frontiers | Frontiers in Earth Science

TYPE Original Research
PUBLISHED 26 May 2023
DOI 10.3389/feart.2023.1182145

OPEN ACCESS

EDITED BY
Sansar Raj Meena,
University of Padua, Italy

REVIEWED BY
Lorenzo Nava,
University of Padua, Italy
Kamal Kumar Sharma,
Lovely Professional University, India

*CORRESPONDENCE
Mingzhe Liu,
liumz@cdut.edu.cn

RECEIVED 08 March 2023
ACCEPTED 15 May 2023
PUBLISHED 26 May 2023

CITATION
Chen X, Liu M, Li D, Jia J, Yang A, Zheng W and Yin L (2023), Conv-trans dual network for landslide detection of multi-channel optical remote sensing images.
Front. Earth Sci. 11:1182145.
doi: 10.3389/feart.2023.1182145

Conv-trans dual network for landslide detection of multi-channel optical remote sensing images

Xin Chen[1], Mingzhe Liu[1]*, Dongfen Li[1], Jiaru Jia[1], Aiqing Yang[1], Wenfeng Zheng[2] and Lirong Yin[3]

[1]State Key Laboratory of Geohazard Prevention and Geoenvironment Protection, Chengdu University of Technology, Chengdu, China, [2]School of Automation, University of Electronic Science and Technology of China, Chengdu, China, [3]Department of Geography and Anthropology, Louisiana State University, Baton Rouge, LA, United States

Landslide detection is crucial for disaster management and prevention. With the advent of multi-channel optical remote sensing technology, detecting landslides have become more accessible and more accurate. Although the use of the convolutional neural network (CNN) has significantly increased the accuracy of landslide detection on multi-channel optical remote sensing images, most previous methods using CNN lack the ability to obtain global context information due to the structural limitations of the convolution operation. Motivated by the powerful global modeling capability of the Swin transformer, we propose a new Conv-Trans Dual Network (CTDNet) based on Swin-Unet. First, we propose a dual-stream module (CTDBlock) that combines the advantages of ConvNeXt and Swin transformer, which can establish pixel-level connections and global dependencies from the CNN hierarchy to enhance the ability of the model to extract spatial information. Second, we apply an additional gating module (AGM) to effectively fuse the low-level information extracted by the shallow network and the high-level information extracted by the deep network and minimize the loss of detailed information when propagating. In addition, We conducted extensive subjective and objective comparison and ablation experiments on the Landslide4Sense dataset. Experimental results demonstrate that our proposed CTDNet outperforms other models currently applied in our experiments.

KEYWORDS

landslide detection, swin transformer, convolutional neural network (CNN), remote sensing (RS), landslide

1 Introduction

As environmental problems become increasingly severe, the frequency of various natural disasters is increasing. As one of the natural disasters, landslides have a substantial negative impact on people's lives and economic development. So, researchers have spent enormous energy studying the process of landslides, including type, location, stability, and triggering factors Zhao and Lu (2018). Landslides are caused by the specific composition of slope movement and have great destructive power, which usually occurs due to the movement of gradual occurrence of rocks, sediments, and soils under the action of gravity Zhang et al. (2021). In the past, landslides were usually discovered by manual exploration. Now, with the emergence of geospatial technologies, such as aerial photogrammetry, satellite remote

sensing images, etc., the methods of finding landslide areas have gradually diversified Mohan et al. (2021). However, with the advancement of technology, there are also many challenges. For example, people with a strong geographical knowledge background need to spend a lot of time and energy to find the landslide area in remote sensing images, which usually has low-efficiency Chae et al. (2017). However, the rapid development of Machine Learning (ML) has created a good condition for efficient landslide detection.

ML is widely used in landslide identification, such as Support Vector Machine (SVM) CHEN and ZHOU (2020); Aslam et al. (2022); Tien Bui et al. (2018), Random Forest (RF) Tehrani et al. (2021); Liu Y. et al. (2021); Yu et al. (2018) and k-Nearest Neighbor (KNN) Mezaal et al. (2018); Ramos-Bernal et al. (2021). Huang et al. CHEN and ZHOU (2020) proposed a remote sensing image landslide detection method based on color feature model and SVM to solve the problem that the target color characterization model is not accurate and poor recognition rate of landslide surface in remote sensing image. Aslam et al. Aslam et al. (2022) used Logistic Regression (LGR), Linear Regression (LR), and Support Vector Machine (SVM) to produce Landslide Susceptibility Maps (LSMs) with weighted overlay techniques using different weights of landslide-related factors and SVM get the highest accuracy. Since AIRSAR data and GIS-based susceptibility mapping are rarely used in landslide detection in tropical environments, Bui et al. Tien Bui et al. (2018) used Support Vector Machine (SVM) and Index of Entropy (IOE) methods for landslide susceptibility assessment in the Cameron Highlands area, Malaysia. Aiming at the problem that many landslide detection works are limited by geographical scope, Tehrani et al. Tehrani et al. (2021) proposed a method based on Random Forest (RF), which achieved good results on Sentinel-2 multi-spectral satellite imagery and ALOS Digital Elevation Model. Liu et al. Liu Y. et al. (2021) used the Geo-detector-RF-integrated model Luo and Liu (2018) to obtain the highest accuracy on thirteen feature datasets, which implied that optimized conditioning factors can effectively improve the prediction accuracy of landslide susceptibility mapping. Yu et al. Yu et al. (2018) calculated six indexes and texture information as features, including water, snow, and vegetation enhancements, and proposed a pixel-level landslide detection model based on change detection using the random forest method in Nepal. Although traditional machine learning methods have made breakthroughs to a certain extent in many fields, there are still many limitations in the field of landslide identification.

The development of Deep Learning (DL) has also greatly promoted the accuracy of landslide detection Yao et al. (2021); Ghorbanzadeh et al. (2019); Nava et al. (2021). DL methods applied to landslides are mainly based on Convolutional Neural Networks (CNN) and Transformer. Compared with the traditional DL method of landslide detection, although CNN has been around for a long time, it did not become mainstream until the advent of AlexNet [30]. Then a large number of CNN variants appear, e.g., VGG Simonyan and Zisserman (2014), ResNet He et al. (2016), GoogleNet Szegedy et al. (2015) and Efficient-Net Tan and Le (2019). CNN has proved that it has excellent feature extraction capabilities and can effectively reduce time consumption. Ullo et al. Ullo et al. (2021) used Mask R-CNN He et al. (2017) for landslide detection, which is a pixel-level segmentation method based on Faster CNN Ren et al. (2015). In their paper, the highest landslide detection accuracy can be obtained by adopting ResNet-50 and ResNet-101 as backbone models. Ji et al. Ji et al. (2020) proposed an attention mechanism for boosting the CNN to extract more distinctive feature representations of landslides from backgrounds and got the best landslide detection F1-Score of 0.9662 in a landslide dataset which is located in Bijie city, China. Nava et al. Nava et al. (2022) is one of the first attempts in which the combination of SAR data and DL algorithms are employed for landslide mapping purposes, using Attention U-Net to SAR data and obtaining competitive results in the experiment. Since the influence of network architecture design and data fusion is still not fully explored in landslide detection, Sameen et al. Sameen and Pradhan (2019) compared a one-layer CNN with two of its deeper counterparts and residual networks with two fusion strategies (layer stacking and feature-level fusion) to detect landslides in Cameron Highlands, Malaysia. Finally, their results show that when using feature-level fusion, the experimental effect could be enhanced with the same network designs.

