DML: Differ-Modality Learning for Building Semantic Segmentation

Junshi Xia, *Senior Member, IEEE*, Naoto Yokoya, *Member, IEEE*, and Gerald Baier, *Member, IEEE*

***Abstract*— This work critically analyzes the problems arising from differ-modality building semantic segmentation in the remote sensing domain. With the growth of multimodality datasets, such as optical, synthetic aperture radar (SAR), light detection and ranging (LiDAR), and the scarcity of semantic knowledge, the task of learning multimodality information has increasingly become relevant over the last few years. However, multimodality datasets cannot be obtained simultaneously due to many factors. Assume that we have SAR images with reference information in one place and optical images without reference in another; how to learn relevant features of optical images from SAR images? We refer to it as differ-modality learning (DML). To solve the DML problem, we propose novel deep neural network architectures, which include image adaptation, feature adaptation, knowledge distillation, and self-training (SL) modules for different scenarios. We test the proposed methods on the differ-modality remote sensing datasets (very high-resolution SAR and RGB from SpaceNet 6) to build semantic segmentation and to achieve a superior efficiency. The presented approach achieves the best performance when compared with the state-of-the-art methods.**

***Index Terms*— Building segmentation, differ-modality, optical, synthetic aperture radar (SAR).**

I. Introduction

With the development of sensor technology and the use of accessible storage facilities, a significant amount of multimodality remote sensing data have been collected in recent years [1]–[3]. These multimodality datasets reflect the same fundamental phenomena from different perspectives [4]. Two of the most used remote sensing modalities are optical imaging and synthetic aperture radar (SAR). In recent years, more and more organizations adopted open data policies, both for SAR (e.g., Sentinel-1) and optical (e.g., Landsat series and Sentinel-2) satellites, significantly increasing the availability of multimodality data. Due to the highly complementary information that optical and SAR observed in the scenes, optical–SAR fusion or multimodality learning on optical and SAR has become an important research area in the field of remote sensing [5]–[7], such as agriculture mapping [8], cloud removal [9], land cover/land use mapping [10], building damage (BD) mapping [11], [12], among others. The subsequent research has contributed to the need to change the approach from conventional uni-/multimodality to a challenging cross-modality knowledge extraction scenario [13]–[17].

For the design of a cross-modality learning framework, the critical challenge is to learn the unified feature space from all modalities, which is both discriminative and classwise compact [18]. Since there are broad differences among the different modalities, learning the common features between the modalities is not easy. Hu *et al.* [19] proposed a classification framework with the combination of manifold alignment and semisupervised fusion of optical and SAR data. The potential of the deep learning methodology has been shown as far as cross-modality learning is concerned [20], [21]. A large volume of low-quality data [e.g., multispectral, multispectral imaging (MSI)] can be easily obtained in the real world. On the other hand, high-quality data (e.g., hyperspectral, HSI) are typically costly and difficult to get. Is it possible to investigate whether a small amount of high-quality data combined with a large number of low-quality data might contribute to meaningful tasks [22]? For this purpose, the authors proposed a novel semisupervised cross-modality learning framework [22]. Furthermore, a learning-shared cross-modality representation using multispectral-light detection and ranging (LiDAR) and hyperspectral is proposed [23]. A novel cross-modal deep-learning framework, called X-ModalNet, with three well-designed modules—the self-adversarial module, the interactive learning module, and the label propagation module—by learning to transfer more discriminative information from a small-scale HSI into the classification task using a large-scale MSI or SAR data was proposed [17]. A cross-modality image-matching network, namely CMM-Net, was proposed to realize thermal infrared and visible image matching by learning a modality-invariant feature representation [24]. In [25], the authors proposed a cross-modality knowledge distillation (CMKD) paradigm for the multimodality (e.g., SAR and optical) aerial view object classification. A simple yet effective cross-modality feature fusion approach, named cross-modality fusion transformer (CFT), was proposed for the object detection of multispectral image pairs [26]. The work addressed the problem of semantic segmentation with limited cross-modality

Manuscript received August 11, 2021; revised December 22, 2021 and January 18, 2022; accepted January 25, 2022. Date of publication February 1, 2022; date of current version March 23, 2022. This work was supported in part by KAKENHI under Grant19K20309, in part by the Japan Society for the Promotion of Science (JSPS) Bilateral Joint Research Project under Grant JPJSBP120203211, and in part by the Open Research Fund of National Earth Observation Data Center under Grant NODAOP2020021. *(Corresponding author: Junshi Xia.)*

Junshi Xia and Gerald Baier are with Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project (AIP), Tokyo 103-0027, Japan (e-mail: junshi.xia@riken.jp).

Naoto Yokoya is with the Department of Complexity of Science and Engineering, The University of Tokyo, Tokyo 277-8561, Japan, and also with Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project (AIP), Tokyo 103-0027, Japan (e-mail: yokoya@k.u-tokyo.ac.jp).

Digital Object Identifier 10.1109/TGRS.2022.3148383

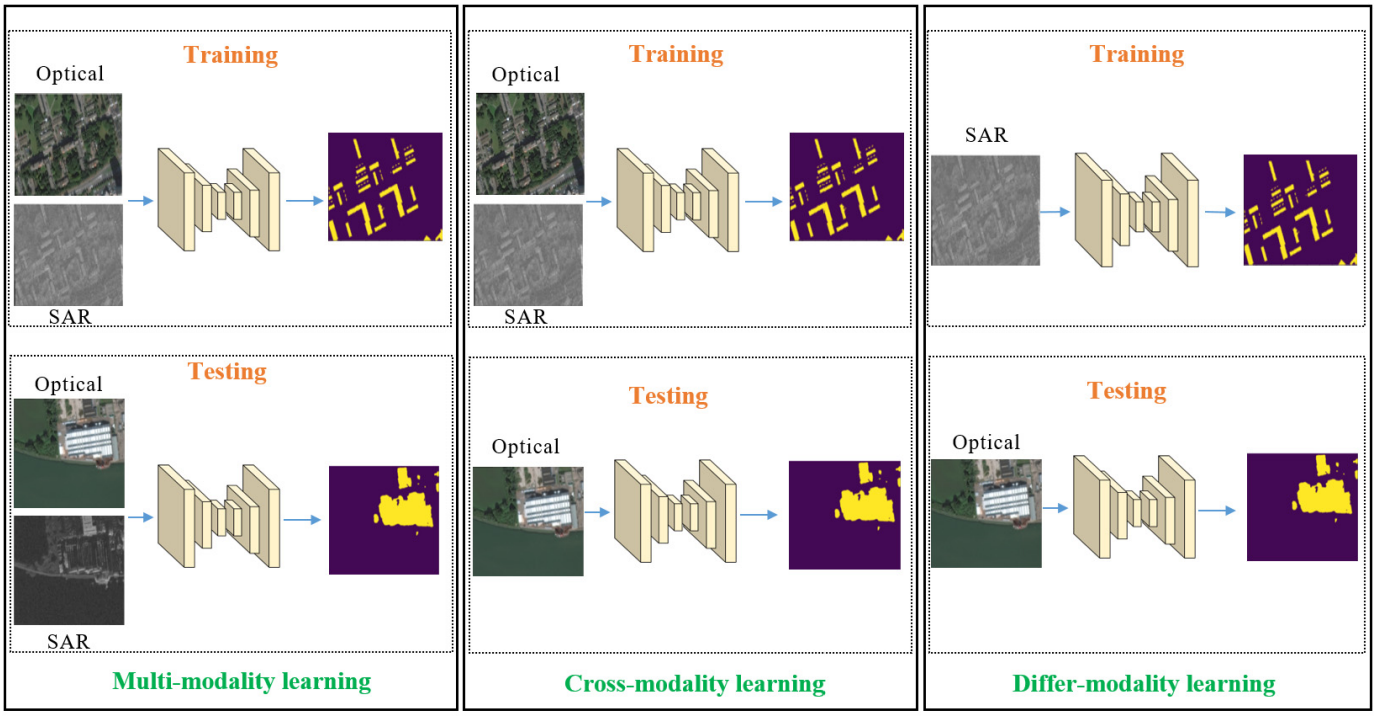


Fig. 1. Multimodality learning (left): the same multimodality datasets are used for both training and testing. Cross-modality learning (middle): the model is trained on multimodalities (e.g., optical and SAR), and only one modality is used in the process of testing. DML (right): the model is trained on one modality (e.g., SAR) with the training labels, and the other modality (e.g., optical) is used in the process of testing.

data in large-scale urban scenes and proposed the multimodality generative adversarial networks (GANs) [27].

Building semantic segmentation is of great importance for urban planning. Progress in the building segmentation techniques has been mainly motivated by the availability of publicly accessible datasets. Most of the publicly available datasets are based on buildings that can be seen through electro-optical imaging. For instance, the SpaceNet Challenges (1, 2, 4) [28] include more than 800 000 building footprints over ten cities, as well as other licensed datasets, including xView [29] and xBD [30]. Yet, these are missing SAR, which complements optical by providing backscattering to spectral information, and can thus help to better distinguish different objects. Public very high-resolution (VHR) (resolution <1 m) SAR datasets for building semantic segmentation are quite restricted. The only datasets can be found from Spacenet 6,[1] which includes airborne full-polarized X-band SAR and spaceborne optical datasets (with a resolution of 0.5 m) over the port of Rotterdam, the Netherlands. In fact, due to load, power supply, imaging mechanism, optical and SAR sensors cannot be mounted on the same platform, and some other factors, it is impossible to guarantee that multimodality datasets were acquired at the same time. We may have different modality datasets at various locations, with the new concept of differ-modality learning (DML).

