Multimodal Co-learning: A Domain Adaptation Method for Building Extraction from Optical Remote Sensing Imagery

1st Yuxing Xie
Remote Sensing Technology Institute (IMF)
German Aerospace Center (DLR)
82234 Wessling, Germany
yuxing.xie@dlr.de
ORCID: 0000-0002-6408-5109

2nd Jiaojiao Tian
Remote Sensing Technology Institute (IMF)
German Aerospace Center (DLR)
82234 Wessling, Germany
jiaojiao.tian@dlr.de
ORCID: 0000-0002-8407-5098

***Abstract*—In this paper, we aim to improve the transfer learning ability of 2D convolutional neural networks (CNNs) for building extraction from optical imagery and digital surface models (DSMs) using a 2D-3D co-learning framework. Unlabeled target domain data are incorporated as unlabeled training data pairs to optimize the training procedure. Our framework adaptively transfers unsupervised mutual information between the 2D and 3D modality (i.e., DSM-derived point clouds) during the training phase via a soft connection, utilizing a predefined loss function. Experimental results from a spaceborne-to-airborne cross-domain case demonstrate that the framework we present can quantitatively and qualitatively improve the testing results for building extraction from single-modality optical images.**

***Index Terms*—building extraction, multimodal data, co-learning, domain adaptation, transfer learning**

I. Introduction

Deep learning-based algorithms have brought remarkable advancements in the task of building extraction from images as well as 3D point clouds [1], [2]. The performance of these algorithms can be largely restricted when the source and target domain data have a large domain shift or domain gaps. On the topic of building extraction, domain gaps exist between multi-sensor and multi-seasonal data sets, also occur due to different building styles and distributions of different cities. As a result, the performance of deep neural networks trained with limited labeled domains drops significantly in unseen unlabeled domains, leading to poor generalization. To address such issues, the topic of domain adaptation is attracting the attention of researchers in this field [3], [4].

One of the main challenges of cross-domain building extraction is the lack of useful information from the target domain data, as labeled target data are unavailable or insufficient. Co-learning [5], which transfers mutual knowledge between different modalities in an unsupervised way, can obtain information from target data pairs (e.g., images and DSMs). This is a potential solution to exploit deep information in the unlabeled target data and furthermore improve the generalization ability of neural networks. In this paper, we will further explore the co-learning framework by introducing two 2D semantic segmentation backbones and investigate their performance in transfer learning between a spaceborne and an airborne dataset for the task of 2D building extraction.

II. Related works

A. Domain adaptation for building extraction

In recent years, a few pioneering studies have investigated potential solutions to domain shift problem that exists in building extraction tasks. For example, [4] proposed a full-level domain adaptation framework for building extraction. It contains an image alignment method, an adversarial learning module, mean-teacher model, as well as a self-training step. In [6], a cross-geolocation attention module was proposed to improve the generalization ability of the CNN for building extraction. The source data and target data are jointly used by a Siamese framework with shared weights. [7] evaluated several state-of-the-art deep learning methods for cross-domain building extraction. In comparison to using single deep learning models, combining probability maps from different approaches can largely improve the results.

B. Multimodal learning with 2D/3D remote sensing data

The majority of existing multimodal learning research within the remote sensing domain primarily focuses on data fusion techniques, encompassing early fusion, middle fusion, and late fusion approaches [8]. In most previous works, 3D data such as DSMs are utilized as rasterized images. They are fused and processed with multispectral images in a two-dimensional domain. Early fusion takes place during the initial data processing phase. The spectral characteristics of images and the elevation information from DSMs are merged to form integrated input features for a single-modal machine learning model. For example, [9] utilized orthophotos, DSMs, and normalized DSMs (nDSMs) as the input data to a CNN for the semantic segmentation task. Middle fusion is employed at the feature embedding step, where features calculated by separate deep neural network pathways are combined into a unified representation. Subsequent operations rely on these

combined features. For example, [10] proposed a hybrid attention-aware fusion network (HAFNet) for the building extraction task. An attention-aware multimodal data fusion block is utilized to fuse multiscale deep features of RGB images and DSMs. Late fusion is utilized at the decision-making stage, where probability maps generated from multiple algorithms are integrated. In [11] a supervised probabilistic framework for building extraction was proposed. Images and DSMs are first processed to different potentials. Then those potentials are fused and processed by a conditional random field to achieve a global optimal labeling.

