, low-frequency sparse backgrounds due to the lack of rich chromatic information and distinct edge contours; moreover, effective high-temperature anomalies tend to be highly localized. Such physical discrepancies imply that when the network relies solely on local, small receptive fields for feature extraction, it becomes highly susceptible to interference from extraneous heat sources within complex forest environments, thereby inducing false positive detections. Consequently, feature extraction within the thermal modality is more dependent on a large receptive field, necessitating extensive feature aggregation to suppress significant background thermal noise and achieve precise target localization. To address the challenges inherent in the thermal modality, we specifically designed the ThermalBlock for the thermal branch of the heterogeneous dual-stream backbone. To maintain consistency across feature scales and optimization paths within the heterogeneous architecture, the ThermalBlock adopts the same Efficient Layer Aggregation Network (ELAN) design as the RGBBlock. However, given the high sparsity of thermal images and the requirement for expansive receptive fields, we eschewed standard convolutions in favor of large-kernel depthwise convolutions. Large-scale convolutional kernels significantly broaden the effective receptive field, enabling the capture of more extensive global thermal distribution characteristics. This provides robust contextual support for distinguishing minute fire points from localized environmental thermal noise. Furthermore, thermal imagery primarily conveys monotonic temperature gradient information, lacking the intricate cross-channel chromatic and textural coupling found in visible-light images. Thus, the application of depthwise convolution aligns with the physical attributes of thermal features—which do not require exhaustive cross-channel interaction—while simultaneously averting the risk of exploding parameter counts and computational complexity associated with large-scale kernels. The mathematical expression for the ThermalUnit, the core component of the ThermalBlock, is as follows: set $F \in \mathbb{R}^{C \times H' \times W'}$ be an arbitrary intermediate feature tensor input to this module. The module first employs a 1×1 standard convolution for channel dimensionality reduction, followed by a 7×7 depthwise convolution to capture global thermal features in the spatial dimension. Finally, a 1×1 standard convolution is utilized to perform cross-channel information integration and dimensionality restoration.

$$\Phi_{\text{Thermal}}(F) = \text{Conv}_{1\times1}^{\frac{1}{2}C\to C}\left(\text{DWConv}_{7\times7}^{\frac{1}{2}C\to\frac{1}{2}C}\left(\text{Conv}_{1\times1}^{C\to\frac{1}{2}C}(F)\right)\right) \tag{7}$$

(c) Target-Enhanced Downsampling

In forest fire scenarios, fire points or incipient smoke typically occupy only a handful of pixels. Whether represented as weak textures in the visible spectrum or localized thermal anomalies in the infrared modality, the feature responses of these minute targets are exceptionally fragile. Convolution is fundamentally a weighted summation within a local receptive field, possessing a strong inherent smoothing effect. When processing expansive forest backgrounds, utilizing convolutions with a stride of 2 for spatial downsampling tends to cause the fragile high-frequency responses of early-stage fire points to be diluted by the surrounding massive background noise. This attenuation often leads to the disappearance of such features during layer-by-layer downsampling propagation, ultimately resulting in severe missed detection issues. To effectively mitigate the feature attenuation of small targets during the downsampling process, this paper proposes the Target-Enhanced Downsampling (TED) module. TED addresses the limitations of traditional downsampling by innovatively constructing a dual-branch parallel architecture and introducing an Adaptive Weighted Fusion (AWF) mechanism to regulate the integration of extremal and local features, thereby maximizing the preservation of fragile features for minuscule fire points. We set the input feature tensor to be $X \in \mathbb{R}^{C_{in}\times H\times W}$. The TED module first directs it into two functionally complementary downsampling paths. The first branch introduces a Max Pooling operation with a 2×2 kernel and a stride of 2. Max pooling is naturally sensitive to local feature extrema, effectively preserving the most salient high-frequency signals within the receptive field (such as the highest thermal points in the infrared modality or sharp smoke edges in the visible spectrum). Subsequently, a 1×1 standard convolution is applied for linear channel projection, yielding the feature tensor X_{pool}:

$$X_{\text{pool}} = \text{Conv}_{1\times1}^{C_{in}\to C_1}(\text{MaxPool}_{2\times2,s=2}(X)) \tag{8}$$

The second branch employs a 3×3 convolution with a stride of 2. While facilitating downsampling, this branch is responsible for smoothly capturing the local spatial context and semantic structural information surrounding minute targets. It effectively offsets the potential loss of local spatial structure inherent in the max pooling operation, yielding the feature tensor X_{conv}:

$$X_{\text{conv}} = \text{Conv}_{3\times3,s=2}^{C_{in}\to C_2}(X) \tag{9}$$

where the output channel counts of the two branches satisfy $C_{out} = C_1 + C_2$.

Conventional concatenation operations overlook the varying relative contributions of extremal and local spatial features across different fire scenarios (for instance, high-temperature fire points exhibit a greater dependence on extremal features, whereas diffuse smoke relies more heavily on spatial structural features). Consequently, this paper designs an AWF mechanism to perform weighted fusion on the outputs of the two branches, enabling the network to adaptively adjust the optimal ratio between max-pooling features and convolutional features. We set the learnable weight parameters as $w = [w_1, w_2]$, corresponding to the importance weights of X_{pool} and X_{conv}, respectively. Simultaneously, a learnable global scaling factor v is introduced to enhance the adaptability of the network representation. To ensure numerical stability and preclude extreme weight distributions, an infinitesimal constant $\epsilon = 0.0001$ is incorporated into the weight normalization process. The normalized weights are calculated as follows:

$$\alpha_i = v_i \cdot \frac{w_i}{\left(\sum_{j=1}^{n} w_j\right) + \epsilon} \tag{10}$$

The final downsampled output feature Y_{AWF} is then expressed as follows:

$$Y_{\mathrm{AWF}} = \mathrm{Concat}\left(\left[\alpha_1 \cdot X_{\mathrm{pool}}, \alpha_2 \cdot X_{\mathrm{conv}}\right]\right) \tag{11}$$

2.2.3. Multi-Scale Semantic Alignment Enhancement Branch

Although the heterogeneous two-stream backbone constructed in this paper can effectively extract the modal features of RGB and Thermal images, cross-modal modeling still faces severe challenges. Since RGB sensors capture reflected light, they predominantly focus on rich texture and color details. In contrast, infrared sensors capture thermal radiation, emphasizing the heat distribution and contours of the targets. The two modalities originate from distinct physical imaging mechanisms, which causes the two-stream features extracted by the backbone network to be inherently independent in the representation space, exhibiting an inherent semantic gap and spatial distribution shift. If these raw heterogeneous features are directly fed into the subsequent multi-modal fusion module without explicit alignment, feature conflicts and semantic interference will be inevitably triggered. Such a direct feature fusion approach shifts the burden of cross-modal alignment onto the fusion module, forcing it to enhance its representational capacity by increasing the number of parameters. This not only increases the computational overhead but also makes the network highly susceptible to overfitting on limited multi-modal samples. Furthermore, we consider the discrepancies in receptive fields and semantic abstraction among feature maps at different depths of the backbone network—for instance, shallow and mid-level features (C_3 layer) typically retain high-resolution spatial details, with semantic features primarily consisting of local characteristics such as low-level textures and edges, whereas deep-level features (C_4, C_5 layers) are characterized by low resolution, global semantics, and target category information [29]. Confronted with such cross-scale semantic variations, single-scale feature alignment struggles to simultaneously accommodate both global and local alignments. Therefore, we design the multi-scale semantic alignment enhancement (MSAE) branch. Through auxiliary supervision, explicit semantic alignment enhancement is performed prior to cross-modal feature fusion, and alignment constraints are simultaneously imposed across multiple scales. This approach alleviates the optimization burden on the fusion module, improves the quality of feature interaction, and ensures the effective alignment of multi-level semantics. The specific architectural design is illustrated in Figure 4.

Since the raw features $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ extracted by the backbone network exhibit highly modality-specific characteristics, directly computing the alignment loss within the original feature space not only impedes convergence but also threatens to impair the backbone network's representational capacity for fundamental modality-specific features. Consequently, a semantic projection head is designed to provide an enhanced non-linear representational capacity and to project the heterogeneous RGB and thermal features into a Shared Latent Semantic Space, where all subsequent alignment loss computations are performed. Specifically, the input features are initially processed by a 1×1 convolution to achieve channel expansion, and are subsequently split equally along the channel dimension:

$$\mathbf{X}_1, \mathbf{X}_2 = \mathrm{Split}(\mathrm{Conv}_{1\times1}(\mathbf{X})) \tag{12}$$

where the number of channels for both $\mathbf{X}_1$ and $\mathbf{X}_2$ is half that of the original output. Serving as an identity branch, $\mathbf{X}_1$ preserves the raw modality details at the current level to ensure

the integrity of the foundational representation, whereas $\mathbf{X}_2$ is fed into sequential projection layers for deep semantic mapping.

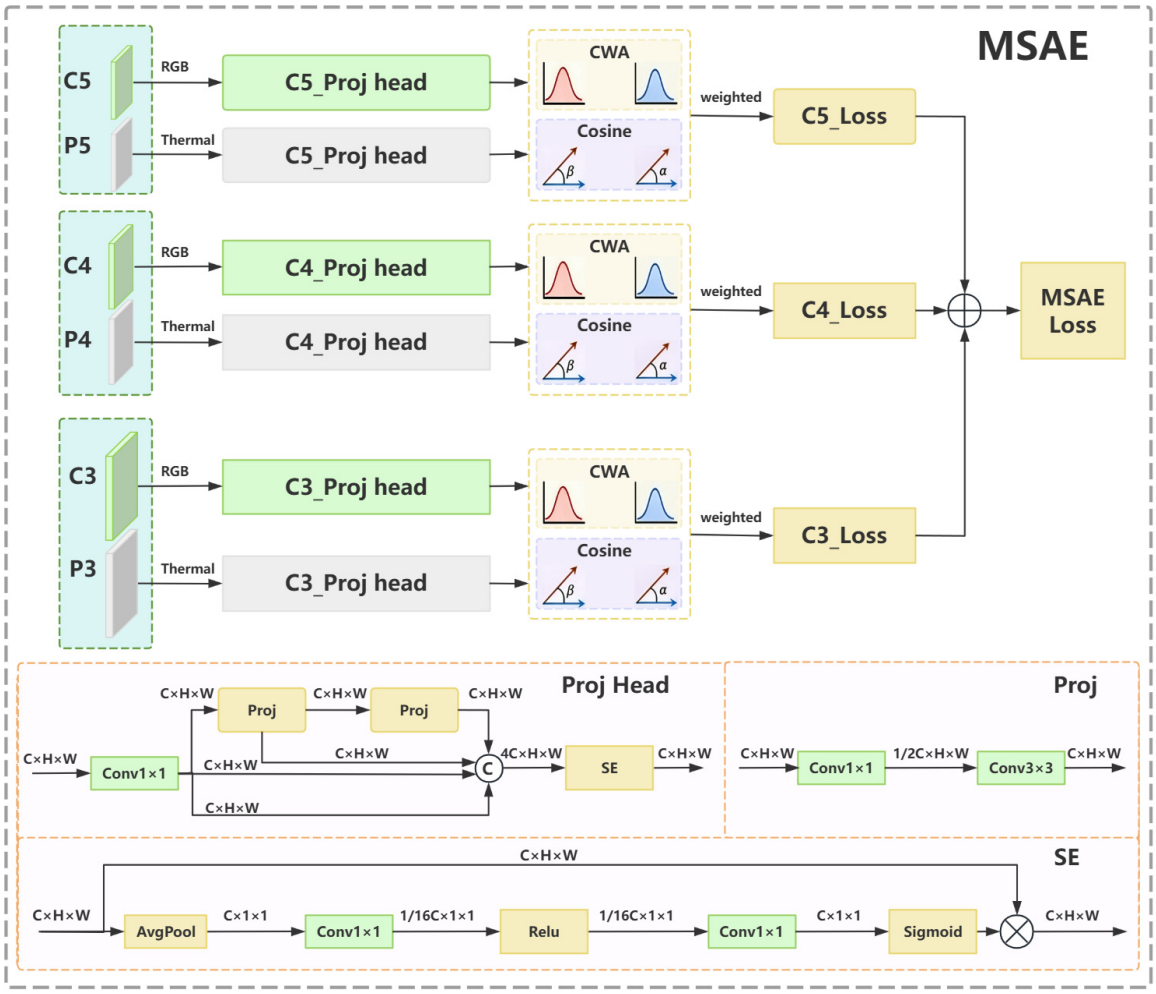


Figure 4. Multi-scale semantic alignment enhancement branch (MSAE) architecture diagram.

