Reg-SA–UNet++: A Lightweight Landslide Detection Network Based on Single-Temporal Images Captured Postlandslide

Chaoyang Niu, Ouyang Gao, Wanjie Lu, Wei Liu, and Tao Lai

Abstract—Landslide detection based on remote sensing images is an effective method for rapidly and accurately detecting landslide regions, which can aid in disaster prevention and mitigation. Landslide detection methods based on semantic segmentation can be used to delineate the scope of landslides while detecting their location. Most existing models use multimodal or geospatial data to improve accuracy. However, the large amount of data introduces additional parameters, consuming significant computing resources. Therefore, this article proposes Reg-SA–UNet++, a model for landslide detection, which uses a single-temporal image captured postlandslide. Reg-SA–UNet++ is based on UNet++ with the following modifications: deep supervised pruning is removed for fewer model parameters and improved detection accuracy; RegNet is employed to replace the convolutional blocks in the encoding process to reduce the number of parameters and improve feature acquisition and attention modules are added at the connection of the convolutional blocks of each layer to strengthen the model’s attention to landslide features. The overall accuracy and F1 score of the Reg-SA–UNet++ model for the constructed landslide dataset (93.37% and 92.41%, respectively) and landslide mapping (97.09% and 96.10%, respectively) verify the effectiveness of the proposed model in detecting landslides from remote sensing images.

Index Terms—Attention mechanism, deep learning, landslide detection, lightweight mode, optical remote sensing image.

I. INTRODUCTION

Landslide is a natural phenomenon caused by various factors, including river erosion, groundwater activity, rain inundation, earthquakes, and artificial slope cutting [1]. Currently, influenced by global climate change, seismic activity, and accelerated urbanization, the number of landslides has gradually increased, and the actual hazards and potential risks have become more prominent. Therefore, timely detection of landslides and determination of their area, scale, and distribution are crucial for disaster management and prevention of secondary disasters [2].

Traditional landslide detection methods are based on field investigation; this process requires significant effort and time. It also poses unanticipated risks to researchers [3]. Remote sensing has several advantages, including fast imaging speed, wide coverage, low cost, and low risk [4]. With the increasing number of remote sensing satellites, landslide detection using remote sensing images has become predominant in disaster prevention and reduction. Computer-based methods have significant advantages in remote sensing image processing. These methods systematically and quantitatively simulate researchers’ visual interpretation process in landslide detection. Early methods are based on statistics or shallow machine learning techniques. However, these statistical methods generally have complex classification rules that require extensive research experience [5], [6]. These methods can achieve acceptable landslide detection accuracy; however, the classification rules need to be adjusted manually for different areas. Therefore, this approach is not universal and has a low degree of automation. Landslide detection methods based on shallow machine learning techniques [7], [8] first compute data features through function mapping or artificial feature design, then use discrimination criteria to determine the landslide category. When landslides occur in simple backgrounds, the classifiers can complete the task efficiently and accurately. However, the results are easily affected by landforms with similar features to landslides because the shallow features cannot effectively describe original data. In 2006, with the advancement of artificial intelligence, the emergence of the deep belief network [9] initiated the research wave of deep learning. Compared with shallow machine learning, the neural networks used in deep learning have deeper structures and more layers of hidden nodes; therefore, they can better utilize a wide range of sample data for training and better exploit the inherent information of the data. To improve image processing performance, neural network models were gradually diversified; deeper and more complex structures were created and lightweight, high-performance models that can be easily deployed were introduced. The landslide detection methods based on deep learning include two main aspects as follows:

1) Methods based on target detection, such as the R-CNN, SSD, and YOLO series, can determine whether there is a landslide in the image. If there is a landslide, the method

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
further determines its location. Ju et al. [10] used Mask R-CNN to detect landslides automatically on the established loess landslide database and achieved a precision of 0.56, a recall of 0.72, and an F1 score of 0.63. This indicates that Mask R-CNN is robust for the automatic detection of loess landslides when using Google Earth imagery as the dataset. It also shows that the method has the potential to provide fast and accurate regional landslide hazard investigation. To improve the accuracy and inference speed, Cheng et al. [11] constructed a model, YOLO-SA, for landslide detection in remote sensing images. The model replaces the corresponding module in YOLOv4 with the group convolution (Gconv) and ghost bottleneck (G-beck) residual modules and adds an attention module. This method achieves an inference speed of 42 f/s while increasing the accuracy rate to 94.08%. Wang et al. [12] improved the feature stitching method of YOLOv5 for landslides of different shapes. They added adaptive spatial feature fusion to combine feature information of different scales and further used the convolutional block attention module (CBAM) to mine shallow feature information. The model had a detection rate of 74.01%.

