A Semantic Segmentation Method for Flame and Water Stream Landing Points Based on Dual-CBAM-MobileNet

Chongshuo Liu[1] and Yaojie Chen[2]

[1]Wuhan University of Science and Technology, Wuhan Hubei 430065, China
[2]Wuhan University of Science and Technology, Wuhan Hubei 430065, China
202313407431@wust.edu.cn

Abstract. Focusing on the problems of insufficiently rapid fire identification and imprecise fire suppression in large-space fire protection systems, this study designs a Dual-CBAM-MobileNet (dual attention-driven feature extraction network) module based on the encoder-decoder architecture of DeepLabV3+ to achieve real-time and accurate segmentation of flames and water stream landing points. First, a light-weight MobileNetV2 feature encoding network is used as the core backbone. Second, the CBAM is introduced and integrated into the MobileNetV2 to enhance the ability to detect flames and water stream landing points.Finally, to further mitigate the issue of category imbalance, Focal Loss is employed as the loss function. Experiments demonstrate that our method enhances image segmentation precision for both flames and water stream landing points while maintaining a high image processing speed. On the self-constructed flame-water stream landing point dataset, it achieves a mIoU of 76.23% and a processing speed of 78.68 frames per second (FPS),satisfying both the real-time demands and accuracy in practical fire protection systems.

Keywords: Semantic Segmentation; DeepLabV3+; Attention Mechanism; MobileNetV2; Lightweight

1 Introduction

Fire, a extremely devastating disasters, consistently constitutes a grave danger to public safety, industrial production, and cultural heritage worldwide. However, the spatial structures of modern buildings are becoming increasingly complex, and occupant density continues to rise. Traditional point-type fire detection systems based on smoke and temperature sensors have limitations such as response delays and vague localization, making it difficult to achieve early, precise warning and efficient suppression in large open spaces like airports, stadiums, and warehousing logistics centers [1]. In contrast, computer vision-based fire protection systems, by analyzing surveillance video sequences, can achieve non-contact, wide-area, and real-time fire perception, demonstrating immense application potential [2].

In recent research, Liu et al. [3] proposed the “Defog DeepLabV3+ model ”for forest fire segmentation in foggy conditions. By introducing a joint defogging learning framework and a Dual Fusion Attention Mechanism (DARA), the model

achieved an mIoU of 89.51% on the synthesized foggy dataset FFLAD, effectively improving segmentation accuracy under low visibility. However, this study focuses on static image recognition, and the model incorporates an independent defogging branch along with complex attention modules, leading to a significant increase in parameters and an inference speed of only about 0.3 FPS. This makes it difficult to achieve the strict time-critical video stream processing demands of real-world firefighting systems. Jasmine et al. [4] proposed a dual-attention DeepLabV3+ model for underwater object detection, employing a MobileNetV2 backbone combined with CBAM and ECA attention mechanisms, which represents a valuable exploration in balancing accuracy and efficiency. Nevertheless, the best version of their model still achieves only 17.4 FPS in inference speed, indicating room for further improvement. Moreover, few recent related studies have addressed the segmentation of water jet landing points, which remains a challenging case.

To tackle these issues, the paper introduces an improved DeepLabv3+ semantic parsing architecture for large-space fire protection systems. The main contributions include:

Adopting MobileNetV2, a lightweight architecture, as the core backbone feature extraction network, replacing the original Xception network. MobileNetV2, based on linear bottlenecks and inverted residual structures, retains decent feature representation ability while substantially cutting down on computational load and parameter quantity. The Convolutional Block Attention Module (CBAM) [5] is embedded into both the deep-level and shallow-level feature layers of the network. This module successively calculates attention weights across both the channel and spatial axes, enabling the model to adaptively focus on salient features related to flames and water streams while suppressing irrelevant background interference, thereby enhancing the capture capability for subtle features (e.g., water streams, landing points). To address the issue of category imbalance, the conventional cross-entropy loss is replaced by Focal Loss[6]. This loss function emphasizes the weights of hard-to-classify samples and rare class samples, effectively mitigating class imbalance during training and improving the model's segmentation accuracy for critical targets.