In fewer data fusion studies, point clouds with color information are processed directly in a three-dimensional domain. Spectral values from images can easily be added as features to point clouds [12]. This operation can be regarded as early fusion. In deep learning-related studies, [13] investigated how the sparse convolutional neural network works on tri-stereo satellite imagery-derived point clouds with color information. However, in this work, the color information from images reduces the performance of point cloud neural networks.

Unlike traditional data fusion strategies, 2D-3D co-learning as a novel idea has not been studied much. Our previous work [5] demonstrated that building extraction tasks can benefit from co-learning, which extended the framework xMUDA by [14]. Different from xMUDA focusing on street-view LiDAR point cloud semantic segmentation, our work mainly investigates the imagery-derived data. Even the orthophoto and DSM pairs applied in our work are generated from the same multiview image source.

III. Methodology

A. Enhanced co-learning for domain adaptation

Two versions of co-learning techniques applicable to building extraction from images and DSM-derived point clouds are presented in [5]. The enhanced version is better suited for cross-domain tasks since it can leverage unsupervised information from the target data, preventing the neural network models from overfitting to the source data. Therefore, we employ it in this work. As illustrated in Fig. 1, the training phase of the enhanced co-learning architecture comprises an image network designed for handling orthophotos and a point cloud network dedicated to processing point clouds. During the training phase, the co-learning method adaptively exploits knowledge from the other modality via the loss function $\mathcal{L}_{CL}$. There are four subsets of co-learning loss functions that enforce consistency constraints between predicted probabilities of 2D image and 3D point cloud modalities, either within the source or the target domain. Among these subsets, two are for the image network, while the remaining two cater to the point cloud network. In the testing phase, both types of networks can function independently without the need for the other modality. This flexibility is one of the advantages of co-learning in comparison to conventional data fusion methods.

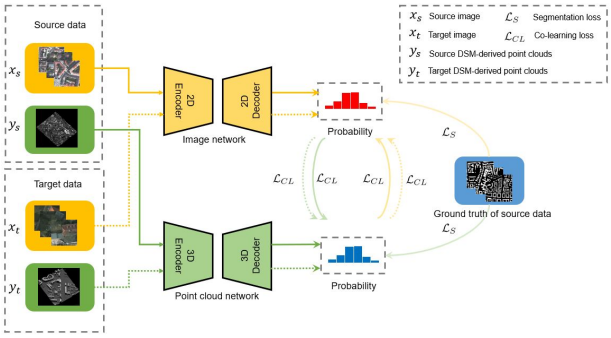

Fig. 1. Framework overview of the enhanced co-learning in the cross-domain building extraction scenario.

B. CNN-based building extraction

1) 2D backbones: We select two widely-used CNN-based 2D semantic segmentation backbones, UNet [15] and HRNet [16] for following experiments, in order to prove that our method is compatible with different mainstream backbones.

- UNet: We employ the UNet architecture with ResNet34 blocks [17], which has been proved effective for generic semantic segmentation tasks.
- HRNet: We use the version of HRNetV2-W48 as the backbone because it has better performance in generic semantic segmentation tasks compared to its peers [16].

2) 3D backbone: In our framework, we select a 3D point cloud network rather than a 2D network to process the DSMs as point clouds. The main reason is that the absolute elevations of terrains in different cities are different. Directly using absolute elevations as input channels in 2D neural networks for cross-domain data sets would lead to significant domain shifts. nDSMs would be a compromise for 2D neural networks. However, automatically calculating accurate nDSMs is an individual and challenging step, which can also introduce unnecessary errors. In our previous work [5], we have proved that 3D networks are more elegant and universal for processing the DSM-derived point clouds, as they use relative coordinates rather than absolute coordinates and can represent the features of objects on the ground more naturally. We employ SparseConvNet [18] as the backbone for extracting buildings from point clouds originating from DSMs.

