


AOSVSSNet: Attention-Guided Optical Satellite Video Smoke Segmentation Network

Taoyang Wang , Jianzhi Hong , Yuqi Han , Guo Zhang , Shili Chen , Tiancheng Dong, Yaping Yang, and Hang Ruan

Abstract—Smoke is more observable than open fires. Optical satellite video has the advantages of a wide monitoring range, fast response speed, and good economy in large-scale surface smoke monitoring tasks. It can be used in wide-area forest wildfire monitoring, battlefield dynamic monitoring, disaster relief decision-making. The smoke segmentation method based on traditional handcrafted features is easily limited by the scene and data. This article introduces the deep learning method to the optical satellite video smoke target segmentation. However, due to the lack of real smoke images and the blurred edges of smoke, there are currently few labeled datasets for smoke segmentation in high-resolution optical satellite imagery scenes, which cannot provide sufficient training data for deep learning models. The smoke image from the satellite perspective also has the characteristics of multiscale features and ground object background interference. To solve the abovementioned problems, we construct a set of high-resolution optical satellite imagery smoke synthesis datasets based on the optical imaging process of smoke targets, which saves the cost of manual labeling. In addition, we design an attention-guided optical satellite video smoke segmentation network model, which can effectively suppress the ground object background's false alarm and extract the smoke's multiscale features. Synthetic data faces the transferability problem in real-world applications, so the physical constraints of the smoke imaging process are introduced into the loss function to improve the generalization of the model in real smoke data. The comprehensive evaluation results show that the method outperforms representative semantic segmentation networks.

Index Terms—Convolutional neural network, moving object segmentation, satellite video, smoke segmentation.

I. INTRODUCTION

SMOKE is more observable than open fires. Compared with traditional inductive detectors that need to be close to the fire source for physical and chemical composition analysis, the smoke sensing technology based on video image processing can respond faster to fire alarms, and the noncontact method can effectively eliminate the loss of the sensor [1]. Compared with the existing smoke coarse localization based on image classification and target detection, smoke segmentation effectively integrates location information and attribute information and obtains accurate pixel-by-pixel information, which helps rescuers effectively identify the source of fire and reduce the possibility of fire alarm delays. In addition, it can dynamically monitor the trend of smoke morphology, effectively reflect the current environmental conditions of the fire scene from the side, and provide instructive data support for predicting the spread trend and speed of the fire, which has significant research value and practical significance.

In recent years, with the continuous innovation of sensor technology and the improvement of the quality of spatial data acquisition, the emergence of optical video satellites that are capable of high frame rate (frame rate ≥ 24 FPS) imaging in the same area with dynamic observation capabilities, has made the research of smoke segmentation method for optical satellite video data more prominent than ground surveillance video, which has the advantages of wider monitoring range, faster response speed, and better economy. It shows great potential in monitoring fire smoke in vast surface spaces such as forests, volcanoes, and large oil tank farms.

However, smoke segmentation is a highly challenging computer vision task because smoke has more substantial intraclass variability than other segmentation targets. It is affected by lighting and shooting perspective and shows different colors, poses, and shapes at different times and under different physical and chemical conditions [2], as shown in Fig. 1(a). Moreover, compared with ground surveillance video, the scene from the satellite perspective has complex and similar ground object backgrounds and multiscale smoke targets, as shown in Fig. 1(b). Therefore, although the traditional artificially designed feature expression method can achieve accurate extraction of smoke to a certain extent, most of these design schemes are relatively

Manuscript received 30 June 2022; revised 3 September 2022; accepted 14 September 2022. Date of publication 26 September 2022; date of current version 12 October 2022. This work was supported in part by the Key Research and Development Program of the Ministry of Science and Technology under Grant 2018YFB0504905, in part by High Resolution Earth Observation Systems National Science and Technology Major Projects under Grant GFZX0404120305, in part by Civil Aerospace Advance Research Project under Grant D040107, in part by Postdoctoral Innovation Talent Support Program under Grant BX2021222, and in part by LIESMARS Special Research Funding. (Corresponding author: Yuqi Han.)

Taoyang Wang, Jianzhi Hong, and Shili Chen are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: wangtaoyang@whu.edu.cn; hongjianzhi@qq.com; sl_chen@whu.edu.cn).

Yuqi Han is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: yuqi_han@tsinghua.edu.cn).

Guo Zhang and Tiancheng Dong are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: guozhang@whu.edu.cn; 2021106190028@whu.edu.cn).

Yaping Yang is with the Navy Research Institute, People's Liberation Army, Beijing 100036, China (e-mail: yang8304@126.com).

Hang Ruan is with the Beijing Institute of Tracking and Telecommunications Technology, Beijing 100094, China (e-mail: dragonhang9@163.com).

Digital Object Identifier 10.1109/JSTARS.2022.3209541

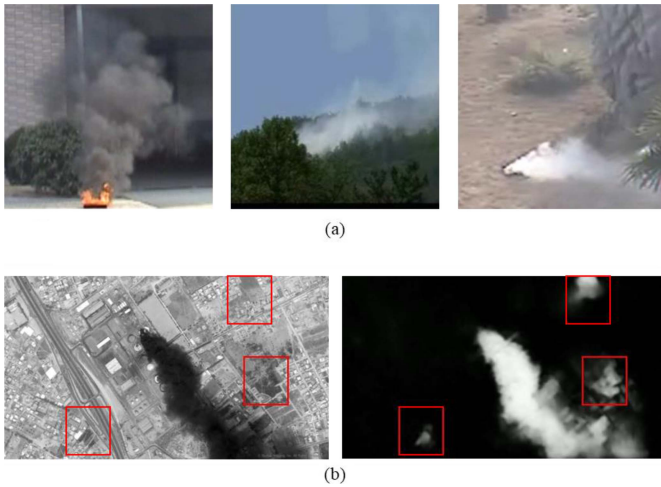


Fig. 1. Challenges of optical remote sensing satellite smoke segmentation. (a) The irregular shape and blurred edges of the smoke make it difficult for pixel-by-pixel manual annotation. (b) Interference of similar or complex background objects in remote sensing images.

complex, and the selection and combination of features lack unified principles and specifications. It is more susceptible to scene and data constraints in smoke target extraction [3]. As an excellent data-driven modeling tool, deep learning can automatically learn the excellent and essential features that conform to the distribution of the current task image dataset and significantly reduce the labor cost of feature modeling, which has attracted the attention of researchers, such as CNN [4], [5], [6], [7], GCN [8], and Transformer [9]. The performance of deep learning models largely depends on large-scale, high-quality labeled training datasets. The slow development of smoke segmentation datasets for high-resolution optical satellite images restricts related research progress. The main reasons include the following two points: 1) Since fire and smoke are accidental emergencies in daily life, there are relatively few fire and smoke scenes on the ground that can be captured by satellites, resulting in a small scale of real datasets for research; 2) Because the smoke target has the characteristics of irregular shape and blurred edge, it is extremely difficult and inaccurate to manually label the boundary of the smoke target pixel by pixel.

This article constructs a set of high-resolution optical satellite image smoke target synthesis datasets based on the optical imaging principle of smoke targets to solve the problem of the lack of reliable training datasets and the difficulty of labeling in the optical satellite video smoke segmentation task, which significantly saves the cost of manual labeling. Compared with typical smoke images, the smoke targets in satellite video have stronger visual saliency than other ground objects. For the problem that the multiscale segmentation results of optical satellite video smoke are easily disturbed by background objects, an attention-guided optical satellite video smoke segmentation network model called attention-guided optical satellite video smoke segmentation network model (AOSVSSNet) is proposed in this article to improve the segmentation accuracy. For the transferability of the synthetic dataset training model on the real

test dataset, this article introduces the physical constraints of the smoke imaging process into the loss function of the segmentation network, which has good generalization.

In summary, this article has the following three main contributions.

- 1) This article constructs a set of high-resolution optical satellite image smog target synthesis datasets. As far as we know, this is the first high-resolution optical satellite smoke dataset, which effectively solves the issues of lack of training samples for smoke segmentation and labeling difficulty.
- 2) This article proposes a convolutional neural network model for smoke segmentation, which to our knowledge is the first deep learning model for smoke segmentation in optical satellite video. It can achieve end-to-end training and prediction, and effectively suppress background interference while extracting smoke pixels.
- 3) In this article, the physical constraints of the smoke imaging process are introduced into the loss function of the segmentation network, which improves the generalization of the synthetic dataset training model on the real test dataset.

The rest of this article is organized as follows. Section II presents related work on semantic segmentation of smoke images. Section III details the optical satellite imagery smoke training data synthesis method and smoke segmentation model for optical satellite video. Section IV presents and analyzes the experimental results of the proposed method. Finally, Section V concludes this article.

II. RELATED LITERATURE

Currently, there is no public report on the research of smoke segmentation for optical satellite video. Therefore, the existing algorithm can be adapted by referring to the research on smoke segmentation based on natural images. The mainstream smoke segmentation methods can be divided into traditional manual features and deep learning methods.

A. Traditional Smoke Segmentation Methods

Smoke targets have rich image information, including static features such as color, texture, and shape, and dynamic characteristics such as diffusion, displacement, and flickering [2]. Therefore, traditional smoke segmentation methods focus on using various smoke image information to form a feature representation method with sufficient recognition. Among them, the focus of research is color, texture, frequency, and motion features.

In terms of color features, the characteristics of smoke in the red-green-blue (RGB) color model are mainly manifested in that the gray values of the R, G, and B channels are relatively similar, roughly distributed in the range of 80–220 [10], [11]. The saliency of the hue-saturation-value (HSV) and hue-saturation-intensity (HSI) color models is mainly focused on the saturation component [12], [13], [14]. In terms of texture features, gray level co-occurrence matrix [15], local binary pattern (LBP) [16], and Pyramid LBP [17], [18] are the more commonly used

methods. Additionally, dynamic textures have the potential to characterize temporal invariance and have also been applied to describe smoke [19]. In terms of frequency features, a single frequency feature can achieve a good recognition effect [20], [21], as different frequency information in the frequency domain corresponds to the image information in the spatial domain. The high, medium, and low-frequency information reflect the image's edge details, structure, and main components. Among them, wavelet transform is the most commonly used frequency feature extraction method [22], [23], [24]. By fusing the target features in the spatial domain and the frequency domain, and using ensemble classifier learning, the translation, and rotation invariant features of the target can be expressed, thereby improving the detection accuracy [25], [26], [27], [28], [29]. In terms of motion features, the drift, diffusion, and other motion characteristics of smoke are the focus of research [30], [31]. The feature extraction mainly adopts statistical features, including optical flow estimation method, area, and centroid change statistics of suspected smoke areas, and movement direction change statistics [11], [22], [32]. Smoke segmentation can also be seen as the process of background and moving foreground segmentation, extracted by methods such as anomaly detection [33], linear unmixing [34], object tracking [35], or modal translation [36].