2 Related Classical Modules

2.1 MobileNetV2 Network Architecture

MobileNetV2 is mainly intended to create a light-weight and computationally efficient neural network architecture tailored for mobile and embedded platforms [7]. The core of this network architecture is the Inverted Residual Block. Figure 1 clearly reveals that the inverted residual structure adopts a "narrow-wide-narrow" design. First, the module uses a 1×1 convolutional layer to significantly increase the channel dimension. Then, this high-dimensional space applies a lightweight 3×3 depthwise separable convolution to derive features. Lastly, a linear 1×1 convolutional layer compresses the channel count back to the target dimension. This series of designs

allows the module to achieve powerful representational capabilities far beyond traditional structures while being extremely lightweight.

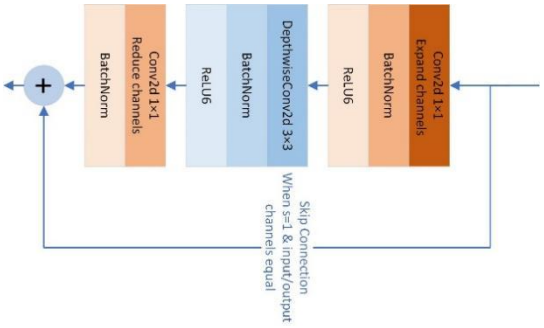

Fig. 1. Inverted Residual Block

2.2 CBAM Attention Mechanism

CBAM is an innovative attention module that enhances the model's representational power through the collaborative work of a dual attention mechanism. This module has two key parts: the “Channel Attention Unit”, responsible for assessing the importance of different feature channels; and the “Spatial Attention Unit”, which focuses on identifying discriminative spatial regions in the feature map. The CBAM attention mechanism module is illustrated in the figure below.

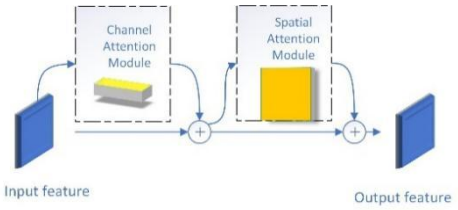

Fig. 2. CBAM attention module

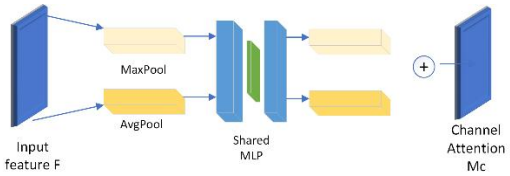

Fig. 3. CAM architecture

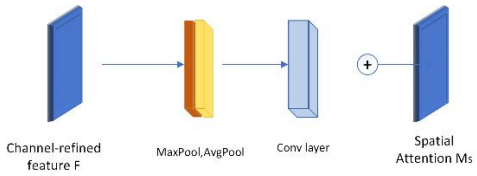

Fig. 4. SAM architecture

Through this carefully designed dual attention mechanism, the CBAM module can adaptively calibrate feature responses, enabling the network to focus more on feature channels and spatial locations relevant to the task, thereby significantly improving the model's performance in various visual tasks.

3 Improved DeepLabv3+ Semantic Segmentation Algorithm Design

3.1 Overall Framework

The semantic segmentation model proposed in this paper is constructed based on the encoder-decoder architecture of DeepLabV3+ [8]. In the encoder part, the model adopts the Dual-CBAM-MobileNet and ASPP modules. The decoder part employs a feature fusion strategy: the high-level semantic features output by ASPP are subjected to 4× upscaling and concatenation with the CBAM-enhanced shallow features. The combined features are reconstructed by a 3×3 convolution, then upscaled 4× via bilinear interpolation to recover the original resolution and yield the final segmentation result.

