

Fire Detection Automation in Search Drones using a Modified DeepLabv3+ Approach

Abhinav Choudhary

Adolfo Perrusquía

Saba Al-Rubaye

Weisi Guo

Abstract—Drones have become a key component in current search and rescue applications such as wildfire detection. The accurate detection of fire in forests plays a crucial global factor to reduce environmental damage and the preservation of wildlife. Current fire detection systems combine the merits of expert-learning systems and light-weight deep learning architectures. The key idea is to introduce color-based rules to identify potential fire pixels and create the associated mask that feeds a light-weight convolutional neural network (CNN) for image segmentation. However, expert learning systems are not robust and suffer of cognitive bias that induce a high number of false positives. In addition, CNN-based architectures cannot capture long-range dependencies reducing the segmentation fidelity. To overcome these gaps, this paper proposes a light weight deep learning (DL) architecture for fire segmentation. The approach is inspired in the Deeplabv3+ architecture for image segmentation. The novelty lies in the incorporation of vision transformers that heavily reduces the model complexity and avoid the usage of color-based rules. Experiments are conducted using open-access fire datasets. The results demonstrate competitive performance and highlight its potential use in drones applications.

I. INTRODUCTION

Wildfire is a global concern that has impacted both ecosystems and human societies [1]. Early detection of fire is crucial to prevent catastrophic events for the wildlife and the surrounding environments including natural ecosystems and cities. Aerial platforms or drones have been used in search and rescue applications [2] due to its bird view capabilities that helps to identify and explore high risk zones [3]–[5]. Hence, drones have become essential devices for early wildfire detection whose ultimate goal is to enhance the protection of both national ecosystems and facilities.

Visual sensors and deep learning (DL) techniques are the main tools used in search drones to detect fire [6] due to their low-computational cost and high detection accuracy. Here, architectures with low-computational cost define a mandatory requirement in drone's applications due to the controller's low-memory capabilities. To satisfy this requirement, the state-of-the-art has focused on detection methods that exploit the capabilities of DL methods for image segmentation enhanced by prior information about fire features and properties.

Most of the fire detection methods are based on a colour-based approach that defines expert rules to identify fire pixels.

We would like to thank Haydn Thompson from Ththink in sponsoring this project.

This work was supported by the Engineering and Physical Sciences Research Council under Grant EP/V026763/1.

Abhinav Choudhary, Adolfo Perrusquía, Saba Al-Rubaye, and Weisi Guo are with the School of Aerospace, Transport and Manufacturing, Cranfield University, MK43 0AL Bedford, UK. email: adolfo.perrusquia-guzman@cranfield.ac.uk

In this context, fire is associated with orange, red, and yellow colours which can be effectively distinguished in forests images or frames. Fire colour extraction is done by applying some expert rules based in the intensity and saturation of the red colour in a RGB colour space [7]. Fuzzy colour models appear as a natural choice for fire detection applications due to the definition of expert rules [8]. In this approach, it is analysed a sequence of images to identify variations of red pixels compared with the mean intensity value of the background red pixels. However, these approaches are not robust due to the diverse colour spectrum in imagery data and variations in either the illumination or lightning conditions.

The YCbCr colour space [9] has been used to overcome the robust problems of the RGB space. The main advantage of using this colour space is that the luminance and chrominance of the image can be better identified and establish new improved expert rules [10], [11]. The use of expert rules can be exploited by unsupervised methods to group fire pixels according to their saturation, colour, and lightness [12].

Fire has a dynamic and changes according to the amount of oxygen and fuel available in the environment. Motion-based techniques have been adopted for fire detection to deal with the dynamic propagation of fire. Some approaches use optical flow method and colour-based decision rules to extract the motion and color features from fire pixels [13], [14]. An improvement of motion-based models is to focus on the pixels within the grey region adjacent to fire [15]. However, motion-based techniques applied in drones are not convenient due to the diverse motion exhibited by drones [16] that causes that the scene changes constantly. This could make difficult to determine if the motion belongs to the fire or to the scene.

Other approaches for fire detection are texture-based methods. Here, fire creates an unique and chaotic texture in imagery data that are completely different from other textures observed in the images or frames. In fire detection applications, texture-based methods focus on areas of the image or frame that exhibit changes in the pixel intensities. Famous methods for fire detection in static images [17] are Gabor filters [18], [19] and Local Binary Patterns (LBP) [20].