Fig. 1 outlines the difference among multimodality learning, cross-modality learning, and DML. Ideally, the same multimodality datasets are used for both training and testing. The need for multimodality approaches in remote sensing data analysis is particularly evident in this context. In cross-modality learning, the training set contains the data from all modalities, while the testing set uses just one of them. In DML, training and testing datasets consist of different modalities. This inevitably leads to a question that is challenging but interesting: *Is it possible to transfer the knowledge from one modality (e.g., VHR SAR) to another modality (e.g., VHR optical) for the task of semantic segmentation?*

It is not surprising that the model trained on SAR data fails when tested on optical data. The straightforward way to recover the model performance is to retrain or fine-tune the models with additional annotations from the optical datasets. However, making annotations for each modality is both time-consuming and costly. Motivated by the aforementioned observations, we want to discuss the following particular challenges in the DML framework.

1) Multimodality datasets typically have different properties. Making the connection between the different modalities is very challenging but very relevant in real applications, such as building segmentation.
2) Given that learning with annotated samples alone is neither scalable nor generalizable from one modality to another, how to utilize the unlabeled datasets in DML?

According to the aforementioned factors, the ultimate aim of this work is to develop new approaches for differ-modality (e.g., SAR and optical) semantic segmentation. In particular, we consider the paired and unpaired cases to design different solutions. We assume that three separated datasets are available for the paired case: 1) paired SAR and optical without labels; 2) SAR datasets with labels; and 3) optical images without labels. Two independent datasets are available for the unpaired

[1]https://spacenetchallenge.github.io/

case: 1) unpaired SAR/optical datasets and only the SAR dataset with label information and 2) optical datasets without labels. We convert the DML to multimodality learning in the paired case based on the paired optical and SAR datasets. Knowledge distillation is then used in these two modalities to ensure that the stream of information is effectively concentrated in the segmentation task. Then, the self-training (SL) [31] is used to raise the robustness to label noise, thereby enhancing the generalization ability. For the unpaired case, we unified the framework by fully considering the image adaptation, the feature adaptation, and the segmentation. We also believe in self-training here to improve the performance as we did in the paired case.

The contributions can be summarized into fourfold.

1) To the best of the authors' knowledge, this is the first time to discuss DML with training and testing in different modalities in the remote sensing community, especially for building semantic segmentation.
2) We proposed two kinds of differ-modality frameworks for the paired and unpaired cases, making the solution more feasible in real applications. More specifically, we introduce the idea of image and feature adaption from GANs [32] to make the connection between two modalities to learn to generate the robust feature representations by learning the real and adversarial features in both the modalities. The knowledge distillation technique is included in the paired case framework, allowing sufficiently complementary information to form discriminative representations for segmentation.
3) The self-training technique is integrated into our proposed framework to jointly digest human-annotated labels and pseudolabels to improve performance.
4) Comprehensive evaluation on the SpaceNet 6 data shows significant improvements by our approach compared with the state-of-the-art methods.

II. Related Works

A. SAR to Optical Translation

GANs [32], which consist of a generative model and a discriminative model, are regarded as a revolution in image synthesis. The discriminative model is used to discriminate the real data from the generated data. Most approaches for SAR-to-optical synthesis are variants of the pix2pix [33], [34], the conditional GAN (cGAN) [35], and the cycle-consistent GAN (CycleGAN) [36]–[38]. The underlying generator and discriminator architectures are similar. To enhance the generator, Turnes *et al.* [34] introduced atrous convolution based on the pix2pix to translate SAR to optical.

However, these methods do not distinguish the features of SAR and optical datasets that significantly influence the translation process, which further consequences segmentation results.

B. Semantic Segmentation

Most of the building segmentation approaches are applied to VHR optical datasets, ranging from the variants of U-Net [39] to the recent High-Resolution Net (HRNet) [40].

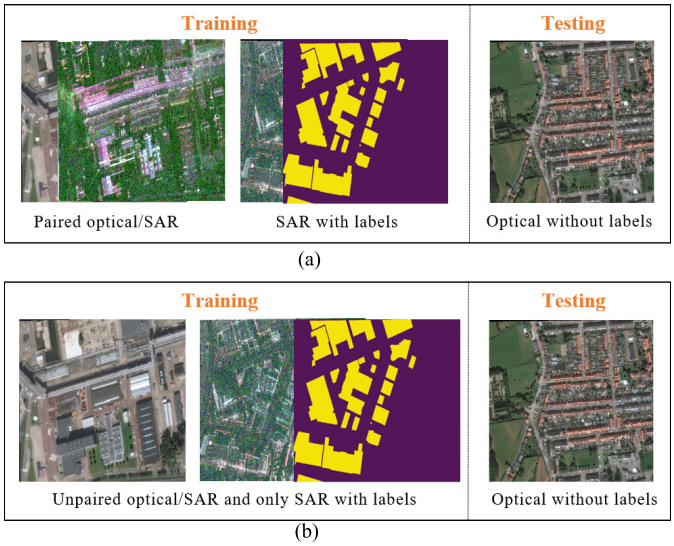


Fig. 2. Paired and unpaired cases in this work.

However, until recently, the particular problem of constructing semantic segmentation from SAR imagery using deep learning has gained relatively little exposure. In SpaceNet 6 challenge, four of the five winners used slight variants of the newly introduced EfficientNet (B5, B7, B8) as the encoder of U-Net.

Interest in unsupervised domain adaptation for the task of semantic segmentation by using adversarial learning [41], [42] has grown in the past few years. Representative methods, including domain-adversarial training of neural networks (DANN) [43], adversarial discriminative domain adaptation (ADDA) [44], and cycle-consistent adversarial domain adaptation (CyCADA) [45], utilize image and feature adaptation through adversarial learning by using a discriminator which differentiates between the feature space of different domains. Nevertheless, it remains challenging to incorporate image and feature adaptation into cross-modality learning/DML between optical and SAR.

III. Differ-Modality Building Segmentation

This section details the architecture and design of the components that make up the proposed differ-modality building segmentation approach. As shown in Fig. 2, two specific cases, paired and unpaired, are considered in this work. In the paired case, we have paired SAR/optical datasets at the training stage without labels but only a few different SAR images with labels. In the more challenging unpaired case, we have the unpaired SAR/optical datasets in the training stage, and only SAR datasets have building labels. The unlabeled optical images are used to predict the building labels in the testing stage. Two different frameworks are developed to address the paired/unpaired cases, detailed in Sections III-A and III-B. It should be emphasized that we used the definition of training SAR and testing optical in the description of the methods. In the next section, we give the experimental results of training optical and testing SAR.

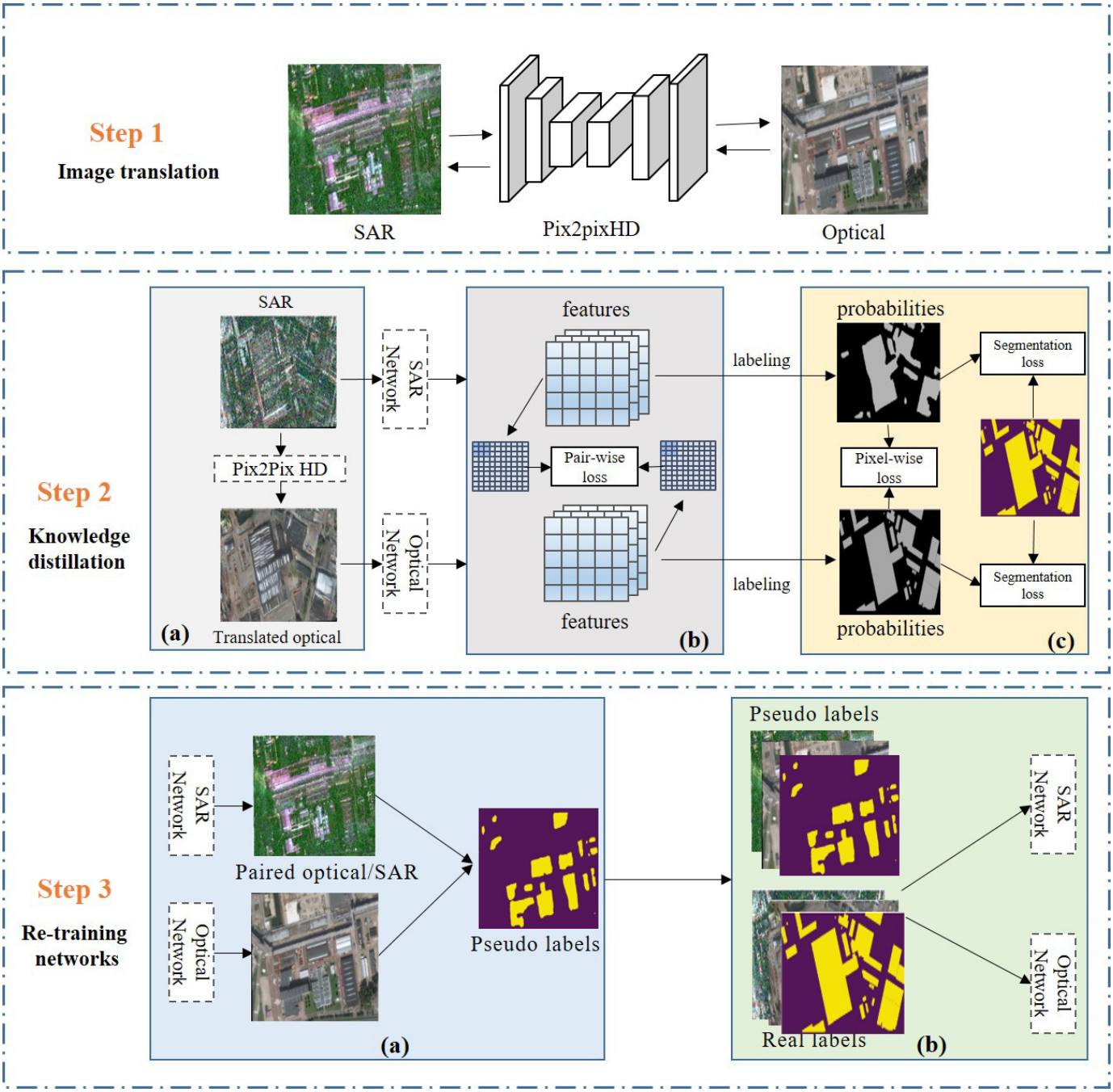


Fig. 3. Overall architecture of the training steps in the paired case. It mainly consists of three steps. Step 1 constructs the translation between the paired SAR/optical datasets. Step 2 applies the knowledge distillation between the SAR and translated optical datasets. Step 3 extracts the pseudolabels from the unlabeled SAR/optical pairs and then combines them with the real labels to retraining the segmentation network.

