frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 29 November 2023
DOI 10.3389/fphy.2023.1206944

OPEN ACCESS

EDITED BY
Anna Cimmino,
ELI Beamlines, Czechia

REVIEWED BY
Jakub Nalepa,
Silesian University of Technology, Poland
Guanqiu Qi,
Buffalo State College, United States

*CORRESPONDENCE
Guiqian Wang,
guiqianw@163.com

RECEIVED 16 April 2023
ACCEPTED 10 October 2023
PUBLISHED 29 November 2023

CITATION
Zhou F, Wen G, Pan H, Wang Y, Wang G and Yuan F (2023), A boundary enhancement and pixel alignment based smoke segmentation network.
Front. Phys. 11:1206944.
doi: 10.3389/fphy.2023.1206944

A boundary enhancement and pixel alignment based smoke segmentation network

Fangrong Zhou[1], Gang Wen[1], Hao Pan[1], Yifan Wang[1], Guiqian Wang[2]* and Feiniu Yuan[3,4]

[1]Joint Laboratory of Power Remote Sensing Technology, Electric Power Research Institute of Yunnan Electric Power Company, Kunming, China, [2]Mathematics and Science College, Shanghai Normal University, Shanghai, China, [3]College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, China, [4]Key Innovation Group of Digital Humanities Resource and Research, Shanghai Normal University, Shanghai, China

Image segmentation methods usually fuse shallow and deep features to locate object boundaries, but it is difficult to improve the accuracy of smoke segmentation by conventional fusion methods. It is a very difficult vision task to perform semantic segmentation of smoke images, because the translucency and irregular shapes of smoke lead to extremely complicated mixtures with background that are adverse to segmentation. To improve the segmentation accuracy of smoke scenes, we propose a Boundary Enhancement and Pixel Alignment based smoke segmentation Network for fire alarms. For the shallow features of the network, an attention mechanism is adopted to capture spatially details of smoke for improving boundary precision. For the deep layers, the Pyramid Pooling Module is used to extract local features and abstract semantic ones simultaneously. Finally, to efficiently merge shallow and deep features, a Pixel Alignment Module is adopted to model the relationship between pixel locations. The experimental results show that the mean Intersection over Union of the proposed method on the three synthetic smoke test datasets is 78.61%, 77.63% and 77.30%, respectively, and it outperforms most of the existing methods. In addition, our method obtains satisfying results on inconspicuous smoke and smoke-like images.

KEYWORDS

smoke segmentation, deep neural network, boundary enhancement module, pixel alignment module, pyramid pooling module

1 Introduction

The frequent occurrence of fires not only causes significant economic losses for society, but more importantly, it will jeopardize social public security and have extremely bad effects on the natural and ecological environment. It is too late to detect the naked flame, so the recognition of smoke early in the fire can control this disaster to a certain extent.

To solve fire detection in open or large spaces, a number of deep models [1–6] have been proposed for fire detection. These deep fire detection methods differ from traditional ones. Deep learning models segment smoke areas at a fine granularity for separating smoke targets from cluttered backgrounds, which are more accurate than those obtained by traditional methods. According to segmented maps, staffs can analyze safe areas and predict fire trends for reducing damages. Detecting smoke from fires at the pixel level is important, but the task is very challenging due to the variability introduced by the visual characteristics of smoke.

Smoke is visually semi-transparent, so it becomes more challenging to separate pixels in smoke edge regions. The problem that needs to be focused on is how to collect semantic information for categorization and localization. Some methods try to increase the depth of network for capturing more semantic features, but this technique also loses local details. The solutions to the above-mentioned conflicts can be divided into three main categories. The first category is to adopt a skip-connection structure [7] between deep and shallow levels. U-Net [7] adopts gradual upsampling to avoid the loss of features. The second one is to use the atrous convolution [8] to capture information at large scales. The third one is to use the multi-scale fusion [9] approach to integrate information from different scales. Spatial attention [10] is able to capture the most informative region of feature maps by globally modeling the relevance of all pixels, thus it effectively solves the misclassification problem of isolated regions that are far away from the main smoke area and strengthens the boundaries of smoke.

Based on the analysis of existing methods, we propose a Boundary Enhancement and Pixel Alignment based smoke segmentation Network (BEPA-Net). For the shallow features, we design a Boundary Enhancement Module (BEM) to model the long-range ability of attention mechanism for obtaining clear target boundaries. For the deep features, we adopt the Pyramid Pooling Module (PPM) [11] to extract global and local contextual information for enhancing semantics. By fusing different resolution feature maps, we propose a Pixel Alignment Module (PAM), which can construct the pixel correspondence relationship between feature maps for better information fusion.

This paper is organized as follows. Section 2 describes related work on image semantic segmentation and smoke segmentation. In Section 3, we describe the main idea of this paper. Section 4 presents experiments and analysis. At last, we conclude this paper in Section 5.

