remote sensing

MDPI

Article

HybriDet: A Hybrid Neural Network Combining CNN and Transformer for Wildfire Detection in Remote Sensing Imagery

Fengming Dong [1] and Ming Wang [2,3,*]

[1] School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai 264209, China; 202100800012@mail.sdu.edu.cn
[2] Inspur Cloud Information Technology Co., Ltd., Jinan 250101, China
[3] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
* Correspondence: oyamingo@whu.edu.cn

Highlights

What are the main findings?

- A novel hybrid neural network architecture named HybriDet is proposed, which effectively integrates the local feature extraction capability of CNNs and the global contextual modeling strength of Transformers. The innovative SwinBottle module and Coordinate-Spatial (CS) dual attention mechanism significantly improve the detection accuracy for wildfires and smoke in complex remote sensing imagery.
- A superior balance between accuracy and efficiency is achieved. The lightweight model after structured pruning contains only 6.45 M parameters. It significantly outperforms state-of-the-art models like YOLOv8 by 6.4% in mAP50 on the FASDD-RS dataset while maintaining real-time inference speed suitable for edge device deployment.

What are the implications of the main findings?

- Provides an efficient and reliable fire detection solution for resource-constrained edge computing environments (e.g., satellites, UAVs). Model compression and optimization techniques enable the practical deployment of high-performance deep learning models on low-power devices, directly contributing to early wildfire warning and emergency response.
- The proposed method demonstrates strong generalization capabilities and broad application prospects. Its superior performance across multiple public datasets (FASDD-UAV, FASDD-RS, VOC) indicates its effectiveness in handling highly heterogeneous remote sensing imagery, providing crucial technical support for intelligent remote sensing monitoring in ecological conservation and socioeconomic security.

Abstract

Early warning systems on edge devices such as satellites and unmanned aerial vehicles (UAVs) are essential for effective forest fire prevention. Edge Intelligence (EI) enables deploying deep learning models on edge devices; however, traditional convolutional neural networks (CNNs)/Transformer-based models struggle to balance local-global context integration and computational efficiency in such constrained environments. To address these challenges, this paper proposes HybriDet, a novel hybrid-architecture neural network for wildfire detection. This architecture integrates the strengths of both CNNs and Transformers to effectively capture both local and global contextual information. Furthermore, we introduce efficient attention mechanisms—Windowed Attention and Coordinate-Spatial (CS) Attention—to simultaneously enhance channel-wise and spatial-wise features in high-resolution imagery, enabling long-range dependency modeling and discriminative feature

check for updates

Academic Editor: Carmen Quintano

Received: 27 August 2025
Revised: 14 October 2025
Accepted: 17 October 2025
Published: 21 October 2025

Citation: Dong, F.; Wang, M. HybriDet: A Hybrid Neural Network Combining CNN and Transformer for Wildfire Detection in Remote Sensing Imagery. *Remote Sens.* **2025**, *17*, 3497. https://doi.org/10.3390/rs17203497

extraction. Additionally, to optimize deployment efficiency, we also apply model pruning techniques to improve generalization performance and inference speed. Extensive experimental evaluations demonstrate that HybriDet achieves superior feature extraction capabilities while maintaining high computational efficiency. The optimized lightweight variant of HybriDet has a compact model size of merely 6.45 M parameters, facilitating seamless deployment on resource-constrained edge devices. Comparative evaluations on the FASDD-UAV, FASDD-RS, and VOC datasets demonstrate that HybriDet achieves superior performance over state-of-the-art models, particularly in processing highly heterogeneous remote sensing (RS) imagery. When benchmarked against YOLOv8, HybriDet demonstrates a 6.4% enhancement in mAP50 on the FASDD-RS dataset while maintaining comparable computational complexity. Meanwhile, on the VOC dataset and the FASDD-UAV dataset, our model improved by 3.6% and 0.2%, respectively, compared to the baseline model YOLOv8. These advancements highlight HybriDet's theoretical significance as a novel hybrid EI framework for wildfire detection, with practical implications for disaster emergency response, socioeconomic security, and ecological conservation.

Keywords: hybrid neural network; swin transformer; dual attention; remote sensing; wildfire detection

1. Introduction

Against the backdrop of global warming, wildfires have become a grave global safety hazard, posing significant risks to human lives, property, and ecosystems. Traditional wildfire detection methods, such as manual observation and smoke-sensor-based technologies, are often limited by inefficiency, restricted coverage, and high false-alarm rates in complex environments [1,2]. Driven by advances in remote sensing, deep learning, and the public availability of large-scale fire datasets, vision-based fire detection has emerged as a solution capable of non-contact, large-scale, and real-time monitoring [3,4].

In recent years, deep learning-based object detection algorithms have been extensively applied to wildfire detection. These methods can be categorized into convolutional neural network (CNN)-based and transformer-based approaches. Owing to the superior capacity for visual feature extraction, CNN-based models (e.g., Fast-RCNN [5], YOLO [6], and Faster-RCNN [7]) have been widely adopted in wildfire detection tasks [8,9]. For instance, Zhu et al. [10] improved YOLOv8 by integrating partial convolution and an AgentAttention module, enhancing multi-angle flame detection in forest environments. Similarly, a recent YOLOv11_MDS model incorporated Multi-Scale Convolutional Attention and Distribution-Shifted Convolution to improve small-target wildfire detection in transmission line corridors, achieving a mAP50 of 88.21% [11]. Benefiting from their localized receptive fields and parallelizable computation, CNN-based models demonstrate notable real-time performance in wildfire detection, as evidenced by architectures like FireNet-CNN [12] for explainable high-speed inference, lightweight CNNs via multi-task distillation for edge deployment [13], and CNN-BiLSTM [14] for near-real-time spread prediction. However, their inherent reliance on local features limits global contextual modeling, hindering performance in large-area smoke dispersion and distributed fire scenarios, while also increasing susceptibility to false alarms under complex environmental conditions like variable illumination and occlusion. In contrast, transformer-based architectures leverage self-attention mechanisms to capture global dependencies [15]. Representative models such as Vision Transformer (ViT) [16,17] and Detection Transformer (DETR) [18] have proven particularly effective for wildfire-related applications, including fire identification in satellite imagery, monitoring along power transmission corridors, and multi-day fire spread forecasting.

Nevertheless, the practical deployment of such transformer-based models in real-time wildfire detection systems remains constrained by their high computational demands, limited local feature modeling capability, and slower training and inference speeds compared to convolutional networks.

To leverage the advantages of both architectures, researchers have developed hybrid CNN-transformer models [19]. CoAtNet [20] integrates convolutional inductive bias with self-attention, while RT-DETR [21] designs an efficient hybrid encoder for real-time end-to-end detection. In particular, several recent studies have further optimized such hybrid designs for fire detection: SAPNet [22] introduced a spatial attention pyramid to improve multi-scale smoke recognition; GPINet [23] integrated Gaussian-process-inspired modules for better uncertainty modeling in wildfire prediction; FDNet [24] employed frequency-domain decomposition to enhance detection under low-visibility conditions; and DDFNet [25] (Information Fusion) proposed a deformable feature fusion mechanism for complex topographic fire monitoring. Despite these advances, deploying deep learning-based fire detection models on edge devices remains challenging due to limited computational resources, storage, and power. To address these issues, researchers have adopted model compression and acceleration techniques such as pruning, quantization, and knowledge distillation. For instance, Mukherjee et al. [26] deployed a quantized fire detection model on an embedded system, while Xie et al. [27] and S. Wang et al. [28] developed novel knowledge distillation methods that significantly improved model accuracy and reduced inference time.

Nevertheless, several critical challenges in wildfire detection remain unresolved, including the absence of an architecture that effectively integrates local feature extraction with global contextual modeling, limited discriminative capability for irregular flame and smoke objects under complex conditions, and high computational complexity hindering edge deployment. To address these issues, we propose HybriDet, a novel lightweight model that fuses multiple attention mechanisms and structured pruning. This design enables accurate wildfire identification and real-time performance on resource-constrained edge devices, achieving an optimal balance of accuracy, generalization, and efficiency. The main contributions of our study are as follows:

- We combined CNN and transformer to design a new wildfire detection model, utilizing the windowed attention of Swin Transformer to facilitate information exchange between image contexts. Simultaneously, incorporating bottleneck residual convolution helps address the deficiency in global perception with lower model parameter costs, effectively enhancing fire detection accuracy.
- We designed a dual attention called Coordinate-Spatial (CS) attention mechanism, which integrates Coordinate and Spatial Attention to enhance feature discrimination. It captures long-range channel dependencies through directional-aware modeling while emphasizing salient spatial regions, enabling comprehensive feature understanding for irregular flame and smoke objects.
- Our comprehensive experiments on the FASDD-UAV, FASDD-RS and Pascal Visual Object Classes (VOC) datasets demonstrate that the HybriDet achieves superior detection performance compared to advanced models, while maintaining a similar level of model complexity. Additionally, ablation studies confirm the effectiveness of each proposed module, and edge deployment experiments validate the model's real-time inference capability on embedded devices.

2. Related Works

2.1. Fire Detection Methods Based on Deep Learning

CNN and transformer models have found wide applications in wildfire detection tasks. This section introduces detection algorithms based on CNN and transformer architectures and their applications in wildfire detection. Convolutional Neural Networks (CNN) are extensively used for feature extraction in computer vision and are widely applied in object detection. Single-stage object detection models include YOLO series, SSD [29], RetinaNet [30], which directly predict class probabilities and coordinates, offering fast speed but sacrificing some accuracy. On the other hand, two-stage object detection models like RCNN [31], SPPNet [32], Faster RCNN [7] locate targets first, ensuring higher accuracy and recall before classification, but at a slightly slower speed compared to single-stage models. Both types of CNN models have been widely applied in wildfire detection. Researchers like Zhang have proposed RCNN models for small object detection in fire detection scenarios [33], capturing multi-scale image features and extracting deep and shallow details. Additionally, pyramid attention-based early forest fire detection models [34] and lightweight forest fire and smoke detection models based on YOLOv7 for unmanned aerial vehicle images [35] have been introduced. Jonnalagadda et al. proposed SegNet [36], which addresses the improvement of UAV image processing speed and detection capability in the context of limited datasets. While these models have made significant progress in real-time detection, they have yet to integrate high-precision functionalities into wildfire detection.

