PiCoCo: Pixelwise Contrast and Consistency Learning for Semisupervised Building Footprint Segmentation

Jian Kang, *Member, IEEE*, Zhirui Wang, Ruoxin Zhu, Xian Sun, *Senior Member, IEEE*, Ruben Fernandez-Beltran, *Senior Member, IEEE*, and Antonio Plaza, *Fellow, IEEE*

***Abstract*—Building footprint segmentation from high-resolution remote sensing (RS) images plays a vital role in urban planning, disaster response, and population density estimation. Convolutional neural networks (CNNs) have been recently used as a workhorse for effectively generating building footprints. However, to completely exploit the prediction power of CNNs, large-scale pixel-level annotations are required. Most state-of-the-art methods based on CNNs are focused on the design of network architectures for improving the predictions of building footprints with full annotations, while few works have been done on building footprint segmentation with limited annotations. In this article, we propose a novel semisupervised learning method for building footprint segmentation, which can effectively predict building footprints based on the network trained with few annotations (e.g., only 0.0324 km^2 out of 2.25-km^2 area is labeled). The proposed method is based on investigating the contrast between the building and background pixels in latent space and the consistency of predictions obtained from the CNN models when the input RS images are perturbed. Thus, we term the proposed semisupervised learning framework of building footprint segmentation as PiCoCo, which is based on the enforcement of Pixelwise Contrast and Consistency during the learning phase. Our experiments, conducted on two benchmark building segmentation datasets, validate the effectiveness of our proposed framework as compared to several state-of-the-art building footprint extraction and semisupervised semantic segmentation methods.**

***Index Terms*—Building footprint segmentation, consistency learning, contrastive learning, missing labels, semantic segmentation, semisupervised learning.**

Manuscript received August 25, 2021; revised September 20, 2021; accepted October 5, 2021. Date of publication October 11, 2021; date of current version October 27, 2021. This work was supported in part by the Jiangsu Province Science Foundation for Youths under Grant BK20210707, in part by the National Natural Science Foundation of China under Grant 62101371 and Grant 62076241, the Priority Academy Program Development of Jiangsu Higher Education Institutions, in part by the Ministry of Science, Innovation and Universities of Spain under Grant RTI2018-098651-B-C54 and Grant PID2019-110315RB-I00 (APRISA), in part by the Valencian Government of Spain under Grant GV/2020/167, in part by FEDER-Junta de Extremadura under Grant GR18060, and in part by the European Union through the H2020 EOXPOSURE Project under Grant 734541. *(Corresponding author: Zhirui Wang.)*

Jian Kang is with the School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China (e-mail: jiankang@suda.edu.cn).

Zhirui Wang is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China and also with the Key Laboratory of Network Information System Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhirui1990@126.com).

Ruoxin Zhu is with the State Key Laboratory of Geo-Information Engineering, Xi'an Research Institute of Surveying and Mapping, Xi'an 710054, China (e-mail: ruoxin.zhu@tum.de).

Xian Sun is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, with the Key Laboratory of Network Information System Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: sunxian@mail.ie.ac.cn).

Ruben Fernandez-Beltran is with the Institute of New Imaging Technologies, Universitat Jaume I, 12071 Castellón de la Plana, Spain (e-mail: rufernan@uji.es).

Antonio Plaza is with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, 10003 Cáceres, Spain (e-mail: aplaza@unex.es).

Digital Object Identifier 10.1109/JSTARS.2021.3119286

I. Introduction

SEGMENTING building footprints from high-resolution remote sensing (RS) images has become a basic task within the field of intelligent RS image interpretation. The footprints of buildings are indispensable elements for the research in urban planning [1], [2], disaster response [3], population density estimation [4], etc. Although distinguishing building pixels from the background is a binary segmentation problem, it is still a challenging and hot topic in the RS community, due to the complex context information present in high-resolution RS images. In recent decades, with the rapid development of aerial and spaceborne sensors, massive volumes of high-resolution RS data have significantly promoted the technical evolution of building footprint extraction [5]–[11].

