A Self-Training Approach Using Benchmark Dataset and Stereo-DSM for Building Extraction

Xiangtian Yuan, Jiaojiao Tian, *Senior Member, IEEE*, and Peter Reinartz, *Member, IEEE*

***Abstract*—Deep learning has been the state-of-the-art solution to numerous remote sensing tasks, especially for building extraction. However, the performance of learning-based building extraction approaches depend to a large extent on the similarity of the source and target domain data. To alleviate the dependence on annotated data, and to exploit the potential of multimodal remote sensing data, a 3-D assisted semisupervised method for building extraction is proposed. The proposed method is based on self-training, a semisupervised method that utilizes both labeled and unlabeled data. In addition, photogrammetric digital surface model and belief function are exploited to bridge the domain gaps between the source and target data. The performance is evaluated with ISPRS Potsdam and Vaihingen benchmark datasets, and a WorldView-2 satellite multimodal dataset. Compared with the direct cross-domain test baseline, improvement of Jaccard score ranging from 8.91% to 21.39% is achieved, demonstrating the efficacy of the proposed 3-D self-training method.**

***Index Terms*—Aerial imagery, DSM, deep learning, domain gap, multimodal, pseudolabeling, satellite imagery, self-training.**

I. Introduction

BUILDING extraction has been a fundamental remote sensing (RS) task for decades, since the results are indispensable for urban planning and management, disaster relief operations, mapping, etc. Before the overwhelming adoption of deep learning methods in building extraction, hand-crafted features and machine learning methods such as support vector machine and random forest (RF) were widely applied. However, the traditional machine learning methods have poor generalization ability with data from different domains, and the hand-crafted features are sometimes susceptible to noise in data. Thereafter, the advancement of deep learning methods and hardware has brought about a paradigm shift in the solutions for many RS tasks. Since 2014, end-to-end trainable fully convolutional network (FCN) [1] has enabled the rapid emergence of research using deep neural networks to tackle semantic segmentation, which assigns a label to each pixel and is one of the solutions to building extraction in RS. FCN replaces the fully connected layer with a fully convolutional layer, enabling pixelwise fine prediction of the same size as the input image. Henceforth, numerous classic network architectures have been developed to tackle semantic segmentation, such as U-Net [2], Deeplabv3+ [3], and HRNet [4]. In addition, to accommodate for the peculiarities of RS data, innovative network structures and modules are proposed [5], [6], [7] and have achieved improved performance on several RS semantic segmentation datasets.

The performance of supervised deep learning approaches is dependent on the amount of data. As networks get deeper, the amount of parameters also grows substantially, requiring more annotated data to avoid overfitting. In addition, for deeper neural networks to learn informative features, training data need to have a high level of diversity and variability. As a consequence, the amount of publicly available building extraction datasets has been spurred by the rapid development of deep learning methods to satisfy the demand for the data-driven approach. Numerous datasets have been curated for RS building extraction task, such as ISPRS Potsdam[1] and Vaihingen,[2] Inria aerial labeling dataset [8], WHU building dataset [9] and SpaceNet,[3] which consist of remotely sensed imagery with corresponding human-annotated ground truth. This emergence of data has significantly elevated the performance of building extraction techniques based on deep learning and promoted research.

Nevertheless, despite the advancements, there remain persistent challenges within the RS community. One critical challenge is the domain gap between source and target domain datasets, which results in deteriorated performance of methods tasked with extracting buildings. Domain gap issues arise when source domain and target domain data are captured by different sensors [10], [11] are from regions with dissimilar building types, or are captured under distinct acquisition conditions [10]. One may argue that the domain gap could be closed with sufficient annotated data. Nevertheless, annotating RS data are another challenge. RS data annotation is a complex, time- and labor-intensive task that demands annotators to possess professional knowledge of the ground objects in RS images or even require region-specific knowledge. Moreover, the amount of data from satellite and aerial earth observation missions is so large that it is impossible to keep data annotation at the speed of data acquisition. As a consequence, the challenges posed by domain gaps and the impracticability of intensive data annotation incentivize RS researchers to explore learning paradigms that do not necessitate large amounts of annotated data.

Manuscript received 5 January 2024; revised 12 April 2024 and 4 June 2024; accepted 5 June 2024. Date of publication 14 June 2024; date of current version 19 June 2024. *(Corresponding author: Jiaojiao Tian.)*

The authors are with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Wessling, Germany (e-mail: xiangtian.yuan@dlr.de; jiaojiao.tian@dlr.de; peter.reinartz@dlr.de).

Digital Object Identifier 10.1109/JSTARS.2024.3412369

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

[2][Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx

[3][Online]. Available: https://spacenet.ai/datasets/

Semisupervised learning (SSL) is a potential solution to the aforementioned problems. SSL methods harness the potential of unlabeled data with the help of annotated data. In the RS community, SSL methods have been adopted and tailored according to the demands of various RS tasks [12], [13], [14], [15]. Among the SSL paradigms, self-training is a pseudo-labeling-based method of SSL and has been widely adopted due to its simplicity and practicality. However, domain shifts will result in inaccurate building pseudolabels for the target domain data. In building extraction specifically, the pseudolabels could contain nonbuildings mistakenly labeled as buildings, or buildings that are not labeled.

Very high resolution (VHR) stereoscopic imagery and its corresponding stereo-matched DSM have enabled more accurate building extraction, thanks to the finer ground sampling distance (GSD) as well as supplemental height information [11], [16]. Intuitively, the resulting 3-D elevation models can contribute to better pseudolabels by fusing additional height information that is not available in 2-D spectral imagery. As DSM contains discernible features of buildings, it can serve to provide supervisory signals for the task of building extraction [17], [18]. While extensively explored in numerous RS tasks, the benefits of the multimodal RS data are not adequately exploited in most of the pseudolabeling-based SSL methods in RS applications. Therefore, it is worthwhile to investigate the potential of stereo DSM for pseudolabeling-based semi-supervised building extraction.

With the above discussion in mind, in this work, we propose a concise and elegant SSL framework for multimodal data (stereo DSM and spectral imagery) to tackle the domain shift issue in RS building extraction. The workflow is conceived based on the real-world issues we encountered, including the performance decrease in cross-domain tests, and the difficulty in annotating incoming RS data. The method is designed to take advantage of publicly available benchmark datasets as source domain data. The main contributions of this work are as follows:

1) We present a 3-D self-training framework for building extraction. The framework exploits the potential of optical imagery and corresponding stereo DSM to reduce the negative impact of domain shift efficiently and intuitively.
2) We use the decision fusion strategy to improve the pseudolabels for self-training. The decision fusion strategy aims to refine overly confident false positives (FPs) in pseudolabel with the help of photogrammetric DSM, which can effectively and accurately reject FPs.
3) We adopt a simple and self-explanatory Tversky loss function that balances FPs and false negatives (FNs), and the parameters can be easily selected for each specific case individually.
4) We evaluate the proposed framework on one private dataset of Munich captured by the WorldView-2 satellite and two aerial multimodal benchmark datasets: ISPRS Potsdam and Vaihingen datasets. The experimental results are analyzed in detail, demonstrating the efficacy of the proposed method for multimodal semisupervised building extraction.

The remainder of the article is structured as follows: Section II reviews in detail various SSL paradigms in computer vision and RS. Section III details the proposed methodology. Section IV describes the experiments and introduces the results of the experiments and ablation study, and compares with other similar methods. In Section V, the results are discussed and analyzed thoroughly. Section VI draws the conclusion and proposes potential future improvements.

II. Related Work

A. SSL and Self-Training

SSL is gaining momentum in the last decades. The core idea is to leverage both labeled and unlabeled data to train a machine learning model. For SSL to work, certain assumptions have to be satisfied, including smoothness, cluster, and manifold assumptions [19], [20]. SSL encompasses various approaches. For example, consistency regularization is a class of methods based on the smoothness assumption stipulating that classes must be separated by low-density regions; or the manifold assumption stating that high-dimensional input data can be represented in its embedding space by multiple lower dimensional manifolds on which neighboring data points have similar labels. It can be understood as a way to leverage the unlabeled data to fit the dataset on a smooth manifold [21]. The core intuition is that perturbations applied to data should not drastically change the model's output, and such regularization can be enforced by minimizing mean squared error or Kullback–Leibler divergence between the outputs [22].

