Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 55-6 (2022) 145–150

A Real-time Fire Segmentation Method Based on A Deep Learning Approach

Mengna Li* Youmin Zhang Lingxia Mu* Jing Xin* Ziquan Yu*** Shangbin Jiao* Han Liu* Guo Xie* Yingmin Yi***

**Shaanxi Key Laboratory of Complex System Control and Intelligent Information Processing Xi'an University of Technology, Xi'an, Shaanxi 710048, China (e-mail: limengna1281@foxmail.com)*

***Department of Mechanical, Industrial and Aerospace Engineering, Concordia University, Montreal, Quebec H3G 1M8, Canada (e-mail: youmin.zhang@concordia.ca)*

****College of Automation Engineering, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, Jiangsu, China (e-mail: yuziquan@nuaa.edu.cn)*

Abstract: As a kind of the forest "fault", fire is highly destructive and difficult to rescue. Fire segmentation is helpful for firefighters to understand the fire scale and formulate a reasonable fire-fighting plan. Therefore, this paper proposes a real-time fire segmentation method based on deep learning. This method is an improved version of deeplbav3+, which is an encoder-decoder structure network. Encoder network is composed of deep convolutional neural network and atrous spatial pyramid pooling. Different from deeplabv3+, in order to improve the segmentation speed, this paper uses the lightweight network mobilenetv3 to build a new deep convolutional neural network and does not use atrous convolution, but it will affect the segmentation accuracy. Therefore, in order to compensate for the loss of segmentation accuracy, on the basis of the original decoder network, this paper adds two different shallow features to make the network contain rich fire feature information. Experimental results show that the comprehensive performance of this method is better than the original deeplabv3+, especially the segmentation speed of the network is greatly improved, which is about 59 FPS.

Keywords: Forest fire, Fire segmentation, Atrous spatial pyramid pooling, Mobilenetv3, Deeplabv3+

1. INTRODUCTION

Forest fire is a "fault" in the forest environment, which is a kind of natural disaster with strong suddenness, great destructiveness, and difficulty in handling and rescue. Forest fire prevention is the foundation of forestry development, which is related to forest resources and ecological security, as well as the safety of people's lives and properties. If we can detect this fault as early as possible and obtain the fault area at this time, firefighters can formulate a reasonable and efficient fire fighting plan, thereby saving manpower and material resources (Yuan et al., 2019).

With the rapid development of technologies, researchers have begun to use remote sensing satellites and aircraft in forest fire detection system and based on these a number of methods have been developed to identify and detect forest fires. According to the relative location in space where the system is located, fire detection system can be divided into three categories: *terrestrial system, airborne system,* and *satellite system* (Barmpoutis et al., 2020). Terrestrial systems consist of a single sensor or sensor network deployed near the ground (Cui et al., 2020). Airborne systems typically use single or multiple aircraft to patrol over forests for early forest fire detection missions (Casbeer et al., 2006; Sherstjuk et al., 2018). Satellite systems use infrared remote sensing technology, infrared imaging principles and visible light photography to observe forest fires from space (Waigl et al., 2020).

In order to develop a general fire segmentation method and apply it to the above-mentioned fire detection system, to help firefighters obtain the fire burning area and spreading direction, and provide effective support for subsequent fire rescue tasks. Therefore, this paper proposes a real-time forest fire segmentation method based on deep learning, which is an improved version of deeplabv3+ (Chen et al., 2018). This method is an encoder-decoder structure network. Encoder network is composed of deep convolutional neural network (DCNN) and atrous spatial pyramid pool (ASPP), which is used to generate semantic features of four different scales. Decoder network is used to fuse these features and upsample them to obtain the result of fire segmentation.