Conventional methods for characterizing building footprints rely on hand-crafted features extracted from high-resolution RS images, such as contours [12], geometry [13], and morphology [14]. In addition, assisted data sources, such as digital surface models, light detection and ranging, or geographic information system, have been widely utilized for improving the building footprint segmentation accuracy with their complementary information [15]–[19]. Although the aforementioned methods have achieved prominent performances in building footprint generation, they often suffer from accuracy degradation on complicated RS images. Nowadays, with the rapid development of deep learning methods, convolutional neural networks (CNNs) have been widely utilized for the intelligent interpretation of high-resolution RS images in a data-driven manner [20]–[23]. By simultaneously and automatically learning low- and high-level features from massive RS images, CNNs can effectively capture both the shape and semantic information for building regions, which significantly improve the generalization and robustness capabilities for the footprint segmentation of

Fig. 1. Semisupervised building footprint segmentation: learning to segment buildings based on the training areas where a small portion is labeled. Within the above 2.25-km^2 area, only an area of 0.0324 km^2 is labeled and the other area of 2.2176 km^2 is without annotations. (The image was obtained from the Inria Aerial Image Labeling Dataset.)

diverse RS images compared to the conventional methods [24]. Therefore, extensive research has been pursued for developing advanced deep learning methods to automatically extract building footprints [24]–[27]. However, to completely release the prediction power of CNNs, large-scale and pixel-level annotations are required. Such a labeling procedure is unrealistic for humans, especially when the RS data are scalable. One alternative approach for annotating ground-truth building regions is based on crowd-sourcing geospatial information, e.g., Google Maps or OpenStreetMap [6], [28]. Nonetheless, under this scenario, missing or incorrect annotations may often appear in the generated ground-truth building mask layer due to some plausible reasons, including urban construction, delayed updating, disagreement among different volunteer annotators, or even low-quality volunteered geographic information. With corrupted annotations as the ground truth, CNNs can be easily overfit to the associated training samples, which severely influences the prediction accuracy of the obtained models [29].

In order to avoid the huge labor cost associated with RS images with full annotations and the performance degradation based on the trained models with label noise, we seek to design a proper semisupervised segmentation method for extracting the building footprints, where limited annotations are required. For example, as shown in Fig. 1, within a 2.25-km^2 area, only a small portion of 0.0324 km^2 is labeled and the other area of 2.2176 km^2 is without annotations. By taking advantages of limited areas with labels and large amounts of unlabeled areas, we propose a novel semisupervised building footprint segmentation framework—PiCoCo, which is based on contrast and consistency learning in a pixelwise manner. Specifically, PiCoCo is constructed on two learning concepts: 1) contrast learning, which is aimed at learning compact and discriminative latent representation space for distinguishing building and background pixels; and 2) consistency learning, which is targeted at imposing the prediction consistency of the models on different perturbations of input images. To this end, the main contributions of the proposed framework can be summarized as follows.

1) A novel semisupervised building footprint segmentation framework, i.e., PiCoCo, is proposed for effectively learning the building footprints with only around 1% pixels existing labels.
2) We first investigate the need for simultaneously imposing the contrast among the features from labeled pixels and the consistency of the predictions from unlabeled pixels within the semisupervised learning framework.
3) When applied to two standard benchmark datasets, PiCoCo outperforms several state-of-the-art building footprint extraction and semisupervised semantic segmentation methods, which demonstrates great potential in real applications. The codes of this paper will be made publicly available in https://github.com/jiankang1991/JSTARS_PiCoCo.

The rest of this article is organized as follows. Section II presents some related work from the perspectives of building footprint and semisupervised semantic segmentation. Section III introduces the overall framework of PiCoCo. Section IV demonstrates the conducted experiments and analyzes the associated results. Section V concludes this article with some remarks and hints for the follow-on research directions.