Since the application of transformers in computer vision, a large number of object detection networks based on transformers have emerged. Neural networks such as VIT(Vision Transformer) [16], Swin Transformer [37], and DETR(Detection TRansformer) [18] have greatly improved the ability of networks to process contextual information. The combination of convolutional neural networks and transformers is gradually becoming the mainstream model in computer vision. Shahid et al. [38] demonstrated the feasibility of using VIT for fire detection. Jiang et al. [39] proposed the SAN-SD model, which uses traditional attention mechanisms combined with K-means clustering algorithms for flame aberration detection, enabling rapid convergence of the network and improving flame detection accuracy. Li et al. [40] proposed an algorithm combining Swin transformer with Bifpn, which can capture fine-grained features at different scales and facilitate recognition of small smoke particles. Liu et al. [41] proposed TFNet, a Transformer-based multi-scale fusion fire detection network. These types of networks have high accuracy but their parameter size and computational requirements are extremely high for edge devices in terms of storage space and computing resources, making it difficult to generalize such algorithms based on existing edge devices. Therefore, based on existing methods, we propose a wildfire detection model called HybriDet that utilizes a CNN+transformer design for a multi-attention-based wildfire recognition algorithm.

2.2. Model Pruning

Model compression is a crucial technique for improving model efficiency in resource-constrained environments, and model pruning is a key technique within this domain. Pruning can be categorized as structured pruning and unstructured pruning. Structured pruning involves removing entire structured patterns, such as convolutional kernels, filter groups, or neuron layers, to reduce redundant structures and, consequently, decrease the model size. Unstructured pruning directly removes individual parameters from the model, usually based on threshold selection, simplifying the model structure. While unstructured pruning is simpler to implement, it may lead to irregular model structures.

Due to limitations in storage and computational resources on edge devices for wildfire detection, model pruning becomes imperative. Previous works, such as Fourier analysis-

based pruning of deep convolutional networks [42], have been proposed for fire detection, saving substantial storage space. Additionally, researchers have introduced flame detection algorithms based on CNN models [43], comparing the impact of different pruning methods on the algorithm and effectively reducing the number of parameters. Therefore, to optimize model performance within the capacity constraints of edge networks and maintain accuracy, minimizing memory and computational resource usage is crucial. In our approach, we perform structured pruning on our model to reduce redundant structures, alleviate edge-side burdens, and achieve efficient deployment.

3. Methodology

3.1. Overall Architecture

We propose a neural network named HybriDet, which adopts a hierarchical design approach and incorporates various attention mechanisms. The architecture of HybriDet consists of three main components: the backbone, neck, and detection head (Figure 1). In the backbone component, we utilize the SliceSamp [44] module for neural network upsampling and downsampling operations. The SwinBottle module is employed for feature extraction by combining Swin Transformer with bottleneck residual convolution to enhance global perception while keeping model parameter cost low. To quantitatively demonstrate the efficiency of our proposed modules, we compare the parameter counts of the SliceSamp and SwinBottle modules against their traditional counterparts—the strided convolution (Conv) and C2f module—across different network layers, as detailed in Table 1. Our design philosophy follows the principle of "using the right resource for the right task": we intentionally reduce the parameter cost in functional modules such as downsampling (SliceSamp), while allocating more parameters to performance-critical feature extraction modules (SwinBottle). As shown in Table 1, the SliceSamp module significantly reduces parameters compared to the standard convolution in downsampling—by approximately 50% across multiple layers—with minimal impact on performance. Conversely, the SwinBottle module, despite its higher parameter count than C2f, substantially enhances global feature modeling capability, leading to improved detection accuracy. This strategic parameter redistribution ensures that the model remains lightweight yet powerful, effectively balancing efficiency and performance.

Additionally, we incorporate the SPPF module to efficiently process images of different sizes accurately. The neck component utilizes YOLOv8's C2f module for feature extraction [45]. We design the CS Attention mechanism to extract both spatial and channel attention, thereby improving the generalization performance of the network. Moreover, we employ the ConcatBifpn module that combines Concatenation with Bifpn [46] to effectively improve mean average precision (mAP) value. Finally, in our detection head component, we adopt YOLOv8's decoupled head structure for Anchor-Free detection. This choice allows us to achieve accurate object localization without relying on predefined anchor boxes commonly used in traditional methods. By incorporating these components together into HybriDet, we aim to create an advanced neural network architecture that can achieve robust object detection performance across various datasets and scenarios.

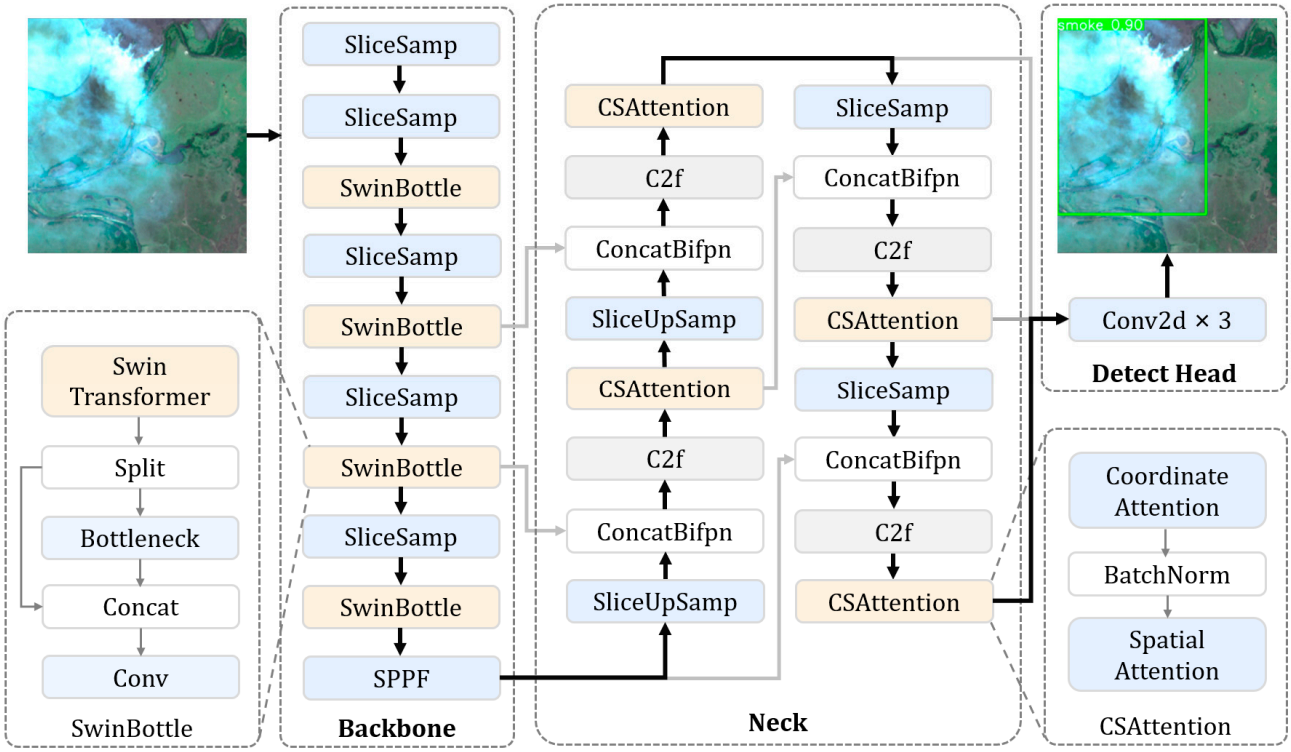

Figure 1. HybriDet Network Structure Model Diagram.

Table 1. Parameter comparison of key modules across different layers in HybriDet.

Layer	Conv Module		Slicesamp Module		C2f Module		SwinBottle Module	
	Number of Module	Params	Number of Module	Params	Number of Module	Params	Number of Module	Params
1	1	464	1	356	1	7360	1	19,426
2	1	4672	1	2816	2	49,664	2	147,208
3	1	18,560	1	9728	2	197,632	2	581,136
4	1	73,984	1	35,840	1	460,288	1	1,187,600
5	1	295,424	1	137,216	-	-	-	-

Specially, in the backbone component, we replace traditional stride convolution with SliceSamp for downsampling operations, which improves computational efficiency while maintaining model accuracy and reducing parameter count. The SwinBottle module is designed to perform feature extraction by integrating the Swin Transformer and C2f modules. It incorporates a bottleneck funnel structure after convolution to downsize features and introduces residual connections in the bottleneck section of the module. Additionally, we integrate SPPF to handle inputs of different sizes and generate fixed-dimensional outputs [47], enhancing flexibility in processing wildfire images. This part allows the neural network to process wildfire images of various sizes, enhancing model flexibility. In the neck component, we employ the C2f module for feature extraction, taking into account computational considerations. We introduce CS Attention based on Coordinate Attention [48] and Global Attention Mechanism (GAM) [49] to enhance the model's generalization performance. Additionally, in this section, we apply the ConcatBifpn module, combining Concat and Bifpn, to learn proper weights for feature fusion and capture essential information during the fusion process. This weight adjustment can adaptively occur based on the characteristics of the training data, enabling the network to better

perform feature fusion. In the head part, we adopt YOLOv8's decoupled detection head, utilizing distribution focal loss to predict targets' coordinates as distributions.

The proposed HybriDet, similar to YOLOv8, adopts a collaborative "divide-and-conquer-aggregate" decision-making mechanism. Its Backbone and Neck components output three parallel feature maps of different scales into the Detect module, which progressively focus on detecting small, medium, and large wildfire and smoke targets, respectively. Each scale's feature map first enters an independent decoupled detection head to generate its own bounding box predictions, classification outputs, and objectness confidence scores. Subsequently, all multi-scale predictions are gathered into a unified proposal set for preliminary screening based on confidence scores. During this process, leveraging its trained discriminative capability, the model naturally assigns higher confidence to genuine targets at their respective optimal detection scales. Finally, all these multi-scale predictions undergo Non-Maximum Suppression (NMS) on a unified competitive platform, effectively integrating multi-scale perceptual advantages to produce a consolidated set of detection results capable of accurately identifying targets across varying sizes.

Overall, these design choices aim at improving both efficiency and effectiveness of wildfire detection by optimizing network architecture components such as downsampling operations, feature extraction modules, attention mechanisms, fusion processes, and target coordinate prediction strategies within HybriDet framework.

