ShapeFormer: A Shape-Enhanced Vision Transformer Model for Optical Remote Sensing Image Landslide Detection

Pengyuan Lv, *Member, IEEE*, Lusha Ma, Qiaomin Li, and Fang Du

***Abstract*—Landslides pose a serious threat to human life, safety, and natural resources. Remote sensing images can be used to effectively monitor landslides at a large scale, which is of great significance for pre-disaster warning and post-disaster assistance. In recent years, deep learning-based methods have made great progress in the field of remote sensing image landslide detection. In remote sensing images, landslides display a variety of scales and shapes. In this article, to better extract and keep the multiscale shape information of landslides, a shape-enhanced vision transformer (ShapeFormer) model is proposed. For the feature extraction, a pyramid vision transformer (PVT) model is introduced, which directly models the global information of local elements at different scales. To learn the shape information of different landslides, a shape feature extraction branch is designed, which uses the adjacent feature maps at different scales in the PVT model to improve the boundary information. After the feature extraction step, a decoder with deconvolutional layers follows, which combines the multiple features and gradually recovers the original resolution of the combined features. A softmax layer is connected with the combined features to acquire the final pixel-wise result. The proposed ShapeFormer model was tested on two public datasets—the Bijie dataset and the Nepal dataset—which have different spectral and spatial characteristics. The results, when compared with those of some of the state-of-the-art methods, show the potential of the proposed method for use with multisource optical remote sensing data for landslide detection.**

***Index Terms*—Landslide detection, remote sensing image, vision transformer (ViT).**

I. Introduction

LANDSLIDES refer to the natural phenomenon of the land cover on a slope sliding down under the effect of natural or human activities such as rainfall, earthquake, flooding, groundwater activity, or destruction of forest [1]. They are one of the most common natural disasters in mountainous areas and represent a serious threat to people's lives and property, as well as the surrounding natural environment [2], [3]. Therefore, it is crucial to identify landslide areas quickly and accurately for postdisaster reconstruction and predisaster warning [4]. The remote sensing technique provides a new perspective for the rapid monitoring of landslide occurrences in large areas and can reduce the cost and improve the efficiency, compared with the traditional geological investigation [5], [6].

According to the data source, the landslide detection task can be achieved through the use of optical data [7], [8], [9], thermal infrared data [10], synthetic aperture radar data [11], [12], LiDAR data [13], [14], or multisource data, such as geographic information system (GIS) and digital elevation model data [15], [16]. Among the above data sources, optical remote sensing images have the advantage of being able to provide abundant spectral information and a high spatial resolution, which can help to distinguish landslides from other land-cover types [17]. The optical remote sensing landslide detection techniques can be classified into traditional methods and deep learning-based methods. The traditional methods typically involve designing features according to prior knowledge and then using a machine learning algorithm to identify the landslides. For example, Zhao et al. [18] used vegetation and soil indices to distinguish landslides in Landsat 8 data. Chen et al. [19] used a morphological operation to extract the spatial features of landslides. Object-oriented analysis can also be introduced to improve the boundaries of the landslide detection results [20]. For example, Li et al. [21] combined object-oriented analysis and a machine-learning algorithm to process LiDAR data for the identification of forested landslides. Rau et al. [22] proposed a semiautomatic object-oriented landslide detection method for multisource data. Multitemporal data can also be used for landslide detection, through considering both the short- and long-term change features. Li et al. [23] used the change detection technique for the landslide detection task. In [24], a random field model was used to improve the boundaries of the results. Travelletti et al. [25] used high-resolution time-series optical imagery to monitor the continuous movement of landslides. Rossi et al. [26] studied the performance of multitemporal unmanned aerial vehicle (UAV) images for landslide recognition. Although the traditional methods have acquired good results in landslide detection, they are still faced with some challenges for more accurate detection. First, landslides display various characteristics, because of the

Manuscript received 30 December 2022; revised 22 February 2023; accepted 4 March 2023. Date of publication 7 March 2023; date of current version 22 March 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42001307 and in part by the Natural Science Foundation of Ningxia under Grant 2022AAC03053. *(Corresponding author: Pengyuan Lv.)*

Pengyuan Lv and Fang Du are with the School of Information Engineering, Ningxia University, Yinchuan 750021, China (e-mail: lpydtc@126.com; dfang@nxu.edu.cn).

Lusha Ma is with the School of Information Engineering, Ningxia University, Yinchuan 750021, China, and also with the School of Information and Media, Yinchuan University of Energy, Yinchuan 750199, China (e-mail: 12021131674@stu.nxu.edu.cn).

Qiaomin Li is with the Ningxia Institute of Remote Sensing Survey, Yinchuan 750021, China (e-mail: nxlqm@foxmail.com).

Digital Object Identifier 10.1109/JSTARS.2023.3253769

effect of different geological environments, weather, hydrology, and other factors. As a result, it is difficult to select and design appropriate features and classifiers under different conditions. Second, the development of remote sensing sensors also brings the problem of data heterogeneity, which requires algorithms that are more robust and have a good generalization ability.

