International Journal of Geo-Information

MDPI

Article

Deep Fusion of DOM and DSM Features for Benggang Discovery

Shengyu Shen [1,2], Jiasheng Chen [3], Shaoyi Zhang [4], Dongbing Cheng [1,2], Zhigang Wang [1,2] and Tong Zhang [3,*]

1 Department of Soil and Water Conservation, Changjiang River Scientific Research Institute (CRSRI), Wuhan 430010, China; shenshengyu@mail.crsri.cn (S.S.); chengdb@mail.crsri.cn (D.C.); wangzg@mail.crsri.cn (Z.W.)
2 Research Center on Mountain Torrent & Geologic Disaster Prevention of the Ministry of Water Resources, Wuhan 430010, China
3 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan 430079, China; chenjs@whu.edu.cn
4 School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; sy.zhang@whu.edu.cn
* Correspondence: zhangt@whu.edu.cn

Abstract: Benggang is a typical erosional landform in southern and southeastern China. Since benggang poses significant risks to local ecological environments and economic infrastructure, it is vital to accurately detect benggang-eroded areas. Relying only on remote sensing imagery for benggang detection cannot produce satisfactory results. In this study, we propose integrating high-resolution Digital Orthophoto Map (DOM) and Digital Surface Model (DSM) data for efficient and automatic benggang discovery. The fusion of complementary rich information hidden in both DOM and DSM data is realized by a two-stream convolutional neural network (CNN), which integrates aggregated terrain and activation image features that are both extracted by supervised deep learning. We aggregate local low-level geomorphic features via a supervised diffusion-convolutional embedding branch for expressive representations of benggang terrain variations. Activation image features are obtained from an image-oriented convolutional neural network branch. The two sources of information (DOM and DSM) are fused via a gated neural network, which learns the most discriminative features for the detection of benggang. The evaluation of a challenging benggang dataset demonstrates that our method exceeds several baselines, even with limited training examples. The results show that the fusion of DOM and DSM data is beneficial for benggang detection via supervised convolutional and deep fusion networks.

Keywords: benggang; deep learning; fusion; CNN; DOM; DSM

Citation: Shen, S.; Chen, J.; Zhang, S.; Cheng, D.; Wang, Z.; Zhang, T. Deep Fusion of DOM and DSM Features for Benggang Discovery. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 556. https://doi.org/10.3390/ijgi10080556

Academic Editor: Wolfgang Kainz

Received: 22 June 2021
Accepted: 15 August 2021
Published: 17 August 2021

1. Introduction

Benggang is a Chinese word for a typical gully erosional landform [1]. Roughly translated, benggang means "slope collapse" or "collapsing gully" in English. Benggang can be found in hilly areas covered by weathered granite crusts in southern and southeastern China. Similar to gullies, the development of benggang is caused by collective impacts of gravity and runoff water, involving complex processes of sediment collapsing and transport [2]. Apart from natural factors, anthropogenic activities that destroy vegetation cover also contribute to the development of benggang [3]. Typically, continuous benggang erosions at gully heads result in chair-like forms with fragmented landscapes. Many studies have investigated the geographical distributions, development mechanisms, and erosion patterns of benggang landscapes [2–5].

In 2015, the United Nations (UN) released the 2030 Agenda for Sustainable Development and introduced 17 global Sustainable Development Goals (SDGs). Goal 15 is

about the protection of land ecosystems, aiming to promote environmental awareness and encourage ecological conservation across the world [6]. Within the framework of SDGs, the UN has defined the concept of Land Degradation Neutrality (LDN) and encouraged the international community to combat land degradation [7]. With a fast-developing erosion mechanism, benggang pose significant risks to local ecological environments and economic infrastructure, as they may destroy forests, fertile lands, roads, and human habitats [8]. In order to achieve LDN and related SDGs, necessary and immediate management and planning actions should be taken to reverse land degradation and restore benggang areas. Before we take appropriate preventive or control measures, it is vital to accurately detect benggang-eroded areas. Traditionally, the primary method to identify benggang is to conduct field surveys, which are costly in terms of resources and time. Recently, researchers have adopted various remote sensing technologies for benggang monitoring, including three-dimensional laser scanning [4] and Unmanned Aerial Vehicle (UAV) photogrammetry [9]. However, benggang are usually of small scales and covered with vegetation in the middle and late development stages, making it challenging to identify the boundary of benggang only based on remote sensing data. Without field investigation, they are very difficult to recognize from remote sensing images by manual interpretation. The current benggang investigation practices mostly start with manual identification of potential benggang areas from remote sensing imagery and are then followed by field surveys to localize benggang units. The entire workflow is time-consuming and error-prone, calling for a robust and automatic benggang discovery approach, especially for large areas. Still, since benggang are widely distributed and characterized by fast development, automatic and accurate detection of benggang areas remains a challenge.

Based on high-resolution remote sensing images, researchers have applied various machine learning methods to detect and monitor specific land deformation phenomena. Recent breakthroughs of deep learning in computer vision have offered many innovative methods and tools for remote sensing image understanding [10]. Among them, convolutional neural networks (CNNs) are the most widely used architecture for high-level image feature representation. Remarkable classification and detector performance has been achieved by either fine-tuning pretrained CNNs [11], modifying CNN frameworks [12], defining novel objective functions [13], or constructing multiple network ensembles [14]. Being powerful deep learning models in computer vision, CNNs have demonstrated their advantages in slope failure detection [15], landslide susceptibility evaluation [16], and landslide mapping [17,18]. It is also beneficial to integrate different machine learning methods for detecting land deformation phenomena. For example, using different earth observation data (satellite images and Digital Elevation Models), Piralilou et al. combined a multilayer perceptron neural network and random forest for landslide detection [19]. Ye et al. leveraged a deep belief network and logistic regression classifier to detect landslides using hyperspectral remote sensing images [20]. As these studies adopted loosely coupled models, we conjecture that integrating different data and models into an end-to-end learning framework may be beneficial for complex landform detection. Some studies have modified vanilla deep learning models to account for specific landform characteristics, such as an improved U-Net model for post-earthquake landslide extraction [21], a progressive CNN training scheme to promote generalization performance [22], and a cascaded deep learning model that accounts for landslide features from limited samples [23]. Compared with common natural or human-made objects, benggang is not a well-defined concept, with large intra-class appearance variations. Benggang comprises complex terrain landscapes without clear boundaries and distinct texture features. Directly applying deep learning detectors for benggang discovery may not achieve satisfactory performance. Therefore, we contend that an effective detection model should account for particular landscape characteristics of benggang.

Other sources of geospatial data such as high-resolution Digital Surface Model (DSM) data can provide complementary information to remote sensing image data. High-resolution DSM data contains rich information on terrain elevation capable of describing fine-grained

characteristics of complex terrains and abrupt edge changes. Despite deep learning-based feature fusion being explored for remote sensing image understanding, most studies focus on visual feature fusion using different feature descriptors [24] or features extracted from multispectral images [25]. In this study, we propose integrating high-resolution Digital Orthophoto Map (DOM) and DSM data for efficient and automatic benggang detection with an integrated end-to-end learning model. We believe this fusion of multi-source monitoring data has the benefits of high detection precision, low cost, and robustness to landform variations, which are favorable for large-scale benggang investigation and studies on the mechanism of benggang erosion.