The loss function of the network consists of two components: classification loss and object recognition loss. The classification loss is computed using cross-entropy, which measures the discrepancy between the predicted probability distribution and the actual labels to evaluate the accuracy of model predictions. The bounding box loss is a combination of L_{DFL} (Distance-IoU Loss) [50] and L_{CIoU} (Complete Intersection over Union Loss) [51]. The detailed formulas are presented below.

$$L_{DFL} = -((y_{i+1} - y)\log(S_i) + (y - y_i)\log(S_{i+1})) \tag{1}$$

$$L_{CIoU} = 1 - \mathrm{IoU} + \frac{\rho^2(b^{pre}, b^{gt})}{c^2} + \alpha v \tag{2}$$

$$\mathrm{IoU} = \frac{b^{pre} \cap b^{gt}}{b^{pre} \cup b^{gt}} \tag{3}$$

$$\mathrm{v} = \frac{4}{\pi^2}[\arctan(\frac{w^{gt}}{h^{gt}}) - \arctan(\frac{w^{pre}}{h^{pre}})]^2 \tag{4}$$

L_{DFL} is based on distance metric and IoU (Intersection over Union), simultaneously considering the differences in position and shape of bounding boxes. It introduces a smoothing parameter to mitigate discontinuity issues during IoU calculation. The core concept of its calculation formula lies in modeling continuous bounding box coordinates as a discrete probability distribution. Here, y represents the target coordinate value, y_i and y_{i+1} denote the two adjacent discretized coordinate bins where the value y falls, and S_i and S_{i+1} are the model's predicted probabilities corresponding to these two bins. This loss optimizes coordinate prediction by maximizing the probabilities of the two nearest neighbor bins associated with the target coordinate value y. Essentially, it minimizes the discrepancy between the predicted coordinate distribution and the true coordinate. On the other hand, L_{CIoU} further improves upon IoU by incorporating factors such as aspect ratio and area when calculating overlap between two bounding boxes. In this formula, IoU represents Intersection over Union, b^{pre} and b^{gt} represent the predicted box and the real box, respectively, ρ denotes the Euclidean distance between the two bounding boxes, c represents the diagonal distance of the closed region of the two bounding boxes, v is used to measure the consistency of relative proportions between the two bounding boxes, and

α is the weight coefficient. The specific calculation method is shown in the formula below. This allows for more accurate assessment of overlap between predicted and ground truth bounding boxes, enhancing the accuracy of object detection tasks. These losses provide robustness in handling scenarios with small objects or high-density scenes where target overlap occurs frequently. By combining these losses with weighted averages along with classification loss, it guides network training to optimize both class prediction accuracy and precise localization of objects. Through integrating classification loss with object recognition losses like L_{DFL} and L_{CIoU}, the overall loss function enhances model performance in object detection tasks by accurately capturing both class information and fine-grained localization details.

Wildfires and smoke are typical detection targets characterized by their irregular shapes, blurry contours, and multi-scale variations. Their edges are often gradual and indistinct, and their morphology is far from a standard rectangle, with extreme aspect ratio variations, posing significant challenges for precise localization. In such tasks, the L_{DFL} loss, based on a Distribution Focal Loss mechanism, enables the model to learn a discrete probability distribution over boundary locations instead of forcibly regressing an absolute coordinate. This approach is particularly effective in handling the ambiguous boundaries of smoke and enables sub-pixel-level precise localization of early-stage, subtle flames or thin smoke traces.

Simultaneously, the L_{CIoU} loss operates from a geometric global perspective. Beyond considering the overlapping area, it introduces additional penalty terms for central point distance and aspect ratio differences, effectively guiding the predicted bounding box to conform to the irregular shapes of smoke plumes. The two components work synergistically: L_{DFL} ensures precision in coordinate prediction at a micro level, while L_{CIoU} optimizes the overall alignment of the bounding box at a macro level. Together, they significantly enhance the model's detection robustness and localization accuracy in complex wildland scenarios.

In summary, this design brings various advantages, including reduced computational complexity, improved model flexibility, enhanced generalization performance, and effective feature fusion. The hierarchical design and integration of multiple attention mechanisms in the entire network structure enable it to better adapt to inputs of different sizes, improving the model's performance in wildfire detection tasks.

3.2. SwinBottle Component

Swin Transformer introduces a windowed attention mechanism that confines computation within local windows while incorporating relative positional encoding. This attention mechanism not only effectively reduces computational complexity but also excels at capturing positional information. This proves particularly crucial when processing large-scale image data, as it enables the model to acquire global information while maintaining precise focus on local details. Through this approach, Swin Transformer achieves a better balance between global and local information acquisition in visual tasks, thereby enhancing model performance.

As shown in Figure 2, the SwinBottle module designed in this paper integrates Swin Transformer with a bottleneck structure, fully leveraging the advantages of both. Specifically, the bottleneck structure—a classic deep neural network architecture—compresses feature dimensions to reduce computational load while preserving information representation capability. The bottleneck structure and Swin Transformer's multi-head self-attention are interconnected through split operations, enabling the module to efficiently transmit and integrate information during feature extraction. This design allows the SwinBottle module to retain critical information while alleviating computational burden and improving overall model performance. This integration also endows the SwinBottle module with

enhanced flexibility, enabling adaptation to features of different scales and receptive fields. Such adaptability makes it better suited for varying image sizes, providing strong transfer capability for future wildfire detection on different edge devices. The feature map is processed through an initial Transformer and then split into three feature tensors of identical dimensions via a channel splitting operation, each directed to three paths with specialized functions. The first path serves as the main feature extraction stream, feeding into a Bottleneck module. It utilizes cross-layer connections and nonlinear activation functions to perform deep feature transformations, forming the core computational graph of the module. The second path establishes an identity mapping shortcut, directly passing the originally split features to the final concatenation layer. This approach preserves spatial structural information while creating a short path for gradient backpropagation, effectively alleviating the gradient vanishing problem caused by increasing network depth. The third path acts as a feature enhancement stream, allowing the untransformed initial features to participate directly in the final fusion. Through concatenation with the output features from the main path and the shortcut features along the channel dimension, it achieves multi-scale feature recalibration and enhances gradient flow diversity. This mechanism significantly improves the model's feature representation capability while maintaining computational efficiency, providing an optimal balance between accuracy and speed for real-time object detection tasks. The bottleneck structure employs multiple convolutional operations to further compensate for the Transformer's limitations in local feature extraction, allowing the model to learn more effective representations during training and enhancing its ability to learn multi-scale and hierarchical features.

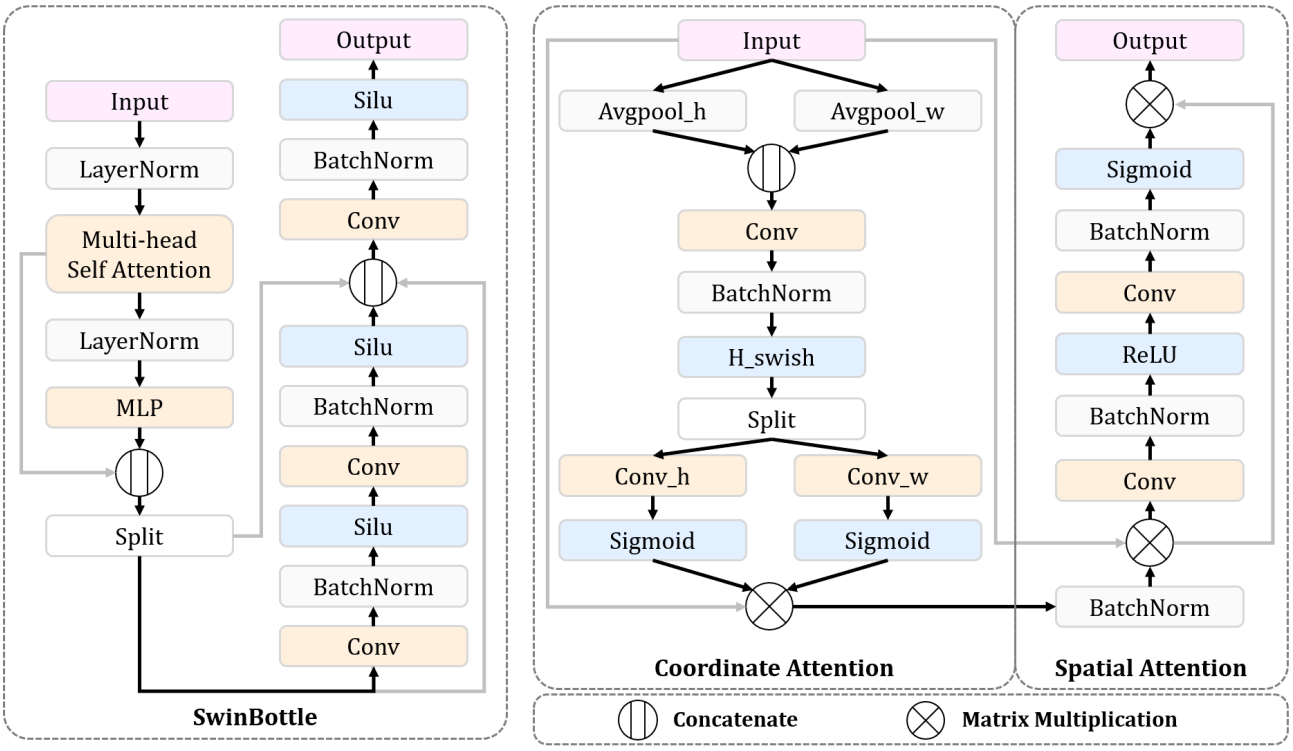

Figure 2. Structure of SwinBottle and CSAttention.

Overall, compared to traditional Transformers that lack sufficient detail extraction capability, the SwinBottle attention module simultaneously addresses global and local feature extraction while improving accuracy through increased attention depth. The design of the SwinBottle module enables Swin Transformer to demonstrate outstanding performance in flame recognition tasks, representing a significant technological innovation. The only

drawback lies in the exponentially increased model complexity due to the Transformer architecture and deepened network structure. To address this high complexity, we conducted experiments on structural pruning and model quantization deployment, effectively reducing model complexity and improving inference speed. These improvements will be detailed in Section 4.5.

3.3. Coordinate-Spatial (CS) Attention Mechanism

CS Attention combines Coordinate Attention and Spatial Attention, significantly improving the performance of networks in flame detection tasks (Figure 2). The input feature map is fed into the CS Attention module, where it first undergoes Coordinate Attention encoding, followed by Spatial Attention weighting. Coordinate Attention captures long-range dependencies of input feature maps in both horizontal and vertical directions through direction-aware feature maps, which aids in modeling inter-channel relationships and incorporates directional and position-sensitive information. On the other hand, Spatial Attention computes attention in the spatial dimension, emphasizing important spatial regions within the feature map, further enhancing the representation capability of positional information. By integrating these two attention mechanisms, CS Attention is capable of capturing rich information across different dimensions, taking into account both inter-channel relationships and the significance of specific spatial regions.