Besides, Transformer-based DL methods also provide a feasible method for landslide detection. Transformer Vaswani et al. (2017) is an encoder-decoder structure based on a multi-head self-attention mechanism, which solves the problem of parallel computing in natural language and, to some extent, alleviates the long-distance dependence of structures such as LSTM Graves and Graves (2012). Due to the Transformer's ability to make full use of global information, some researchers try to apply the Transformer to the field of computer vision. ViT Dosovitskiy et al. (2020) is the first pure transformer-based image classification architecture and achieves comparable performance to state-of-the-art CNN architectures by directly applying sequences of image patches to classification. After that, a large number of models based on Transformer for the computer vision field appeared. There are models for classification, such as MAE He et al. (2022), MViT Fan et al. (2021), for detection, such as DETR Carion et al. (2020), DINO Caron et al. (2021), for segmentation, such as Seg-Former Xie et al. (2021), SETR Zheng et al. (2021). Tang et al. Tang et al. (2022) used the Transformer-based model to compare with the CNN-based model on the landslide dataset, and their findings that the Transformer-based model had better detection performance. There is still little research on landslide detection based on Transformer models, so it is also a challenging direction to introduce Transformer models into landslide detection.

In this article, we explore the application potential of Swin-Unet Cao et al. (2023) for semantic segmentation in the field of landslide and introduce three improvements that enable the network to achieve better performance on publicly available landslide image datasets. Interestingly, few studies have applied Transformer-based models to landslide detection. Although Swin-Unet achieves outstanding results on Medical Image Segmentation, we find that it does not achieve satisfactory results on the landslide dataset. We believe that the reasons are as follows. First, we found that Transformer-based models such as Swin-Unet usually downsamples by 4x in the first stage, resulting in a large loss of image structure information, which makes detailed information cannot be recovered effectively in the subsequent upsampling process. Second, Transformer-based models usually have better global information but may be worse than convolution-based networks in terms of detailed texture information. Finally, we believe that it is too simple

for Swin-Unet to use skip links to directly concatenate feature maps, so shallow texture information and deep semantic information cannot be effectively fused, resulting in poor detection results.

To solve the above problem, we improved Swin-Unet and applied it to Landslide4Sense Ghorbanzadeh et al. (2022), which is a public dataset for landslide detection from remote sensing. Swin-Unet is a model based on Swin Transformer Liu Z. et al. (2021), which uses the shifted window and hierarchy, indicating that it enables the attention mechanism to do hierarchical feature extraction like CNN. Due to the good performance of the Swin Transformer on various computer vision tasks, Liu et al. Liu et al. (2022) achieved comparable accuracy with the Swin Trans-former using a pure convolution-based structure ConvNeXt on the ImageNet-1K dataset, thus proving that the pure convolution structure competes favorably with Transformers in terms of accuracy and scalability. Based on Swin Transformer and ConvNeXt, we propose CTDNet. In order for the model to have better feature extraction ability, we try to improve the attention gate [26] into Swin-Unet. Our main contributions are as follows. 1) We improved Swin-Unet for landslide detection.2) We proposed a dual block based on ConvNeXt and Swin Transformer (CTDBlock), which can establish pixel-level connections and establish global dependencies from the CNN hierarchy to enhance the ability of the model to extract features. 3) We applied an additional gating module (AGM) to effectively fuse the low-level information extracted by the shallow network and the high-level information extracted by the deep network.

The rest of this article is organized as follows. **Section 2** introduces the model we use and some related model structures, and the three improvements we propose will be described in detail. The datasets used in the paper, the training method, and the experimental parameters are described in **Section 3**. **Section 4** presents complete ablation studies and results analysis. The final section presents conclusions.

2 Methods

We propose three main improvements to make Swin-Unet better for landslide remote sensing image detection. First, we changed the downsampling of the Swin-Unet model from 4x to 2x for a more minor loss of detail information. Second, We combine Swin Transformer and ConvNeXt to effectively obtain global semantic context information and spatial context information in remote sensing landslide images. Third, we improve the attention gate in Attention-Unet and apply it to Swin-Unet for a better fusion effect. We will present the ensemble of our CTDNet model in **Section 2.1**. Then, the details of our improvements are introduced in **Section 2.2**, **Section 3**, **Section 4**.

2.1 Model architecture

The architecture of our proposed CTDNet is based on three main modules: 1) Swin Transformer Block. 2) ConvNeXt-Swin Transformer Dual Block. 3) Additional Gate Module. As shown in **Figure 1**, the image $x \in \mathbb{R}^{H\times W\times C}$ is first input into CTDNet. Where H, W, and C represent the height, width, and number of channels in the image, respectively. The channels of the initial image are 14.

As shown in **Figure 1**, the model has a total of eight stages, four down-sampling stages, and four up-sampling stages. There are four stages in the down-sampling process to get $x_{d,1}$, $x_{d,1}$, $x_{d,1}$, $x_{d,1}$ feature maps, respectively. Then, $x_{u,0}$, $x_{u,1}$, $x_{u,2}$, $x_{m,3}$ feature maps are obtained in four up-sampling stages. Features are extracted by Swin-Transformer block in the first three stages, and features are extracted by our proposed CTDBlock in the fourth stage. In the first three stages, each stage consists of a patch merging and two Swin Transformer blocks. Patch merging splits the image into nonoverlapping patches and stacks them together for linear embedding, and then these patches are applied to Swin-Transformer block. The final feature map resolutions are $\frac{H}{2}\times\frac{W}{2}$, $\frac{H}{4}\times\frac{W}{4}$, $\frac{H}{8}\times\frac{W}{8}$, respectively. Next, we use CTDBlock to extract features in bottleneck, and the output feature map resolution is $\frac{H}{16}\times\frac{W}{16}$. There are four stages in the upsampling process. The first three stages contain a patch expanding layer, two consecutive Swin Transformer blocks and an AGM module, and the last stage only has a patch expanding layer and a linear project layer. In the first three stages of upsampling, we aim to better fuse shallow texture features and deep semantic features. The feature maps of the patch expanding layer and the corresponding feature maps in the downsampling process are fused by AGM and then input into two consecutive Transformer blocks. In the last stage, the feature maps are restored to masks through the patch expanding layer and the linear project layer. The feature map resolutions of the four upsampling stages are $\frac{H}{2}\times\frac{W}{2}$, $\frac{H}{4}\times\frac{W}{4}$, $\frac{H}{8}\times\frac{W}{8}$ and $\frac{H}{16}\times\frac{W}{16}$, respectively.

2.2 Swin transformer block

The main difference between Swin Transformer and the origin Transformer is that Swin Transformer uses a module based on shifted window (SW-MSA) instead of the standard multi-head self-attention module (MSA). As illustrated in **Figure 2**, the main innovation of SW-MSA is to rotate the image up and left by half the window size. For model validity, the Swin Transformer performs self-attention operations on a local window. The images are divided into non-overlapping parts. Each individual image patch is equivalent to a token in the Transformer, and the self-attention module in the Transformer calculates the correlation of each token with all other tokens, which enables the model to obtain good global context information. Suppose the window size is M×M, and the computational complexity of MSA and W-MSA are shown as follows, respectively.

$$\Omega MSA = 4hwC^2 + 2(hw)^2C \tag{1}$$

$$\Omega W-MSA = 4hwC^2 + 2M^2hwC \tag{2}$$

Where C represents the dimension of the feature map, and h and w denote the height and width of the feature map, respectively. Since M is a fixed size, the computational complexity of W-SMA is linearly related to h and w, while the computational complexity of MSA is quadratic with h and w. As illustrated in **Figure 2**, the Swin Transformer uses a SW-MSA module instead of the W-MSA module in the Transformer, and the feature map crosses through the traditional Transformer module and the Swin Transformer module to obtain new feature maps. Since the picture is divided

