*Computers, Materials & Continua*
DOI: 10.32604/cmc.2022.026498
***Article***

Tech Science Press

# Optimized Deep Learning Model for Fire Semantic Segmentation

**Songbin Li[1,*], Peng Liu[1], Qiandong Yan[1] and Ruiling Qian[2]**

[1]Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190, China
[2]Loughborough University, Loughborough, LE11 3TT, United Kingdom
*Corresponding Author: Songbin Li. Email: lisongbin@mail.ioa.ac.cn
Received: 28 December 2021; Accepted: 11 March 2022

**Abstract:** Recent convolutional neural networks (CNNs) based deep learning has significantly promoted fire detection. Existing fire detection methods can efficiently recognize and locate the fire. However, the accurate flame boundary and shape information is hard to obtain by them, which makes it difficult to conduct automated fire region analysis, prediction, and early warning. To this end, we propose a fire semantic segmentation method based on Global Position Guidance (GPG) and Multi-path explicit Edge information Interaction (MEI). Specifically, to solve the problem of local segmentation errors in low-level feature space, a top-down global position guidance module is used to restrain the offset of low-level features. Besides, an MEI module is proposed to explicitly extract and utilize the edge information to refine the coarse fire segmentation results. We compare the proposed method with existing advanced semantic segmentation and salient object detection methods. Experimental results demonstrate that the proposed method achieves 94.1%, 93.6%, 94.6%, 95.3%, and 95.9% Intersection over Union (IoU) on five test sets respectively which outperforms the suboptimal method by a large margin. In addition, in terms of accuracy, our approach also achieves the best score.

**Keywords:** Fire semantic segmentation; local segmentation errors; global position guidance; multi-path explicit edge information interaction; feature fusion

## 1 Introduction

Vision-based fire detection is a difficult but particularly important task for public safety. From existing literature, vision-based fire detection methods can be divided into two types. One is to judge whether there is a flame in an image [1–5]. The other regards the flame as an object and uses the object detection based method to detect fire [6–8]. Compared with the first type, the object detection based fire detection method can not only recognize the existence of fire but also locate the fire. However, it lacks accurate flame edge and shape information which makes it hard to accurately and automatically estimate the fire area. In general, due to the lack of precise area, shape, and location of flame, automated fire intensity analysis, prediction, and early warning are difficult to carry out. Therefore, it is necessary to realize the fire semantic segmentation in an image.

The goal of fire semantic segmentation is to recognize whether the pixel belongs to fire (shown in Fig. 1, which is similar to image segmentation tasks. Recently, advances in image processing techniques [9,10] have boosted the state-of-the-art to a new level for many tasks, such as semantic segmentation and salient object detection. However, it is still difficult to accurately resolve flames from a single image. The main reason may be the different backgrounds, multiple scales of fire at different evolving stages, and disturbance by fire-like objects. In this paper, we propose a fire semantic segmentation method based on global position guidance and multi-path explicit edge information interaction. Specifically, to alleviate the problem of local segmentation errors in low-level feature space caused by the disturbance of fire-like objects and background noise, a global position guidance mechanism is proposed. This module uses the accurate top-level position information of top-level features to reconstruct spatial detailed information in a top-down manner. Besides, we employ a multipath explicit edge information interaction module to organically aggregate coarse segmentation results and edge information to refine the fire boundary. In this module, we first explicitly construct edge information extraction through strong supervised learning, and then realize the interaction between edge information and coarse segmentation results through a convolutional layer.

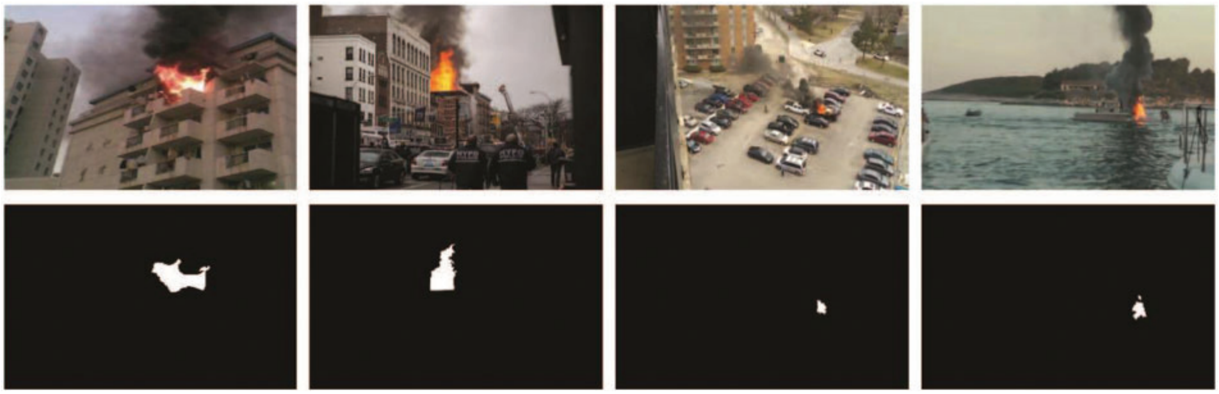


**Figure 1:** The goal of fire semantic segmentation is to recognize whether the pixel belongs to fire. Each column represents an original image and the corresponding fire semantic segmentation map. The pixels belonging to fire are marked as white, and the others are marked as black

The main contributions of this paper can be summarized as follows:

1) We propose a novel fire semantic segmentation method based on global position guidance and multi-path explicit edge information interaction. The experimental results show that our method achieves 94.7% average IoU on five test sets which outperforms the best semantic segmentation method and salient object detection method by 15.9% and 0.8%, respectively. It demonstrates that our method has better performance on fire segmentation than previous state-of-the-art semantic segmentation and salient object detection methods.
2) In this paper, a global position guidance module is proposed to solve the problem of local segmentation errors in low-level feature space. Besides, a multi-path explicit edge information interaction module based on edge guidance is utilized to organically aggregate coarse segmentation results and edge information to refine the fire boundary.
3) A fire semantic segmentation dataset of 30000 images is established, which is currently the first fire semantic segmentation dataset in this area. This dataset is created by synthesizing the real flame region with normal images. We randomly select 1100 images from [5] and label them to obtain the real flame region.