2) Methods based on semantic segmentation, such as the UNet series, DeepLab series, and PSPNet series, separate landslide objects by pixels according to the semantics of the image. As a classic semantic segmentation model, UNet has been adapted by scholars to make it suitable for landslide detection. Liu et al. [13] proposed an automatic model to improve the missing and misplaced landslide detection results of UNet. The proposed model increases the input channel count of UNet from three to six to facilitate the input of geoscience features. A residual learning unit was added to the model to fully extract the sample features in order to deepen the network layer. This model outperforms UNet in terms of feature extraction and classification capabilities, and it solves the problem of gradient disappearance or explosion when the network is deepened; its landslide detection accuracy reaches 91.3%. Using ResNet34 for feature extraction, Wang et al. [14] proposed a landslide detection model based on change detection to avoid confusing landslides and bare soil; it uses pre and postdisaster remote sensing images for landslide detection and can achieve an accuracy of 97.69%. Training degradation may occur when the depth of the network layer increases. To address this issue, Qi et al. [15] proposed ResU-Net, a model for automatic landslide mapping. This model uses ResNet to replace each convolutional module in the UNet encoding process and achieves a landslide detection accuracy of 96%. However, it performs poorly in detecting small landslides and is susceptible to interference from landforms that resemble the spectral signature of landslides. The authors of [13], [14], and [15], the abovementioned scholars, enhanced UNet to meet the needs of landslide detection in remote sensing images. UNet has the advantages of a simple structure, few parameters, high segmentation accuracy, and strong variability, making it one of the most commonly used model frameworks in landslide detection.
Bui et al. [16] also proposed an H-BEMD method for detecting landslide regions. To avoid the influence of lighting conditions, this method detects landslides based on the tonal features of images but cannot avoid interferences, such as landform hue features, because it only uses the hue component to detect landslides. Zhang et al. [17] proposed a deep learning model, SMDRF-Net, based on a semi-supervised multitemporal deep representation fusion network. Deep learning can transform original images into abstract high-level representations. The exact outline of the target becomes more difficult to preserve as the network becomes deeper. To solve this problem, SMDRF-Net incorporates object-level spectral features to exploit deep representations of objects. As a result, it can retain accurate boundaries of landslide objects while reducing noise.

The landslide detection method based on target detection can quickly detect the location of a landslide; however, it does not characterize its shape or coverage. Alternatively, the semantic-segmentation-based method can mark the location of the landslide and accurately characterize its contour and coverage. Current article focuses primarily on three aspects of detecting landslides based on semantic segmentation. The first one employs remote sensing images of pre and postlandslides to detect it based on the change detection principle because landslides produce more noticeable changes than surrounding landforms. The second aspect entails the use of geoscience data, such as the digital elevation model, digital surface model, and digital terrain model, as an aid to enrich landslide features. The third aspect entails the use of only postlandslide remote sensing images. Compared with this method, the data used in the first two methods contain more features. However, some data are challenging to obtain in practical applications, making relevant research difficult. Concurrently, the detection results can easily fall short of expectations if the various data have a weak time correlation.

Moreover, the consumption of computing resources increases with more data input, making the model more demanding on the hardware. Landslide detection requires only a small amount of data through single-temporal images captured postlandslide, which reduce data and hardware resource requirements. However, this method has few available features and high background noise because it uses a single data type.

Given the aforementioned problems, this article uses UNet++ as the basic framework in the proposed model. Nevertheless, the model performs landslide detection based on single-temporal images captured postlandslide, and it is highly practical and shows good accuracy. Compared with UNet++, the enhancements introduced in this article are as follows:

1) The lightweight network, RegNet [18], is used to replace the convolutional blocks in the UNet++ encoding process, which can effectively reduce the number of parameters and improve the feature capture ability of the model.
2) Shuffle attention [19] is introduced between the encoding and decoding structures of each layer to strengthen the model’s understanding of semantic information and make the model prioritize the landslide ontology, enabling it to effectively reduce the interference caused by the background.
3) The deep supervision-pruning operation of UNet++ is removed to further improve the accuracy and reduce the number of parameters.

Compared with existing relevant articles, the expected benefits of the current article are as follows:

1) The landslide detection accuracy and generalization ability of the model are effectively improved.
2) The model only uses single-temporal images captured postlandslide, reducing model dependence on original data, complex data preprocessing steps, and a large amount of computing resources.
3) The model uses fewer parameters; therefore, the hardware requirements of the model are also few.

The rest of this article is organized as follows: The “Dataset” section describes the landslide dataset required to train the model. The “Methodology” section introduces the basic framework of the model, the reconstruction of the model to improve its accuracy, and the loss function used for training. The “Experimental Setup” section describes the experimental parameter settings. The “Evaluation Criteria” section describes the evaluation criteria for the experimental results. The “Results and Discussion” section presents the experimental results and discusses our model. Finally, the “Conclusion” section summarizes the full text.

II. DATASET

A. Construction of Landslide Remote Sensing Image Dataset

Based on the high-precision remote sensing images and interpretation datasets developed based on the landslide disasters that occurred in Sichuan and its surrounding areas published by Zeng et al. [20], we identified several areas where landslides frequently occur, and 1215 remote sensing images of landslides were collected to construct the dataset. The areas mainly include Qiaojia County and Ludian County in North-eastern Yunnan Province, China [11]; Lanzhou and LinxiaHui Autonomous Prefectures in Gansu Province [10]; Wuzhou City in Guangxi Zhuang Autonomous Region [21]; Jiuzhaigou County of Aba Tibetan; and Qiang Autonomous Prefecture in Sichuan Province [22]. All of these areas have experienced different scale landslides at different times. Therefore, we acquired landslide images in different backgrounds to make the dataset coverage more compatible with some representative remote sensing images of landslides, as shown in Fig. 1. The open-source image labeling tool, LabelMe, was used to label the graphic objects of the dataset to create semantic segmentation labels. As a result, the landslide bodies appear white according to landslide detection characteristics and task requirements.

B. Features of Landslide Remote Sensing Images

After a landslide, the surface is commonly covered with bare soil and vegetation. Therefore, these regions have a low vegetation index, high spectral reflectance, and prominent brightness.
They are predominantly khaki and grey, which differ from the background color of the landslide. The landslide movement exhibits an obvious directional texture on the remote sensing images, which significantly differs from the texture of the surrounding landforms. Remote sensing images show that the landslide has a rough, corrugated texture perpendicular to the sliding direction; dappled shadows may appear on the landslide images due to the uneven surface. The deep learning model can complete the landslide detection task by learning the above features.

