remote sensing

MDPI

Article

SAM-CFFNet: SAM-Based Cross-Feature Fusion Network for Intelligent Identification of Landslides

Laidian Xi [1], Junchuan Yu [2,*], Daqing Ge [2], Yunxuan Pang [1], Ping Zhou [1], Changhong Hou [3], Yichuan Li [2], Yangyang Chen [2] and Yuanbiao Dong [2]

[1] School of Earth Sciences and Resources, China University of Geosciences, Beijing 100083, China; 2101210160@email.cugb.edu.cn (L.X.); 2101210161@email.cugb.edu.cn (Y.P.); zhoupx@cugb.edu.cn (P.Z.)

[2] China Aero Geophysical Survey and Remote Sensing Center for Natural Resources, Beijing 100083, China; gedaqing@mail.cgs.gov.cn (D.G.); lyichuan@mail.cgs.gov.cn (Y.L.); chenyangyang@mail.cgs.cov.cn (Y.C.); dongyuanbiao@mail.cgs.gov.cn (Y.D.)

[3] School of Geosciences and Surveying Engineering, China University of Mining and Technology, Beijing 100083, China; bqt2300203041@student.cumtb.edu.cn

* Correspondence: yujunchuan@mail.cgs.gov.cn; Tel.: +86-151-0115-7141

Abstract: Landslides are common hazardous geological events, and accurate and efficient landslide identification methods are important for hazard assessment and post-disaster response to geological disasters. Deep learning (DL) methods based on remote sensing data are currently widely used in landslide identification tasks. The recently proposed segment anything model (SAM) has shown strong generalization capabilities in zero-shot semantic segmentation. Nevertheless, SAM heavily relies on user-provided prompts, and performs poorly in identifying landslides on remote sensing images. In this study, we propose a SAM-based cross-feature fusion network (SAM-CFFNet) for the landslide identification task. The model utilizes SAM's image encoder to extract multi-level features and our proposed cross-feature fusion decoder (CFFD) to generate high-precision segmentation results. The CFFD enhances landslide information through fine-tuning and cross-fusing multi-level features while leveraging a shallow feature extractor (SFE) to supplement texture details and improve recognition performance. SAM-CFFNet achieves high-precision landslide identification without the need for prompts while retaining SAM's robust feature extraction capabilities. Experimental results on three open-source landslide datasets show that SAM-CFFNet outperformed other comparative models in terms of landslide identification accuracy and achieved an intersection over union (IoU) of 77.13%, 55.26%, and 73.87% on the three datasets, respectively. Our ablation studies confirm the effectiveness of each module designed in our model. Moreover, we validated the justification for our CFFD design through comparative analysis with diverse decoders. SAM-CFFNet achieves precise landslide identification using remote sensing images, demonstrating the potential application of the SAM-based model in geohazard analysis.

Keywords: landslide identification; SAM; deep learning; remote sensing; semantic segmentation; cross-feature fusion

check for updates

Citation: Xi, L.; Yu, J.; Ge, D.; Pang, Y.; Zhou, P.; Hou, C.; Li, Y.; Chen, Y.; Dong, Y. SAM-CFFNet: SAM-Based Cross-Feature Fusion Network for Intelligent Identification of Landslides. *Remote Sens.* **2024**, *16*, 2334. https://doi.org/10.3390/rs16132334

Academic Editors: Luca Brocca and Zhong Lu

Received: 23 February 2024
Revised: 20 May 2024
Accepted: 28 May 2024
Published: 26 June 2024

1. Introduction

Landslides are a serious geologic hazard on a global scale and have caused huge losses worldwide in the last few years [1]. They occur when heavy rainfall, earthquakes, and human activities trigger the movement of soil and rock on slopes [2–4]. The frequency and severity of landslide occurrences are on the rise, attributed to factors such as global warming, population growth, resource extraction, and environmental degradation [5,6]. Therefore, conducting landslide hazard studies and accurately identifying landslides are essential for assessing the impact of disasters, guiding post-disaster reconstruction, and preventing secondary disasters [7,8]. With the development of remote sensing and satellite technology, the application of remote sensing in large-scale geohazard investigations has become more and more popular, and great progress has been made in the identification of landslides using remote sensing [9–12].

Currently, there are four main approaches for landslide recognition methods based on remote sensing images: visual interpretation [13,14], pixel-based methods [15], object-based methods [16], and methods based on deep learning (DL) techniques [17–19]. Visual interpretation refers to the manual annotation and classification of landslide areas in images by professionals by directly observing and analyzing the features, morphology, color, texture, and other information of remote sensing images [20,21]. This method has the highest accuracy. However, it is time-consuming and labor-intensive. Pixel-based methods reduce the degree of human intervention through supervised training and improve computational efficiency [22], but it is difficult to obtain clear landslide boundaries and they cannot fully utilize the rich structural and textural information in the images [23]. The object-based method utilizes a variety of discriminative features, such as spectral features, texture features, and morphological features of landslides, for landslide detection [24–26]. This method is capable of categorizing objects with similar features into the same class, reducing salt-and-pepper noise. However, the recognition accuracy of this method largely depends on the initial segmentation precision, and it lacks strong capability for depicting details, resulting in a significant post-processing workload.

With the great progress of DL in the field of computer vision, DL-based methods have been widely applied to landslide recognition tasks and have become the main trend in this field [27,28]. Currently, research on DL models for landslide recognition mainly focuses on three directions: image classification, object detection, and semantic segmentation. Convolutional neural networks (CNNs) are commonly used in this research [29–31], with CNNs being characterized by a relatively complex structure, numerous training parameters, and high demands on training data. Ji et al. [32] employed an attention-boosted CNN model for the recognition of newly occurred landslides in Bijie, China, based on image classification. Ghorbanzadeh et al. [33] compared the performance of CNNs with neural networks, support vector machines, and random forests in the semantic segmentation of landslides and discovered that CNNs perform better when they have enough samples. In addition, other CNN-based models such as PSPNet [34], AlexNet [35], ResNet [36], U-Net [37], and DenseNet [38] have also been utilized for the semantic segmentation of landslides. Furthermore, target detection models represented by faster R-CNN [39] and the YOLO series [40–42] have also been applied to landslide recognition tasks.

With the introduction of the visual transformer (ViT) and its notable successes in computer vision, transformer-based models have been widely used in remote sensing identification tasks [43]. Chen et al. [44] developed sparse token transformers (STTs) for extracting buildings from remote sensing images. Utilizing a novel "sparse token sampler" module to represent buildings as sparse feature vectors, the STT achieves excellent performance on benchmark datasets while reducing computational complexity. Wang et al. [45] introduced a novel ViT architecture named BuildFormer that enables accurate building extraction from remote sensing images. It overcomes the limitations of traditional CNN methods in modeling global dependencies and preserving spatial details, achieving state-of-the-art performance. In the landslide identification task, Huang et al. [46] improved the Swin transformer by incorporating morphological edge analysis to address issues with landslide boundary discretization and irregularity, achieving more accurate landslide boundary extraction in the LuDing area of China. Lu et al. [47] proposed ShapeFormer, a shape-enhanced ViT model designed to effectively handle landslides of various sizes and shapes in remote sensing imagery, enhancing the accuracy of landslide detection. Fu et al. [48] significantly improved the accuracy and recognition capabilities of both YOLOv5 and faster R-CNN by replacing their backbones with Swin transformers. These models mentioned above, while performing well in specific tasks, often require large amounts of data for training. Moreover, their limited generalization, migration, and self-adaptation capabilities result in constraints when adapting to downstream tasks.

Recently, remarkable progress has been made in fundamental models such as GPT-4 [49], Flamingo [50], and SAM [51], which have made important contributions to the development of human society. SAM, a vision foundation model pre-trained on the SA-1B dataset, showcases substantial generalization capabilities across various image and object segmentation tasks without additional training. This creates new ways for intelligent interpretation of natural images [52–54]. SA-1B is the most extensive and diverse image segmentation dataset available. It contains over 11 million high-quality images taken from around the world, covering a wide range of scenes, objects, and environments, and consists of over 1 billion high-quality segmentation masks collected using Meta's data engine [51]. SAM comprises an image encoder, prompt encoder, and mask decoder. The image encoder is a ViT model pre-trained with an MAE [55] that takes an image as input and generates its embedding. The prompt encoder takes prompt information as input and outputs prompt embeddings. The mask decoder maps the image embedding and prompt embeddings to a mask. Since SAM is an interactive model, it can take point, box, or mask prompts when segmenting images. The segmentation results vary depending on the type of prompt used. Currently, some researchers have started to apply SAM to remote sensing data. Chen et al. [56] designed a prompt learning method, RSPrompter, for remote sensing images based on the SAM base model, which generates prompt inputs for SAM to enable it to automatically acquire instance-segmentation-level masks. Sultan et al. [57] introduced GeoSAM by introducing an innovative architecture for fine-tuning SAM using sparse and dense cues, leading to significant enhancements in geographic image segmentation. Zhang et al. [58] proposed RSAM-Seg, introducing adapter-scale in the multi-head attention block of the encoder in SAM, and inserting adapter-feature between ViT blocks. This design aims to generate prompts informed by images and enhance the model's performance in the remote sensing image segmentation tasks.

Figure 1 shows how well SAM recognizes different images with different types of prompts. Without prompt, SAM's performance on remote sensing images is notably weaker compared to other natural images. When prompts are provided, SAM excels at recognizing images with distinct boundaries and minimal background interference, such as airplane and factory images, but struggles with landslide images. Remote sensing images often require specialized spectral and spatial analyses due to their unique acquisition and processing techniques, and there are significant differences between them and natural images, especially for remote sensing landslide images, where common challenges include boundary blurring, complex background perturbations, and diverse morphological features. Although the SA-1B dataset includes images from various sources, such as natural scenes, urban environments, medical images, satellite images, etc., there is still room for further optimization of the SAM algorithm to improve the accuracy and generalization of target recognition in remote sensing imagery, which is in line with what other scholars have recognized [56–58].

In this paper, we propose the SAM-based cross-feature fusion network (SAM-CFFNet). This network is designed to create a novel semantic segmentation model for the high-precision recognition of landslides. We utilize SAM's image encoder to extract multi-level features from remote sensing optical images and design the cross-feature fusion decoder (CFFD) tailored to the characteristics of SAM's image encoder and the requirements of landslide recognition tasks. In the CFFD module, we propose a novel cross-fusion mechanism and demonstrate its effectiveness in subsequent experiments. Furthermore, the CFFD ensures high segmentation accuracy by incorporating a shallow feature extractor (SFE). We comprehensively evaluate the performance of SAM-CFFNet on three open-source landslide datasets, and the experimental results show that our model outperforms the other comparative models.

