Deep Fusion of Local and Non-Local Features for Precision Landslide Recognition

Qing Zhu, Lin Chen, Han Hu, Binzhi Xu, Yeting Zhang, Haifeng Li

***Abstract*—Precision mapping of landslide inventory is crucial for hazard mitigation. Most landslides generally co-exist with other confusing geological features, and the presence of such areas can only be inferred unambiguously at a large scale. In addition, local information is also important for the preservation of object boundaries. Aiming to solve this problem, this paper proposes an effective approach to fuse both local and non-local features to surmount the contextual problem. Built upon the U-Net architecture that is widely adopted in the remote sensing community, we utilize two additional modules. The first one uses dilated convolution and the corresponding atrous spatial pyramid pooling, which enlarged the receptive field without sacrificing spatial resolution or increasing memory usage. The second uses a scale attention mechanism to guide the up-sampling of features from the coarse level by a learned weight map. In implementation, the computational overhead against the original U-Net was only a few convolutional layers. Experimental evaluations revealed that the proposed method outperformed state-of-the-art general-purpose semantic segmentation approaches. Furthermore, ablation studies have shown that the two models afforded extensive enhancements in landslide-recognition performance.**

***Index Terms*—Landslide mapping, U-Net, Attention, Dilated convolution**

I. Introduction

LANDSLIDE is one of the most destructive natural hazards, and may also cause a series of secondary disasters, such as floods from barrier-lake overflows and dam breakages [1]. Due to the complexity of factors that can cause landslides, and the abrupt occurrence of landslides during or after continuous rainfall in landslide-prone areas, pre-hazard mapping of landslide susceptibility is not sufficient for hazard management [2]. Efficient and precision mapping of post-hazard landslide regions is also crucial for successful emergency responses, and the prediction of secondary landslides that may occur due to unstable underlying surfaces.

Manuscript received: February 9, 2020. This work was supported in part by the National Natural Science Foundation of China (Projects No.: 41941019, 41871291, 41871314) and in part by the National Key Research and Development Program of China (Project No.: 2018YFB0505404). (*Mutual corresponding authors: Han Hu and Haifeng Li)*

Qing Zhu, Lin Chen and Binzhi Xu are with the Faculty of Geoscience and Environmental Engineering, Southwest Jiaotong University, Chengdu, China. (e-mail: zhuq66@263.net; chenlin@my.swjtu.edu.cn; xubinzhi@my.swjtu.edu.cn)

Han Hu is with the State Key Laboratory of Rail Transit Engineering Informatization, China Railway First Survey and Design Institute Co. Ltd., Xi'an, China and also with the Faculty of Geoscience and Environmental Engineering, Southwest Jiaotong University, Chengdu, China. (email: han.hu@swjtu.edu.cn)

Yeting Zhang is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China (e-mail: zhangyeting@whu.edu.cn)

Haifeng Li is with the School of Geosciences and Info-Physics, Central South University, Changsha, China (e-mail: lihaifeng@csu.edu.cn)

Currently, many landslide-mapping systems require interactive interpretations, which obviously depend on the experience of human experts, and are thus not sufficient to enable rapid response. Classical approaches tend to use domain-specific knowledge of the spectral characteristic of optical or radar images for the delineation and image-segmentation of landslides, such as textural patterns [3], terrain features [3], [4] and vegetation indexes [5]. The object-based strategy [6] is also widely used to increase the reliability when high-resolution data are used [4], [5], [7], [8], [2].

With the advent of deep learning paradigms, convolutional neural network (CNN)-based approaches have yielded impressive results in many image-processing objectives. Specifically, fully convolutional networks (FCNs) [9] have enabled end-to-end segmentation using a deconvolution module. FCNs and their successors [10], [11] have significantly boosted the development of semantic segmentation of images. However, some challenges remain to be surmounted before such techniques can be applied to the mapping of landslides. (1) *Local receptive field.* CNN-based features only use information in local regions, but landslides often have confusing spectral information generated by background features, such as roads and residential areas, and large contexts are required to remove these ambiguities during segmentation. (2) *Boundary preservation.* FCN essentially uses an up-sampling step to recover the low-resolution feature maps to the original resolution, which results in the loss of large amounts of boundary information. Although pyramid structures such as Deeplab and U-Net [10], [12] can propagate and aggregate information from different scales, this problem persists. The strategy of change detection has also been considered [13] for better boundary preservation, but it is not applicable when no pre-hazard inventory exists.

