remote sensing

MDPI

Article

Weakly Supervised Forest Fire Segmentation in UAV Imagery Based on Foreground-Aware Pooling and Context-Aware Loss

Junling Wang [1,2], Yupeng Wang [3], Liping Liu [2], Hengfu Yin [2], Ning Ye [1] and Can Xu [3,*]

[1] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China; wangjunling@njfu.edu.cn (J.W.); yening@njfu.edu.cn (N.Y.)
[2] State Key Laboratory of Tree Genetics and Breeding, Research Institute of Subtropical Forestry, Chinese Academy of Forestry, Hangzhou 311400, China; huzk@caf.ac.cn (L.L.); fyin@caf.ac.cn (H.Y.)
[3] College of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; yupengwang@njust.edu.cn
* Correspondence: 220106011137@njust.edu.cn

Abstract: In recent years, tragedies caused by forest fires have been frequently reported. Forest fires not only result in significant economic losses but also cause environmental damage. The utilization of computer vision techniques and unmanned aerial vehicles (UAVs) for forest fire monitoring has become a primary approach to accurately locate and extinguish fires during their early stages. However, traditional computer-based methods for UAV forest fire image segmentation require a large amount of pixel-level labeled data to train the networks, which can be time-consuming and costly to acquire. To address this challenge, we propose a novel weakly supervised approach for semantic segmentation of fire images in this study. Our method utilizes self-supervised attention foreground-aware pooling (SAP) and context-aware loss (CAL) to generate high-quality pseudo-labels, serving as substitutes for manual annotation. SAP collaborates with bounding box and class activation mapping (CAM) to generate a background attention map, which aids in the generation of accurate pseudo-labels. CAL further improves the quality of the pseudo-labels by incorporating contextual information related to the target objects, effectively reducing environmental noise. We conducted experiments on two publicly available UAV forest fire datasets: the Corsican dataset and the Flame dataset. Our proposed method achieved impressive results, with IoU values of 81.23% and 76.43% for the Corsican dataset and the Flame dataset, respectively. These results significantly outperform the latest weakly supervised semantic segmentation (WSSS) networks on forest fire datasets.

Keywords: forest fire; UAV imagery; intelligent forestry; weakly supervised learning; semantic segmentation

Citation: Wang, J.; Wang, Y.; Liu, L.; Yin, H.; Ye, N.; Xu, C. Weakly Supervised Forest Fire Segmentation in UAV Imagery Based on Foreground-Aware Pooling and Context-Aware Loss. *Remote Sens.* **2023**, *15*, 3606. https://doi.org/10.3390/rs15143606

Academic Editors: Houbing Song, Weipeng Jing, Huaiqing Zhang, Hua Sun, Qiaolin Ye and Fu Xu

Received: 8 June 2023
Revised: 11 July 2023
Accepted: 17 July 2023
Published: 19 July 2023

1. Introduction

Forest fires pose a significant threat to forest security, leading to atmospheric pollution, global warming, and the loss of animal habitats. This, in turn, results in substantial economic losses and environmental damage [1]. According to statistics [2], forest fires destroy approximately 350–450 million hectares of the soil environment each year. Manual firefighting efforts require a significant allocation of human and material resources and can sometimes result in casualties during the firefighting process [3]. Consequently, the timely detection and prediction of forest fires have garnered considerable research interest in recent years.

The initial method of monitoring forest fires involved the use of manual watchtowers, where towers were positioned on elevated ground, and personnel were stationed to observe the forest environment. However, this approach necessitates significant human and material resources, and the monitoring angle can easily be obstructed by branches and leaves in the forest, making monitoring challenging and unresponsive. With the advancement of hardware devices, the utilization of unmanned aerial vehicles (UAVs) has emerged as a

more cost-effective and flexible means of acquiring images, replacing traditional watch-tower patrols for forest fire monitoring [4,5]. Concurrently, deep learning (DL) has become a fundamental tool for various computer vision tasks, finding applications in autonomous driving, vehicle segmentation, agricultural segmentation, and cityscape segmentation [6–8]. Consequently, the integration of DL techniques into the processing of forest fire images captured by UAVs has become the preferred solution for forest fire detection.

Previous research has predominantly approached fire detection in forest environments as a semantic segmentation problem. Wang et al. [9] conducted semantic segmentation of aerial forest fire images by comparing various classical semantic segmentation models. They selected the most applicable model for forest fire segmentation scenarios to accurately identify fire locations and provide valuable data for firefighting and fire analysis. Classical deep learning-based semantic segmentation networks are typically trained on datasets with pixel-level labels. These models predict the class to which each pixel in the image belongs, enabling the segmentation of target objects in the image. Utilizing a DL framework for semantic segmentation of forest fire images has shown superior performance compared to traditional methods [10,11]. Choi et al. [12] implemented intermediate skip connections using residual networks for forest fire detection on the FiSmo dataset, yielding promising results. However, the aforementioned experimental datasets did not employ aerial forest fire images captured by UAVs. This is noteworthy as UAV-captured images may exhibit dead spots due to the fixed location of surveillance cameras. When employing UAVs for forest fire monitoring, capturing the initial fire spot from a high altitude allows for subsequent close-range photography of the fire location and its surroundings from multiple angles, aiding firefighting efforts.

In this paper, we propose a novel pool module called self-supervised attention foreground-aware pooling (SAP) for WSSS and introduce a new context-aware loss (CAL) for training WSSS network models in the context of forest fire image segmentation. When monitoring forest fires using UAVs, not all captured images depict actual fire incidents. Therefore, before performing forest fire image segmentation, we need to classify whether the images contain fires or not. The subsequent fire segmentation is only performed on the images containing fire points. Thus, our proposed model's classification network has two tasks: (1) classifying images into two categories, with fire and without fire, and (2) conducting various operations such as feature extraction on the images with fire to prepare for pseudo-label generation. In weakly supervised forest fire image segmentation tasks, the bounding box can only provide an approximate location of the foreground but cannot accurately capture the target's exact boundaries, while the bounding box always includes the entire foreground. On the other hand, CAM accurately reflects the specific location of the object during feature extraction. Our SAP method leverages CAM and bounding boxes to generate high-quality pseudo-labels for training network models. However, noise can be introduced during the pseudo-label generation process due to various factors. To address this, we introduce CAL, which utilizes contextual information from the target's image to correct noise in the pseudo-labels, making the entire network less susceptible to such noise. We conducted extensive experiments on two publicly available UAV forest fire datasets, namely the Flame dataset [13] and the Corsican dataset [14]. The results demonstrate that our proposed method outperforms state-of-the-art WSSS methods for UAV forest fire image segmentation.

Our work in this study can be summarized as follows:

- The proposed SAP pooling method is a WSSS pooling method that combines CAM and bounding box annotation to generate pseudo-labels;
- Context-aware loss (CAL) is proposed to generate high-quality pseudo-labels. As an alternative to manual annotation, CAL uses the contextual information of the objects in the image to correct the noise in the pseudo-labels and jointly uses classifier weights to reduce the effect of noise on pseudo-label generation;

- Comparative experiments and ablation experiments on the Flame dataset and the Corsican dataset show that our method outperforms existing models in WSSS of aerial forest fire images at different scales.

2. Related Work

2.1. Fully Supervised Semantic Segmentation

The utilization of UAVs for real-time monitoring of forest fires and capturing aerial forest fire images has emerged as a prominent approach in contemporary forest fire prevention, combining remote sensing, deep learning, and computer vision techniques [15,16].

