Landslide Detection of 6G Satellite Images Using Multi-level Transformer Network

Dong He[1], Liang Xi[1](✉), and Lu Liu[2](✉)

[1] School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China
xiliang@hrbust.edu.cn

[2] College of Biomedical Information and Engineering, Hainan Medical University, Haikou 571199, China
liulu@hainmc.edu.cn

Abstract. The 6G satellite system can help humans detect natural disasters and respond quickly through a low-altitude full-coverage network. Analyzing and identifying landslide images captured by satellites can help humans address the various hazards posed by landslides. Recently, deep learning models have been developed rapidly and demonstrated the effectiveness of landslide detection. Many models use convolutional neural networks (CNN) to extract the features of landslide images for landslide detection. However, CNN-based models cannot obtain global semantic information of images, resulting in low accuracy of landslide detection or some misjudgments. In this paper, we adopt a pre-training feature extraction network and an unsupervised multi-level transformer autoencoder for landslide detection. We first extract multi-scale features from the pre-training network, then reconstruct the image features using an autoencoder transformer network with a U-Net shape, which can better get global semantic information. We use the Bijie landslide dataset captured by satellites for experiments. The experimental results show that, compared with the original CNN model, our method can improve the detection accuracy and effectively distinguish landslide and non-landslide image data.

Keywords: 6G satellite image · landslide detection · autoencoder · vision transformer

1 Introduction

Landslide is a common geological disaster, which can bring unpredictable damage to the natural environment, public infrastructure, and human life [1, 2]. Recently, with the development of 6G satellite communication network and remote sensing technology, people can apply the high-resolution satellite images to disaster management, landslide detection and identification [3, 4]. There are various landslide features in the landslide images, how to comprehensively use them for accurate and effective landslide detection is extremely meaningful to reduce the economic risks and protect human life.

A. Li et al. (Eds.): 6GN 2022, LNICST 504, pp. 100–110, 2023.
https://doi.org/10.1007/978-3-031-36011-4_9

Traditional landslide detection methods usually use statistical features such as image texture and color, to detect landslide. Martha et al. [5] investigated landslide detection using spectral, shape, and contextual information from landslide images. In addition, some machine learning methods, such as support vector machine [6], random forest [7], and genetic algorithm [8], combine statistical methods for landslide detection. However, machine learning methods do not have good feature extraction capability and cannot automatically detect.

Deep learning models have powerful representation learning capability, can automatically extract features and apply them to landslide detection. Specifically, the convolutional neural network (CNN) has surpassed traditional statistical feature extraction and machine learning methods in various image-based tasks [9]. Using optical data and terrain factors from the Rapid Eye satellite, Ghorbanzadeh et al. [10] analyzed the performances of different CNN-based models for landslide detection, and the experimental results showed that CNN-based deep learning methods can achieve better results. Ji S et al. [11] focus on landslide detection from satellite images using CNN-based methods based on the attention mechanism to enhance the CNN to extract more distinctive landslide feature representations. However, CNN focuses on extracting local image features, and it is difficult to learn the global context features of images [12].

Vision transformer (ViT) [13] is the first vision transformer. It proves that the NLP transformer [14] framework can be transferred to the image recognition and detection task [15, 16] with excellent performance. Compared with CNN, ViT can learn image contextual features through patch-level image and global attention mechanism [17]. To capture local and global information of image, we focus on combining transformers on multiple levels for better image detection and discrimination.

Hence, we propose an unsupervised deep learning model for landslide detection. The model mainly contains a feature extraction network and a multi-level transformer autoencoder framework. Firstly, we extract features from a pre-training network; Then the image features can be reconstructed using a multi-level transformer network. In this way, not only fine-grained features but also coarse-grained features can be extracted. At last, we fuse the two kinds of features of satellite images for the landslide detection tasks.

2 Related Work

2.1 6G Satellite Communication

6G satellite communication will play an important role in the global communication and monitoring. Through the integration of satellite communication and terrestrial network, the goal of mobile network coverage in the entire global area will be achieved, realizing user access at any time and business continuity. Satellite networks can provide various services to vast areas, and satellite services have been evolved from traditional voice and broadcast services to broadband Internet services [18], which can monitor the behaviors of mobile users, buildings, aircraft, ships, emergency base stations, and forecast of weather conditions and natural disasters [19]. By transmitting the detected statistical information or captured satellite images to the ground, combining with AI intelligent technology, we can effectively detect weather or geological disasters, such as landslide detection of satellite images.