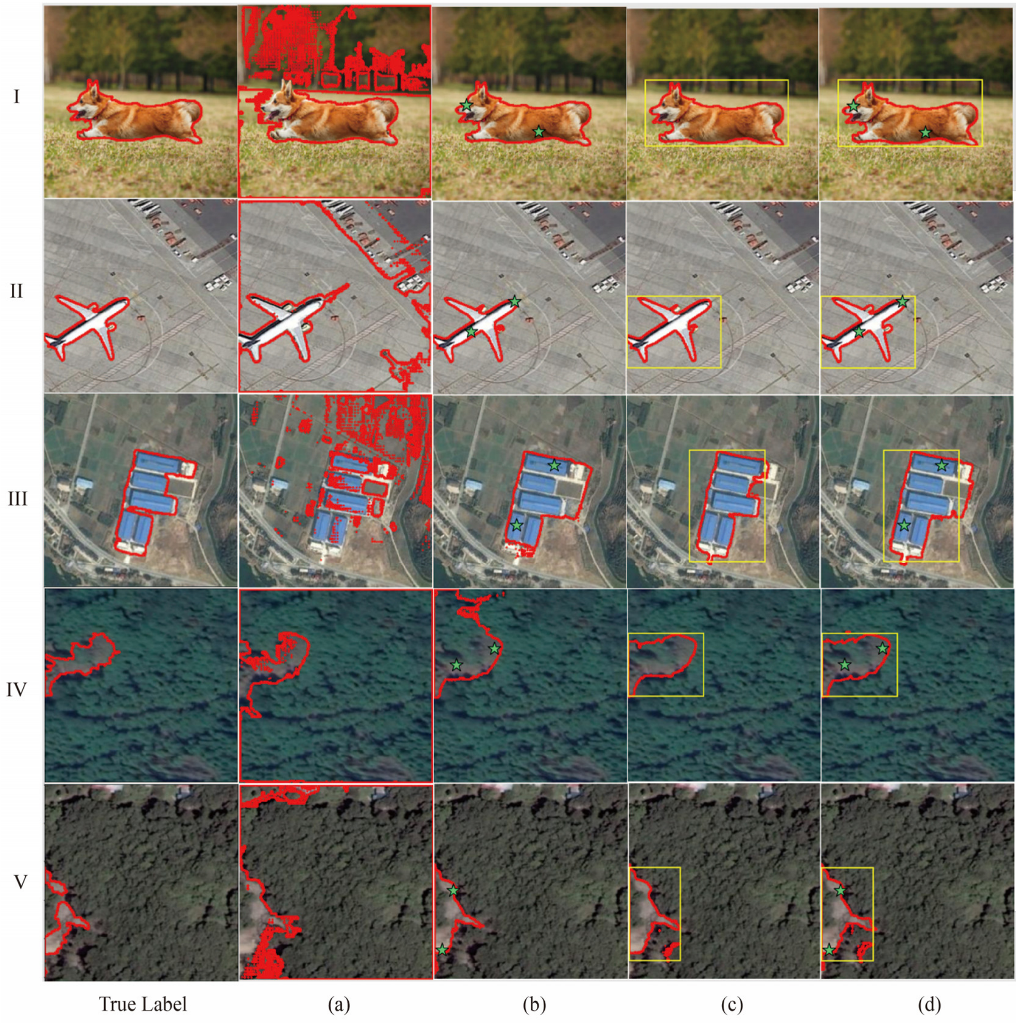

Figure 1. Segmentation effect of the SAM model on different images, where the green pentagram is the prompt point, the yellow box is the prompt box, and the red border is the SAM recognition result; (**a**) no prompt, (**b**) point-based prompt, (**c**) box-based prompt, (**d**) point and a box-based prompt. I–V represent images used for visual comparison, where I is a non-remote-sensing natural image, II and III are non-landslide remote sensing images, and IV and V are landslide remote sensing images.

The innovation of our approach is to leverage the powerful feature extraction capability of SAM to design decoders that are more adapted to the downstream segmentation task, which can achieve high-precision recognition of landslides with smaller trainable parameters without the intervention of human prompts, surpassing other traditional semantic segmentation methods. In summary, the main contributions of this study to the remote sensing landslide identification task are as follows:

(1) We propose a new semantic segmentation model, SAM-CFFNet, which demonstrates excellent performance on three landslide datasets, improving the accuracy of landslide recognition.
(2) Our proposed CFFD fully considers the characteristics of landslide images, can be well adapted to SAM's image encoder, and shows excellent performance in landslide recognition tasks.
(3) The excellent performance of SAM-CFFNet in the landslide identification task highlights the potential application of SAM in this field and provides new ideas and methods for the further application of the SAM base model in the remote sensing field.

2. Materials

2.1. Introduction of Datasets

For this experiment, three open-source remote sensing landslide datasets were selected to evaluate the performance of the research model. These datasets include the Bijie Landslide (BJL) dataset [32], the Landslide4Sense (L4S) dataset [59], and the Global Very-High-Resolution Landslide Mapping (GVLM) dataset [60]. Detailed descriptions of these three datasets are provided below.

2.1.1. BJL Dataset

The study area of the BJL dataset is situated in Bijie, Guizhou, China, at altitudes ranging from 457 to 2900 m. This region is considered a high-risk area for landslides in China. The BJL dataset utilizes remote sensing imagery captured by the TripleSat satellite from May to August 2018 at a resolution of 0.8 m. It comprises 770 landslide images and 2003 other images. The dataset contains satellite optical images and labeled files.

2.1.2. L4S Dataset

The L4S dataset is derived from the Landslide4Sense competition, organized by the Institute of Advanced Research in Artificial Intelligence (IARAI), with data from different landslide-affected areas around the world. It includes Sentinel-2 multi-band data, DEM, and ALOS PAL-SAR slope data adjusted to 10 m resolution with pixel-level masks. The dataset is split into training, validation, and test sets. The training set comprises 3799 images sized at 128 × 128 pixels. In this study, only a training set of the L4S dataset is used as study data.

2.1.3. GVLM Dataset

The GVLM dataset is the first large-scale and open-source very-high-resolution landslide mapping dataset. This dataset comprises 17 sets of landslide sub-datasets from different geographical locations, each including a pair of dual-phase images with a spatial resolution of 0.59 m and the corresponding landslide mask data. With a total coverage area of 163.77 square kilometers, it spans extensive landslide areas across Asia, Africa, North America, South America, Europe, and Oceania.

2.2. Dataset Preparation

To facilitate model training and experimental analysis, we standardized the image size to 1024 × 1024 across all three datasets. Specifically, for the BJL dataset, where individual image sizes vary significantly from 1239 × 1197 to 61 × 61 and have differing aspect ratios, we adjusted them to 1024 × 1024 using equal scaling and zero padding. The image size in the L4S dataset was 128 × 128, so the images were scaled directly to 1024 × 1024 size, and negative sample data that did not contain landslide areas were removed. Because of the large size of individual images in the GVLM dataset, each set of image data and landslide mask is first cropped into multiple small images according to certain rules and then the small images are resized to a size of 1024 × 1024. The data volume of the three datasets processed above is recorded in Table 1, and their data visualization is shown in Figure 2.

Table 1. Datasets partition.

	BJL Dataset	**L4S Dataset**	**GVLM Dataset**
Training	583	1663	1977
Validation/Test	187	568	710

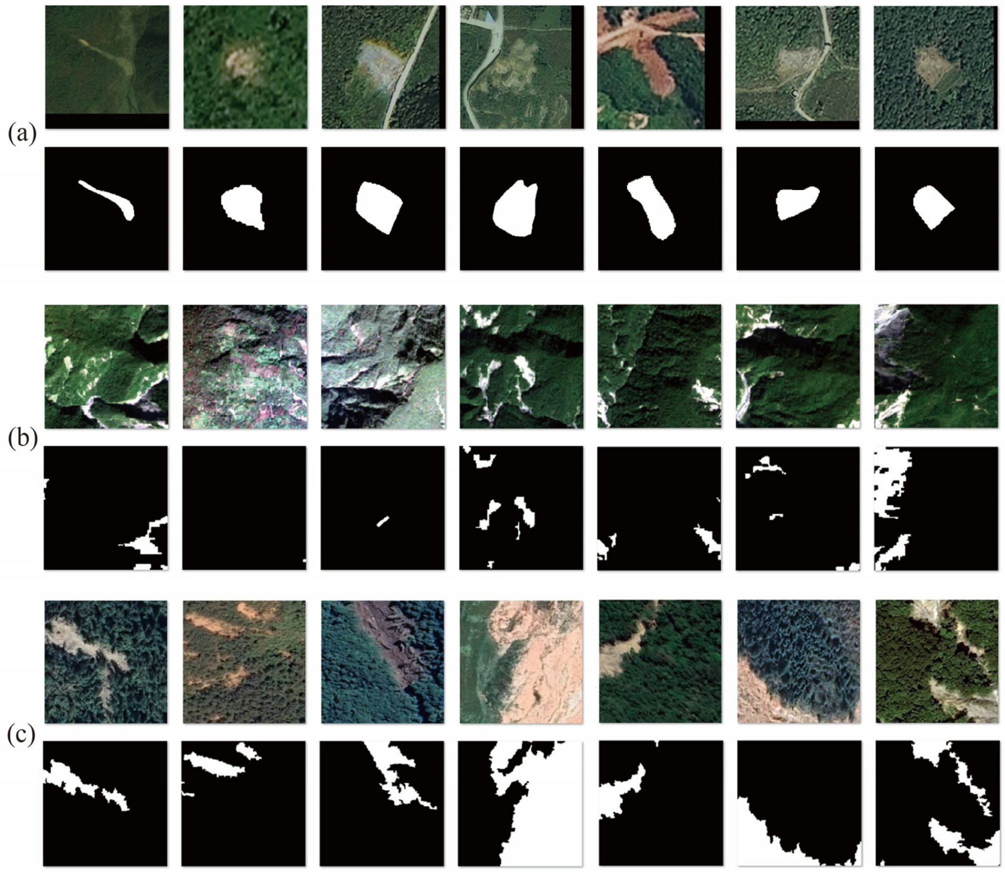

Figure 2. Visualization display of processed dataset: (**a**) BJL dataset, (**b**) L4S dataset, (**c**) GVLM dataset.

3. Methods

3.1. Framework

The SAM-CFFNet proposed in this study is an end-to-end network designed to extract landslide features from remote sensing images and output binary images representing landslide identification results. The structure of SAM-CFFNet is shown in Figure 3, which mainly consists of the image encoder ViT (IEViT) and the CFFD.

The IEViT encoder is tasked with extracting four levels of hierarchical deep features from input images of resolution 1024 × 1024. The CFFD, an adept decoder, is dedicated to integrating the multi-scale semantic features harvested by the IEViT to achieve refined recognition results. It employs the cross-feature fusion module (CFFM) to meticulously fine-tune and cross-fuse the extracted features, thereby amplifying information pertinent to landslide characteristics. These fused features are dimensionally reduced via convolutional layers before entering the Bottle ASPP module, designed to capture background context across disparate receptive fields. The outputs of this module are then upsampled to match the resolution of features processed by the secondary branch, the SFE. The branch, focused on capturing and refining texture details from the input image, leverages an attention module to selectively weigh the shallow features in relevance to the main pathway's deeper features. Finally, the deep features processed and the textured features from the SFE are concatenated. This concatenated feature map is then further upsampled to the original input resolution and passed through a final convolutional layer to produce the prediction output. The specific structure of each module is described in detail next.

