remote sensing

MDPI

Article

A Fire Segmentation Method with Flame Detail Enhancement U-Net in Multispectral Remote Sensing Images Under Category Imbalance

Rui Zou [1], Zhihui Xin [1,*], Guisheng Liao [2], Penghui Huang [3], Rui Wang [1] and Yuhu Qiao [1]

[1] School of Physics and Electronic Information Technology, Yunnan Normal University, Kunming 650500, China; 2223090030@ynnu.edu.cn (R.Z.); 2324090020@ynnu.edu.cn (R.W.); 2324090013@ynnu.edu.cn (Y.Q.)
[2] National Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China; liaogs@xidian.edu.cn
[3] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; huangpenghui@sjtu.edu.cn
* Correspondence: xinzhihui@ynnu.edu.cn

Abstract

Fire poses a serious threat to the global economy, environment, and social stability, highlighting the need for rapid and accurate fire detection. Remote sensing combined with deep learning has outperformed traditional fire assessment methods. However, in early fire stages, small flame areas, class imbalance, and weak feature extraction hinder detection accuracy. This study proposes an end-to-end segmentation model called Flame Detail Enhancement U-Net (FDE U-Net), using Landsat-8 multispectral remote sensing data. The model incorporates the self-Attention and Convolutional mixture (ACmix) module and the Convolutional Block Attention Module (CBAM) into the encoder of the Residual U-Net. ACmix integrates self-attention and convolution to capture global semantic features while maintaining computational efficiency, improving both contextual awareness and local detail. CBAM enhances flame recognition by weighting important channel features and focusing spatially on small flame areas, helping address the class imbalance problem. Additionally, Haar wavelet downsampling is applied to retain image detail and improve the detection of small-scale flame regions. Experimental results show that the FDE U-Net model exhibits robust performance in fire detection, accurately extracting flame regions even when their proportion is low and the background is complex. The F1 score reaches 95.97%, significantly improving the class imbalance problem.

Keywords: FDE U-Net; feature fusion; fire segmentation; multispectral remote sensing images

check for updates

Academic Editor: Salah Bourennane

Received: 12 May 2025
Revised: 17 June 2025
Accepted: 23 June 2025
Published: 25 June 2025

Citation: Zou, R.; Xin, Z.; Liao, G.; Huang, P.; Wang, R.; Qiao, Y. A Fire Segmentation Method with Flame Detail Enhancement U-Net in Multispectral Remote Sensing Images Under Category Imbalance. *Remote Sens.* **2025**, *17*, 2175. https://doi.org/10.3390/rs17132175

Copyright: © 2025 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Fires have recently become more frequent, destroying millions of hectares of forest annually, causing significant economic losses, deteriorating air quality, and reducing biodiversity. High temperatures and dry weather conditions caused by global warming increase the risk of fire [1]. Figure 1 shows the area burned by wildfires worldwide as of 12 June 2025. The figure clearly shows that fire is a major problem. Therefore, detecting forest fires rapidly and accurately is crucial. The key to fire detection is accurately identifying fire locations, a task strongly supported by computer vision technology [2].

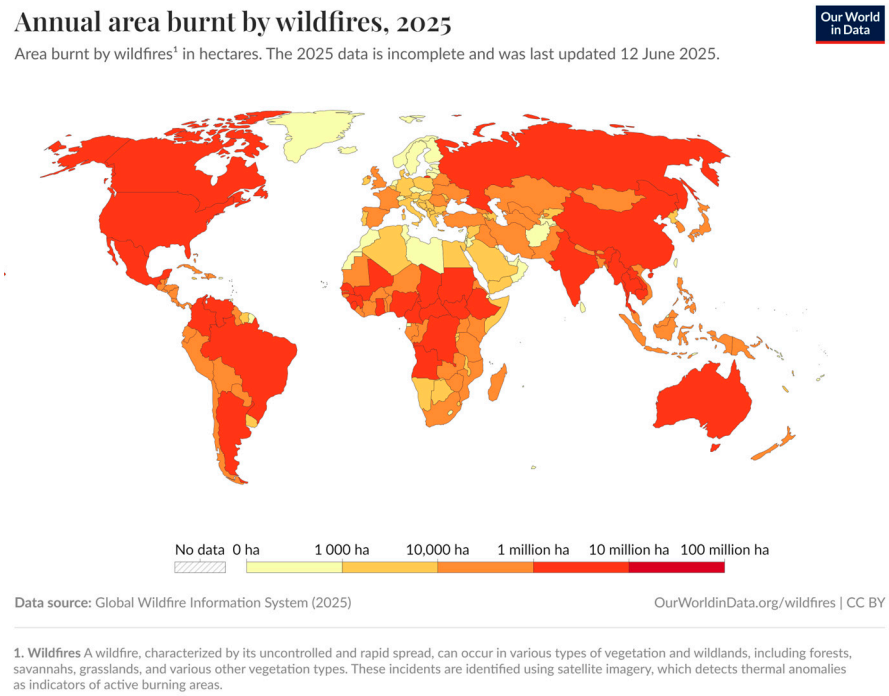


Figure 1. Area burnt by wildfires [3].

Traditional fire detection methods, such as using watchtowers located at high points to observe fire conditions, have improved efficiency and reduced labor intensity compared with manual inspections [4]. However, they are limited by the number of towers, terrain, and topography and have large blind spots, low accuracy, and high costs, making them difficult to use widely. Unmanned aerial vehicles can achieve large-scale, rapid, and effective real-time monitoring with a wide field of view and high efficiency. However, they have safety risks, cannot return images in real time, and are limited by night flights and environmental factors [5,6]. Whether it is UAV inspection or watchtower monitoring, vision-based algorithms have weak detection capabilities for small flame targets owing to camera pixel limitations, and these systems cannot achieve global coverage. Satellite-based systems utilize satellites to provide services. Compared to terrestrial systems, they can achieve global coverage and provide secure connections that are not affected by physical or weather obstacles, making them a viable option for forest fire detection [7,8].

Diverse types of satellite imagery are widely used in fire detection as the basis for monitoring and analysis. The moderate resolution imaging spectroradiometer (MODIS) and the visible infrared imaging radiometer suite (VIIRS) sensors are both capable of capturing multiple spectral bands, enabling effective identification of fire-related thermal radiation and smoke characteristics [9,10]. Cardíl et al. conducted the first systematic analysis of the rate of spread (ROS) of large wildfires in Northwestern Europe using VIIRS satellite active fire data, revealing the influence of different land cover types and seasons on fire spread and providing critical insights for risk management and response strategies in this emerging fire-prone region [11]. Zhang et al. jointly used VIIRS 375 m I-Band and 750 m M-Band data to detect and evaluate fires and fire radiated power (FRP) and proposed a VIIRS-IM method that improved the sensitivity and accuracy of small-area fire detection, which provided more fire pixels and more accurate FRP records compared to MODIS data [12]. Hong et al. proposed the FireCNN model based on Himawari-8 satellite images using multiscale convolution and residual concatenation to efficiently extract the accurate features of fire points with an accuracy rate 35.2% higher than that of the traditional threshold method [13]. These satellite data are used for active fire detection

(AFD). However, their low spatial resolution can reduce AFD accuracy. The Landsat-8 satellite, with its high spatial and temporal resolution, overcomes this problem [14].

In addition to relying on reliable image data, the exploration of accurate and efficient fire detection methods has become a key research topic. Researchers have proposed and applied various popular methods, most of which focus on color and shape features [15,16]. For example, Çelik et al. used the YCbCr color space to better distinguish brightness and chromaticity and constructed a universal flame pixel classification chromaticity model, which effectively reduced illumination change impact with high detection and low false alarm rates [17]. Chen et al. proposed a fire multi-feature fusion recognition algorithm, which outperforms the traditional method in performance, but the recall is slightly lower than the deep learning method [18]. Jang et al. combined thresholding, random forest, and postprocessing techniques to enhance small fire detection accuracy [19]. Although the above methods have achieved good results, they suffer from poor generalization ability and suboptimal performance in fire detection in complex backgrounds, making them insufficient for most fire detection tasks.

Owing to its powerful feature extraction capabilities and the ability to process high-dimensional data, deep learning has achieved remarkable results in areas such as image classification [20], target detection [21,22], and image segmentation [23,24]. The use of deep learning methods for fire detection is considerably effective [25,26]. Sun et al. significantly reduced the computational complexity of YOLO v7 while maintaining high fire detection accuracy by improving activation functions, lightweight modules, and backbone network reconstruction [27]. Hu et al. proposed a network architecture called AF-Net, which significantly improves the detection accuracy of imbalanced fire data in UAV remote sensing images by optimizing the sampling strategy and enhancing the object-contextual representation (OCR) structure. Numerous studies have focused on deep-learning-based approaches and used Landsat-8 data to detect fires [28]. Seydi et al. proposed Fire-Net, a forest fire detection approach based on deep learning, which utilizes residual and separable convolutional modules. The model was trained on 578 Landsat-8 images and tested on multiple scenes, achieving high accuracy in all cases [29]. Han et al. constructed the BA-BS dataset based on Landsat multispectral remote sensing imagery and auxiliary environmental data and proposed a transformer-based change detection model incorporating auxiliary information, which demonstrated excellent performance in burned area identification and burn severity classification tasks [30]. The U-Net model has become an important benchmark model in image segmentation owing to its symmetrical structure, fully convolutional network, multiscale feature fusion, and end-to-end training. It can capture finer details and produce more accurate results [31–33]. Xu et al. proposed the MSU-Net model, which introduced self-attention and multi-scale dilated convolutional structures based on U-Net for high-precision extraction of tea-oil plantation areas in high-resolution UAV images. This method places special emphasis on global modeling and small target recognition capabilities, demonstrating robustness and practicality under hilly and complex terrain conditions [34]. Wang et al. proposed Smoke-Unet by improving the U-Net method and combining spatial-based and channel-based attention with residual blocks. A smoke dataset based on multispectral remote sensing was developed, encompassing multiple years, diverse seasons, geographic regions, and land cover categories. The accuracy rate reached 92.3%, which was better than the basic methods such as U-Net, FCN [35], PSPNet [36], etc [37]. Afsar et al. proposed an encoder–decoder-encoder–decoder structure, ResWnet, which reduced feature loss during sampling and retained more detailed features. The accuracy of ResWnet was 95%, and the F1 score was 94.2% [38]. Compared with traditional methods, the above models achieved higher detection accuracy. However, limited feature expression ability, insufficient capture of microscopic targets, and inadequate processing of

local features remain problems. One of the main reasons for the lack of previous research on the use of satellite imagery for fire detection is the lack of suitable datasets. de Almeida Pereira et al., referring to the method proposed in [39–41], introduced a large-scale multispectral dataset, large-scale active fire detection (LAFD), that can be used for fire detection in 2021 [42]. It contained more than 150,000 image blocks, providing strong support for fire detection.

Owing to the 30 m resolution, wide coverage, and multiple bands of Landsat 8 multispectral data, substantial storage and computational resources are required, which limits the processing speed. Additionally, the images are often covered by clouds and cloud shadows, which affect the observations. The fire area is usually mixed with a complex background. The existing model has limited feature expression ability, and it is difficult to accurately segment the fire area. The shapes and sizes of flames vary, and existing algorithms have limited capabilities in capturing small targets and adequately processing local features. Therefore, this study applies semantic segmentation to fire detection based on an LAFD dataset [42] provided by de Almeida Pereira et al. and proposes a new segmentation network, Flame Detail Enhancement U-Net (FDE U-Net). It achieves the adaptive selection and mixing of input features, effectively capturing subtle features in the early stages of a fire. When analyzing fire images, local information is usually more important than global information, as it can provide more accurate pixel-level information regarding fire areas. The main contributions of this study are as follows:

(1) FDE U-Net comprehensively uses advanced feature extraction and enhancement techniques to increase the depth and breadth of feature extraction, optimize information transfer, and significantly improve detection accuracy and efficiency. Multilevel feature processing and fusion enable the model to more accurately identify and locate flame details when it faces complex multispectral data.