In general, although these traditional methods can achieve accurate smoke extraction to a certain extent, they usually require manual feature design and classifier selection, which requires designers to have solid empirical knowledge in specific fields such as the extraction method and combination of features, the setting of hyperparameters, resulting in high cost. In addition, the migration of artificially designed features is poor, and the testing effect is generally good only on the current task dataset. However, it is difficult to adapt to the smoke targets with different data quality and scene changes, resulting in unstable or poor segmentation accuracy.

B. Deep-Learning-Based Smoke Segmentation Methods

The emerging deep learning algorithm avoids the complex feature design process to the greatest extent. By designing a reasonable neural network structure, people can enable the model to automatically and efficiently learn excellent features adapted to the current task with less manual intervention, and bring significant improvements to visual smoke monitoring tasks of various granularities [37], [38], [39], [40]. The performance of smoke semantic segmentation network models largely relies on large-scale pixel-by-pixel labeled data. At present, the open-source smoke datasets of natural images mainly include the laboratory dataset of Bilken University, Turkey [41], the laboratory dataset of Keimyung University, South Korea [42], the Chino flame smoke image dataset BoWFire [43], the dataset of the State Key Laboratory of Fire Science, University of Science and Technology of China [44], and Jiangxi University of Finance and Economics Yuan Feiniu Laboratory datasets [45]. Among these, only the last two datasets have pixel-by-pixel annotations of smoke. The rest of the labeled datasets are used for classification

or detection, with the scene mainly based on the ground perspective. Remote sensing images have a wide observational perspective and rich and diverse data sources, including optical, SAR, hyperspectral, and video. Data obtained from different platforms can provide diverse and complementary information [46], [47]. The smoke datasets for optical satellite images mostly come from low-resolution multispectral images such as MODIS [48], Himawari-8 [49], LandSat-8 [50], and GOES-16 [51]. The lack of large-scale, open-source, high-resolution labeled datasets for segmentation restricts the development of smoke segmentation network models for high-resolution optical satellite imagery.

In addition, compared to image classification and object detection tasks, fine-grained semantic segmentation tasks rely more on contextual feature information to obtain higher segmentation accuracy. At present, the main ideas of semantic segmentation networks include fully convolutional neural networks (such as FCN [52]), encoder-decoder structures (such as U-Net [53], SegNet [54], PSPNet [55]), and dilated convolutional networks (such as DeepLab series algorithms [56], [57], [58], [59]). Existing smoke segmentation methods are also mainly based on it.

Regarding how the algorithm utilizes the input data stream, video smoke segmentation can be divided into single-frame image smoke segmentation that only uses static appearance features and video smoke segmentation methods that fuse dynamic spatiotemporal features.

In terms of single-frame image smoke segmentation research, Xu et al. [60] proposed an end-to-end framework for smoke saliency detection, which consists of a region proposal network and an autoencoder structure to achieve smoke frame-level recognition and pixel-level fine segmentation. Yuan et al. [45] proposed an end-to-end segmentation network that fuses dual-branch features for blurred, semitransparent, and nonrigid boundaries of smoke targets, which outputs a soft segmentation probability map with 0-1 continuous values and gains pixel-by-pixel density estimation. Yuan et al. [61] believed that the full fusion of information between the high and low layers of the codec could improve the segmentation accuracy of fuzzy objects such as smoke and clouds and proposed a deep neural network with a wave structure using a synthetic smoke dataset for training to achieve smoke density estimation. Yuan et al. [62] proposed a classification-assisted gated regression semantic segmentation network for the problem of interclass similarity of smoke and small smoke segmentation, which can learn long-distance feature relationships and contextual information and improve the accuracy of smoke segmentation. It is not difficult to see from the abovementioned methods that the natural image smoke segmentation network basically innovates and transforms around the goal of how to enhance the contextual features. These strategies include dual-branch feature fusion, high-level and low-level feature fusion, and visual attention mechanisms to improve the accuracy of smoke segmentation, which are worthy of reference and study.

Currently, there are relatively few deep learning smoke segmentation methods for the overall processing of video form. Li et al. [63] applied a 3-D fully convolutional neural network to the video wildfire smoke segmentation task for the first time

and reduced the false detection rate of smoke segmentation by fusing the information between high and low layers and expanding the receptive field. The unsupervised video target segmentation network that has emerged in recent years has also attracted attention. It mainly realizes the classification of the target in the initial frame and the tracking in the subsequent frame from the pixel level according to some salient features of the target to be segmented, such as motion features, which shows potential in video smoke object segmentation that is difficult to manually annotate. Two-stream networks fusing motion and appearance features and recurrent neural networks are two important ideas to achieve unsupervised video object segmentation. The representative methods include MP-Net [64], LVO [65], FSEG [66], PDB [67], CosNet [68], and AGNN [69]. Although these methods perform better in the segmentation of rigid objects with translational motion, the motion pattern of smoke generally presents a diffuse motion from the source point to the surrounding; i.e., the edge pixels move while the interior pixels remain stationary. Therefore, the influence of the video frame sequence is mainly on the edge of the smoke. Although the model that introduces motion information will further refine the edge of the smoke or enhance the feature expression of little smoke, it also introduces more motion noise. The repeated texture inside the smoke makes the description of the motion optical flow feature unreliable, resulting in poor smoke segmentation results or missed detections.

Satellite video processing methods can be divided into multi-frame processing methods using timing information and frame-by-frame processing methods. Considering that video annotation is expensive, to extract the main area of the smoke target as much as possible, this article adopts the idea of frame-by-frame processing of the deframed video, takes the improved version UNet++ of the classic semantic segmentation network UNet as the basic framework, and realizes high and low-level features through a dense skip connection structure. The complete integration of the convolutional attention module guides the model to pay more attention to the smoke target and suppress the background of irrelevant objects to achieve accurate segmentation of the smoke area based on optical satellite video.

III. PROPOSED METHOD

A. Smoke Segmentation Synthetic Dataset Construction

Currently, optical satellite image smoke datasets mainly focus on low-resolution scenes and coarse-grained detection and recognition tasks, lacking large-scale high-resolution open-source segmentation datasets. To solve the problem of the scarcity of training data for the deep learning model of smoke segmentation, we use the existing open-source datasets for natural image smoke segmentation based on the optical smoke imaging principle to construct a rich and diverse optical satellite image smoke target segmentation synthetic dataset and validate the generalization performance on synthetic datasets through real data.

1) *Optical Imaging Principle of Smoke Target:* Smoke is usually composed of incompletely burned tiny solid particles floating in the air. Through scattering and absorption of light

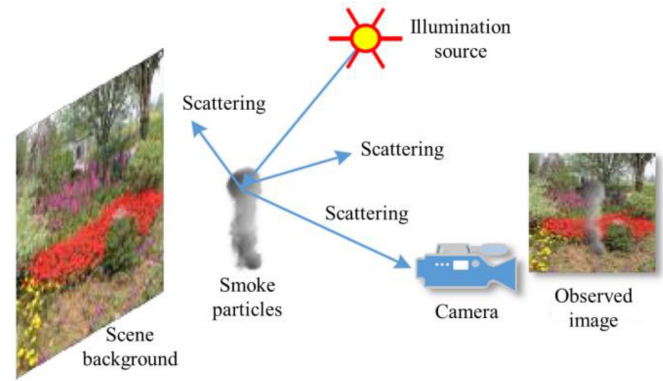


Fig. 2. Optical imaging process of the smoke target [61].

sources or reflected light, the light is continuously weakened during the propagation process and finally imaged in the camera under observation [70]. Fig. 2 shows the optical imaging process of smoke targets [61].

The optical imaging process of a smoke target from a 3-D space to a 2-D plane means each pixel value $i(x)$ can be simplified as a weighted sum of pure background pixel values and pure smoke pixel values in mathematical description

$$i(x) = b(x)(1 - \alpha(x)) + s(x)\alpha(x). \quad (1)$$

In (1), $b(x)$ represents the background color, $s(x)$ represents the smoke color, and $\alpha(x)$ represents the transparency coefficient or alpha channel of the smoke. Since this equation is essentially a linear color synthesis equation in the mathematical form [71], [72], this article regards $\alpha(x)$ as the optical density of smoke, which helps us to synthesize smoke images by quantitative methods later, and incorporate physical constraints into the model to improve segmentation accuracy.

2) *Synthesis Method of the Smoke Target Image:* The smoke optical density $\alpha(x)$ is a value ranging from 0 to 255, and it is neither possible nor accurate to calibrate the transparency of each pixel manually. Existing studies have used computer graphics methods to simulate and visualize smoke based on the principle of fluid dynamics. The most representative one is a set of open-source smoke datasets constructed by the team of Prof. Y. Feiniu from Jiangxi University of Finance and Economics using the open-source 3-D modeling software Blender [45]. The research team has generated a large amount of synthetic smoke data, including background, smoke, and transparency maps, by setting physical parameters such as wind, motion, and gravity. These smokes had different shapes, densities, lighting, and backgrounds, which have a realistic vision of real smoke. It significantly has saved the cost of manually collecting real smoke images and provides a sufficient database for deep learning model training.

