On the Generalization of the Semantic Segmentation Model for Landslide Detection

Fahong Zhang[1], Yilei Shi[2], Qingsong Xu[1], Zhitong Xiong[1], Wei Yao[3] and Xiao Xiang Zhu[13]

[1]*Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany*
[2]*Chair of Remote Sensing Technology (LMF), Technical University of Munich, Munich, Germany*
[3]*Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany*

Abstract

The goal of landslide detection is to detect regions with landslide events. It is critical for emergency response and disaster monitoring. This study is based on the context of Landslide4Sense competition, whose goal is to promote effective and innovative algorithms to detect landslides across different continents, using Sentinel-2 and ALOS PALSAR data. Considering its global-scale coverage, studying the generalization performance of the landslide detection model on unseen regions turns out to be an important task. To this end, we propose a self-training method to improve the generalizability of the landslide detection model by exploiting the pseudo labels of unlabeled samples with low uncertainty. According to experimental results, the proposed self-training method is effective in bridging the shifts between labeled and unlabeled data, and achieves the rank of the 3rd place on the Landslide4Sense competition.

Keywords

Landslide detection, Semantic segmentation, Self-training, Domain adaptation,

1. Introduction

With the ongoing climate change and the rapid urbanization in landslide-prone terrains, Landslides have become an increasingly threatening hazard in mountainous areas and started to affect a large amount of population. In order to accurately and rapidly monitor the landslide events occurred over the world, satellite data are considered as a promising data source owing to their high global coverage, relatively high temporal and spectral resolution.

In a technical point of view, the landslide detection problem based on satellite data can be regarded as a binary semantic segmentation problem, where the learning based model is required to distinguish the landslides with background areas. In the computer vision society, semantic segmentation has always been a popular research topic. From the earlier Fully Convolution Network (FCN) [1, 2] to the currently dominating transformer-based approaches [3, 4], tremendous improvements have been witnessed with the developments of the network architecture. As reported in [5], several baseline semantic segmentation models have demonstrated promising performances in the task of landslide detection.

In addition to designing more sophisticated and task specific network architectures, the research towards the transferability of semantic segmentation model is also of great importance. Due to the different atmospheric conditions, shooting angles and illuminations, satellite data across different regions may have large domain shifts [6]. As a result, the semantic segmentation model trained on specific areas may fail to generalize to different unseen regions across the world in different periods of time.

Self-training approaches have been demonstrated to be effective in promoting the generalizability of deep learning models in the field of semi-supervised learning and domain adaptation [7]. They first generate pseudo labels on the unlabeled data based on a teacher model pre-trained on labeled data. Then the pseudo labels with high confidence will be used to supervise the training of the student model on the unlabeled data. With this considered, we propose a self-training method based on a Monte-Carlo dropout uncertainty [8] and class-balanced thresholding. The contributions of this paper can be listed as follows:

- We propose a self-training method based on Monte-Carlo dropout uncertainty and class-balanced thresholding on the task of landslide detection. The experimental results demonstrate that the proposed method can provide significant improvements over the baseline, and help to improve the generalizability of semantic segmentation models.
- We technically prove the effectiveness of the proposed method on Landslide4Sense competition, where we achieve the 3rd prize with a testing F1 score of 73.50%.

CDCEO 2022: 2nd Workshop on Complex Data Challenges in Earth Observation, July 25, 2022, Vienna, Austria

fahong.zhang@tum.de (F. Zhang); yilei.shi@tum.de (Y. Shi); qingsong.xu@tum.de (Q. Xu); zhitong.xiong@tum.de (Z. Xiong); wei.yao@dlr.de (W. Yao); xiaoxiang.zhu@tum.de (X. X. Zhu)

