remote sensing

MDPI

Article

Forest Fire Segmentation from Aerial Imagery Data Using an Improved Instance Segmentation Model

Zhihao Guan, Xinyu Miao, Yunjie Mu, Quan Sun, Qiaolin Ye and Demin Gao *

College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China; guanzhihao@njfu.edu.cn (Z.G.); miaoxinyu16@njfu.edu.cn (X.M.); muyunjie@njfu.edu.cn (Y.M.); sunquan@njfu.edu.cn (Q.S.); yqlcom@njfu.edu.cn (Q.Y.)

* Correspondence: dmgao@njfu.edu.cn; Tel.: +86-182-0508-4512

Abstract: In recent years, forest-fire monitoring methods represented by deep learning have been developed rapidly. The use of drone technology and optimization of existing models to improve forest-fire recognition accuracy and segmentation quality are of great significance for understanding the spatial distribution of forest fires and protecting forest resources. Due to the spreading and irregular nature of fire, it is extremely tough to detect fire accurately in a complex environment. Based on the aerial imagery dataset FLAME, this paper focuses on the analysis of methods to two deep-learning problems: (1) the video frames are classified as two classes (fire, no-fire) according to the presence or absence of fire. A novel image classification method based on channel domain attention mechanism was developed, which achieved a classification accuracy of 93.65%. (2) We propose a novel instance segmentation method (MaskSU R-CNN) for incipient forest-fire detection and segmentation based on MS R-CNN model. For the optimized model, the MaskIoU branch is reconstructed by a U-shaped network in order to reduce the segmentation error. Experimental results show that the precision of our MaskSU R-CNN reached 91.85%, recall 88.81%, F1-score 90.30%, and mean intersection over union (*mIoU*) 82.31%. Compared with many state-of-the-art segmentation models, our method achieves satisfactory results on forest-fire dataset.

Keywords: fire recognition; instance segmentation; computer vision; deep learning; aerial imagery

Citation: Guan, Z.; Miao, X.; Mu, Y.; Sun, Q.; Ye, Q.; Gao, D. Forest Fire Segmentation from Aerial Imagery Data Using an Improved Instance Segmentation Model. *Remote Sens.* **2022**, *14*, 3159. https://doi.org/10.3390/rs14133159

Academic Editors: Klemen Zakšek, Francesco Marchese, Nicola Genzano and Carolina Filizzola

Received: 26 May 2022
Accepted: 29 June 2022
Published: 1 July 2022

1. Introduction

Forest fires have caused substantial economic losses, air pollution, environmental degradation, and other challenges all over the world, wreaking havoc on human life, animals, and plants [1–3]. According to incomplete statistics, over 200,000 forest fires occur worldwide, destroying approximately 10 million hectares of forest land [4]. Furthermore, the risk of forest fires in China is increasing due to factors such as the wide distribution of forests, the complex topography of forest areas, and the backward monitoring technology [5]. As a result, rapid detection of forest fires can reduce damage to ecosystems and infrastructure. In particular, incipient forest-fire recognition technology can provide forest firefighters with more accurate data on fire behavior, thereby preventing the spread of fires [6].

Traditional forest-fire recognition methods include setting up watchtowers in forest areas for manual monitoring or using infrared instruments carried by helicopters for forest fire detection. These methods are time-consuming, laborious, and generally inefficient in recognition. Thanks to breakthroughs in the field of artificial neural networks, deep neural networks (DNNs) have recently emerged as the most advanced technology in some computer vision challenges [7]. The current success of these types of architectures in very complicated tasks has broadened the scope of their prospective applications and paved the way for their application to real-world problems [8,9]. Nonetheless, recognizing forest fires using aerial imagery is challenging due to fire's various shapes, scopes, and spectral overlaps [10,11].

At the moment, computer vision-based fire-recognition algorithms are primarily concerned with two aspects: (1) flame detection using color and motion features; and (2) feature extraction and classification using deep neural networks [12]. For example, [13] explored the trajectory and irregularity of flames for fire detection in differently colored spaces. [14] separated the chrominance component from luminance using YCbCr and specific rules. This approach detects fires with high accuracy but is only effective for targets with large fire points. [15] tried to detect fires using a multi-channel framework composed of pixels, complex features, and a Bayes classifier. Considering the translucency of smoke, [16] proposed a semantic segmentation model based on a 3D parallel Fully Convolutional Network (FCN) [17] for recognizing smoke regions in videos. The method eliminates background interference to a certain extent and realizes real-time detection. [18] designed several neural networks to recognize forest-fire images with different backgrounds at night and day and focused on analyzing the model's performance under different combinations of network structures. [19] aimed to avoid false alarms from forest fires using recurrent neural networks (RNNs). In contrast to other convolutional neural networks (CNNs), RNN-based models achieve higher fire-detection accuracy, but at the cost of a slower detection speed and higher equipment requirements. At larger scales, satellite technology is routinely utilized to assess forest fires globally, but, usually, at a relatively low resolution, and real-time fire images are limited by satellite trajectory [20,21]. In addition, with the advantages of all-weather, wide field of view, flexibility, and fast deployment, [22] achieved incipient forest-fire monitoring in forest areas using an Unmanned Aerial Vehicle (UAV) equipped with a miniature edge computing platform and visible light lenses.

In specific scenarios, the fire-recognition models mentioned above have obtained good results with respect to accuracy and speed. However, most of the datasets used in these studies are composed of images in the middle and later stages of fires and images taken by ground cameras, characterized by large fires and concentrated fires. Therefore, the recognition effect of incipient forest fires is poor [23–25]. To our knowledge, there are very few aerial imagery datasets for forest-fire recognition, and these are urgently needed in the development of fire-behavior analysis tools for aerial fire forecasting. Note that UAV-based images show different characteristics, such as dynamic blur and top-view perspective, which are fundamentally different from those taken by ground-based cameras [26,27].

In this paper, we designed specific solutions to two computer-vision problems using the FLAME dataset [28] consisting of video frames captured by UAVs: (1) The binary classification of images, i.e., fire and no-fire. In view of the fact that forest fires are characterized by small targets and complex shapes, the module with an attention mechanism is incorporated into the structure of the classification network ResNet [29], aiming to focus on the label-related regions instead of the background; and (2) fire detection using instance segmentation approaches to accurately distinguish the fire individual and pixel information. An improved model based on Mask Scoring R-CNN (MS R-CNN) [30] for forest fire segmentation is proposed.

The main contributions of this paper are summarized as follows:

- We design a novel attention mechanism module, which consists of two independent branches for learning semantic information between different channels to enrich feature representation capability;
- We utilize a U-shaped network to reconstruct the MaskIoU branch of MS R-CNN with the aim of correcting forest-fire edge pixels and reducing segmentation errors; and
- Experimental results show that the proposed MaskSU R-CNN outperforms many existing CNN-based models on forest-fire instance segmentation.

The remainder of this paper is organized as follows. Section 2 introduces experimental materials and methods for two problems, namely, fire classification and fire instance segmentation. Section 3 provides the experimental results and analysis. Section 4 presents the discussion. Finally, Section 5 makes a few concluding remarks.

2. Materials and Methods

This section details the data source and the annotation of the training set, and then two approaches are presented to solve the different challenges. The first challenge is fire and no-fire classification using deep-learning (DL) method. The second challenge is fire instance segmentation, which is complementary to the first problem, and we will further identify fire regions and distinguish pixel classes on images classified as fire-containing.

2.1. Dataset

2.1.1. Data Source

The data obtained through wireless sensor networks and infrared technology has been widely used in the detection, monitoring, and evaluation of forest fires. Furthermore, the aerial superiority of drones allows us to better understand the forest topography structure and the location of the fire [31]. As a result, we selected the FLAME dataset as our data source, which can be obtained from the website (https://ieee-dataport.org/open-access/flame-dataset-aerial-imagery-pile-burn-detection-using-drones-uavs, accessed on 16 December 2021). The FLAME dataset was gathered by drones during the burning of deposits in Arizona pine forests, and it includes video frames and heatmaps taken by infrared cameras, such as the WhiteHot and GreenHot palettes. Figure 1 shows some representative images from fire, no-fire, and thermal videos.

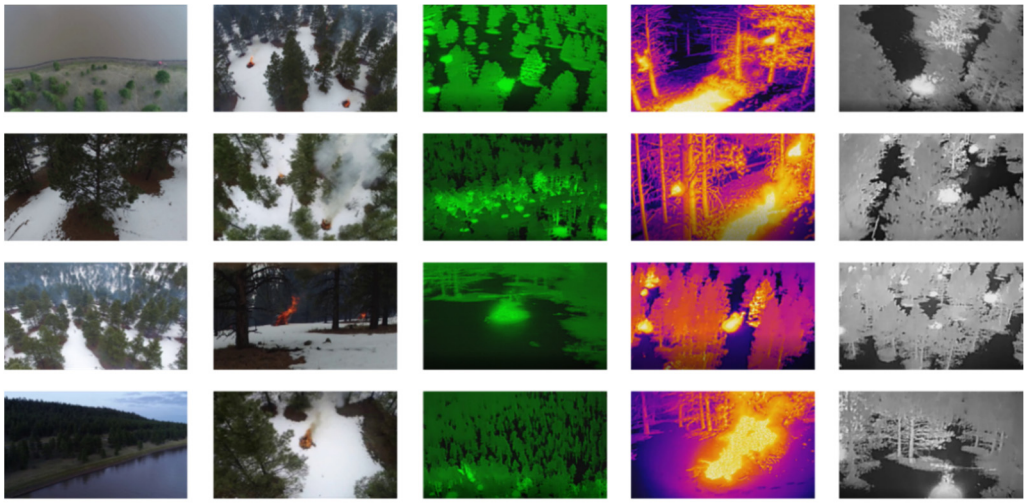

Figure 1. Frame samples of different videos: (**a**,**b**) normal spectrum palette; and (**c**–**e**) heatmaps, including GreenHot, Fusion, and WhiteHot palettes.