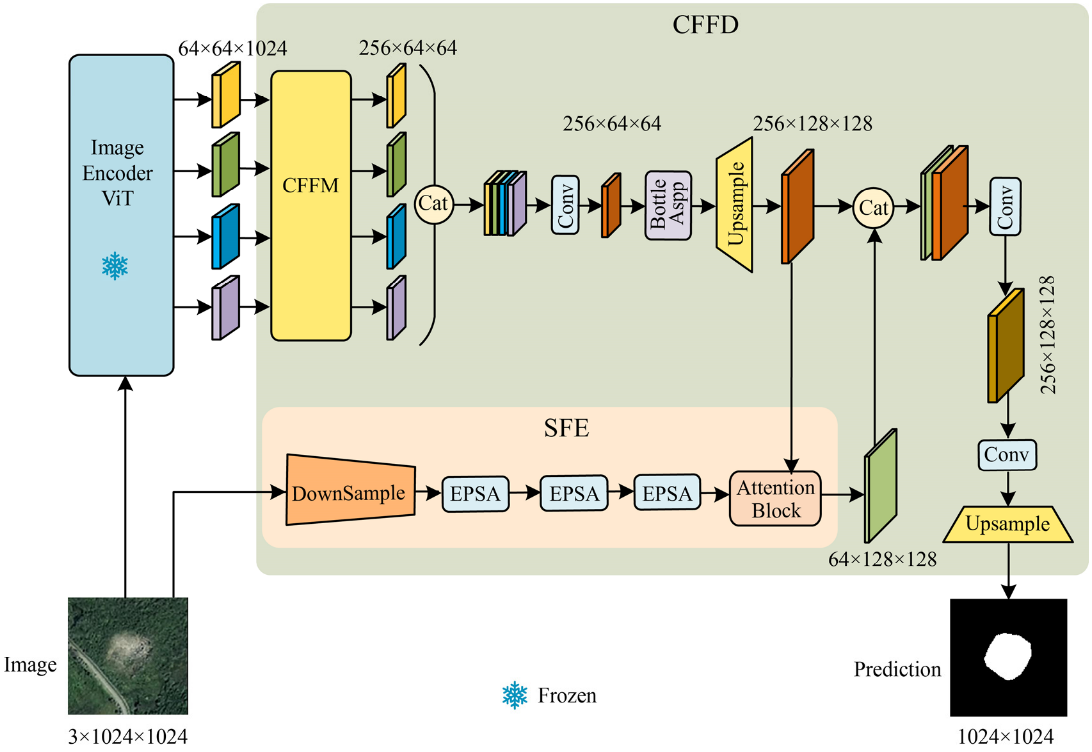

Figure 3. General overview of the SAM-CFFNet.

3.2. Image Encoder ViT

Our IEViT is built upon SAM's image encoder, specifically choosing the ViT-L version for its balanced performance and substantial parameter count while modifying it by removing the neck module positioned at the end of the model. The IEViT loaded pre-trained model weights that are publicly available on the official SAM website, maintaining the freeze on the entire IEViT module throughout the experiment.

As illustrated in Figure 4, the IEViT consists of patch embedding, position embedding, and transformer encoder components. Patch embedding is responsible for dividing the input image of size 1024 $\times$ 1024 into multiple patches of size 64 $\times$ 64. This is achieved by applying a convolutional operation with a kernel size of 16, a stride of 16, and no padding, resulting in patches of the desired size. The position embedding creates a zero tensor with dimensions corresponding to the patches and embedding dimensions. This tensor is then added to the patches via element-wise addition, thereby incorporating positional information into the patches. The output patches are then fed into the transformer encoder, which is the core component of the ViT and is responsible for processing serialized patches to learn global features of the image, efficiently capturing long-range dependencies and complex patterns in the image through the self-attention mechanism and stacking of MLPs. The transformer encoder consists of 24 transformer blocks, each of which maintains the same input and output dimensions and can therefore be used in series. An excessive number of transformer blocks can lead to the problem of information loss when transferring features between transformer blocks at different levels. Moreover, as an interactive visual base model, SAM's depth feature maps may not contain rich semantic information for specific categories. Therefore, to obtain more feature information related to landslides, we output the features of the 6th, 12th, 18th, and 24th transformer blocks. This is reasonable because, by outputting features at different levels, we can capture image representations at various levels, which helps to enhance the model's generalization ability, making it more suitable for various downstream tasks.

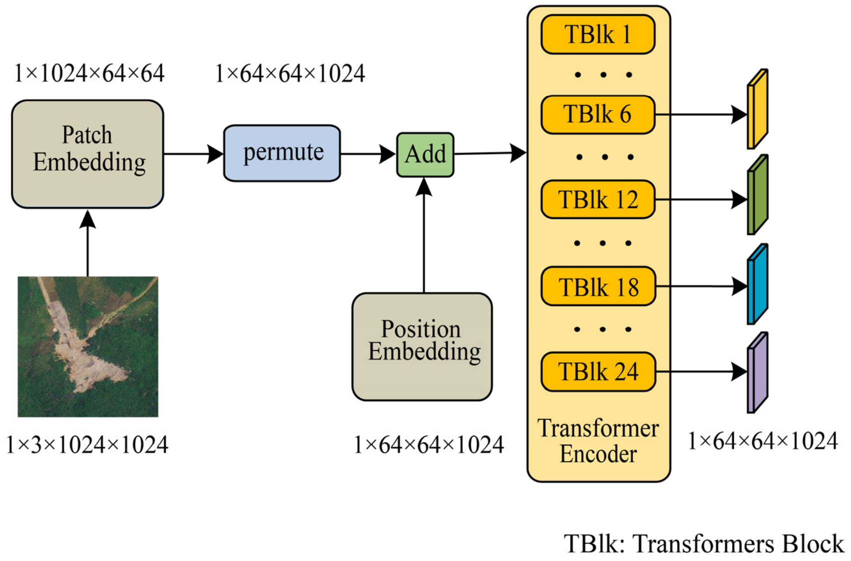

Figure 4. Structure of IEViT.

3.3. Cross-Feature Fusion Decoder

The structure of the CFFD is shown in Figure 3, where the focus will be on the three modules SFE, CFFM, and Bottle ASPP.

3.3.1. Shallow Feature Extractor

In the patch embedding of the IEViT, downsampling the image by a factor of 16 causes a loss of texture information in the model. To address this issue, we introduce the SFE. Comprising three convolutions with a stride of 2, three EPSA modules [61], and an attention block [62], the structure of the SFE is illustrated in Figure 3. The EPSA module, proposed by Zhang et al. [61], extracts fine-grained multi-scale spatial information and establishes long-distance dependencies. Meanwhile, the attention block, introduced by Oktay et al. [62], enhances feature representation by dynamically focusing on crucial features, thereby reducing irrelevant information.

The SFE uses three convolutions to downsample the input image by eight times and then utilizes three EPSA modules to extract shallow information. The attention block [62] suppresses the information in the shallow features that are unrelated to the main branch features in the CFFD, aiming to minimize information loss and confusion resulting from their fusion. The SFE aims to improve the model's ability to represent details and low-level features by supplementing shallow information and to strengthen the model's ability to capture image details and semantic information.

3.3.2. Cross-Feature Fusion Module

The CFFM consists of four feature adjustment modules (FAMs) and three feature cross-fusion structures (FCFSs), as shown in Figure 5. The four FAMs are, respectively, responsible for fine-tuning and resizing the four input features. The FCFS is responsible for the cross-fusion of the four features.

The structure of an FAM is shown in Figure 6. The FAM consists of two multi-layer perceptron (MLP) modules and a neck module. The MLP module contains two linear layers and an activation function. The two linear layers perform the downscaling and upscaling operations on the features, respectively, and this design reduces the number of parameters in the module. Connections are made outside the MLP using a residual network structure to reduce the loss of information from features. The channels of the features are permuted, and then the neck module is used to reduce the dimensionality of the features.

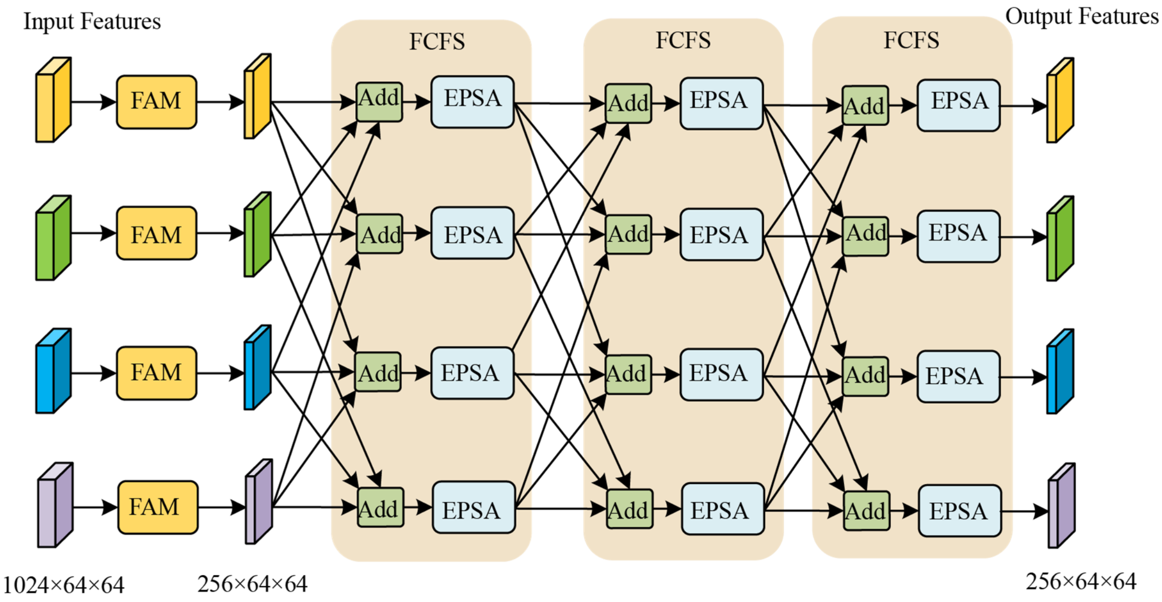

Figure 5. Structure of CFFM.

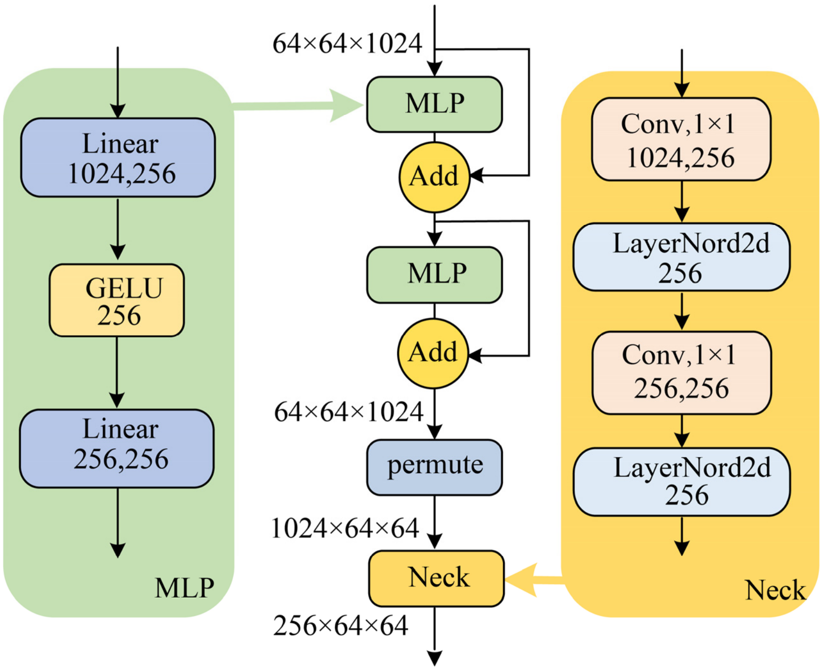

Figure 6. Structure of FAM.