Coordinate Attention enhances the network's feature representation capability. It can take any intermediate feature tensor as input and produce an output with the same size as the tensor. It not only captures inter-channel relationships but also direction-aware and position-sensitive information, assisting the model in more accurately localizing and identifying objects. Furthermore, its simple and lightweight design makes it suitable for real-time detection tasks. Our proposed CS Attention mechanism extends Coordinate Attention through the integration of a dedicated Spatial Attention submodule, enhancing positional encoding capabilities. Within this architecture, the Spatial Attention submodule first processes the fused coordinate embeddings via convolutional feature transformation, thereby enriching representational capacity. Specifically, it operates directly on the spatial dimensions of the feature map, leveraging two convolutional layers to compute spatial attention weights. Unlike other methods, Spatial Attention does not rearrange channels but directly computes attention in the spatial dimension of the feature map. Additionally, the Spatial Attention submodule removes pooling layers to retain more feature map information. In essence, the Spatial Attention mechanism focuses on attention computation in the spatial dimension of the feature map, aiming to highlight certain spatial regions while ignoring less important ones, thereby improving model performance.

Finally, we leverage the advantages of residual networks, using distant residual connections to prevent overfitting. This not only increases the model's representation ability but also enhances its generalization performance. This integrated attention mechanism allows the network to more accurately localize and identify objects while maintaining low model complexity, making it suitable for real-time flame detection tasks. The design of CS Attention not only considers the directional and positional information within the feature map but also fully utilizes spatial relationships, providing the model with more comprehensive perceptual abilities, thus performing exceptionally well in complex scenarios. This combined attention mechanism offers an effective and lightweight solution for object localization and recognition in computer vision tasks.

4. Experiment

We conducted extensive experiments on various deep learning models using two datasets, namely FASDD-UAV and FASDD-RS fire detection datasets (Figures 3 and 4) [41]. In this

section, we compare our network architecture with the YOLOv8 model. Additionally, we compare the algorithm complexity of these methods, using model size as a measure of spatial complexity (i.e., memory) and detection speed per image as a measure of time complexity (i.e., computational latency). We also performed ablation experiments and comparison experiments before and after pruning to demonstrate the effectiveness of the model.

Figure 3. Example images of FASDD-RS Dataset: (**a**) Smoke; (**b**) Neither fire nor smoke.

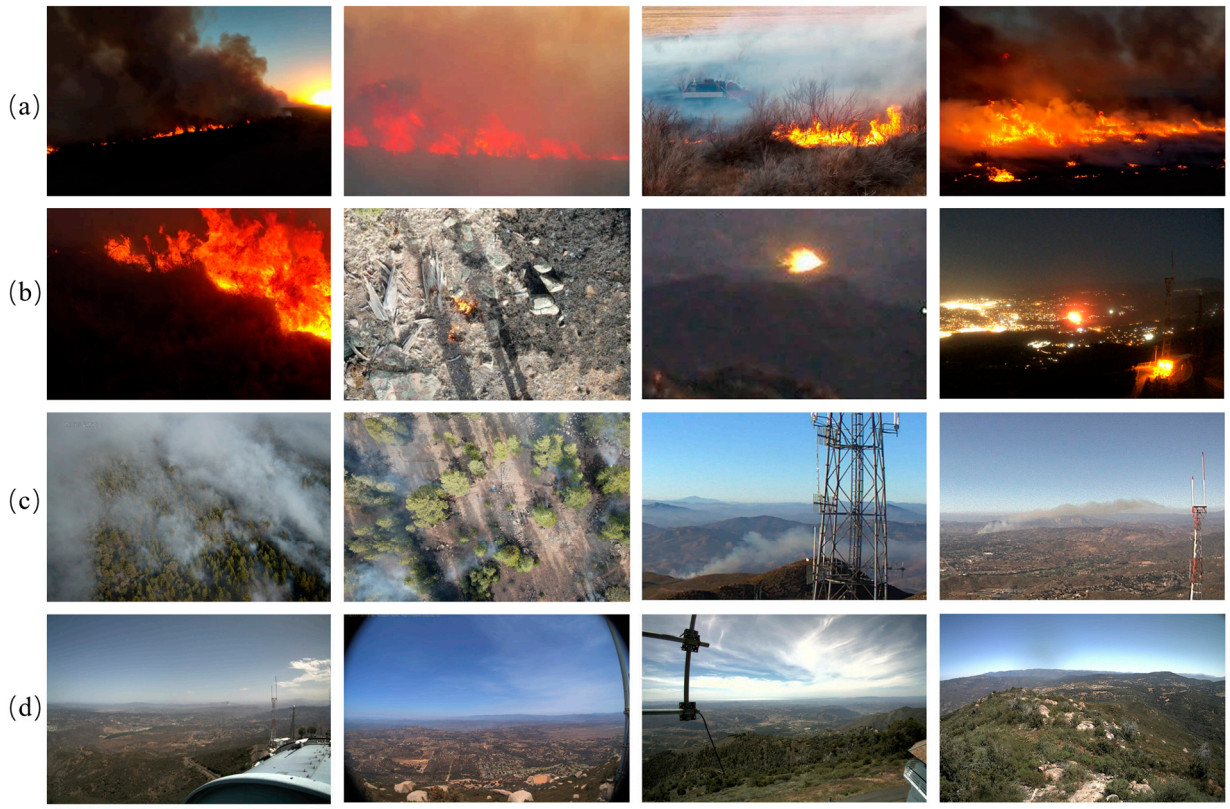

Figure 4. Example images of FASDD-UAV Dataset: (**a**) Both fire and smoke; (**b**) Fire; (**c**) Smoke; (**d**) Neither fire nor smoke.

4.1. Dataset

The FASDD (Flame and Smoke Detection Dataset) is a benchmark dataset designed for fire detection tasks. It comprises a diverse collection of images that depict intricate fire and smoke scenarios, filling the gap in the availability of large-scale benchmark datasets for this specific visual task. With over 120,000 images, the FASDD offers valuable resources for training and validating wildfire detection models. Among them, the FASDD-UAV subset encompasses high-resolution images captured from unmanned aerial vehicles, while the FASDD-RS subset includes imagery with 10 m resolution acquired through multispectral remote sensing satellites. Both subsets play a crucial role in enabling deep learning algorithms to be deployed effectively on both airborne and spaceborne sensors for wildfire detection purposes. Table 2 presents detailed information about this dataset. Since FASDD-RS is primarily designed to detect smoke, there are no flame pictures.

Table 2. FASDD is used in wildfire detection.

Dataset	Images	Train	Val	Test	Fire	Smoke	Both	Neither
FASDD-UAV	25,097	12,551	8365	4181	210	5080	7821	11,986
FASDD-RS	2223	1112	741	370	-	1335	-	888

4.2. Experimental Settings

We chose the CentOS operating system as the experimental environment, with an NVIDIA GeForce RTX 3090 GPU. Model training was performed using dual RTX 3090 GPUs, while validation and finetuning after pruning were conducted with a single GPU configuration. Leveraging high-performance computing devices for training and validation accelerates the experimental process and improves model effectiveness. The number of training epochs for all experiments was set to 1000, with an initial training batch size of 64 and a fine-tuning batch size of 32. The AdamW optimizer was used, with the learning rate set to 0.01, while other configurations were kept consistent with YOLOv8n's default settings. None of the experiments utilized any pre-trained models. To evaluate our trained model's performance in wildfire detection, we employed Precision, Recall, and mean average precision (mAP) as evaluation metrics. These metrics provide a reasonable and scientifically sound assessment of how well our model performs in detecting wildfires.

Regarding the pruning implementation, we adopted a global, filter-level pruning approach targeting the convolutional layers within the backbone and neck networks. The pruning criterion was based on the L1-norm of the filters, operating on the principle that filters with smaller norms contribute less to the final output. A pre-defined, global sparsity ratio of 20% was applied, meaning that 20% of the least important filters across the designated layers were identified and permanently removed. This directly reduced the number of parameters and the computational complexity (FLOPs) of the model. Following the pruning operation, the model underwent a fine-tuning phase using the original training dataset. This critical recovery stage allowed the remaining weights to adapt and compensate for the capacity loss induced by pruning. To further validate the wildfire detection performance of the proposed models on low-computing-power edge devices, we conducted a comparative evaluation on a Raspberry Pi 4B, a platform commonly used in satellite or UAV equipment [52–54]. Specifically, we performed 16-bit quantization and converted all models—including the baseline, our original model, and the pruned-and-fine-tuned model—into TensorRT format to enable efficient inference on the Raspberry Pi 4B.

4.3. Comparative Experiment

As shown in Table 3, we have conducted comparative experiments on the FASDD-RS dataset using today's mainstream object detection models, including YOLOv7, YOLOv8, YOLOv12, Swin Transformer, RT-DETR, and our proposed model, HybriDet. The results demonstrate that the HybriDet model has achieved exceptional accuracy scores, indicating its strong recognition ability in detecting true positive wildfire targets. On the FASDD-RS validation set, our HybriDet improved precision by 5.8 compared to the YOLOv7 model, while recall remained almost the same. In terms of mAP50, our HybriDet outperformed YOLOv7 by 2.6 points. On the FASDD-RS test set, compared to YOLOv7, our model achieved a 12-point higher precision and a 1.4-point higher mAP50, maintaining a significant lead.

Table 3. Comparison Experiment on FASDD-RS.

Model	Dataset	Val			Test		
		Precision	Recall	mAP50	Precision	Recall	mAP50
YOLOv7	FASDD-RS	66.5	63.8	66.6	60.2	68.7	65.2
YOLOv8	FASDD-RS	65.7	60.1	65.0	64.4	60.7	61.7
YOLOv12	FASDD-RS	72.6	62.8	68.1	71.2	63.9	65.6
Swin Transformer	FASDD-RS	26.3	87.2	68.6	29.4	87.5	69.9
RT-DETR	FASDD-RS	69.4	61.1	60.8	68.3	60.2	60.3
HybriDet	FASDD-RS	72.3	63.7	69.2	72.2	62.6	66.6
YOLOv8	VOC	77.1	68.1	75.5	-	-	-
HybriDet	VOC	79.7	71.6	79.1	-	-	-

Compared to YOLOv8, on the FASDD-RS validation set, our model achieved 6.6 points higher precision, 3.6 points higher recall, and 4.2 points higher mAP50. On the test set, precision was higher by 7.8 points, recall by 1.9 points, and mAP50 by 4.9 points. This suggests that by introducing a dual attention mechanism, our model achieves multiple extraction of detailed and global features, effectively improving the mAP50 compared to the YOLO series. Although Swin Transformer achieved a higher mAP50 on the FASDD-RS dataset, its precision is relatively low, mainly due to the significant impact of a higher recall rate on mAP50. This may be due to the model's difficulty in distinguishing between positive and negative samples, leading to a higher false positive rate, which is typically caused by the Transformer architecture's excessive focus on high-confidence features as well as its insufficient sensitivity to small or blurry smoke in the dataset. Additionally, the detection speed of the Swin Transformer architecture is significantly slower than our HybriDet, making it difficult to achieve real-time detection tasks. On the test set, our model also demonstrates a notable performance advantage compared to the latest YOLOv12 [55] and RT-DETR. Specifically, our model achieves 1.0-point and 6.3-point improvements in mAP50 over YOLOv12 and RT-DETR, respectively. To ensure the reliability of our performance comparison, we conducted five independent experimental runs. In all five experimental runs, we observed perfectly consistent results: YOLOv12 consistently achieved an mAP50 of 65.6, while our HybriDet model consistently reached 66.6, yielding a stable performance improvement of +1.0 mAP50 points across all trials. Given the zero variance observed across all five trials, conventional parametric tests such as the t-test are not applicable. Nevertheless, the complete consistency of the results across independent experiments provides strong empirical evidence for the robustness of the observed improvement. Under the null hypothesis that no true performance difference exists, the probability of obtaining five consecutive positive outcomes purely by chance is $0.5^5 = 0.03125$, which is below the conventional significance threshold of $p < 0.05$. This consistent pattern across all runs

substantiates that the performance advantage of the proposed method is genuine and highly reproducible, rather than a result of random variation, thereby reinforcing the effectiveness and reliability of our approach.

