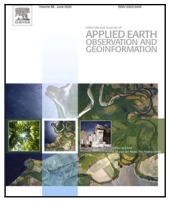
International Journal of Applied Earth Observations and Geoinformation 116 (2023) 103165

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



A co-learning method to utilize optical images and photogrammetric point clouds for building extraction

Yuxing Xie [a,b,*], Jiaojiao Tian [b], Xiao Xiang Zhu [a]

[a] Data Science in Earth Observation, Technical University of Munich, Munich, 80333, Germany
[b] The Remote Sensing Technology Institute, German Aerospace Center (DLR), Muenchener Strasse 20, Wessling, 82234, Germany

ARTICLE INFO

Keywords:
Building extraction
Co-learning
Multimodality learning
Multispectral images
Point clouds
Remote sensing

ABSTRACT

Although deep learning techniques have brought unprecedented accuracy to automatic building extraction, several main issues still constitute an obstacle to effective and practical applications. The industry is eager for higher accuracy and more flexible data usage. In this paper, we present a co-learning framework applicable to building extraction from optical images and photogrammetric point clouds, which can take the advantage of 2D/3D multimodality data. Instead of direct information fusion, our co-learning framework adaptively exploits knowledge from another modality during the training phase with a soft connection, via a predefined loss function. Compared to conventional data fusion, this method is more flexible, as it is not mandatory to provide multimodality data in the test phase. We propose two types of co-learning: a standard version and an enhanced version, depending on whether unlabeled training data are employed. Experimental results from two data sets show that the methods we present can enhance the performance of both image and point cloud networks in few-shot tasks, as well as image networks when applying fully labeled training data sets.

1. Introduction

Automatic building extraction from remotely sensed data is an important task in the photogrammetry and remote sensing field. It plays a vital role in many practical applications, such as building information modeling, urban monitoring and planning, and digital twins. Recently, advanced deep learning algorithms with high-quality data sets have achieved unprecedented performance in building extraction. However, there are still numerous problems restricting the generalization. When the deep neural network is trained with insufficient training samples, overfitting will occur and the network cannot perform accurately against unseen data. To meet the requirements of industry applications, better accuracy and less dependency on annotated training data sets are among the most urgent needs. Annotating a large amount of training data is labor intensive. Hence, studies on automatic building extraction are still ongoing, but researchers' attention has shifted from simply stacking different networks to developing targeted algorithms in order to better regularize the results, as well as designing flexible architectures to efficiently utilize multimodality data in networks, resulting in less dependent on the quantity of annotated data.

Based on the applications and corresponding data types employed, building extraction tasks are usually divided into three categories: 2D image based, 3D geometric data (point clouds/DSMs) based, and multimodality data based. Image-based automatic building extraction is the most widely studied case, as the acquisition cost of optical images is relatively low. In recent years, deep learning-based methods, especially convolutional neural networks (CNNs), have taken the place of the traditional algorithms and became the most widely utilized, as their performance is superior on various data sets (Zhu et al., 2020; Shi et al., 2020; Li et al., 2021).

Although 2D remotely sensed images are widely used in practical applications, they have several obvious limitations. Remotely sensed images captured by airborne or spaceborne sensors usually cover much larger area, which may cause scale variation of buildings, thereby influencing the performance of algorithms. Furthermore, unavoidable reflection of light, shadows, and obstructions can also have negative effects on building extraction results (Tian et al., 2014; Sun et al., 2021b). Due to the development of LiDAR sensors and dense image matching algorithms, 3D geometric data such as point clouds and DSMs have brought new possibilities to the building extraction field and compensate for deficiencies in the images, as they can provide geometric features that are not affected by spectral distortion. Also Driven by the success of deep learning techniques, researchers have recently been keen to apply all kinds of point cloud neural networks to urban point cloud processing (Xie et al., 2020), such as PointNet (Qi et al., 2017; Huang et al., 2020a; Yousefhussien et al., 2018), KPConv (Lin et al.,

* Corresponding author at: The Remote Sensing Technology Institute, German Aerospace Center (DLR), Muenchener Strasse 20, Wessling, 82234, Germany.
E-mail address: yuxing.xie@dlr.de (Y. Xie).

https://doi.org/10.1016/j.jag.2022.103165
Received 25 August 2022; Received in revised form 6 December 2022; Accepted 16 December 2022
Available online 31 December 2022

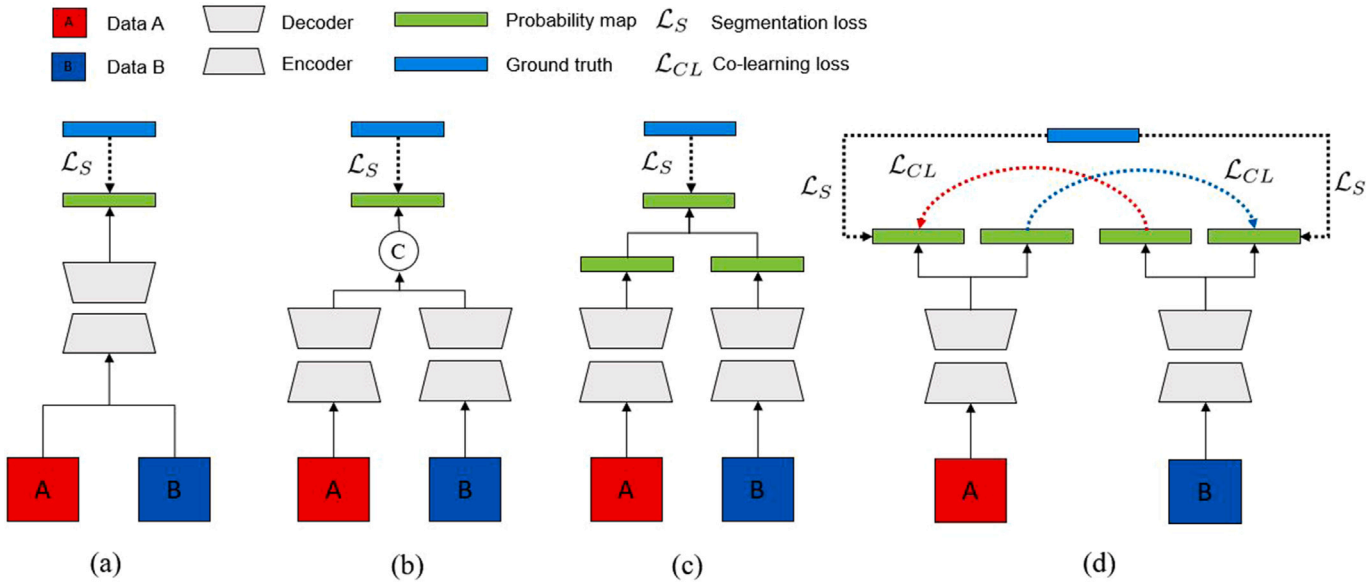


Fig. 1. The difference between conventional data fusion and co-learning. (a) Early fusion. (b) Middle fusion. (c) Late fusion. (d) Multimodality co-learning in our work.

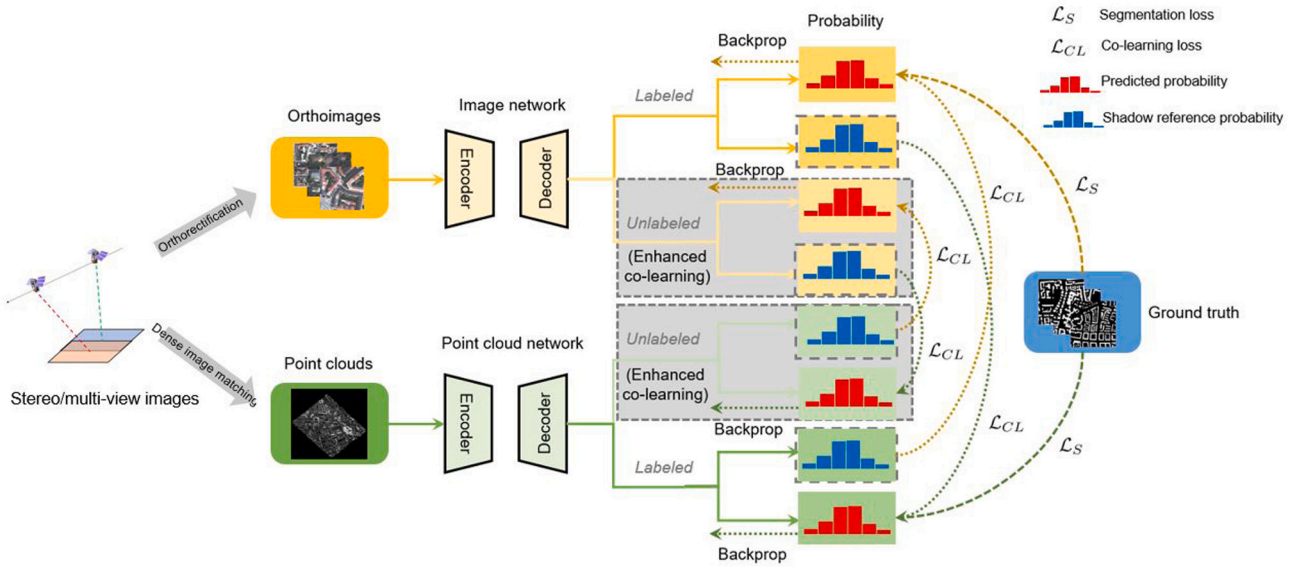


Fig. 2. The training phase of the proposed co-learning framework. In our work, images used for building extraction are orthoimages. The forward propagation, loss functions, and backward propagation of the image network and the point cloud network are indicated by yellow and green arrows, respectively. Point clouds are generated from raw stereo- or multi-view images. In the procedure of optional enhanced co-learning (framed by gray boxes), unlabeled data do not participate in the optimization of the supervised semantic segmentation loss function. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2021; Thomas et al., 2019), and sparse CNN (Graham et al., 2018; Bachhofner et al., 2020).

Unfortunately, 3D data also have limitations in applications to building extraction. Point clouds are discrete, which leads to the problems of missing building structure, as well as boundaries that are not sufficiently sharp. Since 2D images and 3D geometric data can provide information complementary to each other, which could benefit the accuracy of building extraction, methods of multimodality learning have attracted the attention of researchers (Bittner et al., 2018; Sun et al., 2021b). Most available multimodality learning works in the remote sensing field concentrate on data fusion, including early fusion (input fusion or observation-level fusion), middle fusion (feature fusion or feature-level fusion), and late fusion (probability fusion or decision-level fusion) (Schmitt and Zhu, 2016).

As shown in Fig. 1(a), early fusion is usually carried out at the data input stage. In one popular case in the remote sensing field, multispectral images are fused with DSMs for semantic segmentation (Paisitkriangkrai et al., 2015). In this approach, spectral channels of optical images and geometric information such as the height values of DSMs are concatenated as combined input features to a single-modality network. In Fig. 1(b), middle fusion is operated at the stage of feature embedding, concatenating deep features learned by different network streams to a composite stream (Zhou et al., 2021). Following operations are based on the concatenated feature vectors. Late fusion is employed at the decision stage, which operates on the probability maps output from multiple algorithms, as shown in Fig. 1(c).