In this study, we have proposed and developed an approach to alleviate the above problems, using the fusion of both local and non-local information, built upon the U-Net structure [12]. During the down-sampling, we used the atrous convolution [14], [11] with different dilation sizes to simulate multiscale features, prior to down-sampling via a max-pooling operation[12]. In addition, inspired by the Deeplab, we added a fusion step in the bottleneck of U-Net to aggregate multiscale features [10]. In the up-sampling step, inspired by the scale attention module [15], [16], specifically, the alignment model [17], [18], we augmented the U-Net with the attention module to suppress the irrelevant and confusing features from coarse scales by learning a weight map. These two incremental modifications cooperated to improve the accuracy and robustness of the mapping of landslide regions using only post-hazard unmanned aerial vehicle (UAV) images.

II. Methodology

A. Problem setup and overview of the approaches

We formulated the mapping of the post-hazard landslide from UAV images as a semantic segmentation problem [9] that is widely studied in the computer vision community. More specifically, given an image $\mathbf{X}$, the purpose was to assign a binary label l_i ($l_i = \{0, 1\}$) to each pixel i, which comprised the binary segmentation b of the image to background and landslide regions. The objective was to learn this mapping $\mathcal{F} \mapsto b$ in an end-to-end manner [9], using the training binary segmentation samples and corresponding UAV images.

Inspired by previous work on semantic segmentation of remote sensing images [19], we built the system upon the prominent U-Net [12] structure, which fuses CNN maps in a multiscale fashion. U-Net also features a low memory profile, which is critical for large-scale remote sensing applications. In addition, U-Net can make dense semantic predictions by using the encoder-decoder strategy. In the encoder, U-Net grasped both the low-level greyscale and gradient features and high-level contextual features from finer space resolution (shallower channels) to coarser space resolution (deeper channels, respectively). In the decoder, U-Net concatenated the encoder features in the left part to the deconvolved features (Fig. 1). Its use of the above skip layers enabled fusion of multiscale information, which substantially improved the mapping resolution.

However, in our mapping of post-hazard landslide regions, we have often encountered regions with both very small and very large structures. This presents a problem, as due to the computational efficiency, it is not possible to continuously increase the number of layers to enlarge the receptive field to encompass larger objects. In addition, the spectral signatures of small or local features may be obscured by areas with similar spectral information, such as bare earth or roads.

To overcome this barrier, in this study we augmented the U-Net with two modules: (1) the ASPP (atrous spatial pyramid pooling) module for enlarging the receptive field without going too deep and losing spatial resolution in the encoder; and (2) the attention module, to suppress irrelevant or confusing features by exploiting non-local contextual information at the coarse level in the decoder. The overview of the architecture is illustrated in Fig. 1. The two modifications are detailed in the subsections below.

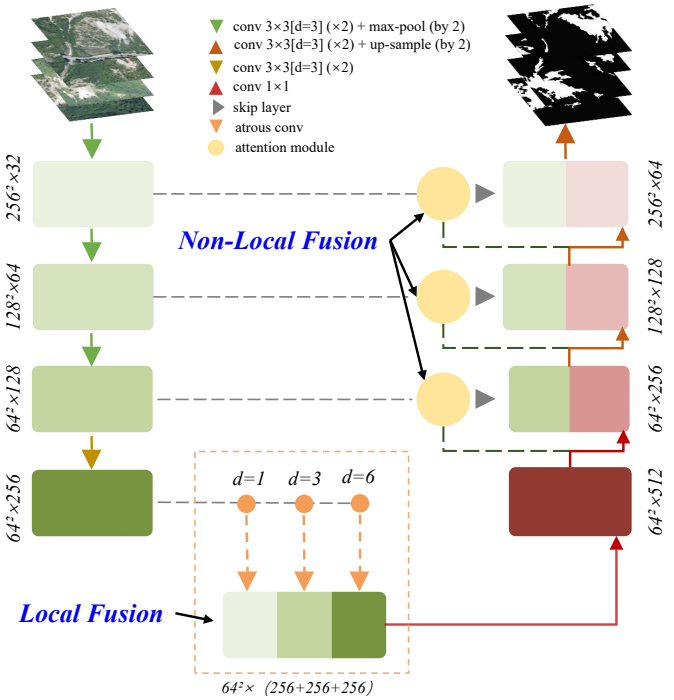


Fig. 1: Overview of the architecture for the landslide mapping.