To the best of our knowledge, we are the first to discover benggang areas using deep learning-driven fusion based on DOM and DSM data. This study makes the following contributions:

(1) We propose using a two-stream CNN framework to integrate aggregated terrain and image features for benggang discovery using high-resolution DOM and DSM data;
(2) We develop a supervised, diffusive convolutional encoding scheme that aggregates local geomorphic features, yielding expressive terrain representations for benggang;
(3) The developed deep fusion model is evaluated with a challenging benggang dataset. Supervised by limited training samples, our approach achieves satisfactory detection performance.

Similar erosional gully-like landforms can also be widely found in other countries, such as "lavaka" in Madagascar [26,27], "vocoroca" in Brazil [28] and "calanchi" in Italy [29,30]. Cost-effective monitoring of these gullies is critical for environmental protection in these countries. However, the current practices are largely limited to manual interpretation of remote sensing images and field surveys which hinge on the domain knowledge of individual experts and the data quality of the images. Machine learning methods have been used to extract specific types of land deformation phenomena (e.g., landslides) based on remote sensing images [15–17,21], but their utilities in detecting complex gully-like landforms are limited because they largely rely on visual features while ignoring terrain features that are specific to gully landforms. We believe the proposed detection approach can also be used in other areas of the Earth, helping local authorities and residents to better monitor and manage erosional gully landscapes.

2. Materials and Methods

2.1. Study Region and Data Description

The proposed approach was tested and evaluated with a DOM and a DSM dataset. The two datasets were produced from a set of aerial images, which were collected in 2018 over a hilly region of Deqing County, Guangdong Province, China. The study region has a subtropical monsoon climate, with a large solar altitude angle, strong radiation, high year-round temperature, and abundant rainfall, which provides sufficient external driving forces for the occurrence of benggang. Mountain soils are formed mainly by the weathering of granite that consists of crystals of quartz and feldspar. The weathered crust is loose and is prone to collapse under the influence of gravity. The original aerial images were acquired with three bands: blue, green, and red. The data were automatically preprocessed by INPHO, including aerial triangulation, image dense matching (for the DSM), and differential correction (for the DOM).

The DOM and DSM data have a spatial resolution of 0.2 and 0.5 m per pixel, respectively. The study region is partitioned by a regular grid with a resolution of 26 pixels, resulting in a cell size of 13 m $\times$ 13 m. We chose this resolution for the grid because it is well suited for providing fine-grained image and terrain information for detecting benggang areas, which have a minimum size of 50 m $\times$ 150 m. In both the training and test datasets, benggang areas were manually annotated by experts who have rich field experience in the study region. The labeled benggang areas were also validated by field observations. In field trips, we paid particular attention to areas that were covered by vegetation and difficult to

interpret based on the DOM data. The data contain complex benggang landscapes that are well representative of the benggang detection problem (Figures 1 and 2).

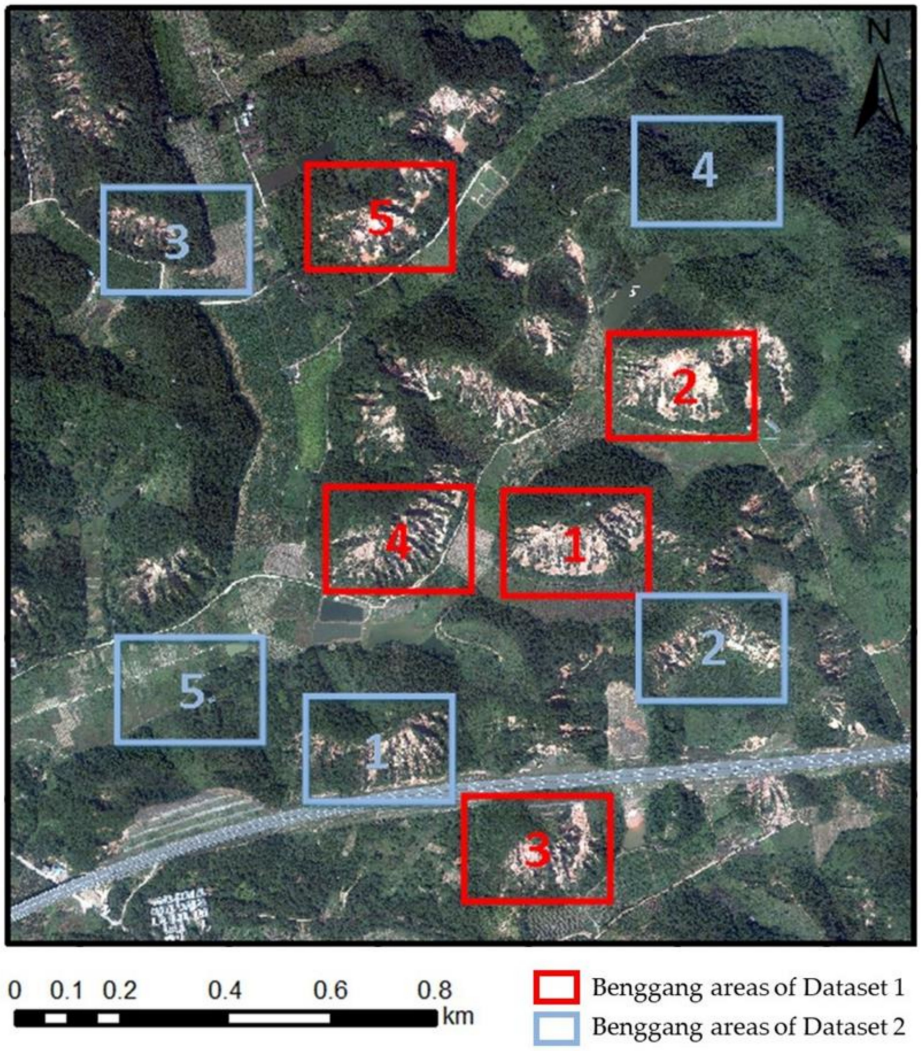


Figure 1. Study region. Benggang areas present heterogeneous structures and appearances (many have been covered with vegetation), making it challenging to detect the base in one single modality of data. The samples used in the first and the second experiments are marked with red and blue rectangles, respectively.

Figure 2. 3D view of a benggang area (produced by ArcScene using DOM and DSM data).

2.2. Detection Approach

Upon the availability of semantically rich feature representations, the benggang discovery task can be casted as a classical object detection problem, which has been extensively researched over the past several decades [31]. The deep fusion-driven benggang discovery framework is shown in Figure 3. The study region was partitioned into regular grid cells, each of which was used as the basic unit for feature extraction and representation. First, we learned a CNN to extract the abstracted image features supervised by detected benggang areas using high-resolution DOM data. The activation of the last hidden layer of the CNN was used as the high-level representation DOM features for the task. Meanwhile, DSM data were used to build a CNN-based high-level encoding scheme that aggregated local low-level geomorphic features. This encoding scheme relies on a diffusive convolutional neural network [32], helping construct high-level geomorphic descriptors, which are also supervised by detected benggang training samples. Upon the availability of the two types of high-level features, we used a two-stream CNN to integrate terrain descriptors and activation image features for benggang detection and localization. With a gated fusion network, both the DOM and DSM features were jointly embedded into a latent semantic space which had much better discriminative capabilities than each type of feature alone. Then, benggang areas could be discovered by a classification model, such as fully connected networks using the binary cross entropy loss function.

