remote sensing

Article

Automatic Detection of Coseismic Landslides Using a New Transformer Method

Xiaochuan Tang [1,2,3,*], Zihan Tu [2], Yu Wang [2], Mingzhe Liu [1,2], Dongfen Li [2] and Xuanmei Fan [1]

[1] State Key Laboratory of Geohazard Prevention and Geoenvironment Protection, Chengdu University of Technology, Chengdu 610059, China; liumz@cdut.edu.cn (M.L.); fanxuanmei2014@cdut.edu.cn (X.F.)

[2] College of Computers and Cyber Security, Chengdu University of Technology, Chengdu 610059, China; tuzihan@stu.cdut.edu.cn (Z.T.); wangyu3@stu.cdut.edu.cn (Y.W.); lidongfen17@cdut.edu.cn (D.L.)

[3] National Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China

* Correspondence: tangchuan@uestc.edu.cn

Abstract: Earthquake-triggered landslides frequently occur in active mountain areas, which poses great threats to the safety of human lives and public infrastructures. Fast and accurate mapping of coseismic landslides is important for earthquake disaster emergency rescue and landslide risk analysis. Machine learning methods provide automatic solutions for landslide detection, which are more efficient than manual landslide mapping. Deep learning technologies are attracting increasing interest in automatic landslide detection. CNN is one of the most widely used deep learning frameworks for landslide detection. However, in practice, the performance of the existing CNN-based landslide detection models is still far from practical application. Recently, Transformer has achieved better performance in many computer vision tasks, which provides a great opportunity for improving the accuracy of landslide detection. To fill this gap, we explore whether Transformer can outperform CNNs in the landslide detection task. Specifically, we build a new dataset for identifying coseismic landslides. The Transformer-based semantic segmentation model SegFormer is employed to identify coseismic landslides. SegFormer leverages Transformer to obtain a large receptive field, which is much larger than CNN. SegFormer introduces overlapped patch embedding to capture the interaction of adjacent image patches. SegFormer also introduces a simple MLP decoder and sequence reduction to improve its efficiency. The semantic segmentation results of SegFormer are further improved by leveraging image processing operations to distinguish different landslide instances and remove invalid holes. Extensive experiments have been conducted to compare Transformer-based model SegFormer with other popular CNN-based models, including HRNet, DeepLabV3, Attention-UNet, U^{2}Net and FastSCNN. SegFormer improves the accuracy, mIoU, IoU and F1 score of landslide detectuin by 2.2%, 5% and 3%, respectively. SegFormer also reduces the pixel-wise classification error rate by 14%. Both quantitative evaluation and visualization results show that Transformer is capable of outperforming CNNs in landslide detection.

Keywords: landslide detection; coseismic landslide; Transformer; self-attention; convolutional neural network; semantic segmentation; deep learning

check for updates

Citation: Tang, X.; Tu, Z.; Wang, Y.; Liu, M.; Li, D.; Fan, X. Automatic Detection of Coseismic Landslides Using a New Transformer Method. *Remote Sens.* **2022**, *14*, 2884. https://doi.org/10.3390/rs14122884

Academic Editors: Wanchang Zhang and Shunping Ji

Received: 25 April 2022
Accepted: 15 June 2022
Published: 16 June 2022

1. Introduction

In recent years, there have been more and more extreme weather events. Natural disasters, such as earthquakes and floods, have caused huge casualties and property losses. Earthquakes in mountain areas usually induce landslides, which are referred to as coseismic landslides or earthquake-triggered landslides. Landslide is one of the most dangerous geological disasters. Earth scientists have manually identified over 300 thousand landslides in China, and even more landslides are discovered every year [1]. However, earth scientists believe that the number of known landslides is a small portion of all existing landslides. Many landslides have not been discovered. There are many challenges in manual landslide detection. For example, it is difficult to manually identify the landslides located in uninhabited and vegetation-covered mountainous areas.

Conventionally, landslide identification is performed by experienced earth scientists [2]. Manual landslide identification has good accuracy but poor efficiency [3]. Remote sensing technologies provide rich data for developing landslide detection models, such as high-resolution optical remote sensing [4], Interferometric Synthetic Aperture Radar (InSAR) [5] and Light Detection and Ranging (LiDAR) [6–8]. Based on these data, many automatic landslide detection methods have been proposed. Automatic landslide detection methods are mainly based on machine learning, which can be classified into non-deep learning and deep learning methods. Figure 1 presents a brief history of automatic landslide detection. The upper timeline summarizes non-deep learning methods for landslide detection, such as support vector machine and random forest. The lower timeline shows the development of deep learning methods for landslide detection. Danneels et al. [9] used pixel classification to detect landslides in the Gultcha area, which is in the southern part of Kyrgyzstan. They compared the accuracies of a maximum likelihood ratio and an artificial neural network classification model. Pradhan et al. [10] used the frequency ratio, logistic regression and an artificial neural network to identify landslide hazard areas in Penang Island, Malaysia. Stumpf et al. [11] combined object-oriented method random forest to detect landslides in Haiti, Wenchuan, Messina and Barcelonnette. They extracted 20 object-oriented features, such as slope, hillshade and grey level co-occurrence matrix. Moosavi et al. [12] used an artificial neural network, support vector machine and object-oriented methods to detect landslides in Kermanshah, Iran. The Taguchi method was used to perform optimization of the structure of the object-oriented classification method. Chen et al. [13] used feature selection and random forest to identify landslides in Three Gorges, China. They extracted the slope, aspect and DTM from LiDAR data. Gorsevski et al. [14] used LiDAR and an artificial neural network approach to detect landslides in Cuyahoga Valley National Park, Ohio. Mezaal et al. [15] used correlation-based feature selection, random forest and ant colony optimization to identify landslides in the Cameron Highlands, Malaysian. Hu et al. [16] employed support vector machine, an artificial neural network and random forest to classify satellite imagery and identify landslides in Jiuzhaigou, China. Tavakkoli et al. [17] incorporated object-based image analysis with multilayer perceptron, logistic regression and random forest to detect landslides in the Rasuwa District of Nepal. Dou et al. [18] used support vector machine with bagging, boosting and a stacking ensemble machine learning framework to improve landslide assessment in a mountainous watershed in Japan. Dias [19] used support vector machine, random forest and maximum likelihood classifiers to recognize landslides in Itaóca, Southeastern Brazil. Rajabi et al. [20] used a back-propagation-type artificial neural network to predict the landslides triggered by the Manjil-Rudbar earthquake, Iran. Wang et al. [4] combined change detection with k-nearest neighbor, support vector machine, random forest and rotation forest for detecting coseismic landslides in Haiti. Ghorbanzadeh et al. [21] compared an artificial neural network, support vector machine and random forest with a convolutional neural network (CNN). They claimed that CNN-based landslide detection methods do not automatically outperform artificial neural networks, support vector machine and random forest. They also pointed out that deep learning methods could improve landslide mapping in the future.

Second, deep learning-based methods may be considered. Deep learning-based image classification [22], semantic segmentation [23–25], object detection [26] and instance segmentation [27] are used for landslide detection. Ding et al. [28] used a convolutional neural network and texture change detection to recognize landslides in Shenzhen, China. Yu et al. [29] proposed an algorithm based on a depth convolutional neural network and an improved region growing algorithm for landslide detection. Ghorbanzadeh et al. [21] introduced a convolutional neural network for landslide detection. Lei et al. [30] proposed FCN-PP for landslide detection, which is a fully convolutional network with a pyramid pooling module. Prakash et al. [31] applied U-Net for landslide detection. Yu et al. [32] introduced

ResNet-101 and pyramid pooling to landslide detection. Qi et al. [33] proposed ResU-Net for landslide detection, which is a combination of residual block and U-Net. Shi et al. [23] proposed a CNN-based semantic segmentation model with change detection for landslide detection. Yi et al. [34] proposed LandsNet for detecting earthquake-triggered landslides. Wang et al. [35] proposed an integrated machine learning and deep learning method to identify natural-terrain landslides. Logistic regression, support vector machine, random forest, boosting methods and a convolutional neural network were utilized and evaluated on a dataset for Lantau and Hong Kong. Ullo et al. [27] used instance segmentation method mask R-CNN to improve landslide detection. Cheng et al. [36] added an attention mechanism on top of YOLOv4. The attention mechanism was used to improve the CNN's focus on the landslide feature and reduce the background noise. Bragagnolo et al. [37] applied U-Net to identify landslides in Nepal. Tang et al. [22] proposed a multi-instance learning model, MILL, for identifying ancient landslides. Ghorbanzadeh [24] conducted a comprehensive transferability evaluation of U-Net and ResU-Net for landslide detection from Sentinel-2 data. Li et al. [38] proposed a two-stage method for landslide detection. They used Fast R-CNN to obtain the bounding box of landslides. Then, the bounding box is fed into U-Net to obtain the boundary of the landslide. Ghorbanzadeh et al. [39] combined a ResU-Net model with object-based image analysis (OBIA). OBIA used rule-based expert knowledge to improve the results of ResU-Net. Meena et al. [40] used fully convolutional U-Net, support vector machines, k-nearest neighbor and random forest to detect landslides in the Rasuwa district, Nepal. Ju et al. [26] added YOLOv3 and RetinaNet models before mask R-CNN. Chen et al. [41] proposed a deep residual shrinkage U-Net to extract potential active landslides in InSAR imagery. Ghorbanzadeh et al. [42] proposed an open landslide dataset, referred to as Landslide4Sense. They used Landslide4Sense to evaluate the landslide detection performance of eleven deep learning-based segmentation models. In the landslide detection community, there is a lack of benchmark datasets and methods. They successfully filled this gap, which provides a great opportunity to attract more scholars in both computer vision and earth science to solve the landslide detection problem. We can see that the existing deep learning models for landslide detection are mainly based on convolutional neural networks and multilayer perceptron.