A. Paired Case

Let us assume a set of paired SAR/optical datasets $\{X^S, X^O\} = \{x_i^S, x_i^O\}_{i=1}^n$, as well as the labeled samples $\{X^{S'}, Y^{S'}\} = \{x_j^{S'}, y_j^{S'}\}_{j=1}^{n'}$ from SAR. The objective is to predict the labels from the unlabeled optical images $\{X^{O'}\} = \{x_k^{O'}\}_{k=1}^m$. The proposed method contains three steps—translation, knowledge distillation, and self-training—as shown in Fig. 3.

In Step 1, we utilize two pix2pixHD [46] models to translate from SAR to optical and vice versa. pix2pixHD consists of two generators: a global generator network (G_1) and a local enhancer network (G_2), and three multiscale discriminators (D_1, D_2, and D_3) with matching features. G_1 and G_2 have three parts: the downsampling layer, the residual block, and the upsampling layer. To improve the model's sensitivity to the details of each image patch, three 3×3 convolution kernels replaced one 7×7 convolution kernel, and the number of residual blocks in G_2 is increased to six. Taking the translation from SAR to optical as an example, the objective is defined as

$$\min_G \left(\left(\max_{D_1, D_2, D_3} \sum_{i=1,2,3} \mathcal{L}_{\text{GAN}}(G, D_i) \right) + \lambda \sum_{i=1,2,3} \mathcal{L}_{\text{FM}}(G, D_i) \right) \tag{1}$$

where $\mathcal{L}_{\mathrm{GAN}}$ is the regular GAN loss function and $\mathcal{L}_{\mathrm{FM}}$ is a feature matching loss function, given as

$$\mathcal{L}_{\mathrm{FM}}(G, D_i) = \mathbb{E}_{x^S, x^O} \sum_{j=1}^{N} \frac{1}{N_j} \left[\left\| D_i^j(x^S, x^O) - D_i^j(x^S, G(x^S)) \right\| \right] \tag{2}$$

where N is the total number of layers in the discriminator and N_j is the number of elements in each layer. D_i^j denotes the ith-layer feature extractor of discriminator D_i.

In Step 2, since we only have the SAR images with training labels, we first translated SAR ($X^{S'}$) to optical-like images ($X^{S' \to O'}$) via the previous pix2pixHD model (seen in Step 2. a of Fig. 3). Consequently, we trained SAR and optical segmentation networks separately by using the aforementioned $X^{S'}$ and $X^{O'}$ datasets. Then, the knowledge distillation strategy [47] is used to enhance the capability of both SAR (T^S) and optical network (T^O). In this case, the pairwise distillation in the features (Step 2. b) and pixelwise distillation in the probabilities (Step 2. c) are introduced [47]. As an example, for using the SAR network (T^S) to improve the optical network (T^O), the overall loss function is defined as

$$\mathcal{L}(T^O) = \mathcal{L}_{\mathrm{seg}}(T^O) + \gamma \left(\mathcal{L}_{\mathrm{pa}}(T^O) + \mathcal{L}_{pi}(T^O) \right) \tag{3}$$

where γ is the tradeoff parameter. $\mathcal{L}_{\mathrm{seg}}$, $\mathcal{L}_{\mathrm{pa}}$, and $\mathcal{L}_{pi}$ are the segmentations.

Inspired by the pairwise Markov random field framework that is widely adopted for improving spatial labeling contiguity, the pairwise relations, especially pairwise similarities among pixels, are used in pairwise distillation [47]. Pairwise loss function is given as

$$\mathcal{L}_{\mathrm{pa}}(T^O) = \frac{1}{(W \times H)^2} \sum_{i \in all} \sum_{j \in all} (a_{i,j}^{T^O} - a_{i,j}^{T^S})^2 \tag{4}$$

where $W \times H$ is the size of the training images. $a_{i,j}^{T^O}$ denote the similarity between ith and jth pixels from T^O, and $a_{i,j}^{T^S}$ denote the similarity between ith and jth pixels from T^S. $a_{i,j}$ is simply calculated from the features $\mathbf{f}_i$ and $\mathbf{f}_j$ in [47], which is given as

$$a_{i,j} = \mathbf{f}_i^\top \mathbf{f}_j / (||\mathbf{f}_i||_2 ||\mathbf{f}_j||_2). \tag{5}$$

In pixelwise distillation, the knowledge distillation can be directly used to align the class probability of each pixel. The pixelwise loss function is given as

$$\mathcal{L}_{pi}(T^O) = \frac{1}{W \times H} \sum_{i \in all} \mathrm{KL}(p_i^{T^O} || p_i^{T^S}) \tag{6}$$

where $p_i^{T^O}$ and $p_i^{T^S}$ are the probabilities of the ith pixel from T^O and T^S. KL is the Kullback–Leibler divergence.

In Step 3, we used the SAR and optical networks trained from Step 2 to predict the unlabeled SAR/optical pairs in Step 1 to generate the pseudolabels by a majority vote (Step 3. a). Then, the combinations of pseudolabels and the real labels in Step 2 are used to fine-tune the SAR and optical segmentation networks. The performance of the networks is often enhanced if the noisy pseudolabels can be properly treated.

In the testing phase, the unlabeled optical images are translated into the SAR domains via the pix2pixHD in Step 1 of training. The optical and translated SAR images are then used to produce the labels from the optical and SAR segmentation networks. The final labels are obtained from the two previous labels by a majority vote.

B. Unpaired Case

Let us assume a set of unpaired SAR and optical images and only SAR images with the labels, i.e., $\{X^S, Y^S\} = \{x_i^S, y_i^S\}_{i=1}^n$ from the SAR domain and $X^O = \{x_j^O\}_{j=1}^{n'}$ from the optical domain, as well as unlabeled samples $X^{O'} = \{x_k^{O'}\}_{k=1}^m$ from the optical domain $X^{O'}$. Fig. 4 presents the overall architecture of the training steps in the unpaired case.

In Step 1, we first applied the GAN with a generator $G_O(x^S) = x^{S \to O}$, which translated the SAR x^S to the realistic optical like $x^{S \to O}$, and a discriminator D_O, which is used to differentiate $x^{S \to O}$ and x^O. Then, the loss function of G_O and D_O is defined as

$$\begin{aligned} \mathcal{L}_{\mathrm{adv}}^O(G_O, D_O) = {} & \mathbb{E}_{x^O \sim X^O} \left[\log D_O(x^O) \right] \\ & + \mathbb{E}_{x^S \sim X^S} \left[\log(1 - D_O(G_O(x^S))) \right]. \end{aligned} \tag{7}$$

The reverse generator E_U is used to reconstruct $x^{S \to O}$ back to x^S with a discriminator D_S, creating a cycle to maintain the original content in the translated images. The cycle-consistency loss $\mathcal{L}_{\mathrm{cyc}}$ is defined as

$$\begin{aligned} \mathcal{L}_{\mathrm{cyc}}(G_O, E\_U) = {} & \mathbb{E}_{x^S \sim X^S} \left[\left\| E\_U(G_O(x^S)) - x^S \right\|_1 \right] \\ & + \mathbb{E}_{x^O \sim X^O} \left[\left\| G_O(E\_U(x^O)) - x^O \right\|_1 \right]. \end{aligned} \tag{8}$$

The loss function of E_U and D_S is defined as

$$\begin{aligned} \mathcal{L}_{\mathrm{adv}}^S(E\_U, D_S) = {} & \mathbb{E}_{x^S \sim X^S} \left[\log D_S(x^S) \right] \\ & + \mathbb{E}_{x^O \sim X^O} \left[\log(1 - D_S(E\_U(x^O))) \right]. \end{aligned} \tag{9}$$

The synthesized optical domain images are paired with the corresponding labels from the SAR domain $\{X^{S \to O}, Y^S\}$ and can be fed into the segmentation task E_C. Then, the segmentation loss $\mathcal{L}_{\mathrm{seg}}$ is given as

$$\mathcal{L}_{\mathrm{seg}}(E_C) = F\left(Y^S, \hat{Y}^{S \to O}\right) + \beta Dice\left(Y^S, \hat{Y}^{S \to O}\right) \tag{10}$$

where $\hat{Y}$ is the predict labels. β is the tradeoff parameter between the first focal and the second Dice loss functions.