Overall, our proposed HybriDet performed best on the FASDD-RS dataset, especially in terms of precision and mAP50. This suggests that HybriDet may be more suitable for handling object detection tasks in wildfire detection datasets. To further validate the scalability of our model, we also conducted cross-dataset comparative experiments on the Pascal VOC dataset, a benchmark dataset for object detection. The results show that our model achieved 2.6 points higher precision, 3.5 points higher recall, and 3.6 points higher mAP50, demonstrating its comprehensive advantages over the baseline model, YOLOv8.

4.4. Performance Evaluation

This study compared HybriDet's parameter count before and after pruning and its performance on the validation and test sets.

The pruning process begins by loading a pretrained model and collecting weight parameters from all BatchNorm layers, sorted by magnitude. Channels are then filtered based on the importance of BatchNorm γ parameters using an adaptive threshold that ensures at least 8 channels are retained per layer, while simultaneously adjusting the channel dimensions of the current convolutional layer and subsequent connected layers. This pruning operation is then applied hierarchically throughout the network. Finally, parameter gradients are enabled to prepare for fine-tuning. Throughout this process, the network structure remains intact while reducing channel counts. After pruning, our model's parameter count increased by only 0.47 M. In the validation set, compared to YOLOv8, our model showed a 1.4% increase in mAP50. On the test set, the pruned model demonstrated a 1.5% increase in mAP50 compared to before pruning, and a 6.4% increase compared to the original YOLOv8 model. This indicates that our model maintains a relatively small parameter count while effectively training feature maps, achieving good generalization performance, making it suitable for deployment in resource-constrained edge computing environments. As shown in Figure 5, Figure 5a demonstrates that our model surpasses YOLOv8 in confidence, Figure 5b indicates our model's more comprehensive performance in smoke detection, and Figure 5c,d reveal YOLOv8's tendency for multiple detections. These results suggest that, thanks to various attention mechanisms in HybriDet, the model can extract image features more accurately, making HybriDet an excellent wildfire detection model.

To further validate the generalization effect of our pruned model, we conducted cross-dataset validation on the FASDD-UAV dataset, and the results are shown in Table 4. The results indicate that both YOLOv8 and our model have an mAP50 exceeding 90%, and the model has reached a performance bottleneck. After pruning, the model size was reduced by 24.7%, while the mAP50 of our model was improved by 0.2% compared to the baseline model, still maintaining a slight increase in accuracy. Moreover, our model has a higher confidence level than YOLOv8, which means better generalization performance. From the prediction effect diagram in Figure 6, it can be seen that YOLOv8 has false detections and missed detections, while our model does not have such issues, and our model's confidence level is higher both before and after pruning. Figure 7 shows the confusion matrix for all pruning comparison experiments. The vertical axis of the confusion matrix is the real label, and the horizontal axis is the predicted label. It can be calculated from the confusion matrix that on the FASDD-RS dataset, the false positive rate of YOLOv8 is 48.6%, while the false positive rate of our model before pruning is 47.9%, and the false positive rate after pruning is 46.8%. It is 1.8% lower than YOLOv8. On the FASDD-UAV dataset, the false positive rate of YOLOv8 is 20.8%, while the false positive rate of our model is 19.6% before pruning and 19.0% after pruning. The false positive rate is 1.8% lower than that of YOLOv8. In

summary, compared to YOLOv8, our model is better suited for wildfire detection tasks; compared to the original model, the pruned model reduces its size while maintaining an improved performance.

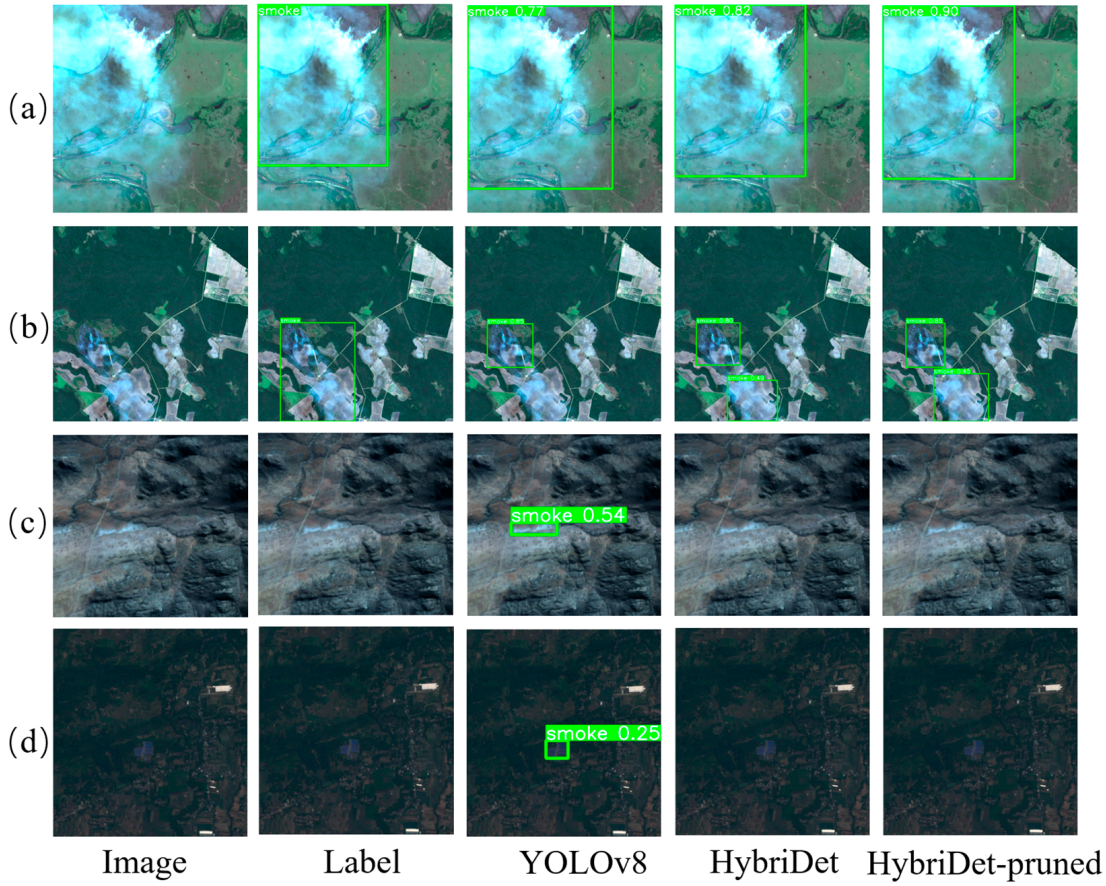

Figure 5. Model Wildfire Prediction Effect Diagram on FASDD-RS: (**a**) High confidence rate; (**b**) Comprehensive detection; (**c**,**d**) Absence of multiple detection tendencies.

Table 4. Pruning Comparison Experiment on FASDD-UAV and FASDD-RS.

Dataset	Model	Size	FASDD Val			FASDD Test		
			Precision	Recall	mAP50	Precision	Recall	mAP50
FASDD-RS	YOLOv8	5.98 M	65.7	60.1	65.0	64.4	60.7	61.7
	HybriDet (original)	8.21 M	72.3	63.7	69.2 (+4.2)	72.2	62.6	66.6 (+4.9)
	HybriDet (pruned)	6.45 M	**66.1**	**65.1**	**66.4 (+1.4)**	**72.1**	**59.4**	**68.1 (+6.4)**
FASDD-UAV	YOLOv8	5.99 M	88.8	87.9	92.2	89.4	87.4	92.3
	HybriDet (original)	8.31 M	90.3	88.3	92.6 (+0.4)	89.8	87.9	92.3
	HybriDet (pruned)	6.26 M	**90.4**	**88.0**	**92.4 (+0.2)**	**90.3**	**88.0**	**92.5 (+0.2)**

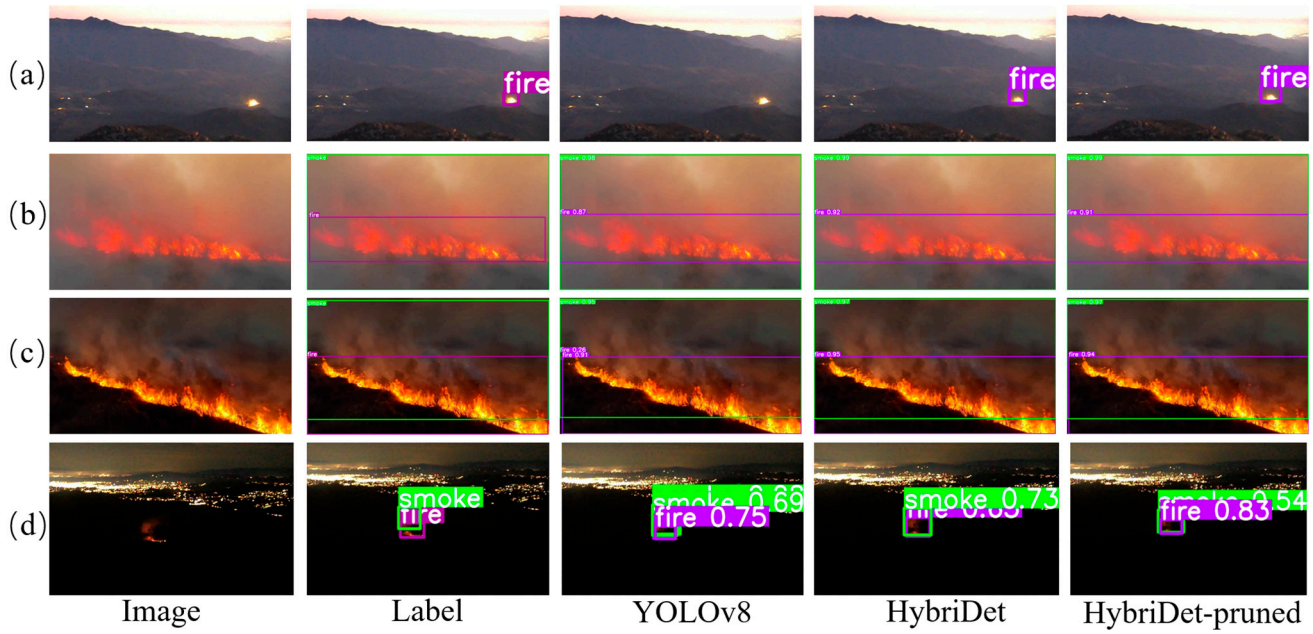

Figure 6. Cross-dataset Validation Experiment on FASDD-UAV: (**a**) No missed detections; (**b**) High confidence rate; (**c**,**d**) No multiple target detection anchor frames.

