

Enhancement of early fire detection using improved YOLOv8-based visual smoke detection model

Roman Kvyetnyy^{*a}, Andrzej Smolarz^b, Roman Maslii^a, Vladyslav Kabachii^a, Ainur Kozbakova^{c,d},
Volodymyr Harmash^a, Nataliia Savina^e, Iryna Shvarts^a, Dinara Ussipbekova^f

^aVinnitsia National Technical University, 95 Khmelnytske Shose, Vinnitsia, Ukraine; ^bLublin University of Technology, Nadbystrzycka 38a, 20-618 Lublin, Poland; ^cInstitute of Information and Computational Technologies CS MES RK, Almaty, Kazakhstan; ^dAlmaty University of Energy and Communications named after G. Daukeyev, Almaty, Kazakhstan; ^eNational University of Water and Environmental Engineering, Rivne, Ukraine; ^fKazakh National University named after S.D. Asfendiyarova, Almaty, Kazakhstan

ABSTRACT

This paper proposes modifications to the YOLOv8 architecture for visual smoke detection in early fire detection systems. To reduce the complexity of the architecture, it is proposed to use the VoVGSCSP block, to improve the detection quality – the efficient multi-scale attention (EMA) block. The proposed changes allowed to reduce the number of neural network parameters by 27% and the computational complexity (GFLOPS) by 12% compared to the YOLOv8n model. The proposed model retained high detection accuracy, compared to the base model, the decrease in detection accuracy by the quality assessment metrics mAP@0.5 and mAP@0.5:0.95 was about 1%. To train and evaluate the model, an own dataset of more than 5000 images was created based on the open datasets D-Fire and WSDY. The obtained results demonstrate the suitability of the model for use on edge devices as part of video surveillance systems for early fire detection.

Keywords: optical smoke detection, early fire detection, YOLOv8, edge device, VoVGSCSP, EMA

1. INTRODUCTION

Early detection of fires is critical for ensuring safety and minimizing damage. Fires spread rapidly, especially in areas with high densities of combustible materials, such as residential zones, airports, and forests, making rapid response challenging. Traditional sensor systems (e.g., smoke, temperature, and particle sensors) are effective in limited areas but require direct contact with smoke or fire, reducing response speed^{1,2}. In contrast, modern video-based fire detection systems using computer vision technologies enable early smoke detection, offering rapid response, broad coverage, and high robustness against environmental factors. Current research on smoke and fire detection from video images is divided into two categories: traditional classifier-based methods and deep learning approaches. The former rely on feature extraction algorithms (e.g., SIFT, HOG, LBP) to analyze brightness, color, texture, and contours, followed by classification using SVM, boosting, or Bayesian networks^{3,4}. However, these methods depend on manual feature engineering, leading to slower processing and false positives due to obstacles or occlusions. Deep neural network-based methods, conversely, automatically learn informative features, achieving better performance in complex environments.

Detection accuracy was improved by training Convolutional Neural Networks (CNNs) on full images and small patch classifiers⁵. Fire detection performance was enhanced using VGG16 and ResNet50 networks⁶. Applying CNNs to video-based fire detection outperformed traditional methods⁷. Recurrent Neural Networks (RNNs) were explored for fire detection, leveraging their ability to capture feature relationships across video frames for long-term memory^{8,9}. Long short-term memory (LSTM) networks, a variant of RNNs, solve the problem of vanishing gradients inherent in traditional RNN models. When applied to fire detection, LSTMs are able to simultaneously extract spatial and temporal features of flames and smoke. This dual extraction provides high accuracy and completeness of recognition while meeting real-time requirements. However, LSTMs are accompanied by problems due to the numerous fully connected layers, large time intervals, deep network architecture, and high computational cost of a large number of parameters, which makes them difficult to train^{8,9}.

