

Landslide Detection for Remote Sensing Images Using a Multilabel Classification Network Based on Bijie Landslide Dataset

Yongxin Li , Zhihui Xin , Guisheng Liao , Senior Member, IEEE, Penghui Huang , Member, IEEE, and Mengting Yuan 

Abstract—To effectively mitigate disaster damage, it is crucial to obtain landslide information quickly and accurately with the abundant remote sensing images. Although related landslide detection research has been carried out a lot in recent years, all of the research articles are still based on single-label image classification. Meanwhile, the existing methods face challenges in balancing global and local information effectively, interpreting and modeling intricate interlabel dependencies. To solve the problems above, we perform multilabel annotation on the Bijie landslide dataset and propose a novel multilabel classification network for landslide detection. The proposed network consists of a feature representation module and a category relation parsing module. The first module is designed to extract feature maps that represent high-level semantic information and convert them into feature sequences. This can fully utilize the capability of the bidirectional long short-term memory network to model label dependencies. The second module is used to model interlabel dependencies to enhance the ability of the model for efficiently detecting landslides. Compared with other recent landslide detecting methods, the results show that the proposed scheme is more effective. F_1 score on landslide detection reaches up to 97.56% and the attention area of categories is more precise in the proposed method. Therefore, the proposed model can provide reliable and effective decision support for disaster emergency response.

Index Terms—Bidirectional long short-term memory (BiLSTM), landslide detection, multilabel classification (MLC), remote sensing images, Swin transformer.

I. INTRODUCTION

WITH climate change and the intensification of human activities, geological disasters, such as landslides, pose

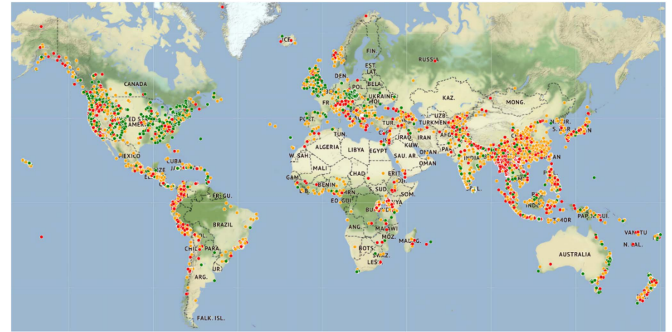


Fig. 1. Global landslide distribution from 2007 to 2022.

a serious threat to the safety of human life and property [1]. As illustrated in the NASA statics, from 2007 to 2022, 39 594 landslide events occurred worldwide, about 2639 cases annually, and the incidence is rising year by year. The global distribution of landslides is shown in Fig. 1. In Fig. 1, red, orange, and green represent the large, medium, and small landslides, respectively. The image clearly reveals the serious impact of landslides on human survival and development. Therefore, it is crucial to understand the causes of landslides and detect them quickly and accurately, which can help mitigate their impact on human society and the natural environment. This will not only provide timely warning and response to emergencies but also effectively guide urban planning and natural resource management to ensure the safety and sustainable development of human life. Landslides often occur in remote mountainous areas, making field surveys very difficult and dangerous [2]. In this regard, achieving rapid and accurate extraction of landslides has always been a challenging problem.

In recent years, with the rapid development of remote sensing technology, various remote sensing data have played an important role in landslide detection [3], [4], [5], [6], [7]. Landslide detection involves extensive data processing. The amount of data generated by remote sensing technology is huge and requires efficient algorithms and computing resources for processing and analysis. Data fusion and analysis require complex data processing processes to ensure the accurate identification of landslide areas. For example, fusion among optical, elevation, and vegetation data. Therefore, the capabilities of heuristics and statistical methods are not sufficient to cope with the increasingly

Manuscript received 10 January 2024; revised 14 March 2024; accepted 7 April 2024. Date of publication 12 April 2024; date of current version 1 May 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62361058 and Grant 62171272, in part by Yunnan Provincial Natural Science Foundation under Grant 202201AT070027, and in part by Yunnan Expert Workstation under Grant 202305AF150012. (Corresponding author: Zhihui Xin.)

Yongxin Li, Zhihui Xin, and Mengting Yuan are with the Yunnan Key Laboratory of Opto-Electronic Information Technology, School of Physics and Electronic Information, Yunnan Normal University, Kunming 650500, China (e-mail: liyx_cdut@163.com; xinzhihui.luncky@163.com; yuanhsp37@163.com).

Guisheng Liao is with the National Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China (e-mail: liaogs@xidian.edu.cn).

Penghui Huang is with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: huangpenghui@sjtu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3387744

complex landslide detection tasks [8]. Deep learning is increasingly used in landslide detection. However, it requires a large amount of labeled data for training and has high requirements for data quality. Therefore, more effective deep-learning-based landslide detection methods are necessary to be further studied.

Some previous studies have utilized deep-learning-based methods to detect landslides [9], [10], [11], [12], [13]. Deep learning combined with change detection has shown potential in the field of landslide detection. For example, Lv et al. [14] effectively improved the performance of deep-learning neural networks in surface change detection through iterative training sample enhancement technology. Furthermore, they utilized a multiscale attention network [15] and adaptive region key point extraction technology [16] to improve the performance of landslide change detection. Some methods utilize deep-learning-based object detection methods to detect landslides. For instance, Han et al. [17] utilized a Yolov3-based detection model to enhance the accuracy of detecting small-scale landslides and complex backgrounds' landslides. Meanwhile, Yang et al. [18] employed a mask R-convolutional neural network (CNN) semantic segmentation model for automated landslide identification to enhance the efficiency of landslide detection. However, studies have indicated that the credibility of the results is highly dependent on the quality of the dataset and the correlation factors [19]. Furthermore, labeling datasets for target detection and semantic segmentation requires significant expertise and human resources [20], [21]. Generally, landslide detection using target detection methods mainly learns the specific characteristics of landslides. However, landslides have diverse spatial and spectral characteristics compared with other land-type features. Previous research work was unable to capture the spatial and spectral characteristics of landslides and failed to explore the potential information from optical remote sensing images in landslide detection.

Landslide detection through scene classification can provide a more comprehensive and in-depth understanding of landslides and use the potential relationship between landslides and ground objects to improve landslide detection performance. Scene classification includes single-label and multilabel classification (MLC). Previous methods only consider landslides as single-label classification. For instance, Ji et al. [22] applied image classification to detect landslides and obtained the visualized landslide locations by class activation mapping (CAM). Hacıfendioglu et al. [23] confirmed the effectiveness of image classification for locating landslides and compared the effectiveness of various visualization methods, such as Grad-CAM [24] and Score-CAM [25]. Li et al. [26] utilized shipboard landslide image data for landslide detection through image classification. They compared various classic neural networks, including VGG, DenseNet, EfficientNet, SEResNet, and vision transformer (ViT) [27], [28], [29], [30], [31], where ViT shows the best performance. These methods demonstrate that landslides can be effectively detected through scene classification. The single-label classification method simply classifies images as “landslides” or “nonlandslides.” Although it can describe the image content from a macroperspective, it ignores the potentially rich landslide area topographic information in the image. The



Fig. 2. Classes present in the example remote sensing image. (a) Tree. (b) Pavement and tree. (c) Building, field, pavement, and tree.

categories existing in three example remote sensing images are illustrated in Fig. 2. As shown in Fig. 2, assigning the three samples to the tree category alone is appropriate for (a). However, for (b) and (c), such a classification cannot completely describe the image content. Landslide area topographic information is very important for landslide detection. This relationship contains the universal law of landslide formation. For example, vegetation can strengthen the soil through its root system and reduce erosion, thus reducing the risk of landslides. Human activities, such as agriculture, mining, road construction, and urban development, change natural landscapes and may increase the risk of landslides. The presence of water can cause soil saturation, increasing the likelihood of landslides. Therefore, considering landslide area topographic information can help improve the accuracy of landslide detection. In addition, combining multi-source data will face problems of heterogeneous data matching and model complexity. Although this approach provides us with a more comprehensive perspective, it also brings new challenges in data processing and analysis.

To address the above problems, MLC is a better strategy. MLC assigns multiple target labels to an image, resulting in a more complete description of the image content. At the same time, the relationship between ground object categories is also considered, and feature information of ground object types is introduced into landslide detection. Making full use of landslide area topographic information can further explore the potential information of optical remote sensing images in landslide detection. MLC methods based on neural networks are mainly divided into five types: based on CNN, based on recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, based on graph neural networks (GNNs), based on transformer networks, and based on hybrid neural network models. Methods based on CNN use hierarchical structures to extract image features, which have powerful visual feature extraction capabilities, but may ignore global contextual information. Methods based on RNN are good at capturing dependencies between sequences. However, when processing long sequences, RNN may suffer from vanishing or exploding gradients. Methods based on GNN can simulate the complex relationships between entities in graph-structured data, but its ability to extract visual features is relatively limited. Methods based on transformer network utilize self-attention (SA) mechanism to deal with long-distance dependencies and can effectively capture global context information. Kaselimi et al. [32] considered monitoring deforestation as an MLC task and proposed the ForestViT MLC model based

on ViT, which performed better than the CNN methods [33]. However, a single model cannot comprehensively capture and handle the various complex features and dependencies that arise in MLC tasks, which limits their adaptability when dealing with complex data scenarios.

Many hybrid model-based methods are proposed to further enhance the accuracy and robustness of complex multilabel tasks. Wang et al. [34] proposed an MLC network based on CNN-RNN, which combines the feature extraction capabilities of CNN and the sequence processing capabilities of RNN to improve MLC performance. Hua et al. [35] proposed the CA-Resnet50 bidirectional LSTM (BiLSTM) model, which revisited the CNN and BiLSTM, and introduced a class attention layer to alleviate the dimensional pressure of BiLSTM. Chen et al. [36] proposed the ML graph convolutional network (GCN), which utilizes GCNs to map label representations into interdependent target classifier. Wang et al. [37] combine Swin transformer [38] and GNN to extract image features and learn label dependencies for multilabel image recognition. All these methods attempt to capture the features of classes in remote sensing images and learn interclass relationships by combining different types of neural networks. Among these methods, the combination of CNN and RNN provides advantages for processing complex data, but it also brings high computational complexity. In addition, due to the essential differences between feature maps and sequences, RNN and its extension BiLSTM have limitations in modeling the relationship between feature maps extracted by CNN without processing. The method of combining CNN or transformer with GNN maps labels relationships to the backbone network. The effect of such methods relies on the construction of knowledge graphs [39], which requires higher standards for the dataset. Therefore, the strategy of network combination is crucial to the performance of the entire model. Efficiently integrating networks to exploit their respective strengths and address their limitations is still a problem that requires in-depth research.

To tackle these problems, in this study, we consider landslide detection as an MLC task, complement the multilabel annotations of the Bijie landslide dataset, and propose a new MLC network for landslide detection. The proposed model effectively balances complexity and performance and solves the imbalance problem of global and local feature extraction, which does not require additional data, such as knowledge graphs. Moreover, we improve the efficiency of connection between models through the conversion of feature maps to feature sequences. The main outstanding contributions of this article are given as follows.

- 1) In this article, landslide detection is considered for the first time as an MLC task. An MLC method is used to detect landslides, which fully considers landslide area topographic information. Achieve more accurate landslide detection by further exploring the potential information of optical remote sensing images. At the same time, it overcomes the heterogeneous matching of multisource data and the complexity of the model.
- 2) Currently, there is no large high-quality multilabel landslide dataset. In this study, we performed multilabel

annotation on the Bijie landslide dataset, which is available online.¹

- 3) An MLC network for landslide detection is proposed, which includes a feature representation (FR) module and a category relation parsing (CRP) module. The backbone of the FR and CRP modules is similar to the Swin transformer and BiLSTM, respectively. The FR module effectively extracts class feature maps of optical remote sensing images, which takes into account both global and local information. The feature map is converted into a feature sequence, which retains the FR ability and is also consistent with the data format processed by CRP. The CRP module captures potential deep information in optical remote sensing images and learns landslide area topographic information. Introducing interclass topographic feature relation into landslide detection can enhance detection performance. The proposed network is an end-to-end network, and the two modules cooperate and complement each other.