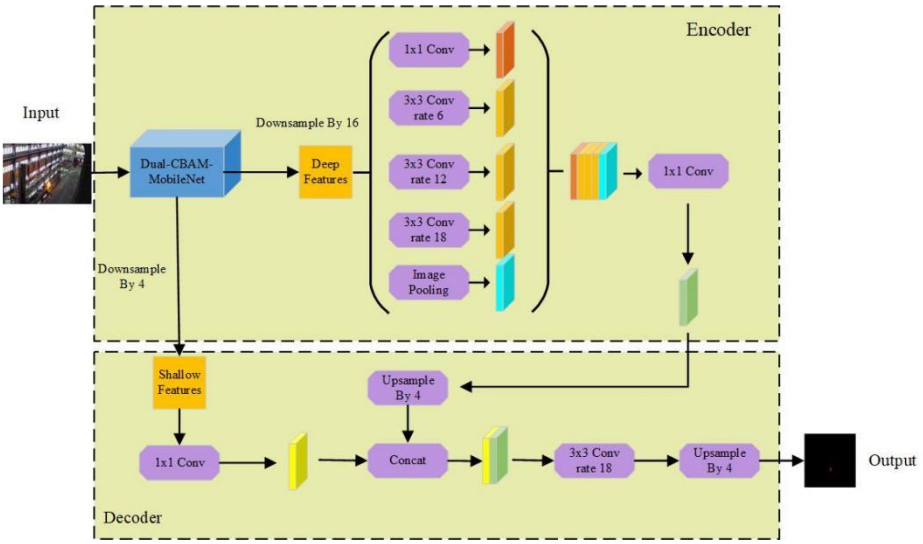

Fig. 5. Improved DeepLabv3+ model structure

3.2 Dual-CBAM-MobileNet Network Architecture

This structure is based on the lightweight MobileNetV2, which progressively extracts multi-level features through cascaded Inverted Residual Blocks. Convolutional Block Attention Modules (CBAM) are integrated at two of its critical stages, forming a dual attention mechanism-driven feature extraction network, as shown in Figure 6. Specifically, two CBAM modules are embedded at different depths of the network: one at the shallow level (output of layer3, with 24 channels) and one at the deep level (output of layer7, with 160 channels). These modules boost key feature extraction through the dual attention scheme operating over channel and spatial axes.

The firefighting process is a dynamically changing scene, and the segmentation of water stream landing points is particularly challenging, often resulting in an IoU value significantly lower than other categories. This difficulty stems from several factors: (1) Irregular morphology and high transparency: The shape of water stream landing points changes rapidly, and they blend highly with the background, resulting in extremely blurry edges. (2) Weak contrast with the background: Especially in complex environments, the features of splashing water areas are similar to those of the ground, smoke, and other backgrounds, making them difficult to distinguish.

Therefore, traditional segmentation models often perform poorly when dealing with these challenges due to insufficient feature extraction capability or inadequate focus on key information. To address the aforementioned problems, this model employs Dual-CBAM-MobileNet to optimize the results.

The shallow CBAM enhances the detailed features from the early encoder stages: (1) Capturing faint edges of water landing points: Effective information for water landing points is often contained in shallow-level detail features. The spatial attention mechanism of the shallow CBAM can amplify the response of these blurry, discontinuous edges, making them "stand out" in the feature map. (2) Improving channel representativeness for hard-case features: The channel attention mechanism optimizes the channel combination of shallow features, assigning higher weights to filters useful for recognizing subtle phenomena like transparent water spray and reflections, thereby constructing more discriminative low-level feature representations specifically for hard cases like "water stream landing points."

The deep CBAM refines the high-level features toward the encoder's final stage: (1) Enhancing target saliency: Through channel attention, the model can autonomously learn and amplify the weights of feature channels related to "flame" and "water," while suppressing unimportant background channels. This makes the semantic information of flames and water streams more prominent in complex fire environments. (2) Focusing on key spatial regions: Through spatial attention, the model can focus on key areas such as flame burning zones and water landing points on the global feature map, providing purer, more discriminative high-level features for the subsequent ASPP module, thereby improving the understanding of the overall scene.

