Landslides
DOI 10.1007/s10346-021-01694-6
Received: 11 December 2020
Accepted: 7 May 2021

Libo Cheng · Jia Li · Ping Duan · Mingguo Wang

A small attentional YOLO model for landslide detection from satellite remote sensing images

Abstract The use of high-spatial-resolution remote sensing image technology on mobile and embedded equipment is an important and effective way for emergency rescue and evaluation decision-makers to quickly and accurately detect landslide areas. Deep learning-based landslide detection models include one-stage and two-stage models. The two-stage landslide detection models are slower. The one-stage landslide detection models are faster but less accurate. Both types of detection models have many parameters. This research aims to improve the speed, accuracy, and parameters of landslide detection models. A you only look once-small attention (YOLO-SA) landslide detection model is proposed. YOLO-SA is an improved version of the one-stage detection model YOLOv4. First, the group convolution (Gconv) and ghost bottleneck (G-bneck) residual modules are used to replace the convolution components and residual module consisting of standard convolution. The purpose is to reduce the parameters of the model. Then, on this basis, an attention mechanism is added to improve the detection accuracy of the model. Finally, the position of the attention mechanism is adjusted to determine the framework of YOLO-SA. Qiaojia and Ludian counties in Yunnan Province, China, are used as the study area to acquire three-channel (red, green, blue) historical landslide optical remote sensing images from Google Earth, with a total of 1818 images, for training the model. YOLO-SA is compared with 11 advanced models, including Faster-RCNN, 3 types of EfficientDet, 2 types of Centernet, SSD-efficient, and 4 types of YOLOv4 models. The results show that the number of YOLO-SA parameters is reduced to 1.472 mb compared to EfficientDet-D0; the accuracy is improved to 94.08% compared to Centernet-hourglass; and the speed is up to 42 f/s. In addition, the effectiveness of the YOLO-SA model for potential landslide detection is verified, with an F1 score of 90.65%.

Keywords Landslide detection · Optical remote sensing images · Convolution neural network · Group convolution · Ghost bottleneck · Attention mechanism

Introduction

Landslides are common and extremely harmful geological disasters that pose a serious threat to global human safety and result in damage to property economies and the surface environment. After a landslide, it is very important to quickly obtain the location information of the landslide by using satellite images for rescue and recovery work (Messeri et al. 2015). It is also helpful to build a landslide database, draw landslide susceptibility maps, and coordinate the harmonious development of humans and nature. Therefore, it is very important to quickly and accurately detect the landslide area.

There are three main types of common landslide detection methods. The first type is the field survey method. This method is inefficient, labor intensive, and costly (Galli et al. 2008). Second is the traditional remote sensing image processing method. The main methods are statistical methods (Yang et al. 2019) and machine learning methods (Di Napoli et al. 2020). For example, Cheng et al. (2013) used K-means to vectorize the descriptors computed by SIFT to obtain the vocabulary of each image, which was then represented by the bag-of-visual-words (BoVW), and in probabilistic late semantic analysis (PLSA), k-NN was applied to distinguish landslide and nonlandslide images. Amatya et al. (2019) used K-means to calculate the NDVI threshold and well distinguished landslide areas from vegetated areas. This type of method requires prior extraction of the features or interpretability factors present in the image and then uses the classifier to perform the classification calculation. This creates complexity in the design of the algorithm and limits the scalability of the algorithm, and the algorithm does not have the performance necessary for real-time applications. The third type is end-to-end deep learning methods. Convolutional neural networks (CNNs), the main component of deep learning, have multilayer nonlinear mapping ability and can better fit (learn) landslide characteristics. In CNNs, the input landslide data are transformed into smaller data through multilayer mapping relationships to extract the important features from the landslide data. CNNs achieve end-to-end detection, and the generalization ability of features is much higher than that of manual features. Compared with the former two type methods, the landslide detection method using deep learning has higher detection speed, fewer human interference factors, and lower cost. As a result, the use of deep learning for landslide detection has become a mainstream method.

At present, the application of deep learning to remote sensing image landslide detection has mainly focused on one-stage and two-stage models. In the two-stage model, first, many candidate areas containing landslide features are extracted from the original image and then input into the CNNs to predict the location and category information of the object. Examples include Faster R-CNN (Ren et al. 2017) and Mask R-CNN (He et al. 2017). Maxwell et al. (2020) used a positive ratio of 33% and negative ratio of 67% as candidate regions in a Mask R-CNN model and then input the candidate regions into CNNs to extract target regions, which resulted in a slower model because image features were processed twice. Hong et al. (2019) improved Faster R-CNN and used a better performance processor to perform the inference calculation, and although the accuracy was higher, the inference time was slower at 0.079 s, (13 f/s—i.e., the number of images processed in 1 s). The two-stage model does not have the ability to detect in real time (minimum standard of 30 f/s). In the one-stage detection model, the real position of the landslide is determined by a regression algorithm with the aid of a prior box of the landslide object, and the whole process involves extracting the landslide only once, which truly realizes end-to-end detection. Examples include the RetinaNet series, SSD series, and YOLO series (Fu et al. 2017; Lin et al. 2017; Liu et al. 2016; Redmon et al. 2015; Redmon and Farhadi 2017; Redmon and Farhadi 2018). In 2015, the

Published online: 22 May 2021

first one-stage detection model YOLOv1 was proposed. SSD improved the localization problem of YOLOv1, but there were many drawbacks, such as the need to manually set the prior box size and the low detection accuracy for small targets. YOLOv2 adopted the idea of cross-scale feature fusion from SSD and automatically calculated the prior box, but there were many problems such as category imbalance. RetinaNet mainly solved the problem of category imbalance but has poor feature extraction ability. YOLOv3-v4 improved this problem. For example, YOLOv4 used a backbone with stronger feature extraction ability; combined spatial pyramid pooling (SPP) and a path aggregation network (PAN) to better fuse features, enhance the perceptual field, and extract features; and used DropBlock to increase the robustness of the model and improve the ability of multisize target detection. One-stage model has strong scalability, reconfigurability, and detection speed and has been applied in remote sensing images. Du et al. (2021) modified YOLOv4 to detect an infrared small target and achieved a high detection speed of 40.74 f/s and an accuracy of 91.92%. Xu et al. (2020) used an optimized SSD to detect landslide hazards and obtained faster detection speeds on an Nvidia GTX 1080. However, the regression algorithm causes the one-stage models such as YOLOv4 to exhibit severe missed detections (Ma et al. 2020). One-stage and two-stage detection models have the problem of many parameters (Bochkovskiy et al. 2020; Ren et al. 2017). Therefore, maintaining a high detection speed without reducing the accuracy of the model, while reducing parameters to meet the needs of smaller storage devices, is a problem to be solved.

The attention mechanism aims to address the degree of association of the image space context and improve the model's attention to important information (Hu et al. 2018). Ji et al. (2020) proposed a 3D attention mechanism to distinguish landslide and nonlandslide images. Cheng et al. (2021) added an attention mechanism to a meta learning task for image classification and achieved better results.

In summary, this paper reconstructed the model framework based on YOLOv4 and designed a new landslide detection model YOLO-SA for the detection of landslide areas on three-channel (RGB) optical remote sensing images. We also compared the performance of 11 advanced detection models with that of the YOLO-SA model for landslide detection from remote sensing images. Finally, the applicability of the model is validated regarding the areas of potential landslides. The following two aspects are specifically addressed in the reconstruction of the model:

(1) Reduce the model parameters. Gconv, with fewer parameters, is used to perform convolutional operations. By using the Ghost module with fewer parameters to build the residual module G-bneck, G-bneck can also increase the number of CNN layers and improve the information fitting ability of the model.
(2) Improve the model accuracy. An attention mechanism model designed with the human eye vision system is used to improve the CNN's focus on the landslide feature and reduce the background noise.>

The rest of this paper is organized as follows: the "Study area and data" section introduces the landslide dataset needed to train the model. The "Method" section introduces the reconstruction of the model framework to reduce parameters, the use of the attention mechanism to improve the accuracy of the model, the framework of YOLO-SA and the experimental parameter settings. The "Results" section describes the experimental results. The "Discussion" section discusses our model. Finally, the "Conclusion" section summarizes the full text.

Study area and data

Study area

The study area is one of the most serious landslide areas in China. It is located in Qiaojia and Ludian counties in northeast Yunnan Province, China. It is between 102°53′–103°40′E and 26°32′–27°32′N, with a total area of 4683 km^2. The Niulan River is the natural dividing line between the two counties, with Qiaojia County located on the west side and Ludian County on the east side. The elevation of the study area is 527–4016 m, which is a typical subalpine deep-cut landform and belongs to the subtropical and temperate coexistence of plateau three-dimensional climate type. The annual precipitation is 674.8–1092.9 mm. The mountainous area accounts for 91.85%, with overlapping mountain ranges, high mountains and deep valleys, and staggered rivers and ditches.