The adversarial losses of segmentation and feature space are considered to alleviate the modality shift between optical and SAR. More specifically, in the segmentation space, the discriminator D_p is used to differentiate the segmentation results via the following adversarial loss:

$$\begin{aligned} \mathcal{L}_{\mathrm{adv}}^p(E\_C, D_p) = {} & \mathbb{E}_{x^{S \to O} \sim X^{S \to O}} \left[\log D_p(E\_C(x^{S \to O})) \right] \\ & + \mathbb{E}_{x^O \sim X^O} \left[\log(1 - D_p(E\_C(x^O))) \right]. \end{aligned} \tag{11}$$

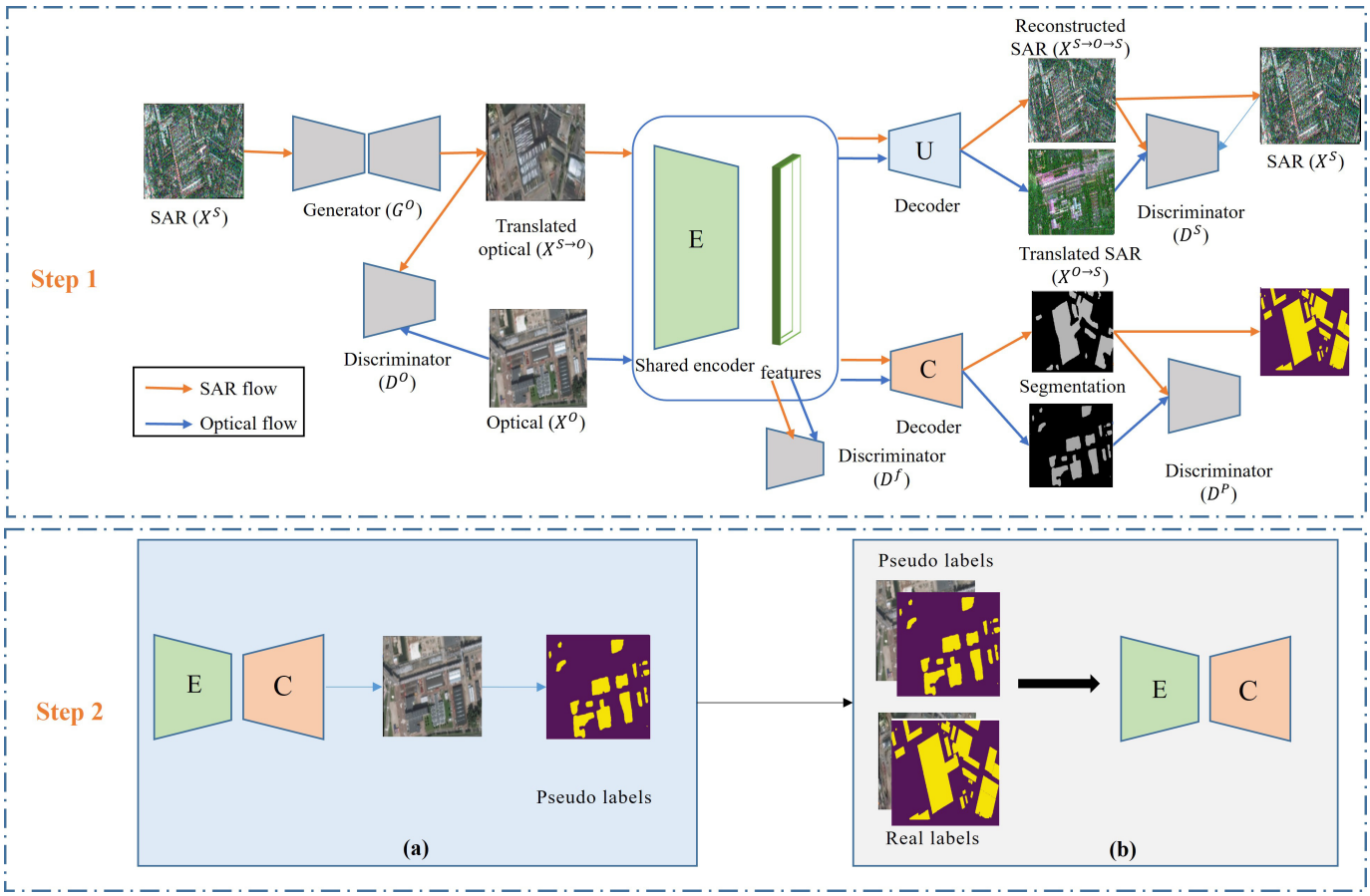


Fig. 4. Overall architecture of the training steps in the unpaired case. It mainly contains two steps. Step 1 adopts the GAN for the image and feature adaptation and integrates the segmentation task into one framework. Here, E represents the shared encoder. U and C represent the decoder of translation and segmentation, respectively. E_U is used for translating optical to SAR. E_C is used to generate the segmentation result. Step 2 extracts the pseudolabels from the unlabeled optical images and then combines with the real labels with translated optical images to retraining the segmentation network.

In the feature space, the following adversarial loss is used in the training process:

$$\mathcal{L}_{\text{adv}}^{f}(E, D_f) = \mathbb{E}_{x^{S\to O}\sim X^{S\to O}}\left[\log D_f(E(x^{S\to O}))\right] + \mathbb{E}_{x^O\sim X^O}\left[\log(1 - D_f(E(x^O)))\right]. \quad (12)$$

In addition, the auxiliary task uses D_S to differentiate whether the generated images are translated from optical domain X^O or reconstructed from $X^{S\to O}$. The subsequent adversarial loss is defined as

$$\mathcal{L}_{\text{adv}}^{\hat{S}}(E\_U, D_S) = \mathbb{E}_{x^{S\to O}\sim X^{S\to O}}\left[\log D_s(E\_U(x^{S\to O}))\right] + \mathbb{E}_{x^O\sim X^O}\left[\log(1 - D_S(E\_U(x^O)))\right]. \quad (13)$$

The overall objective for the framework is defined as

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\text{adv}}^{O}\mathcal{L}_{\text{adv}}^{O}(G_O, D_O) + \lambda_{\text{adv}}^{S}\mathcal{L}_{\text{adv}}^{S}(E\_U, D_S) \\ & + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}}(G_O, E\_U) + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}(E\_C) \\ & + \lambda_{\text{adv}}^{\hat{S}}\mathcal{L}_{\text{adv}}^{\hat{S}}(E\_U, D_S) + \lambda_{\text{adv}}^{p}\mathcal{L}_{\text{adv}}^{p}(E\_C, D_p) \\ & + \lambda_{\text{adv}}^{f}\mathcal{L}_{\text{adv}}^{f}(E, D_f) \end{aligned} \quad (14)$$

where λ_{adv}^{O}, λ_{adv}^{S}, λ_{cyc}, λ_{seg}, $\lambda_{\text{adv}}^{\hat{S}}$, λ_{adv}^{p}, and λ_{adv}^{f} are tradeoff parameters to indicate the importance of each component.

As shown in Step 1, the key issue is that the image/feature adaptation and segmentation shared the same encoder (E). The encoder E is optimized with $\mathcal{L}_{\text{adv}}^{S}$ and $\mathcal{L}_{\text{cyc}}$ and $\mathcal{L}_{\text{seg}}$ via the image/feature adaptation and segmentation perspectives, and also collected the gradients back-propagated from three different discriminators D^p, D^S, and D^f. In this case, the shared encoder extended the framework into multiple-task learning, which emphasizes the pixelwise reconstruction and focuses on structural semantics simultaneously.

In Step 2, we used the segmentation network E_C to generate the pseudolabels from the optical images without labels in Step 1. The combinations of pseudolabels and the real labels are used to fine-tune the segmentation network E_C.

In the testing phase, the unlabeled optical images are directly fed into E_C to produce the final results.

IV. Experiments

A. Dataset Description

The proposed differ-modality framework is validated on the SpaceNet 6 challenges for building semantic segmentation in Rotterdam using VHR optical and SAR images. This dataset features a unique combination of aerial full-polarization SAR imagery (HH, HV, VH, and VV) in the X-band wavelength

from Capella Space and optical imagery (three bands) from Maxar's WorldView 2 satellite [48]. The aerial SAR collection captures the same area of Rotterdam over three days: August 4, 23, and 24, 2019. Data are captured from an off-nadir perspective at a relative look angle of 53.4°–56.6° from both north- and south-facing directions. After the processing steps, including single look complex (SLC) processing, geo-registered, and ortho-rectified, the SAR datasets are resampled with a Lanczos interpolation to a spatial resolution of $0.5m \times 0.5m$ per pixel. For the optical image, a single, cloud-free image strip was collected on August 31, 2019, from a look angle of 18.4° off-nadir. Here, the pansharpend RGB with the resolution of 0.5 m is used. It should be noticed that the optical datasets are atmospherically compensated to surface-reflectance values by Maxar's AComp and ortho-rectified. The original size of the sample is 900×900 pixels.

1) SAR to Optical: For the paired case, we keep the original size and define the numbers of images in the experiments as follows.

1) SAR and optical without labels: 1320
2) SAR with labels used for the training: 1020
3) Optical without labels used for the testing: 1021

For the unpaired case, we mainly focus on urban areas. In this case, we split the images into 256×256 and then discard the images without any buildings. Finally, we randomly selected 900 SAR and 880 optical images, and only SAR with the labels. Seven-hundred and seventy-five optical images are used for the testing.

2) Optical to SAR: In the paired case, we keep the same number as the previous experimental settings of SAR to optical:

1) SAR and optical without labels: 1320
2) Optical with labels used for the training: 1020
3) SAR without labels used for the testing: 1021

Nine-hundred SAR and 880 optical images are randomly selected for the unpaired case, and only optical datasets with the labels. Eight-hundred and fifteen optical images are used for the testing.

B. Implementations and Compared Methods

Our approaches are implemented in the Pytorch framework. All the models are trained on four NVIDIA V100 GPUs with 16 GB memory.

In the paired case, for the pix2pixHD [46], we replaced one 7×7 convolution kernel by three 3×3 kernels and increased the number of residual blocks in the local enhancer network to six. The parameters in pix2pixHD are set to be default in [46]. We adopted the U-Net with ResNeXt50_32 $\times$ 4d and EfficientNet-b5 as the optical and SAR for the knowledge distillation, respectively. The segmentation networks are trained by Adam, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a learning rate of 0.0001, and a batch size of 4 for 100 epochs. Data augmentation, including random crop (size of 512×512) and random flipping, is used for the training. γ is set to be 10 according to [47]. The combination of focal and Dice loss functions is unitized. Specific experimental parameters are shown in Table I.