$$\mathbf{X}_3 = \mathrm{Proj}_1(\mathbf{X}_2) \tag{13}$$

$$\mathbf{X}_4 = \mathrm{Proj}_2(\mathbf{X}_3) \tag{14}$$

where $\mathbf{X}_3$ serves as a preliminary enhancement of $\mathbf{X}_2$, while $\mathbf{X}_4$ represents a further semantic abstraction of $\mathbf{X}_3$. This cascaded structure effectively simulates the progressive expansion of receptive fields within deep networks, thereby extracting higher-dimensional semantic representations. Subsequently, the aforementioned multi-level features are concatenated along the channel dimension, followed by feature fusion via convolution. Ultimately, to alleviate the potential information redundancy introduced by the cascading of multi-level features, the Squeeze-and-Excitation (SE) mechanism [30] is incorporated to modulate the aggregated features. This mechanism suppresses irrelevant background noise and accentuates the core semantic information crucial for cross-modal alignment, thereby generating the final projected feature, $\mathbf{X}_{proj}$, which is directly utilized for the computation of the alignment loss.

$$\mathbf{X}_{proj} = \mathrm{SE}(Conv_{1\times1}(Concat(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4))) \tag{15}$$

After mapping the RGB and Thermal features into a shared latent semantic space via the semantic projection heads, we note that local feature activations within a single image

often suffer from noise, making isolated pixel-level constraints insufficient to guarantee global semantic consistency. Inspired by the statistic alignment mechanisms in traditional domain adaptation and distribution metrics [31], we propose a Channel-level Wasserstein Distribution Alignment Loss (CWA). This loss models the spatial distribution of each channel of the projected features as an independent univariate Gaussian distribution. By constraining the discrepancies in the first-order statistics (mean) and second-order statistics (standard deviation) across channels, it achieves global alignment of the channel-wise marginal distributions of cross-modal features, effectively mitigating the representation collapse induced by modality heterogeneity in deep feature extraction. This strategy parallels the distribution matching concept in domain adaptation, where statistic alignment bridges the distribution shift between source and target domains. In this work, we apply it to bridge the modality gap between RGB and Thermal sensors from different physical sources, implementing cross-modal alignment with zero inference overhead in the form of auxiliary supervision. Let the projected features of the RGB and thermal modalities be denoted as $\mathbf{Z}_l^r, \mathbf{Z}_l^t \in \mathbb{R}^{C\times H\times W}$; the loss is then defined as follows:

$$L_{cwa}^{(l)} = \frac{1}{C}\sum_{c=1}^{C}\left[\left(\mu_{l,c}^r - \mu_{l,c}^t\right)^2 + \left(\sigma_{l,c}^r - \sigma_{l,c}^t\right)^2\right] \tag{16}$$

where $l \in \mathcal{A}$ represents the set of multi-scale feature layers (e.g., $C3$, $C4$, $C5$ in the backbone network), C denotes the total number of channels in the projected feature maps, $\boldsymbol{\mu}$ stands for the mean, and $\boldsymbol{\sigma}$ denotes the standard deviation, with the subscript l, c indicating the c-th channel of the l-th layer. Simultaneously with pixel-level semantic consistency upon the foundation of global distribution alignment, we introduce a cosine similarity loss [32]. By constraining the semantic direction of the two-stream features along the channel dimension, this loss circumvents the susceptibility of isolated pixel-level constraints to noise interference, thereby compelling the network to learn cross-modally consistent core semantic representations. The formulation is as follows:

$$\mathcal{L}_{cos}^{(l)} = 1 - \frac{1}{H\times W}\sum_{i=1}^{H\times W}\frac{\mathbf{Z}_{l,i}^r \cdot \mathbf{Z}_{l,i}^t}{\| \mathbf{Z}_{l,i}^r \|_2 \| \mathbf{Z}_{l,i}^t \|_2} \tag{17}$$

where $H \times W$ denotes the spatial resolution of the feature map, $\mathbf{Z}_{l,i}^r \cdot \mathbf{Z}_{l,i}^t$ represents the inner product of the feature vectors for the RGB and thermal modalities at the l-th layer and the i-th spatial position, and $\| \cdot \|_2$ signifies the L_2 norm. Ultimately, we perform joint alignment optimization across multiple scales, and the multi-scale semantic alignment enhancement loss, $\mathcal{L}_{MSAE}$, is formulated as follows:

$$\mathcal{L}_{MSAE} = \sum_{l\in\mathcal{A}}\left(w_{cwa}\cdot\mathcal{L}_{cwa}^{(l)} + w_{cos}\cdot\mathcal{L}_{cos}^{(l)}\right) \tag{18}$$

where w_{cwa} and w_{cos} denote the weight hyperparameters for the CWA loss and the Cosine loss, respectively. Notably, serving as an auxiliary supervision branch, the MSAE remains active exclusively during the training phase. During the inference phase, the entire branch can be completely decoupled, yielding zero computational overhead for inference and thereby ensuring the model's high efficiency during practical deployment.

2.2.4. Multimodal Spatial Gated Fusion Module

In UAV forest fire monitoring tasks, a severe scale imbalance exists between the extensive background and minute fire targets. Existing multi-modal fusion networks predominantly rely on global attention mechanisms; when computing dense attention matrices, they are highly susceptible to diluting or even overriding the fragile high-frequency local features of minute fire targets due to an overemphasis on redundant forest contextual

backgrounds. Furthermore, flame-like distractors in visible light images, coupled with prominent thermal spillover at target boundaries in infrared images, frequently induce erroneous semantic alignments during conventional cross-modal interactions. Such unconstrained feature fusion inevitably leads to the cross-amplification of interfering features from both modalities, severely degrading the model's detection capability in complex scenes. To address the aforementioned issues, this paper proposes a multi modal spatial gated fusion module, as illustrated in Figure 5. This module achieves high-quality feature fusion of the RGB and thermal modalities through localized precise filtering and enhancement via Spatial Gated Cross-Attention (SGCA), complemented by global contextual compensation via a Mish Gated Linear Unit (MishGLU).

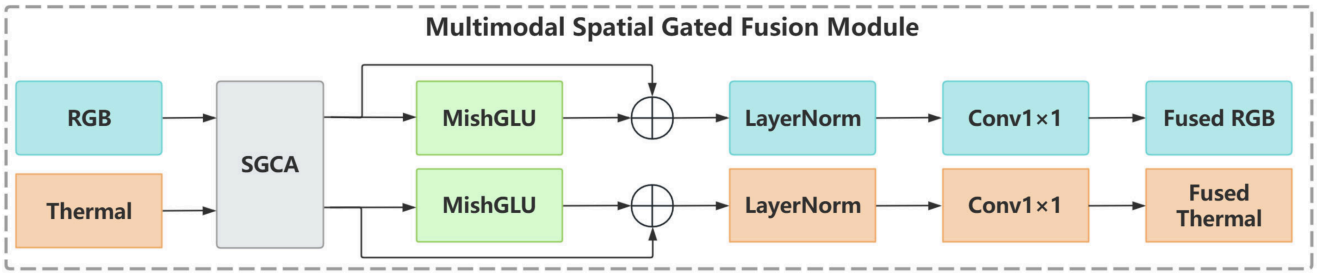


Figure 5. Multimodal spatial gated fusion module (MSGF) architecture diagram.

(a) Spatial Gated Cross Attention

The modality gap between RGB and thermal images is not as substantial as the drastic cross-modal discrepancy observed between text and images. Consequently, employing global attention computation introduces unnecessary computational overhead. Furthermore, it tends to attenuate crucial features due to the prevalence of extensive background noise, thereby compromising the actual fusion efficacy. Therefore, to overcome the inherent deficiency of traditional attention mechanisms—which are prone to diluting minute fire targets in typical forest fire scenarios—and to effectively suppress cross-modal noise between the RGB and thermal modalities, we propose a Spatial Gated Cross-Attention (SGCA) mechanism, as illustrated in Figure 6.

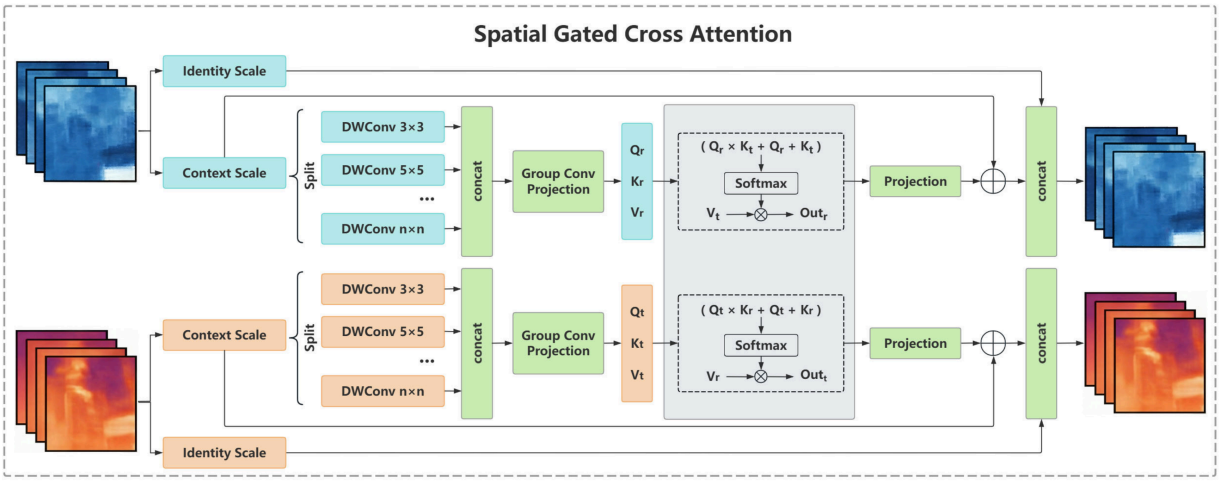


Figure 6. Spatial Gated Cross Attention (SGCA) architecture diagram.