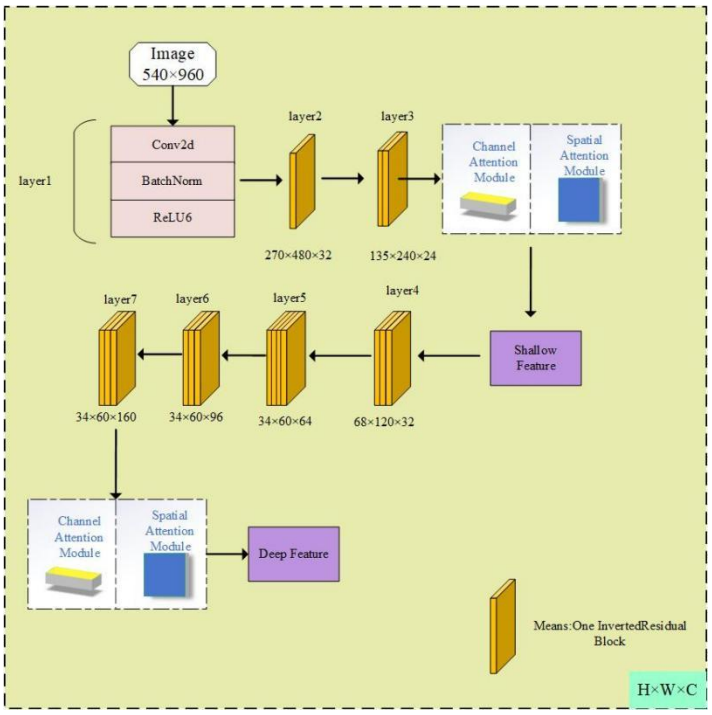

Fig. 6. Dual-CBAM-MobileNet Block

3.3 Loss Function

In semantic segmentation tasks, the uneven distribution of pixel counts across different classes can significantly affect model performance. In large-space firefighting scenarios, background-class pixels typically dominate the image, while targets containing critical semantic information (such as flames, water streams, and their landing points) often constitute only a tiny fraction. This class imbalance phenomenon can easily lead to bias during model training. To address this challenge, this study adopts Focal Loss as the model's optimization objective. It is an improvement upon the traditional multi-class “cross-entropy loss”. By introducing a dynamically adjustable weighting factor, it diminishes the influence of easy-to-classify samples on the total loss, thereby letting the model focus more on hard-to-categorize ones.

The specific mathematical expression of this loss function is as follows:

First, the standard multi-class “cross-entropy loss” is defined as:

$$CE(p,y)=-\log(p_y) \tag{1}$$

Building on this, Focal Loss is improved by introducing two modulating factors. Its complete expression is:

$$FL(p,y)=-\alpha_y(1-p_y)^{\gamma}\log(p_y) \tag{2}$$

In the application scenario of this study, by reasonably setting the γ value and the α_y parameters for each class, the model's segmentation accuracy for challenging cases like water stream landing points has been significantly improved while maintaining its ability to recognize common targets.

4 Experimental Results and Analysis

4.1 Dataset Introduction

This study adopts a self-built intelligent fire fighting dataset. The raw data is collected from video frames captured in water jet fire fighting simulation experiments at multiple shipyards, with a resolution of 540×960 pixels. Frames with clear visual distinguishability are extracted from the original videos and manually annotated to generate corresponding PNG label files. The dataset includes 6,000 images in total, partitioned into three subsets: 4,000 for training, 1,000 for validation, and 1,000 for testing.

4.2 Experimental Setup and Parameter Configuration

The experiments were implemented using the deep learning framework PyTorch on the Windows 11 operating system. The hardware specifications include an Intel® 13th Gen Core™ i5-13500HX CPU, 8 GB of GPU memory, and an NVIDIA GeForce RTX 4060 Laptop GPU. The software environment includes PyCharm 2024, PyTorch 1.12, Python 3.10, CUDA 11.8, among others.

