

Attention on Classification for Fire Segmentation

Milad Niknejad

*Instituto de Sistemas e Robotica,
Instituto Superior Tecnico, University of Lisbon
Lisbon, Portugal
milad3n@gmail.com*

Alexandre Bernardino

*Instituto de Sistemas e Robotica,
Instituto Superior Tecnico, University of Lisbon
Lisbon, Portugal
alex@isr.tecnico.ulisboa.pt*

Abstract—Detection and localization of fire in images and videos are important in tackling fire incidents. Although semantic segmentation methods can be used to indicate the location of pixels with fire in the images, their predictions are localized, and they often fail to consider global information of the existence of fire in the image which is implicit in the image labels. We propose a Convolutional Neural Network (CNN) for joint classification and segmentation of fire in images which improves the performance of the fire segmentation. We use a spatial self-attention mechanism to capture long-range dependency between pixels, and a new channel attention module which uses the classification probability as an attention weight. The network is jointly trained for both segmentation and classification, leading to improvement in the performance of the single-task image segmentation methods, and the previous methods proposed for fire segmentation.

Index Terms—fire detection, semantic segmentation, deep convolutional neural network, multitask learning

I. INTRODUCTION

Every year fire causes severe damage to persons and property all over the world. Artificial intelligence can play an important role to battle the fire incidents by early detection and localization of the fire spots. Many methods have already been proposed for detection of fire and smoke on images and videos in different scenarios such as wildfires. Traditional methods were based on handcrafted features extracted mostly from individual pixel colors [1], [2]. Recently, analogously to many other computer vision areas, the state-of-the-art results have been achieved for fire detection using features from Convolutional Neural Networks (CNN). Methods were mainly proposed for classification of fire in images [3], [4]. Some methods consider the localization of fire in images as well [5], [6]. Localization of fire is important for determining the exact spot of the fire in images which has applications in autonomous systems and geo-referencing of the fire location. Like [21], [22], we consider pixel-wise segmentation for fire localization, which corresponds to the binary semantic segmentation to detect fire in images. Although bounding boxes can be used for localization, pixel-wise segmentation has advantages e.g. it can be used as input for fire propagation models. However, most segmentation methods are localized and do not consider global contextual information in images. In the case of fire detection, even recent well-known segmentation methods produce many incorrect false positive

pixel segmentations for fire-like images due to these localized predictions (see Fig. 4). This false positive prediction is an important issue in fire detection as it may lead to false alarms.

Recently, self-attention mechanisms have attracted lots of interest in computer vision. The purpose of self-attention methods is to use long range information and increase the receptive field sizes of current deep neural networks [23]. Self-attention tends to capture the correlation between different image regions by computing a weighted average of all features in different locations in which the weights are computed based on the similarities between their corresponding embeddings. Some works consider deep architecture composed of only self-attention layers as a replacement for the convolutional networks [24]. Apart from self-attention, which uses the input features itself to compute the attention coefficients, some methods proposed to use attention based on the features extracted from other parts of the network [27].

Multi-task learning methods learn simultaneously multiple correlated computer vision tasks (e.g. semantic segmentation and depth estimation) in a unified network through learning common features [12]–[14]. It has been shown that this multi-task learning leads to improvement in performance and reduction of training complexity compared to using separate networks for each task. In our application, the features in the higher layers of a CNN contain both localization and classification informations [7], [8]. Consequently, in [15], a method is proposed for joint classification and segmentation of medical images, in which a classification network is applied to the features of the last layer of the encoder (the coarsest layer) in a encoding-decoding segmentation CNN.

In this paper, we propose a new CNN that jointly classifies and segments fire in images with improved segmentation performance. We propose an attention mechanism that uses the classification output as the channel attention coefficient of the segmentation output. This allows the overall network to consider the global classification information on the segmentation masks. Furthermore, we use a self-attention model to capture the long-range spatial correlations within each channel. Experiments show that the proposed method with the attention mechanism outperforms other methods in the segmentation metrics. It reduces the false positive results in the segmentation masks, while at the same time, is able to identify small scale fires in images, resulting in state-of-the-art results among fire segmentation methods.