Different from deeplabv3+, this paper first uses the lightweight network mobilenetv3 to reconstruct the deep convolutional neural network in the encoder network, and does not use atrous convolution. Although atrous convolution can increase the receptive field and improve the accuracy of network segmentation, excessive use will reduce the segmentation speed. Second, although the introduction of a lightweight network can reduce the network parameters and improve the segmentation speed, but it will affect the segmentation accuracy. Therefore, in order to compensate for the drop in segmentation accuracy, based on the original

10.1016/j.ifacol.2022.07.120

decoder network, we add two different shallow features to describe the fire phenomenon. The experimental results show that the comprehensive performance of this method is better than the original deeplabv3+. In particular, it greatly improves the segmentation speed of the original deeplabv3+, and its processing speed is about 59 FPS.

The rest of this paper is organized as follows. The second section briefly introduces the related work of forest fire segmentation. The third section introduces the method proposed in this paper and conducts a performance analysis. The fourth section evaluates the fire segmentation performance of the proposed method and gives some fire segmentation results. The fifth section summarizes the work and gives possible future work plans.

2. RELATED WORK

The purpose of fire segmentation is to find a suitable segmentation threshold to extract the fire area from the fire image, so as to segment the fire area and the background. At present, deep learning methods have shown excellent performance in the field of image segmentation, effectively improving the accuracy of fire segmentation (Minaee et al., 2021). Related works include, but are not limited to, Long et al. used a convolutional neural network (CNN) to extract fire regions in images in order to distinguish forest fires from background (Long et al., 2021). Pan et al. used weakly supervised fine segmentation and a lightweight Faster-RCNN to detect and segment fire and smoke areas (Pan et al., 2020). It shows competitive performance compared to other methods. Ghali et al. used three deep convolutional networks, U-Net, U^2-Net, and EfficientSeg, to segment forest fire pixels and detect fire regions (Ghali et al., 2021), but their detection speed was slow.

Deeplabv3+ is a deep learning-based semantic segmentation algorithm with excellent comprehensive performance, and it also shows good performance in the recognition and segmentation of fire images (Liu et al., 2021; Harkat et al., 2021). To further improve the segmentation performance of Deeplabv3+, this paper research and optimizes its network structure, so a forest fire segmentation algorithm based on deep learning is proposed.

3. FIRE SEGMENTSTAION METHOD

In this section, a real-time forest fire segmentation method based on deep learning is proposed by improving deeplabv3+ network. As show in Fig. 1, this method is an encoder-decoder structure network. Encoder network is composed of deep convolutional neural network and atrous spatial pyramid pooling, which is used to generate feature maps with four different resolutions. Decoder network aims to fuse these features and restore the feature size by upsample to get fire segmentation results. Finally, this paper analyzes and discusses the performance of the proposed method.

3.1 Encoder network

The overall architecture of deep convolutional neural network is shown in Table 1. Its first layer is a standard convolutional layer with 16 convolutional filters, and then 15 mobilenetv3 bottlenecks are stacked. These mobilenetv3 (Howard et al., 2019) bottlenecks (that is mbneck as shown in Table 1) are divided into different stages according to the size of input features. In some mobilenetv3 bottlenecks, the authors also uses the sequeeze-and-excite (SE) module. In the Table 1, "exp" means expansion size, "out" means the number of output channels, "SE" denotes whether there is a Squeeze-and-Excite in that block, "NL" denotes the type of indicates the type of nonlinear activation function used, "HS" denotes H-swish and "RE" denotes ReLU. Table 1 shows the structure of DCNN when the output stride is 16.

Table 1. Overall architecture of the DCNN

Input	Operator	exp	out	SE	NL	Stride
$512^2\times3$	Conv2d 3×3	-	16	-	HS	2
$256^2\times16$	mbneck, 3×3	16	16	-	RE	1
$256^2\times16$	mbneck, 3×3	64	24	-	RE	2
$128^2\times24$	mbneck, 3×3	72	24	-	RE	1
$128^2\times24$	mbneck, 5×5	72	40	1	RE	2
$64^2\times40$	mbneck, 5×5	120	40	1	RE	1
$64^2\times40$	mbneck, 5×5	120	80	1	RE	1
$64^2\times40$	mbneck, 3×3	240	80	-	HS	2
$32^2\times80$	mbneck, 3×3	200	80	-	HS	1
$32^2\times80$	mbneck, 3×3	184	80	-	HS	1
$32^2\times80$	mbneck, 3×3	184	80	-	HS	1
$32^2\times80$	mbneck, 3×3	480	112	1	HS	1
$32^2\times112$	mbneck, 3×3	672	160	1	HS	1
$32^2\times160$	mbneck, 5×5	672	160	1	HS	1
$32^2\times160$	mbneck, 5×5	960	160	1	HS	1
$32^2\times160$	mbneck, 5×5	960	160	1	HS	1