As a data-driven method, deep learning has gradually achieved some promising results in remote sensing image processing applications, such as classification [27], [28], segmentation [29], [30], detection [31], and super-resolution [32]. Deep learning-based methods have also shown their potential for remote sensing image landslide detection [33], [34]. Convolutional neural networks (CNNs), as representative deep learning models, have the ability to hierarchically extract multiple features of the images, which is of benefit for optical remote sensing image landslide detection. Wang et al. [35] studied the performance of different convolution operations (i.e., 1-D, 2-D, and 3-D) in CNN models for landslide susceptibility mapping, based on multisource remote sensing and GIS data. High-level features from 16 factors were then automatically extracted based on the proposed method. Ghorbanzadeh et al. [36] used Dempster-Shafer theory to combine the prediction results of CNN models trained on different data sources, to improve the accuracy of landslide detection. Ji et al. [37] proposed a 3-D CNN model with spectral and spatial attention for high-resolution remote sensing image landslide detection. The attention mechanism is used to pay more attention to the important parts in the channel and spatial dimensions, to better recognize landslides. Su et al. [38] proposed an encoder–decoder architecture called LanDCNN, where ResNet-50 [39] is used in the encoder for the extraction of deep features, and the decoder of U-Net [40] follows to recover the original resolution of the imagery. Liu et al. [41] analyzed the performance of a traditional CNN model, ResNet, and DenseNet [42] for landslide detection with remote sensing imagery and conditioning factors. The authors made the conclusion that the combination of DenseNet with remote sensing imagery and conditioning factors could acquire the best results. Gao et al. [43] also studied the performance of DenseNet for landslide detection and proposed the FC-DenseNet model, which improves the feature extraction ability and model efficiency with the help of dense connection and bottleneck layers. Chen et al. [44] introduced a pretrained CNN for feature fusion and landslide detection. Xia et al. [45] introduced atrous spatial pyramid pooling to improve the landslide detection network's ability to preserve details in high-resolution remote sensing images. CNN models can also be combined with multitemporal images and change detection methods for landslide detection. For example, Lei et al. [46] introduced spatial pyramid pooling into a fully convolutional network (FCN) to extract the change features of landslides from bi-temporal aerial images; Lv et al. [47] proposed a dual-branch FCN which extracts the features of bitemporal images from two separate branches to reduce the pseudo-changes; and Shi et al. [48] proposed the CDCNN model, where the CNN is combined with an object-based change detection method and a conditional random field model to take both speed and accuracy into consideration.

Although the CNN models have achieved some satisfactory results, the hierarchical feature extraction strategy based on convolution kernels has some limitations in modeling the global information of the imagery. The transformer models provide a different perspective to model the global information based on a multihead self-attention (MSA) mechanism, where the interaction of local patches is directly considered. The transformer model was first proposed in 2017 [49] in the field of natural language processing (NLP) and was extended to deal with a computer vision task in 2020 [50]. The vision transformer (ViT) models have also been introduced into the field of remote sensing image processing, for applications such as semantic segmentation [51] and object detection [52], and have achieved competitive results, compared with CNN models. Some researchers have also studied the performance of ViT models in remote sensing image landslide detection [53]. However, the ViT structure in [53] used an input of a fixed size, which results in difficulty in exploring some small objects. Landslides display different spectral and spatial characteristics, compared with the artificial features, such as buildings and roads. With the effect of a natural disaster, the scale and shape of landslides are irregular and vary a lot. Thus, it is important to consider the scale and shape characteristics of landslides, to improve the accuracy of the model. Therefore a ViT-based network structure is proposed in this article to better consider the landslide objects with complex scale and shape.

The major contributions of this article are as follows.

1) The shape-enhanced ViT (ShapeFormer) model is proposed to perform the task of optical remote sensing image landslide detection. The features of landslides with different sizes and shapes are considered and enhanced based on two encoder branches.
2) The scale feature extraction branch introduces the pyramid vision transformer (PVT) model to acquire the hierarchical spectral and spatial features of the landslide images, in order to better keep the landslides with a small size. A set of deconvolutional layers follows the output of the PVT model to recover the original resolution of the feature maps. The connection between each scale of feature and the deconvolutional layer is added to help preserve the details.
3) The shape feature extraction branch utilizes the difference information between adjacent features in the PVT model to enhance the boundary information of the landslides, based on an attention mechanism. The shape features and the scale features are connected before the final pixel-wise classifier layer, to improve the feature representation of the network.

The rest of this article is organized as follows. The proposed ShapeFormer model is described in detail in Section II. Section III gives the experimental results and an analysis of the proposed method, compared with some of the state-of-the-art methods. A brief conclusion is given in Section IV.

II. Proposed Method

A. Overview of the ShapeFormer Model

The flowchart of the ShapeFormer model is shown in Fig. 1. The proposed method follows an encoder–decoder structure. The encoder plays the role of feature extractor, and the decoder

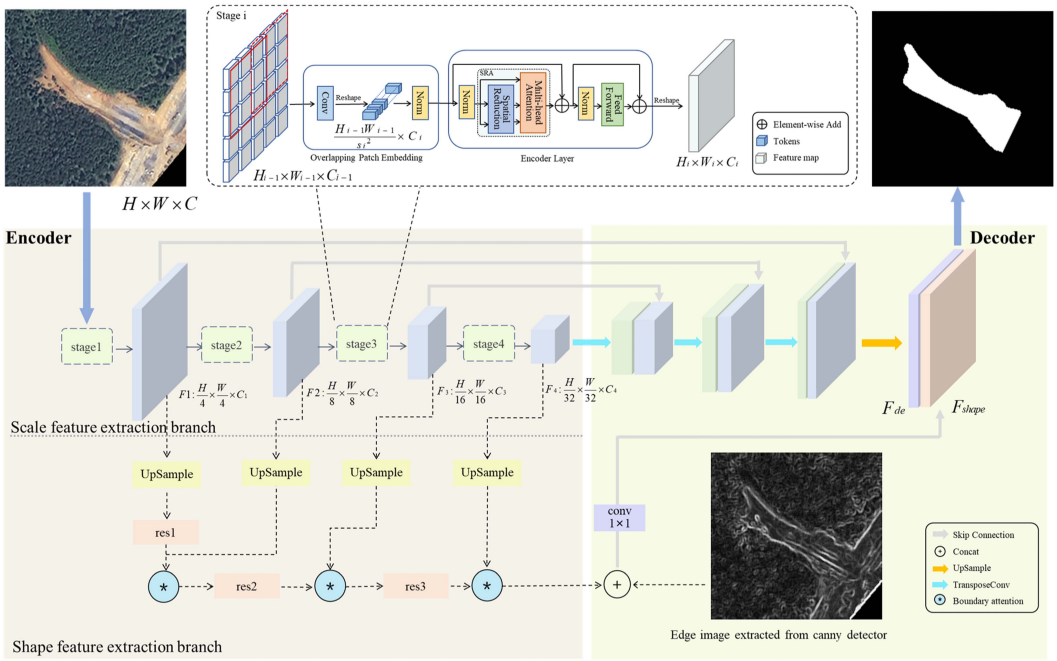

Fig. 1. Flowchart of the ShapeFormer model.