The landslides presented by remote sensing images lack specific shapes and smooth boundaries because of the characteristics of satellite overhead imaging. Therefore, landslide detection is more difficult than regular-shaped target detection. The different locations of landslides result in different background colors, which mainly include green in the background of the forest [Fig. 1(a)] and dark yellow in the background of the loess [Fig. 1(b)]. The background also includes the following landforms: urban built-up regions [Fig. 1(c)], water systems, including rivers and lakes [Fig. 1(d)], clouds [Fig. 1(e)], and regions with similar spectral characteristics to landslides, such as bare soil and fields [Fig. 1(f)]. In addition, landslide images have different brightness due to different light radiation [Fig. 1(g)], resulting in varying degrees of shadows. The above features reflect the complexity and uncertainty of landslide remote sensing images. The dataset created in this article contains different landslide scenes, which increases the difficulty of landslide detection, enabling the development of a detection method with stronger generalization performance.

III. METHODOLOGY

UNet++ is a U-shaped symmetrical mesh structure, as shown in Fig. 2, with an encoder on the left and a decoder on the right. The downward blue solid arrows shown in Fig. 2 indicate downsampling during encoding, which includes max pooling and average pooling. The upward green solid arrow indicates the upsampling of the decoding process, which includes transposed convolution and unpooling. The dotted arrow between the encoder and the decoder is a skip connection, which fuses multiple feature maps through a series of nested convolution blocks. Furthermore, the feature map extracted by the last convolution block is analyzed to obtain the segmentation result. The tightly connected structure can fuse multilevel features and fully parse contextual semantic information, improving semantic segmentation accuracy. However, this structure leads to a dramatic increase in the number of parameters and required computational resources. Zhou et al. [23] used the deep supervision-pruning method to solve this problem; that is, when comprehensive indicators, such as segmentation accuracy and speed, meet the task requirements, part of the network structure is pruned. Although this operation can reduce the number of parameters, it also affects accuracy. The multiscale feature maps generated by different convolution blocks in the dense structure have equal weights; however, the semantic information contained in different scale feature maps differs. Therefore, after further downsampling of remote sensing images, small regions and landslide details may disappear, especially in regions where the features of the landslide and the background are similar. Moreover, fusing feature maps with the same weight makes it difficult to distinguish landslide boundaries and small-area landslides. To solve the above problems and make the model more suitable, this article proposes a new model, Reg-SA–UNet++ (RS–UNet).

As shown in Fig. 3, compared with UNet++, RS–UNet improves the feature extraction module. In the encoding process, RegNet is used as a convolution block, which can reduce model parameters and improve feature extraction capabilities; maximum pooling is used to preserve the texture features of landslides. In the encoding and decoding processes, a shuffle attention module is added to the connection part of each layer to...
improve the model’s semantic parsing ability. To further improve its accuracy and reduce the number of parameters, the deep supervision-pruning operation of UNet++ is removed. RS–UNet inherits the fully connected structure of UNet++ to fuse multiple feature maps, which can enrich feature information. The final result is obtained through softmax classification after the feature map passes through the last convolutional block of the decoding process.

A. RegNet

As a binary classification problem, the single–temporal remote sensing images captured postlandslide contain fewer features. Moreover, remote sensing images often suffer from data imbalance, meaning that some classes in the image occupy larger regions than others. It is frequently observed in the landslide remote sensing images that the landslide is just a small fraction of the image, and the background takes up most of it, making it challenging to extract the landslide features. In the dataset constructed in this article, there are shadows and additional landforms, such as inhabited areas, water systems, and clouds around the landslide, which further interfere with the feature extraction. In the case of limited training samples, complex model structure, and several parameters, it is easy to generate excessive reliance on limited data samples even if a classification model is obtained. As a result, when the model is applied to a new dataset, its detection performance is hardly as good as expected. Moreover, complex models have complex intermediate computations, which consume considerable frame buffers. Therefore, the parameters need to be tuned for optimal performance of the model before training. However, it is challenging to make all parameters reach the optimal value when there are many of them, which may compromise the model’s generalization ability. Therefore, to reduce the number of parameters and improve the feature extraction ability of the model, we replace the convolution block in the UNet++ encoding process with a lightweight network, RegNet.

As shown in Fig. 4, the structure of RegNet comprises three main parts: stem, body, and head [Fig. 4(a)]. The stem is a convolutional layer whose convolutional kernel size is $3 \times 3$; the step distance is 2, with 32 convolutional kernels. The head is a classifier for outputting different classes. As shown in Fig. 4(b), the structure of the body consists of four stages; from stage 1 to stage 4, the resolution of feature maps is halved in turn; the number (i.e., $w_1, w_2, w_3,$ and $w_4$) of feature maps output at each stage is a hyperparameter determined by searching; each stage consists of several blocks (Fig. 5), and the hyperparameters and need to be determined by searching.

B. Shuffle Attention

In this article, shuffle attention modules (SANet) (Fig. 6) are added at the junction of each level of encoding and decoding structures; this improves the model’s ability to understand the semantic information of landslide remote sensing images, enhances the model’s attention to landslides, reduces the interference of other landforms, and guides the model to learn the
features of landslides better. Furthermore, by weighting the feature map, this module suppresses the information that negatively impacts landslide detection and retains the information that helps analyze feature information. As mixed-domain attention, in the spatial domain, SANet transforms the spatial information in the original picture into another space and preserves the key information; in the channel domain, it algorithmically compares channels and allocates resources to channels that require more attention.