This module discards the computationally intensive global attention mechanism, achieving local context enhancement and lightweight perception of the global context through the construction of multi-scale receptive fields. Simultaneously, by establishing a

cross-modal gating logic, the RGB and thermal modalities can mutually serve as guiding priors for semantic purification. This not only explicitly preserves the high response intensity of minute fire targets but also effectively suppresses the mutual amplification of modality-specific noise. Specifically, to meticulously retain the original high-frequency features of minute fire targets while suppressing background noise, SGCA initially employs a channel-splitting strategy that decouples identity mapping from multi-scale extraction. The set $X_{RGB}, X_{Thermal} \in \mathbb{R}^{C\times H\times W}$ denotes the input features for the RGB and thermal modalities, respectively. These features are partitioned into four segments along the channel dimension. The primary segment, $X_m^{(\text{identity})}$, is directly retained as an identity branch to prevent feature degradation or gradient vanishing of the fire targets during subsequent deep non-linear transformations. The remaining contextual features, $X_m^{(context)}$, are evenly divided into three subsets, which are subsequently fed into depthwise separable convolutions with varying receptive fields for local context extraction:

$$F_{m,k} = \text{DWConv}_k\left(X_{m,k}^{(context)}\right) \tag{19}$$

where $m \in \{RGB, Thermal\}$ denotes the distinct input modalities, and $k \in \{3, 5, 7\}$ represents the kernel sizes at varying scales. The multi-scale depthwise convolution compel the network to concentrate its attention within local receptive fields, thereby enabling the more precise capture of morphological boundaries and textural details of fire targets across various scales. Ultimately, the extracted multi-scale features are concatenated along the channel dimension to yield the local context feature F_m^{cat}:

$$F_m^{cat} = \text{Concat}(F_{m,3}, F_{m,5}, F_{m,7}) \tag{20}$$

After acquiring the multi-scale local features, the feature F_m^{cat} for modality m is projected via an efficient 1×1 grouped convolution to generate the Query, Key, and Value vectors:

$$Q_m, K_m, V_m = \text{Split}\big(\text{GroupConv}_{1\times1}\big(F_m^{cat}\big)\big) \tag{21}$$

Conventional cross-modal interactions typically compute the correlation matrix between Q and K directly in the spatial dimension, a practice that is highly susceptible to inducing erroneous semantic alignments across modalities in forest fire scenarios. To address this, SGCA computes cross-modal gating weights at the pixel level. Taking the absorption of thermal information by the RGB branch as an example, the computation of its Gating Activation Map, A_{RGB}, is formulated as follows:

$$A_{RGB} = \text{Softmax}(Q_{RGB} \odot K_{Thermal} + Q_{RGB} + K_{Thermal}) \tag{22}$$

where the Softmax is computed along the multi-scale dimension (i.e., the three receptive fields of 3×3, 5×5, and 7×7) and $\odot$ denotes element-wise multiplication. The multiplicative term $Q_{RGB} \odot K_{Thermal}$ accomplishes the joint activation of the RGB and thermal features at spatial locations. Meanwhile, the bias term $Q_{RGB} + K_{Thermal}$ serves as a residual compensation, preserving the independent confidence of individual modalities. The Softmax operation facilitates adaptive competition among distinct receptive fields, thereby leveraging the high-temperature thermal distribution of the infrared modality to dynamically guide and select the optimal scale response of the visible light features. Ultimately, the feature representation subsequent to cross-modal gating enhancement is formulated as follows:

$$\Delta_{RGB} = A_{RGB} \odot V_{Thermal} \tag{23}$$

Similarly, the thermal branch can utilize the RGB features for gating enhancement, yielding the following feature representation:

$$\Delta_{Thermal} = \text{Softmax}(Q_{Thermal} \odot K_{RGB} + Q_{Thermal} + K_{RGB}) \odot V_{RGB} \quad (24)$$

where the Softmax is computed along the multi-scale dimension (i.e., the three receptive fields of 3×3, 5×5, and 7×7), prompting adaptive competition among different receptive fields at each spatial location. Ultimately, the multi-scale features Δ_m, filtered via cross-modal gating, undergo a 1×1 convolutional projection for cross-channel information modulation. Simultaneously, to ensure stable network optimization, we perform a residual addition between these modulated features and the initial contextual features $X_m^{(context)}$. The result is then concatenated with the initially preserved identity branch $X_m^{(identity)}$ to generate the final output Y_m:

$$Y_m = \text{Concat}\left(X_m^{(identity)}, \text{Conv}_{1\times1}(\Delta_m) + X_m^{(context)}\right) \quad (25)$$

While preserving the independence of the two streams, this architecture achieves high-quality cross-modal complementarity and denoising at the local receptive field level.

(b) Mish Gated Linear Unit

To compensate for the deficiency of Spatial Gated Cross-Attention in comprehending the global context, we design the Mish Gated Linear Unit (MishGLU). Traditional Feed-Forward Networks (FFN) typically rely solely on two cascaded fully connected layers for pixel-wise non-linear mapping, thereby lacking the capacity for contextual perception of spatially adjacent pixels. Consequently, MishGLU introduces a large-kernel 9×9 depth-wise separable convolution upon the foundation of conventional non-linear mapping. By explicitly expanding the receptive field, MishGLU is capable of recapturing extensive global semantic contexts at a marginal parameter cost, subsequently employing a gated activation mechanism to perform further adaptive feature filtering. This structure is illustrated in Figure 7.

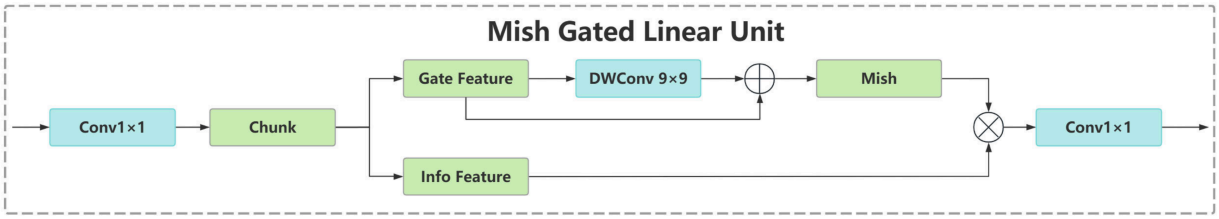


Figure 7. Mish Gated Linear Unit (MishGLU) architecture diagram.

Specifically, given the input feature $Z \in \mathbb{R}^{C\times H\times W}$, MishGLU initially utilizes a 1×1 convolution for dimensional expansion and subsequently splits the feature evenly along the channel dimension into two branches: one serving as the gating stream Z_1 and the other as the information stream Z_2:

$$Z_1, Z_2 = \text{Chunk}(\text{Conv}_{1\times1}(Z), 2) \quad (26)$$

In the gating stream branch, to capture long-range global dependencies without significantly increasing computational complexity, the module incorporates depthwise separable convolutions with a kernel size of 9×9. Such large-kernel convolutions can effectively aggregate global semantics across expansive forest backgrounds. Concurrently, to prevent the large-kernel convolution from blurring the sharp boundaries of tiny fire targets during feature

aggregation, a local residual connection is introduced here. Subsequently, the smooth and non-monotonic Mish activation function [33] is adopted for non-linear mapping. Compared with the conventional ReLU [34] or GELU [35] functions, the Mish activation function possesses a slight gradient in the negative value region, which enables the network to better preserve the gradient information of weak fire targets during backpropagation:

$$F_{Gate} = \text{Mish}(\text{DWConv}_{9\times9}(Z_1) + Z_1) \tag{27}$$

After obtaining the gating feature F_{Gate} with large-range semantic enhancement, the module utilizes it to perform element-wise gating modulation on the information stream Z_2. This allows the network to adaptively control which global contextual information should be propagated and which redundant backgrounds should be suppressed, based on the feature distribution of the other branch. Finally, a 1×1 convolution projects the gated feature back to the original channel dimension, yielding the final output Z_{out}:

$$Z_{out} = \text{Conv}_{1\times1}(F_{Gate} \odot Z_2) \tag{28}$$

2.2.5. Loss Function

The end-to-end training framework of this paper is established upon the set prediction paradigm [26]. To achieve high-precision object detection in complex multi-modal forest fire scenarios, we concurrently employ the Hungarian bipartite matching strategy, albeit with enhancements applied to the label assignment process. Ultimately, a joint objective function is formulated, incorporating both IoU-aware and multi-scale feature alignment constraints.

Given M ground-truth instances within an input image, let their class labels be denoted by c_i and their bounding boxes by b_i. The decoder yields a fixed number of N prediction queries, wherein the j-th prediction comprises a class probability p_j and bounding box coordinates $\hat{b}_j$. To establish the optimal one-to-one assignment permutation σ between the prediction set and the ground-truth set, the Hungarian algorithm is employed to minimize the overall matching cost matrix $\mathcal{C}(i,j)$. Considering the prevalent existence of minute targets with blurred edges in modalities such as infrared, conventional matching costs predicated on the L_1 distance or Intersection over Union (IoU) exhibit excessive sensitivity to marginal positional deviations. Consequently, the Normalized Wasserstein Distance (NWD) [36] is introduced as an auxiliary matching cost. By modeling the ground-truth and predicted boxes as two-dimensional Gaussian distributions, $\mathcal{N}_g$ and $\mathcal{N}_p$, their second-order Wasserstein distance $\mathcal{W}_2^2$ is computed as follows:

$$\mathcal{W}_2^2(\mathcal{N}_p, \mathcal{N}_g) = (cx_p - cx_g)^2 + (cy_p - cy_g)^2 + \frac{(w_p - w_g)^2}{4} + \frac{(h_p - h_g)^2}{4} \tag{29}$$

Based on $\mathcal{W}_2^2$, we can derive a normalized Gaussian distribution similarity and convert it into a distance matching cost $\mathcal{L}_{NWD}$:

$$\mathcal{L}_{NWD}(\hat{b}_j, b_i) = 1 - \exp\left(-\frac{\sqrt{\mathcal{W}_2^2}}{\tau}\right) \tag{30}$$

where τ is the scale smoothing constant for NWD. Ultimately, the overall matching cost $\mathcal{C}(i,j)$ is formulated as a weighted combination of classification loss, bounding box regression, and NWD cost:

$$\mathcal{C}(i,j) = \lambda_{cls}\mathcal{L}_{FL}(p_j, c_i) + \mathcal{L}_{bbox}\left(\hat{b}_j, b_i\right) + \lambda_{nwd}\mathcal{L}_{NWD}\left(\hat{b}_j, b_i\right) \tag{31}$$

where $\mathcal{L}_{FL}$ denotes the Focal Loss [37], employed to mitigate the imbalance between positive and negative samples. $\mathcal{L}_{bbox}\left(\hat{b}_j, b_i\right)$ represents the joint geometric cost for bounding boxes, constructed as a weighted sum of the $L1$ distance and the Generalized Intersection over Union (GIoU) [38] cost, thereby concurrently accounting for the absolute coordinate deviation and the scale overlap of the predicted boxes:

$$\mathcal{L}_{bbox}\left(\hat{b}_j, b_i\right) = \lambda_{box}\mathcal{L}_{L1}\left(\hat{b}_j, b_i\right) + \lambda_{giou}\mathcal{L}_{GIoU}\left(\hat{b}_j, b_i\right) \tag{32}$$

Upon determining the optimal assignment σ, the joint multi-task loss $\mathcal{L}_{total}$ is computed for each decoder layer. Given that the heterogeneous two-stream network constructed in this paper necessitates concurrently addressing end-to-end prediction outputs and the deep alignment of cross-modal features, the overall network loss function is predominantly formulated by combining the prediction loss $\mathcal{L}_{pred}$ and the aforementioned multi-scale semantic alignment enhancement loss $\mathcal{L}_{MSAE}$:

$$\mathcal{L}_{total} = \sum_{l=1}^{D} \mathcal{L}_{pred}^{(l)} + \mathcal{L}_{MSAE} \tag{33}$$

