2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Enhancing Fire and Smoke Segmentation with Cross-Attention Side-Adapter Network

Yuyang Deng[1], Xiying Luan[1], and Fan Zhong[1*]

[1] The School of Computer Science and Technology, Shandong University, Qingdao 266237, China
202200130167@mail.sdu.edu.cn

Abstract. Fire and smoke segmentation is crucial for disaster management and emergency response. Existing fire and smoke segmentation methods predominantly rely on conventional deep learning models such as U-Net. However, a key challenge in fire and smoke segmentation is the inherent uncertainty in target scale—segmentation targets can range from large-scale fire or smoke regions to minute initial flames or smoke particles. This multi-scale characteristic poses additional challenges, as traditional models often struggle to balance global feature extraction for large targets with fine-grained feature representation for small targets, leading to missed or inaccurate detections. Furthermore, existing fire segmentation datasets exhibit limited diversity, resulting in models trained with conventional methods that lack generalization ability in cross-domain applications. To overcome these limitations and enhance model performance, this study proposes an optimized Side Adapter Network (SAN) that integrates cross-attention mechanisms and a CrossViT architecture to improve feature extraction across different target scales. Specifically, the proposed approach employs cross-attention mechanisms to enhance information exchange between CLIP and the side network, while CrossViT effectively strengthens the side network's capability in capturing fine-grained image details. Experimental results demonstrate that, compared to traditional CNN and Transformer-based models, the optimized SAN achieves significant improvements in accuracy for fire and smoke detection and segmentation tasks. Moreover, due to its strong open-vocabulary semantic segmentation capability, the model exhibits robust generalization in cross-domain applications, enabling it to effectively handle complex environments and diverse fire scenarios.

Keywords: Fire And Smoke Image Segmentation, Open-Vocabulary Semantic Segmentation, Cross Attention, Side Adapter Network.

1 Introduction

Fire and smoke semantic segmentation plays a crucial role in disaster management, fire safety, and emergency response applications. However, training deep neural networks for fire and smoke segmentation requires pixel-level annotations for each image, which is a labor-intensive and cost-prohibitive process. Moreover, existing annotated datasets may be incomplete, and real-world fire incidents present highly complex and dynamic scenarios. As a result, models are likely to encounter previously unseen categories, such as red car taillights in Fig. 1, posing challenges in accurate segmentation. With limited annotated data, deep learning models may struggle to precisely delineate fire and smoke boundaries due to insufficient learning.

Fire and smoke sometimes occupy small regions within images, such as small flame in Fig. 1, making it challenging for traditional detection models to effectively capture these small objects' features. Their small scale may lead models to overlook these critical details, resulting in inaccurate segmentation.

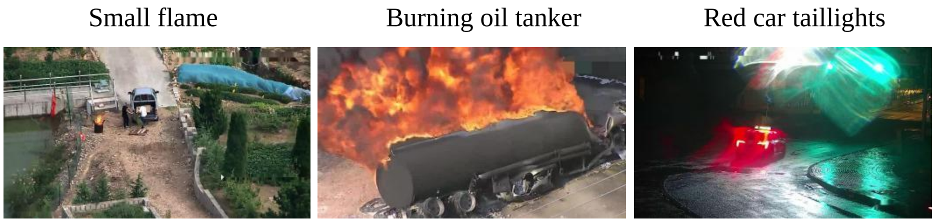

Fig. 1. Several different fire and smoke segmentation scenarios: A metal barrel is burning with an open flame. A overturned oil tanker is engulfed in intense flames. A fast-moving car with its headlights creating blurred light spots and glare.

To address the challenges of small object detection and cross-domain application in fire and smoke semantic segmentation, we propose CASAN (Cross-Attention Side Adapter Network), integrating the following techniques to enhance segmentation performance:

1) Since the Vision Transformer (ViT) [1] has limitations in capturing local features effectively, we propose using CrossViT [2] to replace the traditional ViT. CrossViT leverages cross-layer feature fusion and cross-attention mechanisms to enhance the detection accuracy of small objects. When handling small objects, CrossViT is more efficient in capturing fine-grained features, thereby improving the model's ability to segment small objects.

