IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9884536

SEMI-SUPERVISED LAND-COVER MAPPING BASED ON MULTIMODAL FUSION AND PSEUDO-LABEL

*Yi Gao, Xingyu Ding, Guangyi Yang**

School of Electronic Information, Wuhan University, Wuhan 430079, P.R. China
*Corresponding author: ygy@whu.edu.cn

ABSTRACT

Land-cover mapping is of great significance for remote sensing and earth observation. However, due to the high cost of label acquisition, how to use limited labeled samples and multimodal data to achieve large-scale and high-precision land-cover mapping is still a great challenge. In this paper, a multimodal fusion and pseudo-label based method is proposed for semi-supervised land-cover mapping (SLM). For the problem of domain incompatibility, we use strong data enhancement and multimodal fusion module to strengthen the generalization performance of the method from data level and model level respectively. For a large amount of unlabeled data, we combine the pseudo-label self-training technology and propose Fusion-Finetune-Fusion training strategy to achieve large-scale, high-precision land-cover mapping under semi-supervised conditions. In the track SLM of the 2022 Data Fusion Contest (DFC22-SLM), the proposed method achieves a mean intersection over union (mIoU) of 0.4962 in phase 2, ranking fourth place.

Index Terms—Semi-supervised, Land-cover mapping, Multimodal fusion

1. INTRODUCTION

Land-cover mapping is one of the important issues of earth observation which is crucial to researches such as environmental monitoring, resource survey and urban planning. With the rapid development of remote sensing technology, the acquisition of remote sensing images has become easier, and the spatial resolution has become higher. While the images contain more information, it also brings more cases of the same object with different spectrum and the same spectrum of different objects. It brings greater challenges to land-cover mapping. The rise of deep learning technology has provided powerful means for land-cover mapping. However, deep learning-based land-cover mapping methods need a large amount of labeled remote sensing data. Thus, how to use the large amount of unlabeled data to achieve high-precision land-cover mapping method is still a huge challenge.

Accordingly, the Image Analysis and Data Fusion Technical Committee (IADF TC) of the IEEE GRSS, Universit´e Bretagne-Sud, ONERA, and ESA ϕ-lab organize the DFC22-SLM competition[1]. This competition provides participants with datasets dedicated to semi-supervised semantic segmentation[2,3], which are divided into labeled data and unlabeled data. Both types of data are provided with RGB data and DEM (digital elevation model) data. RGB aerial images are from the French National Institute of Geographical and Forest Information (IGN) BD ORTHO database. The DEM tiles corresponding to each image are from the IGN RGE ALTI database. It is worth noting that the size of the RGB image is 2000×2000 with a resolution of 0.5m, from 2012 to 2014. The size of the DEM is 1000×1000 with a resolution of 1m, acquired in 2019 and 2020. The size of the label is the same as RGB image, and 14 land-use classes are considered corresponding to the second level of the semantic hierarchy defined by UrbanAtlas. In general, the task is to use a small amount of labeled data and a large amount of unlabeled data to make a land-cover map of a new area.

For this task, we propose a semi-supervised land-cover mapping method based on multimodal fusion and pseudo-label. Firstly, we combine strong data augmentation and weak data augmentation techniques to increase the data sample as well as enhance the robustness of the model from the data level. Secondly, a multimodal fusion module is designed to convert DEM data into an attention map, so that the feature map of each level of the single model is weighted. The additional information of DEM is used to improve the model's ability to discriminate ground objects especially the similar objects of different heights. Finally, we predict a large amount of unlabeled data with the two trained models, and fuse the results of the two as pseudo-label. Use pseudo-label[4] to finetune the trained single model separately, and finally fuse the finetuned models to obtain a high-precision land-cover map. The experiment and benchmark test results show the effectiveness of the method, and our method ranks fourth place in the testing phase with a mIoU of 0.4962.