## 2 Related Work and Scope

In this section, we give a summary of related works in Tab. 1. Traditional fire detection methods [11–15] mainly focus on handcraft features, such as color, shape, texture, motion, etc. They have some defects, such as lacking robustness, failing to detect fire at a long distance or in a challenging environment, etc. Recent date-driven based deep learning promoted the progress of fire detection. Fire detection methods based on deep learning can be divided into two categories: classification-based methods [1–5] and object detection-based methods [6–8]. Classification-based approaches treat fire detection as an image classification task. These methods can judge whether there is fire in an image, but cannot locate the fire. The object detection-based fire detection methods can not only recognize the existence of fire but also locate the fire. However, the fire position is marked with rectangular boxes. It is unable to provide flame edge and shape information. The goal of fire semantic segmentation is to recognize whether the pixel belongs to fire, which is similar to image segmentation tasks. However, it is difficult to obtain good results by directly applying the existing deep learning based segmentation methods [16–24] to fire detection. These methods are not specially designed for fire semantic segmentation, so the discrimination ability of fire-like objects is relatively weak, and it is difficult to accurately parse the fire boundary. In addition, they have poor performance on local small-scale fires. To this end, we propose a fire semantic segmentation method based on global position guidance and multi-path explicit edge information interaction. The global position guidance mechanism is proposed to alleviate the problem of local segmentation errors in low-level feature space caused by the disturbance of flame-like objects and background noise. It uses the accurate top-level position information of top-level features to reconstruct spatial detailed information in a top-down manner. Besides, the multipath explicit edge information interaction mechanism is proposed to organically aggregate coarse segmentation results and edge information to refine the fire boundary.

**Table 1:** Summary of related works

| Method | Type | Features | Limitations |
|---|---|---|---|
| [11] | Handcrafted fire detection | Combine fuzzy inference with a color statistical mathematical model to detect fire | These methods focus on handcraft features, and exist some defects such as lacking robustness, failing to detect fire at a long distance or in a challenging environment, etc. |
| [12] | Handcrafted fire detection | Use YCbCr colors statistical information | |
| [13] | Handcrafted fire detection | Employed temporal and spatial wavelet transform to compute flame regions | |
| [14] | Handcrafted fire detection | Combine color, shape variation and motion analysis | |
| [15] | Handcrafted fire detection | Use spatio-temporal flame model and dynamic texture analysis | |

(Continued)

**Table 1:** Continued

| Method | Type | Features | Limitations |
|---|---|---|---|
| [1] | Fire detection based on classification | Use LeNet-5 to detect fire | These methods can judge whether there is fire in an image, but cannot locate the fire |
| [2] | Fire detection based on classification | Use AlexNet to detect fire | |
| [3] | Fire detection based on classification | Use MobileNet to detect fire | |
| [4] | Fire detection based on classification | Use ResNet50 to detect fire | |
| [5] | Fire detection based on classification | Introduce three mechanisms to prove the performance | |
| [6] | Fire detection based on object detection | Use tiny-yolo-voc to locate fire | The fire position is marked with rectangular boxes. Unable to provide flame edge and shape information |
| [7] | Fire detection based on object detection | Use Faster R-CNN to locate fire | |
| [8] | Fire detection based on object detection | Use YOLO v3 to locate fire | |
| [16] | Semantic segmentation method | Use spatial pyramid pooling module to capture rich global contextual information | These methods are not specially designed for fire semantic segmentation, so the discrimination ability of fire-like objects is relatively weak, and it is difficult to accurately parse the fire boundary. In addition, they have poor performance on local small-scale fires |
| [17] | Semantic segmentation method | Use spatial pyramid pooling module to capture rich global contextual information | |
| [18] | Semantic segmentation method | Describe the attention mechanism as expectation maximization | |
| [19] | Semantic segmentation method | Propose two attention mechanisms | |
| [20] | Semantic segmentation method | Propose a strip pooling and a mixed pooling module to predict the shape of the target | |
| [21] | Salient object detection method | Proposed a pixel-wise contextual attention network | |
| [22] | Salient object detection method | Propose a boundary-ware network | |
| [23] | Salient object detection method | Concentrate the differences among feature maps when fusing features | |
| [24] | Salient object detection method | Propose a multi-scale interactive network | |
| Ours | Fire semantic segmentation method | Propose a novel fire semantic segmentation method based on global position guidance and multi-path explicit edge information interaction | None of the above problems |

## 3 Global Position Guidance Mechanism

The encoder based on CNN can extract different feature representations. Top-level semantic features preserve precise fire position information. Low-level spatial detail features contain rich fire boundary information. Both of them are vital to fire segmentation. The progressive fusion of different levels of features has a very significant effect on fire segmentation tasks. However, attacked by background noise and flame-like objects, the low-level fire spatial features may arise local segmentation errors. Consequently, the key to improving the performance of fire semantic segmentation is to restrain the offset of low-level spatial features.