2) To enhance the model's generalization and cross-domain capabilities, we employ a Side Adapter Network, which leverages the features of pre-trained model CLIP [3]. This approach endows the model with strong open-vocabulary semantic segmentation capabilities and improved generalization, enabling it to recognize fire types that were not present in the training dataset. Consequently, the model is well-suited for fire detection tasks across diverse scenarios.

3) We hypothesize a strong correlation between the query tokens used for generating mask proposals and the frozen CLIP CLS token, which encapsulates global information and text alignment features. Therefore, we introduce a cross-attention mechanism to facilitate information exchange between these two tokens, thereby fusing the features from the CLIP model. This mechanism helps the side-adaptation network better capture both global and local features of fire and smoke, enhancing the model’s ability to segment these objects with greater precision.

The integration of these techniques enables CASAN to achieve high-precision and robust performance in complex fire and smoke segmentation tasks. In particular, it significantly enhances segmentation accuracy for small object detection. Moreover, the model exhibits strong transferability, demonstrating excellent generalization capability when encountering objects not previously learned from fire and smoke datasets.

2 Related works

In recent years, deep learning-based image semantic segmentation techniques have garnered widespread attention. FCN (Fully Convolutional Network), proposed by Jonathan Long et al. [4] ,removes the fully connected layers from traditional networks and uses fully convolutional layers to effectively preserve the spatial information of image features. It also employs deconvolution operations to restore object contours in images. The U-Net network, introduced by Olaf Roneberger et al. [5] ,utilizes a symmetric encoder-decoder structure, skip connections, and deconvolution operations to perform high-precision pixel-level segmentation, making it particularly suitable for medical image segmentation and other tasks involving fine structures. The DeepLab network, proposed by Chen et al. [6] , fully leverages dilated convolution to increase the receptive field and replace downsampling operations, thereby maximizing the preservation of spatial features, particularly excelling in small object segmentation tasks. DeepLabv3,proposed by Chen et al. [7] further enhances semantic segmentation performance with innovations like dilated convolution and Atrous Spatial Pyramid Pooling (ASPP). SegFormer, proposed by Xie et al. [8] a model that combines Transformer and Convolutional Neural Networks (CNN), replaces traditional positional encodings with Mix-FFN, uses a fully MLP decoder, and a hierarchical transformer encoder. It is particularly effective at handling objects of varying scales, improving small object segmentation accuracy. SegNext,proposed by Guo et al. [9] a model that combines CNNs and Transformers, further enhances semantic segmentation performance in complex scenarios by merging multi-scale information and efficient feature extraction, particularly demonstrating strong generalization capabilities on large-scale datasets.

In recent years, large-scale vision-language pre-training models such as CLIP and ALIGN have achieved significant success, leading many open-vocabulary semantic segmentation methods to leverage their capabilities. These vision-language models, by jointly learning image and text embedding spaces, enable the model to understand the correspondence between images and text.Building upon this, Muyang Yi proposed a two-stage framework based on the CLIP model called SimSeg [10] . In the first stage, SimSeg utilizes a mask generator to produce masked image crops, representing potential objects or regions within the image. In the second stage, SimSeg employs the CLIP model to perform mask recognition. Notably, the mask generator in SimSeg operates independently of the CLIP model, preventing it from utilizing CLIP's features.MaskCLIP enhances this approach by progressively refining the initial masks using the CLIP encoder and directly applying CLIP's image-text alignment capabilities to the mask generation process. MaskCLIP,proposed by Dong et al. [11] , its CLIP encoder is employed to refine the initially generated masks, rather than solely relying on the mask generator's output.OVSeg [12] , on the other hand, departs from the two-stage framework by directly leveraging CLIP's image-text alignment capabilities without the need for an additional mask generator. By utilizing both the image encoder and text encoder of CLIP, OVSeg efficiently performs open-vocabulary semantic segmentation, employing the joint features of images and text for the segmentation task.In contrast to these methods, SAN [13] adopts a different approach by employing a side adapter network to generate mask proposals, rather than directly producing the final masks. These mask proposals represent potential areas of interest within the image. During training, the mask proposals are combined with CLIP model features and guided by attention bias, ultimately achieving class recognition for these regions.