2 Related works

2.1 Semantic segmentation

Semantic segmentation is an image classification task at the pixel level [12]. proposed a Full Convolution Network (FCN), which is a basic paradigm of semantic segmentation methods. FCN pioneers an end-to-end approach to achieve pixel-by-pixel classifications. In the encoder of FCN, a series of convolutional layers and successive down-sampling ones are often used to extract deep features with large receptive fields, and then the decoder upsamples the extracted deep features to the same resolution as the input. To lessen the loss of spatial information caused by down-sampling, skip connections are used to fuse low-level features with high-level ones from the different scales of encoders and decoders. Two fundamental paradigms have emerged as a result of further researches, including symmetric codec architectures [7, 13] and asymmetric codec ones [8, 14–17]. These symmetric codec structures mainly focus on minimizing the information loss caused by frequent down-sampling for expanding receptive fields. Meanwhile, asymmetric codec structures revolve the problem of extracting the most abstract semantics without reducing the feature map resolutions too much. Thus, a balance of spatial and semantic information can be achieved.

2.2 Smoke segmentation

Traditional smoke segmentation methods make an effort to separate smoke targets from images by extracting the color features of images in different color spaces [18]. Combined color features and shape features to present a fast smoke detection method for video surveillance [19]. Used background removal methods and color features to filter non-smoke pixels [20]. Used the rough set theory to extract candidate smoke regions for video fire detection.

With the rapid development of deep learning in recent years, there are many deep neural networks that have also been proposed for smoke semantic segmentation. Without the need for designing complex hand-crafted features, deep learning based semantic segmentation approaches combine feature extraction and classification for implementing an end-to-end manner [21]. directly used the AlexNet network [22] for smoke recognition [23]. proposed a 3D parallel FCN model to segment smoke regions from videos [24]. proposed a coding and decoding network by designing a dual-path structure to obtain detailed information and semantic information for smoke segmentation [25]. proposed a Wave-shaped deep neural Network (W-Net) for smoke density estimation, which is factually a regression over each pixel [26]. proposed a global smoke attention network that makes a full use of the modeling capabilities of attention mechanism.

Traditional methods rely on manual features, while recent networks only focus on the performance of CNN itself and cannot perform smoke segmentation well. We focus more on the correlation between smoke boundaries and pixels, and complete smoke segmentation by improving the segmentation ability of boundaries and strengthening global modeling capabilities.

3 The proposed method

3.1 The network architecture

We choose the ResNet50 network [27] as our backbone network because it can obtain rich information. As shown in Figure 1, the ResNet50 [27] is used as the backbone of our method to extract features, and the backbone network is divided into four stages. Atrous convolutions [30] increase the receptive field and extract the abundant features of images without significantly reducing spatial resolutions, so we adopt atrous convolutions to compensate for the reduction of spatial resolutions due to down-sampling. The outputs of Stages 2, 3, and 4 greatly improve the semantic representation of deep features after feature fusion and multi-scale context extraction. The shallow features from Stage 1 are first delivered to the proposed Boundary Enhancement Module (BEM) for pixel alignment between different feature maps of different layers. Then the proposed Pixel Alignment Module (PAM) accepts the warping information of the BEM module for information fusion. Thus, we implement pixel alignment for different features. Finally, the fused structure map is upsampled to the original map size for generating the final segmentation map.

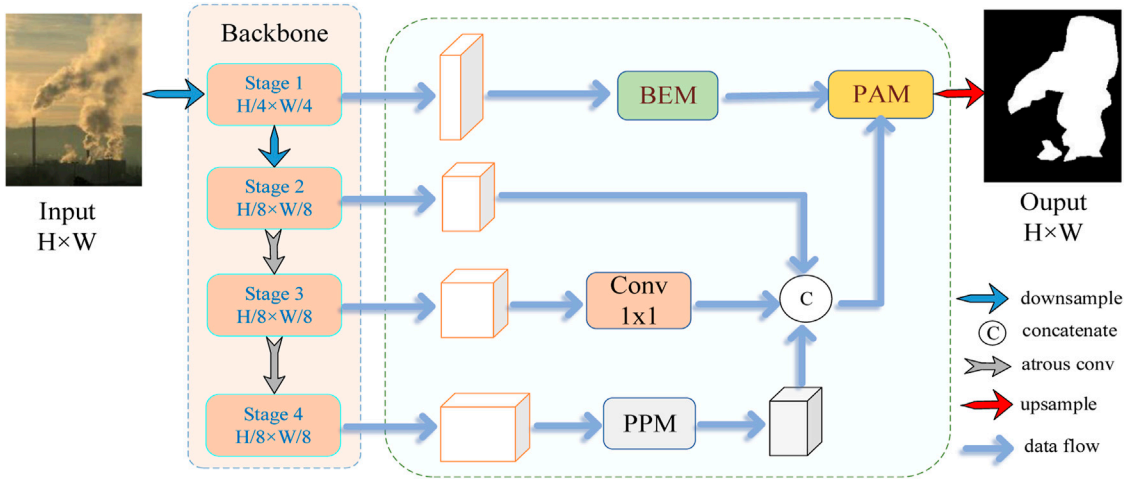

FIGURE 1
Deep and shallow features based Smoke Segmentation Network. BEM, PAM and PPM denotes Boundary Enhancement Module, Pixel Alignment Module and Pyramid Pooling Module, respectively. Conv 1 × 1 is a convolutional kernel with the kernel size equal to 1.