As mentioned above, the receptive field of the top-level features is the largest among these encoded features and the fire position information of them is the most accurate. Besides, when the information progressively flows from the top-level to the low level, the accurate position information contained in top-level features is gradually diluted. Thus a top-down global position guidance mechanism to directly deliver top-level position information to low-level feature space to restrain the local segmentation errors is designed.

In this module, the top-level features $F_e^{(t)}$ are outputted from the last layer of the encoder. Besides, we define the encoded features from $i_{th}$ layer as $F_e^{(i)}$, $i \in (1, t-1)$. First, two pointwise convolution layers $\mathcal{F}$ with batch normalization (BN) and ReLU activation function are performed to change the number of channels of $F_e^{(t)}$ and $F_e^{(i)}$ to $M$. Then, a bilinear interpolation function to up-sample $F_e^{(t)}$ to the same size as $F_e^{(i)}$. The fused features $\hat{F}_e^{(i)}$ could be denoted as:

$$\hat{F}_e^{(i)} = \left[\mathcal{F}\left(F_e^{(i)};\omega_i, b_i\right), U_p\left(\mathcal{F}\left(F_e^{(t)};\omega_t, b_t\right)\right)\right] \tag{1}$$

$$\mathcal{F}(\mathrm{F};\omega,\mathrm{b}) = ReLU\left(BN\left(F \otimes \omega + \mathrm{b}\right)\right) \tag{2}$$

where $(\omega_i, b_i)$ and $(\omega_t, b_t)$ are the kernel weight and bias of $F_e^{(i)}$ and $F_e^{(t)}$ respectively, $U_p$ stands for up-sample, $\otimes$ means convolution operation and $[\ldots]$ means concatenation. Next, a same pointwise convolution layer is used to squeeze the channel of $F_e^{(i)}$ into $M$. So far, we obtain the relative position attention map $\hat{F}_{e-M}^{(i)}$ which has accurate position information.

To further enhance the representation capability of $\hat{F}_{e-M}^{(i)}$, we introduce efficient channel attention. The map $\hat{F}_{e-M}^{(i)}$ is first compressed by a global pooling operation $\mathcal{G}$ to obtain the vector $\mathcal{Y}$ which has global contextual information.

$$\mathcal{Y} = \mathcal{G}\left(\hat{F}_{e-M}^{(i)}\right) = \frac{1}{WH}\sum_{i=1,j=1}^{W,H}\left(\hat{F}_{e-M}^{(i)}\right)_{ij} \tag{3}$$

where $W$, $H$ denotes the width and height of the input respectively. Then, an efficient fully connected layer is utilized to transform the vector $\mathcal{Y}$ into a reconstruction coefficient $\omega$.

$$\omega_m = \sigma\left(\sum_{j=1}^{k}\alpha^j\mathcal{Y}_m^j\right), \mathcal{Y}_m^j \in \Omega_m^j, m \in [1, C] \tag{4}$$

where $\alpha^{\mathrm{j}}$ represents the weight parameters, $\sigma$ is the sigmoid activation function, $\Omega^{\mathrm{j}}_{\mathrm{m}}$ represents the set of k adjacent channels of $\mathcal{Y}_{\mathrm{m}}$, and C is the number of channels. Next, a channel-wise multiplication operation is employed to reconstruct the $\hat{\mathrm{F}}^{(\mathrm{i})}_{\mathrm{e-M}}$,

$$\hat{F}_{e-M-r}^{(i)} = \omega * \hat{F}_{e-M}^{(i)} \tag{5}$$

At last, we multiply $\hat{F}^{(i)}_{e-M-r}$ with $F^{(i)}_{e}$ to restrain the local segmentation errors in low-level feature space.

$$O = \hat{F}^{(i)}_{e-M-r} \cdot F^{(i)}_{e} \tag{6}$$

As shown in Fig. 2, the baseline without the GPG module has some wrong segmentation. With the GPG module applied, the local segmentation errors are restrained.

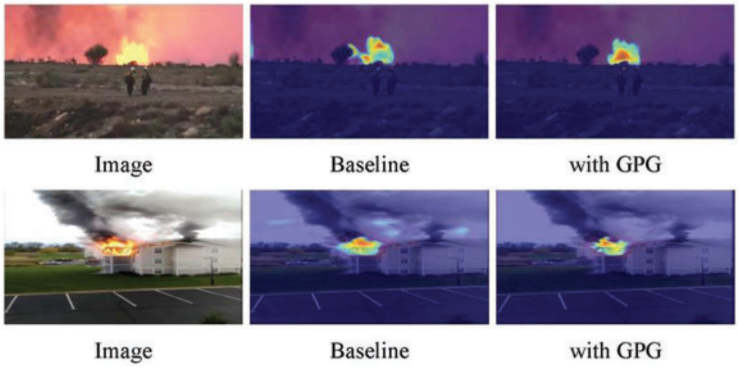


**Figure 2:** The heat map visualization results of baseline and global position guidance module. They demonstrate that the GPG module can effectively restrain the local segmentation errors

## 4 Multi-Path Explicit Edge Information Interaction Mechanism

Another challenge of fire semantic segmentation is edge prediction. Different from central pixels that have higher prediction accuracy due to the internal consistency of the fire, pixels near the boundary are more prone to be misdetected. The main reasons are as follows. Compared with central pixels, the edge of fire contains less information. Besides, diverse and complex backgrounds will suppress edge information. Therefore, to solve the problem of edge segmentation error caused by lack of flame edge information. we need to explicitly utilize flame edge information.

