2024 IEEE International Conference on Image Processing (ICIP) | 979-8-3503-4939-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIP51287.2024.10647443

A CONTEXT-ORIENTED MULTI-SCALE NEURAL NETWORK FOR FIRE SEGMENTATION

Tony Zhang and Robert P. Dick

Electrical and Computer Engineering Department
University of Michigan
Ann Arbor, MI, USA

ABSTRACT

Existing image-based fire segmentation techniques use convolutional neural networks to handle complicated scenes. Such approaches perform poorly when flame sizes vary greatly and when backgrounds are complex. In this paper, we describe a novel Context-Oriented Multi-Scale Network for fire segmentation. We construct a multi-scale aggregation module that combines semantic information at different levels in the neural network in order to recognize fires with different shapes and sizes. We also describe a Context-Oriented Module, which increases the receptive field of the network by utilizing relationships of all pixels in the feature map in order to obtain features that more effectively discriminate between fire and non-fire pixels. Experimental results demonstrate that our proposed model has a 2.7% higher mean Intersection over Union (mIoU) accuracy than previous fire detection methods.

***Index Terms*—** Fire segmentation, multi-scale, context

1. INTRODUCTION

Since wildfires spread quickly and can be difficult to control, detection and suppression speeds are crucial. Wildfires cause millions of dollars of damage and kill thousands of people per year. While traditional fire detection technologies such as smoke sensors are inexpensive, they only detect nearby fire sources. Hence, there is an increasing interest in long-range, image-based fire detection.

The earliest image-based fire detection techniques use hand-crafted features from color, shape, and texture to detect fire regions [1, 2]. With deep learning algorithms achieving remarkable progress in many fields [3, 4], they were also applied to fire detection recently [5, 6].

In image segmentation, deep learning methods have better performance than earlier methods using predetermined features, such as U-Net [7] and PSP-Net [8]. Hossain et al. detect forest fires with a neural network using color space local binary patterns of both flame and smoke signatures [6]. Choi et al. assign pixel-level labels of fire in images via a CNN residual network [9]. A recent study performed fire segmentation using a squeezed fire binary segmentation network with depthwise separable convolutions [10].

Fig. 1. Images contain fire with different kinds of shapes, sizes, and illumination. The left column contains the original image and the right column contains the ground truth segmentation map. It is important to recognize flames that are present and also minimize false alarms.

Despite the progress of fire detection methods, the accuracy of existing models decreases for many difficult scenarios. For example, small or occluded flames are difficult to identify. Also, complex backgrounds make it difficult to distinguish the fire from its surroundings and objects with similar color. Finally, the highly variable sizes, shapes, and colors of flames exacerbate the problem of fire segmentation.

Determining scene context, which refers to relationships among distant pixels, reduces false positives and false negatives. To handle small flame sizes (e.g., less than 5% of the image) as well as differentiate between the flame and background, it is necessary to enlarge the receptive field in order to effectively determine relationships among distant pixels. Also, to handle multiple scales of flames, multi-scale aggregation selectively combines useful information from different network layers. However, existing fire detection methods do not take into account these two important factors.

In this paper, we propose a Context-Oriented Multi-Scale CNN. It does multi-scale aggregation (MSA), which outputs

the segmentation map from multi-scale features and adaptively refines the features. We also introduce a novel Context-Oriented Module (COM) for our fire detection network. It extracts discriminant feature representations by building associations among features with global context, which uses relationships of all pixels in the feature map. In the COM, the input is fed into multiple branches with convolutions, average pooling, and global pooling. Then, the COM integrates the features from all branches.

High-resolution CNNs well model high-resolution relationships among nearby locations in the image, but their inductive biases make modeling long-range relationships difficult. Low-resolution, downsampled CNNs model long-distance relationships effectively, but disallow consideration of short-distance relationships due to downsampling. Our approach considers relationships at multiple length scales, and the additional cost of doing this is low because the downsampled analysis paths need only consider a small fraction of the data in the high-resolution path.