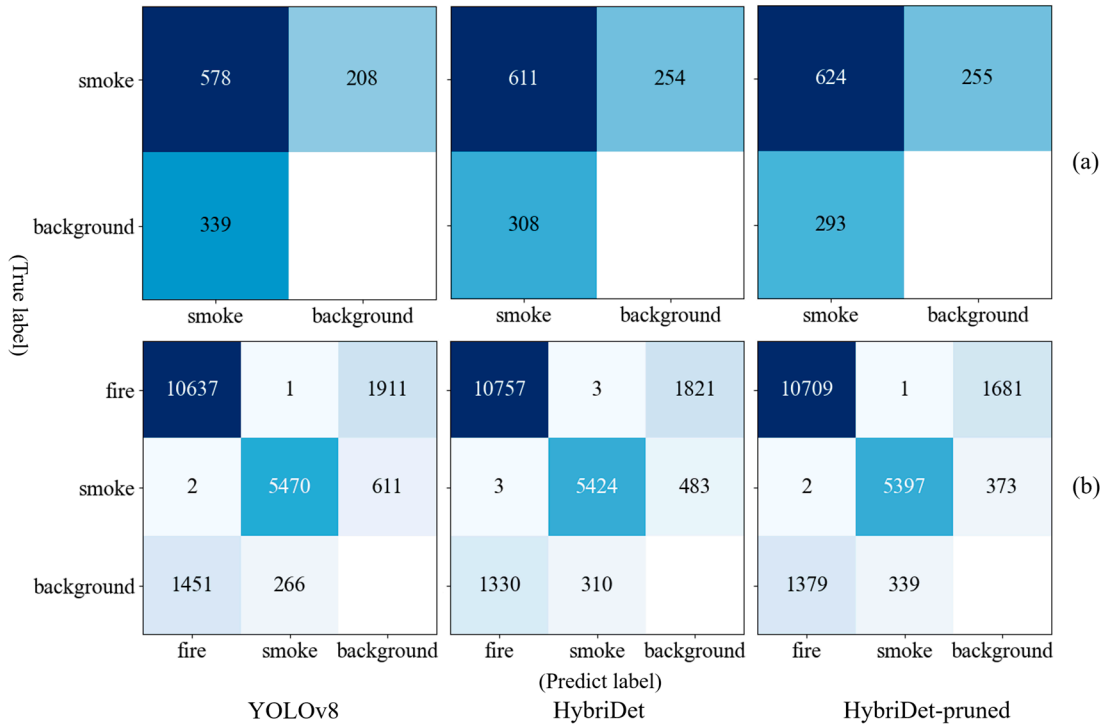

Figure 7. Confusion Matrix of Pruning Comparison Experiment: (**a**) FASDD-RS; (**b**) FASDD-UAV.

Figure 8 presents a comparative analysis of fire smoke detection results, validating the superiority of our attention mechanism in multi-scenario detection. (a) demonstrates robust identification capabilities for both large-scale smoke diffusion and small-target smoke plumes. It is noteworthy that precise detection is maintained in the third sample despite partial occlusion in the aerial imagery. (b) highlights the model's dual advantages in fire detection: high sensitivity to small-scale flame features coupled with the ability to identify accompanying smoke. Given the strong link between fire and smoke in actual

disaster events, this comprehensive detection capability significantly enhances the reliability of early warning systems in practical monitoring scenarios.

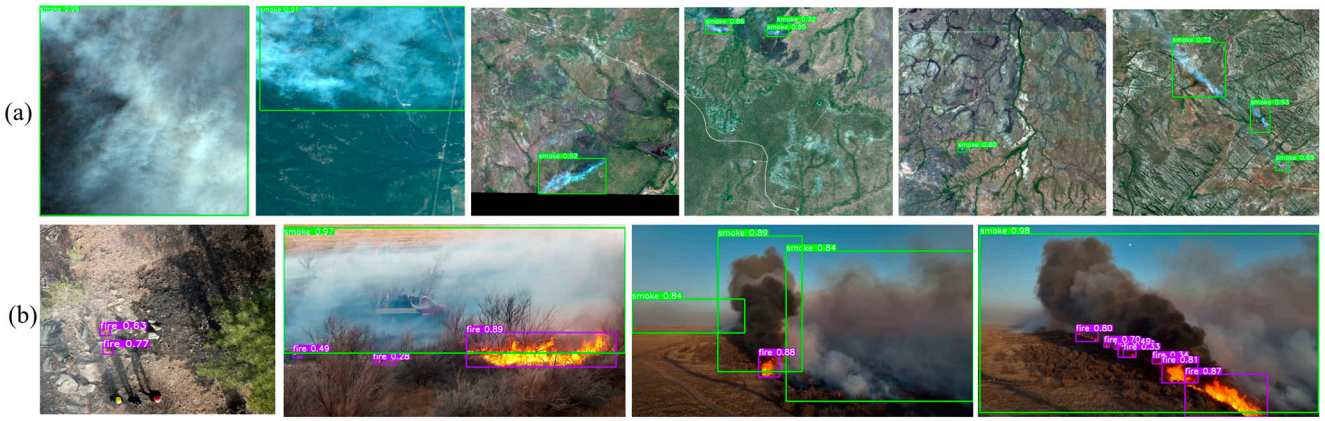

Figure 8. Comparison of Smoke and Fire Results: (**a**) Smoke; (**b**) Fire.

4.5. Edge Deployment Optimization

Table 5 systematically compares the deployment performance of the baseline YOLOv8 model, the proposed HybridDet model, and its pruned version on the FASDD-RS dataset across both RGB and Short-Wave Infrared (SWIR) modalities. All models were optimized using 16-bit quantization and evaluated in terms of parameter size, detection accuracy (mAP50), inference latency, and power consumption on two hardware platforms: an NVIDIA GeForce RTX 3090 and a Raspberry Pi 4B.

Table 5. Deployment Optimization Experiment on FASDD-RS.

Devices	Models	**YOLOv8**		**HybriDet (Original)**		**HybriDet (Pruned)**	
	datasets	FASDD-RS (RGB)	FASDD-RS (SWIR)_	FASDD-RS (RGB)	FASDD-RS (SWIR)	FASDD-RS (RGB)	FASDD-RS (SWIR)
	Patameters (MB)	3.07	3.09	4.14	4.17	3.23	3.24
NVIDIA GeForce RTX 3090	mAP50 (%)	61.6	63.8	66.5	68.3	68.0	69.7
	Latency (ms)	2.1	2.4	3.4	3.6	3.2	3.4
	Power consumption (W)	320–350	320–350	320–350	320–350	320–350	320–350
Raspberry PI 4B	mAP50 (%)	60.3	62.8	65.3	67.4	66.9	68.8
	Latency (ms)	29.8	33.6	38.9	40.7	37.4	39.2
	Power consumption (W)	7–10	7–10	7–10	7–10	7–10	7–10

In terms of model complexity, the pruned HybridDet achieves a compact size of 3.23 MB—only 0.16 MB larger than YOLOv8, while improving mAP50 by 6.4% on the FASDD-RS RGB dataset. This minimal parameter increase, coupled with a substantial gain in accuracy, stems from our architecture's ability to capture both global and local contextual information, enhanced by a dual-attention mechanism that strengthens feature discriminability. These improvements allow the pruned model to significantly outperform the baseline in accuracy, underscoring the efficacy of our structural design. In terms of detection accuracy, the original HybridDet attains mAP50 scores of 66.5% (RGB) and 68.3% (SWIR) on the RTX 3090, exceeding YOLOv8 by 4.9 and 4.5 percentage points, respectively. The pruned version further elevates performance to 68.0% (RGB) and 69.7% (SWIR), indicating that the pruning process not only reduces model size but also slightly enhances generalization, likely due to the elimination of redundant parameters. A similar trend is observed on

the Raspberry Pi 4B, where both HybridDet variants consistently surpass YOLOv8, albeit with a slight decrease in absolute mAP50 due to limited computational precision. Inference latency was evaluated under two hardware settings: a high-performance RTX 3090 GPU and a resource-constrained Raspberry Pi 4B. On the RTX 3090, the pruned HybridDet shows a noticeable speed-up over the original model, with latency decreasing from 3.4 ms to 3.2 ms on RGB and from 3.6 ms to 3.4 ms on SWIR. Although YOLOv8 remains the fastest model, the pruned HybridDet offers a favorable balance between speed and accuracy. On the Raspberry Pi 4B, the pruned model also achieves lower latency—37.4 ms for RGB and 39.2 ms for SWIR—compared to the original HybridDet, highlighting its suitability for resource-constrained environments. Power consumption remains consistent across models under the same hardware configuration, with the RTX 3090 drawing 320–350 W and the Raspberry Pi 4B operating at 7–10 W. This suggests that the use of our model on edge computing devices for fire detection holds significant potential for reducing operational costs and resource consumption. In summary, the pruned HybridDet model achieves an optimal balance among model size, inference speed, and detection accuracy. It not only surpasses the YOLOv8 baseline in accuracy—especially in SWIR-based flame detection—but also maintains competitive speed and reduced size after pruning, demonstrating strong potential for real-time flame detection in satellite or drone-based applications platforms.

4.6. Ablation Study

In this study, we used YOLOv8n as the baseline model and compared the effects of each module on the FASDD-RS dataset, as shown in Table 6. When only the SliceSamp module was introduced, the mAP50 on the FASDD-RS validation set decreased by 1.6%, while it increased by 3.6% on the test set. The SliceSamp module significantly improved the mAP50 on the test set while reducing parameters and computational load, with only a slight decrease in the validation set. CS Attention achieved channel attention and spatial attention extraction, providing the model with more comprehensive perception capabilities. After adding CS Attention, with other modules using only SliceSamp, the mAP50 on the validation set increased by 0.8%. With other modules using only SliceSamp and SwinBottle, the mAP50 on the validation set increased by 3.2%. This experiment indicates that, thanks to CS Attention's consideration of directional and positional information, HybriDet gains representation capabilities in complex scenes, improving wildfire recognition.