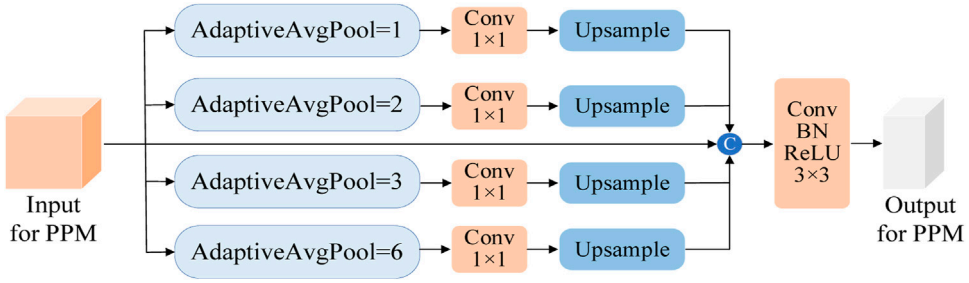

FIGURE 2
Pyramid Pooling Module (PPM). AdaptiveAvgPool is average pooling that uses different pooling kernel. Conv-BN-ReLu 3 × 3 denotes a convolution kernel with batch normalization and ReLU activation function.

3.2 Pyramid pooling module

In deep neural networks, the size of receptive fields can roughly represent the degree of using contextual information. High-level contextual information can be captured by explicitly fusing features of objects at different scales, and it can effectively solve the problem of pixel inconsistency within the objects in the segmentation task and reinforce deep semantics.

To obtain more contextual information, we use the Pyramid Pooling Module (PPM) [11] to obtain the context information of objects, as shown in Figure 2. The PPM module contains global and local information using the adaptive global pooling with the different output sizes of 1, 2, 3, and 6. Then, the pooled feature maps are filtered using a 1 × 1 convolution. Next, bilinear interpolation is used to up-sample these filtered feature maps to the same size as the original input of PPM. The resized feature maps are concatenated together. Three layers of convolution (Conv), batch normalization (BN), and activation (ReLU) are used to obtain a feature map with a large amount of contextual information.

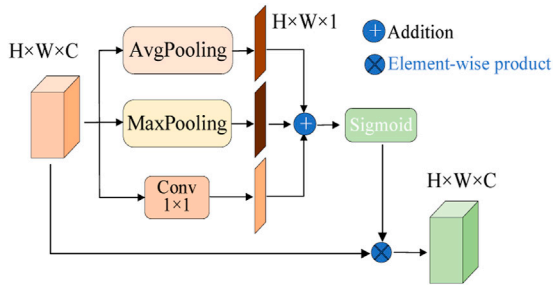

FIGURE 3
Boundary enhancement module (BEM).

3.3 Boundary enhancement module

The embedding of spatial contexts can emphasize the most informative parts and enable the network to selectively focus on more important features. The proposal of attention mechanism

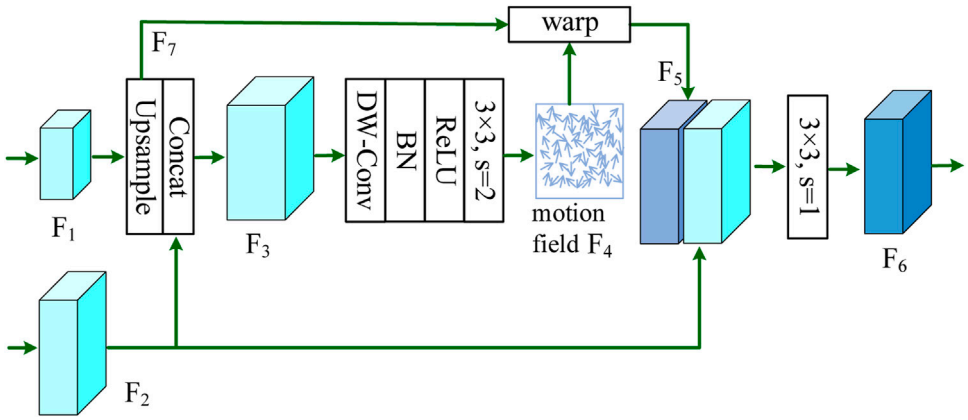

FIGURE 4
Pixel Alignment Module (PAM). Concat means concatenation along channel, DW-Conv is a deeply separable convolutional kernel with a size of 1, BN is a batch normalization operation, ReLU is an activation function, and "3 × 3, s = *k*" represents a convolution with a step size of *k* and a kernel size of 3 × 3.

offers a new direction for extracting powerful spatial and contextual information. Most spatial attention methods adopt matrix operations to capture the relationship between any two pixels in the global scope. To obtain spatial attention maps and improve object boundary localization accuracy, we propose a Boundary Enhancement Module (BEM). As shown in Figure 3, the input feature map of BEM is fed into the three branches for extracting features. *H*, *W* and *C* denote the height, width and channel numbers of the feature map, respectively. Average pooling captures the low frequency components of features, maximum pooling is able to extract the high frequency signals, and a 1 × 1 convolution learns the features about objects. As a result, we obtain three two-dimensional attentional feature maps, which are fused by point-wise summation. The fused feature map is activated by the *sigmoid* function for generating a spatial attention map. Finally, the input to our BEM is weighted with the spatial attention feature map to produce the feature map of the adaptive refinement boundary. Conv1x1 is a convolutional kernel with a kernel of 1.