C. Loss functions

As shown in Fig. 1, two kinds of loss functions are involved in the co-learning framework. Following paper [5], cross-entropy is utilized as the loss function for semantic segmentation:

$$\mathcal{L}_S(P||Q) = H(P||Q) \tag{1}$$

$$= \sum_{x \in \mathcal{X}} P(x) \log(Q(x)) \,, \tag{2}$$

where P and Q are defined within the identical probability space $\mathcal{X}$. P is the distribution of the ground truth, whereas Q is the probability distribution of the predicted output.

Kullback–Leibler (KL) divergence is used for the calculation of the co-learning loss:

$$\mathcal{L}_{CL}(P||Q) = \mathcal{D}_{KL}(P||Q) \quad (3)$$

$$= \sum_{x\in\mathcal{X}} P(x)\log(\frac{P(x)}{Q(x)}) , \quad (4)$$

where P represents the probability distribution of the target data, whereas Q is the probability distribution of the predicted output.

The total loss function of enhanced co-learning is:

$$\mathcal{L}_{total} = \mathcal{L}_S + \lambda_1 \mathcal{L}_{CL}^{labeled}(M_1||M_2) + \lambda_2 \mathcal{L}_{CL}^{unlabeled}(M_1||M_2) , \quad (5)$$

where λ_i is the weighting coefficient. M_1 and M_2 represent two different modalities.

IV. Experiments

A. Data sets

This paper investigates how enhanced co-learning works on cross-domain building extraction via a spaceborne-to-airborne experiment. A spaceborne data set is utilized as the source domain data, while an airborne data set is utilized as the target domain data. Fig. 2 shows examples of above two data sets. As can be observed, these two datasets differ on both spectral style and the quality of the DSMs.

The source data set is the Munich WorldView-2 data set introduced in [5], which is a collection of spaceborne imagery covering Munich, Germany. In this work, the red (5th), green (3th), and blue (2nd) channels are selected to compose the RGB orthophotos, as the target images only have those three channels. We employ labeled pairs of orthophotos and DSMs as the training data. DSMs are used as point clouds and processed by the 3D backbone. The ground sampling distance (GSD) of the orthophotos and DSMs is 0.5m. The training data used in our experiments contain 3 pairs of tiles. Each tile has a size of 6000×6000 pixels. In our experiments, we use 600×600 pixel patches as the input for the deep learning models. The original training data are cropped into 300 patches.

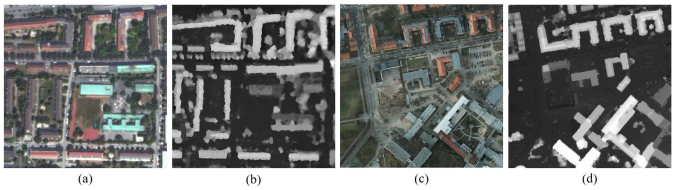

Fig. 2. Examples of the source data and the target data. (a) An image patch by WorldView-2 satellite. (b) A DSM patch derived from WorldView-2 spaceborne images. (c) An image patch of the ISPRS Potsdam airborne data set. (d) A DSM patch of the ISPRS Potsdam airborne data set.

As a pre-processing step of the target domain ISPRS Potsdam data set[1], RGB image and imagery-derived DSM tiles are downsampled to a GSD of 0.5m to maintain consistent resolution with the source data. After the downsampling, each tile has a size of 600 × 600 pixels. As an essential step of enhanced co-learning, the target domain data are used as the unlabeled training pairs in the training phase. The Original training data of the Potsdam benchmark have 24 pairs of tiles. We use 20 tiles of them (ID: 2-10, 2-12, 3-10, 3-11, 3-12, 4-11, 4-12, 5-10, 5-11, 6-7, 6-8, 6-9, 6-10, 6-11, 6-12, 7-7, 7-9, 7-10, 7-11, and 7-12) as the unlabeled training pairs and 4 tiles (ID: 7-8, 4-10, 2-11, and 5-11) as the validation data for locating the optimal models.