The rest of this article is organized as follows. The structure of the two basic components of the proposed network is described in Section II. Comprehensive details concerning the dataset are encapsulated in Section III. Our experimental settings and the results from both quantitative and qualitative perspectives are presented in Section IV. Section V provides the discussion. Finally, Section VI concludes this article.

II. MODEL

A. Network Architecture

In this study, a method combining Swin transformer and BiLSTM is proposed, called Swin-BiLSTM-FS (Swin transformer and BiLSTM-based feature sequence). The network architecture of Swin-BiLSTM-FS is shown in Fig. 3. This network is primarily composed of an FR module and a CRP module. As shown in Fig. 3, the blue grid represents the FR module, while the orange grid represents the CRP module. Let F denotes the feature map, S denotes the feature sequence, and w_i denotes the feature word. Swin transformer is utilized by the FR module to extract the feature map F , which represents high-level semantic information. Following that, it is converted into the feature sequence S . The feature sequence S is initially mapped to the relevant dimension by the CRP module, which ensures that the length of the processed feature vector remains uniform postmapping. Subsequently, BiLSTM is employed by the module to infer the dependencies in the feature sequence.

Swin transformer and BiLSTM are integrated in this study. On the one hand, Swin transformer incorporates a multistage window and sliding window attention mechanism based on the ViT, which enables it to effectively capture both global and local information. This allows Swin transformer to exhibit exceptional performance in capturing image features and managing spatial structure. On the other hand, BiLSTM is a particular type of RNN that can effectively handle long-range dependencies within sequential data. It is worth noting that the relationships between

¹[Online]. Available: <https://github.com/caicaiw/Multi-label-Bijie-landslidedataset.git>

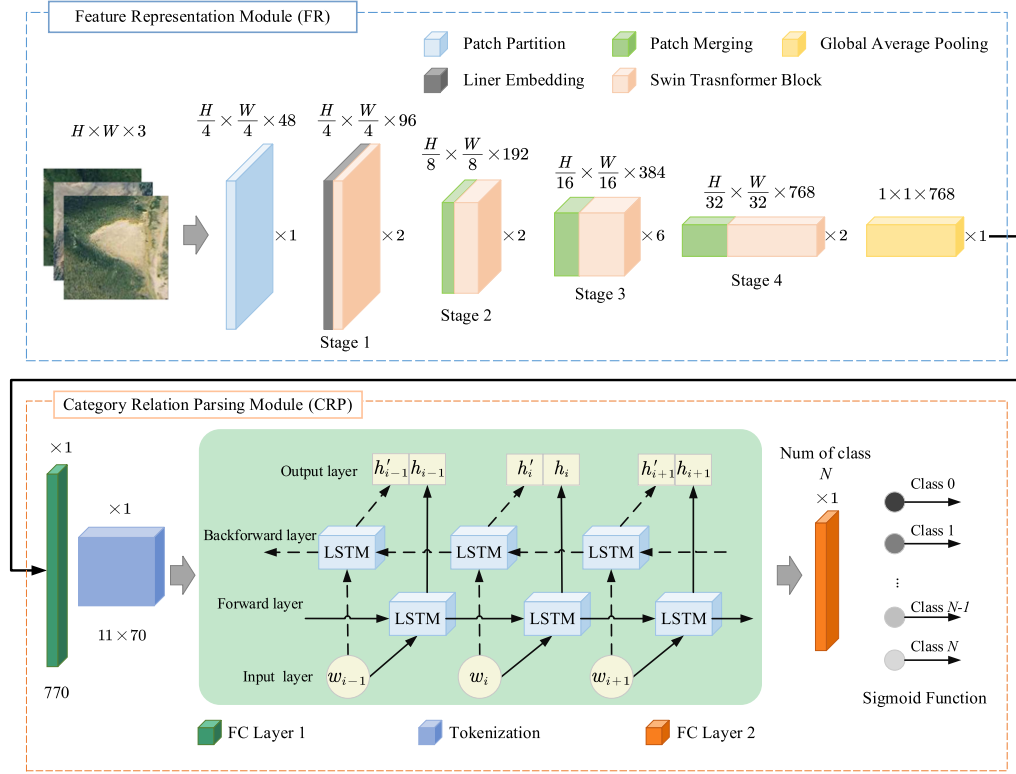


Fig. 3. Architecture of Swin-BiLSTM-FS.

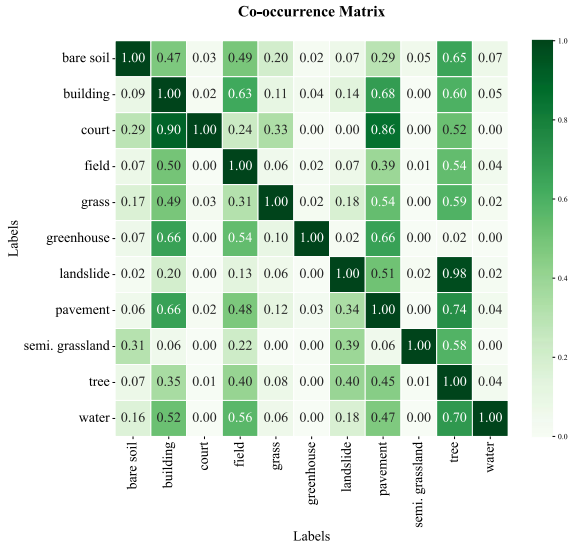


Fig. 4. Label co-occurrence matrix of the Bijie multilabel landslide dataset.

categories are asymmetric in nature. The label co-occurrence matrix of the Bijie multilabel landslide dataset is illustrated in Fig. 4. In this matrix, each row and column represents a specific label, and the elements in the matrix represent the co-occurrence probability of the corresponding two labels. The darker the color of the element, the higher the probability of the two labels co-occurring, and the higher the correlation. On the contrary, the lighter the color, the lower the probability of co-occurrence between the two, and the lower the correlation. As shown in Fig. 4, the probability of a court appearing when

a building is present is 0.02. Conversely, the probability of a building appearing when a court is present is 0.9, indicating that the category relationship is asymmetric. Therefore, inferring the label relationship in both directions helps to improve the model performance.

In conducting a more in-depth exploration of the correlation strategy between the two modules, it is necessary to compress the dimensions of feature map and convert it into a sequence of features appropriate for BiLSTM processing. In order to achieve this goal, global averaging pooling (GAP) is used to compress the dimensions of the feature map. This reduces its size while retaining some FR ability and spatial data. Although conventional methods typically compress feature maps along the channel direction, such compression might diminish the ability of FR. In addition, directly inputting data with spatial structure into the BiLSTM may result in a loss of spatial information. Consequently, in this study, the method of compressing the feature maps along the dimensional direction is adopted. This method is designed to fulfill the input requirements of BiLSTM while preserving spatial information.

Detailed descriptions of the FR and CRP modules are available in Section II-B and II-C, respectively.

B. FR Module

Extracting high-level semantic features is crucial for image classification tasks [40]. CNNs process a fixed perceptual field and, thus, excel at learning local features [41]. To achieve long-range dependencies, the depth of the network needs to be increased. However, this may cause problems, such as gradient

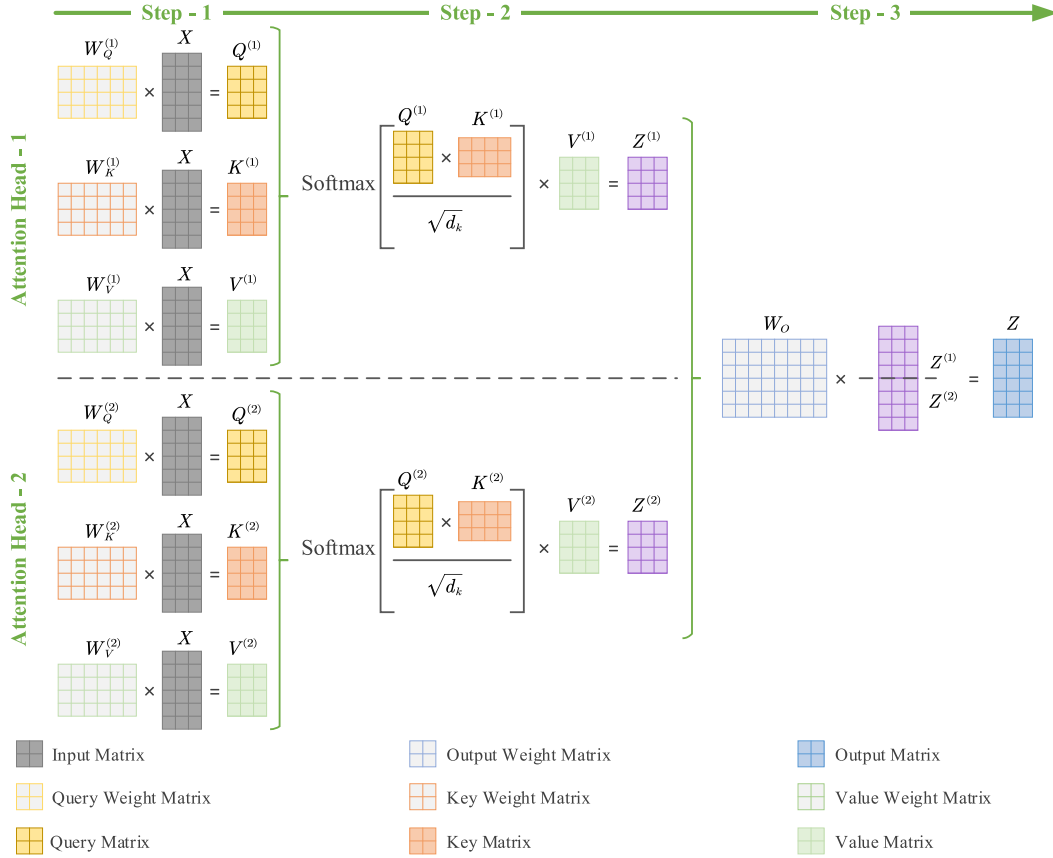


Fig. 5. Detailed structural diagram of the multihead SA mechanism.

disappearance or explosion, at the same time, as increasing the complexity of the model. The ViT is able to learn global information in the spatial domain, which helps to capture contextual relationships in high-resolution remote sensing images [31]. Nevertheless, the texture, shape, and edge features contained in local information are essential for identifying and distinguishing different categories. Therefore, they should be given equal importance. Swin transformer is one of the most popular models of ViT currently. It has been proven to be able to strike a better balance between global and local feature information, making it suitable for processing high-resolution images [38]. In this study, the Swin transformer is adopted as the backbone for feature learning. The architecture of the FR module is illustrated within the blue grid in Fig. 3.

Denote the input remote sensing image by I with size $I_{\text{width}} = 224$ and $I_{\text{height}} = 224$. The Swin transformer is utilized to split the image into nonoverlapping patches of size 4×4 with the number n , as shown in (1). For an image of 224×224 and patch size of 4×4 , 56×56 patches can be obtained

$$n = \frac{I_{\text{width}}}{4} \times \frac{I_{\text{height}}}{4}. \quad (1)$$

Each 4×4 patch is flattened into a 16-D vector and then through a fully connected layer to map it to the embedding dimension D ($D = 96$). This adaptation of the image is necessary for the transformer architecture. For a given patch p_i

$\forall i \in (1, n)$ following this mapping, we obtain a matrix of size $(56 \times 56 \times 96)$, where each row represents the embedded representation e_i of a patch

$$e_i = W_l \times p_i + b_l \quad (2)$$

where W_l and b_l represent the weight matrix and bias vector of the fully connected layer, respectively.

The embedded representation e_i is input into multiple stages of Swin transformer model. In each stage, multiple Swin transformer blocks are included, which are composed of a window multihead SA or sliding window multihead SA sublayer and a multilayer perceptron (MLP) sublayer. The input image is partitioned into numerous local windows, with each window containing pixels that are independently performed for the multiheaded self-attention (MSA) computation. Afterward, the window is slid over the image and the process above is repeated. Such an approach effectively balances local and global feature information and reduces computational complexity. The MSA mechanism, comprised of multiple independent SA mechanisms, is illustrated in Fig. 5. Head is set to 2 in order to display the MSA more concisely. In the proposed model, the number of heads in each MSA is 4. This setting is to fully utilize the capabilities of the multihead SA mechanism while maintaining a balance between model computational complexity and performance, in

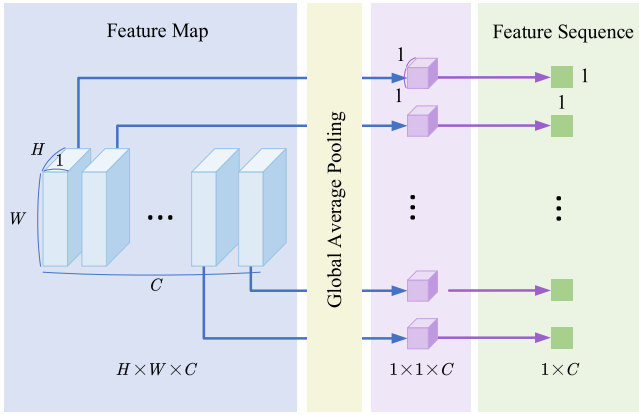


Fig. 6. Detailed block diagram of GAP.