3.4 Pixel alignment module

To aggregate multi-scale features, feature fusion is often performed in different levels of feature maps, but pixel positions corresponding to their objects are different. To solve this problem, we propose a Pixel Alignment Module (PAM). This module first calculates the pixel offsets for pixel alignment, and the feature maps are pixel-aligned for effective fusion. Figure 4 shows the proposed PAM, which accomplishes pixel alignment by obtaining the pixel offset field. The value of each pixel in the offset field can be viewed as a moving distance for the pixel. In other words, the offset field can also be called a motion field. The relationship between a feature map and another one obtained by convolutions can be reconstructed by the pixel motion field. Thus, we actually obtain translation invariance by convolutions.

A bilinear interpolation layer is used to up-sample the feature map F_1 to produce a feature map F_7, which has the same size as the feature map F_2, and feature maps F_7 and F_2 are concatenated to generate a feature map F_3 for channel fusion. A set of depth-wise separable convolutions (DW-Conv) are used to establish the positional relationships between pixels on different feature maps. The pixel motion field $F_4 \in R^{H\times W\times 2}$ is then generated using a 3 × 3 convolution in a similar way to DCN [28]. The field map F_4 contains the spatial translation offsets along x and y-axes, and the feature values at each pixel position ρ_l on F_4 are used to move the pixel of F_1 to a new position, resulting in a warped feature map $F_5 \in R^{H\times W\times 256}$, formulated as:

$$F_5(\rho_l) = \sum_{\rho\in\delta(\rho_l)} \omega_\rho F_4(\rho_l) \tag{1}$$

where ω_ρ denotes the weight of the bilinear kernel on the curved space grid, which is calculated by F_4, and $\delta(\rho_l)$ denotes the adjacent position of pixel position ρ_l.

The warped feature map F_5 is generated by F_1 and F_4, and F_4 is produced from F_1 and F_2, so F_5 certainly has a strong relationship with F_2. Therefore, we concatenate F_5 and F_2 along the channel axis for further feature enhancement. Finally, the concatenated feature map is processed by a 3 × 3 convolution without BN and ReLU layers for feature fusion and dimensionality control to obtain the final output F_6.

4 Experiments and analysis

4.1 Datasets and implementation details

In this paper, the datasets used for training and testing are the same as those in [25], including a virtual synthetic smoke training set with 70,632 images and three virtual synthetic smoke test sets. The three test sets are respectively named DS01, DS02 and DS03, and each set has 1,000 images. The synthetic dataset was created from 8,162 pure smoke images [25]. adopted computer graphics to generate these pure smoke images with a variety of transparency, texture and fluid properties. Each pure smoke sample is an RGBA image with a spatial resolution of 256 × 256, containing RGB color channels $\boldsymbol{S}$ and a opacity channel α, respectively. The opacity α is

TABLE 1 Comparison results of existing algorithms.

Methods	mIoU (%)		
	DS01	DS02	DS03
FCN-8S	64.03	63.28	64.38
SegNet	56.94	56.77	57.18
SMD	62.88	61.50	62.09
TBFCN	66.67	65.85	66.20
Deeplabv1	68.41	68.97	68.71
ESPNet	61.85	61.90	62.77
LRN	66.43	67.71	67.46
DSS	71.04	70.01	69.81
HG-Net2	63.58	62.40	63.61
HG-Net8	63.85	63.27	64.46
W-Net	73.06	73.97	73.36
GSANet	73.13	73.81	74.25
ViT	75.20	75.29	74.10
Swin-Transformer	76.49	75.55	75.80
SegFormer	78.76	78.50	78.03
Our	78.61	77.63	77.30

limited in the range [0, 1]. According to the rules of linear composition, the combination of a pure smoke image $\boldsymbol{S}$ and a background $\boldsymbol{B}$ generates an observation image $\boldsymbol{I}$. This procedure can be mathematically defined as:

$$\mathrm{I} = \alpha \cdot S + (1 - \alpha) \cdot B \tag{2}$$

Using the above method, we are able to construct a large number of training sets without the tedious labeling. Each virtual smoke image is randomly and linearly combined with a real background image for generating a virtual smoke training dataset. The generated virtual dataset is diverse in terms of colors, sizes, textures of smoke and background for simulating most real smoke scenes.