The contributions are (1) a novel fire segmentation model, utilizing global information and multi-scale aggregation, (2) a context-oriented module, which obtains local and global context information to expand the receptive field, and (3) a multi-scale aggregation module, which uses features from low-level and high-level features to capture spatial details better.

2. RELATED WORK

This section discusses related work in semantic segmentation and fire detection.

2.1. Semantic Segmentation

CNNs have achieved state-of-the-art performance in many computer vision fields. For instance, fully convolutional networks are used in image semantic segmentation and perform end-to-end classification of all pixels [11]. However, the receptive field is not large enough for feature representation of all the pixels in the image.

In order to differentiate between objects of different scales and illumination, it is necessary to enhance the discriminative ability of feature representations. One way to improve the performance of FCNs is multi-scale feature aggregation. PSPNet [8] uses spatial pyramid pooling to combine multi-scale information. The Deeplab model uses atrous spatial pyramid pooling (ASPP) with different dilation rates to capture contextual information [12].

In addition, attention mechanisms are applied for pixel-level recognition in order to enhance discriminative features. Zhao et al. introduce a pointwise spatial attention network that encodes relative position information in pixel space [13]. EncNet proposes an encoding layer on top of the network to capture global context [14]. Fu et al. include a self-attention module to model long-range dependencies [15].

Some methods incorporated attention mechanisms to learn feature weights and emphasize important features. OCNet learns feature weights according to object context [16]. Also, CCNet obtains contextual information based on all pixels in the criss-cross path [17]. Furthermore, the Dual Relation-aware Attention Network [18] uses a self-attention mechanism that utilizes different pooling kernels to emphasize certain spatial areas. It also represents associations between channel dimensions to generate channel weights.

AttaNet [19] highlights certain pixels through a strip operation and a cross-level aggregation strategy. BiSeNetV2 [20] incorporates a detail path to preserve the spatial information and a semantic path to process feature maps with a large receptive field. Finally, ConvNeXt [21] constructs a revolutionary convolutional architecture containing inverted bottlenecks, larger kernel sizes, and other architectural differences.

2.2. Fire Detection

Prior image-based fire detection algorithms use the color and features of the fire [22, 23]. The most straight-forward fire detection methods are color-based [24, 25]. They analyze images in the RGB, HSI, or YCbCr color spaces to obtain possible fire regions based on color thresholds [2, 1]. Other past work improves the accuracy of detection by considering additional features as shapes and optical flow [26, 27].

Deep learning algorithms perform automatic extraction of features and can greatly outperform conventional fire detection methods in detection accuracy. For example, Muhammad et al. [5] compared their CNN-based method with other handcrafted fire detection methods and outperforms them in terms of accuracy by 0.88% and false positives by 11.6%. Yin et al. constructed a deep normalization and convolutional neural network attaining smoke detection rates at least 96.4% [28]. Another CNN-based method called the DCNN incorporates a deep dual-channel neural network for smoke detection and has a detection rate of 99.5% on average [29].

Hossain et al. detect forest fires with a neural network using color and multi-color space local binary patterns of both flame and smoke signatures [6]. Saponara et al. implemented a fully real-time CNN for fire detection using the YOLOv2 framework on a NVIDIA Jetson Nano [30]. Muhammad et al. described a framework based on the AlexNet architecture for fire detection and obtain an accuracy of 94.39% and a false positive rate of 9.07% [5, 31].

3. METHODOLOGY

This section provides an overview of the proposed model and describes each of its key components in detail.

