forests MDPI

Article

A Collaborative Region Detection and Grading Framework for Forest Fire Smoke Using Weakly Supervised Fine Segmentation and Lightweight Faster-RCNN

Jin Pan [1,2], Xiaoming Ou [1] and Liang Xu [3,*]

[1] College of Economics and Management, South China Agricultural University, Guangzhou 510642, China; panj@gzpyp.edu.cn (J.P.); ouxm0758@126.com (X.O.)
[2] College of Finance and Economics, Guangzhou Panyu Polytechnic, Guangzhou 511483, China
[3] School of Automation, Guangdong University of Technology, Guangzhou 510006, China
[*] Correspondence: celiangxu@gdut.edu.cn

Abstract: Forest fires are serious disasters that affect countries all over the world. With the progress of image processing, numerous image-based surveillance systems for fires have been installed in forests. The rapid and accurate detection and grading of fire smoke can provide useful information, which helps humans to quickly control and reduce forest losses. Currently, convolutional neural networks (CNN) have yielded excellent performance in image recognition. Previous studies mostly paid attention to CNN-based image classification for fire detection. However, the research of CNN-based region detection and grading of fire is extremely scarce due to a challenging task which locates and segments fire regions using image-level annotations instead of inaccessible pixel-level labels. This paper presents a novel collaborative region detection and grading framework for fire smoke using a weakly supervised fine segmentation and a lightweight Faster R-CNN. The multi-task framework can simultaneously implement the early-stage alarm, region detection, classification, and grading of fire smoke. To provide an accurate segmentation on image-level, we propose the weakly supervised fine segmentation method, which consists of a segmentation network and a decision network. We aggregate image-level information, instead of expensive pixel-level labels, from all training images into the segmentation network, which simultaneously locates and segments fire smoke regions. To train the segmentation network using only image-level annotations, we propose a two-stage weakly supervised learning strategy, in which a novel weakly supervised loss is proposed to roughly detect the region of fire smoke, and a new region-refining segmentation algorithm is further used to accurately identify this region. The decision network incorporating a residual spatial attention module is utilized to predict the category of forest fire smoke. To reduce the complexity of the Faster R-CNN, we first introduced a knowledge distillation technique to compress the structure of this model. To grade forest fire smoke, we used a 3-input/1-output fuzzy system to evaluate the severity level. We evaluated the proposed approach using a developed fire smoke dataset, which included five different scenes varying by the fire smoke level. The proposed method exhibited competitive performance compared to state-of-the-art methods.

Keywords: region detection of forest fire; grading of forest fire; weakly supervised loss; fine segmentation; region-refining segmentation; lightweight Faster R-CNN

check for updates

Citation: Pan, J.; Ou, X.; Xu, L. A Collaborative Region Detection and Grading Framework for Forest Fire Smoke Using Weakly Supervised Fine Segmentation and Lightweight Faster-RCNN. *Forests* **2021**, *12*, 768. https://doi.org/10.3390/f12060768

Academic Editors: Olga Viedma, Stelian Alexandru Borz, Andrea R. Proto, Robert Keefe and Mihai Nita

Received: 8 April 2021
Accepted: 9 June 2021
Published: 10 June 2021

1. Introduction

Forest fires have become one of the major disasters causing serious ecological, social, and economic damage, as well as personal casualty loss [1–3]. In 2013, a forest fire burned a land area of approximately 1042 km^2 in California, causing USD 127.35 million of damage. In China, 214 forest fire events occurred alone in the Huichang County of the JiangXi province from 1986 to 2009, with an area of more than 460 km^2 being affected [4]. The statistical data provided in [5] show that fire disasters alone caused an overall damage of

USD 3.1 billion in 2015. To monitor the fire smoke, numerous image-based surveillance systems have been installed in forests. Therefore, rapid and accurate detection and grading of fire smoke is crucial and helpful for preventing and reducing the forest losses.

Forest fires commonly spread quickly and are difficult to rapidly control. Accurately identifying the fire region and evaluating the fire smoke severity, which helps firefighters to take proper measures and quickly control a fire's spreading, is a very challenging task. Moreover, most firefighters will need to decide how many resources to allocate to a particular forest fire according to useful extinguishing information, which contains the region, the location, and the severity (i.e., the grading of risk) of fire smoke. Therefore, a technique for a fire surveillance system which can give an early region detection of fire or smoke and evaluate the severity of fire or smoke is indispensable.

The traditional technologies for detecting fire and smoke use various sensors. A point sensor [6–8] works best in indoor spaces, as it only covers a small area, and is insensitive to fire in the outdoors and over a large range. An Unmanned Aerial Vehicle (UAV) technique can also be used to monitor forest fires [9]. Additionally, such techniques cannot provide important vision information to help firefighters quickly evaluate the severity of the fire and make appropriate decisions. Satellite sensors [10] only detect a large fire in a wide range, and they are not useful for the early detection of fire and smoke. Currently, with a large amount of image surveillance systems installed in forests, there is an appropriate alternative to the traditional techniques, and vison-based inspecting technologies for fires and smoke have been widely adopted due to their easy deployment and lower cost, insusceptibility to the weather, and long and short availability.

Vision-based technologies make full use of color and motion features for fire detection. Due to the conspicuous color of fire, Chen et al. [11] proposed RGB- and HIS-based color models are used to examine the dynamic behavior of fires, which can be applied to detect the irregular properties of fire. Additionally, the YUV color [12], RGB [13], and YCbCr color space [14], have been explored to classify the pixels in fire and non-fire regions. However, such methods have many limitations in various situations, e.g., due to the complexity of wild scenes, the diversity of fire and smoke in forests, irregular lighting, and low-contrast flame and smoke.

In recent years, convolutional neural networks (CNNs) have attracted attention due to their outstanding performance in image recognition. Some scholars have introduced CNN models—for example, AlexNet [15], GoogleNet [16], ZF-Net [17], VGG [18], and ResNet [19], into the field of the vision detection of fires. Regarding the use of these models, some scholars have also proposed improved CNN-based methods for fire or smoke detection, such as, smoke detection in a video based on a deep belief network using energy and intensity features [20], a video-based detection system using an object segmentation and efficient symmetrical features [21], a two-stream CNN model with the adaptive adjustment of the receptive field [22], and an object detection model incorporating environmental information [23]. Additionally, Liu et al. [24] also proposed a forest fire detection system based on ensemble learning to reduce false alarms. Nevertheless, the aforementioned methods are only applicable for the recognition of whether fire or smoke exists in an image, but such methods cannot provide more detailed information about fires, such as their location, shape, size, etc., which can be used to grade the level of fires or smoke. Sometimes, we need to focus on the fire/smoke spreading or the emerging regions; thus, the region detection of fire or smoke is a better solution to this issue.

Recently, rapid progress in this technology has been made using powerful models, such as DeepLab [25], U-Net [26], and fully convolutional networks (FCNs) [27]. However, the performance of these deep models heavily depends on a large amount of training data with expensive pixel labels. Due to the uncertain, complex, and changeable shape of fires and smoke, annotating such training data has become a bottleneck in applying these models to forest fire detection, as it is a time-consuming and arduous task to label each pixel on a large amount of fire images. One objective of our work was to loosen the supervision, i.e., by performing weakly supervised segmentation for forest fire detection using only image-level supervision.

It is well known that multi-model cooperation can improve the performance of any machine-learning algorithm. Therefore, model selection and collaborative strategies need to be considered carefully. To grade forest fire smoke, we need to collect and consider various factors, such as the fire region size, fire shape, location, etc., which may influence fire smoke and cause fires to change and spread.

Our work mainly concentrated on a region detection of forest fires or smoke at an early-stage and evaluation of the fires' smoke severity, thereby proposing a collaborative region detection and grading framework for forest fire smoke using a weakly supervised fine segmentation and a lightweight Faster R-CNN.

Our other contributions can also be summarized as follows.

(1) We proposed the weakly supervised fine-segmentation method, which consists of a segmentation network (LS-Net), used to simultaneously and accurately detect the region of forest fire smoke, and a novel two-stage weakly supervised leaning strategy, which includes a weakly supervised loss (WSL), a region-refining segmentation (RRS) algorithm, and an attention-based decision network (AD-Net) for fire smoke classification.

(2) To reduce the complexity of the Faster R-CNN, we introduce a knowledge distillation process to compress this model into a simple model. We also proposed a distillation strategy for this model. Moreover, we used a 3-input/1-output fuzzy system based on fuzzy logic to grade forest fire smoke.

(3) We developed a forest fire dataset from public resources to evaluate our method, which is composed of various situations, including large fires, small fires, dense smoke, light smoke, and other scenes.

The remainder of this paper is organized as follows. Section 2 introduces related work. Section 3 describes the proposed framework, including weakly supervised fine segmentation, lightweight Faster R-CNN, a collaborative learning strategy, and the grading method. Section 4 presents the experimental results and a discussion of them. Section 5 concludes the paper.

2. Related Work

The accurate and timely recognition of the region of forest fire smoke is an important task in preventing forest disasters and protecting the environment. To address this issue, many researchers have developed various techniques, such as wireless network sensors and satellite systems [8,10,28], robotic systems [29], intelligent techniques [30], and image processing techniques [11–14]. Due to the factors of deployment, utilization convenience of use, and a high detection rate, the image techniques have been used widely and have attracted the attention of many researchers, as they are more suitable for forest fire detection. Nevertheless, there are some limitations of the traditional image technologies that are used in real-world applications [11–14]. Recently, DL-based methods have become a mainstream technology for intelligent fire detection based on vision [23,31–35].

To improve the performance of fire detection, a deep normalization and convolutional neural network was proposed for automatic feature extraction and classification, avoiding hard-crafted features [31]. Different classic CNN models, Namozov et al. [32], proposed the use of an adaptive piecewise linear unit, instead of using traditional rectified linear units in the hidden layers of the network. In previous work [33], the authors tackled the overfitting and accuracy of CNN-based fire detection using a limited dataset and a deep convolutional generative adversarial network, which achieved a high accuracy in visual fire detection. To improve its implementation in a real-world surveillance network, Baik et al. [34] presented an energy-friendly and computationally efficient CNN architecture for the detection, localization, and scene understanding of fires. This model reduces the computational requirements to a minimum and obtains a better accuracy due to its increased depth. Furthermore, the authors proposed that an efficient CNN model via edge intelligence could be used to detect fires in uncertain surveillance scenarios [35]. This model utilized a lightweight deep network, instead of dense fully connected layers, which require expensive computation. Nevertheless, the CNN models have only been applied to classification tasks to predict whether fires exist in an image or not. Zhang et al. [23] proposed a forest fire smoke recognition method based on an anchor box adaptive generation.

The most important objective in predicting forest fire risk is to obtain more information regarding a fire in an image, such as its shape, size and location, and the CNN-based image segmentation techniques are a better solution for addressing this issue. Recently, several powerful baseline systems, such as the Fast/Faster/Mask R-CNN [36–38], have been proposed to drive the rapid progress in instance segmentation. Semantic segmentation methods highly related to instance segmentation, such as DeepLab [25] and U-Net [26], only recognize the category of each pixel, without distinguishing different detection object instances. While fully supervised methods can achieve an outstanding performance, they require a large amount of training data with expensive annotations, which causes inconvenience in practical applications. Thus, weakly supervised image segmentation using relaxed supervision—i.e., image-level annotations—is a better solution to the issues.

