CLIP Guided UNet for Smoke Segmentation

Xingying Chu[1][0009-0001-7002-3932] and Jinhua Xu[2][0000-0002-8450-3144]

School of Computer Science and Technology, East China Normal University, Shanghai 200062, China
jhxu@cs.ecnu.edu.cn

Abstract. Smoke segmentation aims to classify pixels with the smoke label, which is a downstream task of semantic segmentation. Convolutional neural networks (CNNs) have made great progress in semantic segmentation. UNet is a widely used CNN-based semantic segmentation model. However, the conventional CNNs are inherently limited to local receptive fields that only provide short-range contextual information. Pretrained Vision-Language Models (VLMs) such as CLIP have learned rich semantics from web-scale image-text pairs. Inspired by this, we propose a novel CLIP guided UNet framework (CGUnet) for smoke segmentation, which merits the global and rich context of CLIP and the precise localization of UNet. Specifically, we design a CLIP guided cross attention (CGCA) module, in which the CLIP feature is used as the query, and the visual features of UNet as the Key and the Value. We conduct experiments on two public smoke segmentation datasets. Our method achieves SOTA results on both datasets, outperforming other methods.

Keywords: Smoke segmentation, UNet, CLIP, Visual language model, semantic segmentation.

1 Introduction

Smoke detection and segmentation can provide timely warnings of fire hazards and enhance disaster response. Deep learning methods have achieved great success in semantic segmentation. However, due to the non-rigid morphology and semi-transparent nature of smoke, smoke segmentation remains a challenging task.

Convolutional neural networks (CNNs) have been intensively studied for semantic segmentation. Most of the CNN based segmentation models use some kinds of encoder-decoder architecture. The decoder comprises deconvolution and unpooling layers, which identify pixel-wise class labels and predict segmentation masks [1], [2]. The encoder-decoder model UNet [2] was proposed for efficiently segmenting biological microscopy images. Various extensions of UNet have been developed for different kinds of images and problem domains. However, the CNN models do not account for global context information. To exploit more context, several approaches incorporate into CNN architectures probabilistic graphical models, such as Conditional Random Fields (CRFs) and Markov Random Fields (MRFs) [3]. Some works exploit the attention mechanism

to capture long-range dependencies [4], [5]. Fu et al. [4] proposed a dual attention network for scene segmentation that can capture rich contextual dependencies based on the self-attention mechanism. Huang et al. [5] proposed a criss-cross attention module to capture contextual information from full- image dependencies in an efficient way. Jing et al. [6] proposes a dual-branch segmentation model, named SmokeSeger, which couples a transformer branch and a CNN branch to enhance the representation of both global and local features. Zhang et al. [7] propose a Smoke-Aware Global-Interactive Non-local Network (SAGINN) for smoke segmentation, which harness the power of both convolution and transformer to capture local and global information simultaneously. In [8], Yuan et al design an Attention Convolutional GRU module (Att-ConvGRU) to learn the long-range context dependence of features.

Vision-Language Models (VLMs) have been investigated recently, which learn rich vision-language correlation from web-scale image-text pairs that are almost infinitely available on the Internet. The pre-trained VLMs have been applied to downstream visual recognition tasks without fine-tuning or via transfer learning and knowledge distillation [9]. The contrastive language-image pre-training model (CLIP) [10] employs an image-text contrastive objective and learns by pulling the paired images and texts close and pushing others faraway in the embedding space. It aligns the image and texts on a global level, therefore captures rich global context, but loses the position information. Lüddecke et al employ the pre-trained CLIP model as a backbone and extend with a transformer-based decoder that enables dense prediction [11].

Inspired by the rich semantics learned by CLIP and the localization ability of UNet, we propose to use CLIP to provide the global context and guide the UNet for smoke segmentation. We employ the cross-attention mechanism to integrate the global context from the CLIP into the UNet decoder.

Our main contributions include two aspects. First, we propose a CLIP guided UNet framework for smoke segmentation. Second, we conduct experiments on two public smoke datasets and achieve SOTA results.