All the experiments were performed on a windows 10 PC with an NVIDIA RTX3090 GPU, and the programming environment is the python 3.7 and pytorch 1.7 framework. The Stochastic Gradient Descent (SGD) optimizer was used for training. The learning rate is set to 0.0001, the momentum is set to 0.9, and the learning rate decay is set to 0.95 using step decay. The mean Intersection over Union (mIoU) is used as the segmentation metric. mIoU is widely used to evaluate the overall performance of semantic segmentation algorithms and reflects the degree of overlap between the predicted results and their corresponding true labels. mIoU is formulated as:

$$\mathrm{mIoU} = \frac{1}{N}\sum_{k=1}^{N}\frac{P_k \cap G_k}{P_k \cup G_k} \tag{3}$$

where P_k and G_k are the predicted result of the kth image and the corresponding true label, respectively.

4.2 Comparison experiments

To evaluate the effectiveness of the proposed network, we tested it on three synthetic test datasets and one real smoke dataset, and compared it with several state-of-the-art semantic segmentation methods based on deep learning, including FCN-8S [12], SegNet [13], SMD [29], TBFCN [9], Deeplab v1 [15], ESPNet [31], LRN [32], DSS [24],HG-Net [33], MS-Net [34], W-Net [25], and GSANet [26]. In addition, we also compared it with some Transformer structures ViT [35], Swin-Transformer [36] and SegFormer [37]. To objectively and fairly evaluate the performance of each method, we used the same dataset and experimental configurations to train all the compared methods, and the results are shown in Table 1.

The mIoU metrics achieved by our method on the three virtual smoke test datasets are 78.61%, 77.63% and 77.30%, respectively. Our method achieves the good mIoUs among all the existing methods second only to the SegFormer.

The visualized segmentation results on virtual smoke images are shown in Figure 5, where the first and second columns are the original and labeled images, respectively. According to Table 1 and Figure 5, the models with the mIoU below 70 obtain poor performance, while DSS, W-Net, and GSANet, as specially designed smoke segmentation models, have good performance.

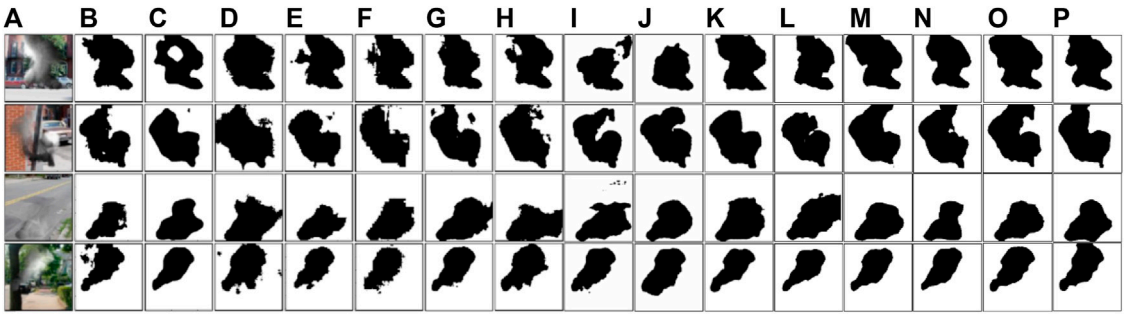

FIGURE 5
Segmentation results for the virtual smoke test dataset. **(A)** Virtual smoke images, **(B)** labeled maps, **(C)** FCN, **(D)** SegNet, **(E)** SMD, **(F)** TBFCN, **(G)** DeepLab v1, **(H)** ESPNet, **(I)** HG-Net 2, **(J)** HG-Net 8, **(K)** W-net, **(L)** GSANet, **(M)** ViT, **(N)** Swin-Transformer, **(O)** SegFormer, **(P)** the proposed method.

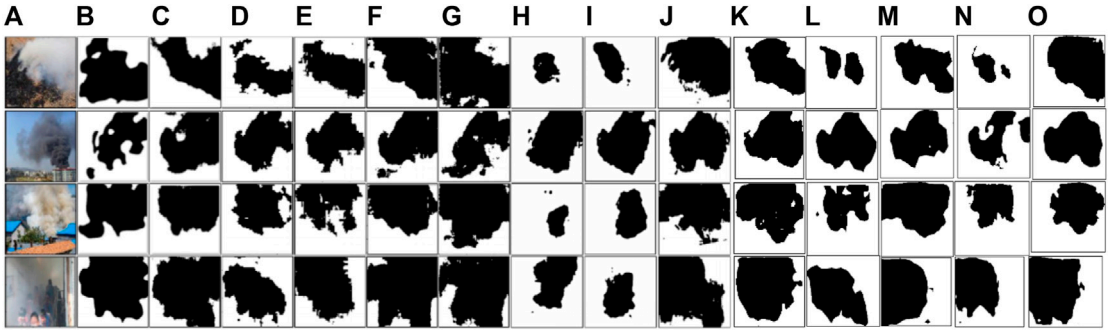

FIGURE 6
The segmentation results on the real dataset. **(A)** Realistic smoke images, **(B)** FCN, **(C)** SegNet, **(D)** SMD, **(E)** TBFCN, **(F)** DeepLab v1, **(G)** ESPNet, **(H)** HG-Net 2, **(I)** HG-Net 8, **(J)** W-net, **(K)** GSANet, **(L)** ViT, **(M)** Swin-Transformer, **(N)** SegFormer, **(O)** the proposed method.