In the following sections, we first mention related works for segmentation, multi-task learning, and self-attention. We then describe our proposed method in detail, and finally compare our method with other segmentation, and multitask classification-segmentation methods.

II. RELATED WORKS

Traditional methods for fire detection mainly use hand-crafted features such as color features [1], [2], covariance-based features [16], wavelet coefficients [17], and then classify the obtained features using a vector classifier e.g. a Support Vector Machine (SVM). Recently, methods based on CNN have improved the performance of fire and smoke detection noticeably. The method in [3] uses simplified structures of the Alexnet [18] and Inception networks [19] for fire image classification. In [4], Faster Region-CNN (R-CNN) [20] is used to extract fire candidate regions, and is further processed by multidimensional texture analysis using Linear Dynamical Systems (LDS) to classify the fire images.

Some works consider localization of fire in images, beyond classification. In [5], a method for classification and patch-wise localization was proposed in which the last convolutional layer of the classification network is used for the patch classification. In [6], a combination of color features, and Faster R-CNN is used to increase the efficiency of the algorithm by disregarding some anchors of R-CNN based on some color features. Some methods consider pixel-wise segmentation for fire images as well. In [21], deep-lab semantic segmentation is adapted for pixel segmentation of fire. In [22], a new CNN architecture is proposed for segmentation of fire in images.

In computer vision, single end-to-end multi-task networks have shown promising results for the tasks that have cross-dependency such as semantic segmentation and depth estimation [12], [13]. They benefit from learning common features. It has been shown that exploiting the cross-dependency between the tasks lead to improvement in performance compared to the networks independently trained for the two tasks [13]. It has other benefits such as reducing the training time. It is known that the features in the last convolutional layer in CNNs trained for classification have also spatial information for localization [7], [8]. Le et. al. [15] proposed a method for joint classification and segmentation for cancer diagnosis in mammography, in which the last convolution layer of the encoder in the segmentation network is used for the global classification.

Self-attention models have recently demonstrated improved results in many computer vision tasks [11], [23], [24]. Self-attention models compute attention coefficients based on the similarities between input features. New features are then obtained by a weighted average of the input features with the self-attention coefficients. Beyond self-attention, there are also attention mechanisms proposed for image classification [25], [26], and semantic segmentation [27], in which the attention weights are computed using the features in other parts of the CNN.

III. PROPOSED METHOD

A simple approach to learn a joint classification and segmentation in a unified CNN is to classify images based on the features after global pooling of the coarsest layer (last encoding layer) of the encoder-decoder segmentation network. The network can be jointly trained with the classification and segmentation labels through a weighted loss. This approach has been previously proposed in [15] for medical imaging application.

In this paper we add two attention modules to consider both global classification score and correlation in the spatial locations, in the segmentation predictions. As the first attention module, we propose to use a channel attention in the segmented output. As shown in Fig. 1, it multiplies the attention weight to the output channel, in which the weight is the probability assigned by the classification branch of the network. The channel is then added to the resulting features, similar to self-attention approaches [23]. Let, $\mathbf{x} \in \mathbb{R}^{W \times H \times 3}$ be the RGB image, and let $s(\mathbf{x}) \in \mathbb{R}$ and $\mathbf{A}(\mathbf{x}) \in \mathbb{R}^{W \times H \times 1}$ be the classification probability and segmentation features extracted by the CNN, respectively. $\mathbf{A}(\mathbf{x})$ could be the output of any segmentation network. In this paper, we use deeplab v3+ encoder and decoder [9] to extract the features. $s(\mathbf{x}) \in [0, 1]$ is computed by the sigmoid function of the classification scores obtained by the classification branch (see Fig. 1). Following [11], we compute the channel attention model as