Weak supervision is an inexact pattern [39]. Every pixel in an image should ideally be annotated in image segmentation. However, we usually have only coarse-grained labels, instead of pixel-wise labels. Weakly supervised image segmentation refers to the training of models with coarse-grained labels to obtain pixel-wise segmentation results. This has been explored in image segmentation primarily to reduce the effort of establishing training data. Many multiple instance learning (MIL) techniques have been investigated for weakly supervised image segmentation. Pathak et al. [40] proposed an MIL formulation of multi-class semantic segmentation learning using a fully convolutional network. Pinheiro et al. [41] investigated a CNN-based model with MIL, which was constrained during training to put more weight on the pixels that were important for classifying an image.

To further accurately segment objects, graphical energy minimization techniques have been extensively used to regularize image segmentation due to their inherent optimality guarantees. In previous studies, many researchers have proposed diverse solutions for image segmentation using this technique. Boykov et al. [42] used graph cuts to find the globally optimal segmentation position. A GrubCut method was proposed to segment images with bounding box annotations by iteratively updating the parameters of a Gaussian mixture model (GMM) [43]. Fully connected CRF models were implemented by an efficient inference algorithm, defining pairwise edge potentials by a linear combination of Gaussian kernels [44]. To overcome the poor localization property of deep networks, Chen et al. [45] proposed a new final layer to be combined with DenseCRF.

3. Methods

In this paper, we focus on the region detection of forest fire smoke and the grading of fire smoke severity in a surveillance image. Considering a better solution which integrates several simple models into a framework to obtain a better performance, we propose a novel collaborative region detection and grading framework for forest fire smoke using a weakly supervised fine segmentation and a lightweight Faster R-CNN (called FireDGWF), as shown in Figure 1. This framework consists of a detection process and a grading process.

In the detection process, we use 3 different models— classification (for example, ShuffleNet [46]), region detection (weakly supervised fine segmentation, also known as WSFS), and region-proposal (lightweight Faster R-CNN) methods, which can locate, segment, and predict the location, region, and category of forest fire or smoke in an input image.

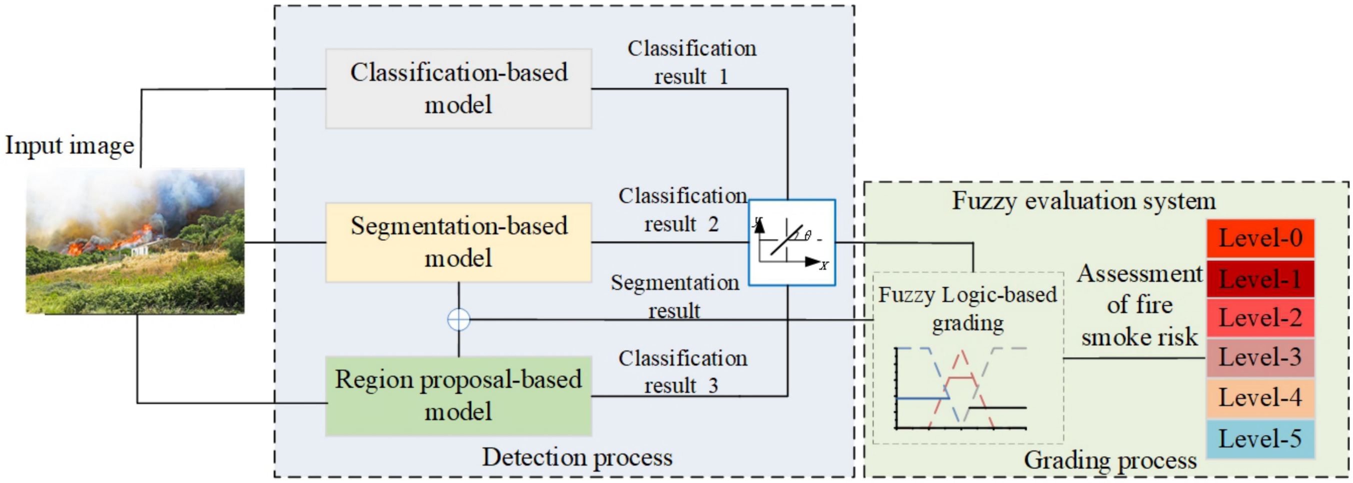

Figure 1. Our proposed framework.

Moreover, compared with the approach that some scholars proposed of using a bounding box to achieve relaxed supervision [47,48], we propose the weakly supervised fine segmentation to obtain a greater segmentation accuracy, which is achieved by applying "pixel-level labels" in our method.

The Faster R-CNN has better detection accuracy than does the one-stage object detection, such as YOLOv3 [49], SSD [50], DSSD [51], and RefineNet [52]. However, the greater number of parameters and time-consuming training has become a bottleneck in applying this method to fire smoke detection in forests. In this paper, we first introduce a knowledge distillation approach, which was proposed by Hinton [53], to reduce the complexity of this model, i.e., the lightweight Faster R-CNN.

For the grading process, we utilize a fuzzy evaluation system, which can synthetically evaluate the fire smoke level based on 3 inputs, including the classification prediction, the region detection (segmentation result), and the location.

In the following section, we separately introduce the region detection using the WSFS, lightweight Faster R-CNN, collaborative learning strategy, and grading method with fuzzy logic.

3.1. Region Detection Using Weakly Supervised Fine Segmentation

The time-consuming effort involved in pixel-wise annotations makes the identification of forest fire regions an inconvenient and challenging task in real-world applications. To address this issue, we propose the novel weakly supervised fine segmentation approach for the detection of the region of forest fire smoke, including the segmentation network (LS-Net) and the decision network (AD-Net), as shown in Figure 2. LS-Net obtains pixel-level segmentation results for fire or smoke regions that are difficult to identify in complex forest scenarios and provides good semantic features for AD-Net. AD-Net uses the deepest feature and the segmentation results of LS-Net and predicts the probability of fire smoke existing in an image. To improve the classification performance, AD-Net focuses on the fire or smoke pixels, the weights of which are determined by the segmentation results provided by LS-Net.

To address the problem that pixel-wise annotations are expensive, we introduce the two-stage weakly supervised learning strategy, including the weakly supervised loss (WSL) and the region-refining segmentation (RRS). LS-Net can only roughly locate fire or smoke regions using the weakly supervised loss only with image-level labels. However, using weak annotation solely at the image level is insufficient for training a high-quality segmentation model. After training the model with the weakly supervised loss, we propose the RRS to fine-tune LS-Net for accurate region segmentation.

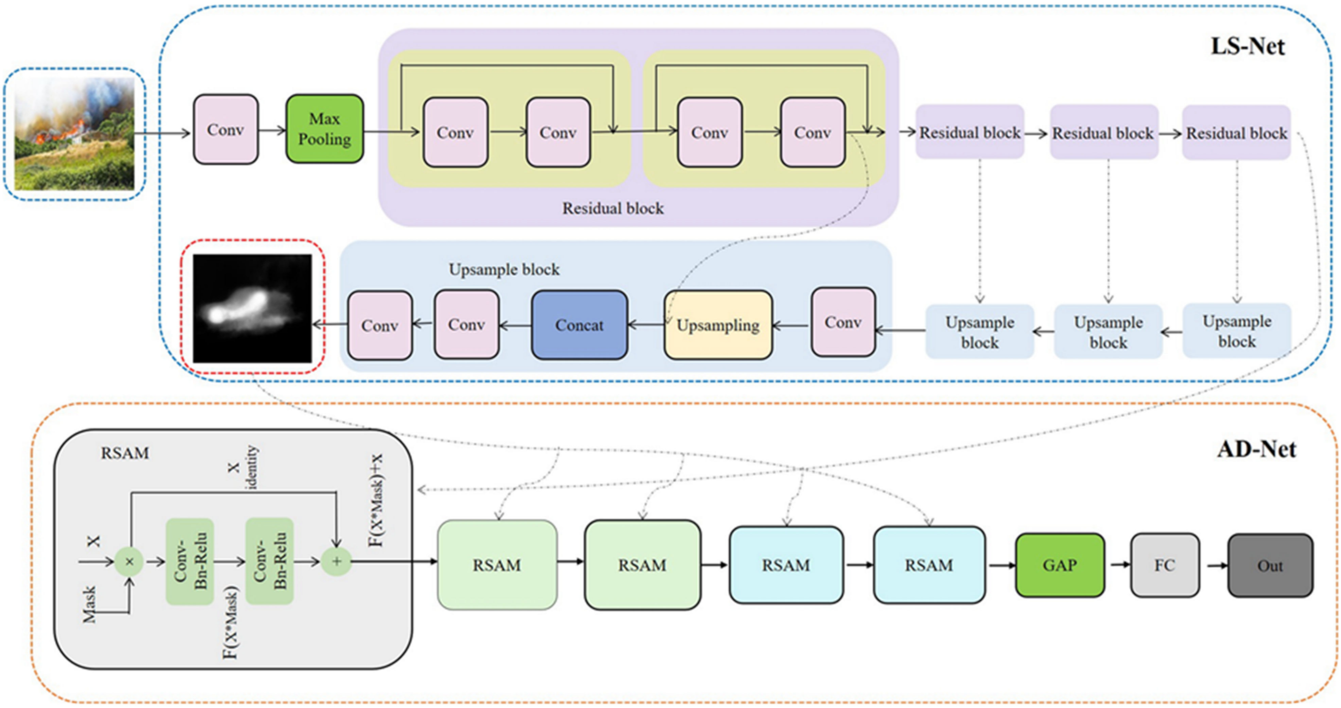

Figure 2. Proposed framework for weakly supervised fine segmentation.

3.1.1. Structure of LS-Net

Table 1 lists the structural parameters of LS-Net. The segmentation network (LS-Net), based on the U-Net structure, is composed of an encoder and decoder, and the encoder is improved using a modified ResNet18 [54]. The encoder uses a hierarchical structure and residual connections to extract multi-level features from an input image, and these have different spatial sizes and semantic information. The first convolutional layer in the original ResNet18 used a 7 × 7 kernel, with a step size of 2. We use 3 3 × 3 convolutional layers, with a step size of 1, to reduce computational costs and maintain the spatial size. The other layers are the same as in the original ResNet18. In the forward process, features of different sizes are outputted from the encoder. The decoder makes full use of the multi-level features of the encoder and outputs a single-channel image to segment fires at the pixel level. The decoder has 4 upsampling layers and multiple convolutional layers. Upsampling is implemented with bilinear interpolation. To capture small fires, each convolutional layer uses a 3 × 3 kernel, with a step size of 1. Batch normalization and nonlinear ReLU layers are included after each convolutional layer.

Table 1. LS-Net architecture.