Within $\mathcal{L}_{pred}^{(l)}$ for the l-th decoder layer, conventional classification confidence frequently fails to authentically reflect the localization quality of bounding boxes. To mitigate this critical inconsistency between classification and localization accuracy, the prediction loss not only incorporates traditional classification and bounding box constraints but also specifically introduces an IoU-aware loss

$$\mathcal{L}_{pred}^{(l)} = \sum_{i=1}^{M}\left[\lambda_{cls}\mathcal{L}_{FL}(p_{\sigma(i)}^{(l)}, c_i) + \mathcal{L}_{bbox}(\hat{b}_{\sigma(i)}^{(l)}, b_i) + \lambda_{iou}\mathcal{L}_{BCE}(iou_{\sigma(i)}^{(l)}, \hat{IoU}_i)\right] \tag{34}$$

where $\mathcal{L}_{FL}$ denotes the Focal Loss employed to alleviate class imbalance, $iou_{\sigma(i)}^{(l)}$ signifies the IoU score yielded by the network's dedicated prediction head, and $\hat{IoU}_i$ represents the actually computed Intersection over Union between the predicted box and the ground-truth box. These two are subjected to regression constraints via the binary cross-entropy loss $\mathcal{L}_{BCE}$. Additionally, $\mathcal{L}_{bbox}$ represents the joint bounding box regression loss, which is formulated as a weighted combination of the $L1$ distance and the Generalized Intersection over Union (GIoU) loss:

$$\mathcal{L}_{bbox}\left(\hat{b}, b\right) = \lambda_{box}\mathcal{L}_{L1}\left(\hat{b}, b\right) + \lambda_{giou}\mathcal{L}_{GIoU}\left(\hat{b}, b\right) \tag{35}$$

Regarding $\mathcal{L}_{MSAE}$ (for the detailed definition, refer to Section 2.2.3), to bridge the physical semantic gap between the RGB and thermal modalities, it utilizes feature-level statistics and directional constraints to execute explicit alignment of the two-stream features. Serving as an auxiliary supervision branch, this loss is computed synchronously during joint optimization and is entirely decoupled during the inference phase, thereby achieving performance enhancement with zero inference overhead.

The loss hyperparameters and decoder configurations of this paper are as follows. Regarding training losses, in the matching cost, the weights for classification, L1 bounding box, GIoU, and NWD are 2.0, 5.0, 3.0, and 2.0, respectively. In the joint training loss, the weights for classification, bounding box, GIoU, IoU-Aware, and MSAE alignment loss are 1.0, 8.0, 5.5, 1.0, and 1.0, respectively. The α and γ for Focal Loss are 0.25 and 2.0, respectively. The scale smoothing constant τ for NWD is 12.8. In MSAE, the weights for both the CWA loss and the cosine loss are 0.5.

2.2.6. Evaluation Metrics

To comprehensively and objectively evaluate the performance differences between FireRGBTNet and other models in complex forest fire scenarios, this paper adopts multiple evaluation metrics, including Precision (P), Recall (R), F1-score (F1), mean Average Precision (mAP), Frames Per Second (FPS), Parameters, and GFLOPs.

Specifically, Precision (P) reflects the proportion of true fire samples among those detected as fire. Recall (R) reflects the proportion of actual fire samples that are correctly detected. The F1-score (F1) is the harmonic mean of Precision (P) and Recall (R), reflecting the overall performance of the model after balancing false alarms and missed detections. Mean Average Precision (mAP) reflects the Average Precision (AP) of the model under different Intersection over Union (IoU) thresholds, evaluating the comprehensive performance of the model in localization and classification. Frames Per Second (FPS) reflects the number of images the model can process per second. Parameters denote the total volume of all learnable parameters in the model, which directly determines the physical storage and memory space required for model deployment and runtime. GFLOPs is used to measure the computational complexity of the model during forward inference, intuitively reflecting the model's demand for underlying hardware computing power.

It is worth noting that for mAP, we primarily adopt mAP@0.5, mAP@0.75, and mAP@0.5:0.95 as the main evaluation criteria. mAP@0.5 represents the average precision at an IoU threshold of 0.5. mAP@0.75 represents the average precision at an IoU threshold of 0.75. mAP@0.5:0.95 represents the mean of the average precisions calculated at IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05.

3. Experiments and Results

3.1. Experimental Setup

The detailed information of the experimental environment is presented in Table 3. To ensure sufficient convergence of all models in the experiments, the number of epochs is set to 250. Regarding the optimization strategy, this paper employs the AdamW optimizer for parameter updates, with the initial base learning rate set to 1×10^{-3}, the weight decay coefficient set to 1×10^{-4}, the first-moment estimation decay coefficient $\beta 1$ set to 0.9, the second-moment estimation decay coefficient $\beta 2$ set to 0.999, and ϵ set to 1×10^{-8}. To enhance the stability of the model during the early training phase and ensure convergence to a better local minimum in later stages, this paper devises a joint learning rate scheduling strategy. Specifically, linear warm-up is performed during the first 2% of the total training steps, allowing the learning rate to smoothly increase to the initially set value. Subsequently, the learning rate gradually decays following the Cosine Annealing schedule, with its minimum learning rate constrained to 0.1% of the initial learning rate. The specific hyperparameter configurations are detailed in Table 4.

Table 3. Experimental environment.

Configuration	Details
Operation system	CentOS7
CPU	Intel(R) Xeon(R) Silver 4116 @ 2.10 GHz
GPU	Tesla V100
Framework	Pytorch 2.4.1 + cu118

Table 4. Hyperparameter configurations.

Parameters	Details
Input Image Size	640 × 640
Epochs	250
Batch Size	32
Workers	16
Initial Learning Rate	0.001
Minimum Learning Rate	1×10^{-6}
LR Scheduler	Cosine Annealing
Warm-up Strategy	Linear(First 2% steps)
Optimizer	AdamW
β1	0.9
β2	0.999
ε	1×10^{-8}
Weight Decay	0.0001

3.2. Ablation Experiment

To verify the effectiveness of each key component in FireRGBTNet and their practical contributions to the detection performance in complex forest fire scenarios, this paper conducts a series of ablation studies on the RGBT-3M dataset. Considering that this paper proposes a novel heterogeneous end-to-end network, to ensure the fairness and interpretability of the experiments, we first construct a baseline model. This baseline adopts a dual-stream architecture, yet both branches uniformly employ basic RGBBlocks stripped of the TED module for feature extraction, and utilize simple 1 × 1 standard convolutions to replace complex attention-based fusion mechanisms during the feature fusion stage. The results of the ablation experiments are presented in Table 5.

Table 5. Results of model ablation experiments.

Model	P	R	F1	mAP@0.5	mAP@0.75	mAP@0.5:0.95	FPS	Parameters (M)	FLOPs (G)
Baseline	87.6	89.3	88.4	93.3	60.4	57.3	54.2	3.07	8.79
+ThermalBlock	88.4	90.9	89.6	93.8	62.5	57.9	54.3	2.91	8.51
+TED	89.7	90.2	90.0	94.1	62.1	57.8	46.1	2.87	7.88
+MSGF	91.7	93.8	92.7	95.4	66.9	61.5	44.7	4.13	13.50
+MSAE	91.8	94.7	93.2	95.9	68.9	62.8	44.7	4.13	13.50

The baseline model, due to its utilization of homogeneous RGBBlocks for processing infrared images, overlooks the low-frequency sparse characteristics inherent in the infrared modality, which leads to unnecessary computational redundancy and constrains the capability to capture heat source targets. Upon substituting the infrared branch with the ThermalBlock—specifically engineered for thermal features—the model's mAP@0.5 increased from 93.3% to 93.8%, while the mAP@0.75 exhibited a significant improvement of 2.1%. Notably, as the ThermalBlock employs large-kernel depthwise separable convolutions in place of standard convolutions, the parameter count decreased from 3.07 M to 2.91 M, accompanied by a reduction in computational complexity. This demonstrates that the ThermalBlock not only provides an expanded receptive field to suppress infrared background noise but also effectively mitigates the computational redundancy introduced by the homogeneous design. Following the integration of the TED module, the model's Precision and F1-score rose to 89.7% and 90.0%, respectively. Although the dual-branch architecture introduces a certain degree of parallel scheduling overhead, resulting in minor fluctuations in FPS, the required theoretical computational complexity and parameter count

are further reduced. More importantly, the TED module maximizes the preservation of extreme features via AWF, effectively mitigating the attenuation of minute early-stage forest fire points during the downsampling process, thereby enhancing the model's capacity to capture high-frequency weak signals.

Following the optimization of the heterogeneous feature extraction, the original simplistic Conv fusion layer was replaced by the MSGF module. As illustrated in Table 5, this modification yielded the most pronounced performance enhancement, with the model's mAP@0.5 increasing to 95.4%, mAP@0.75 at high thresholds rising substantially by 4.8%, and Recall reaching 93.8%. This provides robust evidence that in complex forest scenarios, simple linear superposition or concatenation is insufficient to resolve inter-modal conflicts. Conversely, MSGF effectively suppresses interference in the visible spectrum and the "thermal spillover" phenomenon in the infrared modality through SGCA, while concurrently supplementing global semantics via MishGLU. Although the introduction of MSGF increased the parameter count to 4.13 M, the resulting precision gains achieve an optimal balance between computational overhead and accuracy.

Finally, we integrate MSAE into the network. Experimental results demonstrate that the incorporation of MSAE elevates all metrics of the final model to optimal levels. In particular, the Recall is further boosted to 94.7%, and the mAP@0.5:0.95 is improved to 62.8%. This indicates that performing explicit distribution and direction constraints on heterogeneous features prior to fusion can substantially mitigate the modality semantic gap, alleviate the optimization burden on the MSGF module, and thereby further enhance the complementary capability of cross-modal features. Most crucially, comparing the data before and after introducing MSAE reveals that the FPS, Params, and FLOPs of the final model remain entirely unchanged. This validates the structural superiority of MSAE as an auxiliary supervision branch: it significantly improves the quality of feature alignment during the training phase, while being completely discarded during inference, thereby perfectly accommodating the deployment requirements of resource-constrained UAV edge computing platforms.

To validate the feature extraction advantages of the heterogeneous dual-stream backbone network designed in this paper in complex forest fire scenarios, we conducted comparative replacement experiments against current mainstream feature extraction backbones in the computer vision field (all alternative backbones were instantiated in a dual-stream manner). The comparison benchmarks include the classic convolutional neural network ResNet [39], lightweight networks MobileNet [40] and EfficientNet [41], YOLOv11 [42], the large-kernel convolutional network ConvNeXt [43], and the self-attention-based Swin Transformer [44]. The experimental results are presented in Table 6.

Table 6. Results of backbone comparison experiments.