2.1.2. Data Collection and Annotation

To meet the requirements of our experiment, four researchers preprocessed the original FLAME dataset, including video-frame extraction and image annotation. The study is separated into two parts: forest-fire classification and case segmentation.

For the image-classification task, we extracted 8000 RGB images from the video, which were divided into two classes (fire and no-fire) and saved as the '.jpg' format with the size of 254 × 254. The dataset consists of three components: the training set (6400, 80%), validation set (800, 10%), and testing set (800, 10%). To observe the different effects of the training set on the model's performance, we employed four different proportions of training set to train the model. The validation and testing set in these experiments continue to use the same images, and both are at a proportion of 10%. The implementation details are recorded in Table 1.

Table 1. Information about the datasets used for training, validation, and testing.

Dataset	Number of Images	Proportion
Training set	1600	20%
	3200	40%
	4800	60%
	6400	80%
Validation	800	10%
Testing set	800	10%
Total	8000	100%

For the instance segmentation task, it can be defined as a pixel-level binary classification problem, in which each pixel is labeled as fire or no-fire (background). To complete the segmentation of the forest fires, the images labeled as fire from Table 1 are considered as a training set. In addition, to train the instance segmentation model and guarantee the quality of the segmentation, we extracted the ground truth of each image through Labelme software. Figure 2 shows the annotation result of a training sample.

(a) (b) (c) (d)

Figure 2. Material of annotation result and ground truth for training: (**a**) raw image; (**b**) annotation of the forest-fire target; (**c**) ground truth label; and (**d**) raw image with the extracted mask.

2.2. Fire Image Classification Using DSA-ResNet

The principle of image classification using deep neural networks is different from traditional digital image processing techniques [14,32,33]. Traditional methods mostly use mathematical modeling or shallow networks for processing and then recognition, which often fail to break the recognition rate bottleneck and have the problem of missed and false detections in practical applications [26]. However, training a CNN to realize this image classification task aids in learning elements unrelated to the fire. Among CNNs, ResNet [29], with many residual blocks, allows for smooth gradient-flow and improves classification accuracy. Furthermore, the attention mechanism provides new momentum for the advancement of CNNs to extract more useful information [34,35]. Experiments show that some attention mechanisms based on channel domain or spatial domain, such as SENet [36] and CBAM [37], can significantly improve network recognition ability. Forest-fire recognition is a challenging task due to the interference of smoke and the translucent nature of the flames. Considering the complexity of forest-fire characteristics, we further exploit multi-scale information based on the SE module to enhance the representation of the model. Unlike previous studies, we propose a novel module using attention mechanism for convolution kernels, which can dynamically select and fuse feature maps from different scales of convolution kernels, termed the Dual Semantic Attention (DSA) module. To be more specific, we implement the DSA module via three operators—Separate, Fuse, and Select, as shown in Figure 3.

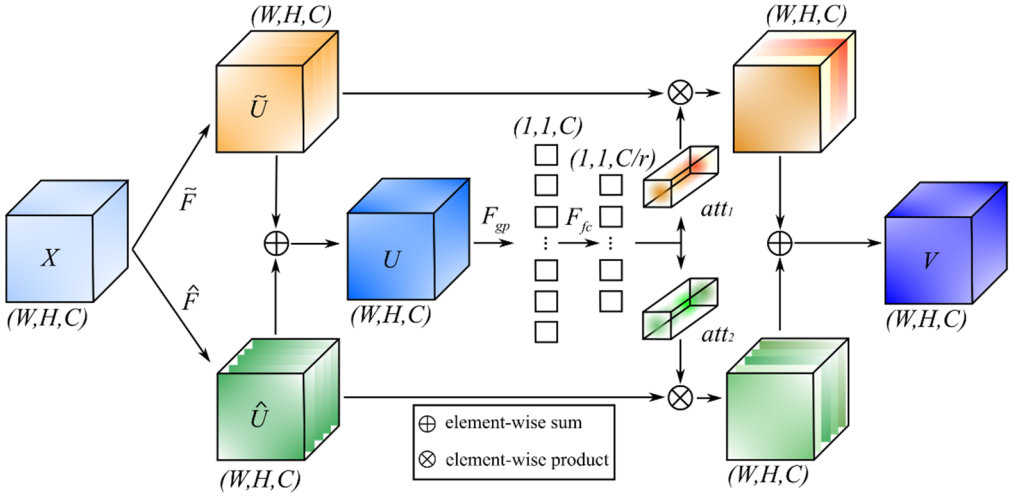

Figure 3. The proposed DSA module.

Separate: An input feature map $X \in R^{H \times W \times C}$ is converted by two transformations with a series of operations, such as grouped convolutions, batch normalization (BN), and Rectified Linear Unit (ReLU) activation function, to achieve the output, denoted as $\widetilde{F} = X \to \widetilde{U} \in R^{H \times W \times C}$ and $\hat{F} = X \to \hat{U} \in R^{H \times W \times C}$. Note that for the purpose of learning the weight relations of different branches, we define convolution kernel size of 3×3 and 5×5 for feature extraction.

Fuse: The purpose of fusion is to learn the channel weights between different feature streams by adaptively adjusting the convolutional kernels (neurons). Firstly, we fuse the output of $\widetilde{F}$ and $\hat{F}$ using element-wise summation:

$$U = \widetilde{U} + \hat{U}, \tag{1}$$

then we obtained the global representation $s \in R^C$ by global average pooling:

$$s_c = F_{gp}(U_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_c(i,j), \tag{2}$$

where U_c denotes the feature map of c-th channel. Furthermore, a fully connected (FC) layer is applied to achieve a compact tensor $z \in R^d$ and reduce computational effort:

$$z = F_{fc}(s) = \delta(BN(Ws)), \tag{3}$$

where δ is the ReLU activation function, $BN(\cdot)$ represents the batch normalization, $W \in R^{\frac{C}{r} \times C}$ is a linear mapping, and r is the descending ratio, which is set to 16 in our experiments.

Select: Two independent FC layers are used to embed the attention information, followed by normalization using softmax function:

$$att_i = softmax(F_{fc}^i(z)), \tag{4}$$

where $att_i \in R^C$ denotes i-th softmax attention, and $i = \{1, 2\}$. Finally, the output of DSA module V is obtained via the channel-based attention weights and their corresponding feature maps, denoted as:

$$V = att_1 \cdot \widetilde{U} + att_2 \cdot \hat{U}. \tag{5}$$

Note that the above formula is implemented in two branches and one can easily derive the case with more branches by extending Equations (1), (4), and (5).

On the basis of ResNet, the above-mentioned DSA module is integrated into the model, and the network structure of DSA-ResNet with 50 layers is shown in Figure 4. Similar

to other CNNs, our DSA-ResNet50 model consists of three primary components: (1) the input feature matrix, (2) the feature extraction layers, and (3) the output layer. During the training phase, the input matrix X is firstly resized to $230 \times 230 \times 3$, which is dependent on the image size and its channels. Then, the data is augmented using random rotation, horizontal flip, and other techniques to improve the generalization ability of the model and to avoid overfitting. The feature extraction layers consist of a large number of convolution blocks (DSA-Residual module), and each block follows identity mapping and a ReLU activation function [38]. The batch normalization helps to accelerate the convergence of the loss function and enables the model to learn different distributions of the data by normalization. The output of the last feature extraction layer is $7 \times 7 \times 2048$, which is later adjusted to $1 \times 1 \times 2048$ by global pooling. Due to the fact that fire classification is a binary classification problem, we use the sigmoid activation function to output its probabilities (fire, no-fire), denoted as:

$$P(fire) = \sigma(fire|\varphi(\theta)) = \frac{1}{1 + e^{-\varphi(\theta)}}, \tag{6}$$

where $\varphi(\theta)$ represents the output of the FC layer, which is obtained using the input matrix X, pixel values for each channel, and all weights across the entire feature extraction layer, and θ is the weight for the last layer. The output is the probability of fire-recognition with a threshold set to 0.5. To train our DSA-ResNet50 model, a loss function is used to improve network accuracy and find the best weight matrix, which is defined as a binary cross-entropy:

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \cdot logp(\hat{y}_i) + (1 - y_i) \cdot log(1 - p(\hat{y}_i))), \tag{7}$$

where N represents the total number of samples used for each epoch, y is the ground truth label for each image labeled as fire $(y = 1)$ or no-fire $(y = 0)$ in the training phase, and $p(\hat{y})$ represents the predicted result of an image classified as the fire class. In addition, training is carried out by Adam optimizer [39] to update gradient flow of the network, with the L_2 regularization set to 1×10^{-4}.