(2) The self-Attention and Convolutional mixture (ACmix) and Convolutional Block Attention Module (CBAM) are integrated into the encoder part of ResU-Net to achieve adaptive selection and mixing of the input features, which significantly improves the category imbalance problem and, thus, the detection accuracy of the model for the flame region.

(3) The Haar wavelet downsampling (HWD) can retain more boundary, texture, and detail information while reducing the size of the feature map using the Haar wavelet transform. The feature representation and detail capture capabilities are enhanced, effectively capturing tiny features in the early stages of a fire.

The following sections further introduce and discuss the proposed network. Section 2 provides a detailed introduction to the basic structure of the proposed network. The details of the dataset are presented in Section 3. Section 4 presents the experimental results and discusses them from both quantitative and qualitative perspectives. Section 5, as the final part, draws the conclusion of this paper.

2. Model

2.1. Network Architecture

We optimize the U-Net architecture and propose a dual-path encoder–decoder-based method, FDE U-Net, to improve its feature learning capability and the efficiency of contextual information acquisition for the accurate segmentation of fire regions. Figure 2 shows the overall network architecture of the FDE U-Net.

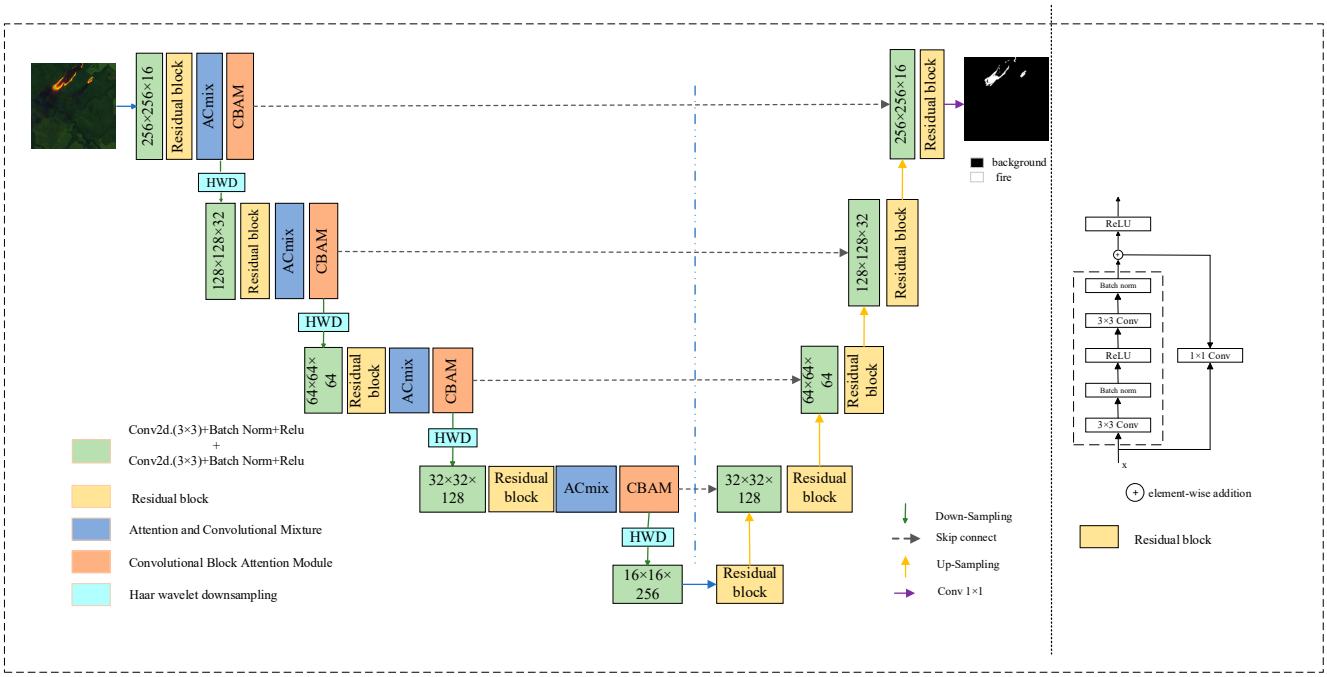


Figure 2. Overall architecture of FDE U-Net.

FDE U-Net acts as an end-to-end encoder–decoder network, with the encoder portion receiving the input image and the number of channels extended to 16, enabling the network to extract richer feature information from the input data, which helps to capture spatial and spectral features in multi-band images. The encoder uses four residual blocks, which effectively mitigates the gradient-vanishing problem in deep network training, improving the model's ability to extract deep features. The ACmix and CBAM are integrated into each downsampling layer. The ACmix module adaptively blends cross-channel features between multispectral bands by combining convolution and self-attention operations to promote effective interaction and fusion of features between different channels. This enables the model to effectively integrate the differentiated spectral characteristics between different bands, enriches the feature expression, and improves the model's ability to understand complex image content. To further improve the model's performance, CBAM refines the weighting of feature maps in two dimensions, space and channel, such that the model can focus on key areas in the image. CBAM can adjust its focus based on the response characteristics of different bands. For example, in the thermal infrared band, it enhances the fire area, reduces background noise, and effectively separates the fire. In the spatial dimension, it focuses on the fire's morphology and boundaries, thereby improving segmentation accuracy. Regarding the downsampling strategy, the traditional maximum pooling operation is abandoned, and HWD is adopted. The HWD utilizes the multiscale analysis capability of the wavelet transform and reduces the resolution of feature maps through the Haar wavelet transform. While reducing the resolution, the fire edges, texture details, and cross-band contextual association information in the multispectral image are retained, which retains more image detail information and provides richer contextual information for subsequent segmentation tasks.

Specifically, the decoder gradually restores the image size through upsampling. After each up-sampling, the spatial resolution of the feature map increased, whereas the number of feature channels is halved, yielding deep features that are skip-connected to the corresponding output features from the encoder. By leveraging skip connections, the model can integrate low-level details with high-level semantics, which supports accurate boundary reconstruction and mitigates the degradation of crucial features when restoring

spatial resolution. In the final layer, classification is performed using a 1×1 convolution followed by a Softmax activation function, resulting in fine segmentation. Simultaneously, batch normalization and dropout layers are added after each convolution layer to avoid overfitting. Finally, a binary cross-entropy (BCE) loss function is used to measure the difference between the model's predicted probability and the actual category to evaluate the model's performance. In binary segmentation tasks, the binary cross-entropy (BCE) loss is formulated as follows:

$$BCE_{loss} = -\frac{1}{n}\sum_{i=1}^{n}(y_i log(p) + (1 - y_i) log(1 - p)) \quad (1)$$

where p is the probability of the fire class, $1 - p$ is the probability of the non-fire class, and y_i is the pixel label.

2.2. Self-Attention and Convolutional Mixture (ACmix)

Convolution and self-attention are two powerful representation-learning techniques. For a standard convolution operation, Figure 3a shows a 3×3 convolution decomposed into nine 1×1 convolutions, followed by shift summation [43]. First, a linear projection of the input feature map is performed using the kernel weights at a specific location; the projected feature map is then shifted and aggregated based on the position. For the self-attention operation, as shown in Figure 3b, the input features are first projected into queries, keys, and values using a 1×1 convolution, after which the attention weights are computed and the values are aggregated. As shown in the figure, the first step of both operations includes a 1×1 convolution. ACmix combines the advantages of convolution and self-attention to achieve the dual functions of the deep extraction of spatial features and efficient aggregation of global information.

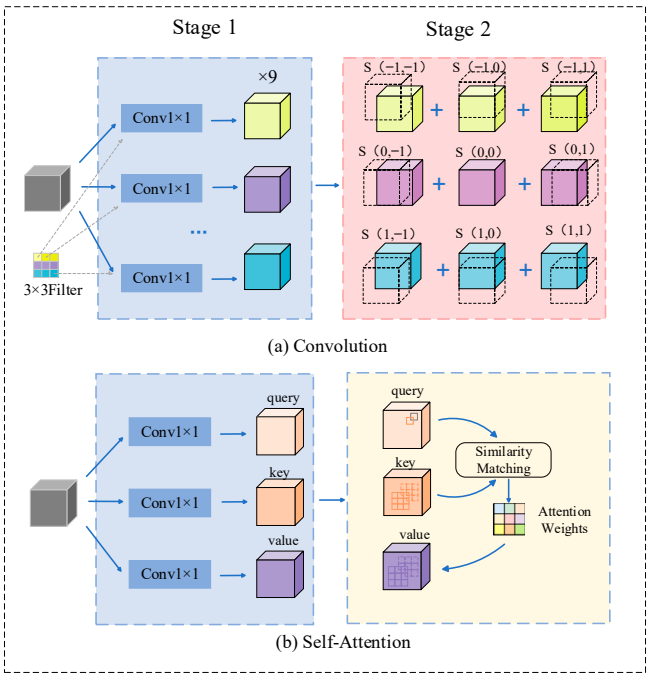


Figure 3. (**a**) A 3×3 convolution decomposition and shift summation process; (**b**) self-attention operation process.

The second step is to reuse and aggregate the intermediate features in different modes (self-attention and convolution). When obtaining the self-attention feature, the intermediate feature set is divided into N groups. Each group contains three feature maps, and the three feature maps are used as *query*, *key*, and *value* matrices to obtain the self-attention feature

$f_{self-attention}$. In obtaining the convolution features, assuming that the convolution kernel size is k, a fully connected layer is used to obtain k^2 feature maps, and the convolution feature f_{conv} is calculated and summed. To eliminate the inefficiency of tensor shift operations, we employ depthwise convolutions with fixed kernels as a lightweight and effective alternative. If we denote the convolution kernel (kernel size $k = 3$) as

$$K_c = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \forall c \tag{2}$$

The corresponding output can be formulated as

$$\begin{aligned} f_{c,i,j}^{(dwc)} &= \sum_{p,q \in \{0,1,2\}} K_{c,p,q} f_{c,i+p-\lfloor k/2 \rfloor, j-q-\lfloor k/2 \rfloor} \\ &= f_{c,i-1,j-1} = \widetilde{f}_{c,i,j}, \forall c, i, j \end{aligned} \tag{3}$$

where c represents each channel of the input feature. $\widetilde{f}_{c,i,j}$ represents the calculation result of tensor shifts.