Backbone	P	R	F1	mAP@0.5	mAP@0.75	mAP@0.5:0.95	FPS	Parameters (M)	FLOPs (G)
ConvNeXt-Tiny	91.0	92.2	91.6	94.8	63.9	59.2	25.4	58.14	155.45
MobileNetV3-small	91.3	93.0	92.1	95.1	64.2	60.0	38.0	5.91	7.61
ResNet-18	91.8	94.3	93.0	95.9	68.6	62.5	39.4	24.66	68.40
Swin-T	87.8	87.7	87.7	92.8	60.3	56.2	21.6	40.22	106.92
YOLO11	89.5	91.6	90.6	94.3	63.5	59.1	36.2	5.16	15.25
EfficientNet	90.6	93.6	92.1	95.5	67.3	61.9	29.7	9.38	21.63
Ours	91.8	94.7	93.2	95.9	68.9	62.8	44.7	4.13	13.50

When Swin-T was adopted as the backbone, the model's metrics experienced a significant decline, with mAP@0.5 reaching only 92.8%. This is attributable to the fact that Transformers lack the local inductive bias inherent in convolutional neural networks, mak-

ing it difficult to effectively focus on local high-frequency features when confronted with the extremely minute targets characteristic of early-stage forest fires. In contrast, the heterogeneous backbone designed in this study, integrated with the TED module, achieved a Recall of 94.7%, demonstrating its absolute advantage in preserving the fragile features of small-scale targets. While the classical ResNet achieved an identical mAP@0.5 of 95.9% to our network, this performance came at the expense of an enormous parameter count (24.66 M) and substantial computational overhead (68.40 G). On the other hand, although lightweight networks such as MobileNet and YOLOv11 reduced the computational burden, their feature extraction capabilities remained constrained, with their mAP@0.5:0.95 lagging behind our method by 2.8% and 3.7%, respectively. The proposed heterogeneous backbone, through its differential extraction and efficient layer aggregation design, surpasses massive networks like ConvNeXt and ResNet using only 4.13 M parameters and 13.50 G FLOPs, thereby manifesting the architectural superiority of our backbone network.

To further demonstrate the superiority of the proposed SGCA in RGB-T feature fusion, we replaced SGCA with multiple mainstream feature fusion mechanisms and attention modules for comparative testing. All compared schemes retain MishGLU and MSAE, with only the fusion attention module replaced. The alternatives mainly include 3×3 convolution, standard cross-attention [45], iAFF [46], CBAM [47], CoordAtt [48], ECA [49], and SE [30]. The experimental results are shown in Table 7.

Table 7. Results of fusion comparison experiments.

Method	P	R	F1	mAP@0.5	mAP@0.75	mAP@0.5:0.95	FPS	Parameters (M)	FLOPs (G)
Conv	91.8	93.1	92.5	95.3	66.1	61.1	49.4	13.23	55.56
Cross-attention	88.1	90.5	89.3	94.0	62.7	58.1	21.9	4.21	14.03
SE	89.8	91.2	90.5	94.2	64.2	59.3	40.0	3.97	12.57
ECA	90.6	91.8	91.2	94.6	64.0	59.1	42.3	8.21	32.41
CoordAtt	90.2	91.6	90.9	94.4	64.2	59.5	48.9	8.27	32.43
CBAM	91.9	93.5	92.7	95.4	66.7	61.0	37.5	3.97	12.57
iAFF	91.9	93.3	92.6	95.4	67.1	61.7	44.7	3.48	9.31
Ours	91.8	94.7	93.2	95.9	68.9	62.8	44.7	4.13	13.50

Employing a plain convolution for fusion merely performs a singular mapping on the feature space; its parameter count expands to 13.23 M and FLOPs surge to 55.56 G, while it struggles to handle inter-modal conflicts. Notably, standard global cross-attention yields the poorest performance, with mAP@0.5 at only 94.0% and a sharp decline in inference speed to 21.9 FPS. This substantiates that within expansive forest backgrounds, global computation not only incurs a significant computational burden but also leads to the severe dilution of minute target features by redundant background noise. Furthermore, although attention mechanisms such as iAFF and CBAM enhance fusion performance to a certain degree—both achieving a mAP@0.5 of 95.4%—they are not specifically engineered to address the modality discrepancies and physical interference issues unique to RGB-T data, thus trailing behind our SGCA. The SGCA module leverages inter-modal gating logic to establish mutual guiding priors, which effectively strengthens semantics and suppresses the amplification of cross-modal noise, ultimately reaching 68.9% under the stringent mAP@0.75 evaluation criterion, surpassing all other approaches. This fusion strategy, specifically customized for forest fire scenarios, achieves high-quality cross-modal feature interaction while preserving the model's lightweight characteristics.

To intuitively validate the superiority of the proposed SGCA module over other mainstream fusion mechanisms and attention architectures (including 3×3 convolution, standard cross-attention, SE, ECA, CoordAtt, CBAM, and iAFF) in terms of feature per-

ception and cross-modal noise suppression, we generated comparative heatmaps using Grad-CAM [50]. As illustrated in Figures 8 and 9, these heatmaps visually present the spatial attention allocation intensity of the model on multispectral image pairs under various typical forest fire challenge scenarios (e.g., nighttime low-illumination fires, complex terraced-field backgrounds, and dense smoke occlusion). Red denotes high-attention regions, while blue indicates suppressed low-attention regions. The visualization comparison clearly demonstrates that, in nighttime low-illumination and small-target scenarios (rows a, b), conventional modules such as CBAM induce severe global pseudo-activation due to their inability to filter thermal infrared sensor noise, whereas FireRGBTNet precisely isolates and locks onto the core ignition points. When confronted with complex terrains like terraced fields and thin smoke (rows c, d), mechanisms such as SE and iAFF are easily misled by background structural contours, producing large scattered pseudo-activation patches; in contrast, our model effectively filters out these structural interferences by virtue of the spatially gated cross-attention. Under severe dense-smoke occlusion scenarios (rows e, f), the attention of global cross-attention completely deviates to non-target backgrounds and image edges, whereas FireRGBTNet, benefiting from the cross-modal guidance mechanism, successfully penetrates the thick smoke layer and accurately localizes active flame fronts in both visible-light and infrared modalities. In summary, this visual evidence sufficiently confirms that FireRGBTNet effectively overcomes the catastrophic background noise amplification observed in other attention mechanisms, and while precisely capturing complex flame edge features, provides solid visual interpretability for high-precision fire detection in real and harsh scenarios.

Furthermore, to investigate the operational mechanism of MSAE across different feature hierarchies in depth, detailed scale-sensitivity ablation experiments were conducted on the three scales' (C3, C4, and C5) output by the heterogeneous backbone network. These experiments encompassed configurations ranging from non-alignment and single-scale alignment to dual-scale combinations and full-scale alignment. The experimental results are summarized in Table 8.

Table 8. Results of multi-scale sensitivity ablation experiments.

Model	P	R	F1	mAP@0.5	mAP@0.75	mAP@0.5:0.95
No MSAE	91.7	93.8	92.7	95.4	66.9	61.5
MSAE-1_C3	89.7	91.7	90.7	94.7	64.0	59.3
MSAE-1_C4	91.0	93.6	92.3	95.3	65.9	60.7
MSAE-1_C5	91.4	93.6	92.5	95.4	66.0	61.1
MSAE-2_C3C4	92.1	94.2	93.2	95.5	66.7	61.4
MSAE-2_C3C5	93.1	93.8	93.5	95.6	67.0	61.7
MSAE-2_C4C5	91.5	94.4	93.0	95.5	66.5	61.3
MSAE-3_C3C4C5 (Ours)	91.8	94.7	93.2	95.9	68.9	62.8

Experimental results for single-scale alignment indicate that when alignment is performed exclusively at the deeper levels of C4 or C5, model performance remains comparable to or even slightly lower than that of the baseline. More notably, when the MSAE constraint is applied solely to the shallow features (C3), all performance metrics exhibit significant degradation, with mAP@0.5 dropping from 95.4% to 94.7%. Given that shallow networks primarily extract basic modal features characterized by strong physical attributes, forcibly aligning the statistical distribution and orientation of these highly divergent low-level heterogeneous features—in the absence of high-level semantic guidance—not only fails to bridge the semantic gap but also undermines the backbone's representational capacity for modality-specific features, thereby introducing severe feature conflicts and noise into the subsequent cross-modal fusion process. As the alignment strategy is extended to dual-scale configurations, such as C3 + C5 or C3 + C4, performance begins to steadily surpass the

baseline. This suggests that the inclusion of high-level semantics from C5 provides the necessary directional guidance for shallow feature alignment, alleviating the structural disruption caused by forced alignment at lower levels. Ultimately, the model achieves peak performance when full-scale alignment (C3 + C4 + C5) is implemented. This result provides robust evidence that effective cross-modal alignment cannot rely on an isolated receptive field hierarchy; rather, high-quality alignment within a shared latent semantic space is only attainable through the multi-level synergy and joint constraint of spatial details (C3), local context (C4), and global semantics (C5).

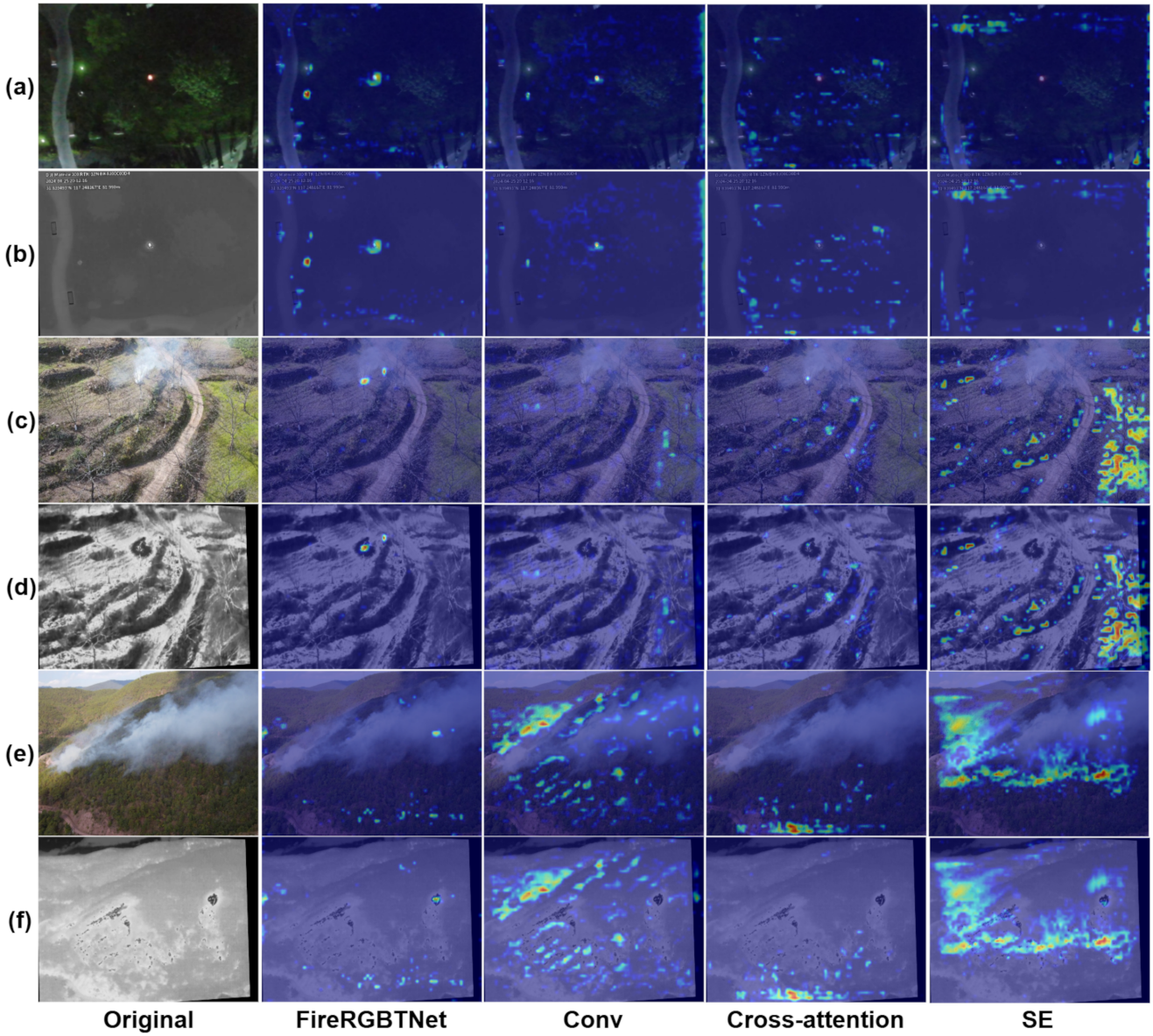


Figure 8. Comparative analysis of heatmap visualizations across different models under typical forest fire challenge scenarios. Subfigures (**a**–**f**) are organized into three groups of RGB-Thermal pairs: (**a**,**b**), (**c**,**d**), and (**e**,**f**). Specifically, (**a**,**b**) correspond to Nighttime low-illumination fires; (**c**,**d**) correspond to Complex terrain and field backgrounds; and (**e**,**f**) correspond to Dense smoke occlusion. Columns from left to right: Original input, FireRGBTNet (Ours), Conv, Cross-attention, and SE. In the heatmaps, red indicates high-attention regions corresponding to fire signatures, while blue represents low-attention areas.