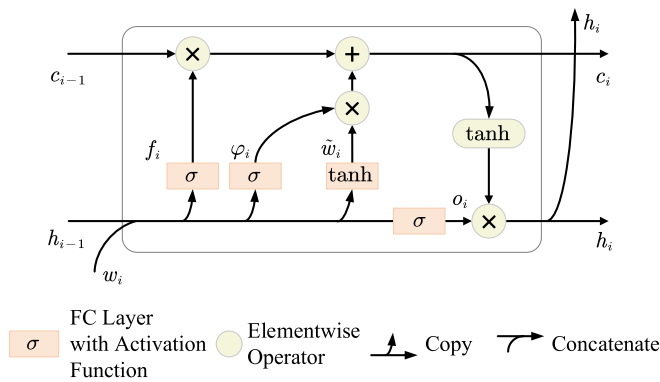


Fig. 7. Detailed structure of LSTM.

which the SA mechanism can be expressed as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) \times V \quad (3)$$

where Q , K , and V denote the query matrix, key matrix, and value matrix, respectively, and d_k is the dimension of Q and K .

After MSA and MLP, the dimensions of the feature map remain unchanged. Therefore, between multiple stages, a patch merging layer is employed to reduce the spatial resolution of the feature map while increasing the depth of the network. In this way, the model can learn more abstract and high-level semantic features while reducing computational cost. The framework of FR is similar to Swin-Tiny. The difference is that the last hidden layer and classification head are removed in order to reduce the computational complexity and minimize overfitting, resulting in an output size of $H = 7$, $W = 8$, and $C = 768$, i.e., $7 \times 7 \times 768$. In the fourth stage of Swin transformer, a GAP layer is added to compress the feature map along the size direction and convert the feature map into a feature sequence. This reduces the dimensionality of the feature map while retaining a certain degree of spatial information and FR capabilities. The GAP process is shown in Fig. 6, where the feature map of size $H \times W \times C$ is transformed into $1 \times 1 \times C$ after GAP to obtain a $1 \times C$ feature sequence. Adjusting the dimensional order of the feature tensor does not change the essential information of

the data [42]. Thus, after altering the dimensional order of the tensor, we calculate the average along the size dimension of the feature map

$$\text{GAP}(F_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W F_c(i, j) \quad (4)$$

where H and W denote the height and width of the feature map, respectively, F_c denotes the feature map of the c th channel, and GAP denotes the global average pooling. At this time, the dimensions of the feature map are reduced to 1×1 , resulting in a 1×768 feature sequence S .

After the conversion of the feature map into a feature sequence, the relationship among features becomes more intuitive, improving the interpretability of model. In addition, such an approach can take full advantage of the ability of BiLSTM to capture contextual relationships and improve the performance of the model in MLC tasks. Therefore, the feature sequence extraction module is referred to as the FR.

C. CRP Module

For complex MLC tasks, the modeling of interlabel dependencies plays a crucial role. Introducing GNN in CNN and representing interlabel relationships by using pretrained label vectors helps to improve the performance of the model on MLC tasks [36]. However, the problem of frequency bias may occur when constructing global graphs [43]. Moreover, the multiple message passing and node updates of GNN may increase the computational cost. To address this issue, Hua et al. [35] proposed a method that employs BiLSTM to model interlabel dependencies and utilizes convolutional layers to aggregate and reduce the dimensionality of channel features. This approach can reduce the computational burden of the BiLSTM while maintaining higher performance.

Therefore, this study utilizes BiLSTM to model interlabel relationships, which is different from [35]. First, we aggregate features over feature map dimensions by GAP instead of using convolutional layers over channels to aggregate features. Consequently, spatial information and feature representability are preserved to a certain degree. Second, feature sequences are mapped into feature vectors prior to being input into the BiLSTM. This process facilitates a more efficient inference of the relationships among them.

In the method described above, the high-resolution remote sensing image is converted into a feature sequence S , following the operation of the FR. Subsequently, based on the number of categories, the feature sequence is mapped into a suitable dimension, which is necessary to maintain a consistent length for each word vector. Moreover, the sequence length must be an integer multiple of the number category. Therefore, when the category number is 11, the feature sequence is mapped to 770 (i.e., 1×770). Before inputting the feature sequence into BiLSTM, it is mapped into a feature vector. Specifically, the 1×770 feature sequence is divided into 11 feature vectors, each of dimension 70, culminating in a 11×70 feature matrix w .

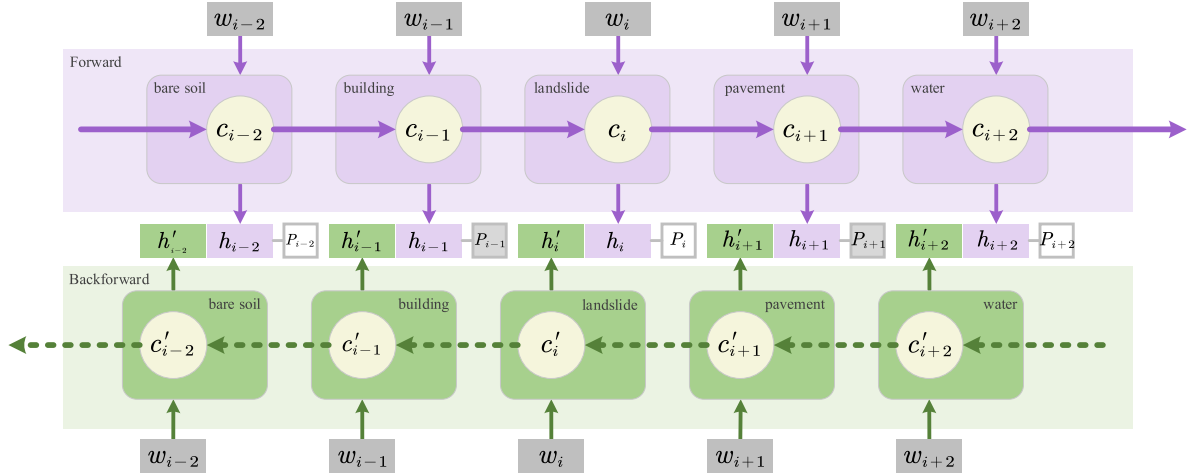


Fig. 8. Detailed structure of BiLSTM.

BiLSTM is composed of two reverse LSTM units, forward and backward. The essence of LSTM resides in its ability to capture and retain long-term dependencies among words through its gated architecture. A comprehensive structure of LSTM is shown in Fig. 7. In Fig. 7, the input vector at the i th time step is represented by w_i . During the i th time step, the relationship between the i th word and the previously processed words is established.

First, a new candidate information $\tilde{w}_i$ is derived based on w_i and the hidden state h_{i-1} of the previous time step

$$\tilde{w}_i = \tanh(W_w \times [h_{i-1}, w_i] + b_w) \quad (5)$$

where W_w and b_w denote the weight matrix and bias vector used to calculate the new candidate information, respectively, as well as $\tanh(\cdot)$ represents the cosine tangent function.

After obtaining the candidate information $\tilde{w}_i$, decisions are made by the LSTM regarding the amount of new information to be introduced to the cell state c_i , influenced by the value of the input gate φ_i . It is calculated by the sigmoid activation function, as shown in (6), with a range of 0 and 1. If the value of input gate is close to 1, a majority of new candidates will be integrated into the cell state. Conversely, if the value is close to 0, most of the new candidate information will be disregarded

$$\varphi_i = \text{sigmoid}(W_i \times [h_{i-1}, w_i] + b_i) \quad (6)$$

where W_i and b_i represent the weight matrix and bias vector used for the computation of the input gate, respectively.

The cell state c_i can be expressed as

$$c_i = f_i \odot c_{i-1} + \varphi_i \odot \tilde{w}_i \quad (7)$$

where c_{i-1} denotes the cell state from the previous time step, $\odot$ represents the dot product, and f_i is the value of the forget gate, expressed as follows:

$$f_i = \text{sigmoid}(W_f \times [h_{i-1}, w_i] + b_f) \quad (8)$$

where W_f and b_f denote the weight matrix and bias vector employed in the computation of the forget gate, respectively. The value of the forget gate determines how much information

is retained from the previous cell state, while the input gate, along with the new candidate information, governs the updating of the cell state to capture new information.

The output gate o_i can be expressed as follows:

$$o_i = \text{sigmoid}(W_o \times [h_{i-1}, w_i] + b_o) \quad (9)$$

where W_o and b_o represent the weight matrix and bias vector used for the computation of the output gate, respectively. The o_i determines the extent of information extracted from the cell state c_i to produce the hidden state h_i

$$h_i = o_i \odot \tanh(c_i). \quad (10)$$

Through this mechanism, relevant outputs are generated by the LSTM at each time step while preserving long-term dependencies.

The detailed structure of BiLSTM is illustrated in Fig. 8. In Fig. 8, BiLSTM consists of forward and backward LSTM units. Consequently, two hidden states are obtained at time step i . Let h_i and h'_i denote the hidden states of the forward and backward LSTM outputs, respectively. From the output $[h_i, h'_i]^T$ at time step i , the probability P_i of the i th feature word is predicted

$$P_i = \text{sigmoid}(a_i[h_i, h'_i]^T + b_i) \quad (11)$$

where a_i and b_i denote the weight vector and bias vector of the fully connected layer, respectively. The P_i denotes the probability at time step i . For enhanced interpretability, it is deliberately aligned with category i .

In this article, the dimension of each feature vector is established at 70. Simultaneously, the hidden state size of the BiLSTM is also set to 70. Through this configuration, the expressiveness and computational complexity of the model are effectively balanced. Given the complexity of the task and the volume of data, the LSTM is set up with 32 cells. After that, probability predictions are generated for all categories.

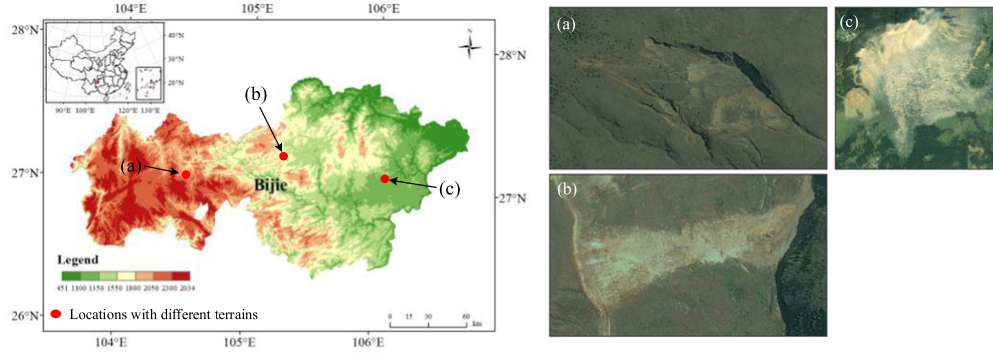


Fig. 9. Geographic information of the study area. (a) Mountainous area. (b) Plateau. (c) Karst Landform.