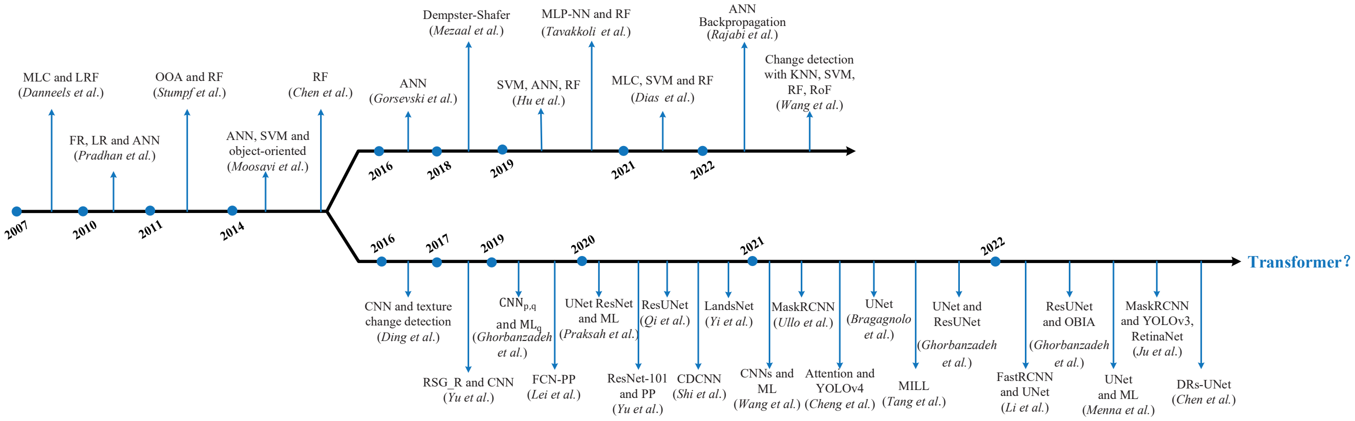

MLC and LRF (*Danneels et al.*)
FR, LR and ANN (*Pradhan et al.*)
OOA and RF (*Stumpf et al.*)
ANN, SVM and object-oriented (*Moosavi et al.*)
RF (*Chen et al.*)
2007
2010
2011
2014
ANN (*Gorsevski et al.*)
Dempster-Shafer (*Mezaal et al.*)
SVM, ANN, RF (*Hu et al.*)
MLP-NN and RF (*Tavakkoli et al.*)
MLC, SVM and RF (*Dias et al.*)
ANN Backpropagation (*Rajabi et al.*)
Change detection with KNN, SVM, RF, RoF (*Wang et al.*)
2016
2018
2019
2021
2022
2016
2017
2019
2020
2021
2022
Transformer?
CNN and texture change detection (*Ding et al.*)
RSG_R and CNN (*Yu et al.*)
CNN$_{p,q}$ and ML$_q$ (*Ghorbanzadeh et al.*)
FCN-PP (*Lei et al.*)
UNet ResNet and ML (*Praksah et al.*)
ResNet-101 and PP (*Yu et al.*)
ResUNet (*Qi et al.*)
CDCNN (*Shi et al.*)
LandsNet (*Yi et al.*)
CNNs and ML (*Wang et al.*)
MaskRCNN (*Ullo et al.*)
Attention and YOLOv4 (*Cheng et al.*)
UNet (*Bragagnolo et al.*)
MILL (*Tang et al.*)
UNet and ResUNet (*Ghorbanzadeh et al.*)
FastRCNN and UNet (*Li et al.*)
ResUNet and OBIA (*Ghorbanzadeh et al.*)
UNet and ML (*Menna et al.*)
MaskRCNN and YOLOv3, RetinaNet (*Ju et al.*)
DRs-UNet (*Chen et al.*)

Figure 1. The development of automatic landslide detection.

Recently, Transformer-based deep learning models have achieved superior performance in many computer vision tasks. Transformer is a type of attention mechanism, which was originally proposed for natural language processing tasks [43]. Whether Transformer is suitable for computer vision tasks is a long-standing problem. In 2020, Dosovitskiy [44] proposed vision Transformer (ViT), which is the first successful application of Transformer in image classification tasks. ViT is a great breakthrough. After ViT, many computer vision tasks have been improved by Transformer. A comprehensive review can be found in [45]. In the following, we survey some typical Transformer-based models in computer vision. Liu et al. [46] proposed Swin Transformer for image classification. Carion et al. [47]

proposed DETR for object detection. Xie et al. [48] proposed SegFormer for semantic segmentation. Chen et al. [49] studied the performance of applying vision Transformer to self-supervised learning. Bazi et al. [50] applied ViT to remote sensing image classification. The performance of applying Transformer for landslide detection remains an unsolved problem.

To fill this gap, this article applies Transformer to landslide detection. Our contributions are summarized as follows: (1) we compare a Transformer-based semantic segmentation model with CNN-based models; the Transformer-based model SegFormer is applied to identify coseismic landslides; (2) a new dataset for landslide detection is constructed; (3) we conduct extensive experiments to compare SegFormer with other popular CNN-based models.

The rest of this article is organized as follows. Section 2 shows the materials and methods for identifying coseismic landslides in 2017 Jiuzhaigou earthquake-triggered landslides; Section 3 presents the experimental results and analysis; Section 4 presents the discussion and conclusions.

2. Materials and Methods

The pipeline of landslide detection can be divided into three parts, i.e., preprocessing, processing and postprocessing, which are shown in Figure 2. (1) Preprocessing. In the preprocessing stage, we prepare data for model training and testing, i.e., image collection, landslide annotation, train/test partition and image cropping. (2) Processing. In the processing stage, we build the Transformer-based landslide detection model, i.e., developing, training and testing of the SegFormer model. (3) Postprocessing. In the postprocessing stage, the semantic segmentation results are improved by morphological operation and instance-wise bounding boxes.

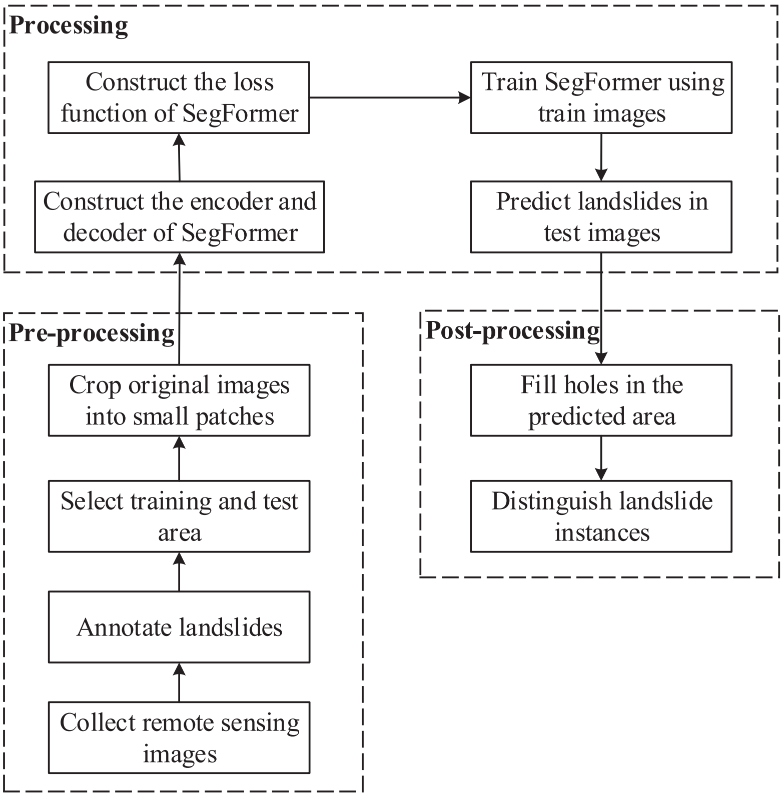

Processing
Construct the loss function of SegFormer
Train SegFormer using train images
Construct the encoder and decoder of SegFormer
Predict landslides in test images
Pre-processing
Post-processing
Crop original images into small patches
Fill holes in the predicted area
Select training and test area
Distinguish landslide instances
Annotate landslides
Collect remote sensing images

Figure 2. The pipeline of coseismic landslide detection using SegFormer.

2.1. Study Area

On 8 August 2017, an earthquake with a magnitude of 7.0 hit Jiuzhaigou County, Sichuan Province, China. The epicenter of the earthquake was located at 33.20° N, 103.82° E.

The earthquake triggered thousands of landslides, which brought enormous losses to the scenic area of the Jiuzhaigou World Natural Heritage site. Fast mapping of earthquake-triggered landslides is important for earthquake emergency search and rescue. In this paper, we study the automatic detection of coseismic landslides with remote sensing imagery. The study area is shown in Figure 3.