In the field of flame and smoke segmentation, conventional methods [14] have primarily focused on extracting the texture and fine details of flames. However, these approaches heavily rely on handcrafted features, which inherently limit their generalization capability.In recent years, several innovative methods and models have been proposed in the field of flame and smoke segmentation. Yuan et al.[15] (2018) introduced a dual-path encoder-decoder network based on Fully Convolutional Networks (FCN) to infer high-quality segmentation masks from blurry smoke images. Yang et al. [16](2023) proposed an effective smoke segmentation method combining KNN background modeling and the SegFormer semantic segmentation algorithm. However, these methods lack open-vocabulary capabilities and exhibit slightly lower accuracy.However, these methods largely rely on a closed vocabulary, resulting in limited generalization capability and poor adaptability to complex environments.

Building upon the core ideas of the aforementioned studies, we propose CASAN, designed to address the challenges of insufficient multi-scale target modeling and poor cross-domain generalization in fire and smoke segmentation.

3 Model Structure

3.1 Cross Attention Side Adapter Network

As shown in Fig. 2,the first part of the model utilizes CLIP to generate logits proposals for image-text alignment. Both the image and text are separately fed into the CLIP model, where the image encoder extracts features in the initial layers. At the ninth layer, the CLS token is duplicated to generate multiple SLS tokens. These SLS tokens, along with the CLIP CLS token and visual tokens, undergo cross-attention updates. Furthermore, the visual tokens and CLS token in the CLIP layer are updated under the influence of attention biases from the CrossViT side network, gradually incorporating fine-grained segmentation features from the side network. Finally, the SLS tokens are multiplied with the features obtained from the CLIP text encoder to generate logits proposals. These proposals represent the degree of alignment between each pixel in the segmented image and the textual description.

By preserving the pretrained parameters of CLIP and leveraging its cross-modal alignment capability, the model maintains its open-vocabulary segmentation ability. Simultaneously, through SLS tokens and Cross-Attention mechanisms, it refines fine-grained feature extraction, ultimately generating logits proposals. This approach enables the model to effectively segment objects that were not present in the training set while maintaining robust performance across diverse environments.

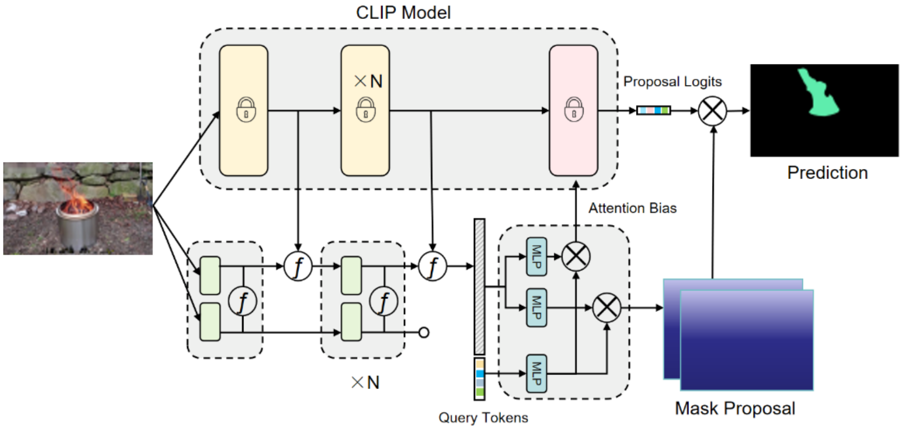

Fig. 2. Cross Attention Side Adapter Network

The side network of the model adopts the CrossViT architecture, consisting of two Vision Transformers (ViTs) with different feature map sizes. A larger 16x16 patch is employed as the backbone network, primarily for capturing the macro-semantic information within the image. Larger patches help the model understand the global structure of the image and the distribution of major objects, thus extracting high-level semantic features. Additionally, this approach efficiently facilitates interaction with CLIP features. To address the limitations of larger patches in capturing small objects and fine details, smaller 10x10 patches are also used. These smaller patches enable the model to more finely capture the textures, details, and local features of small objects, thereby improving detection and segmentation accuracy for small objects. The side network consists of 8 MultiScaleBlocks, and after each block, cross-attention is applied to fuse the information from both networks. Finally, the tokens from the backbone network are output for further utilization.