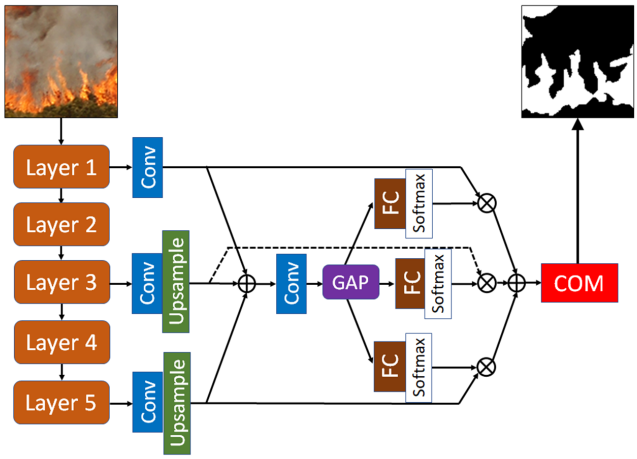

Fig. 2. Our Context-Oriented Multi-Scale Network for fire segmentation uses a Multi-Scale Aggregation (MSA) layer and Context-Oriented Module (COM). MSA considers relationships at multiple layers in the network and performs adaptive feature refinement. COM is explained in the next figure.

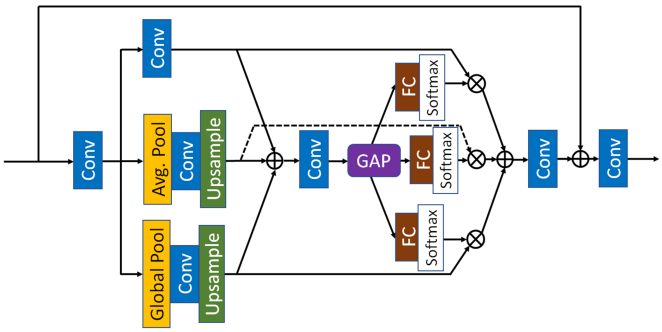

Fig. 3. We propose a Context-Oriented Module (COM) that extracts discriminant feature representations by building associations among features with average and global pooling.

3.1. Overview

Figure 2 shows the architecture of the proposed model. Initially, we use a five-layer ResNet-50 backbone to extract its features, denoted as $f_i(i = 1, 2, ..., 5)$. The backbone maps the input scene to feature representations, but it cannot capture both the local and global information of the scene well.

In order to exploit the multi-scale structure of the flames and deal with different flame sizes, we incorporate a multi-scale aggregation module. We perform adaptive feature refinement at multiple network levels in order to consider relationships at multiple length scales. The implications of this involve enhancing the intra-class and inter-class recognition.

Since contextual information can be used to improve the performance of CNNs, we expand the size of the receptive field by incorporating global contextual information via our Context-Oriented Module (COM). In scenes with diverse backgrounds and varied shapes, the COM can adaptively aggregate global contextual information, which refers to the relationships of all pixels in the feature map, improving feature representation for fire segmentation.

3.2. Multi-Scale Aggregation

We incorporate multi-scale aggregation (MSA) to capture different scales of flames more accurately. Low-level and high-level features are complementary, where low-level features are rich in spatial details but lack semantic information, and vice-versa for high-level features. To bridge the gap between high-level and low-level features, MSA adaptively combines both features with a novel design.

We incorporate a gating mechanism to adjust the level of information from different layers. It adaptively passes important spatial and semantic information at multiple layers in order to improve accuracy. MSA improves accuracy by better capturing spatial details from low-level features. This has not been done before in fire segmentation.

The structure of this module is shown in Figure 2. From the backbone layer, each of f_1, f_3, and f_5 form separate branches, go through a conv layer, and are each upsampled to the dimension of f_1. The outputs are f_1', f_3', and f_5', respectively, containing features at different scales.

We select multiple layers where each layer is downsampled by a different amount. Earlier layers have more spatial information and later layers have more semantic information. Next, we combine all outputs using an element-wise sum as $F = f_1' + f_3' + f_5'$. Afterward, we apply global average pooling (GAP) across the spatial dimension of $F \in R^{W \times H \times C}$ to compute channel-wise statistics $s \in R^{1 \times 1 \times C}$.