2 Related Work

2.1 Semantic Segmentation

Semantic segmentation performs pixel-level labeling with a set of object categories (e.g., human, car, tree, sky) for all image pixels. Deep learning models have yielded remarkable performance improvements in image segmentation [12]. Long et al. [13] proposed Fully Convolutional Networks (FCNs) for semantic image segmentation. The original FCN approach suffers from two main drawbacks including the reduced feature resolution that loses detailed spatial information and the small effective receptive field that fails to capture long-range dependencies. Some methods raise the resolution of feature maps for improving the spatial precision through dilated convolutions [3], decoder network [1], [2]. HRNet [14] maintains high-resolution representations through the encoding process by connecting the high-to-low resolution convolution streams in parallel. Some methods have been proposed to introduce useful contextual information to benefit the semantic segmentation task. Chen et al. [3] proposed an atrous spatial

pyramid pooling module with multi-scale dilation convolutions for contextual information aggregation.

Transformers have also been applied for semantic segmentation [15]. TransUNet [16] merits both Transformers and UNet for medical image segmentation, in which transformer layers encode tokenized features from the backbone for extracting global contexts. Cheng et al. [17] proposes a universal architecture capable of addressing all image segmentation tasks including panoptic, instance or semantic segmentation. It consists of a backbone feature extractor, a pixel decoder and a Transformer decoder. Its key component masked attention of the transformer decoder extracts localized features by constraining cross-attention within predicted mask regions. Kirillov et al. [18] proposed a Segment Anything Model (SAM), which has become a powerful foundational model for segmentation and has been widely applied across various domains.

2.2 Smoke Segmentation

Smoke segmentation is a downstream task of semantic segmentation. Semantic segmentation methods can be applied to smoke segmentation through fine-tuning on smoke datasets. However, smoke segmentation is still a challenging task due to inconspicuousness of small smoke, highly complicated texture by blending semi-transparent smoke and different background, multiple scales of smoke at different evolving stages, and disturbance of smoke-like objects, such as haze and clouds [8]. Yuan et al. [19] proposed a Wave-shaped deep neural Network (W-Net) for smoke density estimation. Li et al. [20] propose a Frequency-Space Interaction with Hierarchical Aggregation Network (FSIHAN) for smoke segmentation. Yao et al. [21] propose prototype scatter learning (PSL), which improves smoke segmentation through unified optimization of feature extractors and prototypes. Yao et al. [22] propose a Focus and Separation Network (FoSp), in which a focus module guides low-resolution and high-resolution features towards mid-resolution to locate and determine the scope of smoke, and a separation module separates smoke images into a pure smoke foreground and a smoke-free background. Cao et al. [23] introduce the SAM-powered Density-Aware Progressive Smoke Segmentation method (DenSiSeg). For smoke regions, smoke mask labels are refined into fine-grained density information using a per-pixel metric. For background regions, knowledge transfer based on the vision foundation model SAM is employed to improve the representation of diverse background regions, which in turn facilitates more reliable learning of smoke representation.

3 Our Method

In this section, we first introduce the overall framework of the network and then introduce the cross-attention mechanism which uses CLIP image features to guide the visual features of UNet for smoke segmentation.

3.1 Architecture

The overall architecture is shown on Fig.1. It consists of a UNet image encoder, a CLIP image encoder and a UNet image decoder.

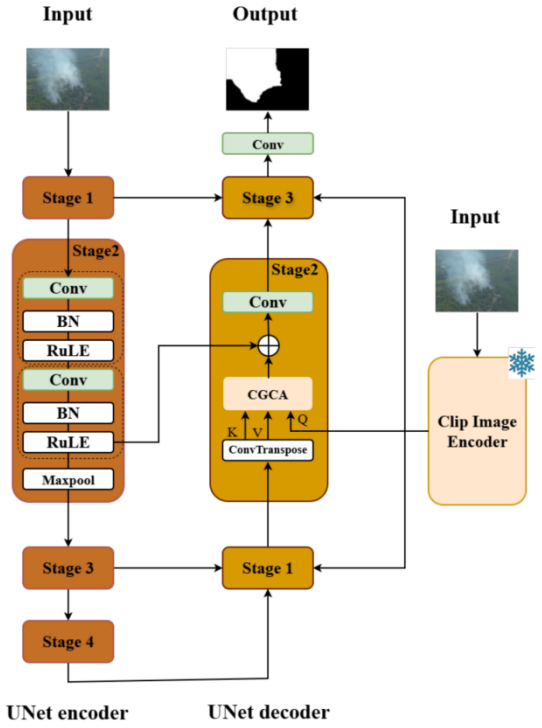

Fig. 1. Overview of the proposed method.

