frontiers | Frontiers in Forests and Global Change

TYPE Original Research
PUBLISHED 17 April 2023
DOI 10.3389/ffgc.2023.1136969

OPEN ACCESS

EDITED BY
Ana Cristina Russo,
University of Lisbon, Portugal

REVIEWED BY
Tomás Calheiros,
University of Lisbon, Portugal
Huaiqing Zhang,
Chinese Academy of Forestry, China

*CORRESPONDENCE
Gui Zhang
✉ csfu3s@163.com

SPECIALTY SECTION
This article was submitted to Fire and Forests, a section of the journal Frontiers in Forests and Global Change

RECEIVED 03 January 2023
ACCEPTED 31 March 2023
PUBLISHED 17 April 2023

CITATION
Zheng Y, Zhang G, Tan S, Yang Z, Wen D and Xiao H (2023) A forest fire smoke detection model combining convolutional neural network and vision transformer. *Front. For. Glob. Change* 6:1136969. doi: 10.3389/ffgc.2023.1136969

A forest fire smoke detection model combining convolutional neural network and vision transformer

Ying Zheng, Gui Zhang*, Sanqing Tan, Zhigao Yang, Dongxin Wen and Huashun Xiao

College of Forestry, Central South University of Forestry and Technology, Changsha, China

Forest fires seriously jeopardize forestry resources and endanger people and property. The efficient identification of forest fire smoke, generated from inadequate combustion during the early stage of forest fires, is important for the rapid detection of early forest fires. By combining the Convolutional Neural Network (CNN) and the Lightweight Vision Transformer (Lightweight ViT), this paper proposes a novel forest fire smoke detection model: the SR-Net model that recognizes forest fire smoke from inadequate combustion with satellite remote sensing images. We collect 4,000 satellite remote sensing images, 2,000 each for clouds and forest fire smoke, from Himawari-8 satellite imagery located in forest areas of China and Australia, and the image data are used for training, testing, and validation of the model at a ratio of 3:1:1. Compared with existing models, the proposed SR-Net dominates in recognition accuracy (96.9%), strongly supporting its superiority over benchmark models: MobileNet (92.0%), GoogLeNet (92.0%), ResNet50 (84.0%), and AlexNet (76.0%). Model comparison results confirm the accuracy, computational efficiency, and generality of the SR-Net model in detecting forest fire smoke with high temporal resolution remote sensing images.

KEYWORDS

forest fire smoke, detection model, convolutional neural network, vision transformer, lightweight model

1. Introduction

Forest fires pose a serious threat to forest resources and people's lives and property. In the early stages of a forest fire, the low temperature makes it difficult for satellites detection. However, inadequate combustion of combustible materials produces large amounts of smoke (Wang Z. et al., 2022), presenting from the ignition to the extinguish of forest fires. Therefore, forest fire smoke could be an important indicator of the occurrence of early forest fire. Timely capture of forest fire smoke allows earlier detection of forest fires compared to the monitoring of infrared reflections of forest fires. Recent development in "high-altitude" satellite remote sensing technology (Zhang et al., 2022) makes it possible to detect forest fire smoke with remote sensing satellites. The application of remote sensing satellites in detecting forest fire smoke not only remedies the defects of "low-altitude" cameras in forest areas, including small monitoring range, poor stability, and high cost (Jia et al., 2016; Wu et al., 2020; Govil et al., 2022), but also solves the issues of "mid-altitude" Unmanned Aerial Vehicles (UAV), including constraints of air traffic controls and weather conditions and short endurance (Allison et al., 2016; Howard et al., 2018; Pérez-Rodríguez et al., 2020). Moreover, satellite remote sensing obtains timely and accurate information on forest fire smoke given

its advantages of large detection range, short response time, and strong anti-interference ability (Li et al., 2015; Filonenko et al., 2018). Although the infrared detection of high-temperature sites has been extensively studied, there is limited research on the application of satellite remote sensing in monitoring forest fire smoke for early-stage forest fire detections.

The essential of forest fire smoke detection with satellite remote sensing is the accurate identification of forest fire smoke, which requires constructing and optimizing the forest fire smoke identification algorithms. Xie et al. (2007) propose a multi-channel threshold method based on MODIS data, which eliminates pixels of other land objects in the research area, by choosing different thresholds, to extract smoke pixels. Compared with their model, the model obtained by network training using deep learning can significantly improve the detection accuracy of the forest fire smoke (Zhu et al., 2017). Based on multi-temporal and multi-spectral features, Chrysoulakis et al. (2007) use the multi-image temporal differentials algorithm to improve forest fire smoke detection. Their method identifies the forest fire smoke by discriminating the images of smoke from other ground objects with spectral differences. Convolutional Neural Networks (CNN), as a representative algorithm of deep learning, has promising applications in the field of image classification (Li et al., 2015; Zheng et al., 2019) and has been applied to forest fire smoke detection of satellite remote sensing images (Zheng et al., 2022). Li et al. (2019) propose a forest fire smoke identification model based on the Back Propagation Neural Network (BPNN). By integrating the multi-threshold approach and the BPNN classification, their method, trained with MODIS data, detects smoke by examining the spectral characteristics among the forest fire smoke and other land objects. Ba et al. (2019) further improve the accuracy of CNN for forest fire smoke detection with remote sensing images by incorporating spatial and channel-wise attentions in CNN to comb spatial features and other information from medium and high spatial resolution satellite remote sensing images. Vision Transformer (ViT), proposed by Google in 2020, is a model that applies Transformer to image classification and recognition (Bazi et al., 2021). Compared to CNN, this model has a better recognition performance with great extensibility, since it learns more comprehensive target features (Han et al., 2022). ViT can outperform CNN given sufficient samples for pre-training. In the area of image classification and recognition, ViT is pre-trained using large-scale datasets (containing ∼1.4–3 billion images) and migrated to small or medium-scale datasets to undertake specific tasks, achieving 94.55% accuracy on the CIFAR-100 dataset (Bazi et al., 2021). Unlike CNN, which has inductive bias, ViT requires more data for training to avoid over-fitting. The inductive bias, also called prior knowledge, of CNN, specifically refers to two main aspects: first is its locality, that is, the CNN considers that adjacent regions on the image have adjacent features; and second is its transitional invariance, which means the detection target always has the same prediction label no matter where it is moved to in the image. Without these two aspects, ViT requires relatively more data to learn a better model than CNN. However, due to the constraints of the in-orbit lifetime of remote sensing satellites, geographical coverage, and other conditions, there is only limited amount of remote sensing image data. It is difficult to obtain a dataset of remote sensing images, containing forest fire smoke, that large enough to avoid overfitting when training the Vit model (Zhang et al., 2018). How to accurately identify forest fire smoke with small-scale remote sensing image datasets is the key research question for effective remote sensing detections of forest fire smoke.

To address this question, this paper proposes a novel forest fire smoke detection model: the SR-Net model by combining CNN and Lightweight ViT. We construct a small-scale remote sensing image dataset using high temporal resolution remote sensing images from the Himawari-8 geostationary satellite. The front part of the SR-Net model uses CNN for inductive bias, and the back part uses the global attention of Lightweight ViT. We confirm that the SR-Net model can detect a forest fire smoke with higher accuracy and less training resources. The study compares and analyzes the performance of SR-Net with benchmark models including: AlexNet, MobileNet, GoogLeNet, and ResNet50 models to comprehensively assess the application potential of SR-Net for forest fire smoke detections. We document supportive evidence that the SR-Net consistently outperform all benchmark models in terms of Accuracy, Precision, Recall, F1-Score, and Kappa Coefficient on both the validation and test sets.

Overall, our paper makes the following contributions to the existing literature:

First, there have been very few studies on forest fire smoke detection with remote sensing satellites. Previous studies like Li et al. (2015) and Ba et al. (2019) use imagery datasets collected from polar orbit satellites. In this paper, we extent existing studies by constructing the dataset originating from the Himawari-8 satellite with high-temporal resolution. Our dataset not only allows forest fire smoke detection with different spatial satellites but also leads to the timely detection of forest fire smoke, which improves the monitoring of early forest fires.

Second, the state of the art in pattern classification and recognition is the CNN and ViT models and both models have limitations in forest fire smoke detection. Unlike CNN, ViT does not have inductive bias. Although ViT outperforms CNN, it requires large amounts of data for pre-processing. However, only limited remote sensing images containing forest fire smoke are available since the remote sensing data collection is limited by conditions such as the in-orbit lifetime of remote sensing satellites.

The proposed SR-Net model is an innovative lightweight model tailored to forest fire smoke detection with limited remote sensing imagery data. The SR-Net model combines the advantages of both CNN and ViT models. The number of parameters of the SR-Net model is lowered to six million, indicating significantly lower computational consumption. Compared to existing models, our model is superior in computational efficiency, generalization capability, robustness to environmental disturbance, and recognition accuracy. Further application of our model in forest fire detection could be promising.

The rest of this paper is organized as follows. Section 2 introduces our new constructed dataset, presents the proposed model, and illustrates the evaluation and visualization methods. Section 3 reports the results of experiments and the comparison of models. Section 4 discusses the empirical results. Finally, we conclude the paper in Section 5.