Khryashchev et al. [17] proposed a convolutional neural network capable of automatically detecting forest fires in high-resolution images. Wang et al. [18] developed a forest fire early segmentation network model called smoke-Unet, which combines attention mechanisms and residual blocks based on an improved U-net architecture. However, the aforementioned methods are fully supervised models that rely on datasets with pixel-level labels for segmenting forest fire images. The process of pixel-level annotation often demands substantial manpower and resources. Annotating forest fire images at the pixel level presents additional challenges compared to other types of images. Aerial forest fire images captured by UAVs are taken from high altitudes and encounter various disturbances in the forest environment where the fire is located, such as occlusions, which make it difficult to accurately distinguish the fire boundaries. Consequently, the process of labeling such images becomes significantly challenging. Although fully supervised semantic segmentation models have demonstrated good performance in segmenting aerial forest fire images, their effectiveness heavily relies on the quality of the dataset labeling.

2.2. Weakly Supervised Semantic Segmentation

The high cost of data annotation poses a significant challenge in the development of semantic segmentation models. To address this challenge, researchers have explored the integration of semi-supervised learning [19] and weakly supervised learning [20] approaches into existing fully supervised semantic segmentation models.

The concept behind weakly supervised semantic segmentation (WSSS) models is to train the model using image-level labels instead of pixel-level labels, aiming to reduce annotation costs. Consequently, weakly supervised learning has emerged as a prominent research direction. In a study by Su et al. [21], a context decoupling augmentation (CDA) method was proposed, which actively removes dependencies between objects and their contextual information by modifying the inherent context in which the target object appears. Amaral et al. [22] employed class activation mapping (CAM) and conditional random fields (CRF) to detect fire masks at the pixel level, applying these techniques to the Corsican dataset as a forest fire segmentation dataset. Although limited studies have applied WSSS techniques to forest fire image segmentation, their application has the potential to assist firefighters in protecting forests. Acquiring aerial forest fire datasets with only image-level labels is considerably less challenging compared to obtaining datasets with precise pixel-level labels. Utilizing WSSS methods can help overcome this challenge while achieving satisfactory segmentation accuracy.

3. Method

We propose an SAP and CAL method for weakly supervised semantic segmentation of UAV forest fire images. Our method involves three stages for forest fire image segmentation, and the entire network structure is illustrated in Figure 1:

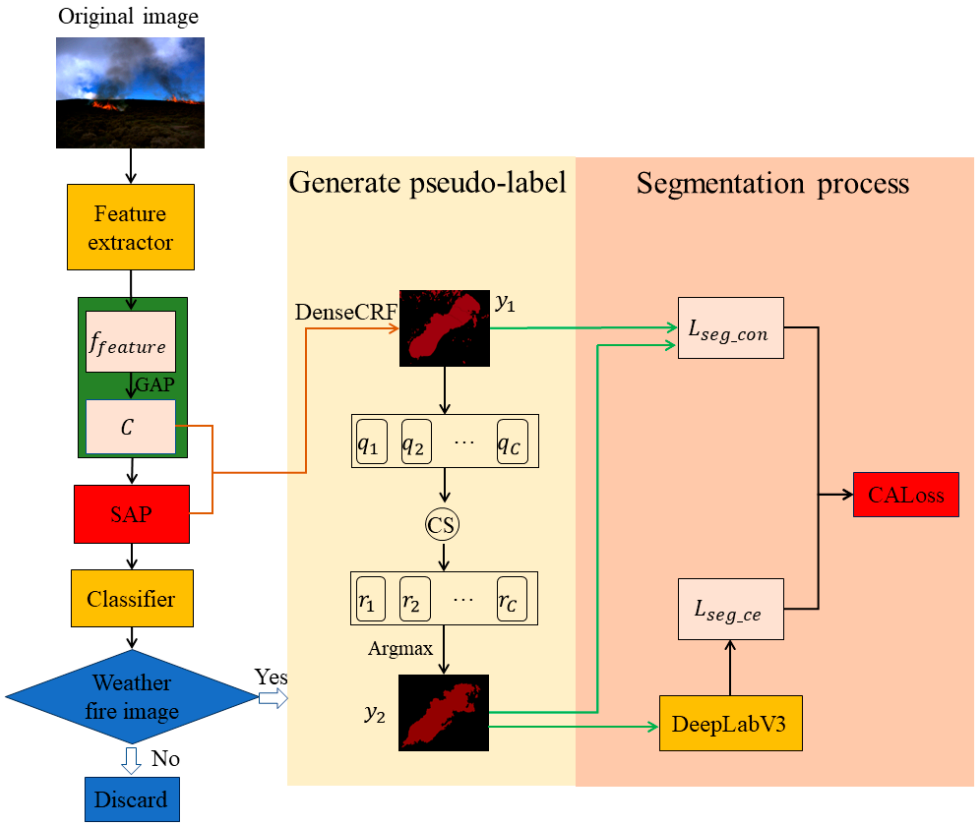

Figure 1. Diagram of the entire network structure.

1. For any of the input images, the image features $f_{feature}$ are first extracted by a feature extractor, and the C is generated by GAP. $f_{feature}$ and C are then fed into the SAP module and a classifier is used to classify the image. If the image is a forest fire image, then continue, otherwise terminate;
2. When this image is a forest fire image, we generate the initial pseudo-label y_1 and the final pseudo-label y_2 from the generated C and the background attention map in the SAP module after DenseCRF;
3. To correct the noise in the pseudo-labels, we introduce CAL in segmentation, which uses contextual semantic information about the target object throughout the image. CAL consists of the contrast loss of the initial pseudo-label y_1 and the final pseudo-label y_2 and the CELoss of y_2 in DeeplabV3, which effectively constrains the human and environmental noise during the generation of the pseudo-labels.

3.1. Self-Supervised Attention Foreground-Aware Pooling

Our classification network has three main components: a feature extractor, the selective pixel correlation module (SPCM) and a 2-way classifier. The overall structure of the classification network is shown in Figure 2. We used the pretrained ResNet50 model as a feature extractor. The feature extractor was trained on the ImageNet dataset with good results and can effectively extract features from images. An image input to a classification network is passed through the feature extractor to generate a feature map $f_{feature}$ and through GAP to generate CAM C. $f_{feature}$ and C are then fed into the SPCM module to obtain $C_{modified}$:

$$C_{modified} = S(f_{feature}, C). \quad (1)$$

where $S(\cdot)$ denotes the overall algorithm in the SPAM module. To obtain the query q_j, we compute a weighted average of the binary masks M representing category-specific

information using $C_{modified}$. Compute a background attentional map A_{back} and a foreground attentional map A_{fore} using the query q_j:

$$q_j = \frac{M_j \cdot C_{modified\_j}}{M_j}, \tag{2}$$

$$A_{back\_j} = \begin{cases} \text{ReLU}(\frac{f_{feature\_j}}{||f_{feature\_j}||} \cdot \frac{q_j}{||q_j||}), j \in \beta \\ 1 \quad , j \notin \beta \end{cases}, \tag{3}$$

$$A_{back} = \frac{1}{N}\sum_{j=0}^{N} A_{back\_j}, \tag{4}$$

$$A_{fore} = 1 - A_{back}.A_{fore} = 1 - A_{back}. \tag{5}$$

where N denotes the total number of q_j, $||\cdot||$ denotes L2 normalization, and Equation (4) quantifies the probability that pixel j is background by calculating the similarity in bounding box β via q_j and controlling A_{back} in [0, 1] via the ReLU function. The closer the pixel is to the background, the closer the value of A_{back} is to 1. Finally, a softmax classifier is applied to each bounding box for foreground features r_i and query q_j, quantifying the probability of each pixel point being a foreground. Additionally, reductive contrast loss is applied to CAMs and masks, while CE loss is used for iterative training on the foreground features r_i and queries q_j.