Later, we feed s into three independent fully connected layers, FC_1, FC_3, and FC_5, and apply softmax to the outputs to obtain w_3, w_4, and w_5. We then perform channel-wise multiplication for $f_1' \cdot w_1$, $f_3' \cdot w_3$, and $f_5' \cdot w_5$ and then fuse them via element-wise summation. This is described as follows:

$$
\begin{aligned}
s &= \text{GlobalPooling}(F), \\
w_1, w_3, w_5 &= \text{softmax}([FC_1(s), FC_3(s), FC_5(s)]), \text{ and} \\
V &= C(F_1 \cdot w_1 + F_3 \cdot w_3 + F_5 \cdot w_5).
\end{aligned}
$$

3.3. Context-Oriented Module

We adopt the Context-Oriented Module (COM) to expand the receptive field to capture richer features. The network initially obtains feature representations by stacking conv layers, but it cannot capture both local and global information simultaneously. COM further improves the network's ability to extract semantic information by using additional length scales. Past work did not consider further expanding the receptive field to additional length scales until ours via COM.

Past work has shown that global context information improves various computer vision tasks [8, 32]. We obtain more discriminative feature representations for better scene understanding by building associations with features through global context. Feature aggregation allows the network to focus on more informative contextual features.

The detailed structure of the Context-Oriented Module is shown in Figure 3. The output of the MSA layer V is fed into an input conv layer to output V'. Next, V' is fed to three branches: one branch contains a conv layer, another branch contains an average pooling layer, followed by a conv layer and upsampling block, and the other branch contains a global pooling layer, followed by a conv layer and upsampling block.

The outputs are F_c, F_l, and F_g, representing local and global features, respectively. All three branches contain features with different receptive fields. Then, we combine both local and global features using an element-wise sum as: $F = F_c + F_l + F_g$. This is described as follows:

$$\begin{aligned} V' &= C(V), \\ F_c &= C(V'), \\ F_l &= U(C(P(V'))), \\ F_g &= U(C(G(V'))), \text{ and} \\ F &= F_c + F_l + F_g, \end{aligned}$$

where C, P, G, and U represent convolution, average pooling, global pooling, and upsampling layers, respectively. We then apply global average pooling (GAP) across the spatial dimension of $F \in R^{W \times H \times C}$ to compute channel-wise statistics $s \in R^{1 \times 1 \times C}$.

Later, we feed s into three independent fully connected layers, FC_c, FC_l, and FC_g, and apply softmax to the outputs to obtain w_c, w_l, and w_g. We then perform channel-wise multiplication for $F_c \cdot w_c$, $F_l \cdot w_l$, and $F_g \cdot w_g$ and fuse them via element-wise summation. The output F' selectively incorporates local and global attention based on their content and characteristics. These operations are described as follows, which is similar to those of the MSA:

$$\begin{aligned} s &= \text{GlobalPooling}(F), \\ w_c, w_l, w_g &= \text{softmax}([FC_c(s), FC_l(s), FC_g(s)]), \\ F' &= C(F_c \cdot w_c + F_l \cdot w_l + F_g \cdot w_g), \text{ and} \\ F' &= C(F' + V). \end{aligned}$$

The two modules serve different purposes. COM further expands the receptive field from the output of the backbone network to additional length scales. MSA uses features from low-level and high-level features to capture spatial details better. Low-level features contain information from a lower receptive field, so MSA does not expand the receptive field but improves the spatial reasoning through local details.

Table 1. Results of fire segmentation with other methods.

Methods	IoU	Dice
U-Net (2015)	0.705	0.792
PSP-Net (2017)	0.653	0.757
DeepLabv3 (2017)	0.755	0.834
CPD (2019)	0.681	0.779
RAS (2020)	0.686	0.780
DRAN (2020)	0.751	0.829
AttaNet (2021)	0.747	0.827
BiSeNetV2 (2021)	**0.781**	**0.852**
ConvNeXt (2022)	0.632	0.741
Ours w/o MSR	0.675	0.771
Ours w/o COM	0.789	0.858
Ours	**0.808**	**0.873**

4. EXPERIMENTAL RESULTS

We first introduce the dataset and the experimental protocol. Next, we evaluate our proposed method on images containing wildfires and compare it with other methods.