The structure of mobilenetv3 bottleneck with SE module is shown in Fig. 2, where 1) standard 1×1 convolution is used for dimension expansion of feature channels; 2) depthwise convolution (or called channel-by-channel convolution, that is DConv in Fig. 1) is used to extract image features on per-channel; 3) pointwise convolution (constructed by standard 1×1 convolution) is used to reduce the number of feature channels to match the shortcut path. Here 2) and 3) constitute a depthwise separable convolution, which can reduce the amount of network parameters and improve the running speed of the network compared with standard convolution. Batch normalization (BN) and nonlinear activation function are are applied in the mobilenetv3 bottleneck.

Two different nonlinear activation functions are used in the mobilenetv3 bottleneck: *rectified linear unit (ReLU)* and *hard*

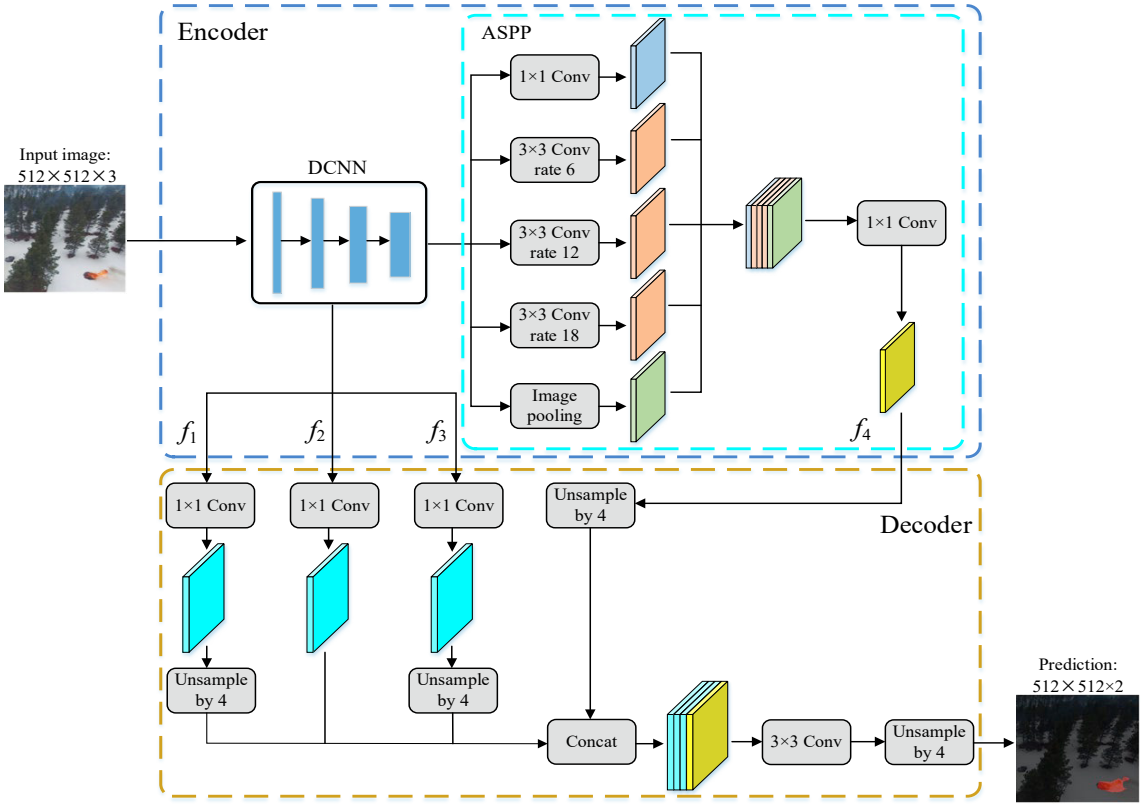

Figure 1. Overall architecture of improve deeplabv3+ network.