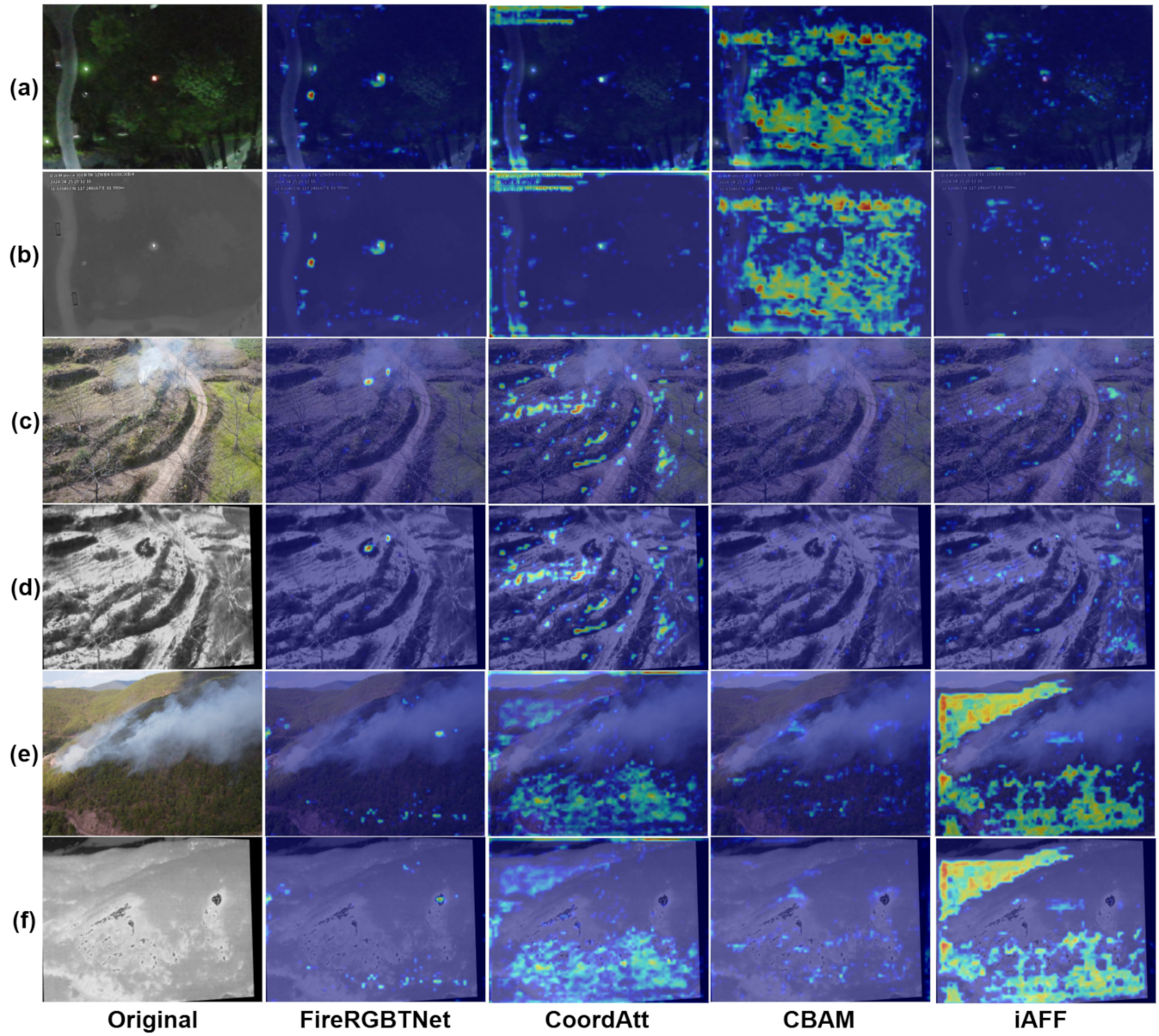


Figure 9. Comparative analysis of heatmap visualizations across different models under typical forest fire challenge scenarios. Subfigures (**a**–**f**) are organized into three groups of RGB-Thermal pairs: (**a**,**b**), (**c**,**d**), and (**e**,**f**). Specifically, (**a**,**b**) correspond to Nighttime low-illumination fires; (**c**,**d**) correspond to Complex terrain and field backgrounds; and (**e**,**f**) correspond to Dense smoke occlusion. Columns from left to right: Original input, FireRGBTNet (Ours), CoordAtt, CBAM, and iAFF. In the heatmaps, red indicates high-attention regions corresponding to fire signatures, while blue represents low-attention areas.

3.3. Comparative Experiment

To comprehensively evaluate the overall performance of the proposed FireRGBT-Net in complex forest fire scenarios, extensive comparative experiments were conducted against a variety of state-of-the-art object detection algorithms in the computer vision field. The unimodal object detection models primarily include the classic two-stage model Faster R-CNN [51], the one-stage model SSD [52], lightweight detection models such as PicoDet [53] and YOLO11, the Transformer-based RT-DETR model, and the unimodal LHRF-YOLO [54] specifically designed for fire scenarios. The multimodal object detection models mainly consist of ICAFusion [19] utilizing cross-fusion, YOLOv5-CFT [55] incorporating a cross-modal Transformer fusion mechanism, the latest multimodal detection

architectures YOLO-RGBT [56] and RT-DETR-RGBT [56], and CP-YOLO11-MF proposed in the original paper of the RGBT-3M dataset. To ensure a fair comparison, the unimodal models were independently trained and tested on RGB and thermal images, respectively. The detailed experimental results are presented in Table 9.

Table 9. Results of comparison experiments.

Model	Mode	P	R	F1	mAP@ 0.5	mAP@ 0.75	mAP@ 0.5:0.95	FPS	Parameters (M)	FLOPs (G)
Faster R-CNN	RGB	84.6	95.2	89.6	93.7	58.7	56.6	38.6	41.36	182.17
	Thermal	81.8	92.1	86.6	91.1	58.1	54.9	38.6	41.36	182.17
SSD	RGB	90.0	88.5	89.2	87.5	48.9	49.8	98.2	24.01	61.05
	Thermal	86.7	84.9	85.8	85.1	46.9	47.5	98.2	24.01	61.05
YOLOv11	RGB	91.5	91.6	91.6	94.4	67.1	62.5	73.1	9.41	21.30
	Thermal	91.5	88.2	89.8	93.5	67.2	61.5	73.1	9.41	21.30
RT-DETR	RGB	91.2	91.2	91.2	93.2	61.8	58.6	12.5	41.94	125.6
	Thermal	87.2	80.7	83.8	85.9	49.5	49.2	12.5	41.94	125.6
PicoDet	RGB	91.2	85.0	88.0	91.3	57.6	55.0	59.7	3.78	51.81
	Thermal	89.3	81.3	85.1	89.3	57.0.	53.0	59.7	3.78	51.81
LHRF-YOLO	RGB	92.2	87.9	90.0	93.3	62.9	59.5	39.2	2.26	5.40
	Thermal	91.0	85.1	87.9	92.1	63.5	58.5	39.2	2.26	5.40
RT-DETR-RGBT	RGB-T	91.2	90.2	90.7	92.8	61.6	58.3	11.3	44.9	158.99
YOLO-RGBT	RGB-T	92.4	90.5	91.4	94.2	66.4	61.8	69.6	9.86	28.50
ICAFusion	RGB-T	92.0	93.1	92.6	95.3	63.0	58.7	42.4	20.15	29.77
YOLOv5-CFT	RGB-T	93.7	94.4	94.0	95.8	69.5	63.1	22.5	205.98	223.49
CP-YOLO11-MF	RGB-T	92.5	93.5	93.0	96.3	-	62.9	-	11.83	-
Ours	RGB-T	91.8	94.7	93.2	95.9	68.9	62.8	44.7	4.13	13.50

Experimental results demonstrate that both the classical Faster R-CNN and the state-of-the-art YOLOv11 exhibit pronounced performance bottlenecks when addressing complex forest fires. Models based on RGB are highly susceptible to severe missed detections under conditions of dense smoke obscuration or low-illumination nocturnal environments. Conversely, while Thermal models can penetrate smoke, their inherent lack of textural details frequently leads to false positives, where heated rocks or specular reflections are misidentified as fire sources. In contrast, multimodal detection models—including the method proposed in this study—achieve substantial improvements across mAP@50 and F1-score metrics. This fully corroborates that in forest fire monitoring tasks, transcending the imaging constraints of individual physical sensors by leveraging the complementary characteristics of RGB and thermal information is an essential strategy for enhancing detection robustness.

Furthermore, although extant multimodal models have attained high detection accuracy, they often incur substantial computational costs. For instance, YOLOv5-CFT incorporates highly complex Transformer modules for cross-modal interaction; while effective in feature fusion, this leads to a sharp escalation in parameter count and computational complexity, alongside a significant decline in FPS, thereby failing to satisfy the stringent real-time requirements of UAV edge computing platforms. Conversely, while YOLO-RGBT attempts to balance efficiency and performance within a multimodal framework, its Recall and performance under the rigorous mAP@0.5:0.95 metric remain suboptimal when detecting minute early-stage fire points, primarily due to the absence of explicit modal alignment and feature preservation mechanisms. Moreover, compared to the CP-YOLO11-MF model originally proposed alongside the dataset, conventional fusion models fail to demonstrate a decisive advantage, which indirectly underscores the limitations of generalized fusion modules when applied to specialized forest fire scenarios.

Compared with all the aforementioned algorithms, the proposed FireRGBTNet achieves the optimal trade-off between accuracy and efficiency. It maintains the lowest parameter count among multimodal models—approaching the scale of the most compact unimodal models—and demonstrates superior overall performance. Regarding detection accuracy, the leading performance of FireRGBTNet across all metrics is primarily attributed to the proposed MSAE, which aligns semantic information across heterogeneous features, and the MSGF, which effectively suppresses multimodal interference noise, thereby enabling the precise capture and fusion of key fire target features. More importantly, in terms of lightweight design and real-time capability, owing to the customized design of the heterogeneous dual-stream backbone for each modality and the superior auxiliary role of the MSAE module, FireRGBTNet achieves high precision with only 4.13 M parameters and 13.5 G FLOPs. Consequently, these results not only significantly surpass the existing multimodal models but also demonstrate efficiency comparable to most lightweight unimodal models.