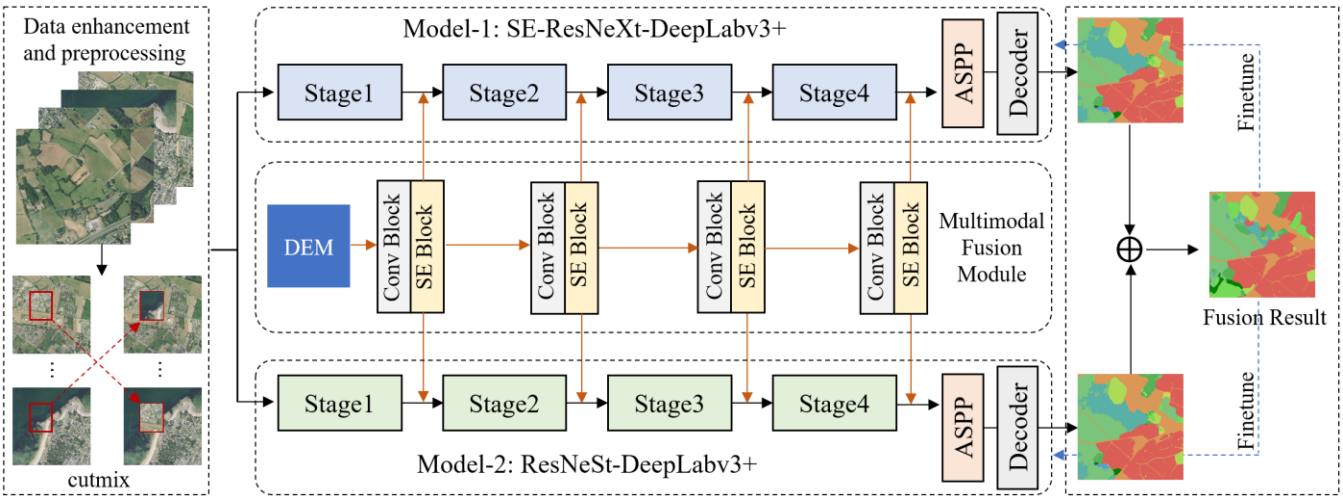


Fig. 1 Flow chart of semi-supervised land-cover mapping method based on multimodal fusion and pseudo-label

2. METHODOLOGY

In this section, the proposed method is introduced, and the flow chart is shown in Fig. 1. It will be introduced in three parts: data enhancement, multimodal fusion module, and Fusion-Finetune-Fusion strategy.

2.1. Data enhancement

In order to extract as much information as possible from the limited labeled data and prevent overfitting, we use both weak data augmentation and strong data augmentation.

Weak data enhancement method is shown in Fig. 2. It includes randomly cropping the 2000×2000 original image into 512×512 input patches with a step size of 128 pixels, which increase the number of training samples and reducing the training cost. On this basis, the patches are randomly flipped, rotation combined to further enrich the training data. In order to capture the features of different scales, we use multi-scale training techniques. The input patches are randomly scaled according to the ratio of 0.5~2.0. Finally, in order to further reduce the influence of different time of samples on the final result, we add color jitter to improve the robustness and generalization ability of the model.

Strong data enhancement method mainly uses cutmix[5] which combines the advantages of cutout and mixup to enhance the robustness of the model. Cutout can make the model focus on the areas where the target is difficult to distinguish. But it introduces non-pixel information, which affects the training efficiency. Although mixup makes full use of pixel information, it introduces unnatural pseudo-pixel information. It is worth mentioning that although images like cutmix do not appear in the real world, its consistency regularization achieves an important breakthrough for semi-supervised semantic segmentation tasks.

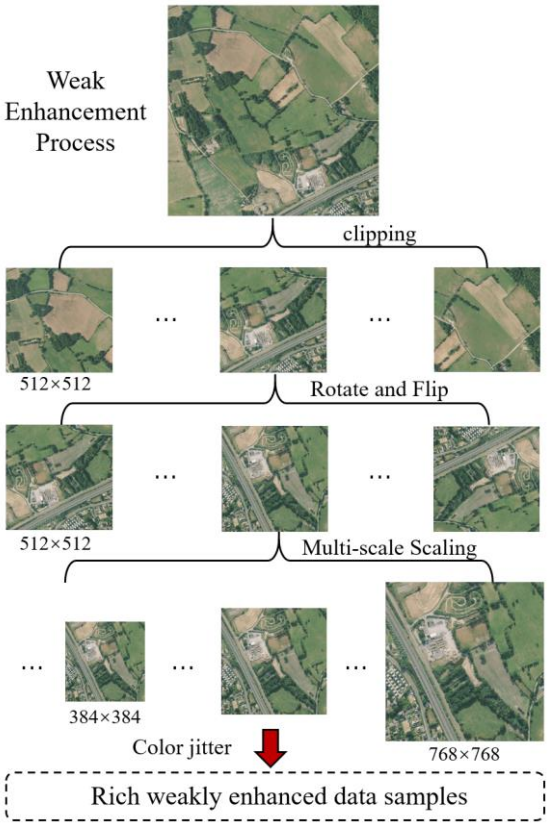


Fig. 2 Weak data augmentation flow chart

2.2. Multimodal Fusion Module

Since DEM data can reflect local topographic features with a certain resolution, a large amount of surface morphological information can be extracted through DEM. In order to further improve the feature discrimination ability of the model, we make full use of DEM data provided by

the organizer. The Multimodal Fusion Module (MFM) is designed as the DEM branch of the network, focusing on solving the problem that the network is insensitive to height information in RGB image. This module uses the SE (Squeeze-and-excitation) module[6] to convert the height feature information provided by DEM data into attention, and further adjust the weights on the feature map of the corresponding RGB image to obtain more discriminative features.