To achieve this, the edge information of the flame needs to be extracted explicitly. A simple approach is to construct an edge information extraction branch and train it through strong supervised learning. First, we apply the edge extraction algorithm (e.g., Canny, Sobel, and Laplace operator, etc.) to label image $\mathcal{Y}_{\text{label}}$ to obtain the corresponding edge annotation $\mathcal{Y}_{\text{edge}}$. To explicitly extract the edge information, the output features $\mathrm{F}_{\mathrm{d}}^{(\mathrm{l})}$ of the last layer of the decoder are inputted into the edge information extraction branch. This branch consists of a 3 × 3 convolution layer, a batch normalization, and an activation function. The edge information $\mathrm{I}_{\text{edge}}$ could be denoted as:

$$I_{edge} = \emptyset\left(BN\left(Conv\left(F_{d}^{(l)}; \omega_{e}, b_{e}\right)\right)\right) \tag{7}$$

where $\omega_e$ and $b_e$ represent the kernel parameters and bias respectively. Ø means activation function. Then, we use three loss functions to train them,

$$\mathcal{L}_{bce} = -\sum_{r,c}\left[G_{r,c}\log\left(I_{r,c}\right) + \left(1 - G_{r,c}\log\left(1 - I_{r,c}\right)\right)\right]$$

$$
\begin{aligned}
&\mathcal{L}_{ssim} = 1 - \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \\
&\mathcal{L}_{iou} = 1 - \frac{\sum_{r=1}^{H}\sum_{c=1}^{W} I_{r,c}G_{r,c}}{\sum_{r,c=1}^{H,W}\left[I_{r,c} + G_{r,c} - S_{r,c}G_{r,c}\right]} \\
&\mathcal{L}_{total} = \mathcal{L}_{bce} + \mathcal{L}_{ssim} + \mathcal{L}_{iou}
\end{aligned}
\tag{8}
$$

where $G_{r,c}$ and $I_{r,c}$ mean the fire edge confidence of the ground truth and prediction map respectively, $\mu_x$ and $\mu_y$ represent the average value of prediction and ground truth respectively, $\sigma_*$ means the variance. $C_1$ and $C_2$ are two small constants.

After the complementary fire edge information is obtained, we aim to aggregate flame edge information and flame object features to achieve information interaction. It is useful for obtaining better flame semantic segmentation results. The decoded features (flame object features) are defined as $F_d^{(i)}$, $i \in (1, l)$. Then, the information interaction can be denoted as:

$$O_d^{(i)} = \mathcal{F}\left[\left(U_p\left(F_d^{(i)}\right)\right), U_p\left(I_{edge}\right)\right] \tag{9}$$

where $O_d^{(i)}$ stands for refined results.

**Algorithm 1:** Multi-path Explicit Edge Information Interaction

```
Input: coarse results $F_d^{(i)}$, $i \in (1, l)$; edge information $I_{edge}$
Output: refined fire prediction map $O_d^i$
if explicit edge extraction then
        $I_{edge} \leftarrow \mathcal{F}(F_d^l)$
        return $I_{edge}$
if edge information interaction then
        while $i = 1; i \le l; i \leftarrow \mathrm{i} + 1$ do
                $F_d^i, I_{edge} \leftarrow U_p\left(F_d^{(i)}\right),\ U_p(I_{edge})$
                $O_d^i \leftarrow Conv\left(\left[F_d^i, I_{edge}\right]\right)$
return $\{O_d^i | i = 1, \ldots\ldots, l\}$
```

## 5 Overview of Global Position Guidance and Multi-Path Explicit Edge Information Interaction Networks

Based on the above ideas, we design a fire semantic segmentation network based on global position guidance and multi-path explicit edge information interaction. The overview of the proposed model is illustrated in Fig. 3. It consists of a deep encoder, four global position guidance modules with feature fusion operation, an explicit edge information extraction module, and a multi-path explicit edge information interaction module. The input image $X$ is fed into the encoder [5] to obtain encoded features $F_e^i$,

$$F_e^i = Encoder\left(X\right), i \in (1, t) \tag{10}$$

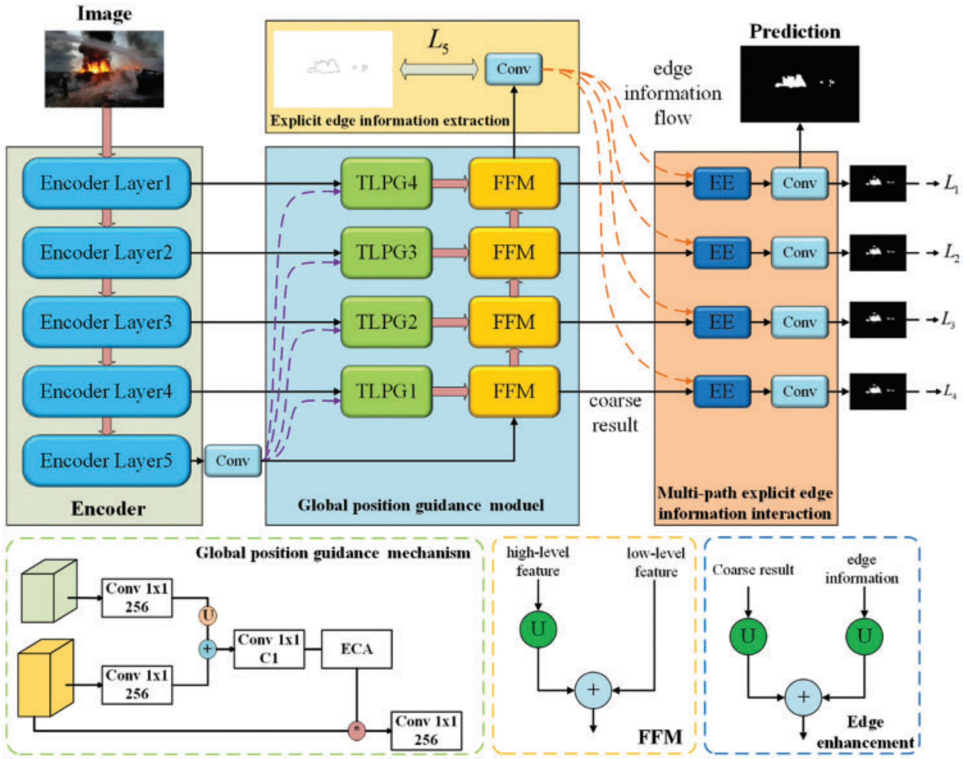


**Figure 3:** The overview architecture of the global position guidance and multi-path explicit edge information interaction based fire semantic segmentation networks