Besides consistency regularization, another SSL paradigm is self-training, which is an inductive method [19] that utilizes labeled data to train an initial model and apply the trained model to a large amount of unlabeled data to generate pseudolabel. Under the taxonomy proposed by Yang et al. [20], self-training is a subcategory of pseudolabeling methods (which also includes cotraining methods that utilize more than one model). A typical self-training scheme follows the pipeline illustrated in Algorithm 1 with possible minor variations. In the first step, a model M_{θ_0} is trained with source data D_s in a supervised manner. In the second step, the trained model M_{θ_s} makes predictions on unlabeled target data D_t. The predictions are selected according to various rules. One of the most commonly used rules is confidence filtering, which can be derived from network softmax output or other confidence measures. After the selection, only predictions with high confidence [23] are kept as pseudolabel and combined with the labeled source data, with which the model is retrained or fine-tuned (D_{new}) in step 3. It could be viewed as a sort of entropy minimization that aims to increase output confidence and has been achieved by using minimum entropy regularizers [24]. Intuitively, selecting high quality pseudolabel is of paramount importance. However, pseudolabels with high-confidence prediction do not necessarily indicate correctness and can propagate the error through training.

SSL approaches are not mutually exclusive and can be used hybrid. For example, MixMatch [25] uses a combination of mixup [26] operations and label sharpening to train on both labeled and unlabeled data. FixMatch exploits both consistency

Algorithm 1: Self-Training Workflow.

Input: Source Domain (D_s), Target Domain (D_t)
Output: Trained model ($M_{\theta_{\text{new}}}$)
  procedure Step 1: (M_{θ_0}, D_s)
    train M_{θ_0} with D_s
    update model weights θ
    return M_{θ_s}
  end procedure
  procedure Step 2: (M_{θ_s}, D_t)
    $D_{\text{new}} = D_s$
    for all $t \in D_t$ **do**
      $Pred_t \leftarrow M_{\theta_s}(t)$ ▷ inference
      if $Confidence(Pred_t) \geq threshold$ **then**
        assign $Pred_t$ as the label of t
        $D_{\text{new}} = D_{\text{new}} \cup (t, Pred_t)$
      else
        discard t
      end if
    end for
    return D_{new}
  end procedure
  procedure Step 3: (M_{θ_s}, D_{new})
    train M_{θ_s} with D_{new}
    update model weights θ_s
    return $M_{\theta_{\text{new}}}$
  end procedure

regularization and pseudolabeling strategies for image classification [27].

In dense prediction tasks such as semantic segmentation, SSL methods have been studied. However, semantic segmentation with SSL is more challenging as the smoothness assumption does not always hold, therefore hindering large-scale applications of consistency regularization in semantic segmentation [28]. Hung et al. [29] is one of the earlier frameworks utilizing self-training for semantic segmentation. They employed a generative adversarial network (GAN) based method with a pixel-based discriminator, the output of which is used to mask out low confidence regions of pseudolabels according to a fixed threshold. Mittal et al. [30] proposed a framework that comprises two network branches: semisupervised segmentation and semisupervised multilabel classification. The segmentation branch employs an adversarial strategy with feature matching loss [31] to select high-quality pseudolabel for self-training; the classification branch employs a mean teacher model to predict image-level semantic classes that are late-fused with the segmentation map to remove nonpresent classes. Ouali et al. [32] leveraged unlabeled data with a cross-consistency training strategy under the cluster assumption for semantic segmentation tasks. The method consists of a common encoder and the main decoder, which are trained with labeled data; and the auxiliary decoders, which take as input different perturbations of output from the encoder. Consistency between the outputs from the common decoder and the auxiliary decoders is enforced to enhance the encoder's representation of the data. The core intuition of this method is that in semantic segmentation, the low-density regions (from the clustering assumption) are more salient in the encoder's output than in the input images.

In the RS community, SSL methods have been adopted and tailored according to the demands of RS tasks. Li et al. [12] proposed a framework combining both self-training and consistency regularization for RS image semantic segmentation. A GAN training strategy is adopted to produce a pixelwise confidence score, which is used to reweight the loss of unlabeled data in self-training. During the self-training, consistency regularization is enforced by calculating cross-entropy loss between perturbations of pseudolabels and unperturbed pseudolabels. Zhang et al. [13] proposed an adversarial training model for road segmentation of RS imagery. First, the target domain is aligned to the source domain by a GAN-based model with a feature pyramid fusion module. In the second stage, before self-training, the pseudolabels of the target domain data are split based on their confidence scores, and the low confidence split is aligned with the high confidence split via adversarial learning. Sun et al. [33] proposed a boundary-aware semisupervised semantic segmentation network for VHR RS images based on adversarial strategy, which also employs a boundary attention module and a channel-weighted multiscale feature module. Peng et al. [14] utilized the Wallis filter and adversarial learning to reduce domain gaps between training and testing data and adopted the mean teacher model and self-training strategy for unsupervised domain adaptation (UDA) building extraction. Liu et al. [34] proposed a UDA method based on self-training for landcover mapping in urban and rural areas. Wang et al. [15] proposed a method named RanPaste, which randomly pastes part of the labeled image into the unlabeled image as a strong perturbation and then the cropped label is merged with the output of the teacher's model as the new ground truth. An adaptive threshold method to weigh the supervised loss and the semisupervised loss was used to improve the pseudolabel quality. Among myriads of SSL methods, pseudolabeling seems to be falling out of favor in the face of new state-of-the-art SSL methods. Conventional pseudolabeling-based methods (which select high-confidence pseudolabels) fared poorly due to incorrect pseudolabeled samples, which lead to noisy ground truth and therefore weak generalization ability. Nevertheless, Rizve et al. [35] argued that in addition to the simplicity, pseudo-labeling-based methods can still perform on par with consistency regularization methods, and proposed to use both confidence and uncertainty for pseudolabel selection. The uncertainty measurement is estimated by using Monte Carlo Dropout [36] and calculating the standard deviation of 10 stochastic forward passes. Cascante-Bonilla et al. [37] also argued in favor of pseudo-labeling methods. The authors stated that pseudo-labeling methods can achieve comparable results to other state-of-the-art methods and are more resilient to out-of-distribution samples. Inspired by curriculum learning [38], an algorithm named Curriculum Labeling was proposed to progressively select harder samples by varying percentile thresholds. Moreover, numerous recent works still exploit the potential of pseudo-labeling-based methods with other SSL paradigms [34], [39], [40]. Therefore, the pseudo-labeling-based methods are still worth attention, especially in multimodal RS

where additional data sources can be used to verify and improve the pseudolabel quality.

B. Stereoscopic Imaging (Spectral and DSM)

RS data encompass multiple modalities that can potentially complement one another, which is one of the very distinct features. Intuitively, additional modalities could have a positive impact on learning tasks by providing additional information. Therefore, multimodal methods have been extensively studied in RS tasks. For example, as representative RS modalities, hyperspectral and light detection and ranging data have been jointly explored and proven to effective in improving the result of the other modality [41], [42]. In satellite and airborne imagery, the photogrammetric digital surface model (DSM) is usually available along with the optical images by stereo matching methods such as semiglobal matching [43]. The derived DSM provides an additional dimension to the optical imagery and can potentially benefit many RS downstream tasks. In orthorectified images, extracting building and other semantic classes from optical images is intrinsically constrained by the lack of height information. The height information from DSM robustly distinguishes objects such as buildings and trees from roads and grass, respectively. Despite good performances in building extraction achieved by solely using optical imagery, errors that can be easily eliminated with height information can still be observed on top of the domain shift issue mentioned before. In fact, before the widespread adoption of deep learning, attempts have already been made to extract buildings from DSM alone [44], [45]. However, this paradigm is constrained by the process of DSM generation. In stereo-derived DSM, the coarse boundary and uncertainty in filled regions pose difficulties for accurate building extraction, which have stimulated the development of RS multimodal methods. Multimodal RS data have been explored in many new learning paradigms. Xie et al. [17] proposed a colearning method that exploits optical imagery and photogrammetric point clouds for building extraction. Zhou et al. [46] proposed an UDA method that fuses image and DSM during both supervised training and UDA.

III. Methodology

A. Overview

In this work, the main purpose of using the self-training scheme is to improve the model's performance on the target domain data, which exhibits a large domain gap between the source domain data. Therefore, the proposed self-training framework also serves the purpose of domain adaptation. The complete workflow of the proposed self-training method is shown in Fig. 1. In the first stage, a semantic segmentation network is trained with labeled benchmark datasets (source domain) under the conventional supervised learning paradigm. In the second stage, target domain data are fed into the trained network for inference. The softmax output is fused with the nDSM using a decision fusion method detailed in the next section. In the third stage, the fused pseudolabel is used to fine-tune the network with the earlier layers frozen.

B. Initial Prediction

The initial prediction of the building is from HRNet [4], the structure of which is illustrated in Fig. 2. The schematic illustration is shown in phase 1 of Fig. 1. The backbone and decoder can be substituted with any semantic segmentation network. In phase 1, the network is trained with the source domain dataset in a supervised manner. Typical loss functions can be used here. The details about the used loss functions are described in the following section.

