fire

MDPI

Article

Smoke Image Segmentation Algorithm Suitable for Low-Light Scenes

Enyu Li and Wei Zhang *

School of Microelectronics, Tianjin University, Tianjin 300072, China
* Correspondence: tjuzhangwei@tju.edu.cn

Abstract: The real-time monitoring and analysis system based on video images has been implemented to detect fire accidents on site. While most segmentation methods can accurately segment smoke areas in bright and clear images, it becomes challenging to obtain high performance due to the low brightness and contrast of low-light smoke images. An image enhancement model cascaded with a semantic segmentation model was proposed to enhance the segmentation effect of low-light smoke images. The modified Cycle-Consistent Generative Adversarial Network (CycleGAN) was used to enhance the low-light images, making smoke features apparent and improving the detection ability of the subsequent segmentation model. The smoke segmentation model was based on Transformers and HRNet, where semantic features at different scales were fused in a dense form. The addition of attention modules of spatial dimension and channel dimension to the feature extraction units established the relationship mappings between pixels and features in the two-dimensional spatial directions, which improved the segmentation ability. Through the Foreground Feature Localization Module (FFLM), the discrimination between foreground and background features was increased, and the ability of the model to distinguish the thinner positions of smoke edges was improved. The enhanced segmentation method achieved a segmentation accuracy of 91.68% on the self-built dataset with synthetic low-light images and an overall detection time of 120.1 ms. This method can successfully meet the fire detection demands in low-light environments at night and lay a foundation for expanding the all-weather application of initial fire detection technology based on image analysis.

Keywords: low-light image enhancement; smoke segmentation; Cycle-Consistent Generation Adversarial Network; vision transformer; attention module

Citation: Li, E.; Zhang, W. Smoke Image Segmentation Algorithm Suitable for Low-Light Scenes. *Fire* **2023**, *6*, 217. https://doi.org/10.3390/fire6060217

Academic Editor: Grant Williamson

Received: 6 April 2023
Revised: 14 May 2023
Accepted: 24 May 2023
Published: 25 May 2023

1. Introduction

Fire accidents are one of the major disasters that seriously endanger the safety of people's lives and property in daily life. Prevention and the timely alarm of fire accidents are the top priorities to protect people's safety. Before most fire accidents, a mass of smoke will be produced at the ignition point. Consequently, monitoring smoke can detect fires quickly to avoid excessive fire spread and critical property damage. With the rapid development of computer vision and artificial intelligence, fire detection technology based on image processing and object detection has been widely studied. Smoke detection technology based on video images has been gradually replacing traditional temperature and smoke sensors with its faster response speed, more comprehensive detection range, lower cost of use, and weaker environmental restrictions.

Object detection mainly relies on color information, edge contour, texture features, and motion information of smoke images to extract features [1,2]. Therefore, it is challenging for object detection algorithms to accurately separate smoke areas from backgrounds during feature marking and extraction due to smoke's variable scale, strong diffusion, blurred boundary, and broad color change with concentration.

As the semantic segmentation methods gradually become a hot spot in image segmentation, this pixel-by-pixel classification method can effectively avoid the influence of

background on smoke features and overcome the weakness that the object detection algorithm is only suitable for rigid targets. The semantic segmentation methods based on deep learning are more efficient in the segmentation task of smoke images than the traditional ones. During the training process, the deep learning models can select the features which need to be learned according to different distributions of the datasets. Therefore, the precise separation of smoke and non-smoke areas can be realized, and the segmentation method is accurate, fast, and robust [3,4].

Current semantic segmentation methods can obtain high accuracy when applied to the segmentation task of bright and clear smoke images during the daytime. However, their segmentation effects for low-light smoke images have difficulty reaching the practical application levels. Images captured in low-light conditions tend to have longer exposure times, which results in more noise and blur, as well as unclear edges and features, affecting the performance of the segmentation model. In addition, many unattended areas such as warehouses and computer rooms are in low-light environments, making it more important to improve the detection ability of segmentation algorithms for early smoke in low-light or nighttime environments, which is especially significant for fire accident detection and alarm tasks.

Considering these points, we proposed an innovative semantic segmentation method which is specifically designed to work in low-light environments for smoke images. It will provide theoretical and technical support for detecting smoke areas in real low-light scenes. The main contributions of our work can be summarized into three key points:

- We have overcome the challenge of extracting features from low-light smoke by separating the entire detection task and completing the low-light smoke image segmentation task through an image enhancement network cascaded with a semantic segmentation network.
- We have designed a low-light image enhancement network based on unsupervised transfer learning methods. Our modified CycleGAN [5] algorithm has significantly improved the brightness and contrast of the smoke image, making it more suitable for subsequent segmentation tasks.
- We have designed a multi-scale feature extraction network based on a Transformer, which is capable of handling smoke feature extraction tasks in complex scenes. By fusing the semantic features of different resolution branches, the extraction ability of our network on the global features of smoke is enhanced.

2. Related Works

Many scholars have proposed various methods to enhance the segmentation ability to overcome the difficulties of image segmentation tasks in low-light environments. Ref. [6] proposed an unsupervised nighttime semantic segmentation model called DANIA. DANIA uses image relighting networks with light loss functions to narrow the image intensity distribution gap in different domains. Additionally, it combines the image relighting network and Convolutional Neural Networks (CNNs) to perform semantic segmentation as a generator. DANIA also designs a discriminator that performs adversarial learning to distinguish whether the segmentation prediction is from the source or target domain. While DANIA has successfully achieved state-of-the-art performance on night driving test datasets, it is not applicable for smoke segmentation due to the variable scales, irregular shapes, and texture information, which are greatly affected by brightness and concentration.

To improve the model's ability to learn from night images, Ref. [7] introduced a self-attention mechanism that considers position information based on Deeplab v3+ [8]. Additionally, a lighting adaptation mechanism was added to reduce the differences in the feature maps extracted by the shallow layers of the network. The model also addressed the differences between normal and low-light feature maps by using an illumination reflection weight map, which improves the feature extraction ability of unevenly illuminated positions. However, this method only considers the illumination and position feature information based on low-light images. Although the attention mechanism enhances the feature

extraction ability during the end-to-end training process, more improvements are needed to solve the problems of inaccurate low-light feature extraction and difficult recovery of edge details.

Ref. [9] proposed a semantic segmentation network that combines visible and infrared images to improve segmentation accuracy in low-light environments. The network design includes two parallel encoders that extract pixels of both modes separately, followed by a fusion process where each pixel of the infrared images is fused into the visible images. This complementary process combines the location information and pixel associations of the two domain images and effectively extracts features from complex nighttime backgrounds. However, the dual-mode segmentation model requires the support of real datasets containing visible and infrared images, which are challenging to construct and prone to class imbalance. The increase in dataset capacity also leads to longer image acquisition and training times, as well as expanded training complexity.

Furthermore, the segmentation network built in Refs. [7–9] mainly utilizes convolutional units. CNNs have advantages in spatial position representation. However, it is difficult for them to capture global feature context information due to the locality of convolution operations [10]. At the same time, CNNs can reduce the amount of computation through pooling operations in the feature extraction process, which will instead lead to the loss of detailed features [11,12]. For smoke targets, the loss of small-scale features can result in poor segmentation of the details of edges. With the application of a Transformer to the visual field, it can capture the long-distance feature dependence of complex spatial transformation domains between feature maps with self-attention and establish the global feature representation of the candidate regions [13]. However, Transformers may ignore the local feature information, which reduces the ability to distinguish between the foreground and the background in images. Therefore, designing a network that integrates the Transformer and CNNs can complete the complementary fusion of global semantic information and local detail features.

3. Method

In order to address the issue of unclear features in low-light images, we cascade an image enhancement network prior to the segmentation task. Our modified CycleGAN algorithm is designed to enhance low-light images, making them more similar to their daytime counterparts. Then, more accurate smoke segmentation results can be obtained with our semantic segmentation network. Figure 1 depicts the entire flow of our low-light smoke image segmentation algorithm.

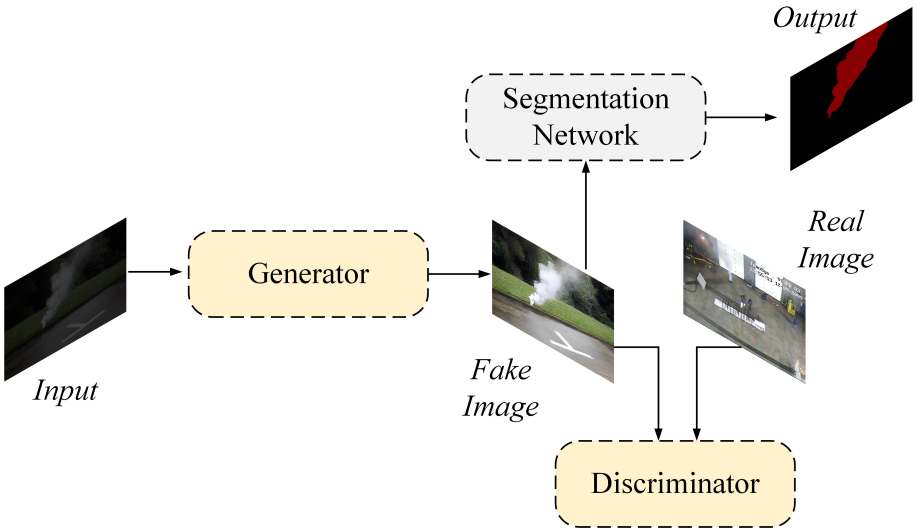


Figure 1. Entire flow of our low-light smoke image segmentation algorithm.

By cascading the two network models, it is possible to avoid the need for complex fusion methods while effectively reducing the model's complexity and computational scale. Furthermore, our approach allows for the enhancement and segmentation networks to be replaced based on different task scenes and requirements, improving the method's generalization and flexibility. Our highly efficient and adaptable approach makes it a valuable tool for applications in low-light environments.

3.1. Image Enhancement Network

To improve the smoke characteristics in low-light environments, we modify the CycleGAN model as the main component of our image enhancement network. The original CycleGAN has difficulty preserving image details during the enhancement process, which results in artifacts, blur, and noise in enhanced images, making the subsequent segmentation task challenging. Inspired by EnlightenGAN [14], we modify the generator network with an encoder–decoder structure to restore the smoke details accurately. We also add a brightness equalization branch to balance the brightness of each part of an image. To ensure that the smoke features are preserved during the enhancement process, we attach a similarity discriminant branch to the discriminator network based on a two-branch form. Figure 2 shows the structure of our low-light smoke image enhancement algorithm.