Simultaneously, to provide a more intuitive demonstration of the feature representation capability and cross-modal balance of the proposed FireRGBTNet in complex environments, we selected three typical fire scenarios from the RGBT-3M dataset to compare the detection performance of FireRGBTNet with other models, as illustrated in Figure 10. Specifically, (a) and (b) represent nighttime scenarios with interfering light sources on the left; (c) and (d) depict wild grassland scenes where thick smoke obscures both flames and personnel; and (e) and (f) show mountainous forest environments characterized by widespread smoke diffusion and partially obscured flames. In (a) and (b), although all models successfully detected the fire points, their confidence scores exhibited significant variations. Certain models that rely excessively on visible light features show suppressed responses to target boundaries when faced with dim backgrounds and halo interference, resulting in relatively low confidence scores. For targets with extremely high thermal radiation, the large-scale YOLOv5-CFT model achieved remarkably high confidence scores. In contrast, the confidence scores of the proposed FireRGBTNet for these fire targets are significantly higher than those of most models, demonstrating robust detection capabilities for high-heat targets and marginally trailing only YOLOv5-CFT in numerical terms. In (c) and (d), the unimodal YOLOv11 and the spatial-feature-oriented YOLO-RGBT yielded very high confidence scores when detecting smoke, which possesses morphological diffusion characteristics. However, due to smoke obscuration, the confidence scores for the occluded flames and personnel experienced significant attenuation. Conversely, while most complex cross-modal fusion models are capable of capturing fire points effectively, their confidence levels when processing diffuse smoke edges are far inferior to those of unimodal models; FireRGBTNet, however, demonstrates superior cross-modal proficiency in this scenario. In (e) and (f), amidst vast smoke backgrounds, most models exhibited varying degrees of decline in confidence scores when detecting faint high-temperature fire points at the bottom. This is primarily because, during the deep downsampling process of conventional networks, the local high-frequency signals of weak fire points are diluted by the extensive background context; nevertheless, FireRGBTNet maintains a very high confidence level for such minute fire points. In these three scenarios, the visual performance of FireRGBTNet strongly validates the effectiveness of the proposed modules, achieving the most robust and balanced performance in complex and volatile forest environments with an extremely lightweight architecture.

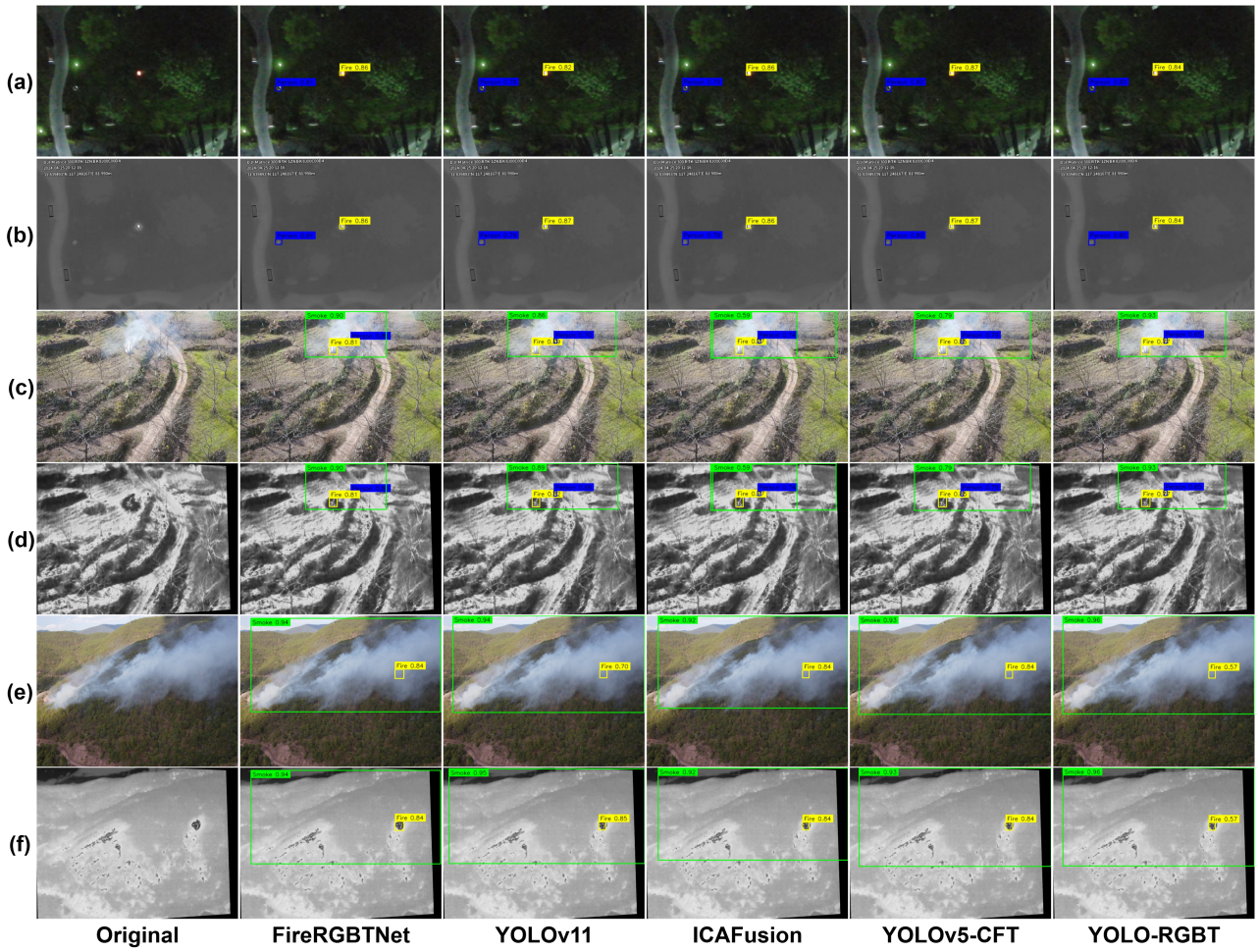


Figure 10. Qualitative comparison of detection performance across diverse forest fire scenarios. Subfigures (**a**–**f**) are organized into three groups of RGB-Thermal pairs: (**a**,**b**), (**c**,**d**), and (**e**,**f**). Specifically, (**a**,**b**) correspond to Nighttime low-illumination fires; (**c**,**d**) correspond to Complex terrain and field backgrounds; and (**e**,**f**) correspond to Dense smoke occlusion. Columns from left to right: Original input, FireRGBTNet (Ours), YOLOv11, ICAFusion, YOLOv5-CFT, and YOLO-RGBT.

4. Discussion

The proposed FireRGBTNet model achieves an optimal balance between accuracy and computational efficiency in complex forest fire detection tasks through the collaborative optimization of multiple modules. First, the heterogeneous dual-stream backbone network fully exploits the essential differences in physical information density between RGB and Thermal modalities, ensuring highly efficient feature extraction at the source. Concurrently, the TED module explicitly preserves the high-frequency features of early-stage tiny fire spots during the downsampling stage, providing high-quality feature inputs for subsequent network layers. On this basis, the MSAE module projects heterogeneous features into a shared latent semantic space via dual constraints of statistical distribution and semantic direction, without incurring any additional inference burden. This cross-modal pre-alignment operation substantially alleviates the optimization pressure on the subsequent fusion module. Ultimately, the MSGF module abandons the global attention mechanism, which is susceptible to massive background noise interference, and instead implements semantic purification and noise suppression of bimodal features through local spatial gating logic. Through the synergistic operation of the aforementioned core modules, the model attains optimal overall detection performance. From a practical application perspective, the method proposed in this paper can serve as an onboard detection module

for UAV patrol monitoring systems, enabling real-time identification of smoke, flame, and personnel targets during forest cruising, thereby assisting in early warning and fire point localization. Compared with traditional watchtowers, fixed camera systems, and satellite remote sensing, UAVs offer the advantages of flexibility, wide visual coverage, and freedom from complex terrain constraints, effectively compensating for the large blind spots, long response cycles, and limited spatial resolution of fixed monitoring means.

However, although FireRGBTNet achieves an excellent balance between accuracy and efficiency, and its theoretical computational load and parameter count have been substantially reduced, several limitations remain at the practical deployment level. First, regarding inference latency, the FPS reported in this paper refers to the model's forward inference speed measured on a Tesla V100 GPU, rather than actual results from onboard edge devices. It should be noted that the parameter count and FLOPs cannot directly predict actual throughput. Although the parameter count (4.13 M) and computational load (13.50 GFLOPs) of FireRGBTNet are both lower than those of YOLOv11, its measured FPS (44.7) remains lower than that of the latter (73.1 FPS). This is mainly attributed to the parallel computing overhead of the dual-stream backbone architecture and the relatively frequent memory access in modules such as SGCA, which may make the architecture more constrained by memory bandwidth rather than computational capacity during inference. Moreover, the FPS discussed in this paper currently addresses only the efficiency of the model itself, excluding the preprocessing stage. In real production environments, raw images (e.g., 1920×1080) must first undergo preprocessing such as resizing before being fed into the model; this preprocessing time constitutes a significant component of end-to-end deployment latency. Due to experimental constraints, it is currently difficult to conduct sufficient measurements of end-to-end latency across different onboard hardware platforms. Second, regarding cross-modal alignment, the M-RIFT registration process mentioned in the text was part of the original RGBT-3M dataset construction phase. This paper directly uses its already registered images; however, the current multimodal interaction heavily relies on ideal physical alignment between RGB and Thermal inputs at the sensor level, whereas high-frequency mechanical vibrations during actual flights can easily induce dynamic spatial misalignment in cross-sensor images. Although the geometric data augmentation (random translation, rotation, and scaling) adopted during the training stage indirectly lays the foundation for coping with mild modal misalignment—as these transformations are applied synchronously to both modalities rather than being explicit mechanisms targeting cross-modal misalignment—and given that the dataset only provides pre-registered images, quantitative robustness evaluation against misalignment is currently challenging. Furthermore, when extreme physical failure occurs in one modality, the model lacks an adaptive modality dropout mechanism, potentially leading to the propagation of erroneous features across the dual streams. It should be emphasized that although constrained by experimental conditions, no deployment verification has yet been conducted on real physical edge devices, the extensive experiments on the public RGBT-3M dataset have validated the algorithm's effectiveness and theoretical feasibility, laying the groundwork for subsequent engineering deployment.

To address the aforementioned limitations, future research will focus on two aspects: model deployment optimization and robustness enhancement. For deployment optimization, structural re-parameterization or operator fusion will be employed to reduce the memory access overhead of dual-stream branches, and model pruning, quantization, and inference acceleration techniques such as TensorRT will be advanced. Ultimately, the model will be deployed on representative embedded platforms such as Jetson and actual UAV hardware for field tests, verifying its end-to-end latency and power consumption performance in real scenarios. For robustness enhancement, the dynamic alignment fusion

mechanism will be evaluated through more in-depth alignment experiment designs; deformable convolutions or feature-level dynamic alignment will be introduced to improve the model's tolerance to spatial misalignment; and model architectures under missing modalities will be investigated to prevent systemic collapse caused by single-sensor failure. Additionally, aligning with development trends in artificial intelligence, future work may draw inspiration from paradigms of vision-language foundation models such as CLIP [57], introducing text prompts to guide visual feature extraction and further enhancing the model's generalization capability to unknown scenes via cross-modal contrastive learning. Furthermore, with the emergence of lightweight multimodal large models, exploring the integration of large model capabilities with lightweight fire detection scenarios also represents a highly promising research direction.