Therefore, with carefully designed kernel weights for specific shift directions, the convolution outputs are equivalent to the simple tensor shifts. This will improve the actual efficiency of the module during inference. Finally, ACmix reuses intermediate features to avoid repeated and complex double-projection operations on the same feature set, thereby reducing redundant calculations, improving computational efficiency, and reducing model complexity while maintaining high performance. Finally, using two learnable scalars, the self-attention and convolution features are organically combined to obtain the final fusion feature. Figure 4 shows the specific operation process of ACmix. The expression is as follows:

$$f_{out} = \alpha \times f_{self-attention} + \beta \times f_{conv} \tag{4}$$

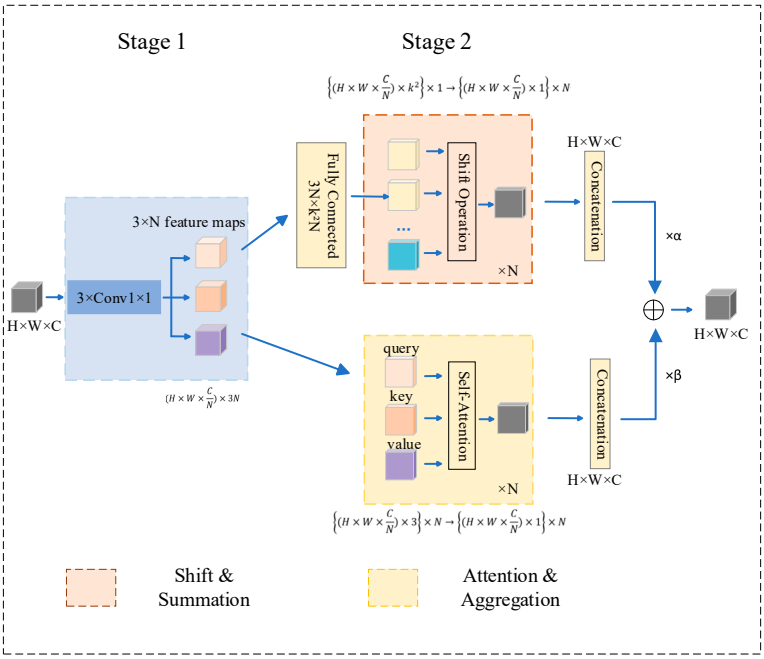


Figure 4. Overall structure of ACmix.

2.3. Convolutional Block Attention Module (CBAM)

During the low-level feature fusion between the encoder and decoder, the CBAM attention mechanism is employed to enhance the extraction of valuable details, guiding

the model to better focus on flame regions and boosting its overall performance. CBAM combines spatial and channel attention in a convolutional block [44]. Figure 5 shows its structure. After being processed by the convolutional layer, the output is first passed through the channel attention mechanism to apply weighting, and then refined by the spatial attention mechanism to produce the final output. The CBAM is introduced to enhance the features associated with the flame area. Background regions unrelated to fire are effectively suppressed in the attention feature map, unlike in the raw features extracted by the base network.

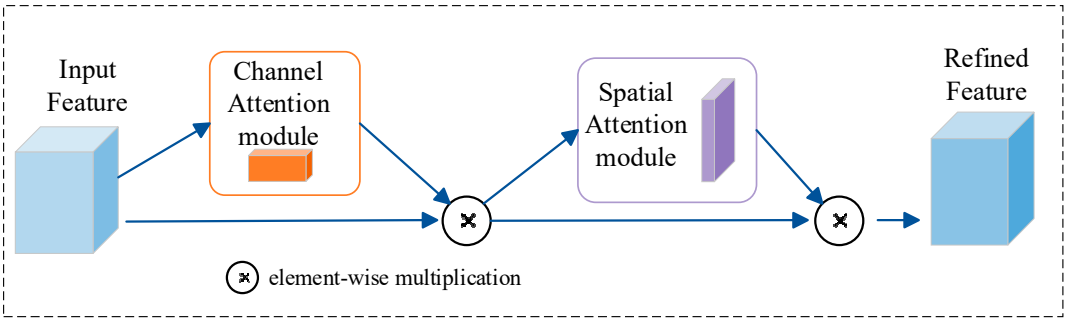


Figure 5. Convolutional block attention module.

Figure 6 shows the details of the channel attention implementation. Input feature map F is compressed in spatial dimension by global average pooling (GAP) and global maximum pooling (GMP) operations to obtain one-dimensional vectors F^c_{avg} and F^c_{max}, corresponding to the feature representations obtained through average and maximum pooling, respectively. The output sizes of the GAP and GMP are $1 \times 1 \times C$. They are then fed into a shared multilayer perceptron (MLP), comprising two fully connected layers, producing outputs denoted as M_{avg} and M_{max}. The outputs of the GAP and GMP after the MLP are summed element-by-element to obtain the channel attention weight map M_c. The original input feature map F is multiplied by the channel attention weight map M_c to adjust the weight of each channel. The channel attention mechanism can selectively enhance the important feature channels and suppress irrelevant or redundant channels by calculating the importance weight of each channel. The computation of channel attention is defined as follows:

$$M_{avg} = W_1(W_0(F^c_{avg})) \quad (5)$$

$$M_{max} = W_1(W_0(F^c_{max})) \quad (6)$$

$$M_c = \sigma(M_{avg} + M_{max}) \quad (7)$$

where σ represents the sigmoid function, and W_0 and W_1 are the two input-shared MLP weight matrices.

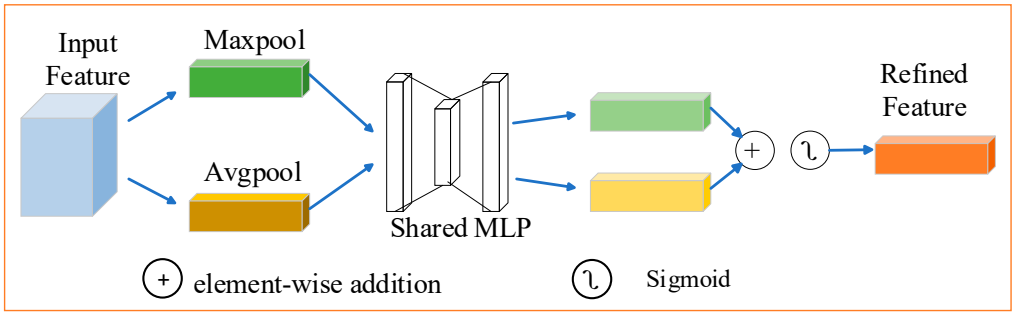


Figure 6. Channel attention module.

The implementation details of spatial attention are shown in Figure 7. It mainly calculates the importance weight of each position to perform weighted adjustment on the input feature map and highlight the key spatial features. The feature map produced by the channel attention mechanism is first utilized as the input feature representation for this module. Global average pooling (GAP) and global maximum pooling (GMP) are performed on the input feature map, and pooling is performed along the channel dimension, respectively, to obtain two single-channel feature maps F^S_{avg} and F^S_{max}. The output sizes of GAP and GMP are both H × W × 1. Then the single-channel feature maps generated by GAP and GMP are concatenated in the channel dimension to obtain a feature map of size H × W × 2. Finally, a 7 × 7 convolutional layer is applied to the concatenated feature map to produce the spatial attention weight map M_S. The original input feature map F is multiplied by the spatial attention weight map M_S to perform a weighted adjustment on each position. Spatial attention is calculated as follows:

$$\begin{aligned} M_S &= \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right) \\ &= \sigma\left(f^{7\times7}\left(\left[F^S_{avg}; F^S_{max}\right]\right)\right) \end{aligned} \tag{8}$$

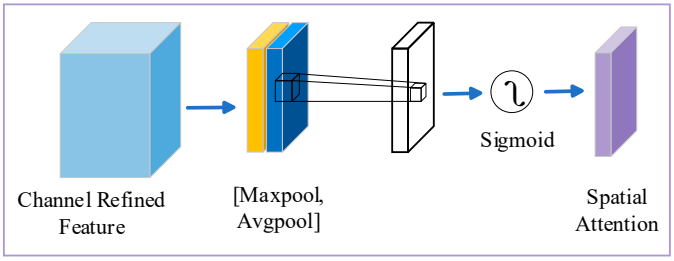


Figure 7. Spatial attention module.

2.4. Haar Wavelet Downsampling (HWD)

Downsampling techniques like max pooling are commonly employed in convolutional neural networks (CNNs) to consolidate local feature information, enlarge the receptive field, and reduce computational complexity. However, in semantic segmentation tasks, performing pooling operations over local neighborhoods can inevitably lead to the loss of crucial spatial information, thereby weakening the model's ability to preserve details. HWD retains more detail and edge features by decomposing the image. Compared to traditional maximum pooling or average pooling, HWD can better preserve the original information of the image and reduce information loss [45].

The HWD consists of two main parts (Figure 8). The first part of the lossless feature-encoding block reduces the spatial resolution of the feature map using the Haar wavelet transform. The input image is divided into four sub-bands using Haar wavelet decomposition: a single low-frequency component, denoted as LL, along with three high-frequency components, namely LH, HL, and HH. By downsampling the above components, the feature map size is reduced to half of the original size, whereas the number of feature map channels is increased fourfold. The size of the output feature map was $\frac{H}{2} \times \frac{W}{2} \times 4C$. In the second stage of the feature representation learning block, redundant information is suppressed using a standard 1 × 1 convolution, followed by batch normalization and ReLU activation. Standard convolution is applied to modify the channel dimensions of the feature map, aligning it with subsequent layers and enabling them to more effectively capture meaningful and representative features.

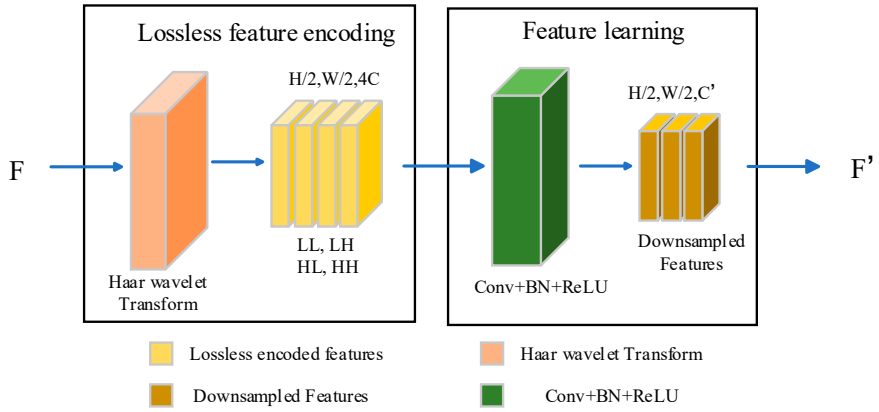


Figure 8. Haar wavelet downsampling.

3. Datasets and Evaluation Metrics

3.1. Datasets

Pereira et al. released a new large multispectral dataset called LAFD in 2021 [40]. This dataset was extracted from the Landsat-8 satellite launched by NASA and included wildfires that occurred at various locations worldwide between August and September 2020. The samples covered various areas, including deserts, rainforests, farmlands, water, snow, clouds, cities, and mountains. All image blocks in the dataset were 256 × 256 pixels. These image blocks were 10-band, 16-bit TIFF images. The 10 bands included the coastal band (COASTAL/AEROSOL), three RGB bands, near-infrared (NIR), shortwave infrared (SWIR 1 and SWIR 2), cirrus band (Cirrus), and thermal infrared band (TIR1 and TIR2). The channels were b1–b11, excluding b8 (panchromatic channel), with a spatial resolution of 30 m.

A significant category imbalance is observed between the fire and non-fire pixels in the dataset. We select image patches with a high and low proportion of fire pixels to detect small fire regions. A total of 1100 image blocks are selected from the huge LAFD dataset, and these data are horizontally and vertically flipped, ultimately generating 3300 image blocks to enhance the diversity of the dataset. The dataset was divided with 80% used for the training and validation processes and 20% reserved for testing to ensure the independence and effectiveness of the model evaluation. Our dataset refers to the three algorithms discussed in [39–41]. We use a "three-round, two-win" voting scheme (if at least two sets of conditions agree that it was a fire pixel, it is considered a fire) to generate the dataset. In Figure 9, (a) shows one of the fire scenes; (b), (c), and (d) represent the masks generated by the three algorithms; and (e) represents the voting result. The three algorithms are as follows:



(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 9. Examples of masks for each algorithm. (**a**) Original image. (**b**) Schroeder. (**c**) Murphy. (**d**) Kumar and Roy. (**e**) Voting.