The terrain of the study area is low in the middle, high on both sides, and low at the edges. It belongs to the combination of subalpine deep-cut landform type, with a variety of geomorphic types. The lithology is mainly limestone, shales, sandstones, and basalts. The study area is located in the uplift zone of the Qinghai-Tibet Plateau, east of the Xiaojiang Fault Zone. The central part of the study area is located in the compound part of the Yaoshan and Lomakou tectonic belts, where frequent neotectonic movements change the inertial forces on the slopes, causing surface deformation and increasing fissures, which promote the formation and sliding of landslides. In the western part, the crust has obvious horizontal torsion and vertical oscillation, with east-west extrusion forming fractures and backslopes, north-south tension forming a large number of larger fissures, and more loose material on the surface, which are potential factors for landslide formation (Gu et al. 2021; Li et al. 2016; Wang et al. 2020). Human activities also exacerbate landslide formation, such as logging, highway construction, mining, construction of hydroelectric projects, and farming.

Building landslide sample sets

In constructing the landslide dataset, the geographical locations of the landslides are determined by optical remote sensing images and data from the geological disaster department. According to the locations of landslides, the satellite image areas of 1818 landslides between 2013 and 2016 are extracted from Google Earth. The landslide images are three-channel (RGB) data. The spatial resolution of the ground in the images is 0.14 m. Each landslide area contains a variety of background information. In Fig. 1, the black triangle represents the geographical location of the landslide area.

The landslide samples are transformed and enhanced. First, the landslide samples are scaled to a 640 × 640 pixel-sized square area. Two basic data enhancement methods are used: geometric transformation and color transformation. (1) Geometric transformation involves random horizontal and vertical panning, rotation, scaling, cropping, up-down and left-right flipping. (2) Color transformation involves random change the hue, saturation, and brightness.

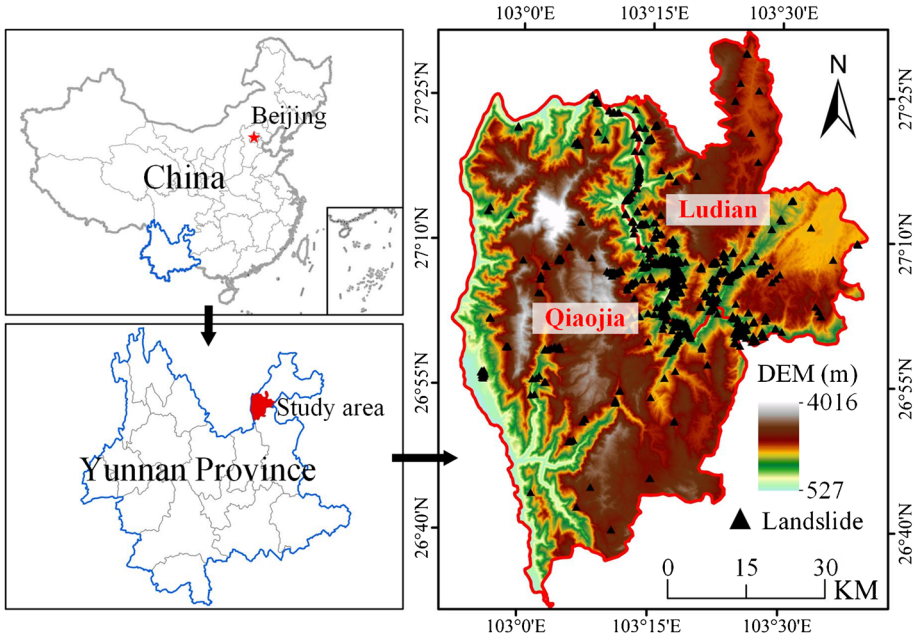

Fig. 1 Study area

Because of the different satellite sensors used and the diversity of the acquisition environment, the collected remote sensing images are different. Therefore, color transformation is needed to eliminate the effects of color deviation on the CNN performance. Moreover, geometric transformation allows the CNNs to look at the landslide area from different perspectives, which improves the robustness of the model.

Finally, the areas of landslides in the sample are marked in PASCAL VOC format using the LabelImg software. Figure 2 shows some sample sets with marked landslide areas. The coordinates of the upper left and lower right corners of the rectangle box indicate the true location of the landslide area.

Method

The main objective of this study is to improve the accuracy, speed, and parameters of the detection model so that it can be applied to the efficient detection of landslides from remote sensing images. The solution includes the following two aspects: (1) Reconstruction of the network structure of YOLOv4 model (Gconv is used to perform the convolution operation, and the G-bneck residual module is used to solve the degradation problem of the model; both Gconv and G-bneck greatly reduce the parameters of the model) and (2) based on the reconstructed model, the attention mechanism is used to improve the detection accuracy of the model. The flowchart of the methodology in this paper is shown in Fig. 3, where the red dotted lines indicate the reconstruction and accuracy optimization of the YOLOv4 model structure.

Gconv

The computational process of Gconv is as follows. First, Gconv groups the input feature maps, and then performs the convolution operation for each group. As shown in Fig. 4, the input feature map $F \in R^{CxHxW}$ (C is the number of channels, and H and W are the height and width, respectively) is divided into two groups, each of which is alone operated by the convolutional kernel $f \in R^{C/_2 \times K \times K}$ ($K \times K$ is the size of the convolution kernel), and the feature map $M(F) \in R^{N/_2 \times H' \times W'}$ of $N/_2$ channels is output (H' and W' are the height and width, respectively). When the number of groups is G ($G \leq C$), the number of input feature maps of each group is $C/_G$, the number of output feature maps of each group is $N/_G$, the size of each convolutional kernel is $\frac{C}{G} \times K \times K$, the total number of convolutional kernels is N, and the number of convolutional kernels of each group is $N/_G$. When $G = C$, $N = C$—i.e., $G = N = C$—the size of each of the N convolutional kernels is $1 \times K \times K$. Gconv is now evolved into DWconv (Howard et al. 2017).

After a convolution operation, the total parameter of Gconv is $N \times \frac{C}{G} \times K \times K$. In the standard convolution operation, the number of channels of the convolution kernel is the same as the number of channels of the input map, the number of convolution kernels is the same as the number of channels of the output map, and the total number of parameters is $N \times C \times K \times K$. Compared with the standard convolution, Gconv has G-fold fewer parameters.

G-bneck residual module

The ghost module points out that there is redundancy in the output feature map when the input feature map is computed by standard convolution. Figure 5 shows the feature maps after the first residual block processing in ResNet-50. There are three similar feature map pairs (He et al. 2016). One feature map in each feature map pair can be used to obtain another feature map by inexpensive operations, and there is a "ghost" relationship between them.

The calculation process of the ghost module is as follows. As shown in Fig. 6, the input feature map $Y' \in R^{H \times W \times m}$ has m