Encoder			Decoder		
Name	**Layers**	**Output Size**	**Name**	**Layers**	**Output Size**
Layer1	Conv3_3 × 3	[h, w]	Layer1	Upsampling, Conv3_3 × 2	[h//8, w//8]
Layer2	Residual block × 2	[h//2, w//2]	Layer2	Upsampling, Conv3_3 × 2	[h//4, w//4]
Layer3	Residual block × 2	[h//4, w//4]	Layer3	Upsampling, Conv3_3 × 2	[h//2, w//2]
Layer4	Residual block × 2	[h//8, w//8]	Layer4	Upsampling, Conv3_3 × 2	[h, w]
Layer5	Residual block × 2	[h//16, w//16]	Out-layer	Conv1_1, Sigmoid	[h, w]

3.1.2. Structure of AD-Net

Table 2 shows the structure of AD-Net. In this study, an attention mask is introduced into the residual branch of a residual spatial attention model (RSAM) to focus more attention on heavily weighted regions, as shown in Figure 2. The RSAM modules utilize the results of LS-Net as an attention mask to cause AD-Net to focus more on small object regions from the segmentation results. The number of RSAM modules is increased from 1 to 5 to enhance the contrast between fire and non-fire regions. In Figure 2, AD-Net includes 5 RSAM modules, 1 global average pooling layer (RRS) and 1 fully connected layer. As in LS-Net, batch normalization and ReLU are used. This pooling layer squeezes the spatial dimensions and extracts abstract semantic information. Because the output features only depend on the channel of the input feature for this pooling layer, a fixed input image size is unnecessary. The fully connected layer with a sigmoid activation function takes the output of this pooling layer as the input and predicts the probability that an object exists in an image.

Table 2. AD-Net architecture.

Name	Layers	Output Size
Layer1	RSAM	[h//16, h//16]
Layer2	RSAM	[h//32, h//32]
Layer3	RSAM	[h//32, h//32]
Layer4	RSAM	[h//64, h//64]
Layer5	RSAM	[h//64, h//64]
Layer6	GAP	-
Layer7	Denser Layer	-

3.1.3. Weakly Supervised Loss

In this section, we introduce the novel weakly supervised loss. To simplify the description, Table 3 firstly presents the variables and descriptions used in this loss.

Table 3. Variables and descriptions.

Variable	Description
X	Input image
$Y(X)$	Image label for X
Φ	Model parameters
$p(Y\|(X,\Phi))$	Predicted model for X
w, h	Height and width of X
χ	Pixel set for X
x_i	A pixel in χ
y_i	Pixel label for x_i
$q(y_i\|(x_i,\Phi))$	Predicted model for x_i

This weakly supervised loss is used to train LS-Net from scratch with only image-level annotations to simultaneously locate and segment the fire or smoke regions. This loss includes a positive loss, negative loss, image-level loss, and pixel-level loss. The positive and negative losses are used for positive (fire or smoke) and negative (non-fire or non-smoke) samples, respectively, during training. The positive and negative loss guides LS-Net to recognize object (fire or smoke) and non-object pixels. The image-level loss allows for an easier and faster convergence [55,56], and the pixel-level loss forces LS-Net to provide a clear prediction for each pixel. The weakly supervised loss can be written as:

$$Loss_{WSL} = Loss_{negative} * \beta + Loss_{positive} + Loss_{image} + Loss_{pixel} \tag{1}$$

where $Loss_{negative}$ is the target loss for negative samples. Since there are only non-object pixels in negative samples, $Loss_{negative}$ is proposed to guide LS-Net to classify the pixels in negative samples as non-objects. The negative samples satisfy $Y(\mathrm{X}) = 0, y_i = Y(\mathrm{X}) = 0$, $\forall(x_i, y_i) \in (\mathcal{X}, \mathcal{Y})$. The cross-entropy loss is used to optimize the negative samples, and the negative loss can be rewritten as:

$$Loss_{negative} = (1 - Y(X)) * E_{(x_i,y_i)\in(\mathcal{X},\mathcal{Y})} - \log(1 - p(y_i|x_i, \Phi)) \tag{2}$$

where $Loss_{positive}$ is the target loss for positive samples. We know that both object and non-object pixels exist in the positive sample image. We sort the pixels in positive samples by their prediction values from LS-Net and assume that the largest α proportion pixels are fire or smoke. α is defined as the ratio of the area of fire or smoke pixels to the total area. We choose the largest double α proportion as the region of interest R. Let R denote the set of pixels in $\mathcal{X}_R$. $\mathcal{Y}_R$ is the set of labels corresponding to $\mathcal{X}_R$. When all the predictions in the positive sample are correct, the expectation entropy corresponding to $\mathcal{X}_R$ reaches its maximum. This can be expressed as:

$$Loss_{positive} = \begin{bmatrix} P(\mathcal{Y}_{\mathrm{R}}|\mathcal{X}_{\mathrm{R}},\Phi)\log(P(\mathcal{Y}_{\mathrm{R}}|\mathcal{X}_{\mathrm{R}},\Phi)) \\ +(1-P(\mathcal{Y}_{\mathrm{R}}|\mathcal{X}_{\mathrm{R}},\Phi))(\log(P(\mathcal{Y}_{\mathrm{R}}|\mathcal{X}_{\mathrm{R}},\Phi))) \end{bmatrix} * \mathrm{Y}(\mathrm{X}), \tag{3}$$

where the prediction expectation corresponding to $\mathcal{X}_R$ is:

$$P(\mathcal{Y}_{\mathrm{R}}|\mathcal{X}_{\mathrm{R}},\Phi) = E_{(x_i,y_i)\in(\mathcal{X}_{\mathrm{R}},\mathcal{Y}_{\mathrm{R}})}\,p(y_i|x,\Phi)\,. \tag{4}$$

where $Loss_{image}$ is the image-level loss. For positive samples, at least one pixel should be labeled as fire or smoke. No pixels in negative samples should be labeled as objects. Based on this idea, we minimize the KL divergence of image-level labels and the maximum of the image's segmentation result by cross-entropy loss. This can be expressed as:

$$Loss_{image} = Y * \log(p_M) + (1-Y) * \log(1-p_M), \tag{5}$$

where $p_M = \mathrm{Max}_{(x_i,y_i)\in(\mathcal{X},\,\mathcal{Y})}(p(y_i|x_i,\Phi))$ is the maximum of an image's pixel predictions.

Ideally, each pixel in an image is an object (fire or smoke) or non-object; i.e., each value segmentation result should be close to 1 or 0. To ensure that LS-Net provides a clear prediction for each pixel, we minimize the entropy values of pixel predictions during training. This can be expressed as:

$$Loss_{pixel} = E_{(x_i,y_i)\in(\mathcal{X},\mathcal{Y})}\,H[\,p(y_i|x_i,\Phi)\,] = \frac{1}{h\,*\,w}\sum_{i=1}^{w*h} H[\,p(y_i|x_i,\Phi)\,], \tag{6}$$

where the entropy of pixel $(\mathrm{x_i},\mathrm{y_i})$ is:

$$\begin{aligned} H[\,p(y_i|x_i,\Phi)\,] &= -p(y_i|x_i,\Phi)\log(p(y_i|x_i,\Phi))- \\ &(1-p(y_i|x_i,\Phi))(\log(1-p(y_i|x_i,\Phi))). \end{aligned} \tag{7}$$

3.1.4. Region-Refining Segmentation

The region detection (segmentation result) using LS-Net only identifies the rough regions of fire and smoke. Furthermore, the fine segmentation results are implemented on the RRS algorithm, which is proposed to fine-tune LS-Net. This can be regarded as an iterative energy minimization method like GrubCut [43]. We build a DenseCRF on LS-Net, and based on these methods, we formulate an energy minimization problem to estimate the latent pixel labels, which are called pseudo-pixel labels. We consider these as the supervision to fine-tune LS-Net with a small learning rate. RRS alternates between estimating pseudo-pixel labels and using them to optimize LS-Net. In other words, DenseCRF improves the segmentation results, and RRS training can be seen as a self-evolution process of LS-Net, as shown in Figure 3.

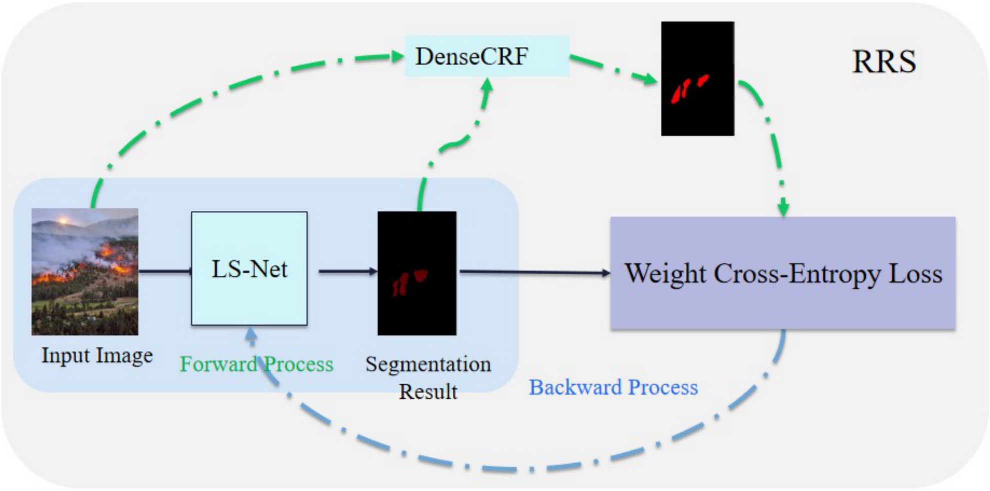

Figure 3. Proposed region-refining segmentation (RRS) algorithm.

The training stage has two key phases: label estimation and model updating.

In the label estimation stage, DenseCRF is built, and pseudo-pixel labels can be acquired by minimizing the energy function [44].

$$E(x) = \sum\nolimits_i \psi_u(f_i) + \sum\nolimits_{i<j} \psi_p(f_i, f_j) \tag{8}$$

The pairwise regularization term $\psi_p(f_i, f_j)$ penalizes label differences for pixels i and j, and typically has the form:

$$\psi_p(f_i, f_j) = \mu(x_i, x_j) * k(v_i, v_j) \tag{9}$$

where μ is a label compatibility function given by the Potts model, $\mu(x_i, x_j) = [x_i \neq x_j]$, and k represents the linear combinations of Gaussian kernels [44].

In the model updating phase, LS-Net is fine-tuned using the pseudo-pixel labels as supervision. We know that all pixels in negative samples are non-fire or non-smoke, so those pixel labels are set to zero, and pseudo-pixel labels are not used. The detailed process of this algorithm is described in Algorithm 1.