TABLE I
NUMBER OF SAMPLES FOR DIFFERENT OBJECT CATEGORIES IN THE BIJIE
MULTILABEL LANDSLIDE DATASET

Category No.	Category name	Training	Test	Total
1	Bare soil	133	78	211
2	Building	744	345	1089
3	Court	14	7	21
4	Field	928	452	1380
5	Grass	163	88	251
6	Greenhouse	43	16	59
7	Landslide	513	257	770
8	Pavement	758	363	1121
9	Semi-arid Grassland	25	11	36
10	Tree	1214	626	1840
11	Water	63	33	96

III. DATA DESCRIPTION

The Bijie landslide dataset [22] is composed of 2773 remote sensing images, which were captured by the TripleSat satellite from May and August 2018 with a ground resolution of 0.8 m. Within them, there are 770 landslides' samples and 2003 nonlandslides' samples. However, the Bijie landslide dataset oversimplifies the complex ground environment into only two categories—landslide and nonlandslide, which creates a disconnect with the reality. Therefore, the Bijie multilabel landslide dataset is conducted by reassigning multiple target labels to the Bijie landslide dataset.

Compared with the original dataset, the total number of samples in the Bijie multilabel landslide dataset remains the same, but each optical remote sensing image includes 11 candidate labels: bare soil, building, field, grass, greenhouse, landslide, pavement, semiarid grasslands, tree, and water. Among them, the categories of field and semiarid grasslands are determined based on the actual target labels present in the Bijie landslide dataset and the ecosystem characteristics of Bijie. Geographic information of Bijie is also taken into account, as shown in Fig. 9. The number of relevant images for each target category is listed in Table I, while the samples of landslide and nonlandslide scene

categories and the class labels they encompass are illustrated in Fig. 10.

During the construction of the Bijie multilabel landslide dataset, we performed four rounds of multilabel annotation iterations. The multilabel annotation iterative process is shown in Fig. 11. As shown in Fig. 11, each iteration includes expert annotation to generate a new dataset, and baseline model experiments are performed on the dataset. The baseline model is multilabel Resnet50. We evaluate the results of the baseline model to decide whether to conduct a bad case study. Bad case study usually includes error type identification, case analysis, cause exploration, etc. If this operation is performed, the dataset is updated based on the results of the baseline model, thus completing an iterative closed loop. Through multiple iterations, the quality of the dataset is optimized. The core of label assignment rules is to handle the size of objects and the differences in features within classes to ensure the clarity and accuracy of labeling decisions. For the size of the target, a minimum size standard needs to be set. Targets whose area is less than 2% of the overall image area are usually not annotated. However, the rule is not absolute. Visibility principles should also be considered. Even if the target is small, if it can still be clearly identified, it should be labeled. When dealing with intraclass feature differences, the key is to determine the multidimensional characteristics of the class.

For example, the characteristics of a building range from size, shape, and roof structure to color and even its distribution in the area. In these aspects, as long as certain conditions are met, they should be marked. Such careful distinction ensures the accuracy and consistency of the annotation process. We used random sampling to divide the dataset. Random sampling ensures that the training and test sets are statistically representative. This approach avoids bias caused by uneven data distribution, allowing a fair assessment of the model's generalization ability in a variety of situations. In addition, the segmentation ratio of training set to test set is 2:1, which is consistent with that of the previous Bijie landslide dataset segmentation mode.

IV. ANALYSIS OF LANDSLIDE DETECTION RESULTS

A. Training Parameter Settings

The initialization methods for the two components included in the proposed Swin-BiLSTM-FS are mainly given as follows.

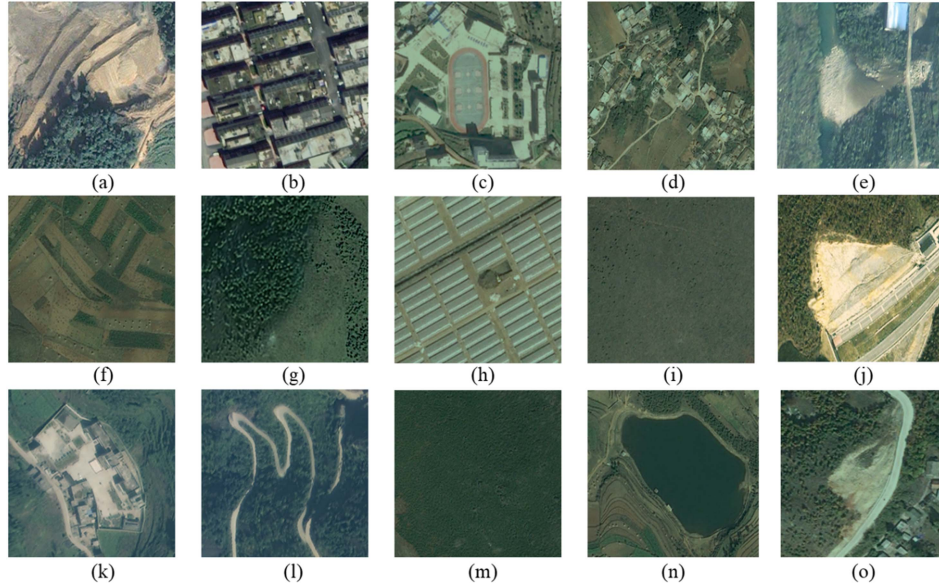


Fig. 10. Samples from the Bijie landslide dataset al.ng with their corresponding multiple object labels after supplementary annotation are provided. Each image has a spatial resolution of 0.8 m, and their scenes and object labels are described as follows. (a) Bare soil and tree. (b) Building and pavement. (c) Court and building. (d) Building, field, tree, and pavement. (e) Building, landslide, pavement, tree, and water. (f) Field. (g) Grass and tree. (h) Greenhouse. (i) Semiarid grassland. (j) Building, grass, landslide, pavement, and tree. (k) Building, field, pavement, and tree. (l) Pavement and tree. (m) Tree. (n) Building, field, tree, and water. (o) Building, landslide, pavement, and tree.

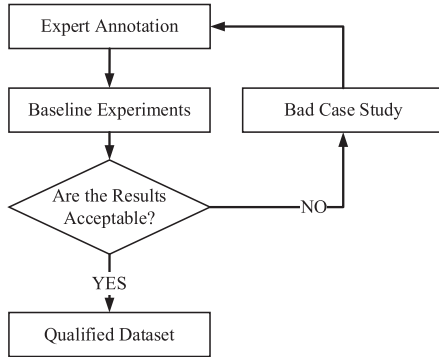


Fig. 11. Process of multilabel annotation iteration.

- 1) Swin transformer-tiny pretrained on the ImageNet dataset is used to initialize the feature extraction backbone [38].
- 2) The Glorot initialization method is used for initializing the weights of the single-unit linear layer, while the bias is initialized to zero.
- 3) The linear layers in BiLSTM (i.e., the weight matrices of the input gate, forget gate, cell gate, and output gate) are orthogonally initialized.

As for the biases' terms, the bias of the forget gate is typically initialized to 1 (to avoid forgetting too much information), and the biases of the other gates are set to 0. All parameters in the network are trainable and these weights are fine-tuned during the training phase.

Regarding the optimizer, we employ adaptive moment estimation (Adam) [44], with ϵ as a small constant to prevent division by zero and control the maximum size of updates; β_1 and β_2 are the hyperparameters of the exponential sliding average

that control the decay rate of the exponential sliding average of gradient. They are set as follows: $\epsilon = 1e - 8$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The learning rate is set to $1e - 4$, and the weight decay is set to $1e - 2$. Binary cross-entropy loss is employed for the loss function. In addition, training images are randomly flipped with a probability of 0.5 to enhance the generalization ability of model. The network was implemented using Pytorch and trained for 100 epochs on an NVIDIA GeForce RTX 3080 Ti 12 GB GPU with a truncation method to prevent overfitting. The training was stopped when the validation loss did not decrease significantly within five epochs. To balance the memory ability and training speed, the training batch size is set to 32. During training, an exponential moving average mechanism [27] was employed, which does not increase the performance of the model but improves the stability and generalization ability of the model. Different network structures and network parameters have different effects on the final results, with several experiments; the optimal basic parameter settings for each model are shown in Table II, and the input and output sizes for each model can be found in Table III.

B. Analysis of MLC Results

1) *Analysis of MLC Performance:* We evaluate the performance of Swin-BiLSTM-FS in landslide detection tasks from two perspectives. On the one hand, the performance of the proposed model on an MLC task, i.e., considering all categories; on the other hand, the performance of the model in the landslide category. For multiclassification problem, the performance of the MLC task is evaluated using macroaveraging, which denotes the category-based average of P , R , and F_1 . The proposed method allocates an equivalent weight to each category, which makes it

TABLE II
BASIC PARAMETERS OF EACH MODEL

Model	Parameters
Resnet50	layers: 50, activation: sigmoid
ViT	D: 768, layers: 12, activation: sigmoid
Swin transformer	C: 96, layer numbers: {2, 2, 6, 2}, activation: sigmoid
Ensemble model	Included models: Vgg19, ViT-Samll, DenseNet121, EfficientNet-B0, SEResNet50, activation: sigmoid
CA-Resnet50-BiLSTM	Step1: layers: 50, dilation: 2 Step2: kernel_size: 1, stride: 1, filters: 11 Step3: hidden_size: 56, units: 32, activation: sigmoid
ML-GCN	Step1: layers: 50 Step2: layers: 2, activation: sigmoid
ML-Decoder	Step1: layers: 50 Step2: layers: 1, heads: 8, activation: sigmoid
Swin-GNN	Step1: C: 96, layer numbers: {2, 2, 6, 2} Step2: layers: 2, activation: sigmoid
Swin-Decoder	Step1: C: 96, layer numbers: {2, 2, 6, 2} Step2: layers: 1, heads: 8, activation: sigmoid
Resnet50-BiLSTM-FS	Step1: layers: 50 Step2: hidden_size: 56, units: 32, activation: sigmoid
Swin-BiLSTM-FS (Multi)	Step1: C: 96, layer numbers: {2, 2, 6, 2} Step2: hidden_size: 56, units: 32, activation: sigmoid
Swin-BiLSTM-FS+DEM (Multi)	Step1: C: 96, layer numbers: {2, 2, 6, 2} Step2: hidden_size: 56, units: 32, activation: sigmoid
Swin-BiLSTM-FS (Single)	Step1: C: 96, layer numbers: {2, 2, 6, 2} Step2: hidden_size: 56, units: 32, activation: sigmoid
Swin-BiLSTM-FS+DEM (Single)	Step1: C: 96, layer numbers: {2, 2, 6, 2} Step2: hidden_size: 56, units: 32, activation: sigmoid

TABLE III
INPUT AND OUTPUT SIZES FOR EACH MODEL

Model	Input size	Output size
Resnet50	$(N \times 3 \times 224 \times 224)$	$(N \times 11)$
ViT	$(N \times 3 \times 224 \times 224)$	$(N \times 11)$
Swin transformer	$(N \times 3 \times 224 \times 224)$	$(N \times 11)$
Ensemble model	$(N \times 3 \times 224 \times 224)$	$(N \times 11)$
CA-Resnet50-BiLSTM	Step1: $(N \times 3 \times 224 \times 224)$ Step2: $(N \times 2048 \times 7 \times 7)$ Step3: $(N \times 11 \times 49)$	$(N \times 11)$
ML-GCN	Step1: $(N \times 3 \times 224 \times 224)$ Step2: $(N \times 11 \times 50)$	$(N \times 11)$
ML-Decoder	Step1: $(N \times 3 \times 224 \times 224)$ Step2: $(N \times 49 \times 2048)$	$(N \times 11)$
Swin-GNN	Step1: $(N \times 3 \times 224 \times 224)$ Step2: $(N \times 11 \times 50)$	$(N \times 11)$
Swin-Decoder	Step1: $(N \times 3 \times 224 \times 224)$ Step2: $(N \times 49 \times 768)$	$(N \times 11)$
Resnet50-BiLSTM-FS	Step1: $(N \times 3 \times 224 \times 224)$ Step2: $(N \times 11 \times 70)$	$(N \times 11)$
Swin-BiLSTM-FS (Multi)	Step1: $(N \times 3 \times 224 \times 224)$ Step2: $(N \times 11 \times 70)$	$(N \times 11)$
Swin-BiLSTM-FS+DEM (Multi)	Step1: $(N \times 6 \times 224 \times 224)$ Step2: $(N \times 11 \times 70)$	$(N \times 11)$
Swin-BiLSTM-FS (Single)	Step1: $(N \times 3 \times 224 \times 224)$ Step2: $(N \times 11 \times 70)$	$(N \times 2)$
Swin-BiLSTM-FS+DEM (Single)	Step1: $(N \times 6 \times 224 \times 224)$ Step2: $(N \times 11 \times 70)$	$(N \times 2)$

more meaningful in the case of unbalanced category distribution. CP stands for the average precision based on category, while CR stands for the average recall rate based on category. They enable a comprehensive assessment of model performance from the perspective of different categories. Their calculation is given as follows:

$$CP = \frac{\sum_{i=1}^N P_{ci}}{N} \quad (12)$$

$$CR = \frac{\sum_{i=1}^N R_{ci}}{N} \quad (13)$$

$$P = \frac{TP}{TP + FP} \quad (14)$$

$$R = \frac{TP}{TP + FN} \quad (15)$$

where TP, FP, and FN represent the quantities of correctly predicted positive class labels (true positives), incorrectly predicted positive class labels (false positives), and incorrectly predicted negative class labels (false negatives) in a category, respectively. P and R indicate the precision and recall rates of a category. Similarly, P_{ci} and R_{ci} represent the P and R of the i th category, respectively.