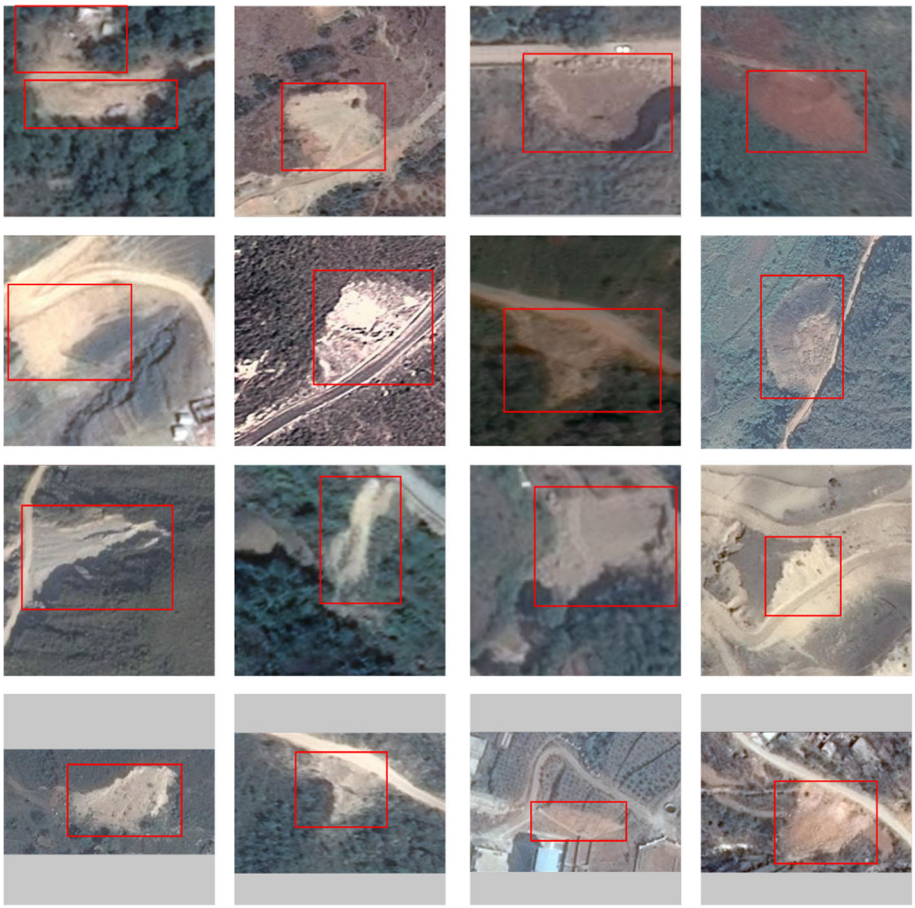

Fig. 2 Examples of marked landslide areas

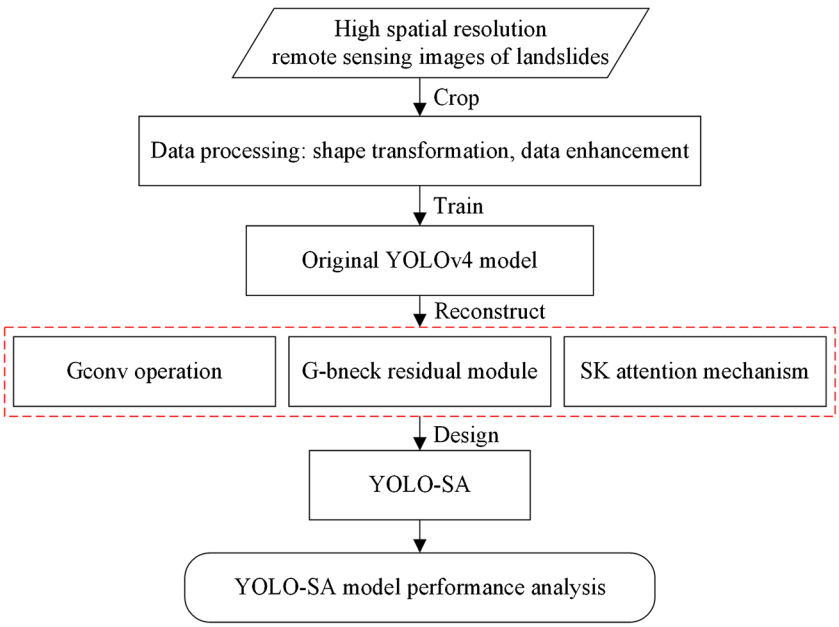

Fig. 3 Model flowchart

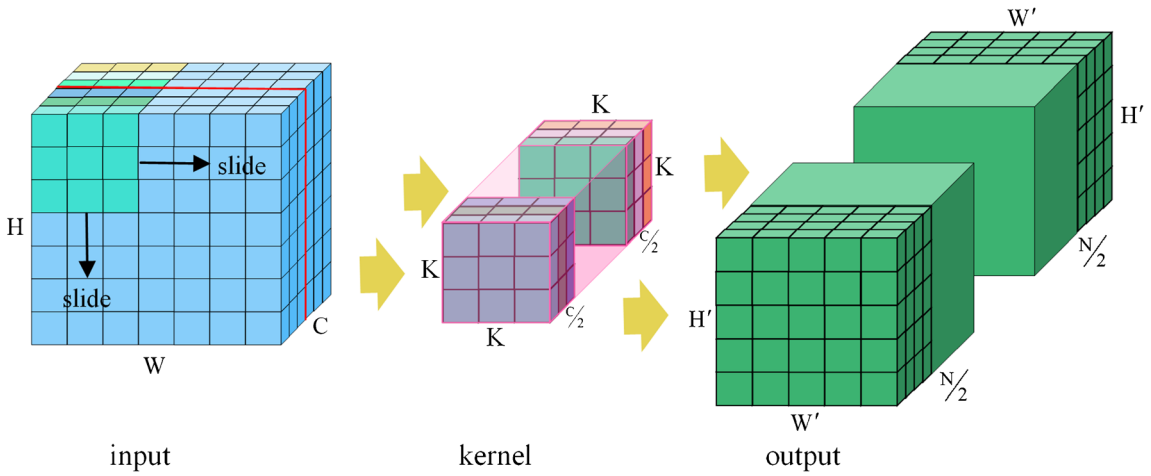

Fig. 4 Gconv

feature maps. To obtain the required n feature maps Y ($m \leq n$), for each input feature map in $Y^{'}$, a series of cheap linear operations are used to generate s "phantom" feature maps. The transformation formula is as follows:

$$y_{ij} = \Phi_{i,j}\left(y_i^{'}\right), \forall i = 1, \ldots, m, j = 1, \ldots, s \quad (1)$$

where $y_i^{'}$ is the i-th input feature map in$Y^{'}$, $\Phi_{i,\,j}$ is the j-th linear operation, and $\Phi_{i,\,j}$ is used to generate the j-th "ghost" feature map y_{ij}. In other words, $y_i^{'}$ can have one or more "ghost" feature maps $\left\{y_{ij}\right\}_{j=1}^{s}$. On this basis, an identity mapping $\Phi_{i,\,s}$ is added. Finally, the input feature map is superimposed with the "ghost" feature map. $n = m \times s$ feature map $Y = [y_{11}, y_{12}, \ldots, y_{ms}]$ is obtained by inexpensive operation.

The parameter evaluation is as follows. The ghost module contains an identity mapping and $m \times (s{-}1) = {}^{n}/_{s} \times (s{-}1)$ linear operations, and the average kernel size of each linear operation is $d \times d$. The final number of parameters for the ghost module is ${}^{n}/_{s} \times H^{'} \times W^{'} \times C \times K \times K + (S{-}1) \times {}^{n}/_{s} \times H^{'} \times W^{'} \times d \times d$. Compared with the standard convolution operation, the ghost module includes s-fold fewer parameters. The parameters of the $d \times d$ kernels are ${}^{n}/_{s} \times C \times K \times K + (S{-}1) \times {}^{n}/_{s} \times d \times d$. Compared with that of the standard convolution operation, the number of

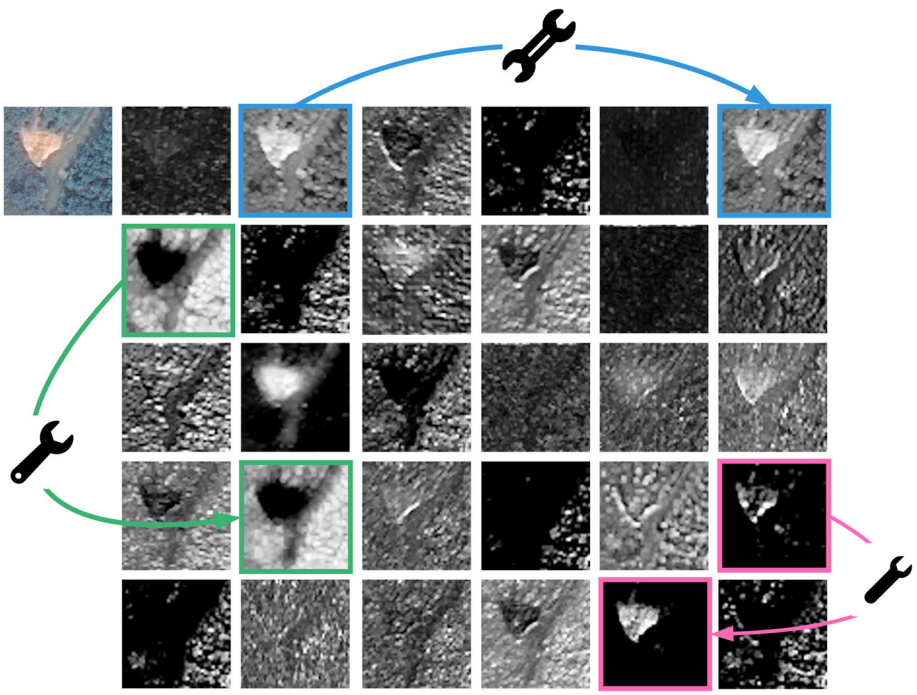

Fig. 5 Feature redundancy

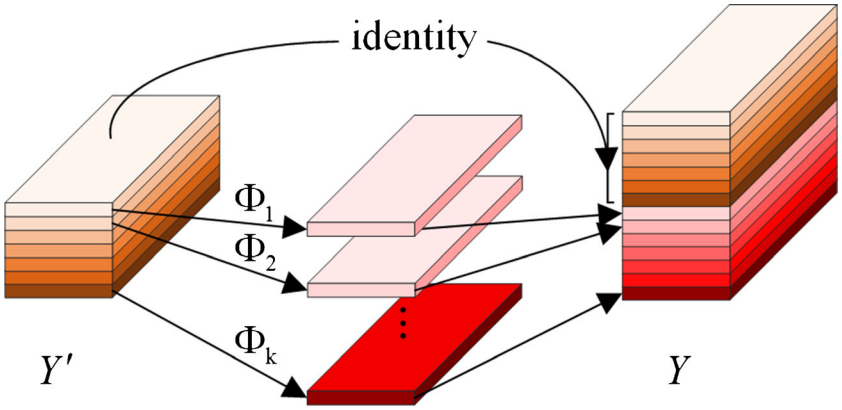

Fig. 6 The ghost module

parameters is $^1/_s$-fold lower. Finally, the G-bneck residual module with moving steps of 1 and 2 for the convolution kernel is designed by taking advantage of Ghost Conv, and the G-bneck is shown in Fig. 7. In the figure, BN is the batch normalization, and ReLU is the activation function.

Attention mechanisms

Gconv can cause a decrease in the accuracy of the model. To compensate for this loss, the selective kernel (SK) attention mechanism is added to the model (Li et al. 2019). The SK attention mechanism helps increase the proportion of useful information in the model and reduce the background noise.