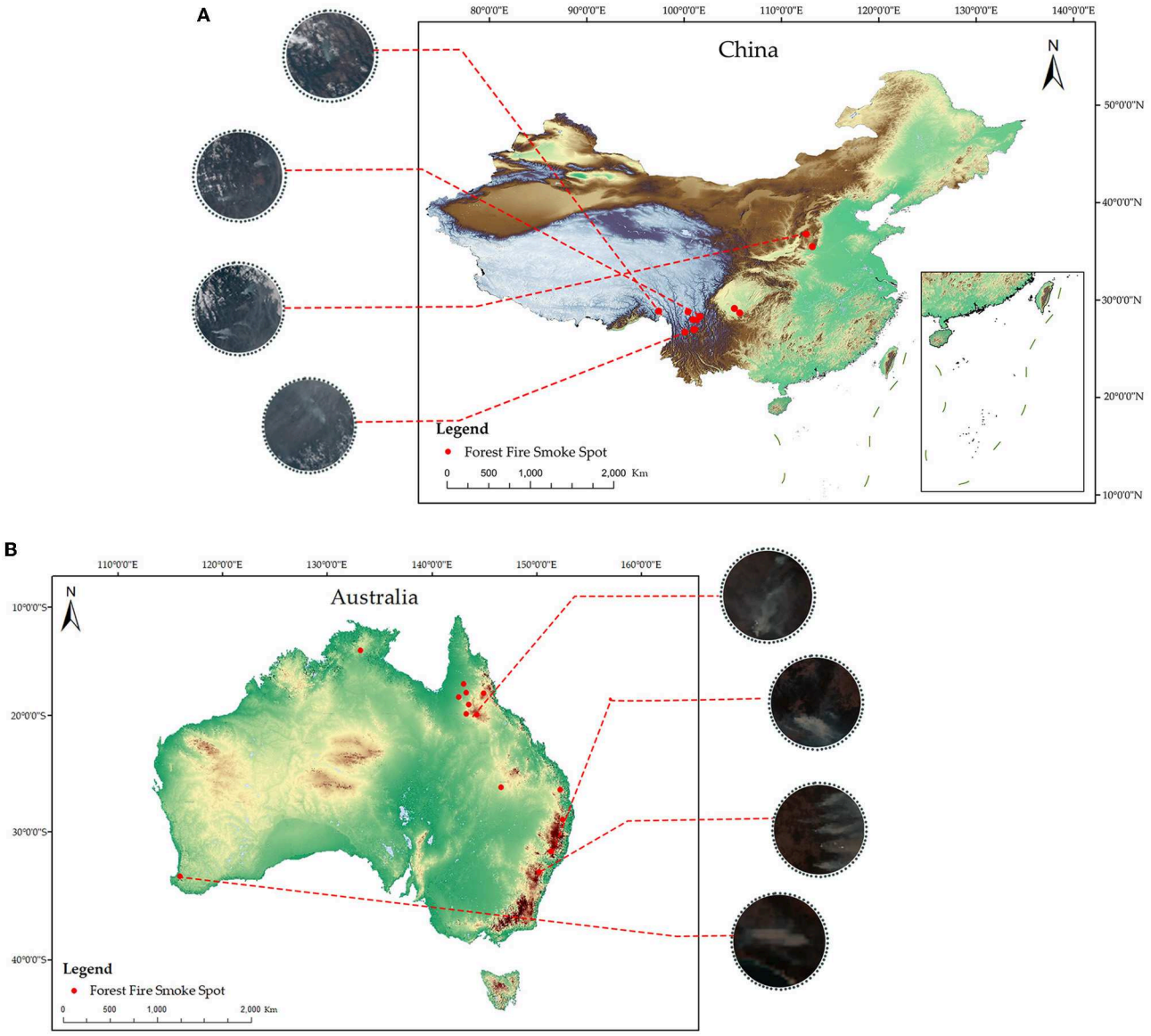

FIGURE 1
Forest fire smoke spots and examples of forest fire smoke images from some of these spots. **(A)** Denotes China, and **(B)** denotes Australia.

2. Materials and methods

2.1. Data acquisition and processing

The remote sensing image data used in the study are derived from the Himawari-8 geostationary satellite. The Himawari-8 satellite has the advantages of high timeliness and stable data quality (Yumimoto et al., 2016). Therefore, compared with polar orbit satellites, the Himawari-8 satellite can provide more timely feedback of remote sensing image information (Jang et al., 2019).

In the study, the full-disk remote sensing images of the Himawari-8 satellite are first acquired. The forest fire smoke spots are marked in Figure 1, which can be seen more directly. And the specific information of remote sensing images containing forest fire smoke is derived from the confirmed historical forest fires, whose specific acquisition date, location, longitude, and latitude are shown in Table 1, in forest areas of China and Australia. What's more, the remote sensing images containing clouds are acquired through the random sampling method. And then, we extract three visible bands: Band1, Band2, and Band3 of remote sensing images (Table 2). Finally, true color remote sensing images are synthesized by these three bands for model training, validation, and testing, and are pre-processed, including geometric correction, radiometric calibration, and atmospheric correction, to compensate for distortions in the imaging process.

As clouds and forest fire smoke are similar in color, shape and other features on true color remote sensing images, the accurate differentiation between clouds and forest fire smoke is crucial for early warning of forest fires. After pre-processing, we clip and classify these true color remote sensing images, of which

TABLE 1 Specific acquisition information of date, location, longitude, and latitude of remote sensing images containing forest fire smoke.

Date	Location	Longitude and latitude
2020.3.29	Lijiang City, Yunnan Province	100.9838E, 26.9710N
		100.0690E, 26.7162N
		101.0687E, 26.9808N
2020.3.30 to 2020.3.31	Xichang City, Sichuan Province	101.3214E, 27.9653N
2021.1.7	Ganzi Tibetan Autonomous Prefecture, Sichuan Province	100.4013E, 28.8039N
		100.8957E, 28.0091N
		101.6373E, 28.3334N
2021.2.20	Border of Henan Province with Shanxi Province	113.1757E, 35.4786N
2019.3.29 to 2019.3.30	Changzhi City, Shanxi Province	112.5494E, 36.7741N
2021.10.27 to 2021.10.28	Linzhi City, Tibet	97.3856E, 28.8735N
2022.8.21	Banan District, Chongqing City	105.1694E, 29.1426N
		105.7187E, 28.6761N
2019.12.21 to 2019.12.31	Queensland, Australia	143.4869E, 18.9582S
2015.11.20 to 2015.11.27	Queensland, Australia	146.6235E, 26.2343S
		142.5751E, 18.3024S
		144.3109E, 19.8391S
		143.2782E, 17.8637S
2019.9.7 to 2019.9.16	Queensland, Australia	143.0585E, 17.0673S 144.9701E, 17.9369S
2019.10.9 to 2019.10.15	Queensland, Australia	152.5122E, 28.9793S
		115.9222E, 33.5917S
		133.1982E, 13.9874S
2019.12.21 to 2019.12.31	New South Wales, Australia	151.3312E, 31.5785S
		150.2216E, 33.2387S
		152.2650E, 26.4607S
2019.12.21 to 2019.12.31	Queensland, Australia	143.4869E, 18.9582S
		143.2535E, 19.8081S

TABLE 2 Information of the Band 1, Band 2, and Band 3 of the Himawari-8 satellite.

Band	Central wavelength (μm)	Temporal resolution (min.)	Numbers of pixels
Band 1	0.46	10	11,000 * 11,000
Band 2	0.51	10	11000 * 11,000
Band 3	0.64	10	22,000 * 22,000

2,000 sample images contained clouds and 2,000 sample images contained forest fire smoke, with fixed size and bit depth. Figure 2 shows typical remote sensing sample images of cloud and forest fire smoke. The forest fire smoke is marked with red arrow and the cloud is marked with green arrow. According to models' training rules based on small-scale datasets, the remote sensing sample images of cloud and forest fire smoke are randomly selected in the ratio of 3(Training Set): 1(Validation Set): 1(Test Set), respectively, each containing 1,200, 400, and 400 remote sensing sample images. The Training Set is used to fit the parameters of the forest fire smoke detection models, the Validation Set is used to adjust the hyper-parameters of the model and evaluate the fitted model, and the Test Set is used to evaluate the performance and verify the generalization ability of the model.

2.2. Model structure and implementation

CNN focuses only on local features with translation invariance and rotation invariance, but there is still room to improve its

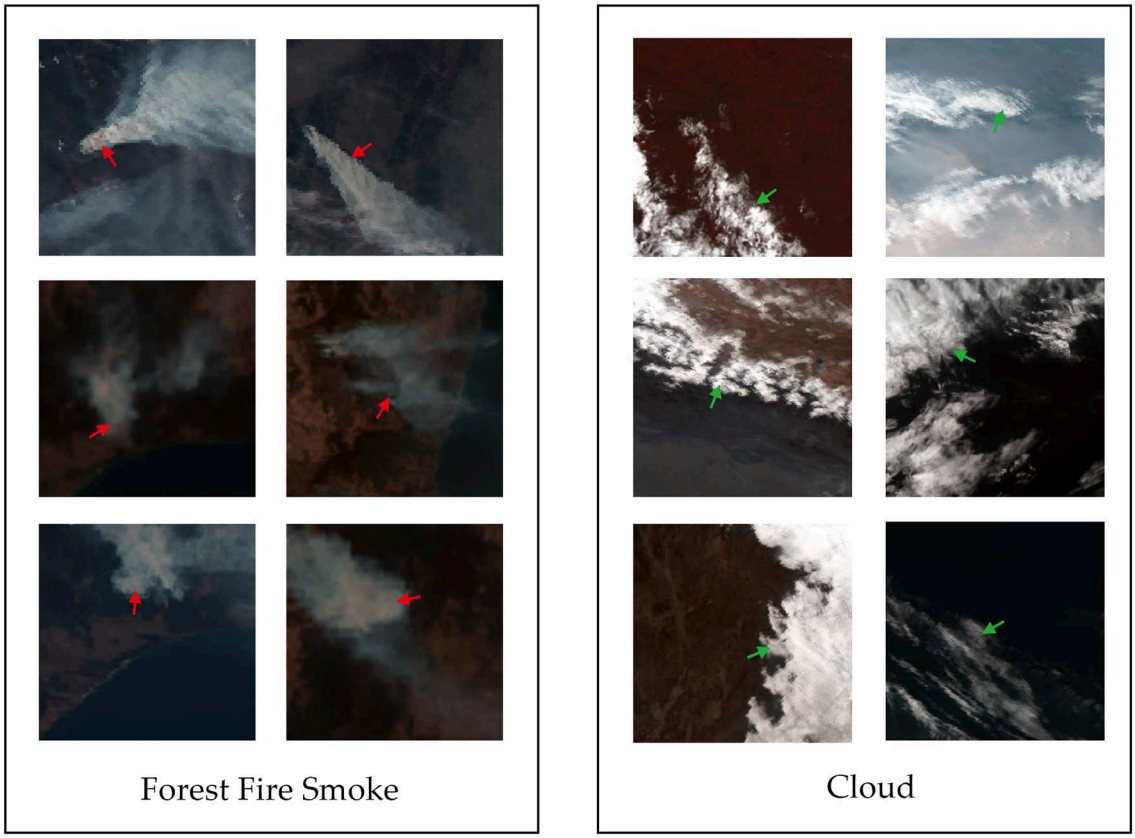

FIGURE 2
Example remote sensing sample images of cloud (marked with green arrow) and forest fire smoke (marked with red arrow).

performance, for instance, it holds a limited amount of spatial information (Kattenborn et al., 2021). Compared with CNN, ViT performs better but has shortcomings, such as the large model size, which makes training more difficult; the need to use large-scale dataset for inductive bias in advance; and the need to add additional decoders for migration to downstream tasks (Wei et al., 2022).