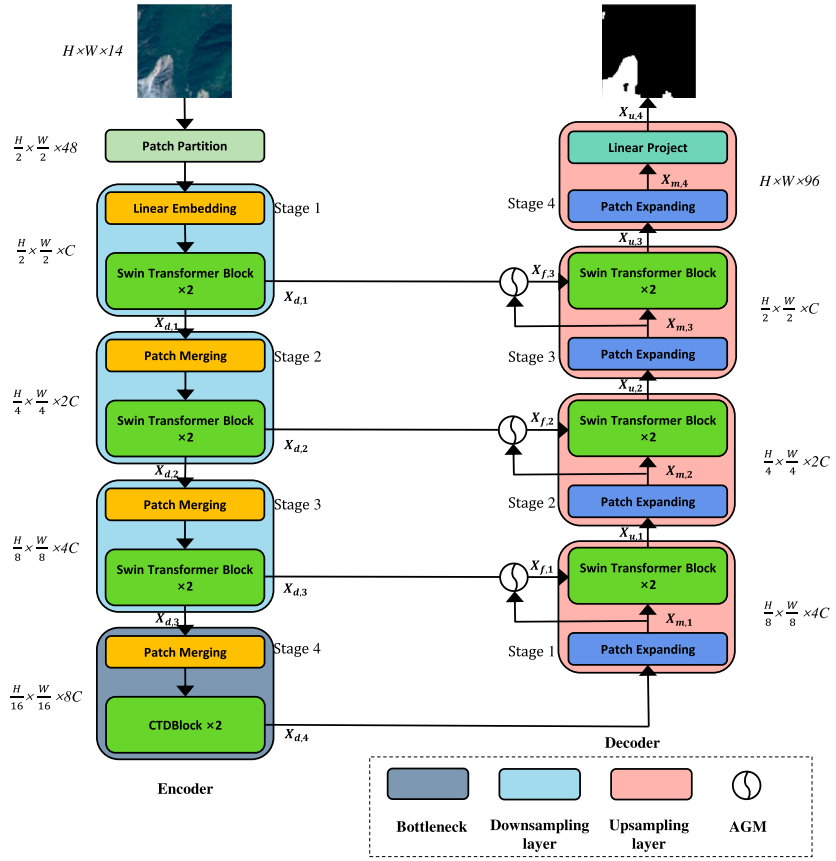

FIGURE 1
The overall structure of CTDNet.

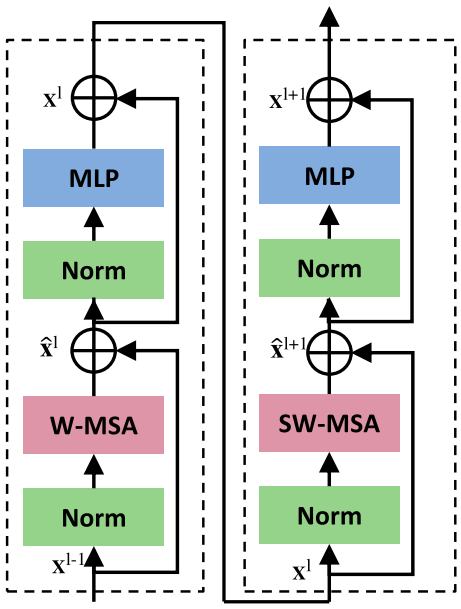

FIGURE 2
Two successive swin transformer blocks.

into non-overlapping windows, which leads to the lack of effective information interaction between different windows, the existence of SW-MSA is necessary for the network to obtain better features. The process of calculating the feature map in consecutive Swin Transformer blocks are computed as follows:

$$\hat{x}^{l} = W - MSA\left(LN\left(x^{l-1}\right)\right) + x^{l-1} \tag{3}$$

$$x^{l} = W - MSA\left(LN\left(x^{l-1}\right)\right) + x^{l-1} \tag{4}$$

$$\hat{x}^{l+1} = SW - MSA\left(LN\left(x^{l}\right)\right) + x^{l} \tag{5}$$

$$x^{l+1} = MLP\left(LN\left(\hat{x}^{l+1}\right)\right) + \hat{x}^{l+1} \tag{6}$$

Where $\hat{x}^{l}$ represents the output characteristics of the W-MSA module for block l, and x^{l} denotes the output characteristics of the MLP module after block l. $\hat{x}^{l+1}$ represents the output characteristics of the SW-MSA module for block l+1, and x^{l+1} denotes the output characteristics of the MLP module after block l+1. LN denotes layer normalization, and W-MSA and SW-MSA denote the multi-head self-attention using the origin window and multi-head self-attention using shifted window.

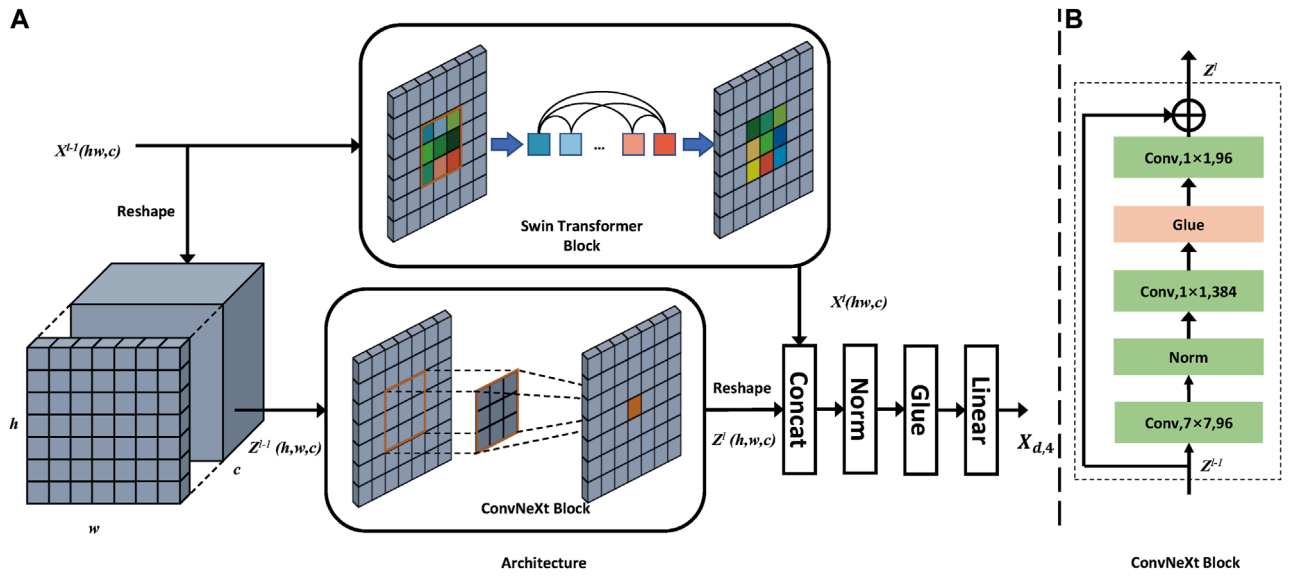

FIGURE 3
(A) The structure of ConvNeXt-Swin Transformer Dual Block (CTDBlock). **(B)** The structure of ConvNeXt Block.

2.3 ConvNeXt-Swin transformer dual Block (CTDBlock)

In order to extract features using CNN and Transformer effectively, we propose a ConvNeXt-Swin Transformer Dual Block. ConvNeXt-Swin Transformer Dual Block is mainly composed of a parallel connection of ConvNeXt block and Swin Transformer block. Although Swin Transformer establishes the relationship between tokens in a small window, it effectively reduces the number of parameters. However, this method weakens the original positional relationship and global modeling ability between the pixels of the image. Besides, remote sensing landslide images usually have a blurred boundary, which leads to the fact that texture information is extremely important in landslide identification. CNN has good spatial information extraction ability. Therefore, we propose CTDBlock to better combine the global relationship between patches extracted by Swin Transformer and the spatial texture information extracted by CNN, which makes the model more effective in image segmentation tasks. The components of CTDBlock are shown in Figure 3.

In the bottleneck stage, we first reshape the output feature $X^{l-1} \in \mathbb{R}^{(h\times w)\times c}$ of the last stage to $Z^{l-1} \in \mathbb{R}^{h\times w\times c}$. Where, $h = H/16$, $w = W/16$ and $c = 8C$. Feature maps X^{l-1}, Z^{l-1} are fed into Swin Transformer block and ConvNeXt block, respectively. Therefore, the process in Swin Transformer block can be represented as follows:

$$x_{ki}^{l} = \sum_{m=0}^{hw-1} f_i\left(x_{km}^{l-1}\right) \tag{7}$$

where i represents the index of the flattened window in the Swin Transformer, and k represents the current channel. $f_i(\cdot)$ represents the mapping relationship between x_{km}^{l-1} and x_{ki}^{l}. Therefore, Therefore, X^l is obtained after each element on the feature map is calculated by (Eq. 7).