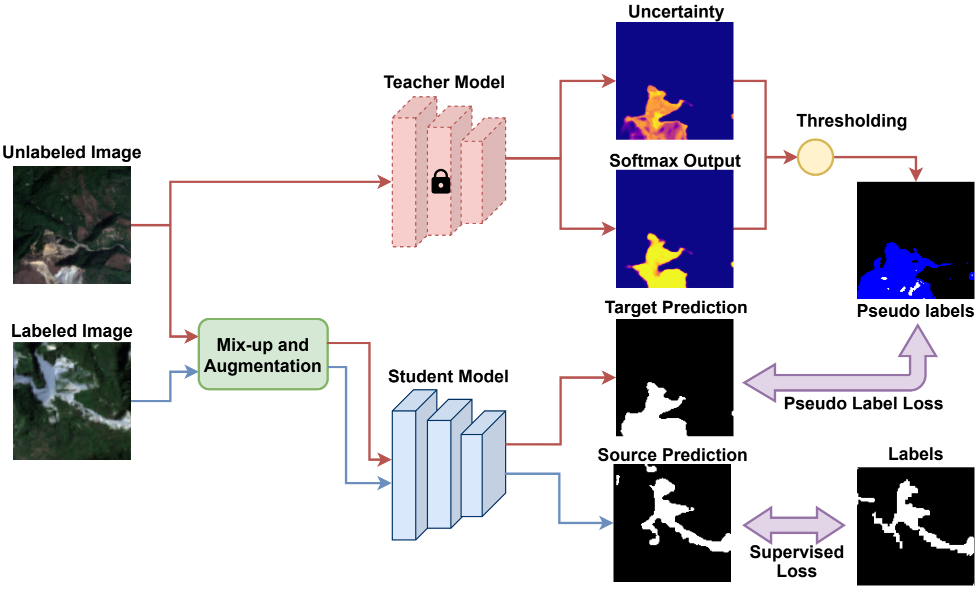

Figure 1: Pipeline of the proposed self-training method. In each training step, a batch of labeled and unlabeled data will be given to the teacher and the student models, where data augmentations and mix-up operation [9] will be applied to the student model branch. For labeled data, supervised losses will be calculated based on the provided labels. For unlabeled data, we first apply Monte-Carlo dropout [8] on the teacher model to estimate the uncertainty of unlabeled predictions, and then generate the pseudo labels based on a class-balanced threshold (see 2.3). The teacher model will be fixed during training.

2. Methodology

We illustrate the pipeline of the proposed method in Fig. 1. The remaining parts of this section will formulate the landslide detection problem and elaborate the methodology in details.

2.1. Problem Formulations

In the landslide detection problem, we are given a set of labeled training data $\mathcal{D}_{train} = \{x_{tr}, y_{tr}\}$, and unlabeled test data $D_{test} = \{x_{te}\}$, where x_{tr}, y_{tr}, and $x_{te} \in \mathbb{R}^{H\times W}$ are each training patch, training label, and test patch, respectively. Our task is to train a semantic segmentation model on D_{train} and D_{test}, and optimize its performance on $\mathcal{D}_{test}$. The overall loss function of the proposed method is:

$$\mathcal{L} = \mathcal{L}_{sup}^{mix} + \mathcal{L}_{pse}^{mix}. \quad (1)$$

The mix supervised loss $\mathcal{L}_{sup}^{mix}$ and pseud label loss $\mathcal{L}_{pse}^{mix}$ will be formulated in Sec. 2.4

2.2. Supervised Losses

We use cross entropy loss and jaccard loss as the supervised losses:

$$\mathcal{L}_{sup}(x_{tr}, y_{tr}) = \mathcal{L}_{cet}(x_{tr}, y_{tr}) + \mathcal{L}_{jac}(x_{tr}, y_{tr}). \quad (2)$$

2.3. Self-training

As shown in Fig. 1, a teacher model pre-trained on the training data will be used to generate pseudo labels for supervising the student model. However, since the raw pseudo labels are usually noisy, a selection strategy is required to filter out the misclassified pixels.

First, we use the Monte-Carlo dropout strategy [8] to estimate an uncertainty map for each input test patch. More specifically, we forward the test patch to the source model with 10 different runs. In each run, random dropout with 0.3 dropping rate will be applied to the feature map obtained by the first convolution layer. The variances of 10 different output logits will be considered as the uncertainty map.

Second, we mask out the uncertain predictions from the teacher model. Inspired by [7], we propose to select a certain proportion of the pixels for each class with the lowest uncertainty among all the test data. To this end, 90% of the background pixels and 70% of the landslide pixels are utilized, and the others will be ignored when calculating the losses. Finally, the pseudo label loss can be formulated by:

$$\mathcal{L}_{pse}(x_{te}, \hat{y}_{te}) = \mathcal{L}_{cet}(x_{te}, \hat{y}_{te}) + \mathcal{L}_{jac}(x_{te}, \hat{y}_{te}). \quad (3)$$

Here $\hat{y}_{te}$ corresponds to the pseudo labels generated by the teacher model.