C. Decision Fusion Based 3-D Pseudolabel Refinement

Since the self-training workflow uses the pseudolabel for fine-tuning, the quality of the pseudolabel is paramount to the final result. Due to the domain gap between source and target data, predictions with high confidence do not indicate correctness. On the contrary, it can indicate FP for *building*, or FN for *other classes*. One typical error in building extraction is predicting ground objects such as tennis courts and stages as buildings, as those objects in the target domain might not be present in the source domain and have similar features (shape and texture). Intuitively, FP can be eliminated with DSM by simple height comparison with surrounding pixels. Therefore, different from the typical pseudolabel filtering based on confidence, we use the local height information from the nDSM to refine the pseudolabel as shown in phase 2 in Fig. 1. Specifically, we adopt decision fusion to fuse the initial pseudolabel with the corresponding DSM. The method is detailed in Algorithm 2. Decision fusion is a powerful and lightweight tool to combine various indicators from multiple sources. It has been successfully introduced to various RS image processing tasks, including landcover classification, building extraction, and change detection [16], [47]. The general introduction of DST can be found in [48], [49], and [50].

Generally, let Θ be a frame of discernment of a problem under consideration. $\Theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$ consists of a list of N exhaustive and mutually exclusive elements $\theta_i, i = 1, 2, \ldots, N$. Each θ_i represents a possible state related to the problem we want to solve. The assumption of exhaustively and mutual exclusivity of elements of Θ is classically referred as *Shafer's model* of the frame Θ. A basic belief assignment (BBA) also called a belief mass function (or just a mass for short), is a mapping $m(.) : 2^\Theta \rightarrow [0, 1]$ from the power set[4] of Θ denoted 2^Θ to [0,1], that verifies [48]

$$m(\emptyset) = 0 \quad \text{and} \quad \sum_{X \in 2^\Theta} m(X) = 1. \tag{1}$$

$m(X)$ represents the mass of belief exactly committed to X. An element $X \in 2^\Theta$ is called a focal element if and only if $m(X) > 0$. In DST, the combination (fusion) of several independent sources of evidence is done with Dempster–Shafer (DS) rule of combination, assuming that the sources are not in total conflict.[5] DS combination of two independent BBAs $m_1(.)$ and

[4]The power set is the set of all subsets of Θ, including empty set.

[5]Otherwise DS rule is mathematically undefined because of 0/0 indeterminacy.

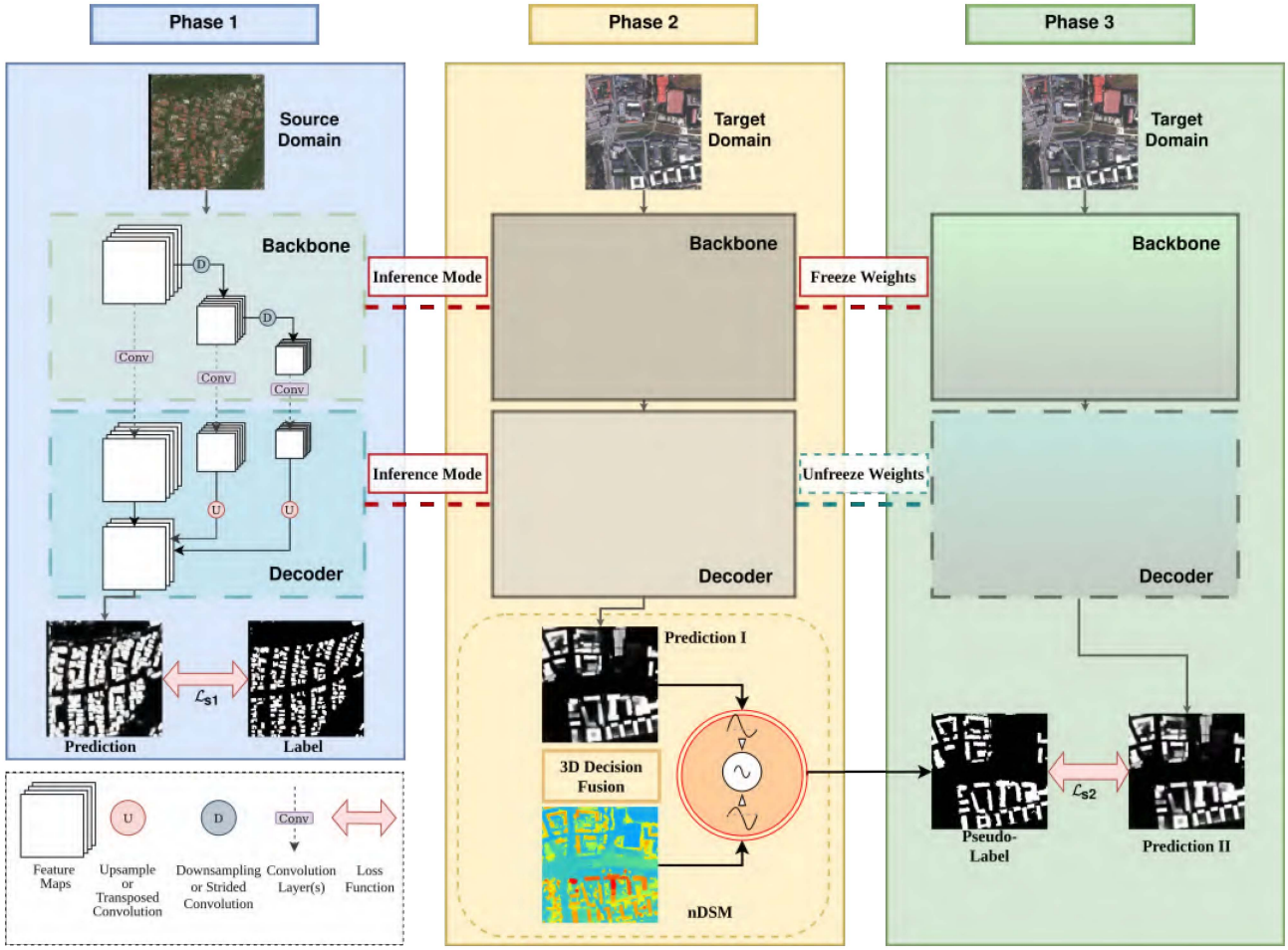

Fig. 1. General framework of the proposed 3-D self-training method. The first phase is conventional supervised training, in which the trained model is then used in phase 2 to generate Prediction (I). Henceforth, Prediction I is fused with the corresponding nDSM by 3-D decision fusion rule to produce the pseudolabel. In phase 3, the weights of certain layers are frozen, and the model is fine-tuned with the pseudolabel.

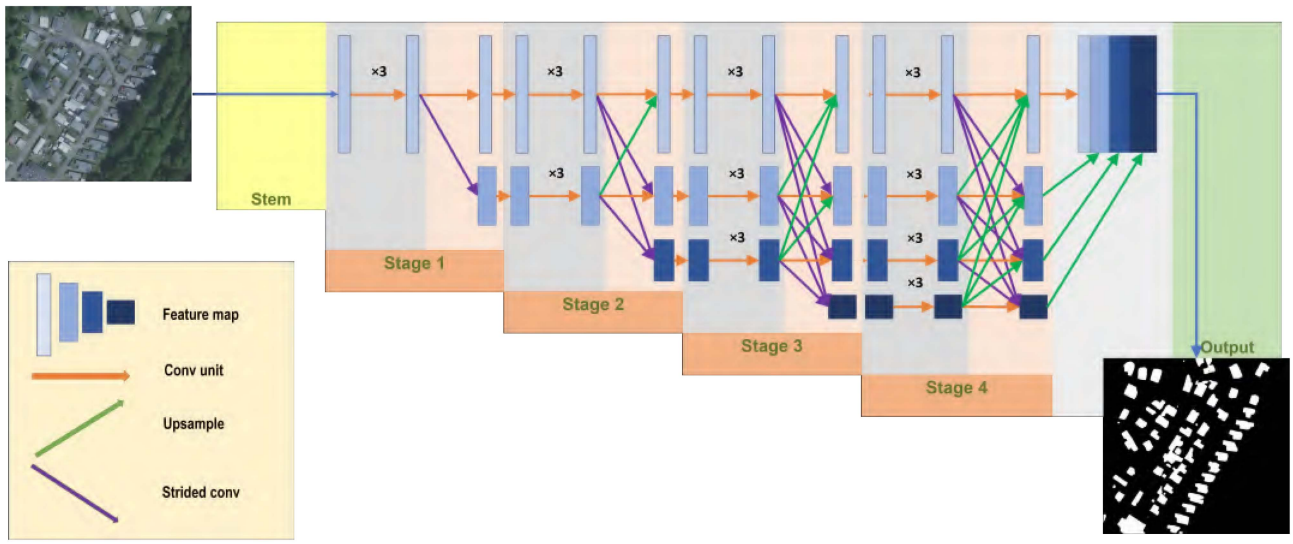

Fig. 2. Network structure of the HRNet. In each stage, the gray part denotes parallel multiresolution streams and the pink area denotes where multiresolution information exchange takes place.