TABLE I
PARAMETERS INVOLVED IN THE PROCESS OF PAIRED CASE

Task	Image translation	Knowledge distillation
Model	pix2pixHD	U-Net (ResNeXt50_32×4d) U-Net (EfficientNet-b5)
Batch size	1	4
Learning rate	0, 0002	0, 0001
Epoch	100	100
Optimizer	Adam (0.5, 0.999)	Adam (0.9, 0.999)
Data augmentation	None	Random crop Random flipping
Weight	Default in [46]	$\gamma = 10$

TABLE II
PARAMETERS INVOLVED IN THE PROCESS OF UNPAIRED CASE

Model	CycleGAN	DRN
Batch size	4	
Learning rate	0.0001	
Epoch	100	
Optimizer	Adam (0.5, 0.999)	Adam (0.9, 0.999)
Data augmentation	Random flipping	
Weight	Default in [36]	$\lambda^O_{adv}, \lambda^S_{adv} = 1$ $\lambda^{\hat{S}}_{adv}, \lambda^p_{adv}, \lambda^f_{adv} = 2$ $\lambda_{cyc}, \lambda_{seg} = 5$

TABLE III
COMPUTATIONAL COMPLEXITY FOR THE PAIRED CASE AND UNPAIRED CASE

	Step 1	Step 2	Step 3
Paired case	112.6 G	126.8G	43.6G
Unparied case	63.4G	11.2G	-

In the unpaired case, followed by the CycleGAN [36], we used U-Net as the generator G^O. For all the discriminators, we use 70×70 PatchGANs to differentiate the real and fake images [33]. The shared encoder used the dilated residual networks (DRNs) [49] to preserve the original spatial resolutions. We constructed one convolutional layer for the decoder U with three residual blocks, two deconvolutional layers, and one convolutional output layer. The classifier consists of an upsampling layer and 1×1 convolutional layer. The Adam optimizer with a learning rate of 0.0001 is used for the adversarial learning losses. The Adam optimizer is parameterized with an initial learning rate of 0.001 and a step decay rate of 0.9 every ten epochs for the segmentation task. A batch size of 4 for 100 epochs, and random flipping augmentation, are used for the training. λ^O_{adv} and λ^S_{adv} are set to be 1. $\lambda^{\hat{S}}_{\text{adv}}$, λ^p_{adv}, and λ^f_{adv} are set to be 2. λ_{cyc} and λ_{seg} are set to be 5. Specific experimental parameters are shown in Table II.

Table III has shown the number of multiply–accumulate (MAC) operations. For the paired and unpaired cases, the image sizes of 512×512 and 256×256 are processed.

C. Results of Training on SAR and Testing on Optical

For the paired case, the following methods are used for the comparisons.

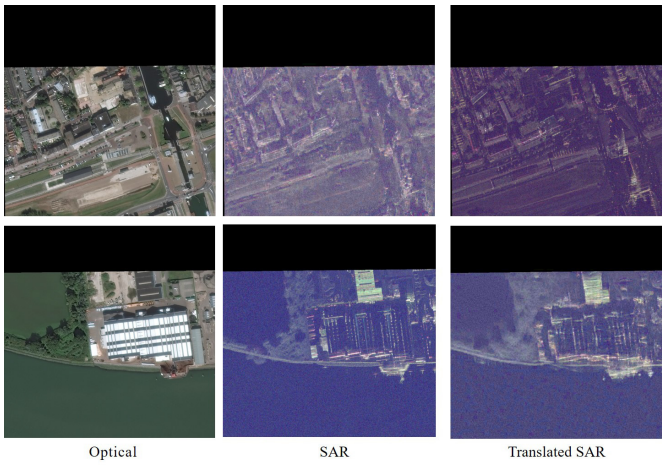


Fig. 5. Examples of optical, SAR (R: HH, G: VV, and B: HV), and translated SAR via pix2pixHD.

1) Baseline 1: Training in real SAR and testing in the translated SAR. We refer to it as the SAR network in Step 2 without knowledge distillation and self-training.
2) Baseline 2: Training in translated optical and testing in the real optical. We refer to it as the optical network in Step 2 without knowledge distillation and self-training.

For the unpaired case, the following methods are used for the comparisons.

1) Baseline 1: Training in real SAR and testing in the translated SAR.
2) Baseline 2: Training in translated optical and testing in the real optical.
3) CyCADA adapts representations at both the pixel level and the feature level for the segmentation task.
4) CrDoCo [50] consists of two main modules: 1) an image-to-image translation network and 2) two domain-specific task networks for the segmentation task.

Fig. 5 shows the examples of optical, SAR, and translated SAR via pix2pixHD from optical. Following the evaluation in [51], we selected structural similarity (SSIM), mean square errors (MSEs), and peak signal-to-noise ratio (PSNR) to evaluate the effectiveness of pix2pixHD. The values of SSIM, MSE, and PSNR are 0.6216, 0.0221, and 18.67, respectively. It should be noticed that any advanced translation techniques, such as feature-guiding GAN (FGGAN) [51], can be used. The main structures of the buildings are well preserved from the optical and SAR domains. However, the boundaries are not well maintained due to the following reasons: 1) difference between the view angle of optical and the incidence angle of SAR and 2) training errors from pix2pixHD.

Table IV lists the quantitative performance of the proposed DML framework and the two baselines. Knowledge distillation and self-training parts are also added to the two baselines. The performance of directly training on three-band SAR and testing on optical is abysmal (intersection over union (IoU) < 0.05). Baselines 1 and 2 can obtain the accuracies of 34.47% and 44.45%. The optical modality is stronger than the SAR modality in building segmentation. With the two distillation modules, the improvements for baseline 1 and baseline 2 are 2.8 and 2.7 percentage points, respectively. The IoUs are further promoted by self-training and reach 40.13% and 49.38% for baseline 1 and baseline 2, respectively. The proposed DML framework makes complementary contributions from SAR and optical segmentation networks. Especially in the production of pseudolabels, DML generates reliable labels compared with only SAR or optical networks. In summary, the proposed method outperforms other methods, demonstrating its superiority with a final score of 52.37%. The qualitative segmentation results in Fig. 6 validate the effectiveness of our approach for preserving the boundaries of the buildings when compared with other methods.

TABLE IV

EFFECTS OF DIFFERENT PARTS ON THE ACCURACY ASSESSMENT IN THE PAIRED CASE (TRAINING ON SAR AND TESTING ON OPTICAL). PI AND PA: PIXELWISE AND PAIRWISE DISTILLATION IN STEP 2. SL: SELF-TRAINING IN STEP 3

Method	IoU(%)
Baseline 1	34.47
+PI	36.78
+PI+PA	37.25
+PI+PA+SL	40.13
Baseline 2	44.45
+PI	45.17
+PI+PA	47.13
+PI+PA+SL	49.38
DML without SL	50.98
DML (paired case)	52.37

TABLE V

PERFORMANCE OF THE PROPOSED AND COMPARED METHODS IN UNPAIRED CASE (TRAINING ON SAR AND TESTING ON OPTICAL)

Methods	IoU(%)
Baseline 1	49.52
Baseline 2	49.51
CyCADA	51.47
CrDoCo	51.98
DML without SL	53.38
DML (unpaired case)	56.78

Table V reports the quantitative results, which show that the segmentation accuracy of our approach is significantly increased when compared with other methods. Like the paired case, the results of directly training on three-band SAR and optical testing are terrible (IoU < 0.03), indicating the severe domain shift between SAR and optical modalities. Baselines 1 and 2, which only considered the image adaptation, improved the results with the IoUs of 49.52% and 49.51%, respectively. Significantly, the IoU of our proposed methods was restored to 56.78%. It should be noted that, compared with the CyCADA and CrDoCo, which also consider both the image and feature adaption, our method without SL achieved a better performance. With the support of SL, our final score reached 56.78%. The results demonstrate the

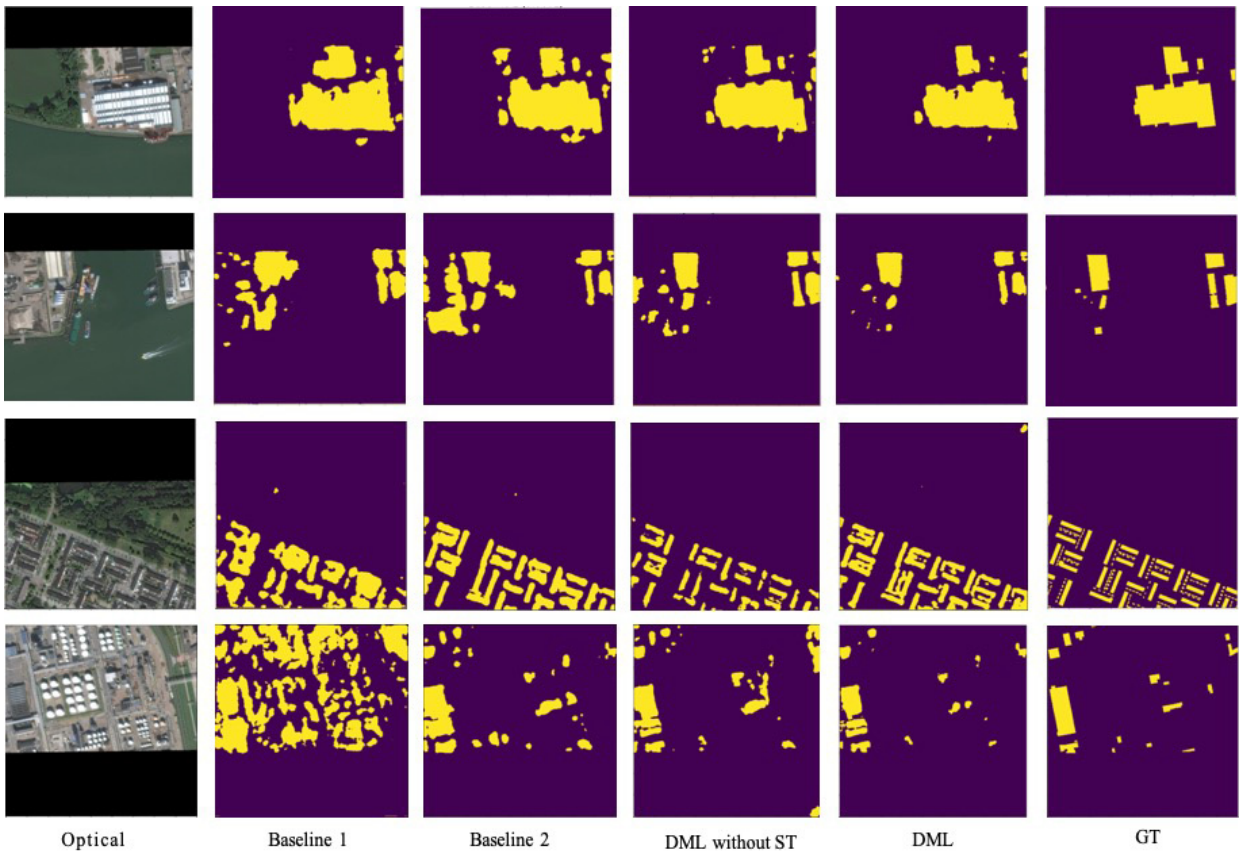


Fig. 6. Qualitative results on the testing set of the paired case (training on SAR and testing on optical). The yellows parts indicate the buildings.

effectiveness of our framework, which integrates the benefits of the complementary image, feature alignment between two modalities, and self-training.