Deep Learning semantic segmentation methods have been used to identify the main characteristics of fire including its shape and texture [21]. The most famous architectures for image segmentation are based on deep convolutional neural networks (CNN) such as U-Net [22], fully convolutional networks (FCNs), Mask R-CNN [23], and Deeplabv3 [24], [25]. Colour-based approach and U-Net have been combined to segment fire pixels from the combination of infra-red and RGB

images [26]. Deeplabv3+ [27] is an autoencoder architecture for image segmentation. This architecture has been applied for semantic segmentation of fire pixels [28]. The main issue of these CNN-based methods is that they cannot capture long-range dependencies between objects in an image. In addition, the most accurate approaches are computationally expensive and cannot be used in drones.

In view of the above, the following gaps are identified:

- 1) Expert methods are not robust and can be confounded by changes in the lighting conditions, obstructed views or objects with similar colours.
- 2) Expert rules have cognitive bias which can lead to a large number of false positives, that is, non-fire pixels are recognized as fire pixels.
- 3) DL segmentation methods are computationally expensive such that they cannot be used in drones. In addition, these methods cannot capture long-range dependencies between objects.

In this paper, we proposed a light-weight DL architecture for fire segmentation whose ultimate goal is to solve the aforementioned gaps. This is done by modifying the deep CNN (DCNN) backbone architecture in Deeplabv3+ with a transformer-based architecture, MobileViT, that reduces the computational requirements of the architecture and avoid the definition of expert rules by preserving the long-range properties of Deeplabv3+ network. With the replacement of the DCNN backbone, the segmentation algorithm will accelerate the detection capabilities, making it highly suitable for real-time wildfire monitoring applications using drones. Experiments are conducted in open-access datasets of fire and compared with the ResNet50 backbone architecture to demonstrate the effectiveness and competitiveness of the proposed approach.

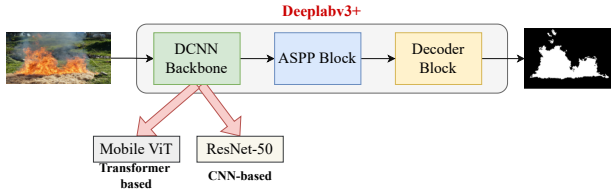


Fig. 1. High-level scheme of the proposed methodology

II. METHODOLOGY

Fig. 1 depicts the general diagram of the proposed fire segmentation architecture. The architecture is based on the Deeplabv3+ for semantic segmentation [29] which has demonstrated good segmentation results in imagery data [30].

A. Deeplabv3+ architecture

Deeplabv3+ (see Fig. 2) has an encoder-decoder structure composed of three main structures: an encoder block, the ASPP (Atrous Spatial Pyramid Pooling) block, and the decoder block. This architecture has shown high accurate results by exploiting the long-range capabilities of the ASPP block

whose main objective is to enlarge the field of view of the convolution filters [31]. Here, the encoder is responsible to uncover contextual nuances in the image by analysing high-dimensional features at multiple scales. On the other hand, the decoder block recovers the spatial information and sharp boundaries of the input images.

However, Deeplabv3+ demands high computational resources due to its DCNN backbone architecture (typically the ResNet 50 architecture is used) and hence, it is not suitable in search drones. To maintain this performance, whilst reducing the computational complexity we proposed to modify the DCNN backbone with a self-attention transformer-based architecture known as MobileViT. This new backbone architecture improves the training of the network and reduces the number of parameters, whilst maintaining high competitive segmentation results.

B. Mobile Vision Transformers (MobileViT)

MobileViT [30] is an architecture that combines the benefits of both CNN and vision transformers (ViT). This particular architecture incorporates inverted residual blocks of mobilenetv2 [32] with a novel “MobileViT block”. Here, the main advantage of the MobileViT block is the global processing of the pixels in an image using transformers rather than local processing using a DCNN. As a result, the MobileViT block will maintain the spatial inductive biases of CNNs and the global processing capabilities of ViTs.