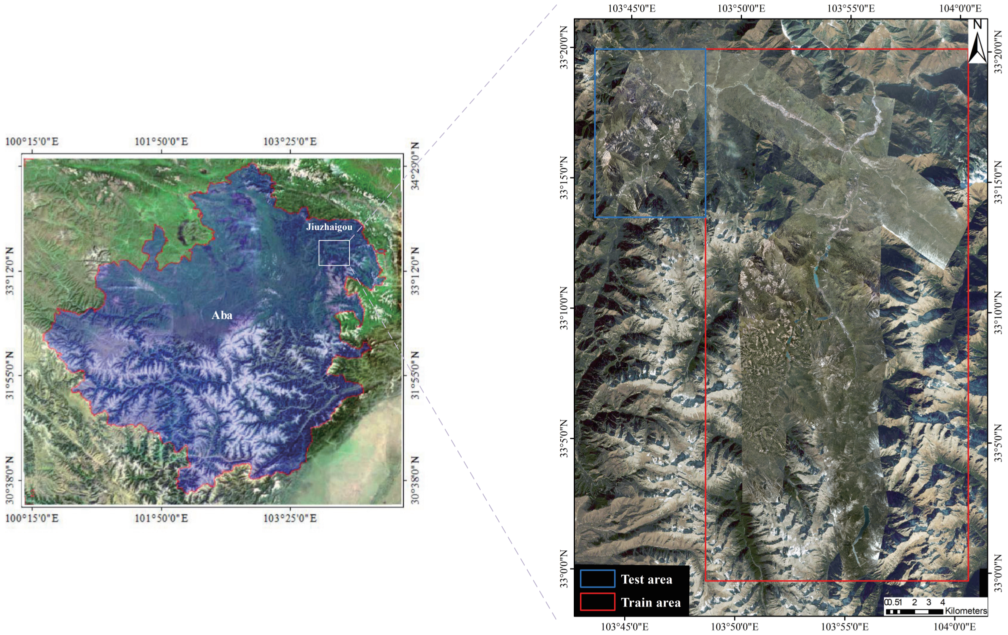


Figure 3. The study area is the Jiuzhaigou scenic spot, which was heavily damaged by the 2017 Jiuzhaigou earthquake.

2.2. *A New Dataset for Landslide Detection*

We propose a new dataset for landslide detection. The whole process can be divided into the following four stages.

2.2.1. Landslide Image Collection

We used an unmanned aerial vehicle to collect airborne remote sensing images of the Jiuzhaigou earthquake area on August 2017. The size of the remote sensing images was 132 K × 189 K pixels. The resolution was 0.2 m.

2.2.2. Landslide Image Annotation

We leverage supervised semantic segmentation to detect the boundaries of landslides in remote sensing images. Semantic segmentation performs pixel-wise classification, i.e., to decide whether each pixel belongs to a landslide or not. Supervised semantic segmentation requires ground-truth labels for training and testing. Errors in labels bring noise to models and lead to inaccurate evaluation results. Therefore, we need to map accurate boundaries of landslides.

Due to different degrees of vegetation cover and deviation in the shapefile, the existing landslide mapping results are inaccurate for the newly collected image. We manually mapped all the coseismic landslides of the airborne remote sensing image. The number of annotated landslides was 1898. The manual landslide mapping was conducted in Arcgis 10.7. The mapping results were the boundaries of landslides, which were stored in a shapefile.

2.2.3. Train/Test Partition

In practice, the generalization ability of many landslide models is poor. One reason is that the training set and test set are not independent. The random division strategy is widely used to prepare training and test datasets, in which samples are randomly divided. Before this, remote sensing data are divided into small cells. Grid clipping and sliding window are two common image cropping strategies. By grid clipping, remote sensing data are divided into grid cells. By sliding window, remote sensing data are divided into cells, which allows a portion of overlap. Due to geographical correlation and overlap, adjacent grid cells are not independent. If grid cells are randomly divided into a training set and test set, there are strong correlations between them. The correlation in the training and test set will lead to overfitting, which results in a model with high prediction accuracy on training and test datasets but poor generalization performance in practical application.

To address the overfitting issue, we divided the study area into two disjoint parts, i.e., the training area and test area (see Figure 3). The image patches in the training area were used to develop the landslide segmentation model. The image patches in the test area were used to test the performance of the model.

2.2.4. Landslide Image Cropping

Deep learning-based landslide detection models require large GPU memory and extensive computation. The size of the original remote sensing image is too large to directly feed into CNN- or Transformer-based landslide detection models. It should be divided into small patches with appropriate size. With the improvement of image resolution, higher-resolution remote sensing images provide more accurate details about landslides. Meanwhile, an image patch with a fixed size covers a small geographical area. In this case, a small patch size cannot cover large-scale landslides, while a large patch size leads to GPU memory overflow. There is a need to balance these factors. By experimental comparison, the patch size of 2048 × 2048 pixels achieves a good balance among landslide coverage, GPU usage, and landslide detection accuracy.

During the image cropping process, a landslide may be divided into several parts, and scattered in adjacent image patches. A partial landslide is even harder to detect, which leads to a decline in prediction accuracy. To address this issue, we adopt different cropping strategies for the training and test images. On the one hand, the training image is cropped by a center clip. That is, for each landslide, we crop an image patch centered on the landslide. The center clip strategy ensures that each landslide lies in at least one image patch, which makes full use of all the landslides of the training area. In deep learning, center clip is a widely used data augmentation operation. Thus, the overlaps among the training image patches usually have positive effects for training deep learning models. On the other hand, we adopt grid clip to avoid overlaps in test image patches, and to achieve a fair comparison of different landslide detection models. Finally, we obtain a new landslide detection dataset with 1295 and 387 samples in the training set and test set, respectively. The train/test ratio is 77% to 23%.

2.3. SegFormer

SegFormer [48] is a Transformer-based semantic segmentation model. The framework of SegFormer is shown in Figure 4, which consists of an encoder and a decoder.

2.3.1. SegFormer Encoder

The encoder of SegFormer has four Transformer blocks, which learn hierarchical feature maps. Similar to Vision Transformer (ViT) [44], each Transformer block can be divided into three main parts, i.e., patch embedding layer→attention layer→position embedding layer. The structure and improvements of each part are shown as follows.

(1) Patch embedding layer. The functionality of patch embedding layer is to divide the image into small patches and convert them into embedding vectors. Conventional Vision Transformer divides the input image into non-overlapping patches, which destroys the local continuity around those patches. For example, in Figure 5a, a landslide may be cut into several pieces and scattered in adjacent image patches. SegFormer introduces overlapped patch embedding and merging to address this issue.

In this study, landslide detection prefers overlapping patch embedding, since a remote sensing image captures the optical information of a whole area.

The overlapped patch embedding layer is implemented by a convolution layer, i.e., "nn.Conv2D" in PyTorch. The overlap ratio is controlled by the stride of the convolution. To learn hierarchical features of high and low resolution, SegFormer constructs four Transformer blocks $\{T_1, T_2, T_3, T_4\}$, whose feature maps are of size $\frac{H}{2^{i+1}} \times \frac{H}{2^{i+1}} \times 2^{i+5} (i \in \{1,2,3,4\})$. The kernel sizes and strides of these blocks are $\{7,3,3,3\}$ and $\{4,2,2,2\}$, respectively.