*e-mail: rkvetny@vntu.edu.ua; <https://vntu.edu.ua/en>

Modern single-stage object detection architectures, particularly the YOLO model series, achieve high accuracy and real-time performance. The YOLO-SMOKE model, incorporating an efficient channel attention module into YOLOv3 and modifying the loss function, enhanced accuracy and robustness¹⁰. A lightweight fire detection algorithm based on YOLOv5 utilized Ghost modules, the Coordinate Attention (CA) mechanism, and weighted feature fusion from the Path Aggregation Network to improve detection speed¹¹. The SimAM attention mechanism was integrated into YOLOv7 to enhance smoke and fire feature detection while suppressing the influence of the background, improving generalization without increasing the number of parameters¹². YOLOv8 was enhanced by modifying the C2f network structure to incorporate Deformable Convolution and SCConv, together with CA modules in the detection head, optimizing feature utilization, enabling adaptive adjustment of the receptive field, and boosting accuracy.¹³ Lightweight models were explored using GhostConv, Depthwise Separable Convolution, and Squeeze-and-Excitation (SENet) networks¹⁴. A Channel Prior Dilatation Attention (CPDA) module, combined with the GSConv module in the neck part and a knowledge distillation strategy, enhanced generalization and reduced false detections¹⁵.

This work aims to optimize the YOLOv8 architecture by minimizing parameters and computational complexity while preserving smoke detection accuracy for deployment on resource-constrained edge devices.

2. MODIFICATION OF YOLOV8 ARCHITECTURE

YOLOv8, introduced by Ultralytics in January 2023, is a single-stage object detection system based on YOLOv5, supporting classification, detection, and tracking. It uses a Backbone for feature extraction, a Neck for feature aggregation, and a Head for prediction. YOLOv8 outperforms its predecessors in accuracy on the COCO dataset and is available in variants (n, s, m, l, x), with YOLOv8n being the lightest, suited for resource-constrained devices. Main improvements include replacing the C3 block with a more efficient C2f, simplifying the Neck by removing convolutions in upsampling, adopting a Decoupled Head for separate localization and classification, and using an Anchor-Free approach to reduce computational load. Its main modules include Conv (convolution, normalization, activation), C2f, Spatial Pyramid Pooling Fast (SPPF), Upsample, Concat, and Detect.

Standard convolutions (Conv) effectively mix cross-channel and spatial information for feature extraction, but their frequent use significantly increases the number of parameters and computational complexity measured in FLOPs. To optimize these characteristics in lightweight models such as MobileNet¹⁶ and ShuffleNet¹⁷, depthwise convolutions (DWConv) are used, which reduce the number of parameters and computations by processing each channel independently, but limit the cross-channel interaction, which can reduce accuracy due to information loss. To overcome this drawback the GSConv module was introduced, which combines a standard 1x1 convolution to reduce the input channels c_1 to half of the given number of output channels c_2 ($c_2/2$) with a 5x5 deep convolution that produces the same number of channels ($c_2/2$) while preserving the spatial dimensions through padding ($p=2$)¹⁸. The features from both branches are combined along the channel dimension, and then a shuffling operation similar to that used in ShuffleNet¹⁷ evenly distributes the inter-channel information, improving its integration and feature salience. Thus, GSConv, whose structure is shown in Figure 1a, provides a balance between computational efficiency and accuracy, optimizing the performance of lightweight models without significantly increasing the cost.

To increase the speed of logical inference while preserving accuracy, the GS-Bottleneck module was introduced, featuring a main branch with two consecutive GSConv blocks for feature processing and a parallel shortcut branch with standard convolution (Conv) that adds the input tensor to the output when a shortcut connection is present¹⁸.

To improve the efficiency of the network, it is proposed to modify the shortcut branch by replacing the standard convolution with a depthwise convolution, which reduces the computational complexity. After combining the results of GSConv with DWConv, it is proposed to add the attention module EMA (Efficient Multi-Scale Attention)¹⁹. EMA uses cross-spatial learning and multi-scale parallel subnetworks to establish short- and long-term dependencies. The module groups channel dimensions into feature subsets, preserving information through cross-dimensional interactions, which ensures a uniform distribution of spatial semantic features. Parallel 3x3 and 1x1 convolutions integrate context at different scales, increasing the accuracy of pixel attention and reducing the number of parameters. EMA preserves accurate spatial information by modeling long-term dependencies, which is especially useful for smoke detection. By being able to recognize diffuse patterns through local textures (short-term dependencies) and global connections in the scene (long-term dependencies), EMA improves detection accuracy in real-world conditions where smoke is heterogeneous and variable. The integration of EMA into the GS-Bottleneck module enhances the aggregation of contextual information at the feature combination stage, contributing to better detail processing and object recognition.

accuracy. The updated block, named GS-BottleneckC-EMA (where "C" stands for "cheap" due to the use of DWConv), has the structure shown in Figure 1b.