Then, the process in ConvNeXt block can be represented as follows:

$$z_{kij}^{l} = \sum_{m=max(0,i-p)}^{min(i+p,h-1)} \sum_{n=max(0,j-p)}^{min(j+p,w-1)} \varphi_k\left(z_{mn}^{l-1}\right) \tag{8}$$

where i, j, and k represent the indexes of width direction, height direction, and channel. $\varphi_k(\cdot)$ represents the mapping relationship between z_{mn}^{l-1} and z_{kij}^{l}, and p is half of the window size. Furthermore, Z^l is obtained after each element on the feature map is calculated by (8). Finally, the feature map S can be expressed as follows:

$$S = \psi\left(Concat\left(Z^l, X^l\right)\right) \tag{9}$$

where $\psi(\cdot)$ means a layer normalization, GELU and a linear projection layer. *Concat* $(\cdot)$ represents concatenating two tensors together.

2.4 Additional gate module (AGM)

In the past, most jobs used the jump link to directly combine the shallow space texture information and deep semantic information when using the U -shaped network. However, we think that this combination method has caused many details and structural information to be lost, especially dense and small-scale targets. Therefore, we proposed an additional gate module (AGM) to try to fuse the features effectively, thereby improving the result of landslide detection.

As shown in Figure 4, we obtain the gate signal by concatenating shallow texture information and deep semantic information. Then we apply the gating signal to the deep semantic information to select spatial regions and apply the selected information to the shallow texture information to get a better fuse feature map. Let $a = x_{d,4-i}$,

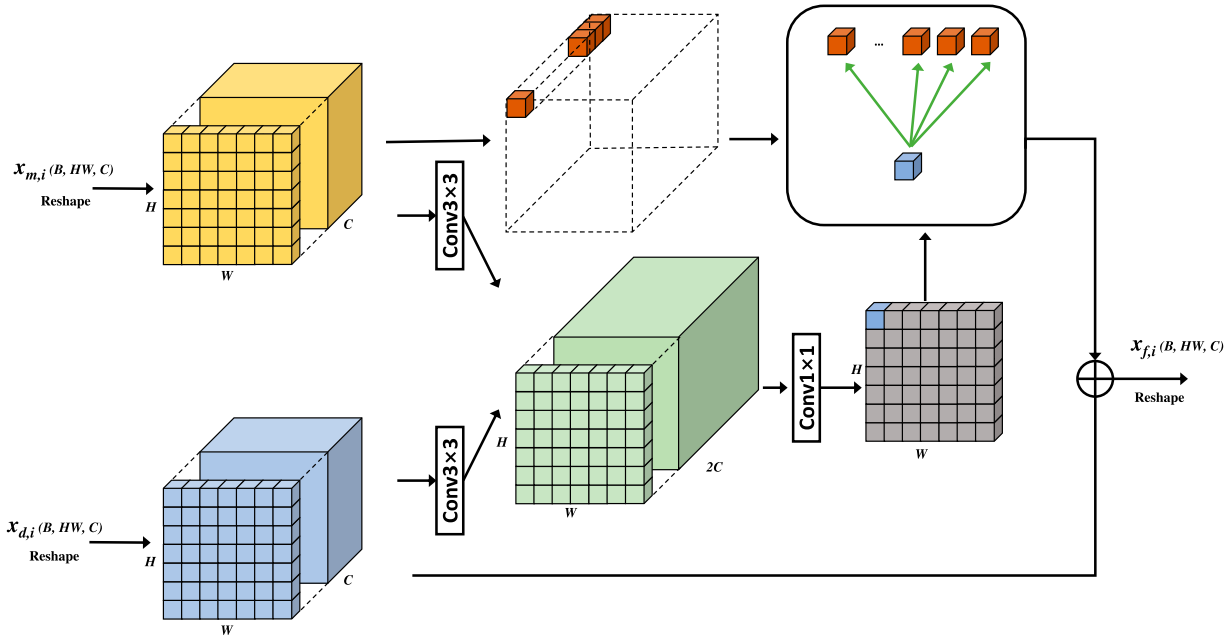

FIGURE 4
The structure of AGM.

$g = x_{m,i}$, AGM is formulated as follows:

$$x_o = Concat\left(\sum_{k=1}^{C} a_k * W_k^a + b_1, \sum_{k=1}^{C} g_k * W_k^g + b_2\right) \tag{10}$$

$$x_t = \sigma\left(max\left(x_o, 0\right)\right) \tag{11}$$

At this point, $x_t \in \mathbb{R}^{h\times w}$, and then, we expand x_t to $\hat{x}_t \in \mathbb{R}^{h\times w\times c}$ along the channel direction.

$$x_{f,i} = a \oplus \left(\hat{x}_t \odot g\right) \tag{12}$$

Where b_1, b_2 represent bias. $x_{d,i}$ denotes the feature map of the output of the ith stage of the down-sampling process, $x_{m,i}$ denotes the feature map of the ith stage of the decoder after patch expanding operation, $x_{f,i}$ denotes the feature map of the output of AGM and $\sigma(\cdot)$ means 1×1 convolution layer with batch normalization. W_k^a and W_k^g are the weights of the convolution kernel on different feature maps in channel k, respectively. Here, $\odot$ stands for element-level multiplication and $\oplus$ represents element-level addition.

3 Dataset and design of experiments

The multi-source landslide benchmark data (Landslide4Sense) Ghorbanzadeh et al. (2022) was utilized to test the effectiveness of the model we used in the experiments. In this section, we first briefly introduce the dataset used in our experiment and then introduce the evaluation index and training configuration.

3.1 Dataset

There are 3,799 image patches fusing optical layers from Sentinel-2 sensors with the digital elevation model and slope layer derived from ALOS PALSAR in the Landslide4Sense dataset. In short, 3799 annotated images are provided without any overlap. Each image is made up of 14 bands and has a resolution of 128× 128 pixels. Bands 1-12 are the multi-spectral data from Sentinel-2 and bands 13-14 belong to slope and digital elevation model (DEM) data from ALOS PALSAR. Bands 1-12 are Coastal aerosol, Blue, Green, Red, Red Edge (short), Red Edge (medium), Red Edge (long), near-infrared (NIR), Water vapor, short-wave infrared (short), short-wave infrared (medium) and short-wave infrared (long), respectively. We used 3040 patches to train all models, while the remaining 759 images are used for evaluation in the experiments. Random horizontal and vertical flips and operations are adopted in the data augment strategy. At the same time, in order to have a better generalization ability of the model, we also add Gaussian random noise to the image.

3.2 Evaluation indices

We used the data publisher's evaluation method, including Recall, Precision, and F1-Score. To introduce these evaluation indices, we will first introduce the confusion matrix. As is shown in **Table 1**, TP (True Positive) indicates that the sample is predicted positive and actually true, FP(False Positive) indicates that the sample is predicted negative and actually true, TN (True Negative) indicates that the sample is predicted positive and actually false, FN(False Negative) indicates that the sample is predicted negative and actually false.

Therefore, precision and recall are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

TABLE 1 Confusion matrix.