4.1. Dataset and Implementation Details

We use a benchmark dataset of wildfires, consisting of 595 images of varying size [33]. The dataset includes annotation of all fire pixels and each is resized from a larger size down to 512×512. We then augment the dataset by applying random cropping five times for each image to size 224×224 to end up with 2,975 images in total.

The training dataset contains 2,000 images, while the testing dataset contains 975 images. During the training phase, we set the learning rate to 2e-4, the batch size to 2, and the number of epochs to 40 for model training. Also, we set the momentum parameter to 0.9 and use Adam to optimize the parameters during training. The binary cross-entropy loss ($\mathcal{L}_{BCE}$) is used to calculate the loss of each pixel in the predicted segmentation map compared to the ground-truth map. All experiments are run on a machine with an NVIDIA GeForce 940MX GPU.

4.2. Model Comparison

We compare our model with past fire segmentation methods, shown in Table 1. For a fair comparison, we calculate each method's accuracy with the same parameters.

Experiments on the benchmark dataset show that our model improves accuracy by 2.7% compared to BiSeNetV2. We report all segmentation results in terms of mean Intersection over Union (mIoU) and Dice error, which are widely used to evaluate the overall performance of semantic segmentation algorithms. The mIoU metric reflects the degree of the overlap between the predicted segmentation and the

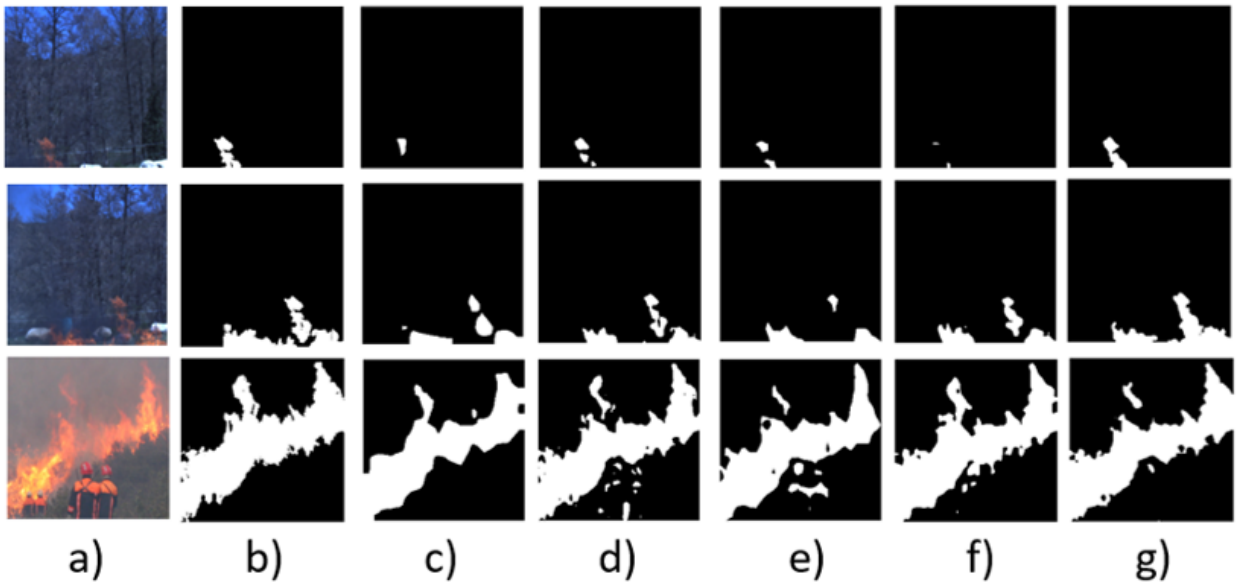

Fig. 4. Visual results of our method and four previous segmentation methods. Our model is effective at segmenting flames of various sizes and distinguishing flames from complex backgrounds. a) Input image, b) Ground-truth image, c) DeepLabv3, d) DRAN, e) AttaNet, f) BiSeNetV2, g) Proposed method.

corresponding ground truth versus their union. In particular, we tested the performance of the Vision Transformer and the performance was lower than the state-of-the-art methods in Table 1.