The category-based average of F_1 , denoted as CF_1 , is calculated as follows:

$$F_1 = \frac{2PR}{(P + R)} \quad (16)$$

$$CF_1 = \frac{\sum_{i=1}^N F_{1i}}{N} \quad (17)$$

where F_{1i} denotes the F_1 of the i th class, and N represents the total number of categories.

The example-based average precision (EP) and average recall (ER) [45] help us to provide insight into the network performance from a holistic perspective

$$EP = \frac{\sum_{j=1}^M P_{ej}}{M} \quad (18)$$

$$ER = \frac{\sum_{j=1}^M R_{ej}}{M} \quad (19)$$

where P_{ej} and R_{ej} represent the P and R of the j th instance, respectively, and M denotes the total number of instances.

In addition, some negative evaluation metrics should be compared to prove the effectiveness of the proposed method. We employ category-based missed detection (CMD) and category-based false alarm (CFA) numbers that clearly show the missed detections and false positives for each method

$$CMD = \sum_{i=1}^N MD_{ci} \quad (20)$$

$$CFA = \sum_{i=1}^N FA_{ci} \quad (21)$$

$$MD = FN \quad (22)$$

$$FA = FP \quad (23)$$

where MD_{ci} and FA_{ci} represent the MD and FA of the i th category, respectively.

To comprehensively assess the performance of the proposed model, the evaluation was conducted in this study from three perspectives.

- 1) In the MLC task, we compare the performance of the proposed method with traditional CNNs, transformer networks, and advanced MLC models to validate the effectiveness of the proposed method in MLC tasks.
- 2) Analyze the performance differences between the proposed method and other advanced technologies at each category level to demonstrate their advantages in different categories.
- 3) Focus on landslide categories to further confirm the effectiveness of the proposed method on landslide detection.

The Softmax function is replaced with the Sigmoid function to calculate the score for each category. After normalizing all scores, the probability of occurrence of each category can be obtained. A threshold of 0.5 is set to produce the predicted binary sequence.

The experimental results on the Bijie multilabel landslide dataset are shown in Table IV. In order to clearly compare the performance of different methods, we indicate the best results in bold and underline the suboptimal results. There are six difficult categories in the Bijie multilabel landslide dataset, accounting for 54.5% of all categories, which results in generally lower performance in category-based indices namely, CF_1 , CP , and CR . To demonstrate that the Swin transformer is a better choice as a feature extraction network, the feature extraction backbone network of ML-GCN [36] and ML-decoder [46] is replaced with the Swin transformer, naming the modified ML-GCN and ML-decoder as Swin-GNN and Swin-decoder, respectively. Compared with the original network, Swin-GNN performs better in CP and CR , and the overall performance is generally improved. Swin-decoder demonstrated improvements on all measures, exhibiting a conspicuous overall performance enhancement. Therefore, in subsequent experiments, we contrast Swin-GNN and Swin-decoder with our novel method, Swin-BiLSTM-FS, to further demonstrate the importance of BiLSTM.

In Table IV, the transformer models outperform CNNs in the task of multilabel image classification, balancing precision, and recall more effectively. Although ViT-small surpasses Resnet50 in some metrics, it tends to underperform in terms of EP due to its focus on capturing global context information. Compared with ViT-small, swin-tiny handles local features more effectively due to its window-based SA mechanism. The ensemble model effectively combines five neural networks to integrate the advantages of different networks. Without considering the model complexity, this ensemble method exhibits excellent performance on multilabel tasks, which is only lower than the proposed method. When dealing with complex multilabel problems, integrating strategies from different networks is also an effective way to improve the overall performance of the model. CA-Conv-BiLSTM makes progress in handling intercategory dependencies, with a maximum EP of 89.87% and a suboptimal CFA of 650. Swin-GNN achieves a more effective balance

TABLE IV
COMPARISON OF CLASSIFICATION PERFORMANCE ON THE BIJIE MULTILABEL LANDSLIDE DATASET

Method	Backbone	CF_1	CP	CR	EP	ER	CMD	CFA
Resnet50 [47]	-	46.09%	53.87%	44.20%	89.02%	81.36%	1270	861
ViT-Small [31]	-	47.50%	55.50%	46.50%	87.70%	83.70%	1218	849
Swin-Tiny [38]	-	48.40%	59.00%	47.30%	88.40%	83.90%	1199	748
Ensemble model [26]	-	<u>58.00%</u>	<u>64.70%</u>	<u>58.00%</u>	86.20%	84.70%	<u>956</u>	720
CA-Resnet50-BiLSTM [35]	Resnet50	43.59%	59.06%	41.17%	89.87%	79.67%	1339	<u>650</u>
ML-GCN [36]	Resnet50	47.86%	57.11%	43.62%	89.12%	78.76%	1283	746
ML-Decoder [46]	Resnet50	51.33%	58.10%	48.42%	87.00%	80.71%	1174	795
Swin-GNN	Swin-Tiny	50.13%	56.81%	48.23%	88.93%	84.63%	1178	835
Swin-Decoder	Swin-Tiny	52.79%	63.32%	50.15%	89.14%	<u>86.49%</u>	1135	661
Resnet50-BiLSTM-FS	Resnet50	49.26%	58.27%	46.03%	<u>89.66%</u>	79.16%	1228	750
Swin-BiLSTM-FS	Swin-Tiny	61.10%	68.70%	59.00%	89.16%	86.80%	933	612

The bold and solid values represent optimal indicators and other indicators respectively.

between P and R , as well as CMD and CFA . However, there is still room for improvement in the overall performance of both methods. Specifically, CA-Resnet50-BiLSTM is limited by the feature extraction ability based on Resnet50 and particularly struggles with balancing global and local information extraction. Meanwhile, the processing pressure of the BiLSTM network may not be effectively reduced by the strategy of compressing the feature maps on the channels. As for Swin-GNN, it might be constrained by the mismatch between content feature space and label space. That is to say, Swin-GNN possesses an advantage in handling interclass dependencies. However, if the correspondence between the feature space and label space is poor, its actual performance in practical applications might be lower than the advantages presented by the evaluation metrics. The Swin-decoder performs well on some metrics, but it typically has lower interpretability. In other words, it is difficult to explain why the model makes specific predictions, which might introduce a degree of uncertainty [45] and make it poorer than optimal.

Our proposed Swin-BiLSTM-FS model is superior to the other compared methods. Its highest F_1 of 61.10% among all methods outperforms Swin-GNN and Swin-Decoder by 10.07% and 8.31%, respectively, which highlights the importance of introducing BiLSTM and matching it to the feature extraction network. Meanwhile, CP and CR are also the highest with 68.70% and 59.00%, respectively, with significant performance improvement. Moreover, Swin-BiLSTM-FS surpasses Swin-tiny in all evaluation metrics, which further validates the essential role of the BiLSTM subnetwork. To elucidate the mechanism of the model, the feature extraction network of the SwinLSTM-FS-Swin-tiny is replaced with Resnet50 to obtain a new model, Resnet50-BiLSTM-FS. Compared to employing Resnet50 alone, the model exhibits superiority on most evaluation metrics, which demonstrates the effectiveness

of transforming feature maps into feature sequences and modeling label dependencies via BiLSTM. However, compared to the benchmark model Resnet50, the ER of Resnet50-BiLSTM-FS falls short. Compared with Resnet50-BiLSTM-FS, Swin-BiLSTM-FS improves CP , CR , CF_1 , and ER while stabilizing EP , which emphasizes the positive impact of Swin-tiny on model performance. Furthermore, Swin-BiLSTM-FS is optimal in CMD and CFA , which are 933 and 612, respectively. This shows the superiority of the proposed method on the MLC task. CA-Resnet50-BiLSTM owns suboptimal CFA , but the CMD of this method is the worst, which shows that CA-Resnet50-BiLSTM cannot effectively balance CFA and CMD . The good performance across various metrics of the proposed model is attributed to the complementary effects of the powerful feature extraction capability of Swin transformer and the relational inference capability of BiLSTM.

For each category in the Bijie multilabel landslide dataset, Fig. 12 provides a detailed display of R and F_1 score. In categories with a higher number of samples, such as building, field, landslide, pavement, and tree, the proposed Swin-BiLSTM-FS model outperforms other methods. In categories with fewer samples, our model performs well in classes, such as bare soil, court, grass, and water, but falls short in the categories of greenhouse and semiarid grassland. This phenomenon might be attributed to several reasons: First, the number of samples for greenhouse and semiarid grassland in the training dataset are 43 and 25, respectively, accounting for only 2.3% and 1.4% of the total training dataset. Second, for the greenhouse category, its color and shape are highly similar to some buildings, such as factory, which might lead to confusion for the Swin-BiLSTM-FS model. When the ensemble model makes classification decisions, it must pass the judgment of five networks. Therefore, its performance in detecting greenhouses is the best. Since the Swin-decoder has high complexity, it has good performance

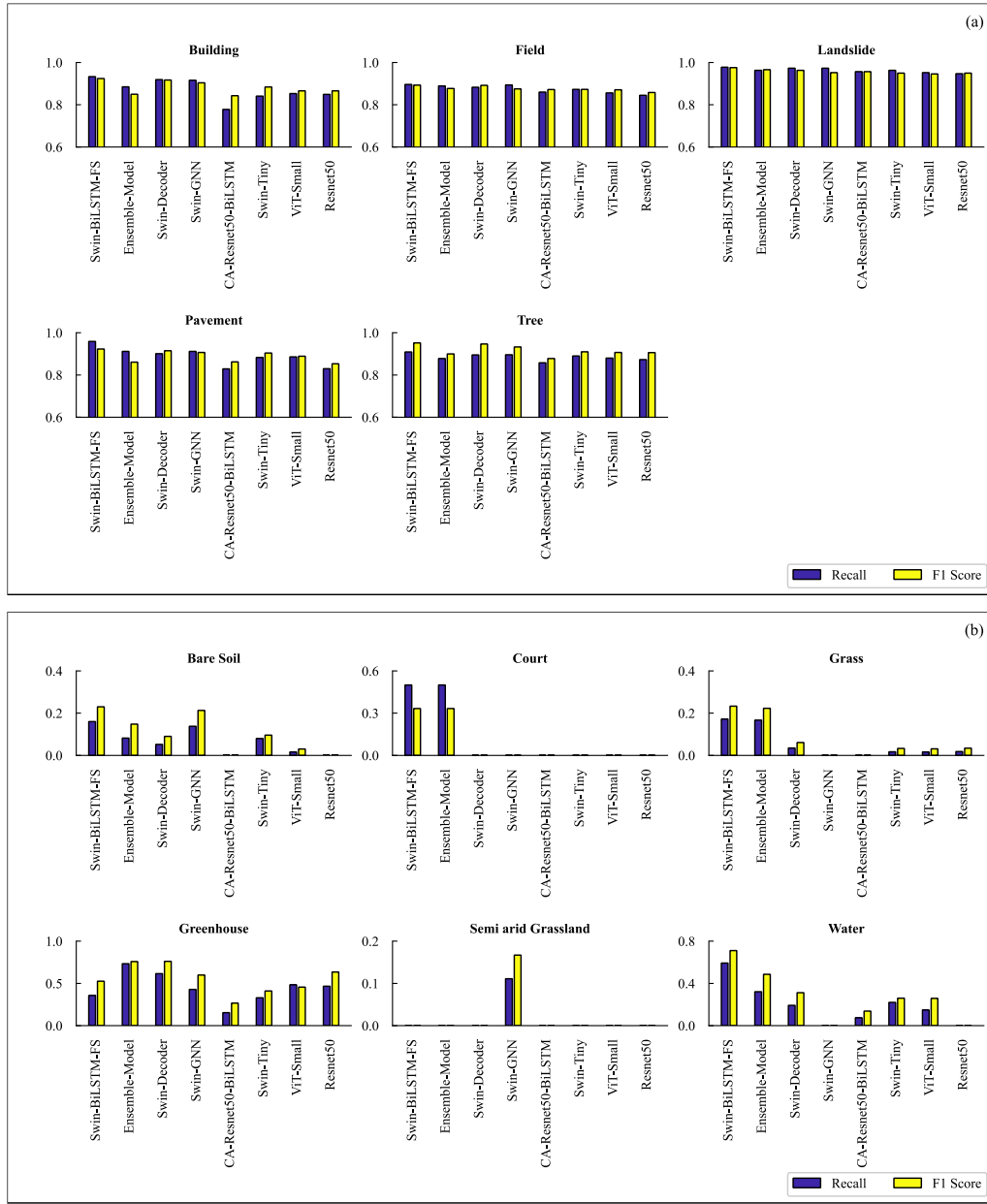


Fig. 12. Evaluation of per-class performance in recall and F_1 -score between Swin-BiLSTM-FS and other MLC methods (Swin-BiLSTM-FS, ensemble model, Swin-decoder, Swin-GNN, CA-Resnet50-BiLSTM, Swin-tiny, ViT-Small, and Resnet50). (a) Performance evaluation for frequent classes in the dataset. (b) Performance evaluation for rare classes.