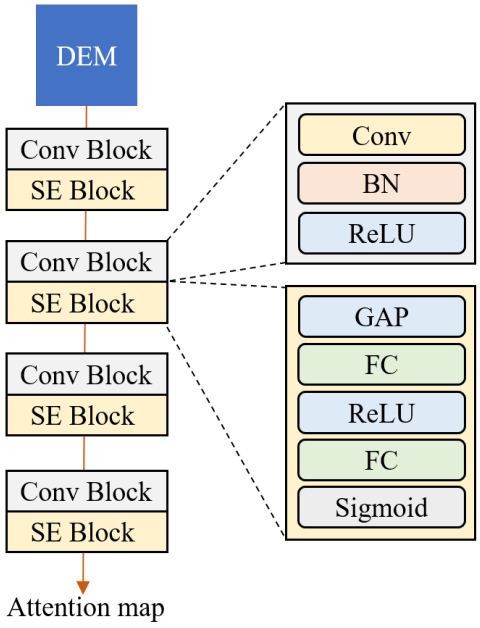


Fig. 3 Multimodal fusion module

The specific structure of the multimodal fusion module is shown in Fig. 3. The module is divided into 4 stages, corresponding to the backbone of deeplabv3+. Each stage is consisted of convolution block and SE block. The convolution block is composed of conv, BN (batch normalization layer), and ReLU, so the high-level semantic features of DEM data are gradually extracted. The SE block is composed of GAP (global average pooling), FC (fully connected layer), ReLU, FC and sigmoid. The implicit information in DEM data is used as the weight of the backbone to redistribute attention, strengthen important information and suppressing redundant information. Finally, the multimodal information is fully utilized, and the extracted features are more abundant.

2.3. Fusion-Finetune-Fusion

For a large amount of unlabeled data, we combine the pseudo-label self-training technology and design a Fusion-Finetune-Fusion training strategy to achieve semi-supervised land-cover mapping method. The specific process is as follows:

a. Model fusion to produce pseudo-label: Through the introduction of the previous two sections, a single model (i.e., SE-ResNeXt, ResNeSt) has been trained well by using RGB images and DEM data with labels. The two trained single models predict unlabeled data separately, and fuse the results in the form of soft voting to obtain pseudo-label.

b. Finetune the single model with pseudo-label: Use the unlabeled data and the corresponding pseudo-label obtained in the previous step to finetune the two trained single models separately, thereby further enhancing the generalization ability of each model.

c. Single-model fusion achieves better results: Re-predict the test data by the finetuned single model and fuse the prediction results again. Finally, we can obtain better-performing land-cover mapping results.

3. EXPERIMENT

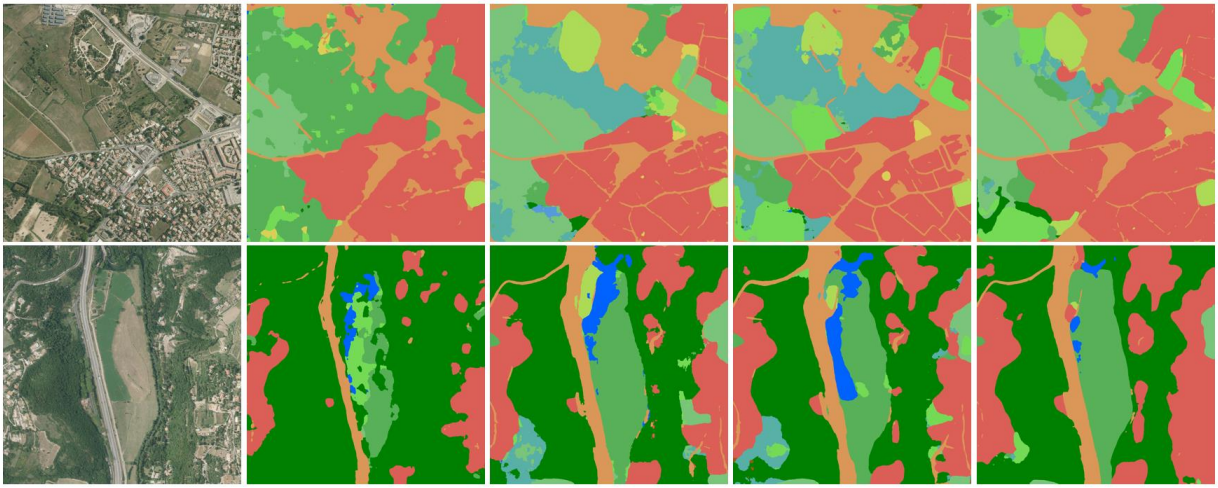
In this section, we list the experiment results of the ablation of our method in Table 1, and show the corresponding visual results in Fig. 4.