The SK attention mechanism enables the model to adaptively adjust the receptive field size according to the multiple scale features of the input information and thus extract useful information from different levels. The SK module is composed of the split, fuse, and select operations.

(1) In the split operation, a multisize convolution operation is performed on the input feature map. The input feature map $F \in R^{CxHxW}$ is convolved by n convolution kernels of different sizes to obtain n feature maps (U_1, U_2, ···U_n. $U_n \in R^{C\times H\times W}$).

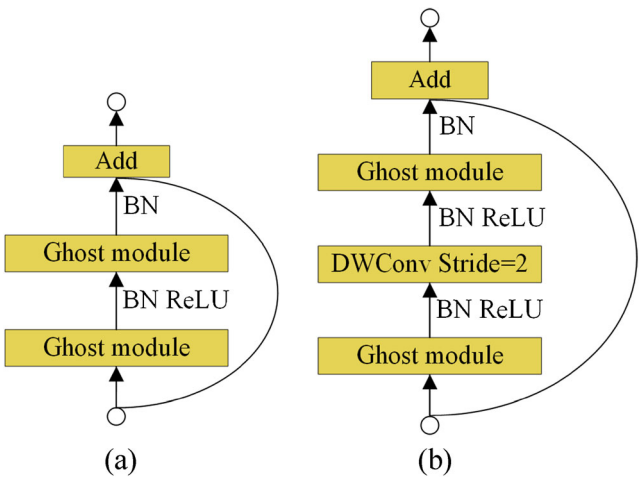

Fig. 7 G-bneck

The convolution operations include Gconv, BN, and ReLU. To further reduce the parameters, two 3 × 3 convolutional kernels are used instead of the standard 5 × 5 convolutional kernels.

(2) In the fuse operation, the output of Split is filtered by the gating mechanism(Srivastava et al. 2015). First, n feature maps are fused by summing elements to obtain feature map $U_c = (U_1+U_2+\ldots U_n) \in R^{C\times H\times W}$ ($n = 1, 2, \ldots, n$). Next, global average pooling (GAP) $F_{gap} \in R^{C\times 1\times 1}$ is used to act on U_c to embed global information $S \in R^{C\times 1}$. Specifically, the c-th element of S is calculated by contracting the c-th channel over the spatial dimensions $H \times W$ in U.>

$$S_c = F_{gp}(U_c) \tag{2}$$

The feature $Z \in R^{d\times 1}$ ($d < C$) is obtained by acting on$S$using the full connection F_{fc}.

$$Z = F_{fc}(s) = \sigma_{relu}(BN(Ws)) \tag{3}$$

where $W = R^{d\times c}$, $d = max\left(\frac{c}{r}, L\right)$, r and L reduce the output dimension to improve the model computational efficiency.

(3) In the selection operation, the new feature map V is calculated based on the soft attention of n cross-channels and n feature maps obtained in Split. Using full connectivity acting on Z, we obtain n channel statistics N_i with expansion dimension C.>

$$n_{i,c} = \frac{e^{N_{i,c}Z}}{e^{N_{i,c}Z} + \cdots + e^{N_{n,c}Z}}, \forall i = 1, \ldots, n \tag{4}$$

where $N_i \in R^{c\times d}$, n_i is the soft attention of U_n, $n_{i,\,c} \in R^{1\times d}$ is the c-th line of N_i, n_i is the c-th element of n, and $n_{i,\,c} = 1$. The new feature map VUC is obtained by applying n_i to U:

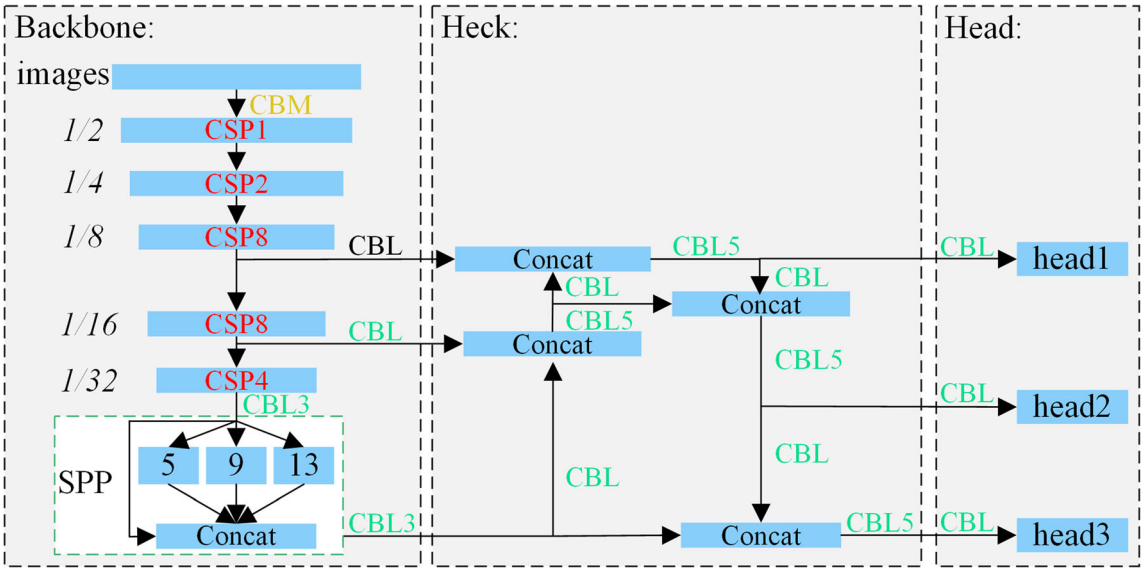

Fig. 8 YOLOv4 model structure

$$V_c = U_c \times n_{i,c} \tag{5}$$

where $V = [V_1, V_2, \ldots, V_c]$, $V_c \in R^{H \times W}$.

The SK attention mechanism also reduces parameters in the model. Gconv operations are used in the SK attention module, which reduces the parameters to G-fold less than that of standard convolutional operations.

YOLO-SA model

Object detectors are composed of three parts: a pretrained backbone, a head for predicting object classes and bounding boxes, and a neck for collecting features by inserting some CNN layers between the backbone and the head. The YOLO series detection model is composed of three parts: the backbone, neck, and head. Figure 8 shows the model structure of YOLOv4 as follows:

(1) Backbone: First, the H × W-sized image is input and scaled down to 32 times its original size after the backbone. Then, the spatial pyramid pooling (SPP) module (He et al. 2014) is used to improve the receptive field of the model.
(2) Neck: Cross-scale information is fused from different receptive fields.
(3) Head: The three detection heads are composed of three feature maps of size $^1/_8$, $^1/_{16}$, and $^1/_{32}$ of the original image, which are used to detect large, medium, and small objects, respectively.

As shown in Fig. 9, in each detection head, each cell corresponds to three types of positioning boxes, predicting the bounding boxes of three sizes of objects and ultimately screening out the most suitable bounding boxes. For each bounding box, the model predicts the four coordinate values (upper left and lower

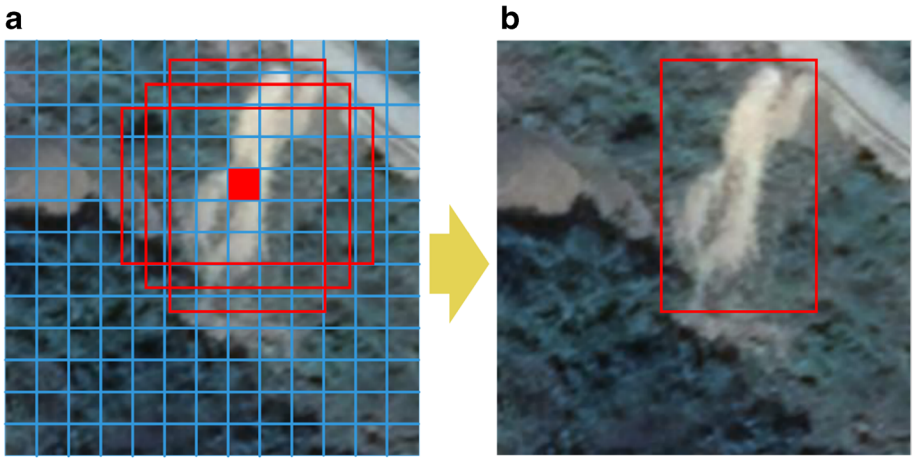

Fig. 9 Schematic diagram of the predicted landslide boundary box from an image of 13 × 13 cells. **a** The detection process on the feature map of 13 × 13 cells. **b** The detection results