The comparison results are visually provided in Fig. 7. It can be easily seen that with only considering image adaption in baselines 1 and 2, the shapes of buildings are very blurred and noisy, and the structures of buildings cannot be well distinguished. CrDoCo and our DML results without self-training, which involved both the feature and image adaptations, can produce semantically accurate predictions for buildings' main structures. However, some buildings are still missing in the prediction process. As shown in the penultimate column, DML considers self-training for fine-tuning the segmentation network, which can recover the missing parts.

D. Results of Training on Optical and Testing on SAR

For the paired case, the following methods are used for the comparisons.

1) Baseline 1: Training in real optical and testing in the translated optical.
2) Baseline 2: Training in translated SAR and testing in the real SAR.

For the unpaired case, the following methods are used for the comparisons.

1) Baseline 1: Training in real optical and testing in the translated optical.
2) Baseline 2: Training in translated SAR and testing in the real SAR.
3) CyCADA.
4) CrDoCo [50].

Fig. 8 shows the examples of SAR, optical, and translated optical via pix2pixHD from SAR. The values of SSIM, MSEs, and PSNR are 0.6324, 0.0207, and 19.13, respectively. From SAR to optical domains, the main structures of the buildings have been adequately retained. The boundaries of small buildings are not well maintained due to the following reasons: 1) disparity in optical and SAR incidence angles, especially for specific sides depending on the geometry and incidence angles and 2) training errors from the translation models.

Table VI presents the results of the proposed DML as well as the baseline for the paired case. To make a fair comparison, we include knowledge distillation (e.g., PI and PA) and self-training parts (e.g., SL) in the two baselines. From the table, it can be seen that the accuracies of baselines 1 and 2 are 42.31% and 32.32%, respectively. With the supports of knowledge distillation and self-learning, the accuracies of baselines 1 and 2 can be further achieved at 46.68% (37.28%) and 50.11% (39.13%). SAR and optical segmentation networks provide complementary contributions to the proposed DML framework. In addition, DML produces trustworthy pseudolabels when compared with only SAR or optical network. In a conclusion, the proposed approach beats the existing methods, with a total score of 52.64% indicating

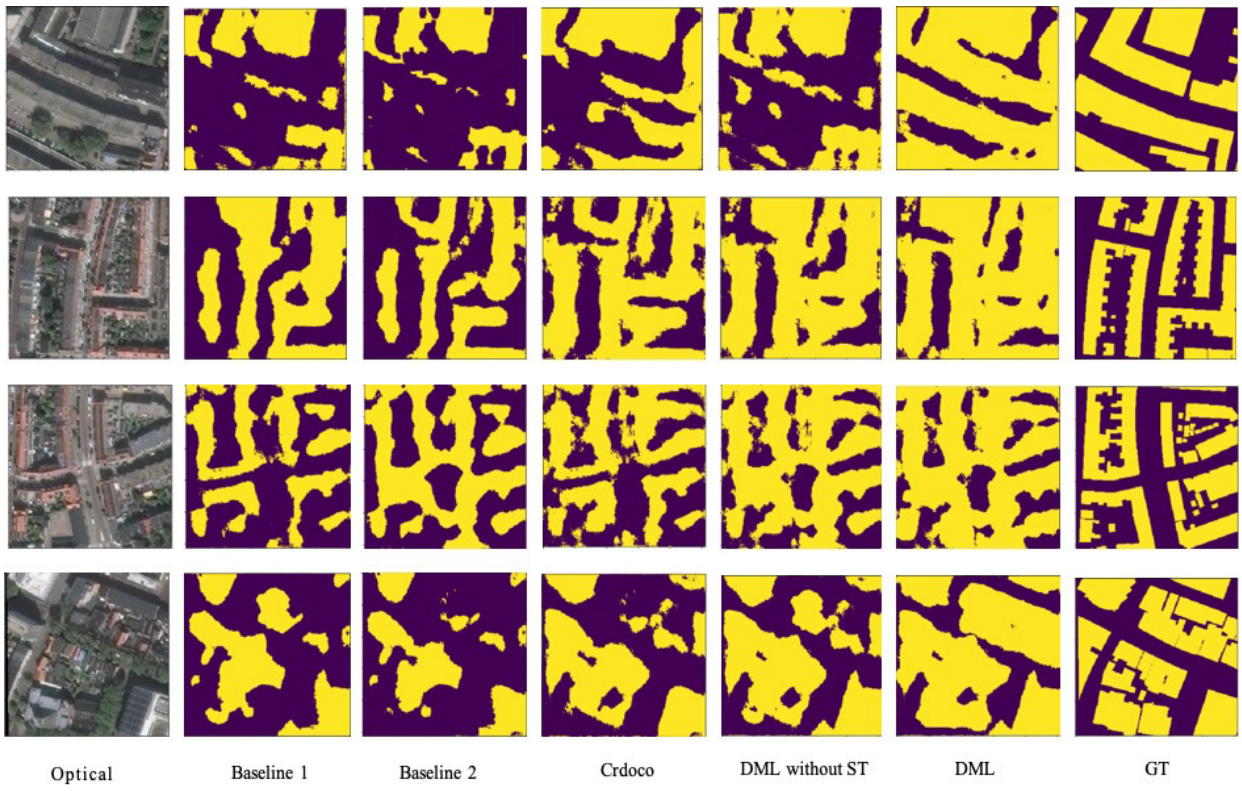


Fig. 7. Qualitative results on the testing set of the unpaired case (training on SAR and testing on optical). The yellows parts indicate the buildings.

TABLE VI
EFFECTS OF DIFFERENT PARTS ON THE ACCURACY ASSESSMENT IN THE PAIRED CASE (TRAINING ON OPTICAL AND TESTING ON SAR). PI AND PA: PIXELWISE AND PAIRWISE DISTILLATION IN STEP 2. SL: SELF-TRAINING IN STEP 3

Method	IoU(%)
Baseline 1	42.31
+PI	43.52
+PI+PA	46.86
+PI+PA+SL	50.11
Baseline 2	32.31
+PI	34.65
+PI+PA	37.28
+PI+PA+SL	39.13
DML without SL	51.01
DML (paired case)	52.64

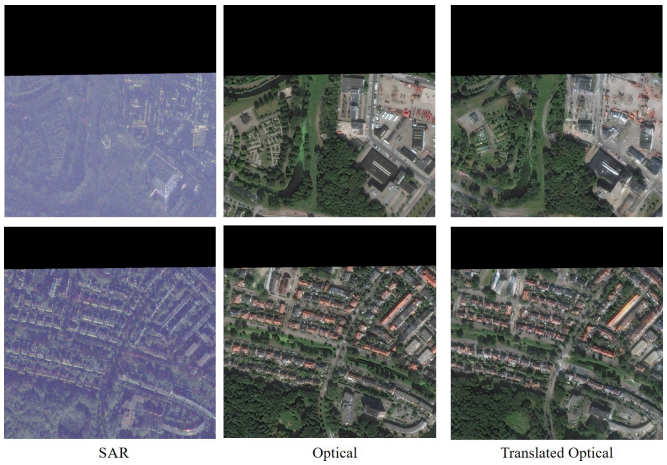


Fig. 8. Examples of SAR (R: HH, G: VV, and B: HV), optical, and translated optical via pix2pixHD.

its superiority. When compared with other methods, the qualitative segmentation results in Fig. 9 validate the effectiveness of our methodology for preserving building boundaries and removing the false segmentation parts.

Table VII shows the quantitative results of the proposed methods in the case of training on optical and testing on SAR. The comparison methods are also listed in the table. This table shows that our approach significantly increased the segmentation performance over the other methods by a large margin of IoUs. Direct training on optical and SAR testing yields very poor results (IoU < 0.02), showing a significant domain shift between SAR and optical modalities. Baselines 1 and 2, which considered image adaptation, improved the performance with IoUs of 48.69% and 47.82%, respectively. It should be highlighted that, when compared with CyCADA and CrDoCo, which consider both the image and feature adaptations, our method without SL still outperformed both the compared methods. Our total score was 55.62%, thanks to SL's assistance. This highlights the feasibility of our approach, which incorporates the advantages of a complimentary image, feature alignment between two modalities, and self-training to produce positive results.