II. Related Work

A. Building Footprint Segmentation

Recently, CNNs have been served as workhorse for effectively learning building footprints from high-resolution RS images. One of the first deep-learning-based building footprint segmentation methods is based on the fully convolutional network (FCN), which learns the building segments based on the skip architecture fusing low- and high-level semantic information in a fully convolutional manner [24], [30]. An encoder–decoder CNN framework with a spatial residual inception module is proposed in [31], which captures and fuses multiple scales of building features to generate the final footprints. Wei *et al.* proposed an FCN architecture with multiscale feature aggregation and the polygon regulation for extracting and refining the building boundaries [32]. By directly exploiting hierarchical features, Li *et al.* introduced a multiple-feature reuse network (MFRN) to accelerate the computational performance of the models applied on very large input RS images [33]. Li *et al.* integrated feature pairwise conditional random field into CNN models for learning sharp building boundaries and fine-grained building segments [34]. Zhu *et al.* proposed a multiple attending path neural network for precisely generating multiscale building footprints based on combining a multiscale feature extraction strategy and attention mechanisms [35]. A modified Pix2Pix [36] framework is proposed in [37] to solve the problem of extracted inaccurate boundaries. Different from the building footprint extraction based on learning segments, extensive research has also been focused on learning vector polygons of buildings. For example, Li *et al.* proposed PolyMapper to directly extract the

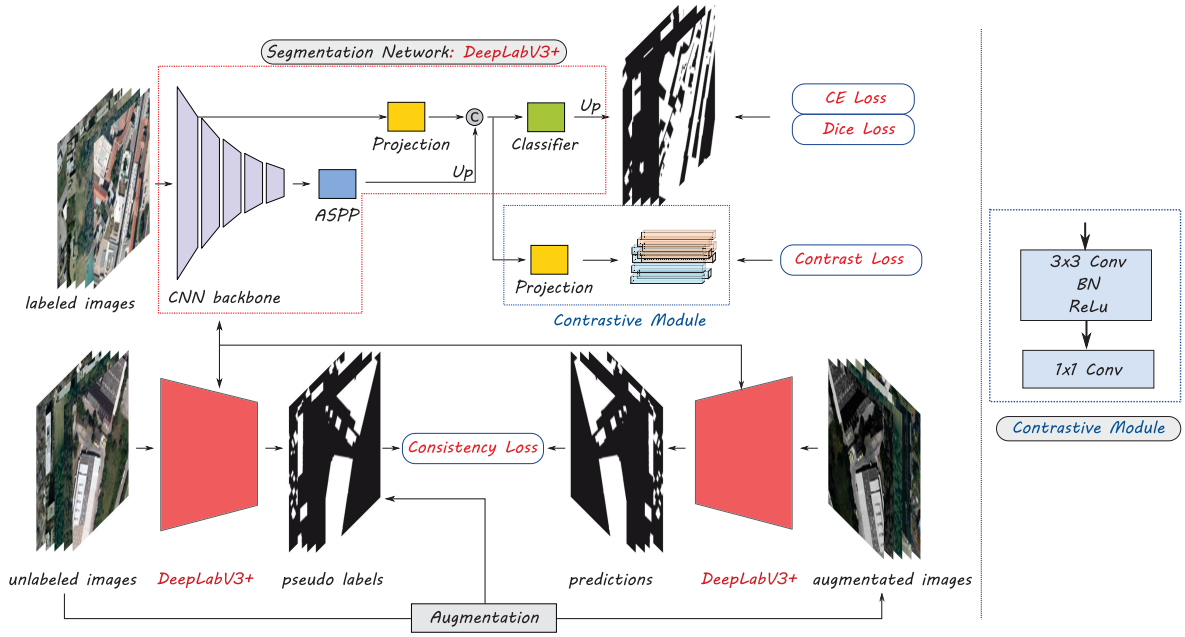

Fig. 2. Graphical illustration of the semisupervised segmentation framework PiCoCo. The labeled images are fed into a segmentation model, such as DeepLabV3+, to calculate the contrast and supervised losses, where the contrastive module is consisted of Conv-BN-ReLu and Conv layers. Then, unlabeled images are fed into the same model to obtain pseudo labels. After that, augmented image and pseudo label pairs are obtained. At last, predictions of the augmented images can be made by the model and the consistency loss is calculated.