The VoVGSCSP block is presented, an improved version of the CSP blocks that uses GSConv in a compact architecture with partial feature path separation for an optimal balance of accuracy and speed¹⁸. It is proposed to replace GS-Bottleneck in VoVGSCSP with GS-BottleneckC-EMA, which results in the VoVGSCSPC-EMA block shown in Figure 1c. This modification combines the advantages of DWConv (reduction of computation) and the EMA attention mechanism (improved context processing), providing more efficient work with features. For example, in the smoke detection task, VoVGSCSPC-EMA can improve the recognition of diffuse patterns due to better aggregation of contextual information^{22–25}.

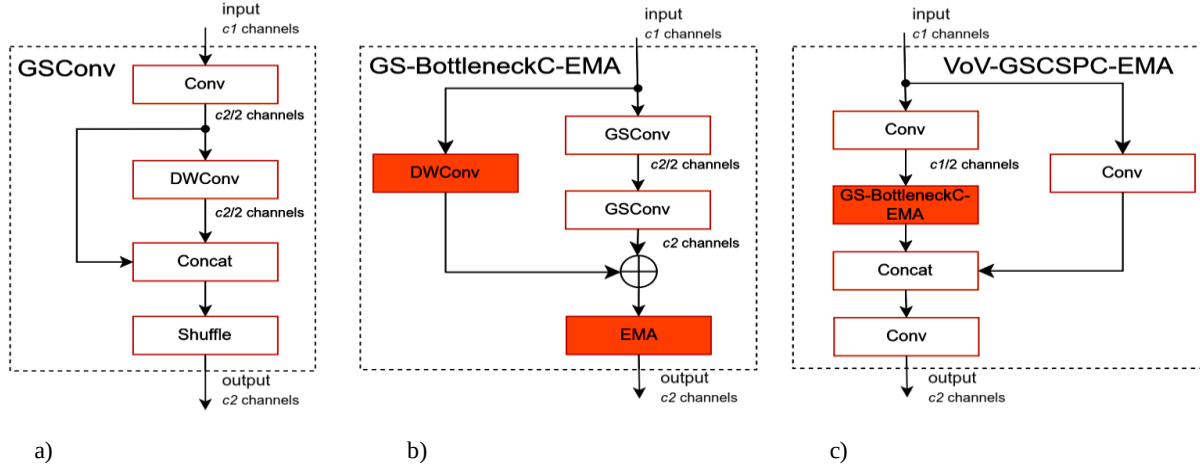


Figure 1. The structure of the GSConv (a), GS-BottleneckC-EMA (b), and VoVGSCSPC-EMA (c) blocks

Figure 2 shows the modifications of the blocks of the original YOLOv8 architecture. In particular, in the Neck part, the standard C2f blocks are proposed to be replaced with more computationally efficient VoV-GSCSPC-EMA modules, with the exception of one C2f block, which is responsible for the detection of small objects – it is retained to maintain high detection accuracy. In the Backbone, one C2f block is replaced with a VoVGSCSPC-EMA module, significantly reducing parameters while maintaining detection quality²⁶.

