SwinLS: Adapting Swin Transformer to Landslide Detection

Dong Zhao[1], Qi Zang[1], Zining Wang[1], Dou Quan[1] and Shuang Wang[1,†]

[1]School of Artificial Intelligence, Xidian University, Xian, 710071, China.

Abstract

Accurate detection of landslides plays an important role in post-disaster search and rescue operations. In this paper, we propose SwinLS for efficient landslide detection in remote sensing images using the swin transformer model. We explore how to efficiently utilize the self-attention mechanism in swin transformer for landslide detection tasks from two aspects. The first aspect is the spectral selection and data enhancement. The second aspect is to reduce imbalanced interference. After that, the performance of the improved swin transformer model is greatly improved, which provides a preliminary exploration for the application of the visual transformer model for remote sensing landslide detection tasks and even anomaly detection tasks. Finally, the proposed SwinLS, achieved the 2nd place in the test leaderboard with 73.99% F1 score, and it differs from the 1st place of 74.54% by only 0.55% F1 score.

Keywords

Landslide detection, remote sensing, swin transformer, multispectral imagery

1. Introduction

Landslides have become more frequent due to drastic climate change, surface activity, and accidents, threatening the lives and properties of residents in these areas. Accurate detection of landslides plays an important role in post-disaster search and rescue operations. As an efficient and convenient solution, automatic interpretation of landslide areas from remote sensing images has received extensive attention from scholars [1]. To advance this research, Ghorbanzadeh and Xu *et al.* [2] release a large-scale landslide detection dataset with pixel-level labels, named Landslide4Sense, and established a related benchmark.

The Landslide4Sense dataset contains multi-spectral imagery from multiple regions and cities collected by Sentinel-2 satellites. The data format is pixel blocks of size 128 with 14 spectrum bands including RGB, VEG (Vegetation Red Edge), NIR, WV (Water vapour), and SWIR. This dataset is finely marked by experts to pinpoint the location of the landslide. In the Landslide4Sense benchmark, Ghorbanzadeh and Xu *et al* [2] tried a series of classic convolution-based semantic segmentation models, such as ResUNet[3], PSPNet[4], ContextNet[5] and DeepLab [6], treating landslide detection as a binary supervised pixel-level classification task. Among these models, they found through experiments that ResUnet achieved the best verification performance on landslide detection tasks, which is due to its reasonable utilization of multi-scale features.

Nonetheless, we believe that this is not enough, because two important issues of landslide detection are ignored. The first is the spatial correlation of landslide data and the second is the imbalanced problem in landslide detection. For the former, we were motivated by the observation that the spectra after the collapse of the slopes exhibited often strong similarities. For the latter, we are inspired by the category statistics in Ghorbanzadeh and Xu's paper[2], showing that the proportion of landslides is much smaller than that of non-landslide, which is in line with the anomaly detection problem. To address these issues, we introduce the swin transformer [7] model to capture the relationship between landslide regions and design a training strategy for it to solve the imbalance problem.

The swin transformer is a recently proposed vision transformer model that has demonstrated strong performance on numerous tasks [7]. The key technology enabling this model is the self-attention mechanism, which aggregates spatial relationships to extract semantic features. However, it is not a good way to directly apply this model to multi-spectral remote sensing data for landslide detection, such as the Landslide4Sense dataset, because all spectral segments in multispectral contain target information. Those useless spectra will introduce massive noise in the feature aggregation process of the self-attention mechanism. Therefore, we first performed spectral selection experiments to determine which spectra are suitable for performing self-attention based feature aggregation. Finally, we use the RGB spectrum to

CDCEO 2022: 2nd Workshop on Complex Data Challenges in Earth Observation, July 25, 2022, Vienna, Austria

† Shuang Wang is the Corresponding author.

zhaodong01@stu.xidian.edu.cn (D. Zhao); qzang@stu.xidian.edu.cn (Q. Zang); 21171213901@stu.xidian.edu.cn (Z. Wang); quandou@xidian.edu.cn (D. Quan); shwang@mail.xidian.edu.cn (S. Wang)

https://github.com/DZhaoXd (D. Zhao)

train the swin transformer model. To complement it, we use CUTMIX [8] and random rotation data augmentation to prevent overfitting of larger capacity models.

To solve the imbalance problem in landslides detection, we design a two-stage balanced training strategy to make the model better focus on foreground (landslides) categories. In the first stage, we train the feature extractor and classifier with weighted cross entropy loss to get better feature representation. In the second stage, we fix the feature extractor and fine-tune the classifier with ordinary cross entropy loss to weaken the bias of the classifier. This strategy better mitigates the misleading of the classifier due to the imbalance between landslide classes and non-landslide classes.