4.2.1. Ablation Analysis

We also evaluate the effectiveness of each module. First, we remove the COM and only keep the MSA module in order to examine the effectiveness of the COM. From Table 1, we observe that our model with the COM outperforms our model by 1.9% without the COM. Hence, the COM improves accuracy by expanding the receptive field in order to consider relationships of longer length scales in the feature map.

We then remove the MSA module and retain the COM. From Table 1, we observe that our model with the MSA module outperforms our model by 13.3% without it.

4.2.2. Visualization of Results

Figure 4 shows the qualitative comparison of our proposed method and past fire segmentation methods. We select some representative examples from the dataset. It can be seen that our method can accurately segmenting flames in challenging scenes and performs significantly better than other models.

In the first row, previous methods were not able to discern the small flame in the image. Some methods in the second row confused the background with the fire. Furthermore, some existing models confused the flames with the background which has similar appearance with the fire. On the other hand, our method can accurately infer the flame region in each case.

This is mainly because Multi-Scale Aggregation (MSA) can handle flames with different scales via adaptive feature refinement at multiple levels of the CNN. Also, the Context-Oriented Module can help discriminate the flames from the background in complex scenes.

5. CONCLUSION

In this paper, we have presented a Context-Oriented Multi-Scale Network for fire segmentation. This network adaptively integrates local and global context and uses multi-scale aggregation in order to give more precise segmentation results. Our proposed method improves IoU accuracy by 2.7% compared to past work. For future work, we plan to decrease the computational complexity and enhance the robustness of the model.

6. ACKNOWLEDGEMENTS

This work was supported, in part, by the National Science Foundation through grant CNS-2008151.

7. REFERENCES

[1] Chao-Ho Chen, Ping-Hsueh Wu, and Yung-Chuen Chiou, "An early fire-detection method based on image processing," *Proceedings of the International Conference on Image Processing*, vol. 3, pp. 1707–1710, 2004.

[2] T. Çelik, Hüseyin Özkaramanli, and H. Demirel, "Fire and smoke detection without sensors: Image processing based approach," *Proceedings of the European Signal Processing Conference*, pp. 1794–1798, 2007.

[3] Xiaopeng Zhang, Hongkai Xiong, Wen gang Zhou, Weiyao Lin, and Qi Tian, "Picking deep filter responses for fine-grained image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1134–1142, 2016.

[4] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[5] Khan Muhammad, Jamil Ahmad, and Sung Wook Baik, "Early fire detection using convolutional neural networks during surveillance for effective disaster management," *Neurocomputing*, vol. 288, pp. 30–42, 2018.

[6] FM Anim Hossain, Youmin M Zhang, and Masuda Akter Tonima, "Forest fire flame and smoke detection from uav-captured images using fire-specific color features and multi-color space local binary pattern," *Journal of Unmanned Vehicle Systems*, vol. 8, no. 4, pp. 285–309, 2020.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," *ArXiv*, 2015.

[8] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6230–6239, 2017.

[9] Han-Soo Choi, Myeongho Jeon, Kyungmin Song, and Myung joo Kang, "Semantic fire segmentation model based on convolutional neural network for outdoor image," *Fire Technology*, pp. 1–15, 2021.

[10] Kyungmin Song, Han-Soo Choi, and Myung joo Kang, "Squeezed fire binary segmentation model using convolutional neural network for outdoor images on embedded device," *Machine Vision and Applications*, vol. 32, 2021.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 834–848, 2018.

[13] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *Proceedings of the European Conference on Computer Vision*, 2018.

[14] Hang Zhang, Kristin J. Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal, "Context encoding for semantic segmentation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7151–7160, 2018.

[15] J. Fu, J. Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu, "Dual attention network for scene segmentation," *2019 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3141–3149, 2019.

[16] Yuhui Yuan and Jingdong Wang, "Ocnet: Object context network for scene parsing," *ArXiv*, vol. abs/1809.00916, 2018.