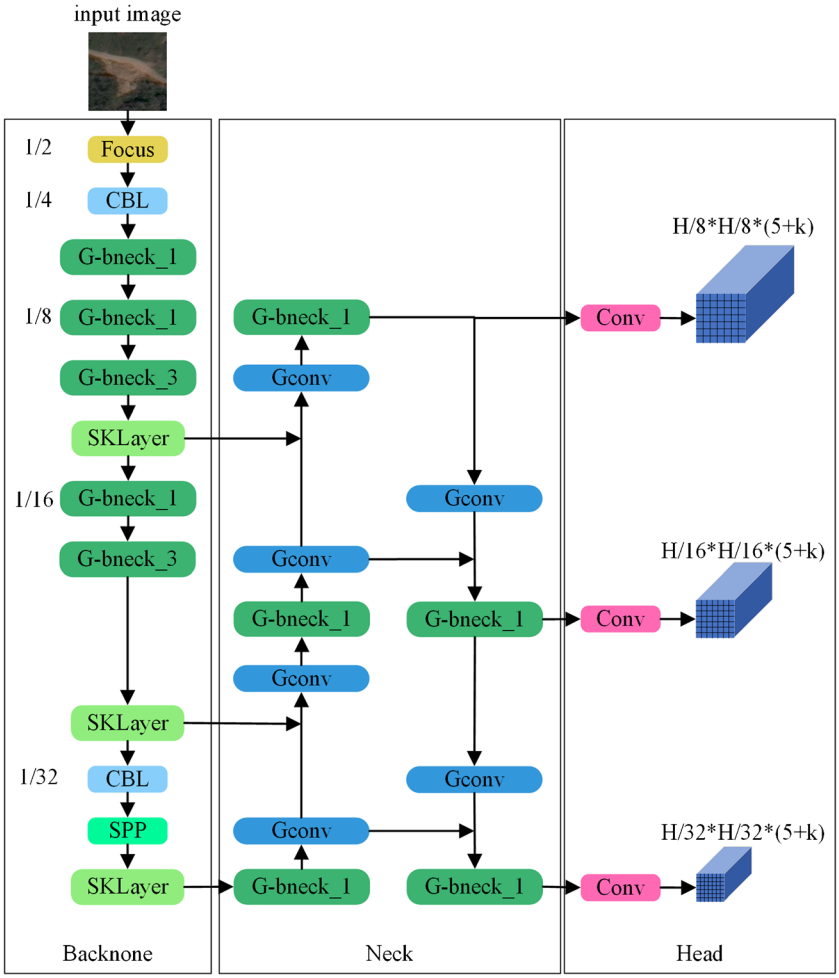

Fig. 10 Structure of the YOLO-SA model

right corners or centroid, width, and height), confidence, and category. For the prediction of the coordinate position, the offset of the coordinates of the positioning box from the corresponding position of the real box is calculated, such as {center x, y offset and W, H offset} or {upper left corner x and y offset and lower right corner x, y offset}. For n input images, the final number of

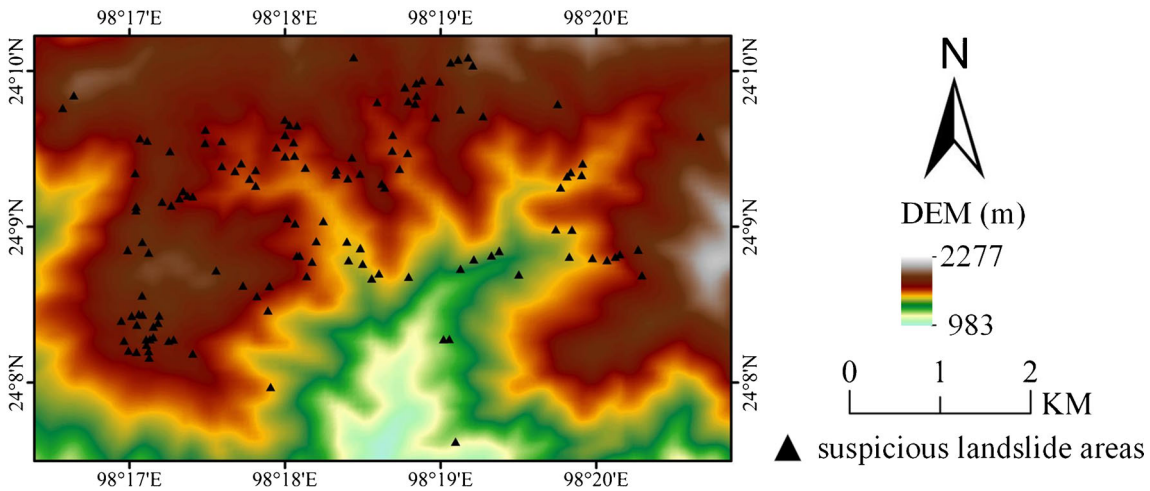

Fig. 11 The potential landslides (black triangles), which need to be further confirmed by comparing them with actual landslides.

Table 1 Results of YOLO-SA detection of potential landslides

Model	Actual existence	Correct detection	omissions
YOLO-SA	128	126	2

channels output by the three detection heads is $n \times \left(3 \times \left({}^{H}/_{32} \times {}^{W}/_{32} + {}^{H}/_{16} \times {}^{W}/_{16} + {}^{H}/_{8} \times {}^{W}/_{8}\right)\right) \times (5 + k)$, where k is the number of object classes.

Model training techniques. Training a detection model with better performance may require adding new data enhancement methods and combining techniques such as Focus from the passthrough layer (Redmon and Farhadi 2017), decay learning rate scheduling strategies, and loss function training to obtain it.

In YOLOv4, the backbone, neck, and head include three components: Two convolution components CBM and CBL and one residual component CSPx, where the component CBM is composed of standard convolution, BN, and Mish. the component CBL replaces the activation function in the CBM with LeakyReLU (Zhang et al. 2017). The residual component CSPx is composed of a number of CSP residual modules, among which the CSP is composed of many CBMs.

In the reconstructed YOLO-SA model structure, Gconv and G-bneck are used to replace the convolutional components CBM and CBL. Focus, CBL, and G-bneck replace the residual component CSPx. Several G-bnecks are added for learning feature information. The structure of the reconstructed YOLO-SA model is shown in Fig. 10, where the attentional mechanism is introduced in the backbone.

Experimental setup

In this study, all experiments use the same data division standard, and the marked landslide dataset is divided into a training set and test set, with a division ratio of 8:2. All experiments are conducted on the same workstation with an NVIDIA Quadro P4000 graphics card, 8 GB of GPU memory, and an Intel Xeon Gold 5118 CPU.

During the training process, all parameters are divided into independent training units: weights of the convolution kernel, deviation terms, and others. All training units use the SGD