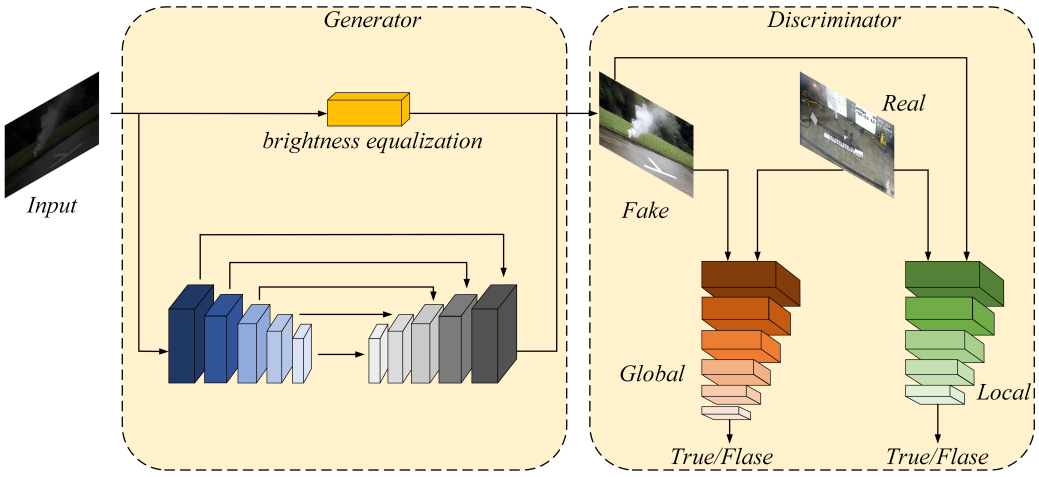


Figure 2. Structure of our low-light smoke image enhancement algorithm.

The encoder follows the same structure as EnlightenGAN, while the decoder part's upsampling process is completed using PixelShuffle [15]. Through convolution and multi-channel recombination, the low-resolution feature map transforms into a high-resolution feature map. PixelShuffle is used mainly to handle the loss of details during upsampling based on a single feature map. The skip connection between the encoding and decoding stages ensures that original image features are transmitted, allowing some of the lost details from downsampling to be recovered. The brightness equalization branch applies additional weights to darker areas of the images, which can make them brighter—at the same time, suppressing the enhancement effect of brighter areas to avoid overexposure problems. The structure of the brightness equalization branch is shown in Figure 3.

To ensure that the brightness equalization branch is sensitive to variable levels of darkness in low-light images, we choose the Parametric Rectified Linear Unit (PReLU) [16] as the activation function. By adjusting the parameters in PReLU adaptively according to different brightness levels in different areas of the images, the equalization branch's sensitivity to different brightness levels is significantly improved. Inspired by EnlightenGAN's dual discriminant branch, our discriminator enhances the global brightness level and corrects image details. However, the generator's skip connection can only partially counteract the loss or change of image features during enhancement. To further recover detailed features, we design a similarity discrimination branch that works in couple with

the subsequent segmentation model in the discriminator network, ensuring that the image enhancement process will not significantly impact the smoke features. The discriminator structure, including the similarity discriminant branch, is illustrated in Figure 4.

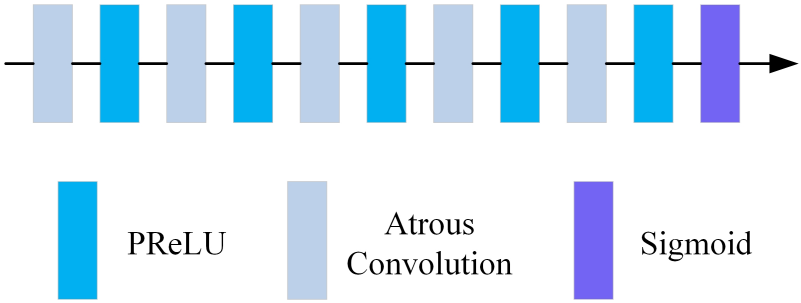


Figure 3. The structure of the brightness equalization branch.

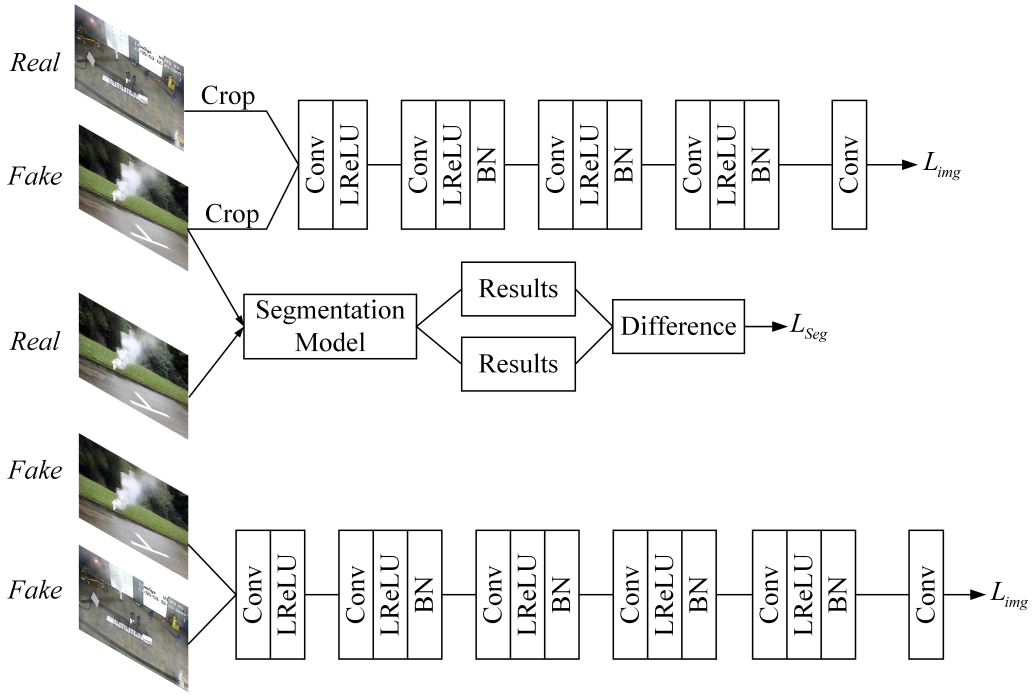


Figure 4. Structure of the discriminator.

For the local discriminator, we modify the loss function of the original LS-GAN [17] with the least squares loss:

$$L_{Local}(G, D_Y^{Local}) = \frac{1}{2} E_{x_f \sim P_{fdata}(x_f)}[(D_Y(x_f) - 1)^2] \tag{1}$$

For the global discriminator, to improve the quality of the generated images and reduce the training time, we refine the original loss function with the standard function of the relativistic adversarial network [18]. The least square loss of the global discriminator according to the corresponding regression target is

$$\begin{aligned} L_{Global}(G, D_Y^{Global}) = & \frac{1}{2} E_{x_r \sim P_{rdata}(x_r)}[(D_{RY}^{avg}(x_r, x_f) - 1)^2] \\ & + \frac{1}{2} E_{x_f \sim P_{fdata}(x_f)}[(D_{RY}^{avg}(x_f, x_r))^2] \end{aligned} \tag{2}$$

where D represents the discriminator network. x_r and x_f represent the distribution of real and fake images.

With unsupervised training, the transfer effect of domains is controlled by the difference between the quadratically generated and real images. To ensure that the characteristic information of the smoke area remains unchanged, we append the cycle consistency loss based on the LS-GAN loss, which helps to prevent the subsequent segmentation process from being adversely affected. The cyclic consistency loss expression is as follows:

$$L_{Cycle}(G,F) = E_{x_r \sim P_{rdata}(x_r)}[\|F(G(x_r)) - x_r\|_1] + E_{y_r \sim P_{rdata}(y_r)}[\|G(F(y_r)) - y_r\|_1] \quad (3)$$

where G and F represent the generators in two directions, x represents the low-light images, y represents the images in the daytime, and the difference between images is measured by 1-norm.

As shown in Figure 4, our similarity discrimination branch effectively mitigates the loss and alteration of image specifics caused by downsampling in enhancement networks. Therefore, we increase the object similarity loss using the segmentation model:

$$L_{seg}(G,F) = \frac{1}{H \times W \times C}\left\{[\Phi(F(G(x_r))) - \Phi(x_r)]^2 + [\Phi(G(F(y_r))) - \Phi(y_r)]^2\right\} \quad (4)$$

where Φ represents the pixel-by-pixel classification results obtained by the subsequent semantic segmentation algorithm of the fake and real images. H, W, and C are the dimensions of the corresponding feature maps. We use Mean Squared Error (MSE) to represent the absolute difference between two image segmentation results.

L_{seg} improves the recovery of details of the images. However, it also leads to blurry output images since quality evaluation indicators such as MSE only consider the difference between pixels at a single point without considering the correlation between them. They ignore the correlation between pixels. Therefore, we incorporate a loss based on the Structure Similarity Index Measure (SSIM) between the real images and the quadratically generated images:

$$L_{Relevance}(G,F) = L_{cyc}(G,F) + L_{one-way}(G,F) \quad (5)$$

$$L_{cyc}(G,F) = [1 - SSIM(x_r, F(G(x_r)))] + [1 - SSIM(y_r, G(F(y_r)))] \quad (6)$$

$$L_{one-way}(G,F) = [1 - SSIM(x_r, F(y_r))] + [1 - SSIM(y_r, G(x_r))] \quad (7)$$

where L_{cyc} indicates the structural similarity loss added to the cyclic generation discriminant, and $L_{one-way}$ represents the structural similarity difference for G and F added in the one-way generation process.

In summary, the loss function of the image enhancement network we proposed is as follows:

$$\begin{aligned} L_{Total} = {} & \frac{1}{2}\left[L_{Local}(G, D_Y^{Local}) + L_{Local}(F, D_X^{Local})\right] \\ & + \frac{1}{2}\left[L_{Global}(G, D_Y^{Global}) + L_{Global}(F, D_X^{Global})\right] \\ & + \lambda_1 L_{Cycle}(G,F) + \lambda_2 L_{seg}(G,F) + \lambda_3 L_{Relevance}(G,F) \end{aligned} \quad (8)$$

where λ_1, λ_2, and λ_3 are the balance parameters that control the proportion of different loss functions. Based on the various experiments we conducted, we found that the parameters $\lambda_1 = 5$, $\lambda_2 = 10$, and $\lambda_3 = 10$ are the most suitable for our enhancement network. We use two branches to calculate the LS-GAN loss based on the two-branch discriminator, while the remaining three loss functions are calculated solely based on the generator network.

3.2. Semantic Segmentation Network

When it comes to smoke segmentation, it can be quite challenging due to the extensive range of sizes, strong diffusion, and changeable shapes of smoke. Background information can also significantly affect the segmentation results, making it even more difficult to obtain satisfactory results. Therefore, the networks which work well for other targets may not

be effective for smoke segmentation tasks. To meet the requirements of video monitoring and analysis systems for detecting and announcing fire accidents, we propose a semantic segmentation network based on HRNet [19] and HRFormer [20]. It allows for the extraction of multi-scale features of smoke, making it possible to segment smoke images enhanced by our image enhancement network accurately. The structure of our semantic segmentation algorithm is shown in Figure 5.