versions of swish (H-swish), as shown in Table 1. The form of ReLU is as follows:

$$ReLU[x] = \max(0, x) \tag{1}$$

The form of H-swish is as follows:

$$H\text{-swish}[x] = x\frac{ReLU6(x+3)}{6} \tag{2}$$

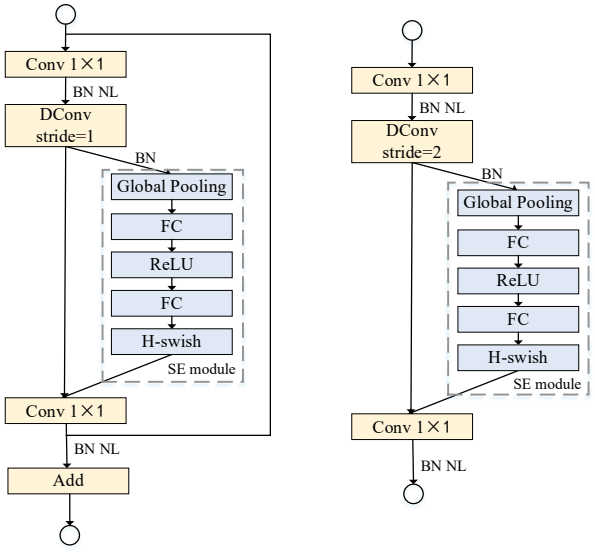

Figure 2. The mobilenetv3 bottleneck with SE module.

Atrous spatial pyramid pooling is inspired by spatial pyramid pooling (SPP), which can effectively resample features of different scales to accurately and effectively classify regions of any scale. ASPP is a variant of SPP that effectively captures multi-scale contextual information by applying four atrous convolutions with different expansion coefficients in parallel on the deep feature maps generated by DCNN.

ASPP is composed of two parts: 1) one 1×1 standard convolution and three 3×3 atrous convolutions are used to generate four different image features. When the output stride is 16, the expansion coefficient of the atrous convolution is set to r=(6, 12, 18); and when the output stride is 8, the expansion coefficient is set to r=(12, 24, 36). All convolution operations have 256 convolution kernels and batch normalization. 2) global average pooling is used to generate an image-level features. Then, these branch features are concatenated and fed into a standard 1×1 convolution to generate the final semantic features. Since ASPP acts on the deepest feature map of DCNN, it can aggregate contextual information of different scales in the deep stage of the network, thereby enhancing the segmentation performance.

Finally, the encoder network consists of a deep convolutional neural network and atrous spatial pyramid pooling, which can be used to generate four different feature maps. When the output stride is 16, the size of the feature map generated by the encoder network are: $f_1 \in \mathbb{R}^{256\times256\times16}$ $f_2 \in \mathbb{R}^{128\times128\times24}$, $f_3 \in \mathbb{R}^{64\times64\times40}$ $f_4 \in \mathbb{R}^{32\times32\times256}$

are: $f_1 \in \mathbb{R}^{256\times256\times16}$, $f_2 \in \mathbb{R}^{128\times128\times24}$, $f_3 \in \mathbb{R}^{64\times64\times40}$ and $f_4 \in \mathbb{R}^{64\times64\times256}$.

3.2 Decoder network

The main purpose of the decoder network is to decode the obtained features by upsample and convolution to restore them to the original image size, resulting in the fire segmentation result.