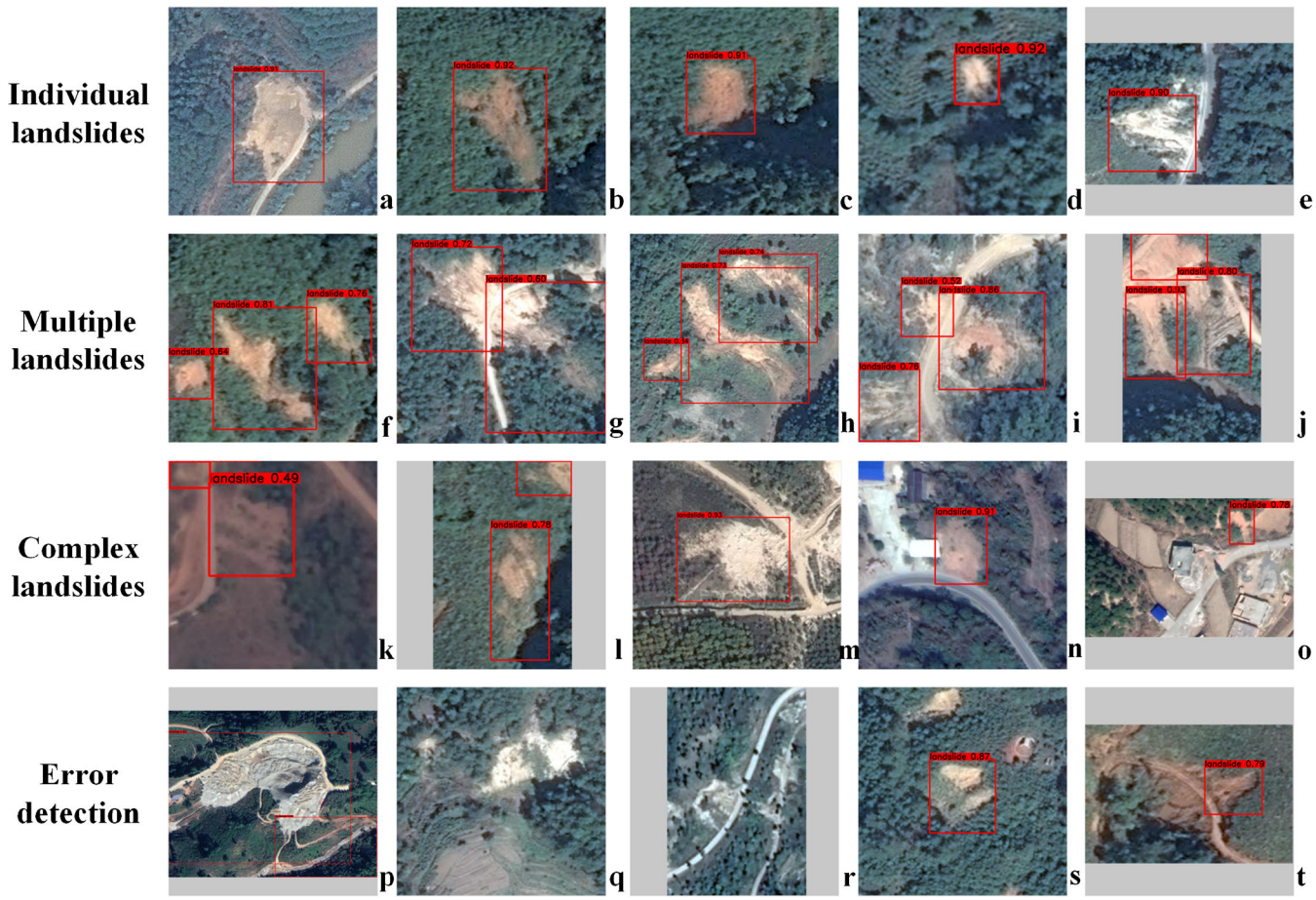

Fig. 12 Detection results using YOLO-SA on areas of potential landslides

Table 2 Performance comparison of YOLO-SA with 11 existing advanced models

Model	Params (mb)	AP	FPS (f/s)
YOLO-SA (**ours**)	**1.472**	**0.9408**	**42**
Faster RCNN	41	0.6810	4
EfficientDet-D0	3.9	0.7634	18
EfficientDet-D1	6.6	0.7946	14
EfficientDet-D2	8.1	0.8852	13
SSD-efficient	97.1	0.8709	37
Centernet-ResNet-50	32	0.7705	37
Centernet-hourglass	191	0.8983	8
YOLOv4	64	0.6560	10
YOLOv4-tiny	5.9	0.6950	17
Mobilenetv2-YOLOv4	46.3	0.7520	18
Mobilenetv3-YOLOv4	48.3	0.7120	17

We aim to highlight the results of our own experiments

optimizer (Polyak and Juditsky 1992) with an initial learning rate of 10^{-2}, and the learning rate takes a range of values of $(10^{-5}, 10^{-2})$. The initial momentum is 0.937, and the momentum takes a range of values (0, 0.7). The nesterov momentum is turned on. When the weights of the convolution kernel use momentum decay, the initial value of decay is 5×10^{-4}, and the decay range is (0, 0.001). In the optimizer, the decay of the learning rate is updated using the cosine annealing scheduler (Loshchilov and Hutter 2016). The threshold of the prediction box is set to 0.5. During the training process, the input image is normalized, and the normalization range is (0, 1). The batch size is 16. The size of the input image is resized to 640×640 pixels. The number of training steps is 300.

Results

Model assessment

To evaluate the performance of the model in terms of the detection speed, accuracy, and number of parameters, the FPS, Params, and AP evaluation metrics are considered. FPS refers to the number of frames per second, or f/s, that a video can display. Here, we define the number of images that the model can continuously process per second. The higher the FPS is, the faster the processing speed. In practice, when the FPS is greater than 30 f/s, the model is judged to provide the ability to process the image in real time. Params refers to the sum of the weights, bias terms, and other parameters to be processed during model training in units of mb. The fewer the parameters, the smaller the calculation cost is. AP is the most important evaluation index for the target detection algorithms and is calculated by the area under the curves of accuracy and recall. The larger the AP value is, the better the model performance. The accuracy rate refers to the ratio of the number of areas correctly predicted to be landslides to the total number of areas predicted to be landslides. The recall rate is the ratio of the number of correctly predicted landslide areas to the number of real landslides on the ground. The formula for AP is as follows:

$$\mathrm{AP} = \frac{1}{\mathrm{r}_{\mathrm{number}}} \sum_{\mathrm{r} \in \{0, x, y, \ldots, 1\}} \mathrm{P}_{\mathrm{interp}}(\mathrm{r}) \tag{6}$$

$$\mathrm{P}_{\mathrm{interp}}(\mathrm{r}) = \max_{\widetilde{r}: \widetilde{r} \geq r} \mathrm{P}\left(\widetilde{r}\right) \tag{7}$$

where $\mathrm{P}_{\mathrm{interp}}(\mathrm{r})$ is the maximum accuracy rate in accordance with specific conditions, and $\mathrm{P}(\widetilde{r})$ is the accuracy rate when the recall rate is $\widetilde{r}$.

Model results for potential landslide detection

To verify the application performance of the YOLO-SA model, the training model is applied to landslide detection in unknown areas. Dehong Autonomous Prefecture is located in the western part of Yunnan Province, China, and landslides are frequent at the junction of Zhafang Town and Manhai Town in its subordinate city of Mangshi. The geological survey found that 128 landslide areas existed in the area as of 2016. As shown in Fig. 11, the model detected a total of 150 suspected landslide areas in the area.

Further analysis revealed that the YOLO-SA model can detect most of the landslide areas in the remote sensing images, and the detection

Table 3 Effects of different components on the model performance

Model	Layers	Params (mb)	AP	FPS (f/s)
G-bneck-CBL	225	4.8623	0.8400	57
G-bneck-Gconv	225	2.4977	0.8922	49

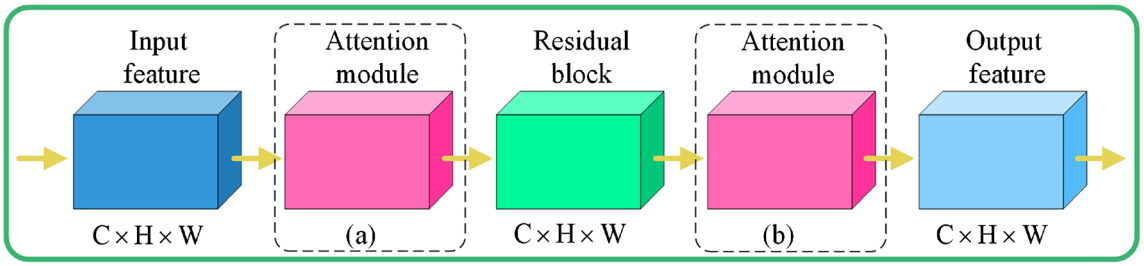

Fig. 13 Different locations ((**a**) and (**b**)) of attentional mechanisms

results are shown in Table 1. To better represent the detection accuracy of the model, the following definitions are provided.

P is the number of landslides detected by the model that are indeed landslides as a percentage of the total number of landslides detected.

R is the number of landslides correctly detected as a percentage of the total number of landslides.

We want the model to have neither false prediction nor omission. The F1 score weighs *P* and *R*. The larger the value of *P* and *R*, the better, and the larger the F1 score. The F1 expression is as follows.

$$F1 = 2\frac{PR}{P+R} \in [0,1] \tag{8}$$

The F1 score of the model is 90.65%, the omission rate is 1.56%, and the error rate is 16%.

In all correctly detected areas, the model is able to accurately determine a landslide that exists alone and assigns a high confidence level (over 90%), as shown in Fig. 12a–e. When faced with complex terrain, even if two or more landslide mounds are present at the same time, the model is able to detect them accurately but gives a lower confidence. The detection results are shown in Fig. 12f–j.

Strong robustness is demonstrated. As shown in Fig. 12k–l, the landslide area is only partially represented in the remote sensing image, and the model is able to find it. In this way, when historical landslide areas are partially covered by surface vegetation, the model may have a higher discrimination ability. The antiinterference ability of the model is shown in Fig. 12m–o. Early soil and rock, landslide-like vertical bars, and mounds are present on the right side of the landslide area, and the model accurately ignores them and detects the landslide area on the left.

However, a comparison shows that the model makes several false identifications and misses some identifications. Figure 12p shows that the model misjudges the lime plant as a landslide area because of the high visual similarity between the two. The surface of the landslide area in Fig. 12q is highly reflective, and the landslide area in Fig. 12r is similar to a mound, which leads to missed detections. Figure 12s and t show two partially missed areas, caused by the landslides being mixed with image shadows and buildings.

Comparison of different detection models

To determine the performance of YOLO-SA, a comparison is made with 11 advanced detection models: Faster RCNN, 3 types of EfficientDet (Tan et al. 2019), 2 types of Centernet (Zhou et al. 2019), SSD-efficient, YOLOv4, YOLOv4-tiny (Bochkovskiy et al. 2020), Mobilenetv2-YOLOv4 and Mobilenetv3-YOLOv4 (Yang and Deng 2020). Original versions of these classical models are not tested due to their poor performance. The comparison results are shown in Table 2. Compared to the EfficientDet-Do model with few parameters, the parameters of YOLO-SA are only 1.472 mb. The detection accuracy increases by 4.25% compared to the optimal Centernet-hourglass model. The detection speed is up to 42 f/s. Although many models reduce the parameter redundancy, they achieve lower detection accuracy.

Discussion

In this section, we discuss the factors that affect the performance of the model in three aspects: The different components, attentional mechanisms, and other training methods.