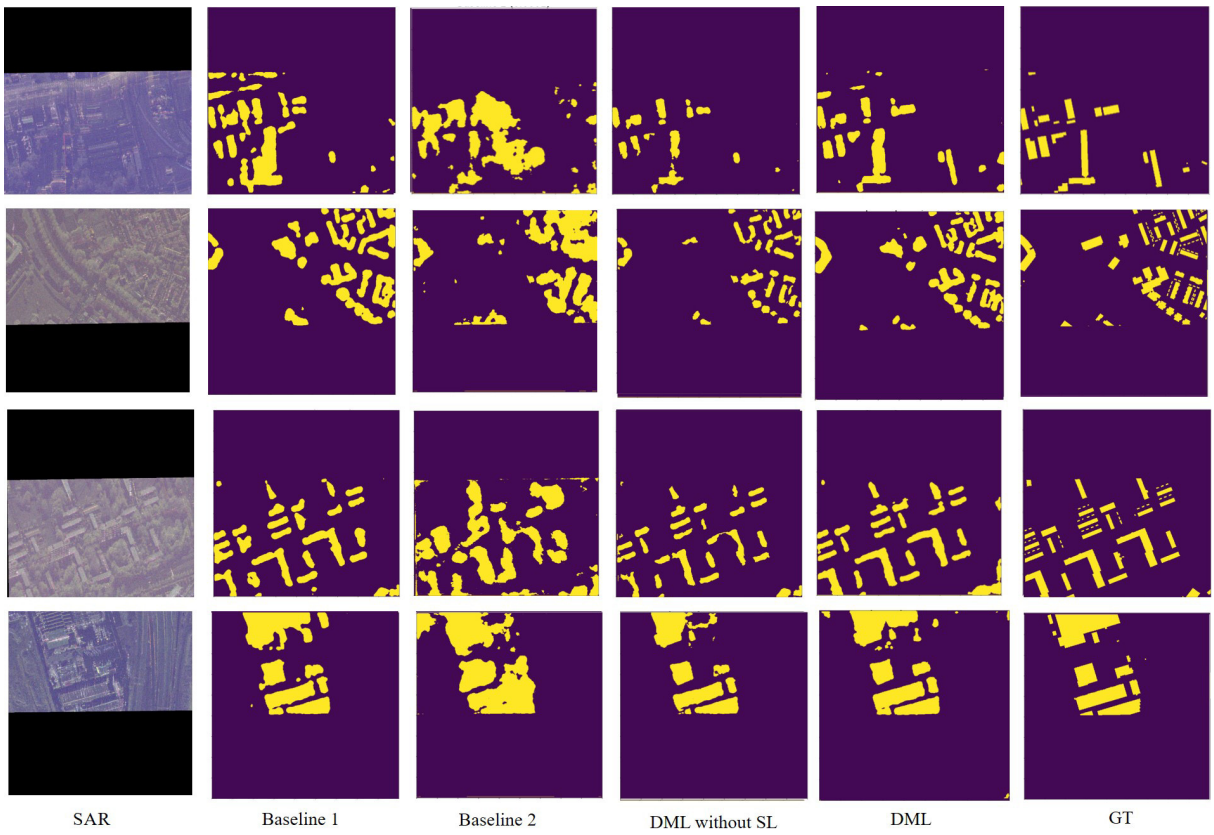


Fig. 9. Qualitative results on the testing set of the paired case (training on optical and testing on SAR).

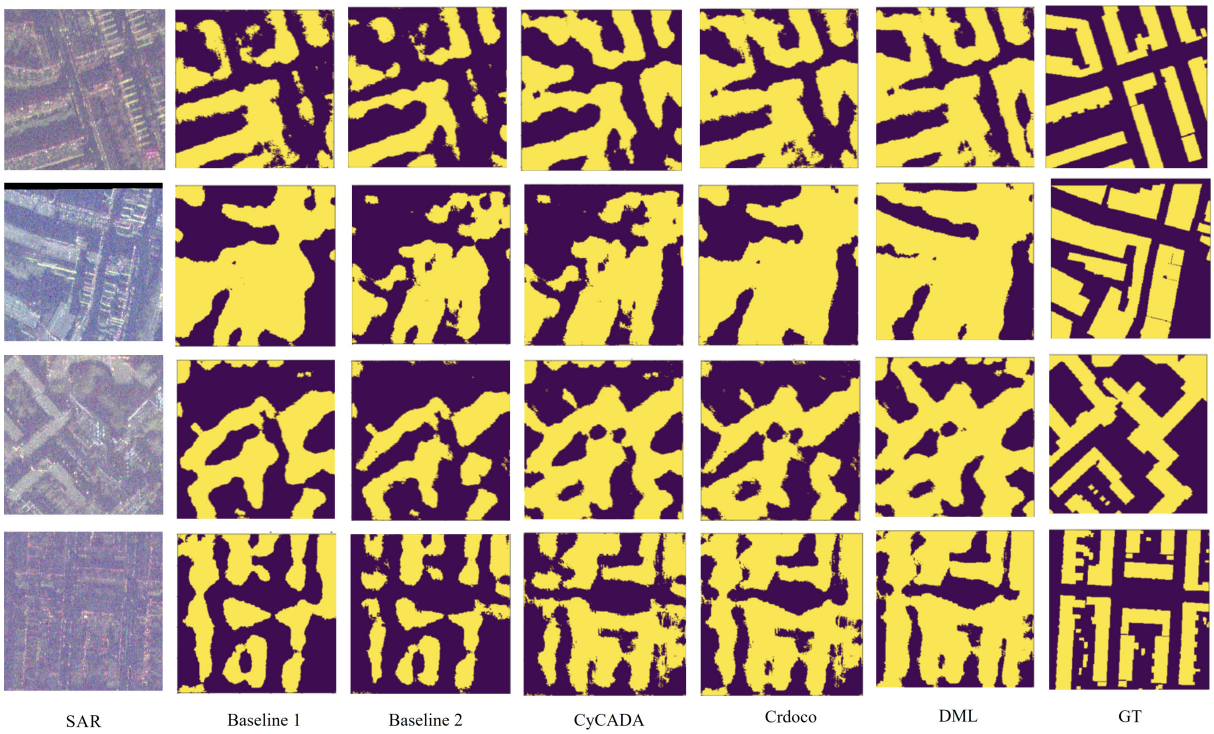


Fig. 10. Qualitative results on the testing set of the unpaired case (training on optical and testing on SAR).

The comparative results are illustrated in Fig. 10. It is readily apparent that when simply image adaptation is taken into account in baselines 1 and 2, the shapes of buildings are very blurred and noisy, and the structures of buildings are difficult to discern between one another. It has been shown that the results of CrDoCo and our DML without

TABLE VII

PERFORMANCE OF THE PROPOSED AND COMPARED METHODS IN UNPAIRED CASE (TRAINING ON OPTICAL AND TESTING ON SAR)

Methods	IoU(%)
Baseline 1	48.69
Baseline 2	47.82
CyCADA	50.12
CrDoCo	52.80
DML without SL	52.98
DML (unpaired case)	55.62

TABLE VIII

INFLUENCE OF EACH COMPONENT IN THE UNPAIRED CASE OF THE PROPOSED DML FRAMEWORK (TRAINING ON SAR AND TESTING ON OPTICAL)

$\mathcal{L}_{adv}^{O}$, $\mathcal{L}_{adv}^{S}$, $\mathcal{L}_{cyc}$, $\mathcal{L}_{seg}$	$\mathcal{L}_{adv}^{\hat{S}}$	$\mathcal{L}_{adv}^{p}$	$\mathcal{L}_{adv}^{f}$	ST	IoU(%)
✓					50.07
✓	✓				50.63
✓	✓	✓			52.75
✓	✓	✓	✓		53.38
✓	✓	✓	✓	✓	56.78

self-training, which involved both feature and image modifications, may yield semantically valid predictions for the key structural elements of buildings. However, there are still some buildings that have not been included in the prediction process. In the penultimate column, it is demonstrated that DML, which is regarded as self-training for fine-tuning the segmentation network, can recover the missing parts.

E. Ablation Studies

We analyzed the performance gain in the unpaired cases by adding the different components step by step. The results in the case of training on SAR and testing on optical are shown in Table VIII. In Step 1, the main part should contain the following loss functions: $\mathcal{L}_{\text{adv}}^{O}$, $\mathcal{L}_{\text{adv}}^{S}$, $\mathcal{L}_{\text{cyc}}$, and $\mathcal{L}_{\text{seg}}$. By removing the adversarial loss of feature adaptation (e.g., $\mathcal{L}_{\text{adv}}^{\hat{S}}$, $\mathcal{L}_{\text{adv}}^{p}$, and $\mathcal{L}_{\text{adv}}^{f}$) and self-training parts, the accuracy reduced to 50.07%. This indicates that images from one modality have been converted closer to the other modality via image adaptation.

Then, we can find that segmentation accuracy gradually improved by adding more feature adaptation components. The results slightly improved when including $\mathcal{L}_{\text{adv}}^{\hat{S}}$. The models including $\mathcal{L}_{\text{adv}}^{p}$ and $\mathcal{L}_{\text{adv}}^{f}$ significantly improve the performance, indicating the importance of semantic prediction space and feature space. The completed model with the self-training parts leads to a further noticeable improvement in the accuracy of segmentation results.

In the paired case, we analyzed the influence of the tradeoff parameter λ (seen in Table IX). It can be seen from the table that the accuracy first increased and then decreased. The optimal value is around 10. In the unpaired case, we already analyzed the influence of the different components related to parameters λ_{adv}^{O}, λ_{adv}^{S}, λ_{cyc}, λ_{seg}, $\lambda_{\text{adv}}^{\hat{S}}$, λ_{adv}^{p}, and λ_{adv}^{f}.

TABLE IX

INFLUENCE OF λ IN THE PAIRED CASE (TRAINING ON SAR AND TESTING ON OPTICAL)

λ	1	5	10	15	20
IoU	52.56	53.64	55.62	55.60	55.21

These parameters adjust the importance of each component in the framework. Based on the experiments, our empirical settings are fine to obtain good results.

V. CONCLUSION

This article proposes a novel technique, DML, and applies it to VHR building semantic segmentation. We design different frameworks for the paired and unpaired SAR and optical datasets. The two frameworks combine image and feature adaptations to transform image appearance and domain-invariant feature learning. The additional self-training module is used to fine-tune the segmentation work to improve the performance. We validate our methods on the SpaceNet 6 from SAR to optical by comparing it with various state-of-the-art methods. Experimental results demonstrate the superiority of our network over the others in terms of IoU. Our frameworks are general and can be easily extended to other semantic segmentation applications in the remote sensing community.

ACKNOWLEDGMENT

The authors would like to thank the SpaceNet 6 challenge for providing the optical and SAR datasets, respectively.

REFERENCES

[1] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.

[2] N. Audebert, B. L. Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[3] Y. Ye, L. Bruzzone, J. Shan, F. Bovolo, and Q. Zhu, "Fast and robust matching for multimodal remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9059–9070, Nov. 2019.