Algorithm 2: 3-D Pseudolabel Refinement.

Input: initial pseudolabel (PL_0), DSM (DSM)
Output: refined pseudolabel (PL_1)
procedure 1. DSM Normalization(DSM, θ_m)
$nDSM = Morph(DSM, \theta_m)$ ▷ θ_m: parameters of Morphological filter
return $nDSM$
end procedure
procedure 2. BBA construction($nDSM$)
$m_1 = PL_0$
$m_2 = Sigmoid(nDSM - \epsilon)$ ▷ ϵ: translation parameter
return m_1, m_2
end procedure
procedure 3. DST Fusion(m_1, m_2)
Calculate K from (3)
Calculate m^{DS} from (2)
return $PL_1 = m^{DS}$
end procedure

$m_2(.)$, denoted symbolically by $DS(m_1, m_2)$, are defined by $m^{DS}(\emptyset) = 0$, and for all $X \in 2^\Theta \setminus \{\emptyset\}$ by

$$m^{DS}(X) = \frac{1}{1 - K^{DS}} \sum_{\substack{X_1, X_2 \in 2^\Theta \\ X_1 \cap X_2 = X}} m_1(X_1) m_2(X_2) \quad (2)$$

where the total degree of conflict K^{DS} is defined by

$$K^{DS} \triangleq \sum_{\substack{X_1, X_2 \in 2^\Theta \\ X_1 \cap X_2 = \emptyset}} m_1(X_1) m_2(X_2). \quad (3)$$

Focusing on building extraction, we use the initial result from the pretrained semantic segmentation model and DSM as two indicators. Two classes, which are *buildings* (B) and *no-buildings* (NB), are considered to define the frame of discernment.

The building probability map is generated by the softmax function applied on output logits. The softmax function is defined as

$$\text{Softmax}(x_i) \triangleq \frac{\exp(x_i)}{\sum_{j=1}^{K} \exp(x_j)} \quad (4)$$

where K is the number of classes and x_i is the logits score of the ith class. The resulting Softmax(x_i) is between 0 and 1. The BBA extraction approach described in [16] has been adopted in this step. A morphological filter is applied to the DSM to derive the normalized DSM (nDSM), which contains the absolute height of the land cover object. We have reprojected the nDSM to a sigmoidal curve with values ranging from 0 to 1 and used it as one set of BBA.

D. Loss Function

In the self-training workflow, the network is trained with real labels in phase 1, and fine-tuned with pseudolabels in phase 3, both in a supervised fashion. Consequently, the loss function is crucial for learning. One of the most common loss functions used is cross entropy (CE), which measures the similarity of the label and prediction probability distributions. In the first phase of the workflow, the online hard example mining (OHEM) CE function is employed; in the third phase, the CE loss, OHEM CE loss, and Tversky loss are tested.

1) Cross Entropy: The CE loss is one of the fundamental loss functions used in classification. The ground truth in training is converted to one-hot encoded vectors of each class. The soft cross-entropy is defined as

$$\text{CE} = -\sum_{c=1}^{M} q_c \log(p_c) \quad (5)$$

where M is the number of classes, q_c is the one-hot label (either 0 or 1) for class c, and p_c is the model's prediction of the probability distribution.

2) OHEM Cross Entropy: Online hard example mining (OHEM) CE loss is first proposed for object detection [51]. OHEM loss is motivated by the overwhelming amount of simple examples in many datasets, and therefore automatically selects difficult samples for more effective and efficient training. The loss function can be adapted for semantic segmentation. It selects the outputs with a confidence score lower than a threshold and only uses those outputs for optimization. The formulation of OHEM loss for semantic segmentation can be written as

$$\text{CE}_{\text{OHEM}} = \frac{1}{K} \sum_{i=1}^{N} H(p_i - T) * l_i \quad (6)$$

$$K = \text{Max}(minkeep, |\{p_i | i \subset N, p_i > T\}|) \quad (7)$$

where $H()$ is the Heaviside step function. l_i is the cross entropy loss of sample i, p_i is the softmax output of sample i, and T is the confidence threshold. N is the total number of samples in each batch. To avoid no sample being used in loss calculation, a minimum number of samples is specified by $minkeep$.

3) Tversky Loss: The 3-D decision fusion of building pseudolabel combines initial prediction with corresponding DSM, which significantly reduces FPs. However, the fused pseudolabel does not correct all FNs when the mass of belief from one indicator (initial prediction) strongly contradicts the other indicator (nDSM). Nevertheless, the problem could be greatly alleviated by simply penalizing more FP than FN in phase three, assuming the network learns effective representation for both *building* and *nonbuilding*. Tversky loss function can satisfy the demand by balancing the penalty of FP and FN with two parameters. It is defined as

$$Tversky = 1 - \frac{TP + \sigma}{TP + \alpha \times FP + \beta \times FN + \sigma}. \quad (8)$$

In (8), TP stands for true positive, and α and β control the relative penalty for FP and FN, respectively. When $\alpha = \beta = 0.5$, the equation is the dice coefficient. σ is added to avoid the 0 denominator.

IV. Experiment and Results

In this section, the details of the experiment and the results are presented. First, source and target domain data are characterized

TABLE I
OVERVIEW OF THE DATASETS

Name	Type	GSD (cm)	Size (pixel)	Tiles	Bands
Train					
xBD	satellite	50	1024×1024	11034	RGB
Test					
MUC	satellite	50	6000×6000	1	RGB
Potsdam	UAV	30	1000×1000	14	RGB
Vaihingen	UAV	20	711×1150 to 1717×1148	17	IRRG

in Table I, followed by descriptions of the experiment setup and the evaluation metrics used. Then the results are presented in detail. Comparisons with other related works are presented in the end.

A. Dataset

1) Source Domain Dataset: In this work, we adopt a large-scale open-source WorldView satellite dataset named xBD as source domain data (training data). The xBD [52] is originally intended for building damage classification and comprises pre- and postdisaster satellite image pairs, in which the building polygons are based on the pre-event image. The dataset encompasses diverse building and land cover types and has extensive coverage on various continents. To repurpose it for building extraction, we use only the pre-event images and the building annotations to train the initial model. The images are cropped into patches of 512×512 pixels with 256 pixels overlapping.

2) Target Domain Datasets: To evaluate the cross-domain adaptability of the proposed method, we select one VHR satellite dataset and two benchmark aerial datasets for testing, all of which come with stereo-DSM. Specifically, we select test data from different modalities, and locations, as well as with different GSDs, and band compositions as the typical contributors to domain shift. The satellite test data is from the WorldView-2 satellite, whose corresponding DSMs are generated by dense stereo matching (see [53] for details). The multispectral image is pan-sharpened in ENVI software using the Gram-Schmidt method. The aerial datasets are the official test sets of ISPRS Potsdam and Vaihingen. The test regions of interest (RoIs) in this study are summarized in Table I. The test images are cropped in the same way as the training data for inference, the predictions of which are then mosaicked back into the original shapes. To alleviate the border effect, each patch is multiplied with a weight matrix, which has a value of 1 for pixels in nonoverlapping regions and linearly decreases to 0 at the patch's border. With the help of the weight matrix, the artefact at the patch border can be effectively eliminated. To summarize, the target datasets are as follows:

1) *Munich Urban Area (MUC):* Shown in Fig. 3, satellite imagery with RGB band, comprises midrise residence buildings, office buildings, large factories and parks. Size: 6000×6000.
2) *ISPRS Potsdam dataset (Potsdam):* Aerial imagery with RGB bands. Downsampled to 30 cm from 5 cm. The official test set contains 14 images with 1000×1000 at 5 cm. The DSM was generated via dense image matching with Trimble INPHO 5.6 software and Trimble INPHO OrthoVista was used to generate the TOP mosaic.
3) *ISPRS Vaihingen dataset (Vaihingen):* Aerial imagery with IRRG bands (near-infrared, red and green bands), down-sampled to 20 cm from 9 cm, official test split with 17 images with different sizes are selected. DSM was generated via dense image matching with Trimble INPHO 5.3 software and Trimble INPHO OrthoVista was used to generate the TOP mosaic.

