Near Real-Time Monitoring of Fire Spots Using a Novel SBT-FireNet Based on Himawari-8 Satellite Images

Zhonghua Hong, *Member, IEEE*, Zhizhou Tang, Haiyan Pan, Yuewei Zhang, Zongsheng Zheng, Ruyan Zhou, Yun Zhang, Yanling Han, Jing Wang, and Shuhu Yang

***Abstract*—Detailed and timely monitoring of the location and intensity of the fire is critical to reducing the destructive impacts of a fire. Satellite imagery platforms, in particular geostationary satellites with high temporal resolution, allow for real-time fire monitoring. However, because of the coarse resolution of geostationary satellite images, even when deep learning models are applied, precision still remains limited. Thus, the prediction models easily fall into a local optimal solution because of the insufficient semantic information from low spatial resolution. Therefore, in this study, we proposed a novel deep learning model, SBT-FireNet, to monitor fire spots from Himawari-8 satellite images. Specifically, the extraction modules of spatial, band, and time-series features were designed and integrated into the model. The spatial feature extraction module served to collate information about fires and their surrounding environment, while the band and time-series features were designed to obtain fire-sensitive band and time information, respectively. The newly structured SBT-FireNet model was tested in four fire-prone areas with high forest cover. The precision of SBT-FireNet in four test areas is 35.2% higher than other methods. The model yielded significant improvements through the combination of the modules of spatial, band, and time-series features and their fire-tailored design. The advantages of the high temporal resolution of geostationary satellite images were fully integrated into the model to ensure that the model monitors the possibility of fire in an automated way in a near-real-time manner.**

***Index Terms*—Convolutional neural network (CNN), deep learning, LSTM, near-real-time monitoring, transformer, wildfire.**

I. Introduction

GLOBALLY, not only are wildfires among the most important natural phenomena but they are also vital to the health and productivity of the ecosystem. In recent years, their frequency has increased [1]. Although there are many types of ecosystem services [2], wildfires can also cause casualties and property losses [3]. The researchers explored trends associated with fire interference and gradual recovery in the global north and temperate forests in the study area and found that between 2001 and 2018, 181 million hectares of the study area were burned [4]. Globally, the annual burned area was estimated at approximately 420 million hectares [5]. Therefore, timely and accurate monitoring of wildfires is of great importance for fire management and policy, as well as for conservation of nature [6]. Due to its significant advantages, such as wide coverage, low cost, and repeated monitoring, satellite remote sensing technology has been widely used to monitor fires over the past decades. Therefore, several algorithms for fire detection using satellite images have been proposed.

Fire detection methods can be categorized into threshold and deep learning approaches [7]. Threshold methods usually use abnormal warming characteristics of fire (in particular, abnormal increase in emissivity) in the mid-infrared and far-infrared bands, with a threshold value to divide all pixels into fire and nonfire spots [8], [9], [10], [11], [12]. To suppress disturbance factors, such as clouds, smoke, and snow, contextual or temporal information is integrated to generate stronger results [13], [14], [15], [16]. The context threshold method sets a pixel as center and a window with a size of X × Y as background and removes a quantity of background interference pixels that can be caused by clouds, water, or other factors, based on a series of thresholds. Finally, after the average temperature (brightness temperature) variance of effective background pixels is calculated as the background feature of the central pixel, fire spots are detected using the difference in the features of the background and fire spots. The National Oceanic and Atmospheric Administration has developed active fire products based on the context threshold method by using the Moderate-Resolution Imaging Spectroradiometer (MODIS) [17]. The MODIS fire products and the global fire spot monitoring system developed using the Medium Resolution Spectrum Imager-II onboard FY-3D are based on typical context threshold methods. The context threshold methods require that the surrounding pixels have background pixels whose certain availability and quality standards are met and not compromised, such as cloud coverage and background pixels of the same nature as the central pixel [18], [19], [20]. In

Manuscript received 17 December 2022; revised 30 May 2023, 9 August 2023, and 19 October 2023; accepted 24 November 2023. Date of publication 1 December 2023; date of current version 20 December 2023. This work was supported by the National Natural Science Foundation of China under Grant 42241164 and Grant 41871325. *(Corresponding author: Haiyan Pan.)*

Zhonghua Hong, Zhizhou Tang, Haiyan Pan, Zongsheng Zheng, Ruyan Zhou, Yun Zhang, Yanling Han, Jing Wang, and Shuhu Yang are with the College of Information Technology, Shanghai Ocean University, Shanghai 201306, China (e-mail: zhhong@shou.edu.cn; m200711468@st.shou.edu.cn; hy-pan@shou.edu.cn; zszheng@shou.edu.cn; ryzhou@shou.edu.cn; y-zhang@shou.edu.cn; ylhan@shou.edu.cn; wangjing@shou.edu.cn; shyang@shou.edu.cn).

Yuewei Zhang is with the Guangzhou Meteorological Satellite Ground Station, Guangzhou, 510000, China (e-mail: weixing1132@126.com).

Digital Object Identifier 10.1109/JSTARS.2023.3338448

general, traditional threshold methods face the issues of high threshold sensitivity and poor applicability. In addition to the poor spatiotemporal generalizability of thresholds, even minor changes in them may result in the omission of small target fires and a large number of false positives [11], [21].

Thanks to the rapid development of deep learning technology, advanced features can now be extracted using a deep learning model, such as convolutional neural networks (CNNs) with convolution kernels as the core, recurrent neural networks with sequence data as input and transformer [22], [23], [24]. He et al. [25] proposed the ResNet model in 2016, which solves the problem of gradient vanishing and exploding that may occur during deep learning training through a residual structure. On the basis of ResNet, a number of variants have been presented, e.g., ResNeSt [26], iResNet [27], and Res2Net [28], by adding attention mechanisms, improving the information flow, and using other methods. Chen et al. [29] proposed FasterNet, which reduces redundant computation problems in CNNs through partial convolution and improves the spatial feature extraction efficiency of the model. In 2020, Dosovitskiy et al. [30] introduced the transformer structure into the field of computer vision and proposed the vision transformer (ViT) model, which had better performance compared to the most advanced CNNs at the time. Compared to CNNs, ViT has better global feature extraction ability, and it possesses stronger sequence data processing ability through the self-attention mechanism. Liu et al. [31] further extended the transformer structure to semantic segmentation tasks and proposed the Swin Transformer. Lee et al. [32] used overlapping convolutional patches to embed and improve ViT, and proposed MPViT. MPViT exhibits excellent performance in object detection, strength segmentation, and semantic segmentation tasks. Due to the intelligence and automation of deep learning models [33], [34], [35], they have been widely used in fire-monitoring tasks. For example, Langford et al. [36] adopted a weight selection strategy to solve the issue of sample imbalance in the training set between fire and nonfire conditions and detected wildfires more accurately using DNNs with this strategy than one without it. Ba et al. [37] designed a novel CNN model, SmokeNet, to detect smoke scenes in satellite images by integrating spatial information and band-wise information attention mechanisms to extract more effective features. Zhang et al. [22] used the Siamese self-attention classification strategy to map burned areas, based on the data synergy of Sentinels 1 and 2. De Almeida Pereira et al. [38] applied the U-Net structure to a fire segmentation task based on Landsat-8 images and obtained accurate results on the global image dataset. Martins et al. [7] proposed an improved U-Net structure for PlanetScope image fire segmentation, using Landsat-8 data for transfer learning. Seydi et al. [39] designed a two-stream feature deep learning framework, Fire-Net, for fire segmentation of Landsat-8 images with optical and thermal modalities. Yolov3, SqueezeNet, and other deep learning models have been used to detect fires from images and surveillance videos of unmanned aerial vehicle [40], [41], [42], [43]. In general, these methods apply the classical semantic segmentation network to satellite remote sensing images, with the assumption that the images can provide sufficient semantic information, that is, high spatial resolution. However, these sun-synchronous satellites usually have a lower temporal resolution of 5–16 days, which is not conducive to wildfire monitoring [44], [45], [46].

The temporal resolution of geosynchronous satellites is significantly higher than that of sun-synchronous satellites and allows for a continuous monitoring and time-series data of a specific geographical location. For example, the 10 min temporal resolution of Himawari-8 spots to the possibility of near-real-time monitoring or early warning of fires. Thus far, few studies have monitored wildfires based on geosynchronous satellite images using deep learning methods. To the best of the authors' knowledge, there exist only two related studies in the related literature. Ding et al. [47] proposed a fully connected CNN model based on the Himawari-8 satellite platform, with its wildfire detection accuracy of $> 80\%$, which was significantly higher than that of the other traditional machine learning algorithms, such as support vector machine, random forest and clustering of k-means. However, the testing data in their study mostly included images of large-area fires, although most small fires occupy only a few pixels; thus, their model performance remains to be validated against small fire events. In addition, in our previous study [48], FireCNN was designed as a novel CNN to detect fire spots from Himawari-8 satellite images, possessed multiscale convolution and residual acceptance to learn the features of fire spots, and resulted in 51.7% higher prediction accuracy than traditional fixed threshold methods. However, its disadvantage is that variance and mean values associated with environmental factors are manually added to the data before processing to enhance the fire feature representation.

Currently, deep learning of wildfires from geosynchronous satellite images remains a challenge for the following reasons. First, the spatial resolution of geostationary satellites is relatively low, causing fire spots occupy only a small portion of the original image. The limited number of fire spots, combined with an overwhelmingly large number of nonfire spots, leading to an imbalanced sample size between the categories, which will inevitably impact the accurate training of deep learning models and increases the risk of false negatives. When the existing semantic segmentation network is implemented, the model tends to identify all pixels as nonfire spots and falls into a local optimal solution. Second, a strategy is needed to screen the samples for the deep learning models to be trained, as the quantity of nonfire spots is higher than that of fire spots. Otherwise, the network is not able to fully learn and causes high false positives or omissions. Finally, given the incomplete semantic information of the fire pixels, they are easily blurred because of the low spatial resolution, suggesting that more features are needed to explore the features of fire spots and improve detection precision.

Therefore, the purpose of this study was to propose a novel deep learning model to detect fire spots using Himawari-8 satellite images. In particular, the following three critical research questions were posed.