It is worth noting that the encoder includes three main parts, namely multi-scale feature extraction, implicit deep supervision, and channel attention mechanism. First, to establish a good feature foundation for the high-level semantic feature and global position information extraction, a multi-scale feature extraction module is used.

$$B = \mathcal{M}(A) = \left[h_{1\times 1}(A), h_{3\times 3}(A), h_{5\times 5}(A), h_{pooling}(A)\right] \tag{11}$$

where $A \in R^{C\times H\times W}$ is the input feature, $h_{k\times k}$ means the convolution operation with a kernel size of $k \times k$, and $B$ is the output. Then, three densely connected structures [25] which permit the gradient to flow directly to earlier layers are employed to enhance the feature representation capability. At last, the channel attention widely used in computer vision tasks is utilized. The process of it can be described as:

$$\begin{aligned}
\tilde{X} &= \mathcal{GP}(\tilde{x}) = \frac{1}{H \times W} \sum_{i,j=0}^{H-1,W-1} x(i,j) \\
x_{lb} &= \emptyset\left(\omega_2 \otimes \emptyset\left(\omega_1 \otimes \tilde{X}\right)\right) \\
o &= x_{lb} * x
\end{aligned} \tag{12}$$

where $o$ is the final output, $\tilde{x}$ means the input, $\tilde{X}$ is a vector that includes the global information. $\omega_2$ and $\omega_1$ are the corresponding weight matrixes. $x_{lb}$ is a reconstruction vector.

When the encoded feature $F_e^i$ is captured, we use a convolution layer to squeeze the channel of top-level feature $F_e^t$ into 256. Then, the feature $F_{e-256}^t$ is fed into the GPG module to restrain the local segmentation errors of low-level feature space. Besides, we aggregate the information progressively from the top level to the low level like the U-Net architecture [26] through a simple feature fusion operation. At last, as mentioned in Section 4, an MEI module is used to refine the coarse segmentation results. The cross-entropy loss based supervision is applied to train the whole network. It can be represented as:

$$\mathcal{L}^{(i)}\left(O_d^i; G\right) = \mathcal{L}_{total}$$

$$\mathbb{L} = \alpha\mathcal{L}\left(\mathcal{Y}_{edge}; I_{edge}\right) + \theta\sum_{i=1}^{4}\mathcal{L}^{(i)}\left(O_d^i; G\right) \tag{13}$$

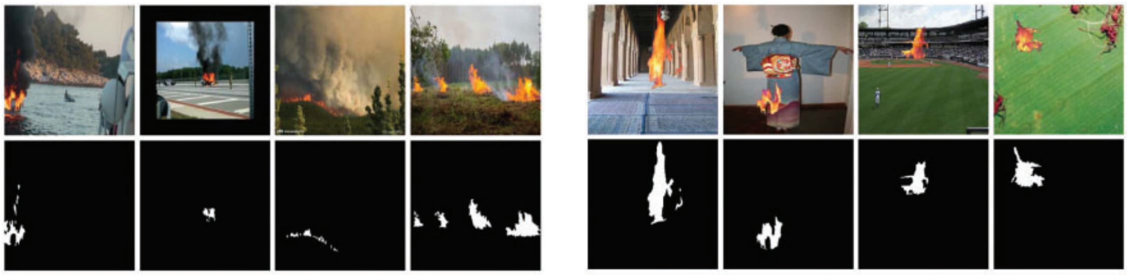
where $\mathbb{L}$ represents the total loss, $O_d^i$ is the fire prediction map, and $j$ is the number of categories. $G$ stands for the ground truth, $\alpha$ and $\theta$ are the weight coefficient.

## 6 Experiments and Analysis

In this section, we first introduce the dataset and evaluation metrics. Then we present the implementation details. Next, a series of ablation studies are conducted to verify the effect of each module. Finally, we carry out reasonable experiments on our created dataset to evaluate the performance of the proposed method. Experimental results demonstrate that our method achieves the best performance compared with the existing semantic segmentation and salient object detection methods.

### *6.1 Dataset and Evaluation Metrics*

In this paper, we create a fire semantic segmentation dataset (FSSD) which consists of 30000 synthetic images and 1100 real fire images. The generation of the dataset is described as follows. First, we randomly select 1100 images from datasets [5] and label them carefully. Then, we extract the real flame region and synthesize them with normal images to create the dataset. Finally, 1000 images are used to generate training datasets, and 100 images are used to generate testing datasets. Some real fire images and synthetic images are shown in Fig. 4. In this paper, 26000 images are used for training (25000 synthetic images and 1000 real images). Besides, we divide the test images into five test sets (each includes 1000 images). To improve the performance of fire semantic segmentation, we use the dataset [5] (except for the 1000 images used to extract the real flame region) to pre-train the encoders of all comparison methods.