B. Training Details

All the experiments are carried out on NVIDIA TITAN Xp GPU with 12 GB memory with the Pytorch framework.[6] For the source domain training phase, the model is trained on two GPUs with a batch size of 16. The learning rate is set to 0.01 and then linearly decreases to 0. The stochastic gradient descent (SGD) optimizer with a Nesterov Momentum of 0.9 and weight decay of 0.001 is selected. Random scaling between 0.5–2.0 and random cropping are used as data augmentation. The objective function used ($\mathcal{L}_{s1}$) is OHEM Loss with a threshold of 0.9. The model is trained for 50 epochs. For the self-training phase, the weights of the first three stages (see Fig. 2) of the network are frozen, and the network is trained on two GPUs with a batch size of 16. The learning rate is set to 0.004 and then linearly decreases to 0. SGD optimizer is selected with a Nesterov Momentum of 0.9 and weight decay of 0.0005. No random scaling or cropping is employed in this stage. The training epoch is 10. All loss functions introduced in Section III-D are tested (for $\mathcal{L}_{s2}$). When using the Tversky loss, α and β combinations from 0 to 1 with an interval of 0.1 are tested, and the best combination is reported.

C. Evaluation Metrics

To quantitatively evaluate the performance, we calculate the following metrics, where TP, FP, FN stand for true positive, FP, and FN, respectively:

1) *Precision* $= \frac{\text{TP}}{\text{TP+FP}}$, which is highest if the model extracted only correct objects (i.e., no FP)
2) *Recall* $= \frac{\text{TP}}{\text{TP+FN}}$, which is highest if the model missed no object (i.e., no FN).
3) *F1 score* $= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$, is the harmonic mean of precision and recall scores
4) *Intersection over union (IoU)* $= \frac{\text{TP}}{\text{TP+FP+FN}}$, which measures the accuracy of our network by quantifying the percentage of overlapping pixels between the ground truth and our predictions.

D. Results

1) Baseline Results: The baseline results are the direct output from the network trained with a large benchmark dataset. It is essentially the result of utilizing the knowledge from the source domain only. The numeric results are shown in Table II. In MUC, the IoU score, precision and recall scores

[6][Online]. Available: https://pytorch.org

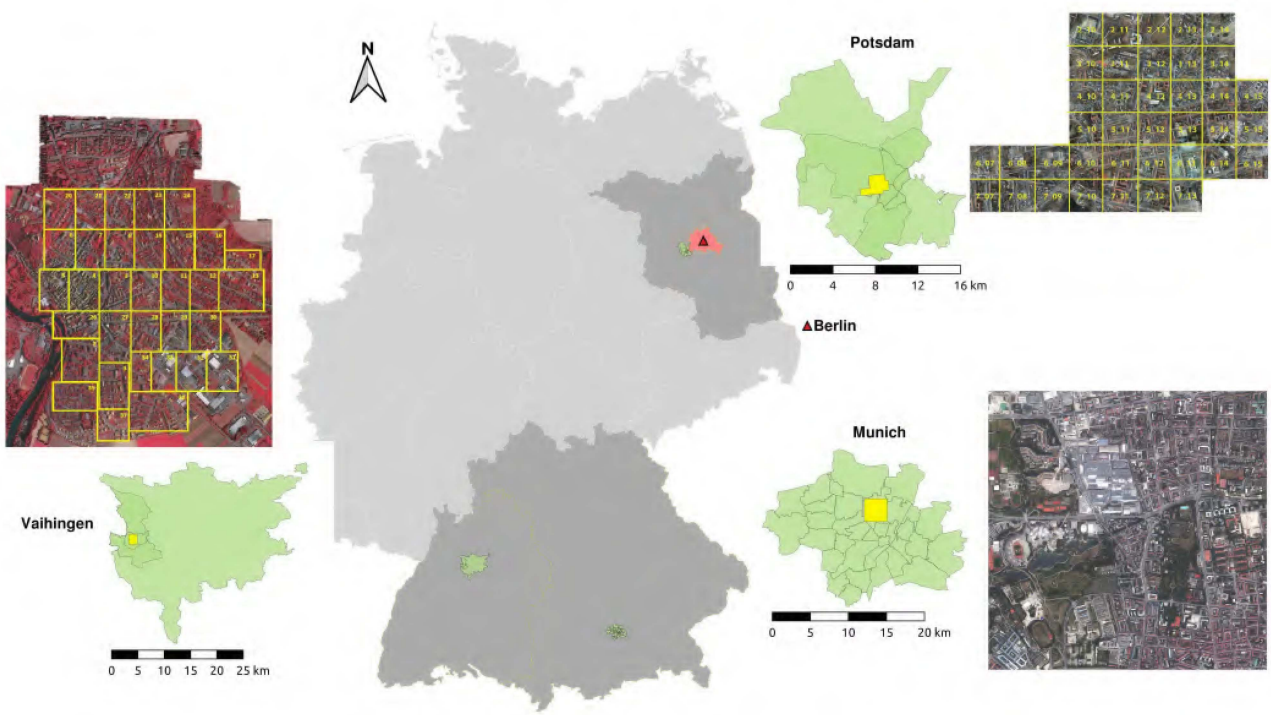

Fig. 3. Visualization of three test areas in Germany (marked by yellow polygons). The images of Potsdam and Vaihingen are taken from the websites noted in Section I. The corresponding federal states are highlighted in darker gray, and the surrounding areas are highlighted in green. The capital city Berlin is highlighted in pink and marked by a red triangle.

TABLE II
BUILDING EXTRACTION PERFORMANCE 3-D SELF-TRAINING

Location	Loss	Precision [%]	Recall[%]	F1 [%]	IoU [%]
MUC	CE	**83.22**	74.90	78.84	65.07
	T (0.4,0.6)	78.32	**82.97**	**80.57**	**67.47**
			Initial Prediction		
	-	68.95	79.52	73.86	58.56
Potsdam	CE	89.40	**90.51**	89.95	81.73
	T (0.6,0.4)	**91.72**	90.13	**90.92**	**83.35**
			Initial Prediction		
	-	79.63	89.69	84.36	72.95
Vaihingen	CE	**98.42**	22.99	37.27	22.90
	T (0.0,1.0)	75.92	**89.33**	**82.08**	**69.61**
			Initial Prediction		
	-	81.02	54.36	65.07	48.22

The best scores are marked in bold.

are 58.56%, 68.95%, and 79.52%, respectively. Higher recall indicates that the network fares better at extracting building pixels than ensuring extracted pixels are buildings, as shown in the lower left of Fig. 4(b) and lower right of Fig. 4(e) where a significant number of *nonbuilding* pixels are falsely classified as *building*, such as the paved surface near a building and the inner courtyard. At the same time, in the aforementioned areas, multiple buildings are not extracted or are only partially extracted as a result of domain gaps.

In the Potsdam scene, despite the imaging modality difference, a similar relationship between precision and recall (recall 10% higher than precision) can be observed. With an 89.69% of recall, only a small fraction of buildings is not extracted. In Fig. 4(h) and (k), the basketball court and the enclosed gardens are mistakenly segmented as *buildings*.

An entirely different phenomenon is observed in the Vaihingen data, where the recall is only 54.36% and around 27% lower than precision. As exemplified in Fig. 4(n), due to the different spectral bands, FNs are prevalent, and only a fraction of building are extracted.

2) 3-D Self-Training Results: The numeric results of the proposed 3-D self-training approach are presented in Table II. *T* refers to *Tversky loss* with the best α and β in the parenthesis. Compared with the baseline, the building extraction IoU scores in MUC, Potsdam, and Vaihingen improved around 8.91%, 10.40%, and 21.39%, respectively. In addition, in terms of precision and recall scores, both MUC and Potsdam see improvements in both metrics while the precision score of Vaihingen decreases slightly. In the MUC scene, as can be seen in Fig. 4(f), FPs such as the inner courtyard and bare ground are correctly classified. In Fig. 4(c) and (f), the *buildings* omitted in the baseline are correctly extracted. In the Potsdam results, similar improvements can be observed. In Fig. 4(i), the basketball court that shares similar geometric and spectral features with a building is not mistakenly segmented as *building* in the 3-D self-training method. As is the case in Fig. 4(l), the ground in the garden, which is misclassified in the baseline due to the domain shift, is correctly classified as *nonbuilding*. In the Vaihingen data, the best IoU is achieved with Tversky loss with $\alpha = 0$ and $\beta = 1$, meaning that only FNs will be punished in the loss calculation. With the 3-D self-training approach, most missing buildings are extracted as can be seen in Fig. 4(o).