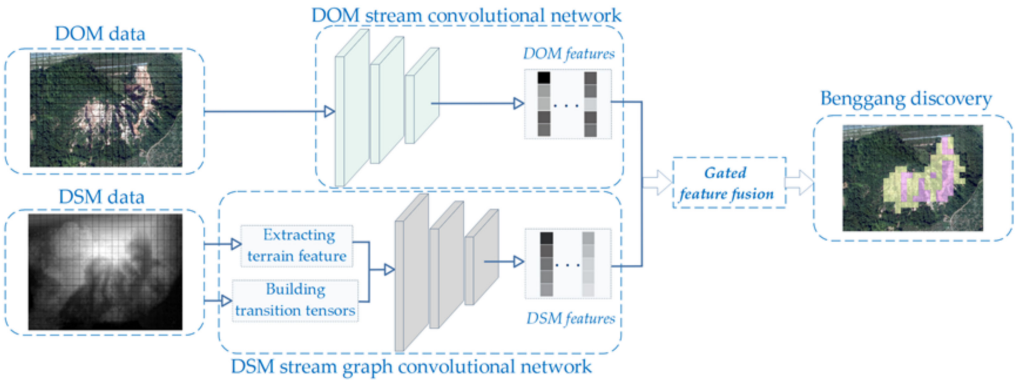


Figure 3. The proposed benggang discovery framework.

2.2.1. Extracting High-Level DOM Features

In computer vision tasks, it has been shown that CNN models trained with a huge amount of data are able to extract deep visual features. Therefore, the VGG network [33] trained with the ImageNet dataset was used to extract representative high-level DOM features in our approach. We used the VGG network to derive 512-dimensional activation feature vectors for DOM images.

2.2.2. Constructing Aggregated DSM Features

Terrain features are critical for benggang recognition and analysis. However, original terrain features are inadequate for complex scene interpretation. In this study, we propose constructing aggregated DSM features based on a diffusive convolutional neural network [32] which is trained by labeled benggang data. The diffusive convolutional neural network has the benefit of extracting semantically meaningful high-level terrain representations. A diffusive convolution was defined to simulate the process of benggang erosion. We considered each grid cell as a graph node. Given a graph $\boldsymbol{G}$ with N nodes, a transition tensor $\boldsymbol{Tr} \in \mathbb{R}^{N \times H \times N}$ can be built that encodes the probability of moving from one node to another one within H hops. $\boldsymbol{G}$ can be described by a terrain feature tensor $\boldsymbol{X} \in \mathbb{R}^{N \times F}$, where F is the size of the feature dimensionality. Our task is then to encode

informative terrain features with diffusive convolutional embeddings for all nodes. For the node i during the tth hop, the output representation can be written as

$$\boldsymbol{h}_i^{(t)} = f\left[\boldsymbol{W}^{(t)} \odot \left(\boldsymbol{Tr}_i^{(t)}\boldsymbol{X}\right)\right] \quad (1)$$

where $\boldsymbol{h}_i^{(t)} \in \mathbb{R}^{(t)\times F}$, $\boldsymbol{W}^{(t)} \in \mathbb{R}^{(t)\times F}$ is a learnable weight tensor, $\boldsymbol{Tr}_i^{(t)} \in \mathbb{R}^{(t)\times N}$ denotes the transition matrix for the tth hop, and $\odot$ denotes the Hadamard product. To enable the computation of $\boldsymbol{h}_i^{(t)}$, we needed to construct aggregated terrain feature vectors $\boldsymbol{X}$ for all graph nodes and derive transition tensor $\boldsymbol{Tr}$.

Based on DSM data, we could extract multi-dimensional terrain feature vectors at the granularity of the grid cells. For each node, a 75-dimensional vector was constructed by concatenating the following features:

(1) Average elevation over all pixels in a grid cell;
(2) Average elevation slope over all pixels in a grid cell;
(3) Maximum elevation difference between pixels;
(4) Maximum slope difference between pixels;
(5) Average gradient orientations;
(6) Maximum elevation from the centroid to four corner points and four edge mid-points;
(7) Average elevations over pixels with the same horizontal coordinates (26 dimensions);
(8) Average slope over pixels with the same horizontal coordinates (26 dimensions);
(9) A 16-dimensional vector that encodes gradient statistics based on the gradient magnitudes and orientations of all pixels. For each pixel, its gradient is weighted by the inverse of the distance between the pixel and the centroid. The 360-degree range of orientation is equally divided into 16 bins. The weighted gradients are accumulated into these 16 orientation bins according to their gradient orientations. After obtaining all the 16 elements, we reset the maximum accumulated gradient as the first element and arranged the rest of the accumulated gradients in clockwise order (16 dimensions);
(10) The normal orientation, which is recorded as the serial number of bin (0–15) that has the maximum accumulated gradients.

The inter-node transition tensor can be computed as follows:

(1) Compute the transition distances between the centroid of each node and the centroids of its eight nearest neighboring nodes (queen-based neighbors) (Figure 4). The transition distance between node o and o′ can be calculated as

$$d_{o\to o'} = \sqrt{(\Delta h)^2 + \left(d_{o_p\to o'}\right)^2} \quad (2)$$

where Δh is the difference of the average elevation between centroids o and o' (i.e, $h_o - h_{o'}$) and $d_{o_p\to o'}$ is the projected distance between the two centroids;
(2) The transition distances are labeled as positive or negative, depending on whether the destination node has a higher average elevation than the origin node. Positive (negative) distances indicate that the origin node is higher (lower) than the destination nodes;
(3) Signed distances are further weighted according to the angle α between the transition link and the normal orientation. The weights are inversely proportional to the range of the angle;
(4) The inter-node transition probabilities of the first hop $\boldsymbol{T}_o^{(1)}$ are calculated as the inverse of the signed transition distance:

$$\boldsymbol{T}_{o\to o'}^{(1)} = Sign(\Delta h) * p_{oo'} * \frac{1}{d_{o\to o'}} \quad (3)$$

$$p_{oo'} = \begin{cases} 1 - \dfrac{\alpha_{oo'} - \frac{\pi}{2}}{\pi}, & \alpha_{oo'} \in [0, \pi) \\ \dfrac{\alpha_{oo'} - \frac{\pi}{2}}{\pi}, & \alpha_{oo'} \in [\pi, 2\pi) \end{cases} \tag{4}$$

where $p_{oo'}$ is a weight to measure the effect of the angle α on transition probability and $Sign\ (\Delta h)$ returns 1 if $\Delta h > 0$; otherwise, it returns -1;

(5) For the tth hop, the transition probabilities $\boldsymbol{T}_o^{(t)}$ can be simply computed by multiplication over all probability matrices of the previous hops, and $\boldsymbol{T}_o^{(t)} = \boldsymbol{T}_o^{(t-1)} * \boldsymbol{T}_o^{(1)} = \left[\boldsymbol{T}_o^{(1)}\right]^t$.