recovers the extracted features into the original resolution and is followed by a softmax layer to obtain the pixel-wise result. The encoder contains two branches, which are called the scale feature extraction branch and the shape feature extraction branch. For the scale feature extraction branch, the PVT model is utilized as the backbone to extract multiscale features from the original remote sensing image. For the shape feature extraction branch, the correlation of adjacent features extracted from each stage in the PVT model is calculated with the help of an attention mechanism, to enhance the boundary information of the landslides. In the decoder, the original spatial size of the features extracted by the PVT model is recovered by several deconvolutional layers, and the features from the two branches are stacked. Finally, the softmax classifier is used to achieve pixel-wise mapping from the stacked features and the labels, which indicates whether a landslide has occurred or not. In the following parts, the encoder and the decoder of the ShapeFormer model are introduced in detail.

B. Encoder of the ShapeFormer Model

1) *Scale Feature Extraction Branch*: In this branch, the PVT model [54], [55] is used to extract features from the original image, based on a ViT model. The original ViT model utilizes nonoverlapping patches of the imagery as the input. The self-attention operation is then performed on these patches. However, patches with a fixed size result in the ViT model having a limited ability to capture multiscale features.

We let $\mathbf{X} \in \mathbf{R}^{H \times W \times C}$ be the original input image, where $H, W, \text{and} C$ indicate the height, width, and channel of the image. The PVT model contains four stages, to hierarchically extract the features at different scales. Each stage consists of several attention and feed-forward network layers. For each stage $i = 1, 2, 3, 4$, the input feature is represented as $\mathbf{F}_{i-1}$ and the output as $\mathbf{F}_i$. Note that, for the first stage, $\mathbf{F}_0 = \mathbf{X}$. The height and width of the output features of each stage are $\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$ of the original spatial size of $\mathbf{X}$.

The input feature $\mathbf{F}_{i-1}$ for each stage is first split into several overlapping patches $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N,\}$, where $\mathbf{x}_j \in p_i \times p_i \times c_{i-1}$ and N indicates the number of patches. The overlapping patches are used to preserve some of the local correlation information. The size of each patch is defined as $p_i = 7, 3, 3, 3$, and the stride of the sliding window is $st_i = 4, 2, 2, 2$. The patches are then flattened and projected to c_i-dimension embeddings to acquire the token series $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N,\}$. In the PVT model, the token is the basic processing unit, which contains some local spatial and spectral information. In each stage, the patch size p_i is adjusted, and thus the PVT model can extract multiscale features without the convolution operation used in a CNN. Each stage i has several encoder layers, which contain an MSA layer and a feed-forward layer. The MSA mechanism considers the correlation between tokens by introducing three queries, called $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$. These queries are initialized by the token $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N,\}$ and can be formulated as follows:

$$\text{MSA}(Q, K, V) = \text{concat}(\text{head}_1, \ldots, \text{head}_{H_j})\mathbf{W} \quad (1)$$

$$\text{head}_k = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{W}_k^Q\left(\mathbf{K}\mathbf{W}_k^K\right)^T}{\sqrt{d_k}}\right)\mathbf{V}\mathbf{W}_k^V \quad (2)$$

where $\mathbf{W} \in \mathbb{R}^{c_i \times c_i}$ is the weight matrix for the linear projection; $\mathbf{W}_k^Q \in \mathbb{R}^{c_i \times d_k}$, $\mathbf{W}_k^K \in \mathbb{R}^{c_i \times d_k}$, and $\mathbf{W}_k^V \in \mathbb{R}^{c_i \times d_k}$ refer to the weights of the three queries; and H_j is the number of heads in the jth encoder layer. In order to reduce the computational cost, a spatial reduction operation is performed on the original MSA

layer. The queries $\mathbf{K}$and $\mathbf{V}$ are reshaped and linear projected into the c_i-dimension embeddings, based on $\mathbf{W}_{SR} \in \mathbb{R}^{(R_i^2 c_i)\times c_i}$

$$\mathrm{SR}(\cdot) = \mathrm{LN}\left(\mathrm{Reshape}\left(\cdot, \mathbf{R}_i\right)\mathbf{W}_{SR}\right) \tag{3}$$

where $\mathbf{R}_i$ is the reduction ratio. The output of the encoder layers is reshaped into a 3-D cube $\mathbf{F}_i$ as the input of the next stage i+1.

2) *Shape Feature Extraction Branch*: The scale feature extraction branch generates hierarchical spatial and spectral features of the original image. However, some details of small objects and boundaries will be lost in the high-level features. In this branch, the boundaries of the features $\mathbf{F}_i$ generated from each stage in the PVT model are enhanced with the help of the boundary attention operation [56], to better consider the shape information of the landslide.

As displayed in Fig. 1, the output feature $\mathbf{F}_i$ for stage i is first upsampled to recover the original resolution $H \times W$ of the image. The output feature $\mathbf{F}_i$ is then fed into a residual block followed by a convolutional layer, to reduce the number of channels and obtain the feature $\mathbf{F'}_i$. The adjacent feature $\mathbf{F}_{i+1}$ from stage i+1 is also processed under the same operation to obtain $\mathbf{F'}_{i+1}$. Because $\mathbf{F'}_i$ and $\mathbf{F'}_{i+1}$ are generated at different scales in the scale feature extraction branch, the boundary information of the objects in these feature maps can be enhanced by considering the difference between these features.