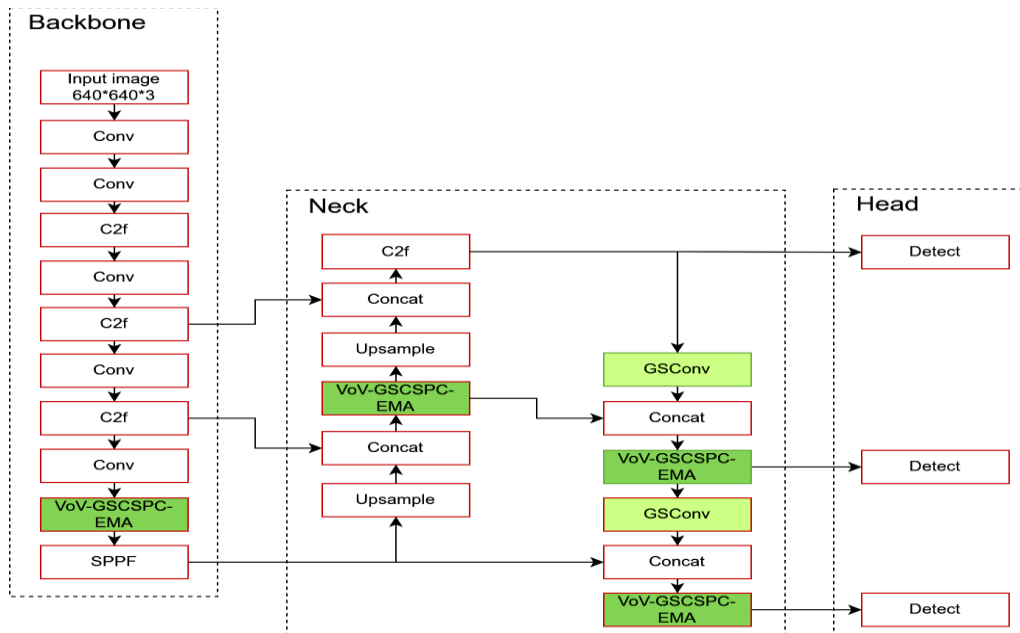


Figure 2. Proposed YOLOv8 modification architecture

3. RESULTS OF EXPERIMENTAL RESEARCH

Dataset creation. The research used two open datasets, D-Fire²⁰ and WSDY²¹, which contain images of smoke in natural environments. D-Fire is a dataset designed to detect smoke and flames in forest and open areas under different weather conditions and shooting angles. WSDY (Wildfire Smoke Detection Dataset) consists of annotated images obtained from video surveillance cameras, with a focus on difficult detection conditions such as fog, clouds and smoke. Both datasets are characterized by a wide variability of scenes and high-quality labeling, which makes them suitable for effective training of early fire detection models.