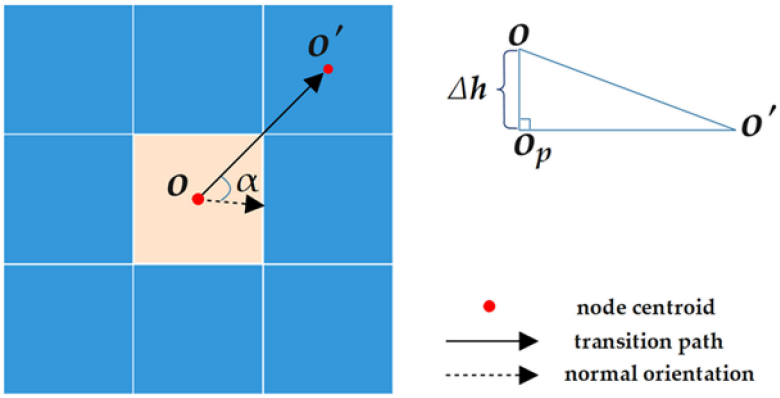


Figure 4. Transition distance calculation.

Upon the availability of the *H-hop* diffusive convolutional neural network embeddings $\boldsymbol{h}_i^{(H)}$ based on Equation (1), we could use them as the aggregated DSM features and fuse them with the DOM features using the labeled benggang data. The workflow of extracting diffusive convolutional embeddings is illustrated in Figure 5.

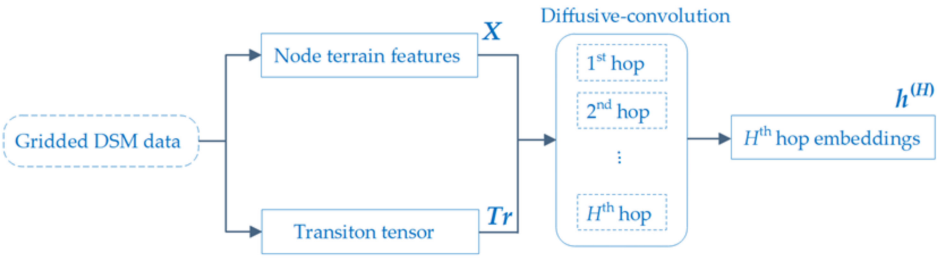


Figure 5. Diffusive convolutional embedding.

2.2.3. Fusing DOM and DSM Features

Following the gated multimodal unit model (GMU) [34], we integrated the extracted high-level image and terrain features in a supervised learning scheme. Linear transformations were applied to two feature tensors, resulting in two vectors with the same dimension for each node. The fusion was performed by a gated unit that combined information from the two modalities. For each node, the resultant fusion vector $\boldsymbol{hf}_i$ is regulated by a gate z:

$$\boldsymbol{hf}_i = z * \tan h\left(\boldsymbol{h}_i^{(I)}\right) + (1 - z) * \tanh\left(\boldsymbol{h}_i^{(H)}\right) \tag{5}$$

$$z = \sigma\left(\boldsymbol{W}_z \cdot \left[\boldsymbol{h}_i^{(I)}, \boldsymbol{h}_i^{(H)}\right]\right) \tag{6}$$

where [,] is a vector concatenation operation and $\boldsymbol{W}_z$ is the trainable gate weight, initialized from a uniform distribution [35]. We use $\boldsymbol{h}_i^{(I)}$ to denote the high-level DOM feature vector

for the ith node. Note that $\boldsymbol{h}_i^{(H)}$ and $\boldsymbol{h}_i^{(I)}$ need to be reshaped into one-dimensional vectors before being used for fusion.

A fully connected layer is used as the classification model to supervise the fusion training, using the binary cross entropy loss function:

$$\boldsymbol{L} = -\frac{1}{N}\sum_{i=1}^{N}\{y_i \log[\sigma(\boldsymbol{\theta hf}_i)] + (1-y_i)\log[1-\sigma(\boldsymbol{hf}_i)]\} \quad (7)$$

where y_i is a binary classification label (benggang or non-benggang) and $\boldsymbol{\theta}$ is a learnable vector.

During testing, the trained image and terrain feature extraction methods were applied to the gridded DOM and DSM data, respectively. The trained GMU model was used to produce fusion node vectors, which were fed into a binary classifier (e.g., a fully connected neural network) to obtain the benggang detection results.

2.3. Implementation Details

To extract the DOM features, before being fed into the DOM stream convolutional network, all training and test images were cropped and scaled to patches of 224 × 224 pixels by maintaining the original aspect ratio. The DOM stream is based on the VGG network [33]. The VGG network comprises 13 convolutional layers (with 3 × 3 convolutional filters), 5 max-pooling layers (with a kernel size of 2 × 2 pixels), and 3 fully connected layers. The learning rate was set to 0.0001.

For constructing aggregated DSM features, the diffusive convolutional neural network consisted of a diffusive convolutional activation layer and a fully connected layer, and the activation functions for the two layers were ReLu and Softmax, respectively. The learning rate was set to 0.05.

As for feature fusion, the fusion training needed at least 10 epochs and reached convergence after the loss remained under 0.01. Using the Adam optimizer [36], the model was trained with a batch size of 32. Before being used for training, the nodes were completely reshuffled. The learning rate was decayed by 0.1 for every 5 epochs.

2.4. Experimental Setting

We conducted three experiments to evaluate the proposed benggang detection approach on two datasets, each of which contained both DOM and DSM data for five samples of rectangular areas (Figure 1). The summaries of the two datasets are given in Table 1. The configurations and results of the three experiments are presented in the following.

Table 1. Summary of the datasets used in the experiments.

		No. of Benggang Grid Cells	No. of Non-Benggang Grid Cells	Total
Dataset 1	Area 1	163	269	432
	Area 2	216	216	432
	Area 3	119	313	432
	Area 4	167	265	432
	Area 5	151	281	432
	Total	816	1344	2160
Dataset 2	Area 1	130	302	432
	Area 2	120	312	432
	Area 3	54	378	432
	Area 4	0	432	432
	Area 5	0	432	432
	Total	304	1856	2160

All experiments were conducted on a desktop machine with an Intel® i7-8700K (3.7 GHz) CPU and a NVIDIA GeForce RTX 2080Ti GPU. The entire feature extraction and fusion method was implemented using PyTorch on a Microsoft Windows 10 operating system.

In the first test, we evaluated the proposed approach by a fivefold cross validation scheme on five continuous benggang areas from the first dataset. Each area consisted of 432 (i.e., 24 × 18) cells. For each run, one area was used as a test set and evaluated by the trained model using data from the other four areas. The average performance results over five runs are reported in Table 2. We used the precision, recall, and F1-score as the performance metrics to compare the proposed approach against the following baselines:

(1) VGG-DOM: a classification model based on the VGG network [33] using only DOM data. VGG16 is a widely used deep convolutional neural network with 13 convolutional layers and small-sized (3 × 3) convolution filters. The DSM data were not used in this baseline, and no data fusion was performed;
(2) DCNN-DSM: a diffusive convolutional neural network (DCNN) [32] using only DSM data. Supervised by the labeled data, the DCNN model can learn integrated representations via diffusive convolutions that leverage both local attribute and graph structure information. Similar to VGG-DOM, only one type of data was used, and no data fusion was performed;
(3) SimpleDSM: a variant of the proposed method using raw terrain features (without using aggregated terrain features that are learned by the diffusive convolutional neural network). Therefore, only the DOM convolutional network is used in the original two-stream CNN model;
(4) Concat-Fusion: a variant of the proposed method using a simple fusion method that is based on feature concatenation. The two-stream CNN architecture was used, but the gated feature fusion was replaced with simple concatenation;
(5) Linear-Fusion: a variant of the proposed method using another simple fusion method that is based on linear summation of the DOM and DSM features. Equal weights are used for the summation of the two modalities. In other words, linear feature summation was used as the fusion method rather than the gated feature fusion in the full model.