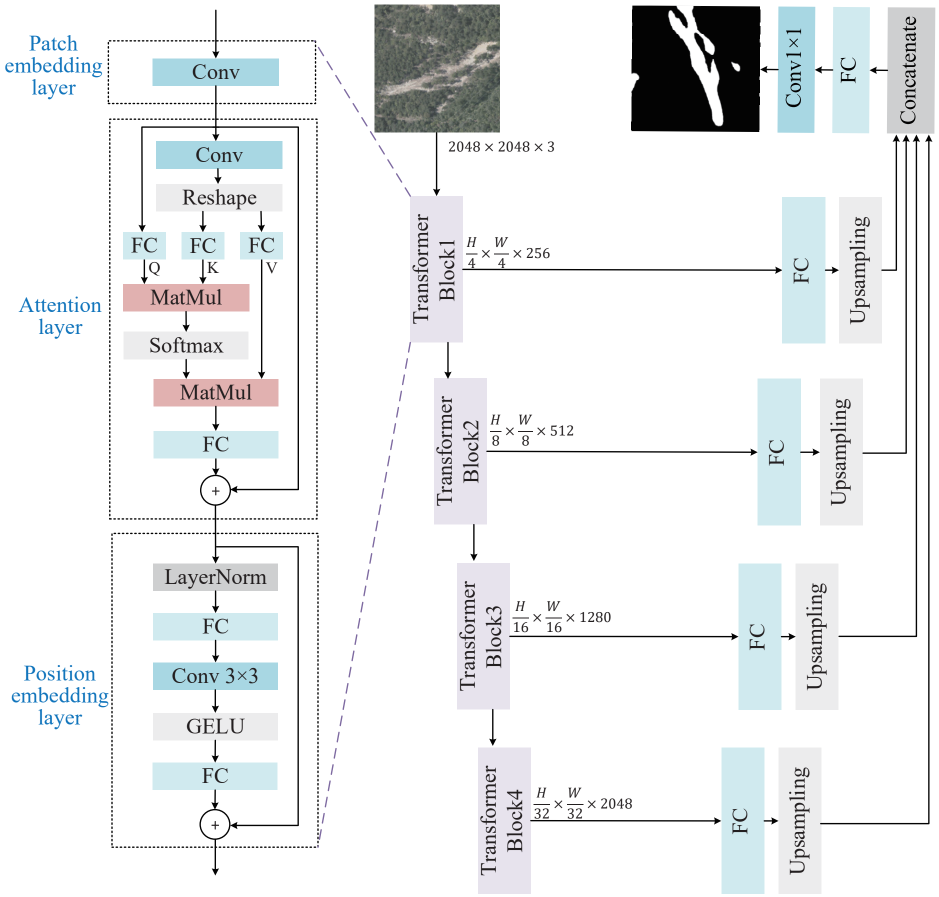
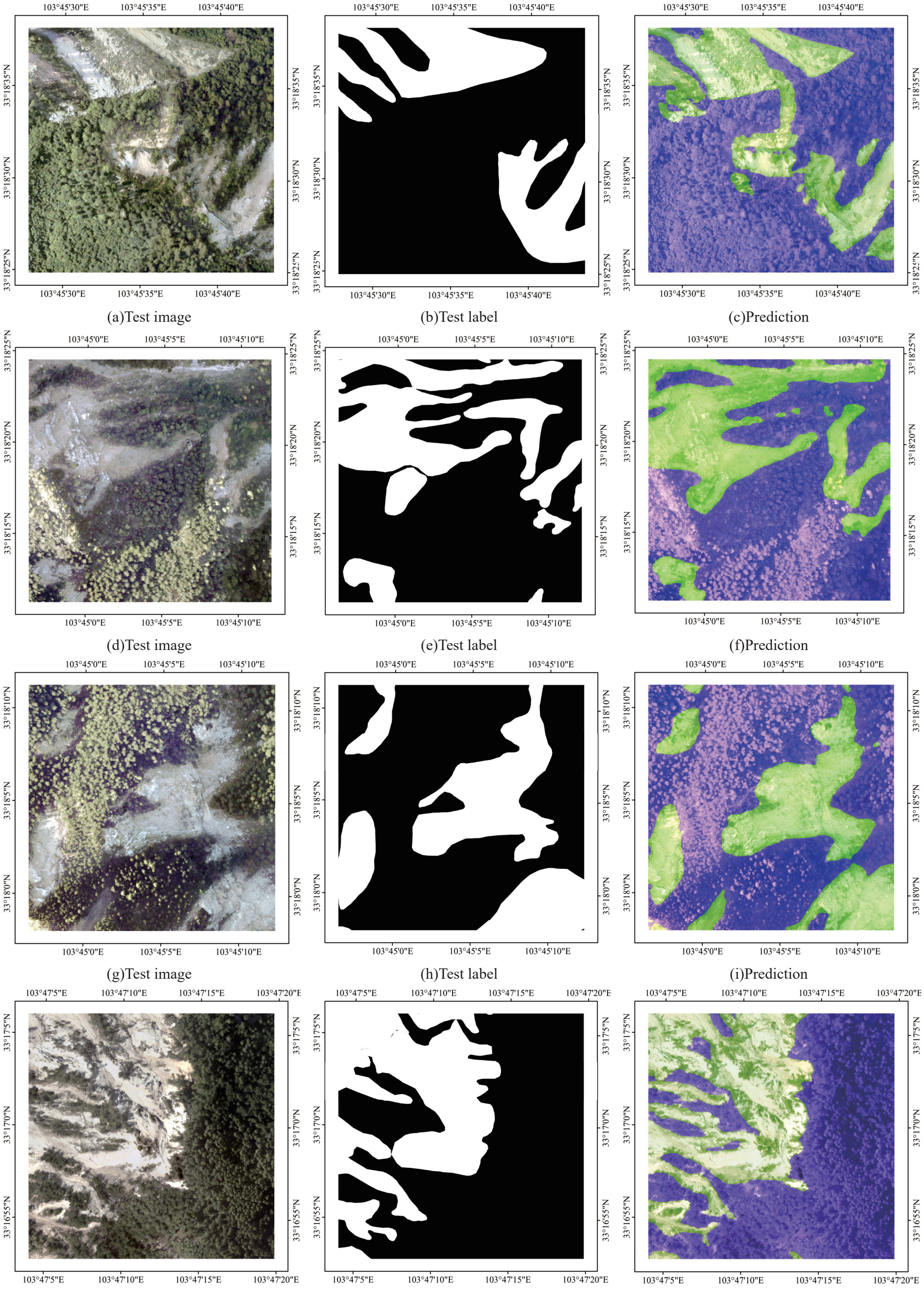


Figure 4. The architecture of SegFormer.

(2) Attention layer. The attention layer of ViT is a multi-head self-attention module (MSA). MSA plays a central role in capturing dependencies among image patches (or embedding vectors), i.e., global dependencies. However, ViT suffers from an extensive computational burden. To address this issue, SegFormer introduces sequence reduction to reduce the number of embedding vectors, which is referred to as efficient multi-head self-attention (EMSA).

EMSA reduces the number of embedding vectors from N to $\frac{N}{r}$, where $N = H \times W$, and r is a hyperparameter called the reduction ratio. EMSA is implemented by a convolution layer, i.e., "nn.Conv2D" in PyTorch. First, a feature map of size $H \times W \times C$ is reshaped to $\frac{N}{r} \times r \cdot C$. The number of embedding vectors is reduced to $\frac{N}{r}$, while the length of the embedding vector is increased to $r \cdot C$. Second, a fully connected layer is used to reduce each embedding vector back to size C. Third, conventional MSA is performed on the reduced feature map, which is shown in Equation (1). The size of Q, K, V is $\frac{N}{r} \times C$.

$$Attention(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{C}})V. \tag{1}$$



(a)Test image (b)Test label (c)Prediction

(d)Test image (e)Test label (f)Prediction

(g)Test image (h)Test label (i)Prediction

(j)Test image (k)Test label (l)Prediction

Figure 5. Visualization results of image patches.

(3) Position embedding layer. In ViT, the positional information of each patch is explicitly encoded and appended to the patch embedding vector. However, it is difficult to directly encode positional information to different levels of hierarchical feature maps. SegFormer introduces a 3 × 3 convolution for implicitly learning patch positional

information. A skip connection is used to add positional information to the feature map. The implementation details of positional embedding are shown in Equation (2).

$$\begin{aligned} X_p &= \mathrm{Linear}(X) \\ X_p &= \mathrm{Conv}_{3\times 3}(X_p) \\ X_p &= \mathrm{GELU}(X_p) \\ X &= \mathrm{Linear}(X_p) + X \end{aligned} \tag{2}$$

2.3.2. SegFormer Decoder

The decoder of SegFormer consists of multilayer perceptrons (MLP). First, each feature map $b_i (i \in \{1,2,3,4\})$ goes through a fully connected layer to unify the channel dimension, and then it is upsampled to $\frac{H}{4} \times \frac{W}{4}$ using bilinear interpolation. Second, the feature maps are concatenated together. Third, an MLP layer is adopted to fuse the channel dimensions of the concatenated features. Finally, another fully connected layer (1 × 1 Conv) takes the fused feature to predict the segmentation mask. The size of prediction P is $\frac{H}{4} \times \frac{W}{4} \times N_c$. Equation (3) shows the implementation details of the SegFormer decoder.

$$\begin{aligned} F_i &= \mathrm{Linear}(M_i),\ \text{where } i \in \{1,2,3,4\} \\ F_i &= \mathrm{Resize}_{\frac{H}{4}\times\frac{W}{4}}(F_i) \\ F &= \mathrm{Concat}([F_1, F_2, F_3, F_4]) \\ F &= \mathrm{Linear}(F) \\ P &= \mathrm{Conv}_{1\times 1}(F) \end{aligned} \tag{3}$$

2.3.3. Morphological-Based Fine Tuning

A semantic segmentation model such as SegFormer is a pixel-level classification model. It usually fails to consider the morphological characteristics of the objects. There are some invalid holes in the predicted object region. To solve this problem, we employ a morphological operation to eliminate the holes. The function "findcontours" of OpenCV is used to find the external contour of each landslide. Then, the holes are removed by filling the region within the external contour. Finally, we draw a rectangle bounding box for each contour based on its center, width and height. Thus, the landslide instances are obtained. In Section 3.6, we will show the visualization results of the morphological operation.

To sum up, the training process of SegFormer is shown in Algorithm 1. The number of iterations is 8000. The batch size is 2.

Algorithm 1 Landslide recognition using SegFormer.

```
Input: (1) Training images and their labels $S_1$. (2) Test images $S_2$.
Output: Predicted mask images $M$ for the test images.
Initialize SegFormer by pre-trained model
#Training
for $i = 1$ to 8000 do
    Randomly select 2 images
    Train the encoder of SegFormer
    Train the decoder of SegFormer
    Calculate the cross-entropy loss
    Backpropagation
    Update the parameters of SegFormer
end for
#Test
Use the trained SegFormer model to predict the test images
Use morphological method for fine tuning the predicted mask images
```

3. Results and Discussion

Extensive experiments have been conducted to compare different image segmentation methods for landslide detection. The experimental configurations and results are shown as follows.

3.1. Experimental Configurations

The experimental environment is a cloud GPU platform, which is called "AI Studio https://aistudio.baidu.com/aistudio/index?lang=en (accessed on 1 April 2022)". SegFormer and the comparison models are implemented in PaddleSeg https://github.com/PaddlePaddle/PaddleSeg (accessed on 20 April 2022). The comparison models are well-known semantic segmentation models, including HRNet [51], DeepLabv3 [52], Attention-UNet [53], U^2Net [54] and FastSCNN [55]. Due to limited GPU memory, the B4 model of SegFormer is adopted, and the input image is resized to 1024 × 1024 × 3. The loss function is the cross-entropy loss. The optimization algorithm is AdamW, whose weight decay factor is 0.01. The learning rate is initialized to 0.00006. If the loss increases, the leaning rate will be decreased by a factor of 0.9. The number of iterations is 8000. The training and test process are shown in Algorithm 1.