SANet is a lightweight mixed-domain attention network. It splits the input data X into G groups along the channel dimension and processes each group of subfeatures in parallel using channel splitting. During the parallel processing, the channel domain branch uses GAP to generate channel statistics and then uses a pair of parameters to scale and shift the channel vector; the spatial domain branch uses the group norm to generate spatial statistics and then creates compact features similar to the channel branch. After the parallel processing is completed, concat fusion and intergroup communication are performed so that the output of SANet is the same size as the input, making SANet perfectly embedded in the model.

### C. Matthews Correlation Coefficient Loss Function

Landslide remote sensing images often suffer from data imbalance. Unless accounted for when training a deep-learning-based segmentation model, such an imbalance can lead to the model converging toward a local minimum of the loss function, yielding suboptimal segmentation results biased toward the background. Loss functions based on overlapping metrics are often the first choice for such problems. Currently, Dice loss (DIC) and IOU loss are frequently used in overlapping metric-based loss functions; neither penalizes misclassifications of true negative pixels, making it challenging to optimize for accurate background prediction. Therefore, a loss function based on the Matthews correlation coefficient (MCC) is proposed [24]. The MCC for a pair of binary classification predictions is defined as

\[
MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{1}
\]

where true positive (TP), true negative (TN), false positive (FP), and false negative (FN) are the elements in the confusion matrix, as shown in Fig. 7. The specific meaning of these elements in the remote sensing image landslide recognition is intuitive. MCC values range from $-1$ to $1$, where $-1$ implies completely irrelevant and $1$ implies complete correlation. Furthermore, the Matthews correlation coefficient loss (MCL) can be defined as follows:

\[
MCL = 1 - MCC. \tag{2}
\]

In a previous article [24], MCL was proved to achieve better results than the above two loss functions, so this article introduces MCL to improve remote sensing image landslide detection results.

### IV. Experiment

In the experimental part of this article, the landslide dataset mentioned in the “Dataset” chapter is divided into training, test, and validation sets in a ratio of 7:2:1; the experimental environment is PyTorch 1.7 framework, the operating system is Windows 10, the CPU is a 12-core Intel Xeon Gold 6226, the primary frequency is 2.70 GHz, the memory is 64 GB DDR4, the GPU is NVIDIA’s Quadro RTX 5000, the frame buffer is 2 × 16 GB, the CUDA version is 11.0, and the cuDNN version is 7.6.5.

During the training process, the input image size was set to $512 \times 512$; the optimizer was Adam, the initial learning rate was 0.001, the weight decay was 0.0005, the loss function was MCL, and the learning rate decay strategy was cosine annealing, which is defined as follows:

\[
\eta_i = \eta_{\text{min}}^i + \frac{1}{2} (\eta_{\text{max}}^i - \eta_{\text{min}}^i) \left( 1 + \cos \left( \frac{T_{\text{cur}}}{T_i} \pi \right) \right) \tag{3}
\]

where $i$ represents the first few runs of the program; $\eta_{\text{max}}^i$ represents the maximum value in the learning rate; $\eta_{\text{min}}^i$ represents the minimum value in the learning rate; $T_{\text{cur}}$ represents the number of cycles executed; and $T_i$ represents the total number of cycles in the $i$th run.

In this experiment, max_epoch was set to 200, and the validation set was used for evaluation after each epoch. The training was terminated if the evaluation index of 10 consecutive epochs did not improve. Notably, in addition to the targeted improvements adopted by some of the mentioned methods, these experimental settings have also been utilized.
Fig. 8. Attention heat map. The closer to red the color in a region in the figure, the more attention the model pays to this region.

To quantitatively illustrate the superior performance of our model, we evaluated the model by the overall accuracy (OA), mean intersection over union (MIOU), balanced score (F1-score), number of parameters (Params), floating-point operations (FLOPs), and frames per second (FPS). OA represents the ratio of the number of pixels with the correct predicted category to the total number of pixels. MIOU is the ratio of the intersection and union of the true value set and the predicted value set. F1 score is the harmonic mean of precision and recall, and it is frequently used as an evaluation metric for imbalanced sample classification. The value range of the above three indicators is between 0 and 1, and the larger the value, the better the result. Params refer to the sum of weights, bias terms, and other parameters processed during model training, and their unit is millions (M); the fewer the parameters, the lower the computational cost. The number of FLOPs is the number of single operations required to run the model, and the unit used here is giga multiply–accumulate operations per second (GMACS); the larger its value, the greater the required amount of computation. FPS represents the number of frames transmitted per second, which reflects the real-time nature of the model. If $p_{ij}$ represents the number of pixels whose true value is class $i$ and predicted value is class $j$, $N$ is the total number of classes, $k$, and $k$th is the class label. The calculation method of each evaluation standard can be described as follows:

$$\text{OA} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij}}$$

### TABLE II

| Loss  | MIOU % | F1-score % | OA % |
|-------|--------|------------|------|
| MCC   | 81.74  | 92.41      | 93.37|
| DIC   | 79.33  | 90.34      | 93.11|
| ASY   | 78.41  | 89.33      | 92.54|
| FTL   | 78.25  | 87.2       | 92.12|
| IOU   | 80.86  | 91.56      | 92.27|
| WCE   | 79.69  | 93.13      | 92.97|
| FOL   | 79.83  | 90.33      | 92.89|