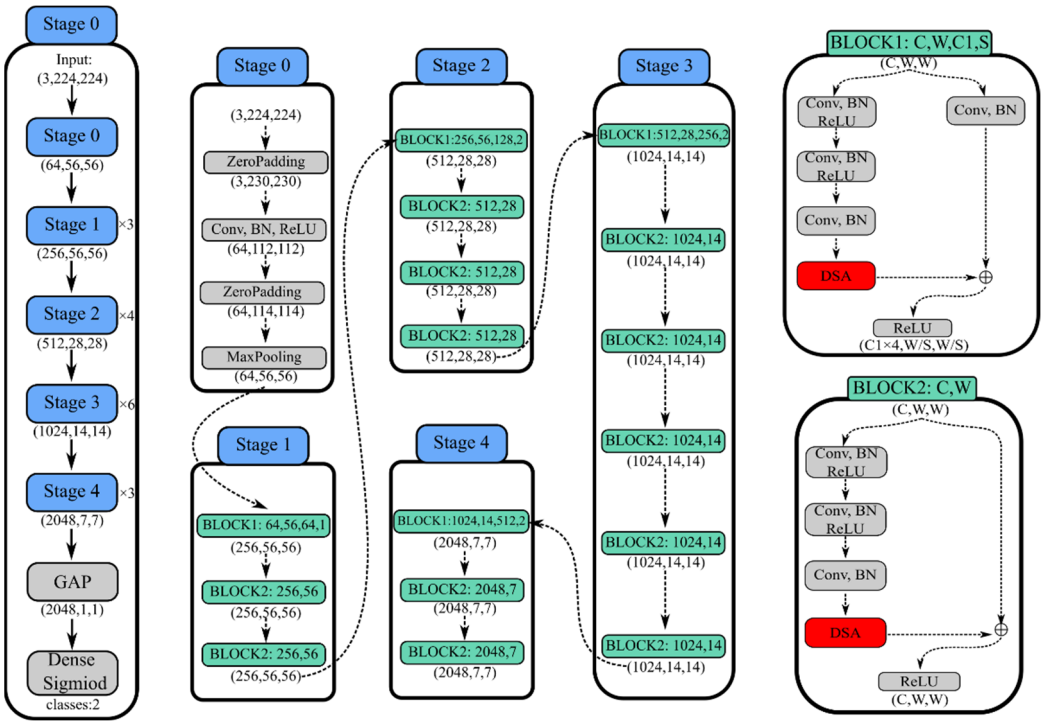

Figure 4. Our proposed DSA-ResNet50 model for fire classification.

2.3. Fire Instance Segementation Using MaskSU R-CNN

As an improvement of Mask R-CNN [40], MS R-CNN is the most advanced instance segmentation method at present. Mask R-CNN regards classification confidence as a criterion for segmentation quality. However, experimental evidence demonstrates that there is no significant correlation between predicted mask quality and classification confidence. To solve this problem, MS R-CNN is obtained by adding a MaskIoU branch to the Mask R-CNN, which is employed to learn and predict the segmentation results, i.e., the segmentation confidence. The MaskIoU branch utilizes related feature map and the mask branch's result as input ($14 \times 14 \times 257$) and then calculates the Intersection Over Union (*IoU*) between the predicted mask and its corresponding ground truth label in the image. Consequently, MS R-CNN is capable of achieving more precise segmentation results than Mask R-CNN.

UNet [41] is a semantic segmentation model based on FCN, which mainly includes three parts: down-sampling, up-sampling, and concatenation operation. Down-sampling consists of convolution-pooling blocks and is used to compress the number of channels. Each block has two convolutional layers, one pooling layer, and ReLU activation function. In addition, up-sampling doubles the size of the feature map and halves the number of channels by transposed convolution. After that, the output is connected with the feature map with the same size obtained by down-sampling. Throughout the entire process, the concatenation of feature map helps to integrate information in both shallow and deep networks.

Inspired by the UNet structure, we reconstructed the MS R-CNN's MaskIoU branch using a U-shaped network called MaskSU R-CNN in this paper, and the model structure is presented in Figure 5.

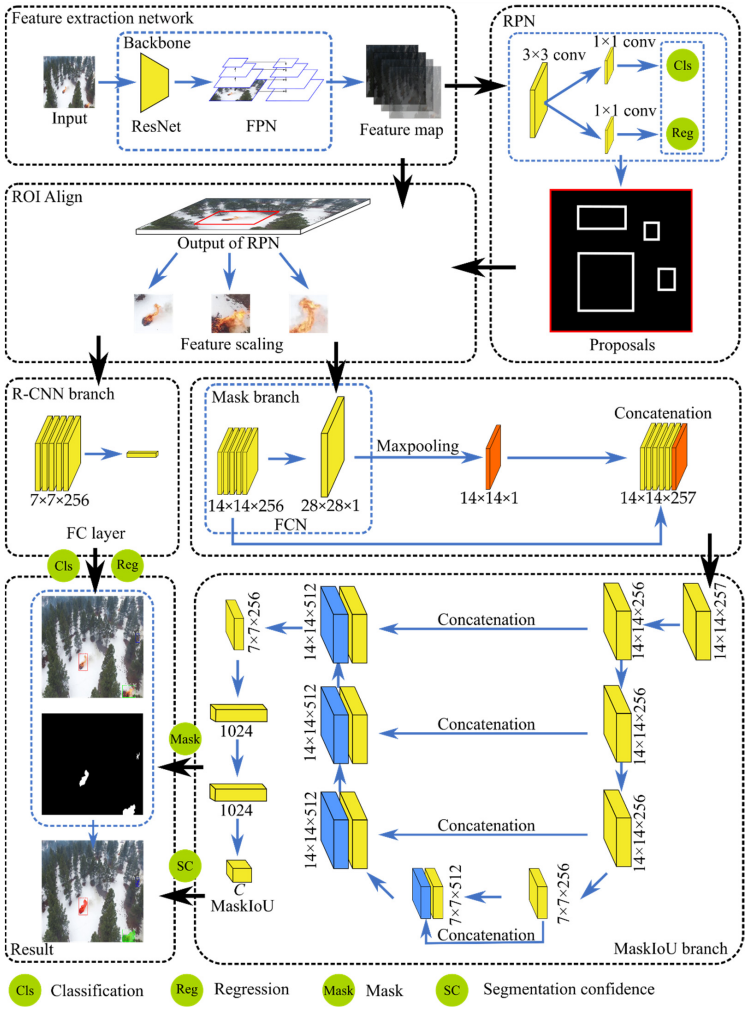

Figure 5. Our proposed MaskSU R-CNN model for fire segmentation.

2.3.1. Feature Extraction Network

Given the outstanding performance of the attention mechanism on the image classification task, we adopt DSA-ResNet50 as the backbone of our MaskSU R-CNN. In addition, the introduction of the feature pyramid network (FPN) [42] for multi-scale fusion contributes to extracting more effective features.

2.3.2. Region Proposal Network (RPN) and Region of Interest (RoI) Align

The RPN [43] is composed of two convolution blocks (3×3, 1×1): The 3×3 convolution block extracts features from the output of the backbone network using 32 convolution kernels; the 1×1 convolution block is used to adjust the parameters of each anchor box and determine whether there is an object in it. During the training phase, the RPN generates nine anchor boxes for each pixel on the feature map, and then a series of initial proposal regions are screened using Non-Maximum Suppression (NMS).

Unlike the RoI Pooling in the work of [43], RoI Align [40] is used here to scale the RPN output to the same size. RoI Align utilizes bilinear interpolation to calculate grid point coordinates, which effectively preserves the edge pixels of the object to obtain predicted masks with high quality.

2.3.3. Multi-branch Prediction for Classes, Bounding Boxes, and Masks

The multi-branch prediction network contains three branches: the R-CNN branch for classification and bounding-box regression, the mask branch for generating predicted masks, and the MaskIoU branch for segmentation evaluation. The R-CNN branch is consistent with most object detection methods, using the softmax function for classification and $Smooth_{L1}$ loss for bounding-box regression.

In Mask R-CNN model, the segmentation quality is equivalent to the classification confidence, which is unscientific in practical situations. To address this problem, MS R-CNN introduces the MaskIoU branch, which evaluates the quality of segmentation by concatenating the feature map ($14 \times 14 \times 256$) with the output of the mask branch ($14 \times 14 \times 1$). During the process of convolution, the features in the shallow network will cause a certain loss. To better perform feature fusion and reduce feature loss, a U-shaped network is adopted in this paper to reconstruct the MaskIoU branch. The novel MaskIoU branch (Table 2) consists of eight convolutional layers, including down-sampling for channel compression, up-sampling for feature expansion, and concatenation operation. Finally, the *IoU* value between the predicted mask and corresponding ground truth is calculated by three FC layers. The feature concatenation in the U-shaped network integrates the information between different feature maps, which is significant for segmenting the edge pixels of fire.

Table 2. The details of the proposed MaskIoU branch with U-shaped network. *C* represents the number of classes in the dataset.

