computers

MDPI

Article

A Weakly Supervised Multi-Scale Cross-Modal Information Fusion Method for Wildfire Detection

Dawei Wen [1,2], Zhoujiang Peng [1,2] and Yuan Tian [3,*]

[1] School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China; daweiwen@wit.edu.cn (D.W.); 22307010030@stu.wit.edu.cn (Z.P.)
[2] Hubei Provincial Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan 430205, China
[3] School of Geographical Sciences, Lingnan Normal University, Zhanjiang 524048, China
* Correspondence: 202031250017@mail.bnu.edu.cn

Abstract

In recent years, wildfires have occurred with increasing frequency. Pixel-level annotation of high-resolution remote sensing wildfire imagery is costly and labor-intensive. Therefore, there is an urgent need for a weakly supervised wildfire detection method that balances detection accuracy and annotation efficiency. To address the key limitations of existing weakly supervised approaches based on class activation maps (CAMs), including imprecise delineation of fire boundaries, insufficient utilization of cross-modal information, and limited capability in modeling temporal characteristics, this paper proposes a dual-branch multi-scale feature fusion framework for weakly supervised wildfire detection. The proposed framework consists of a multispectral branch and a shortwave infrared (SWIR) temporal branch, which are designed to capture the spatial structural information of fire regions and the temporal variation of thermal anomalies, respectively. Attention-guided feature fusion modules are introduced at each network stage to enable complementary integration of cross-modal information. In addition, a multi-scale CAM-weighted fusion strategy is designed to jointly enhance region localization accuracy and semantic discrimination capability. Experimental evaluations are conducted on a high-resolution wildfire dataset covering 29 regions and consisting of 2206 images. The results demonstrate that the proposed method achieves an IoU of 58.7% and an F1-score of 73.5%, outperforming the state-of-the-art methods by 4.6% and 3.2%, respectively. Ablation and comparative experiments further verify that the dual-branch architecture and feature fusion strategy significantly improve fire localization accuracy and effectively reduce the missed detection rate.

Keywords: wildfire detection; Sentinel-2; SWIR; multispectral imagery; cross-modal fusion; weakly supervised segmentation; class activation maps; remote sensing imagery

Check for updates

Academic Editors: Paolo Bellavista and Aditya Kumar Sahu

Received: 1 April 2026
Revised: 3 May 2026
Accepted: 8 May 2026
Published: 14 May 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

Wildfires are uncontrolled burning events occurring in natural ecosystems such as forests, grasslands, and shrublands under natural or anthropogenic triggers, and they may cause substantial ecological, economic, and social losses [1]. Under the background of global climate change, the frequency and severity of extreme heat and drought events have increased, making rapid and accurate wildfire monitoring an important task in disaster prevention and emergency management.

Traditional fire monitoring based on watchtowers and manual inspection is limited by its low efficiency and small area coverage, and severe wildfire damage has been widely

reported in interface communities [2]. In recent years, unmanned aerial vehicle platforms have supported post-fire damage assessment and recovery monitoring at local scales [3,4]; however, UAV-based monitoring still suffers from limited spatial coverage and restricted deployment flexibility during the early stage of fire events. By contrast, remote sensing satellites provide wide-area and all-weather observation capability and therefore play an increasingly important role in wildfire monitoring.

Several global burned-area and active-fire products have been developed based on coarse-resolution sensors, such as MCD64A1 [5], FireCCI50 [6], FireCCI51 [7], and FIRMS [8]. These products are valuable for global wildfire monitoring and post-fire assessment. However, their spatial resolution is often insufficient for accurate delineation of accurate wildfire localization, especially for small-scale or fragmented burned patches.

With the widespread use of high-resolution satellites such as Sentinel-2 and Landsat, wildfire monitoring data with spatial resolutions of 10–30 m have become increasingly available, making fine-scale fire localization more feasible [9–11]. Meanwhile, deep learning-based semantic segmentation methods have shown promise in fire detection tasks [11,12]. However, these fully supervised methods strongly depend on a large amount of pixel-level annotation, which is expensive and difficult to acquire in wildfire scenes because of cloud contamination, smoke interference, and dynamic fire evolution.

Weakly supervised semantic segmentation offers an alternative by reducing annotation cost. Recent studies have also explored weakly supervised forest fire segmentation in remote sensing imagery, demonstrating the potential of low-cost annotation strategies for fire-region localization [13–15]. Among such methods, class activation map (CAM)-based approaches generate spatial response maps from image-level labels and have been widely used as a weak supervision signal [16–18]. Nevertheless, directly applying CAM-based methods to wildfire detection in remote sensing imagery still faces several challenges. First, conventional CAM methods usually rely on low-resolution deep features, leading to coarse localization and blurry fire boundaries. Second, existing CAM frameworks have limited capability in jointly modeling visible, near-infrared, and shortwave infrared (SWIR) information from multispectral remote sensing images. Third, wildfire detection has strong temporal characteristics, while single-temporal CAM methods cannot effectively exploit pre-fire and post-fire differences.

To address these issues, this study proposes a dual-branch multi-scale feature fusion framework for weakly supervised wildfire detection. The proposed architecture separately models the spatial structure information from multispectral imagery and the temporal variation information from SWIR imagery, and progressively fuses them at multiple network stages. In addition, a multi-scale CAM fusion strategy is introduced to generate activation maps with both accurate boundaries and robust semantic responses.

The main contributions of this study can be summarized as follows:

- A dual-branch multi-scale cross-modal feature fusion framework is proposed to jointly exploit multispectral spatial information and SWIR temporal variation information under weak supervision.
- An attention-guided feature interaction mechanism is designed to adaptively enhance discriminative channels and spatial locations, thereby improving the robustness of weakly supervised wildfire detection.
- A multi-scale CAM fusion strategy is developed to balance boundary accuracy and semantic discriminability across different network stages.

2. Methodology

2.1. Class Activation Map

CAM highlights discriminative regions for image classification. For category c, it can be written as

$$M^c = \frac{\mathrm{ReLU}\left(\sum_{k=1}^{K} w_k^c f_k\right)}{\max\left(\mathrm{ReLU}\left(\sum_{k=1}^{K} w_k^c f_k\right)\right)}, \tag{1}$$

where w_k^c denotes the weight associated with class c in channel k, f_k is the k-th feature map in the final convolutional layer, and K is the total number of channels.

Binary cross-entropy loss is used as the primary classification loss:

$$L_{\mathrm{BCE}} = -\frac{1}{C}\sum_{c=1}^{C}[y_c \log(S_c(G(F))) + (1 - y_c)\log(1 - S_c(G(F)))], \tag{2}$$

where C is the number of classes, y_c is the class-c label, $G(\cdot)$ denotes global average pooling, $S_c(\cdot)$ denotes the class-c output, and F denotes the final convolutional feature.