3.2. Evaluation Metrics

To make a systematic comparison, we adopt five widely used metrics to measure the performance of all the semantic segmentation models, i.e., mIoU, precision, recall, F1 score and accuracy. Equation (4) shows the definitions of these metrics.

$$
\begin{aligned}
\mathrm{IoU} &= \frac{\mathrm{TP}}{(\mathrm{FN}+\mathrm{FP}+\mathrm{TP})} \\
\mathrm{mIoU} &= \frac{1}{2}(\mathrm{IoU}_{\mathrm{landslides}}+\mathrm{IoU}_{\mathrm{backgrounds}}) \\
\mathrm{Precision} &= \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})} \\
\mathrm{Recall} &= \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \\
\mathrm{Accuracy} &= \frac{\mathrm{TP}+\mathrm{TN}}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})} \\
\mathrm{F1\_score} &= \frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}
\end{aligned}
\tag{4}
$$

where TP (True Positive), FP (False Positive), TN (True Negative) and FN (False Negative) can be computed by the confusion matrices. The overall performance is evaluated by the average IoU, precision, recall, accuracy and F1 score of the landslide and background classes. For all the metrics in Equation (4), higher is better.

3.3. Comparison of Landslide Detection Accuracy

Table 1 shows the average accuracies of all the semantic segmentation models. The highest mIoU of the comparison models is 0.734, which is achieved by HRNet. SegFormer achieves the highest mIoU of 0.75, which improves the highest mIoU of the comparison methods by 2.2%. SegFormer also achieves the highest F1 score and precision. Therefore, SegFormer achieves a good balance between precision and recall. SegFormer also achieves the highest pixel-wise classification accuracy, i.e., the pixel-wise classification error rate is reduced by 14%.

We experimentally compare the per-class segmentation accuracies of all the segmentation models for landslide detection. The experimental results are shown in Table 2. In remote sensing images, the number of pixels of a landslide is much less than the background. The landslide and background classes are highly imbalanced. The scores of background in Table 2 are close to each other. SegFormer achieves the highest IoU, recall and F1 score for the background class. In the following, the focus is on comparing the detection accuracy of the landslide class. SegFormer achieves the highest IoU for the landslide class, which improves the second highest IoU of landslides by 5%. SegFormer also achieves the highest F1 score and precision for the detection of landslides, which improves the second highest F1 score of landslides by 3%. Although SegFormer achieves the second highest recall, it achieves the highest F1 score. SegFormer achieves a good balance between precision and recall for the detection of landslides.

Table 1. Comparison of the average segmentation accuracy.

Models	mIoU	Precision	Recall	Accuracy	F1 Score
AttentionUNet	0.665	0.730	0.833	0.920	0.769
DeepLabv3	0.675	0.756	0.808	0.930	0.779
FastSCNN	0.690	0.765	0.825	0.933	0.791
LandsNet	0.690	0.749	**0.865**	0.933	0.792
U^2Net	0.696	0.775	0.823	0.936	0.796
HRNet	0.734	0.819	0.839	0.949	0.828
SegFormer	**0.750**	**0.850**	0.833	**0.956**	**0.841**

To sum up, SegFormer achieves the best comprehensive detection accuracy. HRNet comes second in landslide detection, which is also the best model among all the comparison models. We seek to interpret these results from two perspectives, i.e., long-term dependency and high resolution. First, SegFormer uses Transformer for automatic feature extraction. In contrast, all the comparison models use CNN to learn feature maps. Transformer can learn long-term dependencies [43]. Its receptive field is larger than CNN. Thus, SegFormer performs better than all the comparison models. Second, HRNet not only makes use of multi-resolution features, but also maintains high-resolution features. Thus, HRNet achieves the best performance among all the comparison models. SegFormer also makes use of multi-resolution hierarchical features, which is better than ViT. Finally, remote sensing images are logically a whole. There are strong dependencies between patches of remote sensing images, i.e., long-term dependencies. There is no doubt that a high resolution is helpful for identifying landslides in remote sensing images. Therefore, in landslide detection from remote sensing imagery, maintaining a large receptive field and high-resolution feature maps are key for SegFormer to achieve the best performance.

Table 2. Comparison of segmentation accuracy for each class (landslide and background).

Models	Class	IoU	Precision	Recall	F1 Score
AttentionUNet	background	0.921	0.978	0.941	0.959
	landslide	0.408	0.483	0.725	0.580
DeepLabv3	background	0.926	0.971	0.952	0.962
	landslide	0.425	0.541	0.663	0.596
FastSCNN	background	0.936	0.976	0.941	0.966
	landslide	0.445	0.555	0.692	0.616
LandsNet	background	0.928	0.983	0.944	0.963
	landslide	0.451	0.514	0.786	0.622
U^2Net	background	0.939	0.975	0.961	0.968
	landslide	0.454	0.574	0.684	0.624
HRNet	background	0.946	0.975	0.969	0.972
	landslide	0.521	0.662	0.709	0.685
SegFormer	background	**0.954**	0.974	**0.978**	**0.976**
	landslide	**0.545**	**0.725**	0.687	**0.705**

3.4. Comparision of Different Image Resolutions

We study the performance of Segformer under different image resolutions and different input shapes. Table 3 shows the experimental results. Notice that the resolution is 0.2 m. The images of 0.5 m and 1 m are obtained by resampling the original remote sensing image. When the resolution is 0.5 m and the input size is 1024 × 1024, SegFormer achieves the highest mIoU and F1 score. Therefore, we need to balance the image resolution and the input size of landslide detection models.

Table 3. Comparison of segmentation accuracy for SegFormer under different image resolutions and input sizes.

Resolution	Size	mIoU	Precision	Recall	Accuracy	F1 Score
0.2 m	2048 × 2048	0.750	0.850	0.833	0.956	0.841
	1024 × 1024	0.731	0.833	0.842	0.935	0.827
	512 × 512	0.695	0.761	**0.850**	0.932	0.797
0.5 m	2048 × 2048	0.731	**0.858**	0.797	0.950	0.824
	1024 × 1024	**0.753**	0.850	0.837	0.935	**0.844**
	512 × 512	0.728	0.808	0.843	0.946	0.824
1 m	2048 × 2048	0.727	0.842	0.801	0.954	0.820
	1024 × 1024	0.735	0.835	0.821	**0.959**	0.828
	512 × 512	0.743	0.826	0.846	0.956	0.836

3.5. Visualization Results

Figure 5 shows the visualization results of SegFormer. For each row of Figure 5, the images from left to right represent the remote sensing image patch, landslide labels and predicted landslides, respectively. The green areas in the right-hand images represent the predicted landslides. We can see that most of the landslides have been correctly identified. The predicted landslide areas match well with the landslide labels. The boundaries of landslides are clear. Due to image cropping, a few landslide fragments at the image boundaries are missed. There are a few holes in the predicted landslide area. Some landslides near to each other are falsely connected.

To obtain an overview of the predicted landslides, the image patches are combined into a single image according to their geographical coordinates. Figure 6 shows the landslide labels, which have been checked by an earth scientist. Figures 7 and 8 show the visualization results of LandsNet and SegFormer, respectively. We can see that the landslide distribution of ground-truth and prediction is close. In the lower left of Figures 7 and 8, some bare rocks are falsely recognized as landslides. Compared with LandsNet, SegFormer successfully excludes significantly more bare rocks. In the upper right of Figures 7 and 8, some roads are falsely recognized as landslides. Compared with LandsNet, SegFormer reports less invalid landslides in this area. The predicted landslide boundaries of SegFormer are more accurate than LandsNet, e.g., the landslides in the middle of Figures 6 and 8.

3.6. Postprocessing

We employ morphological methods to fine tune the predicted landslide areas. For each row of Figure 9, the images from left to right represent the remote sensing image patch, predicted landslides and the fine-tuned prediction of landslides, respectively. It can be seen that the holes in the predicted landslides are filled. The ResU-Net-OBIA [39] method uses the normalized difference vegetation index, length-to-width-ratio and rule-based expert knowledge to fine tune the segmentation results. In contrast, the proposed method improves the semantic segmentation results without expert knowledge and extra data, which is more automatic. Furthermore, a common drawback of semantic segmentation methods is that they cannot identify object instances. In this article, each instance of the landslide has been automatically detected, i.e., the rectangular bounding boxes in the postprocessed images in Figure 9.