	Operation	Kernel/Stride	Output
Block1	Conv + ReLU	$3 \times 3 \times 256$	$14 \times 14 \times 256$
	Conv + ReLU	$3 \times 3 \times 256$	$14 \times 14 \times 256$
	Conv + ReLU	$3 \times 3 \times 256$	$14 \times 14 \times 256$
Block2	Maxpooling	2×2	$7 \times 7 \times 256$
	Conv + ReLU + Concat	$3 \times 3 \times 256$	$7 \times 7 \times 512$
Block3	Up-sampling	2×2	$14 \times 14 \times 512$
	Conv + ReLU + Concat	$3 \times 3 \times 256$	$14 \times 14 \times 512$
	Conv + ReLU + Concat	$3 \times 3 \times 256$	$14 \times 14 \times 512$
	Conv + ReLU + Concat	$3 \times 3 \times 256$	$14 \times 14 \times 512$
Block4	Conv + ReLU	$3 \times 3 \times 256$	$14 \times 14 \times 512$
	Maxpooling	2×2	$7 \times 7 \times 256$
Block5	FC + ReLU	/	1024
	FC + ReLU	/	1024
	FC + ReLU	/	*C* (MaskIoU)

2.3.4. Model Training and Loss Function

As a large image dataset in deep learning, COCO [44] has more than 220 k images with 80 categories. Before training our MaskSU R-CNN model, a COCO-based pretrained model was applied using transfer learning [45] to train deep neural networks to be more stable and efficient. During training of the model, the method used in the work of [40] has been adopted. Thirty-two anchors are randomly selected from each image in the batch, and the loss is generated based on the positional relationship with the ground truth label. If the *IoU* value is greater than 0.5, the RoI is considered as a positive sample; otherwise, it is a negative sample, and the ratio of positives and negatives is 1:3.

For generating the regression target in MaskIoU branch, the predicted mask is binarized with a threshold of firstly 0.5. Then we regress the MaskIoU between the predicted mask and its corresponding ground truth by l_2 loss. In addition, the mask loss L_{mask} and the MaskIoU loss $L_{maskiou}$ are defined on positive samples only.

During the training phase, the curve of loss function visually reflects the convergence of the model. The total loss of MaskSU R-CNN is composed of two components: (1) the loss in RPN; and (2) the loss generated by the multi-branch prediction network, which can be described as:

$$L_{total} = L_{rpn} + L_{mul-branch}, \tag{8}$$

where the RPN loss L_{rpn} is composed of the classification loss (softmax loss) and the bounding-box regression loss ($Smooth_{L1}$ loss), which is used to generate many proposals (the output of RPN), including the identification of whether or not there are real objects in the anchor and the parameter adjustment of the anchor position. L_{rpn} is computed as follows:

$$L_{rpn} = \frac{1}{N_{cls1}}\sum_i L_{cls}(p_i, p_i^*) + \lambda_1 \frac{1}{N_{reg1}}\sum_i p_i^* L_{reg}(t_i, t_i^*), \tag{9}$$

where $L_{mul-branch}$ is generated by different branches, including the classification loss (softmax loss), the bounding-box regression loss ($Smooth_{L1}$ loss), the mask loss L_{mask} and the MaskIoU loss $L_{maskiou}$. The formula is expressed as follows:

$$\begin{array}{l} L_{mul-branch} = L(p_i, p_i^*, t_i, t_i^*, s_i, s_i^*) \\ = \frac{1}{N_{cls2}}\sum_i L_{cls}(p_i, p_i^*) + \lambda_2 \frac{1}{N_{reg2}}\sum_i p_i^* L_{reg}(t_i, t_i^*) \\ +\gamma_2 \frac{1}{N_{mask}}\sum_i L_{mask}(s_i, s_i^*) + \frac{1}{2N_{mask}}\sum_i L_{maskiou}(s_i, s_i^*) \end{array} \tag{10}$$

where the classification term is normalized by the mini-batch size (i.e., N_{cls} = 256) and regression term is normalized by the number of anchor locations (i.e., N_{reg}~2400); λ_* and γ_* are hyperparameters used to balance the loss of anchors or bounding-boxes regression and the loss of mask generation during the training phase, which are set to 10 and 1 in our implementation. The classification loss L_{cls}, regression loss L_{reg}, mask loss L_{mask}, and MaskIoU loss $L_{maskiou}$ are listed as followed:

$$L_{cls}(p_i, p_i^*) = -log p^* p, \tag{11}$$

$$\begin{array}{l} L_{reg}(t_i, t_i^*) = \sum_i smooth_{L1}(t_i - t_i^*) \\ where\ smooth_{L1}(x) = \begin{cases} 0.5x^2 & if\ |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \end{array} \tag{12}$$

$$L_{mask}(s_i, s_i^*) = -(s_i^* \odot log^\circ(s_i) \oplus (1 - s_i^*) \odot log^\circ(1 - s_i)), \tag{13}$$

$$L_{maskiou}(s_i, s_i^*) = \left(\frac{s_i \cap s_i^*}{s_i \cup s_i^*}\right)^{-2}, \tag{14}$$

where p_i represents the probability that the predicted result of anchor i is the ground truth. Since RPN is used to detect the presence of the target (foreground or background) instead

of classification, the value of p_i^* is 1 when anchor i is a positive sample, otherwise it is 0. $t_i = (t_i^x, t_i^y, t_i^w, t_i^h)$ represents the regression parameters of anchor i, including the center coordinates of the bounding box (x, y), the width w, and height h. t_i^* indicates the ground truth corresponding to anchor i. s and s^* represent the binary matrix of the predicted mask and ground truth, respectively. $\odot$, $\oplus$, and $\log^\circ$ denote the pixel-based product, summation, and logarithm, respectively.

Finally, the segmentation quality of each target can be expressed by mask score S_{mask}:

$$S_{mask} = S_{cls} \cdot S_{maskiou}. \tag{15}$$

where S_{cls} represent the classification confidence obtained from R-CNN branch, and $S_{maskiou}$ is the output of MaskIoU branch.

3. Results

This section presents the different performances of two deep neural network models on forest-fire image classification and segmentation. All the experiments are based on Python 3.6 and Pytorch using the Windows system. The hardware used is AMD R7-5800H and an NVIDIA RTX 3070 with 16 GB memory.

3.1. Fire Image Classification

3.1.1. Accuracy Assessment

To observe the performance of the model at different training-set proportions, we compared five existing deep-classification networks (VGGNet [46], GoogleNet [47], ResNet [29], and SE-ResNet50 [36]) with our DSA-ResNet50. The performance of these models trained with different proportions (20%, 40%, 60%, and 80%) of training images is shown in Figure 6. We evaluate the classification performance using four metrics: accuracy (*Acc*), Kappa coefficient (*K*), Omission Error (*OE*), and Commission Error (*CE*). According to Figure 6, it can be noticed that with the increase of training sets, the *Acc* and *K* keep growing while the *OE* and *CE* decrease. This is because as the training set increases, the model can learn more relevant features. In addition, *OE* is slightly higher than *CE*, which indicates that some of the fine fire points are highly similar to the surrounding soil or obscured by vegetation, making it difficult for the model to classify them accurately. It is worth noting that our DSA-ResNet50 is superior to other models under different proportions of training images. The calculation formulas are as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{16}$$

$$K = \frac{p_o - p_e}{1 - p_e}, \tag{17}$$

$$CE = \frac{FP}{TP + FP}, \tag{18}$$

$$OE = \frac{FN}{TP + FN}. \tag{19}$$

where TP and FP represent the number of fire or no-fire images classified as fire label, respectively; FN and TN represent the number of fire or no-fire images classified as no-fire label, respectively; p_o is the overall classification accuracy, and p_e is the accidental consistency.

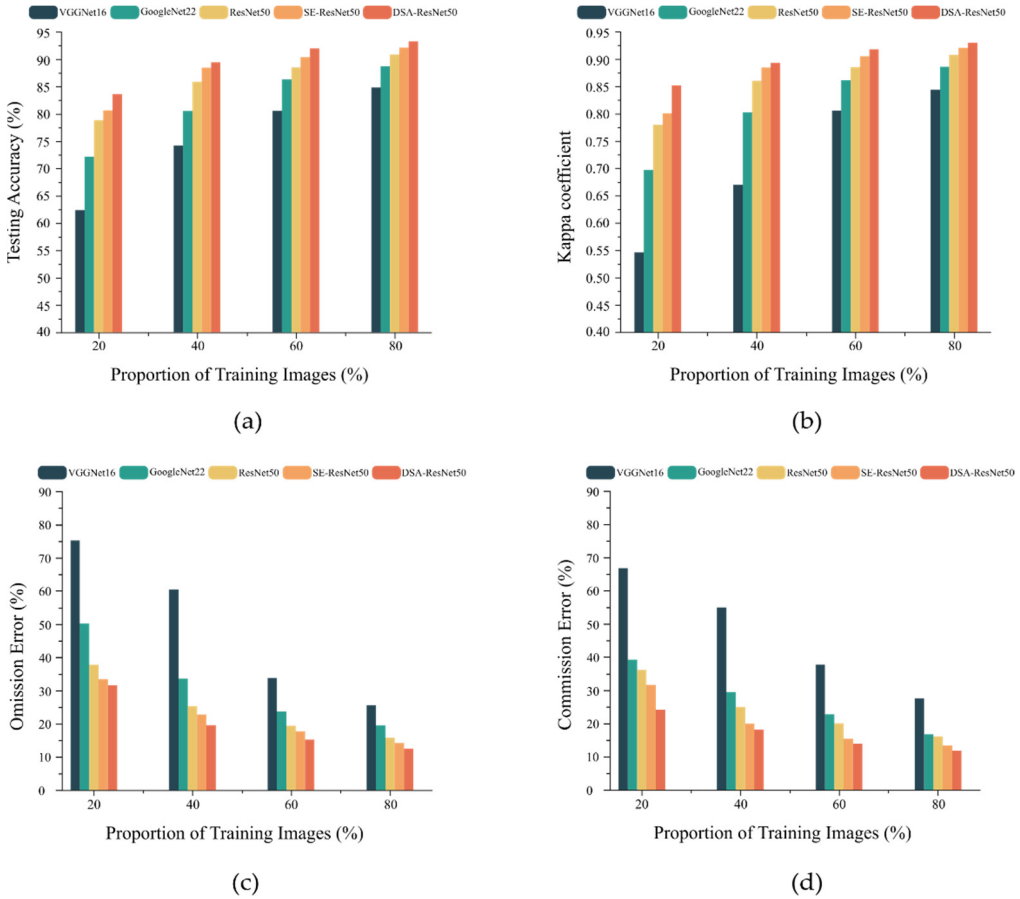

Figure 6. Performances of the existing models (VGGNet16, GoogleNet22, ResNet50, and SE-ResNet50), and DSA-ResNet50 trained with four proportions (20%, 40%, 60%, and 80%) of training images. (**a**) Testing accuracy; (**b**) kappa coefficient; (**c**) omission error; and (**d**) commission error.