The structure of an FCFS is shown in Figure 5. It can be observed that, within each FCFS module, the four input features are partitioned into four groups following the permutation rule C_4^3, where each group consists of three distinct features. Before being fed into the EPSA module, the features in each group undergo channel-wise summation. The CFFM consists of three FCFS submodules; thus, the above process is repeated three times. By utilizing the FCFS to cross-fuse features at different depths, it enables multi-level fusion of information, effectively enhancing the performance and generalization capability of the network.

3.3.3. Bottle ASPP

Building upon the ASPP [63], we introduced the Bottle ASPP inspired by the bottleneck structure, as illustrated in Figure 7. In the Bottle ASPP, the number of channels in the input features of the ASPP module is reduced using 1 × 1 convolutions and, consequently, the channel dimensions of the ASPP module output features are restored using 1 × 1 convolutions. Additionally, the output features are combined with the original input features through a residual structure. Compared to the original ASPP, the Bottle ASPP module reduces information loss and lowers parameter computation. For instance, when the input feature has 256 channels, the ASPP module has 2.13 MB of parameters while the Bottle ASPP has only 0.17 MB of parameters, resulting in a 92% reduction in parameters.

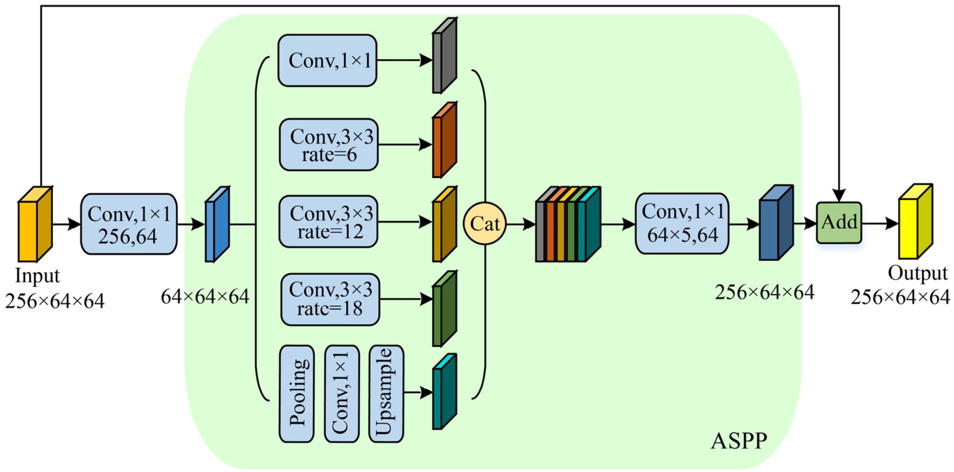

Figure 7. Structure of Bottle ASPP.

3.4. Evaluation Criterion

In this experiment, five performance metrics—precision, recall, F1-score, mean intersection over union (MIoU), and intersection over union (IoU) of landslide targets are used to compare and evaluate the proposed models, which are defined as shown below:

$$\text{Precison} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F1} - \text{score} = \frac{2 \times \text{Precison} \times \text{Recall}}{\text{Precison} + \text{Recall}} \tag{3}$$

$$\text{MIoU} = \frac{\frac{\text{TP}}{\text{TP+FN+FP}} + \frac{\text{TN}}{\text{TN+FN+FP}}}{2} \tag{4}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{5}$$

In the formula, TP, TN, FP, and FN represent the pixels that are correctly predicted as landslides, the pixels that are correctly predicted as non-landslides, the non-landslide pixels that are incorrectly predicted as landslides, and the landslide pixels that are incorrectly predicted as non-landslides, respectively.

3.5. Experimental Settings

When evaluating the performance of SAM-CFFNet in landslide recognition, we conducted comparative and ablation experiments on three landslide datasets. The detailed experimental designs for comparative and ablation experiments will be outlined in Section 4.

Given that the landslide recognition task is essentially a binary classification problem and the non-landslide background of the images in the experimental data occupies a large proportion, it is prone to small target detection problems. Therefore, we use the sum of binary cross-entropy loss and dice loss as the total loss function to train the model to maintain the stability and class balance of the model, and the formula of this loss function is shown below:

$$\mathcal{L} = 0.5 \times \mathcal{L}_B + 0.5 \times \mathcal{L}_D \tag{6}$$

where $\mathcal{L}_B$ denotes the binary cross-entropy loss and $\mathcal{L}_D$ denotes the dice loss, and $\mathcal{L}_B$ can be denoted as

$$\mathcal{L}_B = \frac{1}{N}\sum_i -[y_i \cdot \log(p_i) + (1 - y_i) \times \log(1 - p_i)] \tag{7}$$

$\mathcal{L}_D$ can be represented as

$$\mathcal{L}_D = 1 - \frac{1}{N}\sum_{i=1}^{N} \frac{2\sum_i^N y_i p_i}{\sum_i^N y_i + \sum_i^N p_i} \tag{8}$$

where N denotes the total number of samples, y_i and p_i denote the true label value and the predicted result value of the ith pixel point respectively.

Binary cross-entropy loss is widely used in binary classification and semantic segmentation for its stability and consistency. It quantifies prediction accuracy by comparing predicted probabilities with actual labels, demonstrating good robustness. Dice loss is effective in segmentation tasks and particularly handles class imbalances well. It measures overlap between predicted and truth regions, optimizing the intersection to ensure model sensitivity to object size and shape, resulting in better boundary depiction and more accurate segmentation.

Our experimental environment is based on the Debian operating system, developed using Python 3.7.12, and relies on PyTorch 1.11.0 with CUDA 11.3 for the development framework. Our computer is equipped with an Intel Xeon Gold 5218R processor (Intel Corporation, Santa Clara, CA, USA) and an NVIDIA A100 Tensor Core GPU (Nvidia Corporation, Santa Clara, CA, USA), along with 128 GB of operating memory. During the experiments, all models use stochastic gradient descent (ADAMW) as the optimizer, with an initial learning rate of 0.0002 and epochs set to 30 rounds. The batch size for the SAM-CFFNet and other SAM-based comparative models was set to 8, while the batch size for the other comparative models was set to 64.

4. Results

4.1. Results of Comparative Experiments

In comparative experiments, we selected four classical semantic segmentation models for comparative analysis, namely Attention U-Net [62], DeepLabv3+ [63], HRNet [64], and SegFormer [65]. These models have been widely used and studied in the field of landslide recognition [66–69] and have some structural similarities with the model proposed in this paper, so they can better demonstrate the advantages and features of SAM-based semantic segmentation technology over traditional semantic segmentation technology. A brief introduction to these models is as follows:

(1) Attention U-Net is designed based on the U-Net structure; it introduces attention gates to suppress irrelevant regions, focuses on useful salient features, and improves segmentation accuracy.

(2) DeepLabv3+ employs ASPP modules to process semantic information at different scales and improves the spatial accuracy of the segmentation results with a decoder module.

(3) HRNet, with its high-resolution feature and retention of global contextual information, excels in multi-scale information processing and small target, object, and boundary recognition, achieves excellent results in semantic segmentation tasks, and effectively improves segmentation accuracy.

(4) SegFormer, a semantic segmentation model based on the transformer architecture, achieves race-level performance in semantic segmentation tasks by introducing a self-attention mechanism to establish global dependencies between pixels.

The experimental results on the BJL dataset are recorded in Table 2, and the recognition results of each model are shown in Figure 8. SAM-CFFNet achieves optimal performance among all models, with an MIoU of 87.41%. In comparison, HRNet has an MIoU of 85.93% while SegFormer has an MIoU of only 76.44%. This is due to HRNet maintaining high-resolution details through parallel high- and low-resolution sub-network connections, while SegFormer loses shallow texture information with direct upsampling of deeper features. Compared with HRNet, SAM-CFFNet improves IoU by 2.64%, MIoU by 1.48%, and precision by 1.07%. In terms of visual effects, Attention U-Net and SegFormer struggle with precise contour delineation, exhibiting noticeable noise. In Figure 8IV, many models mistakenly classify the road as a landslide area, whereas SAM-CFFNet excels in accurately distinguishing between the road and landslide boundary, achieving results closer to the actual boundary. Remarkably, SAM-CFFNet excels in superior landslide recognition and precise contour delineation, outperforming others.

Table 2. Comparison results of different models on the BJL dataset.

Model	Precision (%)	Recall (%)	F1-Score (%)	MIoU (%)	IoU (%)
SegFormer	87.29	83.60	85.33	76.44	57.5
Attention U-Net	88.24	89.53	88.87	81.21	66.21
DeepLabv3+	90.53	87.62	89.01	81.44	66.4
HRNet	92.70	91.37	92.02	85.93	74.49
SAM-CFFNet	**93.77**	**92.18**	**92.96**	**87.41**	**77.13**

Bold values indicate optimal scores in each metric.

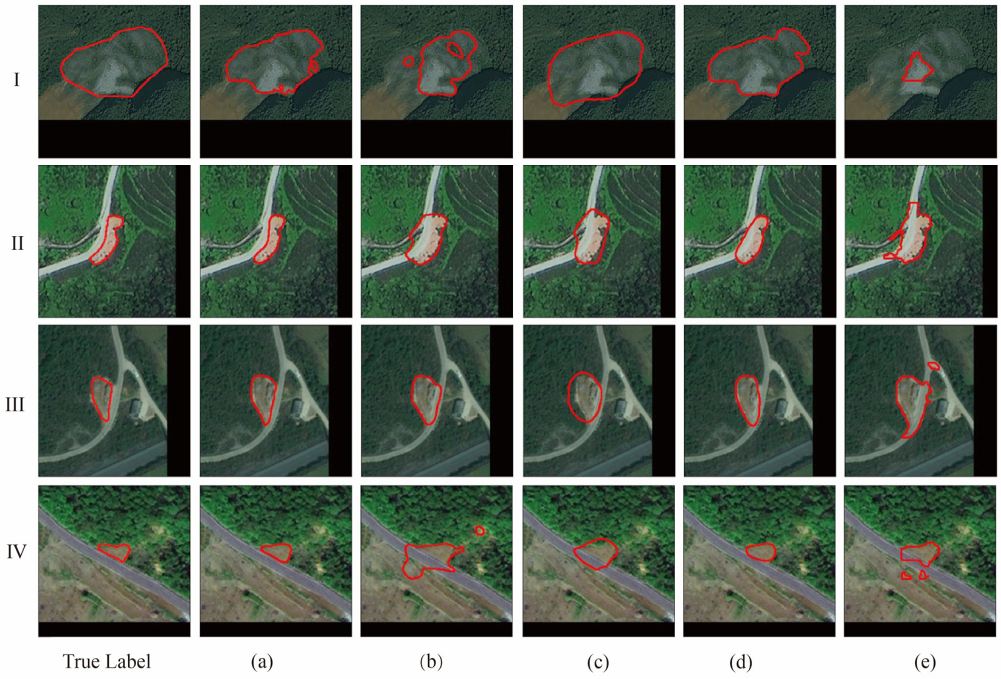

Figure 8. Visualization comparison of SAM-CFFNet with other models on the BJL dataset: (**a**) SAM-CFFNet, (**b**) Attention U-Net, (**c**) DeepLabv3+, (**d**) HRNet, and (**e**) SegFormer. I–IV are images randomly selected from the BJL dataset for visual comparison, the red orbit represents the landslide boundary.