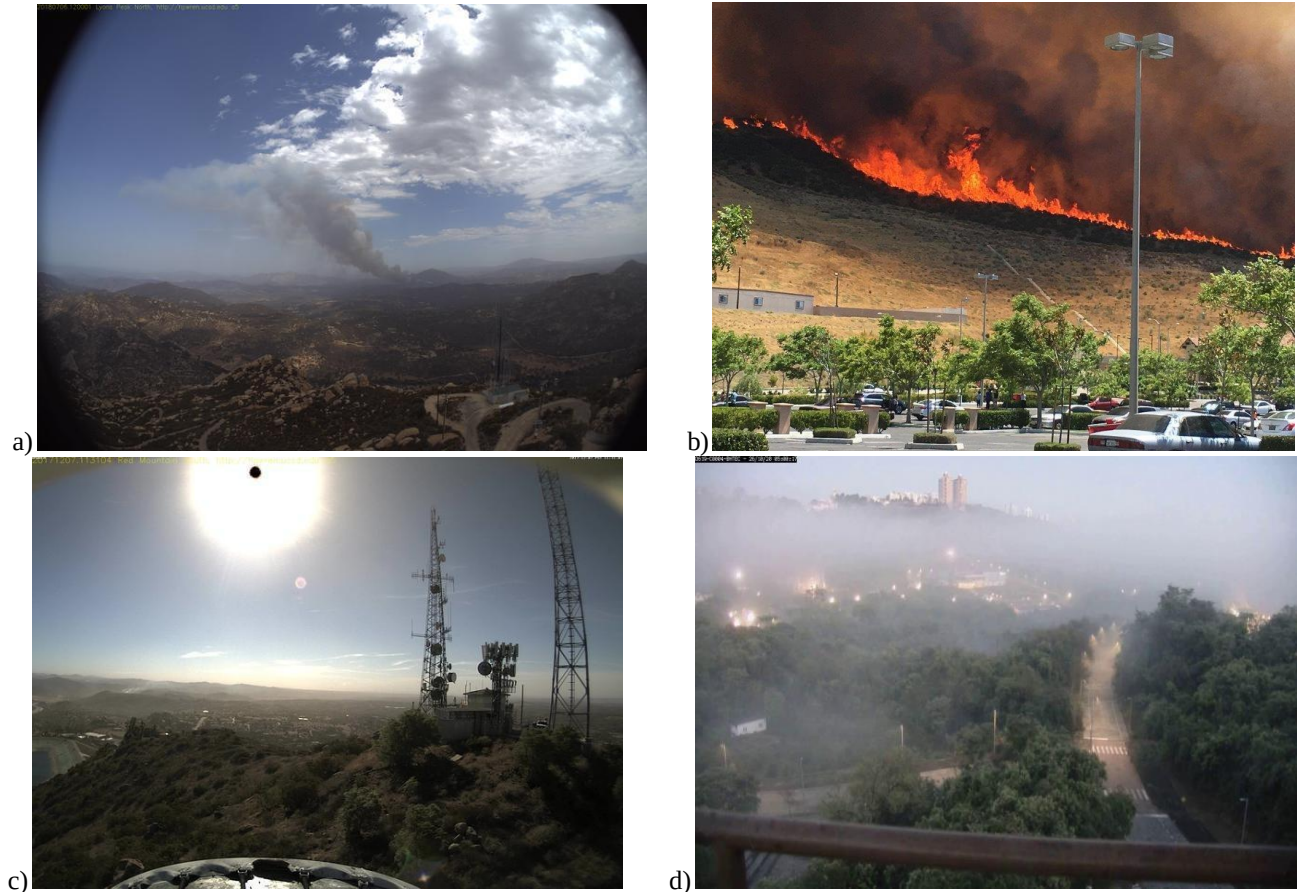


Fig. 3. Representative smoke images selected from the dataset: (a) Smoke on a background of clouds, (b) Smoke with fire, (c) Sun glare with distant smoke, (d) Background image without smoke and with fog

To train a smoke detection model, images with resolution above 480×480 pixels were selected from the D-Fire dataset to enhance data quality and reduce noise. The training subset includes:

- 3409 images labeled as “smoke” without other objects;
- 600 randomly selected images with “smoke” and “fire” labels, using only “smoke” one;
- 600 background images with natural interferences (clouds, fog, sun glare) to improve model robustness.

In total, 4609 D-Fire images were used. Combined with 663 WSDY images, a final dataset of 5272 images was formed, split into training and validation subsets in a 4:1 ratio for model training and evaluation.

Examples of images from the generated dataset are presented in Figure 3.

Evaluation indicators. To evaluate the performance of the smoke detection model in this work, the mean Average Precision (mAP) metric was used. The computational complexity of the model was evaluated by indicators such as the number of parameters and the amount of floating-point operations (FLOPs). Since the study considered only one class (smoke), the mAP value is equivalent to the Average Precision (AP) value for that class.

The AP metric is defined as the area under the Precision-Recall curve and is calculated as follows

$$AP = \int_0^1 Precision(Recall) d(Recall). \quad (1)$$

Precision reflects the proportion of correct detections among all positive predictions of the model, and Recall reflects the proportion of correct detections among all real objects of the “smoke” class. The equations for calculating Precision and Recall are presented in formulas (2) and (3), respectively:

$$Precision = \frac{TP}{TP+FP}, \quad (2)$$

$$Recall = \frac{TP}{TP+FN}, \quad (3)$$

where TP denotes the number of correctly detected positive samples, FN denotes the number of incorrectly detected negative samples, and FP indicates the number of incorrectly detected positive samples.

Two mAP metrics were used in the study: mAP@0.5, which corresponds to the mAP value at an IoU (Intersection over Union) threshold of 0.5, and mAP@0.5:0.95, which is the average mAP value calculated for IoU thresholds from 0.5 to 0.95 in increments of 0.05.