2.4. Mix-up Strategy

To prevent the model from overfitting to the training data, a mix-up strategy [9] is applied to both the training and test data to further increase the generalizability. Given a batch of training and test data, the mixed data can be generated by:

$$\begin{aligned} \tilde{x}_{tr} &= \lambda x_{tr} + (1-\lambda)x_{tr}^{'}, \\ \tilde{x}_{te} &= \lambda x_{te} + (1-\lambda)x_{te}^{'}. \end{aligned} \tag{4}$$

Here $x_{tr}^{'}$ is derived from x_{tr}, where all the image patches in the same batch are shuffled. λ is a scalar randomly sampled from a predefined beta distribution during training. Then we can reformulate the supervised and pseudo label losses as:

$$\begin{aligned} \mathcal{L}_{sup}^{mix} &= \lambda \mathcal{L}_{sup}(\tilde{x}_{tr}, y_{tr}) + (1-\lambda)\mathcal{L}_{sup}(\tilde{x}_{tr}, y_{tr}^{'}), \\ \mathcal{L}_{pse}^{mix} &= \lambda \mathcal{L}_{pse}(\tilde{x}_{te}, \hat{y}_{te}) + (1-\lambda)\mathcal{L}_{pse}(\tilde{x}_{te}, \hat{y}_{te}^{'}). \end{aligned} \tag{5}$$

2.5. Post-processing

We apply the dense conditional random field (DenseCRF) [10] as a post-processing technique to better match the predicted landslide contours with the ground truth.

3. Experiments

3.1. Datasets

The proposed method is developed and evaluated on the Landslide4Sense competition [5]. The provided data consist of 12 Sentinel-2 bands and 2 topological bands including SLOP and DEM, both of which are derived from ALOS PALSAR. Each band is resized to 10 meter resolution per pixel. The data are cropped to 128 × 128 patches. 3799, 245 and 800 patches are provided for training, validation and testing, respectively.

3.2. Implementation Details

For the overall training setting, we use SGD optimizer with Nesterov acceleration to train the network, where the momentum and weight decay are set to 0.9 and 5×10^{-4}, respectively. The batch size is set to 16, and the training lasts for $60,000$ iterations. For data pre-processing, we normalize the first 12 bands by linearly scaling them to the range of $[0, 1]$. For data augmentation, we perform random flipping, random resizing and cropping, and finally resize the patch to the size of 256×256.

The time period of the Landslide4Sense competition includes a validation phase and a test phase. During the validation phase, only validation data are released. During the test phase, the test data will be available, yet the chances for submitting the results for evaluation will be limited. With this as background information, we give the workflow of training our final model as follows.

- **Model** 1. We first train a base model using solely the training data, which means the teacher branch in Fig. 1 is blocked. ResNet50 [11] and Deeplab V3+ [12] are used as the backbone and the decoder, respectively. The ResNet50 backbone is initialized using the ImageNet pretrained weights. The training lasts for only $30,000$ iterations to avoid overfitting.
- **Model** 2. This model is developed during the validation phase, where we use **Model** 1 as the teacher model, and validation data as the unlabeled data. The architecture is based on HRNet [13].
- **Model** 3. Compared to **Model** 2, the only difference of **Model** 3 is that we apply a ResNext50 [14] backbone and a Deeplab V3+ [12] architecture.
- **Final Model**. The final model uses all the validation and test data as unlabeled data. Following Fig. 1, its student model is pre-trained on **Model** 3, and **Model** 2 is considered as the teacher model.

3.3. Results

The final results on the test leaderboard are shown in Tab. 1. For our methods, we plot the results of the **Final Model** and **Model** 2. Due to the limited submission times, the other models were not evaluated. By comparing the results of **Model** 2 to **Final Model**, one can observe that pre-training on a different architecture (**Model** 3) helps to improve the performance of the **Final Model**.

Some qualitative results on the testing data are shown in Fig. 2. According to the results, the proposed method can successfully distinguish the road areas with the landslides, which are similar to each other in RGB appearances. However, some small landslides that fall to the road are also ignored (see the first two rows). By comparing the raw predictions and the post-processed results, we notice that DenseCRF will remove some isolated landslide predictions, but help to shrink them to better fit to the spatial topology (see red rectangles in the last column).

3.4. Ablation Study

We perform the ablation study based on the validation data and list the results in Tab. 2. It can be observed that both **Model** 2 and **Model** 3 are superior to **Model** 1 by a large margin. In addition, if the self-training branch is