Fig. 3 depicts the elements that comprise the MobileViT block. Here, $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{H \times W \times C}$ represent input and output tensors with height H , width W and C channels; and the $n \times n$ Conv block represents the standard $n \times n$ convolution operation. MobileViT Block performs the following tasks

- **Local Representation Operation:** MobileViT block applies a standard $n \times n$ convolution layer followed by a 1×1 convolutional layer to an input tensor $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$. This gives as output a tensor $\mathbf{X}_i \in \mathbb{R}^{H \times W \times d}$, where $d > C$ is a higher-dimensional space.
- **Unfolding Operation:** The tensor $\mathbf{X}_i$ is decomposed and flattened into N disjoint patches, generating a tensor $\mathbf{X}_U \in \mathbb{R}^{P \times N \times d}$, where P is the breakdown of patches having dimension $w \times h$ (width w and height h of each patch). The unfolding operation helps to learn global representations with spatial inductive bias. The unfolded tensor $\mathbf{X}_U$ is then passed through the transformer of L blocks to encode the relationship between patches.
- **Folding Operation:** The tensor $\mathbf{X}_U$ can be seamlessly transformed into a tensor $\mathbf{X}_i \in \mathbb{R}^{H \times W \times d}$ to preserve the spatial information of the image and the sequences of each individual patch. This step allows to obtain accurate spatial feature representations in the subsequent stages of the network.
- **Fusion Operation:** A pointwise convolution is applied to the folded tensor $\mathbf{X}_i$ to project it to a low C -dimensional space. This gives an output tensor of dimension $H \times W \times C$ which is concatenated with the input tensor $\mathbf{X}$ to learn low-level features and hence, the output tensor has

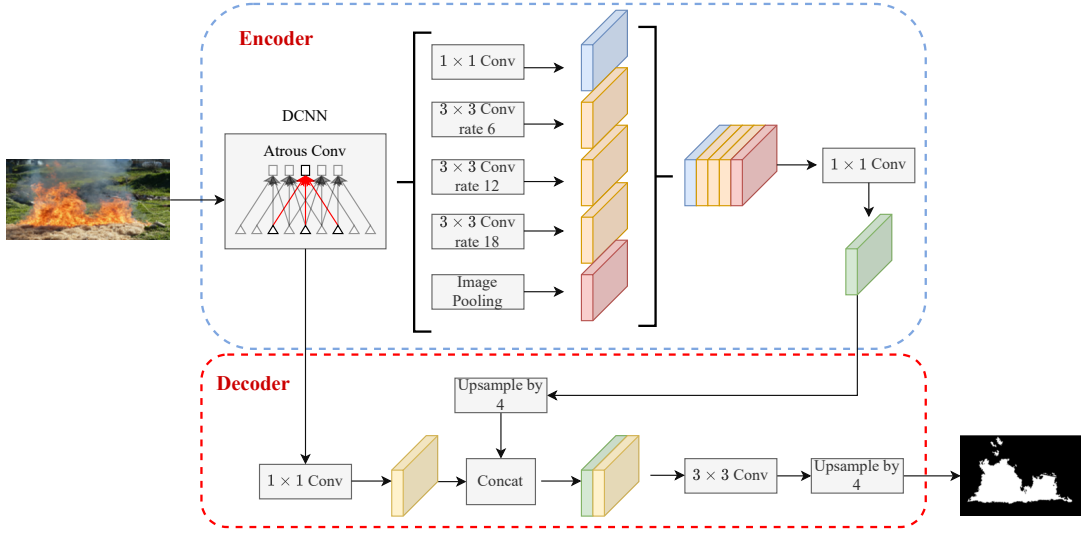


Fig. 2. Deeplabv3+ architecture adapted from [27]