$$\mathbf{A}'(\mathbf{x}) = \mathbf{A}(\mathbf{x}) + \alpha s(\mathbf{x}) \mathbf{A}(\mathbf{x}) \quad (1)$$

where \mathbf{A}' indicates the features after applying the attention module, and α is a parameter which is initialized to zero and learnt during training [11]. The method in [11] used a channel attention model for the segmentation in which the channel weights are obtained by the features themselves using a self-attention approach. However in our method, the weight $s(\mathbf{x})$ is the classification probability which is computed in the classification branch. This approach is supposed to reduce the false positive results as the correct classification output $s(\mathbf{x})$ for a non-fire image is close to zero, so it attenuates the activation of the segmentation output \mathbf{A}' . In the case of fire, a value of $s(\mathbf{x})$ close to one helps to recognize even small portions fire in images. This also encourages the consistency of the results between the segmentation and classification outputs.

Although the above global attention scheme considers the image label information, the performance of the method can be further improved by considering the correlation between the features. To achieve this, besides the classification attention model, we also apply spatial attention model to consider the correlation between features in different locations within each channel. This attention model is exactly the same as proposed in non-local neural networks of [23], in which each feature is replaced by a weighted average of all features based on the similarities between their corresponding embeddings. The general structure of the spatial attention module as shown in Fig. 2 is to apply two 1×1 convolution layers to the input features, and reshape the result to obtain the similarity

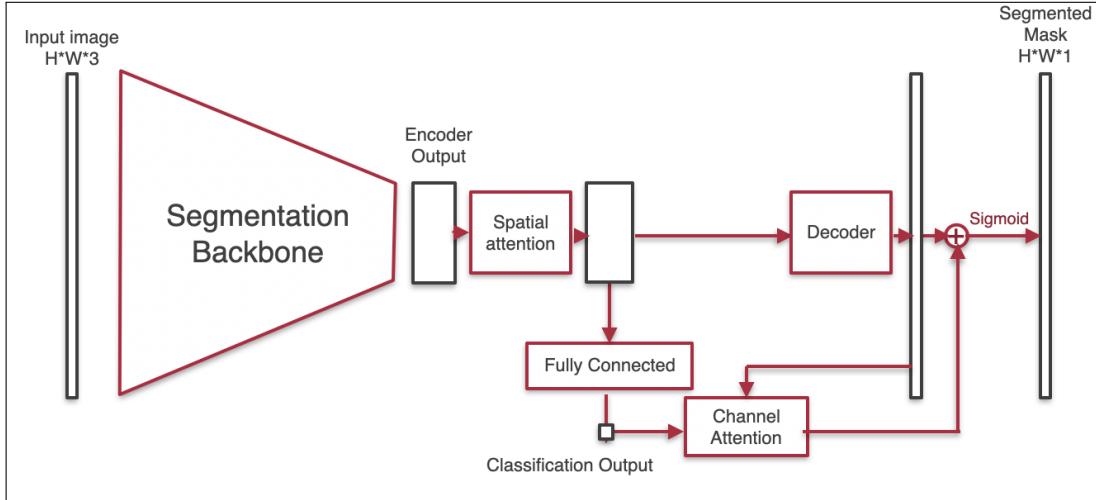


Fig. 1: Proposed CNN architecture for joint classification and segmentation; for the segmentation backbone deeplab v3 [28] is used.