The smoke targets in the ground cameras and remote sensing images have similar diffusion motion patterns, but the scales of the smoke targets in the remote sensing images are more different, and there are complex ground object backgrounds. Therefore, on the basis of the abovementioned open-source

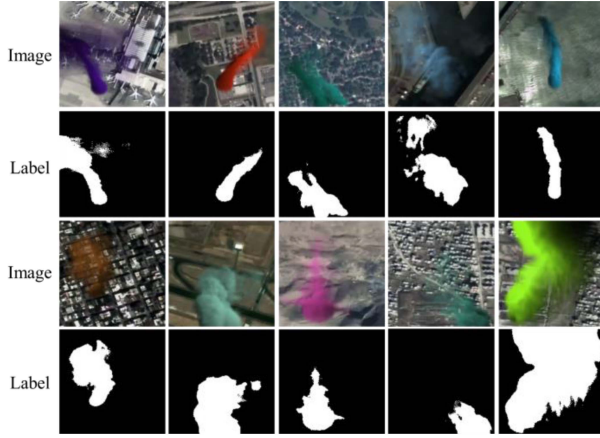


Fig. 3. Example of synthetic smoke dataset from satellite images.

smoke dataset, this article selected remote sensing images of different scenes as the background, and constructed a set of optical satellite image smoke target segmentation synthetic datasets.

- 1) First, deframe and block the optical satellite video to obtain a pure background image $b(x)$.
- 2) Then, for the pure background image $b(x)$ and the existing pure smoke image $s(x)$ and the corresponding transparency image $\alpha(x)$, perform linear stacking and set a random factor γ to generate a composite image $I(x)$ of smoke with different colors. The extended smoke target synthesis equation can be expressed as follows:

$$\begin{cases} I_R(x) = (1 - \alpha(x)) B_R(x) + \alpha(x) \gamma_1 S_R(x) \\ I_G(x) = (1 - \alpha(x)) B_G(x) + \alpha(x) \gamma_2 S_G(x) \\ I_B(x) = (1 - \alpha(x)) B_B(x) + \alpha(x) \gamma_3 S_B(x) \end{cases} \quad (2)$$

Among them, $\gamma_1, \gamma_2, \gamma_3$ are random numbers in the range of [0,1].

- 3) Next, take the data enhancement operation of horizontal and vertical flipping on the composite smoke image $I(x)$, which can reduce the overfitting of the model to a particular feature and improve its robustness and generalization.
- 4) Finally, set the threshold Th for the transparency map $\alpha(x)$ corresponding to the smoke synthesis image $I(x)$, and generate its corresponding binary mask image according to the following equation as the ground-truth map of the semantic segmentation task. Th is set to 128 in this article. Fig. 3 shows the synthetic dataset

$$\beta = \begin{cases} 1 & \text{if } \alpha \geq Th \\ 0 & \text{else.} \end{cases} \quad (3)$$

B. AOSVSSNet Network Structure

Existing semantic segmentation networks are based on U-Net. U-Net includes four-time down-sampling and up-sampling encoders and decoders and a long-skip connection structure, which realizes the splicing of high-level semantic and low-level geometric features and improves segmentation accuracy. However, some questions can still be explored in the design of the U-Net network, including the degree of influence of sampling times on

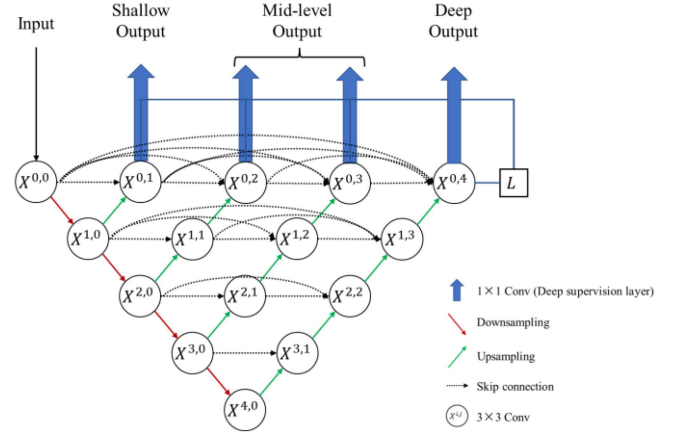


Fig. 4. Framework of the proposed AOSVSSNet algorithm.

feature extraction and the actual performance of long-connection structures in bridging the semantic gap. In response to these problems, Zhou et al. [73] extended the U-Net network and proposed an encoder-decoder structure UNet++ composed of nested dense short-skip connection layers by stacking U-Net networks of different levels, which helps to reduce the semantic gap between the feature map and the decoded feature map. It has a strong ability to capture image feature details, adapt to the high-resolution remote sensing images with rich details, multiscale features, and complex structure of ground objects characteristics, and has better segmentation performance.

Therefore, an attention-guided optical satellite video smoke segmentation network with the pruned version of UNet++ as the basic structure was designed.

- 1) CBAM was introduced between the original encoder layers to adaptively select and enhance features, so that the network could focus more on the smoke target content and global location information, suppress other irrelevant ground objects and noise information, and improve the accuracy of smoke segmentation.
- 2) Select the lightweight network MobileNetV2 as the convolution unit of the network to reduce the number of parameters required for training.
- 3) According to the smoke optical imaging process, a complex loss function with multiple constraints was introduced into the model, which could achieve fine segmentation of smoke targets based on the optical concentration estimation results, and improve the generalization performance of the model tested on real data.

Correspondingly, according to the loss function, the number of channels at the input and output of the network was adjusted. Details are shown in Fig. 4.

1) *UNet++ Network Structure*: The network structure of UNet++ is shown in Fig. 5, which mainly includes five parts: input interface, encoder, decoder, skip connection, and deep supervision.

The encoder part consists of five down sampling layers $X^{00}, X^{10}, X^{20}, X^{30}, X^{40}$. Each downsampling layer is implemented by a VGG block and a pooling layer, and each VGG block is

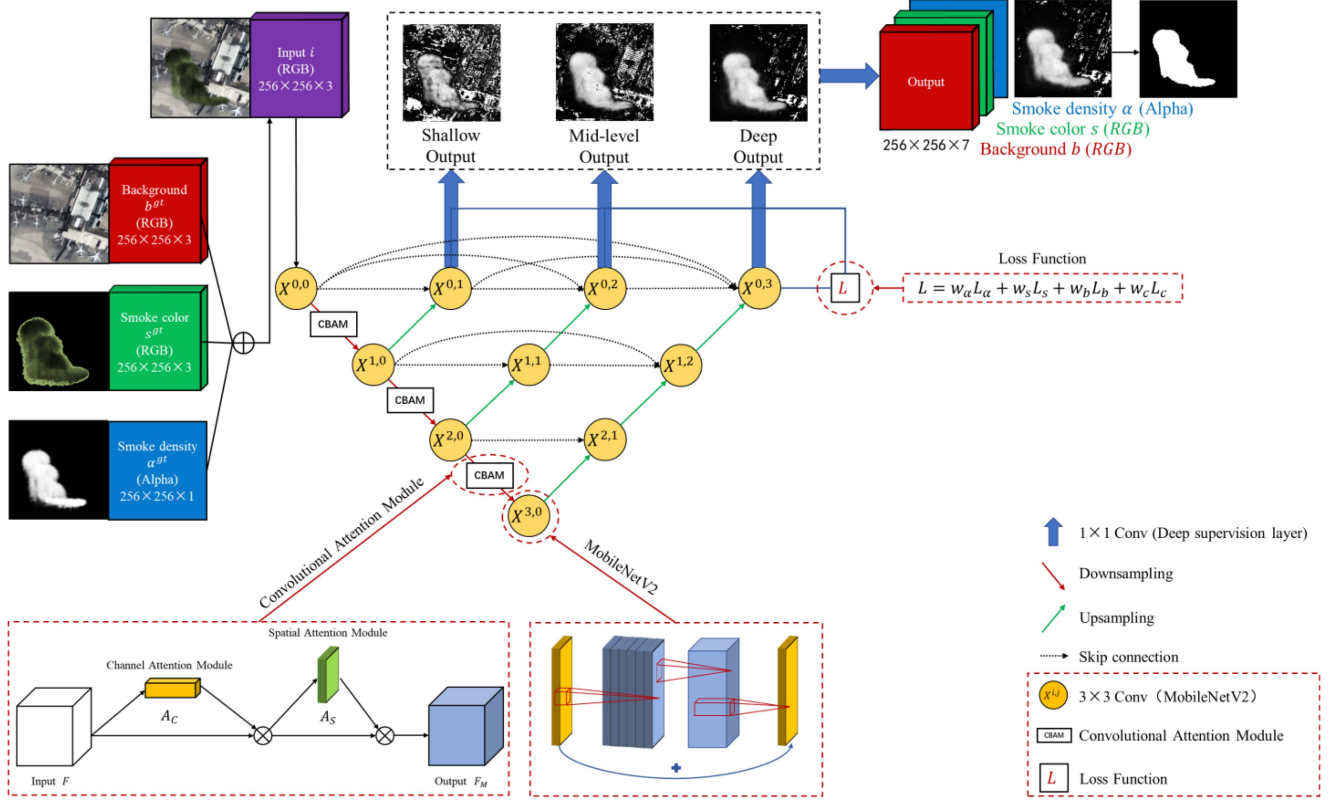


Fig. 5. UNet++ network structure.

concatenated with two convolutional layers with a kernel size of 3×3 pixels and a sliding stride of 1 pixel. The number of VGG block convolution kernels in each layer is 64, 128, 256, 512, and 512, respectively. The implementation of the downsampling layer can choose other convolutional neural network structures according to actual needs, and the number of convolution kernels can also be adjusted as needed.

The decoder part mainly includes four branches. These branches upsample the feature maps extracted by X^{10} , X^{20} , X^{30} , X^{40} , and fuse the shallow features of the same layer, and iteratively process from top to bottom to obtain the output graph of four branches. Similar to the encoder, the specific implementation of each layer unit of the decoder can also be designed as needed. The calculation result of each unit of the codec part can be expressed by the following equation:

$$x^{i,j} = \begin{cases} H(P(x^{i-1,j})) & j = 0, i = 1, 2, \dots, 5 \\ H\left(\left[[x^{i,k}]_{k=0}^{j-1}, U(x^{i+1,j-1})\right]\right) & j > 0, i = 1, 2, \dots, 5 \end{cases} \quad (4)$$

In (4), $H(\cdot)$ represents the convolution computation, $P(\cdot)$ represents the max-pooling computation with a size of 2×2 for downsampling, $U(\cdot)$ represents the deconvolution computation for upsampling, and $[\cdot]$ represents feature connections in the channel dimension.