(1) Schroeder et al. [39]

Seven Landsat-8 channels (c1 to c7) are used to identify the fire pixels. Unambiguous fire pixels must meet the following criteria:

$$\begin{array}{l}[(R_{75} > 2.5)and(\rho_7 - \rho_5 > 0.3)and(\rho_7 > 0.5)]or\\ [(\rho_6 > 0.8)and(\rho_1 < 0.2)and(\rho_5 > 0.4or\rho_7 < 0.1)]\end{array} \quad (9)$$

For each fire candidate pixel, the neighborhood statistics of a 61 × 61 pixel window are adjusted as follows:

$$\begin{array}{l}(R_{75} > 1.8)and(\rho_7 - \rho_5 > 0.17)and(R_{76} > 1.6)and\left(R_{75} > \mu_{R_{75}} + max\left(3\sigma_{R_{75}}, 0.8\right)\right)\\ and\left(\rho_7 > \mu_{\rho_7} + max\left(3\sigma_{\rho_7}, 0.08\right)\right)\end{array} \quad (10)$$

(2) Murphy et al. [41]

A set of context-independent conditions is proposed so that it does not rely on the statistical data calculated from each pixel's neighborhood. Only channels c5, c6, and c7 of Landsat-8 are used. The unambiguous fire pixels must satisfy the following conditions:

$$(R_{76} \geq 1.4)and(R_{75} \geq 1.4)and(\rho_7 \geq 0.15) \quad (11)$$

Potential fires located within the vicinity of an identified fire (within a 3 × 3 window) are classified as fires. A potential fire meets the following conditions:

$$(R_{65} \geq 2)and(\rho_6 \geq 0.5)or((\rho_7\ is\ saturated)or(\rho_6\ is\ saturated)) \quad (12)$$

(3) Kumar and Roy [40]

Using channels c2 to c7, unambiguous fire pixels must meet the following conditions:

$$\rho_4 \leq 0.53\rho_7 - 0.214 \quad (13)$$

Potential fire pixels are identified based on the following conditions:

$$(\rho_4 \leq 0.53\rho_7 - 0.125)or(\rho_6 \leq 1.08\rho_7 - 0.048) \quad (14)$$

$$\left(R_{75} > \mu_{R_{75}} + max\left(3\sigma_{R_{75}}, 0.8\right)\right)and\left(\rho_7 > \mu_{\rho_7} + max\left(3\sigma_{\rho_7}, 0.08\right)\right) \quad (15)$$

where ρ_k represents the reflectivity of channel k. R_{ij} is the ratio of the reflectance of channels i and j. μ and σ represent the mean and standard deviation of the pixels in the 61 × 61 window pixel neighborhood.

3.2. Evaluation Metrics

We evaluate the architecture based on the main metrics used in semantic segmentation: precision (P), recall (R), F1 score (F1), and intersection over union (IoU). Precision refers to the proportion of pixels that belong to the fire region among all the pixels predicted by the model as fire regions. Recall measures the proportion of correctly identified fire pixels relative to the total number of actual fire pixels. The F1 score, defined as the harmonic mean of precision and recall, reflects the balance between these two metrics, with a higher value indicating better overall performance. However, owing to the severe imbalance between fire and non-fire categories, a more convincing criterion is required to understand the actual performance of the model. Therefore, we use the IoU, which measures the degree of overlap between the predicted and real segmentation results. The IoU and F1 score are complementary in evaluating model performance. The IoU focuses on the spatial overlap between the predicted and ground truth results, whereas the F1 score

emphasizes the accuracy and completeness of the detection results. Therefore, we select the IoU and F1 score as the main metrics to evaluate the segmentation performance, as they comprehensively reflect the model's performance from different perspectives. The metrics are defined as follows:

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

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{18}$$

$$IoU = \frac{TP}{TP + FN + FP} \tag{19}$$

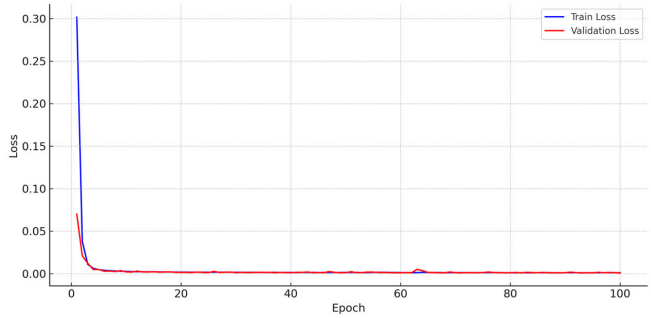
where TP, FP, and FN refer to the number of samples correctly predicted to be in the positive category (correct detection), the number of negative samples incorrectly predicted to be in the positive category (false positive), and the misclassified positive samples categorized as negatives, referred to as false negatives, respectively.

4. Analysis of Fire Detection Results

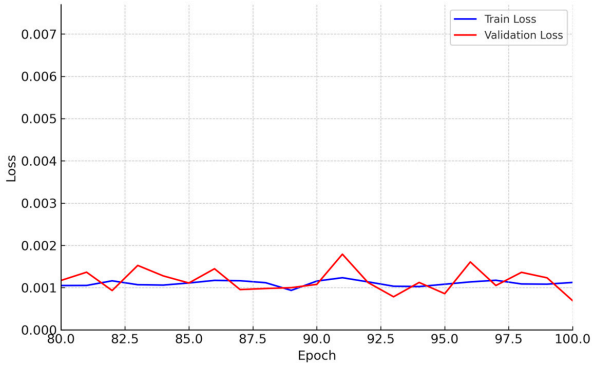
4.1. Training Parameter Settings

All experiments were implemented using PyTorch 1.12.1 and conducted on an NVIDIA GeForce RTX 3080 Ti 12 GB GPU. The training process employed the Adam optimizer, initialized with a learning rate of 1×10^{-3}, and BCE was used as the loss function to train each architecture for 100 epochs.

Figure 10 shows the loss curve of the proposed architecture. It can be seen that the loss curve decreases and eventually stabilizes. Early stopping was applied to prevent overfitting, and the training was stopped when the validation loss did not improve within five epochs. The batch size used during training was set to 32 to balance the memory capacity and training speed.



(**a**) Original image



(**b**) Enlarged image of epoch 80–100

Figure 10. FDE U-Net loss curve.

4.2. Ablation Experiments

We conduct a detailed performance analysis through ablation experiments to verify the effectiveness of the adopted modules. Table 1 presents the ablation experimental results based on ResU-Net. Before each downsampling of the ResU-Net encoder, ACmix, HWD, and CBAM are added separately, and two of them are added at the same time, which are denoted as Improvements 1 to 6, respectively. After adding ACmix and CBAM, the maximum pooling is replaced by HWD and expressed as FDE U-Net. As shown in Table 1, significant improvements are observed in four performance indicators: precision (P), recall (R), IoU, and F1 score. ACmix + CBAM + HWD provides the greatest improvement, with all metrics reaching an optimal level.

Table 1. Ablation experiment results based on ResU-Net.

Method	ACmix	CBAM	HWD	P (%)	R (%)	IoU (%)	F1 (%)
ResU-Net				96.03	84.22	81.31	89.69
Improvement 1	√			95.28	95.28	90.99	95.28
Improvement 2		√		96.58	93.59	90.60	95.06
Improvement 3			√	96.49	93.22	90.16	94.82
Improvement 4	√	√		96.22	94.83	91.43	95.53
Improvement 5	√		√	96.05	95.66	92.04	95.86
Improvement 6		√	√	96.43	94.06	90.89	95.23
FDE U-Net	√	√	√	**97.05**	**95.87**	**92.25**	**95.97**

Specifically, ACmix enhances the model's fusion of cross-channel features between multispectral bands by adaptively selecting and mixing input features. Compared to ResU-Net, Improvement 1 achieves an 11.06% increase in recall, enhancing the model's sensitivity to complex backgrounds and diverse fire characteristics. However, its precision decreases by 0.75%, likely due to the model amplifying non-fire disturbances with similar spectral features. CBAM dynamically adjusts the weights of feature maps by integrating channel and spatial information. Channel attention assigns higher weights to important spectral bands, while spatial attention highlights critical regions, enabling the model to enhance fire-related feature regions and feature channels in fire detection tasks, and suppressing the interference of extraneous information. Improvement 2 improves precision by 0.55% compared to ResU-Net. HWD effectively mitigates feature blurring caused by traditional pooling operations, preserving finer details and boundary information during the downsampling process. Meanwhile, Improvement 3 increases recall by 9% while maintaining high precision, surpassing ResU-Net. Furthermore, the progressive integration of the three modules consistently enhances IoU and F1 scores demonstrating the effectiveness of the modular optimization strategy. This modular optimization process progressively improves model accuracy, strengthens its ability to handle complex scenes and fine-grained segmentation, and ensures a better balance between precision and recall. The specific data show that, compared with the baseline model ResU-Net, precision increases by 1.02%, recall increases by 11.65%%, IoU increases by 10.94%%, and F1 score increases by 6.28%, respectively. The gradual addition of these modules positively and significantly contributes to the overall performance of the model.

Figure 11 shows the segmentation results of the different models for the fire area. We mark false positives in red and false negatives in blue. As shown in Figure 11c–f, all the methods detect the correct location of the fire, with the main errors occurring at the edges of the fire and non-fire areas contained within the large fire area. Figure 11c shows the ResU-Net baseline detection results, with false positives at the edges and non-fire areas between fires misclassified as fire. Figure 11d shows that recall increases and false

negatives decrease with ACmix, indicating its role in feature extraction. Adding only CBAM reduces both false negatives and false positives, highlighting its effectiveness in reinforcing critical information. The combination of ACmix and CBAM balances precision and recall in feature extraction. Haar wavelet downsampling improves edge segmentation. Our method reduces false positives and false negatives, as shown in the elliptical box in Figure 11f, where the non-fire area between fire regions is correctly segmented with minimal false positives. Experimental results further indicate that removing any individual module adversely affects segmentation performance, whereas the combined implementation of these modules significantly enhances segmentation effectiveness.