TABLE 2 Ablation experiments for feature-fused.

Network architecture	mIOU(%)			Statistical analyses		
	DS01	DS02	DS03	t score	*p*-value	Significant
ResNet + Addition	70.68	66.22	67.15	1.8555	0.1371202	no (p>5%)
ResNet + Concat	71.84	67.18	68.49	1.1796	0.3035128	no (p>5%)
ResNet + Concat + PAM	73.35	69.64	70.78	—	—	—
ResNet + Addition + PAM	72.81	68.38	69.46	0.6022	0.5795012	no (p>5%)

TABLE 3 Ablation experiments of different modules.

ResNet50	Atrous-conv	PPM	BEM	PAM	mIOU(%)			Statistical analyses		
					DS01	DS02	DS03	t score	*p*-value	Significant
✓					61.67	60.23	62.09	24.04	0.0000178	yes (p<5%)
✓	✓				64.03	63.28	64.38	27.36	0.0000106	yes (p<5%)
✓	✓	✓			71.83	67.68	69.47	6.474	0.0029330	yes (p<5%)
✓	✓	✓	✓		73.53	69.70	70.16	5.290	0.0061303	yes (p<5%)
✓	✓	✓	✓	✓	78.61	77.63	77.30	—	—	—

However, compared to our method, they are still slightly worse, both in terms of evaluation metrics and visual image quality. Our network has more distinct boundaries that are basically consistent with the original image. Compared with the latest transformer structures in recent years, our model also has good performance, only slightly worse than Segformer.

The segmentation results on the real smoke images are essentially as good as those on the synthetic smoke images. As shown in Figure 6, the predicted results by our BEPA-Net are visually similar to their corresponding real images. To accurately locate smoke edges, feature maps require spatial details, local and global semantic abstractions for delineating smoke. Our BEPA-Net model is proposed to solve these problems. The reasons may be that fusing multi-scale features can easily extract global and local information for better smoke representations.

As shown in Figures 6H, I, the generalization of HG-Net is poor. Although it has achieved certain results on virtual datasets, the performance on real images is poor. The reason may be the lack of skip connections to complete the fusion of deep and shallow features. The segmentation area obtained by our method is basically consistent with the real smoke area. In addition, by comparing the visualized results of virtual smoke datasets and real data, we find that Transformers obtain good results on virtual data, but the results on real data were very poor, as shown in Figures 6L–N. Hence, it may be overfitting.

Compared with DSS, W-Net, and GSANet, our method uses multi-scale fusion and skip connections, resulting in better performance than them. This is because the pixel alignment is performed during feature fusion. This technique greatly improves model performance. We also compared the fusion methods in ablation experiments.

FIGURE 7
Smoke segmentation results in wilderness scenes.

4.3 Ablation experiments

Commonly used methods for feature fusion are pixel-wise addition (Addition) and channel concatenation (Concat). In this paper, we use the technique of pixel alignment followed by feature fusion to bridge the semantic gap between different feature maps during channel fusion. Further comparison results of feature fusion are shown in Table 2. The experiments show that performing pixel alignment first and then fusing features can achieve better results than other configurations. It proves that the proposed modules are powerful for feature representations.

To evaluate the performance of the proposed modules, several ablation experiments were performed on the data set for the different combinations of the proposed modules in our network. The results are shown in Table 3. We adopt ResNet-50 as the backbone network of all the variants of our method for ablation experiments. In Table 3, Atrous-Conv indicates the improved Atrous convolution, PPM is the Pyramid Pooling Module, BEM stands for the Boundary Enhancement Module, and PAM denotes the pixel alignment module.

According to Table 3, we find that the performance of the proposed network can be obviously enhanced by employing Atrous convolutions in the backbone network. After the pyramid pooling module (PPM) is enabled, we achieve the mIoUs of 71.83%, 67.68% and 69.47% on the test datasets of DS01, DS02, and DS03, respectively. The mIoUs are improved by about 2% after using the Boundary Enhancement Module (BEM). Although the boundary pixels often occupy a relatively small portion of the whole image, boundary information plays a key role in improving the accuracy of segmentation. The mIoUs are greatly improved by 7%–8% after the pixel alignment module (PAM) is used.

In addition, we compute the p-values of our results in Tables 2, 3 for analyzing the statistical significance of ablation experiments. Given two random sets X and Y, the t-score of the two sets is computed as follows:

$$t = \frac{\left|\mu_x - \mu_y\right|}{\sqrt{\frac{(n_x-1)\sigma_x^2+(n_y-1)\sigma_y^2}{n_x+n_y-2}\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}} \tag{4}$$

where μ_x, σ_x and n_x are respectively the mean, the standard deviation and the sample number of the best set X, and μ_y, σ_y and n_y are the mean, the standard deviation and the sample number of the tested set Y, respectively.