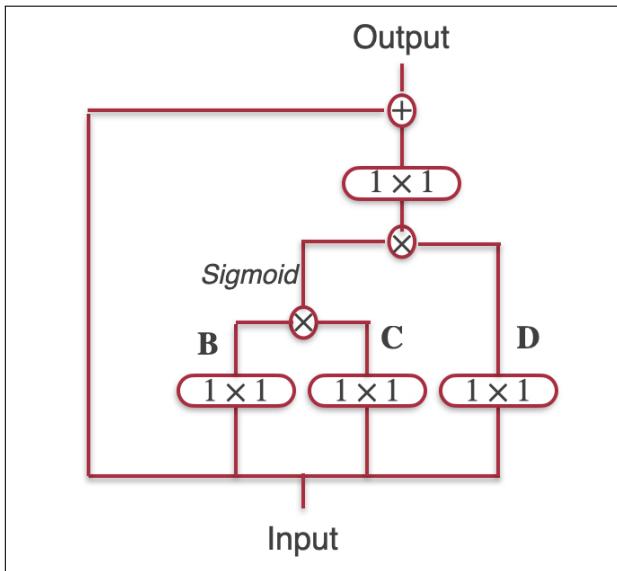


Fig. 2: Spatial self-attention module used on our method to capture long range spatial dependency which is proposed in [23]. The rounded boxes indicate the convolution operator.

embeddings $\mathbf{B} \in \mathbb{R}^{N \times C}$ and $\mathbf{C} \in \mathbb{R}^{N \times C}$, where $N = H \times W$, and C is the number of channels. The two matrices are used to compute a similarity matrix $\mathbf{S} = \mathbf{B}\mathbf{C}^T$. The row-wise softmax of the matrix \mathbf{S} is multiplied to the matrix \mathbf{D} , which results from another 1×1 convolution to the input. The resulting is added to the input. Figure 2 shows the self-attention module in our method.

Spatial and channel attentions have been used for segmentation of general images in [11]. However, our method uses completely different channel attention module based on the classification probabilities, while [11] uses a self-attention module.

Following the common approach in multitask learning, we use a weighted sum of the classification and segmentation losses. Let L_S , and L_C , be the segmentation and classification losses, respectively. The training loss is computed by

$$L = \lambda L_S + (1 - \lambda) L_C \quad (2)$$

where $\lambda \in [0, 1]$ is an appropriate regularization parameter. We use binary cross-entropy loss for both L_C , and L_S .

IV. EXPERIMENTAL RESULTS

In this section, We evaluate our proposed method and compare it to other segmentation methods and multitask methods for joint segmentation and classification. In order to evaluate the performance for false positive segmentation, we compute the label inferred from the segmented image by $\mathbf{1}(\sum_{i,j} \mathbf{M}_{i,j})$ where $\mathbf{M}_{i,j}$ indicates the output mask at pixel i, j , and $\mathbf{1}$ is the indicator function. It is considered zero if all pixels in the output mask are zero, and one otherwise. We use the accuracy between the segmented label and the image label in our comparisons which is called average consistency in Table I.

We create a dataset by combining RGB images and their associated segmentation masks in the Corsican fire dataset [29], and non-fire images in [30], containing some images which are likely to cause false positive results. We divided the dataset into train, validation, and test groups with 60, 20 and 20 percent, respectively.

In this section, our proposed segmentation method is compared to U-net [10], deeplab adapted for fire segmentation [21], and the method in [22] which proposed a new architecture for fire segmentation. We also compare our proposed method with other joint classification-segmentation methods. Inspired by [15], we consider a multi-task approach, which applies a classification network to the output of the encoder of the segmentation network. This method corresponds to our proposed method in which all attention blocks are removed.