The L4S dataset and the GVLM dataset were used in experiments to see how well the models could generalize and how robust they were. The results of the experiments are shown in Tables 3 and 4, and the recognition results for each model are shown in Figures 9 and 10. As can be seen from the two tables, SAM-CFFNet maintains its optimal performance, surpassing other models across most metrics. Particularly noteworthy is

SAM-CFFNet's significant superiority in MIoU and IoU, showcasing improvements of 2.19% and 0.81% for MIoU and 3.82% and 1.99% for IoU on both datasets when compared to Attention U-Net.

Table 3. Comparison results of different models on the L4S dataset.

Model	Precision (%)	Recall (%)	F1-Score (%)	MIoU (%)	IoU (%)
SegFormer	75.48	79.35	77.26	67.7	39.48
DeepLabv3+	83.16	80.71	81.88	72.75	48.41
HRNet	80.88	83.08	81.94	72.77	48.66
Attention U-Net	80.79	**86.06**	83.19	74.19	51.44
SAM-CFFNet	**85.29**	84.63	**84.96**	**76.38**	**55.26**

Bold values indicate optimal scores in each metric.

Table 4. Comparison results of different models on the GVLM dataset.

Model	Precision (%)	Recall (%)	F1-Score (%)	MIoU (%)	IoU (%)
SegFormer	87.59	84.4	85.87	76.44	61.86
DeepLabv3+	89.56	89.21	89.38	81.49	70.36
Attention U-Net	89.10	90.85	89.94	82.31	71.88
HRNet	90.30	90.21	90.21	82.84	72.51
SAM-CFFNet	**90.31**	**91.28**	**90.79**	**83.65**	**73.87**

Bold values indicate optimal scores in each metric.

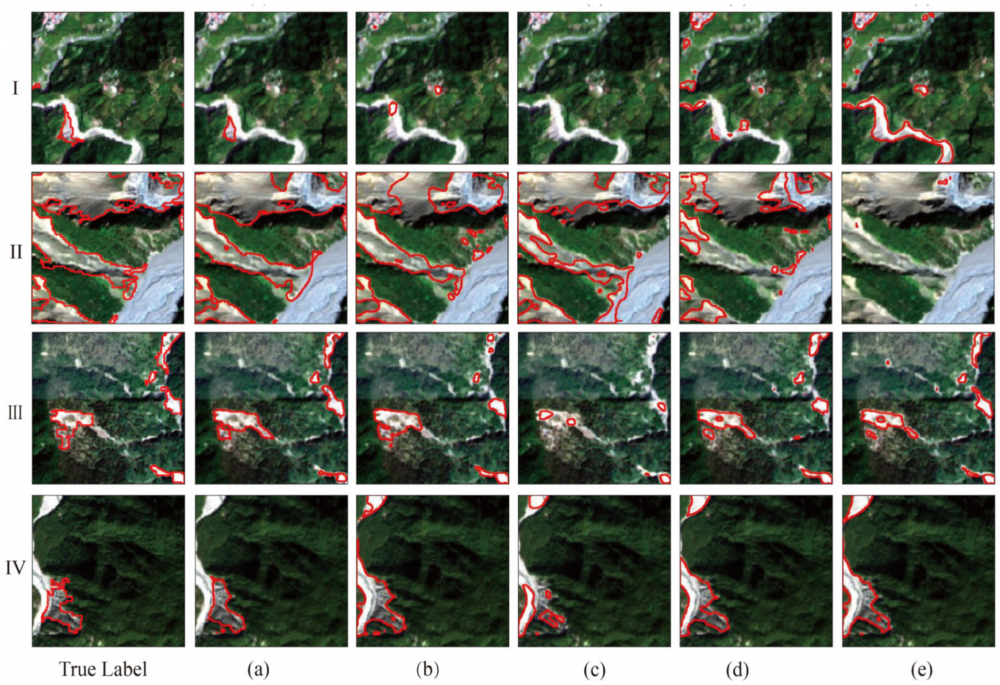

Figure 9. Visualization comparison of SAM-CFFNet with other models on the L4S dataset: (**a**) SAM-CFFNet, (**b**) Attention U-Net, (**c**) DeepLabv3+, (**d**) HRNet, and (**e**) SegFormer. I–IV are images randomly selected from the L4S dataset for visual comparison, the red orbit represents the landslide boundary.

In Figure 9, due to the low spatial resolution of the L4S dataset, Attention U-Net, HRNet, and SegFormer have obvious noise problems on this dataset, and there are more misclassified landslide fragments in the recognition results. SAM-CFFNet, on the other hand, performs better and has a higher recognition rate for landslides. In Figure 10, most of the models have good recognition results on the GVLM dataset. But, when it comes to segmentation details, SAM-CFFNet can correctly identify the clear edges of the landslides while the other models have trouble differentiating between features that look very similar to the landslides. In Figure 10IV, all models except SAM-CFFNet misidentify roads as landslides. This indicates that SAM-CFFNet has excellent generalization ability and robustness on different datasets and can effectively reduce the misclassification rate on lower-resolution images.

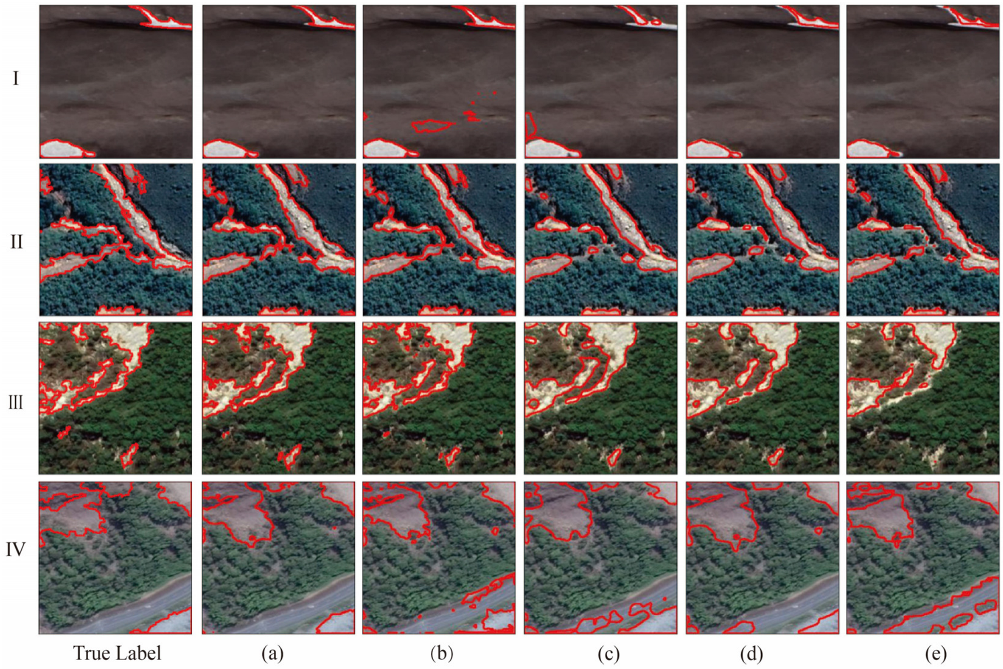

Figure 10. Visualization comparison of SAM-CFFNet with other models on the GVLM dataset: (**a**) SAM-CFFNet, (**b**) Attention U-Net, (**c**) DeepLabv3+, (**d**) HRNet, and (**e**) SegFormer. I–IV are images randomly selected from the GVLM dataset for visual comparison, the red orbit represents the landslide boundary.

4.2. Results of Ablation Experiments

To gain insights into the impact of different components within SAM-CFFNet on the overall model performance, we conducted multiple sets of ablation experiments on the three datasets. In these experiments, we checked what happens to the model's performance when there are different numbers of FCFS settings, Bottle ASPP, and SFEs.

The results of the ablation experiments are documented in Table 5, where, on three datasets, a change in the number of FCFSs leads to a significant decrease in the model performance. For the BJL dataset, IoU decreases by 1.12% when the number of FCFSs is four, while IoU decreases by 3.24% and 3.29% when the number of FCFSs is one and two, respectively. For the L4S dataset and GVLM dataset, Bottle ASPP is the module that has the greatest impact on the model performance, and when Bottle ASPP is removed, the model's IoU decreases by 3.08% and 3.49% for the L4S dataset and GVLM dataset, respectively, and decreases by 1.17% for the BJL dataset.

Figure 11 shows the heat map of the SFE output features, and the features extracted by the SFE have rich texture and edge information, which helps to improve the model's recognition accuracy of the landslide boundary. On the BJL dataset, the model performance degradation after removing the SFE module is small, and the IoU is reduced by 0.36%. However, on the L4S dataset and GVLM dataset, the performance degradation is larger, with 1.88% and 2.93% IoU reductions, respectively. This is because the landslide background of the L4S dataset and GVLM dataset is relatively more complex compared to the BJL dataset, and rich shallow texture information is more needed to improve the segmentation accuracy.

Table 5. Quantitative results of SAM-CFFNet component ablation.

Number of FCFSs	Bottle ASPP	SFE	Recall (%) B	Recall (%) L	Recall (%) G	F1-Score (%) B	F1-Score (%) L	F1-Score (%) G	IoU(%) B	IoU(%) L	IoU(%) G
3	√	√	**92.18**	**84.63**	**91.28**	**92.96**	**84.96**	**90.79**	**77.13**	**55.26**	**73.87**
1	√	√	+0.62	−0.35	−1.15	−1.18	−0.45	−1.00	−3.24	−1.03	−2.45
2	√	√	−0.01	−0.36	−1.41	−1.19	−1.02	−0.46	−3.29	−2.30	−1.28
4	√	√	+0.13	−1.73	−0.66	−0.40	−0.58	−0.54	−1.12	−1.38	−1.34
3	×	√	+0.06	−0.59	−2.04	−0.42	−1.38	−1.40	−1.17	−3.08	−3.49
3	√	×	+1.16	−2.15	−1.12	−0.15	−0.80	−1.22	−0.36	−1.88	−2.93

The SAM-CFFNet results are bolded, with + and − indicating metric increase or decrease compared to it. B is BJL dataset, L is L4S dataset, and G is GVLM dataset. ×denotes removing this module, while √ denotes retention.