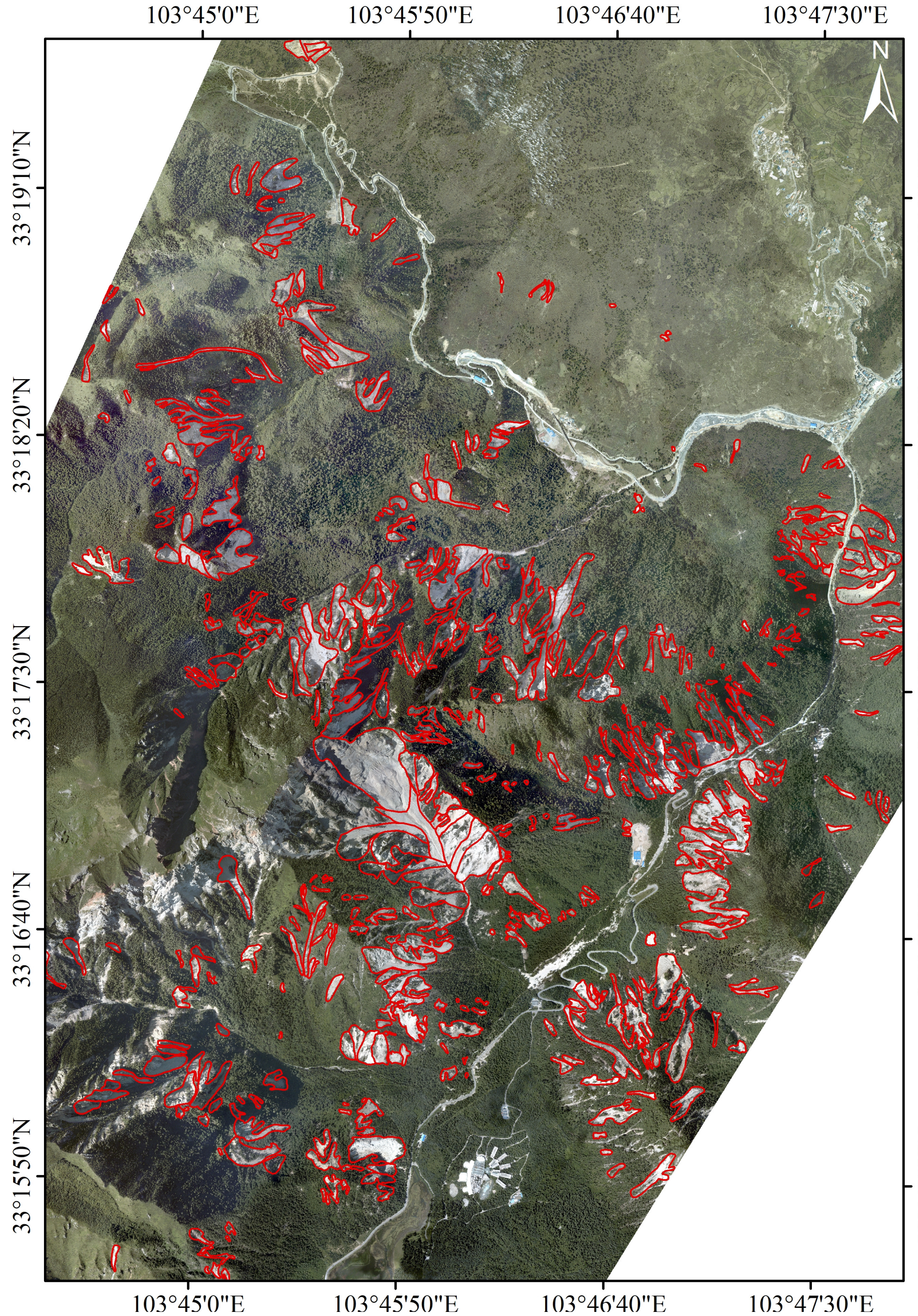


Figure 6. Visualization results of the manually annotated landslide labels. Red lines are the boundaries of landslides.

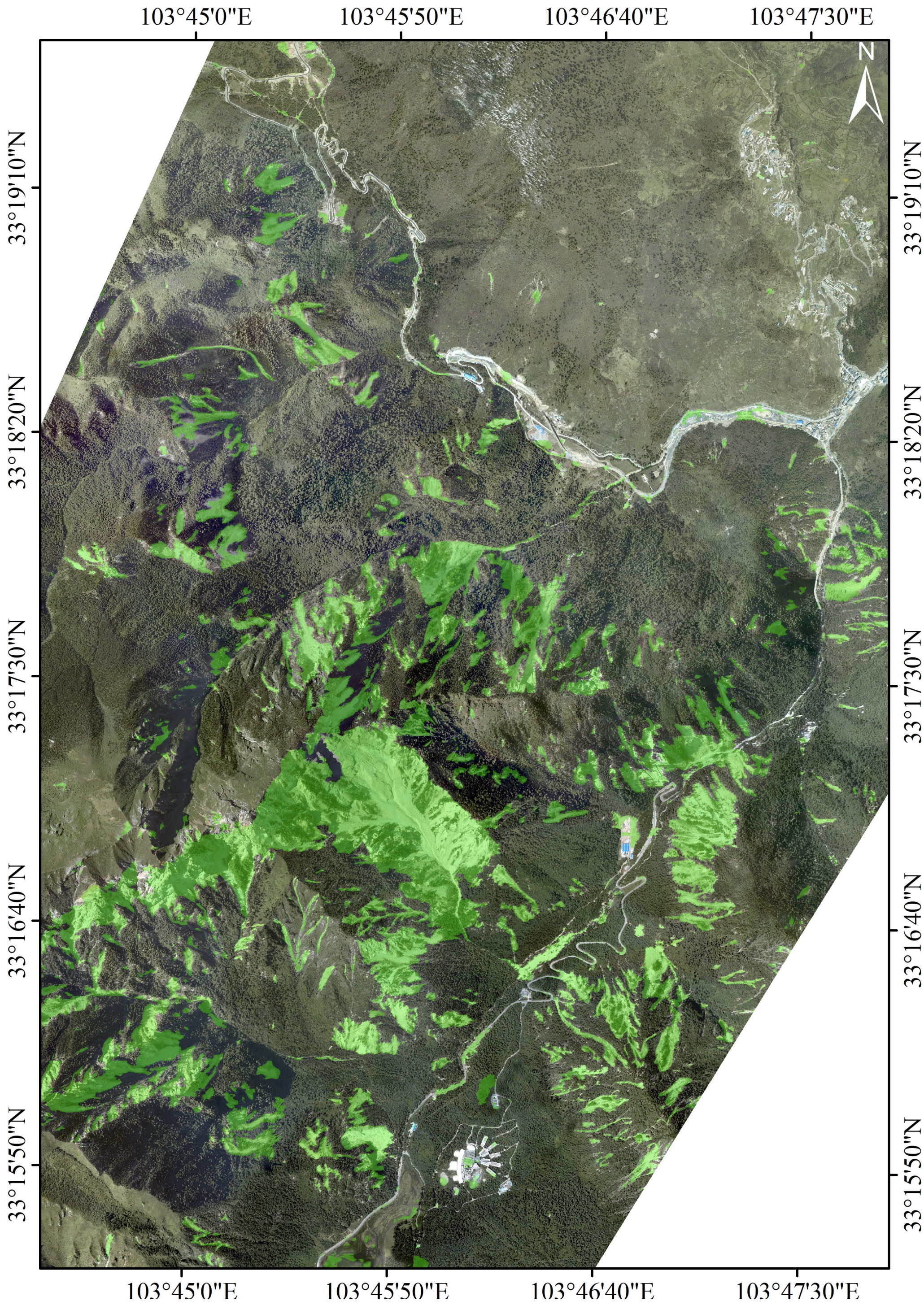


Figure 7. Visualization results of the LandsNet model. Green areas represent the predicted landslides.