In this paper, we draw on the advantages of CNN and ViT to propose a new lightweight forest fire smoke detection model (SR-Net). The proposed SR-Net model takes into account the difficulty of obtaining sufficiently large-scale remote sensing datasets of forest fire smoke and optimizes the learning effect of the model on a small-scale remote sensing dataset of forest fire smoke. The network structure of the SR-Net model is shown in Table 3.

The main body of the SR-Net model uses the Inverted Residual Block, similar to MobileNet, with the input of low-dimensional features and uses Pointwise (PW) Convolution to reduce the computational complexity (Can et al., 2021).

Firstly, the channels of the feature pattern are expanded through the 1 × 1 PW convolution to enrich the number of features. Secondly, features are extracted through Depthwise (DW) Convolution, which can reduce the number of parameters and computational burden. The Depthwise Separable Convolution is one of the important part of Mobilenet V2, whose small number of parameters and computational effort compensates for the large computational effort of the ViT, allowing the model to achieve a balance between efficiency and accuracy. The DW Convolution has three steps. In the first step, a convolution kernel of size $D_k D_k$ is used on an input feature image of size $D_f D_f$ to do the Multiply Accumulation operation. In the second step, the convolution frame is slid in a left-to-right, top-to-bottom order and with a certain step size. The operation in the first step is repeated to obtain a single-channel feature image of size $D_G D_G$. In the third step, the feature images of $D_G D_G M$ dimension are kept as the output features of this layer. And the output feature image of DW convolution is processed by the Batch Normalization layer and activation function and then input to the PW Convolution layer. The PW convolution layer uses N convolution kernels of 1*1 size to map the feature image from the M-dimensional linear space to the N-dimensional space to obtain the output feature image of $D_G D_G N$. From the above process, the precise algorithm is given in Algorithm 1.

Input: Feature map $\mathbf{D_f D_f}$
Step 1 & 2: $\mathbf{D_G D_G M} = \mathbf{D_f D_f}$ *MAC* $\mathbf{D_k D_k}$
Step 3:
$\mathbf{D_G D_G N} = Conv1\_1(ReLU\,(BN\,(\mathbf{D_G D_G M})))$
Output: $\mathbf{D_G D_G N}$

Algorithm 1. Depth wise separable convolution.

TABLE 3 The network structure of the SR-Net model.

Output size	SR-Net Model
128 × 128	*Conv*, 3 × 3, 16, *stride* 2
	$\begin{pmatrix} Conv, 1 \times 1, 16 \\ DWconv, 3 \times 3, 64 \\ Conv, 1 \times 1, 32 \end{pmatrix} \times 1$
64 × 64	$\begin{pmatrix} Conv, 1 \times 1, 32 \\ DWconv, 3 \times 3, 128 \\ Conv, 1 \times 1, 64 \end{pmatrix} \times 3$
32 × 32	$\begin{pmatrix} Conv, 1 \times 1, 64 \\ DWconv, 3 \times 3, 256 \\ Conv, 1 \times 1, 96 \end{pmatrix} \times 1$
	Lightweight VIT Block × 2
16 × 16	$\begin{pmatrix} Conv, 1 \times 1, 96 \\ DWconv, 3 \times 3, 384 \\ Conv, 1 \times 1, 128 \end{pmatrix} \times 1$
	Lightweight VIT Block × 4
8 × 8	$\begin{pmatrix} Conv, 1 \times 1, 128 \\ DWconv, 3 \times 3, 512 \\ Conv, 1 \times 1, 160 \end{pmatrix} \times 1$
	Lightweight VIT Block × 3
	Conv, 1 × 1, 640, *stride* 1
1 × 1	*Average Pool*, 7 × 7, *stride* 1
FC, Softmax, 2	

In this table, "Conv" means convolution, "DWconv" means Depthwise Convolution, "Lightweight VIT" means Lightweight Vision Transformer, and "FC" means Fully Connected Layer.

Finally, convolution is used to downscale the output features to build a highly accurate deep network structure (Figure 3).

Meanwhile, the SR-Net model is alternatively added the Lightweight Vision Transformer (Lightweight VIT) Block to its network (Figure 3). To be specific, after the extraction of local features through convolution layers, the features are embedded into patches. And then the global information is obtained using the Multi-head Attention and Multilayer Perceptron (MLP). The Multi-headed attention is a mechanism that can be used to improve the performance of the general Self-attention layer (Li et al., 2021). The Single-headed attention layer restricts the ability of the model to focus on one or more specific locations without simultaneously affecting the attention to other equally important locations. This is achieved by giving the attention layer a different representation subspace. To be specific, different attention heads use different query, key, and value matrices. These matrices, due to random initialization, can project the trained input vectors into different representation subspaces and are processed by multiple independent attention heads in parallel, with the resultant vectors aggregated and mapped to the final output. The process of the Multi-head Self-attention mechanism can be expressed as Algorithm 2.

```
Input: Feature Map F
Step 1: Patches = Patch Embedding (F)
Step 2: X, Y, Z = Linear Projection(Patches)
Step 3:
Q_i = XW^{Q_i}, K_i = YW^{K_i}, V_i = ZW^{V_i}
Step 4:
Z_i = Attention (Q_i, K_i, V_i), i = 1 ... h
Step 5: MultiHead(Q, K, V) =
Concat (Z_1, Z_2, ..., Z_h) W^o
Output: MultiHead Result
```

Algorithm 2. Vision transformer.

In Algorithm 2, i denotes the header number, the number range is *1* to h. $W^O \in R^{hd_v \times d_{model}}$ denotes the output projection matrix. Z_i denotes the output matrix of each head. $W^{Q_i} \in R^{d_{model} \times d_k}$, $W^{K_i} \in R^{d_{model} \times d_k}$, $W^{V_i} \in R^{d_{model} \times d_v}$ are three different linear matrices. Similar to the Sparse Connectivity of convolution, the Multi-head attention uses a d_{model}/h-dimensional vector to separate the input into h separate attention heads and processes the features of each head in parallel. With no additional computational cost, the Multi-head attention mechanism enriches the diversity of feature subspaces.

The following is that the MLP is applied to integrate the information and the Skip Connection structure is applied to enhance the stability of the training. The integrated information is reassembled into a new feature pattern.

The final part of the SR-Net model uses Global Average Pooling to extract the individual channel information, and uses Softmax Logistic Regression to output the category information. Eventually, different features are extracted by the SR-Net model.

In this experiment, the Adam optimizer is chosen to minimize the cross entropy loss function. The Adam optimizer which can automatically adapt different learning rates for different parameters, is better than SGD optimizer that uses the same learning rate for each parameter update. The parameters of the model are set to A-0.9 and the learning rate is 1e-4. And the model is trained for a total of 100 Epochs.

2.3. Model evaluation and visualization

When the number of positive and negative samples in the dataset is balanced, the confusion matrix, which relates the true labels to the ones detected by each model (De et al., 2022), is a reliable method to count the classification results of the model. By jointly analyzing the amount of correct and mismatched true and detected labels, this method provides a direct assessment of the model's ability to predict both positive and negative cases (Table 4). In this study, the positive case denotes the forest fire smoke, and the negative case denotes the cloud.

When facing large amounts of data or multiple confusion matrices, it is difficult to accurately assess the detection capability of a model with a single confusion matrix. This requires the introduction of secondary indices based on the confusion matrix, including Accuracy (proportion of samples with correct detections out of all samples.), Precision (proportion of samples identified

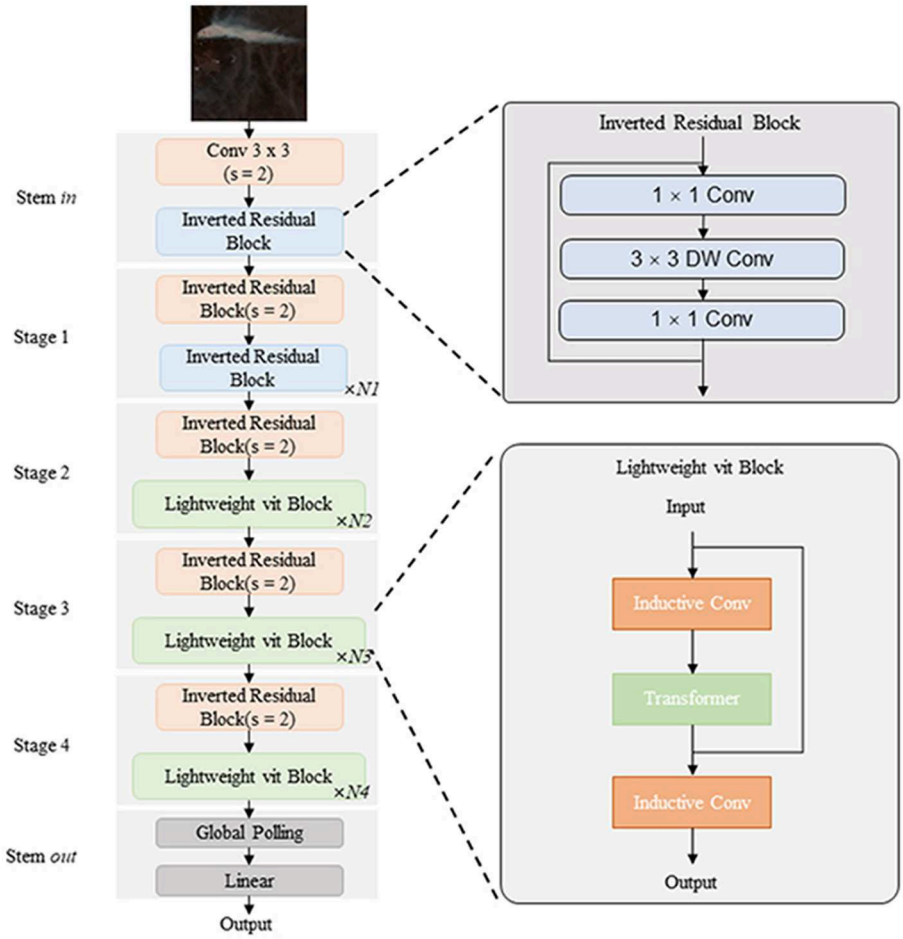

FIGURE 3
The network structure of the proposed SR-Net model. In this figure, "Conv" means convolution, "s" means stride, "Lightweight vit" means Lightweight Vision Transformer, and "Inductive Conv" means convolution layers used to perform inductive bias.