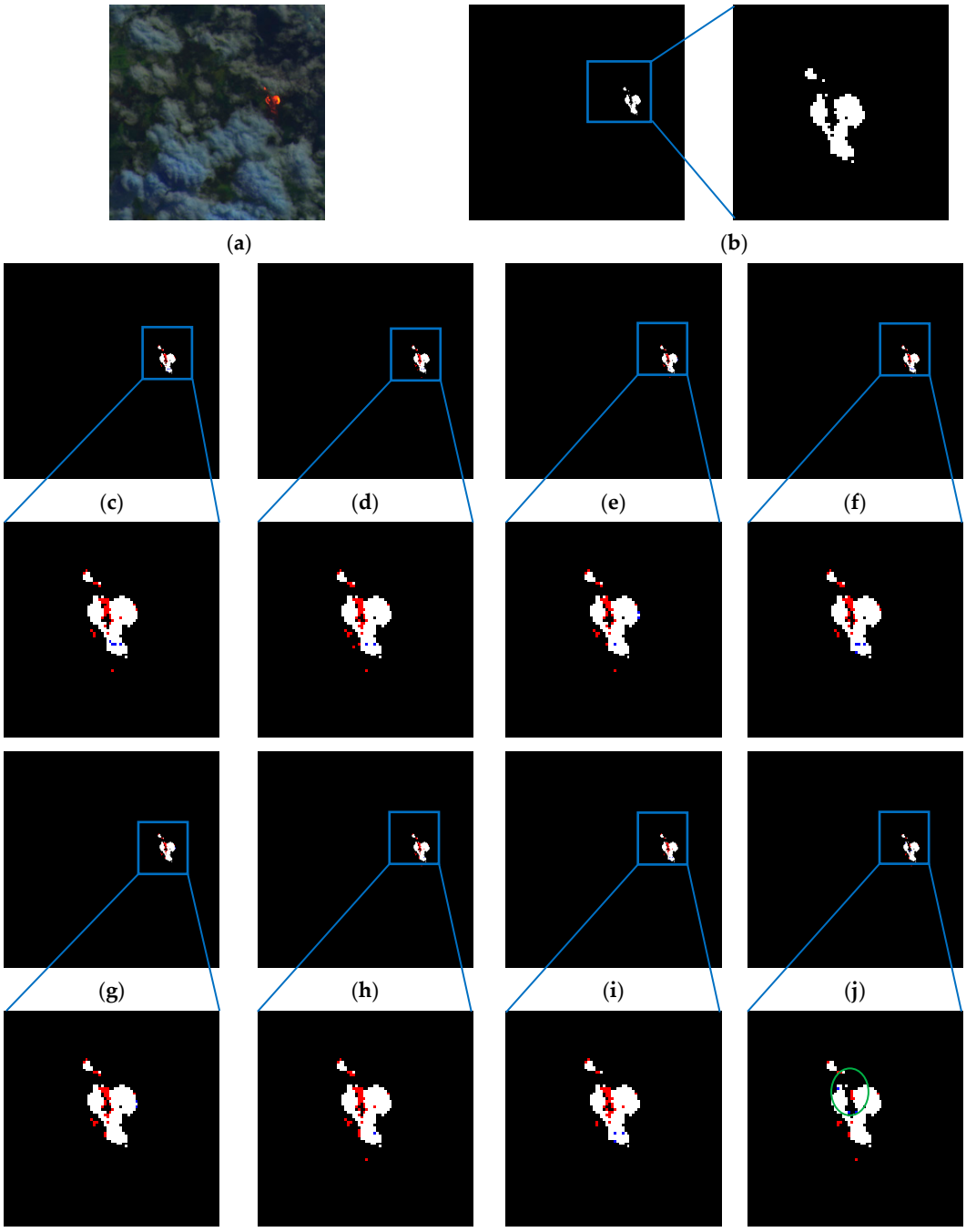


Figure 11. (**a**) Original image. (**b**) Ground truth. (**c**) ResU-Net. (**d**) Improvement 1. (**e**) Improvement 2. (**f**) Improvement 3. (**g**) Improvement 4. (**h**) Improvement 5. (**i**) Improvement 6. (**j**) FDE U-Net.

4.3. Fire Detection Results

4.3.1. Overall Fire Detection Results Analysis

In the ablation experiment described in Section 4.2, the effectiveness of FDE U-Net is demonstrated. To comprehensively evaluate its performance, we compare the proposed method with several state-of-the-art segmentation networks, including U-Net, ResU-Net, U-Net3+, SwinU-Net, and WET-UNet. Table 2 presents the fire detection performances of the different models. The proposed method has the best performance among all models, with IoU and F1 scores of 92.25% and 95.97%, respectively. Compared with the suboptimal SwinU-Net, the IoU increases by 0.39% and the F1 score improves by 0.21%. U-Net, as the basic segmentation network, is lower than the improved model in all metrics, especially in the IoU and F1 metrics, with room for improvement. ResU-Net exhibits improved recall but slightly decreased precision. ResU-Net enhances feature transfer through residual connections, improving model stability, convergence speed, and feature reuse. However, this approach can excessively smooth input features, leading to a reduction in recall. Moreover, the spectral bands of Landsat 8 exhibit varying sensitivities to fire, posing a challenge, as ResU-Net diminishes the influence of key bands such as SWIR-2, thereby weakening the detection of small or subtle fire-related features. U-Net3+ enhances the performance through more skip connections and multiscale feature fusion, with a precision of 96.49%. However, the detection of small objects or edges may not be sufficiently sensitive, resulting in low recall rates. SwinU-Net leverages global information to capture the overall fire extent, enhancing recall over a larger spatial range and detecting more fire areas. However, it may mistakenly classify pixels adjacent to fire zones as fire, leading to a decrease in precision compared to U-Net3+. Additionally, this approach involves a large number of parameters, which can increase computational complexity. By incorporating wavelet transform, WET-UNetcompresses information efficiently and enhances feature extraction, improving edge detection and reducing redundancy, which leads to better precision.

Table 2. Performance comparison of FDE U-Net and other CNN methods.

Method	P (%)	R (%)	IoU (%)	F1 (%)	Para (M)	FLOPS	Inference Time	FPS
U-Net [31]	94.91	84.43	80.71	89.32	1.95	5.44 G	17.33 ms	57.69
ResU-Net [46]	96.03	84.22	81.31	89.69	2.03	4.23 G	19.81ms	50.45
U-Net3+ [47]	96.49	90.16	87.30	93.22	2.31	2.02 G	18.05 ms	55.39
SwinU-Net [48]	95.85	95.66	91.86	95.76	8.73	4.42 G	24.30 ms	41.15
WET-UNet [49]	96.99	93.38	90.76	95.15	6.32	3.16 G	26.10 ms	38.31
FDE U-Net	**97.05**	**95.87**	**92.25**	**95.97**	2.25	4.16 G	20.13 ms	49.7

FDE U-Net offers stronger feature expression and processing. The short-wave infrared band of Landsat 8 is sensitive to fires, and Haar wavelet decomposition preserves high-temperature features during downsampling, enhancing the extraction of smoke edges and fire contours. The synergy between ACmix and CBAM captures the global dependency between multispectral bands, highlights the key bands through channel weight adjustment, and focuses the fire diffusion direction in the spatial dimension. The integration of global and local information enables the model to cope with complex segmentation tasks more effectively, and improves the segmentation precision and detail restoration ability in the case of category imbalance. The parameter amount of FDE U-Net is 2.25M. Compared to SwinU-Net and WET-UNet, the number of parameters is significantly reduced, which gives the proposed method a higher practical value in scenarios with limited computing resources. We further compare inference time, FLOPs (Floating-point Operations), and FPS (frames per second) of different models. FDE U-Net, based on the comprehensive

enhancement of multiple modules, still effectively controls the model's computational workload and inference time. FLOPs are reduced by about 23.53% compared to traditional U-Net, significantly reducing the computational burden of the model. Under the premise of integrating multiple advanced modules, FDE U-Net's inference time is still controlled at around 20 ms, which is faster than SwinU-Net (24.30 ms) and WET-UNet (26.10 ms), indicating that its structural design has good inference efficiency while being compatible with high expressiveness.

Regarding the fire detection task, the main challenges are twofold: (1) the fire area accounts for a small proportion compared to the background, and there is a serious category imbalance problem. (2) Fire regions often have irregular shapes and boundaries, which pose significant challenges for accurate segmentation. We select three representative scenarios, images 1 to 4, which contain a centralized target distribution scenario and a decentralized distribution scenario, as shown in Figure 12. Figures 13–16 show the fire segmentation results of images 1 to 4 using several advanced models. For ease of observation, false positives are marked in red and false negatives are marked in blue. Table 3 presents the statistics for the correctly and incorrectly detected pixels.

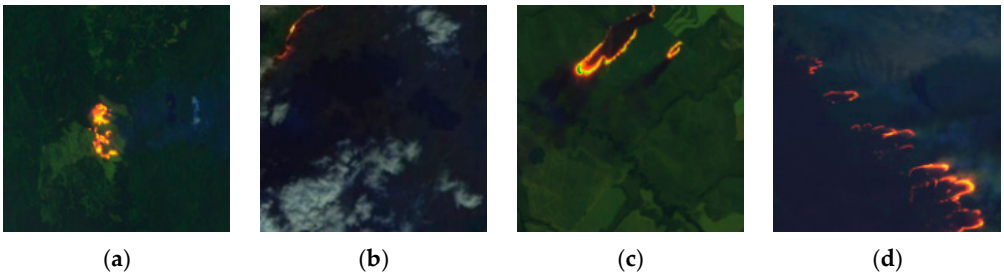


Figure 12. (**a**) Image 1, (**b**) image 2, (**c**) image 3, (**d**) image 4.

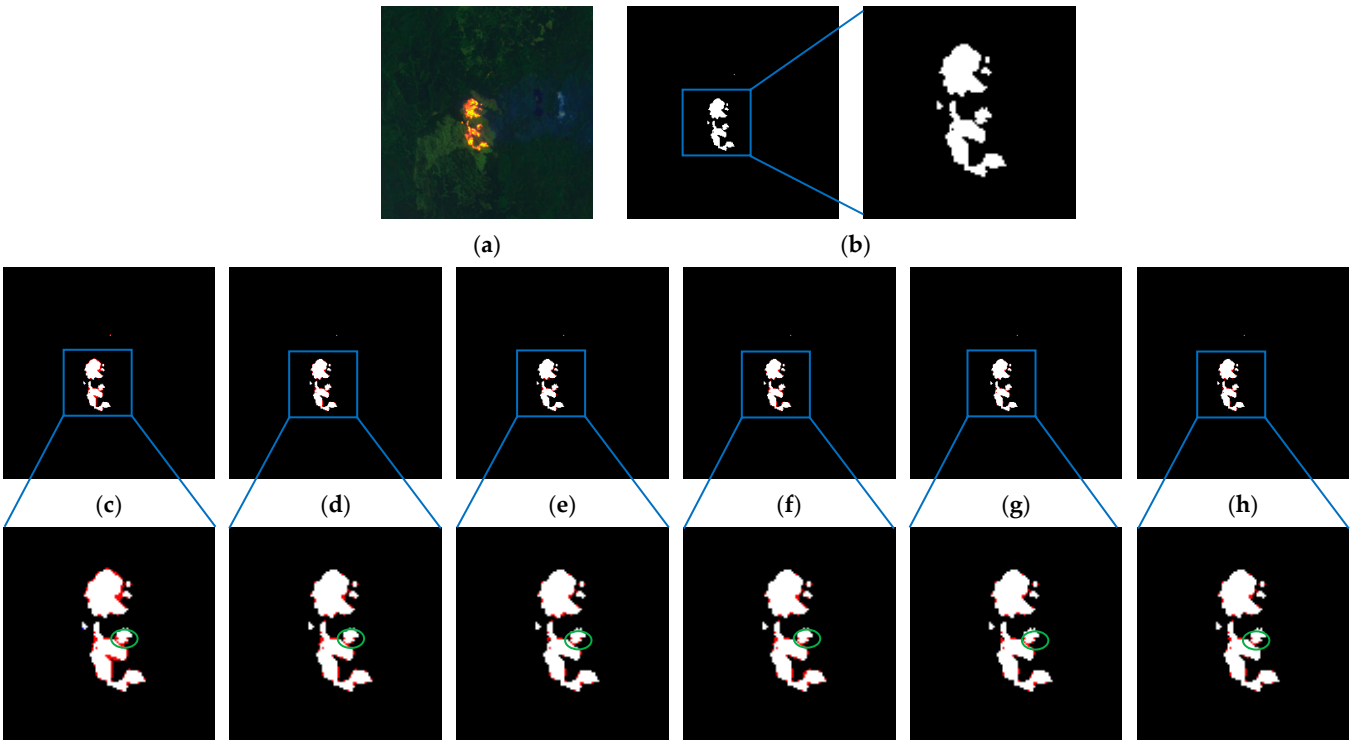


Figure 13. Segmentation results of image 1 by different CNN methods: (**a**) original image, (**b**) ground truth, (**c**) U-Net, (**d**) ResU-Net, (**e**) U-Net3+, (**f**) SwinU-Net, (**g**) WET-UNet, (**h**) FDE U-Net.