To illustrate the superior performance of our model, we evaluated the model by the overall accuracy (OA), mean intersection over union (MIOU), balanced score (F1-score), number of parameters (Params), floating-point operations (FLOPs), and frames per second (FPS).
### Table III
Comparison of Results of Different Attention Mechanisms

| Model          | OA %  | Miou % | F1-score % | Parameters M | GMACS |
|----------------|-------|--------|------------|--------------|-------|
| Reg–UNet-SCSE  | 91.94 | 76.95  | **93.35**  | 5.76         | 22.49 |
| Reg–UNet-SE    | 92.44 | 80.93  | 91.56      | 5.76         | **22.44** |
| Reg–UNet-CBAM  | 92.42 | 78.55  | 91.44      | 5.76         | 22.58 |
| Reg–UNet-SKA   | 83.04 | 53.46  | 50.41      | 61.93        | 576.61 |
| RS–UNet (Ours) | **93.37** | **81.74** | 92.41      | **5.67**     | 22.47 |

Fig. 9. Results obtained via different loss functions: (a) is the landslide remote sensing image, (b) is the label of the landslide in (a), and (c–i) illustrates the result of the using MCC, DIC, ASY, FTL, IOU, WCE, and FOL, respectively.

\[ \text{MIOU} = \frac{1}{k+1} \sum_{i=0}^{k} \left( \sum_{j=0}^{k} p_{ij} + p_{ji} - p_{ii} \right) \]  
\[ F1^{(k)} = \frac{2P^{(k)} \cdot R^{(k)}}{R^{(k)} + R^{(k)}} \]

\[ = \frac{2 \left( p_{kk} / \sum_{j=1}^{N} p_{kj} \right) \cdot \left( p_{kk} / \sum_{i=1}^{N} p_{ik} \right)}{p_{kk} / \sum_{i=1}^{N} p_{ik} + p_{kk} / \sum_{i=1}^{N} p_{ik}} \]  

V. Results and Discussion

A. Ablation Experiments

This section presents a demonstration of the effectiveness of the improvements through ablation experiments. In deep learning, ablation experiments refer to the study of model performance after removing certain modules, which aids in understanding the impact of the removed modules on the entire model. Additionally, ablation experiments require the model to exhibit modest degradation (a model that retains functionality even if certain modules are lost or weakened). Table I lists the ablation experiment results.

To compare the models, we performed four groups of experiments as follows:

1) **Group A**: Using the UNet++ model.
2) **Group B**: Using a model that only changes convolutional blocks without increasing the attention module, referred to as Reg–UNet in Table I.
3) **Group C**: Using a model that only increases the attention module without changing the convolutional blocks, referred to as SA–UNet in Table I.
4) **Group D**: Using the RS–UNet model.
Fig. 11. Results obtained by different semantic segmentation models: (a) is the landslide remote sensing image, (b) is the label of the landslide in (a) and (c–j) correspond to the results using FPN, PSPNet, DeepLabv3+, Res2-UNet, SegFormer, UNet++, ResU-Net, and RS–UNet, respectively.

Table I shows that adding the attention module improves the evaluation index of I more than replacing the convolution block, indicating that the attention module plays a significant role in landslide detection.

To observe the effect of the attention module, we used the Grad-CAM [25] visualization method to conduct a qualitative analysis of the three models (UNet++, Reg–UNet, and RS–UNet). Grad-CAM traces the reverse gradient flow to discover which locations of the feature map have a more significant impact on the final output, as shown in Fig. 8. Although Reg–UNet improves the ability to guide feature learning compared with UNet++, its feature guiding force for small and complex boundary landslides remains weak; RS–UNet can learn the features of landslides more accurately. As shown in Fig. 8, the heat distribution of UNet++ is uneven, and other landforms, such as bare soil and sparse vegetation [Fig. 8(a)–(c)], were also highlighted. As the shape of the landslide becomes more complex, UNet++ pays less attention to it, as shown in Fig. 8. The heatmap of RS–UNet has a clear boundary between the landslide body and the background, as shown in Fig. 8(a)–(d). For the very complex-shaped landslide in Fig. 8(e), RS–UNet can focus most of the attention on the landslide body. Therefore, RS–UNet can better guide the model for feature learning.

B. Effect of Loss Function on the Result

To illustrate the superiority of the loss function chosen in this article, DIC [26], IOU loss (IOU) [27], weighted cross-entropy loss (WCE) [28], focal loss (FOL) [29], asymmetric loss (ASY) [30], and focal Tversky loss (FTL) [31] were selected for comparison experiments. DIC, IOU, and MCL belong to the loss functions based on similarity measures; ASY, FTL, WCE,
and FOL are frequently used for sample imbalanced semantic segmentation. One of the detection results in the test set is selected as an example shown in Fig. 9, indicating that the best visual results can be achieved when MCC is used in training. The result data of the whole test set is shown in Table II. Although the loss function has a minor impact on the model results, as the loss function is MCL, both OA and MIoU achieved the best values, which are 0.48% and 2.54% higher than the suboptimal value, respectively; the F1 score was the suboptimal value, which is 0.72% lower than the best value.

C. Contrast of Different Attention Mechanisms

As detailed in this section, four attention modules (SCSE [32], SE [33], CBAM [34], and SKA [35], [36]) were used for comparative experiments. One of the detection results in the test set is selected as an example shown in Fig. 10, indicating that when the attention module is SA, the result closest to the label image can be obtained, and when the attention module is SKA, the obtained result is considered invalid. The result data of the whole test set is shown in Table III. To distinguish models using different attention modules, they are called Reg–UNet-XXX, where XXX is an abbreviation for different attention modules. SCSE improves upon SE; however, the model using SE achieved better results, suggesting that the improvement of the attention module may not be suitable for landslide detection. Although the model using SKA had the most parameters and FLOPs, it produced the worst results; therefore, the quality of the model does not depend on the number of parameters and computation. Compared with the results of Reg–UNet in Fig. 8, the detection accuracy of the model changes to varying degrees after adding different attention modules, implying that the use of attention modules may not have a positive impact on landslide detection. The parameters and FLOPs of the model using SA did not change significantly; however, both OA and MIoU reached the optimal value, which was 0.93% and 0.81% higher than the suboptimal value, respectively; the F1 score reached the suboptimal value, which was 0.94% lower than the optimal value. In conclusion, the model with SA had the best performance.