Data fusion takes the benefit of multiple information sources and improves the performance of semantic segmentation algorithms, including building extraction algorithms. But these techniques have strict requirements for both data amount and data quality, and assume that all modalities are present, aligned, and noiseless during the training and the test phase (Rahate et al., 2022). However, 2D images and

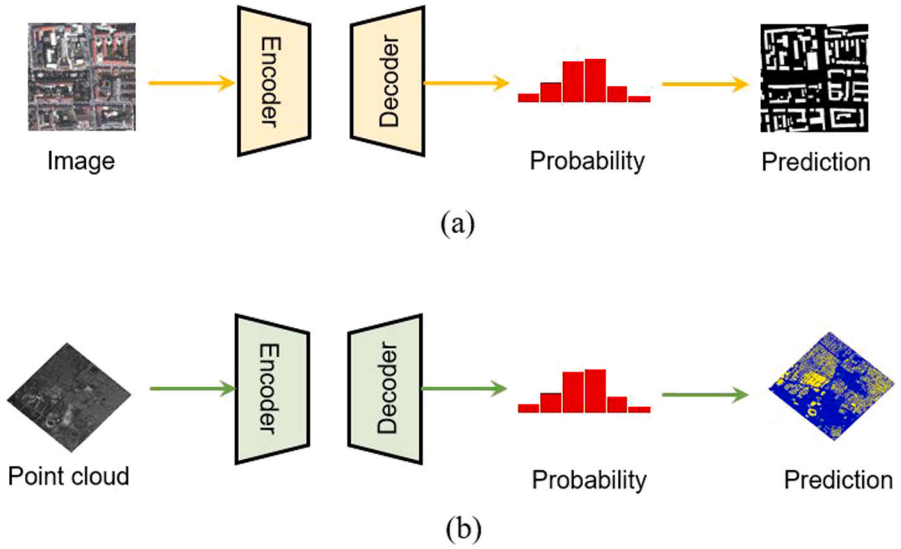


Fig. 3. In the test phase, networks are used individually as normal single-modality networks.

3D data are not always simultaneously available in diverse data sets. In addition, LiDAR-based data are expensive and time consuming to acquire, so are not suitable for projects involving large-scale areas. Imagery-derived 3D data require a certain amount of overlapped high-resolution optical images for the dense image matching algorithm. This is also a challenge for applications involving historic orthophotos in which raw stereo/multi-view images are missing and matched 3D data cannot be obtained. Well-performing single-modality networks are essential for practical applications. On the other hand, the architectures of networks that process fused data are usually complex and bloated, resulting in low efficiency and requiring high computational ability. By contrast, methods with simple and efficient architectures that consume few annotated data would be welcome in practical applications that demand real-time data processing capability.

Recently, co-learning methods are proposed in the generic artificial intelligence field, aiming to aid the modeling of one modality by exploiting knowledge from another and offering a tradeoff between the advantage of multimodalities and strict input data requirements. Co-learning explores how knowledge learning from one modality can help a deep learning model trained on other different modalities, especially when one modality has limited resources, such as missing modality, noisy modality, and lacking annotated data (Rahate et al., 2022; Zheng et al., 2021). As reviewed in (Baltrušaitis et al., 2019) and (Rahate et al., 2022), co-learning-based methods have been employed in several cross-modality applications (e.g., audio-visual Zadeh et al., 2020, visual-text Ma et al., 2021). Fig. 1(d) presents a type of co-learning architectures based on loss functions, which is applicable to multimodality semantic segmentation. The step of knowledge transfer bridging multimodality networks in this architecture is realized by the co-learning loss function rather than direct addition or concatenation. Each single-modality network is trained individually, where corresponding parameters can be better optimized with the help of another modality via co-learning loss. Unlike traditional data fusion approaches, a semantic segmentation model trained with a multimodality data set through this way can be also performed on single-modality test data, thus effectively solving the problem of insufficient availability of multimodality test data.

In computer vision, cross-modality unsupervised domain adaptation (xMUDA) is the first work to adaptively transfer information among multimodality data sets to improve the segmentation results of mobile LiDAR point clouds (Jaritz et al., 2020). As its name suggests, xMUDA aims to address the problem of domain adaptation for point cloud semantic segmentation. In our article, we combine the theoretical background of generic co-learning and xMUDA, and propose an elegant framework applicable to automatic building extraction from spectral images and corresponding photogrammetric point clouds. Fig. 2 shows the architecture of our proposed co-learning model. The architecture of the training model contains a 2D network to process images and a 3D network to work on point clouds. As shown in Fig. 3(a) and (b), these two networks can be used individually in the test phase. As another difference from xMUDA, there is no self-training step involved in our method, so it would be more friendly to software development. In addition, our architecture can utilize unlabeled training data, thereby reducing the dependence on the amount of annotated data. Hence, it is especially suitable for the case with fewer annotated training data. The main contributions of our work are as follows:

- We present a co-learning framework to handle the case in which one modality is missing during the testing time. In particular we exemplify the framework with photogrammetric point clouds and corresponding optical images, because in practice these two modalities are one of the most widely-used pairs.
- We apply the proposed co-learning framework in few-shot tasks to solve the problem of scarcity of labeled training data. We investigate the effects of unlabeled data in our framework.
- We evaluate our co-learning framework on two data sets: the ISPRS Potsdam public airborne data set, and a data set of Munich collected by the WorldView-2 satellite. Experimental results demonstrate the effectiveness of our method for the task of automatic building extraction.

2. Methodology

2.1. Overview

The detailed flowchart of proposed co-learning based network architecture is shown in Fig. 2. As it shows, the co-learning method we applied transfers knowledge from one modality to another based on the probability maps. The intuition behind this approach is that the better results the networks achieve, the smaller the prediction gap between two modalities. To meet this requirement, a co-learning loss function is proposed to learn the similarity between the predictions of the 2D and 3D networks. In the training phase, the target is to minimize two loss functions, a supervised loss function for semantic segmentation purpose, and the unsupervised co-learning loss function to measure the distance between two predictions. In the implementation, each network

outputs two types of probability maps. One, *predicted probability*, is used in the loss functions of the same network, and influences the backward propagation. In order to distinguish from real reference (ground truth), the other is named *shadow reference probability*, and is actually utilized by the other modality network as the reference in the co-learning loss function. The training data involved in two loss functions could be asymmetric, which means only part of the data needs to be annotated. Unlabeled data pairs are also beneficial to the minimization of co-learning loss.

2.2. 2D and 3D feature learning

As building extraction can be regarded as a branch of semantic segmentation, convolutional encoder–decoder neural networks are mainstream architectures applied for feature learning from raw images and/or point clouds. In our work, we employ a 2D U-Net (Ronneberger et al., 2015) with residual blocks of ResNet34 (He et al., 2016) as the backbone to learn 2D features from multispectral images. A U-Net-like sparse convolutional neural network (Graham et al., 2018; Choy et al., 2019) is employed as the backbone to learn 3D features from point clouds.

CNNs are a category of deep learning models that have been successfully utilized in image and point cloud processing, and consist of multiple convolutional layers. In each layer, the input feature maps are convolved by a kernel with learned weights. In image cases, the convolutional kernel is usually naturally dense, and is defined as (Choy et al., 2019)

$$\mathbf{x}_{\mathbf{u}}^{out} = \sum_{\mathbf{i} \in \mathcal{V}^D} W_{\mathbf{i}} \mathbf{x}_{\mathbf{u}+\mathbf{i}}^{in}, \tag{1}$$

where $\mathbf{x}_{\mathbf{u}}^{in} \in \mathbb{R}^{N^{in}}$ is the input feature vector of coordinate $\mathbf{u} \in \mathbb{Z}^D$ in a D-dimensional space. $\mathcal{V}^D$ is the list of offset elements in the hypercube centered at the origin, which is covered by the convolution kernel. $W_{\mathbf{i}}$ is the kernel weight corresponding to the offset element $\mathbf{u}+\mathbf{i}$. For 2D images, $D = 2$.

In the real world, most 3D spaces are not occupied by any objects. As a result, corresponding point clouds and converted voxels contain large empty areas (Xu et al., 2021). If we adopt conventional dense convolutions to process such sparse data, the calculation would be time-consuming and memory-intensive. Sparse convolution presented by Bachhofner et al. (2020) and Choy et al. (2019) is a solution to this problem. Arbitrary kernel shapes instead of conventional dense shapes are utilized in sparse convolutions, which only take those non-empty grids into the convolving calculation. By defining the existing offset grids covered by the convolution as $\mathcal{N}^D(\mathbf{u}, C^{in})$, the output feature vector $\mathbf{x}_{\mathbf{u}}^{out}$ is presented as (Choy et al., 2019)

$$\mathbf{x}_{\mathbf{u}}^{out} = \sum_{\mathbf{i} \in \mathcal{N}^D(\mathbf{u}, C^{in})} W_{\mathbf{i}} \mathbf{x}_{\mathbf{u}+\mathbf{i}}^{in}, \tag{2}$$

where $\mathcal{N}^D(\mathbf{u}, C^{in}) = \{\mathbf{i} | \mathbf{u}+\mathbf{i} \in C^{in}, \mathbf{i} \in \mathcal{N}^D\}$. C^{in} is the predefined sparse tensors to be convolved. In the case of a point cloud, $D = 3$.

2.3. Co-learning

The co-learning method in our work is a flexible framework that makes use of different categories of training data. As mentioned above, both labeled and unlabeled training data can be employed in this framework. The labeled data and unlabeled data can be asymmetric. According to the availability of ground truth and multimodality pairs, the training data in our co-learning framework can be classified into three categories: *labeled pairs*, *unlabeled pairs*, and *labeled singles*, separately named in our work.

- *labeled pairs* refer to the data with the ground truth that are co-registered with another modality; these pairs are involved in both supervised loss function and the co-learning loss function.
- *unlabeled pairs* refer to the samples that are without ground truth but have co-registered multimodalities, which means they can benefit the co-learning loss function.
- *labeled singles* are the single-modality training data with ground truth, that are involved only in the supervised loss function not the co-learning loss function.

In our work, we mainly explore the influence of labeled pairs and unlabeled pairs. The effect of labeled singles is obvious, as they have been widely investigated in works on conventional single-modality learning, which can be regarded as architectures only with labeled singles. We name the setting with only labeled pairs as *standard co-learning*, and the situation trained partly with additional unlabeled pairs as *enhanced co-learning*. Fig. 2 contains the training procedures of both standard and enhanced cases.

2.3.1. Standard co-learning

The intuition behind our co-learning method is that unsupervised mutual information from the other modality would be a positive factor to the target networks. Apart from the difference between the prediction and the ground truth (i.e., supervised segmentation loss function), the similarity between multimodality data could also be potentially valid information benefiting the training phase and helping find more proper deep model parameters. This is realized by a co-learning loss function. As shown in Fig. 2, standard co-learning adopts the labeled training samples in the learning procedure. For each backpropagation step, the gradients of the combination of the supervised segmentation loss function and co-learning function are computed. Algorithm 1 shows how the standard co-learning is implemented. For each iteration, first the predicted probability of images p_{2D} and the predicted probability of point clouds p_{3D} are calculated by the forward propagations of two networks, respectively. Then supervised segmentation loss functions and co-learning functions are computed. In the calculation of co-learning loss for images, p_{3D} is used as the shadow reference probability. In the computation of co-learning loss for point clouds, p_{2D} is employed as the shadow reference probability. Finally, backpropagation operations are carried out and the parameters of the image network W_{2D} as well as the parameters of the point cloud network W_{3D} are updated.

Algorithm 1 Standard co-learning

Input: (D_{2D}, L_{2D}), (D_{3D}, L_{3D})
Output: W_{2D}, W_{3D}
Initialize W_{2D}, W_{3D}
while $i < I$ **do** ▷ I is the number of iterations
  Randomly sample labeled training pairs d_{2D} and d_{3D} from D_{2D} and D_{3D}
  $p_{2D} \leftarrow net_{2D}(d_{2D})$ ▷ forward pass of the image network
  $p_{3D} \leftarrow net_{3D}(d_{3D})$ ▷ forward pass of the point cloud network
  Calculate $\mathcal{L}_S^{2D}(l_{2D}||p_{2D})$ ▷ image segmentation loss
  Calculate $\mathcal{L}_{CL}^{2D}(p_{3D}||p_{2D})$ ▷ image co-learning loss
  Calculate $\mathcal{L}_S^{3D}(l_{3D}||p_{3D})$ ▷ point cloud segmentation loss
  Calculate $\mathcal{L}_{CL}^{3D}(p_{2D}||p_{3D})$ ▷ point cloud co-learning loss
  2D backward pass
  Update W_{2D}
  3D backward pass
  Update W_{3D}
end while
Return W_{2D}, W_{3D}

2.3.2. Enhanced co-learning

Annotating a large amount of training data is always a challenge in deep learning-based tasks, and is both expensive and time-consuming. Thus few-shot learning, which serves as a low-cost solution, is attracting more attention in deep learning related research (Sun et al., 2021a).