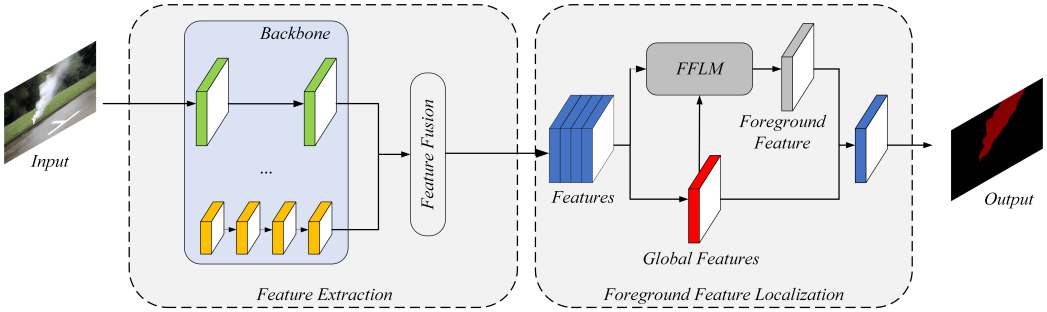


Figure 5. Structure of our semantic segmentation algorithm.

Due to the variable scales and shapes of smoke, feature extraction concentrating on one single scale makes it challenging to obtain accurate segmentation results. We modify HRFormer's Transformer block to handle multi-scale changes in the targets effectively. Additionally, using 3 × 3 depth-wise convolution for information interaction between windows in HRFormer is not sufficient, as it barely covers the internal features of the windows. Meanwhile, increasing the sizes of convolution kernels will import more background information. Therefore, we use the Shifted Window-based Multi-head Self-Attention (SW-MSA) [21] of a Swin Transformer (SW-Trans) instead. The entire structure of our feature extraction module is shown in Figure 6.

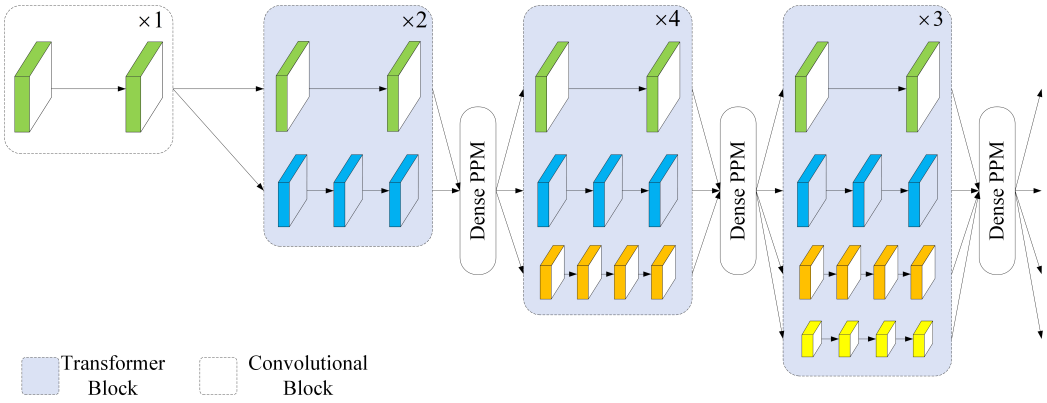


Figure 6. Structure of the feature extraction module.

Compared with the classical network structure of HRNet, our shallow and deep module parts utilize different numbers of feature extraction units. The high-resolution branches generate larger feature maps with fewer output dimensions, which allow for better preservation of pixel spatial positioning information. On the other hand, the low-resolution branches generate smaller feature maps with a larger number of output dimensions, making them better at extracting abstract semantic features. We gradually increase the number of Transformer blocks in our network from high-resolution to low-resolution branches, which allows us to extract both local details and global semantic information in parallel.

The parameters of our feature extraction modules are shown in Table 1. M_1, M_2, M_3, and M_4 represent the numbers of modules in different stages, and B_1, B_2, B_3, and B_4 represent the numbers of Transformer blocks in each branch of different modules.

Table 1. Parameter settings of our feature extraction modules.

Resolution	Stage 1	Stage 2	Stage 3	Stage 4
4×	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times B_1 \times M_1$	$[T-Block] \times B_1 \times M_2$	$[T-Block] \times B_1 \times M_3$	$[T-Block] \times B_1 \times M_4$
8×		$[T-Block] \times B_2 \times M_2$	$[T-Block] \times B_2 \times M_3$	$[T-Block] \times B_2 \times M_4$
16×			$[T-Block] \times B_3 \times M_3$	$[T-Block] \times B_3 \times M_4$
32×				$[T-Block] \times B_4 \times M_4$

Consistent with Figure 5, M_1, M_2, M_3, and M_4 are 1, 2, 4, and 3, and B_1, B_2, B_3, and B_4 are 2, 3, 4, and 4. The numbers of Transformer block channels at different resolutions are, respectively, 32, 64, 128, and 256 at 4, 8, 16, and 32 times. The numbers of self-attention heads are, respectively, 1, 2, 4, and 8. The Multi-Layer Perception (MLP) extension ratio in all transformer blocks is 4.

By incorporating the Local Spatial Attention Module (LSAM) and the Global Channel Attention Module (GCAM), we improve our Transformer block's ability to extract local spatial information and aggregate global context information. The structure of our Transformer block is shown in Figure 7.

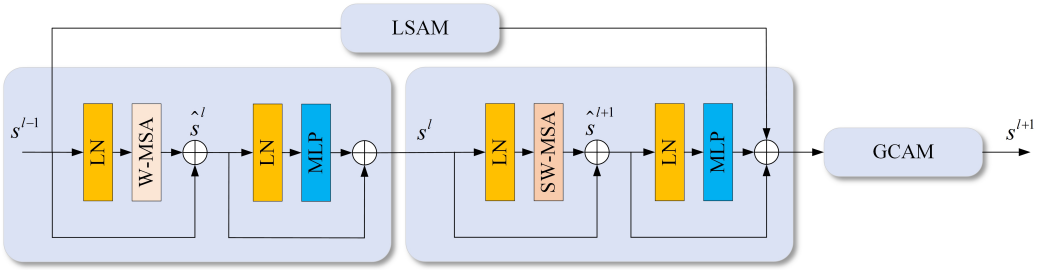


Figure 7. Structure of our Transformer block.

The LSAM structure is shown in Figure 8, and the GCAM structure is shown in Figure 9.

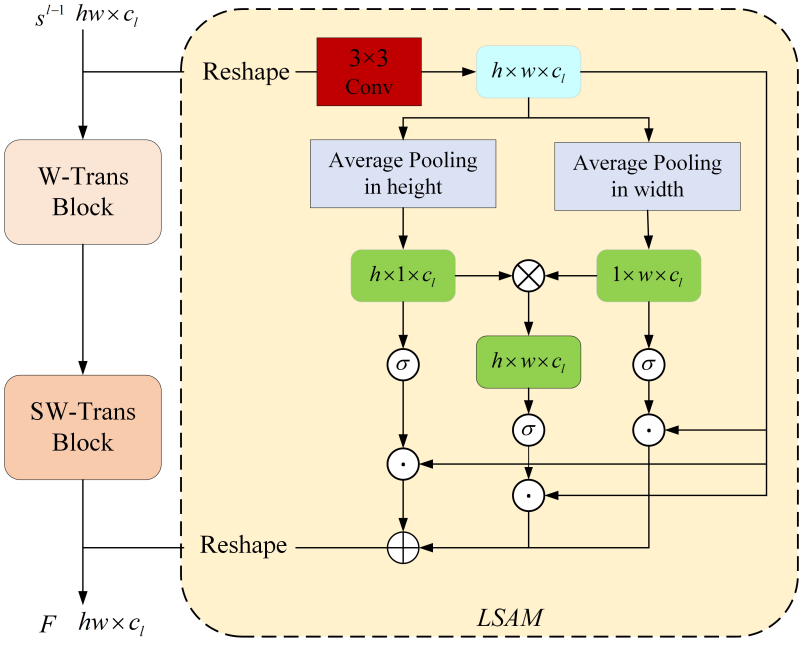


Figure 8. Structure of LSAM.

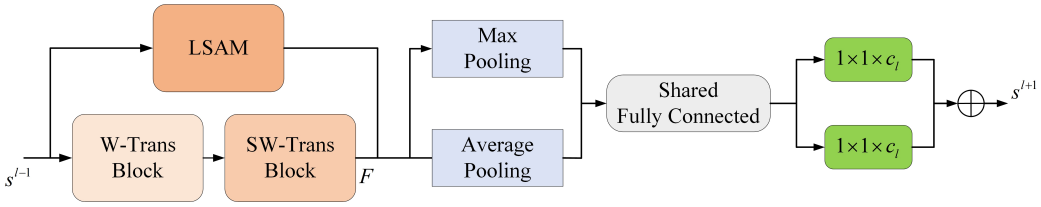


Figure 9. Structure of GCAM.

LSAM reshapes the tokens into feature maps before extracting structural information through a 3 × 3 convolutional layer. Depending on the stage and branch where the transformer block is located, H and W take different values. Once the structural information is extracted, global average pooling is performed in the width and height directions to obtain statistical information:

$$v_{h_i}^k = \frac{1}{w} \sum_{j=0}^{w-1} \hat{z}^k(i,j) \tag{9}$$

$$v_{w_j}^k = \frac{1}{h} \sum_{i=0}^{h-1} \hat{z}^k(i,j) \tag{10}$$

where $v_{h_i}^k \in R^{h \times 1 \times c_l}$ and $v_{w_j}^k \in R^{1 \times w \times c_l}$ are one-dimensional spatial attention vectors in the height and width directions. $\hat{z}$ is the feature map obtained by the convolutional layer, and the Batch Normalization (BN) and Gaussian Error Linear Unit (GELU) activation functions are used after convolution. The corresponding operating ranges are $0 \leq i < h$, $0 \leq j < w$, and $0 \leq k < c_l$.

A feature map of $h \times w \times 1$ is obtained by matrix multiplication using the tensors in the width and height directions within each channel:

$$v_{w_j}^k = \frac{1}{h} \sum_{i=0}^{h-1} \hat{z}^k(i,j) \tag{11}$$

After obtaining two spatial attention vectors in the height and width direction and a two-dimensional spatial attention feature map with information interaction in two directions, the next step is to activate the three attention weight vectors with the sigmoid function. These weight vectors are multiplied point by point to the original feature map. Then the spatial position attention map $M \in R^{h \times w \times c_l}$ of LSAM is obtained. The output feature map of LSAM is calculated by adding M to the output s^{l+1} of the SW-Trans branch yields:

$$F = s^{l+1} \oplus \Phi(\sigma(v_h) \odot \hat{z} + \sigma(v_w) \odot \hat{z} + \sigma(v_{h,w}) \odot \hat{z}) \tag{12}$$

where $\oplus$ represents the point-by-point summation, $\odot$ represents the point-by-point product, and Φ represents the reshaping of the feature maps.