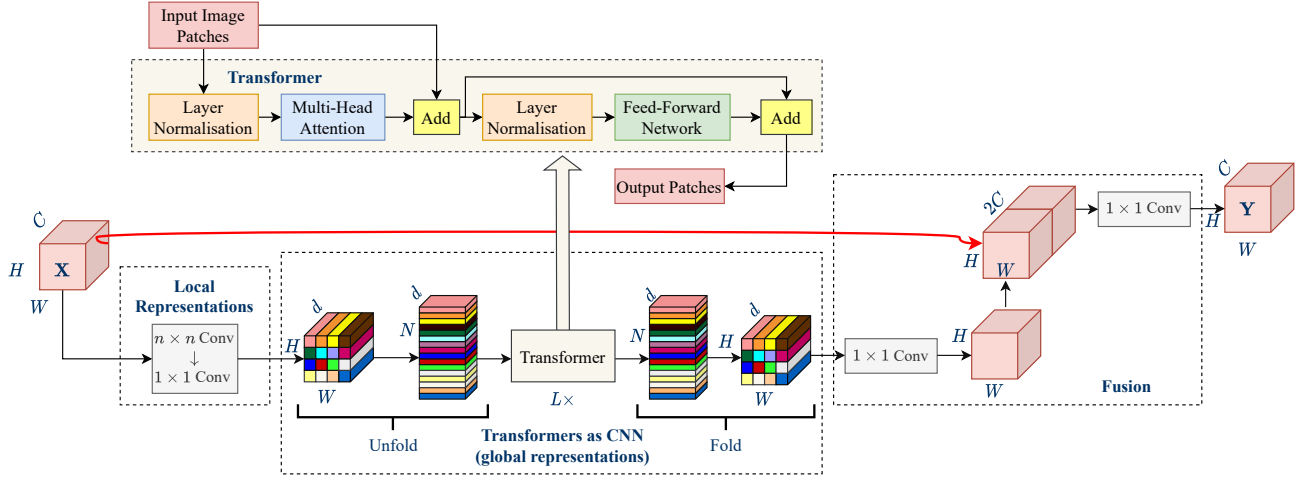


Fig. 3. Mobile ViT block architecture. Adapted from [30]

a dimension of $H \times W \times 2C$. This tensor is followed by a 1×1 convolution to fuse the concatenated tensors and we get the output tensor $Y \in \mathbb{R}^{H \times W \times C}$.

There is a transformer block within the MobileViT block which is mainly based in two sub-layers [30], [33] with residual connections and normalization layers to overcome the vanishing gradient problem and accelerate learning convergence. The first layer is dedicated to a self-attention mechanism with multiple heads that focus on multiple segments of the image simultaneously. A feed-forward layer is then used to process the output of the attention layer in a sequential order.

C. Proposed architecture

The proposed architecture is depicted in Fig. 4. Here, the DCNN block in Deeplabv3+ is replaced by a custom MobileViT network which maintains the spatial properties of Deeplabv3+ with the long-range detection and low-computational power of MobileViT. Here, we incorporate two

variants of the MobileViT architecture known as MobileViT_S and MobileViT_XS which stand for small and extra small architectures of MobileViT, respectively. These variants of MobileViT differ from each other according to the number of channels and projection dimensions. These parameters define the depth of MobileViT model, and the dimension of the output feature maps at different spatial resolutions. On the one hand, MobileViT_S has [144, 192, 240] projection dimensions and [16, 32, 64, 64, 96, 128, 160, 640] channels. On the other hand, MobileViT_XS has [96, 120, 144] projection dimensions and the number of channels are [16, 32, 48, 48, 64, 80, 96, 384]. It is evident that these models have less number of parameters and hence, the computational complexity is reduced.

III. DATASETS

One of the main challenges in wildfire detection is to obtain rich enough data that encompass different fire scenarios - including small and large fires, and top view images of fire

Conversely, in the FLAME dataset, the network cannot detect adequately the fire since most of the training images contain occluded and small fires. This issue is solved by using the combined dataset.

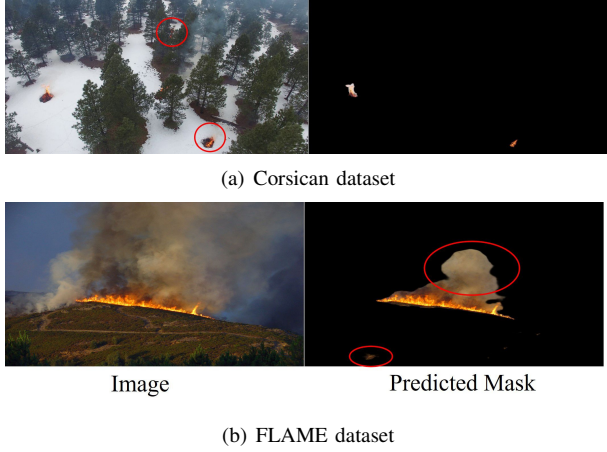


Fig. 5. Segmentation problems under individual datasets

We train the models using the combined dataset under each loss function to observe how they modify the final segmentation outcome. Table I exhibits the final results in the testing set. We can observe, that the MobileViT_XS-Deeplabv3+ exhibits the better performance under the BCE loss. This result indicates the power of vision transformers to replace the DCNN backbone for segmentation tasks. Fig. 6 shows the segmentation results using images of the testing set.