Table 1. Experimental setup and mIoU

Method	W/S	DEM	Finetune	mIoU
Baseline	N/N	N	N	0.1278
DeepLabv3+	Y/N	N	N	0.2781
ResNeSt	Y/N	N	N	0.3023
SE-ResNeXt	Y/N	N	N	0.3273
SE-ResNeXt[1]	Y/Y	N	N	0.3908
ResNeSt[1]	Y/Y	Y	N	0.4271
SE-ResNeXt[2]	Y/Y	Y	N	<u>0.4283</u>
Fusion	Y/Y	Y	Y	**0.4962**

Note: W represents weak enhancement; S represents strong enhancement; Y represents yes; N represents No; SE-ResNeXt and ResNeSt represent Model-1 and Model-2 in Fig. 1, respectively; Fusion refers to the final fusion model; Best results, next best results are highlighted in bold and underlined, respectively.

From the visual comparison of the mIoU score in Table 1 and Fig. 4, it can be seen that conventional training methods such as data enhancement and loss function weighting can increase the officially provided baseline from 0.1278 to 0.2781. At the same time, replacing a more powerful backbone can make the network achieves better feature extraction ability and achieves an increase in mIoU score of 0.03-0.05. The addition of cutmix improves the generalization performance of the model from the data level, and can solve the problems of different time and different domain to a certain extent. On this basis, the MFM we designed integrates DEM data well, and uses multimodal information to enhance the model's ability to discriminate ground objects, and also achieves a score improvement of about 0.03. Finally, the method of model fusion is used to



(a) Image (b) DeepLabv3+ (c) SE-ResNeXt[1] (d) SE−ResNeXt[2] (e) Fusion

Fig. 4 Visual results comparison of test areas

create pseudo-label, and the unlabeled data with pseudo-label are used to further finetune the single model, so as to achieve better fusion prediction results and obtain higher mIoU scores. It can also be seen from the details of the visual effect in Fig. 4 that our method can classify the surface more accurately and greatly reduce the misclassification.

4. CONCLUSION

In this paper, we propose a semi-supervised land-cover mapping method that combines multimodal and pseudo-label to solve the problem of accurate land-cover mapping under limited data. Through the data enhancement of cutmix and the skills of multimodal fusion, the feature discrimination ability and robustness of the model are further improved. The model fusion result is used as a self-training method for pseudo-label, making full use of unlabeled data to realize semi-supervised land-cover mapping. The efficiency of the method is confirmed by the experimental results obtained on the benchmark dataset of 2022 DFC-SLM.

5. ACKNOWLEDGMENT

The authors would like to thank the IEEE GRSS Image Analysis and Data Fusion Technical Committee, Universit´e Bretagne-Sud, ONERA, and ESA ϕ-lab for organizing the Data Fusion Contest. This work is supported in part by the National Natural Science Foundation of China (No.61871298 and 42071322), in part by the Natural Science Foundation of Hubei Province (No. 2020CFA053), and in part by the Wuhan Application Foundation Frontier Project (No. 2020010601012184).

6. REFERENCES

[1] "2022 IEEE GRSS Data Fusion Contest. Online:,"www.grss-ieee.org/technical-committees/image-analysis-and-data-fusion/.

[2] Castillo-Navarro, J., Le Saux, B., Boulch, A. and Lefèvre, S.. Semi-supervised semantic segmentation in Earth Observation: the MiniFrance suite, dataset analysis and multi-task network study. Mach Learn (2021). doi.org/10.1007/s10994-020-05943-yNaoto Yokoya, Pedram Ghamisi, Ronny Hansch et al. "2021 Data Fusion Contest: Geospatial Artificial Intelligence for Social Good [Technical Committees]," *in IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 1, pp. 287-C3, 2021.

[3] Castillo-Navarro, J., Le Saux, B., Boulch, A. and Lefèvre, S.. Semi-supervised semantic segmentation in Earth Observation: the MiniFrance suite, dataset analysis and multi-task network study. Mach Learn (2021).doi.org/10.1007/s10994-020-05943-y

[4] Zhuohong Li, Fangxiao Lu, Hongyan Zhang, ..., Caleb Robinson, Nikolay Malkin, Nebojsa Jojic, Pedram Ghamisi, Ronny H¨ansch, and Naoto Yokoya, "The outcome of the 2021 IEEE GRSS Data Fusion Contest—Track MSD: multitemporal semantic change detection," IEEE Journal of Selected Topics inApplied Earth Observations and Remote Sensing, vol. 15, pp.1643–1655, 2022.

[5] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo and J. Choe, "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6022-6031, doi: 10.1109/ICCV.2019.00612.

[6] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, "Squeeze-and-Excitation Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 8, pp. 2011-2023, 1 Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.