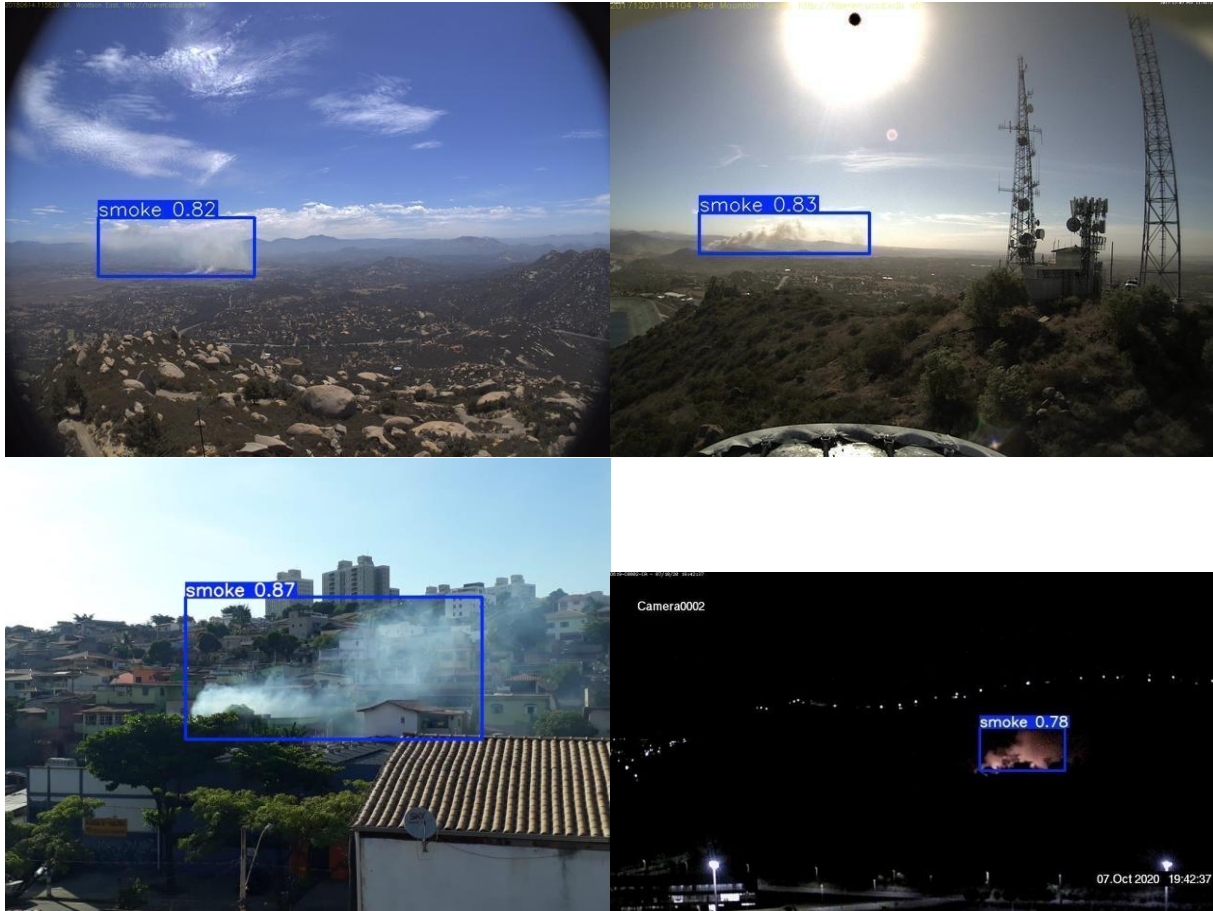


Figure 4. Smoke detection examples

Experimental Platform. The experimental environment for this paper is as follows: 16G RAM, the CPU is an Intel(R) i5-8400, and the GPU is a GeForce GTX 1080 with 8G memory running on an Ubuntu 20.04 operating system. The framework used is Python 3.11, PyTorch 2.5.1, CUDA 11.0.

Training Parameters. Epochs number – 100, batch size – 32, image size – 640*640, initial learning rate – 0.02, optimization algorithm – AdamW, momentum – 0.9.

Ablation experiments. To evaluate the modified YOLOv8 architecture's effectiveness and the impact of the proposed modules on the generated dataset, three sets of ablation experiments were conducted, with results presented in Table 1. The paper proposes an improvement of the YOLOv8 architecture, but experiments were conducted with the YOLOv8n model for the possibility of application on edge devices. Method A replaces C2f blocks with VoVGSCSPC blocks (according to Figure 2), Method B replaces Conv blocks with GSConv in the Neck, and Method C adds the EMA module to VoVGSCSPC, forming VoVGSCSPC-EMA. The “+” sign indicates the presence of the block in the studied architecture.

Table 1. The results of ablation experiments

Method	EMA	GSConv	VoVGSCSPC	mAP@50:95	mAP@50	Parameters	GFLOPS
YOLOv8n	-	-	-	46.3	83.9	3.0	8.1
A	-	-	+	45.1	82.1	2.2	7.2
B	-	+	+	44.9	82.3	2.2	7.1
C	+	+	+	45.7	83.4	2.2	7.1

Table 1 shows that Method A reduces YOLOv8n parameters from 3.0M to 2.2M and GFLOPs from 8.1 to 7.2, with slightly lower mAP@50:95 comparing to base YOLOv8n model (from 46.3 to 45.1) and mAP@50 (from 83.9 to 82.1). Method B further reduces GFLOPs to 7.1 while maintaining mAP@50-95 at 44.9. Method C, incorporating VoVGSCSPC-EMA, maintains parameters and GFLOPs at Method B's levels but enhances feature aggregation across scales, improving mAP@50-95 from 44.9 to 45.7 and mAP@50 from 82.3 to 83.4. Thus, combining GSConv in the Neck and VoVGSCSPC-EMA (Method C) reduces parameters from 3.0M to 2.2M and GFLOPs from 8.1 to 7.1, with a minor accuracy reduction (mAP@50-95 from 46.3 to 45.7, mAP@50 from 83.9 to 83.4) suitable for edge devices. Examples of smoke detection are shown in Figure 4, it shows that the model correctly works in different scenarios, particularly, it may detect smoke against a background of clouds, in the case of sun glare, in urban areas, and at night time.