Table 2. Comparison of cross-validation performance. The means and standard deviations of the three metrics are presented.

Model	Precision		Recall		F1-Score	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
VGG-DOM	0.858	0.075	0.848	0.050	0.849	0.026
DCNN-DSM	0.729	0.106	0.595	0.158	0.634	0.073
SimpleDSM	0.830	0.064	0.873	0.046	0.847	0.018
Concat-Fusion	0.877	0.081	0.861	0.051	0.868	0.058
Linear-Fusion	0.886	0.052	0.864	0.051	0.875	0.050
Ours	0.912	0.024	0.876	0.034	0.894	0.027

The precision was computed as the ratio between the number of correctly detected benggang cells and the total number of cells classified as benggang. The recall was computed as the ratio between the number of correctly detected grid cells and the total number of benggang cells. The F1-score is the geometric mean of the precision and recall.

3. Results

3.1. Comparison with Baselines

Table 2 shows that the proposed deep fusion-based approach achieved better detection performance over the compared baselines. The variations of our approach over different test examples were also relatively small. The empirical improvements over VGG-DOM and DCNN-DSM could be attributed to the fusion of both DOM and DSM information. The performance gain of the proposed approach over SimpleDSM indicates the advantage of

using diffusive convolutional neural networks in training aggregated terrain features. The use of the gated fusion model in the proposed approach was beneficial, as demonstrated by the improvements of the performance metrics over the other two fusion methods (i.e., Concat-Fusion and Linear-Fusion).

The second experiment was to compare the performance of the proposed approach with the other baselines for another five examples in the second dataset (see Figure 1). The five tested examples contained three benggang areas and two non-benggang areas. The tested model was trained using the first dataset. Table 3 indicates that our approach was superior to the other baselines over the three performance metrics. The other baselines incorrectly classified some grid cells as benggang in the two non-benggang areas, but our approach could avoid these mistakes.

Table 3. Performance comparison of benggang and non-benggang samples.

Model	Precision	Recall	F1-Score
VGG-DOM	0.721	0.809	0.763
DCNN-DSM	0.474	0.569	0.517
Concat-Fusion	0.834	0.859	0.846
Ours	0.835	0.931	0.880

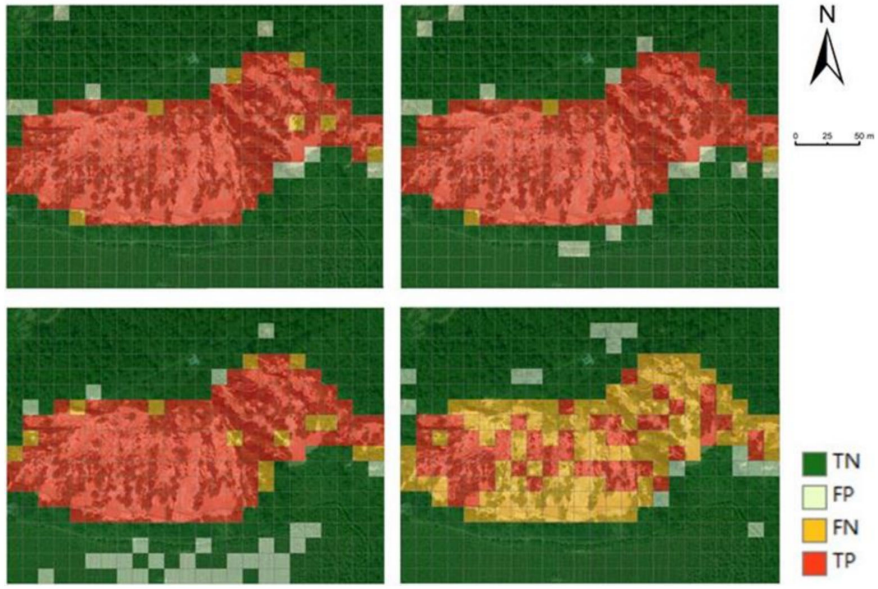
In the last experiment, we used different numbers of training samples from the first dataset and tested over the rest of the samples in the first and second datasets with the goal to evaluate the generalization capabilities of the proposed approach. Table 4 shows that the performance gains of our approach over the other two baselines were more prominent when using a small amount of training data, implying that our approach could generalize well using limited training samples.

Table 4. Detection performance (F1-score) using different numbers of training samples.

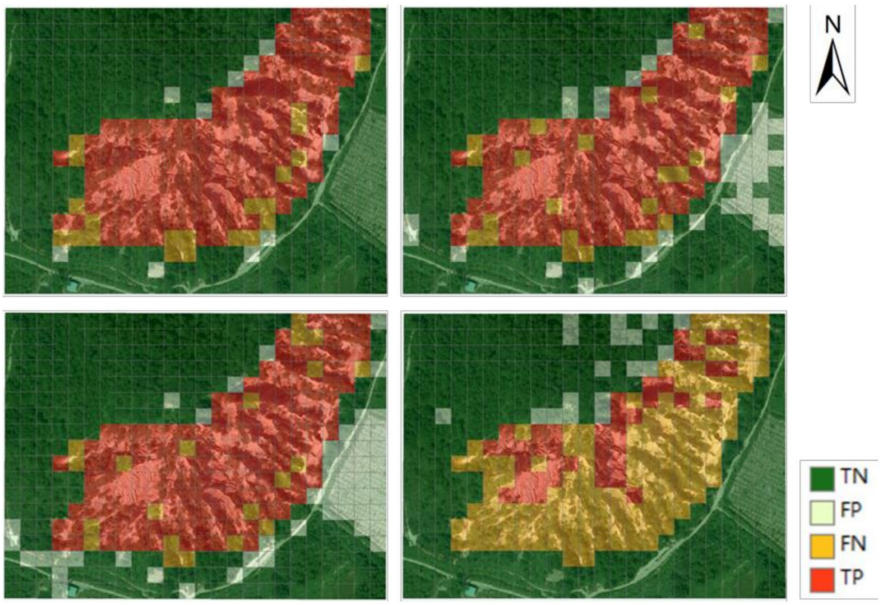
Model	1	2	3	4
VGG-DOM	0.751	0.774	0.797	0.807
Concat-Fusion	0.774	0.797	0.831	0.825
Ours	0.824	0.842	0.833	0.843

3.2. Qualitative Results

Figure 6 presents the detection results of some examples, showing that the proposed approach could distinguish contiguous benggang areas from complex backgrounds. The proposed approach was also robust for complex non-benggang backgrounds. The two areas in Figure 6 contain a mixed set of different landscapes, including forests, roads, and farmland. The proposed approach was able to distinguish them from benggang areas. The VGG-DOM method tends to produce false positive results, since it is not able to distinguish non-benggang areas with similar texture patterns to benggang areas. DCNN-DSM performed the worst, indicating merely relying on terrain features is not robust and should be integrated with image features. Figure 6b shows that the Concat-Fusion method frequently labeled non-benggang areas as benggang since it treated the DOM and DSM equally and may not have chosen the most discriminative local features.