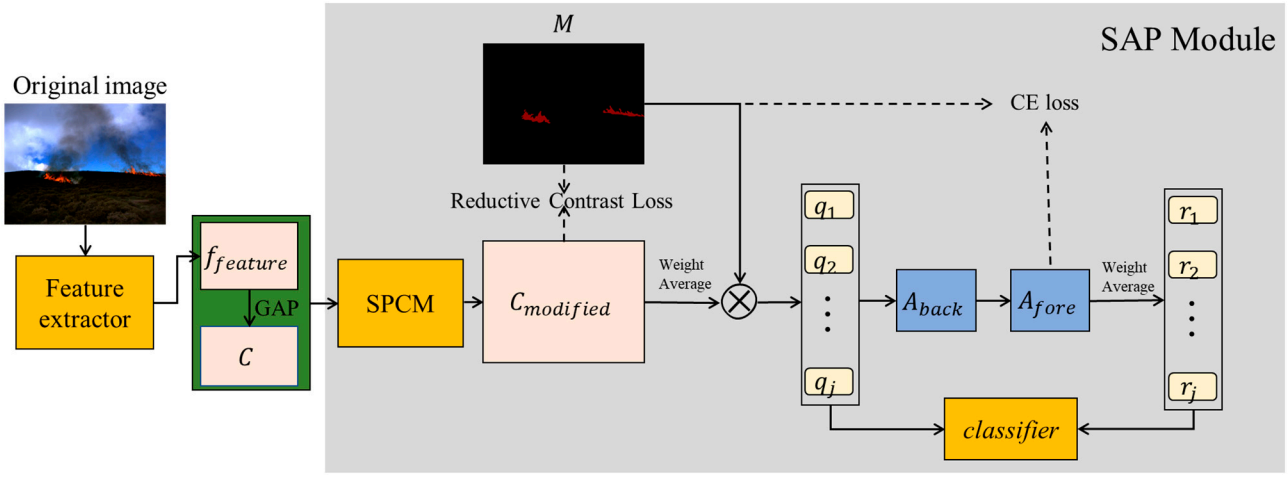

Figure 2. SAP module overall structure diagram.

3.1.1. Selective Pixel Correlation Module

Figure 3 illustrates the overall structure of the SPCM. To efficiently extract channel features from UAV forest fire images, both mean pooling and maximum pooling are utilized. A shared MLP consists of a two-layer neural network (MLP), the number of neurons in the first layer is N/r (where r is the decrement rate), the activation function is Relu, and the number of neurons in the second layer is N, which is shared by the two layers of neural networks. Two inputs $f_{feature}$ and C to the SPCM are performed for the input feature map $f_{feature}$ in the SPCM; average pooling layer compresses the spatial dimension of the feature map while aggregating spatial information, while the maximum pooling layer is used to extract image features and image channel feature information. Generate attention map A_f

after passing the output features of the average pooling and maximum pooling methods to the shared network:

$$\begin{aligned} \mathrm{A}_f\left(f_{feature}\right) &= \sigma(\mathrm{MLP}(\mathrm{AP}(f_{feature})) + \mathrm{MLP}(\mathrm{MP}(f_{feature}))) \\ &= \sigma(\mathrm{W}_1(\mathrm{W}_0\left(f^c_{feature\_avg}\right)) + \mathrm{W}_1(\mathrm{W}_0\left(f^c_{feature\_max}\right))). \end{aligned} \tag{6}$$

where $y^c_{feature\_avg}$ and $y^c_{feature\_max}$ represent the features after the average pooling layer and the maximum pooling layer, respectively, σ represents the sigmoid function, $AP(\cdot)$ and $MP(\cdot)$ represent the average pooling and maximum pooling, respectively. After $y^c_{feature\_avg}$ and $y^c_{feature\_max}$ pass through the shared network, the final channel attention map A_f is generated. The essence of a shared network is a multilayer perceptron (MLP) and contains a hidden layer for reducing the number of parameters. The feature vectors after the shared network are summed element-wise. For A_f, a $HW \times HW$ lift is obtained by matrix multiplication after 2 parallel 1×1 convolutions respectively and is weighted and summed with the original CAM C matrix to obtain the modified CAM $C_{modified}$.

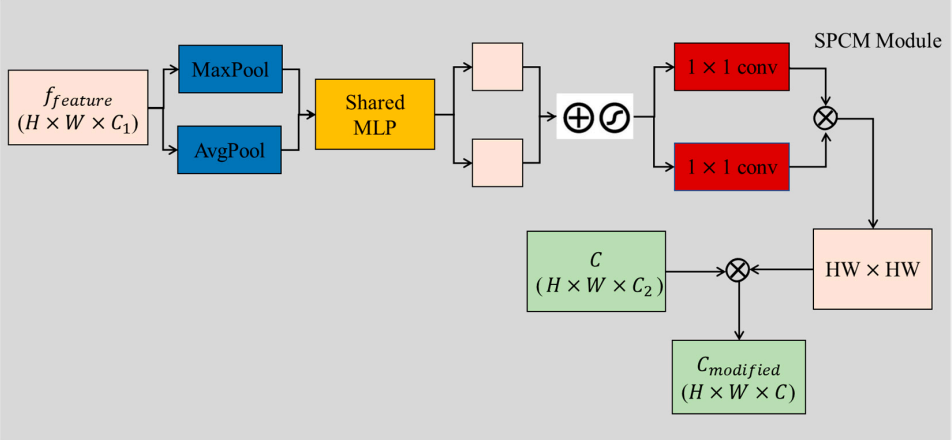

Figure 3. SPCM overall structure diagram.

The SPCM improves on traditional attention mechanisms while retaining the most primitive CAM features, while the features at the bottom of each pixel are required to evaluate the similarity between pixel features using cosine similarity:

$$\theta(x_i, x_j) = \frac{\sum(x_i \times x_j)}{\sqrt{\sum x_i^2} \times \sqrt{\sum x_j^2}}. \tag{7}$$

where i and j indicate the index of a pixel at a spatial location. The normalized cosine similarity is then used and multiplied with the original CAM C to calculate the similarity between the current pixel and other pixels in the feature space; the modified CAM $C_{modified}$ can be expressed as:

$$C_{modified} = \frac{1}{C(x_i)} \sum_{\forall j} \mathrm{ReLU}\left(\frac{\sum(x_i \times x_j)}{\sqrt{\sum x_i^2} \times \sqrt{\sum x_j^2}}\right) C. \tag{8}$$

In this case, ReLU is used to suppress negative values of similarity. In contrast to self-attention alone, the SPCM retains the activation strength of the original CAM while having the ability to self-attention.

3.1.2. SAP

In SAP, the implementation of the correspondence between q_j and r_i focuses on the feature fusion of foreground features in each bounding box β using the A_{fore} generated by q_j:

$$r_i = \frac{\sum_{p \in B} A_{fore} f(p_i)}{\sum_{p \in B} (1 - A(p))}. \tag{9}$$

where $p \in B$ indicates the probability that the current pixel belongs to the foreground.

3.1.3. Loss

After passing through the SPCM, the CAM obtained is a mapping relationship with the original CAM. Points of the same class but far apart in high-dimensional space are mapped to a closer distance in low-dimensional space after passing through the SPCM. Conversely, points of different classes but close together in high-dimensional space will become farther apart in low-dimensional space after being mapped through the SPCM. The reduced contrast loss (L_{RC}) is proposed to address this problem:

$$L_{RC} = \frac{1}{2N} \sum_{n=1}^{N} y d^2 + (1 - y) \max(m - d, 0)^2. \tag{10}$$

where d represents the Euclidean distance between the corresponding two points in the SPCM and the mask, m is a set threshold, N is the number of samples, and y represents the label of whether the two samples match.