1) How can a spatiotemporally robust dataset be built to resolve the issue of imbalanced data between fire spots and nonfire spots?

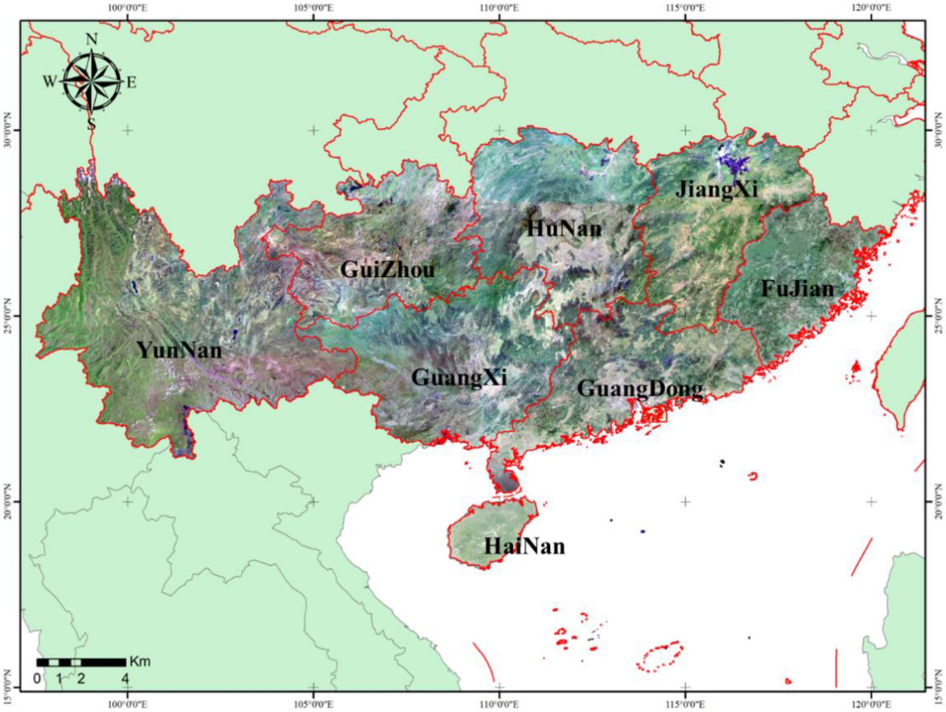


Fig. 1. Study areas.

2) How can we further explore the features of fire spots from low spatial resolution by leveraging the advantages of high temporal resolution?
3) How can a robust deep learning model be constructed to achieve accurate and precise detection of fire spots?

The next is the arrangement for the rest of this article. In Section II, we introduce the study area and the dataset used in this article. Section III describes the methodology, in particular, data processing and extraction modules of spatial, band, and time features. Section IV discusses the experimental results. Section V. discussed the model and conducted ablation experiments. Finally, a conclusion summary is provided in Section VI.

II. Study Area and Datasets

The study area comprises the following eight provinces located in southern China between 18°10’N-30°08’N and 97°31′E-120°30E: Fujian, Guangdong, Guangxi, Hunan, Jiangxi, Hainan, Yunnan, and Guizhou (Fig. 1). The study areas are characterized by a subtropical or tropical monsoon climate, with long-term annual temperature range of 12°C–29°C. The percentage of forest cover in the eight provinces is greater than 55%. From April to December, due to the dry and warm weather, the probability of forest fires increases greatly. Furthermore, local residents with their traditional customs, such as ancestor worship, burning candles, and other sacrificial rituals, accidentally increase the probability of forest fires [53]. The detailed information of the research area is shown in Table I.

The Himawari-8 satellite is the source of data used in this study. As a successor to the MTSAT series of geosynchronous meteorological satellites, the Himawari-8 satellite was launched on 7 October 2014 and is equipped with Advanced Himawari Imagers (AHIs). The AHIs have a total of 16 observation bands: 10 infrared bands, 3 near-infrared, and 3 visible. The time interval between the full-disk observations is 10 min. The detailed information of Himawari-8 satellite is listed in Table II. Bands 3, 4, 6, 7, 14, and 15 were selected. Band 3 is a visible band with a central wavelength of 0.64 μm, band 4 is a near-infrared band with a central wavelength of 0.86 μm, and band 15 has a central wavelength of 12.4 μm. By using combination of these three bands the influence of clouds can be effectively mitigated [5], [49]. In addition, considering that fire spots exhibit abnormal warning characteristics in the mid-infrared and far-infrared bands [15], [19], band 7 with a center wavelength of 3.9 μm and band 14 with a center wavelength of 11.2 μm were also selected. Moreover, band 6 with a wavelength of approximately 2.3 μm was employed to reduce the influence of water [14], [49].

The fire location data (label) is provided by the Meteorological Satellite Ground Station, Guangzhou, Guangdong, China. The fire label was initially generated using the adaptive threshold method [52]. First, real-time histogram statistics are used to obtain thresholds for excluding some nonfire spots. Next, potential fire spots are further filtered through the relative increment of background information, while cloud and water masks are used to exclude the influence of clouds and water. Finally, a field survey was conducted to correct any missed points.

III. Methodology

A. Established Dataset

Our dataset included data from October 1, 2018 to June 30, 2021, two times a day: 3:00 (Coordinated Universal Time, UTC)

TABLE I
DETAILED INFORMATION OF THE STUDY AREA

Province	Latitude/Longitude	Climate	Average temperature	Forest coverage
Fujian	23°30’N-28°20’N 115°40E’-120°30E’	Subtropical monsoon climate	18℃ – 26℃	66%
Guangdong	20°13’N-25°31’N 109°39E’-117°19E’	East Asian monsoon	20℃ – 28℃	53%
Guangxi	20°54'N-26°24'N 104°26'E-112°04'E	Subtropical monsoon climate	20℃ – 27℃	60%
Hunan	24°38’N-30°08’N 108°47’E-114°15’E	Subtropical monsoon climate	16℃ – 23℃	50%
Jiangxi	24°29’N-30°04’N 113°34’E-118°28’E	Subtropical monsoon humid climate	16℃ – 23℃	61%
Hainan	18°10’N-20°10’N 108°37’E-111°05’E	Tropical monsoon climate	23℃ – 29℃	57%
Yunnan	21°8′N-29°15′N 97°31′E-106°11′E	Subtropical monsoon humid climate	12℃ – 22℃	55%
Guizhou	24°37′N-29°13′N 103°36′E-109°35′E	Subtropical monsoon humid climate	12℃ – 19℃	62%

TABLE II
BAND INFORMATION OF HIMAWARI-8

Band #	**Central wavelength (μm)**	**Resolution (m)**	**Band #**	**Central wavelength (μm)**	**Resolution (m)**
1	0.47	1000	**9**	6.9	2000
2	0.51	1000	**10**	7.3	2000
3	0.64	500	**11**	8.6	2000
4	0.86	1000	**12**	9.6	2000
5	1.6	2000	**13**	10.4	2000
6	2.3	2000	**14**	11.2	2000
7	3.9	2000	**15**	12.4	2000
8	6.2	2000	**16**	13.3	2000

and 7:00 (UTC). Data from January 1, 2021 to June 30, 2021 were used to generate the test dataset, whereas the remaining data were used to generate the training dataset. Therefore, there was no overlap between the training and testing datasets.

Taking into account the main features of fires, the input data should include as much information about the band, spatial, and time of all pixels as possible. Our input data primarily included the following three parts for a pixel: context patch, band data, and band data in continuous time-series. As shown in Fig. 2, the context patch of a pixel referred to a window of size M × N centered on the pixel and the environmental information of the pixel was contained in the context patch. The context patch contained the spatial information of the central pixel. In this study, M and N were set to 21. The band data of a pixel refers to the specific value of each band on the pixel. Band data in continuous time-series refer to the sequence data composed of the band values of the pixel within a certain continuous period of time.

The quantity of fire spots in an image was very small in an actual scene; for example, we estimated that the average quantity of fire spots in a 400 × 300 image was 0.3, whereas the average quantity of nonfire spots was 119999.7, pointing to a large imbalance between the quantity of fire spots and nonfire spots. If all pixels were added to the training dataset, the network would easily fall into a local optimal solution. In our preliminary experiments, this caused the network to identify all pixels as nonfire spots.

If all pixels were added to the training dataset, the network easily would fall into a local optimal solution. In our preliminary experiments, this caused the network to identify all pixels as nonfire spots. When the quantity of nonfire spots was significantly larger than that of fire spots, the cost of identifying pixels as nonfire spots was very small. Therefore, in this study, this imbalance was addressed. One common method used to that effect is undersampling. For example, the quantity of nonfire spots in the training dataset was systematically reduced until that of both fire spots and nonfire spots was equivalent to one another. However, because of the small quantity of fire spots, this method greatly reduced the quantity of nonfire spots, causing the network not to be able to fully learn the features of nonfire spots, thus increasing the high false detection rate. Therefore, before the quantity of nonfire spots was decreased, the quantity of fire spots was appropriately increased to ensure that the fire spot data were augmented. However, to the best of the authors' knowledge, there is no widely used data augmentation method for fire spots, based on geosynchronous satellites. Therefore, this study proposed a data augmentation method that can increase

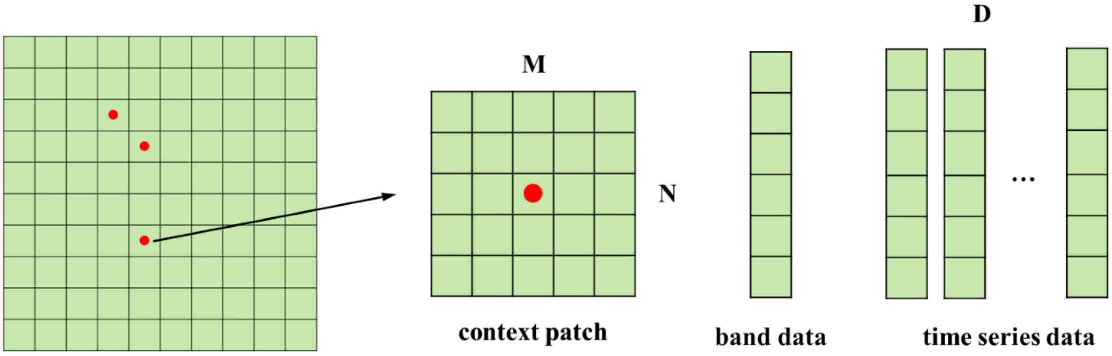


Fig. 2. Input data included context patch, band data, and time-series band data.