3.2 Multi-Scale Feature Extraction

In traditional ViT, an image is typically divided into fixed-size patches, where each patch is flattened into a vector and then fed into the Transformer layers for processing. This approach effectively captures the global information of the image; however, it may lack sensitivity to small objects or fine details, as larger patches tend to lose local information of the image. To address this limitation, we employ CrossViT. Below is the process of feature fusion in CrossViT.

The network architecture of CrossViT consists of two Vision Transformers with different patch sizes. Suppose we have an image, and we process the image using two different patch sizes. The main branch uses a patch size of 16 with an embedding dimension of 240, while the side branch uses a patch size of 10 with an embedding dimension of

120.These patches are then mapped to their respective feature representations through linear transformations.

$$X_{16} = MW_1\,, \quad X_{10} = MW_2 \tag{1}$$

Subsequently, we add 100 query tokens to the tokens obtained from the main branch. These query tokens are used to further strengthen the model's attention mechanism and preserve fine-grained features, which are essential for generating detailed mask proposals. The side branch remains unaffected by this process.$\mathrm{X}_{16} \in \mathrm{R}^{\mathrm{N}\times\mathrm{D}}$,$\mathrm{Q}_{\mathrm{new}} \in \mathrm{R}^{100\times\mathrm{D}}$

$$X_{main} = [\mathrm{X}_{16} \parallel \mathrm{Q}_{\mathrm{new}}] \tag{2}$$

After both the main branch and the side branch are processed through ViT blocks, the output data from both branches is fused using cross-attention. To illustrate this using the main branch, the query tokens from the main branch are first mapped to ensure their dimensions match those of the output from the side network. These mapped query tokens are then concatenated with the visual tokens output from the side branch.

$$\mathrm{X}^{\mathrm{l}} = \left[\mathrm{f}\left(\mathrm{X}_{\mathrm{qry}}^{\mathrm{m}}\right) \parallel \mathrm{X}_{\mathrm{vis}}^{\mathrm{s}}\right] \tag{3}$$

Subsequently, cross-attention is applied to fuse the information, and CA can be represented as follows:

$$Q = X_{qry}^{m} W_q, \quad K = X^l W_k, \quad V = X^l W_v \tag{3}$$

$$A = softmax\left(QK^T/\sqrt{C/h}\right), \quad Z = g\left(A + f\left(X_{qry}^{m}\right)\right) \tag{4}$$

Similar to self-attention, cross-attention employs a multi-head mechanism, where C is the dimensionality, h is the number of heads. A residual connection is used to add $f\left(X_{qry}^{m}\right)$ to A, and the result is passed through function g to map it back to the original dimensionality. Z represents the query tokens of the main branch after fusion.

Finally, the obtained query tokens and visual tokens are projected through two separate MLP layers and then multiplied to generate the mask proposal. The process is similar to the attention bias mechanism, and the resulting attention bias is used to update the SLS tokens across multiple CLIP blocks.

3.3 Cross Attention Fusion Layer

In the previous Side Adapter Network, only the visual tokens from CLIP and the side network were fused. However, given that query tokens are essential for generating mask proposals, we argue that they should also incorporate a certain degree of text-aligned features. To achieve this, we introduce a cross-attention mechanism that facilitates interaction between the sIs token of CLIP and the query tokens of the CASAN main branch. Through this process, the sIs token of CLIP influences and integrates with the query tokens of CASAN, enabling effective information transfer. As a result, the query tokens can simultaneously capture both the overall text-aligned features from CLIP and the essential visual information required for segmentation.