in detecting greenhouses. The Swin-decoder performs best in this category, which may be attributed to its higher complexity. Third, for the semiarid grassland category, whose features lie between bare soil and grass, the Swin-BiLSTM-FS model may have difficulty in adequately recognizing its unique features and, therefore, classifying it accurately. In the situation where the content feature space and label space do not match, Swin-GNN introduces information outside the dataset by using pretrained word vectors, which may have a positive effect on improving model performance in some cases. Nevertheless, in terms of rare and common categories, our model achieved satisfactory results overall. Some prediction examples from the Bijie multi-label landslide dataset are shown in Table V, where our model

correctly predicts the majority of object labels. However, in Table V, there are some mistakes. In the third image, the greenhouse is incorrectly classified as a building; in the fourth image, the semiarid grassland is incorrectly classified as bare soil. These instances are consistent with our previous analysis.

2) *Analysis of Landslide Detection Performance:* In the MLC task, our model performs well and provides strong support for landslide detection. A comparison result for various methods in the landslides category is shown in Table VI, where the first two methods employ single-label image classification. It has been found that the integration of multisource data can enhance the accuracy of landslide detection. Compared to only utilizing optical remote sensing images, F_1 , P , R , and MD are

TABLE V
PREDICTION EXAMPLES FROM THE BIJIE MULTILABEL LANDSLIDE DATASET

Images in dataset					
Ground Truths	Field and tree	Building, field, and pavement	Building, field, greenhouse, and pavement	Landslide, pavement, semiarid grassland, and tree	Bare soil, building, field, tree, and water
Predictions	Field and tree	Building, field, and pavement	Building, field, <i>greenhouse</i> , and pavement	Landslide, pavement, bare soil , <i>semiarid grassland</i> , and tree	Bare soil, building, field, tree, and water

Bolded predictions indicate false positives, while *Italicized* predictions are false negatives.

The bold and solid values represent optimal indicators and other indicators respectively.

TABLE VI
COMPARISON OF LANDSLIDE DETECTION PERFORMANCE

Method	Label	F_1	P	R	MD	FA
Resnet50 [22]	Single	94.05%	95.95%	92.22%	20	10
Resnet50+DEM [22]	Single	94.88%	96.02%	93.77%	16	10
Resnet50+DEM+SCAM [22]	Single	<u>96.62%</u>	98.78%	94.55%	14	3
Ensemble Model [26]	Multi	96.59%	96.84%	96.34%	9	8
Resnet50 [47]	Multi	94.96%	95.21%	94.71%	14	12
ViT-Small [31]	Multi	94.64%	95.20%	94.01%	14	17
Swin-Tiny [38]	Multi	95.72%	96.24%	95.21%	12	10
CA-Resnet50-BiLSTM [35]	Multi	95.68%	95.68%	95.68%	11	11
Swin-GNN	Multi	95.16%	93.16%	97.25%	<u>7</u>	18
Swin-Decoder	Multi	96.32%	95.31%	<u>97.34%</u>	<u>7</u>	12
Swin-BiLSTM-FS	Single	96.17%	96.90%	95.46%	12	8
Swin-BiLSTM-FS+DEM	Single	96.21%	<u>97.40%</u>	95.04%	13	<u>7</u>
Swin-BiLSTM-FS+DEM	Multi	92.12%	93.25%	91.02%	23	17
Swin-BiLSTM-FS	Multi	97.56%	97.30%	97.83%	6	<u>7</u>

The bold and solid values represent optimal indicators and other indicators respectively.

all improved by combining optical remote sensing images and digital elevation models (DEMs). However, creating DEM data requires a substantial amount of a priori knowledge and human resources, which may not be the optimal choice. In addition, the matching between DEM and optical remote sensing data is difficult, and multisource data require higher model complexity. Therefore, we have introduced MLC into the field of landslide detection.

Compared with the single-label image classification, both F_1 and R are increased in the MLC strategy for Resnet50, and the MD and FA are better balanced. The reason is that the loss function of the MLC strategy contains the losses from all binary

classifiers. Although MLC does not directly model interlabel relationships, it indirectly considers the correlation between labels. This validates the effectiveness of further mining optical remote sensing information through MLC. However, P is decreased and FA is increased due to the insufficiency of Resnet50 in extracting global context information. Without considering the model complexity and the problems existing between heterogeneous data, Resnet50+DEM+SCAM achieved the highest P of 98.78% and the lowest FA of 3, which shows that this method has more stringent standards for detecting landslides. However, its performance on R is poor, which can lead to relatively high MD . The ensemble model uses a decision-level fusion strategy

to combine five different networks. This method requires the model to make multiple judgments when making decisions. Therefore, the model obtains a higher F_1 score and maintains the balance between P and R , so are MD and FA. However, each individual network is involved in the final decision, and the errors of each network are added to the overall model. This means that, although the ensemble model performs well overall, its performance is still limited by the performance of the included networks. Compared with Resnet50 based on the MLC method, all indicators of Swin-tiny are improved. Swin-tiny is capable of concurrently extracting global information and local information. This trait makes it ideal for feature extraction in MLC tasks.

In the landslide category, compared with Swin-tiny, both P and F_1 of Swin-GNN are declined, but R is increased. This is consistent with the performance manifestation in MLC tasks, suggesting that employing a GNN to model interlabel dependencies may not be the best strategy. Swin-decoder increases in both F_1 and R , while P decreases. Despite that the ML-decoder of Swin-decoder achieves significant progress in terms of R , the uncertainty in interclass modeling can reduce the P of landslide and increase the FA. The proposed Swin-BiLSTM-FS effectively balances local and global features in single-label classification tasks, which has advantages over other single-label methods. However, the single-label strategy does not consider landslide area topographic information. When faced with landslide detection with diverse spectral and spatial characteristics, the model will miss some real landslides with unclear characteristics. After combining DEM, the model adopts more stringent standards when making predictions. The model predicts a landslide only when the topographic features of the landslide are very obvious. This approach will reduce the number of instances where non-landslides are incorrectly labeled as landslides, thus reducing FA and improving accuracy. However, some real landslides will also be missed, increasing MD and reducing the R . In MLC, Swin-BiLSTM-FS+DEM has poor performance in detecting landslides. The reason is that DEM is mainly used to represent the elevation information of the surface and cannot reflect the characteristics of various types of ground objects. Therefore, in MLC, the fusion of DEM reduces the detection ability of the model for each category. At the same time, landslide area topographic information cannot be used to promote the performance of landslide detection. In addition, the resolution of the DEM in the dataset is 2 m, which also brings difficulties to the MLC task. Our proposed model, Swin-BiLSTM-FS, performs best in multilabel tasks and surpasses other methods in the most critical category—landslide, with the optimal F_1 , R , and MD. The reason is that it efficiently models dependencies between land object types while ensuring the robustness of feature extraction. The model utilizes the inherent connection between the types of land objects presented at landslide sites and the landslides, thereby enhancing landslide detection performance. Although Resnet50+DEM+SCAM and Swin-BiLSTM-FS+DEM also have good performance in FA, they all sacrifice MD. Moreover, in the task of detecting landslides, the consequences of a missed detection may be more serious than a false alarm. Therefore,

TABLE VII
COMPARISON OF THE TIME TO PREDICT A SINGLE IMAGE (MS), THE
PARAMETER VOLUME (M), AND FLOPS (GMAC) OF THE MODEL

Method	Times (ms)	Params (M)	FLOPs (GMac)
Resnet50 [47]	199.47	23.53	4.13
ViT-Small [31]	101.37	21.67	4.25
Swin-Tiny [38]	110.71	27.51	4.50
Ensemble Model [26]	3042.91	263.88	40.68
CA-Resnet50-BiLSTM [35]	258.42	23.58	4.13
Swin-GNN	2725.89	27.73	4.38
Swin-Decoder	115.69	33.63	4.50
Swin-BiLSTM-FS	112.70	28.17	4.37

Swin-BiLSTM-FS can effectively reduce MD while maintaining low FA, reflecting its reliability in practical applications.

An enhancement in model performance often sacrifices computational efficiency [48]. Table VII displays the number of parameters for each model, the time required to predict a single image and floating point operations (FLOPs). Comparing parameter quantities, prediction times and FLOPs help to reach the tradeoff between model performance and computational efficiency, thereby enabling a more thorough evaluation of model in practical application. Resnet50 achieves a balance between performance and efficiency under less parameters and shorter computational time. The FLOPs of ViT-small are slightly higher than ResNet50, but it has fewer parameters and the shortest processing time of 101.37 ms. These differences reflect the characteristics of CNNs and transformer networks. The processing speed of Swin-tiny is similar to ViT-small. It enhances its feature capture capability by moderately increasing the number of parameters. The increment of FLOPs for Swin-tiny is due to the additional computational overhead caused by its hierarchical and moving window SA mechanisms. However, considering its advantages in feature expression, these additional computational overheads are reasonable and effective. The ensemble model improves detection performance by significantly increasing the number of parameters and complexity of the model. While obtaining higher accuracy, it sacrifices the efficiency and maintainability of the model, resulting in limitations in practical applications. The parameter quantity of CA-Resnet50-BiLSTM is similar to that of Resnet50, but its computation time is longer. The reason is that the additional category attention layer and BiLSTM layer increase the model complexity. Swin-GNN owns a moderate parameter quantity, and its calculation time reaches 2725.89 ms, which suggests that GNN subnetworks may add significant computational costs. The number of parameters and FLOPs of Swin-GNN are similar to Swin-tiny, but they are computationally significantly different. Because Swin-GNN combines a GNN, a large number of sparse matrix operations are required. These operations are not as efficient as the dense matrix operations in Swin-tiny on many hardware. The reason