Finally, the proposed method called SwinLS, achieved the 2nd place in the test leaderboard with 73.99% F1 score, and it differs from the 1st place of 74.54% by only 0.55% F1 score.

2. Methods

As shown in Figure 1, SwinLS is a network of codec structure, and there are hop links between codecs. Its encoder E is composed of the base structure of swin transformer, which has a powerful feature representation capability. Its decoder D uses a convolutional structure for decoding and fusing multi-level features for output.

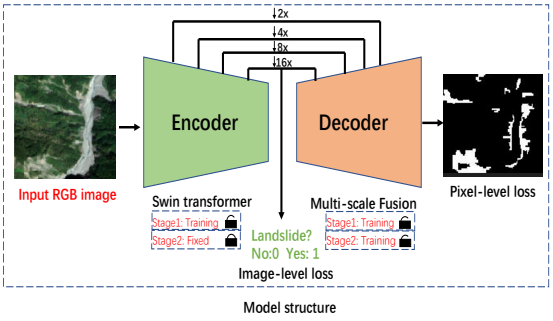

Figure 1: Network structure diagram.

For self-attention mechanism in swin transformer to work better in the landslide detection, we performed spectral selection experiments (see Tabel 1 and Figure 1), Finally, we selected the RGB spectrum from the multi-spectral input into the model. To alleviate the foreground and background imbalance in landslide detection, we design a two-stage training strategy. In the first stage, the codecs are trained simultaneously. For any input samples $x_i \in R^{w\times h\times 3}$, we use weighted cross-entropy loss L^{wce} and Lovasz loss L^{lov} [9] for balanced training as follows,

$$\arg\min_{E,D} L^{wce} + L^{lov} + L^{ice}. \tag{1}$$

The L^{ice} loss is the image-level loss performed in high-level semantic features in the encoder to assist training, which is defined as follows,

$$L^{ice} = -\frac{1}{|\mathcal{X}|}\sum_{x_i\in\mathcal{X}} \delta(y_i)\log MP(E(x_i)), \tag{2}$$

where δ is pointer function. When there is a pixel stand for positive sample (landslide) in y, its value is 1, otherwise it is 0. $MP(\cdot)$ is a fully connected layer with a global pooling operation. $\mathcal{X}$ stands for the total data set. The L^{wce} loss is defined as follows,

$$L^{wce} = -\frac{1}{|\mathcal{X}|}\sum_{x_i\in\mathcal{X}} \frac{N_{neg}}{N_{pos}} y_i \log D(E(x_i)), \tag{3}$$

where N_{neg} stands for the number of negative samples (non-landslides) and N_{pos} stands for the number of positive samples (landslides) in any input image x. As mentioned in [10], this re-weighting method can play a positive role in balancing the feature distribution of positive and negative samples. However, the classifier will still be biased. Therefore, in the second stage, we fix the trained encoder E and use the standard cross-entropy loss L^{ce} to train the decoder D.

$$\arg\min_{D} L^{ce} + L^{ice}. \tag{4}$$

3. Experiment

In this section, we show the performance of the methods proposed above, respectively. Due to the limited number of submissions of test data in the final stage, the data provided in our ablation experiments are all performance on the validation set.

Spectral selection Since the Transformer model needs to perform feature aggregations using self-attention mechanism to extract high-level semantic features. If irrelevant spectral information occupies dominant information, it will have a significant impact on the performance of swin transformer. To this end, we perform a set of experiments verifying the effect of different spectral inputs, as shown in Table 1.

In Table 1, we discovered an interesting phenomenon. With the increase of spectral banks, the performance of the fully convolutional models, such as deeplabv3 and Unet, show a gradually increasing trend, while the performance of swin transformer is severely degraded. We find out it is because the dimensionality enhancement in the fully convolutional model may attenuate the negative effects of irrelevant channels. Swin transformer, on the other hand, uses the dot product to preform the self-attention mechanism. When the spectral content unrelated to the landslide dominates, the attention is seriously dissipated, which makes the aggregated features

Table 1
Spectral selection experiments. In this table, the RGB denotes the red, green, and blue spectral. SWIR denotes the 3-band far infrared in Sentinel-2. NGB denotes the near-infrared, green, and blue spectral. NIR denotes the near-infrared spectral. PCA refers to the use of dimensionality reduction techniques [11] for compressing the original 14 banks into 3 banks. Besides, the encoder of unet model is replaced by resnet-32 and the encoder of deeplab model is also resnet-32. The encoder of swin transformer is swin-B. The metrics reported in the table are F1 scores.