When $y = 0$, it means that the two samples are dissimilar, and in this case:

$$L_{RC} = \max(m - d, 0)^2. \tag{11}$$

If $d > m$, indicating that the distance between the two points is greater than the threshold m, then $L_{RC} = 0$, and no processing is performed on the two sample pairs that exceed the threshold. If $d < m$, indicating that the distance between the two points is less than the threshold m, then $L_{RC} = \max(m - d)^2$ is punished for the sample pairs that are less than the threshold.

When $y = 1$, it means that the two samples are similar, and in this case:

$$L_{RC} = d^2. \tag{12}$$

The penalty increases as the distance between the two sample pairs increases and decreases as the distance between them decreases.

Similarly, 2-way softmax classifiers w are applied to distinguish between the foreground and background regions of individual features (i.e., r_i and q_j). Training the network using standard cross-entropy loss (L_{CE}):

$$L_{CE} = \begin{cases} -\log(p), if(y = 1) \\ -\log(1 - p), otherwise \end{cases}. \tag{13}$$

where p represents the probability of the predicted sample belonging to a certain class, and y represents the sample label. When $y = 1$, the closer p is to 1, the smaller the loss value, and when $y = 0$, the closer p is to 0, the smaller the loss value.

For the entire classification network, a balancing parameter α is defined. The overall loss function in the training process is:

$$L_{classifier} = L_{RC} + \lambda L_{CE}. \tag{14}$$

3.2. Generate Pseudo-Label

In the process of generating pseudo-labels (Figure 4), we input the background attention map C_{back} and CAMs into CRF to obtain the initial pseudo-label y_1. Next, we extract the query q_c from y_1 and retrieve the features r_c in y_1 to obtain the final pseudo-label y_2 using the argmax function.

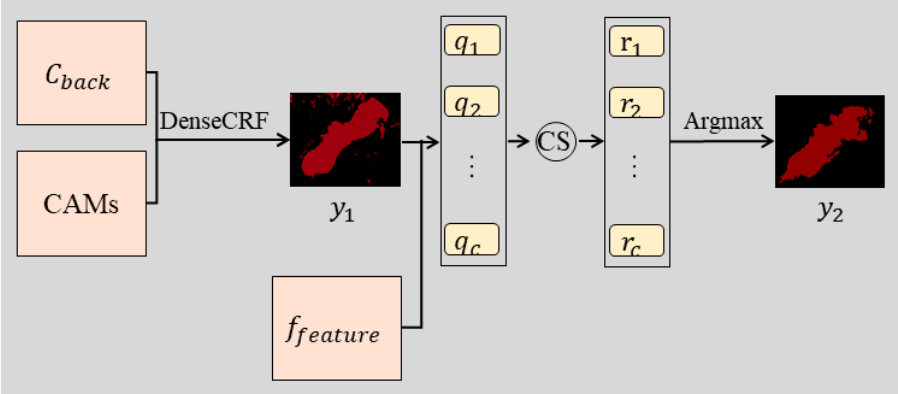

Figure 4. Pseudo-ground truth generation process diagram.

First, using the CAMs and C_{back} obtained from the classification network, we define a unary term $u_c(p)$ for each of these classes c using DenseCRF [23]:

$$u_c(p) = \begin{cases} \frac{CAM_c(p)}{\max_p(CAM_c(p))}, p \in B \\ 0 \quad , p \notin B \end{cases}. \tag{15}$$

where p denotes the probability that the sample is of class c and B denotes the bounding box of the object of class c; $CAM_c(p)$ can be expressed as:

$$CAM_c(p) = \text{ReLU}(C_{back} \cdot w_c). \tag{16}$$

where w_c denotes the classifier weight of class c objects in the classification network. For the background class, we use the background attention map A_{back} to define its unary term u_{back}:

$$u_{back}(p) = A_{back}(p). \tag{17}$$

We stitch the object category terms from Equation (15) and the background category terms from Equation (17) and input them into DenseCRF to obtain y_1. However, this method causes y_1 to contain some low-level features and noise, so we again supplemented the pseudo-label generation process by putting the features $f_{feature}$ obtained by the classification network through a method similar to the query in SAP. In the pseudo-label generation process, the query q_c for each category is defined as:

$$q_c = \frac{1}{|y_{1\_C}|} \sum_{p \in y_{1\_c}} f_{feature}(p). \tag{18}$$

where y_{1_C} is the set of locations of class c in the initial pseudo-label y_1 and $|\cdot|$ denotes the total number of pixels in that class; we use q_c to query similar features from $f_{feature}$ and extract for them the mapping C_2 for that class:

$$C_2(p) = \frac{f_{feature}(p)}{\left|\left|f_{feature}(p)\right|\right|} \cdot \frac{q_c}{||q_c||}. \tag{19}$$

where p indicates the probability of the area being a fire point. Then, we obtain the pseudo-segmentation label y_2 by applying the argmax function on the relevance map $C_2(p)$.

3.3. Context-Aware Loss

We train the network using y_1 and y_2 as the labels of the DeepLabV3 network, and extract the feature map f_d obtained from the penultimate layer of the DeepLabV3 network and input f_d into softmax to obtain a 2-dimensional feature probability map T. We want T to be consistent with the label given by y_2 and therefore the loss function L_{seg_ce} is defined for both:

$$L_{seg\_ce} = -\frac{1}{\sum_c |S_c|} \sum_c \sum_{p \in S_c} \log T_c(p). \tag{20}$$

where T_c represents the probability of class c, S_c denotes the region in which both y_1 and y_2 are of class c. For the same region, but where y_1 and y_2 give different labels, discarding the region directly may result in the loss of some of the true values. Here, we propose that L_{seg_con} once again makes full use of parts of the region with different labels for y_1 and y_2:

$$L_{seg\_con} = \frac{1}{2N} \sum_{n=1}^{N} (x_{y_1\_i} - x_{y_2\_i})^2 + (1 - z)\max(\text{margin} - (x_{y_1\_i} - x_{y_2\_i})^2, 0). \tag{21}$$

where N represents the number of samples, $x_{y_1\_i}$ and $x_{y_2\_i}$ denotes the samples corresponding to y_1 and y_2, respectively, z is a label indicating whether the two samples match, $z = 1$ means the two samples are similar or match, while $z = 0$ means no match, and margin is a set threshold.

Therefore, the overall CALoss is defined as shown below, where η is the equilibrium coefficient of the formula:

$$\text{CALoss} = L_{seg\_ce} + \eta L_{seg\_con}. \tag{22}$$

3.4. Evaluation Metrics

Accuracy (*Acc*) is often used as an evaluation metric for classification networks:

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \times 100\%. \tag{23}$$

We used the fire intersection ratio (IoU) as an indicator for the evaluation of forest fire segmentation (Figure 5).

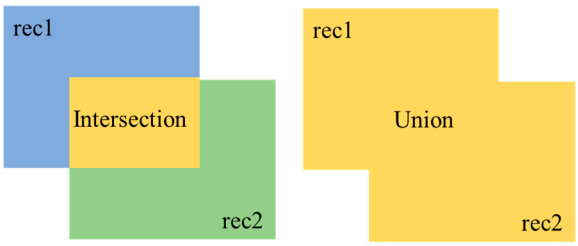

Figure 5. The IoU diagram.

IoU refers to the intersection area divided by its union, which can be expressed as:

$$\text{IoU} = \frac{\text{Intersection}}{\text{Union}} = \frac{S_{rec1} \cap S_{rec2}}{S_{rec1} + S_{rec2} - S_{rec1} \cap S_{rec2}}. \tag{24}$$

4. Experiments

4.1. Dataset

This section provides a detailed introduction to two publicly available aerial forest fire datasets. The Flame dataset and the Corsican dataset are both aerial forest fire images captured by UAVs. Of these, the Flame dataset generally has images with smaller fire

points and the Corsican dataset generally has images with larger fire points. The forest fire smoke dataset from the Kaggle competition was also used.