Table 3 demonstrates the results of comparison models trained with 80% images, where our DSA-ResNet50 performs the best (*Acc* = 93.65%, *K* = 0.864, *OE* = 20.59%, and *CE* = 4.23%). In addition, with a slight increase in network parameters (1.8 million), the addition of the DSA module increased *Acc* and *K* by 2.37%, 0.025, and decreased *OE* and *CE* by 9.28%, 4.12%, respectively, suggesting that the proposed attention mechanism can capture more features and thus improve the classification ability of the model.

Table 3. Performances of different models on testing set using 60% training images.

Model	Layers	*Acc* (%)	*K*	*OE* (%)	*CE* (%)	Params (Million)
VGGNet	16	84.86	0.743	37.54	16.84	138.53
GoogleNet	22	88.23	0.784	34.52	11.61	8.97
ResNet	50	91.28	0.839	29.87	8.35	26.85
SE-ResNet	50	92.46	0.851	25.62	5.62	28.65
DSA-ResNet (ours)	50	93.65	0.864	20.59	4.23	28.43

3.1.2. Visualization Analysis

To better understand CNN's decision on image classification, the visualization method Gradient-weighted Class Activation Mapping (Grad-CAM) [48] was used to generate a heatmap for evaluating important regions in each input image. Figure 7 shows the Grad-CAM visualizations of forest-fire images taken from different UAV angles (shown in the first and third columns), based on DSA-ResNet50 model trained with 80% training images. According to each input image and its corresponding visualization, it can be seen that DSA-

ResNet50 can easily focus on areas with fire points (marked with red boxes), indicating that the DSA module enhances the network's ability to recognize fire areas.

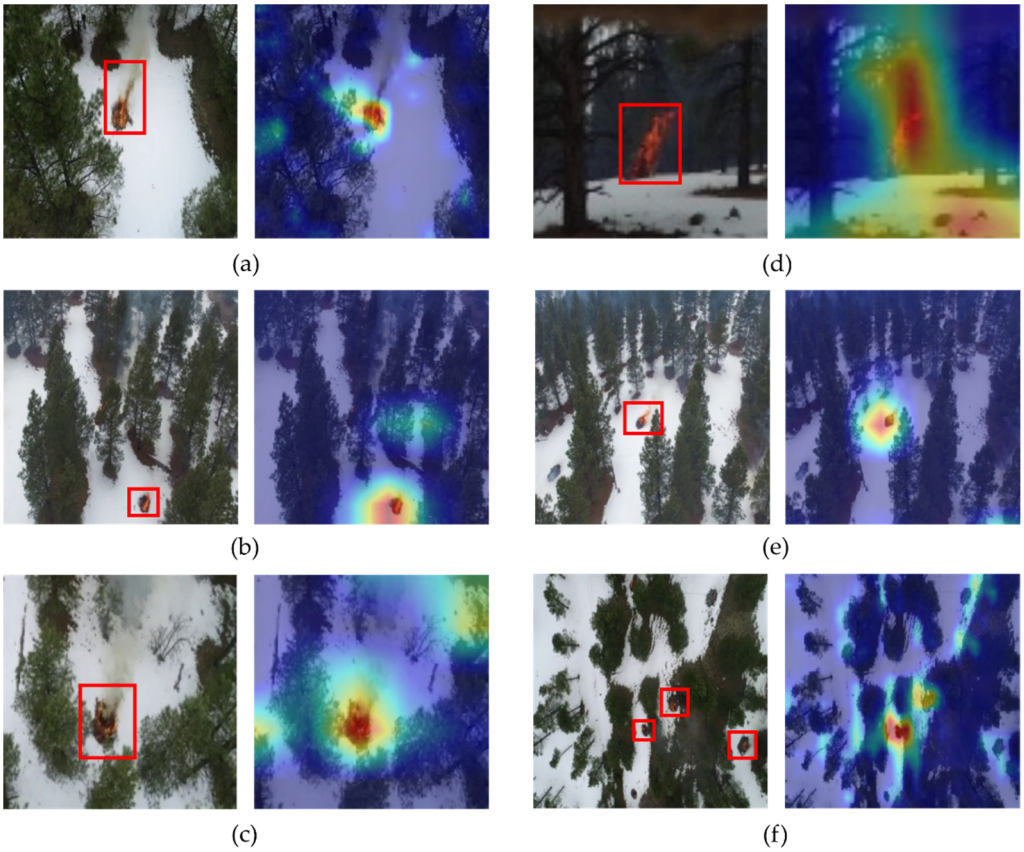

Figure 7. Grad-CAM visualization results on DSA-ResNet50. (**a**–**f**) show the different testing images and their corresponding Grad-CAM results, with the important regions highlighted in red.

3.2. Fire Detection and Segmentation

3.2.1. Evaluation Metrics

The accuracy of segmentation results is often measured by *IoU*, which represents the overlap rate between the predicted result and its corresponding ground truth label, and the closer the value is to 1, the better the segmentation performance is. To be fair, we adopt the mean value of *IoU*s on the testing set to measure the model, denoted as:

$$mIoU = \frac{1}{N}\sum_{i=1}^{N}\frac{P_i \cap G_i}{P_i \cup G_i}, \tag{20}$$

where P_i and G_i denote the predicted result and corresponding ground truth label for i-th image, respectively. In our experiment, if the *IoU* is 0.5 or above, the target is considered a positive sample, otherwise negative. In addition, the F1-score is used as another evaluation metric, denoted as:

$$f = \frac{2TP}{2TP + FP + FN}, \tag{21}$$

where TP, FP, and FN are defined in Section 3.1.1. Obviously, the larger the value of f, the better the accuracy of the model.

3.2.2. Performance Analysis and Comparison

The proposed MaskSU R-CNN is an instance segmentation model that implements parallel processing for object detection and segmentation. From the experimental results (Figure 8b), it can be found that our model can correctly identify forest-fire targets and

achieve good segmentation results. Compared with the ground truth labels (Figure 8c), they remain almost the same except for some defects in flame details, and the minor differences are mainly caused by the translucency of forest fires and the interference of occlusions. Figure 9 demonstrates the loss curves over 120 epochs for both the training and validation sets. It can be seen that our model shows an overall smooth decreasing trend and gradually converges after about 80 epochs.

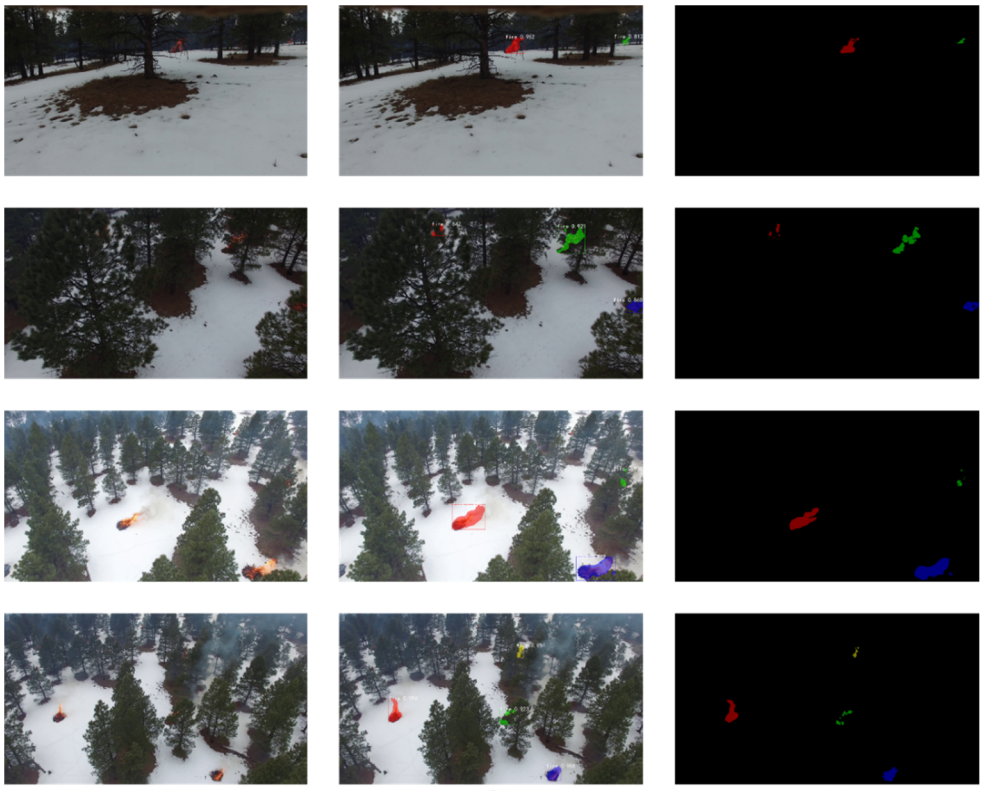

Figure 8. Results of MaskSU R-CNN for forest-fire instance segmentation: (**a**) raw images; (**b**) predicted results; and (**c**) ground truth labels.