Considering the severe class imbalance in wildfire detection, Dice loss is further introduced:

$$L_{\mathrm{Dice}} = 1 - \frac{2\sum_{i=1}^{N} y_i p_i + \varepsilon}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} p_i^2 + \varepsilon}, \tag{3}$$

where $y_i \in \{0, 1\}$ is the ground-truth label of sample i, $p_i \in [0, 1]$ is the predicted probability, and ε is a smoothing factor.

The final loss is formulated as

$$L_{\mathrm{total}} = \alpha L_{\mathrm{BCE}} + \beta L_{\mathrm{Dice}}. \tag{4}$$

2.2. Dual-Branch Multi-Scale Feature Fusion Framework

Wildfire detection under weak supervision is challenging because fire targets are often small, spatially sparse, and easily confused with complex backgrounds [19]. Bare soil, high-temperature non-fire objects, and cloud shadows may produce responses similar to actual fire regions. Therefore, both the physical characteristics of remote sensing data and the representation ability of deep models should be considered when designing the network architecture.

From the perspective of remote sensing imaging, different spectral bands provide complementary information. Visible and near-infrared bands mainly describe surface structure, vegetation cover, and contextual semantics, whereas SWIR is more sensitive to thermal anomalies and burning residues [20]. Existing studies have shown that jointly modeling multispectral information and temporal observations can significantly improve wildfire detection performance [21–23].

In this work, a dual-branch multi-scale feature fusion network is adopted. One branch extracts multispectral spatial-semantic features, and the other branch captures temporal variation from SWIR images. Multi-scale cross-modal feature interaction is introduced to preserve shallow spatial details and deep semantic information simultaneously [24,25].

2.3. Network Architecture

As shown in Figure 1, the proposed framework consists of two parallel branches: a multispectral feature extraction branch and an SWIR temporal feature extraction branch. Both branches are built on ResNet50 backbones, and multi-level feature fusion modules are inserted for cross-modal interaction.

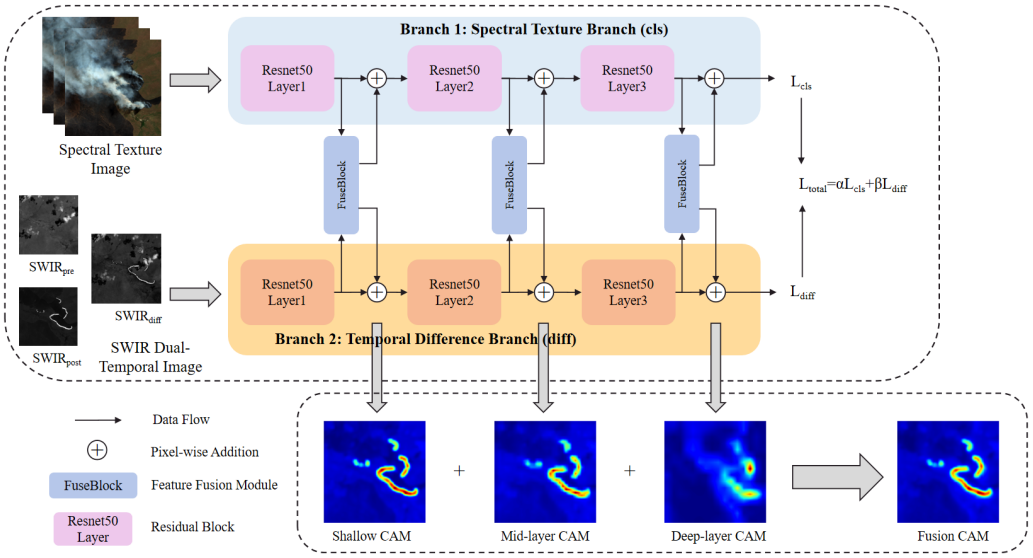


Figure 1. Dual-branch network architecture.

The first branch takes nine non-SWIR Sentinel-2 bands as input, including visible, near-infrared, and red-edge bands, to provide rich spatial and texture cues. The second branch takes pre-fire SWIR, post-fire SWIR, and their difference image as input to capture thermal anomalies and temporal changes.

2.4. Training and Testing Workflow Description

Both branches use ResNet50 as the backbone. To balance receptive field and spatial resolution, the network retains the input stem and the first three residual stages (layer1, layer2, and layer3), while removing the original layer4 and the final fully connected classification head, so as to avoid excessive downsampling and the resulting loss of small-fire targets. The multi-scale feature fusion modules are inserted after these three residual stages; therefore, three feature scales participate in the cross-modal interaction and CAM generation.

During training, 256×256 image patches are used as input, and the dual-branch network is optimized end-to-end using only image-level fire/non-fire labels. During testing, CAMs are generated from the three feature scales, upsampled to a unified spatial size, fused using $w_1 = 0.5$, $w_2 = 0.3$, and $w_3 = 0.2$, and finally binarized using the Otsu adaptive thresholding method to obtain the fire masks. This design preserves clearer boundary details from shallow features while also benefiting from the stronger semantic discrimination provided by deeper features.

2.5. Channel Attention and Spatial Attention

Channel attention is computed from global average pooling and global max pooling statistics:

$$z_{\text{avg}} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} F(c,h,w), \quad z_{\max} = \max_{h,w} F(c,h,w), \tag{5}$$

$$M_c = \sigma\big(W_2 \cdot \text{ReLU}(W_1 \cdot (z_{\text{avg}} + z_{\max}))\big). \tag{6}$$

Spatial attention is computed as

$$s_{\text{avg}} = \frac{1}{C} \sum_{c=1}^{C} F_{\text{ca}}(c,h,w), \quad s_{\max} = \max_{c} F_{\text{ca}}(c,h,w), \tag{7}$$

$$M_s = \sigma\big(\text{Conv}_{7\times7}([s_{\text{avg}}; s_{\max}])\big). \tag{8}$$

2.6. Cross-Modal Fusion and Residual Enhancement

Traditional CAM methods aggregate features only once at the network end, making it difficult to jointly preserve low-level details and high-level semantics. Therefore, progressive multi-scale feature fusion is introduced after each ResNet stage.

As shown in Figure 2, the feature fusion module integrates cross-modal information through attention-guided interaction and residual enhancement.

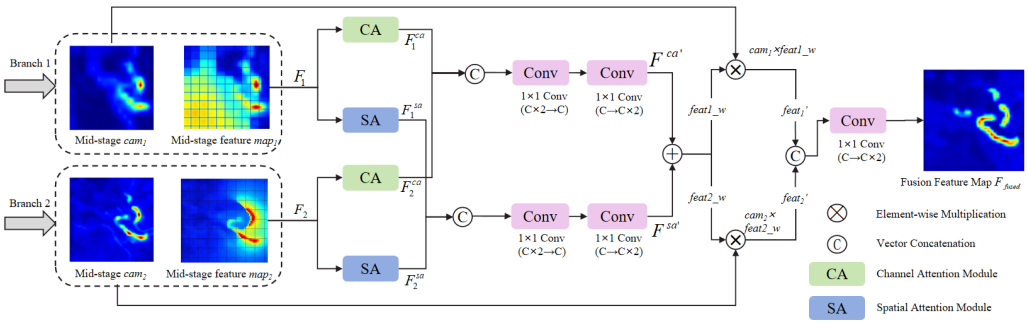


Figure 2. Feature fusion module.