The Flame dataset [13] is a publicly available collection of aerial forest fire imagery, made accessible by researchers from Northern Arizona University and other academic institutions. The dataset comprises images captured by UAVs equipped with cameras at high altitudes. The fire points in the dataset are typically small in size and may be partially concealed. Figure 6 provides a glimpse of the data included in the dataset.

Figure 6. Images from the Flame dataset.

The Corsican dataset [14] is a publicly accessible dataset provided by the Environmental Sciences Laboratory at the University of Corsica. It consists of a collection of images capturing wildfires under various shooting conditions, including visible and near-infrared spectrums of burnt vegetation, diverse climatic conditions, lighting variations, and different fire distances. Unlike the Flame dataset, the images in the Corsican dataset were not taken from very high altitudes, resulting in larger fire sites being depicted. Figure 7 presents a selection of data samples from this dataset.

Figure 7. Images from the Corsican dataset.

The Kaggle dataset (https://www.kaggle.com/datasets/kutaykutlu/forest-fire, accessed on 10 February 2023) selected for this study comprises aerial images captured by UAVs. Specifically, it consists of 2007 images depicting smoke in forest environments. These aerial photographs were carefully selected and included in the dataset. Figure 8 provides a glimpse of some sample images contained within this dataset.

Figure 8. Images from the Kaggle dataset.

4.2. Implementation Details

We present the performance evaluation of our model on each dataset through a comparison of experimental and ablation implementations. The segmentation process of our model consists of three stages: the classification network, pseudo-label generation, and segmentation using DeeplabV3. We evaluate the performance of our model by assessing the results of the classification network and the overall weakly supervised segmentation. All experiments were done on Ubuntu 18.04, using the PyTorch deep learning framework, with an NVIDIA RTX 3090 GPU as the hardware device.

4.3. Forest Fire Image Classification

In practice, in the case of images taken by UAVs, our segmentation task is to default to images with fire points, but not all images taken by UAV have fire points. Before performing semantic segmentation, the images are classified, and only the images classified as containing a fire are subjected to subsequent segmentation operations.

To evaluate the performance of the proposed classification network in forest fire image classification, we compare it with the current mainstream classification networks (VGG16 [24], GoogleNet [25], ResNet50 [26], MobileNetV3 [27]) on two datasets. The Flame dataset consists of 1007 images from the Kaggle dataset, while the Corsican dataset comprises 1000 images from the same Kaggle dataset. We conducted comparisons among classification networks using different proportions (5%, 10%, 30%, 50%, 80%, 100%) of the images in the dataset. This analysis aimed to observe how the model learns additional image features and assess its classification performance as the number of training images increases.

The classifiers of all the networks were trained using the SGD optimizer with a momentum of 0.9, weight decay of 5×10^{-4}, and a batch size of 8. The classifier weights were randomly sampled from a Gaussian distribution with mean zero and standard deviation of 1×10^{-2}. The training process was carried out for 50 epochs. The classification results are summarized below.

The results presented in Tables 1 and 2 and Figure 9 clearly demonstrate that our model surpasses VGG16, GoogleNet, ResNet50, and MobileNetV3 in terms of classification accuracy on both the expanded Corsican dataset and the expanded Flame dataset at various image scales. In the expanded Corsican dataset, where images contain larger fire points and feature extraction is relatively simpler, our model achieves improvements of 0.08%, 0.31%, 0.04%, 0.07%, 0.13%, and 0.22% at 5%, 10%, 30%, 50%, 80%, and 100% scales, respectively, outperforming ResNet50. Conversely, in the expanded Flame dataset, which consists of images with smaller fire points and significant background effects, feature extraction becomes more challenging. In this scenario, our model exhibits a significant advantage over the other models. Across both datasets, our model achieves a highly satisfactory classification accuracy, nearing 100% at a scale of 100%. In practice, our image classification network extracts forest fire features from images with fire points, while images without fire points are filtered out. This enables subsequent image segmentation operations to focus

solely on relevant data. At smaller percentages such as 5% and 10%, our method is optimal on both the extended Corsican dataset and the extended Flame dataset.

Table 1. Classification results for different proportions on the Corsican dataset after incorporating images from the Kaggle dataset.

Networks	Proportions of the Expanded Corsican Dataset					
	5%	10%	30%	50%	80%	100%
VGG16	95.06%	95.23%	95.44%	96.73%	97.19%	98.24%
GoogleNet	95.19%	95.66%	96.17%	96.86%	97.47%	98.99%
ResNet50	96.49%	96.68%	96.99%	97.32%	98.11%	99.45%
MobileNetV3	95.75%	96.22%	96.87%	97.22%	98.05%	99.13%
Ours	96.57%	96.99%	97.03%	97.39%	98.24%	99.67%

Table 2. Classification results for different proportions on the Flame dataset after incorporating images from the Kaggle dataset.

Networks	Proportions of the Expanded Flame Dataset					
	5%	10%	30%	50%	80%	100%
VGG16	94.11%	94.13%	94.23%	95.74%	96.12%	96.99%
GoogleNet	95.13%	95.37%	95.81%	96.41%	96.74%	97.15%
ResNet50	95.87%	96.05%	96.16%	96.77%	98.43%	98.74%
MobileNetV3	94.64%	95.18%	95.44%	96.04%	97.13%	99.01%
Ours	96.33%	96.45%	96.55%	96.89%	98.99%	99.23%

4.4. Forest Fire Image Segmentation

4.4.1. Segmentation Results and Analysis

In the context of forest fire image segmentation, the primary task is to accurately segment the background and fire points in the image. Here, we compare our model with four of the latest weakly supervised models (IRNet [28], Puzzle-CAM [29], SEAM [30], and BABA [31]) for this purpose on the Corsican dataset and the Flame dataset.

From the experimental results (Figures 10 and 11 and Table 3), our model achieved an IoU of 81.23% and 76.43% on the Corsican dataset and the Flame dataset, respectively. From the segmentation visualization, the segmentation results of our model can be generally consistent with GroundTruth, but there are some shortcomings in our model in details such as slight errors in the segmentation results due to the high transparency at the flame boundary.