The blue solid line path is the deep supervision layer, which can combine the output results of each branch of the decoder to obtain the final segmentation result. Combination

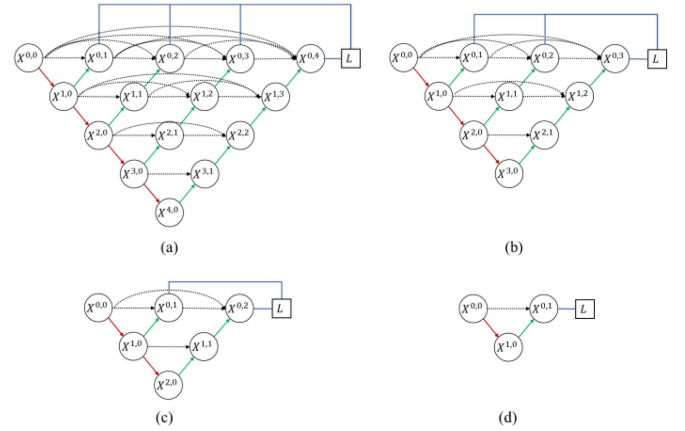


Fig. 6. Pruned form of the UNet++ model.

the four branches of the decoder between the corresponding level of the encoder can be regarded as four subnetworks of different levels. A separate model corresponding to the four versions can be formed through pruning: UNet++L1, UNet++L2, UNet++L3, and UNet++L4, as shown in Fig. 6. Compared with training four subnetworks separately and selecting the model, UNet++ adopts the strategy of training the overall model and then pruning, which has stronger operability and is less time-consuming. When the scale of the subnetwork reaches a specific target prediction accuracy, the model with the smallest memory footprint or calculation amount can be obtained by

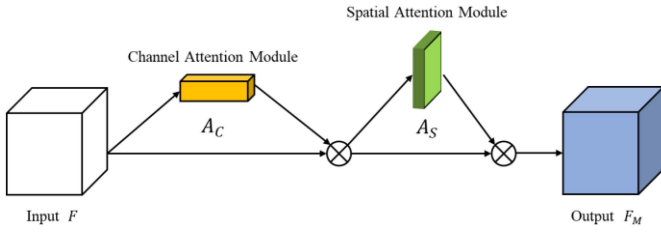


Fig. 7. Calculation process of CBAM.

the approach, which reflects the flexibility and efficiency of the model.

2) *Convolutional Attention Module CBAM*: While providing sufficient, discriminative, multiscale deep features for image classification or regression tasks, convolutional neural networks also introduce more redundant and noisy information, increasing computational cost and affecting segmentation performance. Feature optimization can select the most useful features for the segmentation task from the original feature set. Inspired by human vision research, the convolutional attention mechanism, an excellent deep feature selection method, can learn the weight distribution of output feature maps, highlight the target content and location information, and ignore other irrelevant information. Currently, the convolutional attention mechanism is mainly divided into three categories: spatial attention mechanism, channel attention mechanism, and hybrid attention mechanism. Among them, the hybrid attention mechanism considers spatial and channel similarity. The main methods include CBAM [74], DANet [75].

Optical satellite video is a high-resolution remote sensing time series image. The complex background of ground objects is the primary interference information for the task of smoke segmentation. At the same time, the smoke mainly moves upward and is less constrained by the structure of ground objects. Therefore, a lightweight, efficient, and plug-and-play CBAM module was integrated into the UNet++ model to adjust the feature weights in the spatial and channel directions, improve the semantic expression ability of the network for the smoke target, and realize end-to-end training.

The working principle of CBAM can be shown in Fig. 7. Assuming that the size of the input feature map F is $H \times W$ and the number of channels is C , then CBAM first uses the channel attention module to calculate the feature map F to obtain a 1-D channel attention weight distribution A_C (size is $1 \times 1 \times C$) and then calculate the dot product of the feature map F and A_C to obtain the channel-oriented salient feature map F_C , and the calculation process is expressed by (5). Then, use the spatial attention module to calculate F_C to obtain a 2-D spatial attention weight distribution A_S (the size is $H \times W \times 1$). Finally, the dot product of the feature map F_C and A_S is calculated to obtain the spatially significant feature map F_M , and the calculation process is expressed by (6) as follows:

$$F_C = A_C (F) \otimes F \quad (5)$$

$$F_M = A_S (F_C) \otimes F_C \quad (6)$$

represents the dot product operation in the equation.

TABLE I
ORIGINAL LAYER STRUCTURE OF MOBILENETV2 MODEL

| Layer number | Size of input | Operation | Expansion factor | Number of output channels | Number of repetitions of layer structure |
|--------------|-------------------|----------------------|------------------|----------------------------------|--|
| 1 | $224^2 \times 3$ | 3×3 Conv | — | 32 | 1 |
| 2 | $112^2 \times 32$ | | 1 | 16 | 1 |
| 3 | $112^2 \times 16$ | | 6 | 24 | 2 |
| 4 | $56^2 \times 24$ | Inverse | 6 | 32 | 3 |
| 5 | $28^2 \times 32$ | residual | 6 | 64 | 4 |
| 6 | $14^2 \times 64$ | structure | 6 | 96 | 3 |
| 7 | $14^2 \times 96$ | | 6 | 160 | 3 |
| 8 | $7^2 \times 160$ | | 6 | 320 | 1 |
| 9 | $7^2 \times 320$ | 1×1 Conv | — | 1280 | 1 |
| 10 | $7^2 \times 1280$ | 7×7 Pooling | — | — | 1 |
| 11 | $1^2 \times 1280$ | 1×1 Conv | — | Number of target output channels | — |

3) *Lightweight Convolutional Neural Network MobileNetV2*: Compared with UNet, UNet++ has a stronger multiscale semantic feature expression ability. However, it also has more convolution computing units, which will reduce the processing speed of optical satellite video data. It is challenging to meet the future onboard processing need with limited computing and memory resources. In the context of the needs of embedded mobile devices and real-time processing, lightweight convolutional neural networks have emerged, and MobileNet series algorithms are an excellent representative of them.

MobileNetV2 [76] is a lightweight network proposed by Google in 2018. It inherits the depthwise separable convolution adopted by MobileNetV1 [77] and adds a new structure called bottleneck residual module, mainly composed of two substructures of reverse residual and linear bottleneck. Compared with the accuracy of the ordinary convolution layer, it significantly reduces the number of model parameters and computing resource consumption and has better comprehensive performance. The layer structure of the original MobileNetV2 model is shown in Table I.

a) *Depthwise Separable Convolution*: The most prominent feature of depthwise separable convolution is that it can significantly reduce the number of parameters required for convolution calculation without affecting the performance of the model. The depthwise separable convolution splits the traditional convolution calculation into two stages: depthwise convolution and point convolution. The computation ratio of depthwise separable convolution and traditional standard convolution can be expressed as

$$\begin{aligned} & \frac{O_{(\text{Depthwise Separable Convolution})}}{O_{(\text{Standard convolution})}} \\ &= \frac{O_{(\text{Depthwise convolution})} + O_{(1 \times 1 \text{ convolution})}}{O_{(\text{Standard convolution})}} \\ &= \frac{K \times K \times H \times W \times M + 1 \times 1 \times H \times W \times M \times N}{K \times K \times H \times W \times M \times N} \end{aligned}$$

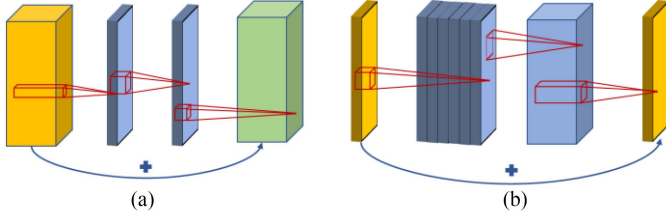


Fig. 8. Residual module and inverse residual module.

$$= \frac{1}{N} + \frac{1}{K^2}. \quad (7)$$

In (7), K represents the size of the convolution kernel, H , W , and M represent the height, width, and the number of channels of the input feature map, respectively, and N represents the number of channels of the output feature map. In reality, since the number of channels N of the output feature map is often large, when using a convolution kernel with a size of 3×3 to calculate the output feature map of 8 channels, the calculation amount of the depthwise separable convolution is reduced by nearly 80%, compared with traditional standard convolution.

b) Inverse Residual Structure: The emergence of the residual structure has well compensated for the difficulty of training caused by the depth of the neural network and brought a significant improvement to the performance of the model. Therefore, MobileNetV2 also draws on this design to form a new structure called the inverse residual structure, as shown in Fig. 8. Considering that the channel compression of the input feature map will reduce the accuracy of the model, its calculation process is designed into three steps: feature enhancement, feature extraction, and feature dimensionality reduction. In addition, to prevent network performance degradation, the structure also replaces the ReLU6 function used for feature mapping from high-dimensional to low-dimensional with a linear function, forming the final linear bottleneck structure. In conjunction with the depthwise separable convolution mentioned previously, it can effectively reduce the computational cost of the reverse residual structure in the high-dimensional feature extraction process and achieve the unification of model performance and efficiency. Therefore, this paper used MobileNetV2 as the basic unit of convolution calculation of the network model to improve the efficiency of the algorithm.

4) Loss Function: In this article, the input interface of the segmentation network was set to a three-channel RGB smoke synthesized image, and the output interface was set to a seven-channel feature map. The first, second, and third channels of the output were used to predict the three-channel pure background pixel values of RGB. The output's fourth, fifth, and sixth channels were used to predict the three-channel pure smoke pixel values of RGB, and the output's seventh channel was used to predict the values of the RGB synthesized smoke images. Correspondingly, the segmentation network adopted a complex loss function [61] containing four error terms with physical constraints, which was defined as the following equation:

$$L = w_\alpha L_\alpha + w_s L_s + w_b L_b + w_c L_c. \quad (8)$$

Among them, L_α , L_s , L_b , and L_c represented the mean square error of the four predicted values including the smoke density, RGB pure smoke pixel value, RGB pure background pixel value, and the RGB synthesized smoke image pixel value, as shown in (9)–(12). w_α , w_s , w_b , and w_c , respectively, represented the weight coefficients of the four error terms in the final error, all of which were taken as 0.25 in this article

$$L_\alpha = \frac{1}{2} \|\alpha - \alpha^{gt}\|^2 \quad (9)$$

$$L_s = \frac{1}{2} \|s - s^{gt}\|^2 \quad (10)$$

$$L_b = \frac{1}{2} \|b - b^{gt}\|^2 \quad (11)$$

$$L_c = \frac{1}{2} \|i - c\|^2 = \frac{1}{2} \|i - b(1 - \alpha) - s\alpha\|^2. \quad (12)$$

In (9)–(11), α , s , and b represented the predicted values of the smoke density, RGB pure smoke pixels and RGB pure background pixels, respectively; α^{gt} , s^{gt} , and b^{gt} represented the corresponding ground truth; In (12), i and c denoted the ground truth and predicted values of the RGB synthesized smoke images, respectively. The setting of these four error terms can constrain the predicted value of each component in a mixed pixel of the smoke image, thereby improving the accuracy of smoke concentration prediction.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Setups

a) Synthetic Dataset and Real Dataset: The experimental data included two parts: synthetic dataset and real dataset. The synthetic dataset was used for training and testing the network model, and the real dataset was used to test the network model trained based on the synthetic data to verify the transferability of the smoke synthetic dataset.