3) 2-D Self-Training Results (Ablation Studies): To quantify how much improvement can be attributed to the DSM-based pseudolabel refinement, self-training experiments using only the 2-D pseudolabels are conducted. All experiment settings are the

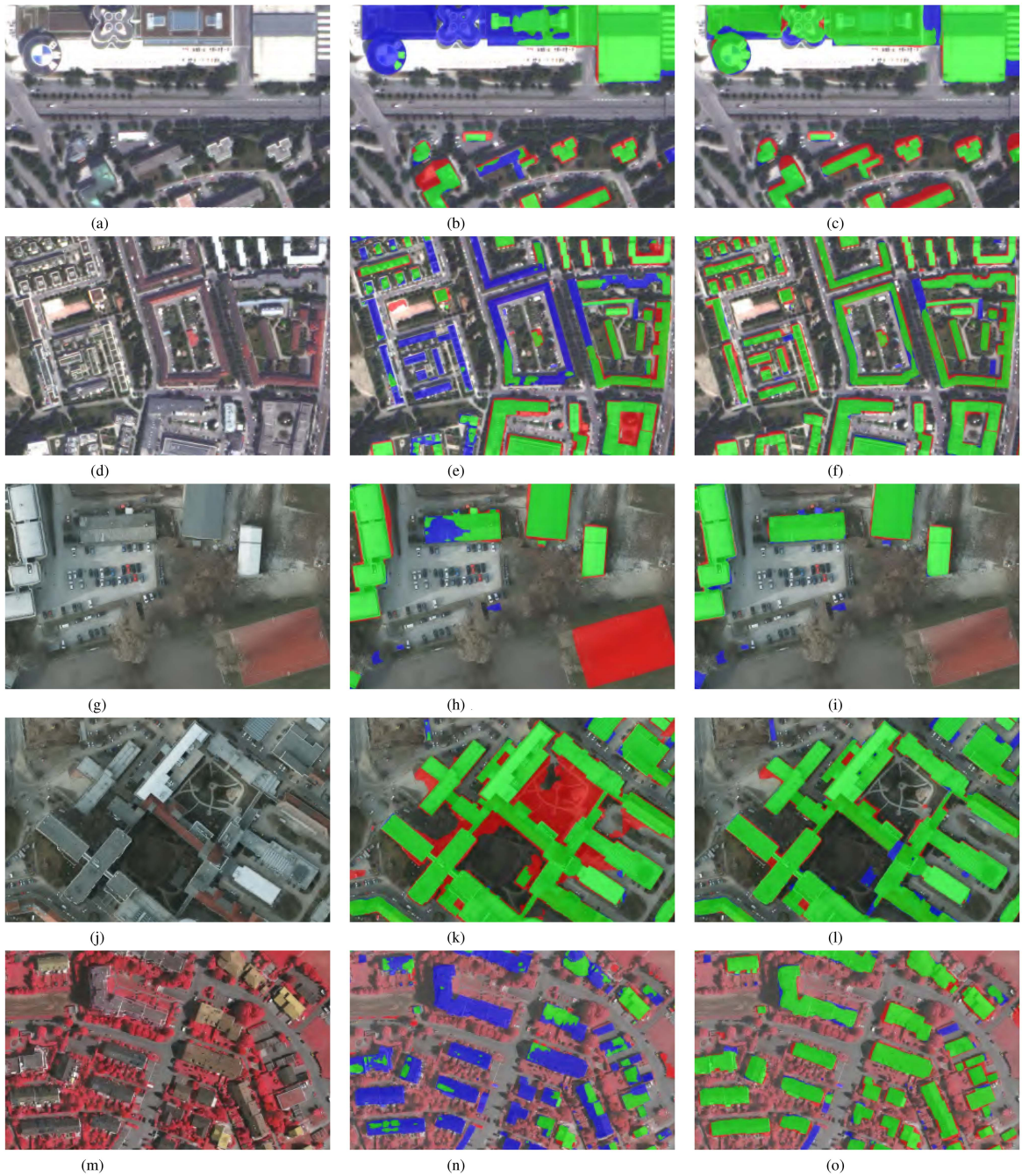

Fig. 4. Results visualization of the three test areas. 3DST denotes the proposed 3-D self-training method. Initial prediction refers to the baseline prediction. Legend: ■ True Positive ■ False Positive ■ False Negative. True Negative is not displayed. (a) Test area in Munich. (b) Initial prediction. (c) 3DST. (d) Test area in Munich. (e) Initial prediction. (f) 3DST. (g) Test area in Potsdam. (h) Initial prediction. (i) 3DST. (j) Test area in Potsdam. (k) Initial prediction. (l) 3DST. (m) Test area in Vaihingen. (n) Initial prediction. (o) 3DST.

same as in 3-D self-training. The results are listed in Table III. The IoU scores in all three test areas are lower than that of the 3-D self-training method by 5.74%, 7.07%, and 3.56% for MUC, Potsdam, and Vaihingen, respectively. The only metric better is the precision score of Vaihingen, where $\alpha = 0.3$ and $\beta = 0.7$. This is in accordance with the definition of Tversky loss. In the 3-D self-training, α is set to zero, which can potentially lead to increased FPs. In addition, the quality evaluation of the pseudolabels is presented in Table IV for analysis. Since the height-aware DSM fusion does not aim to improve the evaluation metrics of the pseudolabel, the pseudolabels actually all have lower evaluation metrics (F1 and IoU) than the final results. Improving the pseudolabel needs many ad-hoc parameters and would be trivial and time-consuming.

4) Upper Limit: To establish the upper limits of the building extraction results in the test regions, we use the ground truth as the pseudolabel for phase 3. The experiment settings are identical to the other experiments, the loss function used is OHEM CE as it yields slightly better results than CE loss. The results of Munich, Potsdam, and Vaihingen are listed in

TABLE III
BUILDING EXTRACTION PERFORMANCE 2-D SELF-TRAINING

Location	Loss	Precision [%]	Recall [%]	F1 [%]	IoU [%]
MUC	CE	69.44	79.66	74.20	58.98
	CE Soft	69.59	80.62	74.40	59.62
	CE Ohem	70.61	80.65	75.30	60.38
	T(0.7,0.3)	74.68	78.06	76.33	61.73
Potsdam	CE	78.91	92.16	85.02	73.95
	CE Soft	81.07	91.22	85.85	75.20
	CE Ohem	78.91	92.16	85.02	73.95
	T(0.7,0.3)	84.08	89.15	86.54	76.28
Vaihingen	CE	83.25	59.93	69.69	53.48
	CE Soft	81.81	56.65	66.94	50.31
	CE Ohem	84.79	56.75	68.00	51.51
	T(0.3,0.7)	80.65	78.49	79.56	66.05

TABLE IV
DSM-FUSED PSEUDOLABELS (PL) EVALUATION AND RESULTS UPPER LIMITS (UL)

Location		Precision [%]	Recall [%]	F1 [%]	IoU [%]
MUC	PL	79.87	75.23	77.48	63.24
	UL	89.68	85.66	87.63	77.98
Potsdam	PL	82.91	87.80	85.28	74.34
	UL	94.11	93.60	93.85	88.42
Vaihingen	PL	96.32	30.73	46.59	30.37
	UP	94.63	88.93	91.69	84.66

the second row for each area in Table IV, which indicates how well the model can learn under the same condition with perfect ground truth. It can be seen that within only 10 epochs, the IoU scores for Potsdam and Vaihingen datasets have increased to well above 80%, while the IoU score of Munich is at around 77%. These numbers indicate that the satellite data, while having larger number of pixels, also encompasses more diverse building types, making it more difficult for the model to converge.

5) Comparison With Other Similar Methods: To demonstrate the effectiveness of the proposed self-training workflow, we compare the building extraction results with other works that have reported results on the Potsdam and Vaihingen datasets. Specifically, we focus on work utilizing semisupervised, weakly supervised methods. Iqbal et al. [54] proposed a weakly supervised domain adaptation method named LT-WAN, which tested different permutations of training and testing data. Besides semantic segmentation, the method employs an additional image-level label to facilitate domain adaptation. Li et al. [55] manually assigned each patch to a *building* or *nonbuilding* based on the percentage of building areas. Patches with 25% or more buildings are assigned an image-level label of *building*. Conditional random field is employed twice for pseudolabel refinement and prediction refinement (denoted as Image label in the table). In the work from [14], an UDA building extraction method termed FDANet was proposed, which exploits image-, feature, and output-level information with a self-training strategy. The comparison is shown in Table V. The source domain datasets are listed in parentheses. MASS, VF, Rwan, and WHU stand for Massachusetts Building Dataset [56], Village Finder [57], Rwanda dataset [54], and WHU Building Dataset [9] respectively. Our method achieves a 37.98% to 54.55% margin in IoU scores compared with LT-WAN. With respect to FDANet with self-training, our method scores 2.12% to 12.85% higher in terms of IoU score. With no postprocessing, compared with the Image label results, our result in Potsdam has a 2.35% higher IoU score, but a 2.89% lower IoU score in Vaihingen.