4.3 Performance Evaluation Metrics

To evaluate the effectiveness of the model, the mean Intersection over Union (mIoU) and mean Pixel Accuracy (mPA) were adopted as evaluation metrics, calculated as shown in Equations (3) and (4).

$$\mathrm{mIoU}=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}(p_{ji}-p_{ui})} \tag{3}$$

$$\mathrm{mPA}=\frac{1}{k+1}\sum_{i=0}^{\kappa}\frac{p_{ii}}{\sum_{j=0}^{\kappa}p_{ji}} \tag{4}$$

Additionally,Frames Per Second is used to evaluate the model's recognition speed and online monitoring capability.

4.4 Ablation Study

To validate the core improvement effect of the proposed Convolutional Block Attention Module (CBAM) and its optimal configuration scheme, this study conducts arguments through systematically designed ablation experiments. The experimental

design encompasses two research dimensions: firstly, analyzing the combined effectiveness of employing CBAM in both deep and shallow layers of the feature extraction network, and secondly, determining the optimal embedding position for CBAM in the shallow features. All comparative tests shared the same training and validation splits, with mIoU, mPA. The Intersection over Union for the "water stream landing point" (droppoint) class as the core evaluation metrics.

Ablation Study on the Number of CBAM Modules in the Feature Extraction Network.

To explore the combined effectiveness of employing CBAM in both deep and shallow layers of the feature extraction network, the baseline model (No. 1) without CBAM achieved an mIoU of 70.16% and a droppoint IoU of 52.75%, providing a benchmark for subsequent improvements. When CBAM was applied only to shallow features (output of MobileNetV2's layer3) (No. 2), the mIoU increased to 72.56%, and the droppoint IoU grew by 1.38 percentage points to 54.13%, indicating that enhancing spatial details in shallow features aids in recognizing small targets. When CBAM was applied solely to deep features (output of layer7 before the ASPP module) (No. 3), the model achieved the highest mIoU (73.42%) and droppoint IoU (55.32%) among the four configurations, demonstrating that the attention mechanism on deep features plays a crucial role in integrating global context. Finally, by deploying both shallow and deep CBAM (No. 4), the model performance reached its optimum, with mIoU, mPA, and droppoint IoU increasing to 74.23%, 86.32%, and 56.57%, respectively, proving the complementary effect of the dual attention mechanism in detail enhancement and context optimization.

Table 1. Impact of Different CBAM Combinations on Prediction Results

No.	shallow	deep	mIoU(%)	mPA(%)	droppointIoU(%)
1			70.16	80.34	52.75
2	√		72.56	84.15	54.13
3		√	73.42	85.34	55.32
4	√	√	74.23	86.32	56.57

Note: √ indicates adding a CBAM module after the shallow features (shallow feature) or deep features (deep feature) in the feature extraction network. 'droppoint' refers to the landing point of the water stream.

Ablation Study on the Position of CBAM for Shallow Feature Enhancement.

For the detailed study on the optimal position for the shallow CBAM, this study sequentially embedded the module after different network stages (layer1 to layer5) for

testing. Experimental data show that performance does not change linearly with network depth but achieves the best result after layer3 (No. 3). Specifically, in stages layer1-2, the features are insufficiently abstract, and the attention mechanism may amplify noise, leading to performance degradation (No. 1-2). In stages layer4-5, due to the reduced resolution of feature maps, detailed information loss is difficult to recover through attention, resulting in performance decline (No. 4-5). This verifies that embedding CBAM after layer3 achieves a balance between feature abstraction and spatial resolution, which is most beneficial for segmenting small targets.

Table 2. Impact of Shallow CBAM at Different Positions on Prediction Results

No.	layer_x	mIoU(%)	mPA(%)	droppointIoU(%)
1	layer1	65.13	74.61	47.85
2	layer2	70.81	84.37	55.92
3	layer3	74.23	86.32	56.57
4	layer4	73.42	83.86	52.35
5	layer5	71.59	80.18	50.77

Note: layer_x indicates adding a CBAM module for shallow feature enhancement after the x-th layer of the feature extraction network. 'droppoint' refers to the landing point of the water stream.