2.2 Deep Learning Models

U-Net. The original intention of U-Net [20] is to perform medical image segmentation. The U-shaped network structure is used to obtain context information and position information. It is a simple and efficient deep learning network and easy to build and train.

Therefore, we employ U-Net to capture both shallow and deep features simultaneously. Because both low-level features and high-level semantic features are important in landslide detection of satellite image. The skip connection structure (feature splicing) of U-shaped structure is quite effective, which can help distinguish abnormal landslide images from normal non-landslide images. The framework of U-Net is shown in Fig. 1.

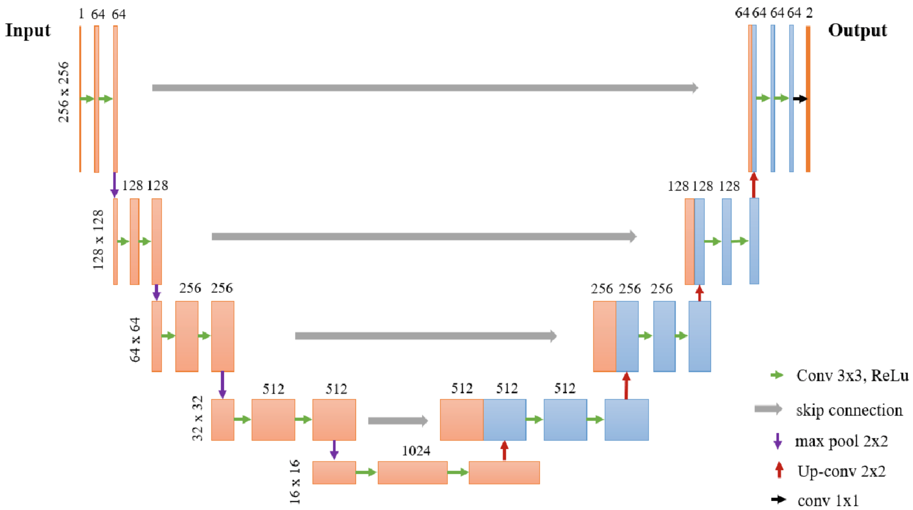

Fig. 1. The U-Net framework

Vision Transformer. ViT mainly processes images by splitting them into multiple patches; then the two-dimensional patches are converted into a one-dimensional vector. The class vector and position vector are also added as the model input. Due to the spatial connections between multiple patches, ViT achieves good performance in downstream tasks [21]. Our framework mainly involves the transformer encoder and decoder with the attention mechanism. The framework of the original vision transformer is shown in Fig. 2.