(a) Area 1 from dataset 1.



(b) Area 4 from dataset 1.

Figure 6. Comparison of detection results for two benggang areas. (**Top left**) The proposed approach. (**Top right**) Concat-Fusion. (**Bottom left**) VGG-DOM. (**Bottom right**) DCNN-DSM. TP: true positive (benggang cells are correctly detected); FP: false positive (non-benggang cells are incorrectly classified as benggang); TN: true negative (non-benggang cells correctly classified as non-nenggang); and FN: false negative (benggang cells incorrectly classified as non-benggang).

3.3. Parameter Selection

We investigated the impacts of one parameter on the detection performance: the number of diffusive hops when constructing aggregated terrain features, following the same setting as the first experiment.

Table 5 shows that when $h = 3$, the model achieved the best and most stable performance. We attribute this optimal selection to the sizes of the benggang in the studied region. The sizes of the benggang areas ranged from 50 m × 150 m to 150 m × 350 m, meaning that three hops (i.e., 26~55 m in transition distance) were suitable for capturing the change patterns of the elevation variations across the benggang boundaries or within

the benggang areas. The cell size and hop number could be adjusted when given different image resolutions and benggang sizes.

Table 5. Performance comparison of different diffusive hops. Means and standard deviations of the three metrics are presented.

Hop Number	Precision		Recall		F1-Score	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
0	0.886	0.053	0.866	0.055	0.874	0.038
1	0.900	0.030	**0.886**	0.037	0.893	0.032
2	0.877	0.051	0.870	0.037	0.873	0.039
3	**0.912**	0.024	0.876	0.034	**0.894**	0.027
4	0.892	0.030	0.871	0.050	0.881	0.036
5	0.884	0.041	0.877	0.041	0.880	0.035

3.4. Computational Effiency

We compared the training and testing time cost of the proposed method and three baselines for the second experiment. Table 6 shows that our approach had approximately similar time costs to VGG-DOM and Concat-Fusion. The DCNN-DSM model performed much faster at both the training and testing stages because it only handled DSM data. The time costs were practically acceptable for the benggang detection task.

Table 6. Comparison of time costs.

Model	Training Time (s)	Test Time (s)
VGG-DOM	316.3	13.7
DCNN-DSM	55.9	7.2
Concat-Fusion	274.1	14.9
Ours	295.9	13.8

4. Discussion

Since benggang areas are surrounded by similar landforms in mountainous southern and southeastern China, it is challenging to detect them by manual inspection or relying on one single source of earth observational data. If benggang areas are covered with vegetation, DOM data may not provide sufficient texture information for benggang detection. Without other sources of information, bare lands or farm lands after harvest would confuse the DOM-based classifier. On the other hand, the development of a benggang is driven by consistent erosion on its gully head, causing significant elevation variations along its boundary. The gully bottom and deposition area have relatively mild elevation changes. Therefore, DSM data can be of help in benggang detection. However, since the spatial resolution of DSM data is usually much lower than that of DOM data, using only DSM data may not produce satisfactory results, as indicated by our tests. The integration of DOM and DSM data thus allowed us to examine three-dimensional landscape models with high-resolution texture, which provided much richer feature information than either DOM or DSM data. We proposed integrating DOM and DSM data under a deep gated fusion framework, taking advantage of the most effective discriminative capabilities of image and terrain features.

To compensate the coarse resolution in the DSM data, we used diffusive convolutions to extract aggregated meaningful terrain features that were able to preserve the variation patterns of elevation for benggang areas. We note that the benggang boundaries have distinct feature vectors from non-benggang areas because they are characterized by significant elevation variations. The diffusive convolutional features thus can capture such variations to facilitate the discovery of benggang boundaries. We used t-SNE [37] to visualize the feature embeddings of the compared detection approaches in 2D space. Figure 7 shows that the proposed deep fusion approach could learn two embedding clusters that could be

easily separated, whereas other baselines failed to distinguish benggang and non-benggang cells since the learned embeddings were significantly overlapped.