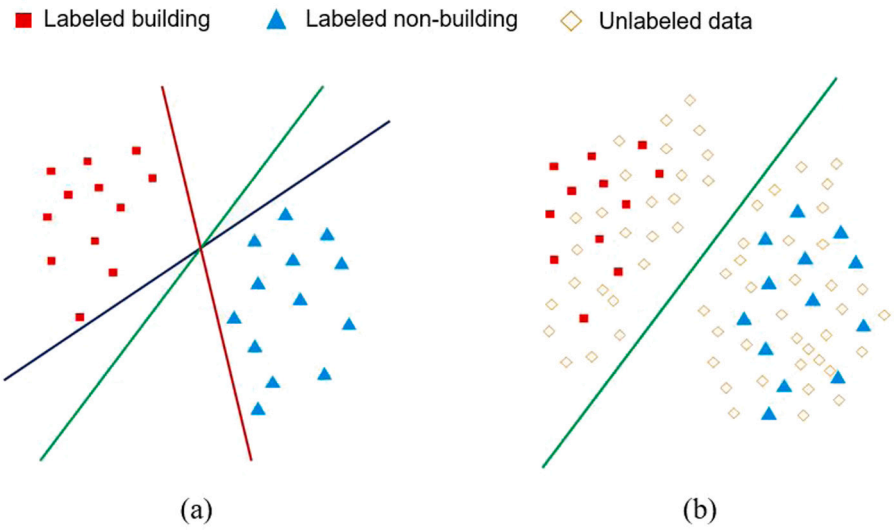


Fig. 4. (a) Learning with few data. (b) Enhanced co-learning. Lines with different colors represent different classifiers/models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

A main drawback of conventional few-shot learning is the restricted beforehand knowledge. As shown in Fig. 4(a), we simulate this problem based on a building extraction task in a simple two-dimensional feature space. If there is no interference on the learning case with fewer training data, multiple models with different parameters can yield a reasonable classification. However, most of those models are prone to overfitting. They may have reasonable prediction results on the training samples, but they are likely to fail to predict unseen test data.

In reality, there is a huge amount of unlabeled data exist, but they are difficult to use directly in supervised learning. One advantage of the co-learning function is that it can employ unlabeled pairs. If unlabeled pairs are able to assist the clustering procedure, more accurate and less ambiguous models with better generalization ability could be obtained, as Fig. 4(b) shows. This is the intuition behind enhanced co-learning. Enhanced co-learning utilizes data in a more efficient way than conventional semi-supervised self-training that employs unlabeled training samples. Self-training is a procedure with several individual steps: training an initial model with a few labeled training samples, predicting on several unlabeled data, and re-training a model with unlabeled data and predicted pseudo labels (Zoph et al., 2020; Zhang et al., 2021). In order to obtain more accurate and stable models, sometimes users have to design extra algorithms to select proper samples with pseudo labels, and the training procedure has to be repeated several times (Zhang et al., 2021; Tong et al., 2020). In contrast, our enhanced co-learning is a one-step operation requiring no extra algorithm, which is much more user-friendly in practice.

For these reasons, our work incorporates an enhanced co-learning structure into our design by adopting both labeled and unlabeled training samples. Algorithm 2 demonstrates the implementation of the enhanced co-learning. For each iteration, enhanced co-learning carries out two forward propagations for each modality. One is with labeled training data. The other is with unlabeled training data.

Algorithm 2 Enhanced co-learning

Input: (D_{2D}, L_{2D}), (D_{3D}, L_{3D}), U_{2D}, U_{3D}
Output: W_{2D}, W_{3D}
Initialize W_{2D}, W_{3D}
while $i < I$ **do** ▷ I is the number of iterations
  Randomly sample labeled training pairs d_{2D} and d_{3D} from D_{2D} and D_{3D}
  $p_{2D}^{labeled} \leftarrow net_{2D}(d_{2D})$ ▷ forward pass of the image network
  $p_{3D}^{labeled} \leftarrow net_{3D}(d_{3D})$ ▷ forward pass of the point cloud network
  Calculate $\mathcal{L}_S^{2D}(l_{2D}||p_{2D}^{labeled})$ ▷ segmentation loss for labeled images
  Calculate $\mathcal{L}_{CL}^{2D-labeled}(p_{3D}^{labeled}||p_{2D}^{labeled})$ ▷ co-learning loss for labeled images
  Calculate $\mathcal{L}_S^{3D}(l_{3D}||P_{3D}^{labeled})$ ▷ segmentation loss for labeled point clouds
  Calculate $\mathcal{L}_{CL}^{3D-labeled}(p_{2D}^{labeled}||p_{3D}^{labeled})$ ▷ co-learning loss for labeled point clouds
  2D backward pass
  3D backward pass
  Randomly sample unlabeled training pairs u_{2D} and u_{3D} from U_{2D} and U_{3D}
  $p_{2D}^{unlabeled} \leftarrow net_{2D}(u_{2D})$ ▷ forward pass of the image network
  $p_{3D}^{unlabeled} \leftarrow net_{3D}(u_{3D})$ ▷ forward pass of the point cloud network
  Calculate $\mathcal{L}_{CL}^{2D-unlabeled}(p_{3D}^{unlabeled}||p_{2D}^{unlabeled})$ ▷ co-learning loss for unlabeled images
  Calculate $\mathcal{L}_{CL}^{3D-unlabeled}(p_{2D}^{unlabeled}||p_{3D}^{unlabeled})$ ▷ co-learning loss for unlabeled point clouds
  2D backward pass
  3D backward pass
  Update W_{2D}
  Update W_{3D}
end while
Return W_{2D}, W_{3D}

2.4. Loss functions

Our method employs two categories of loss functions: the supervised loss function for the purpose of building extraction and the unsupervised loss function to realize co-learning. As mentioned above, we mainly consider two categories of training data: labeled pairs and unlabeled pairs. Hence, we describe our proposed loss functions accordingly.

Building extraction is a branch of supervised semantic segmentation. In our work, a cross-entropy loss function is used for this purpose:

$$\mathcal{L}_S(P \parallel Q) = H(P \parallel Q) \tag{3}$$

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

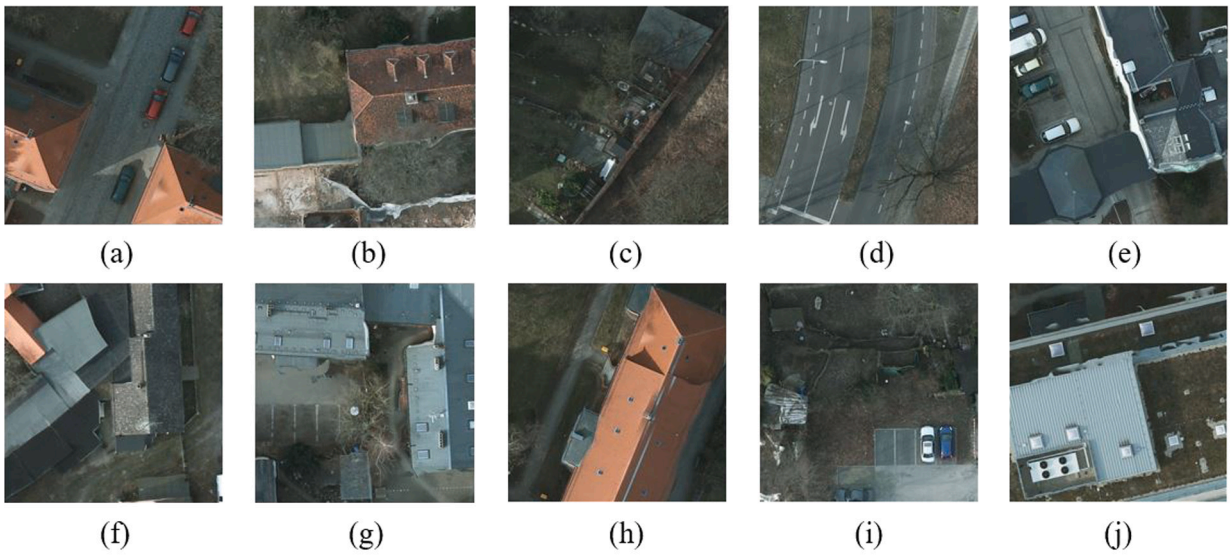


Fig. 5. 10-shot training samples of the ISPRS Potsdam data set.

where P and Q are defined on the same probability space $\mathcal{X}$. The term P denotes the distribution of the ground truth, while Q is the probability distribution of the predicted output.

The co-learning function is designed to transfer mutual information from one modality to another. When both networks are optimized, the difference in the building extraction results between the 2D and 3D modalities should be minimized. In other words, the probability distribution of one modality should be consistent with the distribution of the other. This step can be realized by a similarity loss function. Referring to Jaritz et al. (2020), we adopted KL divergence to realize this optimization.

$$\mathcal{L}_{CL}(P \parallel Q) = \mathcal{D}_{KL}(P \parallel Q) \tag{5}$$

$$= \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right), \tag{6}$$

where P and Q are defined on the same probability space $\mathcal{X}$. The item P denotes the probability distribution of the target data, while Q is the probability distribution of the predicted output. In our co-learning framework, P and Q are from two different modalities. P is the shadow reference probability, while Q is the predicted probability.

Combining a co-learning loss function $\mathcal{L}_{CL}$ with semantic segmentation loss function $\mathcal{L}_S$, the total standard co-learning loss function $\mathcal{L}_{total}$ for each single-modality network is derived. For a 2D image network, the total loss function is:

$$\mathcal{L}_{total} = \mathcal{L}_S + \lambda_1 \mathcal{L}_{CL}(P_{3D} \parallel P_{2D}), \tag{7}$$

where λ_1 is the hyperparameter to weight the co-learning loss function. Here the probability map of point clouds P_{3D} is set as the shadow reference, which is regarded as constant coefficients in the co-learning loss function for the image network.

For a 3D point cloud network, the total loss function is

$$\mathcal{L}_{total} = \mathcal{L}_S + \lambda_1 \mathcal{L}_{CL}(P_{2D} \parallel P_{3D}), \tag{8}$$

where λ_1 is the hyperparameter to weight co-learning loss function. Here the probability map of images P_{2D} is set as the shadow reference, which is regarded as constant coefficients in the co-learning loss function for the point cloud network.

For the case of enhanced co-learning, unlabeled pairs are also taken into consideration by the co-learning loss. The total image network loss function combining enhanced co-learning loss function $\mathcal{L}_{CL}^{unlabeled}$ is:

$$\mathcal{L}_{total} = \mathcal{L}_S + \lambda_1 \mathcal{L}_{CL}^{labeled}(P_{3D} \| P_{2D}) + \lambda_2 \mathcal{L}_{CL}^{unlabelead}(P_{3D} \| P_{2D}), \tag{9}$$

The total loss function combining enhanced co-learning employed in a 3D point cloud network is

$$\mathcal{L}_{total} = \mathcal{L}_S + \lambda_1 \mathcal{L}_{CL}^{labeled}(P_{2D} \| P_{3D}) + \lambda_2 \mathcal{L}_{CL}^{unlabeled}(P_{2D} \| P_{3D}), \tag{10}$$

3. Experiments

In this section, we introduce the data sets utilized for the evaluation of the proposed co-learning methodology, as well as our experimental setup. Two remotely sensed data sets are utilized for the evaluation.