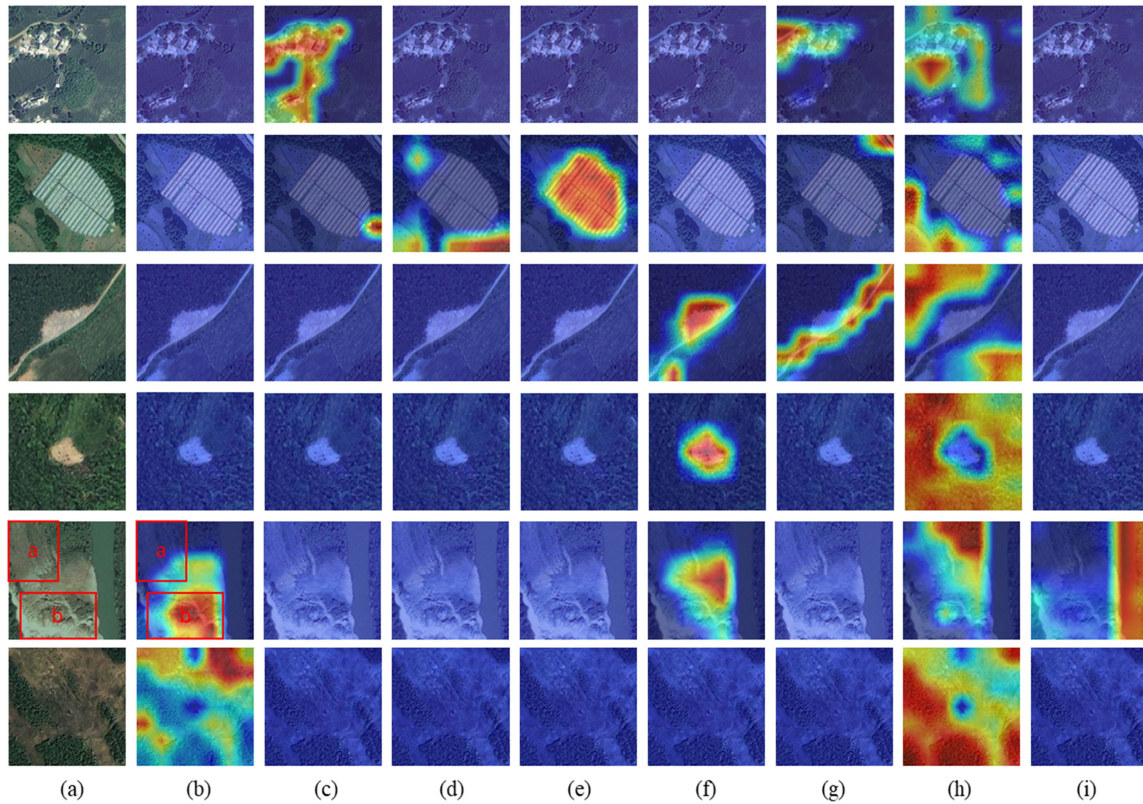


Fig. 13. Example class attention maps of images. (a) is a sample from the Bijie landslide dataset. (b)–(i), respectively, display the class attention maps for bare soil, building, field, greenhouses, landslide, pavement, tree, and water in samples.

is that the irregular memory access pattern of sparse matrices makes it difficult to efficiently utilize hardware resources. As previously analyzed, adopting the ML-GCN strategy may not be the optimal choice for MLC tasks. Swin-decoder has a larger parameter quantity because the multilabel decoder increases the complexity of the model. However, its computation time is shorter, which is approximate to that of Swin-tiny and our proposed model. Our proposed model, Swin-BiLSTM-FS, is closed Swin-tiny in terms of parameter quantity and computation time, achieving faster prediction time under moderate parameter quantity. The design of the FR module and CRP module achieves a balance between computational complexity and model complexity. Therefore, compared with Swin-tiny, the proposed method reduces FLOPs while moderately increasing the number of parameters. This demonstrates that the proposed model reaches a good balance between performance and efficiency.

C. Analysis of Grad-CAM Visualization

In this section, we delve into a qualitative analysis of the performance of our proposed Swin-BiLSTM-FS model on key categories, such as landslides and pavements. We utilize the Grad-CAM visualization technique, which helps identify areas in an image that play a key role in the prediction of a specific category, thereby understanding how the network pays attention to specific areas of the input image. This aids in validating the efficacy of MLC in landslide detection and further elucidating phenomena observed in the quantitative analysis.

In the Bijie multilabel landslide dataset, the Grad-CAM is performed on the output of the norm layer before the GAP of the proposed model. The class attention maps of examples from the Bijie multilabel landslide dataset are illustrated in Fig. 13. Each row represents a different example, with column (a) showing the original image and columns (b)–(i), respectively, representing the CAMs of bare soil, building, field, greenhouse, landslide, pavement, tree, and water. The observation reveals that the CAM for corresponding categories highlights the discriminative areas of the positive category. For instance, when a category (such as landslide, pavement, and water body) does not exist in the original image, the activation mapping of the corresponding category has almost no activated area. In Fig. 13, the object labels of the second row image are building, field, greenhouse, and tree, all of which are significantly activated. The nonexistent categories (such as landslide, pavement, and water) have almost no activation. As can be seen from the third scene of Fig. 13, the CAM maps of landslide, pavement, and tree are obvious, leading to their prediction as positive instances. This is consistent with the high *CR* and *CP* values observed in our quantitative analysis, indicating that our model can effectively identify and distinguish different categories. In the fifth scene in Fig. 13, the proposed model is not complete for the attention area on the bare soil. The reason is that bare soil is a rare category with only 211 samples in total, which makes it difficult for the model to accurately learn its unique features. Bare soil shows greater variability in spectral and spatial characteristics than the landslide category. The cause of its formation also has higher uncertainty. The proposed model

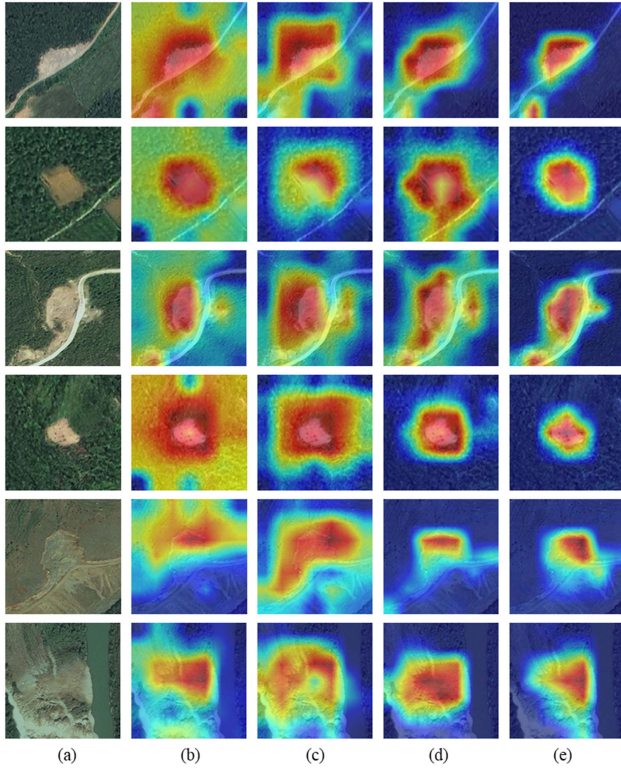


Fig. 14. CAM for landslide categories. (a) Samples in Bijie landslide dataset. (b)–(e) Landslide attention maps of Swin-tiny, Swin-GNN, Swin-decoder, and Swin-BiLSTM-FS, respectively.

learns the landslide area topographic information, which can effectively reduce the misjudgment of bare soil as a landslide when detecting landslides. The bare soil of the fifth scene in Fig. 13 contains two areas, a and b. The bare soil characteristics of areas a and b are different, making it difficult for the model to accurately identify the two areas at the same time. The proposed model did not detect bare soil in area a, but successfully detected bare soil in area b. And it can effectively distinguish between bare soil and landslide areas. In the sixth scene of Fig. 13, the bare soil area is accurately noticed by the proposed method.

In addition, the CAM of different methods on the landslide category is illustrated in Fig. 14. In addition to the proposed model, the Grad-CAM visualization results of the three methods of Swin-tiny, Swin-GNN, and Swin-decoder are all derived from the last norm layer of their respective networks. These layers are the final norm layer of Swin-tiny, the norm layer after the Swin transformer for Swin-GNN, and the norm layer before the ML-decoder classification header. It is shown that the attention of Swin-tiny is more distinct at the location of the landslide, but the overall attention range is too large. On this basis, the addition of the GNN and ML-decoder subnetworks has resulted in an overall area of attention that is closer to the real region. The R of Swin-GNN is better than that of Swin-tiny, which is shown in Table V. As illustrated in Fig. 14, this is because more real regions are included in the attention region. However, the P is reduced, which is due to the attention region that expands too much, which results in that information from other categories being included. The R of the Swin-decoder is better than that of

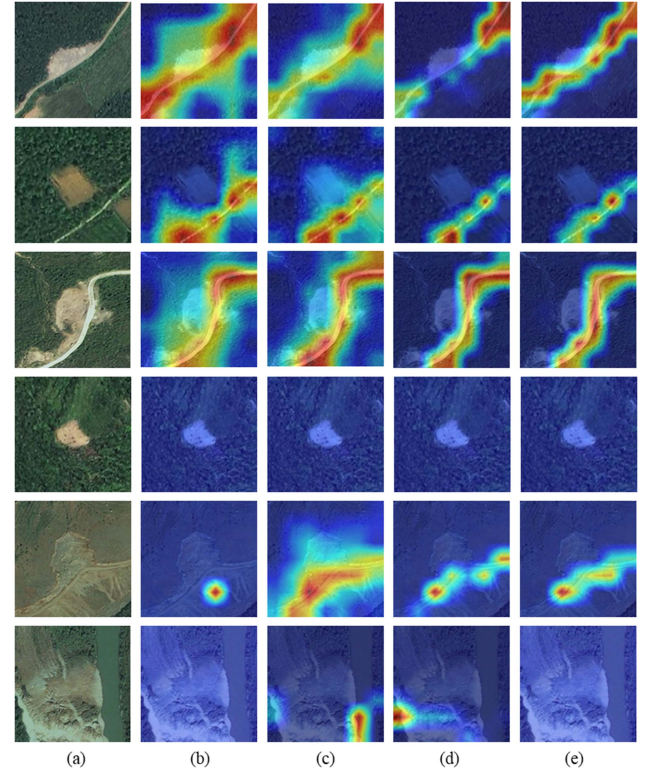


Fig. 15. CAM for pavement categories. (a) Samples in Bijie landslide dataset. (b)–(e) Pavement attention maps of Swin-tiny, Swin-GNN, Swin-decoder, and Swin-BiLSTM-FS, respectively.

Swin-tiny, as shown in Fig. 14, which is because the attention regions are more focused. However, P is slightly lower because of the uncertainty in the modeling of the labels, which makes for slight changes in the attention regions. The attention of our proposed Swin-BiLSTM-FS method on the landslide category is highly concentrated and accurate, further validating the effectiveness of our proposed method in landslide detection. These results confirm that our model has high predictive performance in the task of landslide detection.

Similarly, the CAMs of various methods on the pavement category are illustrated in Fig. 15. It is shown that, in the challenging scenarios, the recognition of Swin-tiny falls not satisfactory. For instance, for the fifth scenario in the fifth row of Fig. 15, there is no remarkable feature difference between the pavement and other categories. Under such a situation, Swin-tiny cannot distinguish well between the pavement and other categories. After the introduction of additional knowledge by Swin-GNN, it fully notices the pavement. However, its attention activation area includes both the landslide and pavement categories. This indicates that it cannot efficiently differentiate between the features of landslides and pavements. In the sixth scenario of Fig. 15, even though the actual label does not encompass the pavement, Swin-decoder produces activation of the pavement category in certain areas. This suggests that this model may have some degree of overfitting. Overall, our proposed method, Swin-BiLSTM-FS, can effectively distinguish the pavement and landslide categories from others. The Grad-CAM visualizations

for the pavement and landslide categories further validate the robustness and effectiveness of the proposed model.

V. DISCUSSION

In this study, we systematically evaluate the performance of the proposed MLC network Swin-BiLSTM-FS in landslide detection in remote sensing images. On the proposed dataset, the experiments included both quantitative and qualitative evaluation dimensions. In quantitative evaluation, the performance of the model on the MLC task is evaluated through a set of standardized metrics. Among them, category-based metrics include CP , CR , CF_1 , CMD , and CFA , while example-based metrics include EP and ER . The P , R , F_1 , MD , and FA of the model on landslide category are also evaluated. Through comparative analysis and ablation studies, we further confirmed the effectiveness of the MLC strategy, as well as the advantages of the model compared with the existing advanced methods and its high efficiency in data processing. In Table IV, Swin-BiLSTM-FS shows better performance than other methods in the MLC task. Swin-BiLSTM-FS achieved the highest CF_1 score of 61.1%, which is 3.1% higher than the next best method of 58.0%. Except for EP , our method is better than the comparison methods in all of the other evaluation indicators. As shown in Table VI, in the landslide detection task, Swin-BiLSTM-FS achieves the highest F_1 score of 97.56%, which is 0.94% points higher than the suboptimal method Resnet50+DEM+SCAM. And the proposed method achieves a better balance between P and R , so are MD and FA . Furthermore, the performance of Swin-BiLSTM-FS in landslide detection using only single-modal images surpasses Resnet50+DEM+SCAM, which uses multimodal data. Its F_1 score and R increased by 0.94% and 3.28%, respectively. To evaluate the efficiency of the proposed model, we compared the number of parameters, single-scene prediction time, and computational complexity. As shown in Table VII, Swin-BiLSTM-FS slightly increases the number of parameters compared with Swin-Tiny while reducing FLOPs. It achieves faster prediction times with a moderate number of parameters, which shows that the model achieves a good balance between performance and efficiency.