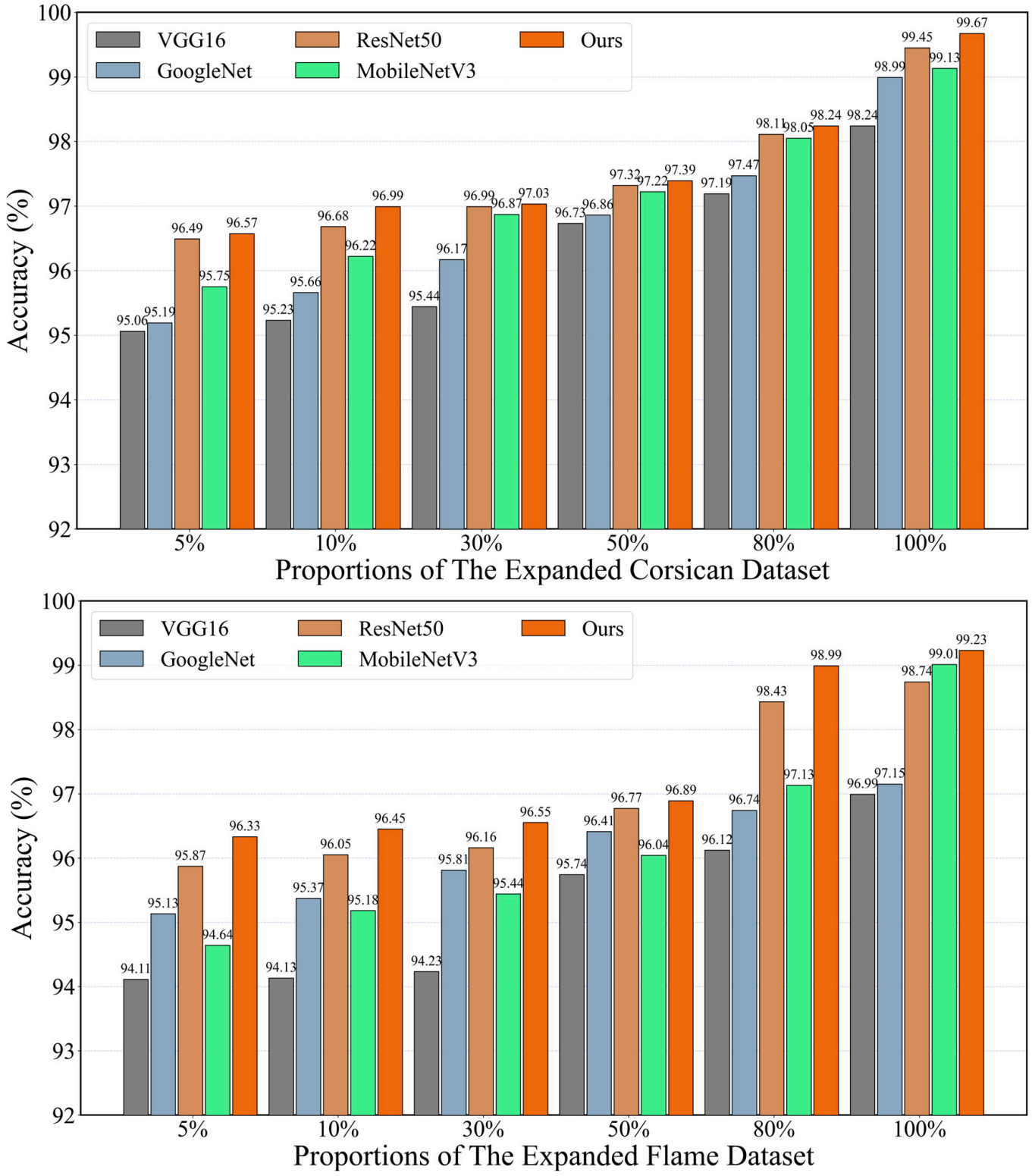

Figure 9. Classification accuracy under different scale datasets on the expanded Corsican dataset and Flame dataset.

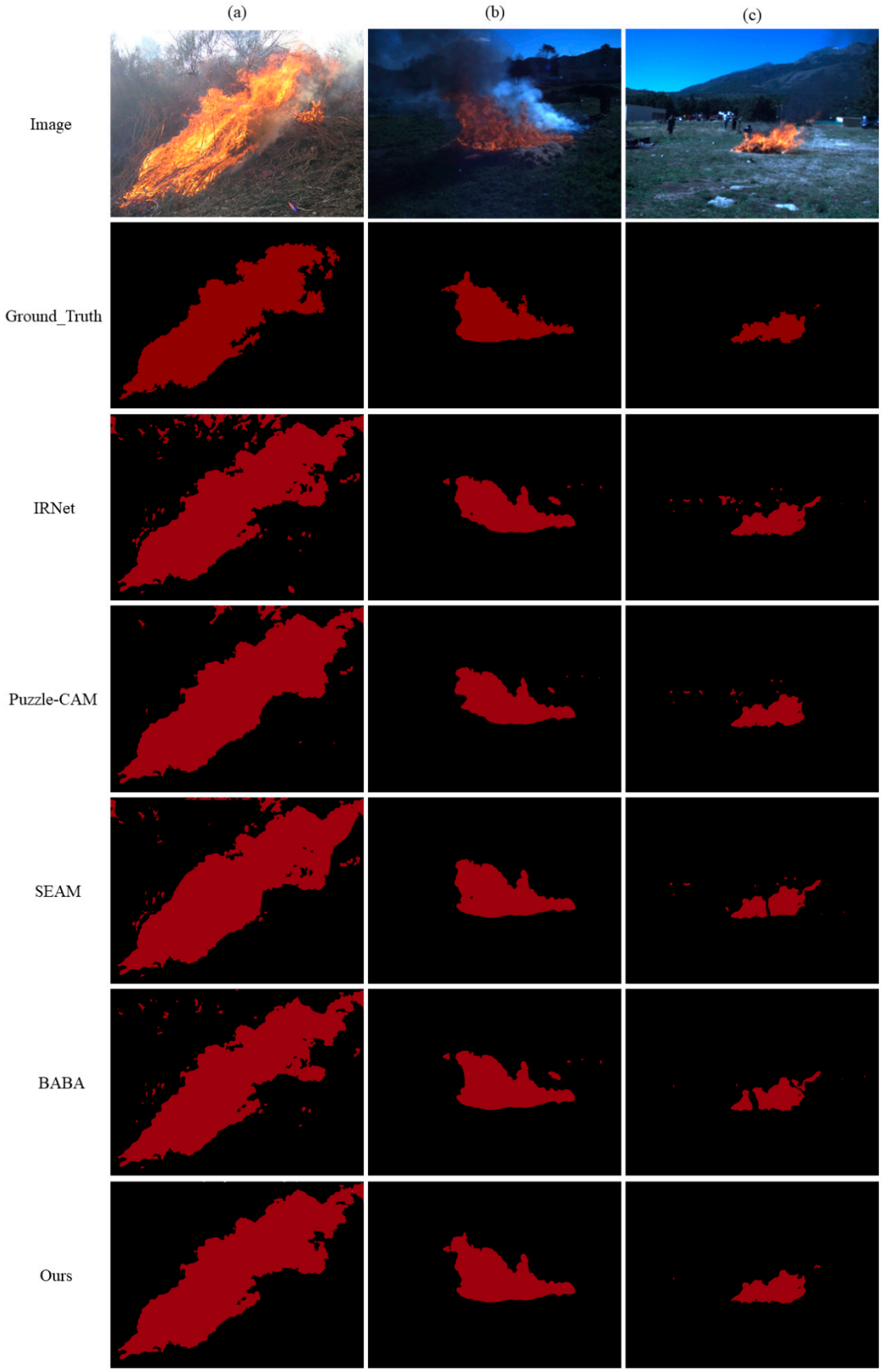

Figure 10. Visualization of semantic segmentation results on the Corsican dataset. (**a**–**c**) represent representative images from the Corsican dataset. Specifically, (**a**) depicts a close-range forest fire image taken during daylight, (**b**) illustrates a close-range forest fire image captured at night, and (**c**) showcases a distant forest fire image taken during the evening.