Algorithm 1. Region-Refining Segmentation algorithm (RRS)

```
Input: dataset $D = \{x_i\}_{i=1}^{N}$, $y(x_i) \in [0,1]$
Initialization: model $\Phi$ trained by the weakly supervised loss; parameters of RRS $\omega_a, \omega_s, \theta_\alpha, \theta_\beta$,
and $\theta_\gamma$; number of steps N; learning rate $lr$; batch size $b$;
For step = $1, 2, \ldots, N$ do:
    $\leftarrow$ sample $b$ images $x$ from D
Predict labels $y(x)$ of samples with $\Phi$
Get pseudo-pixel labels $\acute{y}(x)$ with DenseCRF
update $\Phi$ by minimizing loss of $y(x)$ and $\acute{y}(x)$
End for
```

3.1.5. Training Strategy

Our method has a phased training strategy for the WSFS. The LS-Net is trained with the weakly supervised loss and the region-refining segmentation algorithm successively, and it obtains a good initialization, so it can accurately locate and segment fire or smoke pixels. The next step is to train the AD-Net, during which the LS-Net outputs remain unchanged, and the cross-entropy loss is used.

3.2. *Lightweight Faster R-CNN*

In this section, we introduce a knowledge distillation technique to simplify the complexity of the Faster R-CNN.

3.2.1. Knowledge Distillation

Knowledge distillation [55] aims to compress a complex model into a simpler model that is much easier to deploy. The main goal of knowledge distillation is to train a small network model to imitate a pre-trained effective and complex network. Our proposed lightweight Faster R-CNN is implemented using a teacher stream and a student stream. For the sake of simplicity, we define the key components of the two streams as follows:

The teacher stream uses a complex CNN structure, with a set of parameters pre-trained as a feature extractor. Here, we assume that this model has absorbed the rich knowledge encoded in generous high-resolution forest fire smoke images with labels. Generally, the training dataset is very large and may be invisible to the student stream. The student stream is a much simpler CNN network that does not need too many parameters for recognizing low-resolution forest fire smoke.

In contrast to the classical Faster R-CNN, we use the ResNet 50 as the teacher stream and the ResNet 18 as the student stream to substitute for VGG-16 for proposal and detection. The features extracted by the teacher stream are used to distill the knowledge. The feature loss—i.e., L2 loss—is based on both the eigenvectors obtained from the student stream and the eigenvectors distilled from the teacher stream. In the process of forward propagation, the loss of the whole network includes the feature loss, RPN loss, and RCNN loss.

3.2.2. Loss Function

To train the lightweight Faster R-CNN, we propose a novel loss $Loss_{all}$, which includes $Loss_{RPN}$, $Loss_{RCN}$, and $Loss_{backbone}$. $Loss_{RPN}$ and $Loss_{RCN}$ represent the loss of the RPN module and the loss of the RCNN module, respectively. $Loss_{backbone}$ is the loss in extracting features, which is expressed as

$$Loss_{all} = Loss_{RPN} + Loss_{RCN} + Loss_{backbone} \tag{10}$$

Specially, $Loss_{RPN}$ and $Loss_{RCN}$ are the classification loss and object regression loss, respectively, which are defined as

$$\begin{aligned} Loss_{RPN} &= \tfrac{1}{N_{cls}} \sum Loss_{cls}^{RPN} + \lambda \tfrac{1}{N_{reg}} \sum p^* Loss_{reg}^{RPN} \\ Loss_{RCN} &= \tfrac{1}{M_{cls}} \sum Loss_{cls}^{RCN} + \lambda \tfrac{1}{M_{reg}} \sum p^* Loss_{reg}^{RCN} \end{aligned} \tag{11}$$

where p^* = 1 while an anchor is positive; and p^*=0 while an anchor is negative. N_{cls} and N_{reg} are the batch size of the RPN and anchor location, respectively; M_{cls} and M_{reg} are the batch size of the RCNN and anchor location, respectively.

The classification loss of RPN and RCNN can use a cross-entropy loss. The regression loss of RPN and RCNN is calculated using a smooth L1 loss. A parameter is added to control the smooth area, which is expressed as:

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 \, if|x| < 1\sigma^2 \\ |x| - \frac{0.5}{\sigma^2} \, otherwise \end{cases} \tag{12}$$

$Loss_{backbone}$ is based on a calculation of the KL divergence between the teacher network and the student network. However, before calculating the KL divergence, it is necessary to ensure that the dimensions of the feature map between the teacher network and the student network are consistent:

$$Loss_{backbone}(y_T, y_S) = \sum_{i=1}^{n} y_{S_i} \times \log(\frac{y_{S_i}}{y_{T_i}}) \tag{13}$$

where y_t is the output feature of the teacher network, and y_s is the output feature of the student network.

The distilling process of the lightweight Faster R-CNN is shown in Algorithm 2.

Algorithm 2. Distilling Process of Faster R-CNN

```
Input: Parameters of the features of the teacher network: T_Parameter, Dataset with labels: S_Input{$P, T$}
Procedure Iteration process
1. Get map from Teacher-Net: T_Feature = Teacher_backbone(T_Parameter)
2. Transform dimension: T_Map_trans = Trans (T_ Feature)
3. Generate map from Student-Net: S_Map = Student_backbone(S_Input)
4. Caculate distill_loss with $L2_{loss}$: $Loss_{backbone}$ = $L^{T\to S}$(T_Map, S_Map)
5. Compute the loss of detection with RPN and RCN: $Loss_{rpn}$ = RPN ($P_i$, S_Map, $T_i$), $Loss_{RCN}$ = RCN ($S_i$, S_Map, $T_i$)
6. Reverse the loss $L_{all}$ = $L^{RPN}$ + $L^{RCN}$ + $L^{backbone}$ for the student network
7. Update the parameters and continue
```

3.3. Collaborative Learning Strategy

In the detection process, we adopt 3 different methods, which are the classification-based model (ShuffleNet [46]), segmentation-based model (WSFS), and the region-proposal-based model (lightweight Faster R-CNN), to predict the probability of fire smoke existing in an image. The final result is determined by the estimation of 3 model reasoning results, which are the classification result 1, the classification result 2, and the classification result 3, as shown in Figure 1.

Thus, the final prediction results can be calculated using a logistic regression, which is expressed as

$$P = \alpha + \sum_{i=1}^{n} \beta_i \times x_i \tag{14}$$

where P is the final prediction; x_i (i = 1, 2, . . . , n) is the deterministic variables related to the probability of a model; α is a constant; β_i (i = 1, 2, . . . , n) is a coefficient in Equation (14); and i represents the i-th model. In this paper, the 3 models are combined to predict the results. Therefore, we need to determine the 3 coefficients, β_1, β_2, and β_3, corresponding to the 3 inputs of models x_1, x_2, and x_3. Thus, we proposed a learning strategy, which is based on a stacking method [57], to fit Equation (14) in order to compute these coefficients.

We can perform a 5-fold cross validation on the training dataset, assuming that the training dataset includes 500 images which are divided into 5 subsets represented as $\mathrm{T_{data1}}$, $\mathrm{T_{data2}}$, $\mathrm{T_{data3}}$, $\mathrm{T_{data4}}$, and $\mathrm{T_{data5}}$, and that each subset includes 100 images. The testing dataset also includes 100 images. The models to be combined are represented as C, S, and R corresponding to the classification-based model, the segmentation-based model, and the region proposal-based model, respectively.

Eventually, datasets TC, TS, and TR are considered as the input value. Then, taking a real label as a guide, the models C, S, and R can learn the importance of the different models, and the models are assigned the weight of every algorithm (β_1, β_2, and β_3). The datasets DC, DS, and DR are further used to verify the models to achieve the best results. The detailed process of this strategy is described in Algorithm 3.

Algorithm 3. Collaborative Learning Strategy

```
Input: Five training subsets and one test dataset
Procedure Learning process for model x = C//x indicates C, S, and R.
1: Utilize $T_{data2}$-$T_{data5}$ to train model x. $T_{data1}$ is used as the test data, the obtained result is saved as TC1, and the verified result of the test dataset is saved as $D_{S1}$;
2: Utilize Tdata1 and $T_{data3}$-$T_{data5}$ to train model x. $T_{data2}$ is used as the test data, the obtained result is saved as TC2, and the verified result of the test dataset is saved as $D_{S2}$;
3: Continue to train and verify model x through the above process, and datasets TC3, TC4, and TC5, as well as $D_{S3}$, $D_{S4}$, and $D_{S5}$, are obtained;
4: Compute the average value of TC1 to TC5 stored as $T_{average;}$
5. The results of 5-fold cross validation is saved as TC = [TC1, TC2, TC3, TC4, TC5];
6. Calculate the average value of $D_{S1}$ to $D_{S5}$ and obtain DC;
7: Repeat this process for model S and R and obtain TS, TR, DS, and DR.
```

3.4. Grading Method Using Fuzzy Logic

The influencing factors of the level of fire or smoke include many aspects, such as the category (fire or smoke), region size, location, temperature, and wind, among others. In this paper, we only study the fire or smoke category (CT), region size (RS), and location (LT). The former category is predicted by Equation (14). The value of the region size is computed by Equation (16). The value of the location is divided into two types: dry forest or wet forest, which is determined by the color in the background of the input image, where yellow and green correspond to dry or wet forest, respectively.

The value of region size is expressed as:

$$S_{seg} = Area_{segmentaion} \cap Area_{region-proposal} \tag{15}$$

$$Region_{fire\ or\ smoke} = \begin{cases} FV,\ S_{seg} > \theta \\ 0,\ other \end{cases} \quad j = 1, 2, \ldots, m \tag{16}$$

where S_{seg} represents the size of segmentation. $Area_{fire\ or\ smoke}$ is calculated by the segmentation results using the WSFS. $Area_{region-proposal}$ is taken from the segmentation results using Lightweight Faster R-CNN; $Region_{fire\ or\ smoke}$ represents the region size of fire or smoke; FV is the corresponding value of this region; and FV ∈ [big, middle, small]. θ is an empiric value.

To assess the fire or smoke level, a fuzzy strategy is designed to weigh the variables CT, RS, and LT. This strategy is similar to that employed in our previous work [58,59]. The ambiguity $Level = f(i) \in [0, 1, 2, 3, 4, 5]$ guides the evaluation of the possibility of fire or smoke level. Here, the numbers correspond to the fire or smoke level as follows: 0 = very high, 1 = high, 2 = middle, 3 = low, 4 = very low, and 5 = no fire/no smoke. The CT is developed by defining extreme alarm (i.e., fire, H_fire), alarm (i.e., smoke, M_smoke), and normal situation (Norm) fuzzy sets. The RS is developed by defining large, middle, and small fuzzy sets. The LT is developed by defining high (yellow) and low (green) fuzzy sets. A conventional trapezoid or triangle is selected as the membership function, since they have few parameters and are easily optimized. The ambiguity term *Level* determines the final fire smoke level. The Mamdani model is applied as a reasoning engine [58,59], because it is suitable for capturing and coding expert-based knowledge.

4. Experiments

4.1. Dataset Description

We evaluated our approach on a developed forest fire smoke dataset, named FS-data, which was set up using some image search engines, such as Google and Baidu. The entire fire dataset contains 4856 samples distributed into 5 categories: large fire, small fire, dense smoke, small smoke, and other scenes, e.g., scenes of forests in different seasons. Some examples are shown in Figure 4.