In the qualitative evaluation, Grad-CAM is used to display the attention areas of the Swin-BiLSTM-FS network on different categories to describe the process of model decisions. Compared with state-of-the-art MLC methods on Grad-CAM to evaluate the attention degree of Swin-BiLSTM-FS on important categories, such as landslide and road, the Grad-CAM visualization results in Fig. 13 show that Swin-BiLSTM-FS can accurately focus on specific areas of different categories. As shown in Figs. 14 and 15, compared with other methods, Swin-BiLSTM-FS has more concentrated attention areas on landslide and road. These experimental results demonstrate the advantages of Swin-BiLSTM-FS in visual interpretability and highlight its concentration and accuracy in identifying important features.

In summary, Swin-BiLSTM-FS has excellent performance in MLC and landslide detection tasks. Meanwhile, its advantages are demonstrated in terms of the prediction time and visual interpretability. The application potential of Swin-BiLSTM-FS

is shown in remote sensing image classification and detection tasks.

VI. CONCLUSION

In this article, we first treat landslide detection as an MLC task and propose an MLC network based on Swin-BiLSTM-FS, which uses the MLC method to learn the connection between landslides and ground object categories to improve landslide detection performance. This network mainly consists of an FR module and a CRP module. Specifically, the FR module is responsible for capturing fine-grained high-level semantic features from the original image and transforming them into feature sequences. The CRP module, subsequently, processes these feature sequences, mapping them into feature vectors, and feeding them into the BiLSTM to model interdependencies between labels. Finally, the network predicts multiple labels for the image. However, training MLC models requires high-quality MLC datasets, and there is currently no open-source high-quality multilabel landslide dataset. Therefore, this article expands the Bijie landslide dataset and adds multilabel annotation. The results demonstrated the effectiveness of our model both qualitatively and quantitatively. The performance of the proposed method exceeds the existing methods in both MLC tasks and landslide detection. Moreover, Grad-CAM visualization results further substantiated the accuracy of the attention regions in our proposed method. In future work, we plan to extend this network to the weakly supervised semantic segmentation domain and fine-tune the model to enhance its generalizability on specific categories.

REFERENCES

- [1] IPCC, "Climate change 2021: The physical science basis—Working group I contribution to the sixth assessment report of the intergovernmental panel on climate change," Cambridge Univ. Press, Cambridge, U.K., Rep. AR6, 2021.
- [2] K. Sassa et al., "Landslide risk evaluation and hazard zoning for rapid and long-travel landslides in urban development areas," *Landslides*, vol. 1, no. 3, pp. 221–235, Sep. 2004.
- [3] C. Lissak et al., "Remote sensing for assessing landslides and associated hazards," *Surv. Geophys.*, vol. 41, no. 6, pp. 1391–1435, Nov. 2020.
- [4] S. Chen et al., "Multi-source remote sensing based accurate landslide detection leveraging spatial-temporal-spectral feature fusion," *J. Comput. Res. Develop.*, vol. 57, no. 9, pp. 1877–1887, Nov. 2020.
- [5] C. Niu, H. Zhang, W. Liu, R. Li, and T. Hu, "Using a fully polarimetric SAR to detect landslide in complex surroundings: Case study of 2015 Shenzhen landslide," *ISPRS J. Photogramm. Remote Sens.*, vol. 174, pp. 56–67, Apr. 2021.
- [6] Y. Wang et al., "On the applicability of satellite SAR interferometry to landslide hazards detection in hilly areas: A case study of Shuicheng, Guizhou in Southwest China," *Landslides*, vol. 18, no. 7, pp. 2609–2619, Jul. 2021.
- [7] Y. N. Lin, Y.-C. Chen, Y.-T. Kuo, and W.-A. Chao, "Performance study of landslide detection using multi-temporal SAR images," *Remote Sens.*, vol. 14, no. 10, May 2022, Art. no. 2444.
- [8] X. Shao and C. Xu, "Earthquake-induced landslides susceptibility assessment: A review of the state-of-the-art," *Natural Hazards Res.*, vol. 2, no. 3, pp. 172–182, Sep. 2022.
- [9] A. Mohan, A. K. Singh, B. Kumar, and R. Dwivedi, "Review on remote sensing methods for landslide detection using machine and deep learning," *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 7, Jul. 2021, Art. no. e3998.
- [10] L. Long, F. He, and H. Liu, "The use of remote sensing satellite using deep learning in emergency monitoring of high-level landslides disaster in Jinsha river," *J. Supercomput.*, vol. 77, no. 8, pp. 8728–8744, Aug. 2021.
- [11] S. Sreelakshmi, V. S. S. Chandra, and E. Shaji, "Landslide identification using machine learning techniques: Review, motivation, and future prospects," *Earth Sci. Inform.*, vol. 15, no. 4, pp. 2063–2090, Dec. 2022.

- [12] W. Jiang et al., "Deep learning for landslide detection and segmentation in high-resolution optical images along the Sichuan-Tibet transportation corridor," *Remote Sens.*, vol. 14, no. 21, Nov. 2022, Art. no. 5490.
- [13] Y. Chen et al., "Susceptibility-guided landslide detection using fully convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 998–1018, 2023.
- [14] Z. Lv, H. Huang, W. Sun, M. Jia, J. A. Benediktsson, and F. Chen, "Iterative training sample augmentation for enhancing land cover change detection performance with deep learning neural network," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: [10.1109/TNNLS.2023.3282935](https://doi.org/10.1109/TNNLS.2023.3282935).
- [15] Z. Lv, P. Zhong, W. Wang, Z. You, and N. Falco, "Multiscale attention network guided with change gradient image for land cover change detection using remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Apr. 2023, Art. no. 2501805.
- [16] Z. Lv, P. Zhong, W. Wang, Z. You, J. Benediktsson, and C. Shi, "Novel piecewise distance based on adaptive region key-points extraction for LCCD with VHR remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 5607709.
- [17] Z. Han, Z. Fang, Y. Li, and B. Fu, "A novel Dynahead-Yolo neural network for the detection of landslides with variable proportions using remote sensing images," *Front. Earth Sci.*, vol. 10, Jan. 2023, Art. no. 1077153.
- [18] S. Yang et al., "Automatic identification of landslides based on deep learning," *Appl. Sci.*, vol. 12, no. 16, Aug. 2022, Art. no. 8153.
- [19] A. Dikshit, R. Sarkar, B. Pradhan, S. Segoni, and A. M. Alamri, "Rainfall induced landslide studies in Indian Himalayan region: A critical review," *Appl. Sci.*, vol. 10, no. 7, Apr. 2020, Art. no. 2466.
- [20] M. Cordts et al., "The Cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2016, pp. 3213–3223.
- [21] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 21–37.
- [22] S. Ji, D. Yu, C. Shen, W. Li, and Q. Xu, "Landslide detection from an open satellite imagery and digital elevation model dataset using attention boosted convolutional neural networks," *Landslides*, vol. 17, no. 6, pp. 1337–1352, Jun. 2020.
- [23] K. Hacıfendioglu, G. Demir, and H. B. Basaga, "Landslide detection using visualization techniques for deep convolutional neural network models," *Natural Hazards*, vol. 109, no. 1, pp. 329–350, Oct. 2021.
- [24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.
- [25] H. Wang et al., "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 111–119.
- [26] Y. Li, P. Wang, Q. Feng, X. Ji, D. Jin, and J. Gong, "Landslide detection based on shipborne images and deep learning models: A case study in the Three Gorges Reservoir Area in China," *Landslides*, vol. 20, no. 3, pp. 547–558, Mar. 2023.
- [27] M. Zhang, and Z. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
- [28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 2261–2269.
- [29] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, Long Beach, CA, USA, 2019, pp. 6105–6114.
- [30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 7132–7141.
- [31] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9840.
- [32] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, and A. Doulamis, "A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 7, pp. 3299–3307, Jul. 2023.
- [33] S. H. Khan, X. He, F. Porikli, and M. Bennamoun, "Forest change detection in incomplete satellite images with deep neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5407–5423, Sep. 2017.
- [34] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2016, pp. 2285–2294.
- [35] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 188–199, Mar. 2019.
- [36] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 5172–5181.
- [37] Y. Wang, Y. Xie, L. Fan, and G. Hu, "STMG: Swin transformer for multi-label image recognition with graph convolution network," *Neural Comput. Appl.*, vol. 34, no. 12, pp. 10051–10063, Jun. 2022.
- [38] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.
- [39] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, "A review of relational machine learning for knowledge graphs," *Proc. IEEE*, vol. 104, no. 1, pp. 11–33, Jan. 2016.
- [40] Y. Hua, L. Mou, and X. X. Zhu, "Relation network for multilabel aerial image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4558–4572, Jul. 2020.
- [41] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017.
- [42] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2019, pp. 8026–8037.
- [43] J. Ye, J. He, X. Peng, W. Wu, and Y. Qiao, "Attention-driven dynamic graph convolutional network for multi-label image recognition," in *Proc. 16th Eur. Conf. Glasgow, U.K.*, 2020, pp. 649–665.
- [44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, 2015, pp. 1–15.
- [45] X. Wu and Z. Zhou, "A unified view of multi-label performance measures," in *Proc. 34th Int. Conf. Mach. Learn.*, Sydney, Australia, 2017, pp. 1–14.
- [46] T. Ridnik, G. Sharir, A. Ben Cohen, E. B. Baruch, and A. Noy, "ML-decoder: Scalable and versatile classification head," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Nashville, TN, USA, 2023, pp. 32–41.
- [47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.
- [48] Z. Wang et al., "Land-sea target detection and recognition in SAR image based on non-local channel attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5237316.



Yongxin Li was born in Sichuan, China, in 1998. He received the B.S. degree in information engineering from the Chengdu University of Technology, Chengdu, China, in 2021. He is currently working toward the M.S. degree in electronic information with the School of Physics and Electronic Information, Yunnan Normal University, Kunming, China.

His research interests include synthetic aperture radar image processing and deep learning.



Zhihui Xin was born in Shanxi, China. She received the B.S. and Ph.D. degrees in electrical engineering from Xidian University, Xi'an, China, in 2012 and 2017, respectively.

In 2017, she joined the School of Physics and Electronic Information Technology, Yunnan Normal University, Kunming, China, where she is currently an Associate Professor. Her research interests include space-time adaptive processing, sea clutter modeling and simulation, and maritime target detection.



Guisheng Liao (Senior Member, IEEE) was born in Guilin, China. He received the B.S. degree in mathematics from Guangxi University, Nanning, China, in 1985, and the M.S. and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1990 and 1992, respectively.

He was a Senior Visiting Scholar with The Chinese University of Hong Kong, Hong Kong. He is currently a Professor with Xidian University, where he is also the Dean of the School of Electronic Engineering.

His research interests include synthetic aperture radar (SAR), space-time adaptive processing, SAR ground moving target indication, and distributed small satellite SAR system design.

Dr. Liao is a member of the National Outstanding Person and the Cheung Kong Scholars in China.



Mengting Yuan was born in Yunnan, China, in 1998. She received the B.S. degree in information engineering in 2021 from Yunnan Normal University, Kunming, China, where she is currently working toward the M.S. degree in electronic information with the School of Physics and Electronic Information.

Her research interests include synthetic aperture radar image processing and image change detection.



Penghui Huang (Member, IEEE) was born in Jiangxi, China. He received the B.S. and Ph.D. degrees in electrical engineering from Xidian University, Xi'an, China, in 2012 and 2017, respectively.

In 2017, he joined the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China, where he is currently an Assistant Professor. His research interests include spaceborne radar system design, array signal processing, synthetic aperture radar (SAR)/inverse SAR imaging, and time-frequency analysis.