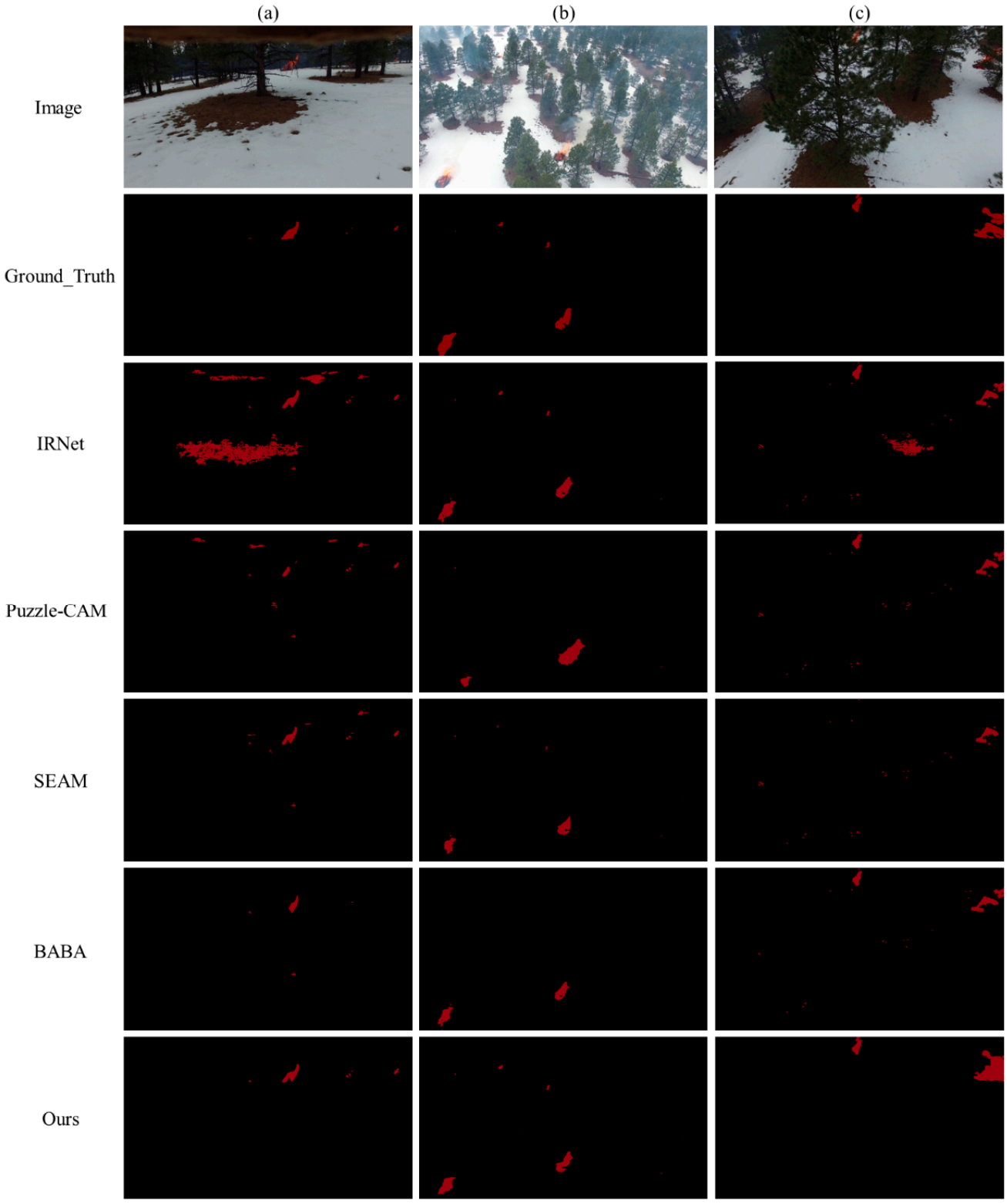

Figure 11. Visualization of semantic segmentation results on the Flame dataset. (**a**–**c**) represent representative images from the Flame dataset. Specifically, (**a**) depicts a forest fire image captured by a horizontal view without obstruction, (**b**) illustrates a forest fire image taken by a UAV at a higher distance with obstruction, and (**c**) showcases a forest fire image captured by a UAV at a lower distance with obstruction.

Table 3. Semantic segmentation results on Corsican dataset and Flame dataset.

NetWorks	IoU (Corsican)	IoU (Flame)
IRNet	77.91%	71.64%
Puzzle-CAM	79.89%	74.99%
SEAM	76.67%	72.89%
BABA	78.60%	74.38%
Ours	81.23%	76.43%

In the close-up UAV images of forest fires (Figure 10 and Table 3), the flame details are obvious, the flames take up a larger proportion of the overall image, and extracting flame features is relatively straightforward. The segmentation results are better for several networks (Table 3), but from the visualization results (Figure 10), our model has a finer edge segmentation of the fire points and the segmentation profile is closer to GroundTruth than those of the other four compared networks. The segmentation profile of the BABA network is also closer to that of GroundTruth, but its treatment of the flame boundary is more ambiguous. Both the IRNet and SEAM segmentation plots have large noise levels, and the Puzzle-CAM segmentation plots are too smooth for the flame boundary and fail to segment the flame boundary in detail.

In UAV images of forest fires taken at a distance (Figure 11 and Table 3), feature extraction and segmentation are more challenging due to the small proportion of flames in the overall image in the Flame dataset. The segmentation results of our model in this case can still outperform the comparison network. Our model is able to locate the fire point accurately, and the background noise has minimal effect on the experimental results (Figure 11). As shown in Figure 11c, our model has the ability to accurately split the outlines of the two fire points, despite the fact that this photograph was taken at a high altitude and is densely wooded and poorly lit. There are a large number of false detections in the IRNet, Puzzle-CAM, and SEAM segmentation results, and the right part of the fire point is clearly missed in BABA. Although the central part of the fire point on the right is not as well segmented by our model compared to that in GroundTruth, our model is much better at segmenting small fire points at long distances compared to the four comparison networks.

4.4.2. Ablation Experiments

To validate the contribution of the proposed SAP and CAL to UAV forest fire image segmentation, we performed ablation experiments on the same datasets (Table 4). Baseline refers to the use of only the pre-trained ResNet50 as the feature extractor during feature extraction and only DeepLabV3 as the segmentation network during segmentation.

Table 4. Results of the ablation experiments conducted on the Corsican dataset and Flame dataset.

Baseline	SAP	CAL	IoU (Corsican)	IoU (Flame)
√			78.54%	71.38%
√	√		80.04%	73.29%
√		√	80.77%	74.35%
√	√	√	81.23%	76.43%

The quality of pseudo-label is decisive for the segmentation effect of WSSS networks. The addition of the SAP module resulted in a significant increase in IoU, indicating a significant improvement in pseudo-label quality, by 1.5% and 1.91% on the Corsican dataset and Flame dataset, respectively. In UAV wildfire imagery, the images are taken in a forested environment and at high altitude, and the image background often contains a lot of noise. Our proposed CAL is a loss function specifically designed for UAV wildfire image segmentation. CAL uses contextual information about objects in the image to correct for the noise in pseudo-label and minimize the effect of ambient noise on feature extraction and

segmentation. By adding CAL, compared with the baseline, the network improved by 2.23% and 2.97% on the Corsican dataset and Flame dataset, respectively. Since environmental noise can cause a large amount of noise to be included in the generated pseudo-labels for UAV wildfire images, CAL is specifically designed to address the environmental noise in pseudo-labels, greatly improving network segmentation accuracy.

The SAP module improves the quality of the generated pseudo-labels and CAL corrects the noise in the pseudo-labels by adding contextual information, confirming both theoretically and experimentally the synergistic effect of the two.

5. Discussion

In this study, we proposed a weakly supervised semantic segmentation (WSSS) model for UAV forest fire images, addressing the challenges of data annotation and noise. Our model utilized the self-supervised attention foreground-aware pooling (SAP) and context-aware loss (CAL) techniques to generate high-quality pseudo-labels and improve segmentation accuracy.

One of the major challenges in forest fire monitoring using UAVs is the need for manual annotation of pixel-level labels, which is time-consuming and resource-intensive. Wang et al. [9] applied a variety of classical segmentation models to segmentation of aerial forest fire images, and Choi et al. [12] also achieved good segmentation results on forest fire datasets. However, all the above methods are based on datasets with pixel-level labels and do not achieve segmentation based on datasets without pixel-level labels. The weakly supervised approach we employed alleviated this challenge by using image-level labels instead, significantly reducing annotation costs.

By utilizing SAP, we incorporated class activation mapping (CAM) and bounding box annotation to generate pseudo-labels that accurately captured the location of fire points. This method improved the segmentation accuracy by effectively distinguishing the foreground fire points from the background. Moreover, we introduced the CAL technique to address the issue of noise in the generated pseudo-labels. By leveraging contextual information from the fire point's surroundings, CAL helped refine the pseudo-labels, making them more accurate and closely resembling the real labels.