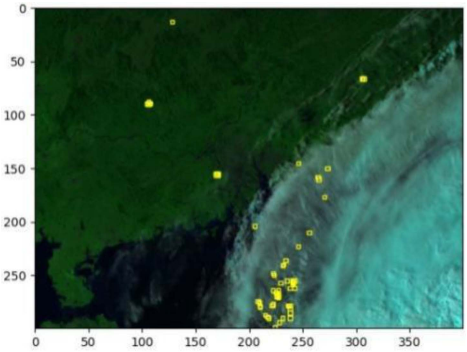


Fig. 3. Initial fire detection results. The yellow box refers to the fire spots extracted from the model, and the background is a false color image.

the quantity of fire spots while maintaining the rationality and diversity of environmental information. Thus, we used cloud, water, and underlying surface masks to copy fire spots to noncloud and nonwater forest/grassland pixels to increase the quantity of fire spots while ensuring the rationality of new fire spots. At the same time, we also keep the underlying surface information of the fire spots unchanged, that is, the fire spots occurring in forest are copied to forest, while fire spots in grassland landforms are transferred to the respective grassland areas. In this way, fire spots were copied to new environments, they could also enrich the spatial information in the training dataset. Using this new data augmentation method, the quantity of fire spots was increased to 20 times that of the original, and a random selection method was then used to extract an equal quantity of nonfire spots to form a preliminary training dataset. The ratio of fire spots to nonfire spots in the preliminary training dataset was 1:1. In early experiments, we used the preliminary training dataset and obtained not ideal experimental results. As shown in Fig. 3, the model identified a large number of cloud edge spots as fire spots. The main reason for this was that, when selecting nonfire spots, to ensure their relative fair representativeness, we adopted a random selection method which in turn led to less training data that were easily confused with fire spots, such as cloud edges and thin clouds. The network could not fully learn the features of nonfire spots/cloud edges/thin clouds and failed to accurately distinguish fire spots from nonfire spots/cloud edges/thin clouds.

To resolve this issue, this study first used the preliminary training dataset for the incomplete model to predict all images (excluding those in the testing dataset), extracted the misclassified pixels in each image, added them to the preliminary training dataset, and set a number limit. Each image can have at most 100 misclassified pixels that could be added to the preliminary training dataset. The upper limit of the number was so set as to prevent too many nonfire spots from being predicted, the local optimal solution. This method was named "rumination." Finally, the training dataset obtained included 1 54 245 fire spots and 8 34 524 nonfire spots.

B. SBT-FireNet Architecture

Taking inspiration from the traditional threshold method, fire spots exhibit the following three main characteristics: 1) Abnormal warming in their mid-infrared and far-infraredbands, evidenced by an increase in radiation value, which means band information is critically important for extracting the fire spots; 2) Distinguishing high temperature and brightness compared to the surrounding environment, which can be described as the spatial feature of the fire spots; and 3) The development of fire exhibits temporal characteristic, that is, at the beginning of the fire, it is relatively weak, then gradually become stronger. Finally, it extinguishes. This imply that the temporal information is extremely important. Therefore, this article designs a novel model namely SBT-FireNet for monitoring fire spots. The flow chart of the proposed SBT-FireNet is shown in Fig. 4. SBT-FireNet is a customized three branch network designed according to the fire features. We transformed the fire detection problem into a classification problem, that is, each pixel in the image was divided into a fire spot and a nonfire spot. The proposed deep learning network mainly includes the following two parts: a fire feature extraction part containing three components and a fully connected layer part to output classification scores.

The fire feature extraction part primarily consists of three modules: the spatial feature extraction module (SFE), the band feature extraction module (BFE), and the time-series feature extraction module (TFE). These modules are designed to extract spatial, band, and temporal features of fire spots, respectively.

The main body of SFE was a ViT structure. We leveraged the powerful global learning ability of ViT to adaptively extract and learn the features and their associations contained in the context patch. The main body of BFE was of a multilayer

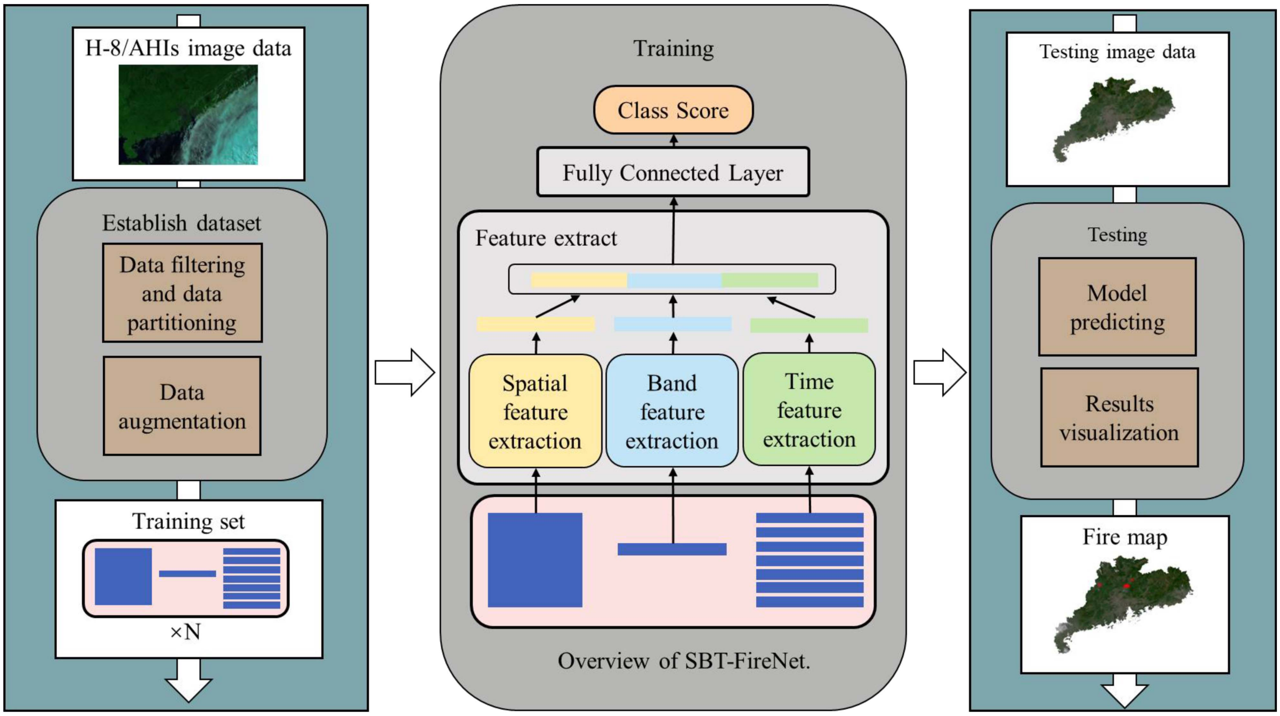


Fig. 4. Flowchart of our model, SBT-FireNet, and the overview of SBT-FireNet, N is the number of sample points.

perceptron (MLP) structure. TFE included an MLP layer and a Bi-long short-term memory (Bi-LSTM) structure. We map the features of each day to a higher dimension and used the Bi-LSTM structure to fully learn the features embedded in the time-series data.

The classification part of the fully connected layer was composed of a fully connected layer with 678 nodes to comprehensively consider the features extracted by the feature extraction process. Pixels with high fire classification scores were classified as fire spots; while those with low fire classification scores are classified as nonfire spots. Based on the combination of these strategies, the proposed method can overcome the issue of deep learning applications to geosynchronous satellites and fully consider the features of fires.

1) Structure of SFE Module: We use the SFE module to extract features from the context patch of the sample point. This context patch consists of the neighboring pixels of the sample point (center pixel) and, to some extent, represents the environmental information related to the sample point. The core component of the SFE module is the ViT structure. Given the small size of the context patch, using 2D convolution or 2D pooling operations for processing can lead to the loss of spatial information. However, ViT does not rely on 2D convolution and 2D pooling operations, making it well-suited for this task. Moreover, compared to CNNs, ViT exhibits superior global feature extraction capability, which is why the ViT structure is chosen as the primary component of the SFE module.

The input of SFE was a context patch $X \in \mathbb{R}^{M\times N\times C}$. The function of SFE was to learn the spatial features of a spot from a context patch. The main component of SFE was ViT. ViT is a successful application of transformer architectures in the field of computer vision [50]. ViT directly applied the transformer to image classification tasks and achieved accurate and precise results [51]. The structure of ViT is shown in Fig. 5.

In the field of natural language processing, transformer is first proposed to complete text translation tasks. Transformers are typically used to process 1D data. To apply transformer to 2D image data, ViT reshaped the image $X \in \mathbb{R}^{H\times W\times C}$ into a series of 2D patches $X_p \in \mathbb{R}^{n\times(p^2\cdot C)}$, which are flattened into 1D, and (H, W) is the size of the 2D image; C is the number of bands of the 2D image; and (p, p) is the size of each image patch. Therefore, the input series length for the transformer was $n = HW/p^2$. To represent the state of the 2D image, a learnable embedding called a class token was applied to the series of flattened patches. In addition, to preserve the position information, a 1D position embeddings is added to the flattened patch. While it is true that some local information may be lost during this process, it significantly enhances the module's capacity to extract global features. In addition, the inclusion of position embeddings helps to partially compensate for the loss of positional information resulting from the flattening process. Given the low spatial resolution of the utilized data (2 km) and the imprecise semantic information provided by the data, we consider the extraction of global features to be of greater importance in this particular dataset.