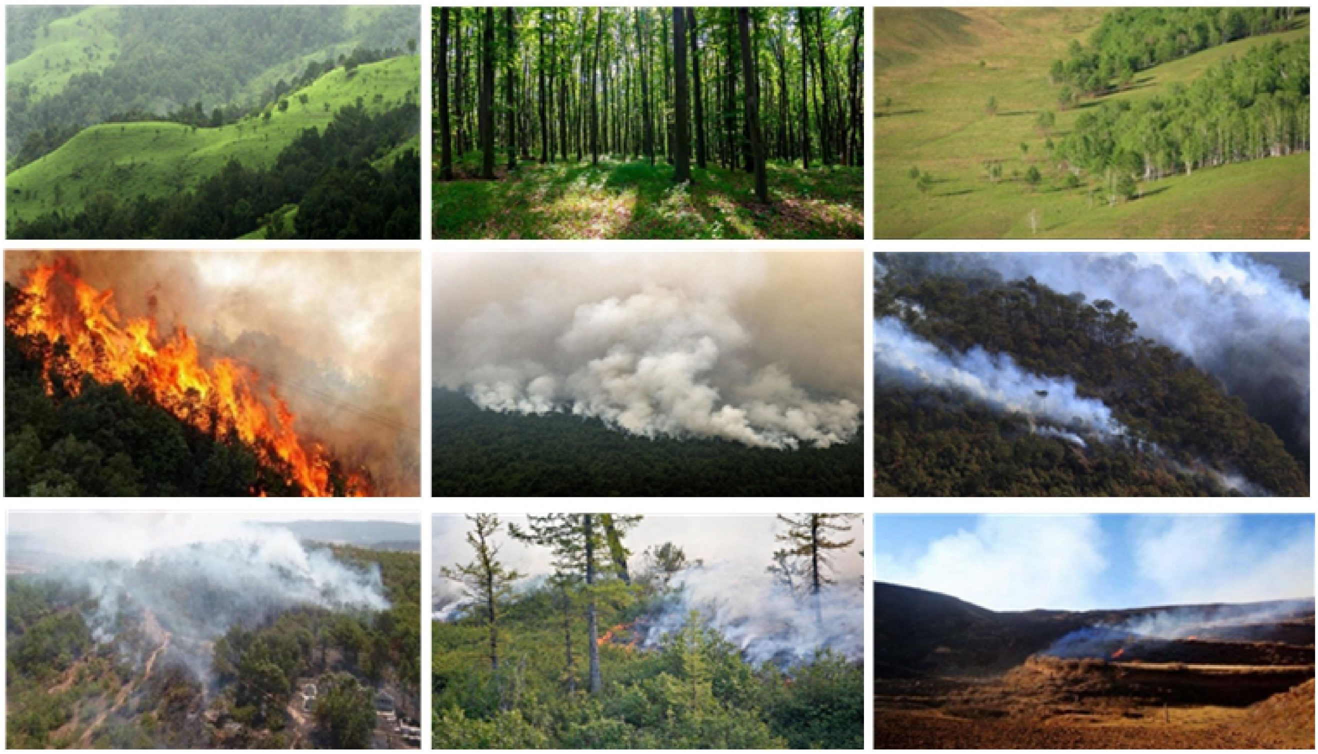

Figure 4. Samples of the forest fire dataset (the first row shows normal images (non-fire), and the other images show fire and smoke, which are, from top left to bottom right, large fire, dense smoke1, dense smoke2, light smoke, small fire, and other scenes).

The first and second rows in Figure 4 contain non-fire negative samples and positive samples for fires and smoke. Each image was labeled at the pixel level. These images had a resolution of 256 × 256 pixels, and 1323 images clearly exhibited visible fire or smoke and served as positive samples. The remaining 3533 images were non-fire negative samples. We divided the dataset into training and testing examples, as shown in Table 4.

Table 4. Distribution of training and testing examples in the datasets.

Dataset	**Training Examples**		**Testing Examples**	
	Positive	**Negative**	**Positive**	**Negative**
Subset A (large fire)	150	309	41	138
Subset B (small fire)	176	454	48	197
Subset C (dense smoke)	245	546	65	207
Subset D (light smoke)	280	755	78	267
Subset E (other scenes)	201	455	39	205

4.2. Performance Metrics

In this section, to obtain a better evaluation of our proposed method, 2 indicators, which are intersection over union (IOU) and average precision (AP), are used for the model performance evolution.

The intersection over union (IOU) [60,61] is a commonly used performance metric for semantic segmentation tasks. The decision prediction is based on the construction of a confusion matrix. True positives (TP) and true negatives (TN) belong to correct predictions. False positives (FP) are negative samples misreported as positive (fire), and false negatives (FN) are positive samples misreported as negative (non-fire). Our segmentation metrics are $IOU = TP/(TP + FN + FP)$.

The average precision (AP) is used to assess the classification accuracy [23]. AP is determined using the calculated area under the precision-recall curve to obtain a precise representation of the comprehensive model performance at different thresholds, particularly when the dataset contains large numbers of negative (non-fire) samples.

4.3. Experiment Environment

We conducted the experiments with the Python language under Pycharm, and the network model was implemented in Pytorch. We used PyDenseCRF (https://github.com/lucasb-eyer/pydensecrf, accessed on 12 September 2020) to construct a DenseCRF model. The simulations were conducted on a PC with an Intel Core i7-7820X CPU running Windows 10, with two GTX1080Ti GPUs (total 22 GHz) and 32 GB of RAM.

4.4. Evaluation of WSFS

In this section, we evaluate the performance of our WSFS approach on FS-data. The WSFS consists of LS-Net and AD-Net. Therefore, we conducted several experiments to validate the performance of LS-Net with WSL and RRS, as well as AD-Net.

4.4.1. The Region Detection Result Using LS-Net with WSL and RRS

We conducted experiments using FS-data to validate the performance of the weakly supervised loss (WSL) by visualizing the segmentation results (region detection) and comparing them with the evaluation index mentioned above. After training with the weakly supervised loss, LS-Net could roughly locate fire areas.

Figure 5 shows the visualization results of FS-data using our method to segment fire or smoke from an original image. In Figure 5, the red region represents the fire region (segmentation result), and the gray region represents the smoke area segmented by LS-Net. The first and second columns show the original images and pixel labels annotated manually, respectively. Columns 3–8 show that the segmentation regions change at different α values of 0.01, 0.02, 0.04, 0.06, 0.08, and 0.1, respectively. The results shown in Figure 5 indicate that LS-Net can predict fire regions under the guidance of the weakly supervised loss, and that the areas of fire regions and α values are positively correlated.

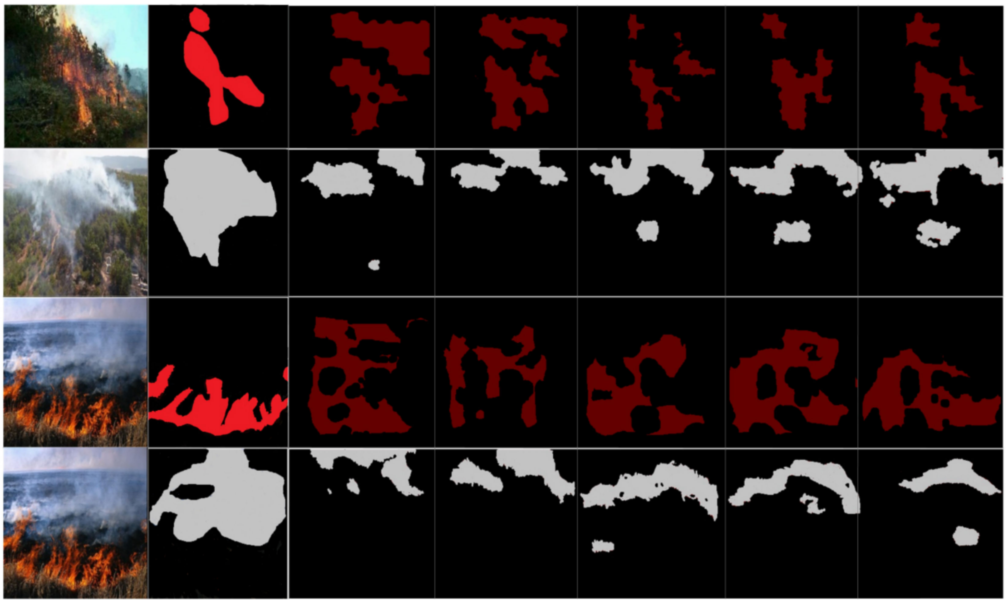

Figure 5. Segmentation results of the weakly supervised loss on FS-data.

We also used the IOU to evaluate the performance (segmentation accuracy). In Figure 6, the graph shows variations in IOU value at different values of α. The IOU value increases initially and then decreases with an increase in α. The IOU value is at a maximum at α = 0.004, which means that the segmentation accuracy is the best. The IOU then decreases with an increase in α, because an increasing number of pixels around the fire regions are wrongly identified as fires. This illustrates that the WSL can effectively identify fire or smoke pixels.

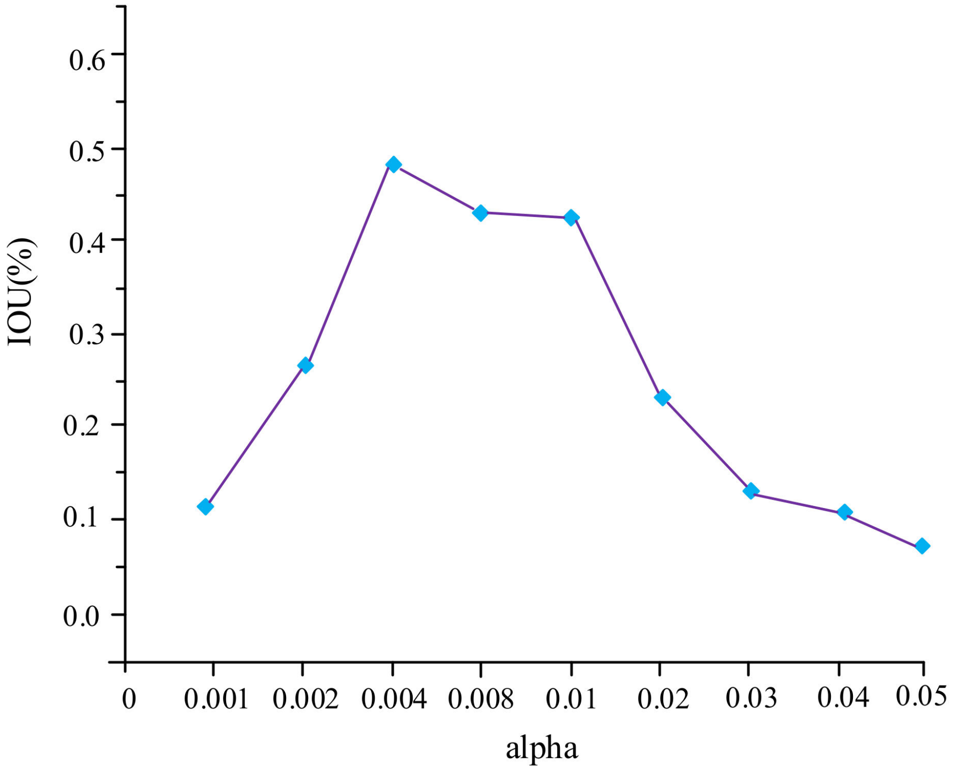

Figure 6. Changes in IOU at different α value.