TABLE 4 The basic confusion matrix of forest fire smoke detection.

True label	Detection label	
	Forest fire smoke	Cloud
Forest fire smoke	Correctly identified	True label is forest fire smoke
	Forest fire smoke samples	Detected label is cloud
Cloud	True label is cloud	Correctly identified
	Detected label is forest fire smoke	Cloud samples

by the model as one class that are actually in that class), Recall (proportion of samples correctly detected by model as one class to the total number of that class), and even the tertiary index F1-Score (relation between Recall and Precision values) (Salih and Abdulazeez, 2021). The above secondary and tertiary indices allow for a standardized evaluation parallel comparison of models by transforming the quantitative results in the confusion matrix into ratio results between 0 and 1. The indices introduced are calculated based on Equations (1–4). Among these indices, to improve Precision, models tend to make predictions only when they are certain enough, which can result in unsure samples being missed due to over-conservatism, resulting in a lower Recall. Therefore, to achieve the best balance between Precision and Recall, the detection ability of the model is better when the result of the F1-score, which is calculated from Precision and Recall, is close to 1.

$$Accuracy = \frac{T_1 + T_2}{T_1 + F_1 + F_2 + T_2}. \quad (1)$$

$$Precision = \frac{\left(\frac{T_1}{T_1+F_2} + \frac{T_2}{T_2+F_1}\right)}{2}. \quad (2)$$

$$Recall = \frac{\left(\frac{T_1}{T_1+F_1} + \frac{T_2}{T_2+F_2}\right)}{2}. \quad (3)$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (4)$$

In Equations (1–4), T_1 represents the number of correctly identified forest fire smoke samples by the model, T_2 represents the number of correctly identified cloud samples by the model, F_1 represents the number of forest fire smoke samples being detected as cloud, and F_2 represents the number of cloud samples being detected as forest fire smoke. The indices are all macro-average, which directly adds and then averages each index of positive and negative cases, giving the same weight to index of each case.

The generalization ability of the model is a key index for evaluating its application range (Caroline and Mariana, 2022), which can be assessed by plotting the Receiver Operating Characteristic (ROC) curve, calculating the Area Under the ROC Curve (AUC), and introducing the Kappa coefficient. The ROC curve provides a visual indication of the model's detecting ability by incorporating both Precision and Recall and is independent of the decision threshold (Obuchowski and Bullen, 2018). The closer the curve is to the upper left (0, 1) coordinate, the better detecting ability the model has. The AUC is a comprehensive measure of the effectiveness of all possible classification thresholds. The closer the AUC is to 1, the more realistic the detection model is, and the higher value it has for application. The Kappa coefficient assesses the consistency of detection results with the actual situation through attempting to renormalize a debiased estimate of Accuracy (Powers, 2020). When the Kappa coefficient is in the range of 0.61–0.80 (Dettori and Norvell, 2020), it means that the detection label is substantially consistent with the true label, and the model detects well. When it is in the range of 0.81–1, it means that the detection label is almost identical to the true label, and the model detects perfectly. The kappa coefficient is calculated according to Equation (5).

$$Kappa\ coefficient = \frac{K_0 - K_e}{1 - K_e}, \quad (5)$$
$$where\ K_0 = Accuracy,$$
$$K_e = \frac{(T_1+F_1)\times(T_1+F_2)+(F_1+T_2)\times(T_2+F_2)}{(T_1+T_2+F_1+F_2)^2}.$$

In Equations (5), T_1 represents the number of correctly identified forest fire smoke samples by the model, T_2 represents the number of correctly identified cloud samples by the model, F_1 represents the number of forest fire smoke samples being detected as cloud, and F_2 represents the number of cloud samples being detected as forest fire smoke.

As for the visualization, CNN, known as black box operations, often has outputs that are difficult to interpret (Wu et al., 2018). However, the Gradient-weighted Class Activation Mapping (Grad-CAM) makes the CNN transparent through visual interpretation without modifying or retraining the model structure. The Grad-CAM can visualize the attention distribution on which the model detection is based. Hence, when the attention distribution appears to be inconsistent with the position of the detection object, such as the forest fire smoke, in original images or the model does not fit well, Grad-CAM can target the reason for model failure. This kind of visual comparative assessment can examine the forest fire smoke detection model for model bias, increase the persuasion of model effects, and enhance confidence from users in model detection results (Selvaraju et al., 2020).

In this study, we evaluate the detection of forest fire smoke by AlexNet, MobileNet, GoogLeNet, and ResNet50 models, which have a wide range of applications in the field of image classification and recognition. Among them, AlexNet deepens the net and replaces the activation function (Krizhevsky et al., 2017). MobileNet uses linear bottlenecks and inverted residuals to reduce the number of parameters and computation (Brijraj et al., 2019). GoogLeNet uses inception block to combine the outputs of convolutional kernels of different sizes for channel merging, which reduces the model complexity (Chen et al., 2022). And ResNet50 uses residual block with residual connections and introduces the Batch Normalization, so that deeper networks will have better performance (Mahdianpari et al., 2018). By comparing the forest fire smoke detection effects of the above four models with the proposed SR-Net model, this study analyzes the potential of the SR-Net model for forest fire smoke detection.

3. Results

This study proposes a forest fire smoke detection model (SR-Net) combining CNN and Lightweight ViT using small-scale remote sensing dataset. We compare and analyze the effectiveness of the SR-Net model with AlexNet, MobileNet, GoogleNet, and ResNet50 models for the detection of forest fire smoke by employing confusion matrices and visual heat images.

3.1. Evaluation of model detecting results

Statistically, the number of input parameters required for the AlexNet, MobileNet, GoogleNet, ResNet50, and SR-Net models are 38 million, 37 million, 6 million, 23 million, and 6 million, respectively.

The confusion matrix of the SR-Net model applied to forest fire smoke detection is shown in Table 5. In the Validation Set, the SR-Net model correctly identified 394 cases of forest fire smoke with a total of 400 and 398 cases of clouds with a total of 400, as well as 378 cases of forest fire smoke with a total of 400 and 397 cases of cloud with total 400 in the Test Set. The above results indicate that the SR-Net model has a higher probability than 95% of accurately detecting positive and negative cases.

Figure 4 compares the secondary indices of the SR-Net model with ResNet50, MobileNet, GoogLeNet, and AlexNet models to further analyze the forest fire smoke detection capability of the SR-Net model. According to the detection results, the SR-Net model has the highest Accuracy, Precision, Recall, F1-Score, and Kappa coefficient for the detection of forest fire smoke. This indicates that the SR-Net model outperforms the ResNet50, MobileNet, GoogLeNet, and AlexNet models in detecting forest fire smoke. On the other hand, the construction purpose of the forest fire smoke detection model is to monitor early forest fires. So among the evaluation indices, the Recall is of great practical significance, because a higher Recall represents less under-reporting of early

TABLE 5 Confusion matrix of the SR-Net model in the test and validation sets.

	Validation set		Test set	
	Detection label: forest fire smoke	Detection label: cloud	Detection label: forest fire smoke	Detection label: cloud
True label: forest fire smoke	394	2	378	3
True label: cloud	6	398	22	397

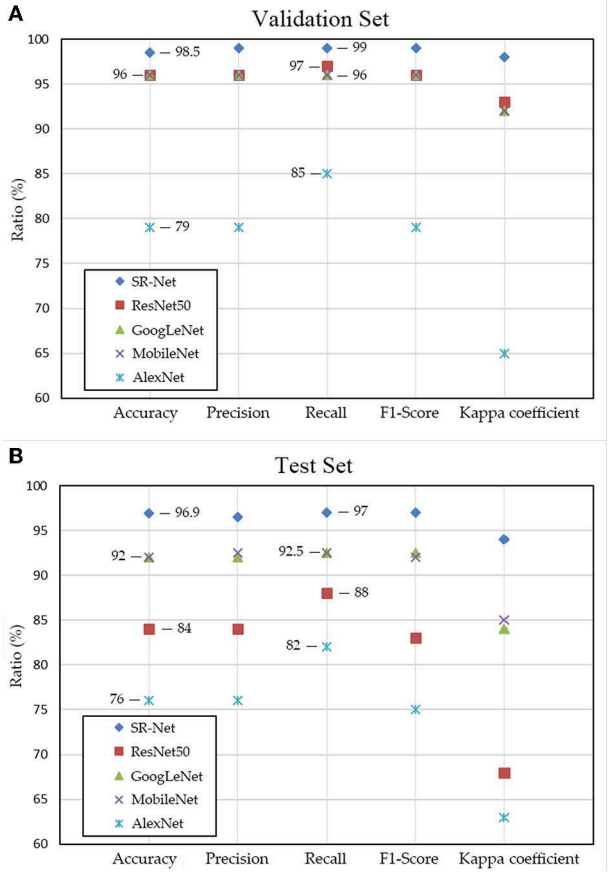

FIGURE 4
Comparison of AlexNet, ResNet50, MobileNet, GoogLeNet and SR-Net models for each evaluation indices based on validation set **(A)** and test set **(B)**.