The transformer encoder is primarily composed of multiheaded self-attention [50], layer-normal, and MLP The class token, the output of the transformer encoder, was inputted into the MLP head. Finally, the MLP head outputted a score for the 2D image in each class. In this task, we set the MLP head to output a 294-dimensional feature vector as the output of the entire SFE module, which is used to represent the spatial features learned by SFE. To make ViT competent for this task, we set p to 7: such an input was split into nine patches. After flattening the

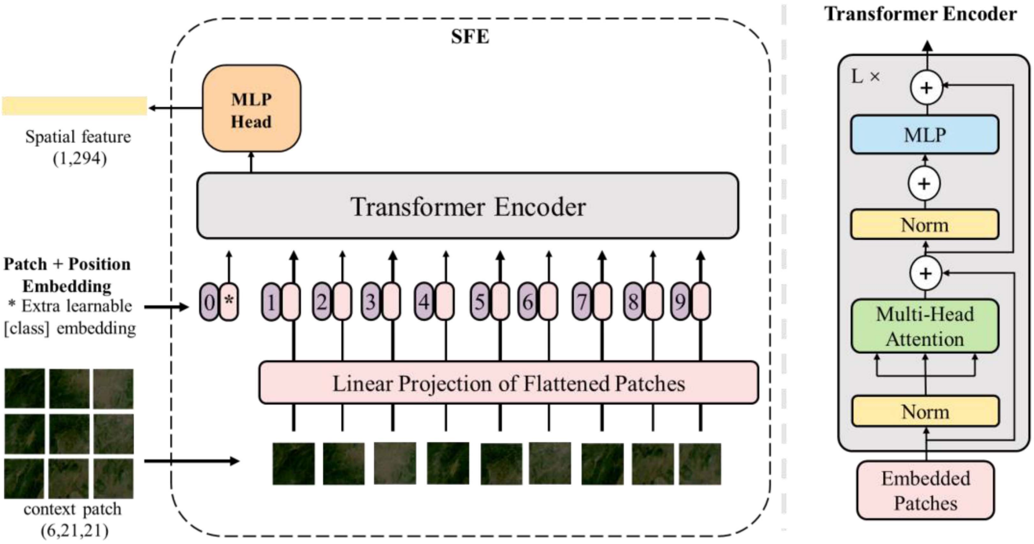


Fig. 5. Structure of SFE, L is the number of times the module is repeatedly stacked.

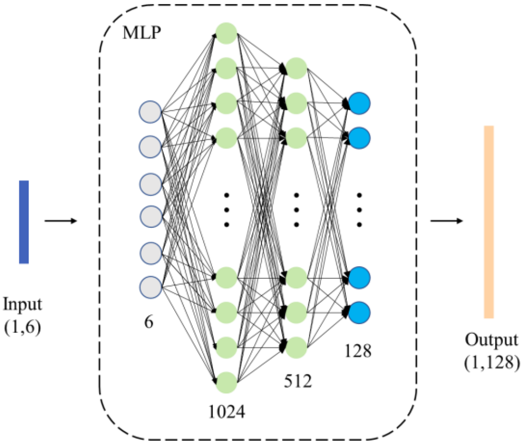


Fig. 6. Structure of BFE.

nine patches and adding a learnable feature token that was the same as the class token, we inputted them into the transformer encoder. The feature token output of the transformer encoder was used as the output of SFE. The feature token was an extracted spatial feature.

2) Structure of BFE Module: The structure of BFE is illustrated in Fig. 6. The input of BFE was a column vector with shape (1, 6) that included the values of six band of the pixel. In BFE, we used MLP to map the input 6D band features to higher dimensions to learn the features in the band. Specifically, MLP consisted of one input layer for receiving input information, two hidden layers for learning hidden features in input information, and an output layer for outputting learned features. The number of neuron nodes in the two hidden layers was 1024 and 512, respectively, and the number of neuron nodes in the output layer was 128. The activation function used in this model is the Gaussian error linear unit (GELU) function. GELU is an enhancement of the linear rectification function (ReLU), which outputs 0 directly for negative values, and can effectively address the issue of gradient disappearance associated with ReLU.

3) Structure of TFE Module: The structure of TFE is illustrated in Fig. 7. TFE consisted of a MLP and a Bi-LSTM. The input of TFE was time-series data of the pixel. Its shape was (D, 6), where D represents the time-series length. Due to TFE is used to process continuous time information, we input band information from the same time point for consecutive D days into TFE, where D is 7 and the number of input bands is 6. Therefore, the actual input shape is (7, 6). It is worth noting that these data (including SFE and BFE inputs) were not averaged or normalized before input. In TFE, we first input the D-day band data into the MLP to learn the band features of the spot in each day, and then concatenate the D-day band features into the Bi-LSTM to learn the time feature of the spot in D days. The structure of MLP is exactly the same as that used in BFE. In this study, D was set to 7.

Bi-LSTM extracted features from time-series data in two ways. Once the data information of the previous time spot was learned, the LSTM structure transferred the learned features to the next LSTM structure, which could fully learn the change features contained in the time-series data. These change features can reflect the change process of the band information of the sample in continuous time-series, thus providing the basis for distinguishing whether or not a pixel was a fire spot.

LSTM was initially used to resolve the issues of gradient disappearance and gradient explosion during sequence data training. The LSTM can transmit the cell state c, and the cell state can continuously accept new information and transmit these new and old information to the next node. Each LSTM node will forget the input of the node (including x^t and h^t), and update more important information to the cell state. At the same time, the hidden state (h^t) of the node is transmitted to the next node by using the input and cell state of the node. LSTM involved the following three steps.

1) Accept the c^{t-1} passed by the previous node, and accept the h^{t-1} output by the previous node and the x^t input by the node. Through a sigmoid unit (σ) to forget the information in h^{t-1} and x^t, and update the unforgettable information

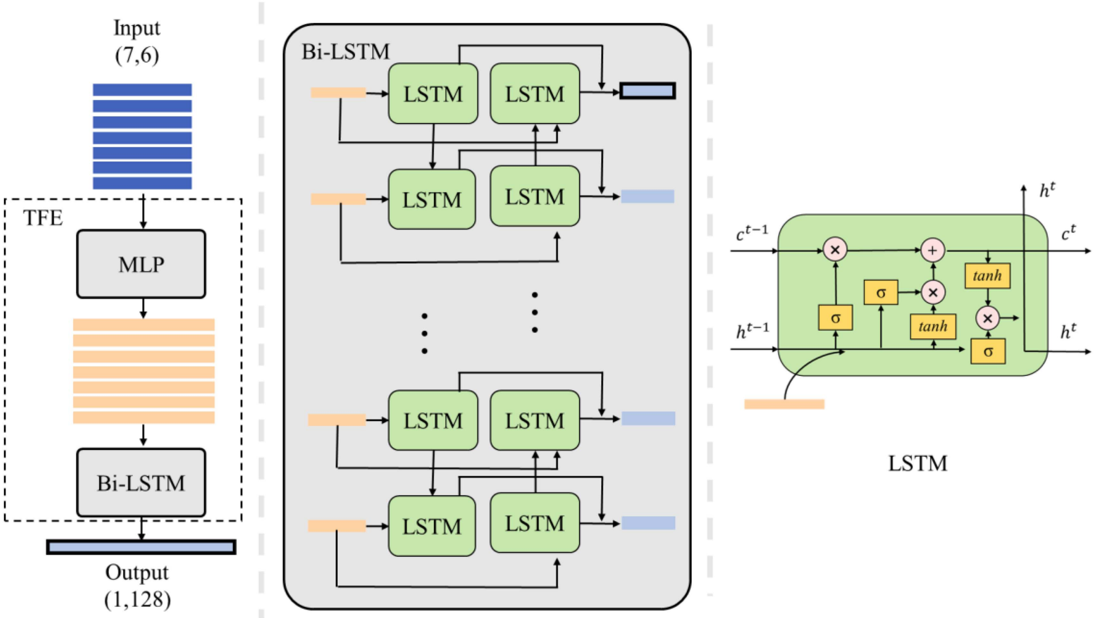


Fig. 7. Structure of TFE.

to c^{t-1}, through the process of "forgetting" unimportant information.

2) The information in h^{t-1} and x^t is learned through a sigmoid unit and tanh unit, and the learned information is updated to c^{t-1}, and the updated c^t is transmitted to the next node.
3) Accept c^t, h^{t-1} and x^t. The c^t is passed through a tanh unit to determine which information in the c^t needs to be output. Then h^{t-1} and x^t are multiplied by the sigmiod unit and the ct passing through the tanh unit to output h^t and pass h^t to the next node.

From the above-mentioned steps, LTSM could learn the input of the current node while retaining the important features learned by the previous node. For LTSM to learn the information contained in the above, Bi-LSTM underwent not only the process of passing from the previous node to the next node but also that of passing from the next node to the previous node to ensure that it could learn the information in the given context more completely. Therefore, the Bi-LSTM was selected to determine the time characteristics of the fire spots.

The input size of Bi-LSTM is 128, and the hidden size of Bi-LSTM is 128. We only used one layer of Bi-LSTM.

IV. Experimental Results

A. Experimental Environment

All the code of this study is written in version 3.6 of Python. In the part of building deep learning model, the learning library used is version 1.2 of PyTorch. All experiments in this article were conducted on an Intel CoreI i9-10900K CPU @ 3.70 GHz, 128 GB of RAM, with an NVIDIA GeForce GTX 3080. In the training process of each deep learning model, we use the Adam optimizer as the parameter optimizer, the loss function as the cross-entropy loss function, epochs set to 500, batch size set to 100, and learning rate set to 10^{-6}. With the exception of the "no-aug" method, all deep learning methods employ the same training set and test set. The test set and training set do not overlap; they are entirely separate from each other. In addition, the accuracy (A), precision (P), recall (R), F1-score (F1), missed detection rate (MD), and error detection rate (ED) were used as the key performance indicators of the model as follows:

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \tag{1}$$

$$\text{Precision} = TP/(TP + FP) \tag{2}$$

$$\text{Recall} = TP/(TP + FN) \tag{3}$$

$$F1 = (2 \cdot P \cdot F)/(P + F)) \tag{4}$$

$$MD = 1 - R \tag{5}$$

$$ED = 1 - P \tag{6}$$

where TP is a true positive; TN is a true negative; FP is a false positive; and FN is a false negative.

B. Comparison Methods and Implementation Details

In order to evaluate the effectiveness of the proposed model, five state-of-the-art methods are selected for comparison purpose, which are ResNet50 (Res50), ResNet50-low (Res50-low), ViT, FasterNet, and MPViT.