Encoders. In the image encoding stage, we employ two parallel processing pathways, a UNet encoder and a CLIP image encoder. Given an input smoke image $I \in \mathbb{R}^{H\times W\times 3}$, we employ a pre-trained CLIP image encoder (ViT-B/32) to extract the global semantic features. During the training process, the weights of the CLIP encoder remain frozen and do not participate in backpropagation updates. A global feature vector $\mathbf{F}_{\text{clip}} \in \mathbb{R}^{D}$ is generated through the CLIP image encoder. Simultaneously, the image I is fed into a customized UNet encoder. This encoder consists of four consecutive stages, and each of them contains two Conv-BN- ReLU blocks. To obtain multi-scale feature representations, max-pooling is applied at each stage with down-sampling factors of 2. Ultimately, the encoder outputs four feature maps with different resolutions, denoted as $\mathbf{F}_1^e, \mathbf{F}_2^e, \mathbf{F}_3^e, \mathbf{F}_4^e$, corresponding to the original input scale, 1/2 scale, 1/4 scale, and 1/8 scale, respectively.

Decoder. As illustrated in Fig.1, the decoder consists of three stages. Each stage comprises the following operations: transposed convolution for up-sampling, a CLIP-

guided cross-attention (CGCA) mechanism which integrates the up-sampled features with the global CLIP semantic features, a skip connection which concatenates the enhanced features with the corresponding encoder-level feature maps along the channel dimension. Finally, a convolutional layer is employed to fuse the concatenated features. This design effectively combines local visual details with global semantic information to enhance feature representation capabilities. After three decoding stages, a 1*1 convolutional layer is used to generate the final segmentation mask.

Clip Guided Cross-Attention. We propose a novel CLIP-Guided Cross-Attention (CGCA) module to enhance visual feature maps using semantic guidance from the CLIP features. The CGCA module is shown on Fig.2. Different from the traditional attention mechanism, this CGCA module uses the visual features from the pre-trained CLIP encoder as the guidance signal, interacts with the visual features from the UNet decoder through the cross-attention mechanism, and uses a gating mechanism to adaptively fuse the two features.

Given the visual feature map from the UNet decoder at the i-th stage $\mathbf{F}_i^d \in \mathbb{R}^{B\times C_i\times H_i\times W_i}, i \in \{1,2,3\}$ and CLIP image embeddings $\mathbf{F}_{\text{clip}} \in \mathbb{R}^{B\times D}$, where B is the batch size of the images, C_i is the number of channels, H_i and W_i are spatial dimensions, and D is the CLIP feature dimension. First, the CLIP embedding $\mathbf{F}_{\text{clip}}$ is projected into a query vector $\mathbf{Q}_i \in \mathbb{R}^{B\times C_i}$ via linear transformation.

$$\mathbf{Q}_i = \mathbf{W}_q \mathbf{F}_{\text{clip}}. \tag{1}$$

The visual feature map $\mathbf{F}_i^d$ is projected into the key **K** and the value **V** through a 1*1 convolution.

$$[\mathbf{K_i}, \mathbf{V_i}] = \text{Reshape}\left(\text{Conv}\left(\mathbf{F}_i^d\right)\right). \tag{2}$$

Then we have the attended feature map $\mathbf{F_i^{att}} \in \mathbb{R}^{B\times C_i\times H_i\times W_i}$:

$$\boldsymbol{F_i^{att}} = Expand\left(Reshape\left(CrossAtt(Q_i, K_i, V_i)\right)\right). \tag{3}$$

Here the cross attention is calculated as follows:

$$CrossAtt(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{C}}\right)V. \tag{4}$$