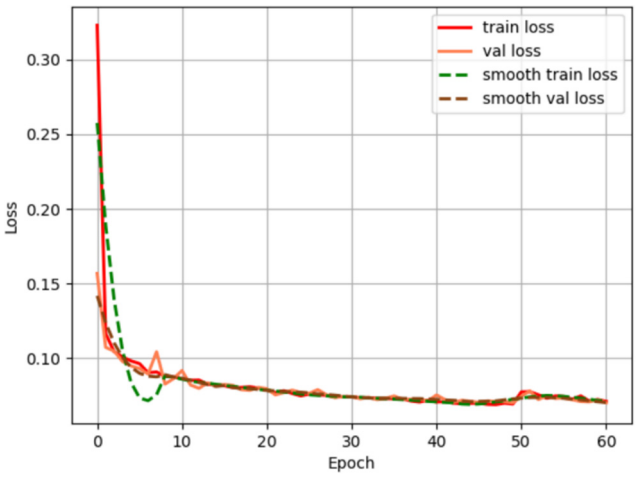

Figure 9. Loss curves of MaskSU R-CNN.

To demonstrate the superiority of MaskSU R-CNN on forest-fire segmentation, we compared our method with several DL-based semantic segmentation models, including SegNet [49], UNet [41], PSPNet [50], and DeepLabv3 [51]. Noticeably, the same dataset and configurations were used to train all models to make the predictions comparable.

We selected a representative part of the images from the testing set for display, and the predicted results are shown in Figure 10. We can visually see that the segmentation results of our MaskSU R-CNN outperform other comparison models, especially on images that are hard for humans to recognize. In terms of apparent forest-fire targets that are significantly different from the background, most methods produced relatively accurate segmentation results. As for those forest fires with small targets and high concealment, as marked with green boxes in Figure 10a, most models generated some degree of under-segmentation, except for DeepLabv3 and our MaskSU R-CNN. It is worth noting that SegNet showed serious mis-segmentation (marked with blue boxes), which was mainly caused by the model without taking full advantage of contextual semantic relationships. Table 4 lists the results of the quantitative analysis with different comparison models. Our model obtained the highest f and $mIoU$ on the testing set. Moreover, unlike the above segmentation methods, our model also achieves the differentiation of individual forest fires, which makes the fire segmentation more interpretable.

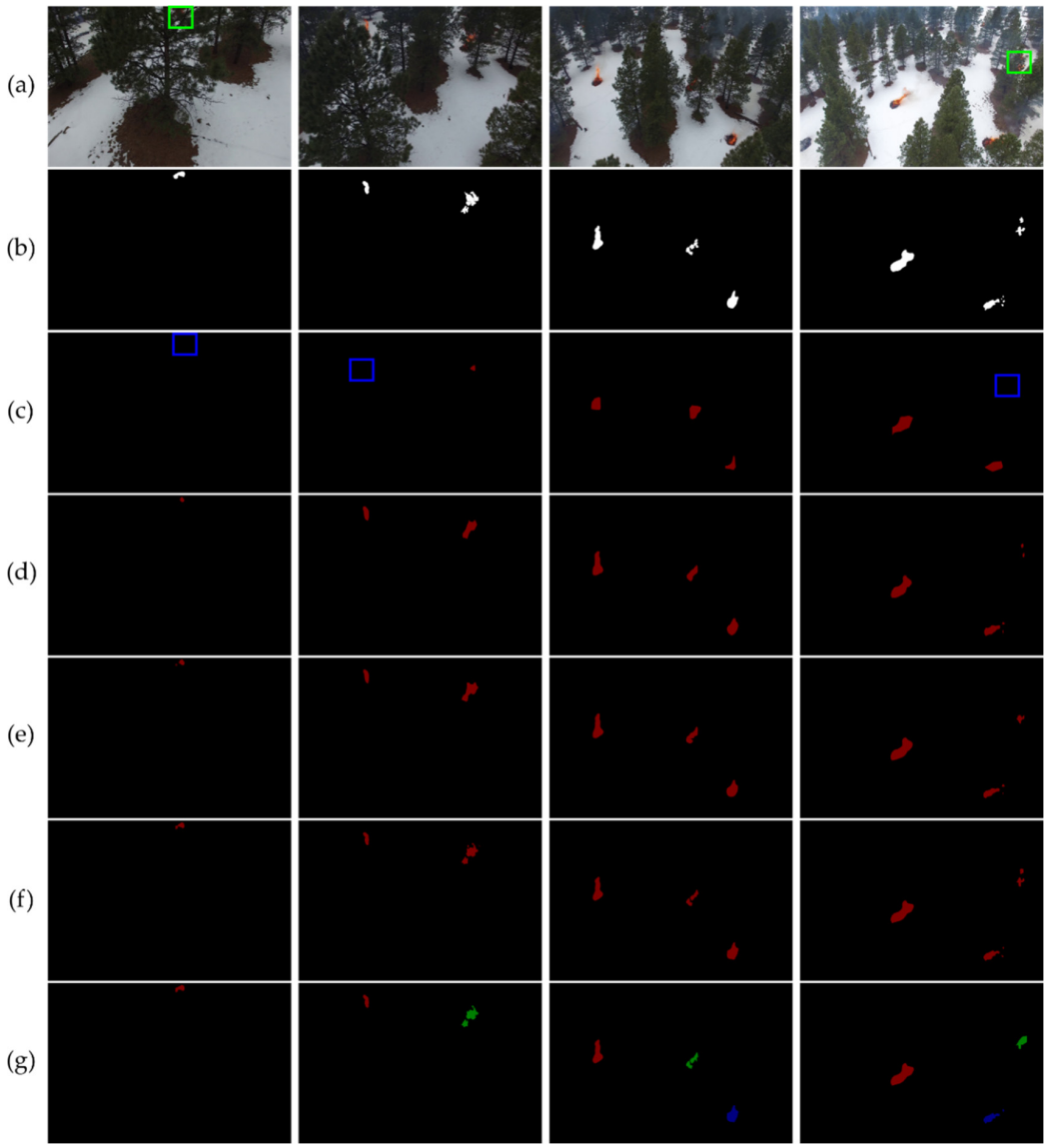

Figure 10. Results on testing images. (**a**) Raw images and (**b**) ground truth labels. Segmented results by (**c**) SegNet, (**d**) UNet, (**e**) PSPNet, (**f**) DeepLabv3, and (**g**) our MaskSU R-CNN.

Table 4. Comparison results between different segmentation models.

Method	Metrics			
	p **(%)**	***r*** **(%)**	***f*** **(%)**	***mIoU*** **(%)**
SegNet	71.37	41.33	52.35	35.45
UNet	86.18	85.96	86.07	77.85
PSPNet	83.12	81.25	82.17	69.74
DeepLabv3	90.95	89.64	90.29	81.12
MaskSU R-CNN (ours)	91.85	88.81	90.30	82.31

Furthermore, in order to verify the effectiveness of our improved model MaskSU R-CNN, we compared it with the original model Mask R-CNN and MS R-CNN, and some of the segmentation results are shown in Figure 11. It can be intuitively found that our MaskSU R-CNN achieves the best segmentation results, followed by the MS R-CNN. The main reason is that these two models both add a new branch MaskIoU on the basis of Mask R-CNN, and we further improve the quality of the predicted mask after reconstructing the MaskIoU branch using a U-shaped network. Therefore, the segmentation results are the most excellent. In particular, our model has a remarkable advantage in the correction of edge pixels in the fire regions, especially on inconspicuous fire images, such as fire points 1, 2, 6, and 8 in Figure 11e. As for those highly occluded fire targets in Figure 12, the segmentation confidence of our method is more reasonable, which is determined by the segmentation quality, rather than directly using the classification confidence. Meanwhile, the novel branch MaskIoU greatly improves the fine-grained characterization capability of our model.