The effects of different components on the model performance

During the experiment, the influence of Gconv and G-bneck on the performance of the model is verified. As shown in Table 3, the model constructed using G-bneck and CBL has lower detection accuracy but a high velocity of 57 f/s. On this basis, after replacing

Table 4 Effects of the location of the attention mechanism on the model performance

Description	Layers	Params (mb)	AP	FPS (f/s)
YOLO-SA_backbone_behind	309	1.7777	0.8953	30
YOLO-SA_backbone_front	309	**1.5594**	**0.8982**	**31**
YOLO-SA_head_behind	337	3.8581	0.8467	29
YOLO-SA_head_front	337	3.8581	0.8778	28
YOLO-SA_behind	379	4.1638	0.8947	25
YOLO-SA_front	379	3.9455	0.8813	25

We aim to highlight excellent experimental results

Table 5 Effects of different attention mechanism modules on the model performance

Description	Layers	Params (mb)	AP	FPS (f/s)
YOLO-SA (**ours**)	**267**	**1.4720**	**0.9408**	42
YOLO-SA-SE	231	2.5407	0.8676	49
YOLO-SA-scSE	234	2.8427	0.7624	50
YOLO-SA-Non-local-concatenation	255	3.1899	0.8380	42
YOLO-SA-Non-local-dot_product	255	3.1899	0.9101	45
YOLO-SA-Non-local-embedded_gaussian	255	3.1899	0.7980	40
YOLO-SA-GC	249	2.5611	0.8707	45
YOLO-SA-Split	234	1.8842	0.8614	**51**

We aim to highlight our own experimental results and other excellent experimental results

the CBL in Neck with Gconv, we are surprised to find that Gconv does not change the number of layers of the model; the accuracy of the model is improved, and the parameters is reduced, but the detection speed is reduced slightly.

The effects of the attention mechanism on the model

The effects of the location of the attention mechanisms on the model performance

In the model structure, the positions of the attention mechanism have different effects on the final result. In the experiments, the attentional mechanisms are placed in different positions, the anterior and posterior sides of the residual module, and according to the different structures of the model, the position of attention is divided into six situations: The anterior and posterior sides of the residual module in the backbone network; the anterior and posterior sides of the residual module in the neck; and the anterior and posterior sides of the residual module in the overall model. As shown in Fig. 13, the attention module is used to refine the mapping features of the input and output parts of the residual module. From Table 4, the model has a high detection effect when an attentional mechanism is used to refine the input mapping of the residual modules in the model's backbone, which is the same conclusion reached by Hu Jie, the author who designed SENet (Hu et al. 2018).

Comparison with other attention modules

Different attentional mechanisms pay different attention to feature maps. In the experiment, we compare seven other advanced attention mechanisms: The SE module, the scSE module (Roy et al. 2018), three versions of the nonlocal module (Wang et al. 2017), the GC module (Cao et al. 2019) and the Split module (Zhang et al. 2020). As seen in Table 5, although the number of layers of the YOLO-SA model is higher than that of the other models, it has the least number of parameters and highest accuracy, and the speed can meet the demand of real-time detection. The improved detection accuracy is due to the advantages of the SK attention mechanism.

Effects of other training methods on the model accuracy

Effects of different training techniques on the model performance

Different training techniques produce different feature mappings that affect the model prediction results. Specifically, these different techniques include basic data enhancement and mosaic data augmentation (Bochkovskiy et al. 2020); cosine annealing schedulers; activation functions such as LeakyReLU, swish, and hard-swish (Howard et al. 2019; Ramachandran et al. 2017; Zhang et al. 2017); and loss functions such as generalized intersection over union (GIoU), distance IoU (DIoU), and complete IoU (CIoU) (Rezatofighi et al. 2019; Zheng et al. 2020). From Table 6, the following training techniques are introduced in the experiment to improve the accuracy of the model: mosaic data augmentation, cosine annealing scheduler, hard-swish, and GIoU.

- B: Basic data augmentation.

Table 6 Effects of different training techniques on the model performance

B	M	CA	LR	Swish	Hs	Mish	Loss	Layers	Params (mb)	AP	FPS (f/s)
√			√				GIoU	267	1.4720	0.3208	40
√	√		√				GIoU	267	1.4720	0.7592	38
√	√	√	√				GIoU	267	1.4720	0.8973	40
√	√	√		√			GIoU	267	1.4720	0.8724	38
√	√	√			√		GIoU	267	1.4720	0.9091	42
√	√	√				√	GIoU	267	1.4720	0.8711	40
√	√	√			√		DIoU	267	1.4720	0.8603	40
√	√	√			√		CIoU	267	1.4720	0.8398	40

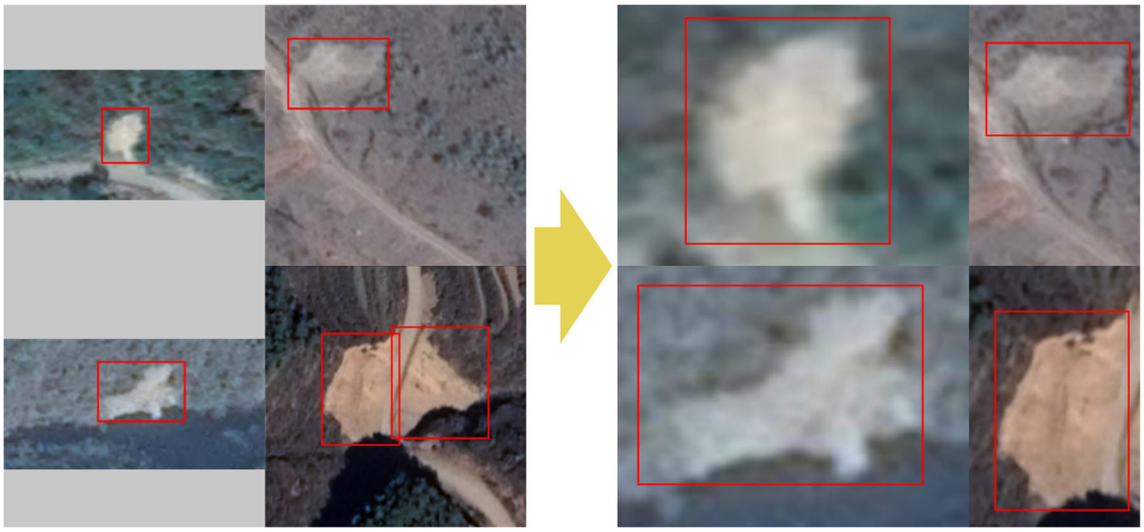

Fig. 14 Mosaic data augmentation

- M: Mosaic data augmentation—the mosaicking of four landslide images into one image increases the background information of the landslide image while reducing the computational cost of BN and the training cost (Fig. 14).
- CA: Cosine annealing scheduler—the cosine annealing scheduler is used to set the learning rate for each parameter group.
- LeakyReLU (LR), swish, hard-swish (hS), and Mish—different activation functions are used to change the nonlinear fitting ability of the model.
- GIoU, DIoU, and CIoU—different loss functions are used to perform regression calculations on the bounding boxes.

Effects of the number of detector heads on the model performance

The number of detector heads affects the inspection accuracy of the model. In this study, an additional detector head is added to each of the above eight attentional mechanism experiments to detect more fine-grained landslide samples: large, medium, general, and small landslide feature maps. The experimental results are shown in Table 7. The model with the fewest parameters is still that using the SK attention mechanism, but this model has lower detection accuracy. The model with the highest detection accuracy has the highest number of parameters and the lowest speed. For the rest of the models, the detection accuracy is below 90%. Compared with Table 5, if the number of detection heads continues to increase, the performance of the model continues to decline, so adding extra inspection heads is not recommended.

Effect of different training batch sizes on the model accuracy

The batch size of the training samples affects the optimization of the model and the speed of detection. We analyze the model performance changes caused by input images of different sizes to three batch size samples on YOLO-SA, and the results are shown in Table 8. As seen in the table, the sample batch sizes are 16, 8, and 4 for the input image sizes of 512×512 pixels and 416×416 pixels, and the model performance is basically unchanged. This result shows that the YOLO-SA model can be used to train excellent models without using expensive GPUs.

Conclusion

In this study, we create a large-scale landslide dataset using satellite remote sensing images after landslides and mark the landslide areas on the dataset. Then, we reconstruct the structure of the deep

Table 7 Effect of adding detector heads on the model performance

Description	Layers	Params (mb)	AP	FPS (f/s)
YOLO-SA-SK-4_head	302	**1.5029**	0.8724	30
YOLO-SA-SE-4_head	266	2.8496	0.8737	33
YOLO-SA-scSE-4_head	269	3.1516	0.7772	31
YOLO-SA-Non-local-concatenation-4_head	290	3.4988	**0.9189**	28
YOLO-SA-Non-local-dot_product-4_head	290	3.4988	0.8013	29
YOLO-SA-Non-local-embedded_gaussian-4_head	290	3.4988	0.8346	29
YOLO-SA-GC-4_head	284	2.6217	0.8777	37
YOLO-SA-Split-4_head	269	2.2092	0.8354	33

We aim to highlight excellent experimental results

Table 8 Effect of different training batch sizes on the detection accuracy of the YOLO-SA model

Image size	Batch size	AP
	16	0.9096
512	8	0.9155
	4	0.9017
	16	0.9226
416	8	0.9570
	4	0.9070

learning detection model YOLOv4 and propose our YOLO-SA model. The experimental results show that the improved YOLO-SA model has high detection accuracy, a fast detection speed and very few parameters. The parameters are reduced to 1.472 mb compared with EfficientDet-D0, the accuracy is improved to 94.08% compared with Centernet-hourglass, and the speed is up to 42 f/s. This study verifies the feasibility of the YOLO-SA model for landslide detection on high-spatial-resolution remote sensing images after landslides: The F1 score of the model is 90.65%, the omission rate is 1.56%, and the error rate is 16%. However, there is some error in the detection accuracy of the model due to the small landslide dataset and human error in the area markings. Therefore, increasing the landslide dataset to include various types of remote sensing images to improve the generalization ability and detection accuracy of the model will be performed in future work.