B. Atrous spatial pyramid pooling for enlarged receptive fields

1) Atrous convolution: Classical convolution is intrinsically a local method, which can only account for a fixed region and relies on pooling operations, e.g. max-pooling, to enlarge the receptive field at the cost of coarser spatial resolution. Atrous convolution [10], [14] has been used to surmount this problem, as it adds holes in the convolution kernel, which can effectively enlarge the receptive field without sacrificing spatial resolution or efficiency, as opposed to enlarging the size of the convolution kernel. The number of holes is controlled by the dilation rate d, and the size of the kernel with holes k_h is also related to the effective kernel size k as $k_h = k + (k-1) \times (d-1)$. Another strength of the atrous convolution approach is that dense feature maps are produced by sliding the window across the whole image. This property substantially simplifies the fusion of multiscale information.

2) Atrous spatial pyramid pooling: Although atrous convolution can solve the receptive field problem, if the dilation rate is sequentially increasing across several dependent layers, the information will inevitably become too sparse [14]. Therefore, we used an ASPP strategy [10] to substitute for the bottleneck part of the U-Net, as shown in Fig. 1. Specifically, the fusion was denoted as

$$y = D_{3,1}(x) \oplus D_{3,3}(x) \oplus D_{3,6}(x), \tag{1}$$

where $D_{d,k}$ indicates the atrous convolution with a dilation rate of d and a kernel size of k and $\oplus$ denotes the concatenation operation of the features. The dimensions of the fused features were resized through convolution with a 1×1 kernel. By use of the ASPP fusion, the receptive field could account for more than a quarter of the entire tile.

C. Attention augmented up-sampling

The attention mechanism originated from natural language processing [20], and was later extended to image classification [17] and semantic segmentation [15], [21], [16], [22]. The premise of the attention mechanism is that the absolute values of feature maps also reveal their importance; therefore, we can learn a weight map $\alpha \in (0, 1)$, with the same spatial resolution as the feature map, to adaptively suppress irrelevant features. Typically, such a weight map considers global information [15], [22] and two models (also known as the compatibility functions [17]) are available, the additive model and the multiplicative model. The success of the dual attention network [22] in general-purpose semantic segmentation motivated us

to extend the additive model to the U-Net structure, i.e., by using the scale attention module introduced in previous works [15], [16].

Specifically, we did not directly concatenate the feature map from the encoder (*i.e.* left part of Fig. 1) with the up-sampled feature map in the coarser level from the decoder (*i.e.* right part of Fig. 1). Instead, we first augmented the encoded features using the features from the coarser level of the decoder. As features in the coarser level contained more contextual information, the non-local information helped in the determination of the weight map.

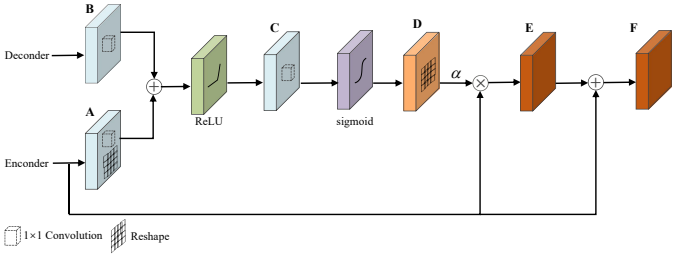


Fig. 2: Attention augmented up-sampling by the non-local fusion of features from coarser levels.

Fig. 2 shows the enlarged representation of the attention module, *i.e.* the circle in Fig. 1. The input encoder and decoder features, *i.e.* **A** and **B** respectively, were first convolved with a 1×1 filter to make the channels compatible; in addition, the encoder features **A** were also reshaped to make the size compatible. The two maps were then connected by an element-wise summation, followed by processing via a rectified linear activation (ReLU) function. Then, another convolution with only 1 channel was used to create the compatibility map **C** [17].