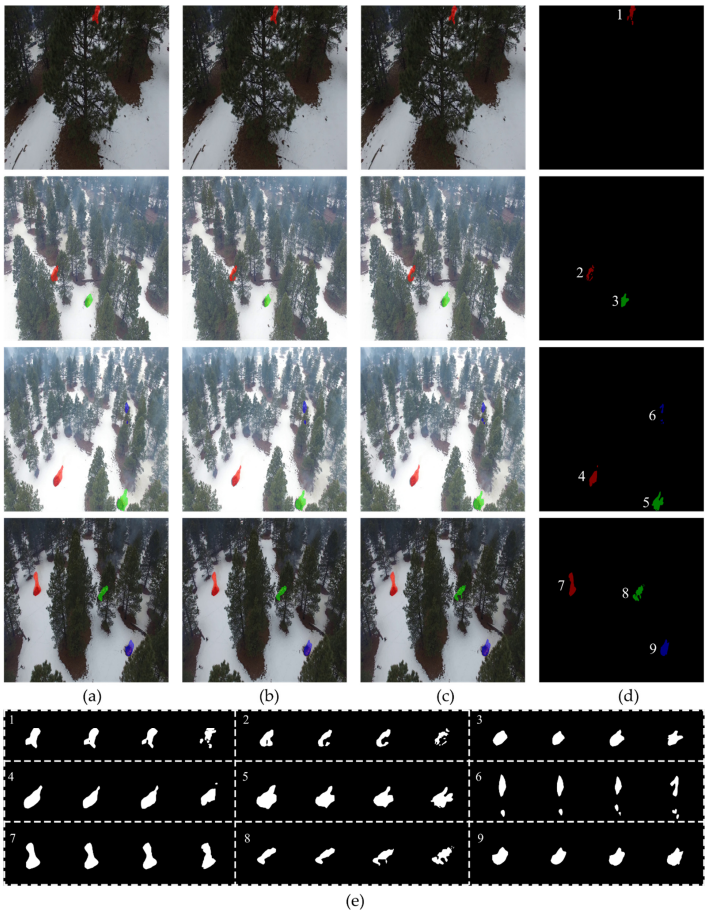

Figure 11. Predicted results of three different instance segmentation models: (**a**) mask R-CNN, (**b**) MS R-CNN, (**c**) our MaskSU R-CNN; (**d**) ground truth labels; and (**e**) segmentation details. Note that numbers 1 to 9 represent the order of fire points, and from left to right are the predicted results generated by Mask R-CNN, MS R-CNN, MaskSU R-CNN, and the corresponding ground truth label.

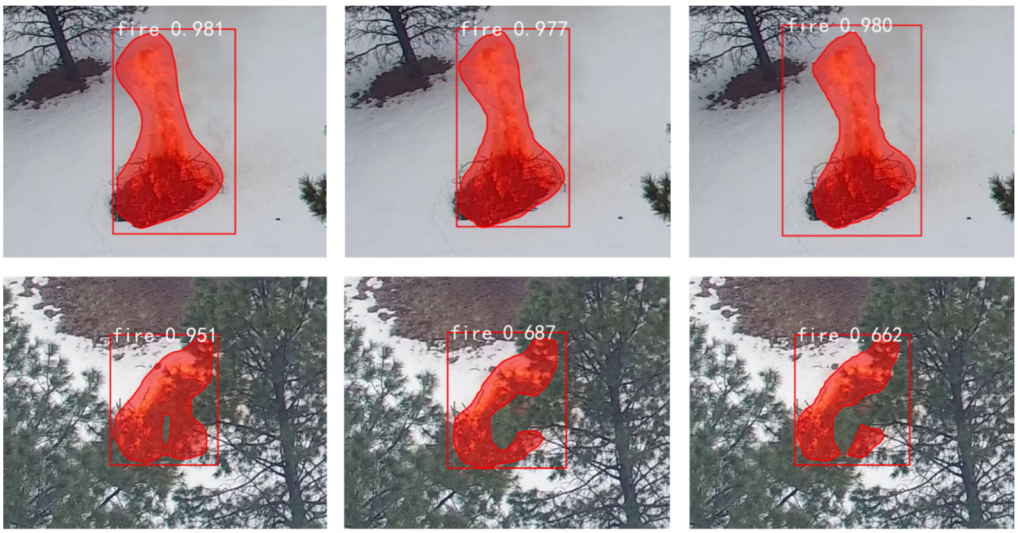

Figure 12. Local visualization and segmentation confidence: (**a**) mask R-CNN, (**b**) MS R-CNN, and (**c**) our MaskSU R-CNN.

To demonstrate the importance of the novel MaskIoU branch in our model, we conducted a series of ablation experiments. According to the ablation results in Table 5, it can be found that the MaskIoU branch can significantly improve the segmentation quality of forest fires (*mIoU*). In particular, after adding the novel MaskIoU branch with the U-shaped network, our *mIoU* reached 80.77%. Meanwhile, introducing the attention mechanism (DSA module) to the backbone ResNet can further mine the intrinsic information of the features and improve the performance of the model.

Table 5. Ablation experiments between different instance segmentation variants.

Model	Backbone	MaskIoU	Metrics			
			p (%)	*r* (%)	*f* (%)	*mIoU* (%)
Mask R-CNN	ResNet	/	85.62	82.61	84.09	75.97
	DSA-ResNet	/	87.94	85.69	86.80	77.55
MS R-CNN	ResNet	FCN	88.95	83.16	85.96	78.61
	DSA-ResNet	FCN	90.15	87.94	89.03	80.42
MaskSU R-CNN (ours)	ResNet	U-shaped network	88.63	88.89	88.76	80.77
	DSA-ResNet	U-shaped network	91.85	88.81	90.30	82.31

In addition to evaluating the segmentation performance of comparative methods, we also calculated the model size and running time. Our novel MaskIoU branch has about 0.63 G FLOPs compared with 0.39 G FLOPs in MS R-CNN. We use one 3070 GPU to test assess the running time (sec./frame). As for DSA-ResNet50, the speed is about 0.235 for Mask R-CNN, and 0.238 for both MS R-CNN and MaskSU R-CNN. Hence, the computation cost of MaskIoU branch is negligible.

4. Discussion

In contrast to other fixed-form objects, forest fires are dynamic objects with variable shapes and hard-to-depict textures [52]. Generally, a forest fire usually begins as a small-scale fire, develops into a medium-scale fire, and then becomes a large-scale fire. Typologically, it starts from ground fire, then spreads to the trunk, and finally to the tree crown [53,54]. Therefore, the detection of incipient fires appear to be particularly important. Unfortunately, there is little research on this aspect, and most of their research data sets are

images of medium-scale or big-scale fires. We focused on this phenomenon and adopted the forest-fire data set based on UAV aerial photography, with minor fire points and strong flame concealment. This can have a high degree of simulating incipient fires. This study proposed a novel method by improving the existing instance segmentation model in order to provide more accurate fire-behavior data, which deeply explores the shallow information and deep higher-order semantics in image features and achieves high-precision recognition of incipient forest fires.

In terms of forest-fire recognition, previous methods have advantages in detecting fires with a faster speed and higher accuracy [55]. However, difficulties arise when complications occur, such as when the capture of fires from a drone's perspective increases the misdetection rate, and inconspicuous fire points with a small target or high concealment are not easily discovered. To address the issues above, we reconstructed the MaskIoU branch of existing MS R-CNN model by adding a U-shaped network. Specifically, the improved branch cascades feature maps of the same size during encoding and decoding phase, allowing for better integration of pixel location features in the shallow network and pixel category features in the deep network, which provides some correction for edge pixels of forest fire targets.

In order to fully illustrate the rationality of the model in this paper, our MaskSU R-CNN is compared with the original models Mask R-CNN and Mask Scoring R-CNN from several perspectives. The convergence comparison in Figure 9 shows that the overall training loss of our method is slightly lower than the other two models with the same training samples. The visualization comparison in Figure 11 reveals that the segmentation mask of our method has the highest matching degree with the actual shape of the forest fire and has obvious advantages in processing forest-fire edge pixels. The quantitative comparison in Table 5 shows that our method achieves SOTA performance in terms of both detection accuracy and segmentation quality. In addition, the fixed structure of our MaskSU R-CNN allows for end-to-end training. Therefore, it is feasible to prune our method and deploy it to mobile devices under the premise of ensuring recognition accuracy.

Future research will focus on the model's recognition capacity in satellite remote-sensing imageries, and the fusion of satellite multimodal data for forest-fire detection.

5. Conclusions

In this study, we present two solutions regarding the classification and segmentation of forest-fire images, with the main contributions as follows: (1) we design a novel attention mechanism (DSA module) to enhance the representation ability of feature channels and further improve the classification accuracy of incipient forest fires; (2) we merge the DSA module into ResNet as the backbone network of the instance segmentation model to improve the feature extraction capability; and (3) we reconstruct the MaskIoU branch of MS R-CNN using a U-shaped network, aiming to reduce the segmentation error. Experiments show that our MaskSU R-CNN outperforms many state-of-the-art segmentation models with a precision of 91.85%, recall 88.81%, F1-score 90.30%, and *mIoU* 82.31% in incipient forest-fire detection and segmentation. Our method, with its flexible structure and excellent performance, represents a shift toward the possibility of unmanned fire monitoring in a large area of forest.

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

Funding: This research was funded by the Future Network Scientific Research Fund Project (grant number FNSRFP-2021-YB-17) and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (grant number SJCX22_0312).

Data Availability Statement: This work uses the publicly available dataset FLAME, see reference [28] for data availability. More details about the data are available under Section 2.1.

Acknowledgments: We are very grateful to all the students assisted with data annotation and the experiments. We also thank the anonymous reviewers for helpful comments and suggestions to this paper.