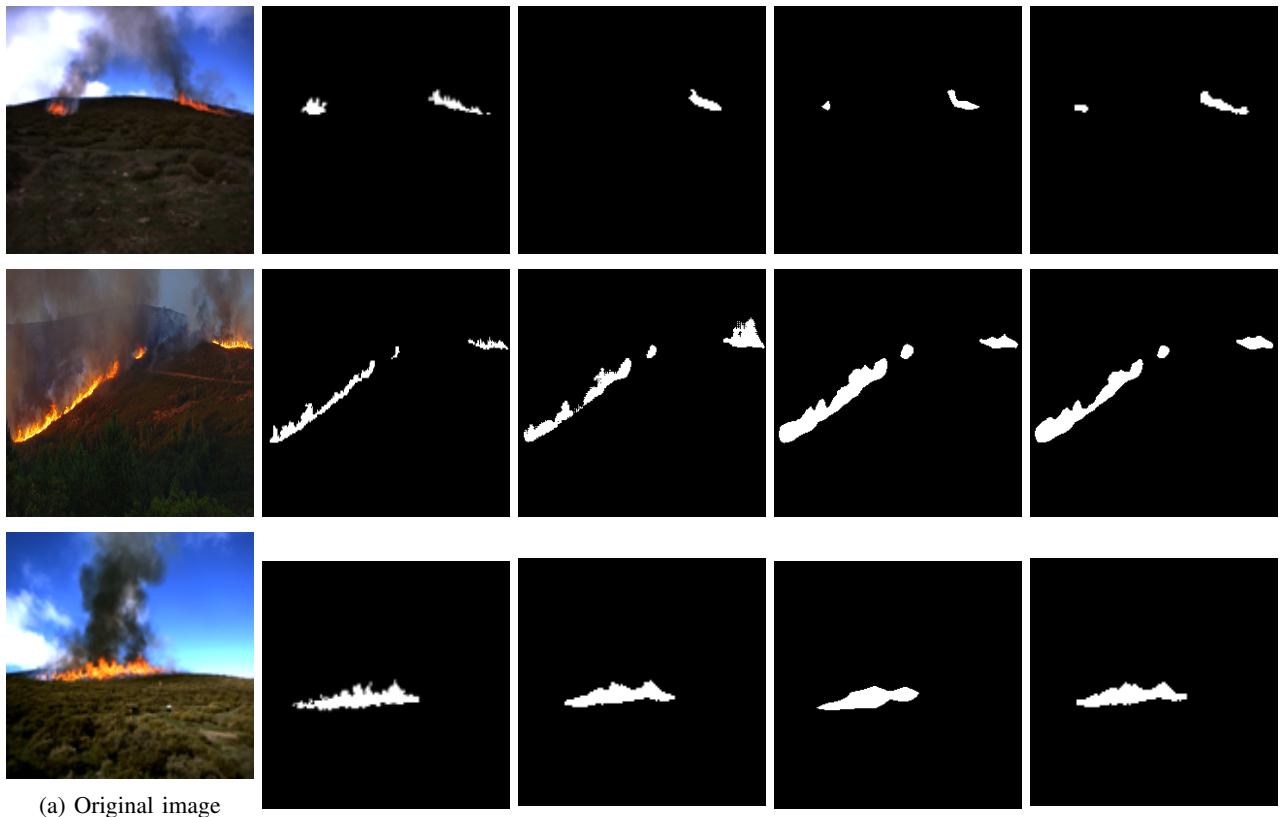


Fig. 3: Examples of the segmentation of fire in images in our proposed method compared to other methods.