Suppose the two branches output features F_1^i and F_2^i at stage i. The fused features are added back by residual connections:

$$F_1^{i\prime} = F_1^i + \alpha_i F_{\text{fused}}^i, \quad F_2^{i\prime} = F_2^i + \beta_i F_{\text{fused}}^i, \tag{9}$$

where α_i and β_i are learnable parameters.

2.7. Multi-Scale CAM Generation and Fusion

To balance localization precision and semantic discriminability, a multi-scale CAM-weighted fusion strategy is used:

$$M_{\text{final}} = w_1 \frac{M_1}{\|M_1\|_\infty} + w_2 \frac{M_2}{\|M_2\|_\infty} + w_3 \frac{M_3}{\|M_3\|_\infty}, \tag{10}$$

where M_i is the CAM generated at stage i. In our experiments, $w_1 = 0.5$, $w_2 = 0.3$, and $w_3 = 0.2$.

3. Data and Experimental Settings

3.1. Dataset

Sentinel-2A imagery acquired from October 2023 to November 2024 was used in this study, and 20 m was adopted as the unified analysis resolution [26]. We did not use 10 m or 60 m as the base resolution mainly because the proposed method explicitly relies on SWIR bands as an important source of thermal-anomaly and temporal-difference information, and the native resolution of Sentinel-2 SWIR bands is 20 m. If 10 m were used as the unified resolution, the 20 m SWIR bands would need to be upsampled to 10 m. This would not add real spatial detail but could introduce interpolation errors and noise, thereby weakening the stability of cross-modal fusion. By contrast, 60 m imagery would significantly reduce small-fire and boundary details and increase mixed-pixel effects. Although we do not separately report additional experiments at 10 m and 60 m in the current manuscript, 20 m represents a more reasonable compromise for this task in terms of both key spectral information and spatial detail.

Based on NASA FIRMS hotspot products [8], more than 40 representative wildfire events from 29 regions worldwide were selected, including Australia, California, the Amazon, Siberia, and Mediterranean coastal areas. The dataset construction procedure includes

coarse cropping of fire-affected areas, splitting images into 256 × 256 patches, hotspot screening with FIRMS, and manual verification using SWIR temperature information and the normalized burn ratio (NBR). Finally, a dataset containing 2206 fire images and 16,173 non-fire images was established.

Most non-fire samples were selected from the same or adjacent scenes as the fire samples, so that they shared similar topographic background, vegetation cover, and imaging conditions, thereby avoiding an artificially easy classification task caused by overly large regional differences. In addition, about one-quarter to one-third of the non-fire samples were collected from other wildfire-free scenes as supplementary negative samples to improve robustness. The selection criteria for non-fire samples were as follows: (1) no overlap with NASA FIRMS hotspot records; (2) no obvious thermal anomaly in the SWIR bands; and (3) manual visual verification to confirm the absence of real fire hotspots or burned areas.

The training set contains 2000 fire images and 8000 non-fire images together with their corresponding pre-fire SWIR images, while the test set contains the remaining 206 fire images and 800 non-fire images. The class ratio is approximately 1:4 in both subsets. To reduce scene leakage as much as possible, the data split was conducted using wildfire events as the basic unit, so that samples from different temporal phases of the same fire event would not appear simultaneously in different subsets.

3.2. Implementation Details

Both branches are based on ResNet50, with the last convolutional stage removed to preserve spatial resolution. CAM maps are normalized and binarized using the Otsu thresholding method. The initial learning rate is set to 10^{-4}, the weight decay is set to 10^{-5}, and cosine annealing is used for optimization [27]. The weights of loss_{cls} and $\text{loss}_{\text{diff}}$ are 0.4 and 0.6, respectively. The coefficients α and β in the BCE-Dice loss are also set to 0.4 and 0.6.

4. Results and Discussion

4.1. Evaluation Metrics

The model is evaluated using overall accuracy (OA), F1-score, intersection over union (IoU), mean IoU (mIoU), recall, precision, and the Kappa coefficient:

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{11}$$

$$\text{F}_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{12}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \tag{13}$$

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \text{IoU}_i, \tag{14}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{15}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{16}$$

$$P_e = \frac{(\text{TP} + \text{FP})(\text{TP} + \text{FN}) + (\text{TN} + \text{FN})(\text{TN} + \text{FP})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})^2}, \tag{17}$$

$$\kappa = \frac{\text{OA} - P_e}{1 - P_e}. \tag{18}$$

4.2. Ablation Study

To verify the effectiveness of the dual-branch network architecture and the cross-modal feature fusion mechanism, three model configurations were designed for ablation experiments: the cls model, using only the multispectral spatial branch; the diff model, using only the SWIR temporal-difference branch; and the full model, enabling both branches together with the multi-scale feature fusion module. The results are shown in Table 1.

Table 1. Ablation results under different model configurations.

Name	OA	Kappa	mIoU	IoU	Precision	Recall	F1
cls	0.880	0.114	0.474	0.070	0.071	0.824	0.128
diff	0.879	0.097	0.469	0.060	0.060	**0.899**	0.112
full	**0.927**	**0.352**	**0.582**	**0.237**	**0.259**	0.855	**0.360**

Note: Bold values indicate the best performance in each metric.

As shown in Table 1, when only the multispectral spatial branch is used, the model can perform preliminary fire localization based on texture, vegetation, and scene context information. However, due to the lack of thermal-anomaly and temporal-change cues, the fire-class IoU is only 0.070 and F1 is only 0.128, indicating limited discriminative capability for true fire regions. This result suggests that multispectral information helps provide boundary morphology and land-cover structure, but under weak supervision it is difficult to independently support stable fire-region recovery.

When only the SWIR temporal-difference branch is used, recall increases to 0.899, showing that SWIR-difference features can more sensitively capture pre-/post-fire thermal anomalies and provide stronger coverage of candidate fire regions. However, precision is only 0.060 and F1 is only 0.112, indicating that relying solely on thermal-change cues is susceptible to interference from hot bare land, noise, and local brightness-temperature anomalies, leading to more false positives.

When dual-branch joint modeling and the feature-fusion module are adopted, the full model achieves the best performance across all key metrics. OA increases to 0.927, Kappa to 0.352, mIoU to 0.582, fire-class IoU to 0.237, and F1 to 0.360. Compared with single-branch models, dual-branch fusion significantly improves precision and IoU while maintaining high recall, indicating that complementary fusion of multispectral spatial information and SWIR temporal information at different scales can effectively reduce false activations, enhance true fire-region responses, and improve region localization quality.