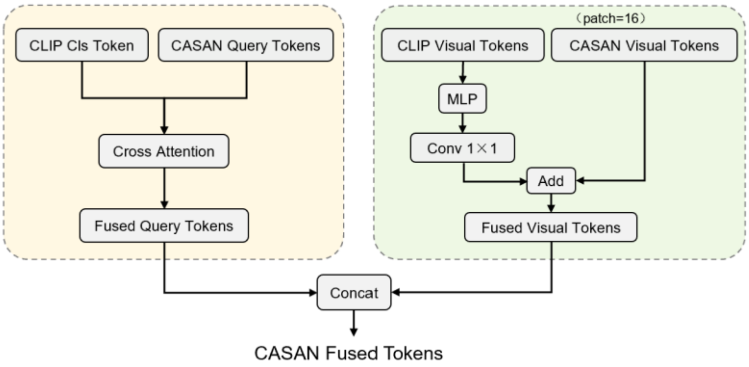

Fig. 3. The fusion of feature representations between the primary pathways of CLIP and CASAN.

The CLIP visual tokens are passed through an MLP layer to align their dimensions, followed by a 1x1 convolution to further extract visual features. These visual features are then added and fused with the CASAN visual tokens, effectively combining the visual features from both the CLIP and CASAN main paths. Finally, the fused query tokens and fused visual tokens are concatenated. The concatenated tokens, which include information from both CLIP and CASAN, capture both visual and query features. This fusion of features allows the model to better capture critical image details and semantic information, thereby enhancing the accuracy of detection and segmentation tasks.

3.4 Loss Function

We employ a mixed loss function consisting of Dice Loss, Binary Cross-Entropy Loss, and Cross-Entropy Loss.

$$\mathrm{L} = \lambda_1 \mathrm{L}_{\mathrm{dice}} + \lambda_2 \mathrm{L}_{\mathrm{bce}} + \lambda_3 \mathrm{L}_{\mathrm{cls}} \tag{5}$$

Dice Loss is utilized to optimize object segmentation and small object detection, while Binary Cross-Entropy Loss aids the model in making accurate pixel-level classifications for segmentation tasks. Cross-Entropy Loss is used to optimize the category prediction. By weighted combination of these three loss functions, we achieve a balance between mask generation and classification tasks, thereby enabling the model to perform effectively in both segmentation and recognition tasks, leading to improved overall performance.

4 Experimental Analysis and Results

The experiments are conducted in the following environment: Ubuntu 20.04 operating system, 12 vCPUs Intel® Xeon® Platinum 8352V processor at 2.10GHz, 90 GB of RAM, with development carried out using PyTorch and Python. The training process is accelerated using an NVIDIA 4080s GPU.

4.1 Dataset and Training Settings

We conduct experiments on four datasets: COCOStuff [17] , PASCAL VOC 2012 Dataset [18] , PA-59 Dataset [19], and the "Jishi" Flame and Smoke Detection Dataset.

COCOStuff is an extended version of the popular COCO dataset, specifically designed for open semantic segmentation tasks. It includes over 170,000 annotated images, labeled with 80 common object categories and multiple "Stuff" categories, which represent background regions such as sky, grass, and ground.

PASCAL VOC 2012 is a classic image segmentation and object detection dataset widely used in object recognition and image segmentation research. It contains approximately 11,000 images, annotated with 20 common object categories.

PA-59 is a dataset designed specifically for open semantic segmentation tasks, comprising 59 categories derived from various objects and backgrounds in daily life, covering common object categories and natural environment backgrounds.

"Jishi" Flame and Smoke Segmentation Dataset is designed for fire and smoke detection and segmentation tasks. It consists of over 80,000 images covering five categories: fire, yellow smoke, black smoke, white smoke, and background. However, the dataset exhibits significant class imbalance and contains certain annotation errors, posing a substantial challenge for model generalization.