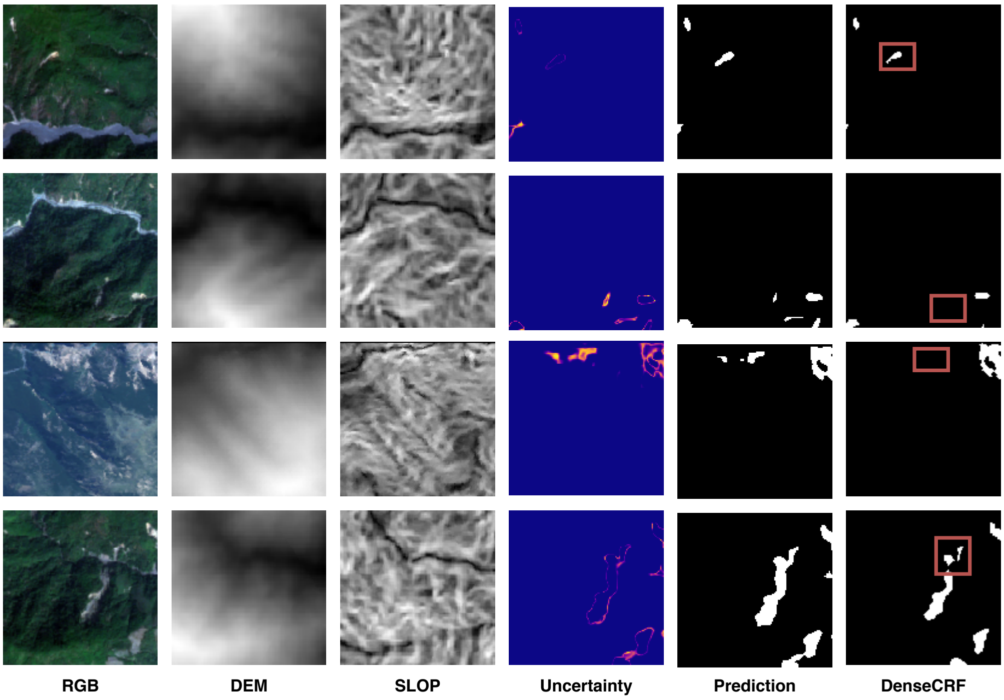

Figure 2: Qualitative results on test data. From left to right columns, we visualize the RGB, DEM and SLOP channels of the data, MC-dropout-based uncertainty maps, predictions from the network and the post-processed results by DenseCRF.

Table 1
F1 score (%) during the test phase.

Team Name	F1
kingdrone	74.54
seek	73.99
ours (**Final Model**)	73.50
ours (**Model 2**)	72.50
sikui	71.87
sklgp	71.29
bao18	70.15

Table 2
Ablation study results during the validation phase (%). "w/o ST" means the self-training or the teacher model branch in Fig. 1 is blocked. "CRF" means DenseCRF is activated as the post-processing method.

Model	Precision	Recall	F1
Model 1	69.70	**82.60**	75.60
Model 1 + CRF	76.82	80.48	78.61
Model 2 (w/o ST)	66.96	81.23	73.41
Model 2	75.60	82.21	78.76
Model 2 + CRF	**82.45**	78.36	**80.35**
Model 3 (w/o ST)	65.63	82.31	73.03
Model 3	73.89	82.34	77.88
Model 3 + CRF	80.19	78.94	79.56

blocked, the performance will be decreased. This demonstrates the effectiveness of the proposed self-training method.

4. Conclusions

This paper studies the landslide detection problem and propose a self-training method to improve the generalizability of the semantic segmentation model. The experimental results on Landslide4Sense dataset demonstrate that the proposed method can help to bridge the gap between labeled and unlabeled data.

Acknowledgments

This work is sponsored by China Scholarship Council.

References

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

[2] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention - MICCAI, volume 9351, 2015, pp. 234–241.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[4] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, Segformer: Simple and efficient design for semantic segmentation with transformers, Advances in Neural Information Processing Systems 34 (2021) 12077–12090.

[5] O. Ghorbanzadeh, Y. Xu, P. Ghamis, M. Kopp, D. Kreil, Landslide4sense: Reference benchmark data and deep learning models for landslide detection, arXiv preprint arXiv:2206.00515 (2022).

[6] O. Tasar, A. Giros, Y. Tarabalka, P. Alliez, S. Clerc, Daugnet: Unsupervised, multisource, multitarget, and life-long domain adaptation for semantic segmentation of satellite images, IEEE Transactions on Geoscience and Remote Sensing 59 (2020) 1067–1081.

[7] Y. Zou, Z. Yu, B. Kumar, J. Wang, Unsupervised domain adaptation for semantic segmentation via class-balanced self-training, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 289–305.

[8] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, PMLR, 2016, pp. 1050–1059.

[9] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, arXiv preprint arXiv:1710.09412 (2017).

[10] P. Krähenbühl, V. Koltun, Efficient inference in fully connected crfs with gaussian edge potentials, Advances in neural information processing systems 24 (2011).

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

[12] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587 (2017).

[13] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., Deep high-resolution representation learning for visual recognition, IEEE transactions on pattern analysis and machine intelligence 43 (2020) 3349–3364.

[14] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.