Considering relative improvement, compared with cls, the full model improves fire-class IoU from 0.070 to 0.237 (about 238.6%); compared with diff, it improves from 0.060 to 0.237 (about 295.0%). For F1, the full model improves by about 181.3% over cls and 221.4% over diff. These results indicate that the performance gain from dual-branch fusion is not a local refinement but an overall enhancement in fire-region recovery quality.

From the precision–recall relationship, the two single-branch models exhibit two typical biases. The cls branch has relatively high recall but very low precision, indicating that it can detect part of potential fire regions but cannot effectively suppress complex backgrounds. The diff branch further increases recall to 0.899, but precision remains low, reflecting over-sensitivity to thermal-anomaly cues. In contrast, while recall in the full model is only slightly lower than that of the diff branch, precision increases to 0.259, indicating that multispectral spatial information effectively constrains false alarms from the SWIR branch, while SWIR temporal information compensates for weak fire responses missed by the multispectral branch.

4.3. Comparison with State-of-the-Art Methods

To further validate the effectiveness of the proposed method for weakly supervised wildfire detection, several representative CAM generation methods were selected for comparison, including CAM, Grad-CAM, Grad-CAM++, LayerCAM, IRNet, and DWSLNet for remote sensing weakly supervised segmentation. All methods were evaluated on the same test set, and Otsu thresholding was uniformly applied as post-processing to ensure fair comparison. The results are shown in Table 2.

Table 2. Comparison of CAM generation results with other methods.

Method	OA	Kappa	mIoU	IoU	Precision	Recall	F1
CAM	0.691	0.093	0.376	0.062	0.081	0.310	0.105
Grad-CAM	0.911	0.093	0.484	0.057	0.060	0.716	0.107
Grad-CAM++	0.886	0.089	0.470	0.055	0.056	0.817	0.102
LayerCAM	0.912	0.101	0.487	0.061	0.065	0.747	0.115
IRNet	0.895	0.084	0.474	0.053	0.054	0.766	0.098
DWSLNet	**0.977**	0.253	0.567	0.158	0.190	0.786	0.261
ours	0.927	**0.352**	**0.582**	**0.237**	**0.259**	**0.855**	**0.360**

Note: Bold values indicate the best performance in each metric.

Traditional CAM-series methods generally show weak performance in this task. CAM yields IoU and F1 of 0.062 and 0.105, respectively; Grad-CAM yields IoU/F1 of 0.057/0.107; Grad-CAM++ yields 0.055/0.102; LayerCAM yields 0.061/0.115; and IRNet yields 0.053/0.098. Although most of these methods achieve OA above 0.886 and some methods obtain relatively high recall, fire-class IoU, F1, and Kappa remain low overall. This indicates that, in remote sensing wildfire detection, these methods tend to focus on only the most discriminative local regions and struggle to recover complete and stable fire bodies and boundary structures.

A closer comparison shows that Grad-CAM, Grad-CAM++, LayerCAM, and IRNet achieve recall values of 0.716, 0.817, 0.747, and 0.766, all clearly higher than CAM (0.310), indicating more aggressive fire-region coverage. However, their precision values are only 0.060, 0.056, 0.065, and 0.054, respectively. The recall increase is accompanied by many false detections, so final F1 and IoU do not improve substantially. This suggests that relying only on gradient weighting, hierarchical activation enhancement, or region propagation cannot fundamentally solve incomplete localization under small targets, fine boundaries, and complex background interference in remote sensing wildfire scenes.

Compared with the traditional CAM family, DWSLNet shows much stronger overall performance, with OA, Kappa, mIoU, IoU, precision, recall, and F1 of 0.977, 0.253, 0.567, 0.158, 0.190, 0.786, and 0.261, respectively. DWSLNet uses a dual-branch structure and introduces coordinate attention in an auxiliary branch to improve activation of small local fire targets, allowing better handling of multi-scale land-cover backgrounds and complex spatial structures. Therefore, it performs significantly better than classic CAM methods, especially in precision, IoU, and F1.

In comparison, the proposed method achieves the best results on multiple key metrics: OA = 0.927, Kappa = 0.352, mIoU = 0.582, IoU = 0.237, precision = 0.259, recall = 0.855, and F1 = 0.360. Fire-class IoU improves from 0.158 to 0.237 over DWSLNet (about 50.0%), F1 improves from 0.261 to 0.360, Kappa improves from 0.253 to 0.352, and recall improves from 0.786 to 0.855. These results indicate stronger fire-region completeness, class-consistency, and missed-detection control, especially for weakly supervised boundary recovery.

It should be noted that DWSLNet has a higher OA (0.977), which is mainly related to class imbalance in the dataset. Because background pixels occupy a much larger proportion, a model can still obtain high OA by performing well on the background class. Therefore, OA alone cannot fully reflect true fire-region detection capability in this task. In contrast, IoU, recall, F1, and Kappa better capture fire-region recovery quality and the overall discrimination ability for positive and negative samples. From these key indicators, the advantage of the proposed method is more evident.

Further examining the precision–recall balance also highlights the difference among methods. CAM has a precision of 0.081 and recall of 0.310, indicating fewer false positives but clearly insufficient fire coverage. Grad-CAM, Grad-CAM++, LayerCAM, and IRNet significantly increase recall but keep precision at a low level, indicating broader heatmap responses and more background false activations. DWSLNet reaches a better precision–recall tradeoff, while the proposed method, with the highest recall and improved precision (0.259), achieves the highest F1 and IoU. This demonstrates that the proposed method not only covers true fire regions more completely but also suppresses pseudo-responses in complex backgrounds to a reasonable extent.

Figures 3–5 show visual comparisons of ablation configurations, heatmap overlays, and Otsu-thresholded CAM results, respectively, for the proposed method, conventional CAM variants, and DWSLNet. CAM, Grad-CAM++, and LayerCAM often respond only to the most salient local fire regions, leading to incomplete activation. In some scenes, they also respond incorrectly to smoke, bright ground areas, or complex texture backgrounds. DWSLNet can generate more complete fire responses, but its activation boundaries are often wider and may introduce extra false positives. In most scenes, the proposed method can preserve both fire bodies and edge details, while alleviating over-wide boundaries and background false activation.

From the binarized results, the masks generated by the proposed method are superior in contour continuity, edge alignment, and small-target preservation, especially in scenes with small fire scales, irregular fire-line shapes, and complex backgrounds. This indicates that the multi-scale CAM fusion strategy and cross-modal feature interaction mechanism effectively enhance weakly supervised wildfire boundary perception and improve the usability of final masks.

Finally, Otsu-binarized CAM outputs were used as pixel-level pseudo labels to train a U-Net, and the second-stage segmentation results are shown in Figure 6.