Two options were available to create the attention map α: the *softmax* $\theta(x_i) = \frac{e^{x_i}}{\sum_t e^{x_t}}$ and *sigmoid* $\theta(x_i) = \frac{e^{x_i}}{e^{x_i}+1}$ function. Although most approaches have used the *softmax* function [15], [17], [22], we have found that the normalisation part in the *softmax* function, *e.g.* $\sum_i \alpha_i = 1$, will make the activation too sparse, which is not good for a process that is applied multiple times. Therefore, the *sigmoid* function was used to generate the feature **D** and resampled to the spatial resolution of the encoder feature A, similar to previous work [16]. The attention-augmented feature **E** comprised element-wise multiplication with weight α. Finally, a skip connection [23] was used for the map **F** before concatenation with the up-sampled features, as below,

$$F_i = A_i + \alpha_i A_i. \tag{2}$$

D. Implementation details

The labels were obtained interactively from the original UAV orthophotos in ArcMap and tiled into clips with a size of 512×512. The images were normalised to $[-0.5, 0.5]$ for both training and testing. As landslides are mainly located in forests, we performed hard mining by intentionally sampling more confusing regions, such as bare earth, large rocks, roads and rivers. Random flipping, rotation and scaling were used for the data augmentation process. Tensorflow 1.9 was used to implement the framework on a machine with four NVIDIA RTX Titan graphics computing units (GPUs). For the hyper-parameters of the training, the batch size was 12 for each GPU, momentum was 0.9, learning rate was $1e^{-3}$ and regularisation was $5e^{-4}$. The training lasted for 100000 iterations and we recorded the model every 20000 iterations. The model with the best testing performance was chosen. Finally, *Softmax* was used, incorporating the binary segmentation loss function. During testing, the original orthophotos were clipped, loaded and mosaicked dynamically.

III. Experimental Evaluations

A. Experimental setup and overall performance

The UAV images covering six counties were obtained for the landslides caused by the earthquake in Jiuzhaigou, China on 8 August 2017. We interactively selected 10^4 tiles for the training, and used a $70-30\%$ spilt. Four entire UAV orthophotos were used for the testing, which comprised approximately 3000 tiles. Three common metrics were used to evaluate the pixel-wise results, namely $precision$, $recall$ and F_1 score.

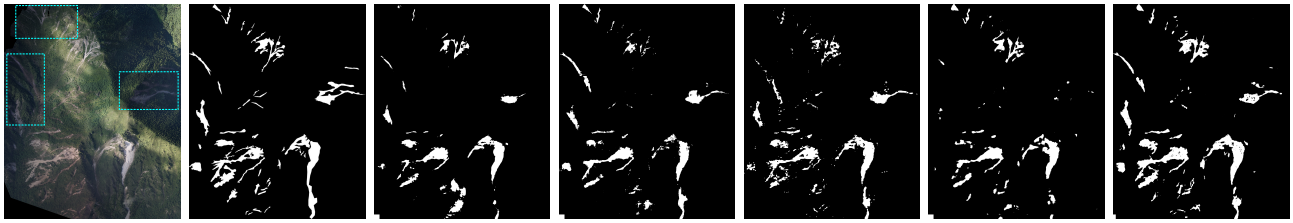
In the following, we use the prefix "D" to denote the dilated convolution and "A" to denote the attention module, such as D-U-Net augmented with only dilated convolution and DA-U-Net with both modified modules. For comparison, we also reimplemented several publicly available methods, such as the FCN [9] with VGG-16 [24] as backbone, PSPNet [25] with ResNet-101 [23] as backbone, the latest DeepLabV3+ [11] with ResNet-101 as backbone and the vanilla U-Net [12].

Fig. 3 compares the performances of the methods above in assessing a typical scene. Notably, in the shadow (eclipse region), almost all of the methods fail to identify the landslide region; this situation could only be improved with use of the attention module, e.g., DA-U-Net. Another interesting finding is that only architectures with a pyramid strategy can satisfactorily identify landslide regions in confusing areas comprising both landslides and bare earth (as indicated by the rectangle). In summary, the proposed methods gave the best overall segmentation results.

Turning to quantitative evaluations, Table I demonstrates the Intersection of Union (IoU), precision, recall and F-scores for all of the methods. The proposed DA-U-Net exceeds the performance of the second-best method, *i.e.* DeeplabV3+ [11], in the most concerned IoU metric and also for recall rate and F-score. Considering that the performance of the current state-of-the-art DeeplabV3+ is also on par with the PSPNet, the modifications on the U-Net were thus effective.

TABLE I: Quantitative comparisons of different methods. The bold cells denote the methods with best performances.