	Predicted positive	Predicted negative
Actual Positive	TP	FN
Actual Negative	FP	TN

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

As in (Eq. 15), F1 is the harmonic mean of Precision and Recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{15}$$

The definition of IoU is shown in the following formula:

$$IoU = \frac{TP}{FN + FP + TP} \tag{16}$$

3.3 Visualization method

To further illustrate the model's ability to capture features efficiently, we visualized the last convolutional layer of the model using Gradient-weighted Class Activation Mapping (Grad-CAM) Selvaraju et al. (2017). The overall structure of Grad-CAM is shown in the **Figure 5**. The formula is expressed as follows:

$$L^c_{Grad-CAM} = ReLU\left(\sum_k a^c_k A^k\right) \tag{17}$$

$$a^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}} \tag{18}$$

Where A represents the feature map, k is the *k*th channel in feature map A. Z represents the product of the width and height of the feature map, and the class is denoted by c. A^k_{ij}, y^c represent the gradient value of backpropagation and the predicted score of class c, respectively.

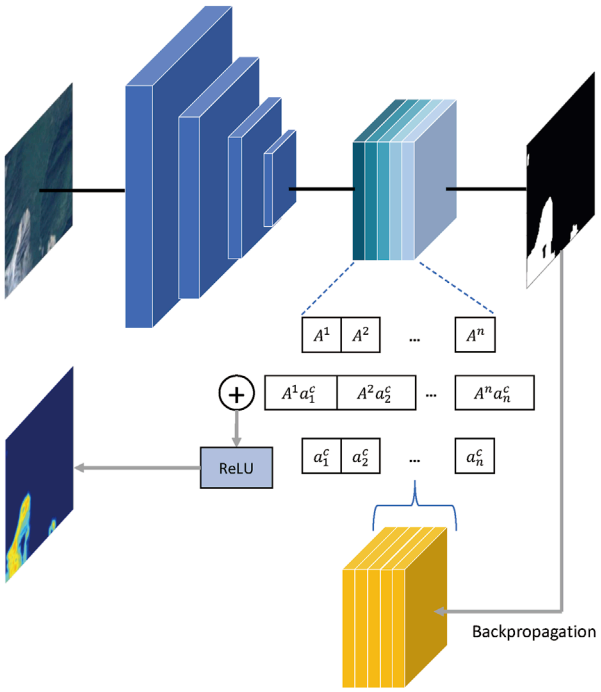

FIGURE 5
The structure of Grad-CAM.

3.4 Train configuration

The experimental environment is PyTorch 1.11.0, Python3.9 and CUDA 11.3. Our optimization algorithm is AdamW, and the weight decay is 0.0005. The number of epochs is 100. The learning rate is initialized as 0.0001. The loss function is the cross-entropy loss. A batch size of 32 is used. The models are used for comparison in our experiment, including DeepLabv3 Chen et al. (2017), FCN Long et al. (2015), PSPNet Zhao et al. (2017), SegNet Badrinarayanan et al. (2017), U-Net Ronneberger et al. (2015), U2-Net Chen et al. (2017), ResU-Net Zhang et al. (2018) and Swin-Unet Cao et al. (2023). We train all models on 3 T P100 GPUs with 16 GB memory. To sum up, the training and test process of our model is represented in **Algorithm 1**.

```
Input: (1)Training images and corresponding labels 2) Test images
Output: Predict masks for test images
Initialize model weights
Data augmentation on training data
#Training process
For epoch in range (100):
Select a batch of data
The data passes through the model's encoder and decoder to get the output
Calculate cross-entropy loss with output and corresponding label
Loss backpropagation
Update model parameters
#Test process
There is no back-propagation process and parameter update process in the testing process, and the others are the same as the training process.
```

Algorithm 1. CTD Net.

4 Results and discussion

We conduct experiments on Landslide4Sense. In the following, we ablate our important improvements based on Swin-Unet. Then, we compare the improved Swin-Unet with other CNN-based and Transformer-based state-of-the-art segmentation models on the Landslide4Sense dataset.

4.1 Ablation studies

We pool a variety of ideas from past work with our novel method to improve Swin-Unet's performance. A summary of the ablation experiments can be found in **Table 2**. We chose the

TABLE 2 Ablation results with different alterations.

Method	Patch size=4 × 4	Patch size=2 × 2	STCBlock	AGM	F1-socre
Swin-Unet	√				72.44
Swin-Ps		√			73.64
Swin-Db		√	√		74.27
CTDNet		√	√	√	74.71

TABLE 3 Swin-Unet and CTDNet performance.

Method	Para(M)	Train-t (min)	Test (FPS)
Swin-Unet	27.18	174	33.7
CTDNet	50.36	362	29.9

original Swin-Unet as the baseline model and performed ablation experiments.

4.1.1 Modify patch size

Ablations of the modified patch-size approach on Landslide4Sense dataset are reported in Table 2. By modifying the patch-size, we make the downsampling multiple of the first stage from four times to two times, which effectively obtains the texture information of the original image without too many additional parameters. The increase of the F1-Score from 72.44% to 73.64% also demonstrates the effectiveness of our method in extracting features.

4.1.2 Introduce STCBlock

Table 2 compares the improvement in model performance when using STCBlock. When we used our proposed STCBlock at the bottleneck, our model performance improved by 0.63% F1-Score. The experimental results prove the effectiveness of our method combining Convnext block and Swin block, which gives the model have stronger ability to extract texture information and semantic information.

4.1.3 Introduce addition gate module

Finally, we show how our addition gate module improves the capabilities of the model. When we use AGM, the F1-score of the model increases by 0.44%. Due to the selection of the solution space by AGM, the shallow texture information and the deep semantic information are fully integrated, which further improves the module effect.

4.1.4 Model performance comparison

To more comprehensively compare the performance difference of our CTDNet model with the original Swin-Unet, we compare their parameter amounts, training time and test FPS under the same conditions. As shown in Table 3, our CTDNet has more overhead in model parameters, training time, and inference time, but our model achieves an improvement of more than 2% with a small increase in cost. Although our Swin-Unet model has more parameters and training time than the original Swin-Unet model, the final test speed is comparable to the original Swin-Unet model. We guess the main reason is that the ConvNeXt and AGM we added are mainly convolution operations, which save time compared to Transformer's matrix multiplication, so the final Test speed is comparable to the original Swin-Unet.

4.1.5 Ablation study visualization

To demonstrate the better segmentation performance of our and inference time, but our model achieves an improvement of more than 2% with a small increase in cost. Compared to Swin-Unet, we visualize the experimental results. As shown in Figure 6, when using the original Swin-Unet, the model does not maintain a high resolution, resulting in a lot of discrete classification results on the boundary, which seriously affects the performance of the model. It is demonstrated that combining the feature maps of CNN and Swin Transformer is more effective than utilizing feature maps of Swin Transformer when comparing Swin-Ps and Swin-Db. Comparing column Swin_Db with column CTDNet, we can find that using the AGM can get better feature maps for landslides. To better interpret the models, we use Grad-CAM [28] to visualize their last convolutional layer. As shown in Figure 7, we compare the regions of interest of the original Swin-Unet and our CTDNet model in landslide detection by using Grad-CAM. By simple upsampling, we keep the size of the feature map consistent with the input image after Grad-CAM. The first row is generated by Swin-Unet and the second row is from the visualization results of CTDNet. In the landslide column, our CTDNet model has a more accurate segmentation of landslides. Since our method utilizes the feature extraction ability of CNN, the model has a better ability to extract texture information, and then it can acquire long-range semantic correlation due to the use of the self-attention mechanism. Therefore, our model can better recognize semantic features and get better results. At the same time, we can see the difference between Swin-Unet and our model in small object recognition. Our model is usually more precise in the recognition of boundaries and small objects, which is due to the better feature extraction and feature fusion used by our model. Compared to Swin-Unet, CTDNet model has a more accurate segmentation result for objects with different sizes and classes.