Table 6. Ablation Experiment Results on FASDD-RS.

Model				FASDD-RS Val			FASDD-RS Test		
SliceSamp	**CS Attention**	**SwinBottle**	**ConcatBifpn**	**Precision**	**Recall**	**mAP50**	**Precision**	**Recall**	**mAP50**
				65.7	60.1	65.0	64.4	60.7	61.7
√				63.6	63.2	63.4	69.4	61.5	65.3
√	√			65.6	61.5	64.2	66.3	59.0	64.6
√		√		67.5	60.4	65.1	67.8	64.5	66.3
√	√	√		67.9	62.5	68.3	68.1	62.4	65.9
√	√	√	√	**72.3**	**63.7**	**69.2**	**72.2**	**62.6**	**66.6**

The addition of SwinBottle better complements global features. With only SliceSamp, the addition of SwinBottle increased mAP50 by 1.7% on the validation set; on the test set, it increased by 1.0%. With SliceSamp and CS Attention, on the validation set, it increased by 4.1%; on the test set, it increased by 1.3%. Since SwinBottle can efficiently transmit and integrate information while extracting features, it allows HybriDet to learn more effective representations during training, enhancing its ability to learn features at different scales and

levels. Finally, after adding ConcatBifpn, mAP50 on the validation set increased by 0.9%; on the test set, mAP50 increased by 0.7%. ConcatBifpn allows features to better capture important information during the fusion process. Overall, our model outperforms the original YOLOv8n model with 6.6% higher precision, 3.6% higher recall and 4.2% higher mAP50. On the FASDD-RS test set, it achieves 7.8% higher precision, 1.9% higher recall, 4.9% higher mAP50. In summary, our designed network achieves real-time detection with lower parameters, enhancing wildfire detection performance.

5. Discussion

The HybriDet framework proposed in this study has achieved significant breakthroughs in edge-device-oriented remote sensing wildfire detection. By integrating the local feature extraction capability of CNNs with the global context modeling strength of Transformers, HybriDet effectively overcomes the inherent limitations of existing models in processing highly heterogeneous remote sensing imagery. The model demonstrates excellent recognition performance in overhead view smoke scenarios, as shown in Figure 8a. However, under certain extreme weather conditions, the model may still exhibit rare false detections. For example, in Figure 9a, extensive dense fog is occasionally misidentified as large-scale fire smoke with low confidence. Figure 9b presents another challenging scenario: during nighttime, distant ground-based yellow lights are mistaken for flames due to the long shooting distance of UAVs. Such images contain targets with subtle features, posing significant challenges for both manual interpretation and automated recognition.

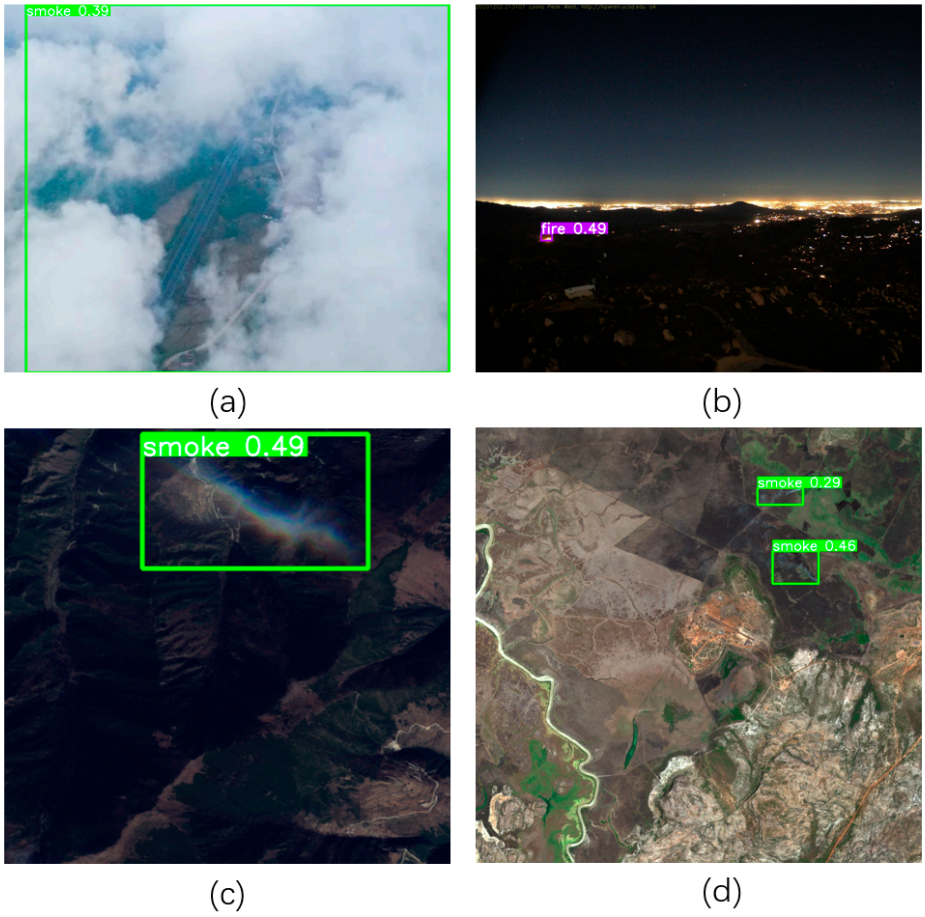

Figure 9. Some Cases of False Detection by HybriDet: (**a**) Fog; (**b**) Lighting; (**c**) Rainbow; (**d**) Cloud.

Furthermore, Figure 9c,d illustrate two additional types of false detections: rainbows being misclassified as smoke, and non-smoke small objects being incorrectly identified as smoke. It is important to emphasize that these false detections occur very infrequently, and in such cases, the model typically outputs recognition confidence scores below 0.50, indicating high uncertainty in its predictions. This provides a feasible pathway for subsequent false positive elimination through confidence-based filtering.

Nevertheless, HybriDet still has room for improvement. Its performance is influenced by the characteristics of annotated datasets such as FASDD, and its generalization capability under low-resolution or extreme weather conditions (e.g., thick clouds, nighttime) requires further validation. The current architecture focuses on static image detection and cannot utilize temporal information for wildfire spread prediction. Although network pruning enhances inference efficiency, it also leads to a slight accuracy loss.

Looking ahead, this research will expand along three interconnected directions to enhance HybriDet's practical deployment capability and system performance. First, we will develop an integrated compression scheme combining knowledge distillation and quantization-aware training, aiming to achieve a better balance between computational efficiency and detection accuracy, thereby ensuring seamless operation on resource-constrained edge devices. Second, to improve the model's robustness and generalization, we will focus on its compatibility with multi-source remote sensing data, particularly through the effective fusion of thermal infrared and RGB data to overcome the current model's limitations in detecting targets under extreme conditions such as nighttime and heavy smoke occlusion. Simultaneously, the introduction of domain adaptation techniques will aim to enhance the model's generalization across data from different sensors and geographical regions. Finally, moving beyond static image detection, we will develop a spatiotemporal sequence-based prediction module. By leveraging continuous satellite or UAV image sequences, combined with geographic information system data such as digital elevation models and wind speed/direction, we will achieve dynamic fire spread trend prediction and comprehensive risk assessment. This will ultimately advance HybriDet from an efficient detector to a forward-looking intelligent early warning system for ecological protection and disaster response.

6. Conclusions

To address the challenge of wildfire detection in edge computing environments, this study proposes a hybrid neural network architecture named HybriDet for fire object detection. The model integrates the strengths of both CNNs and Transformers to effectively capture both global and local contextual information. Furthermore, several attention mechanisms, including SwinBottle and CS Attention, are introduced to simultaneously emphasize channel-wise and spatial features in high-resolution images, thereby enhancing the accuracy and robustness of the network. Additionally, we have compressed HybriDet using pruning techniques to enhance its generalization performance and speed up its operation. Extensive experiments on the FASDD show that HybriDet has significant advantages in accuracy compared to baseline models with similar parameter sizes. On the FASDD-RS dataset, HybriDet is only 6.45 M in size with a 6.4% higher mAP50 than YOLOv8, making it easy to deploy at the edge with outstanding performance. This study is more effective than many existing deep learning methods, improving detection accuracy and efficiency. Our proposed method enhances the capability for wildfire and smoke detection in resource-constrained environments, effectively improves detection efficiency, and provides technical support for wildfire prevention and control. However, our model is mainly oriented to real-time fire detection tasks, and is currently unable to complete tasks such as predicting future fire situations. In future work, we will explore various model

compression methods like knowledge distillation, quantization, reparameterization, etc., to develop lighter-weight and more efficient wildfire detection algorithms. We will also utilize time-series models for fire risk prediction in order to provide strong support for building safer and sustainable urban and social environments.

Author Contributions: Conceptualization and methodology, F.D. and M.W.; validation, investigation, data curation, writing—original draft preparation, and visualization, F.D.; resources, writing—review and editing, supervision, and project administration, M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Shandong Province Technology Innovation Guidance Plan (Central Guidance for Local Scientific and Technological Development Project: Research and Industrialization of Artificial Intelligence Large Model Service Platform), grant number [YDZX2024088].

Data Availability Statement: This study utilized publicly available datasets. The FASDD can be accessed at https://cstr.cn/31253.11.sciencedb.j00104.00103 (accessed on 16 October 2025). The Pascal Visual Object Classes (VOC) dataset is available at https://www.modelscope.cn/datasets/merve/pascal-voc (accessed on 16 October 2025). No new data were generated in this research. The code has been open sourced in https://github.com/dfm021101/HybriDet (accessed on 16 October 2025).

Acknowledgments: We sincerely thank Wuhan University for releasing the Flame and Smoke Detection Dataset (FASDD). This benchmark dataset provides rich and valuable high-quality sample data for our research. Additionally, we express our gratitude to Ultralytics for the open-source YOLOv8 code, which serves as the baseline and offers a powerful object detection framework. Furthermore, we extend our sincere appreciation to all experts and scholars who participated in the review process. Their insightful comments and suggestions greatly contributed to this study. We also acknowledge the support from the Central Guidance for Local Scientific and Technological Development Project.