topology map of building footprints based on the combination of CNNs and recurrent neural networks [38]. Chen *et al.* exploited the FCN-like segmentation method to generate the initial building contours and learn the shape priors of building polygons based on a modified PointNet [39]. By integrating a frame field output into a deep image segmentation model, Girard *et al.* proposed a new polygonization algorithm for learning more accurate building contours [40]. In addition, some works have been introduced for designing novel losses to accurately predict building regions and boundaries. Yuan *et al.* proposed to exploit the signed distance function, which calculates the distance from the pixels to their nearest points on the boundaries, for effectively generating the building contours [41]. Wu *et al.* regularized the region-based cross-entropy (CE) loss based on the boundary loss for extracting building segments and outlines [9]. Bokhovkin and Burnaev proposed a surrogate loss to penalize the misalignment of building boundaries and achieved the out-performance than the commonly utilized CE and Dice losses [42]. Although the aforementioned methods have significantly improved the performance of the building footprint generation, a few works have been investigated to learn the segmentation models based on limited RS images with annotations.

B. Semisupervised Semantic Segmentation

Semisupervised semantic segmentation is aimed at learning the segmentation models based on both labeled and unlabeled images [43]–[45]. Recently, one of the most effective approaches is based on consistency learning. By randomly augmenting the input images and enforcing the consistency on their predictions, the decision function can be learned to lie in low-density regions [46], [47]. Besides the perturbation on the input images, some works focused on developing semisupervised segmentation methods have been based on perturbing latent features and imposing the associated consistency. Ouali *et al.* first observed that the low-density regions were more apparent within the latent feature space than within the input images and proposed a cross-consistency training method, where the invariance of the predictions was enforced over various perturbations of latent image representations [48]. By means of knowledge exchange between two student segmentation models, Peng *et al.* also insisted the consistent predictions on unlabeled images of the two models [49]. Zou *et al.* generated well-calibrated structured pseudo labels for training unlabeled and weakly labeled images and integrated such a strategy into a one-stage consistency training framework [50]. Chen *et al.* developed a cross-consistency training method based on the mutual supervision of two student models, which were independently initialized [51]. Different from these methods, we integrate a contrast learning strategy within the consistency learning framework, where the contrast learning is imposed on labeled images to regularize the compactness and distance of intra- and interclass latent features. Although contrast learning has been investigated for semisupervised segmentation [52], we further show that both the contrast and consistency learning are essential for semisupervised segmentation.

III. Methodology

The proposed PiCoCo mainly consists of two learning strategies: 1) pulling intraclass representations and separating interclass representations in latent space for labeled images; and 2) imposing the prediction consistency of the models on different augmented unlabeled images. To achieve this, the models are reused three times within each learning iteration, and a joint

loss function is proposed, which includes the supervised loss term, the contrast loss term, and the consistency loss term. Fig. 2 graphically illustrates the proposed framework. In the following, we describe all these components in detail.