4. CONCLUSIONS

The proposed modifications to the YOLOv8 architecture, utilizing the VoVGSCSP and EMA blocks, successfully reduced the model's complexity by 27% in parameters and 12% in computational load (GFLOPS), while maintaining high detection accuracy with only a 1% decrease in mAP@0.5 and mAP@0.5:0.95 metrics. The model, trained on a custom dataset of over 5000 images, proves suitable for early fire detection on edge devices in video surveillance systems, effectively balancing efficiency and performance. The integration of the EMA module enhanced the model's ability to capture multi-scale contextual information, contributing to robust smoke detection in challenging environments. Additionally, the use of VoVGSCSP blocks demonstrates a promising approach for developing lightweight models without compromising detection quality. Further investigations will focus on optimizing the model for lower-resource edge devices by exploring additional lightweight architectural blocks and pruning techniques. These efforts aim to enhance real-time performance and scalability, making the model viable for broader applications in autonomous fire detection systems.

REFERENCES

- [1] Zhang, F., Zhao, P., Xu S., Wu, Y., Yang, X., and Zhang, Y., “Integrating multiple factors to optimize watchtower deployment for wildfire detection,” *Sci. Total Environ.*, vol. 737, p. 139561 (2020); <https://doi.org/10.1016/j.scitotenv.2020.139561>.
- [2] Karthi, M et al., “Forest fire detection: a comparative analysis of deep learning algorithms,” in *Proc. Int. Conf. Artif. Intell. Knowl. Discovery Concurrent Eng. (ICECONF)*, pp. 1–6 (2023); <https://doi.org/10.1109/ICECONF57129.2023.10084329>.
- [3] Kvyetnyy, R. et al., “Object detection in images with low light condition,” *Proc. SPIE*, vol. 10445, p. 104450W (2017); <https://doi.org/10.1117/12.2281001>.
- [4] Xiong, D. and Yan, L., “Early smoke detection of forest fires based on SVM image segmentation,” *J. For. Sci.*, vol. 65, pp. 150–159 (2019); <https://doi.org/10.17221/82/2018-JFS>.
- [5] Zhang Q., J. Xu, L. Xu, and H. Guo, “Deep convolutional neural networks for forest fire detection,” in *Proc. Int. Forum Manage., Educ. Inf. Technol. Appl.*, Guangzhou, China, pp. 568–575 (2016); <https://doi.org/10.2991/ifmeita-16.2016.105>.