Here we use the attention mechanism to emphasize the boundary information. First, the attention weight $\mathbf{A}$ is computed by a 1-D convolution operation and nonlinear projection, which is formulated as shown in (4)

$$\mathbf{A}_i = \sigma\left(\mathrm{conv}_{1d}\left(\mathrm{concat}\left(\mathbf{F'}_i, \mathbf{F'}_{i+1}\right)\right)\right) \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function. The shape information in $\mathbf{F'}_i$ is then improved based on the element-wise production, followed by a residual block

$$\mathbf{F''}_i = \left(\left(\mathbf{F'}_i \odot \mathbf{A}_i\right) + \mathbf{F'}_i\right)\mathbf{W}_i \tag{5}$$

where $\mathbf{W}_i$ is a projection matrix to reduce the channels. Note that there are four stages in the scale feature extraction branch, and thus the output features $\mathbf{F}_1$, $\mathbf{F}_2$, and $\mathbf{F}_3$ are enhanced by the boundary attention operation in the shape feature extraction branch. In order to acquire more details from the original image, the boundary information of the original image is calculated by the Canny edge detector [57] and stacked with $\mathbf{F}_3$ to generate the final output of the shape feature extraction branch $\mathbf{F}_{\text{shape}}$.

C. Decoder of the ShapeFormer Model

The objective of the decoder of the ShapeFormer model is to merge the features generated from the two encoder branches and recover the original spatial size of the image, to achieve end-to-end pixel-wise classification. Usually, the upsampling operation can be performed based on interpolation or deconvolution. Inspired by the structure of U-Net [40], we designed the decoder of ShapeFormer under a hierarchical structure, based on deconvolutional layers and skip connections. Although bilinear interpolation is fast and convenient, it will lose some details because the recovered features are simply based on the adjacent values. The advantage of this structure is that some learnable parameters are introduced, and the original spatial size can be recovered without losing much detail. The hierarchical structure of the decoder also makes it more convenient to combine the multistage features generated in the encoder to further keep the details and improve the accuracy.

As displayed in Fig. 1, the output feature of the scale feature extraction branch $\mathbf{F}_4$ is first upsampled by the deconvolution operation to match the spatial size of $\mathbf{F}_3$. The two features are then stacked as the input of a deconvolutional layer. The output of the first deconvolutional layer is stacked with $\mathbf{F}_2$ as the input of the next deconvolutional layer. After the three deconvolutional layers, a bilinear interpolation operation follows to generate feature $\mathbf{F}_{de}$, which has the same spatial size as the original image. $\mathbf{F}_{de}$ is then stacked with the output of the shape feature extraction branch $\mathbf{F}_{\text{shape}}$ to acquire the final feature expression $\mathbf{F}_{\text{final}}$ before the classification layer.

III. Experiments and Analysis

In this part, we describe the experiments conducted with two public landslide detection datasets—the Bijie dataset [38] and the Nepal dataset [58]—where the images were acquired from different remote sensing platforms and areas, to verify the performance of the ShapeFormer model. We also selected some of the related state-of-the-art landslide detection methods as the comparison methods. The methods used in the experiments were the two classical CNN-based methods of ResUNet [39] and DeepLabv3+ [59], and two state-of-the-art landslide detection methods, i.e., TransUNet [54] and FC-DenseNet [43]. We also analyzed the effectiveness of the shape feature extraction branch by introducing the PVT-UNet method, which uses only the PVT model as the encoder. The landslide detection can be viewed as a binary segmentation task where the foreground and background samples are imbalanced. Thus, the precision, recall, F1-score, and intersection over union (IoU) of the landslide class are used as the quantitative indices

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{6}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{7}$$

$$F1 = \frac{2\times \mathrm{TP}}{2\times \mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{8}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{9}$$

where TP, FP, and FN represent the true positives, false positives, and false negatives of the landslide pixels, respectively.

All the experiments were performed on an NVIDIA RTX 3090 GPU, and the deep learning environment was PyTorch. The experimental settings and the result analyses for the two datasets are given as follows.

TABLE I
ACCURACY ASSESSMENT FOR THE BIJIE DATASET (%)

Method	Precision	Recall	F1	IoU
ResUNet	86.91	87.01	86.89	76.57
DeepLabv3+	84.43	88.37	86.35	75.99
FC-DenseNet	74.35	86.12	80.65	67.58
TransUNet	**87.24**	88.23	87.73	78.15
PVT-UNet	83.93	88.86	86.33	75.91
Proposed	86.74	**89.52**	**88.11**	**78.72**

The best results have been highlighted in bold.

TABLE II
ACCURACY ASSESSMENT FOR THE NEPAL DATASET (%)

Method	Precision	Recall	F1	IoU
ResUNet	65.89	66.12	65.82	49.09
DeepLabv3+	**70.67**	63.11	66.65	50.00
FC-DenseNet	65.34	61.54	63.38	46.36
TransUNet	65.39	51.93	68.32	51.93
PVT-UNet	69.69	68.95	69.17	52.89
Proposed	66.06	**73.05**	**69.34**	**53.09**

The best results have been highlighted in bold.

A. Experiment 1: Bijie Dataset

The study area of the first dataset is the city of Bijie, which locates in the northwest of Guizhou province, China. The average altitude of the city of Bijie is 1600 m, and the city is surrounded by mountains and rivers. Because of the geographical position, complex geological features, fragile ecological environment, and frequent rainfall, the area is very prone to landslides. The images in the Bijie dataset were acquired by the TripleSat satellite from May to August 2018, with a spatial resolution of 0.8 m and three visual bands. There are 770 labeled images in this dataset, and the spatial sizes of these images are different. Thus, in the experiment, we resized the images to 224×224 pixels, according to [53], for all the comparison methods.