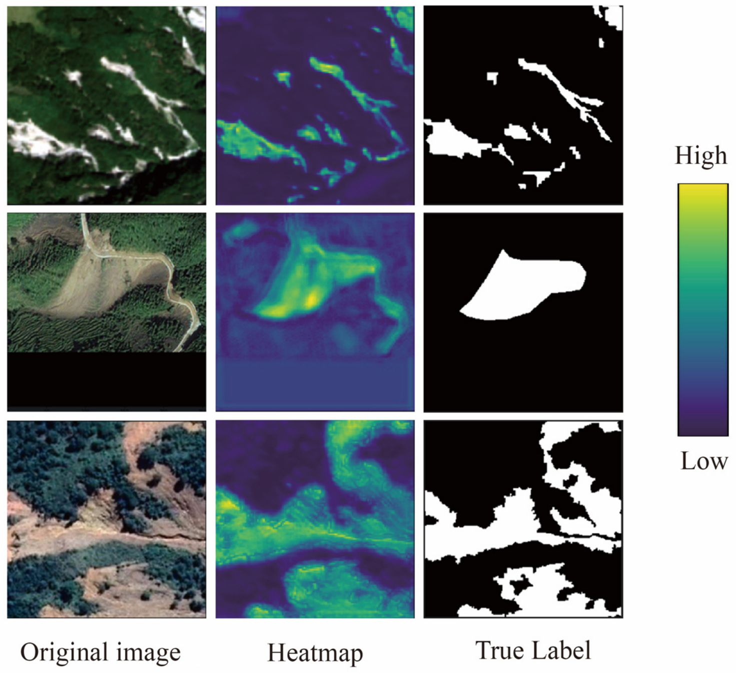

Figure 11. Heat map display of SFE extraction results.

5. Discussion

5.1. Comparison of Different Decoders

Experiments in Sections 4.1 and 4.2 show that our SAM-CFFNet is effective for remote sensing landslide identification, and our model achieves the best accuracy in all three datasets, which shows that the model has good robustness and excellent segmentation performance for landslide data of different types and regions. The superior performance of SAM-CFFNet is not only due to the powerful IEViT but also because the CFFD plays a crucial role. The CFFD is our specially designed decoder that accommodates the IEViT.

To showcase the adaptability of the CFFD to the IEViT, we formulated four models utilizing the frozen IEViT as the encoder. Each model features a distinct decoder architecture, including the mask decoder, PSP decoder [70], ASPP decoder [63], and LawinASPP [71], denoted as Model I, Model II, Model III, and Model IV, respectively. The mask decoder comes from SAM, and since there is no prompt encoder in Model I, the mask decoder is only responsible for processing the output of the image encoder. The PSP decoder and the ASPP decoder are both commonly used decoders for semantic segmentation. LawinASPP is a novel semantic segmentation ViT decoder that can capture rich contextual information at multiple scales through large window attention.

We tested the above models as well as SAM-CFFNet on three datasets, and the experimental results are recorded in Tables 6–8 and the visualization results of the models are shown in Figure 12. Our SAM-CFFNet obtains the best F1-scores, IoU, and MIoU on all three datasets, outperforming the other models, and the MIoU of SAM-CFFNet is generally higher than the other models by more than 2.6%. Model I performs poorly, with the lowest accuracy on both the BJL dataset and the L4S dataset. Model IV performs second only to SAM-CFFNet on all three datasets. LawinASPP demonstrates superior adaptation to the ViT encoder compared to Model I–III decoders. Figure 12 illustrates that SAM-CFFNet achieves a higher accuracy and better alignment with real labels on the BJL dataset, outperforming other models. Similarly, SAM-CFFNet exhibits improved recognition results on the L4S dataset and GVLM dataset, recognizing closer to the true labels compared to the other models.

Table 6. Comparison results of models with different decoders on the BJL dataset.

Model	Decoder	Precision (%)	Recall (%)	F1-Score (%)	MIoU (%)	IoU (%)
Model I	mask decoder	90.02	88.41	89.20	81.70	66.93
Model II	PSP decoder	91.21	90.26	90.73	83.95	70.97
Model III	ASPP decoder	**94.48**	88.97	91.51	85.15	72.94
Model IV	LawinASPP	91.23	**93.70**	92.42	86.54	75.71
SAM-CFFNet	CFFD	93.77	92.18	**92.96**	**87.41**	**77.13**

Bold values indicate optimal scores in each metric.

Table 7. Comparison results of models with different decoders on the L4S dataset.

Model	Decoder	Precision (%)	Recall (%)	F1-Score (%)	MIoU (%)	IoU (%)
Model I	mask decoder	79.82	77.65	78.69	69.27	41.96
Model II	PSP decoder	82.22	77.40	79.60	70.26	43.67
Model III	ASPP decoder	83.09	80.56	81.77	72.62	48.17
Model IV	LawinASPP	83.98	83.62	83.80	74.97	52.64
SAM-CFFNet	CFFD	**85.29**	**84.63**	**84.96**	**76.38**	**55.26**

Bold values indicate optimal scores in each metric.

Table 8. Comparison results of models with different decoders on the GVLM dataset.

Model	Decoder	Precision (%)	Recall (%)	F1-Score (%)	MIoU (%)	IoU (%)
Model II	PSP decoder	80.74	73.57	76.31	64.55	42.9
Model I	mask decoder	79.78	76.97	78.25	66.57	46.99
Model III	ASPP decoder	88.86	88.39	88.62	80.34	68.54
Model IV	LawinASPP	89.87	88.28	89.05	80.99	69.46
SAM-CFFNet	CFFD	**90.31**	**91.28**	**90.79**	**83.65**	**73.87**

Bold values indicate optimal scores in each metric.

The poor performance of Model I shows that it is not feasible to directly fine-tune SAM's mask decoder for landslide identification tasks. The performance of Model II-IV, on the other hand, shows that excellent performance cannot be achieved by directly using the decoders of other models without modifying and adapting them for specific tasks and specific encoders. Our design of the CFFD fully considers the characteristics of the IEViT as well as the requirements of the remote sensing landslide identification task, which can better enhance the landslide features, suppress the noise, and obtain more accurate landslide boundaries.

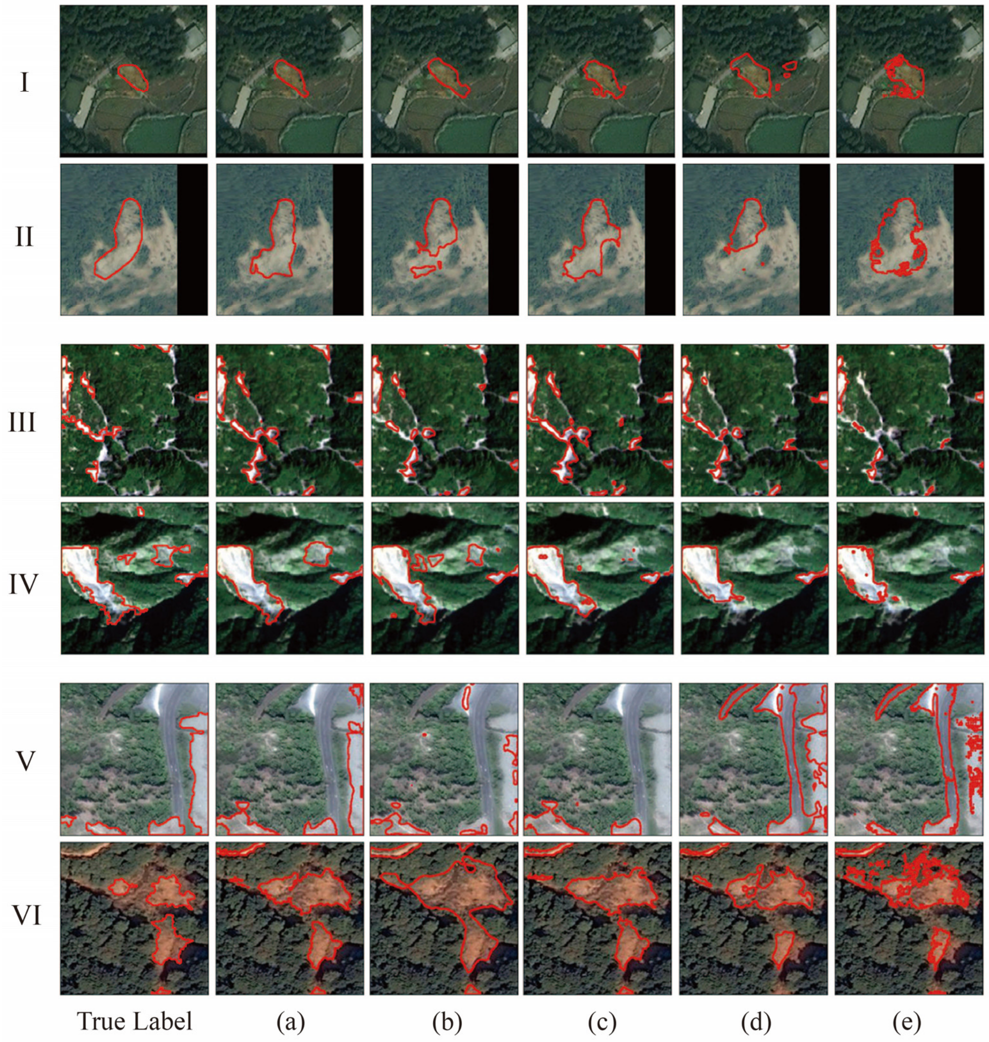

Figure 12. Visual comparison of SAM-CFFNet with other image decoder ViT-based models on three datasets: (**a**) SAM-CFFNet, (**b**) Model I, (**c**) Model II, (**d**) Model III, (**e**) Model IV; I–VI are images randomly selected from the test sample for visual comparison, where I and II are from the BJL dataset, III and V are from the L4S dataset, and V and VI are from the GVLM dataset, the red orbit represents the landslide boundary.

5.2. Model Advantage Comparison

In our prior investigations, we extensively verified the exceptional segmentation accuracy of SAM-CFFNet in landslide recognition tasks. However, assessing model performance and practicality also requires considering computational efficiency and resource consumption. In Table 9, we present a comparison of SAM-CFFNet with other models on the GVLM dataset in terms of the total number of parameters, trainable parameters, FLOPs, and accuracy metrics. It is important to highlight that the accuracy metrics are calculated based on the GVLM dataset, and all models use images of size 1024 × 1024 for FLOP computation.

The model structure is pivotal in enhancing the precision of recognition tasks. In models with more than 81% of MIoU, Attention U-Net boosts the detection of essential features through an innovative combination of skip connections and attention mechanisms. Similarly, DeepLabv3+ refines segmentation accuracy by integrating the ASPP module with techniques that enhance shallow features. HRNet stands out by utilizing a cross-feature fusion strategy that maintains high-resolution feature maps. This approach optimizes the model's detail capture and leads to superior segmentation accuracy. Building on these strengths, our SAM-CFFNet adopts the advanced transformer structure IEViT to extract features, combined with cross-feature fusion and shallow feature enhancement techniques that not only preserve detail information in depth but also fuse different levels of features effectively. In addition, the processing capability of multi-scale information is further enhanced by introducing ASPP. This series of comprehensive structural designs enhances the proposed model's performance in complex landslide identification scenarios.

Table 9. Comparison results of various models for landslide detection on the GVLM dataset.