(a) real fire image with annotation

(b) synthetic image with annotation

**Figure 4:** Some visual examples of our created fire semantic segmentation dataset. Each column represents an original image and the corresponding annotation

We use three measurements to evaluate all methods. Mean Absolute Error (MAE) is described as the average pixel-wise absolute difference between the prediction map and the ground truth. Therefore, the mathematical formula of the MAE can be expressed as:

$$mae = \frac{1}{H \times W} \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} \left| P_{x,y} - G_{x,y} \right| \tag{14}$$

where $P$ denotes the fire semantic segmentation map, $G$ is the corresponding ground truth. Interaction over Union (IoU) is widely used in semantic segmentation [27] to evaluate the performance of the algorithm. It represents the degree of overlap between the prediction map and the ground truth. The IoU can be computed by

$$IoU = \frac{\sum_{x=0,y=0}^{H-1,W-1} \left\{ P_{x,y} == G_{x,y} \right\}}{N}$$
$$N = \sum_{x=0,y=0}^{H-1,W-1} P_{x,y} + G_{x,y}, P, G \neq 0 \tag{15}$$

The third evaluation metric is accuracy, which is defined as the ratio of the number of correctly predicted images (The IoU threshold is set to 0.4) to the total images. The accuracy can be illustrated as:

$$accuracy = \frac{M}{N} \tag{16}$$

where $M$ indicates the images correctly predicted, $N$ is the total images.

### *6.2 Implementation Details*

In this paper, we adopt EFDNet [5] pre-trained on FSSD (only for encoder) as our backbone. In the training stage, we resize each image to 320 × 320 with random flipping, then randomly crop a patch with the size of 288 × 288 for training. We utilize Pytorch to implement our method. The Adaptive moment estimation is applied to optimize the whole parameters of the network with a batch size of 8. The hyperparameter values are shown in Tab. 2, referring to the settings in [5]. To avoid the model failing into suboptimal, we adopt the "poly" learning rate policy with the initial learning rate $1e{-}5$ for the backbone and 0.001 for the other parts to train our model. Like [21], the maximum iterative epoch of all methods is set to 30.

**Table 2:** Hyperparameter values

| Betas | Eps | Weight decay | Learning rate |
|---|---|---|---|
| (0.9, 0.999) | 1e−8 | 5e−4 | 1e−5 for the backbone, 0.001 for the other parts |

### *6.3 Ablation Study*

In this section, to investigate the effect of the proposed GPG and MEI modules, a series of ablation studies are performed. As illustrated in Tab. 3, the baseline which does not contain any optimization achieves 0.008% and 88.3% in terms of MAE and IoU, respectively. With the GPG module applied, both IoU and MAE are improved, where the MAE score is decreased by 50.0% compared with the baseline. The IoU of GPG is 91.5% which outperforms the baseline by 3.2% demonstrating that the idea of using top-level accurate position information to restrain the local fire segmentation errors is very efficient. Besides, when we aggregate MEI and GPG, the performance of the proposed approach is enhanced further. In terms of MAE, the final model achieves 0.002 which brings a 50.0% improvement compared with the baseline. It also outperforms GPG. Furthermore, the final model improves the IoU from 91.5% to 94.1% based on GPG.

**Table 3:** The quantitative results of the ablation experiment with different components on the DS01

| Baseline | GPG | MEI | MAE | IoU |
|---|---|---|---|---|
| √ | | | 0.008 | 88.3% |
| √ | √ | | 0.004 | 91.5% |
| √ | √ | √ | 0.002 | 94.1% |

### *6.4 Compared with Existing Deep Learning Based Segmentation Methods*

In this section, to demonstrate the performance of our method, 9 segmentation methods (5 semantic segmentation methods [16–20] and 4 salient object detection methods [21–24]) are compared. For a fair comparison, the fire semantic segmentation results of different methods are obtained by running their released codes under the default parameters. Moreover, we pre-train all encoders on FSSD.

The quantitative comparison results on our created benchmark are illustrated in Tabs. 4 and 5. Compared with other methods, our method achieves the best performance. In terms of MAE, the proposed method achieves a better performance on five test sets which outperforms the other methods by a large margin. The IoU evaluation metric is widely used in the semantic segmentation task. Our method improves it from 93.2% to 94.1% on DS01. Besides, we use accuracy as an evaluation metric for image-level fire detection. From the results, we can see that our method achieves an accuracy of 96.2% which outperforms other methods by a large margin (Threshold $T$ is set to 0.6).

**Table 4:** The quantitative comparison results with existing semantic segmentation methods on the FSSD dataset. The best result of each evaluation metric is highlighted in boldface

| Methods | DS01 | | | DS02 | | | DS03 | | |
|---|---|---|---|---|---|---|---|---|---|
| | MAE | IoU | Accuracy | MAE | IoU | Accuracy | MAE | IoU | Accuracy |
| [16] | 0.015 | 77.6% | 89.4% | 0.012 | 74.8% | 86.0% | 0.013 | 80.4% | 94.6% |
| [17] | 0.019 | 73.2% | 83.8% | 0.018 | 69.0% | 77.1% | 0.019 | 72.9% | 85.2% |
| [18] | 0.023 | 68.4% | 77.1% | 0.013 | 68.4% | 73.6% | 0.018 | 73.2% | 90.4% |
| [19] | 0.019 | 73.8% | 85.1% | 0.018 | 68.8% | 75.8% | 0.019 | 73.9% | 85.9% |
| [20] | 0.025 | 65.3% | 69.0% | 0.028 | 56.0% | 60.6% | 0.027 | 63.5% | 51.5% |
| [21] | 0.005 | 84.9% | 95.0% | 0.004 | 83.1% | 93.9% | 0.005 | 85.9% | 98.6% |
| [22] | 0.004 | 93.2% | 97.4% | 0.003 | 92.8% | 97.9% | 0.002 | 94.4% | 99.8% |
| [23] | 0.005 | 86.8% | 97.0% | 0.004 | 85.4% | 97.0% | 0.004 | 87.8% | 99.7% |
| [24] | 0.012 | 77.2% | 87.7% | 0.007 | 77.5% | 91.4% | 0.009 | 81.2% | 94.4% |
| Ours | 0.002 | 94.1% | 98.5% | 0.002 | 93.6% | 98.5% | 0.002 | 94.6% | 99.9% |