4.2 Evaluation and comparisons on the landslide4Sense dataset

In this section, we compare the effect of CTDNet with other state-of-the-art models (Unet, U2net, ResU_net, etc.) on the Landslide4Sense Dataset. Furthermore, we compare the performance of state-of-the-art model in Ghorbanzadeh et al. (2022) and our CTDNet using the same dataset as ours. We let these models all use the same

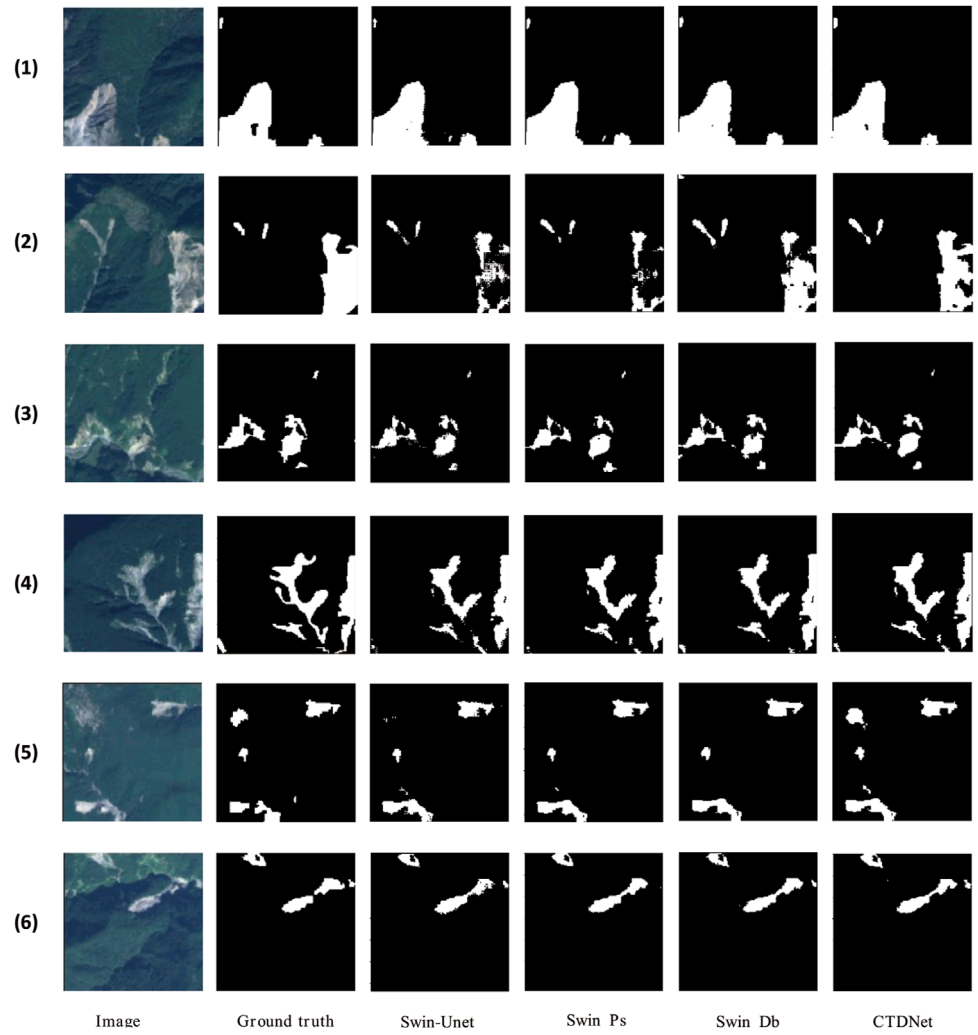

FIGURE 6
The result figure of ablation studies visualized in different images.

experimental settings and do not use other kinds of fine-tuning methods.

4.2.1 Comparison of experimental results on the landslide4Sense dataset

Table 4 shows the results of all the semantic segmentation models in the comparative experiments. From the table, we can see that CTDNet has the highest mF1-Score. In the landslide images, the number of landslides and the background are severely unbalanced, so we can see that the background usually has a higher F1-Score. First, we look at the landslide class. ResU-net was used by Ghorbanzadeh et al. (2022) with the best results obtained on this dataset. In our experiments, ResU-Net also achieves a competitive result, only lower than our proposed CTDNet method, which also demonstrates the effectiveness of our method. Our CTDNet achieves the highest Recall, F1-Score and mF1-Score. Compared with the original Swin-Unet, we have made great improvements. Although the precision of our method is a little lower than the original Swin-Unet, there is a huge increase in Recall, so the F1-score of the model finally has an increase of about 2%. Our proposed CTDNet achieves a good balance between Precision and Recall in landslide detection.

In summary, CDTNet achieves state-of-the-art segmentation accuracy on the Landslide4Sense Dataset. We try to analyze the experimental results. First, we can see that the better-performing networks (U-Net, U2-Net, ResU-Net, etc.) all have skip connections, which shows the effectiveness of skip connections in landslide detection. Then, we compare these models with Swin-Unet, and we can find that these models maintain a feature map of the original resolution using a skip connection while Swin-Unet uses a skip connection to maintain a feature map after downsampling four times, which leads to the poor performance of Swin-Unet in landslide detection. Therefore, it is crucial to maintain a high-resolution feature map for the model to achieve a good performance in landslide detection from remote sensing imagery. It is no doubt that Transformer can usually learn better long-term dependencies Vaswani et al. (2017), and CNN can usually learn better local features Geirhos et al. (2018). So, our proposed CTDNet model

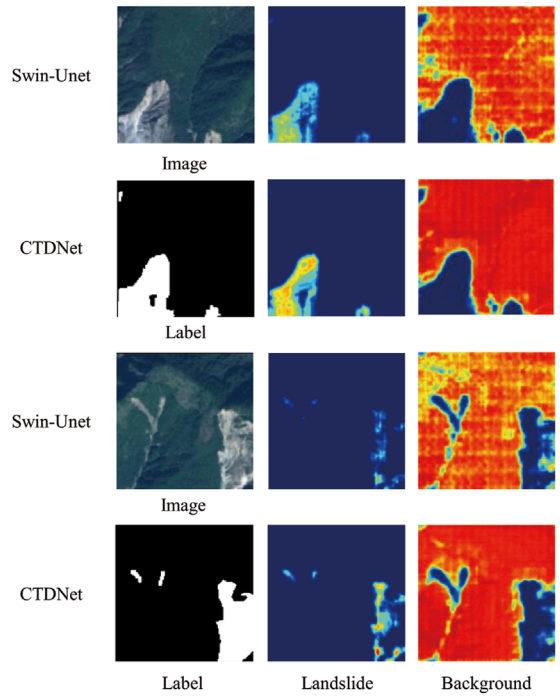

FIGURE 7
The Grad-CAM figure of different classes with different landslide images.

can achieve the best performance by combining the advantages of Transformer and CNN to extract feature maps with long-term dependencies and local information.

4.2.2 Comparison of experimental results on the HR-GLDD dataset

To further verify the effectiveness of our model, we also conducted experiments on the HR-GLDD dataset Meena et al. (2022). All our parameter settings are kept consistent with those on the Landslide4Sense dataset. **Table 5** shows the semantic segmentation models' results in the comparative experiments on the HR-GLDD dataset. Although both Swin-Unet and U-Net performed well, our model still achieves the highest mF1-Score of 83.79%, which fully demonstrates the effectiveness of the architecture proposed by our model.

4.2.3 Model performance analysis

In order to better show the advantages and disadvantages of the model, we summarize the model parameters, training time and test time used in the experiment in **Table 6**. As shown in the table, ResU-Net has the smallest number of model parameters, and U-Net has the least training time and inference time. Our model combines CNN and Transformer, resulting in a larger number of parameters than CNN-based models. Transformer-based models have self-attention modules and MLP modules, so Transformer-based models usually have larger time and space costs. CNN-based models usually use a local 3× 3 convolution kernel, while Transformer's self-attention

TABLE 4 Experimental results of different models on the Landslide4Sense dataset.