TABLE V
BUILDING EXTRACTION PERFORMANCE, 3-D SELF-TRAINING COMPARED WITH OTHER METHODS ON POTSDAM AND VAIHINGEN DATASETS

Data	Method	F1 Score [%]	IoU Score [%]
Potsdam	LT-WAN (MASS) [54]	62.43	45.38
	LT-WAN (VF) [54]	44.70	28.80
	LT-WAN (Rwan) [54]	57.30	40.20
	Image label [55]	-	81.00
	FDANet (WHU) [14]	89.65	81.23
	FDANet (MASS) [14]	82.70	70.50
	ours (xbd)	90.92	83.35
Vaihingen			
	Image label [55]	-	72.50
	ours	82.08	69.61

In the parenthesis after each method, training data are specified, where MASS, VF, Rwan, and WHU represent the Massachusetts building dataset, Village Finder dataset, Rwanda dataset, and WHU building dataset, respectively.

V. DISCUSSION

The proposed method bridges the domain gap between training and testing data by exploiting the intrinsic information from photogrammetric DSM, as opposed to other methods that center on image-level features. In addition, contrary to the common practice in self-training that eliminates unreliable pseudolabels by confidence, our method utilizes the belief function to fuse DSM with pseudolabels and adopts the Tversky loss to offset the negative effects of inaccurate pseudolabels. The proposed method has significant practical implications for many RS applications where the test data have large domain gaps from the training data. Our method is fast to implement, which is crucial for time-sensitive tasks such as disaster relief operations. In addition, the network in the framework can be substituted with any semantic segmentation network, making it flexible and adaptable. In this chapter, details of domain gaps will be analyzed. To elucidate the most pivotal aspects of the method, the influence of the pseudolabel quality and the crucial role of DSM are discussed.

A. Domain Gaps Analysis

In RS tasks, the domain gaps can be attributed to differences in modality, sensor design, GSD, band compositions; as well as variations in data location, urban types, building styles, etc. Generally speaking, the common domain gaps come from varying acquisition locations of the source and target domain datasets. In our experiments, the different locations result in dissimilarities in landcover, urban types, and building styles. In addition, differences in modality, GSD, and band composition are considered. Within the experiments, different modalities do not necessarily lead to worse performance, as evidenced by MUC and Potsdam, where the baseline result of MUC is lower than that of Potsdam. In fact, the difference in GSDs plays a bigger role according to our experiments using the original 5 cm resolution Potsdam data (numeric results not shown). It is intuitive that the extracted *building* features in high-resolution imagery are different from

those extracted from lower resolution imagery, as more details are available. As the GSD difference decreases, the resulting domain gap also shrinks. Band composition is a factor that contributes to the largest domain shift, as evidenced by the results of Vaihingen, where the baseline IoU is the lowest. The responses of ground object to electromagnetic waves with distinct wavelengths can be drastically different. Consequently, the model could potentially fail to extract meaningful spectral features that characterize *building*, resulting in a low recall score. As a consequence, the Vaihingen dataset is less used than the Potsdam dataset in literature.

B. Influence of the Quality of Pseudolabels

Different from methods that filter pseudolabel based on the confidence score, which is actually not a true measurement of confidence, our method exploits the dichotomy in prediction, i.e., either hit or miss. Looking at the evaluation metrics of the pseudolabels, it can be observed that the best α and β values are related to the relationship between precision score (Precision) and recall score (Recall). Specifically, when $Precision \gg Recall$, the best performance is achieved with $\alpha < \beta$ and vice versa. When the difference between Precision and Recall is small, α and β are both closer to 0.5. With different ratios between the penalty factors on FP and FN, drastic improvements can be observed in all test regions, especially in Vaihingen where the baseline result is the worst. In Tables II and IV, the pseudolabel DSM fusion does not necessarily improve the overall evaluation metrics of the pseudolabel. Instead, the DSM fusion drastically increases the precision scores at the cost of the recall score. Nevertheless, the performance of the 3-D self-training method triumphs over the self-training without DSM fusion, despite that the pseudolabels of the latter have higher F1 and IoU scores. Therefore, we can conclude from the experiments that with respect to building extraction, the overall quality of the pseudolabel is not pivotal to the final results, when the errors can be accounted for during the optimization. Nevertheless, the role of the quality of the pseudolabel should not be totally downplayed, as evidenced in the Vaihingen results, which have the lowest IoU and F1 scores among all the test sets, as well as the evaluation metrics of the pseudolabel. Therefore, it should be in mind that the quality of the pseudolabel still plays a nonnegligible role in the final results, as it is constrained by the number of correct supervision signals.

C. Impact of DSM Accuracy on the Fusion Model

In the ablation study, the positive influence of the DSM on the pseudolabel has been verified. The improvement introduced by the height information through decision fusion seems counterintuitive when looking at the decreased IoU and F1 scores of the fused pseudolabel. Nevertheless, the decision fusion serves to accurately remove FPs in the pseudolabel, which in many cases are difficult to remove without referring to height information. With a more precise but less complete pseudolabel, and a loss function penalizing more FNs, the network can learn meaningful building representation while discounting false supervisory signals. Therefore, as can be seen in the result visualization, the DSM-fused pseudolabels help to eliminate FPs in the baseline predictions. Moreover, the DSM-fused pseudolabels can also contribute to removing FNs, as the features extracted by the fine-tuned network in areas with correct building pseudolabels tend to be similar to features extracted in areas where buildings are not annotated in pseudolabels, assuming that the target data do not exhibit large intravariance.

TABLE VI
TRAINING TIME

Location	Time (s)	Speed (km^2/s)	Speed (mpx/s)
MUC	472	0.019	0.076
Potsdam	415	0.003	0.034
Vaihingen	420	0.0017	0.043

D. Advantages

This work is driven by real-world demands where accurate building prediction is needed within a short period of time. The training time for all three area is summarized in Table VI, where the speed with respect to both square kilometer and megapixel is calculated. The proposed workflow requires around 330 s for building extraction of 1 km^2 with 30 cm GSD. In addition, the workflow is clean and requires minimum ad-hoc parameters and hyperparameter tuning to derive reasonable results. Moreover, the height information and the Tversky loss function enable an explainable interpretation of the building extraction performance, which is difficult for GAN-based models. The workflow can also work without DSM and achieve results better than the baseline. With the development of real-time DSM generation methods, the workflow could be further expedited. For example, dAngelo and Kurz [53] proposed a real-time sliding window-based bundle adjustment method, which significantly improves image orientations and DSM quality and allows generating detailed DSMs with a resolution of $2\times$GSD.

VI. CONCLUSION

In this work, we propose a fast and effective 3-D self-training method for building extraction. The method exploits publicly available benchmark datasets by tackling the domain shift issues with DSM, which has significant implications for real-world applications. To our knowledge, it is the first work that adopts DSM for self-training and pseudolabel refinement.

We conduct comprehensive experiments to evaluate the performance of the proposed workflow with one private dataset and two public benchmark datasets, and the results show significant improvements compared with the baseline and other domain adaptation methods on the ISPRS benchmark datasets.

The method can handle the domain shift issues attributed to different sensors, modalities, or spectral bands, without needing ad-hoc adjustment for specific domain issues. Contrary to domain adaptation methods that employ consistency regularization and GAN that require often different perturbations, our method has the advantage of saving computation power and time. The method is evaluated with HRNet, which can be substituted with any semantic segmentation network.

In the future, we intend to determine the parameters α and β in an automatic manner depending on the target domain data. In addition, we will exploit the potential of DSM within a self-supervised framework for domain adaptation on RS applications.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, Munich, Germany, 2015, pp. 234–241.

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018.

[4] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[5] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 135, pp. 158–172, 2018.

[6] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5412012.

[7] F. Zhou, R. Hang, and Q. Liu, "Class-guided feature decoupling network for airborne image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2245–2255, Mar. 2021.

[8] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 3226–3229.

[9] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

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

[11] J. Tian, X. Zhuo, X. Yuan, C. Henry, P. d'Angelo, and T. Krauss, "Application oriented quality evaluation of Gaofen-7 optical stereo satellite imagery," *ISPRS Ann. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. 1, pp. 145–152, 2022.

[12] J. Li, B. Sun, S. Li, and X. Kang, "Semisupervised semantic segmentation of remote sensing images with consistency self-training," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5615811.

[13] L. Zhang, M. Lan, J. Zhang, and D. Tao, "Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5609413.

[14] 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 Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5607317.