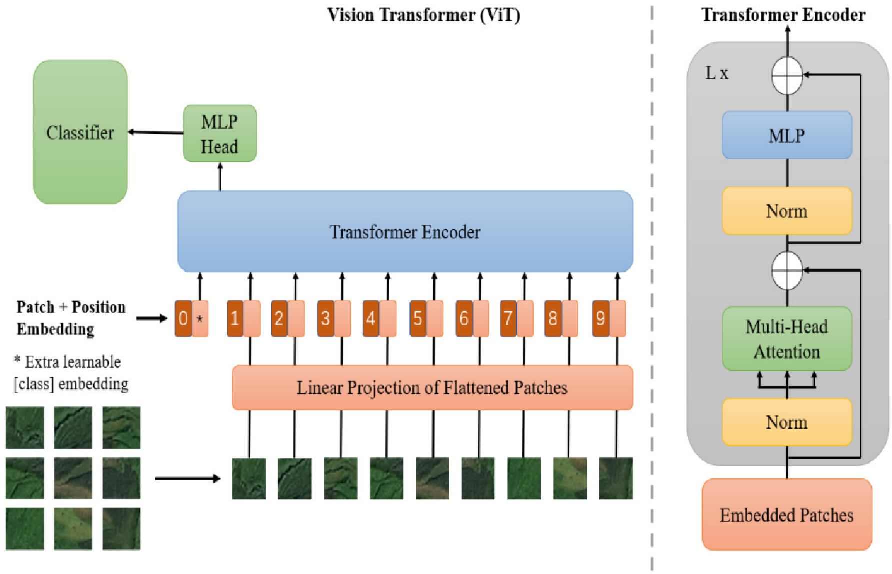

Fig. 2. The ViT framework

Considering the ability of the transformer to handle contextual semantics, we utilize transformers of different granularities to extract image features to provide multi-scale features for image reconstruction. In addition, due to the skip connection structure of U-Net that can reduce information loss, we apply multiple transformers with different granularities to U-Net to obtain a U-shaped transformer autoencoder, which can effectively extract the image features, and use the reconstruction error map to distinguish landslide and non-landslide images.

3 Methodology

We apply an unsupervised multi-level transformer encoder-decoder framework to landslide detection from satellite images, shown in Fig. 3. Our model contains a pre-training CNN network and a multi-level autoencoder with transformer structure. First, we introduce the structure of pre-training network to extract latent features from images; Next, we present the specific implementation details of the multi-level transformer autoencoder with U-Net to reconstruct the features; Finally, we give the loss function and the evaluation method of anomaly score.

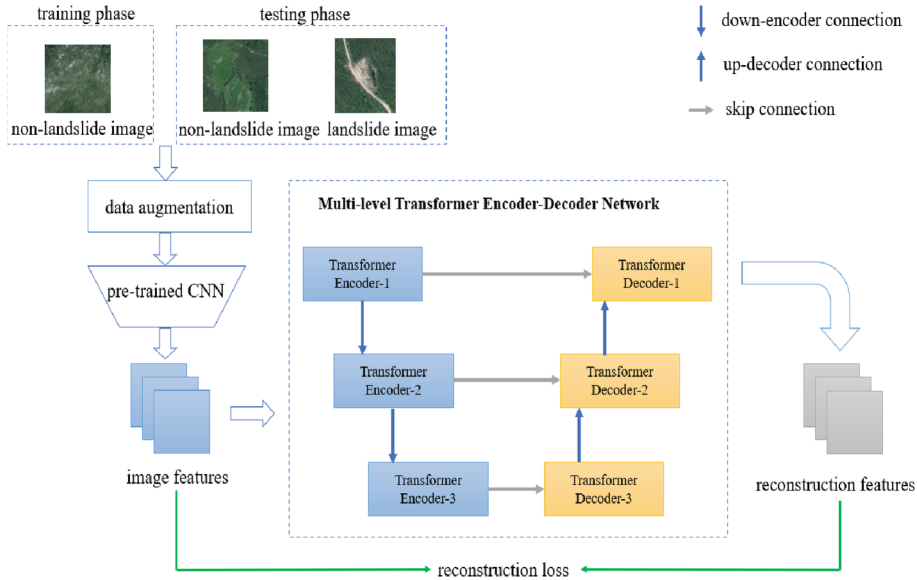

Fig. 3. Model framework

3.1 Pre-training Network

Due to the different distribution of landslide and non-landslide images in the feature space, we perform feature extraction through a pre-training network based on CNN, which can map the original images to the feature space.

We utilize the pre-training Resnet-18 [22] on ImageNet [23]. And the reason we use Resnet-18 as a pretrained network is because of its simplicity and effectiveness. We not only use the output features of the last layer, but also connect the features of the middle three layers along the channel to form feature representations of different levels. Then, the fused features are taken as the whole image features and put into the multi-level auto-encoder for reconstruction. The formula is as follows:

$$z = F_{Resnet18}(x) \tag{1}$$

3.2 Multi-level Transformer Network

Compared with the traditional convolutional neural network, the transformer has greater abilities of the extraction and reconstruction of image features due to its attention mechanism. To capture both coarse-grained and fine-grained image features, we use multiple scale transformer encoders and decoders. In addition, we combine the advantages of U-Net, the encoder and decoder in the same level use skip connections for information transfer. The compression or decompression information of the previous level is sequentially transferred between the encoders and decoders in different levels. The following is an introduction to the entire transformer encoder and decoder framework.