Consistent with the previous experiments, the proposed method achieves the best results in second-stage semantic segmentation. The predicted masks are highly consistent with labels and can accurately mark fire regions and contours. Grad-CAM++ and LayerCAM tend to over-focus on background information during training and fail to produce precise fire responses. DWSLNet performs better than these methods, but because fire boundaries are not accurately localized in its CAM generation stage, the thresholded pseudo-label contours remain relatively broad, limiting final segmentation precision.

4.4. Discussion

The results show that the advantage of the proposed method lies not only in metric improvement, but also in how it exploits fire-related discriminative cues. Existing studies can roughly be divided into three categories: coarse-resolution hotspot monitoring products such as MODIS- and VIIRS-based systems, which provide high temporal efficiency but limited boundary detail; high-resolution fully supervised segmentation methods based on sensors such as Sentinel-2, which offer better spatial localization but require expensive pixel-level annotations; and general CAM-based weakly supervised methods, which reduce annotation cost but mainly rely on visual saliency, gradient responses, or region propaga-

tion, and therefore make limited use of thermal anomalies and pre-/post-fire changes. In contrast, the proposed method jointly models multispectral spatial structure information and SWIR temporal-difference information, enabling simultaneous improvement in fire-region localization quality and boundary completeness under weak supervision. This is why the proposed method performs better than competing methods on IoU, F1, and Kappa, which more directly reflect fire-region recovery quality.

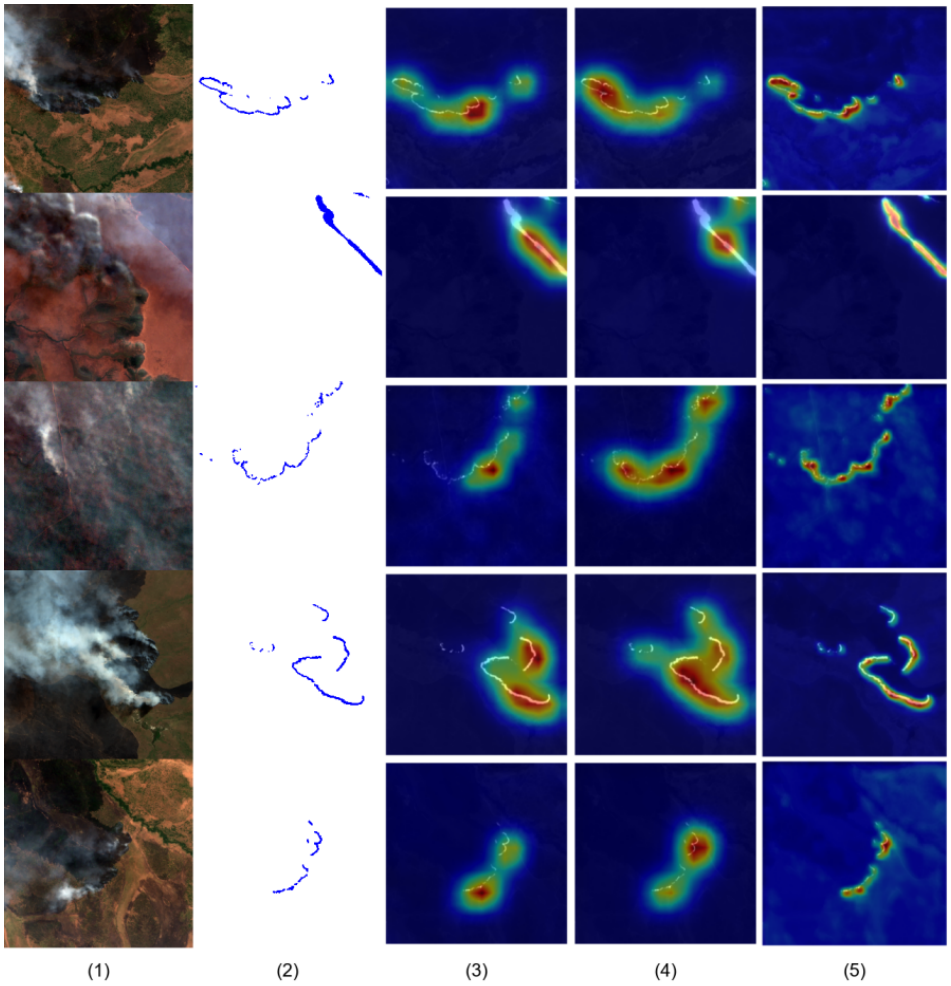


Figure 3. Visual comparison of the three ablation configurations: (**1**) RGB fire image, (**2**) fire label, (**3**) cls branch, (**4**) diff branch, (**5**) cls+diff+fuse block. In the heatmaps, warmer red colors indicate higher activation responses.

It should also be emphasized that the higher OA of DWSLNet (0.977) is largely related to class imbalance in the dataset. Because background pixels dominate the scene, a model may still achieve high OA by performing well mainly on the background class. Therefore, OA alone cannot fully reflect true fire-region detection capability in this task. By contrast, IoU, recall, F1, and Kappa better reflect fire-region recovery quality and the overall discrimination ability for positive and negative samples. Although the proposed method does not achieve the highest OA, its Kappa reaches 0.352, clearly exceeding the 0.253 of DWSLNet, indicating that the improvement does not mainly come from accidental agreement on the background class, but from a more stable overall distinction between fire and background.

In addition, the current evaluation protocol is based on post-annotated data and is intended to objectively validate detection accuracy and generalization capability rather than to directly simulate a minute-level real-time early-warning scenario. Because the

proposed method relies on two Sentinel-2 images acquired at different times over the same region, the operational prerequisite is the timely availability of a current image together with a historical reference image. Therefore, the proposed method is better suited as a high-spatial-resolution tool for fine confirmation and boundary localization after hotspot alerting, rather than as a complete replacement for strict real-time early-warning systems based on high-temporal-resolution sensors. In practical monitoring systems, high-temporal-resolution hotspot products such as VIIRS, MODIS, or NASA FIRMS can serve as front-end alert sources for rapidly identifying potential thermal anomalies, while the proposed method can use higher-spatial-resolution Sentinel-2 imagery for alert verification, false-alarm filtering, and boundary refinement. These two types of systems should therefore be regarded as complementary rather than alternative, forming a coordinated workflow of coarse high-temporal alerting followed by fine high-spatial analysis.