3.1. Data description

ISPRS Potsdam is a public benchmark for 2D/3D semantic labeling (ISPRS, 2022). It is also widely used as a building detection benchmark (Li et al., 2021, 2022). This data set provides airborne orthoimages and corresponding DSMs generated via dense image matching. The ground sampling distance of images and DSMs is 5 cm. In our experiment, we convert these DSMs to 3D point clouds. Thus we can evaluate our methodology on a public benchmark, as there is no well-known public data set providing both annotated airborne images and well-matched original point clouds. Furthermore, we crop images from this data set into patches with a size of 512 × 512 pixels. The overlap between two up-and-down or left-and-right neighboring patches is 256 pixels. In our main experiments, a 10-shot learning case is investigated, which means only 10 randomly selected labeled patches of images and point clouds are used as the training samples. The training samples used in our 10-shot learning experiments are shown in Fig. 5.

Munich WorldView-2 is a collection of WorldView-2 satellite imagery captured over the city center of Munich, Germany. It contains two parts: orthoimages with only RGB channels, and unrasterized colorless 3D point clouds. The 3D point clouds are generated from the stereo WorldView-2 panchromatic images using the improved semi-global matching approach (Tian et al., 2013; d'Angelo, 2016). Rasterized DSMs from point clouds are adopted to orthorectify the multispectral and panchromatic images. After pansharpening, we select the red (5th), green (3th), and blue (2nd) channels from multispectral images to generate the orthoimages. The ground sampling distance of the orthoimages is 0.5 m. As Fig. 6 shows, the test region marked as A4 has a size of 6000 × 6000 pixels. The images denoted as A1, A2, and A3, each with a size of 6000 × 6000 pixels, comprise the full training data. The images marked as A5 and A6 are used as the validation sets, each of which has a size of 6000 × 3200, respectively. The building masks

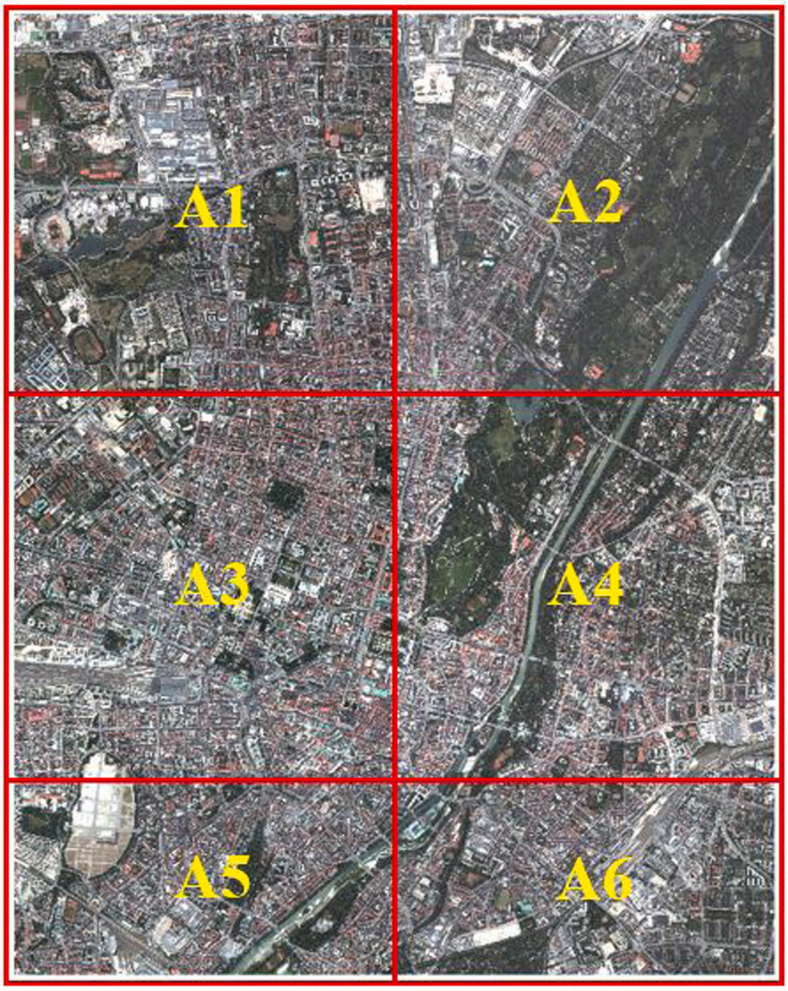


Fig. 6. The coverage of the Munich data set used in our experiment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(ground truths) of original images are manually annotated by using the open street map as a basis. The ground truths of point clouds are obtained through an affine transformation from the building masks. To satisfy the limitation of GPU memory, the full training data set has been cropped into patches with a size of 512 × 512 pixels and an overlap of 256 pixels in rows and columns. Fig. 7 shows those 10 training samples utilized in our 10-shot learning experiments on the Munich WorldView-2 data set.

3.2. Experiment setup

Our experiments are conducted within the PyTorch deep learning framework. We adopt the SparseConvNet library presented by Graham et al. (2018) to implement the sparse convolutional neural network. Training and testing are performed on a Geforce RTX 2080 Ti GPU with 11 GB RAM. All models are trained with the Adam optimizer until convergence is achieved. The scaling factor controlling the input resolution of voxels is an important parameter for sparse convolutional neural networks. Referring to the resolution of the original images, we set the input voxel size of Munich WorldView-2 to 0.5 m, and the input size of ISPRS Potsdam to 0.05 m. The learning rate is set to 0.001. The batch size of the training models is set as 4. In our experiments, the input features to the image network are red, green, and blue channels. Because in real applications the expected point cloud test data sometimes have no spectral information, only coordinate values (X, Y, and Z) are employed as input features to the point cloud neural network, ignoring potential color information provided by multispectral images.

We test both of the standard and enhanced co-learning approaches in our experiments. In order to explore the learning ability of the co-learning architecture, we do not carry out any pre-training or data augmentation operations. In the experiments of 10-shot labeled training pairs, baseline methods and standard co-learning only utilizes 10 labeled patches in the training phase. Enhanced co-learning employs 10 labeled patches as well as all remaining patches of original training data as unlabeled pairs. In short, for the ISPRS Potsdam 10-shot experiment, we used 10 labeled and 10,570 unlabeled training pairs. While for the Munich WorldView-2 experiment we employ 10 labeled and 1577 unlabeled training pairs.

Following Li et al. (2021), the F1-score and intersection over union (IoU) of the building class are selected as the evaluation metrics. In order to better evaluate the confusion between the background and buildings, overall accuracy (OA), false negative rate (FNR), and false positive rate (FPR) are reported in our work. These metrics are calculated as follows:

$$OA = \sum_{i=1}^{n} \left(\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \right), \tag{11}$$

$$F1 = \frac{2TP}{2TP + FP + FN}, \tag{12}$$

$$IoU = \frac{TP}{TP + FP + FN}, \tag{13}$$

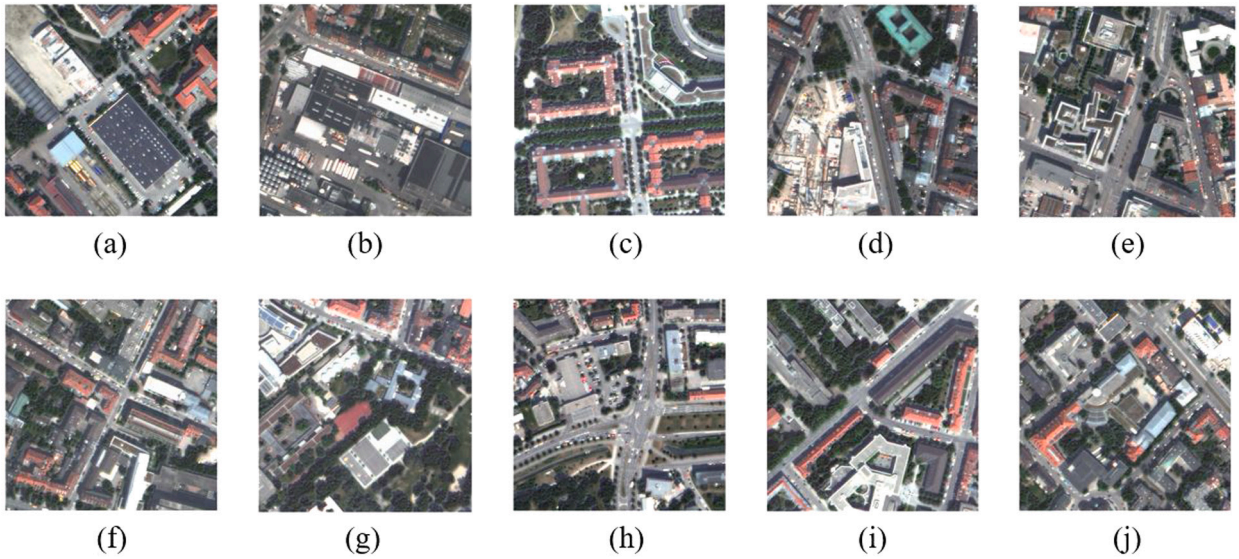


Fig. 7. 10-shot training samples of the Munich WorldView-2 data set.

Table 1
Performance of different methods for building extraction in the 10-shot ISPRS Potsdam data set.

	Methods	*OA*	IoU	F1	FNR	FPR
Image	Single-modality U-Net (baseline)	0.8795	0.5633	0.7202	0.3502	0.0483
	Early fusion U-Net (RGB + elevation)	0.9004	0.6471	0.7857	0.2364	0.0566
	Co-learning U-Net (standard)	0.8850	0.6018	0.7514	0.2734	0.0652
	Co-learning U-Net (enhanced)	**0.9370**	**0.7439**	**0.8532**	**0.2349**	**0.0089**
Point clouds	Single-modality SparseConvNet (baseline)	0.9409	0.7773	0.8747	0.1379	0.0343
	Early fusion SparseConvNet (colorized point clouds)	0.9167	0.6958	0.8206	0.2034	0.0455
	Co-learning SparseConvNet (standard)	0.9450	0.7906	0.8831	**0.1321**	0.0307
	Co-learning SparseConvNet (enhanced)	**0.9504**	**0.8059**	**0.8925**	0.1390	**0.0215**

$$FNR = \frac{FN}{TP + FN}, \tag{14}$$

$$FPR = \frac{FP}{TN + FP}, \tag{15}$$

where i is the class index and n is the total number of classes; in our case $n = 2$. *TP* refers to the number of true positives, *FP* the false positives, *TN* the true negatives, and *FN* the false negatives.

4. Results and discussion

In this section, the results of experiments for single-modality learning (as the baseline), and proposed co-learning methods are presented on the two data sets. In the experiments using the ISPRS Potsdam data set, the point cloud network is superior to the image network. In the Munich WorldView-2 experiment, the image network has a better performance. Therefore, we also explore a late fusion operation by averaging probabilities to improve the initial result of the weaker modality. Furthermore, we investigate how co-learning works on the full data set.

4.1. Comparison on the 10-shot ISPRS potsdam data set

We perform four approaches on the 10-shot ISPRS Potsdam data set and compare their results. The first is the baseline approach trained with the single-modality network. The second is with the standard co-learning. The third is with the enhanced co-learning strategy utilizing 10 labeled training pairs and all unlabeled pairs. These three approaches are conducted separately on the 2D images and 3D point clouds. The fourth approach, probability fusion, is performed only on the image modality, which has a inferior performance compared to 3D point cloud modality.

Table 2
Performance of probability enhanced image results in the 10-shot ISPRS Potsdam data set.