Method	Class	IoU	Precision	Recall	F1-score	mF1-score
DeepLabv3 Chen et al. (2017)	background	98.64	99.19	99.44	99.32	82.12
	landslide	48.06	69.51	60.90	64.92	
FCN Long et al. (2015)	background	98.60	99.17	99.42	99.29	81.56
	landslide	46.86	68.34	59.85	63.82	
PSPNet Zhao et al. (2017)	background	98.80	99.33	99.46	99.39	84.70
	landslide	53.86	72.32	67.85	70.01	
SegNet Badrinarayanan et al. (2017)	background	98.65	99.07	99.57	99.32	81.02
	landslide	45.67	72.85	55.04	62.71	
U-Net Ronneberger et al. (2015)	background	98.96	99.37	**99.58**	**99.48**	86.55
	landslide	58.23	**77.80**	69.84	73.61	
U2-Net Qin et al. (2020)	background	98.96	99.40	99.55	99.47	86.59
	landslide	58.36	76.68	70.95	73.70	
ResU-Net Zhang et al. (2018)	background	**98.97**	99.40	99.56	**99.48**	86.76
	landslide	58.79	77.23	71.11	74.04	
Swin-Unet Cao et al. (2023)	background	98.91	99.37	99.53	99.45	85.95
	landslide	56.79	75.39	69.71	72.44	
CTDNet	background	**98.97**	**99.45**	99.51	**99.48**	**87.10**
	landslide	**59.63**	75.85	**73.61**	**74.71**	

Bold values represent the best results among all models.

TABLE 5 Experimental results of different models on the HR-GLDD dataset.

Method	Class	IoU	Precision	Recall	F1-score	mF1-score
DeepLabv3 Chen et al. (2017)	background	93.50	96.05	97.24	96.64	83.08
	landslide	53.28	73.63	65.84	69.51	
FCN Long et al. (2015)	background	93.07	95.01	97.85	96.41	80.34
	landslide	47.35	75.30	56.05	64.26	
PSPNet Zhao et al. (2017)	background	93.48	95.97	97.30	96.63	82.90
	landslide	52.86	73.83	65.05	69.16	
SegNet Badrinarayanan et al. (2017)	background	92.54	94.28	**98.04**	96.12	77.85
	landslide	42.42	74.84	49.47	59.57	
U-Net Ronneberger et al. (2015)	background	**93.99**	95.83	98.00	**96.90**	83.65
	landslide	54.31	**78.86**	63.56	70.39	
U2-Net Qin et al. (2020)	background	93.97	95.88	97.92	96.89	83.58
	landslide	54.17	78.14	63.84	70.27	
ResU-Net Zhang et al. (2018)	background	93.88	95.93	97.77	96.84	83.54
	landslide	54.13	77.11	64.49	70.24	
Swin-Unet Cao et al. (2023)	background	93.74	96.26	97.28	96.77	83.73
	landslide	54.66	74.26	67.44	70.68	
CTDNet	background	93.65	**96.39**	97.05	96.72	**83.79**
	landslide	**54.85**	73.05	**68.77**	**70.85**	

Bold values represent the best results among all models.

TABLE 6 Performance of different models.

Method	Para(M)	Train-t (min)	Test (FPS)
DeepLabv3 Chen et al. (2017)	39.67	291	34.14
FCN Long et al. (2015)	32.98	217	34.44
PSPNet Zhao et al. (2017)	65.61	1177	19.00
SegNet Badrinarayanan et al. (2017)	53.55	621	27.97
U-Net Ronneberger et al. (2015)	32.98	**126**	**45.70**
U2-Net Qin et al. (2020)	44.03	301	23.40
ResU-Net Zhang et al. (2018)	**13.06**	325	39.72
CTDNet	50.36	362	29.90

Bold values represent the best results among all models.

mechanism is global, which is why the Transformer needs more parameters. Convolution has good translation invariance and scale invariance, but it is not so good at learning the relationship between large objects. So we think it has the potential to combine CNN and Transformer for landslide detection, and the final experimental results also confirm our idea.

4.2.4 Visualization results

The comparative visualization results of CTDNet model and other model are displayed in **Figure 8**. The first two rows in the first column of **Figure 8** are the original remote sensing image, and the right and bottom is the visualization result of each model on the Landslide4Sense Dataset. Although most of the models can correctly identify the landslide location, and the prediction results can match the landslide label, the recognition results in some boundaries and small landslide areas are still not effective. From the experimental results, we can find that the model based on the use of skip connections usually has a better landslide detection effect, and has a clearer outline in the division of the boundary, too. We further found that DeeplabV3, FCN and PSPNet models did not maintain a high-resolution feature map, resulting in the final landslide identification effect on the boundary being usually poor. Unet, U2Net and ResU-Net maintain a high-resolution feature map because of skip connections so that these models usually have a better recognition result in boundary regions. Our CTDNet combines the Transformer's better ability to extract contextual information while maintaining a high-resolution feature map and skip connections, which improves the model's ability to detect small objects. Finally, the experimental visualization results prove the effectiveness of our method, which still has good recognition ability in the complex and variational landslide environment. Our improved Swin-Unet achieves state-of-the-art results on the Landslide4Sense Dataset, and the landslide category F1-Score is 74.71%, which is 2% higher than the original Swin-Unet's 72.44%.

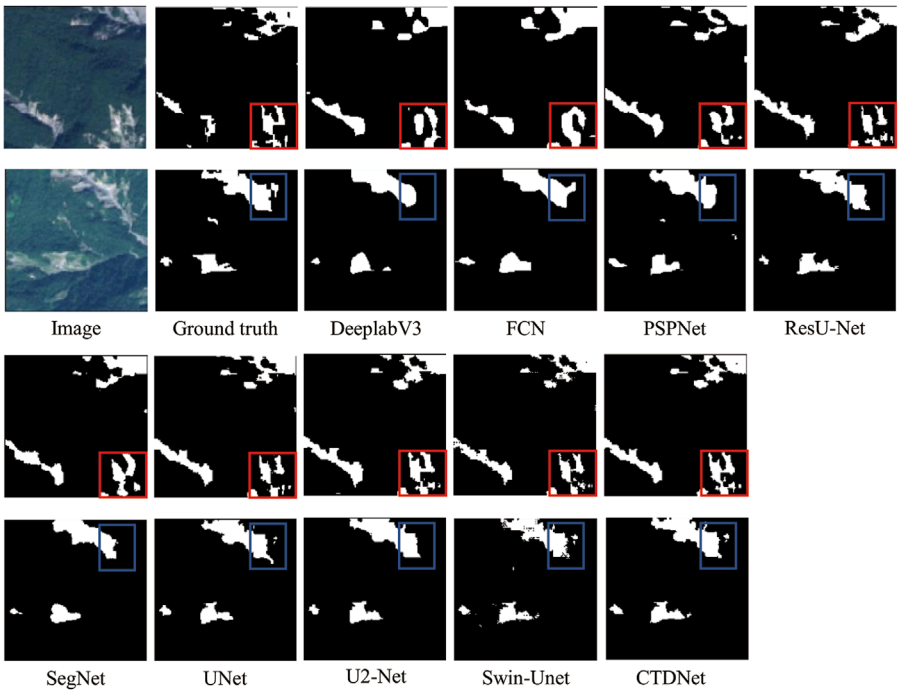

FIGURE 8
Visualization of experimental results for different models.

5 Conclusion

In this article, we propose a model of fusing ConvNeXt and Swin Transformer, CTDNet. We changed the patch size of the model to maintain a high resolution feature map which is better for the model to recognize small landslides and boundaries. We propose the ConvNeXt-Swin Transformer Dual Block, which combines the different advantages of Swin-Transformer and ConvNeXt to extract features so that the global semantic information extracted by Swin Transformer and the spatial texture information extracted by ConvNeXt can be fused to obtain stronger feature information. Besides, we also propose an additional gating module, which fully integrates the features of the skip connections and the features of the model upsampling process. Through the above improvements, the ability of the model to obtain rich semantic and spatial texture features is further enhanced. CTDNet achieves state-of-the-art performance compared to other advanced models on the Landslide4Sense Dataset. The combination of Transformer and CNN still has excellent potential to be applied in the field of computer vision, and we will continue to learn how to better apply Transformer and CNN to landslide detection in the future.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.iarai.ac.at/.