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

References

1. Ryu, J.-H.; Han, K.-S.; Hong, S.; Park, N.-W.; Lee, Y.-W.; Cho, J. Satellite-based evaluation of the post-fire recovery process from the worst forest fire case in South Korea. *Remote Sens.* **2018**, *10*, 918. [CrossRef]
2. Yun, T.; Jiang, K.; Li, G.; Eichhorn, M.P.; Fan, J.; Liu, F.; Chen, B.; An, F.; Cao, L. Individual tree crown segmentation from airborne LiDAR data using a novel Gaussian filter and energy function minimization-based approach. *Remote Sens. Environ.* **2021**, *256*, 112307. [CrossRef]
3. Lucas-Borja, M.; Hedo, J.; Cerdá, A.; Candel-Pérez, D.; Viñegla, B. Unravelling the importance of forest age stand and forest structure driving microbiological soil properties, enzymatic activities and soil nutrients content in Mediterranean Spanish black pine (*Pinus nigra* Ar. ssp. *salzmannii*) Forest. *Sci. Total Environ.* **2016**, *562*, 145–154. [CrossRef] [PubMed]
4. Burrell, A.L.; Sun, Q.; Baxter, R.; Kukavskaya, E.A.; Zhila, S.; Shestakova, T.; Rogers, B.M.; Kaduk, J.; Barrett, K. Climate change, fire return intervals and the growing risk of permanent forest loss in boreal Eurasia. *Sci. Total Environ.* **2022**, *831*, 154885. [CrossRef]
5. Wu, Z.; He, H.S.; Keane, R.E.; Zhu, Z.; Wang, Y.; Shan, Y. Current and future patterns of forest fire occurrence in China. *Int. J. Wildland Fire* **2020**, *29*, 104. [CrossRef]
6. Yang, X.; Chen, R.; Zhang, F.; Zhang, L.; Fan, X.; Ye, Q.; Fu, L. Pixel-level automatic annotation for forest fire image. *Eng. Appl. Artif. Intell.* **2021**, *104*, 104353. [CrossRef]
7. Chai, J.; Zeng, H.; Li, A.; Ngai, E.W. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Mach. Learn. Appl.* **2021**, *6*, 100134. [CrossRef]
8. Fu, L.; Li, Z.; Ye, Q.; Yin, H.; Liu, Q.; Chen, X.; Fan, X.; Yang, W.; Yang, G. Learning robust discriminant subspace based on joint L2, p-and L2, s-Norm distance metrics. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *33*, 130–144. [CrossRef]
9. Ye, Q.; Huang, P.; Zhang, Z.; Zheng, Y.; Fu, L.; Yang, W. Multiview learning with robust double-sided twin SVM. *IEEE Trans. Cybern.* **2021**, 1–14. [CrossRef]
10. Zhan, J.; Hu, Y.; Zhou, G.; Wang, Y.; Cai, W.; Li, L. A high-precision forest fire smoke detection approach based on ARGNet. *Comput. Electron. Agric.* **2022**, *196*, 106874. [CrossRef]
11. Yu, Z.; Zhang, Y.; Jiang, B.; Yu, X. Fault-tolerant time-varying elliptical formation control of multiple fixed-wing UAVs for cooperative forest fire monitoring. *J. Intell. Robot. Syst.* **2021**, *101*, 48. [CrossRef]
12. Peng, Y.; Wang, Y. Real-time forest smoke detection using hand-designed features and deep learning. *Comput. Electron. Agric.* **2019**, *167*, 105029. [CrossRef]
13. Yan, Y.; Wu, X.; Du, J.; Zhou, J.; Liu, Y. Video fire detection based on color and flicker frequency feature. *J. Front. Comput. Sci. Technol.* **2014**, *8*, 1271–1279.
14. Çelik, T.; Demirel, H. Fire detection in video sequences using a generic color model. *Fire Saf. J.* **2009**, *44*, 147–158. [CrossRef]
15. Borges, P.V.K.; Izquierdo, E. A probabilistic approach for vision-based fire detection in videos. *IEEE Trans. Circuits Syst. Video Technol.* **2010**, *20*, 721–731. [CrossRef]
16. Li, X.; Chen, Z.; Wu, Q.M.J.; Liu, C. 3D Parallel Fully Convolutional Networks for Real-Time Video Wildfire Smoke Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *30*, 89–103. [CrossRef]
17. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
18. Wang, Y.; Zhao, J.-H.; Zhang, D.-Y.; Ye, W. Forest fire image classification based on deep neural network of sparse autoencoder. *Comput. Eng. Appl.* **2014**, *50*, 173–177.
19. Yin, M.; Lang, C.; Li, Z.; Feng, S.; Wang, T. Recurrent convolutional network for video-based smoke detection. *Multimed. Tools Appl.* **2018**, *78*, 237–256. [CrossRef]
20. Friedlingstein, P.; Jones, M.W.; O'sullivan, M.; Andrew, R.M.; Hauck, J.; Peters, G.P.; Peters, W.; Pongratz, J.; Sitch, S.; le Quéré, C.; et al. Global carbon budget 2019. *Earth Syst. Sci. Data* **2019**, *11*, 1783–1838. [CrossRef]
21. Huang, Q.; Razi, A.; Afghah, F.; Fule, P. Wildfire spread modeling with aerial image processing. In Proceedings of the 2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM), Cork, Ireland, 31 August–3 September 2020; pp. 335–340.
22. De Sousa, J.V.R.; Gamboa, P.V. Aerial forest fire detection and monitoring using a small UAV. *KnE Eng.* **2020**, *5*, 242–256. [CrossRef]
23. Ciprián-Sánchez, J.F.; Ochoa-Ruiz, G.; Gonzalez-Mendoza, M.; Rossi, L. FIRe-GAN: A novel deep learning-based infrared-visible fusion method for wildfire imagery. *Neural Comput. Appl.* **2021**. [CrossRef]
24. Pan, H.; Badawi, D.; Zhang, X.; Cetin, A.E. Additive neural network for forest fire detection. *Signal Image Video Process.* **2019**, *14*, 675–682. [CrossRef]

25. Zhang, J.; Zhu, H.; Wang, P.; Ling, X. ATT squeeze U-Net: A lightweight network for forest fire detection and recognition. *IEEE Access* **2021**, *9*, 10858–10870. [CrossRef]
26. Yuan, C.; Liu, Z.; Zhang, Y. UAV-based forest fire detection and tracking using image processing techniques. In Proceedings of the 2015 International Conference on Unmanned Aircraft Systems (ICUAS), Denver, CO, USA, 9–12 June 2015; pp. 639–643.
27. Sudhakar, S.; Vijayakumar, V.; Kumar, C.S.; Priya, V.; Ravi, L.; Subramaniyaswamy, V. Unmanned Aerial Vehicle (UAV) based Forest Fire Detection and monitoring for reducing false alarms in forest-fires. *Comput. Commun.* **2019**, *149*, 1–16. [CrossRef]
28. Shamsoshoara, A.; Afghah, F.; Razi, A.; Zheng, L.; Fulé, P.Z.; Blasch, E. Aerial imagery pile burn detection using deep learning: The FLAME dataset. *Comput. Netw.* **2021**, *193*, 108001. [CrossRef]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
30. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask scoring R-Cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6409–6418.
31. Xu, Y.-H.; Li, J.-H.; Zhou, W.; Chen, C. Learning-empowered resource allocation for air slicing in UAV-assisted cellular V2X communications. *IEEE Syst. J.* **2022**, 1–4. [CrossRef]
32. Chelali, F.Z.; Cherabit, N.; Djeradi, A. Face recognition system using skin detection in RGB and YCbCr color space. In Proceedings of the 2015 2nd World Symposium on Web Applications and Networking (WSWAN), Sousse, Tunisia, 21–23 March 2015; pp. 1–7.
33. Umar, M.M.; Silva, L.C.D.; Bakar, M.S.A.; Petra, M.I. State of the Art of Smoke and Fire Detection Using Image Processing. *Int. J. Signal Imaging Syst. Eng.* **2017**, *10*, 22–30. [CrossRef]
34. Hackel, T.; Usvyatsov, M.; Galliani, S.; Wegner, J.D.; Schindler, K. Inference, learning and attention mechanisms that exploit and preserve sparsity in CNNs. *Int. J. Comput. Vis.* **2020**, *128*, 1047–1059. [CrossRef]
35. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. *Adv. Neural Inf. Process. Syst.* **2014**, *2*, 2204–2212.
36. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
37. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
38. Li, Y.; Yuan, Y. Convergence analysis of two-layer neural networks with relu activation. *Adv. Neural. Inf. Process. Syst.* **2017**, *30*, 597–607.
39. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
40. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-Cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
41. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
42. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
43. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
44. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft Coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
45. Huh, M.; Agrawal, P.; Efros, A.A. What makes ImageNet good for transfer learning? *arXiv* **2016**, arXiv:1608.08614.
46. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
47. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
48. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-Cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
49. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
50. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
51. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
52. Ba, R.; Chen, C.; Yuan, J.; Song, W.; Lo, S. SmokeNet: Satellite smoke scene detection using convolutional neural network with spatial and channel-wise attention. *Remote Sens.* **2019**, *11*, 1702. [CrossRef]
53. Zhang, F.; Dong, Y.; Xu, S.; Yang, X.; Lin, H. An approach for improving firefighting ability of forest road network. *Scand. J. For. Res.* **2020**, *35*, 547–561. [CrossRef]
54. Xu, R.; Lin, H.; Lu, K.; Cao, L.; Liu, Y. A forest fire detection system based on ensemble learning. *Forests* **2021**, *12*, 217. [CrossRef]
55. Ye, Q.; Li, Z.; Fu, L.; Zhang, Z.; Yang, W.; Yang, G. Nonpeaked Discriminant Analysis for Data Representation. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3818–3832. [CrossRef]