However, the rough regions were only implemented using LS-Net. To further refine the detection region, we also implemented region-refining segmentation (RRS) on the FS-data. To evaluate the performance of RRS, the related algorithms were tested for comparison. These algorithms included 2 supervised learning segmentation methods, DeepLabV3+ [25] and U-Net [26]. The third method was the use of the weakly supervised loss alone. The fourth method adopted RRS to post-process the results of the weakly supervised loss. The same configurations were used in all experiments.

The detection results and metrics are shown in Figure 7. The first and second columns show the original images and pixel labels, respectively. The third and fourth columns show the results of a segmentation network trained in a supervised mode (U-Net) and a new supervised segmentation model, e.g., DeepLab V 3+. The segmentation results of LS-Net trained only using WSL are shown in column 5. The last column shows the segmentation results of WSFS trained successively using the 2-stage training strategy (WSL and RRS). The results show that our method, WSFS, has obtained a competitive performance in the segmentation result of fire or smoke.

The performance of these methods, evaluated using Boxplot descriptions, is shown in Figure 8. According to Figure 8, the U-Net method using pixel-wise labels obtains the maximum IOU for FS-data. The use of WSL alone returns the worst results, and RRS greatly improves the segmentation results of training with WSL; its results are closest to those of supervised training.

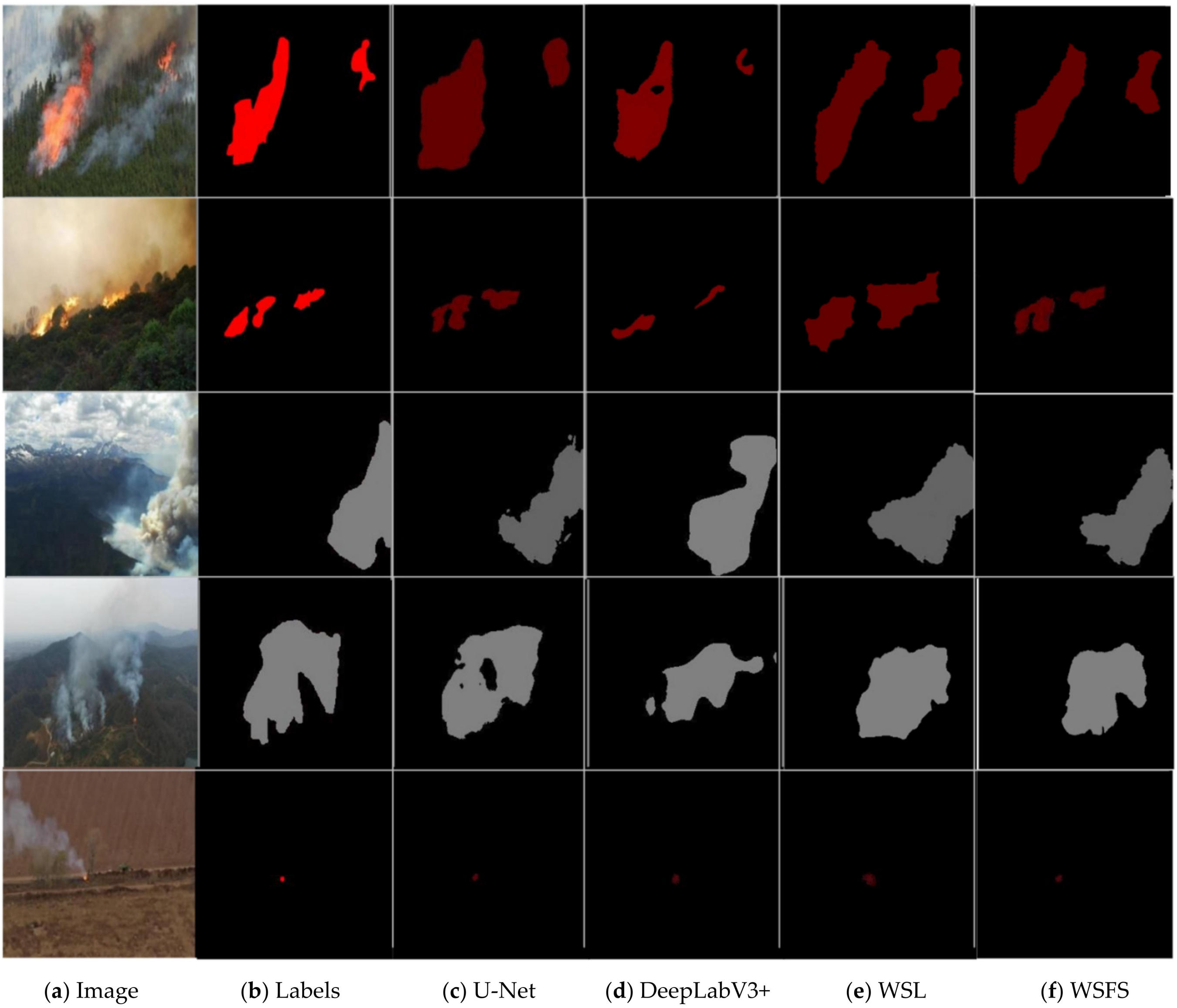

Figure 7. Visualization of the segmentation results of different methods. The (**a**,**b**) show the original images and pixel labels, respectively. (**c**–**f**) represent the experimental results using various methods which are U-Net, DeepLabV3+, WSL, and WSFS from the third column to the last column, respectively. (The red region represents the fire region, and the gray region represents the smoke area, which regions are segmented using the methods).

4.4.2. AD-Net

In this section, we performed controlled trials on FS-data to evaluate the classification performance of the AD-Net, while the ResNet18 and the ShuffleNet were selected as comparison methods. AP was used as an evaluation indicator, with values from the FS-data, as shown in Figure 9.

In Figure 9, the AP values of ResNet18, ShuffleNet, and our proposed method show better results for Subset A, Subset B and Subset C, in which the AP values are 99.9%,98.8%, and 98.9%, respectively. However, the AP value of these methods show worse results for Subset E and Subset D. The reason is that the image from Subset D and Subset E has a complex background, with low-contrast between objects and the background compared with other subsets. The AD-Net with the RSAM module built according to the result of LS-Net have shown the best performance for the five subsets.

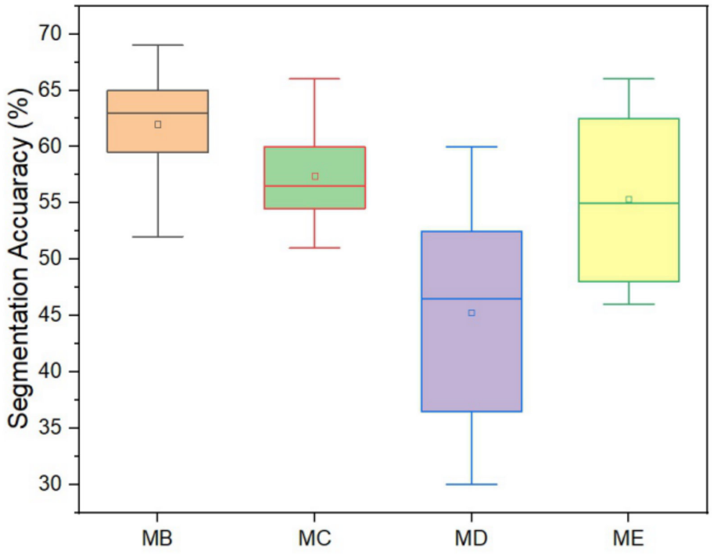

Figure 8. Boxplots of the segmentation results of different training methods: MB-U-Net, MC-DeepLabV3+, MD WSL (Ours), and ME- WSFS (Ours).

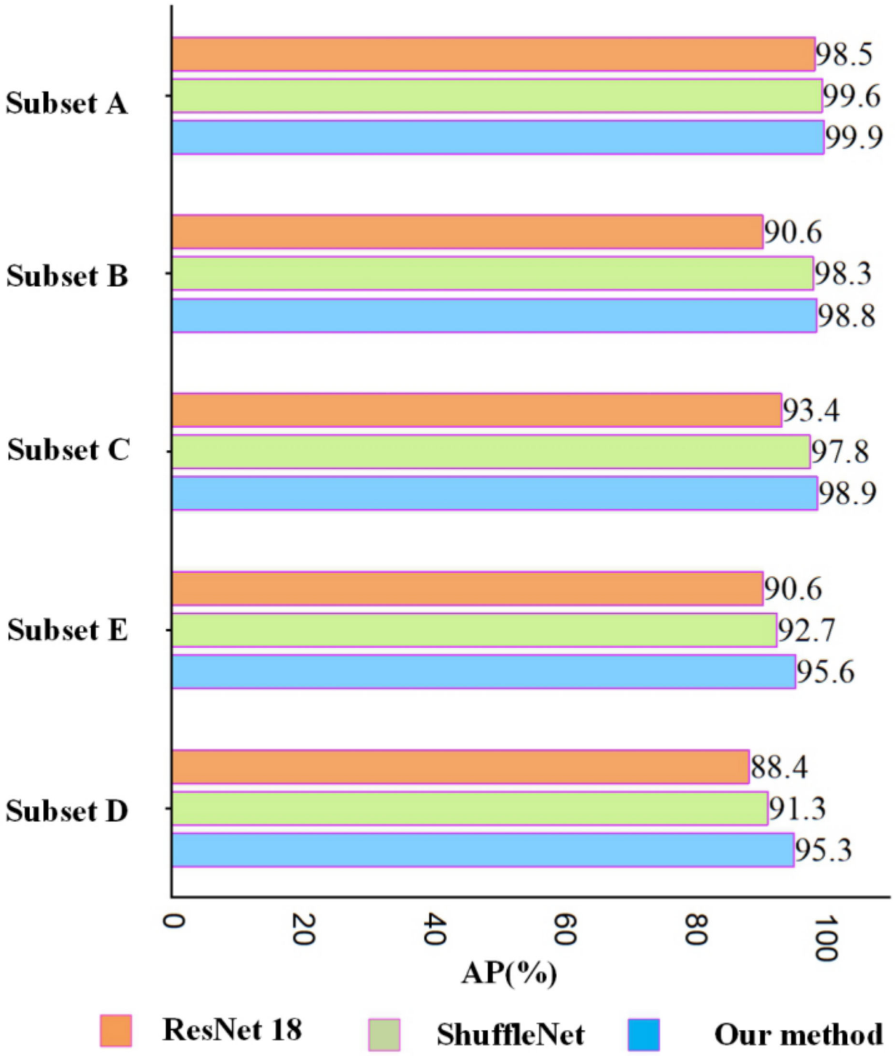

Figure 9. Comparison of the classification results of the different models on the FS-data.