Acknowledgements

We thank all the authors for their great contribution to this study. We thank Zhoujie Luo and Peng Lai for their help in the experimental data collection and labeling process. Thanks for the valuable landslide data provided by the Yunnan Geological Disaster Department.

Author contribution

Conceptualization: [Jia Li], [Ping Duan]; methodology: [Libo Cheng], [Jia Li]; formal analysis and investigation: [Libo Cheng]; writing - original draft preparation: [Libo Cheng]; writing - review and editing: [Libo Cheng], [Jia Li]; funding acquisition: [Jia Li]; resources: [Mingguo Wang], [Ping Duan]; supervision: [Jia Li], [Ping Duan]; accuracy evaluation: [Mingguo Wang], [Ping Duan].

Funding

This research was funded by the National Natural Science Foundation of China (No. 41961061) and Yunnan Fundamental Research Projects (grant NO. 202001AT070057).

Availability of data and material

The datasets used or analyzed during the current study are available from the corresponding author on reasonable request.

Code availability

The code used during the current study is available from the corresponding author on reasonable request.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication All authors have read and agreed to the published version of the manuscript.

Competing interests The authors declare that they have no competing interests.

References

Amatya P, Kirschbaum D, Stanley T (2019) Use of very high-resolution optical data for landslide mapping and susceptibility analysis along the Karnali Highway, Nepal. Remote Sens 11:2284. https://doi.org/10.3390/rs11192284

Bochkovskiy A, Wang CY, Liao HYM (2020) YOLOv4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934

Cao Y, Xu J, Lin S, Wei F, Hu H (2019) GCNet: non-local networks meet squeeze-excitation networks and beyond. arXiv preprint arXiv:1904.11492

Cheng G, Guo L, Zhao T, Han J, Li H, Fang J (2013) Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA. Int J Remote Sens 34:45–59. https://doi.org/10.1080/01431161.2012.705443

Cheng G, Li R, Lang C, Han J (2021) Task-wise attention guided part complementary learning for few-shot image classification. Sci China Inform Sci 64:120104

Di Napoli M et al (2020) Machine learning ensemble modelling as a tool to improve landslide susceptibility mapping reliability. Landslides 17:1897–1914. https://doi.org/10.1007/s10346-020-01392-9

Du S, Zang P, Zang B, Xu H (2021) Weak and occluded vehicle detection in complex infrared environment based on improved YOLOv4. IEEE Access 9:25671–25680. https://doi.org/10.1109/ACCESS.2021.3057723

Fu CY, Liu W, Ranga A, Tyagi A, Berg AC (2017) DSSD: deconvolutional single shot detector. arXiv preprint arXiv:170106659

Galli M, Ardizzone F, Cardinali M, Guzzetti F, Reichenbach PJG (2008) Comparing landslide inventory maps. Geomorphology 94:268–289. https://doi.org/10.1016/j.geomorph.2006.09.023

Gu T, Li J, Wang M, Duan P (2021) Landslide susceptibility assessment in Zhenxiong County of China based on geographically weighted logistic regression model. Geocarto Int:1–23. https://doi.org/10.1080/10106049.2021.1903571

He K, Zhang X, Ren S, Sun J (2014) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal 37:1904–1916

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, p 770-778. https://doi.org/10.1109/CVPR.2016.90

He K, Gkioxari G, Dollár P, Girshick R Mask R-CNN (2017) In: arXiv preprint arXiv:1703.06870

Hong et al (2019) Improved faster R-CNN with multiscale feature fusion and homography augmentation for vehicle detection in remote sensing images. IEEE Geosci Remote Sens Lett 16:1761–1765. https://doi.org/10.1109/LGRS.2019.2909541

Howard AG et al. (2017) MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:170404861

Howard A et al. (2019) Searching for MobileNetV3. Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1314-1324

Hu J, Shen L, Albanie S, Sun G, Wu E (2018) Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell:7132–7141

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

Li T-T, Pei X-J, Huang R-Q, Jin L-D (2016) The formation and evolution of the Qiaojia pull-apart basin, North Xiaojiang Fault Zone, Southwest China. J Mt Sci Engl 13:1096–1106. https://doi.org/10.1007/s11629-015-3778-1

Li X, Wang W, Hu X, Yang J (2019) Selective Kernel Networks. Proceedings of the IEEE conference on computer vision and pattern recognition, p 510-519

Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. IEEE Trans Pattern Anal:2999–3007

Liu W, Anguelov D, Erhan D (2016) SSD: Single shot multibox detector. arXiv preprint arXiv:151202325: 21-37

Loshchilov I, Hutter F (2016) SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983

Ma H, Liu Y, Ren Y, Yu J (2020) Detection of collapsed buildings in post-earthquake remote sensing images based on the improved YOLOv3. Remote Sens 12:44. https://doi.org/10.3390/rs12010044
Maxwell AE, Pourmohammadi P, Poyner JD (2020) Mapping the topographic features of mining-related valley fills using mask R-CNN deep learning and digital elevation data. Remote Sens 12:547. https://doi.org/10.3390/rs12010044
Messeri A, Morabito M, Messeri G, Brandani G, Petralli M, Natali F, Grifoni D, Crisci A, Gensini G, Orlandini S (2015) Weather-related flood and landslide damage: a risk index for Italian Regions. PLoS One 10:e0144468
Polyak BT, Juditsky AB (1992) Acceleration of stochastic approximation by averaging. SIAM J Control Optim 30:838–855. https://doi.org/10.1137/0330046
Ramachandran P, Zoph B, Le QV (2017) Swish: a self-gated activation function. arXiv preprint arXiv:1710.05941
Redmon J, Farhadi A (2017) YOLO9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6517-6525
Redmon J, Farhadi A (2018) YOLOv3: an incremental improvement. arXiv preprint arXiv:180402767
Redmon J, Divvala S, Girshick R, Farhadi A (2015) You only look once: unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, p 779-788
Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal 39:1137–1149
Rezatofighi H, Tsoi N, Gwak JY, Sadeghian A, Savarese S (2019) Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the IEEE conference on computer vision and pattern recognition
Roy AG, Navab N, Wachinger C (2018) Concurrent spatial and channel squeeze & excitation in fully convolutional networks. International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp 421–429
Srivastava RK, Greff K, Schmidhuber J (2015) Training very deep networks. arXiv preprint arXiv:150500387
Tan M, Pang R, Le QV (2019) EfficientDet: scalable and efficient object detection. arXiv preprint arXiv:191109070
Wang X, Girshick R, Gupta A, He K (2017) Non-local neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition, p 7794-7803
Wang H, Liu S, Xu W, Yan L, Qu X, Xie W-C (2020) Numerical investigation on the sliding process and deposit feature of an earthquake-induced landslide: a case study. Landslides 17:2671–2682. https://doi.org/10.1007/s10346-020-01446-y
Xu Z, Chen Y, Yang F, Chu T, Zhou H (2020) A postearthquake multiple scene recognition model based on classical SSD method and transfer learning. ISPRS Int J Geo-Inf 9:238. https://doi.org/10.3390/ijgi9040238
Yang Y, Deng H (2020) GC-YOLOv3: you only look once with global context block. Electronics 9:1235. https://doi.org/10.3390/electronics9081235
Yang Y, Yang J, Xu C, Xu C, Song C (2019) Local-scale landslide susceptibility mapping using the B-GeoSVC model. Landslides 16:1301–1312. https://doi.org/10.1007/s10346-019-01174-y
Zhang X, Zou Y, Wei S (2017) Dilated convolution neural network with LeakyReLU for environmental sound classification. In: Digit Signal Process, pp 1-5
Zhang H et al. (2020) ResNeSt: split-attention networks. arXiv preprint arXiv:2004.08955
Zheng Z, Wang P, Liu W, Li J, Ren D (2020) Distance-IoU loss: faster and better learning for bounding box regression. In: Distance-IoU loss: faster and better learning for bounding box regression, pp 12993-13000
Zhou X, Wang D, Krhenbühl P (2019) Objects as points. arXiv preprint arXiv:191109070

L. Cheng · J. Li (✉) · P. Duan (✉)
Faculty of Geography,
Yunnan Normal University,
Kunming, 650500, Yunnan, China
Email: keguigiser@163.com
Email: dpgiser@163.com

L. Cheng
e-mail: chenglibogis@163.com

M. Wang
Yunnan Institute of Geological Sciences,
Kunming, 650501, Yunnan, China
e-mail: wangmingguo86@hotmail.com