Subsequently, a gating mechanism generates a spatial attention map $\mathbf{G_i} \in \mathbb{R}^{1\times H_i\times W_i}$, which is derived from the visual feature map via a 1*1 convolution followed by a sigmoid activation function.

$$\mathbf{G_i} = \sigma\left(\text{Conv}\left(\mathbf{F}_i^d\right)\right). \tag{5}$$

This gating map adaptively controls the injection strength of semantic features. Finally, the original visual features are combined with the gated attention features through addition, yielding the enhanced visual feature map.

$$\mathbf{F}_i^{\text{fused}} = \mathbf{F_i^d} + \mathbf{G_i} \odot \mathbf{F_i^{att}}. \tag{6}$$

where $\odot$ denotes element-wise multiplication.

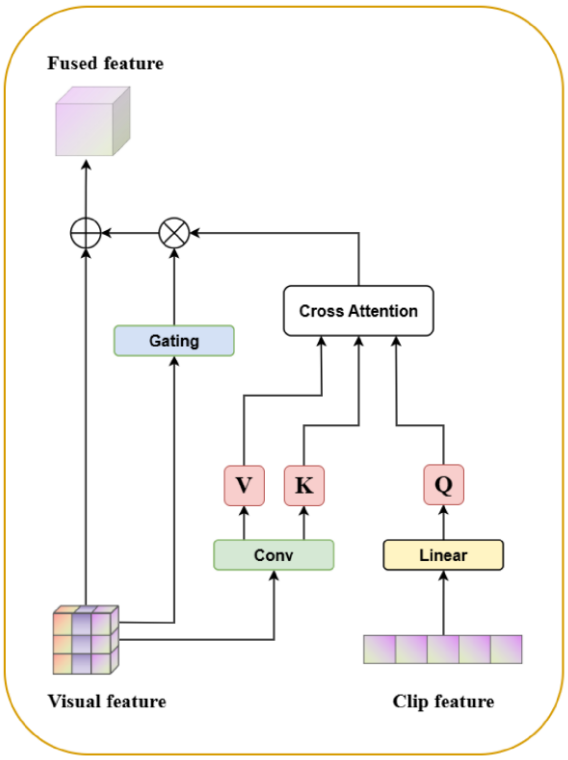

Fig. 2. Architectures of the CGCA module

3.2 Loss Function

We employ a combined loss function that integrates both Dice loss and Binary Cross-Entropy (BCE) loss to leverage their complementary advantages. The Dice loss is particularly effective for handling class imbalance between foreground and background regions, while the BCE loss provides stable gradient propagation throughout training. The combined loss is formulated as a weighted sum of these two components.

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2|X \cap Y| + \epsilon}{|X| + |Y| + \epsilon}, \tag{7}$$

$$\mathcal{L}_{\text{BCE}} = -[Y\log(X) + (1 - Y)\log(1 - X)], \tag{8}$$

$$\mathcal{L}\text{total} = \lambda\mathcal{L}_{\text{Dice}} + (1 - \lambda)\mathcal{L}_{\text{BCE}}. \tag{9}$$

where X denotes the predicted probability for pixel, Y represents the ground truth value

for pixel, and ϵ is a smoothing constant (set to 10^{-5}) to avoid division by zero. In our implementation, λ is set to 0.5, ensuring equal contribution from both loss components.

4 Experiments

In this section, we first introduce implementation details, two public datasets used in experiments and evaluation metrics, then present the experimental results and compare them with existing methods. Finally, we discuss the ablation study and demonstrate the effects of each proposed component.

4.1 Implementation Details

All experiments are conducted on a server equipped with an NVIDIA GeForce RTX 3090 GPU and an Intel Xeon E5-2620 v3 CPU. Our implementation is based on PyTorch. We employ the Adam optimizer with an initial learning rate of 1e-4. The learning rate is scheduled by ReduceLROnPlateau, in which the rate is reduced by a factor of 0.5 when the validation loss does not improve for 5 consecutive epochs. The model is trained for 200 epochs with a batch size of 4. Input images are resized to 224×224.