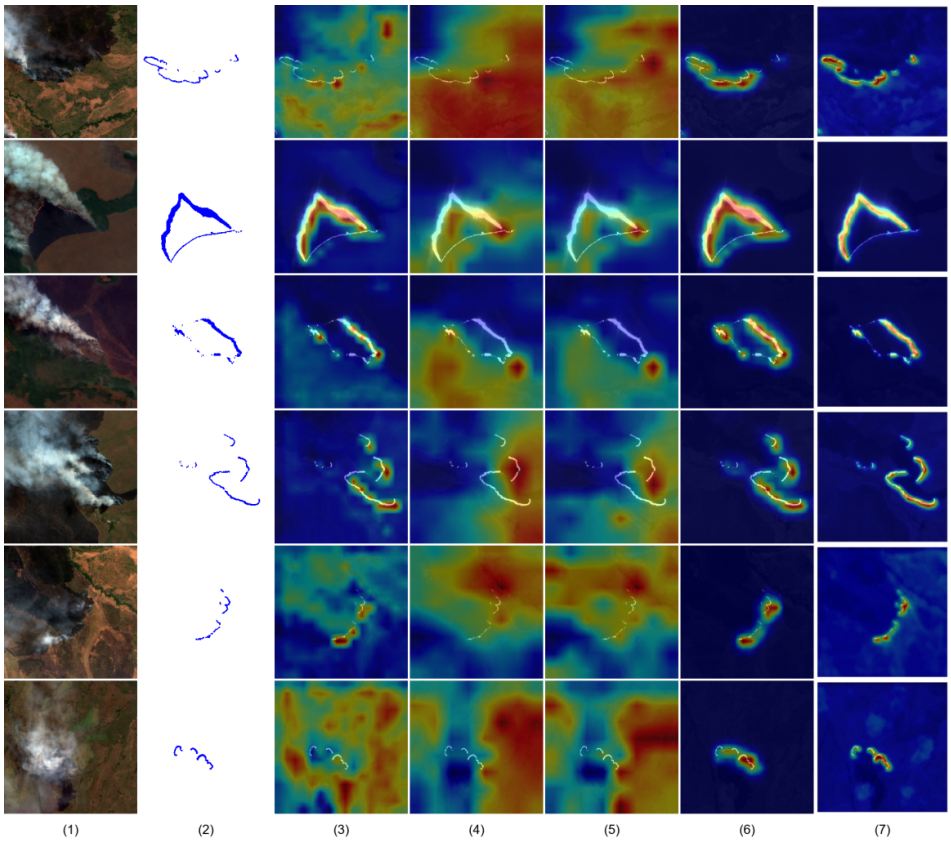


Figure 4. CAM heatmap overlays on SWIR background imagery: (**1**) RGB fire image, (**2**) fire label, (**3**) CAM, (**4**) Grad-CAM++, (**5**) LayerCAM, (**6**) DWSLNet, (**7**) ours. In the heatmaps, warmer red colors indicate higher activation responses.

From the perspective of fire-type applicability, the proposed method is more suitable for fire types with relatively clear SWIR thermal anomalies and evident pre-/post-fire changes, such as steppe fires, shrubland fires, agricultural fires, and surface or crown fires in relatively open-canopy regions. For low-intensity ground fires, soil smoldering, and thermally active zones under dense tree canopies, the model response may degrade because the thermal anomaly signal can be weak or partially masked by the canopy. Since the current dataset does not yet include a unified fine-grained fire-type annotation scheme, we do not present fire-type-specific quantitative statistics in the current study. This has been explicitly identified as an important direction for future work.

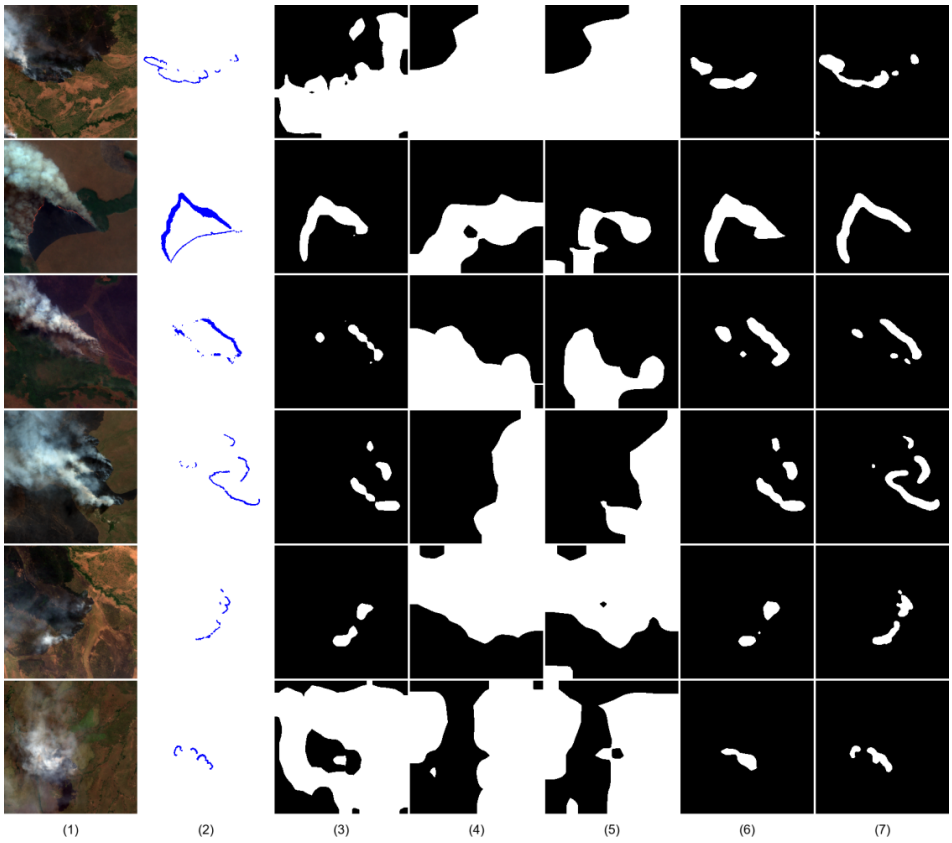


Figure 5. Comparison of Otsu-thresholded CAM results: (**1**) RGB fire image, (**2**) fire label, (**3**) CAM, (**4**) Grad-CAM++, (**5**) LayerCAM, (**6**) DWSLNet, (**7**) ours.