The attention feature vectors in different directions can capture long-distance feature dependencies in their respective direction. By integrating the spatial position information of two directions after multiplying, a two-dimensional spatial mapping between the features and the pixels can be established. LSAM can effectively extract the regional features of smoke, which helps suppress background information and noise to a certain extent.

GCAM performs global average pooling and global max pooling of LSAM output features in the spatial dimension. By learning the weight distributions of the obtained max pooling features and average pooling features in the channel dimension through a shared fully connected layer, GCAM can reduce the dimensionality of the channel features and acquire two feature vectors. After applying GELU activation, the fully connected layer upgrades the feature maps and restores the initial number of channels. Finally, the two

attention vectors of features are added and activated by the sigmoid function before being applied to the output features of LSAM to obtain the output of the Transformer block:

$$s^{l+1} = F \otimes \sigma[f_0(f_1(\text{AvgPool}(F))) + f_0(f_1(MaxPool(F)))] \quad (13)$$

where $\otimes$ represents the channel attention weights multiplied by the corresponding feature map, f_0 and f_1 represent the two shared fully connected layers that perform dimensionality operations, and AvgPool and MaxPool represent global average pooling and global max pooling.

The Transformer block can adaptively assign feature weights on the channel and spatial domains according to the degree of correlation of features in the smoke image. It also allows for essential features to be enhanced and invalid information to be suppressed.

The segmentation task for the enhanced smoke images is challenging due to the blurred appearances. We modify the Pyramid Pooling Module (PPM) [22] and extend it into a dense style using DenseNet [23]. After each module of our backbone network, we implement a fusion module to upsample the low-resolution feature maps. By concatenating the feature maps of different scales and repeating the upsampling process, our network can increase the receptive field of the feature maps and extract more context information effectively. The structure of Dense PPM is shown in Figure 10.

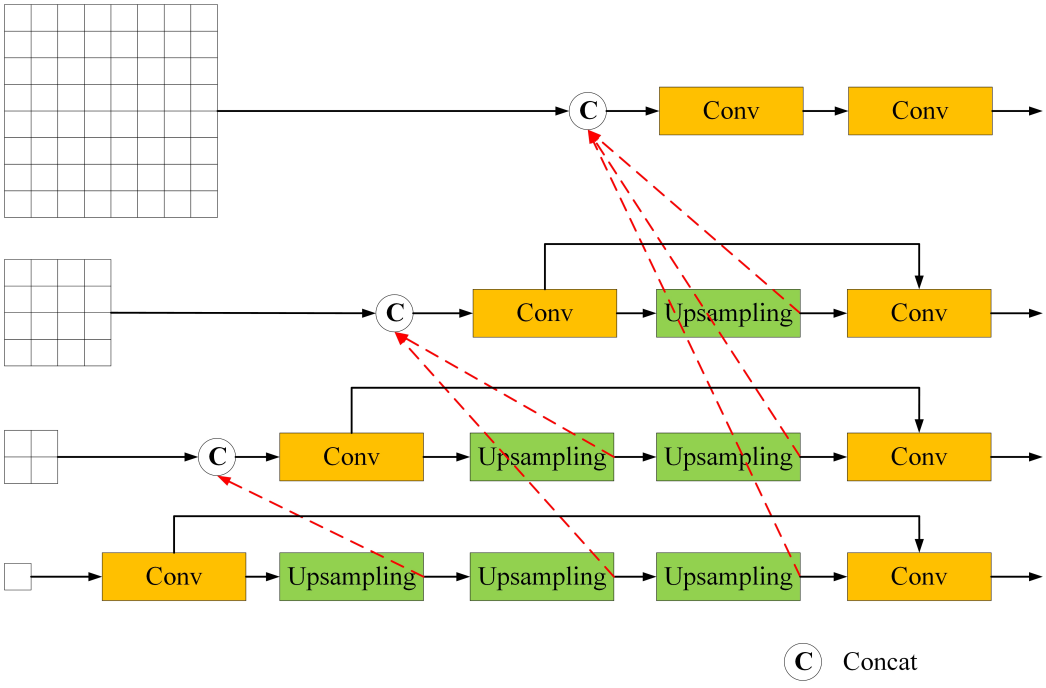


Figure 10. Structure of Dense PPM.

In the proposed network, we implement depth-wise separable convolution [24] instead of original convolutional units to reduce the computation amount during the training process, which results in a notable improvement in the network's training speed.

To accurately distinguish smoke from the background areas and ensure that the segmentation effect is not disturbed by external factors, we propose a foreground feature localization module called FFLM, which can precisely segment the thin smoke areas in the images through the calculation of correlation between each pixel and the smoke foreground. FFLM also helps to increase the differentiation between smoke areas and background information and prevent confusion. The specific structure of FFLM is shown in Figure 11.

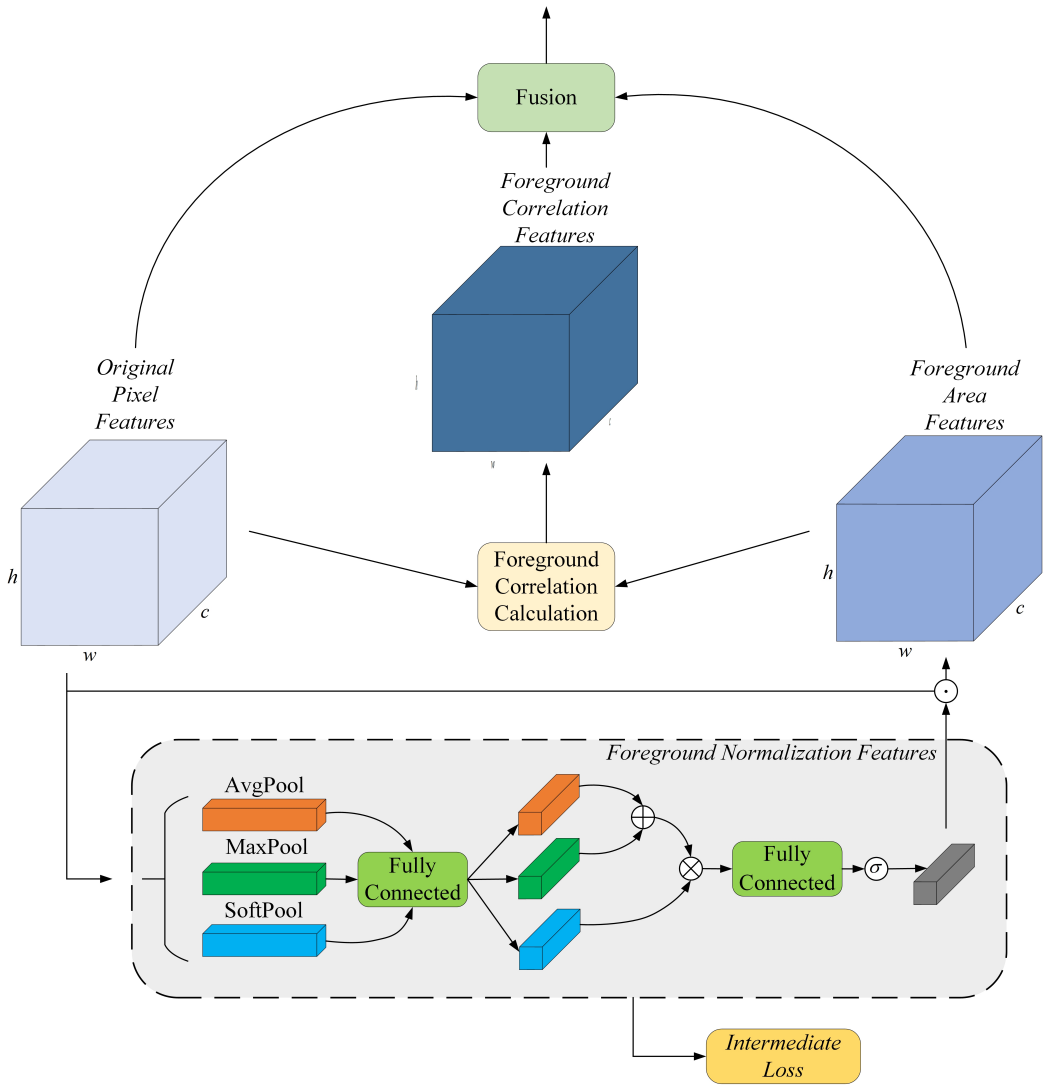


Figure 11. Structure of FFLM.

The output feature maps $F_{in} \in R^{h \times w \times c}$, which contain rich and detailed features and semantic information, are processed with average pooling, max pooling, and soft pooling [25] techniques in the channel dimension. The statistics obtained from these techniques are combined through a shared fully connected layer. The sum of average and max pooling is then multiplied by the index weights of soft pooling. With the obtained global weight description passing through a fully connected layer and a sigmoid function, the foreground normalized features of each channel of the output feature maps are presented.

Finally, the foreground normalized features are multiplied with the original pixel features F_{in} to obtain the foreground area features:

$$P_{avg,max} = f_1(\text{AvgPool}(F_{in})) + f_1(\text{MaxPool}(F_{in})) \tag{14}$$

$$P_{soft} = f_1(\text{SoftPool}(F_{in})) \tag{15}$$

$$\widehat{F} = \sigma[f_2(P_{avg,max} \odot P_{soft})] \odot F_{in} \tag{16}$$

where $\widehat{F}$ represents the foreground area features, $P_{avg,max} \in R^{1 \times 1 \times c}$ represents the statistical feature description after the sum of average pooling and max pooling, P_{soft} represents the description of the statistical features obtained by soft pooling. AvgPool, MaxPool, and SoftPool correspond to the operations of average pooling, max pooling, and soft pooling

in the channel dimension. f_1 and f_2 are fully connected layers. $\odot$ is the point-by-point product operation.

The degree of association between each pixel and the foreground areas can be calculated with the foreground area features and the original pixel features:

$$\Sigma = \frac{\exp(\varphi(\widehat{F}))}{\exp(\varphi(F_{in}))} \tag{17}$$

where Σ is the foreground correlation representation. φ is the foreground correlation calculation function implemented by convolution, BN, and Rectified Linear Units (ReLUs).

The output features enhanced by FFLM can be obtained by fusing foreground correlation features, foreground area features, and original pixel features:

$$F_{out} = F_{in} + \rho(\Sigma \cdot \delta(\widehat{F})) \tag{18}$$

where ρ and δ are the fusion functions implemented by convolution, BN, and ReLUs. F_{out} is the enhanced output features. FFLM can help improve the overall quality of images by enhancing the features in the foreground area while reducing the impact of the background on smoke feature extraction. Additionally, FFLM can improve the ability of segmentation in thin smoke areas.