Multi-Level Transformer Encoder. Consistent with the traditional vision transformer processing, we convert the feature maps to patches, and add zero padding token and positional encoding as the input of the transformer encoder. We divide the feature map of the H × W landslide image into patches with a block size of $P_H \times P_W$, and such patches have a total of $N_H \times N_W$ blocks. The output of the level-1 encoder includes a latent head vector and a latent feature vector. The input of the level-2 encoder is the latent head vector of the level-1 encoder, and the latent feature vector is used as the input of the level-1 decoder for feature reconstruction. Similar to the level-1 encoder, the level-2 and level-3 encoders also follow this transfer process. The formula of the encoder of the level-*i* transformer is as follows,

$$\mathrm{E}_0^{(\mathrm{n})}, \left(\mathrm{E}_1^{(\mathrm{n})}, \ldots, \mathrm{E}_{P_i \times P_i}^{(\mathrm{n})}\right) = \mathrm{F}_{encoder}\left(z_0, z_1^{(\mathrm{n})}, \ldots, z_{P_i \times P_i}^{(\mathrm{n})}\right) \tag{2}$$

where $\mathrm{z}_1^{(\mathrm{n})}, j = \{1, ..., P_i \times P_i\}$ are the feature vectors after pre-training. $\mathrm{E}_0^{(\mathrm{n})}, \mathrm{E}_\mathrm{j}^{(\mathrm{n})}, j = \{0, ..., P_i \times P_i\}$ are the output of the level-*i* encoder. z_0 is the zero-padding token, *i* is the number of levels of the encoder, and *n* is the number of patches.

Multi-Level Transformer Decoder. The input of the decoder consists of two parts, one is the latent feature vector from the output of the same level encoder, and the other is the output of the previous level decoder, which is the latent head vector. In addition, the output of the decoder is the reconstruction feature for each patch, and they are all reconstructed into a feature map of $H \times W$ shape. The formula of the decoder of the level-*i* transformer is as follows,

$$\hat{z}_1^{(\mathrm{n})}, \ldots, \hat{z}_{P_i \times P_i}^{(\mathrm{n})} = \mathrm{F}_{decoder}\left(E_0^{(n)}, E_1^{(n)}, \ldots, E_{P_i \times P_i}^{(n)}\right) \tag{3}$$

where $E_j^{(n)}, j = \{1, \ldots, P_i \times P_i\}$ are the outputs of the same level encoder using skip connections, and $E_0^{(n)}$ is the output from the previous level transformer decoder. $\hat{z}_j^{(n)}, j = \{0, ..., P_i \times P_i\}$ are the outputs of the current decoder.

Based on the image features extracted by pre-training, the feature map can be reconstructed by a symmetric multi-level U-shaped transformer to get the reconstruction error map $\hat{z}$:

$$\hat{z} = R(z) \tag{4}$$

where $R(z)$ represents the multi-level transformer autoencoder network. z is the image features extracted by the pre-training CNN network.

3.3 Loss Function

The reconstruction loss makes the output features reconstructed from the multi-level transformer similar to its input image features. Specifically, we calculate the $L2$ distance between z and the reconstructed, $\hat{z}$, which ensures that our model can reconstruct image features contextually similar to non-landslide images in the training phase. The reconstruction loss is shown below,

$$L_{rec} = \mathbb{E}\|z - \hat{z}\|_2 \tag{5}$$

3.4 Landslide Detection

Since the reconstruction error can reflect the differences between the landslide and non-landslide images, we can directly use the reconstruction error as the anomaly score. When the normalized anomaly score exceeds a predefined threshold, this image is classified as a landslide image.

$$S(z) = L_{rec}(z) \tag{6}$$

$$N(S(z)) = (S(z) - \min(S(z)))/(\max(S(z)) - \min(S(z))) \tag{7}$$

S(z) represents the anomaly scores, and $N(S(z))$ represents the min-max normalization over the whole samples.