Method	IoU	Precision	Recall	F-Score
FCN	48.15	75.96	56.81	65.00
U-Net	48.18	75.42	57.15	65.03
PSPNet	52.63	77.56	62.08	68.97
DeeplabV3+	57.25	**79.6**	67.1	72.81
DA-U-Net	**59.41**	70.06	**79.62**	**74.54**



(a) Orthophoto (b) Label (c) FCN (d) PSPNet (e) DeeplabV3+ (f) U-Net (g) DA-U-Net

Fig. 3: Qualitative comparisons with other methods.

B. Study of confusing areas

For efficient hazard mitigation and rapid emergency response, the most critical landslide regions that require precision detection are those that occur in confusing areas. As most landslide regions in the training were in non-confusing regions, we privileged confusing areas by choosing imbalanced samples that preferred the confusing regions. Fig. 4 shows two typical confusing areas, namely roads (top) and bare earth (bottom), with the interesting regions in each row highlighted with cyan polygons. As these regions are hard to distinguish without inferring from a large context, it is almost impossible to identify these confusing areas without fusion of multiscale features. Unfortunately, FCNs have no mechanism of handling multiscale features and therefore would be expected to have inferior performance. U-Net propagates and aggregates limited multiscale features by down-sampling, which means that it cannot go too deep in the contextual information without loss of spatial resolution. PSPNet constructs the spatial pyramid and fuses different layers and DeeplabV3 embeds the ASPP module for multiscale features; both of these can exploit larger contextual information without loss of spatial resolution, but do not sufficiently preserve fine-grained structures. The proposed DA-U-Net has the best performances, thanks to its ASPP module and attention-guided up-sampling.

C. Ablation studies

The DA-U-Net augments the vanilla U-Net with two modules: 1) the dilated convolution and ASPP in the bottleneck for the exploitation of larger contextual information; and 2) the attention module for guided up-sampling. Table II compares different variants against the vanilla U-Net, namely D-U-Net, A-U-Net and DA-U-Net. Notably, it can be seen that both the dilation and attention modules clearly and extensively improve the overall performance.

Fig. 5 compares different U-Net architectures in three confusing areas with roads. Both the D-U-Net and A-U-Net show improved detection of false landslide regions due to interference from roads, and attention-guided sampling shows better efficiencies than the dilation module and ASPP bottleneck. This is also consistent with the quantitative evaluations shown in Table II. Thus, in combination, these two modules demonstrated the best overall performances.

TABLE II: Ablation studies of different modules based on U-Net. The bold cell denotes the best performance.

Method	Attention Module	Dilated Convolution+ASPP	IOU
U-Net	×	×	48.18
D-U-Net	×	✓	54.61
A-U-Net	✓	×	52.44
DA-U-Net	✓	✓	**59.41**

IV. Conclusion

In this study we have proposed and developed an improved landslide inventory mapping methods that were based on augmenting the U-Net structure, *i.e.* the dilated convolution [10] and attention-guided up-sampling [15], [16]. As landslide or probable landslide regions generally co-exist with regions that also have similar spectral to landslides, contextual information should be considered to remove the feature-ambiguities produced by CNNs. The two modules are designed to fuse both local and non-local information to alleviate this issue. The two modules were used to simultaneously enlarge the receptive field of local convolution and preserve dense high-resolution feature maps. Future work may be devoted to the combined use of orthophotos and digital elevation models for more accurate and robust landslide mapping. In addition, pre-hazard susceptibility mapping of landslide-prone regions [26] is also crucial for hazard mitigation. The code corresponding to this paper is made publicly available[1].

References

[1] T. R. Martha, N. Kerle, C. J. van Westen, V. Jetten, and K. V. Kumar, "Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 12, pp. 4928–4943, 2011.

[2] J. Dou, K.-T. Chang, S. Chen, A. Yunus, J.-K. Liu, H. Xia, and Z. Zhu, "Automatic case-based reasoning approach for landslide detection: integration of object-oriented image analysis and a genetic algorithm," *Remote Sensing*, vol. 7, no. 4, pp. 4318–4342, 2015.

[3] A. Stumpf and N. Kerle, "Combining random forests and object-oriented analysis for landslide mapping from very high resolution imagery," *Procedia Environmental Sciences*, vol. 3, pp. 123–129, 2011.