[1]https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx

B. Implementation Details

In our experiments, the open-source library PyTorch is utilized to implement the deep learning models. Both training and testing stages are conducted on an NVIDIA GeForce RTX 2080 Ti. Each training session lasts for 30 epochs, and the checkpoint that performs best on the validation data is selected as the final model. Adam optimizer is adopted with a learning rate of 0.001. The batch size of the training models is set as 3. For the point cloud network, the input voxel size is set to 0.5m, consistent with the GSD of DSMs. The input point clouds fed into the network do not include spectral features. Regarding the loss function, λ_1 is set to 0 and the co-learning loss function is not applied to the source data, as the crossmodal constraint of the source data could make the deep learning models easier to over-fitting according to our experience. λ_2 is set to 0.1.

To evaluate and compare the results, we employ two commonly used metrics: the F1-score and intersection over union (IoU). They are calculated by the number of true positives (*TP*), false negatives (*FN*), true negatives (*TN*), and false positives (*FP*). Positives are pixels classified as buildings, while negatives represent pixels classified as non-buildings.

$$F1 = \frac{2TP}{2TP + FP + FN} , \quad (6)$$

$$IoU = \frac{TP}{TP + FP + FN} , \quad (7)$$

C. Results and analysis

Table I quantitatively lists the results achieved by different methods. Compared to the baseline UNet and HRNet, enhanced co-learning with DSM-derived point clouds has a large improvement in performance. With HRNet, enhanced co-learning increases F1 and IoU by 9.36% and 9.72%, respectively. For the UNet, the improvement from enhanced co-learning is even more obvious, 17.97% in the F1 score, and 21.52% in IoU.

The visualization results in Fig. 3 also indicate that enhanced co-learning has a remarkable improvement in the performance of 2D CNNs. From Fig. 2, the color style of the Munich WorldView-2 image and that of the Potsdam airborne image are noticeably different. This could explain why 2D CNNs trained only with single-modality source images fail to deliver satisfactory performance on the unseen target data.

Comparing Fig. 3 (b) and (e), HRNet trained with single-modality source images has several notable limitations. For instance, as highlighted in red circles, it overlooks a majority of the building structures. In contrast, HRNet trained using enhanced co-learning can capture more details. As circled in blue, HRNet trained with enhanced co-learning generates considerably fewer false positives. In Fig. 3 (c), UNet trained only with source images also results in noticeable issues. The generated prediction appears visually noisy due to a large number of non-building pixels being misclassified as buildings. In Fig. 3 (f), however, most false positives are eliminated by employing enhanced co-learning with unlabeled training pairs.

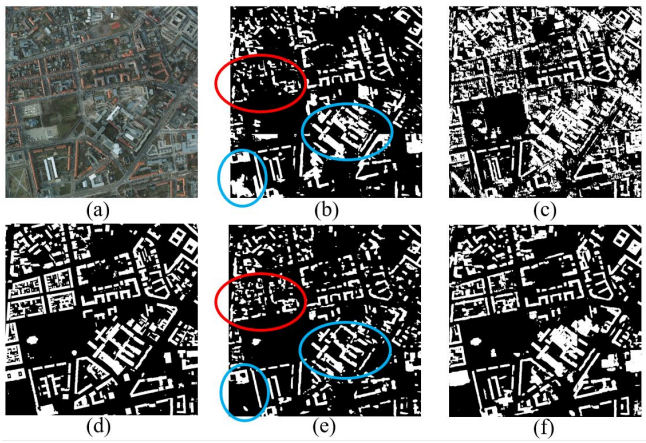

Fig. 3. Building extraction results of ISPRS Potsdam target data. (a) RGB image. (b) HRNet (source only). (c) UNet (source only). (d) Ground truth. (e) HRNet (enhanced co-learning). (f) UNet (enhanced co-learning).