FIGURE 8
Experiments on images of electric power transmission lines. Reproduced from the Yunnan Electric Power Company, with permission from the Company.

According to the *t*-scores and degree of freedom $d_f = n_x + n_y$-2, we can find the ranges of the *p*-values in the *t*-score lookup table. For the sake of convenience, we use the *Excel* software to automatically compute the *p*-values. By observing Table 2, we find that all the *p*-values are greater than 5%, so it is not significant in statistics. In vision fields, although existing methods achieving only 1% increase of mIoUs are not significant statistically, they are often considered as excellent algorithms, and they do not perform *p*-value analyses. In Table 3, we find that all the *p*-values are far less than 5%, so the best variant is statistically significant.

4.4 Testing on wild scenes with electric power transmission lines

The proposed method also achieves good results for real fires in wild scenes, as shown in Figure 7. The proposed method can accurately segment smoke regions. Although there are smoke-like objects, such as clouds, our model easily discriminates between clouds and smokes.

In order to ensure the safety of electric power transmission lines, it is necessary to check fire safety around the electric lines. We tested the proposed method on several images captured from iron towers of electric power transmission lines, as shown in Figure 8. From the experimental results, we can see that the proposed method not only detects smoke successfully, but also obtain a relatively accurate smoke contours. Segmented smoke contours can allow the relevant personnel in the electric power department to have a more accurate judgment of the possible fire spread trend, and take corresponding countermeasures in advance to determine the safety of power lines.

5 Conclusion

In this paper, a deep neural network is proposed to improve the performance of smoke semantic segmentation. To learn the spatial details and contextual information about objects, we design a spatial attention mechanism to enhance the localization accuracy of object boundaries for improving the representation ability of the network. To improve the segmentation performance of blurry smoke objects, we use Atrous convolutions with different rates and the Pyramid Pooling Module (PPM) to obtain contextual and abstract information. To effectively aggregate features, we propose a Pixel Alignment Module (PAM) to recalibrate the position of features and produce more powerful features. Compared with other excellent semantic segmentation algorithms, the proposed method consistently outperforms existing algorithms on the three synthetic smoke datasets and real smoke images. In addition, our method also achieves very good segmentation results on images captured from wild scenes with electric power transmission lines. However, our method is still not lightweight enough and requires a lot of computational resources. Compared to the existing transformer structure, its performance cannot achieve optimal results. In future work, we will focus on lightweight and Transformer structures to further improve accuracy.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

FZ was responsible for scheme designs and method requirements, GaW and YW collected and annotated training data, HP completed the testing experiments of the proposed model, GuW designed, trained the network and drafted the paper, and FY created the test dataset and revised the paper. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Major Scientific and Technological Projects of Yunnan Province (202202AD080010).

Conflict of interest

Authors FZ, GW, HP, and YW were employed by Electric Power Research Institute of Yunnan Electric Power Company.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Yuan F, Shi J, Xia X, Zhang L, Li S. Encoding pairwise Hamming distances of Local Binary Patterns for visual smoke recognition. *Computer Vis Image Understanding* (2019) 178:43–53. doi:10.1016/j.cviu.2018.10.008

2. Yuan F, Shi J, Xia X, Fang Y, Fang Z, Mei T. High-order local ternary patterns with locality preserving projection for smoke detection and image classification. *Inf Sci* (2016) 372:225–40. doi:10.1016/j.ins.2016.08.040

3. Yuan C, Liu Z, Zhang Y. Learning-based smoke detection for unmanned aerial vehicles applied to forest fire surveillance. *J Intell Robotic Syst* (2019) 93(1):337–49. doi:10.1007/s10846-018-0803-y

4. Mahmoud M, Ren H. Forest fire detection and identification using image processing and SVM. *J Inf Process Syst* (2019) 15(1):159–68. doi:10.3745/JIPS.01.0038

5. Tian H, Li W, Ogunbona P, Wang L. Detection and separation of smoke from single image frames. *IEEE Trans Image Process* (2017) 27(3):1164–77. doi:10.1109/tip.2017.2771499

6. Yuan F, Zhang L, Xia X, Huang Q, Li X. A gated recurrent network with dual classification assistance for smoke semantic segmentation. *IEEE Trans Image Process* (2021) 30:4409–22. doi:10.1109/tip.2021.3069318

7. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. *Int Conf Med image Comput computer-assisted intervention* (2015) 9351:234–41. doi:10.1007/978-3-319-24574-4_28

8. Chen L, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation (2017). Available at: https://arxiv.org/abs/1706.05587.