TABLE I
DETECTION RESULTS OF THE PROPOSED MODELS ACROSS ALL METRICS AND LOSS FUNCTIONS

Model	Loss	MPA	F1	IoU	Recall	Precision	MCC
MobileViT_S	Dice	0.9930	0.8954	0.8157	0.8904	0.9060	0.8921
	Tversky	0.9932	0.8959	0.8165	0.9243	0.8741	0.8930
	Focal Tversky	0.9928	0.8765	0.7883	0.8846	0.8778	0.8740
	Combo	0.9931	0.8926	0.8119	0.8823	0.9092	0.8896
	BCE	0.9928	0.8755	0.7863	0.8268	0.9398	0.8746
MobileViT_XS	Dice	0.9931	0.8892	0.8066	0.8654	0.9210	0.8866
	Tversky	0.9931	0.8971	0.8175	0.9095	0.8887	0.8935
	Focal Tversky	0.9930	0.8900	0.8081	0.9031	0.8827	0.8867
	Combo	0.9927	0.8950	0.8145	0.8884	0.9066	0.8914
	BCE	0.9869	0.9286	0.8697	0.9122	0.9507	0.9207
ResNet50	Dice	0.9942	0.9087	0.8371	0.9067	0.9151	0.9057
	Tversky	0.9941	0.9068	0.8340	0.9386	0.8812	0.9043
	Focal Tversky	0.9934	0.9018	0.8263	0.9371	0.8737	0.8992
	Combo	0.9942	0.9052	0.8314	0.9225	0.8924	0.9023
	BCE	0.9938	0.8837	0.7988	0.8581	0.9173	0.8815

The results of Fig. 6 demonstrate the proficiency of the proposed architecture to detect fire in diverse scenarios including large, small and occluded fires. Here, the weakness of DCNN to capture long-range dependencies is overcome and enhanced by the global processing capabilities of transformers. Furthermore, the proposed approach notably reduces the number of parameter of the network which is a fundamental requirement for search and rescue drones. Here, the number of parameters of each architecture are: MobileViT_S-Deeplabv3+ with 6,624,689 parameters, MobileViT_XS-Deeplabv3+ with 3,064,401 parameters, and ResNet50-Deeplabv3+ with 17,869,697 parameters. This

clearly demonstrate that the proposed models can achieve competitive results with a much smaller computational expense.

V. CONCLUSIONS

In this paper, a light-weight DL model for fire segmentation in search drones is proposed. The model is based on the image segmentation architecture Deeplabv3+. The deep convolutional neural network backbone of Deeplabv3+ is modified by a vision transformer architecture which reduces the computational complexity and enables global processing of the image rather than local processing. The vision transformer is based on two variants of the MobileViT network which highly reduces the number of parameters, whilst maintaining the processing properties for accurate image segmentation. The Corsican and FLAME datasets are combined to cover different fire scenarios simultaneously without focusing only in a single case. Different loss functions are used to observe the behaviour of the network and obtain the best model. Experiments are carried out under different metrics using the proposed combined dataset. The results show that the proposed architecture achieves competitive results with much less number of parameters in comparison with the well-used ResNet50-Deeplabv3+ architecture.

Future work will address the extension of the proposed model to detect both fire and smoke to improve the detection capabilities of the network and reduce the damage to both national facilities and ecosystems.