The synthetic dataset was made according to the method described in Section III. The experiments in this chapter construct a smoke synthesis dataset containing 10000 synthetic images and labels with a size of 256×256 . Its background types include airports, highways, forests, built-up areas, and water bodies from the perspective of satellite remote sensing. The synthetic smoke targets had different colors and shapes. In total, 80% of the samples were randomly selected as the training set in the experiments in this chapter, and the remaining 20% were equally divided into the validation set and the test set. The samples are placed according to the file structure of the Pascal VOC public dataset.

The real dataset adopted a set of real-shot optical satellite videos (a total of 200 frames of images) and the corresponding artificially labeled data to test the effectiveness of the synthetic data and the methods in this chapter, as shown in Table II and Fig. 9.

b) Evaluation Metric: Moving object segmentation algorithms can evaluate their performance in terms of both accuracy and efficiency.

In terms of accuracy, the most commonly used evaluation index for image segmentation is intersection over union (IoU),

TABLE II
SPECIFICATIONS OF OPTICAL SATELLITE VIDEO REAL DATA

| Data | Video1 |
|-----------------|--|
| Location | A built-up area where the oil tank caught fire |
| Resolution(m) | 1.1 |
| Band | Panchromatic band |
| Size(pixels) | 640×368 |
| Duration(s) | 29 |
| Frame Rate(FPS) | 24 |
| Sensor | SkySat |

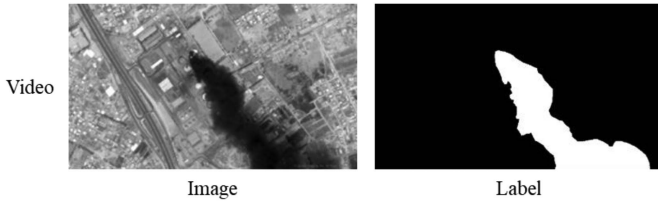


Fig. 9. Example on real data segmentation labels.

which represents the ratio of the area of the intersection between the prediction and the label area to the total area covered by the two, indicating the accuracy of the algorithm prediction. PT represents the set of pixel locations within the prediction area, and GT represents the set of pixel locations within the true label area. The IoU can be represented by the following equation:

$$\text{IoU} = \frac{|PT \cap GT|}{|PT \cup GT|}. \quad (13)$$

In the case of multiclassification, the abovementioned equation can be extended to mean IoU (mIoU):

$$mIoU = \frac{1}{c} \sum_{i=1}^c \text{IoU}_i. \quad (14)$$

In (14), c is the number of categories, and IoU_i is the intersection ratio of the i th category. The task of smoke segmentation is a binary classification problem, so this article took the average of IoU of smoke pixels and background pixels as the processing accuracy of a single-frame image, and calculated the average of mIoU of multiple frames to obtain the processing accuracy of our method.

In terms of efficiency, this chapter adopted the number of predicted frames per second (FPS) to evaluate the segmentation speed of our method.

c) Environment: The experimental environment was the Ubuntu 20.04 operating system. The PyTorch environment was configured, the Visual Studio Code editor was used, and the NVIDIA GeForce RTX 3090 graphics card was used to complete the algorithm implementation, training, and prediction.

We adopted the end-to-end training method and the Adam parameter optimization algorithm. The hyperparameters were set as follows: the initial learning rate was set to $1e-4$, the learning rate decay coefficient was 0.98, the number of batches was set to 8, the training epoch was set to 50 times, the momentum was set to 0.9, and the weight decay was $1e-4$.

B. Ablation Experiments

To verify the effectiveness of each module in the method in this chapter, this section took UNet++ as the basic framework to design a series of ablation experiments, as shown in Table III. The first line is the UNet++ network. The second line embeds the attention mechanism module CBAM into the UNet++ encoder. The third line replaces all VGGNet used for feature extraction in UNet++ with MobileNetV2. The fourth line replaces the binary cross-entropy loss function in UNet++ is the composite loss function described in Section III. The fifth line is AOSVSSNet and the sixth line prunes it.

The test performance of each model in the ablation experiment on the synthetic smoke dataset is shown in Table IV. It can be seen that models 5, 2, and 6, namely ASSNet without pruning, UNet++&CBAM, and AOSVSSNet with pruning, have higher segmentation accuracy, which are 72.58%, 72.23%, and 70.51%, respectively. Models 6, 1, and 2, namely AOSVSSNet with pruning, UNet++, UNet++ & CBAM, have higher segmentation efficiencies 227FPS, 190FPS, and 185FPS, respectively.

As shown in Fig. 10, model 2 added the convolutional attention module CBAM to UNet++, which could incorporate global context information in training, eliminate the interference of irrelevant ground object background information, enhance the distinguishability of smoke areas. Compared with model 1, namely UNet++, its segmentation accuracy was improved by 0.56%. CBAM only introduced a small number of parameters, so that the segmentation efficiency of model 2 was slightly lower than that of model 1, taking into account both accuracy and efficiency.

Model 3 used the lightweight module MobileNetV2 in the feature extraction. The segmentation accuracy of Model 3 on the test data was low, and overfitting occurs. This might be because the expansion coefficient of the inverse residual module is set larger - when the coefficient was set to 6, there were more features used for model fitting, and its segmentation accuracy and efficiency were 61.50% and 104FPS, respectively; when it was set to 2, its segmentation accuracy and efficiency were 67.54% and 245FPS, respectively. In Model 2, CBAM could alleviate the overfitting effect of training data by adjusting the feature weights of space and channels. To fully express the characteristics of smoke images, the expansion coefficient of the inverse residual module in this method was still set to 6, and the pruning operation improves segmentation efficiency.

Model 4 redesigned the UNet++ loss function to obtain a more refined segmentation edge for pixel-by-pixel smoke density estimation and also introduced noise caused by similar ground object backgrounds, resulting in a decrease in the overall segmentation accuracy to 61.50%. The introduction of multiple loss terms computation also reduced the segmentation efficiency to 164FPS.

Model 5 introduced the convolutional attention module, lightweight module, and composite loss function into the basic framework of UNet++, which realized fine segmentation of smoke edges and reduced false alarms caused by incorrect

TABLE III
STRUCTURE DESIGN OF EACH MODEL IN THE ABLATION EXPERIMENT

| No. | Model | UNet++ | CBAM | MobileNetV2 | Composite loss function | Pruning |
|-----|---|--------|------|-------------|-------------------------|---------|
| 1 | UNet++ | ✓ | | | | |
| 2 | UNet++&CBAM | ✓ | ✓ | | | |
| 3 | UNet++&MobileNetV2 | ✓ | | ✓ | | |
| 4 | UNet++&Composite loss function (UNet++&CLF) | ✓ | | | ✓ | |
| 5 | AOSVSSNet without pruning | ✓ | ✓ | ✓ | ✓ | |
| 6 | AOSVSSNet with pruning (Our method) | ✓ | ✓ | ✓ | ✓ | ✓ |

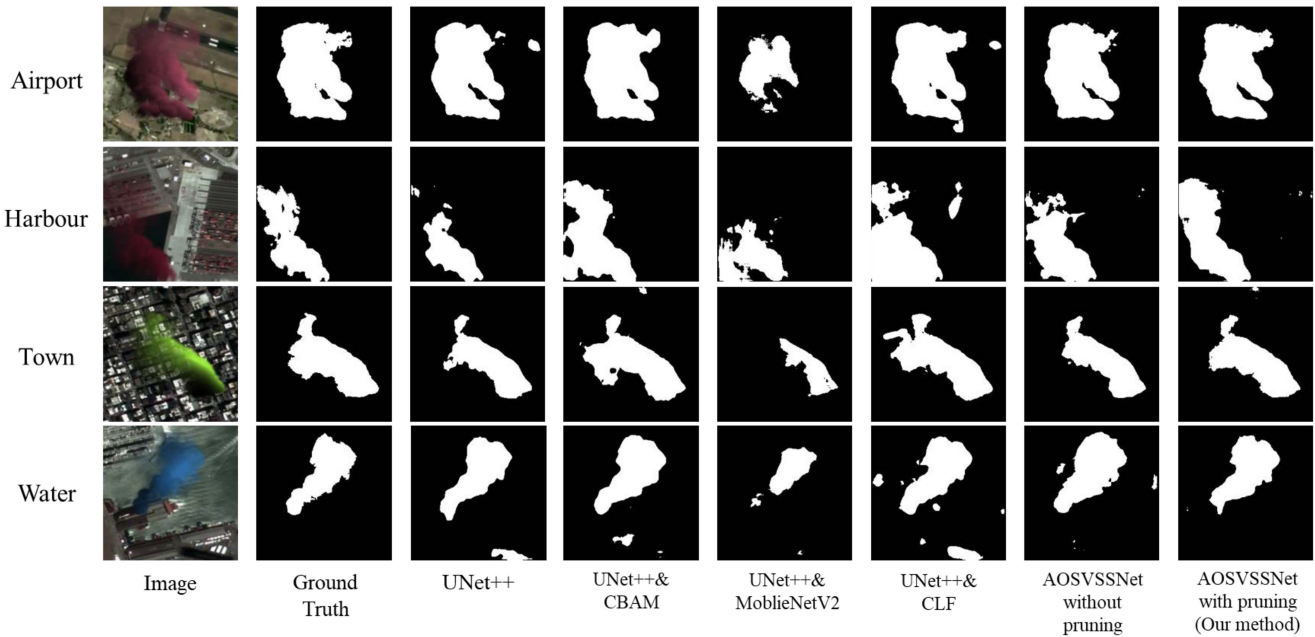


Fig. 10. Synthetic smoke dataset segmentation results for each model in ablation experiments.