4.2 Datasets

We evaluate our method on two widely used smoke segmentation datasets, SmokeSeg and SMOKE5K. SmokeSeg comprises 6,144 real-world smoke images characterized by high transparency. The test set is partitioned into small, medium, and large subsets based on the smoke size to assess model performance across different scales. SMOKE5K is a hybrid dataset containing 1,360 real-world images and 4K synthetic smoke images. Following the setup in [24], 400 images are designated as the test set.

4.3 Evaluation Metrics

To comprehensively evaluate the model performance, we employ the F1-Score, Mean Squared Error (mMSE), mean Intersection over Union (mIoU), and Accuracy (ACC) as the core evaluation metrics. The definitions of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are adopted from the standard confusion matrix. Based on these, Precision (P), Recall (R) and ACC are calculated as follows:

$$\mathrm{P} = \frac{TP}{TP+FP+\epsilon}, \tag{10}$$

$$\mathrm{R} = \frac{TP}{TP+FN+\epsilon}, \tag{11}$$

$$\mathrm{ACC} = \frac{TP+TN}{TP+TN+FP+FN}. \tag{12}$$

We set $\epsilon = 10^{-7}$ in this work. Accordingly, the F1-Score, mMSE, mIoU are defined as:

$$F_1 = \frac{2\times P\times R}{P+R}, \tag{13}$$

$$\text{mMSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{14}$$

$$\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c+FP_c+FN_c}. \tag{15}$$

Here y_i is ground truth value of the i-th sample, $\hat{y}_i$ is Predicted value of the i-th sample, N is number of samples and C is Number of classes.

4.4 Results

We evaluated our proposed method CGUnet on the SMOKE5K and SmokeSeg datasets and compared it with other smoke segmentation methods.

Results on SmokeSeg. As shown in Table 1, our method achieves total $F1$ score of 79.70 and mIoU of 66.25, which are 5.95 and 4.75 higher than the second best of PSL [21] respectively, and outperforms all other methods significantly. Especially on the small smoke segmentation, our method achieves $F1$ 78.92 and mIoU 65.18, 11.31 and 12.13 higher than $F1$ 67.61 and mIoU 53.05 of PSL [21].

Table 1. Experimental results on the smokeseg dataset

Method	Backbone	Small			Medium			Large			Total			Params
		F_1	mIoU	mMSE	F_1	mIoU	mMSE	F_1	mIoU	mMSE	F_1	mIoU	mMSE	
EncNet [25]	ResNet-50	58.09	44.86	0.0012	71.16	57.74	0.0042	71.54	57.96	0.0232	66.54	53.15	0.0088	35.9
DeepLabV3+	ResNet-50	57.00	43.70	0.0013	73.34	59.80	0.0042	72.79	59.39	0.0219	67.26	53.85	0.0084	43.6
Trans-BVM[24]	ResNet-50	60.64	47.01	0.0013	72.97	59.51	0.0040	71.77	58.79	0.0223	69.12	55.57	0.0085	47.3
PSL [21]	ResNet-50	61.29	46.92	0.0012	74.60	61.04	0.0039	74.15	62.23	0.0219	70.87	57.89	0.0083	33.6
Mask2Former	Swin-B	58.71	45.78	0.0018	73.81	60.06	0.0041	72.33	58.97	0.0251	67.02	53.42	0.0089	99.4
PSL [21]	Swin-B	59.32	46.18	0.0014	74.62	60.22	0.0039	74.13	62.53	0.0232	69.73	57.21	0.0086	98.5
SegFormer [26]	MiT-B3	58.32	45.72	0.0017	74.34	61.37	0.0038	72.50	58.95	0.0235	67.99	54.98	0.0088	44.6
ProtoSeg [27]	MiT-B3	63.59	50.27	0.0011	72.41	59.80	0.0039	74.67	62.15	0.0216	70.88	57.46	0.0082	45.3
FoSp [22]	MiT-B3	66.03	52.68	0.0010	75.90	62.99	0.0037	74.98	62.24	0.0208	72.05	59.03	0.0079	47.5
PSL [21]	MiT-B3	67.61	53.05	0.0010	76.91	64.26	0.0035	77.59	64.53	0.0203	73.75	61.50	0.0072	45.3
CGUnet (ours)		**78.92**	**65.18**	0.0052	**79.60**	**66.11**	0.0055	**79.78**	**66.36**	0.0255	**79.70**	**66.25**	0.0098	7.4∗