REFERENCES

- [1] J. G. Pausas and J. E. Keeley, "Wildfires and global change," *Frontiers in Ecology and the Environment*, vol. 19, no. 7, pp. 387–395, 2021.
- [2] C. Gruffeille, A. Perrusquía, A. Tsourdos, and W. Guo, "Disaster area coverage optimisation using reinforcement learning," in *2024 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE, 2024, pp. 61–67.
- [3] H. Yao, R. Qin, and X. Chen, "Unmanned aerial vehicle for remote sensing applications—a review," *Remote Sensing*, vol. 11, no. 12, p. 1443, 2019.
- [4] M. A. Akhloufi, A. Couturier, and N. A. Castro, "Unmanned aerial vehicles for wildland fires: Sensing, perception, cooperation and assistance," *Drones*, vol. 5, no. 1, p. 15, 2021.
- [5] M. El Debeiki, S. Al-Rubaye, A. Perrusquía, C. Conrad, and J. A. Flores-Campos, "An advanced path planning and uav relay system: Enhancing connectivity in rural environments," *Future Internet*, vol. 16, no. 3, p. 89, 2024.
- [6] M. Yandouzi, M. Grari, I. Idrissi, O. Moussaoui, M. Azizi, K. Ghomid, and A. K. Elmiad, "Review on forest fires detection and prediction using deep learning and drones," *J. Theor. Appl. Inf. Technol.*, vol. 100, no. 12, pp. 4565–4576, 2022.
- [7] T.-H. Chen, P.-H. Wu, and Y.-C. Chiou, "An early fire-detection method based on image processing," in *2004 International Conference on Image Processing, 2004. ICIP'04.*, vol. 3. IEEE, 2004, pp. 1707–1710.
- [8] T. Celik, H. Demirel, and H. Ozkaramanli, "Automatic fire detection in video sequences," in *2006 14th European signal processing conference*. IEEE, 2006, pp. 1–5.
- [9] S. Wu and L. Zhang, "Using popular object detection methods for real time forest fire detection," in *2018 11th International symposium on computational intelligence and design (ISCID)*, vol. 1. IEEE, 2018, pp. 280–284.
- [10] T. Celik, H. Demirel, H. Ozkaramanli, and M. Uyguroglu, "Fire detection in video sequences using statistical color model," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 2. IEEE, 2006, pp. II–II.