TABLE IV
PERFORMANCE OF MODELS IN ABLATION EXPERIMENTS ON SYNTHETIC SMOKE DATASETS

| No. | Model | mIoU(%) | Speed(FPS) |
|-----|---|------------------|---------------|
| 1 | UNet++ | 71.67 (-1.16) | 190 (-37) |
| 2 | UNet++&CBAM | 72.23 (-1.72) | 185 (-42) |
| 3 | UNet++&MobileNetV2 | 61.50 (-9.01) | 104 (-123) |
| 4 | UNet++&Composite loss function (UNet++&CLF) | 64.93 (-5.58) | 164 (-63) |
| 5 | AOSVSSNet without pruning | 72.58 (+2.07) | 94 (-133) |
| 6 | AOSVSSNet with pruning (Our method) | 70.51 | 227 |

TABLE V
PERFORMANCE OF MODELS IN ABLATION EXPERIMENTS ON REAL SMOKE DATASETS

| No. | Model | mIoU(%) | Speed(FPS) |
|-----|---|-------------------|------------|
| 1 | UNet++ | 41.35 (-31.49) | 15 (-3) |
| 2 | UNet++&CBAM | 65.67 (-7.17) | 16 (+4) |
| 3 | UNet++&MobileNetV2 | 59.42 (-13.42) | 10 (-2) |
| 4 | UNet++&Composite loss function (UNet++&CLF) | 73.56 (+0.72) | 16 (+4) |
| 5 | AOSVSSNet without pruning | 68.81 (-4.03) | 10 (-2) |
| 6 | AOSVSSNet with pruning (Our method) | 72.84 | 12 |

segmentation of ground objects and backgrounds. Compared with UNet++, the accuracy was improved by 0.91%. Its pruned version, Model 6, significantly improves the segmentation efficiency, reaching 227FPS, while slightly reducing the segmentation accuracy, which was 19.47% and 141.49% higher than Model 1 and Model 5, respectively.

The test performance of each model in the ablation experiment on the real smoke dataset is shown in Table V. It could be seen that the segmentation accuracy of models 4, 6, and 5, namely UNet++ and composite loss function, AOSVSSNet without pruning, and AOSVSSNet with pruning, are 73.56%, 72.84%, and 68.81%, respectively. Models 2, 4, 1, and 6,

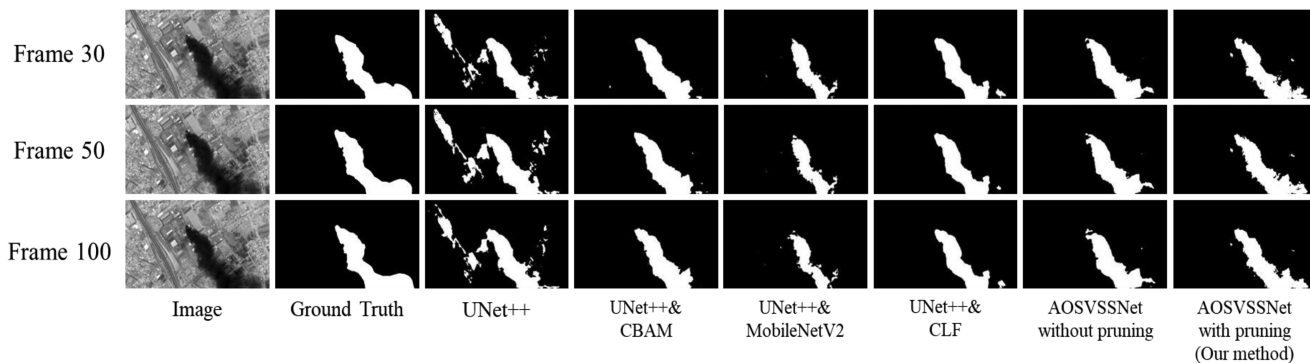


Fig. 11. Segmentation results of real smoke datasets for each model in ablation experiments.

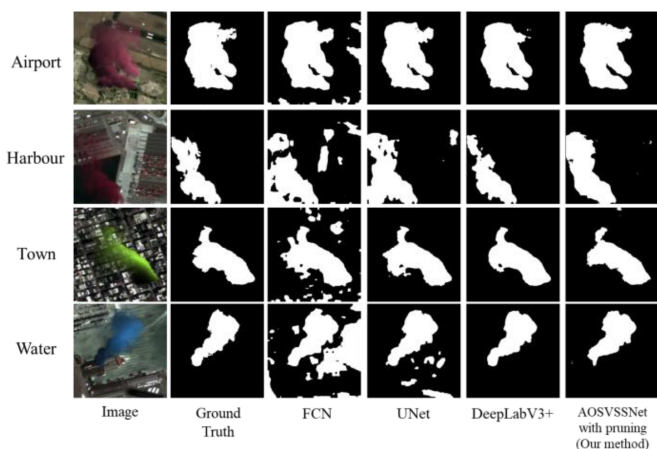


Fig. 12. Synthetic smoke dataset segmentation results for each model in comparative experiments.

namely UNet++&CBAM, UNet++&composite loss function, UNet++, and AOSVSSNet with pruning had higher segmentation efficiency, 16FPS, 16FPS, 15FPS, and 12FPS, respectively.

Among them, Model 4 achieved the highest segmentation accuracy on the real smoke dataset, indicating that the composite loss function based on the physical process constraints of the optical imaging of the smoke target could effectively improve the generalization of the UNet++ network model, and the synthetic smoke dataset had better performance. The segmentation edge of model 4 was relatively stable between video frames, indicating that the composite loss function based on concentration estimation helped to enhance the smoothness of video object segmentation. As shown in Fig. 11, Model 2 with CBAM module could effectively focus on the smoke information and eliminate the interference of similar ground objects. Compared with Model 1, its segmentation accuracy was significantly improved by 24.32%, indicating that the enhancement of spatial dependencies between pixels played an important role in aggregating homogeneous smoke pixels and enhancing the distinguishability of similar ground objects. Model 3 had an overfitting problem, and the training accuracy was high but the test accuracy on synthetic and real smoke datasets was lower, at 59.42%, and loosed more smoke pixels. The segmentation

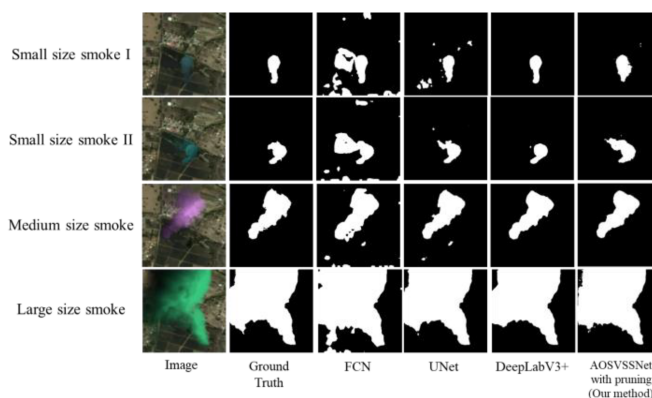


Fig. 13. Segmentation results of different scales of synthetic smoke datasets for each model in the comparative experiment.

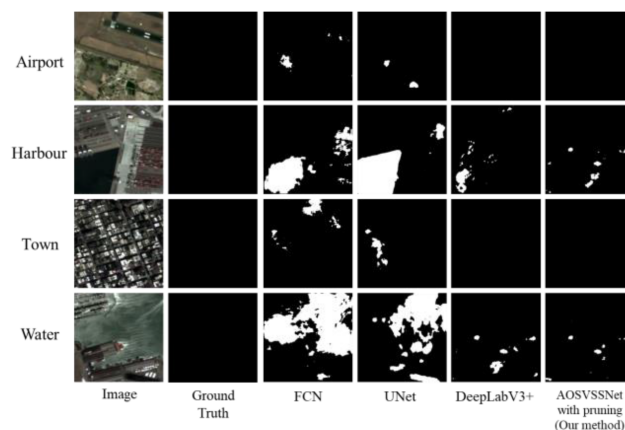


Fig. 14. Smoke-free image segmentation results of each model in the comparative experiment.

accuracy of model 6 was 4.03% higher than that of model 5, and it could segment the pixels with lower concentration at one end of the smoke diffusion. This might be because the pruning operation not only reduces the computational complexity of the model, but also alleviated the degree of overfitting of the model on the training dataset.

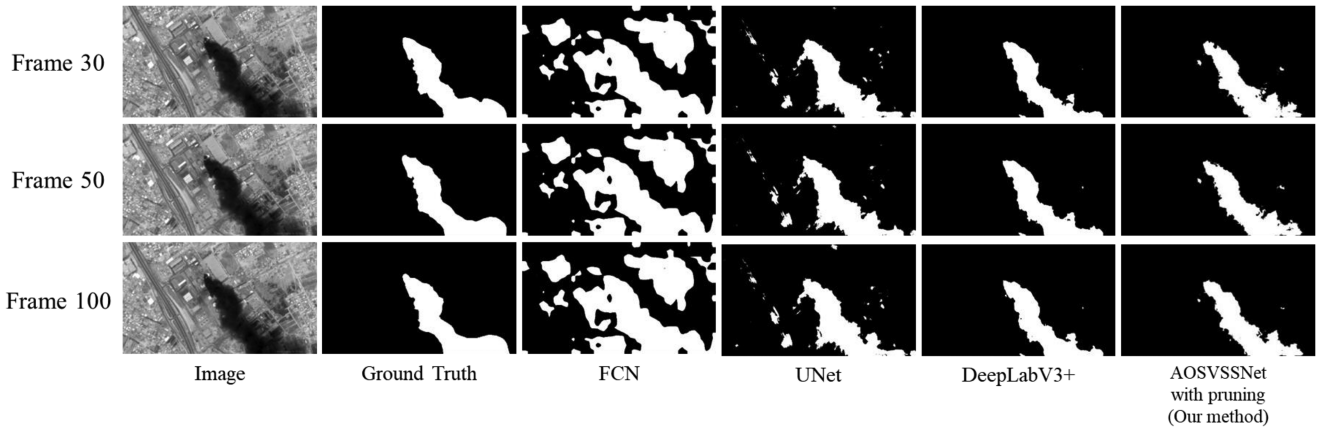


Fig. 15. Segmentation results of real smoke datasets for each model in comparative experiments.