For the experimental settings, the ratio of the training set, validation set, and test set was 0.7, 0.2, and 0.1 for all the experiments. Each experiment was run five times, and we calculated the mean value of each quantitative index. The same data augmentation was conducted in all the experiments, i.e., random flips horizontal and vertical, random rotation, and salt-and-pepper noise. For ShapeFormer and PVT-UNet, cross-entropy loss was selected to be the loss function during the training stage. Moreover, in order to reduce the influence of sample imbalance, we set the weight of landslide and background to 10 and 3 in the loss function. The AdamW optimizer was used, and the weight decay was set to 0.0005. The initial learning rate started from 0.001. The training epochs numbered 230 and the batch size was 48. The above parameters were acquired according to a trial-and-error strategy. For TransUNet, the parameters were set according to the original paper.

The accuracy results are listed in Table I. As shown in Table I, for the CNN-based methods, ResUNet acquires a better performance, compared with DeepLabv3+ and FC-DenseNet, which indicates that using ResNet as the backbone can achieve good feature representation on the Bijie dataset. When the accuracy results of ResUNet and TransUNet are compared, TransUNet shows an obvious improvement in these four quantitative measurements. This is because TransUNet uses the output feature of ResNet as the input of the ViT encoder, which extracts more useful semantic information for recognizing the landslide areas. When the PVT model is selected as the backbone, the results of PVT-UNet are also satisfactory, compared with ResUNet and DeepLabv3+, which shows that multiscale features can also be extracted well from the pure ViT backbone. The accuracy of TransUNet is higher than that of PVT-UNet because the input of TransUNet is the multiple CNN features, while the input of PVT-UNet is the original image. The proposed ShapeFormer model acquires the best performance in recall, F1-score, and IoU, at 89.52%, 88.11%, and 78.72%, respectively. The precision of the ShapeFormer model is also acceptable. The prominent improvement of the F1-score of ShapeFormer indicates that the proposed method can keep a good balance between the truly detected landslides and the false alarms. The highest value of IoU on the landslide type also shows the effectiveness of the proposed shape feature extraction branch in keeping the boundary information.

In Fig. 2, several images from the test set with different spectral and shape characteristics are selected and displayed. As can be seen in Fig. 2, ResUNet can find the major area of the landslides, but loses some details. DeepLabv3+ can preserve some local details with the help of the atrous convolution. However, it also introduces some false alarms. TransUNet can keep a good balance between the details and the noise, but it still has some limitations when the shape of the landslide is complex. For the results of the ShapeFormer model, it performs well in the landslide areas with complex shapes and various sizes. This suggests that the shape feature in the proposed framework is very effective, especially on landslides of a small size.

B. Experiment 2: Nepal Dataset

The study area of the second dataset is Nepal, which is located in the southern foothills of the Himalayas, with an altitude between 4877 and 8844 m. The complex terrain, frequent seismic activity, and torrential monsoon rains make Nepal a landslide-prone area. The dataset consists of 275 landslide images, all from the Landsat 8 satellite. The spatial size of each image in this dataset is 256×256. Because the spatial resolution of the images in the Nepal dataset is lower than that of the Bijie dataset, it brings more challenges for the algorithms to find clear boundaries of the landslides.

In the training process, according to [58], there were 230 images contained in the training set, 35 images in the validation set, and 10 images in the test set. The hyperparameters of ShapeFormer and PVT-UNet were set to be the same values as used in the Bijie dataset. The parameters were again selected based on the grid search strategy.

The quantitative results for the Nepal dataset are listed in Table II. Differing from the Bijie dataset, where the images have a high spatial resolution, the images in the Nepal dataset have a medium resolution and cover more complex scenes. As a result, it is difficult for the networks to extract the landslides based on the spatial details, and the accuracies are not as good as for the Bijie dataset. The networks should utilize the multiple spectral and spatial features in these images to acquire good

Fig. 2. Visual results for the Bijie dataset. Rows 1–8 indicate the original images, the labels, the results of ResUNet, the results of DeepLabv3+, the results of FC-DenseNet, the results of TransUNet, the results of PVT-UNet, and the results of ShapeFormer.

results. For the CNN-based methods, DeepLabv3+ obtains a better performance in F1-score and IoU, which indicates that DeepLabv3+ can better use the spectral features contained in the multispectral images. For TransUNet, the F1-score and IoU results are also acceptable because this method combines the advantages of both CNN and ViT models, but the accuracy could still be improved. The higher precision and lower recall indicate that TransUNet introduces more false alarms in the Nepal dataset. For PVT-UNet, the precision and recall values are 69.69% and 68.95%, which shows that it is quite accurate with regard to the detected landslides, but it misses more true landslides than the other methods. The proposed ShapeFormer model obtains the highest values in the recall, F1-score, and IoU, at 73.05%, 69.34%, and 53.09%, respectively. The recall of ShapeFormer is about 10% higher than that of the other methods, showing the ability of the proposed method to find true landslides in complex scenes when the shape information is emphasized. The highest IoU of ShapeFormer also shows that it