The implementation process of the decoder network is as follows: 1) apply a standard 1×1 convolution with 256 convolution kernels to adjust the number of channels of features f_1, f_2 and f_3; 2) upsample features f_1, f_3 and f_4 to have the same feature size as f_2; 3) concatenate these features and use 3×3 convolution to re-adjust the channel number of features; 4) upsample the last obtained features to get the final segmentation result.

3.3 Discussion and analysis

In the original deeplabv3+, the author used the xception network to build the DCNN and introduced atrous convolution. However, this paper uses a lightweight network mobilenetv3 in the encoder network to reconstruct the DCNN, and does not use atrous convolution. Although the use of atrous convolution can improve the segmentation accuracy of the network by increasing the receptive field, using atrous convolution reduces the segmentation speed. Second, using a lightweight network can reduce the network parameters and improve the network speed, but it will reduce the segmentation accuracy.

Since the deep features (that is, the feature f_4 in Fig. 1) contains abstract semantic information, which enables the model to understand the fire phenomenon. The shallow feature (that is, the feature f_1, f_2 and f_3 in Fig. 1) contains some detailed information of the object, such as the color, shape, and outline information of the fire. The decoding method including both depth features and shallow features can improve the segmentation accuracy.

In the deeplabv3+, the authors only used convolution and upsample operation for the features f_2 and f_4. In order to make up for the above-mentioned loss of accuracy caused by the use of lightweight networks, this paper adds two shallow features, i.e., features f_1 and f_3, on the basis of the original decoder network, so that decoder network contain more fire feature information. This helps to improve the segmentation accuracy of the network.

4. EXPERIMENT EVALUATION

In this section, quantitative and qualitative evaluations are used to measure the performance of the proposed method. First, three evaluation indicators are introduced to measure the segmentation accuracy and processing speed of the algorithm. Second, in a quantitative evaluation, the comprehensive segmentation performance of the proposed method is quantified. Finally, in the qualitative evaluation, the segmentation results of forest fire areas obtained using the method in this paper are presented and analyzed.

4.1 Evaluation indicators

The evaluation metrics used in the experiments are described as follows.

MPA: MPA is the mean pixel accuracy. The calculation is as follows:

$$MPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}} \tag{3}$$

where k represents the number of classes (in this paper, $k=1$, including fire and background), and p_{ij} represents the number of pixels that belong to class i but are predicted to be class j; p_{ii} represents the number of pixels that belong to class i and are correctly predicted as class i; p_{ji} represents the number of pixels that belong to class j but are predicted to be class i.

MIoU: MIoU is the mean intersection over union. The calculation is as follows:

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{4}$$

FPS: FPS is the number of frames processed per second, which is used to measure the processing speed of the proposed method. When the processing speed is higher than 20 FPS, it is considered that real-time segmentation is achieved.

In this experiments, MPA and MIoU are used to measure the fire segmentation precision of the proposed method, and FPS is used to measure the processing speed of the proposed method.

4.2 Quantitative evaluation

This paper uses the aviation fire segmentation dataset developed by Shamsoshoara et al. to train and test the performance of the proposed method (Shamsoshoara et al., 2020). The images in the dataset are collected by using the DJI Phantom UAV and its default camera, with a total of 2003 frames, which are divided into training set and testing set according to 0.85:0.15. In the experiments, total training is 30 epochs, and the batch size is set to 2. The experimental equipment is a computer with windows 10 system, i7-9700k and NVIDIA RTX 2080 Ti.

Table 2 reports the segmentation accuracy and processing speed of deeplabv3+, xception-deeplabv3+, and our method on the test set. In the experiment, the output stride is set to 16. The second column represents the test results using mobilennetv3 only in the encoder network.