Algorithm 1 Multi-level Transformer Autoencoder for Landslide Detection

Input: image, $x = \{(x_i, y_i)\}_{i=1}^{N}$, Iterations, L, threshold, τ
Initialize l=0, S = 0

1. Get z by feeding x into the pre-training network
2. while l<L do
3. Sample $\{(x_1, y_1), ..., (x_m, y_m)\}$ from non-landslide and landslide images
4. For i = 1 to 3 do
5. Split the feature map, z, to patches, $P_i \times P_i$, and add position encoding, then fed into the level-i encoder
6. The latent vectors, e_i, is calculated using Equation (2)
7. For i = 3 to 1 do
8. Fed the latent feature vector into the level-i decoder
9. The reconstructed feature map, z_i, is calculated using Equation (3)
10. The reconstruction loss is calculated using Equation (5)
11. Calculate the anomaly score, S, using Equations (6) and (7)
12. Classify each image as a landslide image if S >τ, otherwise the non-landslide image.
13. End

4 Experiment

4.1 Datasets Description

Ji S et al. [12] created an open-source large landslide dataset named the Bijie landslide dataset. So, we use it to evaluate our model. The dataset consists of satellite optical images, shapefiles of landslides' boundaries and digital elevation models which contains 770 landslide images and 2003 non-landslide images. They were cropped from the TripleSat satellite images. Some image examples of the Bijie landslide dataset are shown in Fig. 4.

a) Satellite non-landslide images

b) Satellite landslide images

Fig. 4. The Bijie landslide dataset

We treat the experiment as a one-classification task to detect landslide images. The training phase uses only non-landslide images, and the test phase contains both landslide and non-landslide images. We divide 70% of the 2003 non-landslide images as normal training data, and the rest of the non-landslide images and landslide images are used as the test data. Our purpose is to identify landslide images from a wide variety of landslide and non-landslide images to verify the effectiveness of our method. Details of the datasets and corresponding data distributions used in our experiments are shown in Table 1.

4.2 Performance Metrics

We use five metrics to evaluate the model, including AUC, Accuracy, Precision, Recall, and F1-score. AUC is the area under the receiver operating characteristic curve. The other metrics are described and presented as follows.

Table 1. Details of Bijie dataset

Dataset	Training data	Testing data
Bijie [12]	1402 non-landslide images	601 non-landslide images and 770 landslide images

Accuracy indicates the ratio of correctly predicting landslide and non-landslide images in the overall images. The formula is shown as follows,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Precision indicates the ratio of correctly predicting landslide images in the overall images. The formula is shown as follows,

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

Recall indicates the ratio of correctly predicting landslide images among all landslide images. The formula is shown as follows,

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

F1-score is calculated by considering both precision and recall, and is the weighted average of both. The formula is shown as follows,

$$F1 - score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

where false negative (FN) indicates the number of non-landslide images that were falsely judged to be landslide images. The false positive (FP) represents the number of landslide images that were falsely identified as non-landslide images. The true positive (TP) represents the number of non-landslide images correctly identified as positive. The true negative (TN) is the number of landslide images correctly identified as negative.

4.3 Implementation Detail

Before training and testing, we randomly rotate the images by 10 degrees and crop them to the same size of 256 × 256. The entire network contains three transformer encoders and decoders, respectively. And the patch size is set as 4. Different from the original transformer structure, we set the attention layer dimension to 512 and the feedforward layer dimension to 1024. We used the Adam optimizer to train the multi-level transformer, and the learning rate is set to 0.0001. The model is trained by 200 epochs, and the batch size is 2. We use U-net as a comparison method, which adopts the conventional CNN network structure to reconstruct the pre-trained image features. We conduct our experiments on an NVIDIA RTX 3090 GPU on Pytorch 1.9.1.

5 Result and Discussion

Table 2 shows the results of one-class landslide detection on the Bijie dataset. Compared to CNN and U-Net as an autoencoder, our multi-level transformer autoencoder model has a significant improvement in AUC due to the multi-level feature extraction. In addition, our model has a slight improvement over the U-Net in accuracy, precision, recall, and F1-score. Especially, the recall can reach 92.5% on the basis of certain accuracy performance, which shows that our model can be effectively used for landslide detection.