Author contributions

Conceptualization, ML, XC, and JJ; methodology, XC, DL, and WZ; software, JJ, LY, and WZ; validation, XC, AY, and JJ; formal analysis, ML and DL; investigation, XC, ML, and AY; resources, ML and LY; data curation, XC and JJ; writing—original draft preparation, XC and WZ; writing—review and editing, ML and JJ; visualization, XC and DL; supervision, ML; project administration, AY, LY, and JJ; funding acquisition, ML. All authors contributed to the article and approved the submitted version.

Conflict of interest

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

Publisher's note

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

References

Aslam, B., Maqsoom, A., Khalil, U., Ghorbanzadeh, O., Blaschke, T., Farooq, D., et al. (2022). Evaluation of different landslide susceptibility models for a local scale in the chitral district, northern Pakistan. *Sensors* 22, 3107. doi:10.3390/s22093107

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. pattern analysis Mach. Intell.* 39, 2481–2495. doi:10.1109/tpami.2016.2644615

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2023). "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Proceedings, Part III computer vision–ECCV 2022 workshops: Tel aviv, Israel, october 23–27, 2022* (Springer), 205–218.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). "End-to-end object detection with transformers," in *Proceedings, Part I computer vision–ECCV 2020: 16th European conference, glasgow, UK, august 23–28, 2020* (Springer), 16, 213–229.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., et al. (2021). "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chae, B.-G., Park, H.-J., Catani, F., Simoni, A., and Berti, M. (2017). Landslide prediction, monitoring and early warning: A concise review of state-of-the-art. *Geosciences J.* 21, 1033–1070. doi:10.1007/s12303-017-0034-4

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). *Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587*.

Chen, S., and Zhou, R. (2020). Landslide detection based on color feature model and svm in remote sensing imagery. *Spacecr. Recovery and Remote Sens.* 40, 89–98.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929*.

Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., et al. (2021). "Multiscale vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 6824–6835.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2018). *Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231*.

Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S. R., Tiede, D., and Aryal, J. (2019). Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. *Remote Sens.* 11, 196. doi:10.3390/rs11020196

Ghorbanzadeh, O., Xu, Y., Ghamis, P., Kopp, M., and Kreil, D. (2022). *Landslide4sense: Reference benchmark data and deep learning models for landslide detection. arXiv preprint arXiv:2206.00515*.

Graves, A., and Graves, A. (2012). Long short-term memory. *Supervised sequence Label. Recurr. neural Netw.*, 37–45.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ji, S., Yu, D., Shen, C., Li, W., and Xu, Q. (2020). Landslide detection from an open satellite imagery and digital elevation model dataset using attention boosted convolutional neural networks. *Landslides* 17, 1337–1352. doi:10.1007/s10346-020-01353-2

Liu, Y., Zhang, W., Zhang, Z., Xu, Q., and Li, W. (2021a). Risk factor detection and landslide susceptibility mapping using geo-detector and random forest models: The 2018 hokkaido eastern iburi earthquake. *Remote Sens.* 13, 1157. doi:10.3390/rs13061157

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021b). "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Luo, W., and Liu, C.-C. (2018). Innovative landslide susceptibility mapping supported by geomorphon and geographical detector methods. *Landslides* 15, 465–474. doi:10.1007/s10346-017-0893-9

Meena, S. R., Nava, L., Bhuyan, K., Puliero, S., Soares, L. P., Dias, H. C., et al. (2022). Hr-gldd: A globally distributed dataset using generalized dl for rapid landslide mapping on hr satellite imagery. *Earth Syst. Sci. Data Discuss.*, 1–21.

Mezaal, M. R., Pradhan, B., and Rizeei, H. M. (2018). Improving landslide detection from airborne laser scanning data using optimized dempster–shafer. *Remote Sens.* 10, 1029. doi:10.3390/rs10071029

Mohan, A., Singh, A. K., Kumar, B., and Dwivedi, R. (2021). Review on remote sensing methods for landslide detection using machine and deep learning. *Trans. Emerg. Telecommun. Technol.* 32, e3998.

Nava, L., Bhuyan, K., Meena, S. R., Monserrat, O., and Catani, F. (2022). Rapid mapping of landslides on sar data by attention u-net. *Remote Sens.* 14, 1449. doi:10.3390/rs14061449

Nava, L., Monserrat, O., and Catani, F. (2021). Improving landslide detection on sar data through deep learning. *IEEE Geoscience Remote Sens. Lett.* 19, 1–5. doi:10.1109/lgrs.2021.3127073

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., and Jagersand, M. (2020). U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognit.* 106, 107404. doi:10.1016/j.patcog.2020.107404

Ramos-Bernal, R. N., Vázquez-Jiménez, R., Cantú-Ramírez, C. A., Alarcón-Paredes, A., Alonso-Silverio, G. A., Bruzón, G. A., et al. (2021). Evaluation of conditioning factors of slope instability and continuous change maps in the generation of landslide inventory maps using machine learning (ml) algorithms. *Remote Sens.* 13, 4515. doi:10.3390/rs13224515

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. neural Inf. Process. Syst.* 28.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings, Part III medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, munich, Germany, october 5-9, 2015*, 18(Springer), 234–241.

Sameen, M. I., and Pradhan, B. (2019). Landslide detection using residual networks and the fusion of spectral and topographic information. *IEEE Access* 7, 114363–114373. doi:10.1109/access.2019.2935761

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 618–626.

Simonyan, K., and Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Tan, M., and Le, Q. (2019). "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning* (PMLR), 6105–6114.

Tang, X., Tu, Z., Wang, Y., Liu, M., Li, D., and Fan, X. (2022). Automatic detection of coseismic landslides using a new transformer method. *Remote Sens.* 14, 2884. doi:10.3390/rs14122884

Tehrani, F. S., Santinelli, G., and Herrera Herrera, M. (2021). Multi-regional landslide detection using combined unsupervised and supervised machine learning. *Geomatics, Nat. Hazards Risk* 12, 1015–1038. doi:10.1080/19475705.2021.1912196

Tien Bui, D., Shahabi, H., Shirzadi, A., Chapi, K., Alizadeh, M., Chen, W., et al. (2018). Landslide detection and susceptibility mapping by airsar data using support vector machine and index of entropy models in cameron highlands, Malaysia. *Remote Sens.* 10, 1527. doi:10.3390/rs10101527

Ullo, S. L., Mohan, A., Sebastianelli, A., Ahamed, S. E., Kumar, B., Dwivedi, R., et al. (2021). A new mask r-cnn-based method for improved landslide detection. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 14, 3799–3810. doi:10.1109/jstars.2021.3064981

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. neural Inf. Process. Syst.* 30.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* 34, 12077–12090.

Yao, G., Zhou, W., Liu, M., Xu, Q., Wang, H., Li, J., et al. (2021). An empirical study of the convolution neural networks based detection on object with ambiguous boundary in remote sensing imagery—A case of potential loess landslide. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 15, 323–338. doi:10.1109/jstars.2021.3132416

Yu, B., Chen, F., and Muhammad, S. (2018). Analysis of satellite-derived landslide at central Nepal from 2011 to 2016. *Environ. earth Sci.* 77, 331–412. doi:10.1007/s12665-018-7516-1

Zhang, H., Liu, M., Wang, T., Jiang, X., Liu, B., and Dai, P. (2021). "An overview of landslide detection: Deep learning and machine learning approaches," in *2021 4th international conference on artificial intelligence and big data (ICAIBD)* (IEEE), 265–271.

Zhang, Z., Liu, Q., and Wang, Y. (2018). Road extraction by deep residual u-net. *IEEE Geoscience Remote Sens. Lett.* 15, 749–753. doi:10.1109/lgrs.2018.2802944

Zhao, C., and Lu, Z. (2018). Remote sensing of landslides—A review. *Remote Sens.* 10, 279. doi:10.3390/rs10020279

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., et al. (2021). "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6881–6890.