Table 2. Comparison of MPA, MIoU, and FPS among Deeplabv3+, Xception-deeplabv3+, and our method on the test set

	Deeplabv3+	Xception-deeplabv3+	Our method

MPA (%)	92.09	91.40	92.46
MIoU (%)	86.75	86.49	86.98
FPS	24	62	59

It can be seen from Table 2 that the processing speed of this method is about 59 FPS, which is much faster than the original deeplabv3+. This is because we use the lightweight network mobilenetv3 in the construction of the encoder network and do not use atrous convolution. Although using atrous convolution will increase the receptive field to improve detection accuracy, excessive use of atrous convolution will slow down the detection speed. Second, although the introduction of a lightweight network helps to reduce network parameters and improve the segmentation speed, it will reduce the segmentation accuracy. The second column of Table 2 shows that using mobilennetv3 can improve the running speed of the network, but it can slightly reduce the segmentation accuracy.

Therefore, to compensate for the drop in segmentation accuracy, this paper introduces two different shallow features based on the original decoder network. This makes the network contain rich fire information, which helps the network to better describe the fire phenomenon. Therefore, the MPA and MIoU values of our method are similar to (or slightly higher than) the original deeplabv3+.

4.3 Qualitative evaluation

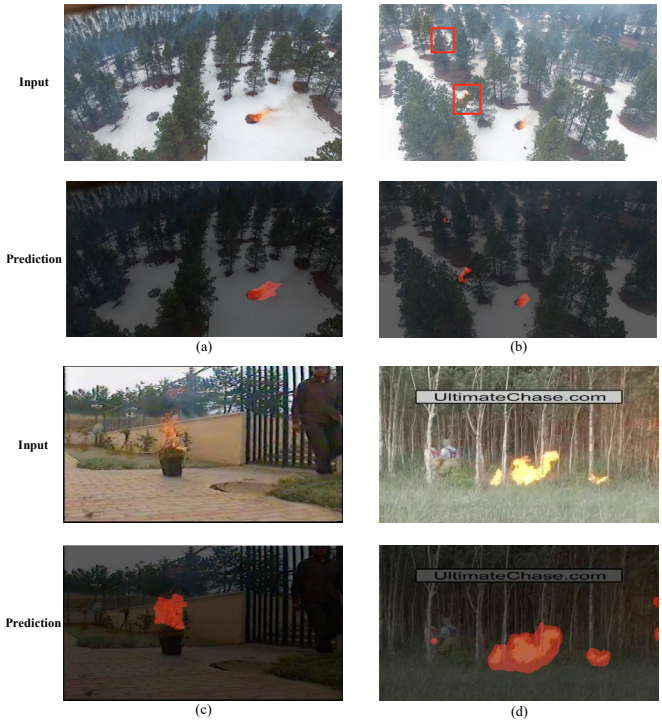

Figure 3. Segmentation results of forest fire.

Fig. 3 shows the forest fire segmentation results obtained using our method, where the red boxes mark some fire areas occluded by trees. The first row is the input image of the network, and the second row is the segmentation result of the network output. It can be seen that for some occluded fire areas, the proposed method can identify and finely segment the fire areas. This shows that our method can handle forest fires with occlusion. The input images shown in Figure 3(c) and (d) are fire images outside of the training and testing datasets, and it can be seen that our method can still detect and segment fire regions despite some misclassification in the fire edge regions. This shows that the proposed method has certain generalization performance and robustness.

Through the analysis of the above quantitative evaluation and qualitative evaluation results, it can be concluded that the method in this paper is a fire segmentation method with excellent performance. It not only maintains a similar and slightly better segmentation accuracy than the original deeplabv3+, but also greatly improves the processing speed of the network.

5. CONCLUSIONS

In order to solve the “fault” of forest, that is, forest fire, this paper proposes a real-time fire segmentation method based on deep learning. This method is an improved version of deeplabv3+, which is an encoder-decoder structure network. Encoder network is composed of deep convolutional neural network and atrous spatial pyramid pooling, which is used to generate four semantic features with different resolutions. Encoder network consists of a deep convolutional neural network and atrous spatial pyramid pooling to generate semantic features at four different resolutions. Different from deeplabv3+, this paper first uses the lightweight network mobilenetv3 to reconstruct the deep convolutional neural network in the encoder network without atrous convolution. While this helps speed up the processing of the network, it reduces segmentation accuracy. Therefore, to compensate for the loss of segmentation accuracy, based on the original decoder network, we introduce two different image features to enrich fire information. Experiments show that the method in this paper is a segmentation model with better performance. Not only does it have slightly better segmentation accuracy than the original deeplabv3+, but its segmentation speed is much faster than the original deeplabv3+, about 59 FPS. In the future, we intend to deploy the proposed algorithm on a UAV platform for fire detection and area segmentation.