TABLE I
QUANTITATIVE RESULTS ACHIEVED BY DIFFERENT METHODS.

Methods	F1	IoU
HRNet (source only)	0.5648	0.3935
HRNet (Enhanced co-learning)	0.6584↑	0.4907↑
UNet (source only)	0.6150	0.4441
UNet (Enhanced co-learning)	0.7947↑	0.6593↑

V. CONCLUSION

In this paper, we investigate how the enhanced co-learning framework presented in [5] helps to improve the transfer learning ability of 2D CNNs. The framework is tested with two popular image CNN-based backbones, HRNet and UNet. Our experiment is carried out on an optical spaceborne data set and an airborne data set for building extraction. It demonstrates that 2D-3D co-learning with optical images and point clouds is an effective way to improve the generalization ability of 2D CNNs for optical imagery and improve the building extraction results in cross-domain scenarios. Both HRNet and UNet benefit from the DSM modality embedded in the point cloud network SparseConvNet via co-learning. Future efforts will focus on investigating more variants of co-learning frameworks and extending this method to other remote sensing applications.

REFERENCES

[1] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[2] Y. Xie, J. Tian, and X. X. Zhu, "Linking points with labels in 3d: A review of point cloud semantic segmentation," *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 4, pp. 38–59, 2020.

[3] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 41–57, 2016.

[4] D. Peng, H. Guan, Y. Zang, and L. Bruzzone, "Full-level domain adaptation for building extraction in very-high-resolution optical remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2021.

[5] Y. Xie, J. Tian, and X. X. Zhu, "A co-learning method to utilize optical images and photogrammetric point clouds for building extraction," *International Journal of Applied Earth Observation and Geoinformation*, vol. 116, p. 103165, 2023.

[6] Q. Li, L. Mou, Y. Hua, Y. Shi, and X. X. Zhu, "Crossgeonet: A framework for building footprint generation of label-scarce geographical regions," *International Journal of Applied Earth Observation and Geoinformation*, vol. 111, p. 102824, 2022.

[7] H. Li, J. Tian, Y. Xie, C. Li, and P. Reinartz, "Performance evaluation of fusion techniques for cross-domain building rooftop segmentation," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 501–508, 2022.

[8] M. Schmitt and X. X. Zhu, "Data fusion and remote sensing: An ever-growing relationship," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 4, pp. 6–23, 2016.

[9] S. Paisitkriangkrai, J. Sherrah, P. Janney, V.-D. Hengel *et al.*, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 36–43.

[10] P. Zhang, P. Du, C. Lin, X. Wang, E. Li, Z. Xue, and X. Bai, "A hybrid attention-aware fusion network (hafnet) for building extraction from high-resolution imagery and lidar data," *Remote Sensing*, vol. 12, no. 22, p. 3764, 2020.

[11] D. Chai, "A probabilistic framework for building extraction from airborne color image and dsm," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 3, pp. 948–959, 2016.

[12] M. Weinmann, B. Jutzi, S. Hinz, and C. Mallet, "Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 105, pp. 286–304, 2015.

[13] S. Bachhofner, A.-M. Loghin, J. Otepka, N. Pfeifer, M. Hornacek, A. Siposova, N. Schmidinger, K. Hornik, N. Schiller, O. Kähler *et al.*, "Generalized sparse convolutional neural networks for semantic segmentation of point clouds derived from tri-stereo satellite imagery," *Remote Sensing*, vol. 12, no. 8, p. 1289, 2020.

[14] M. Jaritz, T.-H. Vu, R. d. Charette, E. Wirbel, and P. Pérez, "xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 605–12 614.

[15] 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.

[16] 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*, vol. 43, no. 10, pp. 3349–3364, 2020.

[17] 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.

[18] B. Graham, M. Engelcke, and L. Van Der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9224–9232.