Regarding Mean Squared Error (mMSE), we observe that our method reports 0.0098, which is higher than several compared methods (e.g., FoSp achieves 0.0079, PSL achieves 0.0072). We attribute this to a known characteristic of mMSE in class-imbalanced segmentation: mMSE averages squared errors across all pixels, and in smoke segmentation, the background class dominates the pixel count. Methods that predict high-precision smoke masks (fewer false positives but potentially more false negatives) can achieve lower mMSE by correctly predicting the abundant background pixels. Our

method, optimized primarily for F1 and mIoU, which introduces additional variance in boundary pixels and consequently increases mMSE. We argue that for the safety-critical application of smoke detection, higher recall (capturing more true smoke regions) is preferable to lower pixel-wise mMSE.

The visualization of segmentation results of the PSL model and the proposed model CGUnet are compared in Fig.3. Our segmentation masks are more similar with the ground truths. Both quantitative assessment and visual comparison confirm that our model can segment small smoke regions into images more accurately and completely. This indicates that the proposed UNet+CGSA architecture can more effectively model subtle features and global contextual dependencies of smoke, thereby significantly enhancing the model's segmentation accuracy for small-scale smoke.

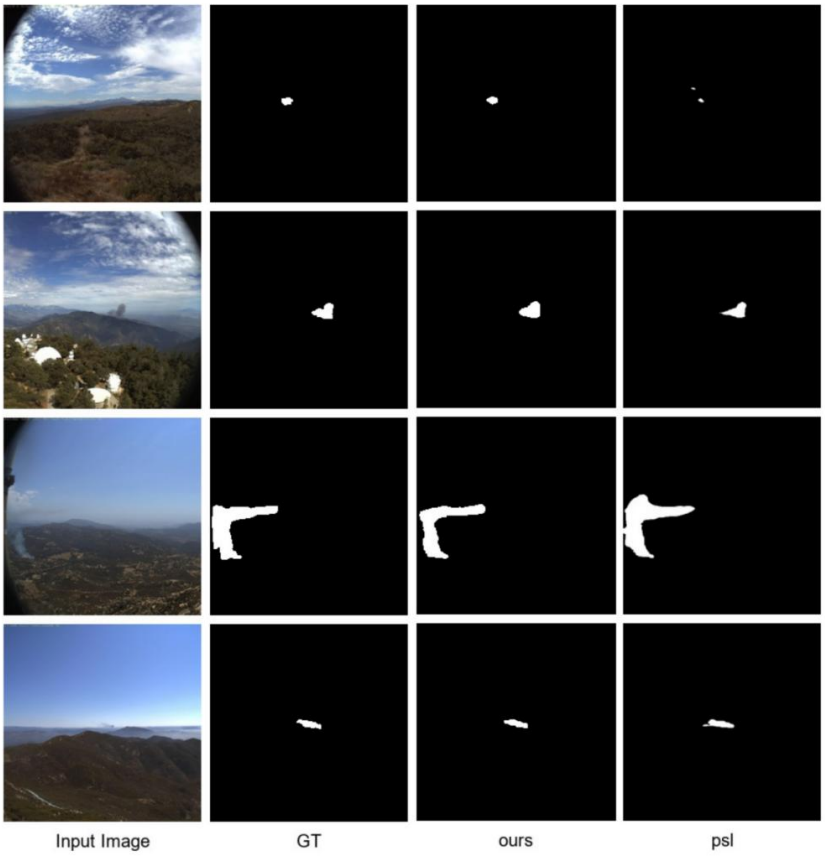

Fig. 3. Visualization of the smoke segmentation results on SmokeSeg dataset.

Results on SMOKE5K. As shown in Table 2 our method achieves F_1 score of 88.5, 6.6 higher than the second best of PSL [21] and outperforms all other methods significantly.