ResNet50 is a classic CNN classification model that uses a multilayer convolutional pooling structure to gradually learn the semantic information of the image and uses a residual structure to prevent gradient disappearance or gradient explosion.

The structure of ResNet50-low is the same as ResNet. The only difference is that ResNet50-low will also classify pixels with low fire spot scores as fire spots. Specifically, all pixels with high fire spot scores above 0.95, except for ResNet-low,

TABLE III
COMPARISON OF EXPERIMENTAL RESULTS

	A	P	R	F1	ED	MD	Time(s)
SBT-FireNet	0.999	0.747	0.761	0.754	0.253	0.239	12.505
ViT	0.999	0.395	0.096	0.154	0.605	0.904	11.698
Res50	0.999	–	0.000	–	–	1.000	11.803
Res50-low	0.878	0.001	0.952	0.002	0.999	0.048	11.612
FasterNet	0.991	0.007	0.958	0.014	0.993	0.009	17.419
MPViT	0.830	0.001	0.986	0.002	0.999	0.170	23.529
No-aug	0.999	0.643	0.551	0.593	0.357	0.449	13.128

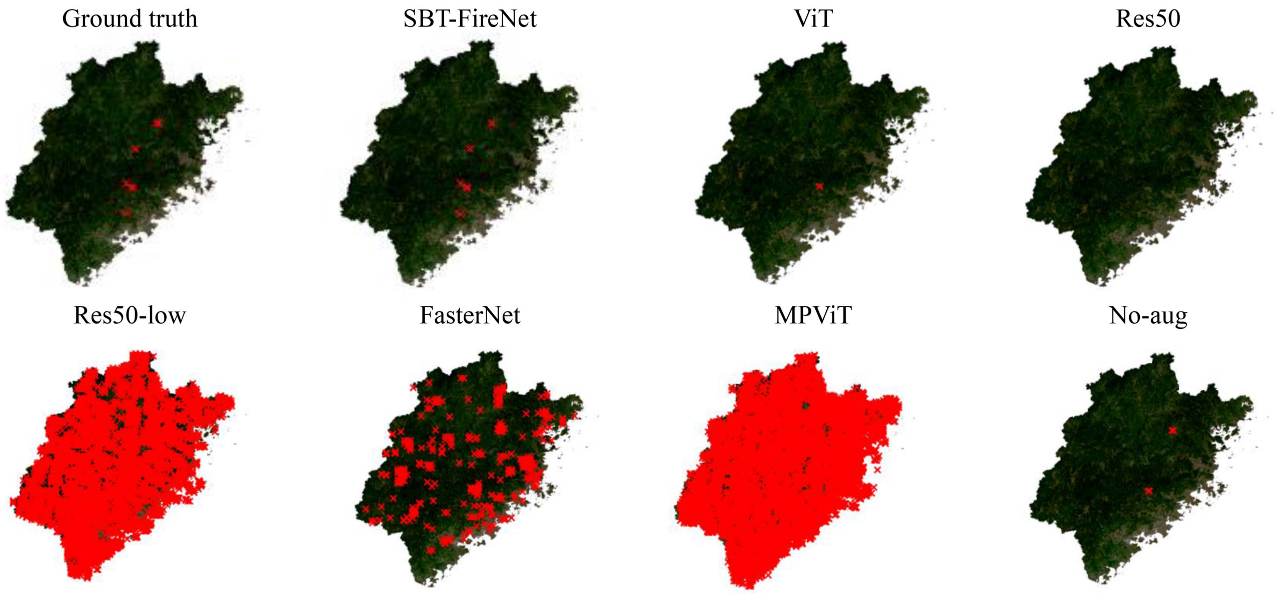


Fig. 8. Fujian fire map and ground truth generated by each method, where red "×" is the fire spot.

will be classified as fire spots. For ResNet-low, pixels with fire spot scores surpassing 0.8 are classified as fire spots.

ViT is the successful application of transformers in image classification, which transforms image data into blocks and then compresses them. It uses a multilayer attention mechanism to iteratively learn the features of each block and finally outputs the classification of the image by head.

FasterNet and MPViT are relatively new deep learning networks, which were proposed in 2023 and 2022, respectively. FasterNet incorporates partial convolutions to reduce redundant computations and memory access. This approach not only accelerates network training speed but also enhances performance.

MPViT, on the other hand, leverages multiscale path embedding and multipath structures to enhance the representation of both fine and coarse features. In addition, it utilizes global-to-local feature interaction to improve the model's ability to process global context effectively.

It should be noted that all the comparison methods are conducted in the same environment, and data augmentation is also employed except "no-aug."

C. Results Analysis

As indicated in Table III, SBT-FireNet had the best comprehensive performance. Res50-low had an extremely low precision and identified a large number of pixels as fire spots. The MD value of Res50 up to 1.000 and the ED value of Res50-low up to 0.999 indicated that the two models were completely unqualified for the task of fire spot detection. SBT-FireNet achieved a precision of 0.747, recall rate of 0.761, and F1-score of 0.754. Compared to those of ViT, the precision and recall rate of SBT-FireNet improved by 0.352 and 0.665, respectively. SBT-FireNet achieved the highest F1-score, indicating that SBT-FireNet was more suitable for the fire detection. This result was expected because ViT and Res50 were not designed for the fire detection mission based on geostationary satellites and failed to comprehensively utilize of all information aspects of fire. Both FasterNet and MPViT exhibit a high number of false positives in fire detection.

A fire map for the different methods is presented in Figs. 8–11. ViT missed most of the fire spots. Res50-low

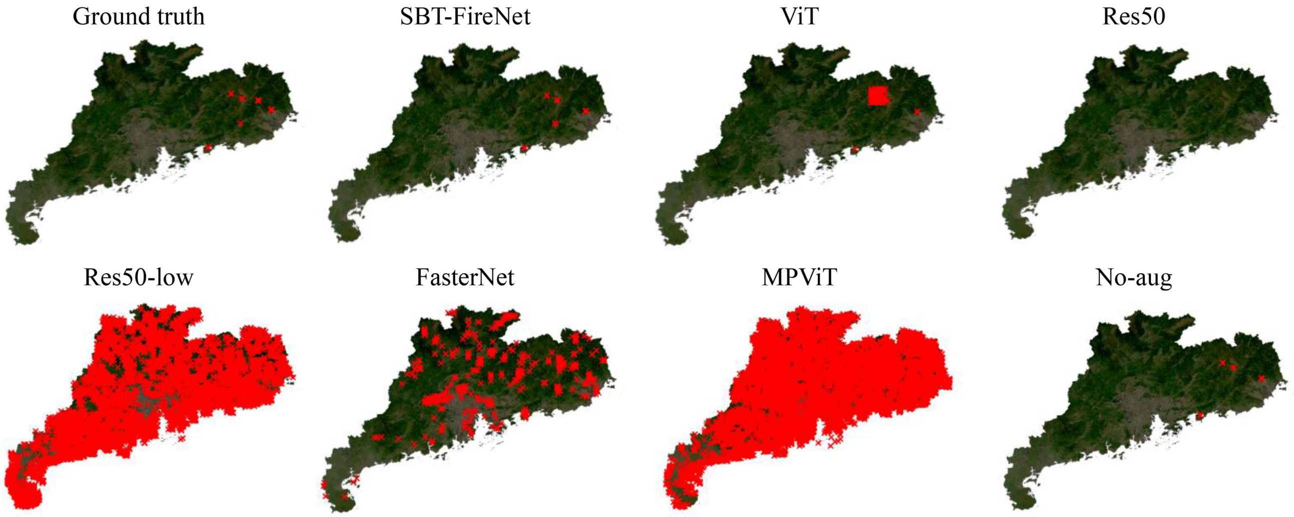


Fig. 9. Guangdong fire map and ground truth generated by each method, where red “×” is the fire spot.

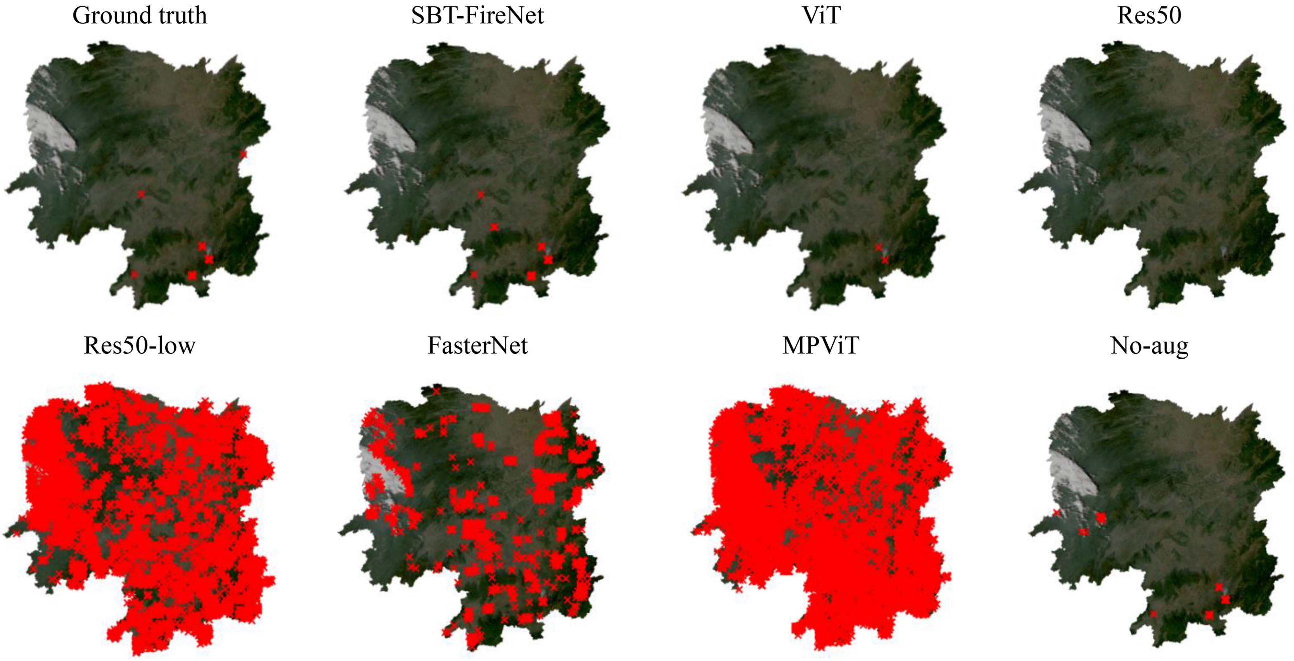


Fig. 10. Hunan fire map and ground truth generated by each method, where red “×” is the fire spot.

detected almost all the pixels as fire spots, whereas Res50 identified all the pixels as nonfire spots. ViT, Res50, and Res50-low were incapable of detecting fire spots. FastNet and MPViT both have a large number of false positives, and as shown in Fig. 10, MPViT has poor robustness to the cloud, while FastNet has poor robustness to cloud edges. However, the fire map generated by SBT-FireNet was consistent with the ground truth.