- [6] Sharma, J., Granmo, O.C., Goodwin, M., and Fidge, J.T., "Deep convolutional neural networks for fire detection in images," in Proc. 18th Int. Conf. Eng. Appl. Neural Netw. (EANN), Athens, Greece, pp. 183–193 (2017); https://doi.org/10.1007/978-3-319-65172-9_16.
- [7] Frizzi, S. et al., "Convolutional neural network for video fire and smoke detection," in Proc. IECON 2016 – 42nd Annu. Conf. IEEE Ind. Electron. Soc., Florence, Italy, pp. 877–882 (2016); <https://doi.org/10.1109/IECON.2016.7793196>.
- [8] Cao, Y., Yang, F., Tang, Q., and Lu, X., "An Attention Enhanced Bidirectional LSTM for Early Forest Fire Smoke Recognition," IEEE Access, vol. 7, pp. 154732–154742 (2019); <https://doi.org/10.1109/ACCESS.2019.2946712>.
- [9] Nguyen, M.D, Vu, H.N., Pham, D.C., Choi, B., and Ro, S., "Multistage Real-Time Fire Detection Using Convolutional Neural Networks and Long Short-Term Memory Networks," IEEE Access, vol. 9, pp. 146667–146679 (2021); <https://doi.org/10.1109/ACCESS.2021.3122346>.
- [10] Cai W., Wang C., Huang H., and Wang T., "A Real-Time Smoke Detection Model Based on YOLO-SMOKE Algorithm," in 2020 Cross Strait Radio Sci. Wireless Technol. Conf. (CSRSWTC), Fuzhou, pp. 1–3 (2020); <https://doi.org/10.1109/CSRSWTC50769.2020.9372453>.
- [11] Yang, J. et al., "Lightweight forest smoke and fire detection algorithm based on improved YOLOv5," PLoS ONE, vol. 18, e0291359 (2023); <https://doi.org/10.1371/journal.pone.0291359>.
- [12] Z. Xiao et al., "FL-YOLOv7: A Lightweight Small Object Detection Algorithm in Forest Fire Detection," Forests, vol. 14, p. 1812 (2023); <https://doi.org/10.3390/f14091812>.
- [13] Z. Yang, Y. Shao, Y. Wei, and J. Li, "Precision-Boosted Forest Fire Target Detection via Enhanced YOLOv8 Model," Appl. Sci., vol. 14, p. 2413 (2024); <https://doi.org/10.3390/app14062413>.
- [14] Huang, J., He, Z., Guan, Y., and Zhang, H., "Real-time forest fire detection by ensemble lightweight YOLOX-L and defogging method," Sensors, vol. 23, p. 1894 (2023); <https://doi.org/10.3390/s23041894>.
- [15] Yun, B., Zheng, Y., Lin, Z., and Li, T., "FFYOLO: A Lightweight Forest Fire Detection Model Based on YOLOv8," Fire, vol. 7, p. 93 (2024); <https://doi.org/10.3390/fire7030093>.
- [16] Howard, A. G et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv preprint arXiv:1704.04861 (2017); <https://doi.org/10.48550/arXiv.1704.04861>.
- [17] Zhang, X., Zhou, X., Lin, M., and Sun, J., "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6848–6856 (2018); <https://doi.org/10.48550/arXiv.1707.01083>.
- [18] Li, H. et al., "Slim-neck by GSConv: A lightweight-design for real-time detector architectures," J. Real-Time Image Process., vol. 21, no. 3, p. 62 (2024); <https://doi.org/10.48550/arXiv.2206.02424>.
- [19] Ouyang, D. et al., "Efficient multi-scale attention module with cross-spatial learning," in Proc. IEEE, Speech Signal Process. (ICASSP) (2023); <https://doi.org/10.48550/arXiv.2305.13563>.
- [20] de Venancio, P.V.A.B, Lisboa, A.C., and Barbosa, A.V., "An automatic fire detection system based on deep convolutional neural networks for low-power, resource-constrained devices," Neural Comput. Appl., vol. 34, no. 18, pp. 15349–15368 (2022); <https://doi.org/10.1007/s00521-022-07467-z>.
- [21] Vasilevskiy, O., Voznyak, O., Didych, V., Sevastianov, V., Ruchka, O., and Rykun, V., "Methods for Constructing High-precision Potentiometric Measuring Instruments of Ion Activity," 2022 IEEE 41st International Conference on Electronics and Nanotechnology (ELNANO), 247–252, (2022).
- [22] Kukharchuk, V.V., Holodiuk, V.V., et al., "Information conversion in measuring channels with optoelectronic sensors", Sensors, 22(1), 271 (2022).
- [23] Romanyuk, O.N., Pavlov, S.V., Romanyuk, O.V., et al., "A function-based approach to real-time visualization using graphics processing units", Proc. SPIE 11581, 115810E (2020).
- [24] Hemateja, A.V.N.M., "WildFire Smoke Dataset YOLO," Kaggle, (Apr. 2025). [Online]. Available: <https://www.kaggle.com/datasets/ahemateja19bec1025/wildfiresmokedatasetyolo>.
- [25] Grądz, Ż., Wójcik, W., Gromaszek, K., Kotyra, A., Smailova, S., Iskakova, A., Yeraliyeva, B., Kumargazhanova, S. and Imanbek, B. "Application of Fuzzy Neural Networks in Combustion Process Diagnostics," Energies 17(1), 212 (2024). <https://doi.org/10.3390/en17010212>
- [26] Bisikalo, O., Kharchenko, V., Kovtun V., et al., "Parameterization of the Stochastic Model for Evaluating Variable Small Data in the Shannon Entropy Basis. Entropy. 25(2):184. (2023); <https://doi.org/10.3390/e25020184>