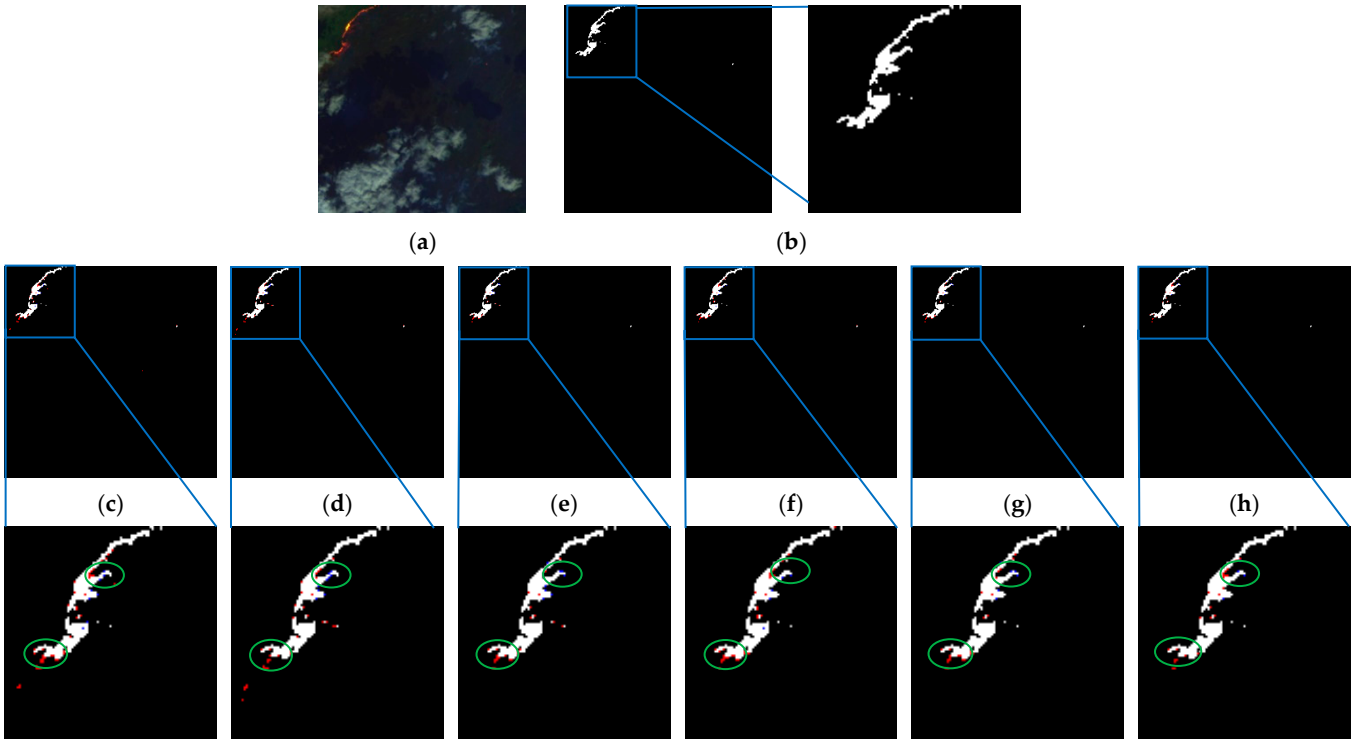


Figure 14. Segmentation results of image 2 by different CNN methods: (**a**) original image, (**b**) ground truth, (**c**) U-Net, (**d**) ResU-Net, (**e**) U-Net3+, (**f**) SwinU-Net, (**g**) WET-UNet, (**h**) FDE U-Net.

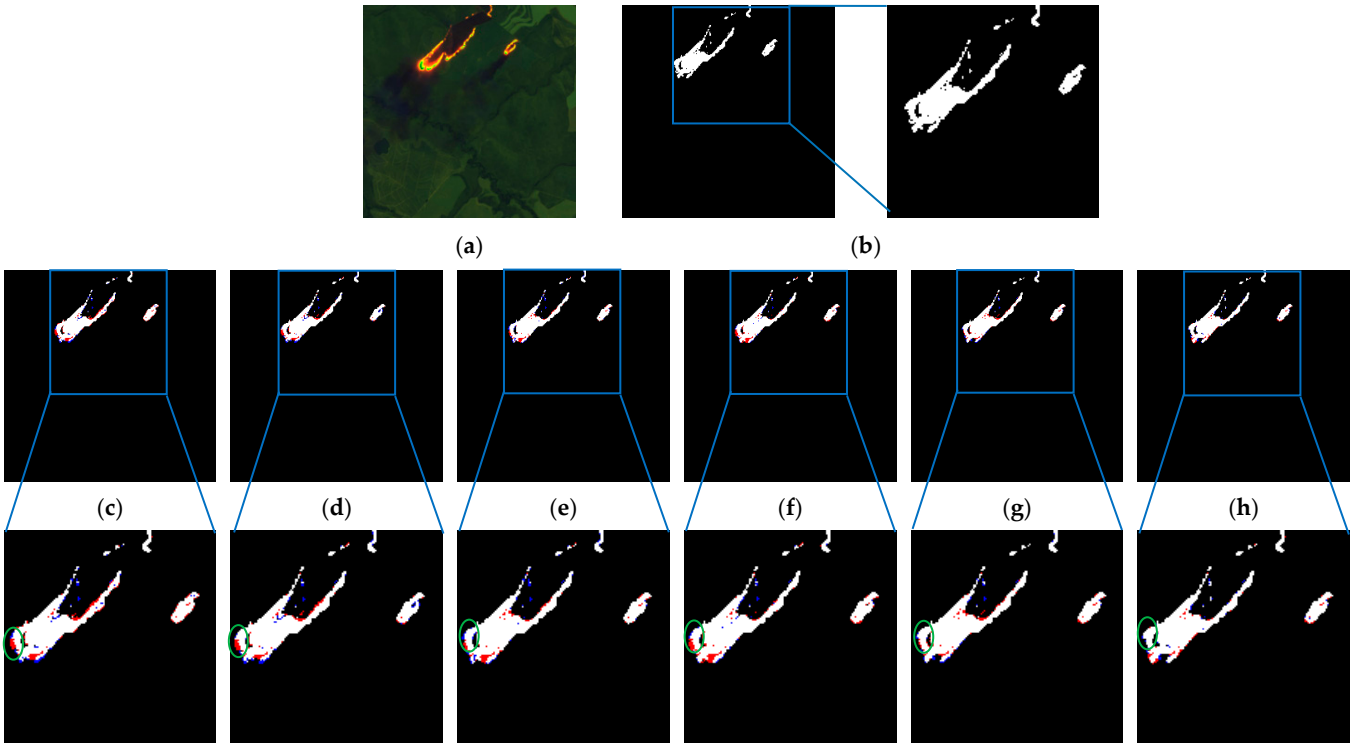


Figure 15. Segmentation results of image 3 by different CNN methods: (**a**) original image, (**b**) ground truth, (**c**) U-Net, (**d**) ResU-Net, (**e**) U-Net3+, (**f**) SwinU-Net, (**g**) WET-UNet, (**h**) FDE U-Net.

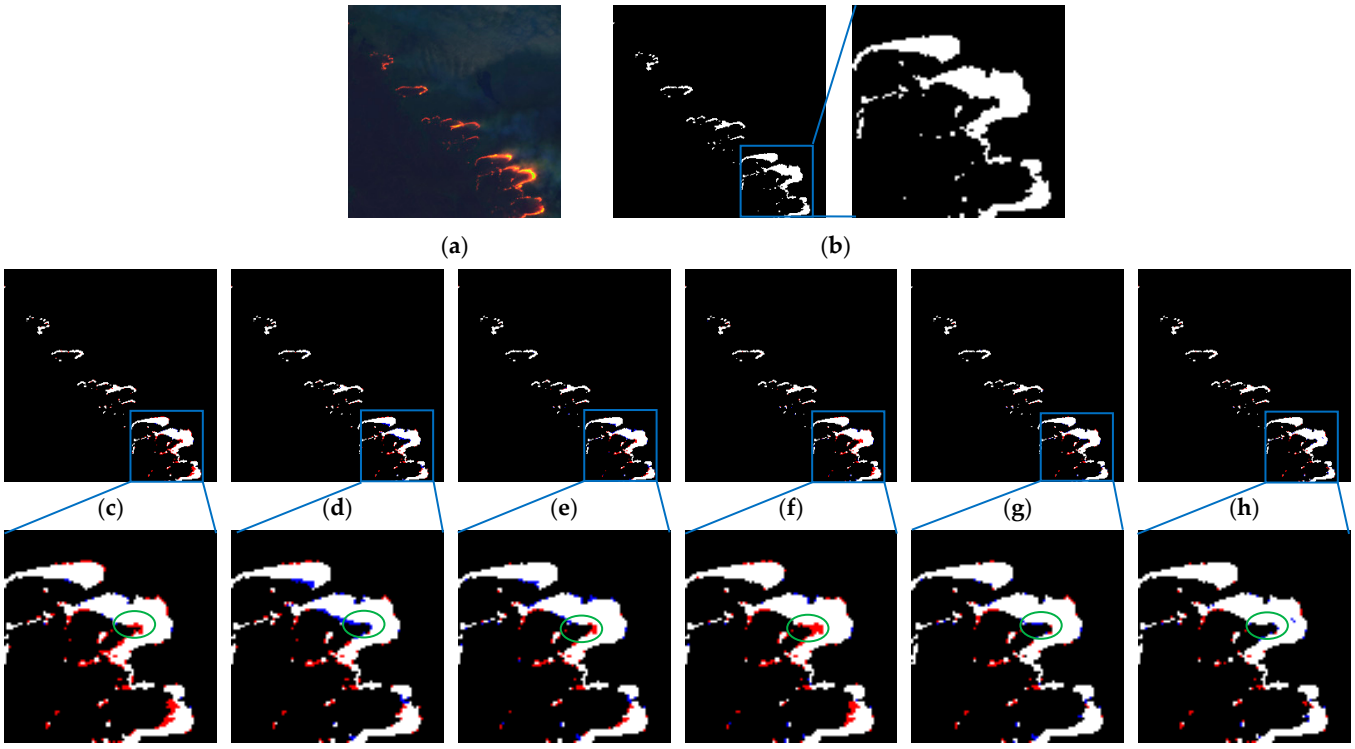


Figure 16. Segmentation results of image 4 by different CNN methods: (**a**) original image, **(b)** ground truth, (**c**) U-Net, (**d**) ResU-Net, (**e**) U-Net3+, (**f**) SwinU-Net, (**g**) WET-UNet, (**h**) FDE U-Net.

Table 3. Detection results of FDE U-Net and other CNN methods.

Method	Image 1				Image 2				Image 3				Image 4			
	GT	TP	FN	FP	GT	TP	FN	FP	GT	TP	FN	FP	GT	TP	FN	FP
U-Net	781	780	1	124	410	402	8	47	1347	1269	78	150	1711	**1693**	**18**	250
ResU-Net		781	0	74		401	9	36		1273	74	111		1632	79	146
U-Net3+		781	0	56		402	8	25		1281	66	95		1640	71	128
SwinU-Net		781	0	78		405	5	38		1300	47	121		1678	33	186
WET-UNet		781	0	53		404	6	25		1296	51	89		1664	47	106
FDE U-Net		**781**	**0**	**52**		**406**	**4**	**25**		**1306**	**41**	**68**		1683	28	**102**

As shown in Figures 13 and 14, the fire areas in the two scenarios are more concentrated, and most of the models are able to accurately find the exact location of the fire, the significant difference in the outputs is that the edges of the fire areas are not clearly segmented, such as in the elliptical box area in the figure, with more false positives as well as a small number of missed detections. The proposed method overcomes this problem by enhancing feature extraction, feature fusion, and detail processing. As shown in Figure 14c–g, this scene shows that there are many clouds in the image, and the fire area is relatively small. The high brightness of clouds easily confuses the fire area with the background area. The classic segmentation models U-Net, ResU-Net, U-Net3+, SwinU-Net, and WET-UNet generate a significant number of false positives, which are concentrated in the elliptical region in the lower-left corner of the fire area. As shown in Figure 14h, the numbers of false positives and false negatives are 25 and 4, respectively, which are the lowest among the compared methods. The area within the elliptical shows that the FDE U-Net performs the best boundary preservation for fire. Thus, the proposed method extracts rich, high-level semantic features and preserves the main information, which improves the network's detection ability in the case of a category imbalance. Irrelevant information and noise

are suppressed so that the model can effectively distinguish between fire boundaries and background pixels in cloudy scenes.