4.5 Ablation Study

To validate the effectiveness of our proposed model, we conducted ablation studies on the SMOKE5K dataset. Two baseline models were tested. The first baseline is the UNet model. The second baseline is the pre-trained CLIP image encoder, for which we add a segmentation head. To test the effectiveness of CGCA, we conducted an experiment (UNet+GSA) with a gated self-attention (GSA) module, in which the query is also from the visual feature, rather than from the CLIP feature in CGCA.

During training, all other parameters remain unchanged. The quantitative comparison results of the different models are presented in Table 3. The baseline UNet achieves F_1 score of 82.9, higher than the baseline CLIP. This means the pretrained CLIP encoder does not extract detailed and discriminative smoke features for segmentation. With GSA integrated into the UNet, $F1$ score decreases from 82.9 to 79.47, meaning the context features from the UNet encoder are not helpful. With the guidance of the CLIP feature, F_1 score improves significantly from 82.9 to 88.51. This demonstrates that the proposed CGCA module is effective, which merits the global semantics of CLIP and the precise localization of UNet.

Table 2. Experimental results on the smoke5k dataset.

Method	Backbone	$F_1(\uparrow)$	mMSE(↓)
BASNet [28]	ResNet-50	73.3	0.005
UCNet [29]	ResNet-50	78.7	0.003
Trans-BVM [24]	ResNet-50	79.1	0.002
PSL [21]	ResNet-50	80.8	0.003
Mask2Former [17]	Swin-B	76.8	0.004
PSL [21]	Swin-B	78.2	0.003
SegFormer [26]	MiT-B3	78.6	0.003
ProtoSeg [27]	MiT-B3	80.8	0.003
FoSp [22]	MiT-B3	81.6	<u>0.002</u>
PSL [21]	MiT-B3	<u>81.9</u>	**0.001**
CGUnet(ours)		**88.5**	0.005

Table 3. Ablation studies on the smoke5k dataset.

Method	F_1	mMSE	Acc
UNet	82.90	0.009	99.07
CLIP	72.00	0.016	98.37
UNet + GSA	79.47	0.008	99.13
UNet + CGCA	88.51	0.005	99.48

To investigate the impact of employing cross-attention layers in the decoding stage, we conducted layer ablation experiments on the SMOKE5K dataset. The results are presented in Table 4. The 0-layer (no attention) model achieves F_1 score of 79.23 and

mIoU of 65.60. However, introducing a single attention layer leads to performance degradation (F_1:75.30, mIoU: 60.39), indicating that a single-layer attention mechanism struggles to effectively capture cross-modal semantic information. When the number of layers increases to 2, performance improves substantially, with $F1$ reaching 86.88 and mIoU reaching 76.81—improvements of 9.7% and 17.1%, respectively, over the 0-layer baseline. The 3-layer configuration achieves the optimal performance (F_1: 88.51, mIoU: 79.39). These experimental results demonstrate that shallow attention (1 layer) fails to leverage the advantages of cross-modal interaction; Deep attention (2–3 layers) effectively fuses CLIP semantic features with multi-scale visual features, and the 3-layer setup achieves the best performance. Accordingly, we adopt the 3-layer cross-attention design as the final model configuration.

Table 4. Experimental results on the number of cross-attention layers.

Layers	F_1	mIoU	Acc	mMSE
0-layer	79.23	65.60	99.12	0.008
1-layer	75.30	60.39	99.01	0.009
2-layers	86.88	76.81	99.41	0.005
3-layers	88.51	79.39	99.48	0.005

4.6 Model Complexity

Since the CLIP encoder is frozen during training, only parameters of UNet need to be trained. The number of parameters to be trained is 7.4M, much smaller than other methods, as shown in table 1. The total number of parameters including the CLIP image encoder is 15.8M, which is still smaller than other models.

5 Conclusion

In this paper, we propose a novel CLIP guided UNet framework for smoke segmentation. The rich semantics of CLIP learned from huge amounts of web-scale image-text pairs on the Internet is used as the guidance of UNet decoder for smoke segmentation. Since UNet is a popular semantic segmentation model and CLIP is a general visual language model, we think the success of our method on smoke segmentation can be extended to other semantics. Although CLIP is frozen during training and only UNet parameters need to be trained, the model complexity is increased during inference, which may limit its applicability on edge devices.