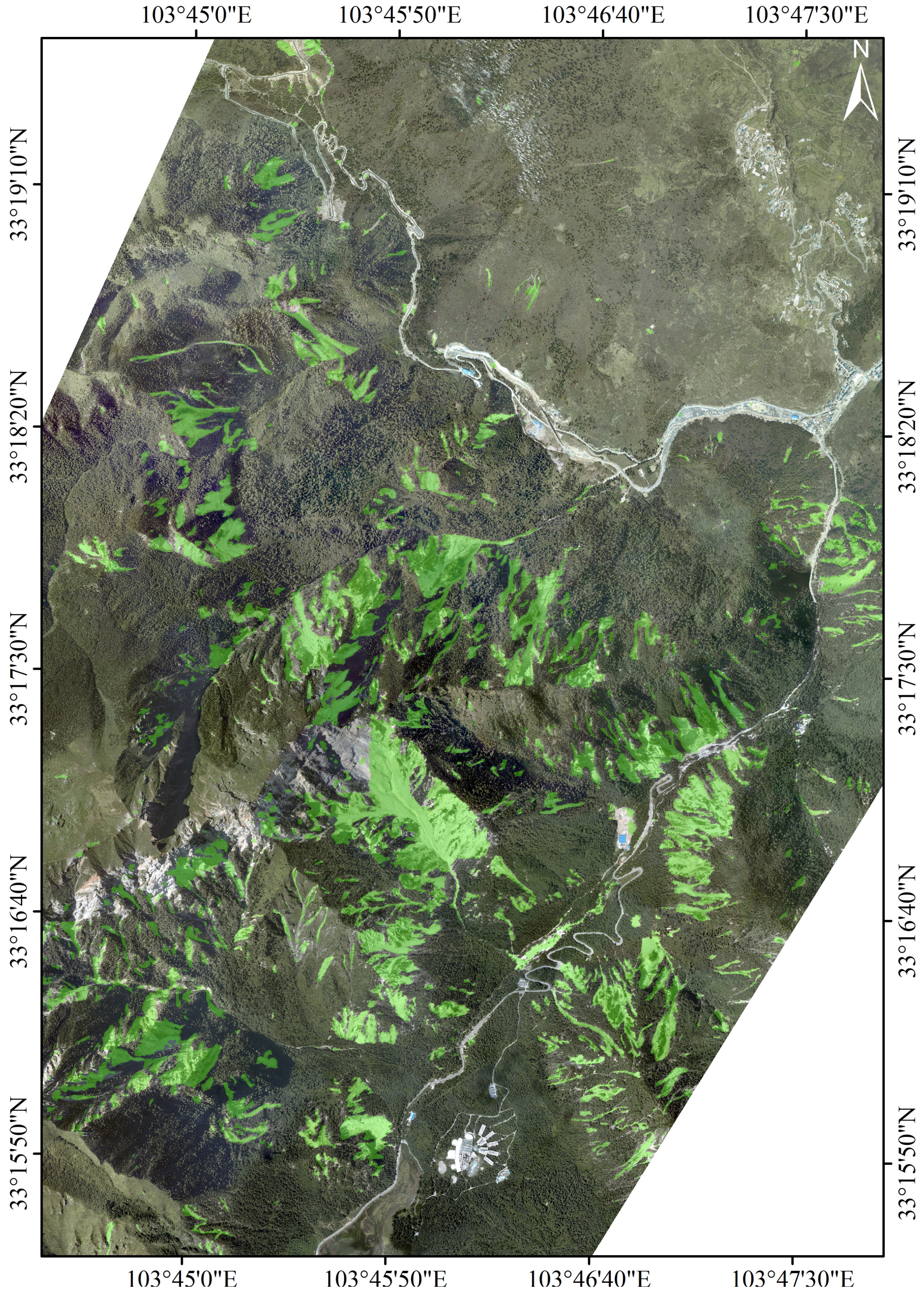
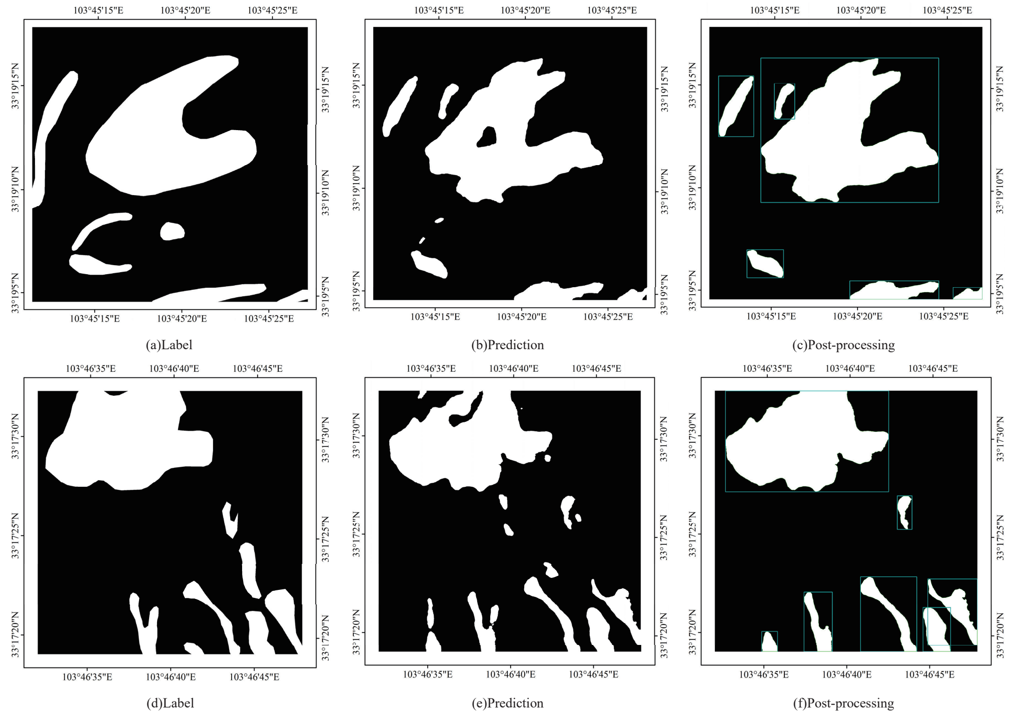


Figure 8. Visualization results of the SegFormer model. Green areas represent the predicted landslides.



(a)Label (b)Prediction (c)Post-processing

(d)Label (e)Prediction (f)Post-processing

Figure 9. Visualization results of postprocessing.

4. Conclusions

We studied the performance of using Transformer to identify coseismic landslides. The semantic segmentation model SegFormer was applied to detect coseismic landslides, which is based on Transformer. High-resolution remote sensing images were collected and used to create a new landslide detection dataset. The training and test datasets were strictly separated, i.e., they were located in two distinct areas. In contrast, conventional random division of train/test datasets suffers from overfitting and inflated accuracies, since there are strong correlations between the remote sensing image patches of the training and test datasets.

Extensive experiments have been conducted to compare SegFormer with a landslide detection model, LandsNet, and many popular CNN-based semantic segmentation models, including HRNet, DeepLabv3, Attention-UNet, U^{2}Net and FastSCNN. SegFormer outperforms LandsNet, e.g., SegFormer improves the mIoU, precision and F1 score by 8.7%, 13.5% and 6.2%, respectively. Both mean accuracy and per-class accuracy were evaluated. The evaluation metrics included Iou, mIoU, precision, recall, accuracy and F1 score. The experimental results demonstrated that SegFormer outperformed all the competing models. SegFormer improved the mIoU, IoU and F1 score of landslides by 2.2%, 5% and 3%, respectively. Visualization results also showed that SegFormer successfully identified most of the landslides. Therefore, SegFormer achieved a good balance between precision and recall. Moreover, Transformer-based model SegFormer outperformed other CNN-based models. A possible reason is that Transformer is capable of capturing long-term dependencies within each image patch. The reception field of Transformer is larger than CNN. A large reception field is the key to developing an accurate landslide detection model. In addition, semantic segmentation does not distinguish object instances. To distinguish different landslide instances, we used image processing operations to find the bounding box of every landslide. Morphological methods could improve the landslide detection performance by removing holes in the semantic segmentation results.

A landslide is a type of natural phenomenon, which is challenging to identify by experts and machines. A drawback of SegFormer is that it requires large GPU memory. We need to balance the image resolution and the input size of image patches. In future works, it would be worthwhile to develop a lightweight Transformer model for landslide detection, which would enable larger input images. In addition, landslides threaten the property and lives of the people under the slope. We need to understand how landslide detection models make decisions. The interpretability of landslide detection models is an important research direction.

Author Contributions: Conceptualization, X.T. and Z.T.; methodology, X.T. and Z.T.; software, Z.T. and Y.W.; validation, X.T. and Z.T.; formal analysis, X.T.; investigation, Z.T. and X.T.; resources, X.F. and X.T.; data curation, Z.T., X.T. and X.F.; writing—original draft preparation, Z.T. and X.T.; writing—review and editing, X.T.; visualization, Z.T.; supervision, X.F.; project administration, X.T., M.L. and D.L.; funding acquisition, M.L., D.L. and X.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the China Postdoctoral Science Foundation (Grant No. 2021M690024), Research Funding of SKLGP (Grant No. SKLGP2019Z014), National Natural Science Foundation of China (Grant No. 62172060) and Sichuan Science and Technology Program (Grant No. 2022YFG0316).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
LiDAR	Light Detection and Ranging
InSAR	Interferometric Synthetic Aperture Radar
MSA	Multihead Self-Attention
MLP	Multilayer Perceptron

References

1. Xu, Q.; Dong, X.; Li, W. Integrated space-air-ground early detection, monitoring and warning system for potential catastrophic geohazards. *Geomat. Inf. Sci. Wuhan Univ.* **2019**, *44*, 957–966.
2. Huang, R.; Li, W. Analysis of the geo-hazards triggered by the 12 May 2008 Wenchuan Earthquake, China. *Bull. Eng. Geol. Environ.* **2009**, *68*, 363–371. [CrossRef]
3. Li, Z.; Shi, W.; Lu, P.; Yan, L.; Wang, Q.; Miao, Z. Landslide mapping from aerial photographs using change detection-based Markov Random Field. *Remote Sens. Environ.* **2016**, *187*, 76–90. [CrossRef]
4. Wang, X.; Fan, X.; Xu, Q.; Du, P. Change detection-based co-seismic landslide mapping through extended morphological profiles and ensemble strategy. *ISPRS J. Photogramm. Remote Sens.* **2022**, *187*, 225–239. [CrossRef]
5. Zhang, L.; Dai, K.; Deng, J.; Ge, D.; Liang, R.; Li, W.; Xu, Q. Identifying potential landslides by stacking-InSAR in southwestern China and its performance comparison with SBAS-InSAR. *Remote Sens.* **2021**, *13*, 3662. [CrossRef]
6. Xu, Q.; Guo, C.; Dong, X.; Li, W.; Lu, H.; Fu, H.; Liu, X. Mapping and characterizing displacements of landslides with InSAR and airborne LiDAR technologies: A case study of Danba county, southwest China. *Remote Sens.* **2021**, *13*, 4234. [CrossRef]
7. Pawluszek, K. Landslide features identification and morphology investigation using high-resolution DEM derivatives. *Nat. Hazards* **2019**, *96*, 311–330. [CrossRef]
8. Syzdykbayev, M.; Karimi, B.; Karimi, H.A. Persistent homology on LiDAR data to detect landslides. *Remote Sens. Environ.* **2020**, *246*, 111816. [CrossRef]
9. Danneels, G.; Pirard, E.; Havenith, H.B. Automatic landslide detection from remote sensing images using supervised classification methods. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Barcelona, Spain, 23–28 July 2007; pp. 3014–3017.
10. Pradhan, B.; Lee, S. Delineation of landslide hazard areas on Penang Island, Malaysia, by using frequency ratio, logistic regression, and artificial neural network models. *Environ. Earth Sci.* **2010**, *60*, 1037–1054. [CrossRef]
11. Stumpf, A.; Kerle, N. Object-oriented mapping of landslides using Random Forests. *Remote Sens. Environ.* **2011**, *115*, 2564–2577. [CrossRef]
12. Moosavi, V.; Talebi, A.; Shirmohammadi, B. Producing a landslide inventory map using pixel-based and object-oriented approaches optimized by Taguchi method. *Geomorphology* **2014**, *204*, 646–656. [CrossRef]
13. Chen, W.; Li, X.; Wang, Y.; Chen, G.; Liu, S. Forested landslide detection using LiDAR data and the random forest algorithm: A case study of the Three Gorges, China. *Remote Sens. Environ.* **2014**, *152*, 291–301. [CrossRef]
14. Gorsevski, P.V.; Brown, M.K.; Panter, K.; Onasch, C.M.; Simic, A.; Snyder, J. Landslide detection and susceptibility mapping using LiDAR and an artificial neural network approach: a case study in the Cuyahoga Valley National Park, Ohio. *Landslides* **2016**, *13*, 467–484. [CrossRef]