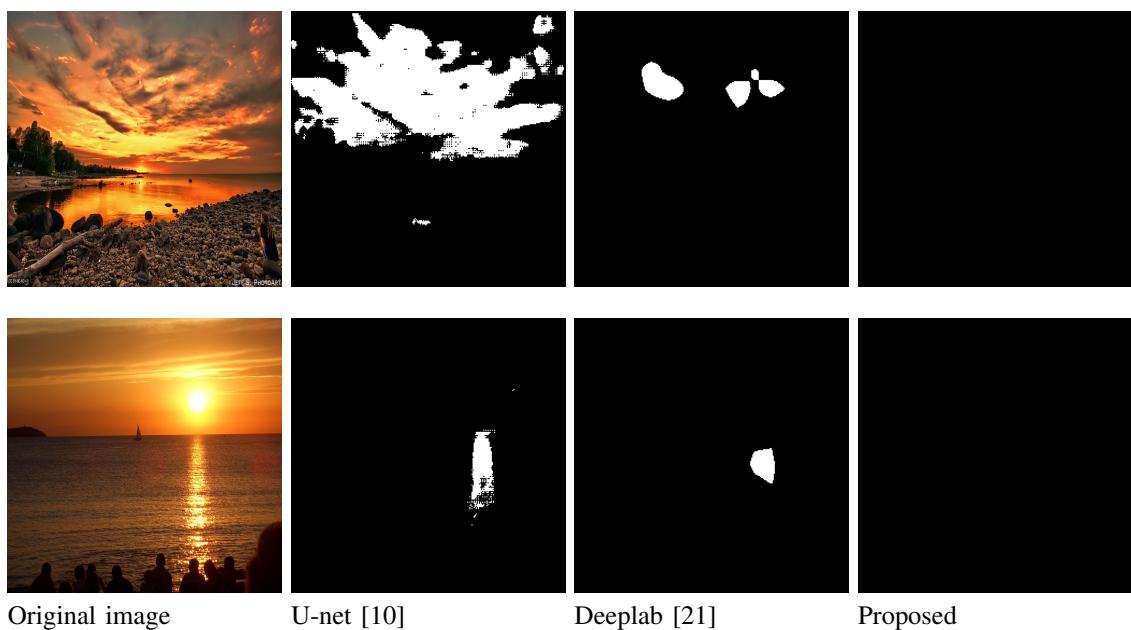


Fig. 4: Examples of segmentation of fire part for images that are likely to produce false positive results.

	Classification metrics	Segmentation Metrics		Consistency
		mean Accuracy	mean IOU	
U-net [10]	-	96.88	87.45	.8521
Deeplab [21]	-	97.18	88.34	.8876
Fire segmentation in [22]	-	97.06	88.02	.8623
multi-task network in [15]	98.75	96.55	87.22	.8912
Naive multi-task	98.75	97.21	90.02	.9654
Proposed	99.12	98.02	92.53	.9823

TABLE I: Comparison of the proposed method with the baseline for classification-segmentation, and U-net for segmentation.

We also consider a simple approach for removing false positives in segmentation in which the segmentation mask is set to zero if the classification output is zero (without using any attention). This basically relies on the classification output for segmentation. We call this method the naive approach.

The proposed method is implemented in the following settings. The encoder of Deeplab-v3+ is used as the encoder backbone for the segmentation network. The network is then initialized by pre-trained ImageNet weights for the segmentation backbone, and i.i.d normal random weights with mean zero and standard deviation of .05 for the classification branch. The weights are learned during the training by the ADAM algorithm [31] with initial learning rate of 5×10^{-4} , and a weight decay of 10^{-5} . The loss regularization parameter λ is set to .6, empirically, to achieve the best performance in terms of the overall validation loss. All other methods were trained in our dataset with ADAM algorithm with the parameters which performs the best in the validation set. For fire segmentation method in [22], we trained our own implementation as the source codes are not available online.

We report the result of our proposed method and other mentioned methods, on the test set, in Table I. The main metric for assessing the performance of semantic segmentation methods is intersection over union (IOU). This value is computed in this table for the classes of background and fire. In the classification metric, we compare the accuracy between the ground truth image labels and the predicted labels. This metric is obviously valid for multitask networks that have the image classification branch. In the segmentation metrics, the pixel accuracy (averaged over all tests images) and mean IOU are reported. As it can be seen, in the IOU segmentation metric, the proposed method outperforms segmentation methods of U-net, Deeplab adapted for fire segmentation in [21], and the fire segmentation method in [22]. IOU is also improved over the joint segmentation and classification method of [15], and the naive approach described above. Some examples of fire segmentation on test dataset are shown in Fig. 3. In the first row, it can be seen that our method could capture small portion of fire in the image. Besides that, based on the results on the table, the proposed method performs better in the consistency metric which is defined in the first paragraph of this section (the accuracy between the inferred label from the segmentation output and the ground truth label). This metric shows that the segmented image better corresponds to the image label in our method, i.e. reducing the cases in which some pixels are assigned as fire in non-fire images. This is a common problem

in fire segmentation as illustrated in Fig. 4 for two images which are prone to false positive outputs. As it can be seen, other methods mistakenly select some parts of both images as fire, while the proposed method correctly does not segment any pixel as fire.