TABLE VI
PERFORMANCE OF MODELS IN COMPARATIVE EXPERIMENT ON REAL SMOKE DATASETS

| No. | Model | mIoU(%) | Speed(FPS) |
|-----|--|-------------------|-------------|
| 1 | FCN | 29.63 (-43.21) | 14 (-2) |
| 2 | UNet | 65.67 (-7.17) | 25 (-13) |
| 3 | DeepLabV3+ | 72.14 (-0.7) | 9 (-3) |
| 4 | AOSVSSNet with pruning (Our method) | 72.84 | 12 |

TABLE VII
PERFORMANCE OF MODELS IN COMPARATIVE EXPERIMENT ON SYNTHETIC SMOKE DATASETS

| No. | Model | mIoU(%) | Speed(FPS) |
|-----|--|-------------------|--------------|
| 1 | FCN | 54.13 (-16.38) | 233 (+6) |
| 2 | UNet | 68.79 (-1.72) | 185 (-42) |
| 3 | DeepLabV3+ | 69.33 (-1.18) | 166 (-61) |
| 4 | AOSVSSNet with pruning (Our method) | 70.51 | 227 |

C. Comparative Experiments

This section selected FCN, UNet, and DeepLabV3+ as the representatives of the three main structures of the classic semantic segmentation network to verify the effectiveness of AOSVSSNet with pruning.

In the comparison experiment, the test performance of each model in the synthetic smoke dataset is shown in Table VI. It can be seen that models 4 and 3, namely AOSVSSNet with pruning and DeepLabV3+, had higher segmentation accuracy, 70.51% and 69.33%, respectively. Models 1 and 4, namely FCN and AOSVSSNet with pruning, had higher segmentation efficiency, 233FPS and 227FPS, respectively.

As shown in Fig. 12, in addition to the smoke target, the segmentation results of Model 1, FCN, more background objects present, and the segmentation accuracy was 54.13%. This was because although FCN combined the segmentation results of the rough layer and the fine layer, so that the prediction of local pixels followed the global ground object distribution structure to a certain extent, it still didn't fully consider the relationship between pixels and lacks local information consistency. This led to the appearance of a large number of missegmented patches of similar ground objects. Model 2, UNet, fused the high and low-level features output by the same layer of the encoder and decoder and upsampled them layer-by-layer, which narrowed the semantic gap and greatly improved the segmentation accuracy to 68.79%, which was 14.66% higher than

Model 1. Model 3, DeepLabV3+, could fuse information of various scales without reducing the feature resolution through the hole convolution pyramid, equivalent to incorporating more fine global context information, and its segmentation accuracy was 69.33%. The accuracy of AOSVSSNet with pruning was 70.51%, but the segmentation efficiency was higher, which was 36.75% higher than that of DeepLabV3+, realizing the unity of accuracy and efficiency.

The synthetic smoke dataset shown in Fig. 13 contains small, medium, and large-sized smoke targets. The segmentation results showed that our method could adapt to the segmentation of smoke targets of different sizes and could simulate different spatial scales from the perspective of satellite remote sensing and the actual smoke scene at different periods. In addition, as shown in Fig. 14, our method and model 3 of this article, namely DeepLabV3+, had lower false alarms in smoke-free images. It resulted from the enhancement of the global context information feature make the false alarms of similar ground objects suppressed, thereby improving smoke segmentation accuracy of the target.

The test performance of each model in the comparison experiment on the real smoke dataset is shown in Table VII and Fig. 15. It can be seen that models 4 and 3, AOSVSSNet with pruning and DeepLabV3+, had higher segmentation accuracy, 72.84% and 72.14%, respectively, indicating good generalization. Models 2 and 1, U-Net and FCN, had higher segmentation efficiency but lower accuracy. Combined with the ablation experiments,

the reason was that the methods of Model 4 and 3 integrated more spatial context information in the training process in different ways. DeepLabV3+ utilized atrous convolution pyramid structure to fuse multilevel fine feature information. Our method combined the convolutional attention module and the complex loss function to enhance the expression of spatial dependencies between pixels, making the features of the smoke target and the background features more distinguishable, and improving the segmentation accuracy.

V. CONCLUSION

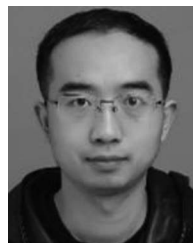
In this article, a deep learning method is innovatively introduced into optical satellite video smoke object segmentation. An attention-guided optical satellite video smoke segmentation network model for optical satellite video is proposed to aim at the multiscale segmentation of satellite video smoke targets and the background interference of complex and similar ground objects. Based on UNet++, a lightweight attention module CBAM enhances the smoke target features, effectively suppresses the false alarm of the ground object background, and achieves high segmentation accuracy on the synthetic dataset. A synthetic dataset is constructed based on the optical imaging process of smoke targets to solve the difficulty of manual labeling and model segmentation of deep learning samples due to blurred smoke edges, which saves manual labeling costs. In addition, it introduces the physical constraints of the smoke imaging process into the loss function and improves the generalization of the model to real smoke data.

Future work mainly focuses on optimizing the context feature extraction method, improving the network's ability to fuse global and local features, further reducing the missegmentation and missing pixels of smoke, and testing in different real smoke scenes. In addition, there are differences in the imaging process between satellite video and natural images. Remote sensing physical mechanisms such as atmospheric radiative transfer models can be considered as constraints and integrated into the model to enhance the interpretability and generalization of the smog segmentation network.

REFERENCES

- [1] L. Zhang, "Image smoke semantic segmentation based on deep learning," Ph.D. dissertation, Jiangxi Univ. of Finance and Economics, Nanchang, China, 2020.
- [2] J. Shi, F. Yuan, and X. Xia, "Video smoke detection: A literature survey," *Chin. J. Image Graph.*, vol. 23, no. 3, pp. 303–302, 2018.
- [3] X. Xia, F. Yuan, L. Zhang, L. Yang, and J. Shi, "From traditional methods to deep ones: Review of visual smoke recognition, detection, and segmentation," *Chin. J. Image Graph.*, vol. 24, no. 10, pp. 1627–1647, 2019.
- [4] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.
- [5] C. Deng, D. Jing, Y. Han, S. Wang, and H. Wang, "FAR-Net: Fast anchor refining for arbitrary-oriented object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6505805.
- [6] W. Wang, Y. Han, C. Deng, and Z. Li, "Hyperspectral image classification via deep structure dictionary learning," *Remote Sens.*, vol. 14, no. 9, 2022, Art. no. 2266.
- [7] L. Tang, W. Tang, X. Qu, Y. Han, W. Wang, and B. Zhao, "A scale-aware pyramid network for multi-scale object detection in SAR images," *Remote Sens.*, vol. 14, no. 4, 2022, Art. no. 973.
- [8] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.
- [9] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.
- [10] T.-H. Chen, Y.-H. Yin, S.-F. Huang, and Y.-T. Ye, "The smoke detection for early fire-alarming system based on video processing," in *Proc. Int. Conf. Intell. Inf. Hiding Multimedia*, 2006, pp. 427–430.
- [11] K. Dimitropoulos, P. Barmoutis, and N. Grammalidis, "Higher order linear dynamical systems for smoke detection in video surveillance applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 5, pp. 1143–1154, May 2017.
- [12] Y. Zhao, Q. Li, and Z. Gu, "Early smoke detection of forest fire video using cs adaboost algorithm," *Optik-Int. J. Light Electron Opt.*, vol. 126, no. 19, pp. 2121–2124, 2015.
- [13] L. Wang, A. Li, X. Wang, and Y. Yu, "Early fire smoke detection based on multi-feature fusion," *J. Dalian Maritime Univ., Natural Sci. Ed.*, vol. 40, no. 1, pp. 97–100, 2014.
- [14] J. Liu, H. Zhao, and T. Zhao, "Research of flame detection on visual saliency method," *J. Comput.*, vol. 8, no. 12, pp. 3264–3271, 2013.
- [15] Y. Han, C. Deng, B. Zhao, and D. Tao, "State-aware anti-drift object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4075–4086, Aug. 2019.
- [16] C. Deng, Y. Han, and B. Zhao, "High-performance visual tracking with extreme learning machine framework," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2781–2792, Jun. 2020.
- [17] L. Tang, W. Tang, X. Qu, Y. Han, W. Wang, and B. Zhao, "A scale-aware pyramid network for multi-scale object detection in SAR images," *Remote Sens.*, vol. 14, no. 4, p. 973, 2022.
- [18] Y. Wang, "Smoke detection based on computer vision in coal mine," *J. Liaoning Univ. Eng. Technol., Natural Sci. Ed.*, vol. 35, no. 11, pp. 1230–1234, 2016.
- [19] L. Dong and J. Yu, "Smoke detection method in video based on image separation," *Comput. Eng.*, vol. 41, no. 9, pp. 251–254, 2015.
- [20] H. Li and F. Yuan, "Image based smoke detection using pyramid texture and edge features," *Chin. J. Image Graph.*, vol. 20, no. 6, pp. 772–780, 2015.
- [21] F. Yuan, X. Xia, J. Shi, H. Li, and G. Li, "Non-linear dimensionality reduction and Gaussian process based classification method for smoke detection," *IEEE Access*, vol. 5, pp. 6833–6841, 2017.
- [22] C. E. Prema, S. Vinsley, and S. Suresh, "Multi feature analysis of smoke in YUV color space for early forest fire detection," *Fire Technol.*, vol. 52, no. 5, pp. 1319–1342, 2016.
- [23] W. Ye, J. Zhao, S. Wang, Y. Wang, D. Zhang, and Z. Yuan, "Dynamic texture based smoke detection using Surfacelet transform and HMT model," *Fire Saf. J.*, vol. 73, pp. 91–101, 2015.
- [24] W. Ye, J. Zhao, Y. Zhao, and Y. Wang, "Smoke detection based on Surfacelet transform and dynamic texture," *Comput. Eng.*, vol. 41, no. 2, pp. 203–208, 2015.
- [25] S. Ye, Z. Bai, H. Chen, R. Bohush, and S. Ablameyko, "An effective algorithm to detect both smoke and flame using color and wavelet analysis," *Pattern Recognit. Image Anal.*, vol. 27, no. 1, pp. 131–138, 2017.
- [26] J. Sun and R. Yang, "Smoke detection in video based on color histogram and wavelets," *Comput. Sci.*, vol. 41, no. 12, pp. 251–254, 2014.
- [27] H. He, L. Peng, D. Yang, and X. Chen, "Smoke detection based on a semi-supervised clustering model," in *Proc. Int. Conf. Multimedia Model.*, Springer, 2014, pp. 291–298.
- [28] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, and Y. Wang, "Fourier-based rotation-invariant feature boosting: An efficient framework for geospatial object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 302–306, Feb. 2020.
- [29] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, "ORSim detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5146–5158, Jul. 2019.
- [30] Y. Jia, J. Yuan, J. Wang, J. Fang, Q. Zhang, and Y. Zhang, "A saliency-based method for early smoke detection in video sequences," *Fire Technol.*, vol. 52, no. 5, pp. 1271–1292, 2016.
- [31] Q. Liu, G. Cui, P. Wu, and C. Li, "Remove of fix interference based on machine learning in smoke detection," *Comput. Meas. Control*, vol. 23, no. 3, pp. 880–881, 2015.
- [32] I. F. Ince, M. E. Yildirim, Y. B. Salman, O. F. Ince, G.-H. Lee, and J.-S. Park, "Fast video fire detection using luminous smoke and textured flame features," *KSII Trans. Internet Inf. Syst.*, vol. 10, no. 12, pp. 5485–5506, 2016.