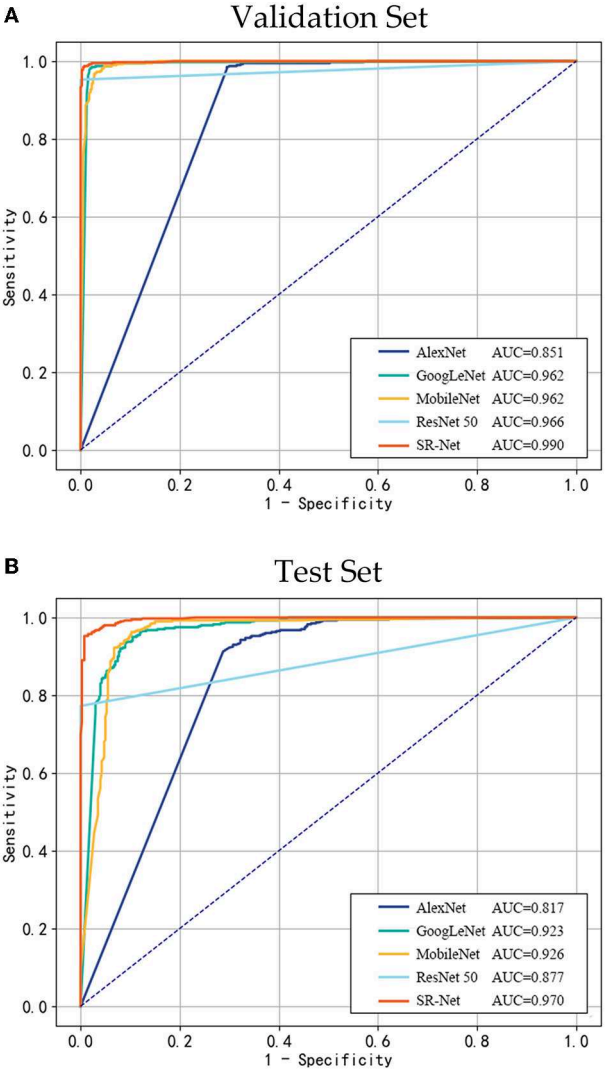

FIGURE 5
ROC curves and AUC for AlexNet, ResNet50, MobileNet, GoogLeNet, and SR-Net models on the validation set **(A)** and test set **(B)**.

forest fires in practice through forest fire smoke detection by remote sensing. Figure 4 shows that the proposed model SR-Net is superior to the other four models in terms of Recall because according to the results of the Test Set, the Recall of AlexNet model is 82%, ResNet50 model is 88%, MobileNet and GoogLeNet models are 92.5%, while that of proposed SR-Net model reached 97%. This means that the SR-Net model has a lower probability of missing the detection of forest fire smoke than the other four models. Furthermore, because of the interaction between Recall and Precision, the F1-Score has emerged to measure the balance condition of the two indices. The Recall and Precision need to be optimally balanced to avoid missing or miss-detecting of forest fire smoke to reduce missed and false forecasts in early forest fire monitoring in practical use. The F1-Score of the SR-Net model is closer to 1 (>0.95), which indicates a good balance between Recall and Precision. What's more, compared to other four models, the SR-Net model has the largest Kappa coefficient, which implies that the model's detections for forest fire smoke are highly consistent with the actual situations.

Figure 5 shows the ROC curves of AlexNet, ResNet50, MobileNet, GoogLeNet, and SR-Net models and their AUC results

for the Validation and Test sets. As shown in Figure 5, the SR-Net model corresponds to the ROC curve closest to the (0, 1) coordinate, which reveals that the SR-Net model has the best capability for forest fire smoke detection and performs consistently across different datasets. In addition, by comparing the AUC, it can be concluded that the SR-Net model has the best generalization than the other four models, and can be applied to other small-scale forest fire smoke datasets.

3.2. Visualization of model detecting effects

Figure 6 shows the original remote sensing images of forest fire smoke and heat images processed by Grad-CAM based on the SR-Net model. The white areas in both the original and the heat images represent forest fire smoke (marked with red arrows in the original images), and the attention degree from high to low is colored blue, yellow, and red in the heat images. Figure 6 compares and analyzes the smoke detection results of the SR-Net model with different proportions of smoke in an image, where (A) represents the proportion of smoke area >30%, (B) represents the proportion of smoke area <20%, and (C) represents situations when there is a small portion of clouds (marked with red rectangles) in the original smoke images. The results show that the percentage of smoke area in an image has little effect on the forest fire smoke detection of the SR-Net model, because of the global attention added to the SR-Net model. However, when the forest fire smoke is partially obscured by clouds in the original images, the attention distribution scope of the SR-Net model to detect forest fire smoke is reduced (Figure 6C). This is due to the narrowing of the distinction between forest fire smoke and background clouds when the smoke is obscured by point-like clouds, which increases the difficulty of model detection. Overall, the attention of the SR-Net model can largely avoid areas where cloud points are present. This indicates that the model has the ability of resistance to interference and can identify forest fire smoke under complex meteorological environmental conditions.

The original remote sensing images of forest fire smoke are shown in Figure 7 (A-1, B-1, C-1), along with the visualized heat images based on the AlexNet, ResNet50, MobileNet, GoogLeNet, and SR-Net models processed by Grad-CAM. The white areas marked with red arrows in the original images are forest fire smoke, and the distribution of attention when the model detects forest fire smoke is marked with blue (high), yellow (medium), and red (low) in descending order of weight.

Compared to other models, first, the SR-Net is more stable than the GoogLeNet model and is more likely to focus on the target object—the forest fire smoke. The attention distribution for detecting forest fire smoke obtained by the SR-Net model is square in shape. It largely matches the contour of the background and forest fire smoke, which occupy a larger area in images (A-6, B-6). The yellow area outside the square is the part that the model does not focus on. In addition, when the remote sensing image has a strong background texture, the attention of the SR-Net model only wraps around the target—forest fire smoke to analyze features of it (C-6). The attention distribution of the GoogLeNet model is focused on the forest fire smoke presenting an ellipse to include the target (A-5, B-5). However, there are still cases of bias in its attention distribution as can be seen in its heat images (C-5), which may cause errors when detecting forest fire smoke. Second, the performance of the ResNet50 and MobileNet models is erratic. Specifically, the shape of the attention distribution of the ResNet50 model is generally consistent with the outline of the target object—forest fire smoke (A-3). But its attention distribution is fragmented into patches and some of them are scattered to other parts of the image (B-3, C-3), which may lead to inaccurate results of forest fire smoke detection. The attention distribution of the MobileNet model shows a blocky distribution, which can almost cover the target object—forest fire smoke in most cases (A-4). But it can sometimes be shifted and completely cannot overlap with the target object (B-4, C-4). Third, the distribution of attention based on the AlexNet model is mostly blurred, with the focal (blue) areas being lightly and irregularly colored (A-2, B-2). And the focal areas only cover an extremely small proportion of the target object—forest fire smoke (C-2). In conclusion, the SR-Net model has a stable attention distribution state, outstanding detecting performance on different datasets, and good generalization performance.

A comparative analysis of the results in Figure 7 shows that the SR-Net model outperforms AlexNet, ResNet50, MobileNet, and GoogLeNet in terms of both the fitness and stability of the attention distribution state, and has a better adaptability and generalization for forest fire smoke detection.

Figure 8 presents a very small number of anomalies in the visualization of forest fire smoke images based on the SR-Net model. The first row is original forest fire smoke images and the second row is the heat images processed by Grad-CAM based on the SR-Net model. The white areas in images are the forest fire smoke sample (marked with red arrows) and the attention weight of the SR-Net model is marked from high, medium, to low with the color from blue, yellow, to red.

As shown in Figure 8, in a few cases, the attention distribution of the SR-Net model does not show a blue square but chooses to ignore the forest fire smoke roots. This may be because the SR-Net model considers the overall features in the Middle and rear of the smoke and the surrounding background to be more important than the individual target features in the thickest part of the forest fire smoke. Therefore, the SR-Net model does not choose to use the forest fire smoke as the only basis for detection and identification.

4. Discussions

At present, forest fire monitoring by meteorological satellites are mainly through infrared technology to detect high-temperature points of forest fires. The limitation is that in the early stage of forest fires, combustible materials are not fully burnt, so their temperature is not high enough to be detected. This is why the infrared band of meteorological satellites cannot receive enough energy of infrared radiation for imaging, which makes it difficult to detect in time. Therefore, there is a risk of delayed detection of forest fires. In the meanwhile, however, incomplete combustion produces a large amount of smoke. The method of forest fire smoke detection by remote sensing satellites can forecast forest fires much earlier than the method using infrared technology. However, there is scant studies on detecting forest fire smoke to forecast early forest fires. And existing researches of forest fire smoke models are mostly