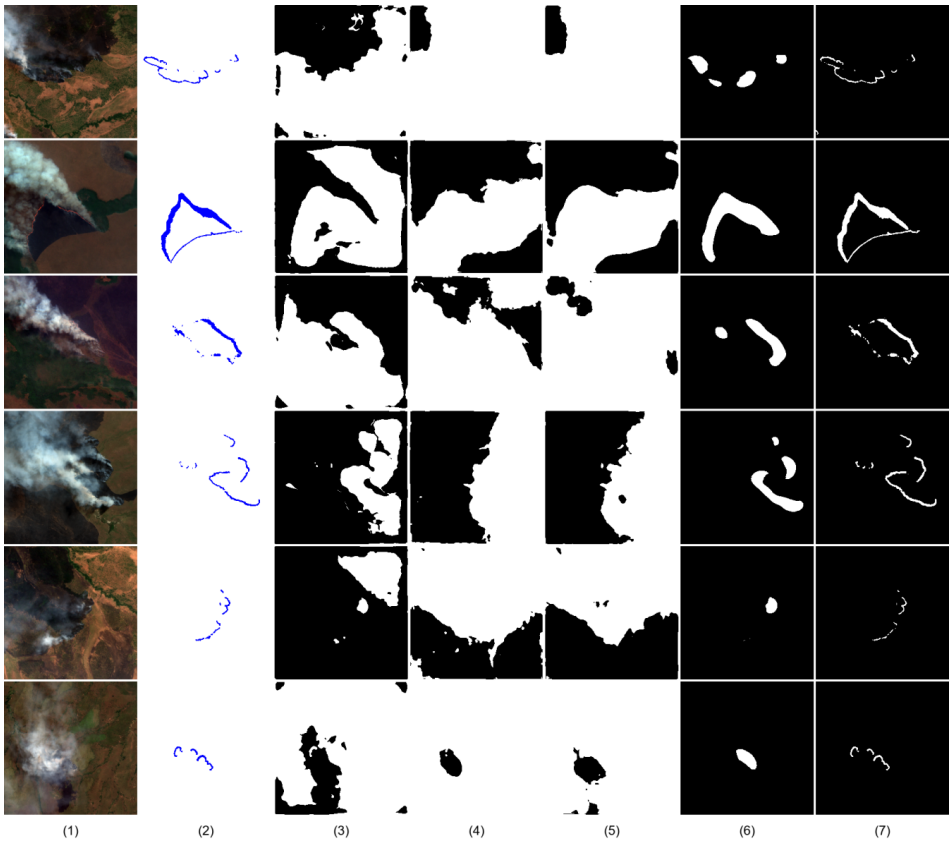


Figure 6. Comparison of second-stage semantic segmentation results: (**1**) RGB fire image, (**2**) fire label, (**3**) CAM, (**4**) Grad-CAM++, (**5**) LayerCAM, (**6**) DWSLNet, (**7**) ours.

The results also suggest that the proposed method has a certain degree of cross-regional adaptability. The dataset covers 29 regions with considerable differences in vegetation type, topographic conditions, climatic background, and fire morphology. The consistent performance of the proposed method on the mixed multi-region test set, especially on IoU, F1, and Kappa, indicates that the model is not merely memorizing the texture pattern of a specific region, but is learning discriminative cues more closely related to thermal anomalies, burn-change characteristics, and boundary structure. Nevertheless, we also acknowledge that we have not yet conducted a strict leave-one-region-out validation or cross-dataset transfer evaluation. Therefore, the current evidence for generalization is mainly reflected in stable performance under mixed multi-region testing rather than true zero-shot transfer to unseen regions. Future work will further incorporate leave-one-region-out validation, cross-year validation, and cross-sensor validation to more systematically assess the generalization upper bound of the proposed method in real deployment scenarios.

From a practical application perspective, the proposed method may also provide useful support for forest management agencies and emergency decision makers. In routine monitoring and emergency duty, high-temporal-resolution hotspot products such as VIIRS and FIRMS can be used as rapid alert sources, after which Sentinel-2 imagery and the proposed model can be called to conduct fine verification of the target region, thereby improving hotspot confirmation efficiency and reducing false alarms. During firefighting resource deployment, the predicted fire-region contours can help identify fire-line locations, expansion boundaries, and priority control zones. In post-fire assessment and recovery planning, the CAMs and binary fire masks generated by the proposed method can serve as prior information for rapid burned-area extraction, fire-scar statistics, ecological loss assessment, and restoration zoning. However, in scenes with heavy cloud cover, strong smoke interference, or dense forest canopies, combining the proposed method with higher-temporal-resolution thermal infrared monitoring is still recommended in order to reduce the risk of missed detections and improve operational stability.

5. Conclusions

This study proposes a dual-branch multi-scale feature fusion framework for weakly supervised wildfire detection in high-resolution remote sensing imagery. By jointly modeling multispectral spatial-semantic features and SWIR temporal variation features and introducing attention-guided multi-stage fusion, the proposed method generates more accurate and stable CAMs for wildfire detection.

Experiments on a dataset covering 29 regions and 2206 wildfire images demonstrate that the proposed method outperforms representative CAM-based methods and a remote sensing-oriented weakly supervised baseline. The method achieves an IoU of 0.237 and an F1-score of 0.360, showing clear advantages in small-fire localization and boundary delineation. The results also indicate that, under the severe class imbalance of wildfire detection, OA can be easily inflated by the large background class, whereas Kappa, IoU, and F1 more faithfully reflect the recovery quality of the minority fire class. Although the proposed method does not achieve the highest OA, it performs better on these more task-relevant metrics, indicating that the improvement mainly comes from a more stable overall distinction between fire and background.

Despite the promising results, several limitations remain. The current framework still depends on the quality of SWIR imagery and uses fixed fusion weights across different wildfire types and regions. In addition, the current evaluation is still based on post-annotated data and is mainly intended to validate detection accuracy and generalization capability. Therefore, the proposed method is better suited as a high-resolution refinement tool after hotspot alerting than as a standalone minute-level early-warning system.

Future work can be pursued in three directions. First, more robust temporal modeling and noise-suppression mechanisms may further improve stability under complex meteorological and background conditions. Second, adaptive cross-modal fusion and region-aware strategies may enable the model to dynamically adjust the contributions of multispectral and temporal information according to different scenes. Third, stricter leave-one-region-out validation, cross-year validation, and cross-sensor validation should be conducted, and the proposed method can be integrated with high-temporal-resolution sensors such as VIIRS to build a coordinated workflow of rapid alerting plus fine identification for more realistic operational deployment.

Author Contributions: Conceptualization, D.W. and Y.T.; methodology, D.W. and Z.P.; software, Z.P.; validation, Z.P. and Y.T.; formal analysis, D.W. and Z.P.; investigation, Z.P.; resources, D.W. and Y.T.; data curation, Z.P.; writing—original draft preparation, D.W. and Z.P.; writing—review and editing, D.W. and Y.T.; visualization, Z.P.; supervision, D.W. and Y.T.; project administration, D.W.; funding acquisition, D.W. and Y.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant Nos. 41901279 and 42505175) and the Science Foundation Research Project of Wuhan Institute of Technology (Grant No. K202239).

Data Availability Statement: Data are available from the corresponding author upon reasonable request.