**Table 5:** The quantitative comparison results with existing semantic segmentation methods on the FSSD dataset. The best result of each evaluation metric is highlighted in boldface

| Methods | DS04 | | | DS05 | | |
|---|---|---|---|---|---|---|
| | MAE | IoU | Accuracy | MAE | IoU | Accuracy |
| [16] | 0.015 | 80.0% | 90.8% | 0.018 | 81.0% | 96.6% |
| [17] | 0.016 | 79.2% | 93.4% | 0.020 | 79.7% | 96.4% |
| [18] | 0.026 | 70.7% | 78.8% | 0.030 | 72.8% | 84.6% |
| [19] | 0.016 | 78.7% | 92.0% | 0.020 | 80.9% | 97.3% |
| [20] | 0.020 | 72.0% | 78.4% | 0.026 | 75.2% | 94.6% |
| [21] | 0.004 | 88.3% | 99.1% | 0.006 | 88.5% | 99.8% |
| [22] | 0.003 | 94.4% | 98.6% | 0.006 | 94.6% | 99.7% |
| [23] | 0.004 | 89.1% | 98.7% | 0.006 | 89.1% | 99.9% |
| [24] | 0.011 | 79.1% | 89.3% | 0.014 | 81.5% | 94.6% |
| Ours | 0.002 | 95.3% | 99.3% | 0.002 | 95.9% | 100.0% |

To comprehensively compare the performance of different methods, we present some visual results of different methods. As illustrated in Fig. 5, our method has a better performance than the previous semantic segmentation methods. Specifically, the proposed method not only highlights the correct fire regions clearly but also well suppresses the background noises. Besides, it is robust in dealing with flame-like objects (row 1) and low contrast background (row 4). Moreover, compared with other methods, the fire boundary generated by the proposed method is more accurate.

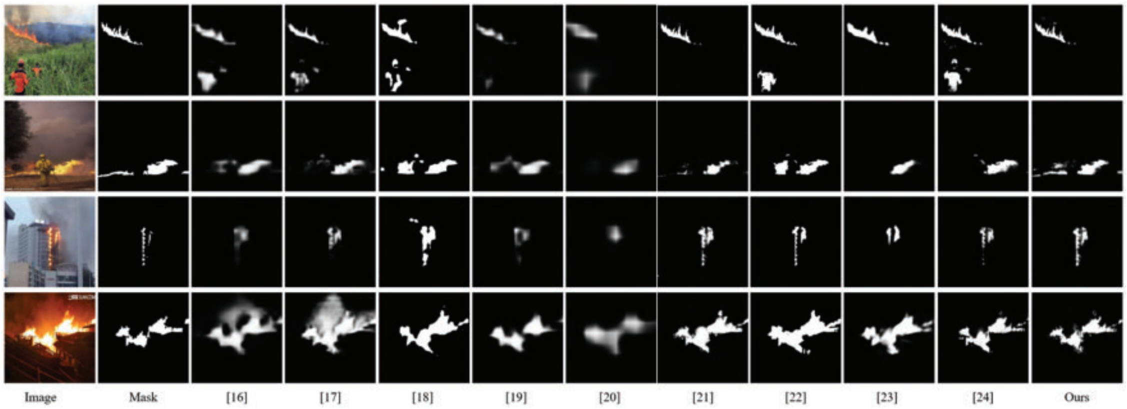


**Figure 5:** Some visual results of different methods. Each row stands for one original image and corresponding fire semantic segmentation maps. Each column represents the predictions of one method

### 6.5 *Analysis of Model Parameters*

In this subsection, we analyze the parameters of different methods. The results are illustrated in Tab. 6. We can see that the proposed method has only 6.9 MB parameters which is suitable for resource-constrained devices. Compared with the suboptimal method, it decreases 72.9%.

**Table 6:** The parameter size of different methods

| Methods | Parameters (MB) | Methods | Parameters (MB) |
|---|---|---|---|
| [16] | 53.6 | [21] | 67.1 |
| [17] | 39.0 | [22] | 332.4 |
| [18] | 34.7 | [23] | 25.5 |
| [19] | 47.4 | [24] | 162.4 |
| [20] | 41.5 | Ours | 6.9 |

## 7 Conclusion

In this paper, a method based on global position guided and multi-path explicit edge information interaction is proposed for fire semantic segmentation. First, existing literature shows that it is challenging to accurately separate the fire from diverse backgrounds and flame-like objects. To this end, considering the accurate position information contained in top-level features, we propose a global position guidance module to restrain the feature offset in low-level feature space thereby correcting the local segmentation errors. Besides, to further get more accurate boundary prediction, we first explicitly extract the edge information through strong supervision. Then, a multi-path information interaction is designed to refine the coarse segmentation. Experimental results on FSSD datasets show that the proposed method outperforms previous state-of-the-art methods under three evaluation metrics.