4.5 Comparative Experiments.

To verify the efficacy of our proposed enhanced model, we carried out comparison tests under identical experimental settings against several conventional semantic segmentation models, including UNet [9], PSPNet [10], and DeepLabV3+ [8]. Additionally, different backbone networks—ResNet50 [11], MobileNetV2 [7], and Xception [12]—were employed in the semantic segmentation models for further comparison. Furthermore, recent state-of-the-art (SOTA) lightweight segmentation models, DDRNet23[13] and BiSeNetV2[14], were also included in the comparison.The results are presented in Table 3.

Table 3. Comparative Results of Different Models on the Test Set

Model	mIoU(%)	mPA(%)	droppointIoU(%)	FPS

Model	mIoU(%)	mPA(%)	droppointIoU(%)	FPS
DeepLabV3+(Xception)	74.37	83.84	56.81	23.67
DeepLabV3+(ResNet50)	74.33	85.26	56.24	32.36
DeepLabV3+(MobileNetV2)	70.16	80.34	52.75	86.86
UNet (ResNet50)	73.15	81.72	52.58	15.34
PSPNet (ResNet50)	68.49	78.49	46.27	41.98
DDRNet23	74.92	82.69	56.85	60.20
BiSeNetV2	73.23	80.76	54.78	105.49
Ours	74.23	86.32	56.57	78.68

According to the quantitative findings, our enhanced model strikes an excellent trade-off between precision and speed. Regarding segmentation performance, the mIoU achieved by our model is 74.23%, falling slightly below DDRNet23's 74.92% but exceeding BiSeNetV2's 73.23%, and it stays on par with DeepLabV3+(Xception) (74.37%) and DeepLabV3+(ResNet50) (74.33%). It represents a 4.07% improvement over the original DeepLabV3+(MobileNetV2) (70.16%). Notably, the model achieves an mPA of 86.32%, outperforming all compared models, including DDRNet23 (82.69%) and BiSeNetV2 (80.76%), and yields a 6% increase over the original MobileNetV2 version, validating the effectiveness of the attention mechanism and decoder optimization.

Regarding key target detection, the model achieves a droppointIoU of 56.57%, which is very close to the best result of 56.85% obtained by DDRNet23 and higher than BiSeNetV2' s 54.78%. It yields a 3.82% improvement over the original version. This indicates that through feature enhancement by the CBAM attention modules and multi-scale information fusion by the ASPP module, the model' s perceptual ability for detailed features is effectively enhanced while maintaining a lightweight advantage.

In terms of inference efficiency, the model achieves 78.68 FPS, which is lower than BiSeNetV2 (105.49 FPS) and the original MobileNetV2 version (86.86 FPS), but significantly faster than DDRNet23 (60.20 FPS) and other heavy backbone models. Specifically, the inference speed is 2.43 times that of DeepLabV3+(ResNet50), 5.13 times that of UNet(ResNet50), and 1.31 times that of DDRNet23. Although there is a slight speed reduction compared to the fastest variants, this trade-off brings

comprehensive improvements in all accuracy metrics, demonstrating a favorable performance balance. Specific segmentation results are compared in Figure 7:

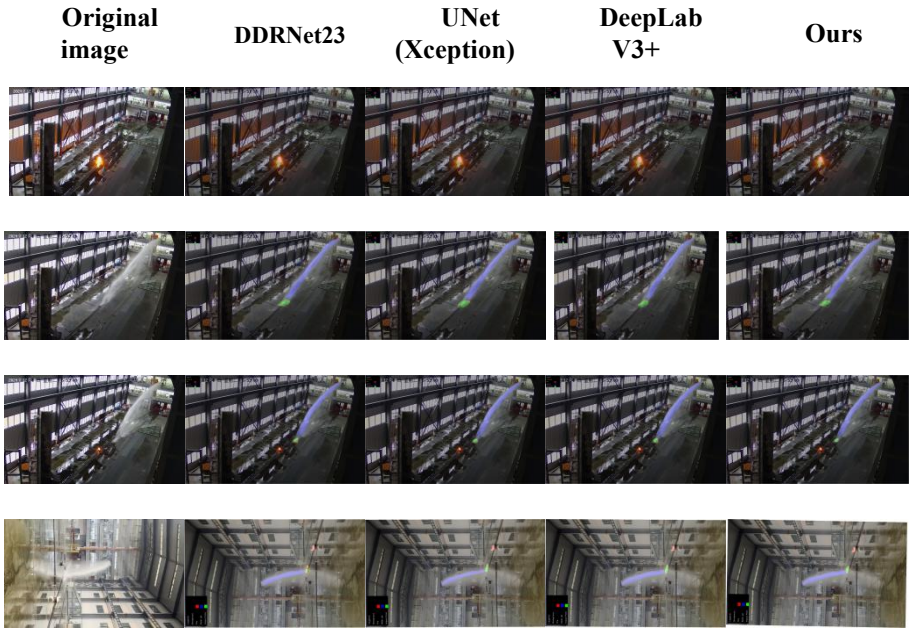

Fig. 7. Segmentation effects of different models on flames, water streams, and water stream landing points

To conclude, our enhanced model attains segmentation precision on par with or even beyond that of more complex architectures, all while sustaining near-real-time inference speed, with its standout performance on the mPA metric being especially noteworthy. This excellent performance balance makes the model particularly suitable for practical scenarios like intelligent firefighting that have high requirements for both real-time performance and accuracy.

5 Conclusion

Aiming to meet the key demands for both real-time performance and precision in smart fire protection systems, we present a segmentation architecture built upon an enhanced DeepLabV3+. Through the integration of Dual-CBAM-MobileNet and Focal Loss, the model strikes an effective trade-off between precision and speed.

Nevertheless, this work still has certain shortcomings. The model's effectiveness relies heavily on the quality and variety of the training samples, and its robustness in extreme scenarios requires further verification. In the future, we will put more attention on expanding the scale and scene coverage of the dataset and exploring

efficient deployment schemes for the model on edge devices, to promote the practical application of this technology in real-world fire protection systems.

References

1. Jin, C., Wang, T., Alhusaini, N., et al.: Video fire detection methods based on deep learning: datasets, methods, and future directions. Fire **6**(8), 315 (2023)
2. Secilmis, A., Aksu, N., Dael, F.A., et al.: Machine learning-based fire detection: a comprehensive review and evaluation of classification models. JOIV: Int. J. Inform. Visual. **7**(3-2), 1982–1988 (2023)
3. Liu, T., Chen, W., Lin, X., et al.: Defogging learning based on an improved DeepLabV3+ model for accurate foggy forest fire segmentation. Forests **14**(9), 1859 (2023)
4. Jasmine, J.J., Sherryl, R.M.R., Jarin, T., et al.: Dual-attention DeepLabv3+ with MobileNetV2 for efficient underwater object detection. Int. J. Inf. Technol. (2025). https://doi.org/10.1007/s41870-025-03053-3
5. Woo, S., Park, J., Lee, J.Y., et al.: CBAM: convolutional block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018)
6. Lin, T.Y., Goyal, P., Girshick, R., et al.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988. IEEE (2017)
7. Sandler, M., Howard, A., Zhu, M., et al.: Mobilenetv2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520. IEEE (2018)
8. Wang, P., Chen, P., Yuan, Y., et al.: Understanding convolution for semantic segmentation. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1451–1460. IEEE (2018)
9. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015)
10. Zhao, H., Shi, J., Qi, X., et al.: Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2881–2890. IEEE (2017)
11. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE (2016)
12. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251–1258. IEEE (2017)
13. Yu, C., Gao, C., Wang, J., et al.: BiSeNet V2: bilateral network with guided aggregation for real-time semantic segmentation. In: 2021 IEEE International Journal of Computer Vision (IJCV), pp. 1562–1583. IEEE (2021)
14. Hong, Y., Pan, H., Sun, W., et al.: Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. In: 2022 IEEE Transactions on Intelligent Transportation Systems (T-ITS), pp. 7156–7167. IEEE (2022)