We conducted ablation experiments on publicly available datasets such as COCO-Stuff to validate the effectiveness of our method. For the fire and smoke segmentation task, we trained and evaluated our model on the 'Jishi' Fire and Smoke Segmentation dataset.

4.2 Performance Comparison of Different Models

Models presents a comparison of segmentation accuracy among different models for the fire and smoke segmentation task, evaluating categories including fire, black smoke, white smoke, yellow smoke, and the overall mean Intersection over Union (mIoU).

Our model, CASAN, achieves the highest segmentation accuracy across all categories. Specifically, the segmentation accuracy for the fire category reaches 76.97, while the segmentation accuracy for black smoke, white smoke, and yellow smoke all exceed 63, significantly outperforming other benchmark models. Notably, CASAN achieves an mIoU of 71.69, which is a substantial improvement over Deeplabv3+ (51.82), SegFormer (66.96), and SegNext (69.28).

Table 1. Segmentation Accuracy of CASAN Compared to Other Models

Methods	Fire	Smoke_black	Smoke_white	Smoke_yellow	miou
Deeplabv3+	58.54	32.08	57.44	59.21	51.82
Segformer	74.95	53.8	67.44	71.66	66.96
SegNext	73.21	60.8	70.9	72.21	69.28
CASAN	**76.97**	**63.01**	**72.94**	**73.86**	**71.69**

Fig. 4. demonstrates that CASAN outperforms SegFormer in flame and smoke segmentation tasks. The incorporation of the CrossViT module and the feature fusion module effectively enhances the detection capability for objects of different scales, leading to more precise segmentation results for both small and large flames. Furthermore, the frozen CLIP model enables CASAN to achieve superior adaptability and robustness in complex environments compared to the SegFormer model trained on a closed dataset. This allows CASAN to accurately segment previously unseen objects with higher precision.

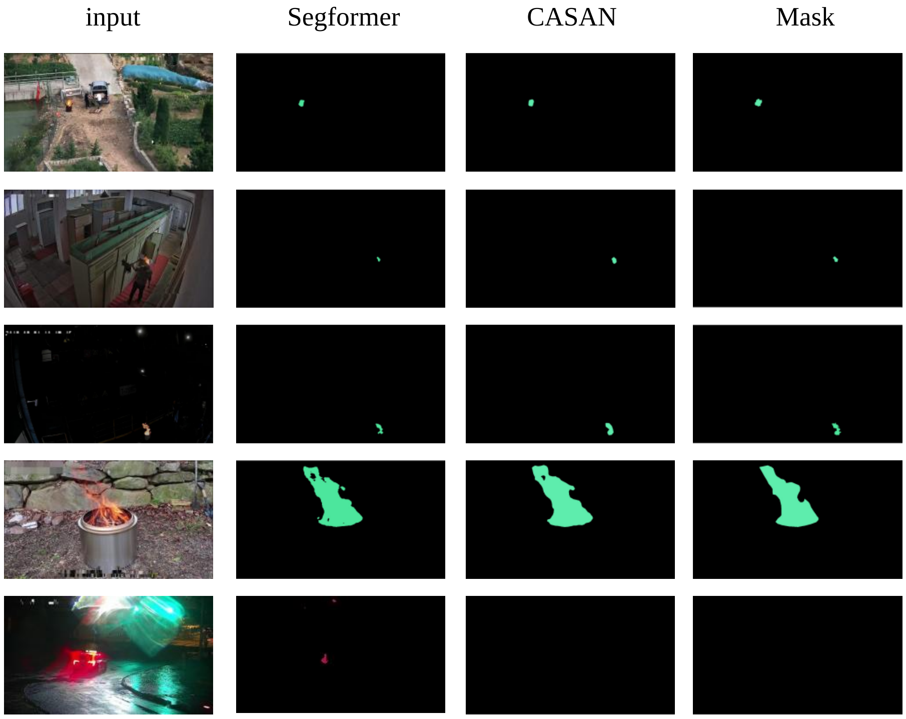

Fig. 4. Comparison of Segmentation Results from Different Models

These results demonstrate that CASAN exhibits superior segmentation capability in fire and smoke detection tasks, enabling more precise identification of different types of smoke and fire regions. Given that the dataset itself presents challenges such as class imbalance and annotation errors, CASAN still achieves leading segmentation performance, proving its strong generalization ability and robustness.