[17] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, Humphrey Shi, and Wenyu Liu, "Ccnet: Criss-cross attention for semantic segmentation," *2019 IEEE International Conference on Computer Vision*, pp. 603–612, 2019.

[18] J. Fu, Jing Liu, Jie Jiang, Yong Li, Yongjun Bao, and Hanqing Lu, "Scene segmentation with dual relation-aware attention network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, pp. 2547–2560, 2020.

[19] Qi Song, Kangfu Mei, and Rui Huang, "Attanet: Attention-augmented network for fast and accurate scene parsing," in *AAAI Conference on Artificial Intelligence*, 2021.

[20] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, pp. 3051–3068, 2021.

[21] Zhuang Liu, Hanzi Mao, Chaozheng Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," *2022 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11966–11976, 2022.

[22] T Wittkopp, C Hecker, and D Opitz, "Cargo fire monitoring system (CFMS) for the visualisation of fire events

in aircraft cargo holds.," in *International Conference on Automatic Fire Detection "AUBE '01", 12th. Proceedings. National Institute of Standards and Technology, Gaithersburg, MD*, 2001-03-25 2001.

[23] C.L. Lai, Yang Jie-Ci, and Y.H. Chen, "A real time video processing based surveillance system for early fire and flood detection," in *2007 IEEE Instrumentation and Measurement Technology Conference IMTC 2007*, 06 2007, pp. 1 – 6.

[24] Dongil Han and Byoungmoo Lee, "Development of early tunnel fire detection algorithm using the image processing," in *Advances in Visual Computing*, George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Paolo Remagnino, Ara Nefian, Gopi Meenakshisundaram, Valerio Pascucci, Jiri Zara, Jose Molineros, Holger Theisel, and Tom Malzbender, Eds., Berlin, Heidelberg, 2006, pp. 39–48, Springer Berlin Heidelberg.

[25] Chao-Ching Ho, "Machine vision-based real-time early flame and smoke detection," *Measurement Science and Technology*, vol. 20, pp. 045502, 03 2009.

[26] X. Han, J. Jin, Mingjie Wang, Wei Jiang, Lei Gao, and L. Xiao, "Video fire detection based on gaussian mixture model and multi-color features," *Signal, Image and Video Processing*, vol. 11, pp. 1419–1425, 2017.

[27] Konstantin Trambitckii, K. Anding, V. Musalimov, and G. Linß, "Colour based fire detection method with temporal intensity variation filtration," in *2014 Joint IMEKO TC1-TC7-TC13 Symposium: Measurement Science Behind Safety and Security*, 2015.

[28] Zhijian Yin, Boyang Wan, Feiniu Yuan, Xue Xia, and Jinting Shi, "A deep normalization and convolutional neural network for image smoke detection," *IEEE Access*, vol. 5, pp. 18429–18438, 2017.

[29] Ke Gu, Zhifang Xia, Junfei Qiao, and Weisi Lin, "Deep dual-channel neural network for image-based smoke detection," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 311–323, 2019.

[30] Sergio Saponara, Abdussalam Elhanashi, and Alessio Gagliardi, "Real-time video fire/smoke detection based on CNN in antifire surveillance systems," *Journal of Real-Time Image Processing*, vol. 18, no. 3, pp. 889–900, 2021.

[31] Khan Muhammad, Jamil Ahmad, Zhihan Lv, Paolo Bellavista, Po Yang, and Sung Wook Baik, "Efficient deep CNN-based fire detection and localization in video surveillance applications," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 7, pp. 1419–1434, 2018.

[32] Hang Zhang, Kristin J. Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal, "Context encoding for semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7151–7160, 2018.

[33] Tom Toulouse, Lucile Rossi, Antoine Campana, Turgay Çelik, and Moulay A. Akhloufi, "Computer vision for wildfire research: An evolving image dataset for processing and analysis," *Fire Safety Journal*, vol. 92, pp. 188–194, 2017.