To evaluate the effectiveness of our proposed data augmentation methods, we conducted performance tests on the SBT-FireNet model using a “no data augmentation” strategy (referred to as “no-aug”). As shown in Table III, the precision, recall rate, and F1 value are significant lower compared to that of SBT-FireNet, which decreased 0.104, 0.21, and 0.161 lower, respectively. “No-aug” method exhibits a significant number of false positives and omissions. As depicted in Fig. 10, “no-aug” shows poor fire detection performance at cloud edges and in areas with thin clouds, leading to the occurrence of false positives. These results indicate that utilizing the enhanced dataset can significantly imporve the accuracy and recall rate of the model.

Table III also includes the average time consumption of each method for detecting fire spots in an image. As can be seen from Table III, the time cost of FasterNet and MPViT is 23.529 s and 17.541 s, which is significantly higher than the other methods. ResNet50, ResNet50-low, ViT, and the proposed SBT-FireNet generate comparable time efficiency, which can provide nearly real-time monitoring of fire spots.

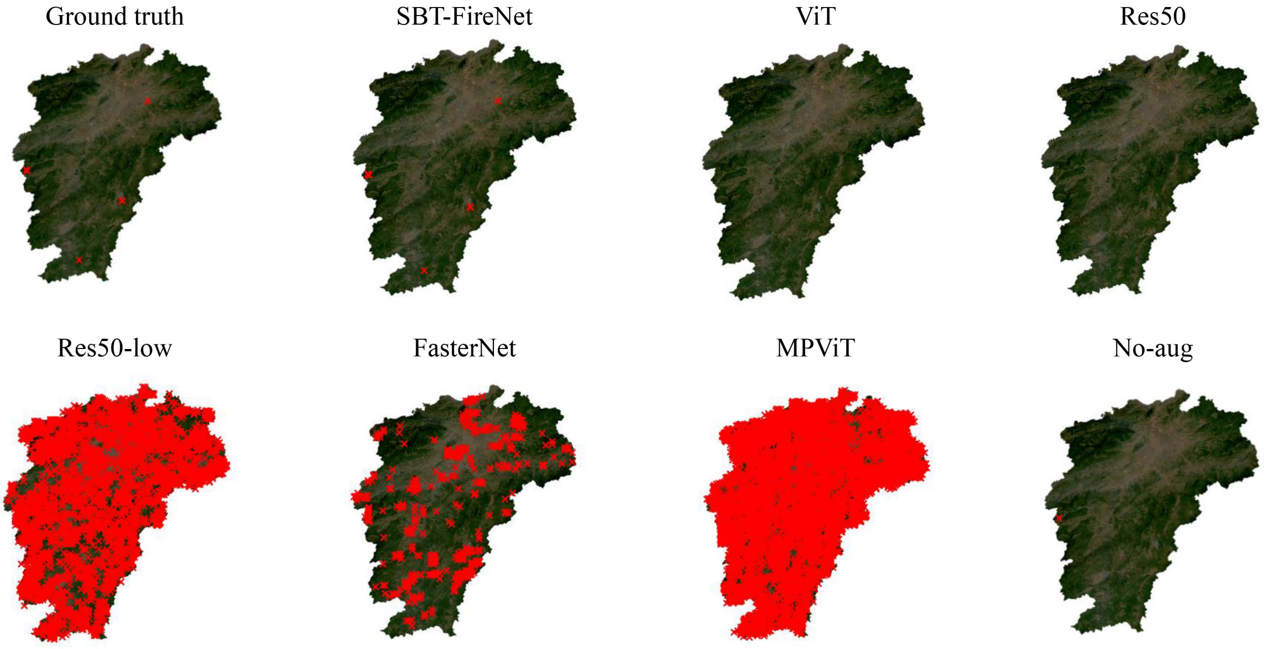


Fig. 11. Jiangxi fire map and ground truth generated by each method, where red “×” is the fire spot.

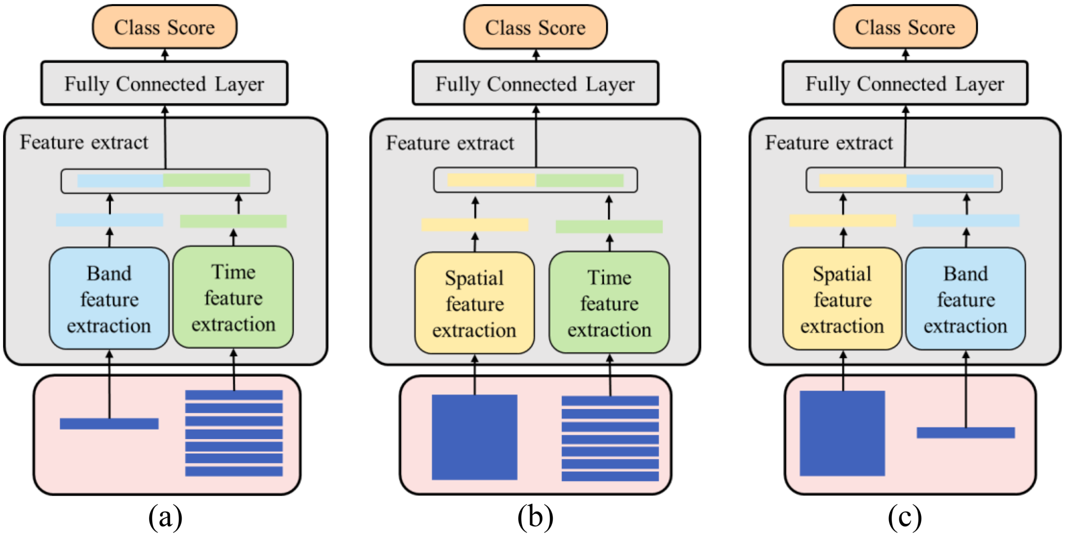


Fig. 12. Structures of (a) BT, (b) ST, and (c) SB.

V. Discussion

SBT-FireNet consisted of the three feature extraction modules designed based on the fact that fire spots exhibit significant differences from the surrounding pixels, such as extremely high temperatures. In addition, the intensity and severity of fires change spatiotemporally. Therefore, continuous time-series data play a vital role in fire monitoring. Based on the above-mentioned fire characteristics, we proposed a fire-customized deep learning model. To verify the effectiveness of the designed network, the function of each module was explored and discussed based on the inclusion or exclusion of each for the original SBT-FireNet model. We compared SBT-FireNet to SBT-FireNet without SFE (BT), SBT-FireNet without BFE (ST), and SBT-FireNet without TFE (SB). The structure is shown in Fig. 12. We used SFE to learn the context patch of the fire spot and output the spatial features of the fire spot. If there was no SFE module, the model would lose the perception of the spatial features. For example, if the pixel was located in the cloud area, the edge of the cloud, or the thin cloud, its band and time features would be similar to those of the fire pixel. As there was no spatial feature to distinguish, the model could easily classify such pixels as fire spots, reducing the model accuracy. From Table IV, it can be seen that BT (without SFE) had lower precision than SBT-FireNet for Fujian, Guangdong, Hunan, and Jiangxi (0.074, 0.063, 0.120, and 0.101, respectively). Low precision led to a low F1-score of SB. In SFE, we used ViT as the main body as it had a strong global feature learning ability, did not involve pooling

TABLE IV
Testing Results of Ablation Experiment

	Models	A	P	R	F1	ED	MD
Fujian	SBT	0.999	0.797	0.829	0.813	0.203	0.171
	BT	0.999	0.723	0.844	0.778	0.277	0.156
	ST	0.999	0.751	0.713	0.732	0.249	0.268
	SB	0.999	0.838	0.793	0.815	0.162	0.207
Guangdong	SBT	0.999	0.705	0.706	0.705	0.295	0.294
	BT	0.999	0.642	0.739	0.687	0.358	0.261
	ST	0.999	0.668	0.706	0.686	0.332	0.294
	SB	0.999	0.745	0.641	0.689	0.255	0.359
Hunan	SBT	0.999	0.743	0.778	0.760	0.257	0.222
	BT	0.999	0.623	0.810	0.704	0.377	0.190
	ST	0.999	0.671	0.680	0.675	0.329	0.320
	SB	0.999	0.715	0.739	0.727	0.285	0.261
Jiangxi	SBT	0.999	0.774	0.828	0.800	0.226	0.172
	BT	0.999	0.673	0.871	0.759	0.327	0.129
	ST	0.999	0.559	0.796	0.657	0.441	0.204
	SB	0.999	0.750	0.799	0.774	0.250	0.201

TABLE V
Experimental Results of Fireless Areas

Area	Images	TP	TN	FP	FN	Sum
Fujian	103	0	7789370	5	0	7789375
Guangdong	85	0	10199984	16	0	102000000
Hunan	117	0	10529986	14	0	10530000
Jiangxi	107	0	8827424	76	0	8827500

operations, and did not cause the loss of detailed information because of pooling in the learning process. These characteristics rendered ViT suitable for SFE.

We used an MLP as the main body of BFE as it could map band information to a higher dimension and learn the abstract features embedded in it through several hidden layers. The final output band features characterized the characteristics of the fire spot in the band. ST (without BFE) was not able to distinguish between the fire and nonfire spots when pixels were compared with the environment or past data. As can be seen from Table IV, the precision and recall of ST were not higher than those of SBT-FireNet.