- [33] M. Wang, Q. Wang, D. Hong, S. K. Roy, and J. Chanussot, "Learning tensor low-rank representation for hyperspectral anomaly detection," *IEEE Trans. Cybern.*, to be published, doi: [10.1109/TCYB.2022.3175771](https://doi.org/10.1109/TCYB.2022.3175771).
- [34] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.
- [35] Z. Zhao, Y. Han, T. Xu, X. Li, H. Song, and J. Luo, "A reliable and real-time tracking method with color distribution," *Sensors*, vol. 17, no. 10, 2017, Art. no. 2303.
- [36] X. Liu, D. Hong, J. Chanussot, B. Zhao, and P. Ghamisi, "Modality translation in remote sensing time series," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5401614.
- [37] C. Tao, J. Zhang, and P. Wang, "Smoke detection based on deep convolutional neural networks," in *Proc. Int. Conf. Ind. Inform.-Comput. Technol., Intell. Technol., Ind. Inf. Integration*, 2016, pp. 150–153.
- [38] Z. Yin, B. Wan, F. Yuan, X. Xia, and J. Shi, "A deep normalization and convolutional neural network for image smoke detection," *IEEE Access*, vol. 5, pp. 18 429–18 438, 2017.
- [39] A. S. Pundir and B. Raman, "Deep belief network for smoke detection," *Fire Technol.*, vol. 53, no. 6, pp. 1943–1960, 2017.
- [40] M. D. Nguyen, D. Kim, and S. Ro, "A video smoke detection algorithm based on cascade classification and deep learning," *KSH Trans. Internet Inf. Syst.*, vol. 12, no. 12, pp. 6018–6033, 2018.
- [41] T. B. University, "Computer vision based fire detection software." [Online]. Available: <http://signal.ee.bilkent.edu.tr/VisiFire/index.html>
- [42] K. University. Korea, "Computer vision and pattern recognition laboratory." [Online]. Available: <http://cvpr.kmu.ac.kr/>
- [43] D. Y. Chino, L. P. Avalhais, J. F. Rodrigues, and A. J. Traina, "Bowfire: Detection of fire in still images by integrating pixel color and texture analysis," in *Proc. 28th SIBGRAPI Conf. Graph., Patterns Images*, 2015, pp. 95–102.
- [44] U. of Science and T. of China, "State key laboratory of fire science." [Online]. Available: <http://sklfs.ustc.edu.cn/>
- [45] F. Yuan, L. Zhang, X. Xia, B. Wan, Q. Huang, and X. Li, "Deep smoke segmentation," *Neurocomputing*, vol. 357, pp. 248–260, 2019.
- [46] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5517010.
- [47] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, "Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model," *ISPRS J. Photogrammetry Remote Sens.*, vol. 178, pp. 68–80, 2021.
- [48] R. Ba, C. Chen, J. Yuan, W. Song, and S. Lo, "SmokeNet: Satellite smoke scene detection using convolutional neural network with spatial and channel-wise attention," *Remote Sens.*, vol. 11, no. 14, 2019, Art. no. 1702.
- [49] Z. Hong et al., "Active fire detection using a novel convolutional neural network based on Himawari-8 satellite images," *Front. Environ. Sci.*, vol. 10, p. 102, 2022.
- [50] Z. Wang et al., "Semantic segmentation and analysis on sensitive parameters of forest fire smoke using smoke-UNET and landsat-8 imagery," *Remote Sens.*, vol. 14, no. 1, p. 45, 2021.
- [51] M. Ramasubramanian et al., "Pixel level smoke detection model with deep neural network," in *Proc. Image Signal Process. Remote Sens. XXV*, 2019, vol. 11155, pp. 376–386.
- [52] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
- [53] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, Springer, 2015, pp. 234–241.
- [54] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
- [55] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.
- [56] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.
- [57] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
- [58] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
- [59] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
- [60] G. Xu et al., "Video smoke detection based on deep saliency network," *Fire Saf. J.*, vol. 105, pp. 277–285, 2019.
- [61] F. Yuan, L. Zhang, X. Xia, Q. Huang, and X. Li, "A wave shaped deep neural network for smoke density estimation," *IEEE Trans. Image Process.*, vol. 29, pp. 2301–2313, 2020.
- [62] F. Yuan, L. Zhang, X. Xia, Q. Huang, and X. Li, "A gated recurrent network with dual classification assistance for smoke semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 4409–4422, 2021.
- [63] X. Li, Z. Chen, Q. J. Wu, and C. Liu, "3d parallel fully convolutional networks for real-time video wildfire smoke detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 89–103, Jan. 2020.
- [64] P. Tokmakov, K. Alahari, and C. Schmid, "Learning motion patterns in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3386–3394.
- [65] P. Tokmakov, K. Alahari, and C. Schmid, "Learning video object segmentation with visual memory," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4481–4490.
- [66] S. Dutt Jain, B. Xiong, and K. Grauman, "FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3664–3673.
- [67] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper ConvLSTM for video salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 715–731.
- [68] X. Lu, W. Wang, J. Shen, D. Crandall, and J. Luo, "Zero-shot video object segmentation with co-attention Siamese networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2228–2242, Apr. 2022.
- [69] W. Wang, X. Lu, J. Shen, D. J. Crandall, and L. Shao, "Zero-shot video object segmentation via attentive graph neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9236–9245.
- [70] N. Max, "Optical models for direct volume rendering," *IEEE Trans. Vis. Comput. Graph.*, vol. 1, no. 2, pp. 99–108, Jun. 1995.
- [71] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.
- [72] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2970–2979.
- [73] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning For Clinical Decision Support*. New York, NY, USA: Springer, 2018, pp. 3–11.
- [74] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.
- [75] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.
- [76] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.
- [77] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.



Taoyang Wang received the B.E. and Ph.D.E. degrees in photogrammetry and remote sensing from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2007 and 2012, respectively.

His doctoral dissertation concerned the block adjustment of high-resolution satellite remote sensing imagery. He has been with the School of Remote Sensing and Information Engineering, Wuhan University, since 2014, where he became an Associate Research Fellow in 2015. His research interests include space photogrammetry, geometry processing of spaceborne optical/SAR/InSAR imagery, target detection, and recognition based on satellite video.



Jianzhi Hong received the B.S. degree in geographic information science from South China Normal University, Guangzhou, China, in 2020, and the M.E. degree in resources and environment in 2022 from Wuhan University, Wuhan, China, where he is currently working toward the Ph.D. degree in photogrammetry and remote sensing.

His main research interests include intelligent interpretation of moving objects in optical satellite video, including detection, tracking, and segmentation.



Yuqi Han received the B.Eng. degree in the field of information engineering from the Beijing Institute of Technology, Beijing, China, in 2015, the B.Sc. degree in the field of economy from the National School of Development, Peking University, Beijing, China, in 2015, and the Ph.D. degree in information and communication engineering from the School of Information and Electronics, Beijing Institute of Technology, Beijing, China, in 2021.

He is a Research Fellow with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include computer vision, remote sensing, and UAV.



Guo Zhang received the B.E. and Ph.D.E. degrees in photogrammetry and remote sensing from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2000 and 2005, respectively.

His doctoral dissertation concerned the rectification of high-resolution remote sensing imaging under lack of ground control points. He has been with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, since 2005, where he became a Professor in 2011. His research interests include space photogrammetry, geometry processing of spaceborne optical/SAR/InSAR imagery, altimetry, and high-accuracy image matching.



Shili Chen received the B.E. degree in remote sensing and information engineering in 2021 from Wuhan University, Wuhan, China, where she is currently working toward the M.E. degree in photogrammetry and remote sensing with the School of Remote Sensing and Information Engineering.

Her research interests include object tracking and machine learning.

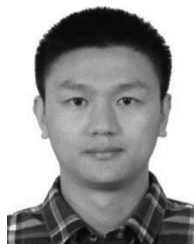


Tiancheng Dong received the B.S. degree in remote sensing science and technology and the M.S. degree in geophysical prospecting and information technology from the China University of Geosciences, Wuhan, China, in 2018 and 2021, respectively. He is currently working toward the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China.



Yapeng Yang received the M.S. and Ph.D. degrees in control science and engineering from the College of Mechatronics and Automation, National University of Defense Technology, Changsha, China, in 2010 and 2015, respectively.

He is currently with the Navy Research Institute of PLA, Beijing, China. His research interests include remote sensing information processing.



Hang Ruan received the B.S. degree in electronic information engineering from Zhejiang University, Hangzhou, China, in 2008, and the Ph.D. degree in communication and information systems from the Department of Photoelectric Equipment, Academy of Equipment, Beijing, China, in 2013.

He is currently an Assistant Researcher with the Beijing Institute of Tracking and Telecommunications Technology, Beijing, China. His main research interests include radar imaging and intelligent interpretation of SAR images.