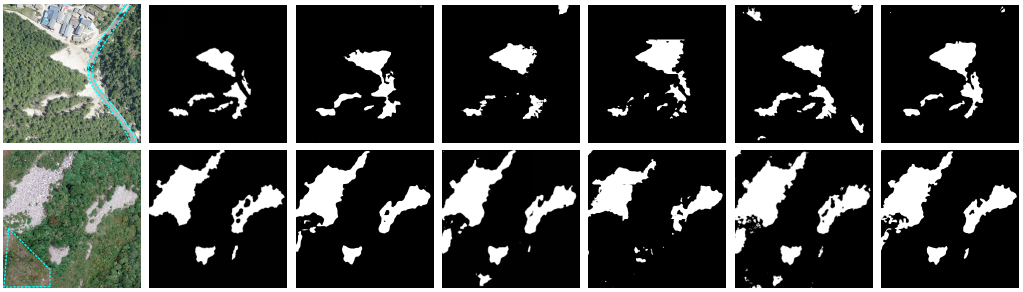
[4] J.-Y. Rau, J.-P. Jhan, and R.-J. Rau, "Semiautomatic object-oriented landslide recognition scheme from multisensor optical imagery and dem," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 2, pp. 1336–1349, 2013.

[5] A. Othman and R. Gloaguen, "Automatic extraction and size distribution of landslides in kurdistan region, ne iraq," *Remote Sensing*, vol. 5, no. 5, pp. 2389–2410, 2013.

[6] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS journal of photogrammetry and remote sensing*, vol. 65, no. 1, pp. 2–16, 2010.

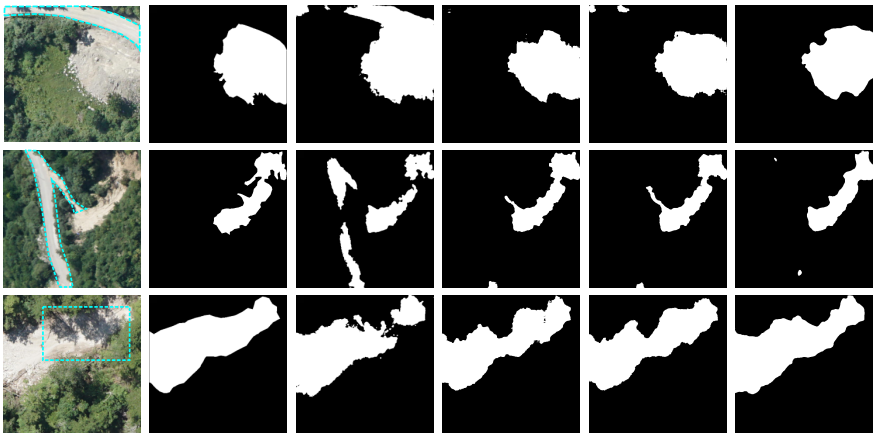
[7] R. N. Keyport, T. Oommen, T. R. Martha, K. Sajinkumar, and J. S. Gierke, "A comparative analysis of pixel-and object-based detection of landslides from very high-resolution images," *International journal of applied earth observation and geoinformation*, vol. 64, pp. 1–11, 2018.

[1] https://github.com/saedrna/DA-U-Net



(a) Orthophoto (b) Label (c) FCN (d) PSPNet (e) DeeplabV3+ (f) U-Net (g) DA-U-Net

Fig. 4: Performances of different methods on typical confusing areas. The highlighted are common confusing regions.



(a) Image (b) Label (c) U-Net (d) D-U-Net (e) A-U-Net (f) DA-U-Net

Fig. 5: Ablation studies of the DA-U-Net architecture. The highlighted are regions prone to errors.

[8] N. Casagli, F. Cigna, S. Bianchini, D. Hölbling, P. Füreder, G. Righini, S. Del Conte, B. Friedl, S. Schneiderbauer, C. Iasio *et al.*, "Landslide mapping and monitoring by using radar and optical remote sensing: Examples from the ec-fp7 project safer," *Remote sensing applications: society and environment*, vol. 4, pp. 92–108, 2016.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[13] T. Lei, Y. Zhang, Z. Lv, S. Li, S. Liu, and A. K. Nandi, "Landslide inventory mapping from bitemporal images using deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 6, pp. 982–986, 2019.

[14] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[15] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.

[16] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[17] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, "Learn to pay attention," *arXiv preprint arXiv:1804.02391*, 2018.

[18] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[19] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[21] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[22] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[25] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[26] B. Pradhan, E. A. Sezer, C. Gokceoglu, and M. F. Buchroithner, "Landslide susceptibility mapping by neuro-fuzzy approach in a landslide-prone area (cameron highlands, malaysia)," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 12, pp. 4164–4177, 2010.