A. Notations

Let $\mathcal{X}^{\mathrm{L}} = \{\mathbf{X}_1^{\mathrm{L}}, \ldots, \mathbf{X}_{N^{\mathrm{L}}}^{\mathrm{L}}\}$ denote a building extraction dataset with N^{L} labeled images with binary masks, i.e., $\mathcal{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_{N^{\mathrm{L}}}\}$, where each element of $\mathbf{Y}_i$ is either 0 or 1. In this article, 1 indicates the building region and 0 is considered as background. Within the framework of semisupervised learning, there also exists unlabeled dataset $\mathcal{X}^{\mathrm{U}} = \{\mathbf{X}_1^{\mathrm{U}}, \ldots, \mathbf{X}_{N^{\mathrm{U}}}^{\mathrm{U}}\}$ for assisting to learn segmentation models. We exploit an encoder model denoted as $f(\cdot)$ to learn a latent representation map $\mathbf{Z}_i$ of the input image $\mathbf{X}_i$. For example, within DeepLabV3+ [53], $\mathbf{Z}_i$ is constructed by concatenating the low-level features from the first convolutional layer and the high-level features from Atrous Spatial Pyramid Pooling layer. Then, $\mathbf{Z}_i$ is fed into one classifier head $\phi_c(\cdot)$ and one projection head $\phi_p(\cdot)$ to generate predicted building mask $\hat{\mathbf{Y}}_i$ and dense vectors $\mathbf{F}_i$, respectively.

B. PiCoCo

1) Pixelwise Contrast Learning: With annotations, contrast learning aims at characterizing the latent representation space by decreasing and increasing the distances of intra- and interclass representations, respectively [52], [54]. To achieve this, we minimize the following loss:

$$
\begin{aligned}
L_{\mathrm{contrast}} &= \sum_{c\in\{0,1\}} \sum_{\mathbf{f}_q \in \mathcal{R}_q^c} \\
&\quad - \log \frac{\exp(\mathbf{f}_q^T \mathbf{f}_k^{c,+}/\tau)}{\exp(\mathbf{f}_q^T \mathbf{f}_k^{c,+}/\tau) + \sum_{\mathbf{f}_k^- \in \mathcal{R}_k^c} \exp(\mathbf{f}_q^T \mathbf{f}_k^-/\tau)} \\
\mathcal{R}_q^c &= \bigcup_{(u,v)} \mathbb{1}(\mathbf{Y}[u,v] = c)\mathbf{F}[u,v,:] \\
\mathbf{f}_k^{c,+} &= \frac{1}{|\mathcal{R}_q^c|} \sum_{\mathbf{f}_q \in \mathcal{R}_q^c} \mathbf{f}_q \\
\mathcal{R}_k^c &= \bigcup_{(u,v)} \mathbb{1}(\mathbf{Y}[u,v] \neq c)\mathbf{F}[u,v,:]. \qquad (1)
\end{aligned}
$$

Here, c denotes the class indicator and $\mathbf{f}_q$ are the query vectors sampled from $\mathbf{F}$. Within one labeled mini-batch, we utilize class information to constrain the sampling of M_q query vectors, i.e., all query vectors are from the same class. It can be achieved by resizing the spatial dimension of the building mask $\mathbf{Y}$ same as $\mathbf{F}$ and then sampling from $\mathcal{R}_q^c$. $\mathbf{f}_k^{c,+}$ represents the key vector, which is the mean vector of class c. In order to discriminate the positive representations from the negative ones, we also sample M_k vectors from $\mathcal{R}_k^c$, which is the set of key vectors from different classes with respect to $\mathbf{f}_q$. τ denotes the temperature parameter, which controls the concentration level of the distributions. The objective of (1) is to learn compact and discriminative classwise representations $\mathbf{Z}$ in latent space. As suggested in [54], it is beneficial to apply contrast loss on $\mathbf{F}$ obtained from the projection head $\phi_p(\cdot)$ rather than $\mathbf{Z}$, where $\phi_p(\cdot)$ is composed by two stacked convolutional layers.

Hard query sampling: It is time consuming and computationally expensive to sample all the dense vectors from $\mathcal{R}_q^c$ in each training iteration under the above pixelwise contrast learning framework. To avoid such issue, it is necessary to sample *informative* query vectors. Following [52], we sample hard queries whose associated pixel prediction confidences are below a defined threshold δ. Specifically, in practice, $\mathcal{R}_q^c$ is replaced by

$$
\mathcal{R}_q^{c,\mathrm{hard}} = \bigcup_{(u,v)} \mathbb{1}(\mathbf{Y}[u,v] = c, \hat{\mathbf{Y}}[u,v] \leq \delta)\mathbf{F}[u,v,:]. \qquad (2)
$$

We ignore those samples with high classification confidence, since they make less contributions to the contrast loss.