In the future work, we intend to introduce multitask learning to further improve the performance of the model and multi-scale feature extraction to deal with small flame segmentation. Besides, the fast and small model which can be easily implemented on resource-limited mobile devices will be also considered.

**Funding Statement:** This work was supported in part by the Important Science and Technology Project of Hainan Province under Grant ZDKJ2020010, and in part by Frontier Exploration Project Independently Deployed by Institute of Acoustics, Chinese Academy of Sciences under Grant QYTS202015 and Grant QYTS202115.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

[1] S. Frizzi, R. Kaabi, M. Bouchouicha, J. -M. Ginoux, E. Moreau *et al.,* "Convolutional neural network for video fire and smoke detection," in *IECON 2016-42nd Annual Conf. of the IEEE Industrial Electronics Society*, Florence, Italy, pp. 877–882, 2016.

[2] K. Muhammad, J. Ahmad and S. W. Baik, "Early fire detection using convolutional neural networks during surveillance for effective disaster management," *Neurocomputing*, vol. 288, pp. 30–42, 2018.

[3] K. Muhammad, S. Khan, M. Elhoseny, S. H. Ahmed and S. W. Baik, "Efficient fire detection for uncertain surveillance environment," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 5, pp. 3113–3122, 2019.

[4] J. Sharma, O. -C. Granmo, M. Goodwin and J. T. Fidje, "Deep convolutional neural networks for fire detection in images," in *Int. Conf. on Engineering Applications of Neural Networks*, Athens, Greece, pp. 183–193, 2017.

[5] S. Li, Q. Yan and P. Liu, "An efficient fire detection method based on multiscale feature extraction, implicit deep supervision and channel attention mechanism," *IEEE Transactions on Image Processing*, vol. 29, pp. 8467–8475, 2020.

[6] S. Wu and L. Zhang, "Using popular object detection methods for real time forest fire detection," in *2018 11th Int. Symp. on Computational Intelligence and Design (ISCID)*, Hangzhou, China, pp. 280–284, 2018.

[7] P. Barmpoutis, K. Dimitropoulos, K. Kaza and N. Grammalidis, "Fire detection from images using faster r-cnn and multidimensional texture analysis," in *ICASSP 2019-2019 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, pp. 8301–8305, 2019.

[8] P. Li and W. Zhao, "Image fire detection algorithms based on convolutional neural networks," *Case Studies in Thermal Engineering*, vol. 19, pp. 100625, 2020.

[9] S. Yadav, "Vision-based detection, tracking, and classification of vehicles," *IEIE Transactions on Smart Processing & Computing*, vol. 9, no. 6, pp. 427–434, 2020.

[10] S. Yadav and S. Yadav, "Image fusion using hybrid methods in multimodality medical images," *Medical & Biological Engineering & Computing*, vol. 58, no. 4, pp. 669–687, 2020.

[11] T. Celik, H. Ozkaramanlt and H. Demirel, "Fire pixel classification using fuzzy logic and statistical color model," in *2007 IEEE Int. Conf. on Acoustics, Speech and Signal Processing-ICASSP'07*, Honolulu, Hawaii, USA, pp. I–1205, 2007.

[12] T. Celik and H. Demirel, "Fire detection in video sequences using a generic color model," *Fire Safety Journal*, vol. 44, no. 2, pp. 147–158, 2009.

[13] B. U. Töreyin, Y. Dedeoğlu, U. Güdükbay and A. E. Cetin, "Computer vision based method for real-time fire and flame detection," *Pattern Recognition Letters*, vol. 27, no. 1, pp. 49–58, 2006.

[14] P. Foggia, A. Saggese and M. Vento, "Real-time fire detection for video-surveillance applications using a combination of experts based on color, shape, and motion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 9, pp. 1545–1556, 2015.

[15] K. Dimitropoulos, P. Barmpoutis and N. Grammalidis, "Spatiotemporal flame modeling and dynamic texture analysis for automatic video-based fire detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 2, pp. 339–351, 2014.

[16] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid scene parsing network," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, pp. 2881–2890, 2017.

[17] L. -C. Chen, G. Papandreou, F. Schroff and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv:1706.05587, 2017.

[18] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin *et al.,* "Expectation-maximization attention networks for semantic segmentation," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Seoul, Korea, pp. 9167–9176, 2019.

[19] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao *et al.,* "Dual attention network for scene segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 3146–3154, 2019.

[20] Q. Hou, L. Zhang, M. -M. Cheng and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 4003–4012, 2020.

[21] Z. Chen, Q. Xu, R. Cong and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," arXiv:2003.00651, 2020.

[22] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan *et al.,* "Basnet: Bundary-aware salient object detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 7479–7489, 2019.

[23] J. Wei, S. Wang and Q. Huang, "F$^3$net: Fusion, feedback and focus for salient object detection," *In Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12321–12328, 2020.

[24] Y. Pang, X. Zhao, L. Zhang and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 9413–9422, 2020.

[25] Q. Liu, X. Xiang, J. Qin, Y. Tan, J. Tan *et al.,* “Coverless steganography based on image retrieval of DenseNet features and DWT sequence mapping,” *Knowledge-Based Systems*, vol. 192, no. 1, pp. 105375–105389, 2020.

[26] R. Rajaragavi and S. P. Rajan, “Optimized u-net segmentation and hybrid res-net for brain tumor mri images classification,” *Intelligent Automation & Soft Computing*, vol. 32, no. 1, pp. 1–14, 2022.

[27] R. A. Naqvi, D. Hussain and W. Loh, “Artificial intelligence-based semantic segmentation of ocular regions for biometrics and healthcare applications,” *Computers, Materials & Continua*, vol. 66, no. 1, pp. 715–732, 2021.