The smoke segmentation task is a pixel-level dense binary classification, and each pixel needs to be classified between the foreground and background areas. Therefore, we use a binary cross-entropy loss to design the loss function of our segmentation network:

$$L = -\frac{1}{N}\sum_{i=1}^{n}[q_i \log(p_i) + (1 - q_i)\log(1 - p_i)] \tag{19}$$

where N is the number of pixels in the feature map. p_i is the probability that the pixel is predicted as smoke foreground. q_i is the ground truth of the pixel.

When analyzing images with smoke areas, it is important to consider the proportion of the smoke areas in relation to the background regions. If the loss function treats these areas with a consistent weight, the party with a larger proportion will play a more dominant role in the backward propagation process, resulting in a higher weight during the prediction. Therefore, we introduce weighted coefficients to the two parts of the loss function based on the relative sizes of the smoke areas and background regions. The coefficients allow our model to balance the feature learning process between the two regions and adapt to the unique characteristics of each image. The modified loss function includes a foreground weight coefficient α_f:

$$L = -\frac{1}{N}\sum_{i=1}^{n}[\alpha_f \cdot q_i \log(p_i) + (1 - q_i)\log(1 - p_i)] \tag{20}$$

We add an intermediate layer loss to the foreground feature localization process and weigh it against the loss on the segmentation results to balance the supervision throughout the network training process. Both the intermediate layer loss and the final loss use a foreground-weighted binary cross-entropy loss function:

$$L_m = -\frac{1}{N}\sum_{i=1}^{n}[\alpha_f \cdot q_i \log(p_i^m) + (1 - q_i)\log(1 - p_i^m)] \tag{21}$$

$$L_{out} = -\frac{1}{N}\sum_{i=1}^{n}[\alpha_f \cdot q_i \log(p_i^{out}) + (1 - q_i)\log(1 - p_i^{out})] \tag{22}$$

$$L_{total} = \alpha_u L_m + (1 - \alpha_u) L_{out} \tag{23}$$

where L_m and L_{out} are the intermediate layer loss and final loss. p_i^m and p_i^{out} are the probability value that the intermediate layer pixel and output feature pixel are classified to the foreground. α_u is the balance weight parameter of the union loss. After conducting numerous experiments, it has been determined that our segmentation model achieves optimal performance with a value of 0.25 for α_u when applied to our smoke dataset.

4. Results and Discussion

4.1. Dataset Settings

There is a lack of public image datasets available for the smoke segmentation task, and even fewer public datasets that include smoke images in low-light environments. Therefore, we utilize the smoke video dataset from the State Key Laboratory of Fire Science (SKLFS) [25] and add our own collected images to construct a dataset in the PASCAL VOC format. The training set of the image enhancement network includes 1000 smoke images in the daytime and 1000 synthetic low-light images. The test set includes 200 synthetic low-light smoke images. The semantic segmentation network's training set consists of 4000 smoke images in the daytime, and its test set consists of 200 enhanced synthetic low-light smoke images and 400 images in the daytime. Overall, our datasets include 70 scenes, and some of the images in the daytime are shown in Figure 12.

Figure 12. Examples of images in the daytime.

The differences in brightness between low-light images and daytime images are not constant. The values increase as brightness increases, meaning high-brightness pixels in an image are reduced in brightness more than low-brightness ones when transitioning to low-light environments. To simulate these conditions, we use gamma correction [26] methods based on existing synthetic low-light image generation techniques [27–29]. We transform the V channel of the images in HSV space and add Gaussian and Poisson noise to simulate the blur and noise captured by the camera in low-light environments. Our image conversion formula is as follows:

$$X_{out} = B_G(S \cdot (X_{in})^{\gamma}) + N_G + N_P \tag{24}$$

where X_{in} is the value of the V channel of an HSV image. X_{out} is an output synthetic low-light image. B_G is a Gaussian blur function that sets the standard deviation to a random value between 1.5 and 2. S and γ are the correction parameters corresponding to gamma correction and are, respectively, set to 0.8 and 0.65. N_G is Gaussian noise, whose kernel size is (5, 5), and the standard deviation is 1.25. N_P is Poisson noise, whose λ is 1.0.

The synthetic low-light images are shown in Figure 13. The detailed features of each part can still be obtained, but it may take some time to distinguish each area. Our network configurations during training and testing are shown in Table 2. The size of images is normalized to 640 × 480. The network batch size is set to 1. The epoch is set to 300. The initial value of the learning rate is 0.0005. The learning rate decays to 0.0001 when the epoch reaches 150.

Figure 13. Examples of synthetic low-light images.

Table 2. Configurations of network during training and testing.

Hardware/Software	Configurations
CPU	Intel Xeon Silver 4214R 12CPU@2.40 GHz
GPU	NVIDIA TITAN XP
Programming Language	Python 3.7
Deep Learning Framework	Pytorch 1.13.1

4.2. Comparison Experiments of Image Enhancement Algorithm

In order to verify the effectiveness and superiority of our enhancement algorithm, we select several methods such as LIME [30], SRIE [31], RetinexNet [32], EnlightenGAN [14], LLFlow [33], Zero-DCE [34], and CycleGAN [5] for comparative experiments. Signal Noise Ratio(SNR), Peak Signal Noise Ratio(PSNR), and SSIM are used as image quality evaluation indicators. The average evaluation results on our synthetic low-light smoke image dataset are shown in Table 3. Our enhanced network has advantages over traditional methods SIRE and LIME in various evaluation indicators. Compared with fully supervised algorithms such as RetinexNet, LLFlow, and Zero-DCE, our algorithm also has a certain degree of advantage. In comparison to the unsupervised EnlightenGAN and CycleGAN, although there is not much difference in SNR, our algorithm has excellent advantages in SSIM and NIQE indicators. The results in Table 3 clearly demonstrate that our proposed enhancement algorithm is superior to other low-light image enhancement algorithms in various evaluation indicators.

Table 3. Comparison of experimental results of image enhancement networks.

Networks	MSE	SNR	PSNR	SSIM	NIQE
LIME [30]	931.994	14.350	18.989	0.884	6.154
SIRE [31]	877.623	12.906	19.043	0.837	6.073
RetinexNet [32]	519.397	15.515	21.302	0.909	5.760
EnlightenGAN [14]	73.162	24.314	30.508	0.935	4.856
LLFlow [33]	353.466	17.322	23.436	0.927	4.997
Zero-DCE [34]	424.280	16.599	23.532	0.915	5.015
CycleGAN [5]	44.154	25.929	32.111	0.937	5.024
Ours	35.996	26.617	32.816	0.964	4.699

Figure 14 shows the enhancement results on our synthetic low-light smoke image dataset.

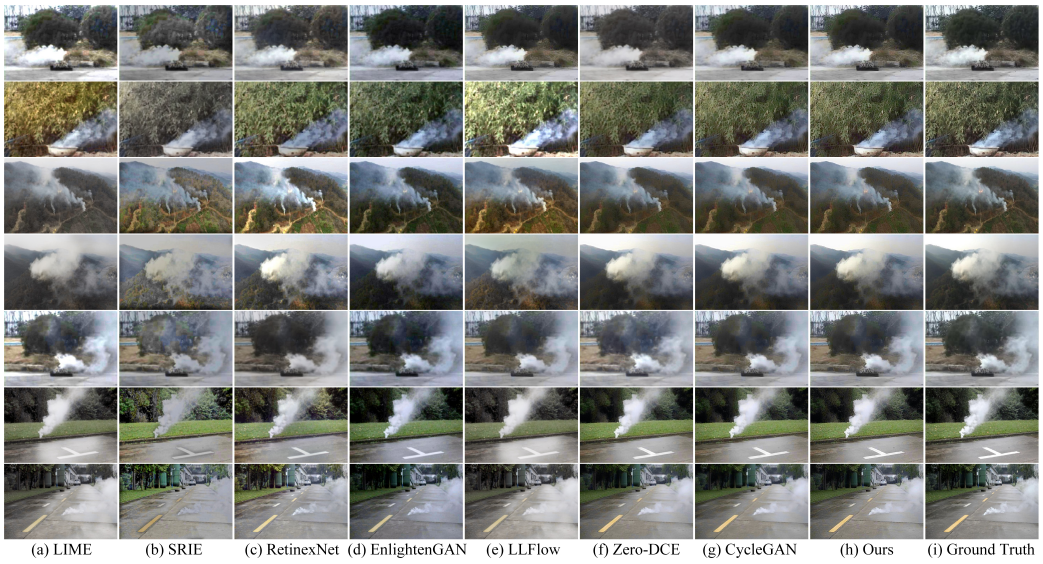


Figure 14. Examples of enhancement results on our synthetic low-light smoke image dataset.

The enhancement effects of SRIE and LIME on low-light smoke images are not satisfactory. The enhanced images tend to be overexposed or underexposed, and the boundaries between areas of similar colors appear blurred. Similarly, RetinexNet and Zero-DCE also do not produce desired results. Their enhanced images show severe blur and color distortion, as shown in the fifth row of column (c) and the first row of column (f). CycleGAN can recover the color of images effectively and suppress noise in some uncomplicated scenes. However, when the background information becomes complex, the images enhanced by CycleGAN tend to produce more severe blurring, such as the third and fourth rows of column (g), which can negatively impact subsequent segmentation tasks.

LLFlow and EnlightenGAN can cause slight chromatic aberration, and their enhanced images have high contrast. LLFlow tones appear warmer, while EnlightenGAN tones are colder than real images. However, our method stands out as it can restore image details better, and the enhanced images contain a lower noise level. Most importantly, the enhanced images obtained by our method have the highest structural similarity, indicating that our enhancement operation has the most negligible impact on subsequent segmentation tasks, which further demonstrates the advantages of our method over other segmentation algorithms in low-light smoke image segmentation tasks.

4.3. Ablation Experiments of the Image Enhancement Algorithm

The proposed low-light image enhancement algorithm is subjected to ablation experiments in order to effectively verify the contribution of each individual part. The ablation objects are mainly aimed at the brightness equalization branch and the object similarity

discrimination between the generated images and the original real samples during the loop generation. The quantitative results of the ablation experiments are shown in Table 4, and the qualitative results are shown in Figure 15.

Table 4. Ablation experimental results of our image enhancement algorithm.

Networks	Brightness Equalization Branch	Object Similarity Discrimination	SNR	PSNR	SSIM
Net1			12.703	16.881	0.871
Net2	✓		18.247	23.372	0.938
Net3/Ours	✓	✓	26.617	32.816	0.964

(a) Net1 (b) Net2 (c) Net3/Ours (d) Real Images

Figure 15. Examples of the ablation experimental results.