Input spectral banks	Input banks	Swin	Deeplabv3	Unet
RGB	3	**65.6**	58.0	59.2
SWIR	3	55.6	50.2	52.1
NGB	3	60.8	**59.2**	58.9
PCA [11]	3	49.5	46.8	52.4
RGB + NIR	4	63.3	57.2	59.4
RGB + SWIR	6	58.2	55.9	59.8
RGB + NIR + SWIR	7	54.8	57.5	60.0
All banks	14	55.8	57.8	**61.1**

contain a lot of noise and are less discriminative. Through the above experiments, we selected the RGB spectrum as the input of swin transformer. Moreover, we clearly show a visualization of the dissipation of swin transformer's attention as the spectrum increases, as shown in Figure 2. This figure further verifies the above conclusion.

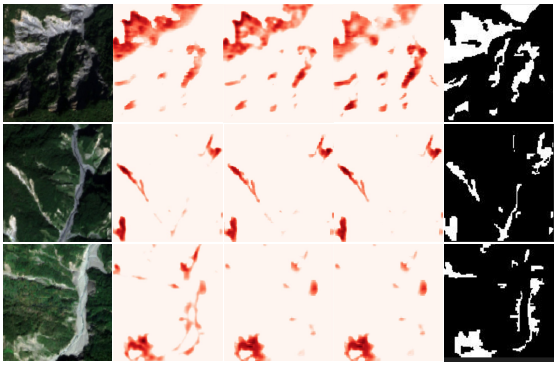

Figure 2: Visualization of the feature activation map of the swin transformer when inputting different spectral banks. We show the features from the last layer of swin transformer model in the training set.The redder the feature activation diagram, the greater the response.

In addition, Table 1 also shows that the swin transformer without any enhancements shows a very good baseline performance after properly selecting the spectrum. Therefore, our subsequent implementations rely on this strong baseline model to further improve the performance for detecting landslide.

Data augmentation When the task of landslide detection only uses RGB spectrum, the data pattern will be relatively simple, which increases the risk of overfitting. In addition, the swin transformer model has a large capacity and is easier to memorize and lose generalization under such simple data. To this end, we designed data augmentation experiments to verify the transformation methods for landslide detection using only RGB spectral information, as shown in Table 2. We also add the Unet model that uses all banks to compare with it.

Table 2
Data augmentation experiments. For both models, we randomly flip the input data as the baseline. The metrics reported in the table are F1 scores.

Transformation	Swin transformer	Unet
None (baseline)	65.6	61.1
color enhancement	62.1	60.3
cutout [12]	65.9	62.1
cutmix [8]	66.0	62.6
rotate and shift	**69.8**	**63.7**

Table 2 shows that random color augmentation degrades the performance of swin transformer, while it improves the Unet model. We analyze that this is because the RGB samples to be tested are also collected from mountainous areas, and the color space is not rich, so color enhancement leads to invalid generalization. The purpose of these two strategies, cutout and cutmix, is to disrupt the spatial layout of images so that the model can learn robust representations, and both slightly improve the performance of the two models. For swintransformer, the most effective way to enhance the data is to rotate and translate the data, which directly improves the F1 score by 4.2%. This augmentation increases the difficulty of capturing the relationship between landslides, which is very effective for swin transformer model. For unet, although this method is effective, the overall improvement strength is not as good as that of swin transformer. In general, after the data enhancement of rotation and translation, the F1 score of transformer is 6.1% higher than that of unet.

Balanced training We tried multiple sets of methods for balanced training, to verify the effectiveness of these methods, as shown in Table 3. Among them, normal training is a one-stage training method using cross entropy loss. For weighted cross entropy loss, we use the scale coefficients of positive samples and negative samples as the loss weighting coefficient of negative samples. This method has achieved a certain improvement by weighting the positive and negative pixels, but the improvement is relatively limited. Focal loss [13] balances easy and hard samples by modifying their gradients for back propagation, and is also used in many unbalanced scenarios. But on this task, the performance degrades when this loss is added. Our analysis is that it has a great influence on the gradient, and inappropriate hyperparameters will greatly affect the performance. Lovasz loss

[9] is a loss that directly optimizes the IoU coefficients, which is efficient and used as the first stage loss for our balanced training. Balanced training achieves the best performance, which further corrects the bias of the classifier. Finally, balanced training improves the F1 score by 4.1% on the basis of baseline. The results of this strategy are visualized in Figure 4.

Table 3
Balance training experiments. We use swin transformer with data augmentation and normal training (only using cross entropy loss) as the baseline model. The metrics reported in the table are F1 scores.

Training	Swin transformer	Unet
Normal training	69.8	63.7
Weighted cross entropy	70.8	64.9
Focal loss [13]	68.2	61.8
Lovasz loss [9]	72.3	66.4
Balance training	**73.9**	**67.7**

Self-training. We also use self-training techniques to further improve the model performance, as shown in Table 4. We verify who to select pseudo-labels is suitable for landslide detection. We sorted the output probabilities predicted in the previous stage, selected the top $\lambda\%$ high-confidence pixel-level pseudo-labels and added them to the training data for self-training. The number of percentages selected should be explored, *i.e.* λ.