4.5. Comparison with State-of-the Art Methods

We compared several state-of-the-art approaches with our method. For a fair comparison, the data and training setups were kept the same. The numerical experimental results for the fire dataset are summarized in Table 5, which involved the same experimental configurations.

Table 5. Comparison of our method with other methods using the FS-data.

Method	mIOU(%)	AP (%)
U-Net [26]	69.6	96.1
DeepLab v3+ [25]	67.4	98.4
Segmentation decision approach [62]		98.2
WSL (Ours)	60.5	97.9
WSFS (Ours)	68.8	98.6
FireDGWF (Ours)	70.2	99.6

Table 5 shows that the performance of region detection for DeepLabe V3+ is the worst of the tested methods and 67.4% in mIOU. The WSL methods achieve at 60.5% in mIOU. A small improvement of 8.3% is obtained by WSFS, due to RRS adopted by this method. However, the best performance of region detection, which reaches 70.2% in mIOU, is achieved by FireDGWF because the intersection between the detection regions obtained with the two methods (WSFS and lightweight Faster R-CNN) corresponds to the results.

The U-Net-based segmentation decision network produces an AP of 96.1%. The segmentation decision method based on DeepLab v3 + produces an average accuracy of 98.4%. Tabernik et al. [62] proposed a supervised 2-stage segmentation-based network that was trained with pixel labels to improve its classification performance. The WSL method obtains 97.9% in AP. Compared to these methods, the proposed method (WSFS) achieves 98.6% in AP. Moreover, FireGDWF shows a better ability to discriminate between fire-objects and non-fire objects, with an accuracy of 99.6% for 763 images from the FS-data. These results show that multi models can improve the performance of the algorithms.

4.6. Result for Grading of Forest Fire and Smoke

To evaluate forest fire severity, we designed a 3-input/1-output fuzzy evaluation system. These 3-inputs included the classification results, the segmentation region size, and the location. The single output was the forest fire smoke level: level 0, level 1, . . . , level 5. In our fuzzy evaluation system, the membership function must be decided first, which is similar to our previous work [58,59]. Once the membership degree of every input has been determined, a set of rules, which are defined in the system, as shown in Table 6, can be used to explain the evaluation results. Then, the detailed procedure for the example using the fuzzy evaluation system to grade the forest fire smoke level can be provided.

Table 6. Examples of the knowledge rules.

Rule No	Knowledge Rule
Rule1	*If* CT is H_fire and RS is Big and LT is Low, *then Level* = 0
Rule2	*If* CT is H_fire and RS is Big and LT is High, *then Level* = 1
Rule3	*If* CT is H_fire and RS is Middle and LT is Low, *then Level* = 1
Rule4	*If* CT is H_fire and RS is Middle and LT is Hight, *then Level* = 2
Rule5	*If* CT is H_fire and RS is Small and LT is Low, *then Level* = 2
Rule6	*If* CT is H_fire and RS is Small and LT is High, *then Level* = 1
Rule7	*If* CT is M_smoke and RS is Big and LT is Low, *then Level* = 2
Rule8	*If* CT is M_smoke and RS is Middle and LT is Low, *then Level* = 3
Rule9	*If* CT is M_smoke and RS is Small and LT is High, *then Level* = 3
Rule10	*If* CT is M_smoke and RS is Small and LT is Low, *then Level* = 4
Rule11	*If* CT is Norm, *then Level* = 5

When the system was run at a point of CT = fire, RS = 21,845 (pixel number), and LT = dry forest, and Rule 1 was activated (see Table 6), then, the result was put into the implication process and used to determine the output fuzzy set. The result given by the defuzzication process indicated that the output was *Level* = 0, as shown in Figure 10. Thus, we used the fuzzy evaluation system to determine different scenes of forest fire smoke, which could assess these fire smoke levels from level 1 to level 4. Level 5 indicates non-fire/non-smoke. These processes are shown in Figure 10.

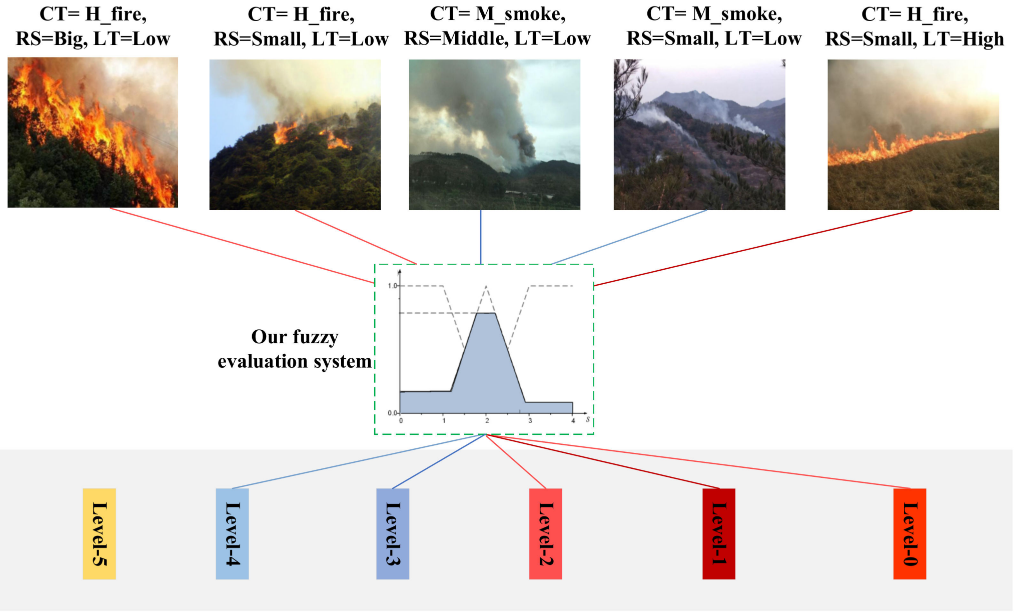
CT= H_fire, RS=Big, LT=Low
CT= H_fire, RS=Small, LT=Low
CT= M_smoke, RS=Middle, LT=Low
CT= M_smoke, RS=Small, LT=Low
CT= H_fire, RS=Small, LT=High
Our fuzzy evaluation system
1.0
0.0
0
1
2
3
4
5
Level-5
Level-4
Level-3
Level-2
Level-1
Level-0

Figure 10. An example of grading forest fire smoke using the fuzzy evolution system.

Additionally, the response time to region detection of forest fire using our FireDGWF was evaluated on FS-data. The experimental results are shown in Table 7. Our method, FireDGWF, included 2 processes: detection and grading. The detecting time of FireDGWF for an image is 0.128 s, which is the maximum time among the 3 methods, and the grading time for an image is 0.023 s. Therefore, a total response time to region detection and grading of forest fire is 0.151 s.

Table 7. Response time to detection and grading of forest fire using our method.

Method	Detecting Time	Grading Time
FireDGWF	0.128 s	0.023 s

4.7. Analysis and Discussion

Several experiments were conducted using our proposed approach, and its performance was evaluated, indicating that our approach performed competitively for the region detection and grading of forest fire and smoke. In the detection process, we proposed a weakly supervised fine segmentation method to effectively detect fire regions in a coarse-to-fine way, similar to a human-like recognition process. A two-stage weakly supervised learning strategy consisting of a weakly supersized loss and a region-refining segmentation algorithm was proposed to train the segmentation network. The negative and positive loss allowed the segmentation network to identify fire areas that differed from the background (forest) in training. Our method (WSFS) has achieved a better segmentation result, with a 68.8% mIOU. Experiments on the FS-data demonstrated that the region-refining segmentation algorithm obviously improved the performance of the segmentation network, without increasing the inference time. To further improve segmentation performance, we combined the Lightweight Faster R-CNN with the model to obtain a small improvement of 8.3%, for the mIOU values.

We also evaluated the decision network on the FS-data and compared it to state-of-the-art CNN-based models. Our method has obtained a competitive performance in the prediction of the category of an input image. This is because the RSAM module utilizes the segmentation result of LS-Net as an attention mask to cause AD-Net to focus more on object regions.

In the grading process, a 3-input and 1-out fuzzy system was developed to assess the level of forest fire, with the classification result, the region detection results, and the location as input values of the system.

The experimental results showed that multi models can improve the performance of our algorithm in terms of the detection accuracy and segmentation accuracy. We synthetically utilized three values: the classification result, the segmentation region, and the location. These were fed into our designed fuzzy evaluation system to obtain the forest fire smoke level, which could help humans to take proper precautionary measures in a timely manner.

5. Conclusions

In this paper, we introduced a collaborative region detection and grading framework for forest fire and smoke using a weakly supervised fine segmentation and a lightweight Faster R-CNN. This framework can detect the region and grade of fire and smoke in forests. To obtain the accurate region of fire and smoke, we propose a weakly supervised fire-segmentation model, which is trained using only image-level labels. A distillation strategy is used to reduce the complexity of the Faster R-CNN. Our proposed method has achieved an excellent performance and outperformed state-of-the-art CNN-based models in terms of detection accuracy (99.6%) and segmentation accuracy (70.2%). The final latency of our proposed method is only 151ms, which shows an excellent balance between detection performance and efficiency. Moreover, our fuzzy evaluation system can be used to assess the forest fire smoke level in a timely manner.

In future works, we plan to study an attention mechanism to improve the weakly supervised fine-segmentation method in the detection performance. To overcome the insufficient training data in a real-world application, a data augmentation technique, which is based on a Generative Adversarial Networks [63], will be introduced into our model. Furthermore, a possible improvement to our method is the incorporation of Multi-scale Adversarial Erase to substantially improve the detection rate. Additionally, we will work on developing a forest fire and smoke assessment system for risk level, which can identify different types, locations, sizes, and levels of fires or smoke. This system can track the evolution, spread, and grade of forest fires and smoke.

Author Contributions: J.P.: Data curation, Investigation, Resources, Method, Visualization, and Writing. X.O.: Supervision, Project Administration, and Funding Acquisition. L.X.: Conceptualization, Software, and Funding Acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Key project of National Social Science Foundation of China (grand number 17AGL018) and National Natural Science Foundation of China (grant number 71773031 and 21376091).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We thank MDPI for its linguistic assistance during the preparation of this manuscript.