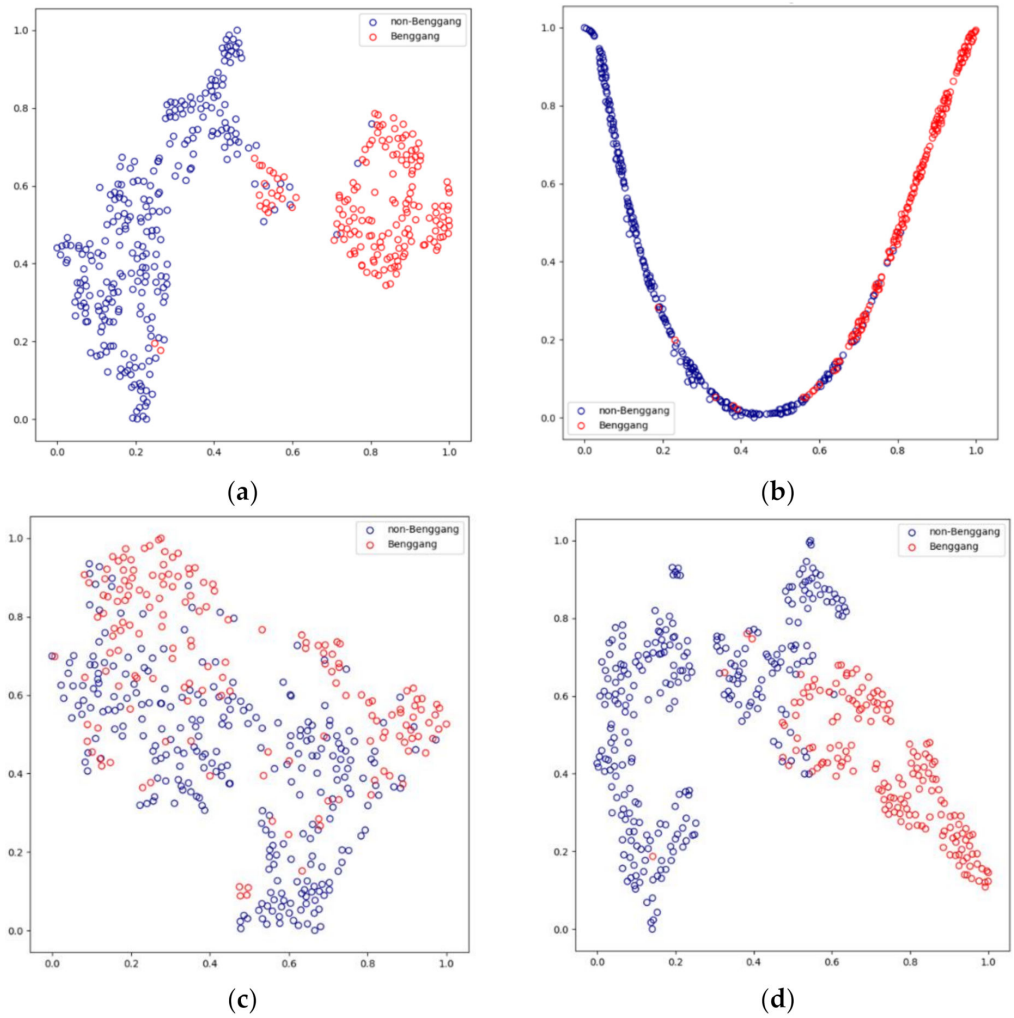
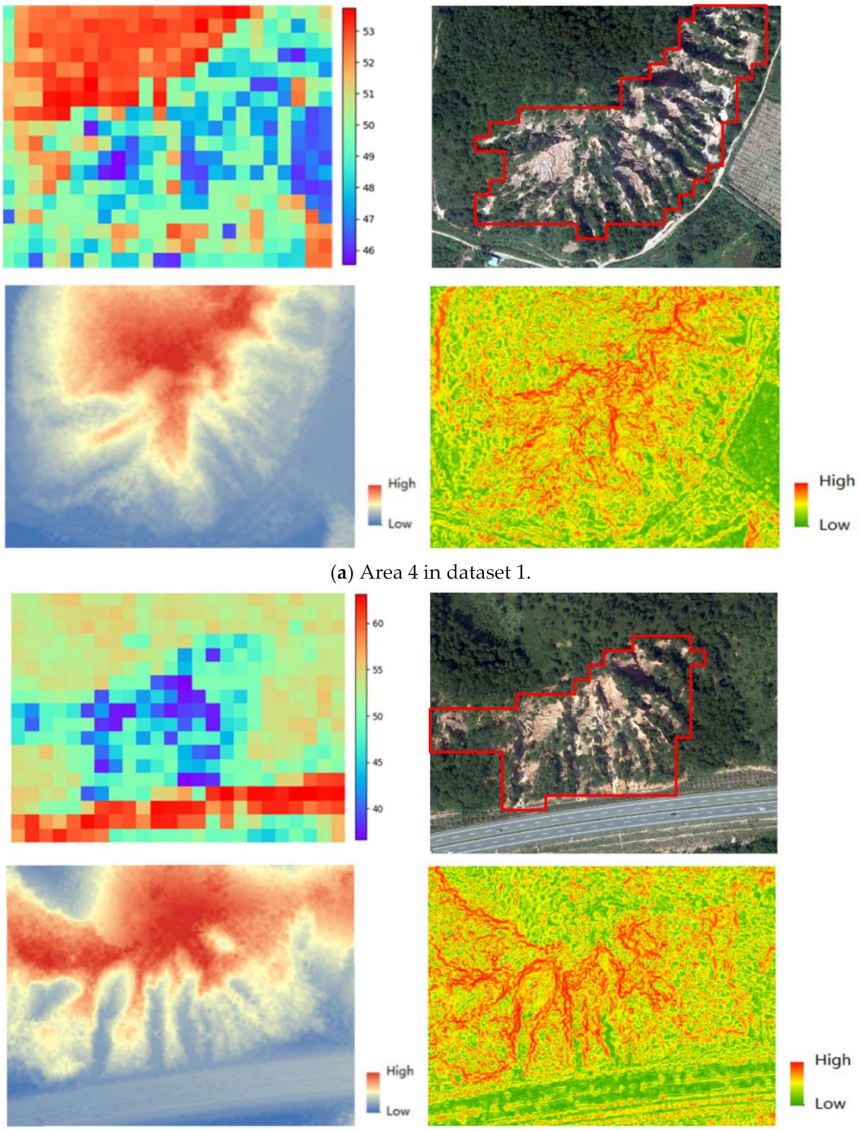


Figure 7. Visualization of feature embeddings for each grid cell using t-SNE for area 1 in dataset 1. (**a**) Our deep fusion approach. (**b**) VGG-DOM. (**c**) DCNN-DSM. (**d**) Concat-Fusion.

Being totally data-driven, the gated fusion mechanism facilitates the interpretation of the most informative integrated features based on DOM and DSM features. According to Equation (5), the gate activations z regulates the influences of DOM and DSM data on benggang detection. Thus, we could use the averages of z to see which data modality had greater effects on the test results. Figure 8 shows the quantitative scores that describe which data modality was more influential for each detected cell for two areas. The grid cells with a blue (red) color show that DSM (DOM) data played a more important role in the fusion model. According to the two samples, we can see that the terrain features were more useful in detecting benggang areas or identifying non-benggang areas if they had similar image features to those of benggang areas (e.g., farmlands to the right side of Figure 8a). Image features are more helpful when we try to distinguish non-benggang areas from benggang areas if these areas present distinct texture features (e.g., roads in Figure 8b).



(**a**) Area 4 in dataset 1.

(**b**) Area 1 in dataset 2.

Figure 8. Visualization of the effects of DOM and DSM on detection results. (**Top left**) Influential score for each grid cell (0–100, with higher scores indicating large influences of DOM features). (**Top right**) Corresponding image showing the boundaries of benggang areas. (**Bottom left**) Shaded relief map. (**Bottom right**) Slope map.

5. Conclusions

This study explores the possibility of combining DOM and DSM data for detecting benggang, a common erosional landform in southern and southeastern China. Diffusive convolutional neural networks are used to extract representative terrain features, which are then integrated with CNN-derived image features to label benggang landscapes. We have demonstrated that the proposed detection approach achieved performance superior to several baselines, showing that the fusion of DOM and DSM data is beneficial for benggang detection via supervised convolutional and deep fusion networks. Future work will focus on the detection of different development stages of benggang and the evaluation of erosion risk for the surrounding environments. We also plan to collect DOM and DSM data from other areas in southern China and perform extensive evaluations on the proposed fusion-based detection approach.