Methods	*OA*	IoU	F1	FNR	FPR
Enhanced co-learning	0.9370	0.7439	0.85326	0.2349	0.0089
Enhanced co-learning (fusion)	**0.9581**	**0.8291**	**0.9066**	**0.1509**	**0.0076**

4.1.1. Quantitative evaluation

The performance metrics are shown in Table 1. The best results are achieved with enhanced co-learning. Compared to the results obtained by single-modality learning, the best results of images achieved by enhanced co-learning gain increments of 5.75%, 18.06%, and 13.30% in OA, IoU, and F1, respectively. Enhanced co-learning also demonstrates an improvement over the results achieved by standard co-learning. In addition, the best FNR and FPR scores are also obtained by enhanced co-learning. When testing on point clouds, the differences among the three models are rather limited. Compared to the baseline result, the best performance achieved by enhanced co-learning strategy has an improvement of 0.95%, 2.86%, and 1.78% in OA, IoU, and F1, respectively. Enhanced co-learning also achieves the best FPR among all the methods and an FNR score very close to the best.

It should be noted that our experiments on the ISPRS Potsdam data set proved that 3D point clouds outperform images. To further explore whether the results of the weaker data type could be improved by probability fusion, we average the 2D building probability and 3D building

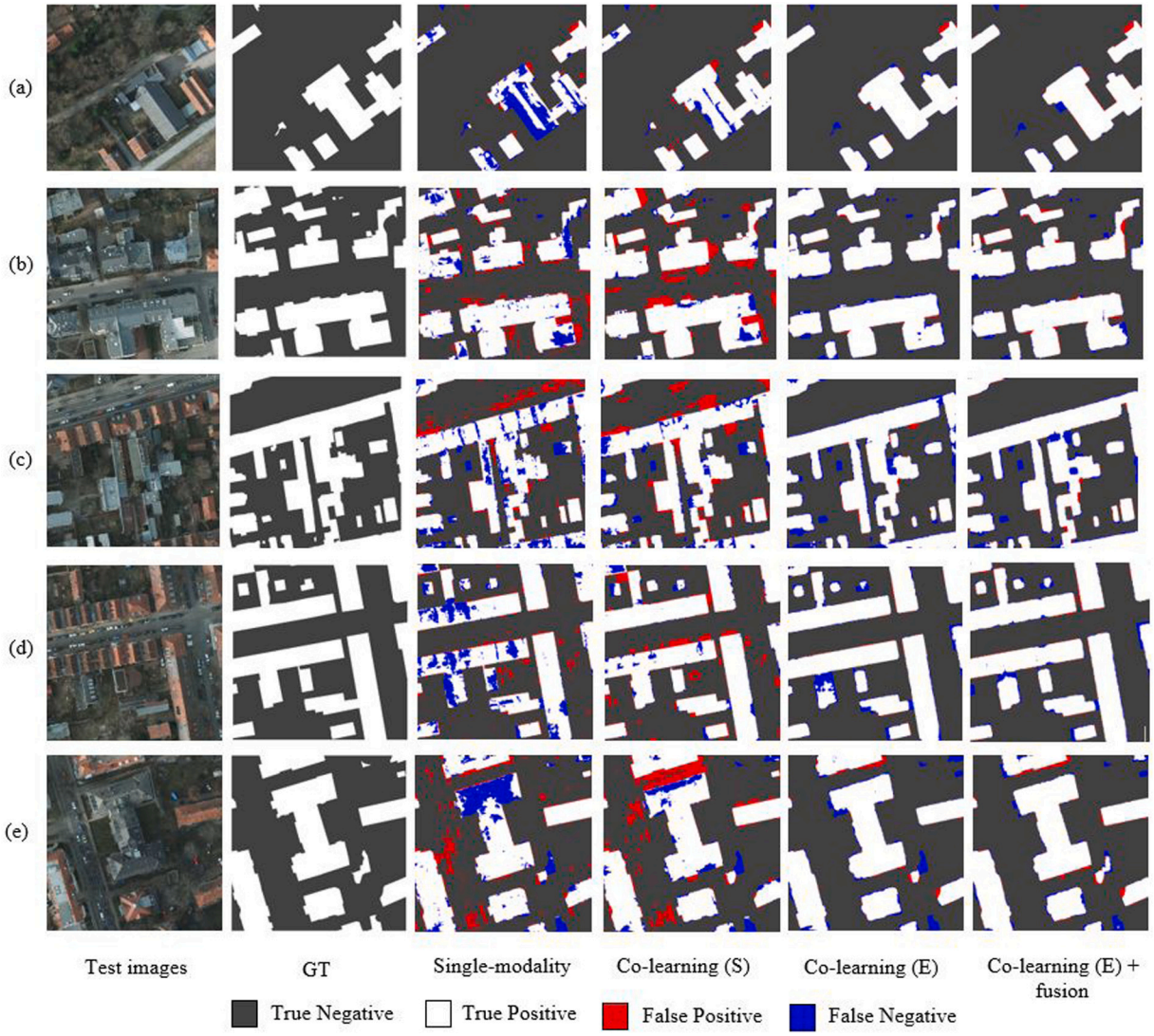


Fig. 8. 2D building extraction results obtained from the ISPRS Potsdam data set using 10 labeled training samples and various training strategies. GT: Ground truth. Co-learning (S): standard co-learning. Co-learning (E): enhanced co-learning. Co-learning (E) + fusion: enhanced co-learning and probability fusion. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

probability maps as a new probability map for the images. An image network and point cloud network trained by enhanced co-learning are used. Table 2 compares the best 2D building extraction result achieved by the image network (enhanced co-learning) and the result obtained from the fused probability map with image and DSM-derived point clouds. The probability fusion operation has a further enhancement on building extraction result, which gains an improvement of 2.11% on OA, 8.52% on IoU, and 5.34% on F1, as well as a decrease of 8.4% on FNR and 0.13% on FPR, compared with the results without fusion.

4.1.2. Qualitative evaluation

Single-modality learning is more sensitive to the quantity and quality of the training samples: thus the performance of deep learning models is restricted by the limited amount of training samples. In all five examples presented in Fig. 8, almost every building has defects, due to the poor features learned from only 10 annotated samples. Standard co-learning shows some improvement on buildings. However, many background pixels are wrongly classified as buildings. In contrast, the enhanced co-learning strategy with a large quantity of unlabeled training data achieves excellent results. In those examples, only building boundaries, small buildings, and auxiliary structures have apparent flaws. The probability fusion approach with enhanced co-learning is superior to all three of the abovementioned cases, especially at recognizing small-sized buildings, as presented in (d) and (e), which are ignored by the enhanced co-learning without fusion operation.

For building extraction from DSM-derived point clouds, a main drawback shared by all three methods is that some points of high objects are easily misclassified as buildings, since there is no spectral textural information as a constraint. Fortunately, with the mutual knowledge transferred from the image neural network, such errors are eliminated. Fig. 9 is one typical example. As shown in the circled area, both results by two types of co-learning strategies have fewer false positive points than what single-modality learning achieves. Enhanced co-learning performs the best among the training strategies.

4.2. Comparison on 10-shot Munich WorldView-2 data set

The proposed approach was also applied and evaluated on Munich WorldView-2 data set with the same experimental setting.

4.2.1. Quantitative evaluation

Table 3 shows the performance of co-learning strategies in 10-shot settings, as well as the performance of the baseline. As with the first

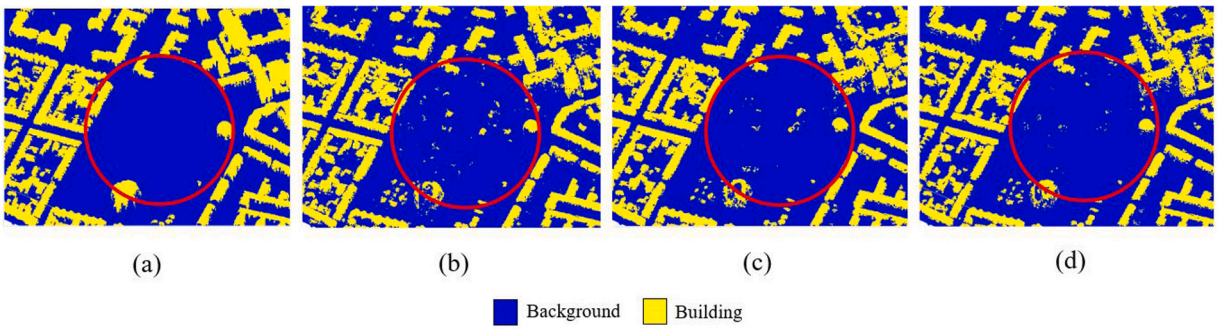


Fig. 9. Point cloud segmentation results obtained from the ISPRS Potsdam data set using 10 labeled training samples and various training strategies. (a) Ground truth. (b) Single-modality. (c) Standard co-learning. (d) Enhanced co-learning. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Performance of different methods for building extraction in the 10-shot Munich WorldView-2 data set.

	Methods	*OA*	IoU	F1	FNR	FPR
Image	Single-modality U-Net (baseline)	0.8903	0.5979	0.7484	0.1940	0.0883
	Co-learning U-Net (standard)	**0.9245**	**0.6847**	**0.8129**	**0.1899**	0.0465
	Co-learning U-Net (enhanced)	0.9224	0.6682	0.8011	0.2282	**0.0393**
Point clouds	Single-modality SparseConvNet (baseline)	0.8465	0.4753	0.6443	0.3958	0.0811
	Early fusion SparseConvNet (colorized point clouds)	0.7938	0.4756	0.6446	**0.1874**	0.2118
	Co-learning SparseConvNet (standard)	0.8492	0.5024	0.6688	0.3388	0.0946
	Co-learning SparseConvNet (enhanced)	**0.8790**	**0.5746**	**0.7298**	0.2902	**0.0703**

Table 4
Performance of probability enhanced point cloud results in the 10-shot Munich data set.

Methods	*OA*	IoU	F1	FNR	FPR
Enhanced co-learning	0.8790	0.5746	0.7298	0.2902	0.0703
Enhanced co-learning (fusion)	**0.9371**	**0.7456**	**0.8543**	**0.1984**	**0.0224**

experiment, the trained models are separately tested on images and 3D point clouds. According to the comparison results, both standard and enhanced co-learning strategies can largely improve building extraction results. For the image-based results, in comparison to the baseline method, standard co-learning achieves a 3.42% higher OA, an 8.68% higher IoU, and a 6.45% higher F1, while FNR and FPR are reduced by 0.41% and 4.18%, respectively. However, the enhanced co-learning model trained by involving unlabeled training pairs as well as labeled pairs is slightly inferior to the standard version in overall performance.

For point clouds, the improvement achieved by standard co-learning includes 0.27% in OA, 2.71% in IoU, and 2.45% in F1 score, respectively. The best performance is achieved by the enhanced co-learning strategy, where IoU and F1 are increased by 9.93% and 8.55%, and FNR and FPR are decreased by 10.56% and 1.08% in comparison with the results by the single-modality method.

Unlike the ISPRS Potsdam data set, image results are better than point cloud results in the Munich WorldView-2 data set. At this point in the experiment, we fused the probability map of point clouds and corresponding image pixels to improve the building extraction results of the point clouds. In the probability fusion experiment of the Munich WorldView-2 data set, an image network and a point cloud network trained by enhanced co-learning are utilized. As reported in Table 4, the probability fusion operation improves the point cloud results by 5.81%, 17.1%, 12.45%, 9.18%, and 4.79% on OA, IoU, F1, FNR, and FPR, respectively.

4.2.2. Qualitative evaluation

As shown in Fig. 10, many non-building areas are distinguished as buildings by the single-modality baseline method. Some of those errors are continuous areas, while others are presented as dispersed spots, so the corresponding predicted building mask looks quite noisy. For example, inside the red oval marked area, low vegetation and partial water with a light color and regular boundary can easily be distinguished as buildings by the baseline method. The explanation is that using only 10 labeled images cannot provide sufficient spectral and textural information to the deep learning models. With the help of the co-learning strategy's transferred geometric knowledge from corresponding point clouds, such false positives can be largely eliminated.