It is worth noting that we only make contrast of the pixels from labeled inputs ($\mathbf{X}^{\mathrm{L}}$) under the semisupervised learning framework. Differently, in [52], the contrast is made between the positive and negative pixels from both the truly annotated pixels and the pixels with pseudo labels predicted on the unlabeled inputs ($\mathbf{X}^{\mathrm{U}}$). One consideration is that the contrast loss is only applied on the latent space of truly annotated pixels in order to improve the generalization capability of the learned classifiers for unseen images. In addition, the generated pseudo labels may contain label noise, which influences such metric learning performance of classwise representations.

2) Pixelwise Consistency Learning: To leverage a large number of unlabeled images $\mathbf{X}^{\mathrm{U}}$, inspired by [55], we apply pixelwise consistency learning on the class predictions from the input images and their augmented versions. Specifically, we first generate pseudo labels $\hat{\mathbf{Y}}^{\mathrm{U}}$ given the inputs $\mathbf{X}^{\mathrm{U}}$. Then, heavy augmentations are applied on the pairs of ($\mathbf{X}^{\mathrm{U}}$, $\hat{\mathbf{Y}}^{\mathrm{U}}$) and yield the augmented pairs ($\mathbf{X}_{\mathrm{pert}}^{\mathrm{U}}$, $\hat{\mathbf{Y}}_{\mathrm{pert}}^{\mathrm{U}}$). After the class predictions of $\mathbf{X}_{\mathrm{pert}}^{\mathrm{U}}$ are obtained through $f \circ \phi_c$, the consistency of the models on those predictions can be achieved by increasing the agreement between $\hat{\mathbf{Y}}_{\mathrm{pert}}^{\mathrm{U}}$ and probabilities $\mathbf{P}(\hat{\mathbf{Y}}^{\mathrm{U}}|\mathbf{X}_{\mathrm{pert}}^{\mathrm{U}})$. In our work, we minimize CE and Dice losses to achieve this. Specifically, such consistency learning strategy can be described by

$$
\begin{array}{ccccc}
\mathbf{X}^{\mathrm{U}} & \rightarrow & f \circ \phi_c & \rightarrow \hat{\mathbf{Y}}^{\mathrm{U}} & \rightarrow \hat{\mathbf{Y}}_{\mathrm{pert}}^{\mathrm{U}} \\
\downarrow & & \updownarrow & & \\
\mathbf{X}_{\mathrm{pert}}^{\mathrm{U}} & \rightarrow & f \circ \phi_c & \rightarrow \mathbf{P}(\hat{\mathbf{Y}}^{\mathrm{U}}|\mathbf{X}_{\mathrm{pert}}^{\mathrm{U}}). &
\end{array} \qquad (3)
$$

The consistency loss is defined as

$$
\begin{aligned}
L_{\mathrm{consistency}} &= L_{\mathrm{CE}}(\mathbf{P}(\hat{\mathbf{Y}}^{\mathrm{U}}|\mathbf{X}_{\mathrm{pert}}^{\mathrm{U}}), \hat{\mathbf{Y}}_{\mathrm{pert}}^{\mathrm{U}}) \\
&\quad + L_{\mathrm{Dice}}(\mathbf{P}(\hat{\mathbf{Y}}^{\mathrm{U}}|\mathbf{X}_{\mathrm{pert}}^{\mathrm{U}}), \hat{\mathbf{Y}}_{\mathrm{pert}}^{\mathrm{U}}). \qquad (4)
\end{aligned}
$$

The objective of (4) is to impose pixelwise consistency between the predictions of the original input images and their perturbed versions. In this way, we would like to learn segmentation models that can achieve smooth predictions on the unlabeled images with perturbations rather than the models that achieve high predictions on those images without stable performances when perturbations are applied. The former models are more robustly adapted to unlabeled images than the latter ones, which may be better generalized to unseen images.