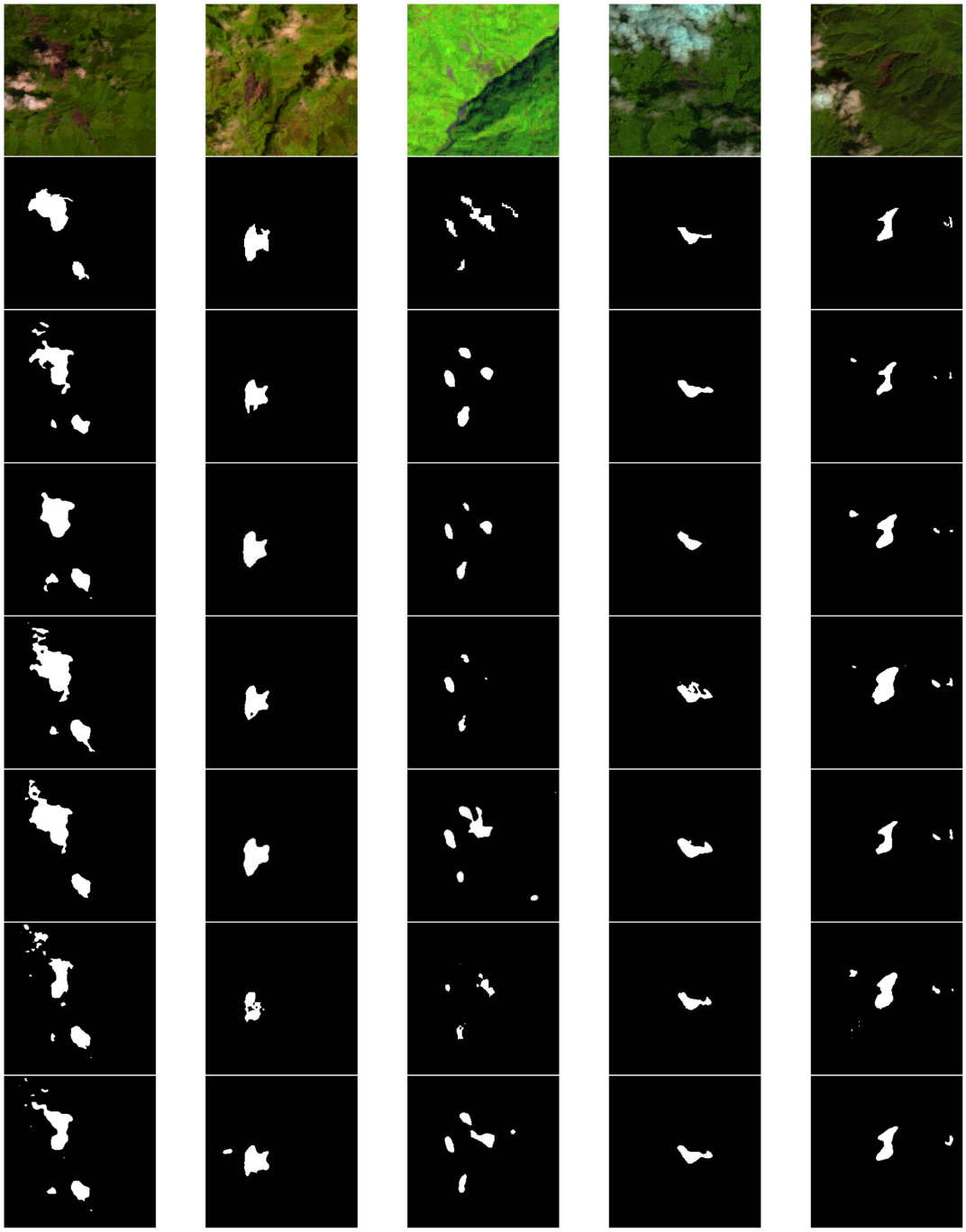

Fig. 3. Visual results for the Nepal dataset. Rows 1–8 indicate the original images, the labels, the results of ResUNet, the results of DeepLabv3+, the results of FC-DenseNet, the results of TransUNet, the results of PVT-UNet, and the results of ShapeFormer.

can effectively distinguish the boundaries of the landslides from the background.

The visual results are displayed in Fig. 3. The first row shows the false-color images selected from the test set of the Nepal dataset. It can be seen that, differing from the Bijie dataset, it is very difficult to find the landslides in the images by visual interpretation. For the CNN-based methods, DeepLabv3+, which considers the multiresolution features, has the ability to find more detailed landslides in this dataset. There are more false alarms in the results of TransUNet. When the PVT model is used as the backbone in PVT-UNet, it loses some true landslide areas, although the detected parts are accurate. The visual results of TransUNet and PVT-UNet are consistent with their quantitative accuracy in Table II. The proposed ShapeFormer model acquires the best visual results, keeping a balance between the landslides and the background types.

C. Model Analysis

The parameters and floating point operations per second (FLOPs) of the models used in the experiments are listed in Table III. From Table III, it can be seen that the CNN-based models have fewer parameters than the ViT-based methods. When the accuracy values in Tables I and II are also considered,

TABLE III
PARAMETERS AND FLOPS OF THE DIFFERENT MODELS

Method	Parameters	FLOPs
ResUNet	51.51 M	11937.37 M
DeepLabv3+	45.67 M	10798.10 M
FC-DenseNet	1.38 M	10195.47 M
TransUNet	93.40 M	24680.46 M
PVT-UNet	48.48 M	8365.51 M
Proposed	48.58 M	13364.93 M

it can be concluded that, although the size of the TransUNet model is very large, it does result in some improvement in accuracy. Overall, the proposed ShapeFormer keeps a good balance between the model size and the running efficiency.

IV. CONCLUSION

In this article, we have mainly focused on the problem of landslide detection based on optical remote-sensing images through deep learning-based methods. We analyzed the spectral and spatial characteristics of the landslides in multisensor satellite imagery and proposed solving this problem from a new perspective. The performance of ViT models was studied in depth for the landslide detection task, and the ShapeFormer model was proposed to take the complexity of the different shapes of landslides into account. In the proposed ShapeFormer model, a PVT model is used to extract multiscale deep features from the original imagery, based on a self-attention operation. The shape information of the features extracted from each stage in the PVT model is enhanced by taking the difference of the adjacent features into consideration, with the help of a boundary attention operation. The results, compared with those of some of the state-of-the-art deep learning methods on two public datasets, showed the potential of the proposed method, which indicates the benefit of ViT models when dealing with the remote sensing image landslide detection problem.

REFERENCES

[1] H. Wang, G. Liu, W. Xu, and G. Wang, "GIS-based landslide hazard assessment: An overview," *Prog. Phys. Geogr., Earth Environ.*, vol. 29, no. 4, pp. 548–567, 2016.