Close-up views of several image segmentation examples are presented in Fig. 11. The prediction results of the single-modality network contain a significant amount of false positive pixels. Although enhanced co-learning does not achieve the best scores in evaluation metrics, it shows better performance on complex buildings. As can be observed in (a), (b), and (c), there are more missing building structures predicted by single-modality and standard co-learning methods. However, enhanced co-learning is prone to ignore small individual houses in our experiment. As shown in (d), a large number of small-sized buildings are classified as the background by the enhanced co-learning strategy. This phenomenon commonly happens in the full test image. That is why the standard co-learning strategy has a slightly better performance than the enhanced version in the quantitative evaluation.

For point clouds, three examples are presented in Fig. 12. The second, third, and fourth columns compare results obtained by the single-modality baseline method, standard co-learning, and enhanced co-learning, respectively. As shown in (a), (b), and red and green circled areas of (c), many building structures are ignored by single-modality learning. The standard co-learning-based network can recognize more building points correctly. Enhanced co-learning achieves better accuracy in identifying complete building structures than standard co-learning. However, it sometimes results in more false positive points, as highlighted in (a) by the red and (c) by the yellow. The fourth and fifth columns of Fig. 12 qualitatively analyze the probability fusion approach and the corresponding original enhanced co-learning method on point clouds. With the help of probability fusion, many of the abovementioned errors can be eliminated, such as those circled in (a) and (c). In addition, the probability fusion approach can benefit several inconspicuous buildings, such as those highlighted by the green oval in (b).

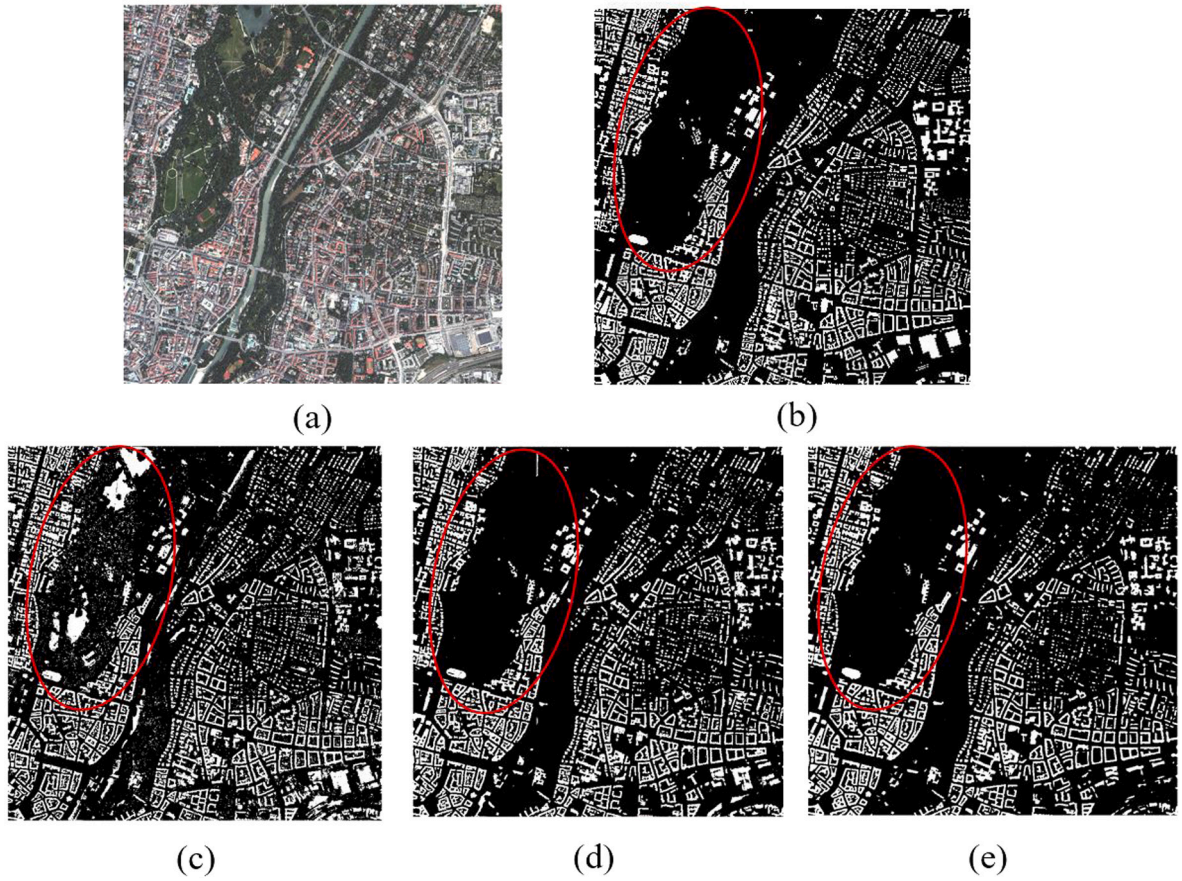


Fig. 10. The overview of image results obtained from the Munich WorldView-2 data set using 10 labeled training samples and various training strategies. (a) Original image. (b) Ground truth. (c) Single-modality. (d) Standard co-learning. (e) Enhanced co-learning. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Performance of single-modality learning and co-learning results in the ISPRS Potsdam data set with full labels. The results of EPUNet and ESFNet are from Li et al. (2022).

Methods	*OA*	IoU	F1	FNR	FPR
EPUNet (Guo et al., 2021)	–	0.7941	0.8852	–	–
ESFNet (Lin et al., 2019)	–	0.8023	0.8865	–	–
RegGAN (Li et al., 2022)	–	0.8248	0.9040	–	–
SegNet-8s-AFM (Li et al., 2021)	–	0.8275	0.9056	–	–
Single-modality U-Net	0.9486	0.7928	0.8844	0.1770	0.0120
Early fusion U-Net (RGB + elevation)	0.9678	0.8686	0.9297	0.1092	**0.0080**
Co-learning U-Net (standard)	0.9623	0.8484	0.9180	0.1183	0.0123
Co-learning U-Net (enhanced + test)	0.9673	0.8676	0.9291	0.1048	0.0100
Co-learning U-Net (enhanced + test + fusion)	**0.9759**	**0.9025**	**0.9488**	**0.0683**	0.0102

4.3. Comparison on data sets with fully labeled training data

To further investigate the potentials of the co-learning framework and compare it with the state-of-the-art single-modality networks, we conduct the experiment for building extractions based on 2D images, using fully labeled training data from the ISPRS Potsdam and Munich WorldView-2 data sets.

4.3.1. 2D building extraction from fully labeled ISPRS potsdam data set

We follow the data splitting settings of Li et al. (2021, 2022). No pre-training operation or data augmentation is carried out. Table 5 describes our results and state-of-the-art results reported by Li et al. (2021, 2022). Compared with the result achieved by our single-modality learning, the OA of the standard co-learning method is 1.37% higher, and the IoU and F1 of the building class is increased by 5.56% and 1.63%, respectively. Our 2D U-Net trained with the standard co-learning strategy achieves higher scores than the state-of-the-art results by single-modality networks reported by Li et al. (2021, 2022).

In addition, we investigate the enhanced co-learning and probability fusion operation employing the test data with images and point clouds as the unlabeled pairs. Among them, the enhanced co-learning slightly outperforms the standard co-learning approach, while the probability fusion operation achieves the best scores of OA, IoU, and F1 among all co-learning strategies. The main problem in single-modality learning with fully annotated training data is that a few building structures are classified incorrectly as the background. It has the highest FNR among all the methods. Fig. 13 gives four examples. In (b) and (d), co-learning-based methods are capable of successfully recognizing more building structures. Fig. 13(a) is an area with several industrial buildings. Due to the lack of valid training samples, it is quite challenging for the single-modality 2D U-Net model to detect these buildings correctly. It should be noted that co-learning strategies have a better performance, especially on small-sized objects. Fig. 13(c) is an extreme example: the color of two buildings is close to the color of vegetation, so they are completely wrongly classified as "background" by the model trained

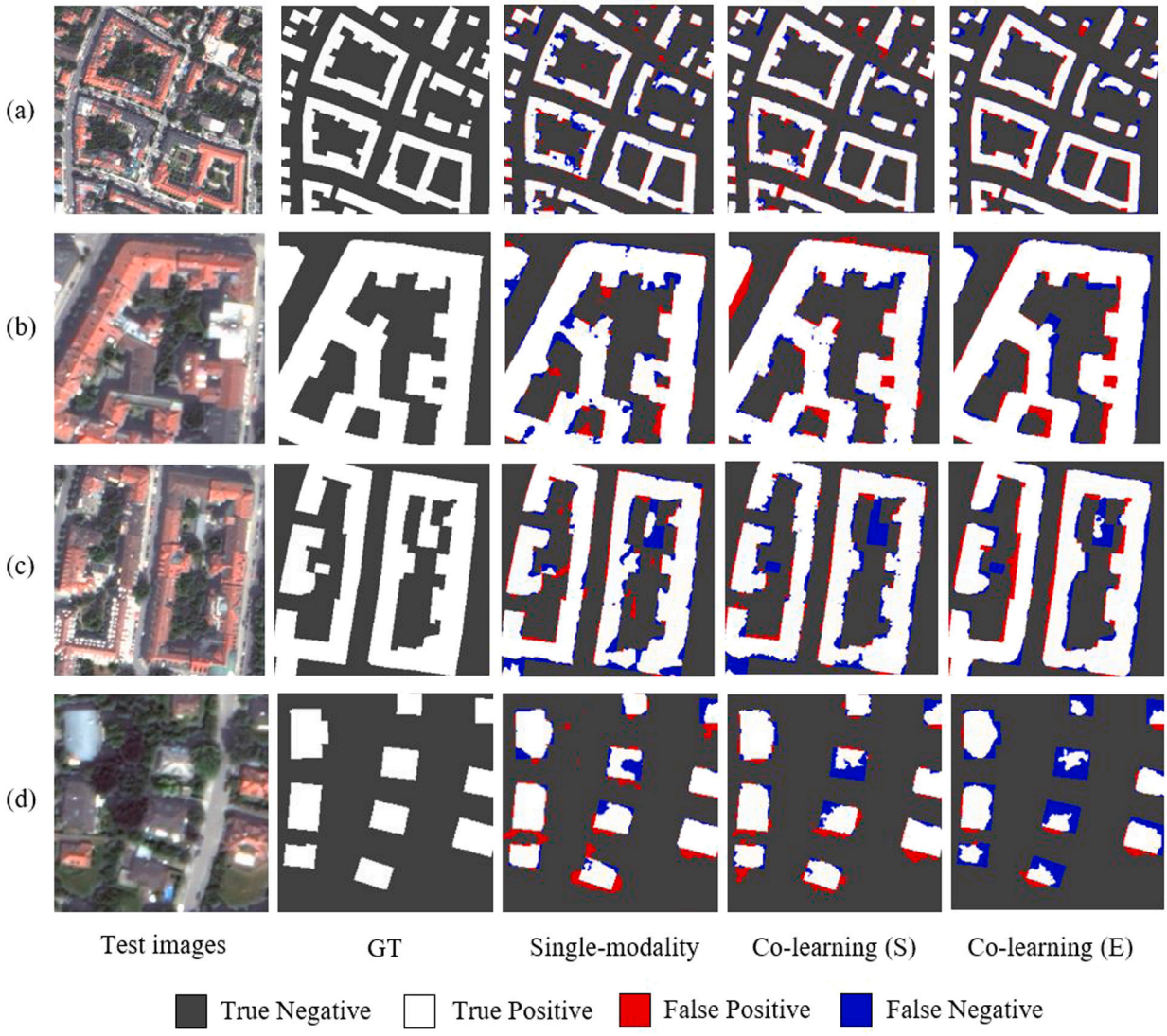


Fig. 11. Close-up views of image results obtained from the 10-shot Munich WorldView-2 data set using various training strategies. GT: Ground truth. Co-learning (S): standard co-learning. Co-learning (E): enhanced co-learning. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

by single-modality learning. With the transferred geometric knowledge from standard or enhanced co-learning, the results are slightly improved. Benefiting from the point cloud network, the probability fusion operation successfully eliminates most false negative pixels and outperforms other methods.