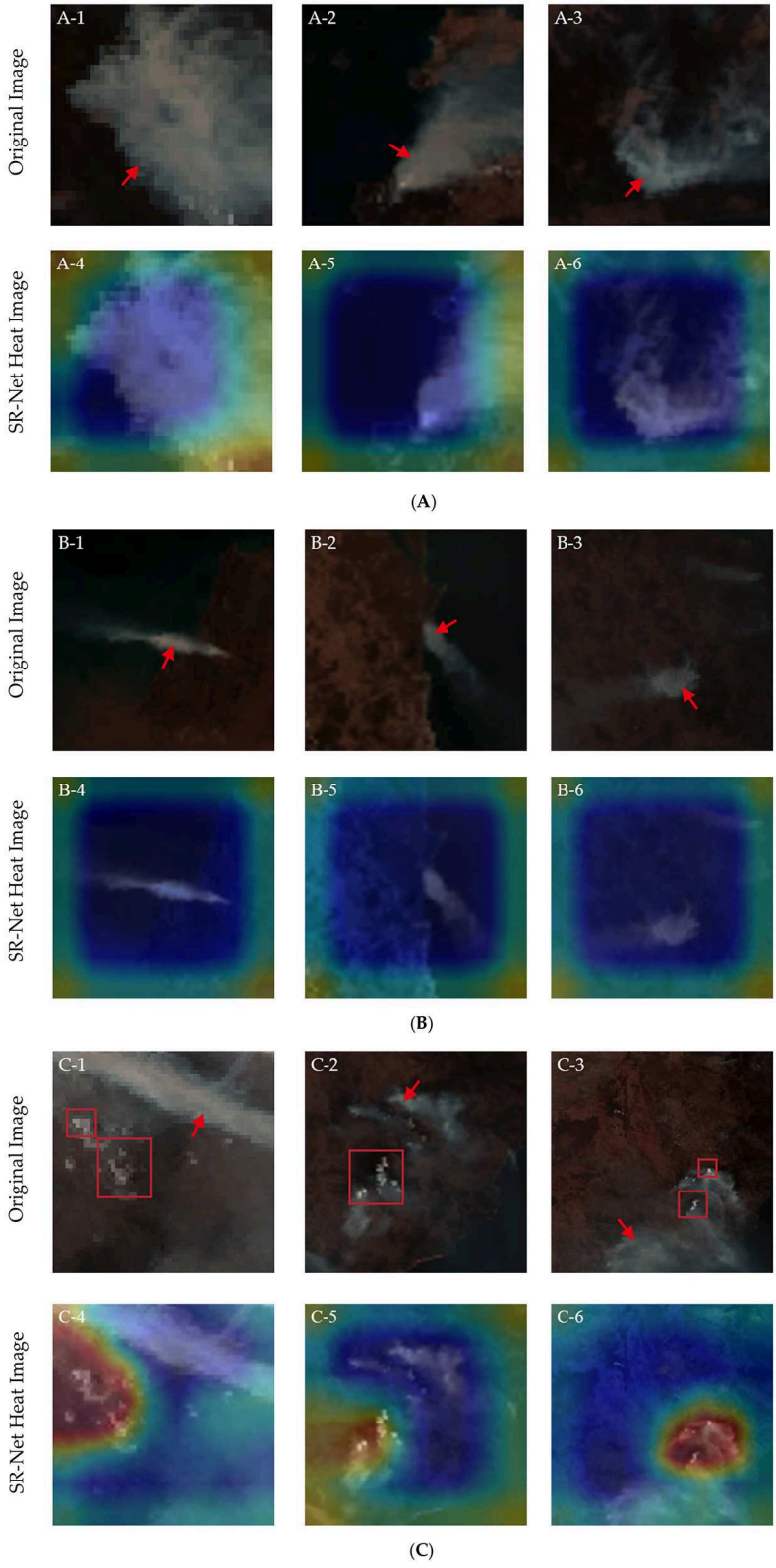

FIGURE 6
Comparison between the original remote sensing images of forest fire smoke and heat images processed by Grad-CAM based on the SR-Net model. **(A)** The proportion of smoke area >30%, **(B)** the proportion of smoke area <20%, **(C)** situations when there is a small portion of clouds (marked with red rectangles) in the original smoke images. The attention distribution when the model detects forest fire smoke is marked with blue, yellow, and red in descending order of weight from high to low.

10.3389/ffgc.2023.1136969

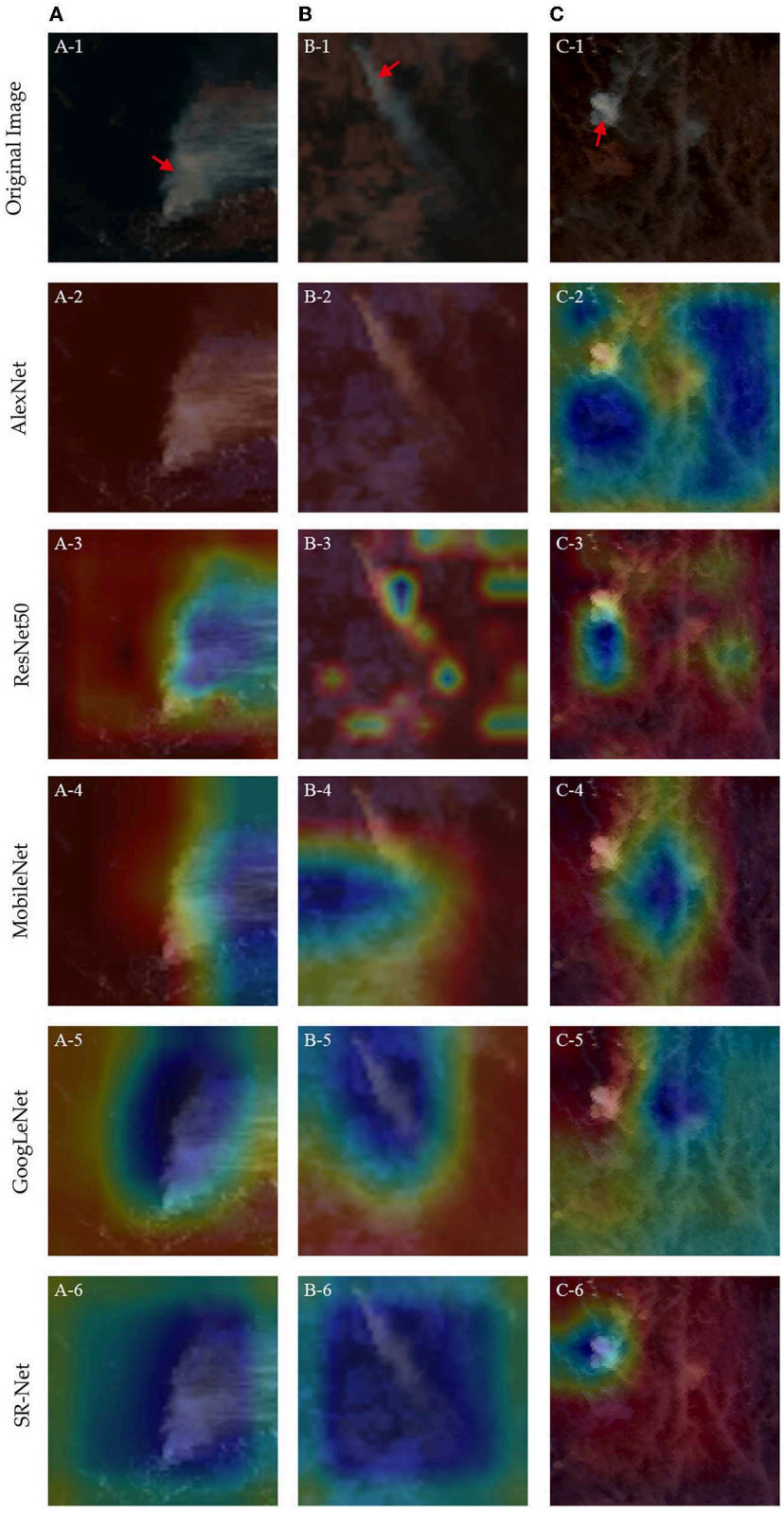

FIGURE 7
Original forest fire smoke images (A-1, B-1, C-1), comparison of heat images processed by Grad-CAM based among AlexNet (A-2, B-2, C-2), ResNet50 (A-3, B-3, C-3), MobileNet (A-4, B-4, C-4), GoogLeNet (A-5, B-5, C-5), and SR-Net (A-6, B-6, C-6) models. **(A-C)** Represents three different sets of forest fire smoke images and heat images based on each model. The attention distribution when the model detects forest fire smoke is marked with blue, yellow, and red in descending order of weight from high to low.

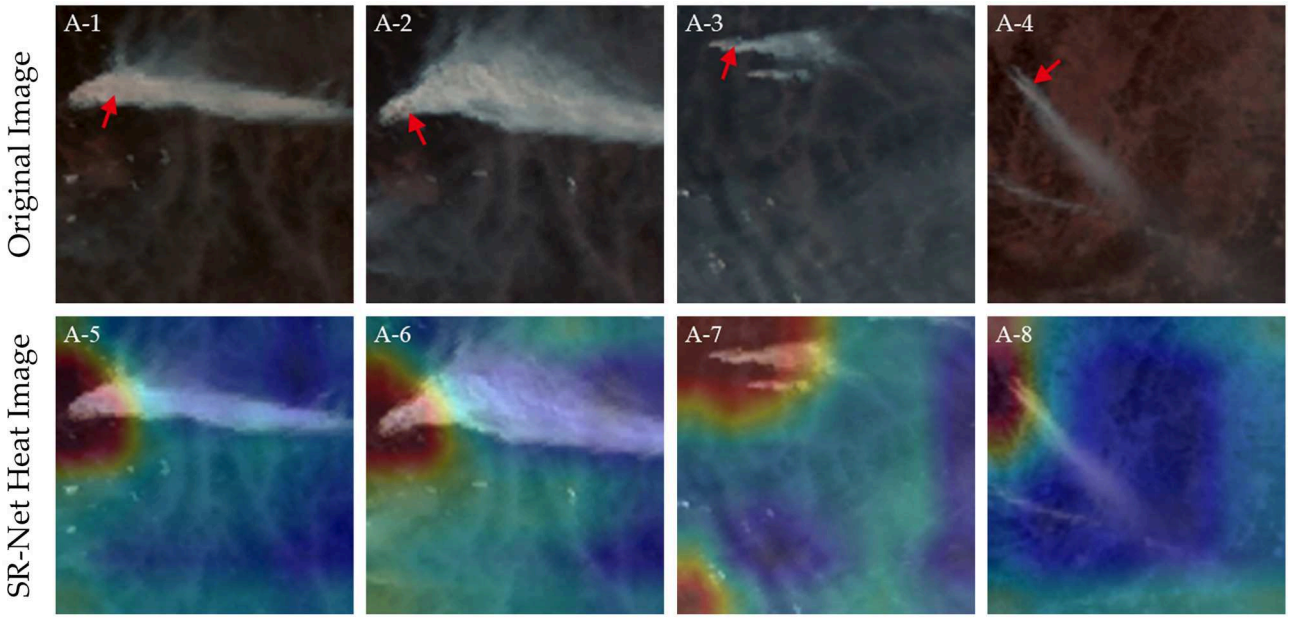

FIGURE 8
Examples of some unusual conditions of visualizations based on the SR-Net. The **(Top)** row shows the original forest fire smoke images, and the **(Bottom)** row shows heat images processed by Grad-CAM based on the SR-Net. The attention distribution when the model detects forest fire smoke is marked with blue, yellow, and red in descending order of weight from high to low.

devoted to developing CNN, for example, increasing the scale of the dataset (Zhang et al., 2018), adding different mechanisms (Xie et al., 2018), and improving the structure of CNN (Khan et al., 2021), to improve the forest fire smoke detection accuracy of models. The common problems with the above development methods are as follows: first, it is difficult to achieve an effective balance between data scale and detection accuracy. A small number of parameters will affect the detection accuracy, while a large number of parameters will require sufficient data to solve the overfitting problem (Krizhevsky et al., 2017). Second, it is hard to collect a large amount of forest fire smoke data from remote sensing satellites (Zhang et al., 2018). However, CNNs may not perform well when there are not enough datasets available for its pre-training (Sathishkumar et al., 2023). And third, the increase in data scale will cause an increase in computational resource cost. To address the above issues, we introduce the ViT and propose a forest fire smoke detection model (SR-Net) for small-scale remote sensing forest fire smoke datasets. It has improved detection accuracy and reduced resource consumption compared with traditional CNN.