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

References

1. McLauchlan, K.K.; Higuera, P.E.; Miesel, J.; Rogers, B.M.; Schweitzer, J.; Shuman, J.K.; Tepley, A.J.; Varner, J.M.; Veblen, T.T.; Adalsteinsson, S.A.; et al. Fire as a fundamental ecological process: Research advances and frontiers. *J. Ecol.* **2020**, *108*, 2047–2069. [CrossRef]
2. Kramer, H.A.; Mockrin, M.H.; Alexandre, P.M.; Radeloff, V.C. High wildfire damage in interface communities in California. *Int. J. Wildland Fire* **2019**, *28*, 641–650. [CrossRef]
3. Zhu, D.; Yang, P. Study on the Evolutionary Characteristics of Post-Fire Forest Recovery Using Unmanned Aerial Vehicle Imagery and Deep Learning: A Case Study of Jinyun Mountain in Chongqing, China. *Sustainability* **2024**, *16*, 9717. [CrossRef]
4. Vetrita, Y.; Diwyacitta, K.; Sukarno, K.M.; Albar, I.; Usman, A.B.; Ritonga, R.P.; Santoso, I.; Lestari, A.I.; Kartika, T.; Novresiandi, D.A.; et al. Evaluating the capabilities of high-resolution PlanetScope and Sentinel-2 images for mapping burned area and vegetation regrowth in Indonesia's savannas. *Int. J. Remote Sens.* **2025**, *46*, 6803–6825. [CrossRef]
5. Hall, J.V.; Argueta, F.; Zubkova, M.; Chen, Y.; Randerson, J.T.; Giglio, L. GloCAB: Global cropland burned area from mid-2002 to 2020. *Earth Syst. Sci. Data* **2024**, *16*, 867–885. [CrossRef]
6. Chuvieco, E.; Lizundia-Loiola, J.; Pettinari, M.L.; Ramo, R.; Padilla, M.; Tansey, K.; Mouillot, F.; Laurent, P.; Storm, T.; Heil, A.; et al. Generation and analysis of a new global burned area product based on MODIS 250 m reflectance bands and thermal anomalies. *Earth Syst. Sci. Data* **2018**, *10*, 2015–2031. [CrossRef]
7. Franquesa, M.; Stehman, S.V.; Chuvieco, E. Assessment and characterization of sources of error impacting the accuracy of global burned area products. *Remote Sens. Environ.* **2022**, *280*, 113214. [CrossRef]
8. NASA. Fire Information for Resource Management System (FIRMS). Available online: https://firms.modaps.eosdis.nasa.gov/ (accessed on 1 January 2025).
9. Xiang, J.; Xing, Y.; Wei, W.; Yan, E.; Jiang, J.; Mo, D. Dynamic Detection of Forest Change in Hunan Province Based on Sentinel-2 Images and Deep Learning. *Remote Sens.* **2023**, *15*, 628. [CrossRef]
10. Zhou, B.; Gao, S.; Yin, Y.; Zhong, Y. Enhancing active fire detection in Sentinel 2 imagery using GLCM texture features in random forest models. *Sci. Rep.* **2024**, *14*, 31076. [CrossRef]
11. Al-Dabbagh, A.M.; Ilyas, M. Uni-temporal Sentinel-2 imagery for wildfire detection using deep learning semantic segmentation models. *Geomat. Nat. Hazards Risk* **2023**, *14*, 1. [CrossRef]
12. Hong, Z.; Tang, Z.; Pan, H.; Zhang, Y.; Zheng, Z.; Zhou, R.; Ma, Z.; Zhang, Y.; Han, Y.; Wang, J.; et al. Active fire detection using a novel convolutional neural network based on Himawari-8 satellite images. *Front. Environ. Sci.* **2022**, *10*, 794028. [CrossRef]

13. Wang, J.; Wang, Y.; Liu, L.; Yin, H.; Ye, N.; Xu, C. Weakly Supervised Forest Fire Segmentation in UAV Imagery Based on Foreground-Aware Pooling and Context-Aware Loss. *Remote Sens.* **2023**, *15*, 3606. [CrossRef]
14. Hu, Z.; Gao, J.; Yuan, Y.; Li, X. Contrastive Tokens and Label Activation for Remote Sensing Weakly Supervised Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5620211. [CrossRef]
15. Chen, Z.; Lian, Y.; Bai, J.; Zhang, J.; Xiao, Z.; Hou, B. Weakly Supervised Semantic Segmentation of Remote Sensing Images Using Siamese Affinity Network. *Remote Sens.* **2025**, *17*, 808. [CrossRef]
16. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Las Vegas, NV, USA, 2016; pp. 2921–2929. [CrossRef]
17. Li, J.; Zhu, L.; Wang, W.; Gong, Y. Attention Guided Class Activation Maps for Boosting Weakly Supervised Semantic Segmentation. *Eng. Appl. Artif. Intell.* **2025**, *151*, 110556. [CrossRef]
18. Amaral, B.; Niknejad, M.; Barata, C.; Bernardino, A. Weakly Supervised Fire and Smoke Segmentation in Forest Images with CAM and CRF. In *2022 26th International Conference on Pattern Recognition (ICPR)*; IEEE: Montreal, QC, Canada, 2022; pp. 442–448. [CrossRef]
19. White, L.A.; Gibson, R.K. Comparing fire extent and severity mapping between Sentinel-2 and Landsat 8 satellite sensors. *Remote Sens.* **2022**, *14*, 1661. [CrossRef]
20. Gui, L.; Li, R.; Tao, M.; Jiang, Z.; Wang, L.; Wang, J.; Huang, J.; Wang, Y.; Chen, L.; Salinas, S.V. A General and Robust Active Fire Detection Method Using Sentinel-2 and Landsat Shortwave Infrared Measurements. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 4100314. [CrossRef]
21. Wickramasinghe, C.H.; Jones, S.; Reinke, K.; Wallace, L. Development of a multi-spatial resolution approach to the surveillance of active fire lines using Himawari-8. *Remote Sens.* **2016**, *8*, 932. [CrossRef]
22. Roy, D.P.; Huang, H.; Boschetti, L.; Giglio, L.; Yan, L.; Zhang, H.H.; Li, Z. Landsat-8 and Sentinel-2 burned area mapping: A combined sensor multi-temporal change detection approach. *Remote Sens. Environ.* **2019**, *231*, 111254. [CrossRef]
23. Mazzeo, G.; Falconieri, A.; Filizzola, C.; Genzano, N.; Pergola, N.; Marchese, F. Wildfire Detection and Mapping by Satellite With an Enhanced Configuration of the Normalized Hotspot Indices: Results From Sentinel-2 and Landsat 8/9 Data Integration. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5606121. [CrossRef]
24. Cao, Y.; Lei, R. Progress and prospect of forest fire monitoring based on multi-source remote sensing data. *Natl. Remote Sens. Bull.* **2024**, *28*, 1854–1869. [CrossRef]
25. Ding, Z.; Yang, Y.; Zhang, Y.; Luo, X.; Huang, M.; Xiang, X. Cross-Modal Feature Calibration and Fusion Network for Remote Sensing Optical–SAR Joint Object Detection Under Cloud Occlusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 27302–27319. [CrossRef]
26. Sali, A.; Nomqupu, S.; Nyamugama, A.; Ndou, N. Smoke characterization for incipient wildfire detection from Sentinel-2 sensor based on sigmoid activation function and momentum gradient descent optimizer. *Earth Sci. Inform.* **2025**, *18*, 488. [CrossRef]
27. Loschilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2019**, arXiv:1711.05101. [CrossRef]

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.