4.3.2. 2D building extraction from fully labeled Munich WorldView-2 data set

As shown in Table 6, the image network trained by co-learning is superior to the one trained by single-modality learning. OA, IoU, and F1 scores of the building class are increased by 1.08%, 5.64%, and 3,73%, respectively. In addition, co-learning method contributes an 8.45% lower FNR, which means it can correct many building pixels classified as non-buildings by the baseline U-Net. In Fig. 14, four visualization examples of predicted results are given. In (a) and (b), the co-learning-based network achieves greater completeness on buildings, especially at boundaries. Example (c) is an example of small-sized buildings, where the co-learning method is able to detect building structures that are more complete, although it also presents a few false positives. Example (d) is a rare case that includes round buildings and a multi-tiered square building, where standard co-learning approaches also have better performances and predict building structures that are more complete than single-modality learning.

Table 6
Performance of single-modality learning and co-learning results in the full 2D Munich data set.

Methods	*OA*	IoU	F1	FNR	FPR
Single-modality U-Net	0.9370	0.7099	0.8304	0.2384	**0.0185**
Co-learning U-Net (standard)	**0.9478**	**0.7663**	**0.8677**	**0.1539**	0.0264

4.4. Discussion

Our experiments have clearly demonstrated the advantages of the proposed co-learning framework. At first, it reduces the dependence on the quantity of annotated data. Another advantage of the proposed co-learning framework is its flexibility. First, the training data and the test data can be asymmetric. Co-learning utilizes multimodality data to train the neural networks, while the test data can be single-modality. Second, both labeled and unlabeled training pairs can be fed to the neural network, and they can be asymmetric. There is no specific requirement for the ratio of labeled to unlabeled training samples. Third, the framework can also accept conventional single-modality labeled data. As this is a generally accepted strategy to improve the generalization ability of networks, it is not tested in our paper.

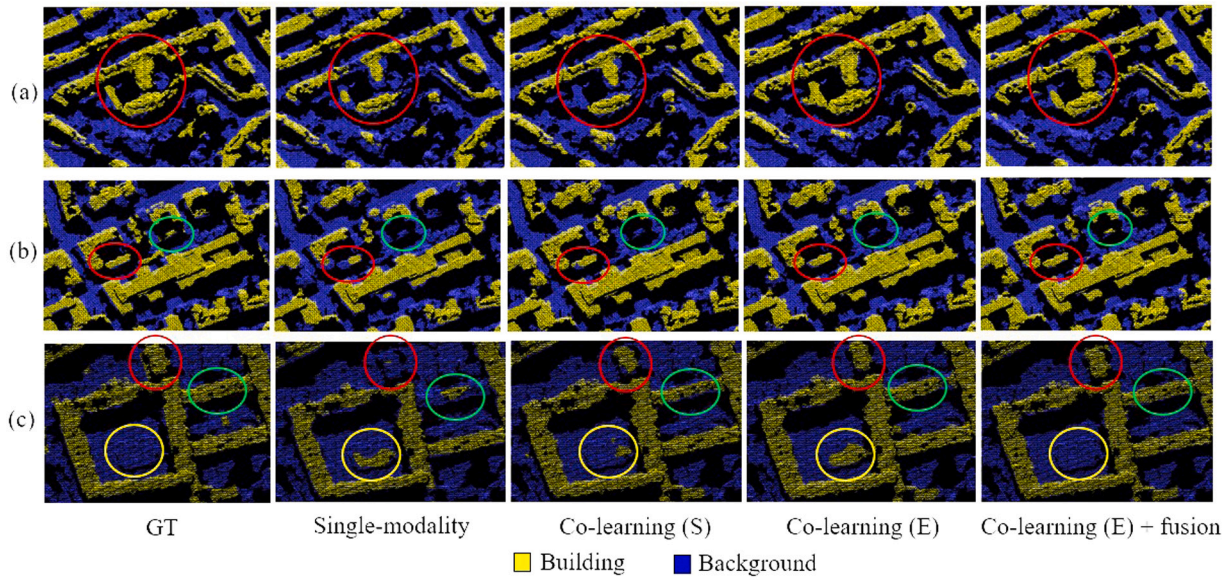


Fig. 12. Close-up views of point cloud results obtained from 10-shot Munich WorldView-2 data set using various training strategies. GT: Ground truth. Co-learning (S): standard co-learning. Co-learning (E): enhanced co-learning. Co-learning (E) + fusion: enhanced co-learning and probability fusion. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

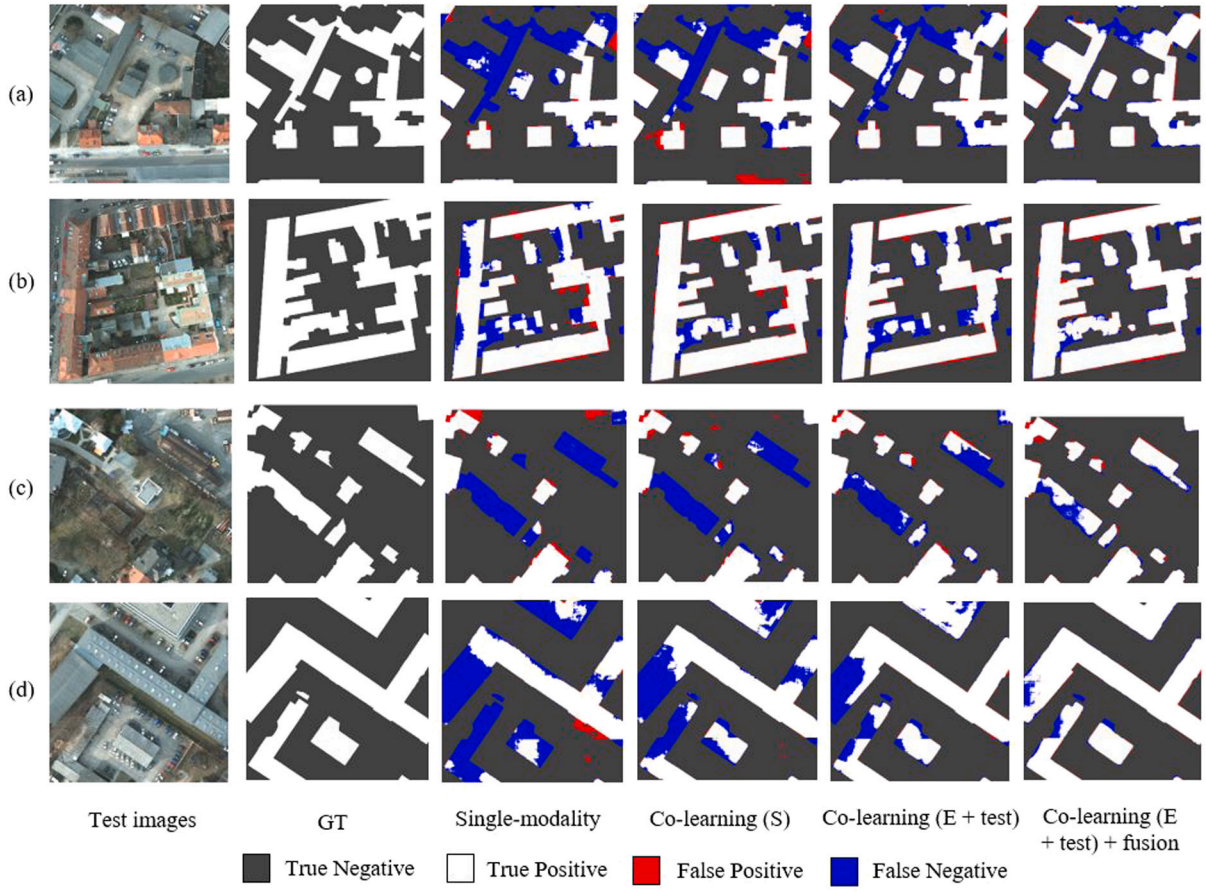


Fig. 13. 2D building extraction results obtained from the ISPRS Potsdam data set using fully labeled training data and various training strategies. Co-learning (S): standard co-learning. Co-learning (E + test): enhanced co-learning with test data. Co-learning (E+test) + fusion: enhanced co-learning with test data, and probability fusion. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

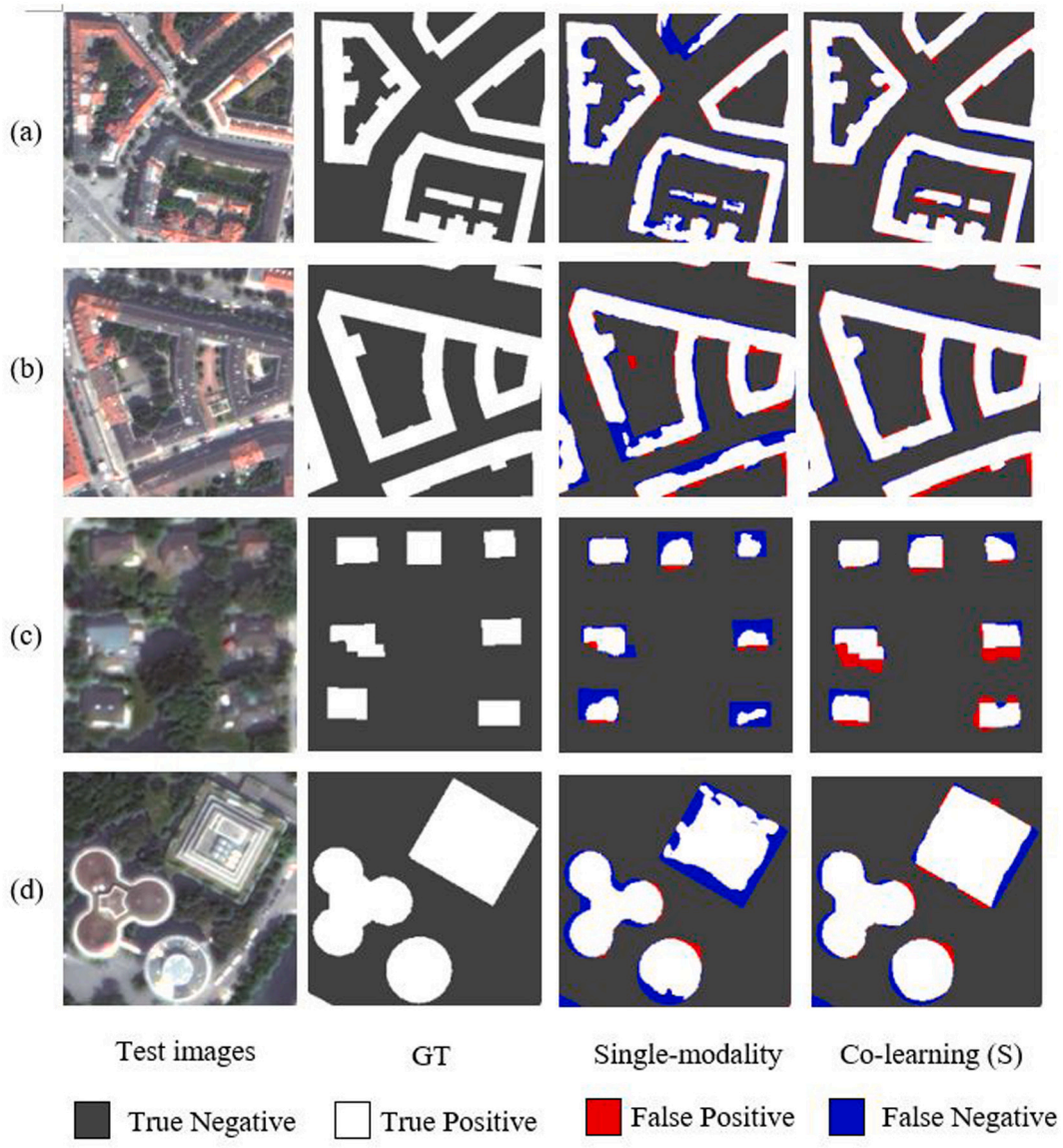


Fig. 14. 2D building extraction results obtained from the WorldView-2 Munich data set using fully labeled training data and various training strategies. GT: Ground truth. Co-learning (S): Standard co-learning. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Depending on the available data sets, co-learning can be performed in various ways. Our experiments demonstrate that co-learning is well suited for few-shot tasks. In 4.1 and 4.2, we conducted four groups of experiments with 10 labeled 2D and 3D training samples. Both standard and enhanced co-learning methods achieve superior performance compared to single-modality learning. In three of them, enhanced co-learning is superior to standard co-learning. Only in one study case is the result of enhanced co-learning slightly worse than standard co-learning. These results demonstrate that mutual information by unsupervised learning can benefit building extraction to a large extent. According to the example in 4.3, standard co-learning is also able to improve the capacity of the image network trained with full training samples. Benefiting from transferred geometric knowledge from DSM-derived point clouds, even an essential U-Net has better performance than state-of-the-art networks on ISPRS Potsdam benchmark. As presented in Section 4.3.1, test data can be used as unlabeled training data by the enhanced co-learning framework, further improving the performance of image models.