V. CONCLUSION

In this paper, we proposed a method for joint classification and segmentation of fire in images based on CNN using attention. We used a channel attention mechanism in which the weight is based on the output in the classification branch. A self-attention mechanism is used for spatial attention. Our method shows improved segmentation results over other fire segmentation methods, and other multitask CNN structures.

REFERENCES

- [1] 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.
- [2] 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.
- [3] A. J. Dunnings and T. P. Breckon, “Experimentally defined convolutional neural network architecture variants for non-temporal real-time fire detection,” in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1558–1562.
- [4] P. Barmpoutis, K. Dimitropoulos, K. Kaza, and N. Grammalidis, “Fire detection from images using faster r-cnn and multidimensional texture analysis,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8301–8305.
- [5] Q. Zhang, J. Xu, L. Xu, and H. Guo, “Deep convolutional neural networks for forest fire detection,” in *2016 International Forum on Management, Education and Information Technology Application*. Atlantis Press, 2016.
- [6] C. Chaoxia, W. Shang, and F. Zhang, “Information-guided flame detection based on faster r-cnn,” *IEEE Access*, vol. 8, pp. 58 923–58 932, 2020.
- [7] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [8] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.
- [9] 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.
- [10] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [11] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

- [12] O. H. Jafari, O. Groth, A. Kirillov, M. Y. Yang, and C. Rother, “Analyzing modular cnn architectures for joint depth prediction and semantic segmentation,” in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 4620–4627.
- [13] T. Dharmasiri, A. Spek, and T. Drummond, “Joint prediction of depths, normals and surface curvature from rgb images using cnns,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 1505–1512.
- [14] Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, and J. Yang, “Joint task-recursive learning for semantic segmentation and depth estimation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 235–251.
- [15] N. Thome, S. Bernard, V. Bismuth, F. Patoureaux *et al.*, “Multitask classification and segmentation for cancer diagnosis in mammography,” in *International Conference on Medical Imaging with Deep Learning–Extended Abstract Track*, 2019.
- [16] Y. H. Habiboglu, O. Günay, and A. E. Çetin, “Covariance matrix-based fire and flame detection method in video,” *Machine Vision and Applications*, vol. 23, no. 6, pp. 1103–1113, 2012.
- [17] B. U. Töreyin, Y. Dedeoğlu, U. Güdükbay, and A. E. Cetin, “Computer vision based method for real-time fire and flame detection,” *Pattern recognition letters*, vol. 27, no. 1, pp. 49–58, 2006.
- [18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [20] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, 2015, pp. 91–99.
- [21] H. Harkat, J. M. Nascimento, and A. Bernardino, “Fire detection using residual deeplabv3+ model,” in *2021 Telecoms Conference (ConfTELE)*. IEEE, 2021, pp. 1–6.
- [22] S. Frizzi, M. Bouchouicha, J.-M. Ginoux, E. Moreau, and M. Sayadi, “Convolutional neural network for smoke and fire semantic segmentation,” *IET Image Processing*, vol. 15, no. 3, pp. 634–647, 2021.
- [23] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.
- [24] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, “Attention augmented convolutional networks,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3286–3295.
- [25] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.
- [26] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, “Learn to pay attention,” *arXiv preprint arXiv:1804.02391*, 2018.
- [27] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, “Attention u-net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.
- [28] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” *arXiv preprint arXiv:1706.05587*, 2017.
- [29] 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.
- [30] 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.
- [31] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.