ACKNOWLEDGMENT

This work is partially supported by the National Natural Science Foundation of China (Grants 61833013, 61873200, 61903297, and 62003162), Natural Science Foundation of Shaanxi Provincial Department of Education (19JK0569), and Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Yuan, C., Zhang, Y.M., and Liu, Z.X. (2015). A survey on technologies for automatic forest fire monitoring, detection, and fighting using unmanned aerial vehicles and remote sensing techniques. *Canadian Journal of Forest Research*, 45(7), pp.783-792.

Barmpoutis, P., Papaioannou, P., Dimitropoulos, K., and Grammalidis, N. (2020). A review on early forest fire detection systems using optical remote sensing. *Sensors*, 20(22), p. 6442

Cui, F.M. (2020). Deployment and integration of smart sensors with IoT devices detecting fire disasters in huge forest environment. *Computer Communications*, 150, pp. 818-827.

Casbeer, D., Kingston, D., Bear, R., McLain, T., Li, S. (2006). Cooperative forest fire surveillance using a team of small unmanned air vehicles. *Int. J. Syst. Sci*., 37(6), 351–360

Sherstjuk, V., Zharikova, M., and Sokol, I. (2018). Forest fire monitoring system based on UAV team, remote sensing, and image processing. *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, 2018, pp. 590-594.

Waigl, C.F., Prakasha, A., Stuefera, M., Verbylab, D., and Dennisonc, P. (2019). Fire detection and temperature retrieval using EO-1 Hyperion data over selected Alaskan boreal forest fires. *International Journal of Applied Earth Observation and Geoinformation*, 81, pp. 72-84.

Chen, L.C., Zhu, Y., Papandreou G., Schroff, F., and Adam H., (2018). Encoder-Decoder with atrous separable convolution for semantic image segmentation. *Springer, Cham.*

Minaee, S., Boykov, Y.Y., Porikli, F., Plaza, A.J., Kehtarnavaz, N. and Terzopoulos, D. (2021). Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, pp. 1-1.

Long, N.T., Alexander A.D., and Huong N.T. (2021) Segmentation of forest fire images based on convolutional neural networks. *International Journal of Artificial Intelligence*, 19(1), pp. 21-35.

Pan, J., Ou, X.M., and Xu, L. (2021). A collaborative region detection and grading framework for forest fire smoke using weakly supervised fine segmentation and lightweight Faster-RCNN. *Forests*, 12(6), pp. 768-768.

Ghali, R., Akhloufi, M.A., Jmal, M., Mseddi W.S., and Attia, R. (2021). Forest Fires Segmentation using Deep Convolutional Neural Networks. *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2021, pp. 2109-2114.

Liu, Z.Y., Xie, C.S., Li, J.J., and Sang, Y. (2021). Smoke region segmentation recognition algorithm based on improved Deeplabv3+. *Systems Engineering and Electronics*, 43(2), pp. 328-335.

Harkat, H., Nascimento J.M.P., and Bernardino A. (2021) Fire Detection using Residual Deeplabv3+ Model. *2021 Telecoms Conference (ConfTELE)*, 2021, pp. 1-6.

Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.C., and Tan, M. (2019). Searching for MobileNetV3. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314-1324

Shamsoshoara, A., Afghah, F., Razi, A., Zheng, L., Fulé, P.Z., and Blasch, E. (2021). Aerial imagery pile burn detection using deep learning: The FLAME dataset, *Computer Network*s, 193, p. 108001.