Model	Structural Features	Complexity			Performance		
		Total Params (MB)	Trainable Params (MB)	FLOPs (G)	Recall (%)	F1-Score (%)	MIoU (%)
Attention U-Net	Skip connection + Attention	33.26	33.26	992.89	90.85	89.94	82.31
DeepLabv3+	ASPP + Shallow feature enhancement	5.54	5.54	98.46	89.21	89.38	81.49
HRNet	Cross-feature fusion	9.19	9.19	69.49	90.21	90.21	82.84
SegFormer	Transformer	7.36	7.36	48.77	84.4	85.87	76.44
Model I	Transformer	297.87	3.87	1224.58	76.97	78.25	66.57
Model II	Transformer + Pyramid pooling	297.92	3.93	1280.95	73.57	76.31	64.55
Model III	Transformer + ASPP	297.56	3.56	1250.74	88.39	88.62	80.34
Model IV	Transformer + Large window attention	302.66	8.66	1251.07	88.28	89.05	80.99
SAM-CFFNet	Transformer + Attention + ASPP + Shallow feature enhancement + Cross-feature fusion	298.06	4.06	1282.04	91.28	90.79	83.65

For model training efficiency, SAM-CFFNet and Model I-IV use the IEViT as the encoder, so the total number of parameters of the model is larger than 297 MB, and the FLOPs are larger than 1220 G, which is much larger than the other models. In terms of trainable parameters, SAM-CFFNet has relatively few (4.06 MB), and Model I-III perform poorly in terms of accuracy, although the trainable parameters are lower than SAM-CFFNet. The training loss curves of these models for 30 epochs of training on the three datasets are recorded in Figure 13, and it can be seen that the fitting speeds of these models, SAM-CFFNet and Model I–IV, are generally faster than the other types of models, and SAM-CFFNet is able to complete the fitting within 10 epochs in all three datasets with smooth curves.

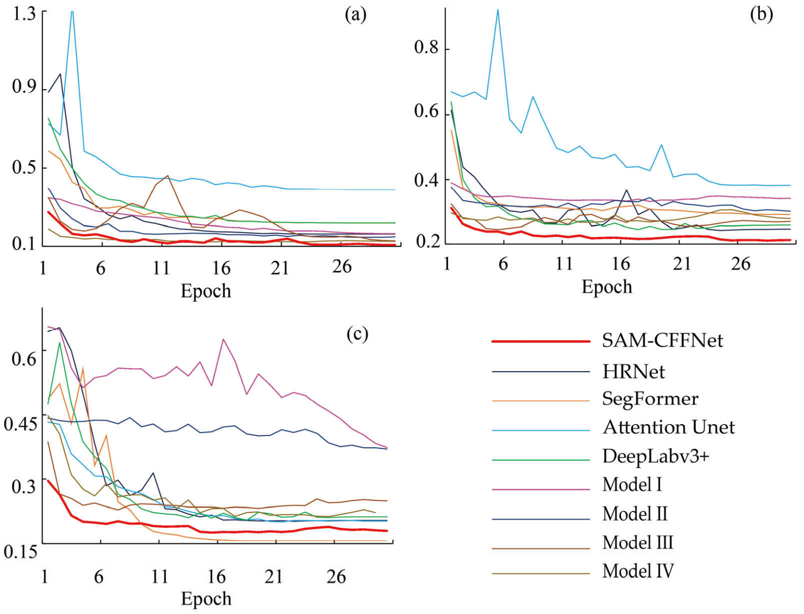

Figure 13. The training process curves of loss for different landslide detection models on three datasets: (**a**) BJL dataset, (**b**) L4S dataset, and (**c**) GVLM dataset.

In summary, SAM-CFFNet demonstrates notable structural and performance benefits for landslide identification tasks. The model's integration of the advanced IEViT, cross-feature fusion, and ASPP modules allows for efficient feature integration across levels, deep retention of detailed information, and strong multi-scale information processing capabilities. Moreover, SAM-CFFNet offers clear advantages in training efficiency, with a large parameter count yet a relatively low number of trainable parameters, coupled with high stability. However, its substantial parameter and FLOP count does increase training costs to some extent, necessitating a GPU with ample memory for effective training.

5.3. Limitations

This study explores the performance of SAM-CFFNet in landslide detection tasks from various perspectives, yet there are still some shortcomings and areas for improvement.

(1) Although SAM-CFFNet excelled in landslide detection, it sacrificed SAM's interactive segmentation and generalization capabilities in other image domains. In future work, we will explore how the model can be adapted to specific downstream tasks while still retaining the interactive segmentation features and generalization capabilities of SAM.

(2) SAM-CFFNet currently exclusively relies on optical remote sensing images and does not incorporate geospatial data such as DEM, geological data, and rainfall data. In the future, we aim to integrate these multi-source data to explore the potential of the SAM model in multi-source data fusion and cross-disciplinary applications to comprehensively address the challenges and complexities of geological disaster identification.

(3) While SAM-CFFNet demonstrates excellent performance, its large total number of parameters leads to high training costs and makes it challenging to deploy on small-scale devices. Therefore, future research will focus on further optimizing the model structure to achieve model lightweighting, enhancing its versatility and flexibility.

6. Conclusions

In this study, SAM-CFFNet is proposed as a novel and effective application of SAM. The objective is to improve the landslide recognition accuracy using SAM and to address its performance degradation and dependence on prompt information in the task of landslide recognition from remote sensing images. Notably, our specially designed CFFD effectively improves the model's adaptability for downstream tasks. During the training process, the IEViT reads the pre-training weights and keeps them frozen, and this strategy fully utilizes the powerful feature extraction capability of SAM. This effectively improves the convergence speed and training efficiency of the model and enhances its generalization ability and adaptability on the landslide identification task.

We train and validate SAM-CFFNet against several other reference models on three landslide datasets and evaluate the model's effectiveness in recognizing landslides using precision, recall, F1-score, MIoU, and IoU. Our results show that SAM-CFFNet performs optimally in terms of accuracy on all three landslide datasets, significantly outperforming the other compared models. SAM-CFFNet demonstrates excellent generalization ability and robustness on different datasets. Furthermore, we substantiated the rationale behind our designed CFFD through comparative analysis with various decoders. Additionally, we deliberated on the model's training efficiency and outlined forthcoming research directions.

The results of this study highlight the excellent performance of SAM-CFFNet in landslide identification tasks and the importance of this model in assessing the impact of landslides after a disaster as well as in guiding post-disaster reconstruction efforts. The SAM-based model represented by SAM-CFFNet shows great potential in the field of landslide detection and monitoring, and the insights gained from this study will help to promote the further development of SAM-based models in the field of geohazard monitoring.

Author Contributions: Conceptualization, J.Y. and L.X.; methodology, J.Y. and L.X.; software, L.X.; validation, L.X. and Y.P.; formal analysis, P.Z.; investigation, L.X. and Y.P.; resources, Y.L.; data curation, Y.C. and Y.D.; writing—original draft preparation, L.X.; writing—review and editing, J.Y.; visualization, L.X.; supervision, C.H.; project administration, D.G.; funding acquisition, D.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the National Key Research and Development Program of China (2021YFC3000400).

Data Availability Statement: The BJL dataset, L4S dataset, and GVLM dataset can be openly accessed at: http://gpcv.whu.edu.cn/data/Bijie_pages.html (accessed on 10 December 2023), https://github.com/iarai/Landslide4Sense-2022 (accessed on 10 December 2023), and https://github.com/zxk688/GVLM (accessed on 10 December 2023), respectively. The code for this experiment has been posted at: https://github.com/JunchuanYu/SAM-CFFNet (accessed on 20 May 2024).

Acknowledgments: We would like to express our heartfelt gratitude to the researchers who created and shared the open-source landslide datasets used in this study. Additionally, we extend our heartfelt thanks to Heywhale, who provided the arithmetic platform for this work.

Conflicts of Interest: The authors declare no conflicts of interests.

References

1. Zhang, J.; Cui, Q.; Ma, X. Deep Evidential Remote Sensing Landslide Image Classification with a New Divergence, Multiscale Saliency and an Improved Three-Branched Fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 3799–3820. [CrossRef]
2. Zhou, S.Y.; Gao, L.; Zhang, L.M. Predicting debris-flow clusters under extreme rainstorms: A case study on Hong Kong Island. *Bull. Eng. Geol. Environ.* **2019**, *78*, 5775–5794. [CrossRef]
3. Iverson, R.M. Landslide Triggering by Rain Infiltration. *Water Resour. Res.* **2000**, *36*, 1897–1910. [CrossRef]
4. Sato, H.; Hasegawa, H.; Fujiwara, S.; Tobita, M.; Koarai, M.; Une, H.; Iwahashi, J. Interpretation of Landslide Distribution Triggered by the 2005 Northern Pakistan Earthquake Using SPOT 5 Imagery. *Landslides* **2007**, *4*, 113–122. [CrossRef]
5. Qiang, X.U.; Xiujun, D.; Weile, L.I. Integrated Space-Air-Ground Early Detection, Monitoring and Warning System for Potential Catastrophic Geohazards. *Geomat. Inf. Sci. Wuhan. Univ.* **2019**, *44*, 957–966.
6. Carrión-Mero, P.; Montalván-Burbano, N.; Morante-Carballo, F.; Quesada-Román, A.; Apolo-Masache, B. Worldwide Research Trends in Landslide Science. *Int. J. Environ. Res. Public Health* **2021**, *18*, 9445. [CrossRef] [PubMed]
7. Samia, J.; Temme, A.; Bregt, A.; Wallinga, J.; Guzzetti, F.; Ardizzone, F.; Rossi, M. Do landslides follow landslides? Insights in path dependency from a multi-temporal landslide inventory. *Landslides* **2017**, *14*, 547–558. [CrossRef]
8. Kamp, U.; Growley, B.; Khattak, G.; Owen, L. GIS-Based Landslide Susceptibility Mapping for the 2005 Kashmir Earthquake Region. *Geomorphology* **2008**, *101*, 631–642. [CrossRef]
9. Antoine, R.; Lopez, T.; Tanguy, M.; Lissak, C.; Gailler, L.-S.; Labazuy, P.; Fauchard, C. Geoscientists in the Sky: Unmanned Aerial Vehicles Responding to Geohazards. *Surv. Geophys.* **2020**, *41*, 1285–1321. [CrossRef]
10. Lu, H.; Ma, L.; Fu, X.; Liu, C.; Wang, Z.; Tang, M.; Li, N. Landslides Information Extraction Using Object-Oriented Image Analysis Paradigm Based on Deep Learning and Transfer Learning. *Remote Sens.* **2020**, *12*, 752. [CrossRef]
11. Ullo, S.L.; Mohan, A.; Sebastianelli, A.; Ahamed, S.E.; Kumar, B.; Dwivedi, R.; Sinha, G.R. A New Mask R-CNN-Based Method for Improved Landslide Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3799–3810. [CrossRef]
12. Ghorbanzadeh, O.; Meena, S.R.; Blaschke, T.; Aryal, J. UAV-Based Slope Failure Detection Using Deep-Learning Convolutional Neural Networks. *Remote Sens.* **2019**, *11*, 2046. [CrossRef]
13. Dong, J.; Niu, R.; Li, B.; Xu, H.; Wang, S. Potential Landslides Identification Based on Temporal and Spatial Filtering of SBAS-InSAR Results. *Geomat. Nat. Hazards Risk* **2023**, *14*, 52–75. [CrossRef]
14. Fang, K.; Tang, H.; Li, C.; Su, X.; An, P.; Sun, S. Centrifuge modelling of landslides and landslide hazard mitigation: A review. *Geosci. Front.* **2022**, *14*, 101493. [CrossRef]
15. Dai, K.; Feng, Y.; Zhuo, G.; Tie, Y.; Deng, J.; Balz, T.; Li, Z. Applicability Analysis of Potential Landslide Identification by InSAR in Alpine-Canyon Terrain—Case Study on Yalong River. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 2110–2118. [CrossRef]
16. Bhuyan, K.; Tanyaş, H.; Nava, L.; Puliero, S.; Meena, S.R.; Floris, M.; van Westen, C.; Catani, F. Generating Multi-Temporal Landslide Inventories through a General Deep Transfer Learning Strategy Using HR EO Data. *Sci. Rep.* **2023**, *13*, 162. [CrossRef] [PubMed]
17. Shao, X.; Xu, C. Earthquake-Induced Landslides Susceptibility Assessment: A Review of the State-of-the-Art. *Nat. Hazards Res.* **2022**, *2*, 172–182. [CrossRef]
18. Catani, F. Landslide Detection by Deep Learning of Non-Nadiral and Crowdsourced Optical Images. *Landslides* **2020**, *18*, 1025–1044. [CrossRef]
19. Scardigli, A.; Risser, L.; Haddouche, C.; Jatiault, R. Integrating Unordered Time Frames in Neural Networks: Application to the Detection of Natural Oil Slicks in Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4202914. [CrossRef]