References

1. Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481–2495(2017)

2. Olaf Ronneberger, Philipp Fischer, and Thomas Brox: U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention. Springer*, 234–241(2015)
3. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4),834–848(2018)
4. Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu: Dual attention network for scene segmentation. in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3141–3149(2019)
5. Zilong Huang, Xinggang Wang, Yunchao Wei, Lichao Huang, Humphrey Shi, Wenyu Liu, and Thomas S. Huang: Ccnet: Criss-cross attention for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6),6896–6908(2023)
6. Tao Jing, Qing-Hao Meng, and Hui-Rang Hou: Smokeseger: A transformer-cnn coupled model for urban scene smoke segmentation. *IEEE Transactions on Industrial Informatics*, 20(2),1385– 1396(2024)
7. Lin Zhang, Jing Wu, Feiniu Yuan, and Yuming Fang: Smoke-aware global-interactive non-local network for smoke semantic segmentation. *IEEE Transactions on Image Processing*, 33, 1175–1187(2024)
8. Feiniu Yuan, Lin Zhang, Xue Xia, Qinghua Huang, and Xuelong Li: A gated recurrent network with dual classification assistance for smoke semantic segmentation. *IEEE Transactions on Image Processing*, 30, 4409–4422(2021)
9. Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu: Vision- language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8), 5625– 5644(2024)
10. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever: Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, Marina Meila and Tong Zhang, Eds., 18–24 Jul 2021, vol. 139 of *Proceedings of Machine Learning Research*, 8748–8763.
11. Timo Lüddecke and Alexander Ecker: Image segmentation using text and image prompts. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7076–7086(2022)
12. Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos: Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7), 3523–3542(2022)
13. Jonathan Long, Evan Shelhamer, and Trevor Darrell: Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431– 3440(2015)
14. Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao: Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10), 3349–3364(2021)

15. Xiangtai Li, Henghui Ding, Haobo Yuan, Wenwei Zhang, Jiangmiao Pang, Guangliang Cheng, Kai Chen, Ziwei Liu, and Chen Change Loy: Transformer-based visual segmentation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 10138–10163(2024)
16. Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, and Yuyin Zhou: Transunet: Transformers make strong encoders for medical image segmentation. (2021)
17. Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar: Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1280–1289(2022)
18. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick: Segment any- thing. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3992–4003(2023)
19. Feiniu Yuan, Lin Zhang, Xue Xia, Qinghua Huang, and Xuelong Li: A wave-shaped deep neural network for smoke density estimation. *IEEE Transactions on Image Processing*, 29, 2301–2313(2020)
20. Kang Li, Feiniu Yuan, and Chunmei Wang: Frequency-space interaction with hierarchical aggregation network for lightweight smoke image segmentation. *IEEE Transactions on Consumer Electronics*,71(2), 2632–2643(2025)
21. Lujian Yao, Haitao Zhao, Zhongze Wang, Kaijie Zhao, and Jingchao Peng: Prototype-based scatter learning for smoke segmentation. *Pattern Recognition*, 172, 112605(2026)
22. Lujian Yao, Haitao Zhao, Jingchao Peng, Zhongze Wang, and Kaijie Zhao: Fosp: Focus and separation network for early smoke segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7), 6621–6629(2024)
23. Yichao Cao, Feng Yang, Xuanpeng Li, Xiaolin Meng, and Xiaobo Lu: Refining the granularity of smoke representation: Sam-powered density-aware progressive smoke segmentation framework. *Pattern Recognition*, 172, 112517(2026)
24. Siyuan Yan, Jing Zhang, and Nick Barnes: Transmission-guided bayesian generative model for smoke segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3), 3009–3017(2022)
25. Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal: Context encoding for semantic segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7151–7160(2018)
26. Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo: Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, 15, 12077–12090(2021)
27. Tianfei Zhou and Wenguan Wang: Prototype-based semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10), 6858–6872(2024)
28. Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand: Basnet: Boundary-aware salient object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7471–7481(2019)

29. Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes: Uc-net: Uncertainty inspired rgb- d saliency detection via conditional variational autoencoders. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8579–8588(2020)