Net1 is a generator network without a brightness equalization branch or attention module, which only contains the encoder–decoder structure of EnlightenGAN. Its discriminator still uses the two-branch discriminator structure of PatchGAN to fuse local features with global information. Net2 adds a parallel brightness equalization branch to Net1. Net3 is our network for the low-light smoke image enhancement task.

Based on the experimental data in the SSIM column of the table, it appears that the structural control of the generated images with the subsequent segmentation network

is quite useful, and Net3 achieves an SSIM increase of 0.026 compared to Net2. The ablation experiment shown in Figure 15 compares the effects of different networks on low-light image enhancement. Net2's brightness equalization branch helps to balance the image's brightness enhancement, avoiding the problem of local brightness being too high or too low and the problem of overexposure or underexposure. Net3's similarity discrimination branch helps to segment the local details more finely. The results of the ablation experiments visually demonstrate the effectiveness of each part of our low-light image enhancement algorithm.

4.4. Comparison Experiment of Semantic Segmentation Algorithm

In order to verify the effectiveness of our segmentation algorithm in multiple scenes in low-light environments, we compare PSPNet [22], HRNet [19], Deeplab v3+ [10], SegNet [35], SW-Trans [21], and HRFormer [20] with our algorithm on the self-built test set. Quantitative comparison results on test images in the daytime are shown in Table 5. We use Mean Intersection over Union (mIoU) as the primary evaluation criterion. At the same time, Floating Point Operations (FLOPs), parameter amount (Params), and detection time (T) are considered as auxiliary evaluation indicators.

Table 5. Comparison of experimental results of segmentation networks.

Networks	mIoU/%	FLOPs/G	Params/M	T/ms
PSPNet [22]	90.12	300.3	68.0	89.90
HRNet [19]	90.83	110.2	65.9	54.46
Deeplab v3+ [10]	90.34	298.1	62.6	89.57
SegNet [35]	88.91	35.8	29.0	31.04
SW-Trans [21]	91.97	97.9	59.9	51.33
HRFormer [20]	92.08	74.4	43.2	44.75
Ours	92.93	125.7	63.4	58.16

The pooling operation of PSPNet can lead to the loss of local features between layers, which negatively affects its ability to identify the edges of smoke. SegNet eliminates the fully connected layer and uses pooled indexes to replace feature map concatenating operations, which significantly reduces the number of network operations. Although SegNet achieved a segmentation accuracy of 88.91%, it is the fastest in comparison experiments. Deeplab v3+ uses the Atrous Spatial Pyramid Pooling (ASPP) module to fuse multi-scale features. However, atrous convolution introduces many background contexts while expanding the receptive field. It is unsuitable for smoke, whose texture features will be greatly affected by background information. Therefore, the improvement in segmentation accuracy brought by ASPP modules is not apparent. Additionally, the computational cost of Deeplab v3+ has increased significantly due to the use of the Xception modules instead of ResNet-101, which results in a slower segmentation speed than other CNN models. HRNet is more effective for smoke segmentation than other CNN models, achieving an mIoU of 90.83%. The better performance indicates that using multiple resolution branches for parallel feature extraction for smoke is effective.

SW-Trans and HRFormer achieve higher segmentation accuracy than the CNN models, which indicates that in the segmentation task for smoke, better global feature extraction ability can bring higher segmentation performance. By introducing Window-based Multi-Head Self-Attention (W-MSA) into the self-attention stage, the calculation is limited to a small scale, and the calculation complexity of the model is effectively controlled while introducing the CNN effect locally. HRFormer uses 3×3 depth-wise convolution at the window's interaction, while SW-Trans moves the windows by the masks and feature shift operations, making its Transformer blocks more computationally intensive than HRFormer. However, HRFormer still acquires higher segmentation accuracy than SW-Trans because the medium and low-resolution branches supplemented the semantic information of different scales for high-resolution streams.

According to Table 5, our segmentation network has demonstrated a remarkable accuracy of 92.93% on our self-built smoke dataset, outperforming all the other algorithms. In order to further enhance the accuracy, we incorporate LSAM and GCAM attention modules into Transformer blocks, allowing us to extract smoke features. Additionally, we enhance the smoke foreground areas to better differentiate texture from the background in thinner areas of the smoke edges. These improvements enable us to achieve a significantly higher segmentation accuracy than HRFormer.

Since our segmentation model modified the Transformer blocks and added a foreground enhancement module after the feature extraction, our model's complexity is higher than that of SW-Trans and HRFormer, leading to a decline in speed. Nevertheless, our segmentation network still holds practical value at the application level, particularly in situations where real-time demands are not excessive.

The segmentation results of the classical segmentation algorithms and ours on the images of daytime are shown in Figure 16.

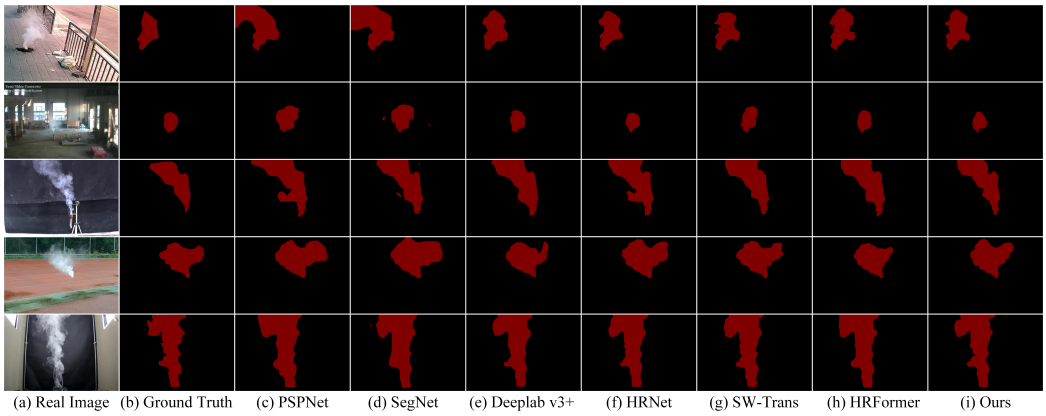


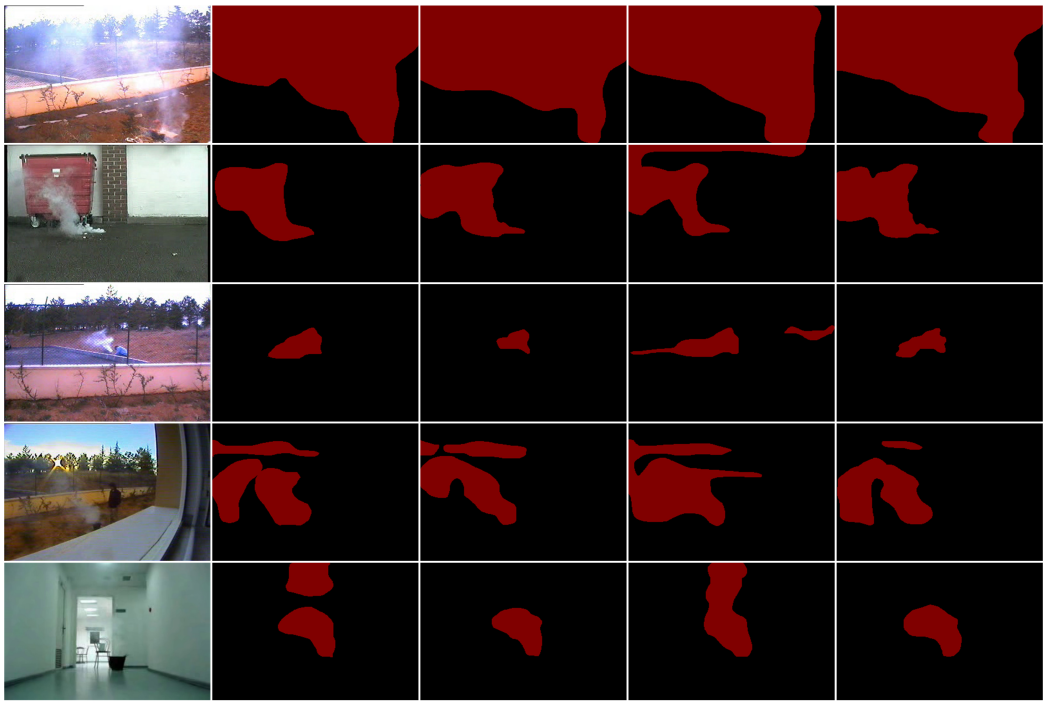
Figure 16. Examples of segmentation results of the images in the daytime.

As mentioned earlier, PSPNet has difficulties with identifying the background near the edges of smoke, as seen in too many smoke areas in the first, second, and third rows. On the other hand, SegNet performs worse in segmentation, with more non-smoke areas being divided in the segmentation results and the outline being too smooth in the first, fourth, and fifth rows. Moreover, there are isolated misjudgment segmentation areas in the second row. Compared with the previous two algorithms, Deeplab v3+ has improved its segmentation ability on smoke, reflected in the noticeable improvement in the second and fifth rows. However, Deeplab v3+ still has problems accurately recognizing some detailed texture information. Finally, HRNet achieves better segmentation results than other CNN models. However, due to its focus on local feature information, it may divide thin areas around smoke into more extensive ranges, such as the left area of the third row.

Due to their exceptional global feature extraction capabilities, SW-Trans and HRFormer have significantly improved segmentation effects for large-scale smoke compared with CNN models. However, in the second row, the foreground target becomes muddled with background information, leading to poor segmentation effects. Some of the bright window areas near the smoke become erroneously classified as smoke targets. Our segmentation algorithm incorporates a foreground feature localization module, which enhances the segmentation accuracy of smoke by highlighting foreground pixels. With this, our algorithm achieves superior segmentation results on the smoke in the second row and other smoke images, with results that come close to the ground truth.

To verify the effectiveness and generalization of our segmentation network and prove that it surpasses other existing algorithms considering segmentation performance, we conduct a comparison experiment using the public dataset [36] proposed by Bilkent University. We compared our segmentation network with three smoke segmentation algorithms [37–39],

which perform well on daytime smoke datasets. The results of the comparison segmentation are presented in Figure 17.



(a) Real Images (b) AOSVSSNet (c) CGRNet (d) W-Net (e) Ours

Figure 17. Examples of segmentation results on the public datasets.

AOSVSSNet [37] introduces a new plug-and-play Convolutional Block Attention Module (CBAM) based on the U-Net++ [40] network. The new module focuses more on the spatial location information of smoke areas, which results in improved segmentation results for smoke areas with high concentrations, such as the source areas in the second and third rows. However, the CBAM and improved loss function introduced by AOSVSSNet tend to focus more on the global location information of smoke rather than local characteristics. Therefore, AOSVSSNet is more suitable for optical satellite smoke images than rarefied ones and has poor segmentation results for thin smoke areas.