In TFE, we first used an MLP to map each day data to a higher dimension and then used Bi-LSTM for the time feature extraction. The fire had different brightness temperatures and radiation values before and after the combustion. By observing this change process, the fire spot (abnormally high temperature or high bright spot) could be identified; however, the time feature also received an image of the cloud. The change characteristics before and after the cloud coverage were similar to those before and after the fire. If TFE was omitted, the time feature extraction could not be performed, and the model accuracy might be accordingly improved, whereas the recall rate would be reduced. According to Table IV, SB (without TFE) achieved higher precision in the Fujian and Guangdong regions; however, its recall F1-score was lower than those of SBT-FireNet. SBT-FireNet obtained the highest F1-score in the training area and obtained a F1-score of 0.813 in the remaining testing area, slightly lower from the best F1 score (0.815) in this area. In the absence of SFE, BT exhibited the worst P score and the high false detection rate in the three test areas. In the absence of BFE, the ST showed the lowest F1-score, with poor comprehensive ability in all the four regions. In the absence of TFE, SB yielded the best F1-score in the Fujian test area and had a good F1-score in the other test areas. Overall, SB was second only to SBT-FireNet, resulted in good precision and recall, and even achieved the best F1-score in the Fujian region. ST achieved the lowest F1-score in all the four regions. Given the above-mentioned results and the performance of ST, testing whether or not to detect fire spots by comparing the band information of the central pixel with the environmental information or by comparing its band information with the time information proved to be an effective strategy.

Most geostationary satellite images show no fires or few fires. We also tested the classification performance of SBT-FireNet in the absence of fire spots. As the data used were images without any fire spots, the recall rate and accuracy were invalid. It is important to note that in the previous test and training phases, the data utilized included both fire spots and nonfire spots. However, in this experiment, the data used exclusively consisted of images without any fire spots. We computed the specific values of TP, TN, FP, and FN in the four test areas as well as the number of testing images, as shown in Table V.

According to Table V, our network still exhibited accurate and robust classification performance in the absence of fire spots.

The limitation of this study is that the best combination of various spectral bands for fire point detection task is not explored. This issue will be addressed in our future work.

VI. Conclusion

In this study, a novel deep learning model named SBT-FireNet was proposed to detect fire spots from Himawari-8 satellites. The proposed model successfully resolved the issue that the existing sematic segmentation in the traditional deep learning models cannot accurately detect fire spots from geosynchronous satellite images because of the incomplete semantic information of the low spatial resolution images. The proposed model transformed the fire detection problem into a classification problem and includes the following two parts: 1) feature extraction and 2) fully connected layer classification. To comprehensively learn the features of fires, the three extraction modules of spatial, band, and time-series feature were designed and integrated in the model. A series of strategies were adopted to alleviate the sample imbalance issue between fire and nonfire spots.

We trained SBT-FireNet based on the constructed dataset and tested it in the four provinces in southern China with high forest coverage. The results demonstrated that the best precision was achieved by SBT-FireNet, which was 35% higher than that of the other models.

The main contribution of this study can be summarized as follows.

1) We built a new fire spot database, using numerous Himawari-8 satellite images. To alleviate the issue of category imbalance between fire and nonfire spots, geosynchronous satellite data and data augmentation and "rumination" strategies were adopted. A robust dataset was constructed, which supported the full and high-quality learning of fire features.
2) This study customized a novel deep learning model to detect fire spots from Himawari-8 geosynchronous satellite images. Specifically, SFE, BFE, and TFE modules were designed to strengthen the feature representation of fires. Finally, a powerful deep learning model, SBT-FireNet, was obtained.
3) The proposed method fully exploited the advantages of the high temporal resolution of geosynchronous satellites. Also, SBT-FireNet was automatized for the near-real-time monitoring of fire spots.

The limitation of the proposed model is that it was only validated using Himawari-8 satellite images. In the future, we will further explore the generalizability of this model by applying it to other satellite images.

References

[1] S. Hantson, M. Padilla, D. Corti, and E. Chuvieco, "Strengths and weaknesses of MODIS hotspots to characterize global fire occurrence," *Remote Sens. Environ.*, vol. 131, pp. 152–159, 2013, doi: 10.1016/j.rse.2012.12.004.

[2] S. O. Sunderman and P. J. Weisberg, "Remote sensing approaches for reconstructing fire perimeters and burn severity mosaics in desert spring ecosystems," *Remote Sens. Environ.*, vol. 115, no. 9, pp. 2384–2389, 2011, doi: 10.1016/j.scitotenv.2016.03.129.

[3] J. Parente, M. G. Pereira, and M. Tonini, "Space-time clustering analysis of wildfires: The influence of dataset characteristics, fire prevention policy decisions, weather and climate," *Sci. Total Environ.*, vol. 559, pp. 151–165, 2016, doi: 10.1016/j.jag.2019.102034.

[4] B. Sha, C. Ah, A. Sj, A. Sb, D. Asb, and A. Thn, "A satellite data driven approach to monitoring and reporting fire disturbance and recovery across boreal and temperate forests," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 87, 2020, Art. no. 102034, doi: 10.1016/j.jag.2019.102034.

[5] G. Louis, B. Luigi, D. P. Roy, M. L. Humber, and C. O. Justice, "The collection 6 MODIS burned area mapping algorithm and product," *Remote Sens. Environ.*, vol. 217, pp. 72–85, 2018, doi: 10.1016/j.rse.2018.08.005.

[6] C. Maffei and M. Menenti, "Predicting forest fires burned area and rate of spread from pre-fire multispectral satellite measurements," *ISPRS J. Photogrammetry Remote Sens.*, vol. 158, pp. 263–278, 2019, doi: 10.1016/j.isprsjprs.2019.10.013.

[7] V. Martins, D. Roy, H. Huang, L. Boschetti, H. Zhang, and L. Yan, "Deep learning high resolution burned area mapping by transfer learning from Landsat-8 to PlanetScope," *Remote Sens. Environ.*, vol. 280, 2022, Art. no. 113203, doi: 10.1016/j.rse.2022.113203.

[8] J. San-Miguel-Ayanz and N. Ravail, "Active fire detection for fire emergency management: Potential and limitations for the operational use of remote sensing," *Natural Hazards*, vol. 35, no. 3, pp. 361–376, 2005, doi: 10.1007/s11069-004-1797-2.

[9] X. Wang, T. Swystun, and M. D. Flannigan, "Future wildfire extent and frequency determined by the longest fire-conducive weather spell," *Sci. Total Environ.*, vol. 830, 2022, Art. no. 154752, doi: 10.1016/j.scitotenv.2022.154752.

[10] M. J. Wooster et al., "Satellite remote sensing of active fires: History and current status, applications and future requirements," *Remote Sens. Environ.*, vol. 267, 2021, Art. no. 112694, doi: 10.1016/j.rse.2021.112694.

[11] M. J. Wooster, W. Xu, and T. Nightingale, "Sentinel-3 SLSTR active fire detection and FRP product: Pre-launch algorithm development and performance evaluation using MODIS and ASTER datasets," *Remote Sens. Environ.*, vol. 120, pp. 236–254, 2012, doi: 10.1016/j.rse.2011.09.033.

[12] M. Liu, X. Wang, A. Zhou, X. Fu, Y. Ma, and C. Piao, "UAV-YOLO: Small object detection on unmanned aerial vehicle perspective," *Sensors*, vol. 20, no. 8, 2020, Art. no. 2238, doi: 10.1016/j.rse.2016.02.027.

[13] S. W. Murphy, C. R. de Souza Filho, R. Wright, G. Sabatino, and R. C. Pabon, "HOTMAP: Global hot target detection at moderate spatial resolution," *Remote Sens. Environ.*, vol. 177, pp. 78–88, 2016, doi: 10.1016/j.rse.2016.02.027.

[14] W. Schroeder, P. Oliva, L. Giglio, B. Quayle, E. Lorenz, and F. Morelli, "Active fire detection using Landsat-8/OLI data," *Remote Sens. Environ.*, vol. 185, pp. 210–220, 2016, doi: 10.1016/j.rse.2015.08.032.

[15] Z. Lin et al., "An active fire detection algorithm based on multi-temporal FengYun-3C VIRR data," *Remote Sens. Environ.*, vol. 211, pp. 376–387, 2018, doi: 10.1016/j.rse.2018.04.027.

[16] X. Hu, Y. Ban, and A. Nascetti, "Sentinel-2 MSI data for active fire detection in major fire-prone biomes: A multi-criteria approach," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 101, 2021, Art. no. 102347, doi: 10.1016/j.jag.2021.102347.

[17] W. Xu, M. J. Wooster, J. He, and T. Zhang, "First study of Sentinel-3 SLSTR active fire detection and FRP retrieval: Night-time algorithm enhancements and global intercomparison to MODIS and VIIRS AF products," *Remote Sens. Environ.*, vol. 248, 2020, Art. no. 111947.

[18] E. J. Fusco, J. T. Finn, J. T. Abatzoglou, J. K. Balch, S. Dadashi, and B. A. Bradley, "Detection rates and biases of fire observations from MODIS and agency reports in the conterminous United States," *Remote Sens. Environ.*, vol. 220, pp. 30–40, 2019, doi: 10.1016/j.rse.2020.111947.

[19] J. Chen et al., "The Fengyun-3D (FY-3D) global active fire product: Principle, methodology and validation," *Earth Syst. Sci. Data*, vol. 14, no. 8, pp. 3489–3508, 2022, doi: 10.5194/essd-14-3489-2022.

[20] W. Xu, M. Wooster, G. Roberts, and P. Freeborn, "New GOES imager algorithms for cloud and active fire detection and fire radiative power assessment across North, South and Central America," *Remote Sens. Environ.*, vol. 114, no. 9, pp. 1876–1895, 2010, doi: 10.1016/j.rse.2010.03.012.

[21] Y. J. Kaufman, A. Setzer, C. Justice, C. Tucker, M. Pereira, and I. Fung, "Remote sensing of biomass burning in the tropics," in *Fire in the Tropical Biota*, Berlin, Germany: Springer-Verlag, 1990, pp. 371–399, doi: 10.1007/978-3-642-75395-4_16.