D. Comparison With Other Methods

A landslide detection model based on remote sensing image, ResU-Net [15] and six classical semantic segmentation models (FPN [34], PSPNet [37], DeepLabv3+[38], UNet++ [23], Res2-UNet [39], and SegFormer [40]) were selected for comparative experiments. Some landslide detection results were selected and shown in Fig. 11. Fig. 11(a) and (b) is, respectively, landslides and their tag images in different scenes. As shown in Line I of Fig. 11, all methods can obtain accurate results when the landslide area is large and the background is forest. With the presence of water bodies and roads in the background (as shown in Line II of Fig. 11), the results of FPN, PSPNet, SegFormer, and UNet++ [Fig. 11(c), (d), (g), and (h)] obviously missed a part of the landslide area; the ResU-Net’s result [Fig. 11(i)] mistook the road as landslide; the results of DeepLabv3+, Res2–UNet, and RS-UNet [Fig. 11(e), (f) and (j)] almost sketched out the right area of the landslide. For the landslide with a small area (as shown in Line III of Fig. 11), a large number of areas are wrongly identified as landslide areas by FPN and Res2–UNet [Fig. 11(c) and (f)]; PSPNet, SegFormer, and UNet++ only determined a small part of the landslide area [Fig. 11(d), (g), and (h)]; the results of DeepLabv3+ and ResU-Net [Fig. 11(e) and (i)] are acceptable for they identified the landslide area with a small part of spare forest; and RS-UNet obtained the most satisfying result [Fig. 11(j)] among all the methods since it draws the right outline.
of the landslide with a quite small mistake. When there are a large amount of trees in the landslide area (as shown in Line IV of Fig. 11), DeepLabv3+ and SegFormer can depicted more accurate landslide areas [Fig. 11(e) and (g)] than other methods. It can be inferred from the above analysis that the detection results of RS–UNet were better than those of other models.

The results of evaluation metrics for the test set are listed in Table IV. RS–UNet showed enhancement for all metrics, namely, 20.42-M fewer parameters, 51.56 fewer GMACs, 3.94% more OA, 10.05% more MIOU, and 10.07% higher F1 score than those of UNet++. The OA, MIOU, and F1 scores of RS–UNet were 2.3%, 2.38%, and 2.59% higher than those of ResU-Net, respectively; moreover, parameters and FLOPs dropped by 18.77 M and 9.24 GMACS, respectively. Compared with the metrics of other models in Table IV, the number of parameters of RS–UNet decreased considerably. Compared with the suboptimal values of various indicators, the OA, MIOU, and F1 scores of RS–UNet increased by 1.03%, 1.60%, and 1.45%, respectively; additionally, its parameters reduced by 10.75 M, and FLOPs reached an acceptably low level. Although Res2-UNet achieved landslide detection results close to RS–UNet, the Params and FLOPs of Res2-UNet greatly exceeded those of RS–UNet with 27.01 M and 21.43 GMACS, respectively. Among all models, the FLOPs of PSPNet are only 9.59 GMACS. It is speculated that models with fewer FLOPs cannot meet the computational requirements of feature maps. In contrast, UNet++ had the highest number of
parameters and substantial computation; however, its indicators could not reach suboptimal values, demonstrating that more parameters and calculations do not determine the efficiency of the model. In all the aforementioned methods, the FPS value of PSPNet reached the highest (158.38), but it has quite low landslide detection accuracy. Res2-UNet has the smallest FPS value of 21.19, while RS–UNet has a higher FPS value of 2.43, which makes the whole reasoning process short. Since there is no hard requirement for real-time landslide detection after disaster, the reasoning speed of RS–UNet is acceptable under the condition of high landslide detection accuracy.

To analyze the impact of image size on landslide detection results, we changed all the images of the test dataset to 0.1× (represent 10% of the original image scale), 0.2×, 0.3×, 0.4×, 0.5×, 0.6×, 0.7×, 0.8×, and 0.9× of its original size, and tested the performance of the landslide detection model at these nine scales. We selected Res2–UNet, SegFormer, and ResU-Net with better performance in Table IV for comparative experiments. Their quantitative indicators are calculated and plotted as a broken line chart (Fig. 12).

As shown in Fig. 12, almost all models maintain good performances when the image scale was changed from 1.0× to 0.5×, except that the F1-score and MIoU of the SegFormer decrease as the image scale decreases. Furthermore, the OA, MIoU, and F1-score values of RS–UNet were the highest within a certain range of image scale. With the development of remote sensing technology, the resolution of remote sensing images is improving. Under the condition of higher resolution, the image scale will not excessively interfere with the performance of models.

E. Application in Landslide Mapping

To verify the practicability of RS–UNet in landslide mapping, the landslides caused by earthquakes in Hokkaido, Japan, were selected [41] for landslide mapping tests. The basic situation of the landslide is shown in Fig. 13; landslides occurred mainly in vegetation-covered mountains; in addition, farmland and residential areas were present in the background. The image of the landslide region was obtained by Bigemap software. However, the image size is large, making it difficult to directly input it into the network model. Therefore, it was first cropped into multiple subimages with a size of 512 × 512 pixels and then inputted into the network model for training and testing; subsequently, the test results from the subsections were stitched together to obtain the detection results for the entire area. Furthermore, 20% of the cropped images were randomly selected as the training set, and the mapping test was performed on the whole image.