CGRNet [38] designs attention convolution modules based on Gated Recurrent Units (GRUs) to identify spatial correlation and global context dependence of smoke. Additionally, the Multi-scale Context Contrasted Local (MCCL) calculates the difference of smoke features at different resolutions to enhance the model's ability to segment small-scale smoke. The results of smoke segmentation on the public dataset show that CGRNet outperforms AOSVSSNet regarding the segmentation effect on the second and third rows of small targets and the first and fourth rows of the thin smoke boundary. In the fifth row especially, CGRNet has no false division for the light area. However, in the classification of the location of smoke in the second and third rows, AOSVSSNet performs better than CGRNet. Overall, CGRNet has shown significant improvements in smoke segmentation.

According to Ref. [39], the W-Net architecture utilizes multiple asymmetric encoder–decoder structures to create a waveform structure. The semantic information of images is mainly contained in the trough position, while the peak position contains local and mesoscale information. The use of skip connections between the peak and trough positions and the decoding layers enhances the accuracy of smoke segmentation. However, since W-Net adopts an upsampling and downsampling path similar to U-Net, the details lost in pooling operations are difficult to recover. W-Net divides excessive smoke areas in the first and fourth rows, which are highly diffuse smoke images, and misjudges the smoke in the second, third, fourth, and fifth rows. Overall, the segmentation effect of W-Net is relatively poor.

As shown in the results of comparison experiments, our smoke segmentation network is highly effective in identifying the edges of smoke with high concentrations. Moreover, it can accurately distinguish the background information from the smoke characteristics in the thin smoke areas. In particular, our algorithm can delineate clear smoke boundaries for the first row. For the second and third rows, which involve small-scale and high-concentration smoke targets, our algorithm performs better than Refs. [37–39]. Although our network has a small-scale misjudgment in the fourth row, our algorithm correctly judges the light in the fifth row to the background. These results demonstrate the generalization ability and practicality in multiple scenes of our segmentation algorithm, as well as its superior performance compared to other smoke segmentation algorithms.

4.5. Ablation Experiments of Semantic Segmentation Algorithm

In order to assess the contributions of the modules to performance improvement, we perform ablation experiments on each module on our synthetic self-built dataset. The ablation modules consist of the Transformer blocks with W-MSA and SW-MSA, the LSAM and GCAM, Dense PPM with feature fusion between branches of different resolutions, and FFLM after the feature extraction network. Table 6 presents the results of the ablation experiment, providing valuable insights into the functioning of our network.

Table 6. Ablation experimental results of our smoke segmentation algorithm.

Networks	W-MSA and SW-MSA	LSAM and GCAM	Dense PPM	FFLM	mIoU/%	FLOPs/G	T/ms
HRNet*		✓	✓	✓	91.35	196.5	72.72
Net1					90.41	85.1	47.86
Net2	✓				91.19	65.7	42.05
Net3	✓	✓			91.81	79.3	46.20
Net4	✓	✓	✓		92.26	90.6	49.38
Net5/Ours	✓	✓	✓	✓	92.93	125.7	58.16

In Table 6, the row of HRNet* represents the original HRNet with modified LSAM and GCAM, the same Dense PPM after each stage, and FFLM after segmentation. Net1 is a modified version of HRNet with reduced convolutional modules in each branch. Net2 replaces the convolutional units in Net1 with Transformer blocks. Net3 adds LSAM and GCAM to the Transformer blocks in Net2. Net4 is based on Net3 but with a different feature fusion method using Dense PPM. Finally, Net5 is our proposed segmentation network, which combines all the modifications made in the previous rows. ✓ in the table indicates that the module is selected.

The comparisons between the mIoU of 90.83% with the detection time of 54.46 ms of HRNet in Table 5 and Net1 in Table 6 indicate that reducing the number of feature extraction units of each branch can decrease the complexity and calculation scale while still improving the accuracy. The comparison results between Net2 and Net1 and between HRNet* and Net5 indicate that the convolution modules containing W-MSA and SW-MSA can reduce the computational scale and improve the segmentation performance of the network. These results also demonstrate that the better the global feature extraction ability of the model, the better the ability to segment smoke will be.

Based on the comparison of Net3 and Net2, adding LSAM and GCAM resulted in an increase of 13.6 G FLOPs and a 0.62% improvement in segmentation accuracy, which proves that GCAM and LSAM can reduce feature loss and suppress irrelevant feature information related to smoke. Furthermore, the integration of Dense PPM contributes to more efficient incorporation of smoke features at different resolutions, resulting in a 0.45% improvement in segmentation accuracy between Net3 and Net4. On the other hand, FFLM can attenuate the influence of background information on the extraction process of smoke features, particularly in the thin positions of smoke where texture and background information can be easily confused. The data in Table 6 also support the effectiveness of FFLM, as Net5 shows a 0.67% improvement in segmentation accuracy compared to Net4.

Our modifications to HRNet can improve the segmentation accuracy of smoke without causing a significant increase in network computation. Our network can still meet the actual detection requirements. The comparison results of the ablation experiment are shown in Figure 18.

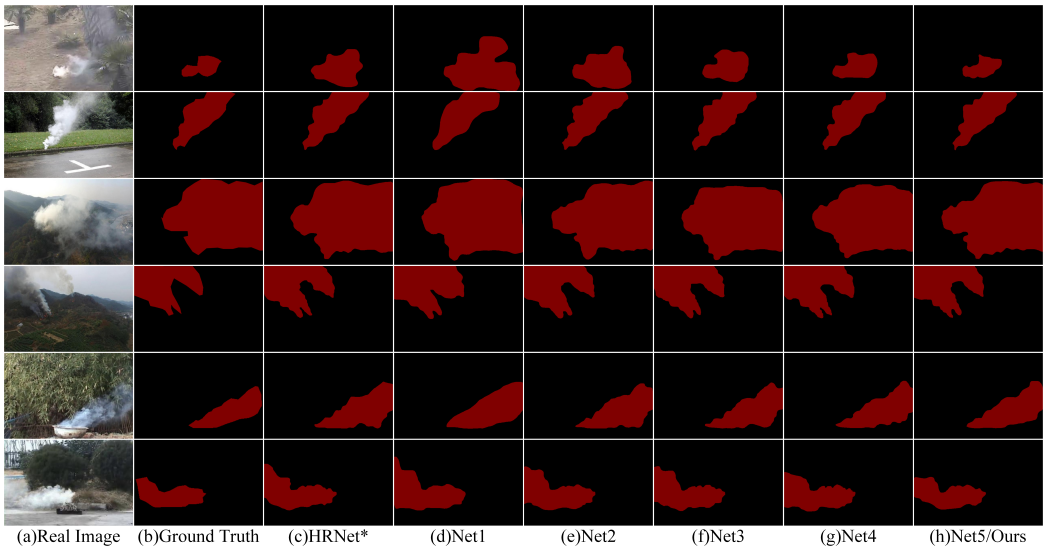


Figure 18. Examples of the ablation experimental results.

As shown in Figure 18, the smoke segmentation performance of Net5 is superior to that of HRNet*. Specifically, Net5 achieves more accurate segmentation results with fewer smoke edge miscalculations, as evidenced by the results of the third and fourth rows. Moreover, compared to Net2, Net3 exhibits better segmentation performance in regions with high smoke concentrations, such as the lower areas of the first and fifth rows. By incorporating FFLM, Net5 can more effectively differentiate between the thin edges of smoke and background information, as demonstrated in the first, third, and fourth rows. Our ablation experiment confirms the efficacy of our modifications for improving smoke segmentation.

4.6. Enhancement Segmentation Experiments

In order to accurately segment low-light smoke, our segmentation network needs to be highly precise when enhancing smoke images. We test our semantic segmentation algorithm on enhanced smoke images acquired in Section 4.2. The results are presented in Table 7, where the T/ms column represents the overall detection time achieved after we cascade the image enhancement network and the smoke segmentation network. The partial segmentation results are shown in Figure 19.

Table 7. Comparison of experimental results of our cascaded enhancement segmentation network.

Networks	mIoU/%	T/ms
PSPNet [22]	89.32	153.94
HRNet [19]	90.27	118.45
Deeplab v3+ [10]	89.71	153.50
SegNet [35]	88.09	94.11
SW-Trans [21]	90.24	112.89
HRFormer [20]	91.05	107.58
Ours	91.86	120.10

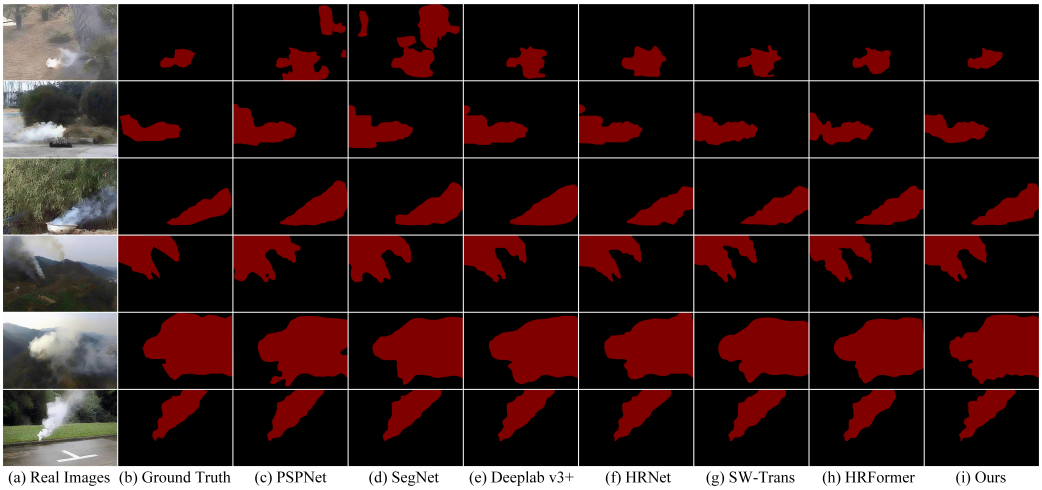


Figure 19. Examples of the comparison experimental results of our cascaded enhancement segmentation network.

The above results show that our algorithm performs better than other networks in the low-light smoke image segmentation task. For the image in the first row, PSPNet and SegNet misjudge the tree area in the upper right corner and do not accurately divide the smoke boundary, which is confused with the background information. Similarly, for the fourth-row image, SegNet and PSPNet have connectivity of smoke segmentation and are unable to distinguish between the background area and smoke. HRNet and Deeplab v3+ divide too few or too many smoke areas for highly diffuse smoke in the fifth row, indicating a low ability to distinguish between texture features and background information around edges. Our algorithm achieves superior segmentation results for the smoke images in the second, third, and sixth rows compared to SW-Trans and HRFormer.

In summary, our method can accurately segment smoke areas in daytime images and provide exceptional segmentation performance for low-light images enhanced by our enhancement network. While our approach has slightly increased detection time due to the complexity of the model, it still meets low-light smoke segmentation and alarm requirements in scenes that do not require high real-time performance.