Although ViT has been less studied in the field of forest fire monitoring, existing research is attempting to compare the classification accuracies of the latest CNN and ViT models on the ImageNet dataset, aimed at the image classification task (Xu et al., 2022). Part of the results is in Table 6 (Xu et al., 2022).

Table 6 indicates that ViT models have the potential to achieve comparable performance or even outperform state-of-the-art CNN architectures (Xu et al., 2022). And an increasing number of researches on ViT or progressively merging CNN and ViT have come out in various fields (Xu et al., 2022). In the field of Remote Sensing, the network combined CNN and ViT is proposed to do Hyperspectral image (HSI) classification tasks (Li et al., 2022) and to solve cross-resolution issues conducted on IKONOS and WorldView 2 with 4- and 8-band multispectral (MS) images (Wang N. et al., 2022). In the field of Scene Classification, ViT is used to distinguish scenes and obtains an average classification accuracy of 98.49% on Merced datasets (Bazi et al., 2021). In the field of Medicine, CNN and ViT are combined to diagnose Novel Corona Virus Pneumonia (COVID-19) and its result is obviously better than that of the typical CNN network (ResNet-152) (95.2%) and Transformer network (Deit-B) (75.8%) (Fan et al., 2022). And the combination is also applied to diagnose Acute lymphocytic

TABLE 6 Flogs, parameters, and accuracy of each model on the ImageNet dataset (Xu et al., 2022).

Model	Flogs (G)	Parameters (M)	Accuracy (%)
Convolution-based neural network			
ResNet	4.1	25.6	76.2
RegNetY-16G	16.0	84	82.9
EfficientNet-B7	37.0	66	84.3
Visual transformer			
ViT	55.4	86	77.9
	190.7	307	76.5
ConViT	5.4	27	81.3
	17.0	86	82.4
Swin transformer	4.5	29	81.3
	47.0	88	84.2

leukemia, and the accuracy reached 99.03% (Jiang et al., 2021). The above researches provide theoretical support for the application prospects of models combining CNN and ViT.

What's more, while previous researches of forest fire smoke detection has focused on using high spatial-resolution satellites, this paper chooses to use high time-resolution satellite. The SR-Net model, benefitting from the high temporal resolution feature of the Himawari-8 satellite, can be applied to detect forest fire smoke promptly. Considering the difficulty of collecting remote sensing sample data of forest fire smoke, the front part of SR-Net uses CNN to make inductive bias of forest fire smoke samples, complementing the missing priori knowledge of ViT due to being based on a small-scale dataset. The back part uses lightweight ViT, which adds a global attention mechanism compared to CNN, allowing the model to achieve better performance on small-scale remote sensing datasets of forest fire smoke. SR-Net simplifies the structure and reduces the number of parameters to reduce computational costs and resource consumption while maintaining the detection accuracy of forest fire smoke. Other benchmarks, like GoogLenet et al., are still essentially convolutional models. The advantage of convolutions lies in its inductive bias, while the disadvantage lies in its inability to effectively obtain and construct the global information, as the size of the convolutional kernel is finite. However, our proposed model introduces the ViT, which has the multi-headed attention mechanism, compensating for the disadvantages of all kinds of convolutional models.

By comparing the results of different indices, it is found that the SR-Net model has the highest accuracy, precision, recall, F1-Score, and Kappa coefficient of forest fire smoke detection, outperforming AlexNet, MobileNet, ResNet50, and GoogLeNet models. In forest fire smoke detection, the SR-Net model has a higher accuracy and lower false rate than other models as well as a better capacity of balance between the two, allowing the model to make the best measurements and avoid being too 'conservative' or 'confident' in its judgments. This balance allows the model to be used more reliably in real-time scenarios of forest fire smoke detection and helps the human and material resources needed to confirm forest fires to be deployed more efficiently and effectively, without time and resource consuming.

In addition, heat images of SR-Net attention distribution drawn by Grad-CAM show that SR-Net presents a wider attention distribution in images because of the global attention mechanism. This mechanism allows a more comprehensive exploration of forest fire smoke features and is less affected by the proportion of forest fire smoke in remote sensing images. What's more, when there is interference from point-like clouds in forest fire smoke images, the difficulty of classification and detection increases for the reason that the distinction between forest fire smoke and the cloud is narrowed. In this case, the attention scope of the SR-Net model is reduced. But the reduced scope manifests that the SR-Net model has better resistance to interference for it can avoid the areas where cloud points are present when detecting forest fire smoke. By comparing the heat images of AlexNet, ResNet50, MobileNet, and GoogLeNet models, this study find that the SR-Net model is more stable and fit for forest fire smoke detection than other models. It has a more fixed pattern of attention distribution for forest fire smoke detection, showing a square shape, which can include the target object and its background. With this pattern, the detection by the SR-Net model does not tend to miss the target object—forest fire smoke. When the background has strong textural features, the detection capability of the model is disturbed and there is a reduction in the scope of attention distribution. The GoogLeNet model, in general, performs well with regard to the attention distribution state but is less stable than the SR-Net model. The attention distribution of it is affected by the presence of background interference, resulting in a shift. The attention distribution of ResNet50 and MobileNet models is unstable and can cover most of the target object—forest fire smoke in most cases, however, there are also cases where the attention distribution is scattered or only covers a small portion of the target object. The attention distribution of AlexNet is blurred, which cannot cover forest fire smoke, resulting in the poor effect of forest fire smoke detection. To sum up, the SR-Net model is more effective in detecting forest fire smoke under complex environmental conditions with better accuracy and greater generalization.

The major limitation of our study is that the model needs to be put to further practical use to explore what contingencies exist in real-time applications and to refine the model by developing emergency pre-solutions. Notwithstanding the limitation, the SR-Net model is more effective and stable than traditional CNNs in detecting forest fire smoke with high timeliness. Therefore, it has the potential to be used in practical applications to help monitor forest fire smoke or as a complement to monitoring forest fire smoke through near-infrared bands.

5. Conclusions

In this paper, we propose a lightweight forest fire smoke detection model (SR-Net) combining the merits of CNN and ViT models and construct a new small-scale remote sensing dataset, containing cloud and forest fire smoke, collected from the Himawari-8 satellite. We conduct a comprehensive evaluation of the of SR-Net and benchmark models, including AlexNet, MobileNet, GoogLeNet, and ResNet50. We conclude our findings as follows:

(1) The combination of CNN and ViT allows for a lightweight forest fire smoke detection model (SR-Net), and reduces the number of the model parameters to six million. The model can be applied to small-scale remote sensing datasets of forest fire smoke images.
(2) The results of the confusion matrix manifest that the SR-Net model is more than 95% likely to accurately detect positive and negative cases of forest fire smoke samples. On both the validation and test sets, the SR-Net model is superior to AlexNet, MobileNet, GoogLeNet, and ResNet50 models in comparison criteria including: Accuracy, Precision, Recall, F1-Score, and Kappa Coefficient.
(3) Visualization of the model attention in detecting forest fire smoke by Grad-CAM revealed that the SR-Net model has a wide range of attentions, which could comprehensively explore the features of the remote sensing images, leading to an accurate detection of the forest fire smoke with less interference from environmental factors. The comparison of the heat images further confirms the outperformances of the SR-Net model over

benchmark models in adaptability and stability for forest fire smoke detections.

This paper sheds a light on the lightweight models of forest fire detection with small-scale datasets. The documented model performance calls for further application of the proposed model on broader sets of imagery data from multiple satellites to test the model generality. As it is difficult for existing remote sensing satellites to achieve the coexistence of high temporal and high spatial resolution, future research may focus on processing spatial resolution information collected from high temporal resolution remote sensing images. Moreover, recent developments in Computer Vision (CV) could further improve forest fire smoke detections by exploring the migration and scalability of the new networks.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

GZ conceived and designed the study. YZ wrote the first draft, collected the study data, and performed the experiment. GZ, ZY, and ST provided critical insights in editing the manuscript. GZ and ST are responsible for the project management. HX and DW assisted in the supervision. All authors have read and agreed to the published version of the manuscript.

Funding

This work was funded by the National Natural Science Foundation Project of China (Grant No. 32271879) and the Science and Technology Innovation Platform and Talent Plan Project of Hunan Province (Grant No. 2017TP1022).