As shown in Fig. 14(a) and (b), both the methods completed large-scale landslide mapping with high accuracy, but different degrees of misjudgments were noted in the results; these errors mainly occurred in two types of areas: 1) the boundary of the landslide; and 2) bare soil, fields, and low vegetation in the background. The incorrect judgment of the boundary may be because of the error between the manual annotation and the actual landslide area; the same phenomenon can also cause some landslides to be unsuccessfully detected; these errors are minor and have little effect on the detection results. Misjudgments in the background are caused by landforms resembling landslide features (fields and bare soil resembling a landslide mass; low vegetation resembling a landslide-destroying vegetated area). The missing areas in the background appear mostly in the vegetation areas damaged by landslides; it is difficult for the model to determine whether these areas belong to the landslides because of the different degrees of damage. As a result, fewer misjudged and missed areas can be observed in the RS–UNet results, as shown in Fig. 14(c) and (d); therefore, RS–UNet is more effective than ResU-Net in landslide mapping.

The OA, MIoU, and F1 scores of the two methods are shown in Table V. Both methods achieved high detection rates (up to 95% or more) in the landslide mapping with a relatively simple background. Compared with ResU-Net, the OA, MIoU, and F1 scores of RS–UNet’s landslide detection results were improved by 0.33%, 1.25%, and 1.86%, respectively.

VI. Conclusion

In this article, we developed a remote sensing image landslide detection model, RS–UNet, based on the attention mechanism and the lightweight network (RegNet). The model was tested on the constructed dataset with different loss functions and attention modules to observe their impact on model performance. The conclusions of this article are outlined as follows:

1) RegNet was used to replace the convolution block in the UNet++ encoding process, and an improved landslide recognition performance was achieved.
2) The addition of the attention module enabled the model to better guide the feature learning process.
3) MCL utilized all the elements of the confusion matrix to achieve the best landslide detection performance.

RS–UNet effectively accomplished landslide detection using the remote sensing images and showed superior performance compared with the existing models. However, the small amount of landslide data used in this article and the human errors in regional labeling affected the detection accuracy of the model. In future articles, we aim to establish a larger dataset to further improve the generalization ability and detection accuracy of the model. At the same time, the occlusion of clouds and fog is also a significant factor that interferes with landslide detection. Thus, in future article, one of the problems to be solved is how to accurately detect landslides under the interference of clouds and fog.

REFERENCES

[1] H. Zhang, “Landslide deformation monitoring and analysis with GB-SAR,” Master’s thesis, Dept. Remote Sens., Geospatial Inf. Eng., Southwest Jiao tong Univ., Chengdu, China, 2015.
[2] C. Niu, H. Zhang, W. Liu, R. Li, and T. Hu, “Using a fully polarimetric SAR to detect landslide in complex surroundings: Case study of 2015 Shenzhen landslides,” ISPRS J. Photogramm. Remote Sens., vol. 174, pp. 56–67, 2021.
[3] Y. Wang, “Research on landslide information extraction algorithm of remote sensing image based on deep learning,” Master’s thesis, Dept. Electron. Eng., College, Heilongjiang Univ., Harbin, China, 2020.
[4] S. Zhang, “The method of landslide extraction with high resolution remote sensing image combining change detection and object oriented method,” Master’s thesis, Fac. Geosci. Environ. Eng., Southwest Jiao tong Univ., Chengdu, China, 2014.
[5] T. R. Martha, N. Kerle, V. Jetten, C. J. van Westen, and K. V. Kumar, “Characterising spectral, spatial and morphometric properties of landslides for semi-automatic detection using object-oriented methods,” Geomorphology, vol. 116, no. 1–2, pp. 24–36, 2010.

[6] Y. Yang, Y. Li, G. Feng, and L. Zhang, “A method aimed at automatic landslide extraction based on background values of satellite imagery,” Int. J. Remote Sens., vol. 35, no. 6, pp. 2247–2266, 2014.

[7] X. Duan, F. Han, and Y. Du, “Automatic identification technology of landslide based on UAV image,” J. Lanzhou Jiaotong Univ., vol. 39, no. 6, pp. 1–4, 2020.

[8] S. Cheng, Q. Kang, Z. Shen, and R. Zhou, “Landslide detection based on color feature model and SVM in remote sensing imagery,” Spacecraft Recovery Remote Sens., vol. 40, no. 6, pp. 89–99, 2019.

[9] G. E. Hinton, S. Osindero, and Y. Teh, “A fast learning algorithm for deep belief nets,” Neural Comput., vol. 18, no. 7, pp. 1527–1554, 2006.

[10] Y. Ju, Q. Xu, S. Jin, W. Li, X. Dong, and Q. Guo, “Automatic object detection of loess landslide based on deep learning,” Geomatics Inf. Sci. Wuhan Univ., vol. 45, no. 11, pp. 1747–1750, 2020.

[11] L. Cheng, J. Li, P. Duan, and M. Wang, “A small attentional YOLO model for landslide detection from satellite remote sensing images,” Landslides, vol. 18, no. 8, pp. 2751–2765, 2021.

[12] T. Wang, M. Liu, H. Zhang, X. Jiang, Y. Huang, and X. Jiang, “Landslide detection based on improved YOLOv3 and satellite images,” in Proc. 4th Int. Conf. Pattern Recognit. Artif. Intell., 2021, pp. 567–571.

[13] P. Liu, Y. Wei, Q. Wang, Y. Chen, and J. Xie, “Research on post-earthquake landslide extraction algorithm based on improved U-net model,” Remote Sens., vol. 12, no. 5, 2020, Art. no. 894.