In Figures 15 and 16, the fire regions are irregular and scattered. In scene image 3, the segmentation results differ mainly around the edge of the fire region, within the elliptical box. This may be due to high-temperature fire pixels affecting the spectral reflectivity of nearby pixels through thermal radiation, causing misclassification. U-Net and ResU-Net have the highest false negatives, 78 and 74, respectively, with U-Net having more false positives. While U-Net3+ improves feature extraction through multi-scale fusion, reducing false positives to 95 and false negatives to 66, its contextual information fusion may be insufficient in complex or boundary-ambiguous scenarios. SwinU-Net combines the strengths of Swin Transformer to capture global information and long-range dependencies, making it particularly effective for detecting large fire regions. This leads to an improvement in recall and a reduction in false negatives to 47. However, it may still misidentify pixels in adjacent fire regions as part of the fire area, especially in edge regions where the boundaries are less clear, resulting in more false positives. WET-UNet takes advantage of wavelet transform in multi-scale analysis and combines it with the modeling capabilities of the transformer to achieve comprehensive extraction and refined processing of image features, optimizing the balance between precision and recall. The fire areas in the image 4 scene of Figure 16 are of different sizes and unevenly distributed, and there are separate fires detached from the main fire area that can be accurately detected by all models. The main difference lies in the densely distributed fire areas, such as the region within the elliptical box in Figure 16c–g, where U-Net, ResU-Net, U-Net3+, SwinU-Net, and WET-UNet all have varying negatives and false positives. In contrast, our method optimizes the capture and fusion of contextual information in the structural design and generates results closest to the ground truth. The results in Figures 15h and 16h show that both false positives and false negatives are minimal. It shows that FDE U-Net extracts and integrates information-rich high-level features from the input. By fusing different bands, the model's focus area is adjusted according to the response characteristics of different bands. The feature fusion and embedded attention mechanism enhance the model's ability to capture the fire area. Through wavelet transform, the low-frequency and high-frequency information extracted between multiple bands helps enhance the correlation between features in different bands, allowing the model to better utilize information from different bands for fire identification, pay more attention to the fire body and edge details, and, ultimately, achieve accurate segmentation of the fire area.

To further validate the generalization ability of our proposed model, we conduct additional experiments on the FLAME dataset with 2003 images, which is collected by drones during planned burns in the pine forests of Arizona. Figure 17 presents representative samples from the dataset, highlighting its key visual characteristics. The images in this dataset have higher spatial resolution and differ significantly from the Landsat-8 remote sensing images used in the main experiment in terms of data source and observation scale. The detection results on this dataset are shown in Table 4.

Table 4. Detection results on FLAME dataset.

Method	P (%)	R (%)	IoU (%)	F1 (%)
U-Net	89.03	84.94	76.89	86.94
ResU-Net	91.72	82.73	76.97	86.99
U-Net3+	90.07	87.10	79.46	88.56
SwinU-Net	92.61	88.84	83.04	90.72
WET-UNet	93.45	87.56	82.58	90.45
FDE U-Net	**93.82**	**89.73**	**84.69**	**91.71**



Figure 17. Examples from the FLAME dataset.

In the experiment on the FLAME dataset, the images are captured by close-range drones, which provide higher spatial resolution and more complex background texture information, such as ground vegetation, smoke occlusion, and buildings. These factors make it more difficult for the model to accurately extract pure fire area features, thus affecting detection accuracy. Table 4 shows the detection results on the FLAME dataset. As a result, the overall evaluation metrics are slightly lower than those on the Landsat-8 dataset. Despite this, the proposed FDE U-Net model still achieves the best detection performance on this dataset, demonstrating strong robustness and generalization ability. Specifically, the Precision, Recall, IoU, and F1 scores of FDE U-Net reach 93.82%, 89.73%, 84.69%, and 91.71%, respectively. Compared with other models, FDE U-Net shows the smallest performance drop. This observation indicates that the HWD, CBAM, and ACMix modules introduced into the model have good adaptability to changes in fire point distribution and target size.

4.3.2. Analysis of Small Targets and Cloud Cover Results

In order to further verify the role of the model in small target detection, we construct a dataset containing only small flame areas. We select 100 images with only small fires and apply horizontal and vertical flipping to generate a total of 300 small-fire images, several examples of which are shown in Figure 18. The experimental results show that the overall performance of the model decreases compared to the large-scale dataset. However, despite the overall more difficult task, FDE U-Net still performs better than other comparison models in this scenario, with an IoU of 87.89% and an F1 score of 93.49%, as shown in Table 5, maintaining its leading performance in small target recognition. This result indirectly proves that the edge detail preservation mechanism provided by the HWD module plays a key role in small-scale flame detection.

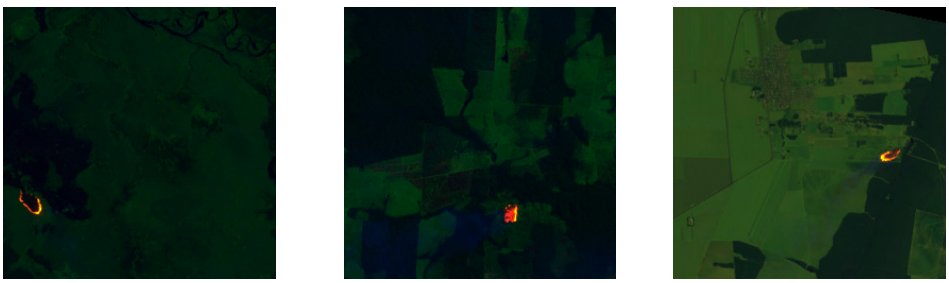


Figure 18. Small target scene examples.

Table 5. Detection results on small objects.

Method	P (%)	R (%)	IoU (%)	F1 (%)
U-Net	92.76	85.31	80.36	88.88
ResU-Net	93.18	84.07	79.55	88.39
U-Net3+	91.95	91.45	84.85	91.70
SwinU-Net	94.19	91.32	86.58	92.71
WET-UNet	94.35	90.66	86.36	92.62
FDE U-Net	**95.33**	**92.36**	**87.89**	**93.49**

Cloud cover also has an interfering effect on remote sensing fire detection tasks. Clouds block the spectral and thermal radiation characteristics of the fire source area, while introducing a large amount of invalid background information and destroying the edge clarity of the image, thus reducing the model's discrimination ability. To evaluate the robustness of each model under complex weather interference, we construct a dataset containing cloud-occluded scenes by increasing the proportion of cloud-containing images to 70%, several examples of which are shown in Figure 19. Table 6 shows the detection results on this dataset.

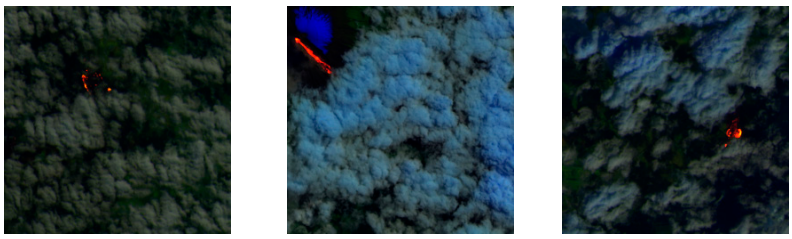


Figure 19. Cloud cover scene examples.

Table 6. Detection results with cloud cover.

Method	P (%)	R (%)	IoU (%)	F1 (%)
U-Net	93.52	84.21	79.89	88.64
ResU-Net	94.85	83.66	80.12	88.91
U-Net3+	95.28	86.17	82.87	90.50
SwinU-Net	96.21	91.31	88.00	93.59
WET-UNet	95.89	91.54	88.19	93.72
FDE U-Net	**96.41**	**92.06**	**88.98**	**94.18**

The experimental results show that all indicators of all models decrease to varying degrees under cloud interference. Despite this, FDE U-Net still achieves the best performance in precision, recall, IoU, and F1 score, demonstrating strong anti-interference capability. For example, the traditional U-Net model only reaches an intersection over union of 79.89, whereas FDE U-Net reaches 88.98, as shown in Table 6, representing an improvement of nearly nine percentage points. This indicates that its structure is more effective in maintaining edge information and suppressing background interference. This suggests that the model not only accurately detects fire regions but also effectively reduces the false positives and false negatives caused by clouds. These results show that FDE U-Net exhibits stronger robustness and generalization ability in complex environments such as cloud cover.

5. Discussion

In this study, we systematically evaluate the performance of the proposed FDE U-Net network for fire detection in remote sensing images. The experiments include both

quantitative and qualitative evaluation dimensions. In the quantitative evaluation, the performance of the model on fire detection is evaluated by a set of standardized metrics: Precision, Recall, IoU, and F1 scores. Through comparative analysis and ablation studies, we further confirm the effectiveness of the model, and compared with existing advanced methods, our proposed method performs better in all scenarios. To evaluate the efficiency of the proposed model, we compare the number of parameters, inference time, and FLOPS. It achieves faster inference time with a moderate number of parameters, which shows that the model achieves a good balance between performance and efficiency. In the qualitative evaluation, by showing the detection results of the model in different scenarios, it can be seen that the proposed model can accurately locate the fire area with minimal false positives and false negatives. In summary, FDE U-Net has excellent performance in fire detection tasks.

Although the overall performance of the model has been improved, there are still some limitations. Firstly, the innovation of FDE U-Net is mainly reflected in the combination and integration of existing modules. In the future, it is possible to consider building modules and feature extraction methods that can be used for fire detection based on the characteristics of remote sensing fire images. Secondly, relying on the traditional U-Net framework, it is still slightly insufficient in multi-scale modeling and cross-resolution feature fusion. Compared with emerging architectures based on transformer or multi-branch structures, the model may have information loss in processing complex fire forms. In the future, we will improve the innovation and practicality of the model from the following directions: firstly, it is necessary to explore new network backbone structures, such as transformer or Mamba, as well as dedicated extraction modules for fire features to enhance global perception capabilities; secondly, we can integrate multi-source remote sensing and ground data (such as MODIS, Sentinel, LiDAR, ground monitoring) to improve scene understanding capabilities; thirdly, we will develop lightweight models and semi-supervised learning methods to adapt to the needs of data scarcity and real-time detection.

6. Conclusions

We propose a novel FDE U-Net model for fire detection using Landsat-8 multispectral remote sensing data. The ACmix, CBAM, and HWD modules are integrated into each encoder layer in ResU-Net to extract the information-rich high-level features. ACmix combines convolution and self-attention mechanisms, where convolution effectively captures local features and self-attention dynamically captures global contextual information. By integrating the two, the model can handle local details while focusing on global features, enhancing its ability to understand complex images. CBAM uses channel and spatial attention mechanisms to focus on key areas in fire images by applying weights to features. This improves the model's ability to recognize essential regions and enhances the detection accuracy. The Haar wavelet downsampling technique can effectively capture detailed features, avoid the loss of important information during the downsampling process, and retain more detail and boundary information. The experimental results indicate that the proposed FDE U-Net performs exceptionally well in fire detection tasks, accurately extracting fire regions, even when the fire area is small and the background is complex. For the terrain interference, although the terrain will affect the distribution of background information of local images, such as vegetation occlusion and mountain shadows, the microscopic interference of the terrain will not significantly affect the model's recognition results of the fire area due to the spatial resolution of Landsat-8 of 30m. Therefore, in this study, terrain interference is not the main influencing factor, but we recognize that it may have an impact on the detection process, which can be further studied in the future.