Table 2. Landslide detection results

Method/Metric	AUC	Accuracy	Precision	Recall	F1-score
CNN	0.564	0.573	0.544	0.885	0.674
U-Net	0.597	0.602	0.589	0.907	0.726
Our model	**0.658**	**0.622**	**0.61**	**0.925**	**0.735**

6 Conclusion

We apply a feature extraction network and a multi-level transformer network to feature extraction and reconstruction of landslide images to perform landslide detection task. Based on the U-Net network structure, we built a U-shaped transformer autoencoder, which can extract and reconstruct image features from both local and global perspectives. Experiments show that compared with the U-Net model using the traditional convolutional neural network, our model has different degrees of improvement in indicators such as AUC and F1-score. In addition, our method can reach a higher recall result based on ensuring a certain precision, which is exceptionally suitable for natural hazard detection like landslides. This will be of great help in the identification of landslide images captured by 6G satellites.

Acknowledgment. This work was supported by Heilongjiang Province Natural Science Foundation under Grant LH2022F034.

References

1. Chen, W., et al.: Spatial prediction of landslide susceptibility using an adaptive neuro-fuzzy inference system combined with frequency ratio, generalized additive model, and support vector machine techniques. Geomorphology **297**, 69–85 (2017)
2. Cullen, C.A., Al-Suhili, R., Khanbilvardi, R.: Guidance index for shallow landslide hazard analysis. Remote Sens. **8**, 866 (2016)
3. Liu, Y., Gross, L., Li, Z., Li, X., Fan, X., Qi, W.: Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling. IEEE Access **7**, 128774–128786 (2019)

4. Liu, Y., Li, Z., Wei, B., Li, X., Fu, B.: Seismic vulnerability assessment at urban scale using data mining and GIScience technology: application to Urumqi (China). Geomat. Nat. Haz. Risk **10**(1), 958–985 (2019)
5. Martha, T.R., Kerle, N., Jetten, V., van Westen, C.J., Kumar, K.V.: Characterising spectral, spatial and morphometric properties of landslides for semi-automatic detection using object-oriented methods. Geomorphology **116**(1–2), 24–36 (2010)
6. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
7. Pradhan, B., Seeni, M.I., Nampak, H.: Integration of LiDAR and QuickBird data for automatic landslide detection using object-based analysis and random forests. In: Pradhan, B. (ed.) Laser Scanning Applications in Landslide Assessment, pp. 69–81. Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-319-55342-9_4
8. Dou, J., et al.: Automatic case-based reasoning approach for landslide detection: integration of object-oriented image analysis and a genetic algorithm. Remote Sens. **7**(4), 4318–4342 (2015)
9. Huang, G., Liu, Z., Pleiss, G., et al.: Convolutional networks with dense connectivity. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
10. 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. **11**(2), 196 (2019)
11. Ji, S., Yu, D., Shen, C., et al.: Landslide detection from an open satellite imagery and digital elevation model dataset using attention boosted convolutional neural networks. Landslides **17**(6), 1337–1352 (2020)
12. Raghu, M., Unterthiner, T., Kornblith, S., et al.: Do vision transformers see like convolutional neural networks? Adv. Neural. Inf. Process. Syst. **34**, 12116–12128 (2021)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, p. 30 (2017)
15. Yuan, L., Hou, Q., Jiang, Z., et al.: Volo: vision outlooker for visual recognition. arXiv preprint arXiv:2106.13112 (2021)
16. Wang, W., Xie, E., Li, X., et al.: Pvt v2: improved baselines with pyramid vision transformer. Comput. Vis. Media **8**(3), 415–424 (2022)
17. Lee, Y., Kang, P.: AnoViT: unsupervised anomaly detection and localization with vision transformer-based encoder-decoder. IEEE Access **10**, 46717–46724 (2022)
18. Botta, A., Pescapé, A.: On the performance of new generation satellite broadband internet services. IEEE Commun. Mag. **52**(6), 202–209 (2014)
19. Chini, P., Giambene, G., Kota, S.: A survey on mobile satellite systems. Int. J. Satell. Commun. Network. **28**(1), 29–57 (2010)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Liu, Z., Lin, Y., Cao, Y., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
22. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
23. Deng, J., Dong, W., Socher, R., et al.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)