9. Zhang Z, Zhang C, Shen W, Yao C, Liu W, Bai X. Multi-oriented text detection with fully convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; June 2016; Las Vegas, NV, USA (2016). p. 4159–67. doi:10.1109/CVPR.2016.451

10. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; June 2018; Salt Lake City, UT, USA (2018). p. 7794–803. doi:10.1109/CVPR.2018.00813

11. Zhao H, Shi J, Qi X, Jia J. Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition; July 2017; Honolulu, HI, USA (2017). p. 6230–9. doi:10.1109/CVPR.2017.660

12. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Machine Intelligence* (2017) 39(4):640–51. doi:10.1109/TPAMI.2016.2572683

13. Badrinarayanan V, Kendall A, Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Tra*ns pattern Anal machine intelligence* (2017) 39(12):2481–95. doi:10.1109/TPAMI.2016.2644615

14. Wang L, Li D, Zhu Y, Tian L, Shan Y. Dual super-resolution learning for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 2020; Seattle, WA, USA (2020). p. 3774–83.

15. Chen L, Papandreou G, Kokkinos I, Yuile A. Semantic image segmentation with deep convolutional nets and fully connected crfs (2014). Available at: https://arxiv.org/abs/1412.7062.

16. Chen L, Papandreou G, Kokkinos I, Yuile A, Yuille AL. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans pattern Anal machine intelligence* (2017) 40(4):834–48. doi:10.1109/TPAMI.2017.2699184

17. Chen L, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proc Eur Conf Comput Vis* (2018) 11211:833–51. doi:10.1007/978-3-030-01234-2_49

18. Filonenko A, Hernández D, Jo K. Fast smoke detection for video surveillance using CUDA. *IEEE Trans Ind Inform* (2017) 14(2):725–33. doi:10.1109/TII.2017.2757457

19. Dimitropoulos K, Barmpoutis P, Grammalidis N. Higher order linear dynamical systems for smoke detection in video surveillance applications. *IEEE Trans Circuits Syst Video Tech* (2017) 27(5):1143–54. doi:10.1109/TCSVT.2016.2527340

20. Zhao Y. Candidate smoke region segmentation of fire video based on rough set theory. *J Electr Comp Eng* (2015) 2015:1–8. doi:10.1155/2015/280415

21. Tao C, Zhang J, Wang P. Smoke detection based on deep convolutional neural networks. In: Proceedings of the 2016 International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII); December 2016; Wuhan, China (2016). p. 150–3. doi:10.1109/ICIICII.2016.0045

22. Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks. *Assoc Comput Machinery* (2017) 60(6):84–90. doi:10.1145/3065386

23. Li X, Chen Z, Wu Q, Liu C. 3D parallel fully convolutional networks for real-time video wildfire smoke detection. *IEEE Trans Circuits Syst Video Tech* (2020) 30(1):89–103. doi:10.1109/TCSVT.2018.2889193

24. Yuan F, Zhang L, Xia X, Wan B, Huang Q, Li X. Deep smoke segmentation. *Neurocomputing* (2019) 357:248–60. doi:10.1016/j.neucom.2019.05.011

25. Yuan F, Zhang L, Xia X, Huang Q, Li X. A wave-shaped deep neural network for smoke density estimation. *IEEE Trans Image Process* (2019) 29:2301–13. doi:10.1109/TIP.2019.2946126

26. Dong Z, Yuan F, Xue X. Improved spatial and channel information based global smoke attention network. *J Beijing Univ Aeronautics Astronautics* (2022) 48(8):1471–9. doi:10.13700/j.bh.1001-5965.2021.0549

27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; June 2016; Las Vegas, NV, USA (2016). p. 770–8. doi:10.1109/CVPR.2016.90

28. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Han H, et al. Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision; October 2017; Venice, Italy (2017). p.764–73. doi:10.1109/ICCV.2017.89

29. Wang W, Shen J, Shao L. Video salient object detection via fully convolutional networks. *IEEE Trans Image Process* (2017) 27(1):38–49. doi:10.1109/TIP.2017.2754941

30. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans pattern Anal machine intelligence* (2017) 40(4):834–48. doi:10.1109/tpami.2017.2699184

31. Mehta S, Rastegari M, Caspi A, Shapiro L, Hajishirzi H. Espnet: efficient spatial pyramid of dilated convolutions for semantic segmentation. *Eur Conf Comput Vis* (2018) 11214:561–80. doi:10.1007/978-3-030-01249-6_34

32. Islam MA, Naha S, Rochan M, Bruce N, Wang Y. Label refinement network for coarse-to-fine semantic segmentation (2017). Available at: https://arxiv.org/abs/1703.00551.

33. Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. *Eur Conf Comput Vis* (2016) 9912:483–99. doi:10.1007/978-3-319-46484-8_29

34. Yuan F, Zhang L, Wan B, Xia X, Shi J. Convolutional neural networks based on multi-scale additive merging layers for visual smoke recognition. *Machine Vis Appl* (2019) 30:345–58. doi:10.1007/s00138-018-0990-3

35. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations; May 2021; Vienna, Austria (2021).

36. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; October 2021; Montreal, QC, Canada (2021). p. 9992–10002.

37. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez J, Luo P. SegFormer: simple and efficient design for semantic segmentation with transformers (2021). Available at: https://arxiv.org/abs/2105.15203.