[2] S. D. Pardeshi, S. E. Autade, and S. S. Pardeshi, "Landslide hazard assessment: Recent trends and techniques," *SpringerPlus*, vol. 2, no. 1, Art. no. 523.

[3] M. Scaioni, L. Longoni, V. Melillo, and M. Papini, "Remote sensing for landslide investigations: An overview of recent achievements and perspectives," *Remote Sens.*, vol. 6, no. 10, pp. 9600–9652, 2014.

[4] C. Zhong et al., "Landslide mapping with remote sensing: Challenges and opportunities," *Int. J. Remote Sens.*, vol. 41, no. 4, pp. 1555–1581, 2019.

[5] F. Mantovani, R. Soeters, and C. Van Westen, "Remote sensing techniques for landslide studies and hazard zonation in Europe," *Geomorphology*, vol. 15, no. 3–4, pp. 213–225, 1996.

[6] Y. Hong, R. Adler, and G. Huffman, "Use of satellite remote sensing data in the mapping of global landslide susceptibility," *Natural Hazards*, vol. 43, no. 2, pp. 245–256, 2007.

[7] J. Nichol and M. S. Wong, "Detection and interpretation of landslides using satellite images," *Land Degradation Develop.*, vol. 16, no. 3, pp. 243–255, 2005.

[8] A. C. Mondini, F. Guzzetti, P. Reichenbach, M. Rossi, M. Cardinali, and F. Ardizzone, "Semi-automatic recognition and mapping of rainfall induced shallow landslides using optical satellite images," *Remote Sens. Environ.*, vol. 115, no. 7, pp. 1743–1757, 2011.

[9] W. Li et al., "Precursors to large rockslides visible on optical remote-sensing images and their implications for landslide early detection," *Landslides*, vol. 20, pp. 1–12, 2022.

[10] M. Melis et al., "Thermal Remote sensing from UAVs: A review on methods in coastal cliffs prone to landslides," *Remote Sens.*, vol. 12, no. 12, 2020, Art. no. 1971.

[11] Y. Zhu, X. Yao, L. Yao, and C. Yao, "Detection and characterization of active landslides with multisource SAR data and remote sensing in western Guizhou, China," *Natural Hazards*, vol. 111, no. 1, pp. 973–994, 2022.

[12] P. Confuorto et al., "Post-failure evolution analysis of a rainfall-triggered landslide by multi-temporal interferometry SAR approaches integrated with geotechnical analysis," *Remote Sens. Environ.*, vol. 188, pp. 51–72, 2017.

[13] M. Jaboyedoff et al., "Use of LIDAR in landslide investigations: A review," *Natural Hazards*, vol. 61, no. 1, pp. 5–28, 2010.

[14] M.V. D. Eeckhaut et al., "Use of LIDAR-derived images for mapping old landslides under forest," *Earth Surf. Processes Landforms*, vol. 32, no. 5, pp. 754–769, 2007.

[15] K. Pawluszek, "Landslide features identification and morphology investigation using high-resolution DEM derivatives," *Natural Hazards*, vol. 96, no. 1, pp. 311–330, 2018.

[16] D. Kawabata and J. Bandibas, "Landslide susceptibility mapping using geological data, a DEM from ASTER images and an artificial neural network (ANN)," *Geomorphology*, vol. 113, no. 1–2, pp. 97–109, 2009.

[17] F. Fiorucci, D. Giordan, M. Santangelo, F. Dutto, M. Rossi, and F. Guzzetti, "Criteria for the optimal selection of remote sensing optical images to map event landslides," *Natural Hazards Earth Syst. Sci.*, vol. 18, no. 1, pp. 405–417, 2018.

[18] W. Zhao, A. Li, X. Nan, Z. Zhang, and G. Lei, "Postearthquake landslides mapping from Landsat-8 data for the 2015 Nepal earthquake using a pixel-based change detection method," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 1758–1768, May 2017.

[19] T. Chen, J. Trinder, and R. Niu, "Object-oriented landslide mapping using ZY-3 satellite imagery, random forest and mathematical morphology, for the Three-Gorges Reservoir, China," *Remote Sens.*, vol. 9, no. 4, 2017 Art. no. 333.

[20] A. Stumpf and N. Kerle, "Object-oriented mapping of landslides using Random Forests," *Remote Sens. Environ.*, vol. 115, no. 10, pp. 2564–2577, 2011.

[21] X. Li, X. Cheng, W. Chen, G. Chen, and S. Liu, "Identification of forested landslides using LiDar data, object-based image analysis, and machine learning algorithms," *Remote Sens.*, vol. 7, no. 8, pp. 9705–9726, 2015.

[22] J.-Y. Rau, J.-P. Jhan, and R.-J. Rau, "Semiautomatic object-oriented landslide recognition scheme from multisensor optical imagery and DEM," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1336–1349, Feb. 2014.

[23] Z. Li, W. Shi, P. Lu, L. Yan, Q. Wang, and Z. Miao, "Landslide mapping from aerial photographs using change detection-based Markov random field," *Remote Sens. Environ.*, vol. 187, pp. 76–90, 2016.

[24] Z. Lv, W. Shi, X. Zhang, and J. A. Benediktsson, "Landslide inventory mapping from bitemporal high-resolution remote sensing images using change detection and multiscale segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1520–1532, May 2018.

[25] J. Travelletti et al., "Correlation of multi-temporal ground-based optical images for landslide monitoring: Application, potential and limitations," *ISPRS J. Photogramm. Remote Sens.*, vol. 70, pp. 39–55, 2012.

[26] G. Rossi, L. Tanteri, V. Tofani, P. Vannocci, S. Moretti, and N. Casagli, "Multitemporal UAV surveys for landslide mapping and characterization," *Landslides*, vol. 15, no. 5, pp. 1045–1052, 2018.