To evaluate the efficacy of our approach in real low-light environments, we perform enhancement and segmentation of low-light smoke images in real nighttime environments. The results are shown in Figure 20.

Since low-light images in real-world scenes are different from synthetic low-light ones, the characteristics of non-smoke objects are unclear. As a result, our enhancement network is more capable of recovering smoke features, while non-smoke areas still show unknown black features like the original images. The results indicate that our enhancement network can effectively restore the smoke features in low-light images, and our segmentation model can complete the segmentation task of the enhanced images. Therefore, our proposed method is highly effective and practical in restoring smoke features and segmenting smoke areas in low-light images.

(a) Example 1 (b) Example 2 (c) Example 3

Figure 20. Examples of enhancement segmentation results on real low-light images.

5. Conclusions

We propose a low-light smoke image segmentation method utilizing a cascaded image enhancement algorithm with a semantic segmentation algorithm. The method shows excellent performance in accurately segmenting smoke images in low-light environments. To address the challenge of unclear smoke features in low-light environments, we propose a low-light smoke image enhancement network based on CycleGAN. Furthermore, we propose a multi-scale smoke semantic segmentation network based on HRNet and HRFormer to segment smoke areas in enhanced images accurately. Through a series of experiments, our method's ability to effectively complete the segmentation task of smoke in low-light environments is proven.

In addition, the real images collected at nighttime have unclear color information, making many details difficult to recover. Therefore, improving the suppression ability of irrelevant information in the semantic segmentation algorithm will be the direction of our research in the future.

Author Contributions: Conceptualization, E.L. and W.Z.; methodology, E.L. and W.Z.; software, E.L.; validation, E.L. and W.Z.; formal analysis, E.L.; investigation, E.L.; resources, E.L. and W.Z.; data curation, E.L.; writing—original draft preparation, E.L.; writing—review and editing, W.Z.; visualization, E.L.; supervision, W.Z.; project administration, W.Z.; funding acquisition, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CycleGAN	Cycle-Consistent Generative Adversarial Network
CNN	Convolutional Neural Network
PReLU	Parametric Rectified Linear Unit
MSE	Mean Squared Error
SSIM	Structure Similarity Index Measure
SW-MSA	Shifted Window-based Multi-Head Self-Attention
MLP	Multi-Layer Perception
LSAM	Local Spatial Attention Module
GCAM	Global Channel Attention Module
GELU	Gaussian Error Linear Unit
BN	Batch Normalization
PPM	Pyramid Pooling Module
FFLM	Foreground Feature Localization Module
ReLUs	Rectified Linear Units
SKLFS	State Key Laboratory of Fire Science
SNR	Signal Noise Ratio
PSNR	Peak Signal Noise Ratio
mIoU	Mean Intersection over Union
FLOPs	Floating Point Operations
Params	Parameter amount
T	Detection time
ASPP	Atrous Spatial Pyramid Pooling
W-MSA	Window-based Multi-Head Self-Attention
CBAM	Convolutional Block Attention Module
GRU	Gated Recurrent Unit
MCCL	Multi-Scale Context Contrasted Local
SW-Trans	Swin Transformer

References

1. Li, J.; Zhou, G.; Chen, A.; Lu, C.; Li, L. BCMNet: Cross-Layer Extraction Structure and Multiscale Downsampling Network with Bidirectional Transpose FPN for Fast Detection of Wildfire Smoke. *IEEE Syst. J.* **2023**, *17*, 1235–1246. [CrossRef]
2. Cao, Y.; Wang, G.; Wen, H.; Liu, X.; Yang, Z. Enhanced Receptive Field Smoke Detection Model Embedded with Attention Mechanism. In Proceedings of the China Automation Congress (CAC), Xiamen, China, 25–27 November 2022; pp. 5122–5126.
3. Chaturvedi, S.; Khanna, P.; Ojha, A. Comparative Analysis of Traditional and Deep Learning Techniques for Industrial and Wildfire Smoke Segmentation. In Proceedings of the Sixth International Conference on Image Information Processing (ICIIP), Shimla, India, 26–28 November 2021; pp. 326–331.
4. Khan, S.; Muhammad, K.; Hussain, T.; Del Ser, J.; Cuzzolin, F.; Bhattacharyya, S.; Akhtar, Z.; de Albuquerque, V.H.C. DeepSmoke: Deep Learning Model for Smoke Detection and Segmentation in Outdoor Environments. *Expert Syst. Appl.* **2021**, *182*, 115–125. [CrossRef]
5. Güzel, S.; Yavuz, S. Infrared Image Generation From RGB Images Using CycleGAN. In Proceedings of the 2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Biarritz, France, 8–10 August 2022; pp. 1–6.
6. Wu, X.; Wu, Z.; Ju, L.; Wang, S. A One-Stage Domain Adaptation Network With Image Alignment for Unsupervised Nighttime Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 58–72. [CrossRef] [PubMed]
7. Peng, J.; Su, J.; Sun, Y.; Wang, Z.; Lin, C.W. Semantic Nighttime Image Segmentation via Illumination and Position Aware Domain Adaptation. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 1034–1038.
8. Mahmud, M. Altitude Analysis of Road Segmentation from UAV Images with DeepLab V3+. In Proceedings of the 12th International Conference on Control System, Computing and Engineering (ICCSCE), Penang, Malaysia, 21–22 October 2022; pp. 219–223.
9. Xie, X.; Xu, Z.; Tao, J.; Yuan, J.; Wu, S. Semantic Segmentation Algorithm for Night Traffic Scene Based on Visible and Infrared Images. In Proceedings of the 2022 3rd Asia Conference on Computers and Communications (ACCC), Shanghai, China, 16–18 December 2022; pp. 103–108.
10. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
11. Zhao, J.; Cees, G. LiftPool: Bidirectional ConvNet Pooling. *arXiv* **2021**, arXiv:2104.00996.

12. Chen, X.; Li, Z.; Jiang, J.; Han, Z.; Deng, S.; Li, Z.; Fang, T.; Huo, H.; Li, Q.; Liu, M. Adaptive Effective Receptive Field Convolution for Semantic Segmentation of VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 3532–3546. [CrossRef]
13. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
14. Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; Wang, Z. EnlightenGAN: Deep Light Enhancement without Paired Supervision. *IEEE Trans. Image Process.* **2021**, *30*, 2340–2349. [CrossRef]
15. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
16. Wu, Y.; Lin, W.; Huang, S. Low-Power Hardware Implementation for Parametric Rectified Linear Unit Function. In Proceedings of the IEEE International Conference on Consumer Electronics—Taiwan (ICCE-Taiwan), Taoyuan, Taiwan, 28–30 September 2020; pp. 1–2.
17. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least Squares Generative Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2813–2821.
18. Zhang, S.; Cheng, D.; Jiang, D.; Kou, Q. Least Squares Relativistic Generative Adversarial Network for Perceptual Super-Resolution Imaging. *IEEE Access* **2020**, *8*, 185198–185208. [CrossRef]
19. Wang, J. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3349–3364. [CrossRef]
20. Yuan, Y.; Fu, R.; Huang, L.; Lin, W.; Zhang, C.; Chen, X.; Wang, J. HRFormer: High-Resolution Transformer for Dense Prediction. *arXiv* **2021**, arXiv:2110.09408.
21. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.
22. Li, Z.; Guo, Y. Semantic segmentation of landslide images in Nyingchi region based on PSPNet network. In Proceedings of the 2020 7th International Conference on Information Science and Control Engineering (ICISCE), Changsha, China, 18–20 December 2020; pp. 1269-1273.
23. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. DenseASPP for Semantic Segmentation in Streets Scenes. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 3684–3692.
24. Wan, S.; Hsu, C.Y.; Li, J.; Zhao, M. Depth-Wise Convolution with Attention Neural Network (DWA) for Pneumonia Detection. In Proceedings of the 2020 International Conference on Intelligent Computing, Automation and Systems (ICICAS), Chongqing, China, 11–13 December 2020; pp. 136–140.
25. Research Webpage about Smoke Detection for Fire Alarm of State Key Laboratory of Fire Science (SKLFS). Available online: http://smoke.ustc.edu.cn/datasets.htm (accessed on 20 February 2022).
26. Gonzalez, R.; Woods, R. Digital Image Processing. *J. Biomed. Opt.* **2009**, *14*, 66–67. [CrossRef]
27. Guo, Y.; Ke, X.; Ma, J.; Zhang, J. A Pipeline Neural Network for Low-Light Image Enhancement. *IEEE Access* **2019**, *7*, 13737–13744. [CrossRef]
28. Loh, Y.; Liang, X.; Chan, C. Low-light Image Enhancement Using Gaussian Process for Features Retrieval. *Signal Process. Image Commun.* **2019**, *4*, 175–190. [CrossRef]
29. Liu, C.; Wu, F.; Wnag, X. EFINet: Restoration for Low-Light Images via Enhancement-Fusion Iterative Network. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 8486–8499. [CrossRef]
30. Guo, X.; Li, Y. LIME: Low-Light Image Enhancement via Illumination Map Estimation. *IEEE Trans. Image Process.* **2017**, *26*, 982–993. [CrossRef]
31. Fu, X.; Zeng, D.; Huang, Y.; Zhang, X.P.; Ding, X. A Weighted Variational Model for Simultaneous Reflectance and Illumination Estimation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2782–2790.
32. Wei, C.; Wang, W.; Yang, W.; Liu, J. Deep Retinex Decomposition for Low-Light Enhancement. In Proceedings of the British Machine Vision Conference, Newcastle, UK, 3–6 September 2018.
33. Johnson, J.; Alahi, A.; Li, F. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. *arXiv* **2016**, arXiv:1603.08155.
34. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976.
35. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
36. Sample Fire and Smoke Video Clips of Bilkent University in Merkez Yerleşkesi, Ankara. Available online: http://signal.ee.bilkent.edu.tr/Visi-Fire/Demo/SampleClips.html (accessed on 20 November 2021).
37. Wang, T.; Hong, J.; Han, Y.; Zhang, G.; Chen, S.; Dong, T.; Yang, Y.; Ruan, H. AOSVSSNet: Attention-Guided Optical Satellite Video Smoke Segmentation Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 8552–8566. [CrossRef]
38. Yuan, F.; Zhang, L.; Xia, X.; Huang, Q.; Li, X. A Gated Recurrent Network With Dual Classification Assistance for Smoke Semantic Segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 4409–4422. [CrossRef]

39. Yuan, F.; Zhang, L.; Xia, X.; Huang, Q.; Li, X. A Wave-Shaped Deep Neural Network for Smoke Density Estimation. *IEEE Trans. Image Process.* **2020**, *29*, 2301–2313. [CrossRef] [PubMed]
40. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *Deep. Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support* **2018**, *11045*, 3–11.