Table 4
Self-training experiments with different λ values. ST denotes the self-training.

λ	Precision (%)	Recall(%)	F1(%)
- (Before ST)	73.4	74.7	73.9
50%	65.2	**80.5**	72.7
70%	69.3	79.5	73.7
90%	72.4	77.1	74.9
100%	**78.2**	74.2	**76.1**

In Table 4, we found that when λ is small, the accuracy rate after self-training will degrade seriously, but the recall rate will improve significantly. This is because when the λ is small, the selected landslide area is only located in the center of the landslide, and the pixels in the surrounding area will be ignored due to low confidence. This makes the self-trained model tend to predict all surrounding similar blocks as landslides, resulting in increased over-detection of landslides. As the selected landslide area continues to increase, the accuracy of the model continues to rise, and the recall rate begins to decline. This shows that with the addition of many inaccurate pseudo-labels, it has played a strong role in preventing over-detection. And the model can learn more knowledge about the samples to be tested from the noisy training data, which increases the accuracy.

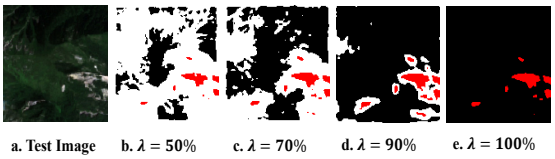

Figure 3: Visualization of pseudo labels when different lambda values are selected. In the pseudo labels, black represents Class 0 (non landslide), red represents class 1 (landslide), and white represents ignored classes.

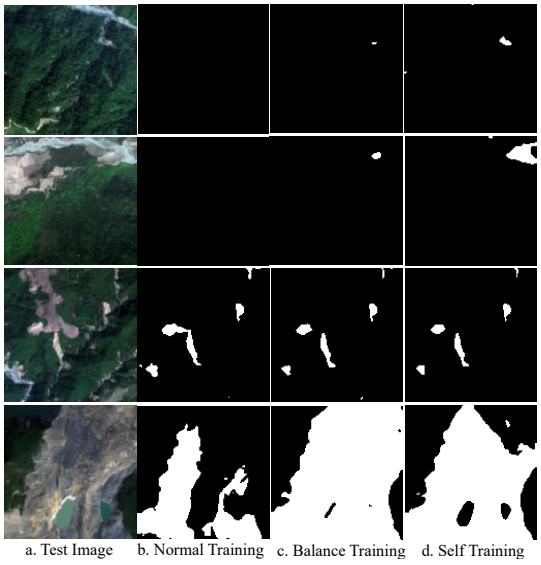

Figure 4: Visualization of model output after adding different strategies.

In practical application, we can reasonably design this parameter according to the requirements. When we need to roughly find more areas that may be landslides, we design a smaller lambda. When we need to detect the landslide area more accurately, we design a larger lambda.

Furthermore, we visualize example plots for picking pseudo-labels with different λ values, as shown in Figure 3. We also visualize the output of the self-trained model in Figure 4, which further supports the above conclusion.

Acknowledgments

This work is in part supported by Key Research and Development Program of Shannxi (Program No.2021ZDLGY01-06), Key Research and Development Program of Shannxi (Program No. 2022ZDLGY01 -12) and National Key R&D Program of China under Grant No. 2021ZD0110404.

References

[1] J. Gawlikowski, S. Saha, A. Kruspe, X. X. Zhu, An advanced dirichlet prior network for out-of-distribution detection in remote sensing, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–19.

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

[3] F. I. Diakogiannis, F. Waldner, P. Caccetta, C. Wu, Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data, ISPRS Journal of Photogrammetry and Remote Sensing 162 (2020) 94–114.

[4] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.

[5] R. P. Poudel, U. Bonde, S. Liwicki, C. Zach, Contextnet: Exploring context and detail for semantic segmentation in real-time, arXiv preprint arXiv:1805.04554 (2018).

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE transactions on pattern analysis and machine intelligence 40 (2017) 834–848.

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[8] G. French, T. Aila, S. Laine, M. Mackiewicz, G. Finlayson, Semi-supervised semantic segmentation needs strong, high-dimensional perturbations (2019).

[9] M. Berman, A. R. Triki, M. B. Blaschko, The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4413–4421.

[10] B. Zhou, Q. Cui, X.-S. Wei, Z.-M. Chen, Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9719–9728.

[11] A. M. Martinez, A. C. Kak, Pca versus lda, IEEE transactions on pattern analysis and machine intelligence 23 (2001) 228–233.

[12] T. DeVries, G. W. Taylor, Improved regularization of convolutional neural networks with cutout, arXiv preprint arXiv:1708.04552 (2017).

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.