[27] P. Lv, W. Wu, Y. Zhong, F. Du, and L. Zhang, "SCViT: A spatial-channel feature preserving vision transformer for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4409512.

[28] Z. Lv, G. Li, Z. Jin, J. A. Benediktsson, and G. M. Foody, "Iterative training sample expansion to increase and balance the accuracy of land classification from VHR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 139–150, Jan. 2020.

[29] J. Wang, A. Ma, Y. Zhong, Z. Zheng, and L. Zhang, "Cross-sensor domain adaptation for high spatial resolution urban land-cover mapping: From airborne to spaceborne imagery," *Remote Sens. Environ.*, vol. 277, 2022, Art. no. 113058.

[30] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 60–77, 2018.

[31] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 3–22, 2018.
[32] P. Wang, B. Bayram, and E. Sertel, "A comprehensive review on deep learning based remote sensing image super-resolution methods," *Earth-Sci. Rev.*, vol. 232, 2022, Art. no. 104110.
[33] 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, 2020, Art. no. e3998.
[34] O. Ghorbanzadeh, T. Blaschke, K. Gholamnia, S. Meena, D. Tiede, and J. Aryal, "Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection," *Remote Sens.*, vol. 11, no. 2, 2019, Art. no. 196.
[35] Y. Wang, Z. Fang, and H. Hong, "Comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County, China," *Sci. Total Environ.*, vol. 666, pp. 975–993, 2019.
[36] O. Ghorbanzadeh, S. R. Meena, H. S. S. Abadi, S. T. Piralilou, L. Zhiyong, and T. Blaschke, "Landslide mapping using two main deep-learning convolution neural network streams combined by the Dempster–Shafer model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 452–463, 2021.
[37] 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, 2020.
[38] Z. Su, J. K. Chow, P. S. Tan, J. Wu, Y. K. Ho, and Y.-H. Wang, "Deep convolutional neural network–based pixel-wise landslide inventory mapping," *Landslides*, vol. 18, no. 4, pp. 1421–1443, 2020.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[40] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.
[41] T. Liu, T. Chen, R. Niu, and A. Plaza, "Landslide detection mapping employing CNN, ResNet, and DenseNet in the Three Gorges Reservoir, China," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 11417–11428, 2021.
[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
[43] X. Gao, T. Chen, R. Niu, and A. Plaza, "Recognition and mapping of landslide using a fully convolutional DenseNet and influencing factors," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7881–7894, 2021.
[44] Y. Chen, D. Ming, X. Ling, X. Lv, and C. Zhou, "Landslide susceptibility mapping using feature fusion-based CPCNN-ML in Lantau Island, Hong Kong," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3625–3639, 2021.
[45] W. Xia, J. Chen, J. Liu, C. Ma, and W. Liu, "Landslide extraction from high-resolution remote sensing imagery using fully convolutional spectral–topographic fusion network," *Remote Sens.*, vol. 13, no. 24, 2021, Art. no. 5116.
[46] T. Lei, Y. Zhang, Z. Lv, S. Li, S. Liu, and A. K. Nandi, "Landslide inventory mapping from bitemporal images using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 982–986, Jun. 2019.
[47] Z. Lv, T. Liu, X. Kong, C. Shi, and J. A. Benediktsson, "Landslide inventory mapping with bitemporal aerial remote sensing images based on the dual-path fully convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4575–4584, 2020.
[48] W. Shi, M. Zhang, H. Ke, X. Fang, Z. Zhan, and S. Chen, "Landslide recognition by deep convolutional neural network and change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4654–4672, Jun. 2021.
[49] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[50] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.
[51] L. Gao et al., "STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10990–11003, 2021.
[52] H. Gong et al., "Swin-transformer-enabled YOLOv5 with attention mechanism for small object detection on satellite images," *Remote Sens.*, vol. 14, no. 12, 2022, Art. no. 2861.
[53] Z. Yang, C. Xu, and L. Li, "Landslide detection based on ResU-net with transformer and CBAM embedded: Two examples with geologically different environments," *Remote Sens.*, vol. 14, no. 12, 2022, Art. no. 2885.
[54] W. Wang et al., "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, 2022.
[55] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578.
[56] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape CNNs for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5229–5238.
[57] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
[58] L. Bragagnolo, L. Rezende, R. da Silva, and J. Grzybowski, "Convolutional neural networks applied to semantic segmentation of landslide scars," *Catena*, vol. 201, 2021, Art. no. 105189.
[59] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

Pengyuan Lv (Member, IEEE) received the B.S. degree in photogrammetry and remote sensing from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2019.

He is an Associate Professor with the School of Information Engineering, Ningxia University, China. His main research interests include high spatial resolution remote sensing image change detection, deep learning, and random field algorithms.

Lusha Ma received the B.Eng. degree in software engineering from the School of Dalian Maritime University, Dalian, China, in 2014. She is currently working toward the master's degree in computer technology with the School of Information Engineering, Ningxia University, Yinchuan, China.

Her research interests include remote sensing image semantic segmentation and computer vision.

Qiaomin Li received the Master's degree in geological engineering (remote sensing geology) from the China University of Geosciences, Beijing, China, in 2016.

He serves as an Engineer with Ningxia Remote Sensing Survey Institute. His main research interests include remote sensing of resources and environment, ecological geological remote sensing, and geological disaster remote sensing.

Fang Du received the B.S. degree in computer application technology from the Ningxia University, Yinchuan, China, in 1996, the M.S. degree in computer application technology from the Lanzhou University, Lanzhou, China, in 2002, and the Ph.D. degree in computer application technology from the Renmin University of China, Beijing, China, in 2013.

She is a Professor and Deputy Dean of the School of Information Engineering, Ningxia University, Yinchuan, China. She has authored more than 20 research articles for international conferences and journals, including *Journal of Software*, IEEE BigData 2016, and IEEE ICME 2020. Her major research interests include big data management and artificial intelligence.