4.3 Ablation Study

To validate the effectiveness of our experiments, we conducted evaluations on public datasets such as COCO-Stuff, PASCAL VOC 2012, and PC-59. All experiments were based on the CLIP ViT-B/16 model. Our baseline utilized the standard SAN model, which includes only fully connected and convolutional layers in its fusion layer without additional enhancement modules. Subsequent experiments incorporated enhanced feature fusion modules (fuse) and CrossViT structures into the baseline. All models were trained on the COCO-Stuff dataset and subsequently tested on the test sets of COCO-Stuff, PASCAL VOC 2012, and PC-59.

Table. 2. Ablation Study Results

Methods	VL-Model	Training Dateset	COCO	VOC	PC-59
Original SAN	CLIP VIT-B/16	COCO	41.6	93.8	53.5
Ours(fuse)	CLIP VIT-B/16	COCO	41.7	**94.6**	53.2
Ours(crossvit)	CLIP VIT-B/16	COCO	41.8	94.0	**53.6**
Ours(fuse+crossvit)	CLIP VIT-B/16	COCO	**41.9**	94.4	**53.7**

These results ,as shown in Table. 2, Results indicate that the introduction of the improved feature fusion module (fuse) and the CrossViT structure has varying impacts on model performance. The enhanced feature f usion module led to slight improvements on the COCO and VOC datasets but resulted in a minor decline on PC-59. We hypothesize that the more complex feature fusion module enables the model to extract features more comprehensively; however, for more intricate scenes, its generalization may be slightly reduced. In contrast, the incorporation of CrossViT aids the model in better extracting multi-scale features, leading to modest performance gains across the tested datasets. Combining both approaches allows the model to retain its original generalization capabilities while effectively capturing multi-scale features, resulting in more precise segmentation.

Building upon this, Table .3 further validates the effectiveness of our model, particularly in small object segmentation. The results demonstrate that CASAN consistently achieves higher segmentation accuracy compared to the Baseline, highlighting its superior ability to capture fine-grained details in small objects. The data in Table 3 is based on small object categories selected from the COCO-Stuff 171K dataset, ensuring a targeted evaluation. This further confirms that our approach enhances multi-scale feature extraction, ultimately leading to more effective segmentation of small objects.

Table .3. Small object segmentation results

Methods	Tie	Baseball bat	Sports ball	Knife	Spoon	Baseball gloves
Original SAN	9.52	37.01	**58.76**	15.21	24.07	50.29
CASAN	**9.89**	**39.46**	56.87	**18.69**	**25.06**	**54.49**

By integrating an enhanced feature fusion module and the CrossViT architecture, our model's capacity for multi-scale feature extraction has been significantly improved. This advancement enables the model to excel in smoke and flame segmentation tasks, facilitating more precise identification and delineation of these regions. Furthermore, the model demonstrates robust open-vocabulary semantic segmentation capabilities, allowing it to effectively handle previously unseen categories and thereby enhancing its generalization performance. In contrast to traditional models, our approach exhibits greater resilience to data discrepancies absent from the training set, maintaining accurate segmentation of smoke and flames even in complex and dynamic environments.

5 Conclusion

This study proposes an optimized Side-Adaptive Network (SAN) model, which incorporates a cross-attention mechanism and a CrossViT architecture to significantly enhance the performance of flame and smoke detection and segmentation. Compared to conventional Convolutional Neural Networks (CNNs) and Transformer-based models, the optimized SAN model demonstrates superior performance in complex backgrounds and small-object scenarios, particularly in its adaptability to underrepresented or previously unseen objects.