15. Mezaal, M.R.; Pradhan, B. An improved algorithm for identifying shallow and deep-seated landslides in dense tropical forest from airborne laser scanning data. *Catena* **2018**, *167*, 147–159. [CrossRef]
16. Hu, Q.; Zhou, Y.; Wang, S.; Wang, F.; Wang, H. Improving the accuracy of landslide detection in "off-site" area by machine learning model portability comparison: a case study of Jiuzhaigou earthquake, China. *Remote Sens.* **2019**, *11*, 2530. [CrossRef]
17. Tavakkoli Piralilou, S.; Shahabi, H.; Jarihani, B.; Ghorbanzadeh, O.; Blaschke, T.; Gholamnia, K.; Meena, S.R.; Aryal, J. Landslide detection using multi-scale image segmentation and different machine learning models in the higher himalayas. *Remote Sens.* **2019**, *11*, 2575. [CrossRef]
18. Dou, J.; Yunus, A.P.; Bui, D.T.; Merghadi, A.; Sahana, M.; Zhu, Z.; Chen, C.W.; Han, Z.; Pham, B.T. Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan. *Landslides* **2020**, *17*, 641–658. [CrossRef]
19. Dias, H.C.; Sandre, L.H.; Alarcón, D.A.S.; Grohmann, C.H.; Quintanilha, J.A. Landslide recognition using SVM, Random Forest, and Maximum Likelihood classifiers on high-resolution satellite images: A case study of Itaóca, southeastern Brazil. *Braz. J. Geol.* **2021**, *51*, e20200105. [CrossRef]
20. Rajabi, A.M.; Khodaparast, M.; Mohammadi, M. Earthquake-induced landslide prediction using back-propagation type artificial neural network: Case study in northern Iran. *Nat. Hazards* **2022**, *110*, 679–694. [CrossRef]
21. Ghorbanzadeh, O.; Blaschke, T.; Gholamnia, K.; Meena, S.R.; Tiede, D.; Aryal, J. Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. *Remote Sens.* **2019**, *11*, 196. [CrossRef]
22. Tang, X.; Liu, M.; Zhong, H.; Ju, Y.; Li, W.; Xu, Q. MILL: Channel attention–based deep multiple instance learning for landslide recognition. *ACM Trans. Multimed. Comput. Commun. Appl.* **2021**, *17*, 76. [CrossRef]
23. Shi, W.; Zhang, M.; Ke, H.; Fang, X.; Zhan, Z.; Chen, S. Landslide recognition by deep convolutional neural network and change detection. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4654–4672. [CrossRef]
24. Ghorbanzadeh, O.; Crivellari, A.; Ghamisi, P.; Shahabi, H.; Blaschke, T. A comprehensive transferability evaluation of U-Net and ResU-Net for landslide detection from Sentinel-2 data (case study areas from Taiwan, China, and Japan). *Sci. Rep.* **2021**, *11*, 14629. [CrossRef] [PubMed]
25. Huang, F.; Tao, S.; Chang, Z.; Huang, J.; Fan, X.; Jiang, S.H.; Li, W. Efficient and automatic extraction of slope units based on multi-scale segmentation method for landslide assessments. *Landslides* **2021**, *18*, 3715–3731. [CrossRef]
26. Ju, Y.; Xu, Q.; Jin, S.; Li, W.; Su, Y.; Dong, X.; Guo, Q. Loess landslide detection using object detection algorithms in northwest China. *Remote Sens.* **2022**, *14*, 1182. [CrossRef]
27. Ullo, S.L.; Mohan, A.; Sebastianelli, A.; Ahamed, S.E.; Kumar, B.; Dwivedi, R.; Sinha, G.R. A new mask R-CNN-based method for improved landslide detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3799–3810. [CrossRef]
28. Ding, A.; Zhang, Q.; Zhou, X.; Dai, B. Automatic recognition of landslide based on CNN and texture change detection. In Proceedings of the Youth Academic Annual Conference of Chinese Association of Automation, Wuhan, China, 11–13 November 2016; pp. 444–448.
29. Yu, H.; Ma, Y.; Wang, L.; Zhai, Y.; Wang, X. A landslide intelligent detection method based on CNN and RSG_R. In Proceedings of the IEEE International Conference on Mechatronics and Automation, Takamatsu, Japan, 6–9 August 2017; pp. 40–44.
30. Lei, T.; Zhang, Y.; Lv, Z.; Li, S.; Liu, S.; Nandi, A.K. Landslide inventory mapping from bitemporal images using deep convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 982–986. [CrossRef]
31. Prakash, N.; Manconi, A.; Loew, S. Mapping landslides on EO data: Performance of deep learning models vs. traditional machine learning models. *Remote Sens.* **2020**, *12*, 346. [CrossRef]
32. Yu, B.; Chen, F.; Xu, C. Landslide detection based on contour-based deep learning framework in case of national scale of Nepal in 2015. *Comput. Geosci.* **2020**, *135*, 104388. [CrossRef]
33. Qi, W.; Wei, M.; Yang, W.; Xu, C.; Ma, C. Automatic mapping of landslides by the ResU-Net. *Remote Sens.* **2020**, *12*, 2487. [CrossRef]
34. Yi, Y.; Zhang, W. A new deep-learning-based approach for earthquake-triggered landslide detection from single-temporal RapidEye satellite imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 6166–6176. [CrossRef]
35. Wang, H.; Zhang, L.; Yin, K.; Luo, H.; Li, J. Landslide identification using machine learning. *Geosci. Front.* **2021**, *12*, 351–364. [CrossRef]
36. Cheng, L.; Li, J.; Duan, P.; Wang, M. A small attentional YOLO model for landslide detection from satellite remote sensing images. *Landslides* **2021**, *18*, 2751–2765. [CrossRef]
37. Bragagnolo, L.; Rezende, L.; da Silva, R.; Grzybowski, J. Convolutional neural networks applied to semantic segmentation of landslide scars. *Catena* **2021**, *201*, 105189. [CrossRef]
38. Li, H.; He, Y.; Xu, Q.; Deng, J.; Li, W.; Wei, Y. Detection and segmentation of loess landslides via satellite images: A two-phase framework. *Landslides* **2022**, *19*, 673–686. [CrossRef]
39. Ghorbanzadeh, O.; Shahabi, H.; Crivellari, A.; Homayouni, S.; Blaschke, T.; Ghamisi, P. Landslide detection using deep learning and object-based image analysis. *Landslides* **2022**, *19*, 929–939. [CrossRef]
40. Meena, S.R.; Soares, L.P.; Grohmann, C.H.; van Westen, C.; Bhuyan, K.; Singh, R.P.; Floris, M.; Catani, F. Landslide detection in the Himalayas using machine learning algorithms and U-Net. *Landslides* **2022**, *19*, 1209–1229. [CrossRef]

41. Chen, X.; Yao, X.; Zhou, Z.; Liu, Y.; Yao, C.; Ren, K. DRs-UNet: A deep semantic segmentation network for the recognition of active landslides from InSAR imagery in the three rivers region of the Qinghai–Tibet Plateau. *Remote Sens.* **2022**, *14*, 1848. [CrossRef]
42. Ghorbanzadeh, O.; Xu, Y.; Ghamis, P.; Kopp, M.; Kreil, D. Landslide4Sense: Reference benchmark data and deep learning models for landslide detection. *arXiv* **2022**, arXiv:2206.00515.
43. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
44. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 3–7 May 2021.
45. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *in press*. [CrossRef]
46. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
47. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
48. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 7–10 December 2021; pp. 12077–12090.
49. Chen, X.; Xie, S.; He, K. An empirical study of training self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 9640–9649.
50. Bazi, Y.; Bashmal, L.; Rahhal, M.M.A.; Dayil, R.A.; Ajlan, N.A. Vision transformers for remote sensing image classification. *Remote Sens.* **2021**, *13*, 516. [CrossRef]
51. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5693–5703.
52. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1804.03999.
53. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning where to look for the pancreas. In Proceedings of the Medical Imaging with Deep Learning, Amsterdam, The Netherlands, 4–6 July 2018.
54. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U^2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404. [CrossRef]
55. Poudel, R.P.; Liwicki, S.; Cipolla, R. Fast-SCNN: Fast semantic segmentation network. In Proceedings of the British Machine Vision Conference, Cardiff, UK, 9–12 September 2019.