[4] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "CMIR-NET: A deep learning based model for cross-modal retrieval in remote sensing," *Pattern Recognit. Lett.*, vol. 131, pp. 456–462, Mar. 2020.

[5] X. Blaes, L. Vanhalle, and P. Defourny, "Efficiency of crop identification based on optical and SAR image time series," *Remote Sens. Environ.*, vol. 96, nos. 3–4, pp. 352–365, Jun. 2005.

[6] F. Tupin and M. Roux, "Detection of building outlines based on the fusion of SAR and optical features," *ISPRS J. Photogramm. Remote Sens.*, vol. 58, nos. 1–2, pp. 71–82, 2003.

[7] H. Sportouche, F. Tupin, and L. Denise, "Extraction and three-dimensional reconstruction of isolated buildings in urban scenes from high-resolution optical and SAR spaceborne images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3932–3946, Oct. 2011.

[8] K. Van Tricht, A. Gobin, S. Gilliams, and I. Piccard, "Synergistic use of radar Sentinel-1 and optical Sentinel-2 imagery for crop mapping: A case study for Belgium," *Remote Sens.*, vol. 10, no. 10, p. 1642, Oct. 2018.

[9] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 333–346, Aug. 2020.

[10] S. Liu, Z. Qi, X. Li, and A. Yeh, "Integration of convolutional neural networks and object-based post-classification refinement for land use and land cover mapping with optical and SAR data," *Remote Sens.*, vol. 11, no. 6, p. 690, Mar. 2019.

[11] B. Adriano, J. Xia, G. Baier, N. Yokoya, and S. Koshimura, "Multi-source data fusion based on ensemble learning for rapid building damage mapping during the 2018 Sulawesi earthquake and tsunami in Palu, Indonesia," *Remote Sens.*, vol. 11, no. 7, p. 886, Apr. 2019.

[12] B. Adriano *et al.*, "Learning from multimodal and multitemporal Earth observation data for building damage mapping," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 132–143, May 2021.

[13] R. Touati, M. Mignotte, and M. Dahmane, "Multimodal change detection in remote sensing images using an unsupervised pixel pairwise-based Markov random field model," *IEEE Trans. Image Process.*, vol. 29, pp. 757–767, 2020.

[14] A. Zampieri, G. Charpiat, N. Girard, and Y. Tarabalka, "Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 657–673.

[15] R. Fernandez-Beltran, J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "Remote sensing image fusion using hierarchical multimodal probabilistic latent semantic analysis," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4982–4993, Dec. 2018.

[16] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, Aug. 2020.

[17] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 12–23, Sep. 2020.

[18] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[19] J. Hu, D. Hong, and X. X. Zhu, "MIMA: MAPPER-induced manifold alignment for semi-supervised fusion of optical image and polarimetric SAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9025–9040, Nov. 2019.

[20] W. Xiong, Z. Xiong, Y. Zhang, Y. Cui, and X. Gu, "A deep cross-modality hashing network for SAR and optical remote sensing images retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5284–5296, 2020.

[21] L. H. Hughes, D. Marcos, S. Lobry, D. Tuia, and M. Schmitt, "A deep learning framework for matching of SAR and optical imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 166–179, Nov. 2020.

[22] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu, "Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 193–205, Jan. 2019.

[23] D. Hong, J. Chanussot, N. Yokoya, J. Kang, and X. X. Zhu, "Learning-shared cross-modality representation using multispectral-LiDAR and hyperspectral data," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1470–1474, Aug. 2020.

[24] S. Cui, A. Ma, Y. Wan, Y. Zhong, B. Luo, and M. Xu, "Cross-modality image matching network with modality-invariant feature representation for airborne-ground thermal infrared and visible datasets," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[25] L. Yang and K. Xu, "Cross modality knowledge distillation for multi-modal aerial view object classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 382–387.

[26] F. Qingyun, H. Dapeng, and W. Zhaokui, "Cross-modality fusion transformer for multispectral object detection," 2021, *arXiv:2111.00273*.

[27] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, "Multimodal GANs: Toward crossmodal hyperspectral–multispectral image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5103–5113, Jun. 2021.

[28] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "SpaceNet: A remote sensing dataset and challenge series," 2018, *arXiv:1807.01232*.

[29] D. Lam *et al.*, "XView: Objects in context in overhead imagery," 2018, *arXiv:1802.07856*.

[30] R. Gupta *et al.*, "Creating xbd: A dataset for assessing building damage from satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2019.

[31] Y. Zhu *et al.*, "Improving semantic segmentation via self-training," 2020, *arXiv:2004.14960*.

[32] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014, pp. 2672–2680.

[33] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[34] J. Noa Turnes, J. D. B. Castro, D. L. Torres, P. J. S. Vega, R. Q. Feitosa, and P. N. Happ, "Atrous cGAN for SAR to optical image translation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[35] W. He and N. Yokoya, "Multi-temporal Sentinel-1 and -2 data fusion for optical image simulation," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 10, p. 389, Sep. 2018.

[36] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[37] J. D. Bermudez, P. N. Happ, R. Q. Feitosa, and D. A. B. Oliveira, "Synthesis of multispectral optical images from SAR/optical multitemporal data using conditional generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1220–1224, Aug. 2019.

[38] M. F. Reyes, S. Auer, N. Merkle, C. Henry, and M. Schmitt, "SAR-to-optical image translation based on conditional generative adversarial networks—Optimization, opportunities and limits," *Remote Sens.*, vol. 11, no. 17, p. 2067, Sep. 2019.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015* (Lecture Notes in Computer Science), vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[40] 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.

[41] J. Choi, T. Kim, and C. Kim, "Self-ensembling with GAN-based data augmentation for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6829–6839.

[42] S.-W. Huang, C.-T. Lin, S.-P. Chen, Y.-Y. Wu, P.-H. Hsu, and S.-H. Lai, "AugGAN: Cross domain adaptation with GAN-based data augmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 718–731.

[43] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1–35, Jan. 2016.

[44] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7167–7176.

[45] J. Hoffman *et al.*, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. 35th Int. Conf. Mach. Learn.*, vol. 80, J. Dy and A. Krause, Eds. Stockholm, Sweden: Stockholmsmässan, Jul. 2018, pp. 1989–1998. [Online]. Available: http://proceedings.mlr.press/v80/hoffman18a.html

[46] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[47] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2604–2613.

[48] J. Shermeyer *et al.*, "SpaceNet 6: Multi-sensor all weather mapping dataset," 2020, *arXiv:2004.06500*.

[49] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 472–480.

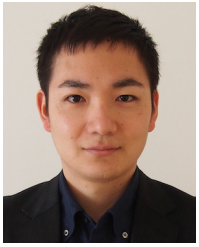
[50] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, "CrDoCo: pixel-level domain transfer with cross-domain consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3–19.

[51] J. Zhang, J. Zhou, M. Li, H. Zhou, and T. Yu, "Quality assessment of SAR-to-optical image translation," *Remote Sens.*, vol. 12, no. 21, p. 3472, Oct. 2020.

Junshi Xia (Senior Member, IEEE) received the B.S. degree in geographic information systems and the Ph.D. degree in photogrammetry and remote sensing from the China University of Mining and Technology, Xuzhou, China, in 2008 and 2013, respectively, and the Ph.D. degree in image processing from the Grenoble Images Speech Signals and Automatics Laboratory, Grenoble Institute of Technology, Grenoble, France, in 2014.

From 2014 to 2015, he was a Visiting Scientist with the Department of Geographic Information Sciences, Nanjing University, Nanjing, China. From 2015 to 2016, he was a Post-Doctoral Research Fellow with the University of Bordeaux, Bordeaux, France. From 2016 to 2018, he was the Japan Society for the Promotion of Science (JSPS) Post-Doctoral Overseas Research Fellow with The University of Tokyo, Tokyo, Japan. Since 2018, he has been a Research Scientist with the RIKEN Center for Advanced Intelligence Project (AIP), Tokyo. His research interests include multiple classifier systems in remote sensing, hyperspectral remote sensing image processing, and deep learning in remote sensing applications.

Dr. Xia was a recipient of the first place prize in the IEEE Geoscience and Remote Sensing Society Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee in 2017. Since 2019, he has been an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL) and the Guest Editor for *Remote Sensing* and the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS).



Naoto Yokoya (Member, IEEE) received the M.Eng. and Ph.D. degrees from the Department of Aeronautics and Astronautics, The University of Tokyo, Tokyo, Japan, in 2010 and 2013, respectively.

He was an Assistant Professor with The University of Tokyo from 2013 to 2017. From 2015 to 2017, he was an Alexander von Humboldt Fellow with the German Aerospace Center (DLR), Oberpfaffenhofen, Germany, and Technical University of Munich (TUM), Munich, Germany. He is a Lecturer with The University of Tokyo and a Unit Leader with the RIKEN Center for Advanced Intelligence Project, Tokyo, where he leads the Geoinformatics Unit. His research is focused on the development of image processing, data fusion, and machine learning algorithms for understanding remote sensing images, with applications to disaster management.

Dr. Yokoya won the first place in the 2017 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee (IADF TC). He was the Chair from 2019 to 2021, a Co-Chair of the IEEE GRSS IADF TC from 2017 to 2019, and also the Secretary of the IEEE GRSS All Japan Joint Chapter from 2018 to 2021. He was an Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS) from 2018 to 2021. He has been an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING since 2021.

Gerald Baier (Member, IEEE) received the M.Sc. degree in electrical engineering from the Université catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium, and the Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2012, and the Ph.D. degree in synthetic aperture radar interferometry from the Technical University of Munich, Munich, Germany, in 2018.

From 2014 to 2018, he was with the Department for SAR Signal Processing, German Aerospace Center's (DLR), Remote Sensing Technology Institute, Munich. He was a Post-Doctoral Researcher from 2018 to 2020 with Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan. In 2020, he joined Synspective, Tokyo, which is building a small SAR satellite constellation and became a Visiting Scientist with the RIKEN Center for Advanced Intelligence Project. His research interests include signal processing, machine learning, synthetic aperture radar, and high-performance computing.