[14] Y. Wang, X. Wang, and J. Jian, “Remote landslide recognition based on convolutional neural network,” Math. Prob. Eng., vol. 2019, pp. 1–12, 2019.

[15] W. Qi, M. Wei, W. Yang, C. Xu, and C. Ma, “Automatic mapping of landslides by the Resu-Net,” Remote Sens., vol. 12, no. 15, 2020, Art. no. 2487.

[16] T. Bui, P. Lee, K. Lum, C. Loh, and K. Tan, “Deep learning for landslide recognition in satellite architecture,” IEEE Access, vol. 8, pp. 143663–143678, 2020.

[17] X. Zhang, N. Pan, and M. Lian, “Semi-supervised multi-temporal deep representation fusion network for landslide mapping from aerial orthophotos,” Remote Sens., vol. 13, no. 4, 2021, Art. no. 548.

[18] I. Radosavac, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, “Designing network design spaces,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 10425–10433.

[19] Q. Zhang and Y. Yang, “SA-Net: Shuffle attention for deep convolutional neural networks,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2021, pp. 2235–2239.

[20] C. Zeng, Z. Cao, F. Su, Z. Zeng, and C. Yu, “High-Precision aerial imagery and interpretation dataset of landslide and debris flow disaster in Sichuan and surrounding areas,” China Sci. Data, vol. 7, no. 2, pp. 195–205, 2020.

[21] B. Li, R. Tang, Q. Zhang, Y. Xie, and Y. Zhang, “Extraction of deep structure and salient features of convolutional neural network from landslide,” Sci. Serv. Mapping, vol. 47, no. 1, pp. 154–164, 2022.

[22] Q. Li, W. Zhang, and Y. Yi, “An information extraction method of earthquake-induced landslide: A case study of the Jiuzhaigou earthquake in 2017,” J. Univ. Chin. Acad. Sci., vol. 37, no. 1, pp. 93–102, 2020.

[23] Z. Zhou, M. R. Siddiquie, N. Tujakhkash, and J. Liang, “U-Net++: Redesigning skip connections to exploit multiscale features in image segmentation,” IEEE Trans. Med. Imag., vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[24] K. Abhishek and G. Hamameh, “Matthews correlation coefficient loss for deep convolutional networks: Application to skin lesion segmentation,” in Proc. IEEE 18th Int. Symp. Biomed. Imag., 2021, pp. 225–229.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 618–626.

[26] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in Proc. 4th Int. Conf. 3D Vis., 2016, pp. 565–571.

[27] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, “UnitBox: An advanced object detection network,” in Proc. 24th ACM Int. Conf. Multimedia, 2016, pp. 516–520.

[28] X. Ren, Z. Xing, X. Xia, D. Lo, X. Wang, and J. Grundy, “Neural Network-based detection of self-admitted technical debt: From performance to explainability,” ACM Trans. Softw. Eng. Methodol., vol. 28, no. 3, pp. 1–45, 2019.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2999–3007.

[30] T. Rishik et al., “Asymmetric loss for multi-label classification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 82–91.

[31] N. Abraham and N. M. Khan, “A novel focal tversky loss function with explainability,” in Proc. 2nd Int. Conf. Workshop Neural Inf. Process. Syst., 2021, pp. 10425–10433.

[32] A. G. Roy, N. Navab, and C. Wachinger, “Concurrent spatial and channel ‘Squeeze & excitation’ in fully convolutional networks,” in Proc. Int. Conf. Med. Image Comput. Assist. Intervention, 2018, pp. 421–429.

[33] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[34] S. W. Woo, J. Park, J. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 3–19.

[35] X. Li, W. Wang, X. Hu, and J. Yang, “Selective kernel networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 510–519.

[36] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 936–944.

[37] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6230–6239.

[38] L. Chen, Y. Zhu, G. Papandreou, S. Schorff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 3–19.

[39] F. Chen, N. Wang, B. Yu, and L. Wang, “Res2-unet, a new deep architecture for building detection from high spatial resolution images,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 15, pp. 1494–1501, Jan. 2022.

[40] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “SegFormer simple and efficient design for semantic segmentation with transformers,” in Proc. Conf. Workshop Neural Inf. Process. Syst., 2021, pp. 12077–12090.

[41] P. Zhang, C. Xu, S. Ma, X. Shao, Y. Tian, and B. Wen, “Automatic extraction of seismic landslides in large areas with complex environments based on deep learning: An example of the 2018 Ibari earthquake, Japan,” Remote Sens., vol. 12, no. 23, 2020, Art. no. 3992.
Wanjie Lu received the B.S. degree in photogrammetry and remote sensing and the Ph.D. degree in surveying and mapping from the Information Engineering University, Zhengzhou, China, in 2016 and 2020, respectively. He is currently a Lecturer with the Data and Target Engineering Institute, Information Engineering University, Zhengzhou. His research interests include UAV remote sensing image processing, deep learning algorithm, and spatial information service.

Wei Liu received the B.S., M.S., and Ph.D. degrees in information and communication engineering from the Information Engineering University, Zhengzhou, China, in 2001, 2003, and 2016, respectively. He is currently an Associate Professor with the Information Engineering University, Zhengzhou. His research interests include pattern recognition, remote sensing information processing, and deep learning.

Tao Lai received the B.S. degree in electronic engineering and the M.S. and Ph.D. degrees in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2002, 2004, and 2010, respectively. Since 2019, he has been with the Sun Yat-Sen University, Guangzhou, China, where he is currently an Associate Professor with the School of Electronics and Communication Engineering. His research interests include radar design, and signal and data processing algorithms for GB-SAR and UAVSAR.