Author Contributions: Conceptualization, Shengyu Shen and Tong Zhang; methodology, Shengyu Shen and Tong Zhang; software, Jiasheng Chen and Shaoyi Zhang; validation, Jiasheng Chen and Dongbing Cheng; formal analysis, Shengyu Shen, Tong Zhang, Jiasheng Chen and Zhigang Wang; investigation, Dongbing Cheng and Zhigang Wang; resources, Shengyu Shen; data curation, Jiasheng Cheng and Shaoyi Zhang; writing—original draft preparation, Tong Zhang; Shengyu Shen; writing—review and editing, Shengyu Shen; visualization, Jiasheng Chen; supervision, Shengyu Shen and Zhigang Wang; project administration, Shengyu Shen. and Dongbing Cheng; funding acquisition, Shengyu Shen. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant numbers 41601298 and 41871308; the Basic Scientific Research Operations Fund of Central Public Welfare Institutions under grant CKSF2014024/TB; the National Key R&D Program of China (International Scientific & Technological Cooperation Program) under grant 2019YFE0106500; and the Fundamental Research Funds for the Central Universities.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the reviewers and editors for valuable comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zeng, Z. *Topographic Principles: Book I*; South China Normal University Press: Guangzhou, China, 1960. (In Chinese)
2. Lin, J.; Huang, Y.; Wang, M.; Jiang, F.; Zhang, X.; Ge, H. Assessing the sources of sediment transported in gully systems using a fingerprinting approach: An example from South-east China. *Catena* **2015**, *129*, 9–27. [CrossRef]
3. Xu, J. Benggang erosion: The influencing factors. *Catena* **1996**, *27*, 249–263.
4. Liu, H.; Qian, F.; Ding, W.; Gómez. Using 3D scanner to study gully evolution and its hydrological analysis in the deep weathering of southern China. *Catena* **2019**, *183*, 104218. [CrossRef]
5. Liu, H.; Hörmann, G.; Qi, B.; Yue, Q. Using high-resolution aerial images to study gully development at the regional scale in southern China. *Intl. Soil Water Conserv. Res.* **2020**, *8*, 173–184. [CrossRef]
6. United Nations. Transforming Our World: The 2030 Agenda for Sustainable Development. Available online: https://sdgs.un.org/2030agenda (accessed on 10 August 2021).
7. United Nations Convention to Combat Desertification (UNCCD). Integration of the Sustainable Development Goals and Targets into the Implementation of the United Nations Convention to Combat Desertification and the Intergovernmental Working Group Report on Land Degradation Neutrality (ICCD/COP(12)/20/Add.1). Available online: https://www.unccd.int/sites/default/files/inline-files/dec3-COP.12eng.pdf (accessed on 10 August 2021).
8. Luk, S.; Yao, Q.; Gao, J.; Zhang, J.; He, Y.; Huang, S. Environmental analysis of soil erosion in Guangdong Province: A Deqing case study. *Catena* **1997**, *29*, 97–113. [CrossRef]
9. Shen, S.; Zhang, T.; Zhao, Y.; Wang, Z.; Qian, F. Automatic Benggang recognition based on latent semantic fusion of UHR DOM and DSM feature. *Isprs Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *5*, 331–338. [CrossRef]
10. Gu, Y.; Wang, Y.; Li, Y. A survey on deep learning-driven remote sensing image scene understanding: Scene classification, scene retrieval and scene-guided object detection. *Appl. Sci.* **2019**, *9*, 2110. [CrossRef]
11. Wang, J.; Luo, C.; Huang, H.; Zhao, H.; Wang, S. Transferring Pre-Trained Deep CNNs for Remote Scene Classification with General Features Learned from Linear PCA Network. *Remote Sens.* **2017**, *9*, 225. [CrossRef]
12. Xie, J.; He, N.; Fang, L.; Plaza, A. Scale-Free Convolutional Neural Network for Remote Sensing Scene Classification. *IEEE Trans. Geos. Remote Sens.* **2019**, *57*, 6916–6928. [CrossRef]
13. Cheng, G.; Yang, C.; Yao, X.; Guo, L.; Han, J. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2811–2821. [CrossRef]
14. Zhang, F.; Du, B.; Zhang, L. Scene Classification via a Gradient Boosting Random Convolutional Network Framework. *Ieee Trans. Geos. Remote Sens.* **2015**, *54*, 1793–1802. [CrossRef]
15. Ghorbanzadeh, O.; Meena, S.; Blaschke, T.; Aryal, J. UAV-Based slope failure detection using deep-learning convolutional neural networks. *Remote Sens.* **2019**, *11*, 2046. [CrossRef]
16. Wang, Y.; Fang, Z.; Hong, H. Comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County, China. *Sci. Total Environ.* **2019**, *666*, 975–993. [CrossRef]
17. Prakash, N.; Manconi, A.; Loew, S. Mapping landslides on EO data: Performance of deep learning models vs. traditional machine learning models. *Remote Sens.* **2020**, *12*, 346. [CrossRef]

18. Bragagnolo, L.; Rezende, L.; da Silva, R.; Grzybowski, J. Convolutional neural networks applied to semantic segmentation of landslide scars. *Catena* **2021**, *2*, 105189. [CrossRef]
19. Piralilou, S.; Shahabi, H.; Jarihani, B.; Ghorbanzadeh, O.; Blaschke, T.; Gholamnia, K.; Meena, S.; Aryal, J. Landslide detection using multi-scale image segmentation and different machine learning models in the higher Himalayas. *Remote Sens.* **2019**, *11*, 2575. [CrossRef]
20. Ye, C.; Li, H.; Cui, P.; Liang, L.; Pirasteh, S.; Marcato, J.; Gonçalves, W.; Li, J. Landslide detection of hyperspectral remote sensing data based on deep learning with constrains. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 5047–5060. [CrossRef]
21. Liu, P.; Wei, Y.; Wang, Q.; Chen, Y.; Xie, J. Research on post-earthquake landslide extraction algorithm based on improved U-Net model. *Remote Sens.* **2020**, *12*, 894. [CrossRef]
22. Prakash, N.; Manconi, A.; Loew, S. A new strategy to map landslides with a generalized convolutional neural network. *Sci. Rep.* **2021**, *11*, 9722. [CrossRef]
23. Yi, Y.; Zhang, W. A new Deep-learning-based approach for earthquake-triggered landslide detection from single-temporal RapidEye satellite imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 6166–6176. [CrossRef]
24. Chaib, S.; Liu, H.; Gu, Y.; Yao, H. Deep feature fusion for VHR remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *55*, 4775–4784. [CrossRef]
25. Yang, J.; Zhao, Y.; Chan, J. Hyperspectral and Multispectral Image Fusion via Deep Two-Branches Convolutional Neural Network. *Remote Sens.* **2018**, *10*, 800. [CrossRef]
26. Wells, N.; Andriamihaja, B.; Solo Rakotovololona, H. Patterns of development of lavaka, Madagascar's unusual gullies. *Earth Surf. Process. Landf.* **1991**, *16*, 189–206. [CrossRef]
27. Cox, R.; Zentner, D.; Rakotondrazafy, A.; Rasoazanamparany, C. Shakedown in Madagascar: Occurrence of lavakas (erosional gullies) associated with seismic activity. *Geology* **2010**, *38*, 179–182. [CrossRef]
28. Bacellar, L.; Coelho Netto, A.; Lacerda, W. Controlling factors of gullying in the Maracujá catchment, southeastern Brazil. *Earth Surf. Process. Landf.* **2005**, *30*, 1369–1385. [CrossRef]
29. Moretti, S.; Rodolfi, G. A typical "calanchi" landscape on the Eastern Apennine margin (Atri, Central Italy): Geomorphological features and evolution. *Catena* **2000**, *40*, 217–228. [CrossRef]
30. Neugirg, F.; Stark, M.; Kaiser, A.; Vlacilova, M.; Della Seta, M.; Vergari, F.; Schmidt, J.; Becht, M.; Haas, F. Erosion processes in calanchi in the Upper Orcia Valley, Southern Tuscany, Italy based on multitemporal high- resolution terrestrial LiDAR and UAV surveys. *Geomorphology* **2016**, *269*, 8–22. [CrossRef]
31. Zou, Z.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. Available online: https://arxiv.org/abs/1905.05055.2019 (accessed on 10 June 2021).
32. Atwood, J.; Towsley, D. Diffusion-convolutional neural networks. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 5–10 December 2016.
33. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
34. Arevalo, J.; Solorio, T.; Montes-y-Gómez, M.; González, F. Gated multimodal unit for information fusion. In Proceedings of the ICLR 2017, Toulon, France, 24–26 April 2017.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on Imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
36. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference for Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
37. Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.