20. Chen, Y.; Wei, Y.; Wang, Q.; Chen, F.; Lu, C.; Lei, S. Mapping Post-Earthquake Landslide Susceptibility: A U-Net Like Approach. *Remote Sens.* **2020**, *12*, 2767. [CrossRef]
21. Dao, D.V.; Jaafari, A.; Bayat, M.; Mafi-Gholami, D.; Qi, C.; Moayedi, H.; Phong, T.V.; Ly, H.-B.; Le, T.-T.; Trinh, P.T.; et al. A Spatially Explicit Deep Learning Neural Network Model for the Prediction of Landslide Susceptibility. *Catena* **2020**, *188*, 104451. [CrossRef]
22. Li, S.; Hua, H. Automatic Recognition of Landslides Based on Change Detection. In Proceedings of the International Symposium on Photoelectronic Detection and Imaging 2009, Beijing, China, 17–19 June 2009.
23. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
24. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. Yolact: Real-Time Instance Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9157–9166.
25. Chen, K.; Li, W.; Lei, S.; Chen, J.; Jiang, X.; Zou, Z.; Shi, Z.X. Continuous Remote Sensing Image Super-Resolution Based on Context Interaction in Implicit Function Space. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4702216. [CrossRef]
26. Ying, H.; Huang, Z.; Liu, S.; Shao, T.; Zhou, K. EmbedMask: Embedding Coupling for Instance Segmentation. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), Montreal, QC, Canada, 19–27 August 2021.
27. Zhang, L.; Zhang, L.; Du, B. Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [CrossRef]
28. Zhao, Z.-Q.; Zheng, P.; Xu, S.-T.; Wu, X. Object Detection with Deep Learning: A Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef] [PubMed]
29. Längkvist, M.; Kiselev, A.; Alirezaie, M.; Loutfi, A. Classification and Segmentation of Satellite Orthoimagery Using Convolutional Neural Networks. *Remote Sens.* **2016**, *8*, 329. [CrossRef]
30. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef] [PubMed]
31. Radovic, M.; Adarkwa, O.; Wang, Q. Object Recognition in Aerial Images Using Convolutional Neural Networks. *J. Imaging* **2017**, *3*, 21. [CrossRef]
32. Ji, S.; Dawen, Y.; Shen, C.; Li, W.; Xu, Q. Landslide Detection from an Open Satellite Imagery and Digital Elevation Model Dataset Using Attention Boosted Convolutional Neural Networks. *Landslides* **2020**, *17*, 1337–1352. [CrossRef]
33. Ghorbanzadeh, O.; Blaschke, T.; Gholamnia, K.; Meena, S.R.; Tiede, D.; Aryal, J. Evaluation of Different Machine Learning Methods and Deep-Learning Convolutional Neural Networks for Landslide Detection. *Remote Sens.* **2019**, *11*, 196. [CrossRef]
34. Yu, B.; Chen, F.; Xu, C. Landslide Detection Based on Contour-Based Deep Learning Framework in Case of National Scale of Nepal in 2015. *Comput. Geosci.* **2020**, *135*, 104388. [CrossRef]
35. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Stateline, NV, USA, 3–8 December 2012; pp. 1097–1105.
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *arXiv* **2015**, arXiv:1512.03385.
37. Soares, L.P.; Dias, H.C.; Grohmann, C.H. Landslide Segmentation with U-Net: Evaluating Different Sampling Methods and Patch Sizes. *arXiv* **2020**, arXiv:2007.06672.
38. H Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
39. Qin, H.; Wang, J.; Mao, X.; Zhao, Z.; Gao, X.; Lu, W. An Improved Faster R-CNN Method for Landslide Detection in Remote Sensing Images. *J. Geovisualization Spat. Anal.* **2023**, *8*, 2. [CrossRef]
40. Cheng, L.; Li, J.; Duan, P.; Wang, M. A small attentional YOLO model for landslide detection from satellite remote sensing images. *Landslides* **2021**, *18*, 2751–2765. [CrossRef]
41. Liu, Q.; Wu, T.; Deng, Y.; Liu, Z. SE-YOLOv7 Landslide Detection Algorithm Based on Attention Mechanism and Improved Loss Function. *Land* **2023**, *12*, 1522. [CrossRef]
42. Li, Y.; Ding, M.; Zhang, Q.; Luo, Z.; Huang, W.; Zhang, C.; Jiang, H. Old Landslide Detection Using Optical Remote Sensing Images Based on Improved YOLOv8. *Appl. Sci.* **2024**, *14*, 1100. [CrossRef]
43. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
44. Chen, K.; Zou, Z.; Shi, Z. Building Extraction from Remote Sensing Images with Sparse Token Transformers. *Remote Sens.* **2021**, *13*, 4441. [CrossRef]
45. Wang, L.; Fang, S.; Li, R.; Meng, X. Building extraction with vision transformer. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5625711. [CrossRef]
46. Huang, Y.; Zhang, J.; He, H.; Jia, Y.; Chen, R.; Ge, Y.; Ming, Z.; Zhang, L.; Li, H. MAST: An Earthquake-Triggered Landslides Extraction Method Combining Morphological Analysis Edge Recognition with Swin-Transformer Deep Learning Model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 2586–2595. [CrossRef]
47. Lv, P.; Ma, L.; Li, Q.; Du, F. ShapeFormer: A Shape-Enhanced Vision Transformer Model for Optical Remote Sensing Image Landslide Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 2681–2689. [CrossRef]
48. Fu, R.; He, J.; Liu, G.; Li, W.; Mao, J.; He, M.; Lin, Y. Fast Seismic Landslide Detection Based on Improved Mask R-CNN. *Remote Sens.* **2022**, *14*, 3928. [CrossRef]

49. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
50. Alayrac, J.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A Visual Language Model for Few-Shot Learning. *arXiv* **2022**, arXiv:2204.14198.
51. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. *arXiv* **2023**, arXiv:2304.02643.
52. Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; Wang, B. Segment anything in medical images. *Nat. Commun.* **2023**, *15*, 654. [CrossRef]
53. Zhang, R.; Jiang, Z.; Guo, Z.; Yan, S.; Pan, J.; Dong, H.; Gao, P.; Li, H. Personalize Segment Anything Model with One Shot. *arXiv* **2023**, arXiv:2305.03048.
54. Cen, J.; Zhou, Z.; Fang, J.; Shen, W.; Xie, L.; Jiang, D.; Zhang, X.; Tian, Q. Segment Anything in 3D with NeRFs. *arXiv* **2023**, arXiv:2304.12308.
55. He, K.; Chen, X.; Xie, S.; Li, Y.; Doll'ar, P.; Girshick, R.B. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 15979–15988.
56. Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; Shi, Z.X. RSPrompter: Learning to Prompt for Remote Sensing Instance Segmentation Based on Visual Foundation Model. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 1–17. [CrossRef]
57. Sultan, R.I.; Li, C.; Zhu, H.; Khanduri, P.; Brocanelli, M.; Zhu, D. GeoSAM: Fine-tuning SAM with Sparse and Dense Visual Prompting for Automated Segmentation of Mobility Infrastructure. *arXiv* **2023**, arXiv:2311.11319.
58. Zhang, J.; Yang, X.; Jiang, R.; Shao, W.; Zhang, L. RSAM-Seg: A SAM-based Approach with Prior Knowledge Integration for Remote Sensing Image Semantic Segmentation. *arXiv* **2024**, arXiv:2402.19004.
59. Ghorbanzadeh, O.; Xu, Y.; Zhao, H.; Wang, J.; Zhong, Y.; Zhao, D.; Zang, Q.; Wang, S.; Zhang, F.; Shi, Y.; et al. The Outcome of the 2022 Landslide4Sense Competition: Advanced Landslide Detection from Multisource Satellite Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 9927–9942. [CrossRef]
60. Zhang, X.; Yu, W.; Pun, M.-O.; Shi, W. Cross-Domain Landslide Mapping from Large-Scale Remote Sensing Images Using Prototype-Guided Domain-Aware Progressive Representation Learning. *ISPRS J. Photogramm. Remote Sens.* **2023**, *197*, 1–17. [CrossRef]
61. Zhang, H.; Zu, K.; Lu, J.; Zou, Y.; Meng, D. EPSANet: An Efficient Pyramid Squeeze Attention Block on Convolutional Neural Network. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–2 December 2021.
62. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
63. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
64. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *43*, 3349–3364. [CrossRef]
65. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
66. Lu, W.; Hu, Y.; Zhang, Z.; Cao, W. A dual-encoder U-Net for landslide detection using Sentinel-2 and DEM data. *Landslides* **2023**, *20*, 1975–1987. [CrossRef]
67. Gao, O.; Niu, C.; Liu, W.; Li, T.; Zhang, H.; Hu, Q. E-DeepLabV3+: A Landslide Detection Method for Remote Sensing Images. In Proceedings of the 2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 17–19 June 2022; Volume 10, pp. 573–577.
68. Li, D.; Tang, X.; Tu, Z.; Fang, C.; Ju, Y. Automatic Detection of Forested Landslides: A Case Study in Jiuzhaigou County, China. *Remote Sens.* **2023**, *15*, 3850. [CrossRef]
69. Tang, X.; Tu, Z.; Wang, Y.; Liu, M.; Li, D.; Fan, X. Automatic Detection of Coseismic Landslides Using a New Transformer Method. *Remote Sens.* **2022**, *14*, 2884. [CrossRef]
70. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
71. Yan, H.; Zhang, C.; Wu, M. Lawin Transformer: Improving Semantic Segmentation Transformer with Multi-Scale Representations via Large Window Attention. *arXiv* **2022**, arXiv:2201.01615.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.