However, our model still exhibited some limitations. In certain scenarios, such as when there was high transparency at the flame boundary, slight errors in segmentation results were observed. Further improvements are needed to address these challenges and enhance the accuracy of fire point segmentation.

In terms of practical applications, our model provides a cost-effective and accurate solution for UAV-based forest fire monitoring. By automating the segmentation process and reducing the need for laborious manual annotation, our approach can significantly improve the efficiency of forest fire prevention and control efforts.

6. Conclusions

We present SAP in WSSS of UAV forest fire images for the generation of high precision pseudo-labels. In forest fire image semantic segmentation, it is difficult to accurately label the boundaries of fire point objects, which are located in the foreground of bounding boxes. The CAM generated during feature extraction can accurately reflect the specific location of the fire point object. SAP uses CAM and bounding box annotation to generate high quality pseudo-labels for training semantic segmentation networks for forest fire images. However, during the process of generating pseudo-labels, there is inevitable background noise due to the large interference of the environmental background. We introduce CAL to correct the noise in the generated pseudo-labels by using contextual information from the image where the fire point is located. We experimented on two publicly available UAV forest fire datasets and compared our model with other models. The IoU of the semantic segmentation results of this model on the Corsican dataset and Flame dataset can reach 81.23% and 76.43% respectively. Our results show a significant breakthrough in WSSS of UAV forest fire images.

Author Contributions: J.W.: Conceptualization, software, original draft preparation; C.X.: review & editing, supervision; Y.W.: methodology, data curation; L.L., H.Y. and N.Y.: visualization, formal analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Nonprofit Research Projects (CAFYBB2021QD001-1) of Chinese Academy of Forestry and the Zhejiang Science and Technology Major Program on Agricultural New Variety Breeding (2021C02070-1) and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX22_1105).

Data Availability Statement: The Flame dataset is available at https://ieee-dataport.org/open-access/flame-dataset-aerial-imagery-pile-burn-detection-using-drones-uavs (accessed on 19 January 2022). The Corsican dataset is available at http://cfdb.univ-corse.fr/index.php?menu=1 (accessed on 19 January 2022). The Kaggle dataset is available at https://www.kaggle.com/datasets/kutaykutlu/forest-fire (accessed on 10 February 2023).

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

References

1. Han, Z.; Geng, G.; Yan, Z.; Chen, X. Economic Loss Assessment and Spatial–Temporal Distribution Characteristics of Forest Fires: Empirical Evidence from China. *Forests* **2022**, *13*, 1988. [CrossRef]
2. Dimitropoulos, S. Fighting Fire with Science. *Nature* **2019**, *576*, 328–329. [CrossRef] [PubMed]
3. Feng, L.; Zhou, W. The Forest Fire Dynamic Change Influencing Factors and the Impacts on Gross Primary Productivity in China. *Remote Sens.* **2023**, *15*, 1364. [CrossRef]
4. Osco, L.P.; Junior, J.M.; Ramos, A.P.M.; de Castro Jorge, L.A.; Fatholahi, S.N.; de Andrade Silva, J.; Matsubara, E.T.; Pistori, H.; Gonçalves, W.N.; Li, J. A Review on Deep Learning in UAV Remote Sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102456. [CrossRef]
5. 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]
6. Kang, H.; Wang, X. Semantic Segmentation of Fruits on Multi-Sensor Fused Data in Natural Orchards. *Comput. Electron. Agric.* **2023**, *204*, 107569. [CrossRef]
7. Chen, Z.; Deng, L.; Luo, Y.; Li, D.; Junior, J.M.; Gonçalves, W.N.; Nurunnabi, A.A.M.; Li, J.; Wang, C.; Li, D. Road Extraction in Remote Sensing Data: A Survey. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102833. [CrossRef]
8. Zhang, H.; Liu, M.; Wang, Y.; Shang, J.; Liu, X.; Li, B.; Song, A.; Li, Q. Automated Delineation of Agricultural Field Boundaries from Sentinel-2 Images Using Recurrent Residual U-Net. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102557. [CrossRef]
9. Wang, Z.; Peng, T.; Lu, Z. Comparative Research on Forest Fire Image Segmentation Algorithms Based on Fully Convolutional Neural Networks. *Forests* **2022**, *13*, 1133. [CrossRef]
10. Park, M.; Bak, J.; Park, S. Advanced Wildfire Detection Using Generative Adversarial Network-Based Augmented Datasets and Weakly Supervised Object Localization. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *114*, 103052. [CrossRef]
11. Flood, N.; Watson, F.; Collett, L. Using a U-Net Convolutional Neural Network to Map Woody Vegetation Extent from High Resolution Satellite Imagery across Queensland, Australia. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *82*, 101897. [CrossRef]
12. Choi, H.-S.; Jeon, M.; Song, K.; Kang, M. Semantic Fire Segmentation Model Based on Convolutional Neural Network for Outdoor Image. *Fire Technol.* **2021**, *57*, 3005–3019. [CrossRef]
13. 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]
14. Toulouse, T.; Rossi, L.; Campana, A.; Celik, T.; Akhloufi, M.A. Computer Vision for Wildfire Research: An Evolving Image Dataset for Processing and Analysis. *Fire Saf. J.* **2017**, *92*, 188–194. [CrossRef]
15. Novac, I.; Geipel, K.R.; de Domingo Gil, J.E.; de Paula, L.G.; Hyttel, K.; Chrysostomou, D. A Framework for Wildfire Inspection Using Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/SICE International Symposium on System Integration (SII), Honolulu, HI, USA, 12–15 January 2020; pp. 867–872.
16. Peng, Y.; Wang, Y. Real-Time Forest Smoke Detection Using Hand-Designed Features and Deep Learning. *Comput. Electron. Agric.* **2019**, *167*, 105029. [CrossRef]
17. Khryashchev, V.; Larionov, R. Wildfire Segmentation on Satellite Images Using Deep Learning. In Proceedings of the 2020 Moscow Workshop on Electronic and Networking Technologies (MWENT), Moscow, Russia, 11–13 March 2020; pp. 1–5.
18. Wang, Z.; Yang, P.; Liang, H.; Zheng, C.; Yin, J.; Tian, Y.; Cui, W. Semantic Segmentation and Analysis on Sensitive Parameters of Forest Fire Smoke Using Smoke-Unet and Landsat-8 Imagery. *Remote Sens.* **2022**, *14*, 45. [CrossRef]
19. Van Engelen, J.E.; Hoos, H.H. A Survey on Semi-Supervised Learning. *Mach. Learn.* **2020**, *109*, 373–440. [CrossRef]
20. Zhang, D.; Han, J.; Cheng, G.; Yang, M.-H. Weakly Supervised Object Localization and Detection: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 5866–5885. [CrossRef] [PubMed]

21. Su, Y.; Sun, R.; Lin, G.; Wu, Q. Context Decoupling Augmentation for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 7004–7014.
22. Amaral, B.; Niknejad, M.; Barata, C.; Bernardino, A. Weakly Supervised Fire and Smoke Segmentation in Forest Images with CAM and CRF. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 442–448.
23. Zhang, Z.; Sabuncu, M. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 8792–8802.
24. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
25. 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 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V. Searching for Mobilenetv3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
28. Ahn, J.; Cho, S.; Kwak, S. Weakly Supervised Learning of Instance Segmentation with Inter-Pixel Relations. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2209–2218.
29. Jo, S.; Yu, I.-J. Puzzle-Cam: Improved Localization via Matching Partial and Full Features. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 639–643.
30. Wang, Y.; Zhang, J.; Kan, M.; Shan, S.; Chen, X. Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 12275–12284.
31. Oh, Y.; Kim, B.; Ham, B. Background-Aware Pooling and Noise-Aware Loss for Weakly-Supervised Semantic Segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 6913–6922.