[22] Q. Zhang et al., "Deep-learning-based burned area mapping using the synergy of Sentinel-1&2 data," *Remote Sens. Environ.*, vol. 264, 2021, Art. no. 112575, doi: 10.1016/j.rse.2021.112575.

[23] L. Collins, G. McCarthy, A. Mellor, G. Newell, and L. Smith, "Training data requirements for fire severity mapping using Landsat imagery and random forest," *Remote Sens. Environ.*, vol. 245, 2020, Art. no. 111839, doi: 10.1016/j.rse.2020.111839.

[24] D. V. V. Prasad et al., "Analysis and prediction of water quality using deep learning and auto deep learning techniques," *Sci. Total Environ.*, vol. 821, 2022, Art. no. 153311, doi: 10.1016/j.scitotenv.2022.153311.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778, doi: 10.1109/cvpr.2016.90.

[26] H. Zhang et al., "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2736–2746, doi: 10.1109/cvprw56347.2022.00309.

[27] I. C. Duta, L. Liu, F. Zhu, and L. Shao, "Improved residual networks for image and video recognition," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 9415–9422, doi: 10.1109/icpr48806.2021.9412193.

[28] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021, doi: 10.1109/tpami.2019.2938758.

[29] J. Chen et al., "Run, don't walk: Chasing higher FLOPS for faster neural networks," 2023, *arXiv:2303.03667*, doi: 48550/arXiv.2303.03667.

[30] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*, doi: 10.48550/arXiv.2010.11929.

[31] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022, doi: 10.1109/iccv48922.2021.00986.

[32] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, "MPViT: Multi-path vision transformer for dense prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7287–7296, doi: 10.1109/cvpr52688.2022.00714.

[33] M. Xie, "Development of artificial intelligence and effects on financial system," *J. Phys.: Conf. Ser.*, vol. 1187, no. 3, 2019, Art. no. 032084, doi: 10.1088/1742-6596/1187/3/032084.

[34] S. Mishra and A. K. Tyagi, "The role of machine learning techniques in Internet of Things-based cloud applications," in *Proc. Artif. Intell.-Based Internet Things Syst.*, 2022, pp. 105–135, doi: 10.1007/978-3-030-87059-1_4.

[35] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[36] Z. Langford, J. Kumar, and F. Hoffman, "Wildfire mapping in Interior Alaska using deep neural networks on imbalanced datasets," in *Proc. IEEE Int. Conf. Data Mining Workshops*, 2018, pp. 770–778, doi: 10.1109/ICDMW.2018.00116.

[37] 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, doi: 10.3390/rs11141702.

[38] G. H. de Almeida Pereira, A. M. Fusioka, B. T. Nassu, and R. Minetto, "Active fire detection in Landsat-8 imagery: A large-scale dataset and a deep-learning study," *ISPRS J. Photogrammetry Remote Sens.*, vol. 178, pp. 171–186, 2021, doi: 10.1016/j.isprsjprs.2021.06.002.

[39] S. T. Seydi, V. Saeidi, B. Kalantar, N. Ueda, and A. A. Halin, "Fire-Net: A deep learning framework for active forest fire detection," *J. Sensors*, vol. 2022, pp. 1–14, 2022, doi: 10.1155/2022/8044390.

[40] Z. Jiao et al., "A deep learning based forest fire detection approach using UAV and YOLOv3," in *Proc. 1st Int. Conf. Ind. Artif. Intell.*, 2019, pp. 1–5, doi: 10.1109/ICIAI.2019.8850815.

[41] Z. Jiao et al., "A YOLOv3-based learning strategy for real-time UAV-based forest fire detection," in *Proc. Chin. Control Decis. Conf.*, 2020, pp. 4963–4967, doi: 10.1109/CCDC49329.2020.9163816.

[42] Y. Zhao, J. Ma, X. Li, and J. Zhang, "Saliency detection and deep learning-based wildfire identification in UAV imagery," *Sensors*, vol. 18, no. 3, p. 712, 2018, doi: 10.3390/s18030712. [Online]. Available: https://www.mdpi.com/1424-8220/18/3/712

[43] L. Zhang, M. Wang, Y. Fu, and Y. Ding, "A forest fire recognition method using UAV images based on transfer learning," *Forests*, vol. 13, no. 7, p. 975, 2022, doi: 10.3390/f13070975. [Online]. Available: https://www.mdpi.com/1999-4907/13/7/975

[44] W. Xu, M. J. Wooster, T. Kaneko, J. He, T. Zhang, and D. Fisher, "Major advances in geostationary fire radiative power (FRP) retrieval over Asia and Australia stemming from use of Himarawi-8 AHI," *Remote Sens. Environ.*, vol. 193, pp. 138–149, 2017, doi: 10.1016/j.rse.2017.02.024.

[45] W. Xu, M. J. Wooster, J. He, and T. Zhang, "Improvements in high-temporal resolution active fire detection and FRP retrieval over the Americas using GOES-16 ABI with the geostationary fire thermal anomaly (FTA) algorithm," *Sci. Remote Sens.*, vol. 3, 2021, Art. no. 100016, doi: 10.1016/j.srs.2021.100016.

[46] M. Lu et al., "An improved cloud detection method for GF-4 imagery," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1525, doi: 10.3390/rs12091525.

[47] C. Ding, X. Zhang, J. Chen, S. Ma, Y. Lu, and W. Han, "Wildfire detection through deep learning based on Himawari-8 satellites platform," *Int. J. Remote Sens.*, vol. 43, no. 13, pp. 5040–5058, 2022, doi: 10.1080/01431161.2022.2119110.

[48] Z. Hong et al., "Active fire detection using a novel convolutional neural network based on Himawari-8 satellite images," *Front. Environ. Sci.*, vol. 10, 2022, Art. no. 794028.

[49] G. Xu and X. Zhong, "Real-time wildfire detection and tracking in Australia using geostationary satellite: Himawari-8," *Remote Sens. Lett.*, vol. 8, no. 11, pp. 1052–1061, 2017, doi: 10.1155/2022/8044390.

[50] A. Vaswani et al., "Attention is all you need," in *Proc. Advances Neural Inf. Process. Syst.*, 2017, vol. 30, doi: 10.48550/.arXiv.1706.03762. [Online]. Available: https://www.mdpi.com/about/announcements/784

[51] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10, pp. 1–41, 2022, doi: 10.1145/3505244.

[52] Q. He and C. Liu, "Improved algorithm of self-adaptive fire detection for MODIS data," *J. Remote Sens.*, vol. 3, pp. 448–453, 2008, doi: 10.3321/j.issn:1007-4619.2008.03.010.

[53] Q. Liu, Y. Shan, L. Shu, P. Sun, and S. Du, "Spatial and temporal distribution of forest fire frequency and forest area burnt in Jilin Province, Northeast China," *J. Forestry Res.*, vol. 29, no. 5, pp. 1233–1239, 2018, doi: 10.1007/s11676-018-0605-x.

Zhonghua Hong (Member, IEEE) received the Ph.D. degree in GIS from Tongji University, Shanghai, China, in 2014.

Since 2022, he has been a Professor with the College of Information Technology, Shanghai Ocean University, Shanghai, China. His research interests include satellite/aerial photogrammetry, high-speed videogrammetric, planetary mapping, 3D emergency mapping, GNSS-R, deep learning, and processing of geospatial Big Data.

Zhizhou Tang received the B.S. degree in educational technology from the School of Educational Science, Hunan Normal University, Changsha, China, in 2020. He is currently working toward the master's degree in computer technology with Shanghai Ocean University, Shanghai, China.

His research interests include the fire detection of geosynchronous satellites.

Haiyan Pan received the Ph.D. degree in surveying and mapping from Tongji University, Shanghai, China, in 2020.

She has been a Lecturer with the College of Information Technology, Shanghai Ocean University, Shanghai, China. Her research interests include multitemporal remote sensing data analysis, change detection, multispectral/hyperspectral image classification.

Yuewei Zhang received the M.S. degree in computer technology from the South China University of Technology, Guangzhou, China, in 2012.

He is currently a Senior Engineer with the Guangzhou Meteorological Satellite Ground Station, Guangzhou, China. His research interests include satellite data transmission, image composition, and forest fire detection.

Zongsheng Zheng received the Ph.D. degree in physical geography from East China Normal University, Shanghai, China, in 2008.

Since 2012, he has been an Associate Professor with Shanghai Ocean University, Shanghai, China. His research interests include ocean remote sensing, deep learning, and processing of geospatial Big Data.

Ruyan Zhou received the Ph.D. degree in agricultural bioenvironment and energy engineering from Henan Agricultural University, Zhengzhou, China, in 2007.

From 2007 to 2008, she worked with Zhongyuan University of Technology. She is currently with Shanghai Ocean University, Shanghai, China.

Yun Zhang received the Ph.D. degree in applied marine environmental studies from Tokyo University of Maritime Science and Technology, Tokyo, Japan, in 2008.

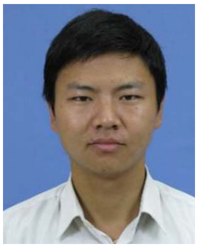
Since 2011, he has been a Professor with the College of Information and Technology, Shanghai Ocean University, Shanghai, China. His research interests include the study of navigation system reflection signal technique and its maritime application.

Yanling Han received received the M.S. degree in mechanical automation from Sichuan University, Chengdu, China, in 1999, and the Ph.D. degree in engineering and control theory from Shanghai University, Shanghai, China, in 2005.

She is currently a Professor with the Shanghai Ocean University, Shanghai, remote sensing image classification and fishery Big Data mining.

Jing Wang received the Ph.D. degree in biomedical engineering from the Department of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2014.

Since 2021, she has been an Associate Professor with the College of Information Technology, Shanghai Ocean University, Shanghai, China. Her research interests include computer vision, medical image processing.



Shuhu Yang received the Ph.D. degree in physics from the School of Physics, Nanjing University, Nanjing, China, 2012.

Since 2012, he has been a Lecturer with the College of Information Technology, Shanghai Ocean University, Shanghai, China. His research interests include evolution of the Antarctic ice sheet, hyperspectral remote sensing, and the use of navigational satellite reflections.