5. Conclusions

Early warning and real-time monitoring of forest fires are of paramount importance for maintaining global ecological balance and safeguarding life and property. In response to challenges such as diminutive fire targets in complex forest environments, severe background interference, and the constrained deployment capabilities of UAV edge devices, traditional unimodal detection models and conventional high-parameter multimodal models are often inadequate. To this end, this paper proposes FireRGBTNet, a lightweight forest fire detection model based on efficient RGB–Thermal fusion.

This study conducts network design based on the physical characteristics of multiple modalities. First, by integrating a customized heterogeneous dual-stream backbone with a Target-Enhanced Downsampling (TED) module, the model effectively mitigates low-frequency redundancy in the infrared modality while preventing the attenuation of fragile high-frequency features of early fire targets in deep network layers. Second, to address the inherent semantic gap between RGB and thermal features, this paper innovatively introduces a multi-scale semantic alignment enhancement (MSAE) branch, which achieves cross-modal pre-alignment within a shared semantic space through dual constraints of statistical distribution and orientation, thereby achieving a performance enhancement with zero inference overhead. Finally, the Multimodal Spatial Gating Fusion (MSGF) module leverages a Spatial Gated Cross-Attention (SGCA) mechanism to suppress unimodal-specific noise and incorporates MishGLU to supplement global semantics, achieving high-quality feature fusion.

The experimental results demonstrate that our method exhibits significant superiority using the public RGBT-3M UAV forest fire dataset. When confronted with challenging scenarios such as nighttime low illumination, heavy smoke occlusion, and high-altitude wide field-of-view, FireRGBTNet consistently showcases exceptional robustness and target detection precision. More importantly, compared with current state-of-the-art (SOTA) object detection models, FireRGBTNet achieves the highest detection accuracy while maintaining a minimal computational footprint of only 4.13 M parameters and 13.50 G FLOPs. FireRGBTNet not only provides a forest fire monitoring solution with substantial engineering deployment value for resource-constrained UAV devices but also establishes a solid theoretical and architectural foundation for future explorations integrating vision–language foundation models.

Author Contributions: Y.M. conceived the study, designed the methodology, conducted the investigation and data curation, drafted the manuscript, created visualizations, and administered the project. W.S. supervised the research, administered the project, and reviewed and edited the manuscript. M.W. (Maofa Wang) contributed to methodology design, performed formal analysis, and reviewed and edited the manuscript. Y.S. and M.W. (Mengyu Wang) performed validation of the results. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Innovation Program for Postgraduate students in IDP subsidized by Fundamental Research Funds for the Central Universities (ZY20260306), the Hebei Natural Science Foundation (D2026512004), and the Scientific Research and Development Plan Project of Langfang Science and Technology Bureau (2025011070).

Data Availability Statement: The RGBT-3M dataset used in this study is a publicly available benchmark. The data can be accessed through the original source cited in the references [24].

Acknowledgments: The authors appreciate the editors and anonymous reviewers for their valuable recommendations.

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

References

1. Jones, M.W.; Abatzoglou, J.T.; Veraverbeke, S.; Andela, N.; Lasslop, G.; Forkel, M.; Smith, A.J.P.; Burton, C.; Betts, R.A.; van der Werf, G.R.; et al. Global and regional trends and drivers of fire under climate change. *Rev. Geophys.* **2022**, *60*, e2020RG000726. [CrossRef]
2. Li, Z.; Kaufman, Y.J.; Ichoku, C.; Fraser, R.; Trishchenko, A.; Giglio, L.; Jin, J.-Z.; Yu, X. A review of AVHRR-based active fire detection algorithms: Principles, limitations, and recommendations. In *Global and Regional Vegetation Fire Monitoring from Space, Planning and Coordinated International Effort*; SPB Academic Publishing: Amsterdam, The Netherlands, 2001; pp. 199–225.
3. Barmpoutis, P.; Papaioannou, P.; Dimitropoulos, K.; Grammalidis, N. A review on early forest fire detection systems using optical remote sensing. *Sensors* **2020**, *20*, 6442. [CrossRef] [PubMed]
4. Bao, S.; Xiao, N.; Lai, Z.; Zhang, H.; Kim, C. Optimizing watchtower locations for forest fire monitoring using location models. *Fire Saf. J.* **2015**, *71*, 100–109. [CrossRef]
5. Bouguettaya, A.; Zarzour, H.; Taberkit, A.M.; Kechida, A. A review on early wildfire detection from unmanned aerial vehicles using deep learning-based computer vision algorithms. *Signal Process.* **2022**, *190*, 108309. [CrossRef]
6. Li, X.; Diao, W.; Mao, Y.; Gao, P.; Mao, X.; Li, X.; Sun, X. OGMN: Occlusion-guided multi-task network for object detection in UAV images. *ISPRS J. Photogramm. Remote Sens.* **2023**, *199*, 242–257. [CrossRef]
7. Munir, A.; Siddiqui, A.J.; Anwar, S. Investigation of UAV Detection in Images with Complex Backgrounds and Rainy Artifacts. In Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 1–6 January 2024.
8. McEnroe, P.; Wang, S.; Liyanage, M. A Survey on the Convergence of Edge Computing and AI for UAVs: Opportunities and Challenges. *IEEE Internet Things J.* **2022**, *9*, 15435–15459. [CrossRef]
9. Emmy Prema, C.; Vinsley, S.S.; Suresh, S. Multi Feature Analysis of Smoke in YUV Color Space for Early Forest Fire Detection. *Fire Technol.* **2016**, *52*, 1319–1342. [CrossRef]
10. Pan, J.; Ou, X.; Xu, L. A Collaborative Region Detection and Grading Framework for Forest Fire Smoke Using Weakly Supervised Fine Segmentation and Lightweight Faster-RCNN. *Forests* **2021**, *12*, 768. [CrossRef]
11. Majid, S.; Alenezi, F.; Masood, S.; Ahmad, M.; Gündüz, E.S.; Polat, K. Attention based CNN model for fire detection and localization in real-world images. *Expert Syst. Appl.* **2022**, *189*, 116114. [CrossRef]
12. Polivka, T.N.; Wang, J.; Ellison, L.T.; Hyer, E.J.; Ichoku, C.M. Improving Nocturnal Fire Detection With the VIIRS Day–Night Band. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 5503–5519. [CrossRef]
13. Sousa, M.J.; Moutinho, A.; Almeida, M. Thermal Infrared Sensing for Near Real-Time Data-Driven Fire Detection and Monitoring Systems. *Sensors* **2020**, *20*, 6803. [CrossRef] [PubMed]
14. Zhao, M.; Li, L.; Li, W.; Tao, R.; Li, L.; Zhang, W. Infrared Small-Target Detection Based on Multiple Morphological Profiles. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6077–6091. [CrossRef]
15. Wang, Q.; Sun, Y.; Chi, Y.; Shen, T. RGB-T Object Detection With Failure Scenarios. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 3000–3010. [CrossRef]
16. Meng, F.; Hong, A.; Tang, H.; Tong, G. FQDNet: A Fusion-Enhanced Quad-Head Network for RGB-Infrared Object Detection. *Remote Sens.* **2025**, *17*, 1095. [CrossRef]
17. Liu, R.; Mi, L.; Chen, Z. AFNet: Adaptive Fusion Network for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7871–7886. [CrossRef]
18. Sun, N.; Zhou, J.; Hu, K.; Wei, C.; Wang, Z.; Song, L. Multimodal Fire Salient Object Detection for Unregistered Data in Real-World Scenarios. *Fire* **2025**, *8*, 415. [CrossRef]
19. 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]

20. Zhang, J.; Liu, H.; Yang, K.; Hu, X.; Liu, R.; Stiefelhagen, R. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 14679–14694. [CrossRef]
21. Li, J.; Selvaraju, R.R.; Gotmare, A.D.; Joty, S.; Xiong, C.; Hoi, S.C.H. Align before fuse: Vision and language representation learning with momentum distillation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9694–9705.
22. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 121–137.
23. Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; Wu, Y. Coca: Contrastive captioners are image-text foundation models. *arXiv* **2022**, arXiv:2205.01917.
24. Zhang, Y.; Rui, X.; Song, W. A UAV-Based Multi-Scenario RGB-Thermal Dataset and Fusion Model for Enhanced Forest Fire Detection. *Remote Sens.* **2025**, *17*, 2593. [CrossRef]
25. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017*; IEEE: New York, NY, USA, 2017; pp. 936–944.
26. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024*; IEEE: New York, NY, USA, 2024; pp. 16965–16974.
27. Felzenszwalb, P.; McAllester, D.; Ramanan, D. A discriminatively trained, multiscale, deformable part model. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008*; IEEE: New York, NY, USA, 2008; pp. 1–8.
28. Wang, C.-Y.; Liao, H.-Y.M.; Yeh, I.H. Designing network design strategies through gradient path analysis. *arXiv* **2022**, arXiv:2211.04800.
29. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 818–833.
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: New York, NY, USA, 2018; pp. 7132–7141.
31. Sun, B.; Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 443–450.
32. Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; Liu, W. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: New York, NY, USA, 2018; pp. 5265–5274.
33. Misra, D. Mish: A self regularized non-monotonic activation function. *arXiv* **2019**, arXiv:1908.08681.
34. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*; Omnipress: Madison, WI, USA, 2010; pp. 807–814.
35. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
36. Wang, J.; Xu, C.; Yang, W.; Yu, L. A normalized Gaussian Wasserstein distance for tiny object detection. *arXiv* **2021**, arXiv:2110.13389.
37. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international Conference on Computer Vision, Venice, Italy, 22–29 October 2017*; IEEE: New York, NY, USA, 2017; pp. 2980–2988.
38. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. *arXiv* **2019**, arXiv:1902.09630.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016*; IEEE: New York, NY, USA, 2016; pp. 770–778.
40. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
41. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019*; PMLR: New York, NY, USA, 2019.
42. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725.
43. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
44. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
45. 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*, 6000–6010.

46. Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; Barnard, K. Attentional feature fusion. In *Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021*; IEEE: New York, NY, USA, 2021; pp. 3559–3568.
47. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2018.
48. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
49. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
50. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017*; IEEE: New York, NY, USA, 2017; pp. 618–626.
51. 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*, 91–99.
52. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016.
53. Yu, G.; Chang, Q.; Lv, W.; Xu, C.; Cui, C.; Ji, W.; Dang, Q.; Deng, K.; Wang, G.; Du, Y.; et al. PP-PicoDet: A better real-time object detector on mobile devices. *arXiv* **2021**, arXiv:2111.00902.
54. Ma, Y.; Shan, W.; Sui, Y.; Wang, M.; Wang, M. LHRF-YOLO: A Lightweight Model with Hybrid Receptive Field for Forest Fire Detection. *Forests* **2025**, *16*, 1095. [CrossRef]
55. Qingyun, F.; Dapeng, H.; Zhaokui, W. Cross-modality fusion transformer for multispectral object detection. *arXiv* **2021**, arXiv:2111.00273.
56. Wan, D.; Lu, R.; Fang, Y.; Lang, X.; Shu, S.; Chen, J.; Shen, S.; Xu, T.; Ye, Z. Yolov11-rgbt: Towards a comprehensive single-stage multispectral object detection framework. *arXiv* **2025**, arXiv:2506.14696.
57. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning, Shenzhen, China, 26 February–1 March 2021*; PMLR: New York, NY, USA, 2021; pp. 8748–8763.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.