Tables 1, 3, and 5 have also reported the quantitative results by conventional early fusion. In the experiments of images, early fusion is to concatenate the elevation values from DSMs with RGB channels of corresponding images as a 4-channel input for the 2D U-Net. In the experiments of point clouds, early fusion is to utilize RGB channels projected from images as extra initial features of point clouds for the 3D SparseConvNet. When applying the early fusion strategy, both two modalities including images and point clouds/DSMs are also required in the testing phase. It has a more stringent data requirement than our co-learning methods without probability fusion. In Tables 1 and 3, early fusion is inferior to the same image backbone and point cloud backbone trained with standard and/or enhanced co-learning. For the results of point clouds in Table 1, early fusion even causes an obviously negative effect, inferior to the single-modality baseline. Color information does not lead to an increase in general performance for deep learning-based point cloud semantic segmentation. Sometimes it even reduces the performance of point cloud neural networks (Huang et al., 2020b; Bachhofner et al., 2020). In Table 5, the results obtained by early fusion are close to corresponding scores achieved by standard co-learning and enhanced co-learning. However, if multimodality test data pairs are involved, enhanced co-learning with probability fusion has a better performance than early fusion. The above phenomenons indicate co-learning methods are comparable to conventional data

fusion strategies. They can even replace conventional data fusion in some cases, with lower requirements for the test data.

Apart from the data fusion, previous deep learning-related works for building extraction mostly introduce extra modules to enhance the recognition ability of backbones (Lin et al., 2019; Guo et al., 2021; Li et al., 2021). Such methods usually target specific issues such as blur building boundaries (Guo et al., 2021; Li et al., 2021) and can achieve considerable enhancement compared with backbones. The cost of doing so is introducing more parameters for models and causing bigger model sizes as well as lower efficiency. Co-learning does not influence the structure of backbones. It exploits hidden knowledge via the communication between different modalities to optimize backbones, but it does not create redundant parameters. The usage method of models trained by co-learning is the same as the usage method of single-modality backbones. The main drawback of co-learning is more GPU memory usage and more training time, as there are two neural networks for different modalities that are trained in one GPU in parallel.

Our experiments also suggest the novel idea of utilizing photogrammetric point clouds or DSMs, which are incomplex and cheap to obtain when stereo- or multi-view high-resolution imagery is available. By comparing the results between the ISPRS Potsdam airborne data set and the Munich WorldView-2 spaceborne data set, we find that a point cloud network trained by the former yields better performance than the WorldView-2 data set. The image resolution directly influences the stereo matching results (Tian et al., 2017). With 5 cm resolution, the Potsdam point cloud data present not only sharper building boundaries, but also rich geometric features. Therefore, the buildings and trees can be well separated without the assistance of spectral information, which is the reason that the 3D point cloud in the Potsdam data set contributes to co-learning better than the Munich WorldView-2 satellite data. The absorption of more reliable transferred point cloud information by enhanced co-learning has a greater improvement on the image results of the ISPRS Potsdam data set.

5. Conclusion

In this paper, we proposed a co-learning framework for automatic building extraction from remotely sensed images and corresponding stereo/multi-view point clouds. The experiments indicate that co-learning is able to enhance the ability of a single-modality neural network by transferring mutual information from another modality with spaceborne or airborne data, and therefore is especially suitable for situations with insufficient labels. Enhanced co-learning, which is superior to standard co-learning in most experiments, shows great potential in learning with unlabeled data pairs. Fusing the prediction results from the multimodality data sets can further improve the building extraction results. Using a fully labeled data set, our method is able to further enhance the capability of the image network with the help of knowledge from corresponding photogrammetric point clouds. The experiments also show that both dense-image-matching and DSM-derived point clouds can benefit a 2D image network via co-learning. In the future, we will explore more architectures of co-learning, and introduce our framework to more diverse remote sensing tasks, such as multi-class semantic segmentation and change detection. In addition, more advanced fusion strategies will be investigated to combine the prediction results from multimodality data.

CRediT authorship contribution statement

Yuxing Xie: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Jiaojiao Tian:** Supervision, Resources, Writing – original draft, Writing – review & editing, Project administration, Methodology. **Xiao Xiang Zhu:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by a DLR-DAAD research fellowship (57424731) funded by German Aerospace Center (DLR) and German Academic Exchange Service (DAAD). The authors would like to thank Prof. Dr. Peter Reinartz for the provision of necessary data and hardware. The authors would like to thank the German Society for Photogrammetry and Remote Sensing for providing the Potsdam data set. The authors thank Dr. Pablo d'Angelo for generating point clouds from WorldView-2 images, and Xiangtian Yuan for proofreading the manuscript.

References

Bachhofner, S., Loghin, A.-M., Otepka, J., Pfeifer, N., Hornacek, M., Siposova, A., Schmidinger, N., Hornik, K., Schiller, N., Kähler, O., et al., 2020. Generalized sparse convolutional neural networks for semantic segmentation of point clouds derived from tri-stereo satellite imagery. Remote Sens. 12 (8), 1289.

Baltrušaitis, T., Ahuja, C., Morency, L.-P., 2019. Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 41 (2), 423–443.

Bittner, K., Adam, F., Cui, S., Körner, M., Reinartz, P., 2018. Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11 (8), 2615–2629.

Choy, C., Gwak, J., Savarese, S., 2019. 4D spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3075–3084.

d'Angelo, P., 2016. Improving semi-global matching: cost aggregation and confidence measure. In: XXIII ISPRS Congress, Technical Commission I, Vol. 41. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, pp. 299–304.

Graham, B., Engelcke, M., Van Der Maaten, L., 2018. 3D semantic segmentation with submanifold sparse convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9224–9232.

Guo, H., Shi, Q., Marinoni, A., Du, B., Zhang, L., 2021. Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images. Remote Sens. Environ. 264, 112589.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Huang, S., Usvyatsov, M., Schindler, K., 2020b. Indoor scene recognition in 3D. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 8041–8048.

Huang, R., Xu, Y., Hong, D., Yao, W., Ghamisi, P., Stilla, U., 2020a. Deep point embedding for urban classification using ALS point clouds: A new perspective from local to global. ISPRS J. Photogramm. Remote Sens. 163, 62–81.

ISPRS, 2022. 2D semantic labeling contest - potsdam. URL https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx.

Jaritz, M., Vu, T.-H., Charette, R.d., Wirbel, E., Pérez, P., 2020. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12605–12614.

Li, Q., Mou, L., Hua, Y., Shi, Y., Zhu, X.X., 2021. Building footprint generation through convolutional neural networks with attraction field representation. IEEE Trans. Geosci. Remote Sens..

Li, Q., Zorzi, S., Shi, Y., Fraundorfer, F., Zhu, X.X., 2022. RegGAN: An end-to-end network for building footprint generation with boundary regularization. Remote Sens. 14 (8), 1835.

Lin, J., Jing, W., Song, H., Chen, G., 2019. ESFNet: Efficient network for building extraction from high-resolution aerial images. IEEE Access 7, 54285–54294.

Lin, Y., Vosselman, G., Cao, Y., Yang, M.Y., 2021. Local and global encoder network for semantic segmentation of Airborne laser scanning point clouds. ISPRS J. Photogramm. Remote Sens. 176, 151–168.

Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., Peng, X., 2021. SMIL: Multimodal learning with severely missing modality. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35. pp. 2302–2310.

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

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660.

Rahate, A., Walambe, R., Ramanna, S., Kotecha, K., 2022. Multimodal Co-learning: Challenges, applications with datasets, recent advances and future directions. Inf. Fusion 81, 203–239.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Schmitt, M., Zhu, X.X., 2016. Data fusion and remote sensing: An ever-growing relationship. IEEE Geosci. Remote Sens. Mag. 4 (4), 6–23.

Shi, Y., Li, Q., Zhu, X.X., 2020. Building segmentation through a gated graph convolutional neural network with deep structured feature embedding. ISPRS J. Photogramm. Remote Sens. 159, 184–197.

Sun, Y., Fu, Z., Sun, C., Hu, Y., Zhang, S., 2021b. Deep multimodal fusion network for semantic segmentation using remote sensing image and LiDAR data. IEEE Trans. Geosci. Remote Sens. 60, 1–18.

Sun, X., Wang, B., Wang, Z., Li, H., Li, H., Fu, K., 2021a. Research progress on few-shot learning for remote sensing image interpretation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 2387–2402.

Thomas, H., Qi, C.R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L.J., 2019. Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6411–6420.

Tian, J., Cui, S., Reinartz, P., 2014. Building change detection based on satellite stereo imagery and digital surface models. IEEE Trans. Geosci. Remote Sens. 52 (1), 406–417.

Tian, J., Reinartz, P., d'Angelo, P., Ehlers, M., 2013. Region-based automatic building and forest change detection on Cartosat-1 stereo imagery. ISPRS J. Photogramm. Remote Sens. 79, 226–239.

Tian, J., Schneider, T., Straub, C., Kugler, F., Reinartz, P., 2017. Exploring digital surface models from nine different sensors for forest monitoring and change detection. Remote Sens. 9 (3), 287.

Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sens. Environ. 237, 111322.

Xie, Y., Tian, J., Zhu, X.X., 2020. Linking points with labels in 3D: A review of point cloud semantic segmentation. IEEE Geosci. Remote Sens. Mag. 8 (4), 38–59.

Xu, Y., Tong, X., Stilla, U., 2021. Voxel-based representation of 3D point clouds: Methods, applications, and its potential use in the construction industry. Autom. Constr. 126, 103675.

Yousefhussien, M., Kelbe, D.J., Ientilucci, E.J., Salvaggio, C., 2018. A multi-scale fully convolutional network for semantic labeling of 3D point clouds. ISPRS J. Photogramm. Remote Sens. 143, 191–204.

Zadeh, A., Liang, P.P., Morency, L.-P., 2020. Foundations of multimodal co-learning. Inf. Fusion 64, 188–193.

Zhang, L., Lan, M., Zhang, J., Tao, D., 2021. Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Zheng, Z., Ma, A., Zhang, L., Zhong, Y., 2021. Deep multisensor learning for missing-modality all-weather mapping. ISPRS J. Photogramm. Remote Sens. 174, 254–264.

Zhou, W., Jin, J., Lei, J., Hwang, J.-N., 2021. CEGFNet: Common extraction and gate fusion network for scene parsing of remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–10.

Zhu, Q., Liao, C., Hu, H., Mei, X., Li, H., 2020. MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery. IEEE Trans. Geosci. Remote Sens. 59 (7), 6169–6181.

Zoph, B., Ghiasi, G., Lin, T.-Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q., 2020. Rethinking pre-training and self-training. Adv. Neural Inf. Process. Syst. 33, 3833–3845.