Conflicts of Interest: Author Ming Wang was employed by the company Inspur Cloud Information Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Özel, B.; Alam, M.S.; Khan, M.U. Review of Modern Forest Fire Detection Techniques: Innovations in Image Processing and Deep Learning. *Information* **2024**, *15*, 538. [CrossRef]
2. Wu, Q.; Cao, J.; Zhou, C.; Huang, J.; Li, Z.; Cheng, S.-M.; Cheng, J.; Pan, G. Intelligent Smoke Alarm System with Wireless Sensor Network Using ZigBee. *Wireless Commun. Mobile Comput.* **2018**, *2018*, 8235127. [CrossRef]
3. Wang, M.; Yue, P.; Jiang, L.; Yu, D.; Tuo, T.; Li, J. An Open Flame and Smoke Detection Dataset for Deep Learning in Remote Sensing Based Fire Detection. *Geo-Spat. Inf. Sci.* **2024**, *28*, 511–526. [CrossRef]
4. Khan, A.; Hassan, B.; Khan, S.; Ahmed, R.; Abuassba, A. DeepFire: A Novel Dataset and Deep Transfer Learning Benchmark for Forest Fire Detection. *Mobile Inf. Syst.* **2022**, *2022*, 5358359. [CrossRef]
5. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
6. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
7. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.
8. Chen, Y.; Li, J.; Sun, K.; Zhang, Y. A Lightweight Early Forest Fire and Smoke Detection Method. *J. Supercomput.* **2024**, *80*, 9870–9893. [CrossRef]
9. Dalal, S.; Lilhore, U.K.; Radulescu, M.; Simaiya, S.; Jaglan, V.; Sharma, A. A Hybrid LBP-CNN with YOLO-v5-Based Fire and Smoke Detection Model in Various Environmental Conditions for Environmental Sustainability in Smart City. *Environ. Sci. Pollut. Res.* **2024**. [CrossRef] [PubMed]
10. Zhu, W.; Niu, S.; Yue, J.; Zhou, Y. Multiscale Wildfire and Smoke Detection in Complex Drone Forest Environments Based on YOLOv8. *Sci. Rep.* **2025**, *15*, 2399. [CrossRef]

11. Lei, G.; Dong, J.; Jiang, Y.; Tang, L.; Dai, L.; Cheng, D.; Chen, C.; Huang, D.; Peng, T.; Wang, B.; et al. Wildfire Target Detection Algorithms in Transmission Line Corridors Based on Improved YOLOv11_MDS. *Appl. Sci.* **2025**, *15*, 688. [CrossRef]
12. Mohammad Imdadul Alam, G.; Tasnia, N.; Biswas, T.; Hossen, M.J.; Arfin Tanim, S.; Saef Ullah Miah, M. Real-Time Detection of Forest Fires Using FireNet-CNN and Explainable AI Techniques. *IEEE Access* **2025**, *13*, 51150–51181. [CrossRef]
13. El-Madafri, I.; Peña, M.; Olmedo-Torre, N. Real-Time Forest Fire Detection with Lightweight CNN Using Hierarchical Multi-Task Knowledge Distillation. *Fire* **2024**, *7*, 392. [CrossRef]
14. Marjani, M.; Mahdianpari, M.; Mohammadimanesh, F. CNN-BiLSTM: A Novel Deep Learning Model for near-Real-Time Daily Wildfire Spread Prediction. *Remote Sens.* **2024**, *16*, 1467. [CrossRef]
15. Li, R.; Hu, Y.; Li, L.; Guan, R.; Yang, R.; Zhan, J.; Cai, W.; Wang, Y.; Xu, H.; Li, L. SMWE-GFPNNet: A High-Precision and Robust Method for Forest Fire Smoke Detection. *Knowl.-Based Syst.* **2024**, *289*, 111528. [CrossRef]
16. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale 2021. *arXiv* **2020**, arXiv:2010.11929.
17. Cruz, M.V.; Kumar, N.; Subramanian, A.; Sethuraman, S.C. Wildfire Detection Using Vision Transformer(ViT). In Proceedings of the 2025 6th International Conference on Artificial Intelligence, Robotics and Control (AIRC), Savannah, GA, USA, 7–9 May 2025; pp. 402–406.
18. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
19. Chen, J.; Han, H.; Liu, M.; Su, P.; Chen, X. IFS-DETR: A Real-Time Industrial Fire Smoke Detection Algorithm Based on an End-to-End Structured Network. *Measurement* **2025**, *241*, 115660. [CrossRef]
20. Dai, Z.; Liu, H.; Le, Q.V.; Tan, M. CoAtNet: Marrying Convolution and Attention for All Data Sizes. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3965–3977.
21. Lv, W.; Zhao, Y.; Chang, Q.; Huang, K.; Wang, G.; Liu, Y. RT-DETRv2: Improved Baseline with Bag-of-Freebies for Real-Time Detection Transformer 2024. *arXiv* **2024**, arXiv:2407.17140.
22. Wei, Z.; Chen, P.; Yu, X.; Li, G.; Jiao, J.; Han, Z. Semantic-Aware SAM for Point-Prompted Instance Segmentation. *arXiv* **2023**, arXiv:2312.15895.
23. Li, X.; Xu, F.; Liu, F.; Tong, Y.; Lyu, X.; Zhou, J. Semantic Segmentation of Remote Sensing Images by Interactive Representation Refinement and Geometric Prior-Guided Inference. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5400318. [CrossRef]
24. Li, X.; Xu, F.; Yu, A.; Lyu, X.; Gao, H.; Zhou, J. A Frequency Decoupling Network for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5607921. [CrossRef]
25. Li, X.; Xu, F.; Zhang, J.; Yu, A.; Lyu, X.; Gao, H.; Zhou, J. Dual-Domain Decoupled Fusion Network for Semantic Segmentation of Remote Sensing Images. *Inf. Fusion* **2025**, *124*, 103359. [CrossRef]
26. Mukherjee, A.; Mondal, J.; Dey, S. Accelerated Fire Detection and Localization at Edge. *ACM Trans. Embed. Comput. Syst.* **2022**, *21*, 1–27. [CrossRef]
27. Xie, J.; Zhao, H. Forest Fire Object Detection Analysis Based on Knowledge Distillation. *Fire* **2023**, *6*, 446. [CrossRef]
28. Wang, S.; Zhao, J.; Ta, N.; Zhao, X.; Xiao, M.; Wei, H. A Real-Time Deep Learning Forest Fire Monitoring Algorithm Based on an Improved Pruned + KD Model. *J. Real-Time Image Process.* **2021**, *18*, 2319–2329. [CrossRef]
29. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
30. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
31. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
33. Zhang, L.; Wang, M.; Ding, Y.; Bu, X. MS-FRCNN: A Multi-Scale Faster RCNN Model for Small Target Forest Fire Detection. *Forests* **2023**, *14*, 616. [CrossRef]
34. Zhang, Y.; Chen, S.; Wang, W.; Zhang, W.; Zhang, L. Pyramid Attention Based Early Forest Fire Detection Using UAV Imagery. *J. Phys. Conf. Ser.* **2022**, *2363*, 012021. [CrossRef]
35. Chen, G.; Cheng, R.; Lin, X.; Jiao, W.; Bai, D.; Lin, H. LMDFS: A Lightweight Model for Detecting Forest Fire Smoke in UAV Images Based on YOLOv7. *Remote Sens.* **2023**, *15*, 3790. [CrossRef]
36. Jonnalagadda, A.V.; Hashim, H.A. SegNet: A Segmented Deep Learning Based Convolutional Neural Network Approach for Drones Wildfire Detection. *Remote Sens. Appl. Soc. Environ.* **2024**, *34*, 101181. [CrossRef]

37. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
38. Shahid, M.; Hua, K. Fire Detection Using Transformer Network. In Proceedings of the 2021 International Conference on Multimedia Retrieval, Taipei, Taiwan, 6–19 November 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 627–630.
39. Jiang, M.; Zhao, Y.; Yu, F.; Zhou, C.; Peng, T. A Self-Attention Network for Smoke Detection. *Fire Saf. J.* **2022**, *129*, 103547. [CrossRef]
40. Li, A.; Zhao, Y.; Zheng, Z. Novel Recursive BiFPN Combining with Swin Transformer for Wildland Fire Smoke Detection. *Forests* **2022**, *13*, 2032. [CrossRef]
41. Liu, H.; Zhang, F.; Xu, Y.; Wang, J.; Lu, H.; Wei, W.; Zhu, J. TFNet: Transformer-Based Multi-Scale Feature Fusion Forest Fire Image Detection Network. *Fire* **2025**, *8*, 59. [CrossRef]
42. Pan, H.; Badawi, D.; Cetin, A.E. Computationally Efficient Wildfire Detection Method Using a Deep Convolutional Network Pruned via Fourier Analysis. *Sensors* **2020**, *20*, 2891. [CrossRef]
43. de Venâncio, P.V.A.B.; Lisboa, A.C.; Barbosa, A.V. An Automatic Fire Detection System Based on Deep Convolutional Neural Networks for Low-Power, Resource-Constrained Devices. *Neural Comput. Appl.* **2022**, *34*, 15349–15368. [CrossRef]
44. He, L.; Wang, M. SliceSamp: A Promising Downsampling Alternative for Retaining Information in a Neural Network. *Appl. Sci.* **2023**, *13*, 11657. [CrossRef]
45. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLO 2023. Available online: https://docs.ultralytics.com/zh/models/yolov8/ (accessed on 16 October 2025).
46. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
47. Jocher, G. Ultralytics/Yolov5: V3.1—Bug Fixes and Performance Improvements 2020. Available online: https://docs.ultralytics.com/zh/yolov5/quickstart_tutorial/ (accessed on 16 October 2025).
48. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design 2021. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021.
49. Liu, Y.; Shao, Z.; Hoffmann, N. Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions 2021. *arXiv* **2021**, arXiv:2112.05561.
50. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
51. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. *IEEE Trans. Cybern.* **2021**, *52*, 8574–8586. [CrossRef]
52. Bin Jabar, A.H.; Noordin, N.H.; Samad, R. Image Recognition System for Pico Satellite Earth Surface Analysis (50–75 M). In Proceedings of the 2025 IEEE 8th International Conference on Electrical, Control and Computer Engineering (Inecce), Kuantan, Malaysia, 27–28 August 2025; pp. 224–228.
53. Jangirova, S.; Jankovic, B.; Ullah, W.; Khan, L.U.; Guizani, M. Real-Time Aerial Fire Detection on Resource-Constrained Devices Using Knowledge Distillation 2025. *arXiv* **2025**, arXiv:2502.20979.
54. Titu, M.F.S.; Pavel, M.A.; Michael, G.K.O.; Babar, H.; Aman, U.; Khan, R. Real-Time Fire Detection: Integrating Lightweight Deep Learning Models on Drones with Edge Computing. *Drones* **2024**, *8*, 483. [CrossRef]
55. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv* **2025**, arXiv:2502.12524.