This enhanced information interaction not only accelerates the training process but also improves the model's capability in fine-grained segmentation. Moreover, by utilizing a frozen CLIP model, the proposed approach retains open-set semantic segmentation capabilities, demonstrating strong resistance to interference when handling underrepresented or previously unseen objects in complex environments, thereby enhancing the model's robustness. As evidenced by Table 1 and Fig. 4, integrating open-set semantic segmentation into flame and smoke segmentation is both a rational and necessary enhancement to this task.

In addition, the adoption of the CrossViT architecture further improves the side-adaptive network’s ability to capture fine details. By integrating multi-scale feature fusion and cross-domain information exchange, CrossViT enhances the model’s ability to extract both global and local information, leading to more precise segmentation of flames and smoke of varying sizes. Experimental results demonstrate that, compared to conventional deep learning approaches, the optimized model achieves superior accuracy and robustness in flame and smoke segmentation tasks.

Acknowledgments. This work is supported by National Key Research and Development Program of China (No. 2022YFB2603303).

References

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020, October 22). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv.org. https://arxiv.org/abs/2010.11929

[2] Chen, C., Fan, Q., & Panda, R. (2021, March 27). CROSSVIT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. arXiv.org. https://arxiv.org/abs/2103.14899v2

[3] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021, February 26). Learning transferable visual models from natural language supervision. arXiv.org. https://arxiv.org/abs/2103.00020

[4] Long, J., Shelhamer, E., & Darrell, T. (2014, November 14). Fully convolutional networks for semantic segmentation. arXiv.org. https://arxiv.org/abs/1411.4038

[5] Ronneberger, O., Fischer, P., & Brox, T. (2015, May 18). U-NET: Convolutional Networks for Biomedical Image Segmentation.

[6] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2016, June 2). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. arXiv.org. https://arxiv.org/abs/1606.00915

[7] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., & Google Inc. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation [Journal-article]. https://arxiv.org/pdf/1802.02611

[8] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P., The University of Hong Kong, Nanjing University, NVIDIA, & Caltech. (2021). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In 35th Conference on Neural Information Processing Systems (NeurIPS 2021) [Journal-article]. https://proceedings.neurips.cc/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf

[9] Guo, M., Lu, C., Hou, Q., Liu, Z., Cheng, M., & Hu, S. (2022, September 18). SegNEXT: Rethinking Convolutional Attention Design for Semantic Segmentation. arXiv.org. https://arxiv.org/abs/2209.08575

[10] A simple framework for Text-Supervised semantic segmentation. (2023, June 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/10203335

[11] Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., Wen, F., & Yu, N. (2022, August 25). MaskCLIP: Masked Self-Distillation Advances Contrastive Language-Image Pretraining. arXiv.org. https://arxiv.org/abs/2208.12262

[12] Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., & Marculescu, D. (2022, October 9). Open-Vocabulary Semantic Segmentation with Mask-adapted CLIP. arXiv.org. https://arxiv.org/abs/2210.04150

[13] Xu, M., Zhang, Z., Wei, F., Hu, H., & Bai, X. (2023, February 23). Side adapter network for Open-Vocabulary Semantic segmentation. arXiv.org. https://arxiv.org/abs/2302.12242

[14] Forest Fire Detection and Identification using Image Processing and SVM (Mubarak Adam Ishag Mahmoud & Honge Ren, 2019)

[15] Yuan, F., Zhang, L., Xia, X., Wan, B., Huang, Q., & Li, X. (2018, September 4). Deep smoke segmentation. arXiv.org. https://arxiv.org/abs/1809.00774

[16] Yang, J., Liu, H., Lv, M., Wang, M., & Li, Q. (2023b). Effective smoke segmentation using KNN Background Modeling and SegFormer. In Lecture notes in electrical engineering (pp. 371–377). https://doi.org/10.1007/978-981-99-4882-6_52

[17] Caesar, H., Uijlings, J., & Ferrari, V. (2016, December 12). COCO-Stuff: thing and stuff classes in context. arXiv.org. https://arxiv.org/abs/1612.03716

[18] The role of context for object detection and semantic segmentation in the wild. (2014, June 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/6909514

[19] Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2014). The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision, 111(1), 98–136. https://doi.org/10.1007/s11263-014-0733-5