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

References

1. Eugenio, F.C.; Santos, A.D.; Fiedler, N.C.; Ribeiro, G.A.; Silva, A.; Santos, Á.D.; Paneto, G.G.; Schettino, V.R. Applying GIS to develop a model for forest fire risk: A case study in Espírito Santo, Brazil. *J. Environ. Manag.* **2016**, *173*, 65–71. [CrossRef]
2. de Azevedo, A.R.; Alexandre, J.; Pessanha, L.S.P.; da S.T. Manhães, R.; de Brito, J.; Marvila, M.T. Characterizing the paper industry sludge for environmentally-safe disposal. *Waste Manag.* **2019**, *95*, 43–52. [CrossRef]
3. Çolak, E.; Sunar, F. Evaluation of forest fire risk in the Mediterranean Turkish forests: A case study of Menderes region, Izmir. *Int. J. Disaster Risk Reduct.* **2020**, *45*, 101479. [CrossRef]
4. Tang, X.; Machimura, T.; Li, J.; Liu, W.; Hong, H. A novel optimized repeatedly random undersampling for selecting nega-tive samples: A case study in an SVM-based forest fire susceptibility assessment. *J. Environ. Manag.* **2020**, *271*, 1–13. [CrossRef] [PubMed]
5. Guha-Sapir, D.; Philippe, H. Estimating Populations Affected by Disasters: A Review of Methodological Issues and Research Gaps. In *Global Sustainable Report*; School of Public Health Université Catholique de Louvain: Brussels, Belgium, 2015; pp. 1–15.
6. Marshall, S.C.A.D. Fire detection using smoke and gas sensors. *Fire Saf. J.* **2007**, *42*, 507–515.
7. Qiu, X.; Wei, Y.; Li, N.; Guo, A.; Zhang, E.; Li, C.; Peng, Y.; Wei, J.; Zang, Z. Development of an early warning fire detection system based on a laser spectroscopic carbon monoxide sensor using a 32-bit system-on-chip. *Infrared Phys. Techn.* **2019**, *96*, 44–51. [CrossRef]
8. Varela, N.; Díaz-Martinez, J.L.; Ospino, A.; Zelaya, N.A.L. Wireless sensor network for forest fire detection. *Procedia Comput. Sci.* **2020**, *175*, 435–440. [CrossRef]
9. 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.* **2020**, *149*, 1–16. [CrossRef]
10. Fernandes, A.M.; Utkin, A.B.; Lavrov, A.V.; Vilar, R.M. Development of neural network committee machines for automatic forest fire detection using lidar. *Pattern Recognit.* **2004**, *37*, 2039–2047. [CrossRef]
11. Chen, T.-H.; Wu, P.-H.; Chiou, Y.-C. An early fire-detection method based on image processing. In Proceedings of the 2004 International Conference on Image Processing, Singapore, 24–27 October 2004; Volume 3, pp. 1707–1710.
12. Marbach, G.; Loepfe, M.; Brupbacher, T. An image processing technique for fire detection in video images. *Fire Saf. J.* **2006**, *41*, 285–289. [CrossRef]
13. Habiboğlu, Y.H.; Günay, O.; Çetin, A.E. Covariance matrix-based fire and flame detection method in video. *Mach. Vis. Appl.* **2012**, *23*, 1103–1113. [CrossRef]
14. Çelik, T.; Demirel, H. Fire detection in video sequences using a generic color model. *Fire Saf. J.* **2009**, *44*, 147–158. [CrossRef]
15. Tao, C.; Jian, Z.; Pan, W. Smoke Detection Based on Deep Convolutional Neural Networks. In Proceedings of the 2016 International Conference on Industrial Informatics—Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII), Wuhan, China, 3–4 December 2016; pp. 150–153.
16. Muhammad, K.; Ahmad, J.; Mehmood, I.; Rho, S.; Baik, S.W. Convolutional Neural Networks Based Fire Detection in Sur-veillance Videos. *IEEE Access* **2018**, *6*, 18174–18183. [CrossRef]
17. Yin, Z.; Wan, B.; Yuan, F.; Xia, X.; Shi, J. A Deep Normalization and Convolutional Neural Network for Image Smoke Detec-tion. *IEEE Access* **2017**, *5*, 18429–18438. [CrossRef]
18. Sharma, J.; Granmo, O.-C.; Goodwin, M.; Fidje, J.T. Deep Convolutional Neural Networks for Fire Detection in Images. *Commun. Comput. Inf. Sci.* **2017**, *744*, 183–193. [CrossRef]
19. Hu, Y.; Lu, X. Real-time video fire smoke detection by utilizing spatial-temporal ConvNet features. *Multimed. Tools Appl.* **2018**, *77*, 29283–29301. [CrossRef]
20. Kaabi, R.; Bouchouicha, M.; Mouelhi, A.; Sayadi, M.; Moreau, E. An Efficient Smoke Detection Algorithm Based on Deep Belief Network Classifier Using Energy and Intensity Features. *Electronics* **2020**, *9*, 1390. [CrossRef]
21. Islam, R.; Amiruzzaman; Nasim, S.; Shin, J. Smoke Object Segmentation and the Dynamic Growth Feature Model for Video-Based Smoke Detection Systems. *Symmetry* **2020**, *12*, 1075. [CrossRef]
22. Lu, P.; Zhao, Y.; Xu, Y. A Two-Stream CNN Model with Adaptive Adjustment of Receptive Field Dedicated to Flame Region Detection. *Symmetry* **2021**, *13*, 397. [CrossRef]
23. Zhao, E.; Liu, Y.; Zhang, J.; Tian, Y. Forest Fire Smoke Recognition Based on Anchor Box Adaptive Generation Method. *Electronics* **2021**, *10*, 566. [CrossRef]
24. Xu, R.; Lin, H.; Lu, K.; Cao, L.; Liu, Y. A Forest Fire Detection System Based on Ensemble Learning. *Forests* **2021**, *12*, 217. [CrossRef]
25. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
26. 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, Cham, Switzerland, 5–9 October 2015; pp. 234–241.
27. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651.
28. Alkhatib, A.A.A. A Review on Forest Fire Detection Techniques. *Int. J. Distrib. Sens. Networks* **2014**, *10*, 1–14. [CrossRef]
29. Yuan, C.; Liu, Z.; Zhang, Y. Vision-based forest fire detection in aerial images for firefighting using UAVs. In Proceedings of the 2016 International Conference on Unmanned Aircraft Systems (ICUAS), Arlington, VA, USA, 7–10 June 2016; pp. 1200–1205.

30. Ren, X.; Li, C.; Ma, X.; Chen, F.; Wang, H.; Sharma, A.; Gaba, G.; Masud, M. Design of Multi-Information Fusion Based Intelligent Electrical Fire Detection System for Green Buildings. *Sustainability* **2021**, *13*, 3405. [CrossRef]
31. Zhao, Y.; Ma, J.; Li, X.; Zhang, J. Saliency Detection and Deep Learning-Based Wildfire Identification in UAV Imagery. *Sensors* **2018**, *18*, 712. [CrossRef]
32. Namozov, A.; Cho, Y.I. An Efficient Deep Learning Algorithm for Fire and Smoke Detection with Limited Data. *Adv. Electr. Comput. Eng.* **2018**, *18*, 121–128. [CrossRef]
33. Xu, Z.; Guo, Y.; Saleh, J.H. Tackling Small Data Challenges in Visual Fire Detection: A Deep Convolutional Generative Ad-versarial Network Approach. *IEEE Access* **2020**, *9*, 3936–3946. [CrossRef]
34. Muhammad, K.; Ahmad, J.; Lv, Z.; Bellavista, P.; Yang, P.; Baik, S.W. Efficient Deep CNN-Based Fire Detection and Locali-zation in Video Surveillance Applications. *IEEE Trans. Syst. Man Cybern. Syst.* **2019**, *49*, 1419–1434. [CrossRef]
35. Muhammad, K.; Khan, S.; Elhoseny, M.; Ahmed, S.H.; Baik, S.W. Efficient Fire Detection for Uncertain Surveillance Environment. *IEEE Trans. Ind. Inform.* **2019**, *15*, 3113–3122. [CrossRef]
36. 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]
37. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
38. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
39. Zhou, Z.-H. A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **2018**, *5*, 44–53. [CrossRef]
40. Pathak, D.; Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Multi-Class Multiple Instance Learning. *arXiv* **2014**, arXiv:1412.7144.
41. Pinheiro, P.O.; Collobert, R. From image-level to pixel-level labeling with Convolutional Networks. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1713–1721.
42. Boykov, Y.Y.; Jolly, M.-P. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In Proceedings of the Eighth IEEE International Conference on Computer Vision ICCV 2001, Vancouver, BC, Canada, 7–14 July 2001; pp. 105–112.
43. Rother, C.; Kolmogorov, V.; Blake, A. "GrabCut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.* **2004**, *23*, 309–314. [CrossRef]
44. Krähenbühl, P.; Koltun, V. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. *Adv. Neural Inf. Process. Syst.* **2012**, *24*, 109–117.
45. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *arXiv* **2014**, arXiv:1412.7062.
46. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
47. Khoreva, A.; Benenson, R.; Hosang, J.; Hein, M.; Schiele, B. Simple Does It: Weakly Supervised Instance and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 876–885.
48. Hsu, C.C.; Hsu, K.J.; Tsai, C.C.; Lin, Y.Y.; Sinica, A. Weakly Supervised Instance Segmentation using the Bounding Box Tightness Prior. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 6582–6593.
49. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
50. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 21–37.
51. Fu, C.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. DSSD: Deconvolutional Single Shot Detector. *arXiv* **2017**, arXiv:1701.06659.
52. Zhang, S.; Wen, L.; Lei, Z.; Li, S.Z. RefineDet++: Single-Shot Refinement Neural Network for Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 674–687. [CrossRef]
53. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
54. 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.
55. Bearman, A.; Russakovsky, O.; Ferrari, V.; Fei-Fei, L. What's the Point: Semantic Segmentation with Point Supervision. In *European Conference on Computer Vision*; Springer: Amsterdam, The Netherlands, 2016; pp. 549–565.
56. Laradji, I.H.; Rostamzadeh, N.; Pinheiro, P.O.; Vazquez, D.; Schmidt, M. Where are the Blobs: Counting by Localization with Point Supervision. In *Computer Vision—ECCV 2018*; Springer: Munich, Germany, 2018; pp. 560–576.
57. Kim, J.; Kim, J.; Kwak, N. StackNet: Stacking feature maps for Continual learning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, Seattle, WA, USA, 14–19 June 2020; pp. 1–8. [CrossRef]
58. Xu, L.; He, X.-M.; Li, X.X.; Pan, M. A machine-vision inspection system for conveying attitudes of columnar objects in packing processes. *Measurement* **2016**, *87*, 255–273. [CrossRef]

59. Xu, L.; Wang, Y.; Li, R.; Yang, X.; Li, X. A Hybrid Character Recognition Approach Using Fuzzy Logic and Stroke Bayesian Program Learning with Naïve Bayes in Industrial Environments. *IEEE Access* **2020**, *8*, 124767–124782. [CrossRef]
60. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Garcia-Rodriguez, J. A review on deep learning tech-niques applied to semantic segmentation. *arXiv* **2017**, arXiv:1704.06857.
61. Branco, P.; Torgo, L.; Ribeiro, R. A Survey of Predictive Modeling on Imbalanced Domains. *ACM Comput. Surv.* **2016**, *49*, 1–50. [CrossRef]
62. Tabernik, D.; Ela, S.; Skvar, J.; Skoaj, D. Segmentation-based deep-learning approach for surface-defect detection. *J. Intell. Manuf.* **2020**, *31*, 759–776. [CrossRef]
63. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Adv. Neural Inf. Process. Syst.* **2014**, *3*, 2672–2680. [CrossRef]