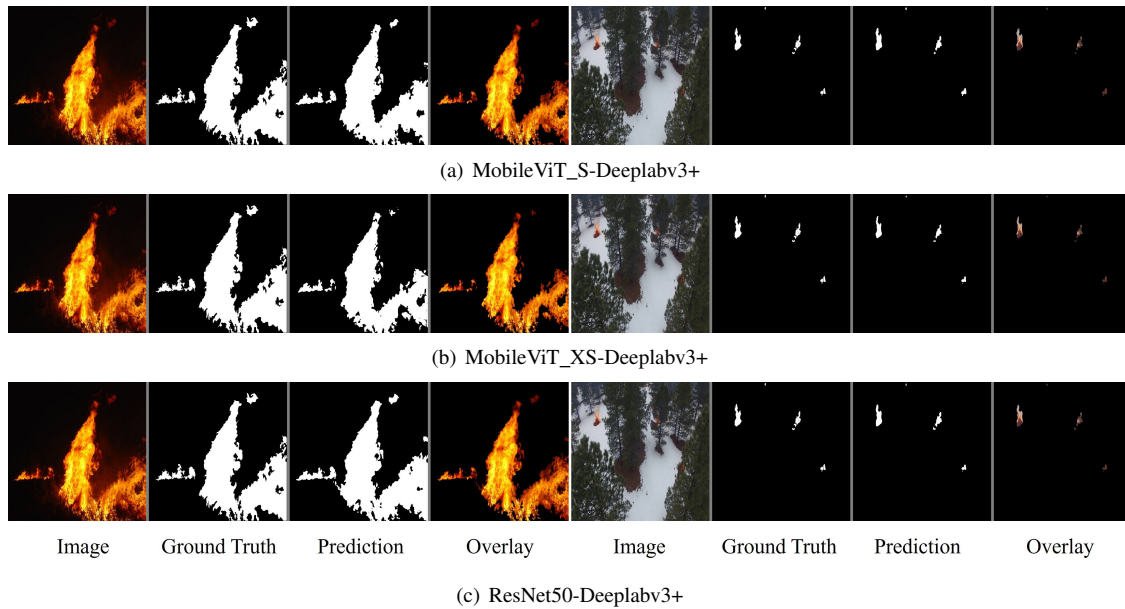


Fig. 6. Fire detection results using the combined dataset

- [11] T. Celik and H. Demirel, "Fire detection in video sequences using a generic color model," *Fire safety journal*, vol. 44, no. 2, pp. 147–158, 2009.
- [12] D. Dzigal, A. Akagic, E. Buza, A. Brdjanin, and N. Dardagan, "Forest fire detection based on color spaces combination," in *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)*. IEEE, 2019, pp. 595–599.
- [13] C. Yuan, K. A. Ghamry, Z. Liu, and Y. Zhang, "Unmanned aerial vehicle based forest fire monitoring and detection using image processing technique," in *2016 IEEE Chinese Guidance, Navigation and Control Conference (CGNCC)*. IEEE, 2016, pp. 1870–1875.
- [14] P. Patel and S. Tiwari, "Flame detection using image processing techniques," *International Journal of Computer Applications*, vol. 58, no. 18, 2012.
- [15] G. Yadav, V. Gupta, V. Gaur, and M. Bhattacharya, "Optimized flame detection using image processing based techniques," *Indian Journal of Computer Science and Engineering*, vol. 3, no. 2, pp. 202–211, 2012.
- [16] A. Perrusquia, W. Guo, B. Fraser, and Z. Wei, "Uncovering drone intentions using control physics informed machine learning," 2023.
- [17] D. Y. Chino, L. P. Avalhais, J. F. Rodrigues, and A. J. Traina, "Bowfire: detection of fire in still images by integrating pixel color and texture analysis," in *2015 28th SIBGRAPI conference on graphics, patterns and images*. IEEE, 2015, pp. 95–102.
- [18] W. Li, K. Mao, H. Zhang, and T. Chai, "Selection of gabor filters for improved texture feature extraction," in *2010 IEEE International Conference on Image Processing*. IEEE, 2010, pp. 361–364.
- [19] N. Sasikala and G. Shruthi, "Texture analysis of a color image using traditional and circular gabor filters," *International Journal of Advanced Engineering Research and Science*, vol. 3, no. 9, p. 236842.
- [20] R. K. YB and R. K. CN, "Local binary pattern: An improved lbp to extract nonuniform lbp patterns with gabor filter to increase the rate of face similarity," in *2016 Second International Conference on Cognitive Computing and Information Processing (CCIP)*. IEEE, 2016, pp. 1–5.
- [21] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, 2019.
- [22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
- [23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
- [24] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation. arxiv 2017," *arXiv preprint arXiv:1706.05587*, vol. 2, 2019.
- [25] A. Kumar, A. Perrusquia, S. Al-Rubaye, and W. Guo, "Wildfire and smoke early detection for drone applications: A light-weight deep learning approach," *Engineering Applications of Artificial Intelligence*, vol. 136, p. 108977, 2024.
- [26] T. Kacker, A. Perrusquia, and W. Guo, "Multi-spectral fusion using generative adversarial networks for uav detection of wild fires," in *2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*. IEEE, 2023, pp. 182–187.
- [27] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.
- [28] H. Harkat, J. Nascimento, and A. Bernardino, "Fire segmentation using a deeplabv3+ architecture," in *Image and signal processing for remote sensing XXVI*, vol. 11533. SPIE, 2020, pp. 134–145.
- [29] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.
- [30] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021.
- [31] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.
- [32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
- [33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [34] T. Toulouse, L. Rossi, A. Campana, T. Celik, and M. A. Akhloufi, "Computer vision for wildfire research: An evolving image dataset for processing and analysis," *Fire Safety Journal*, vol. 92, pp. 188–194, 2017.
- [35] A. Shamsoshoara, F. Afghah, A. Razi, L. Zheng, P. Z. Fulé, and E. Blasch, "Aerial imagery pile burn detection using deep learning: The flame dataset," *Computer Networks*, vol. 193, p. 108001, 2021.



DSpace

<https://dspace.org/>

Fire detection automation in search drones using a modified DeepLabv3+ approach

Choudhary, Abhinav; Perrusquía, Adolfo; Al-Rubaye, Saba; Guo, Weisi

2024-10-29

Institute of Electrical and Electronics Engineers (IEEE)

<https://doi.org/10.1109/isc260477.2024.11004293>

Choudhary A, Perrusquía A, Al-Rubaye S, Guo W. (2024) Fire detection automation in search drones using a modified DeepLabv3+ approach. In: 2024 IEEE International Smart Cities Conference (ISC2), 29 Oct - 1 Nov 2024, Pattaya, Thailand