[15] J.-X. Wang, S.-B. Chen, C. H. Q. Ding, J. Tang, and B. Luo, "RanPaste: Paste consistency and pseudo label for semisupervised remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 2002916.

[16] J. Tian, S. Cui, and P. Reinartz, "Building change detection based on satellite stereo imagery and digital surface models," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 406–417, Jan. 2014.

[17] Y. Xie, J. Tian, and X. X. Zhu, "A co-learning method to utilize optical images and photogrammetric point clouds for building extraction," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 116, 2023, Art. no. 103165.

[18] X. Yuan, J. Tian, T. Krauß, X. Zhuo, and P. Reinartz, "Multi-layer thematic map representation for urban understanding," in *Proc. Joint Urban Remote Sens. Event*, 2023, pp. 1–4.

[19] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[20] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 9, pp. 8934–8954, Sep. 2023.

[21] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, no. 11, pp. 2399–2434, 2006.

[22] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 3239–3250.

[23] D.-H. Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Int. Conf. Mach. Learn., Workshop Challenges Representation*, 2013, pp. 1–6. [Online]. Available: https://openreview.net/pdf?id=3iGjh_NmoG

[24] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 529–536.

[25] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "MixMatch: A holistic approach to semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, Art. no. 454.

[26] H. Zhang, M. Cissè, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. 6th Int. Conf. Learn. Representations*, Vancouver, BC, Canada, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[27] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 596–608.

[28] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, varied perturbations," in *Proc. 31st Brit. Mach. Vis. Conf.*, BMVA Press, 2020, pp. 1–14. [Online]. Available: https://www.bmvc2020-conference.com/assets/papers/0680.pdf

[29] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *Proc. Brit. Mach. Vis. Conf.*, Newcastle, UK: BMVA Press, 2018, pp. 1–12. [Online]. Available: http://bmvc2018.org/contents/papers/0200.pdf

[30] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high-and low-level consistency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1369–1379, Apr. 2021.

[31] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[32] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12674–12684.

[33] X. Sun, A. Shi, H. Huang, and H. Mayer, "Bas ^{4} net: Boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5398–5413, 2020.

[34] W. Liu, J. Liu, Z. Luo, H. Zhang, K. Gao, and J. Li, "Weakly supervised high spatial resolution land cover mapping based on self-training with weighted pseudo-labels," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, 2022, Art. no. 102931.

[35] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," in *Proc. 9th Int. Conf. Learn. Representations*, Virtual Event, Austria, May 3–7, 2021. [Online]. Available: https://openreview.net/forum?id=-ODN6SbiUU

[36] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2016, pp. 1050–1059.

[37] P. Cascante-Bonilla, F. Tan, Y. Qi, and V. Ordonez, "Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 6912–6920.

[38] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.

[39] Y. Wang, J. Peng, and Z. Zhang, "Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9092–9101.

[40] Y. Wang et al., "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4248–4257.

[41] R. Hang, X. Qian, and Q. Liu, "Cross-modality contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532812.

[42] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, "Deep encoder–decoder networks for classification of hyperspectral and LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 5500205.
[43] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.
[44] H. Arefi, "From LiDAR point clouds to 3D building models," Ph.D. dissertation, Inst. Appl. Comput. Sci., Bundeswehr Univ. Munich, Munich, Germany, 2009. [Online]. Available: https://elib.dlr.de/60168/
[45] H. Arefi, M. Hahn, and H. Reinartz, "Ridge based decomposition of complex buildings for 3D model generation from high resolution digital surface models," in *Proc. Int Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.: WG I/4 ISPRS Istanbul Workshop "Model. Opt. Airborne Space Borne Sensors"*, vol. XXXVIII-1/W17, Oct. 2010, pp. 1–7. [Online]. Available: https://www.isprs.org/proceedings/XXXVIII/1-W17/15_Arefi.pdf
[46] S. Zhou et al., "DSM-assisted unsupervised domain adaptive network for semantic segmentation of remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5608216.
[47] J. Tian and J. Dezert, "Fusion of multispectral imagery and DSMs for building change detection using belief functions and reliabilities," *Int. J. Image Data Fusion*, vol. 10, no. 1, pp. 1–27, 2019.
[48] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, NJ, USA: Princeton Univ. Press, 1976.
[49] J. Dezert and A. Tchamova, "On the validity of dempster's fusion rule and its interpretation as a generalization of Bayesian fusion rule," *Int. J. Intell. Syst.*, vol. 29, no. 3, pp. 223–252, 2014.
[50] F. Smarandache and J. Dezert, *Advances and Applications of DSmT for Information Fusion*. Rehoboth, NM, USA: American Research Press, 2004–2015. [Online]. Available: http://www.onera.fr/staff/jean-dezert?page=2
[51] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 761–769.
[52] R. Gupta et al., "Creating xBD: A dataset for assessing building damage from satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019.
[53] P. d'Angelo and F. Kurz, "Aircraft based real time bundle adjustment and digital surface model generation," *Int. Arch. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. XLII-2/W13, pp. 1643–1647, 2019. [Online]. Available: https://isprs-archives.copernicus.org/articles/XLII-2-W13/1643/2019/
[54] J. Iqbal and M. Ali, "Weakly-supervised domain adaptation for built-up region segmentation in aerial and satellite imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 263–275, 2020.
[55] Z. Li, X. Zhang, P. Xiao, and Z. Zheng, "On the effectiveness of weakly supervised semantic segmentation for building extraction from high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3266–3281, 2021.
[56] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Dept. Comput. Sci., Univ. Toronto, 2013. [Online]. Available: https://www.cs.toronto.edu/~vmnih/docs/Mnih_Volodymyr_PhD_Thesis.pdf
[57] K. Murtaza, S. Khan, and N. Rajpoot, "VillageFinder: Segmentation of nucleated villages in satellite imagery," in *Proc. Brit. Mach. Vis. Conf.*. BMVA Press, 2009, pp. 83.1–83.11, doi: 10.5244/C.23.83.

Xiangtian Yuan received the B.Eng. degree in civil and environmental engineering from East China Normal University, Shanghai, in 2015, and the M.Sc. degree in civil engineering from the University of Washington, Seattle, CA, USA, in 2017. He is also currently working toward the Ph.D. degree in geoinformatics with German Aerospace Center, Wessling, Germany.

Since 2018, he has been working with the Photogrammetry and Image Analysis Department, Remote Sensing Technology Institute, German Aerospace Center. His research interests include the application of deep learning in multimodal remote sensing, with focus on urban remote sensing, disaster monitoring, and 3-D change detection.

Jiaojiao Tian (Senior Member, IEEE) received the B.S. degree in geoinformation systems from the China University of Geoscience, Beijing, China, in 2006, the M.Eng. degree in cartography and geoinformation from the Chinese Academy of Surveying and Mapping, Beijing, China, in 2009, and the Ph.D. degree in mathematics and computer science from Osnabrück University, Osnabrück, Germany, in 2013.

Since 2009, she has been with the Photogrammetry and Image Analysis Department, Remote Sensing Technology Institute, German Aerospace Center, Wessling, Germany, where she is currently the Head of the 3-D and Modeling Group. She was a Guest Scientist with the Institute of Photogrammetry and Remote Sensing, ETH Zürich, Switzerland, in 2011. She is a Co-Chair of the ISPRS Commision WG I/8: Multi-sensor Modelling and Cross-modality Fusion. She is a member of the editorial board of the *ISPRS Journal of Photogrammetry and Remote Sensing* and of the *International Journal of Image and Data Fusion*. Her research interests include 3-D change detection, digital surface model generation, 3-D point cloud semantic segmentation, object extraction, and DSM-assisted building reconstruction, forest monitoring, and classification.

Peter Reinartz (Member, IEEE) received the Diploma (Dipl.-Phys.) in theoretical physics from the University of Munich, Munich, Germany, in 1983 and the Ph.D. (Dr.-Ing) degree in civil engineering from the University of Hannover, Hannover, Germany, in 1989.

His dissertation is on optimization of classification methods for multispectral image data. He is currently the Department Head of the department "Photogrammetry and Image Analysis," German Aerospace Centre (DLR), Remote Sensing Technology Institute (IMF) and holds a professorship for computer science at the University of Osnabrück. He has more than 35 years of experience in image processing and remote sensing and over 500 publications in these fields. His main interests are in machine learning, stereo-photogrammetry, and data fusion using space borne and airborne image data, generation of digital elevation models and interpretation of very high resolution data from satellite cameras like WorldView, and Pleiades. He is also engaged in using remote sensing data for disaster management and using high frequency time series of airborne image data for real-time image processing and for operational use in case of disasters as well as for traffic monitoring.