Author Contributions: R.Z.: conceptualization, investigation, methodology, validation, visualization, formal analysis, writing (original draft). Z.X.: conceptualization, funding acquisition, investigation, resources, supervision, formal analysis, article revision. G.L. and P.H.: project administration, supervision, resources, formal analysis. R.W. and Y.Q.: conceptualization, methodology, validation, formal analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China under Grant 62361058 and 62171272, Yunnan Expert Workstation under Grant 202305AF150012 (Corresponding author: Zhihui Xin).

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

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

References

1. Adams, M.A.; Shen, Z. Introduction to the characteristics, impacts and management of forest fire in China. *For. Ecol. Manag.* **2015**, *356*, 1. [CrossRef]
2. Ramos, L.; Casas, E.; Bendek, E.; Romero, C.; Rivas-Echeverría, F. Computer vision for wildfire detection: A critical brief review. *Multimed. Tools Appl.* **2024**, *83*, 83427–83470. [CrossRef]
3. Global Wildfire Information System (2025)—With Minor Processing by Our World in Data. Available online: https://ourworldindata.org/grapher/annual-area-burnt-by-wildfires?tab=maps (accessed on 22 June 2025).
4. Abdusalomov, A.; Umirzakova, S.; Bakhtiyor Shukhratovich, M.; Mukhiddinov, M.; Kakhorov, A.; Buriboev, A.; Jeon, H.S. Drone-Based Wildfire Detection with Multi-Sensor Integration. *Remote Sens.* **2024**, *16*, 4651. [CrossRef]
5. Fraser, R.H.; Van der Sluijs, J.; Hall, R.J. Calibrating Satellite-Based Indices of Burn Severity from UAV-Derived Metrics of a Burned Boreal Forest in NWT, Canada. *Remote Sens.* **2017**, *9*, 279. [CrossRef]
6. Tong, H.; Yuan, J.; Zhang, J.; Wang, H.; Li, T. Real-Time Wildfire Monitoring Using Low-Altitude Remote Sensing Imagery. *Remote Sens.* **2024**, *16*, 2827. [CrossRef]
7. Thangavel, K.; Spiller, D.; Sabatini, R.; Amici, S.; Sasidharan, S.T.; Fayek, H.; Marzocca, P. Autonomous satellite wildfire detection using hyperspectral imagery and neural networks: A case study on Australian wildfire. *Remote Sens.* **2023**, *15*, 720. [CrossRef]
8. Sun, Y.; Jiang, L.; Pan, J.; Sheng, S.; Hao, L. A satellite imagery smoke detection framework based on the Mahalanobis distance for early fire identification and positioning. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *118*, 103257. [CrossRef]
9. 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]
10. Pourshakouri, F.; Darvishsefat, A.A.; Samadzadegan, F.; Attarod, P.; Amini, S. An improved algorithm for small and low-intensity fire detection in the temperate deciduous forests using MODIS data: A preliminary study in the Caspian Forests of Northern Iran. *Nat. Hazards* **2023**, *116*, 2529–2547. [CrossRef]
11. Cardíl, A.; Tapia, V.M.; Monedero, S.; Quiñones, T.; Little, K.; Stoof, C.R.; Ramirez, J.; de Miguel, S. Characterizing the rate of spread of large wildfires in emerging fire environments of northwestern Europe using Visible Infrared Imaging Radiometer Suite active fire data. *Nat. Hazards Earth Syst. Sci.* **2023**, *23*, 361–373. [CrossRef]
12. Zhang, T.; Wooster, M.J.; Xu, W. Approaches for synergistically exploiting VIIRS I- and M-Band data in regional active fire detection and FRP assessment: A demonstration with respect to agricultural residue burning in Eastern China. *Remote Sens. Environ.* **2017**, *198*, 407–424. [CrossRef]
13. 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]
14. Thompson, D.K.; Morrison, K. A classification scheme to determine wildfires from the satellite record in the cool grasslands of southern Canada: Considerations for fire occurrence modelling and warning criteria. *Nat. Hazards Earth Syst. Sci.* **2020**, *20*, 3439–3454. [CrossRef]
15. Ghali, R.; Jmal, M.; Mseddi, W.S.; Attia, R. Recent advances in fire detection and monitoring systems: A review. In Proceedings of the 10th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, Hammamet, Tunisia, 20–22 December 2020. [CrossRef]
16. Zhang, Z.; Shen, T.; Zou, J. An improved probabilistic approach for fire detection in videos. *Fire Technol.* **2014**, *50*, 745–752. [CrossRef]
17. Celik, T.; Demirel, H. Fire detection in video sequences using a generic color model. *Fire Saf. J.* **2009**, *44*, 147–158. [CrossRef]
18. Chen, X.; An, Q.; Yu, K.; Ban, Y. A novel fire identification algorithm based on improved color segmentation and enhanced feature data. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 3075380. [CrossRef]

19. Jang, E.; Kang, Y.; Im, J.; Lee, D.-W.; Yoon, J.; Kim, S.-K. Detection and monitoring of forest fires using Himawari-8 geostationary satellite data in South Korea. *Remote Sens.* **2019**, *11*, 271. [CrossRef]
20. Tennant, E.; Jenkins, S.F.; Miller, V.; Robertson, R.; Wen, B.; Yun, S.-H.; Taisne, B. Automating tephra fall building damage assessment using deep learning. *Nat. Hazards Earth Syst. Sci.* **2024**, *24*, 4585–4608. [CrossRef]
21. Zhu, J.; Zhang, H.; Li, S.; Wang, S.; Ma, H. Cross teaching-enhanced multispectral remote sensing object detection with transformer. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 2401–2413. [CrossRef]
22. Zhang, S.; Zhang, S.; Guo, T.; Xu, R.; Liu, Z.; Du, Q. Clutter Modeling and Characteristics Analysis for GEO Spaceborne-Airborne Bistatic Radar. *Remote Sens.* **2025**, *17*, 1222. [CrossRef]
23. Yu, L.; Qin, H.; Huang, S.; Wei, W.; Jiang, H.; Mu, L. Quantitative study of storm surge risk assessment in an undeveloped coastal area of China based on deep learning and geographic information system techniques: A case study of Double Moon Bay. *Nat. Hazards Earth Syst. Sci.* **2024**, *24*, 2003–2024. [CrossRef]
24. Hwang, G.; Jeong, J.; Lee, S.J. SFA-Net: Semantic Feature Adjustment Network for Remote Sensing Image Segmentation. *Remote Sens.* **2024**, *16*, 3278. [CrossRef]
25. Kukuk, S.B.; Kilimci, Z.H. Comprehensive analysis of forest fire detection using deep learning models and conventional machine learning algorithms. *Int. J. Comput. Exp. Sci. Eng.* **2021**, *7*, 84–94. [CrossRef]
26. Jin, C.; Wang, T.; Alhusaini, N.; Zhao, S.; Liu, H.; Xu, K.; Zhang, J. Video fire detection methods based on deep learning: Datasets, methods, and future directions. *Fire* **2023**, *6*, 315. [CrossRef]
27. Sun, H.R.; Shi, B.J.; Zhou, Y.T.; Chen, J.H.; Hu, Y.L. A smoke detection algorithm based on improved YOLO v7 lightweight model for UAV optical sensors. *IEEE Sens. J.* **2024**, *24*, 26136–26147. [CrossRef]
28. Hu, X.; Liu, W.; Wen, H.; Yuen, K.-V.; Jin, T.; Junior, A.C.N.; Zhong, P. AF-Net: An Active Fire Detection Model Using Improved Object-Contextual Representations on Unbalanced UAV Datasets. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 13558–13569. [CrossRef]
29. Seydi, S.T.; Saeidi, V.; Kalantar, B.; Ueda, N.; Halin, A.A. Fire-Net: A deep learning framework for active forest fire detection. *J. Sens.* **2022**, *2022*, 8044390. [CrossRef]
30. Han, Y.; Zheng, C.; Liu, X.; Tian, Y.; Dong, Z. Burned Area and Burn Severity Mapping with a Transformer-Based Change Detection Model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 13866–13880. [CrossRef]
31. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015. [CrossRef]
32. Siddique, N.; Paheding, S.; Elkin, C.P.; Devabhaktuni, V. U-Net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access* **2021**, *9*, 82031–82057. [CrossRef]
33. Zhang, P.; Ban, Y.; Nascetti, A. Learning U-Net without forgetting for near real-time wildfire monitoring by the fusion of SAR and optical time series. *Remote Sens. Environ.* **2021**, *261*, 112467. [CrossRef]
34. Xu, Z.; Li, H.; Long, B. MSU-Net: Multi-scale self-attention semantic segmentation method for oil-tea camellia planting area extraction in hilly areas of southern China. *Expert Syst. Appl.* **2025**, *263*, 125779. [CrossRef]
35. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015. [CrossRef]
36. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
37. Wang, Z.; Yang, P.; Liang, H.; Zheng, C.; Yin, J.; Tian, Y.; Cui, W. Semantic segmentation and analysis on sensitive parameters of forest fire smoke using smoke-unet and landsat-8 imagery. *Remote Sens.* **2021**, *14*, 45. [CrossRef]
38. Afsar, R.; Sultana, A.; Abouzahra, S.N.; Aspiras, T.; Asari, V.K. Using ResWnet for semantic segmentation of active wildfires from Landsat-8 imagery. In Proceedings of the SPIE—The International Society for Optical Engineering, National Harbor, MD, USA, 19–24 August 2024. [CrossRef]
39. Schroeder, W.; Oliva, P.; Giglio, L.; Quayle, B.; Lorenz, E.; Morelli, F. Active fire detection using Landsat-8/OLI data. *Remote Sens. Environ.* **2016**, *185*, 210–220. [CrossRef]
40. Kumar, S.S.; Roy, D.P. Global operational land imager Landsat-8 reflectance-based active fire detection algorithm. *Int. J. Digit. Earth* **2018**, *11*, 154–178. [CrossRef]
41. Murphy, S.W.; de Souza Filho, C.R.; Wright, R.; Sabatino, G.; Pabon, R.C. HOTMAP: Global hot target detection at moderate spatial resolution. *Remote Sens. Environ.* **2016**, *177*, 78–88. [CrossRef]
42. de Almeida Pereira, G.H.; Fusioka, A.M.; Nassu, B.T. Active fire detection in Landsat-8 imagery: A large-scale dataset and a deep-learning study. *ISPRS J. Photogramm. Remote Sens.* **2021**, *178*, 171–186. [CrossRef]
43. Pan, X.; Ge, C.; Lu, R.; Song, S.; Chen, G.; Huang, Z.; Huang, G. On the Integration of Self-Attention and Convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 805–815. [CrossRef]

44. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018. [CrossRef]
45. Xu, G.; Liao, W.; Zhang, X.; Li, C.; He, X.; Wu, X. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern Recognit.* **2023**, *143*, 109819. [CrossRef]
46. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
47. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020. [CrossRef]
48. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022. [CrossRef]
49. Zeng, Y.; Li, J.; Zhao, Z.; Liang, W.; Zeng, P.; Shen, S.; Zhang, K.; Shen, C. WET-UNet: Wavelet integrated efficient transformer networks for nasopharyngeal carcinoma tumor segmentation. *Sci. Prog.* **2024**, *107*, 00368504241232537. [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.