Acknowledgments

Many thanks to the editor and reviewers for their valuable comments.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Allison, R. S., Johnston, J. M., Craig, G., and Jennings, S. (2016). airborne optical and thermal remote sensing for wildfire detection and monitoring. *Sensors* 16, 1310. doi: 10.3390/s16081310

Ba, R., Chen, C., Yuan, J., Song, W., and Lo, S. (2019). SmokeNet: satellite smoke scene detection using convolutional neural network with spatial and channel-wise attention. *Remote Sens*. 11, 1702. doi: 10.3390/rs11141702

Bazi, Y., Bashmal, L., Rahhal, M. M. A., Dayil, R. A., and Ajlan, N. A. (2021). Vision transformers for remote sensing image classification. *Remote Sens*. 13, 516. doi: 10.3390/rs13030516

Brijraj, S., Durga, T., and Sharan, K. A. (2019). Shunt connection: an intelligent skipping of contiguous blocks for optimizing MobileNet-V2, *Neural. Netw*. 118, 192–203. doi: 10.1016/j.neunet.2019.06.006

Can, H. K., M., Cagri, K., Turker, T., Sengul, D., and, U., et al. (2021). Automated classification of remote sensing images using multileveled MobileNetV2 and DWT techniques. *Expert Syst. Appl*. 185, 115659. doi: 10.1016/j.eswa.2021.115659

Caroline, M. G., and Mariana, B. (2022). Assessing the generalization capability of deep learning networks for aerial image classification using landscape metrics. *Int. J. Appl. Earth Obs*. 114, 103054. doi: 10.1016/j.jag.2022.103054

Chen, H., Fu, X., and Dong, J. (2022). SAR target recognition based on inception and fully convolutional neural network combining amplitude domain multiplicative filtering method. *Remote Sens*. 14, 5718. doi: 10.3390/rs14225718

Chrysoulakis, N., Herlin, I., Prastacos, P., Yahia, H., Grazzini, J., and Cartalis, C. (2007). An improved algorithm for the detection of plumes caused by natural or technological hazards using AVHRR imagery. *Remote Sens. Environ*. 108, 393–406. doi: 10.1016/j.rse.2006.11.024

De, D. I. M., Redondo, A. R., Fernández, R. R., Jorge, N., and Javier, M. M. (2022). General performance score for classification problems. *Appl. Intell.* 52, 12049–12063. doi: 10.1007/s10489-021-03041-7

Dettori, J. R., and Norvell, D. C. (2020). Kappa and beyond: is there agreement? *Glob. Spine J.* 10, 499–501. doi: 10.1177/2192568220911648

Fan, X. L., Feng, X. F., Dong, Y. Y., and Hou, H. C. (2022). COVID-19 CT image recognition algorithm based on transformer and CNN. *Displays* 72, 102150. doi: 10.1016/j.displa.2022.102150

Filonenko, A., Hernandez, D. C., and Jo, K. H. (2018). Fast smoke detection for video surveillance using CUDA. *IEEE Trans. Industr. Notify* 14, 725–733. doi: 10.1109/TII.2017.2757457

Govil, K., Welch, M. L., Ball, J. T., and Pennypacker, C. R. (2022). Preliminary results from a wildfire detection system using deep learning on remote camera images. *Remote Sens.* 12, 166. doi: 10.3390/rs12010166

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., et al. (2022). A survey on vision transformer. *IEEE T. Pattern Anal.* 45, 87–110. doi: 10.1109/TPAMI.2022.3152247

Howard, J., Murashov, V., and Branche, C. M. (2018). Unmanned aerial vehicles in construction and worker safety. *Am. J. Ind. Med*. 61, 3–10. doi: 10.1002/ajim.22782

Jang, E., Kang, Y., Im, J., Lee, D.-W., Yoon, J., and Kim, S.-K. (2019). Detection and monitoring of forest fires using himawari-8 geostationary satellite data in South Korea. *Remote Sens*. 11, 271. doi: 10.3390/rs11030271

Jia, Y., Yuan, J., Wang, J. J., Fang, J., Zhang, Q. X., and Zhang, Y. M. (2016). A saliency-based method for early smoke detection in video sequences. *Fire Technol.* 52, 1271–1292. doi: 10.1007/s10694-014-0453-y

Jiang, Z. C., Dong, Z. X., Wang, L. Y., and Jiang, W. P. (2021). Method for diagnosis of acute lymphoblastic leukemia based on ViT-CNN ensemble model. *Comput. Intel. Neurosc.* 2021. doi: 10.1155/2021/7529893

Kattenborn, T., Leitloff, J., Schiefer, F., and Hinzb, S. (2021). Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS J. Photogramm*. 173, 24–49. doi: 10.1016/j.isprsjprs.2020.12.010

Khan, S., Muhammad, K., Hussain, T., Ser, J. D., and Albuquerque, V. H. C. D. (2021). DeepSmoke: deep learning model for smoke detection and segmentation in door environments. *Expert Syst. Appl.* 182, 115125. doi: 10.1016/j.eswa.2021.115125

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM*. 60, 84–90. doi: 10.1145/3065386

Li, C. M., Wang, X. Y., Chen, Z. H., Gao, H. M., and Xu, S. F. (2022). Classification of hyperspectral image based on dual-branch feature interaction network. *Int. J. Remote Sens.* 43, 3258–3279. doi: 10.1080/01431161.2022.2089069

Li, J., Wang, X., Tu, Z. P., and Michael, R. L. (2021). On the diversity of multi-head attention. *Neurocomputing* 454, 14–24. doi: 10.1016/j.neucom.2021.04.038

Li, W., Chen, C., Zhang, M., Li, H. C., and Du, Q. (2019). Data augmentation for hyperspectral image classification with deep CNN. *IEEE Geosci. Remote Sens. Lett*. 16, 593–597. doi: 10.1109/LGRS.2018.2878773

Li, X., Song, W., Lian, L., and Wei, X. (2015). Forest fire smoke detection using back-propagation neural network based on MODIS data. *Remote Sens*. 7, 4473–4498. doi: 10.3390/rs70404473

Mahdianpari, M., Salehi, B., Rezaee, M., Mohammadimanesh, F., and Zhang, Y. (2018). Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery. *Remote Sens.* 10, 1119. doi: 10.3390/rs10071119

Obuchowski, N. A., and Bullen, J. A. (2018). Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Phys. Med. Biol.* 63, 07TR01. doi: 10.1088/1361-6560/aab4b1

Pérez-Rodríguez, L. A., Quintano, C., Marcos, E., Suarez-Seoane, S., Calvo, L., and Fernández-Manso, A. (2020). Evaluation of prescribed fires from unmanned aerial vehicles (UAVs) imagery and machine learning algorithms. *Remote Sens.* 12, 1295. doi: 10.3390/rs12081295

Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint.2010.16061*. doi: 10.48550/arXiv.2010.16061

Salih, A. A., and Abdulazeez, A. M. (2021). Evaluation of classification algorithms for intrusion detection system: a review. *J Soft Comput Data Min*. 2, 31–40. doi: 10.30880/jscdm.2021.02.01.004

Sathishkumar, V., Cho, J., Subramanian, M., and Naren, O. (2023). Forest fire and smoke detection using deep learning-based learning without forgetting. *Fire Ecol*. 19, 1–17. doi: 10.1186/s42408-022-00165-0

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* 128, 336–359. doi: 10.1007/s11263-019-01228-7

Wang, N., Meng, X. J., Meng, X. C., and Shao, F. (2022). Convolution-embedded vision transformer with elastic positional encoding for pansharpening. *IEEE T. Geosci. Remote* 60, 1–9. doi: 10.1109/TGRS.2022.3227405

Wang, Z., Yang, P., Liang, H., Zheng, C., Yin, J., Tian, Y., et al. (2022). Semantic segmentation and analysis on sensitive parameters of forest fire smoke using smoke-unet and landsat-8 imagery. *Remote Sens*. 14, 45. doi: 10.3390/rs14010045

Wei, H. P., Deng, Y. Y., Tang, F., Pan, X. J., and Dong, W. M. (2022). A comparative study of CNN- and transformer-based visual style transfer. *J. Comput. Sci. Technol.* 37, 601–614. doi: 10.1007/s11390-022-2140-7

Wu, Q., Shen, C., Wang, P., Anthony, D., and Anton, V. D. H. (2018). Image captioning and visual question answering based on attributes and external knowledge. *IEEE T. Pattern Anal*. 40, 1367–1381. doi: 10.1109/TPAMI.2017.2708709

Wu, X., Lu, X., and Leung, H. (2020). A motion and lightness saliency approach for forest smoke segmentation and detection. *Multimed. Tools Appl*. 79, 69–88. doi: 10.1007/s11042-019-08047-5

Xie, Y., Qu, J. J., Xiong, X., Hao, X., Che, N., and Sommers, W. (2007). Smoke plume detection in the eastern United States using MODIS. *Int. J. Remote Sens.* 28, 2367–2374. doi: 10.1080/01431160701236795

Xie, Z., Song, W., Ba, R., Li, X., and Xia, L. (2018). A spatiotemporal contextual model for forest fire detection using himawari-8 satellite data. *Remote Sens*. 10, 1992. doi: 10.3390/rs10121992

Xu, Y. F., Wei, H. P., Lin, M. X., Deng, Y. Y., Sheng, K. K., Zhang, M. D., et al. (2022). Transformers in computational visual media: a survey. *Comp. Visual Media*. 8, 33–62. doi: 10.1007/s41095-021-0247-3

Yumimoto, K., Nagao, T. M., and Kikuchi, M. (2016). Aerosol data assimilation using data from Himawari-8, a next-generation geostationary meteorological satellite. *Geophys. Res. Lett.* 43, 5886–5894. doi: 10.1002/2016GL069298

Zhang, B., Wu, Y. F., Zhao, B., Jocelyn, C., Dan, F. H., Jing, Y., et al. (2022). Progress and challenges in intelligent remote sensing satellite systems. *IEEE J-STARS* 15, 1814–1822. doi: 10.1109/JSTARS.2022.3148139

Zhang, Q. X., Lin, G. H., Zhang, Y.-M., Xu, J., and Wang, J.-J. (2018). Wildland forest fire smoke detection based on faster R-CNN using synthetic smoke images. *Proc. Eng*. 211, 441–446. doi: 10.1016/j.proeng.2017.12.034

Zheng, X., Chen, F., Lou, L., Cheng, P., and Huang, Y. (2022). Real-time detection of full-scale forest fire smoke based on deep convolution neural network. *Remote Sens*. 14, 536. doi: 10.3390/rs14030536

Zheng, Y. P., Li, G. Y., and Li, Y. (2019). Survey of application of deep learning in image recognition. *Comput. Eng. Appl.* 55, 20–36. Available online at: en.cnki.com.cn/Article_en/CJFDTotal-JSGG201912004.htm

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L. P., Xu, F., et al. (2017). Deep learning in remote sensing: a comprehensive review and list of resources. *IEEE Geosc. Rem. Sen. M* 5, 8–36. doi: 10.1109/MGRS.2017.2762307