3) Joint Loss Function: Besides the above two loss terms, given the true annotations $\mathbf{Y}^{\mathrm{L}}$ and class probabilities

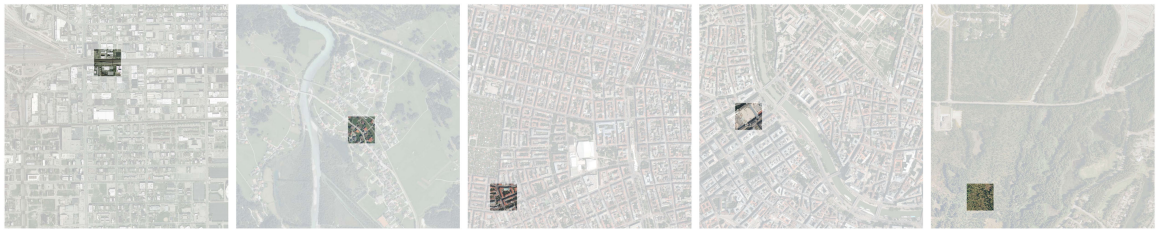

Fig. 3. Training set construction for the Inria dataset. For each tile of 2.25-km^2 area, we only randomly crop 0.0324-km^2 area with the spatial size of 600 × 600 pixels to be labeled. The others are regarded as unlabeled images in the training set.

Algorithm 1: Optimization Scheme for PiCoCo.

Require: $\mathcal{X}^{\mathrm{L}}$, $\mathcal{Y}$, and $\mathcal{X}^{\mathrm{U}}$

1: Initialize the parameters of the encoder f, the classifier head ϕ_c, and the projection head ϕ_p, along with M_q, M_k, δ, and τ.
2: **for** The epoch number $t = 0$ to maxEpoch **do**
3: Sample one mini-batch of labeled and unlabeled images, i.e., $(\mathcal{X}_B^{\mathrm{L}}, \mathcal{Y}_B^{\mathrm{L}})$ and $\mathcal{X}_B^{\mathrm{U}}$, respectively.
4: Calculate the supervised loss $L_{\mathrm{supervised}}$ based on $\mathcal{Y}_B^{\mathrm{L}}$ and $f \circ \phi_c(\mathcal{X}_B^{\mathrm{L}})$.
5: Calculate the contrast loss L_{contrast} based on (1).
6: Obtain the pseudo labels $\hat{\mathbf{Y}}_B^{\mathrm{U}}$ based on $f \circ \phi_c(\mathcal{X}_B^{\mathrm{U}})$
7: Augment the unlabeled image and pseudo label pairs $(\mathbf{X}_B^{\mathrm{U}}, \hat{\mathbf{Y}}_B^{\mathrm{U}})$ and yield $(\mathbf{X}_{B_{\mathrm{pert}}}^{\mathrm{U}}, \hat{\mathbf{Y}}_{B_{\mathrm{pert}}}^{\mathrm{U}})$.
8: Calculate the consistency loss $L_{\mathrm{consistency}}$.
9: Obtain the joint loss L and backpropagate the gradients.
10: **end for**

Ensure: f, ϕ_c and ϕ_p.

$\mathbf{P}(\hat{\mathbf{Y}}^{\mathrm{L}}|\mathbf{X}^{\mathrm{L}})$, we minimize the following supervised loss term to improve the confidences of the class predictions on the labeled images:

$$L_{\mathrm{supervised}} = L_{\mathrm{CE}}(\mathbf{P}(\hat{\mathbf{Y}}^{\mathrm{L}}|\mathbf{X}^{\mathrm{L}}), \mathbf{Y}^{\mathrm{L}}) + L_{\mathrm{Dice}}(\mathbf{P}(\hat{\mathbf{Y}}^{\mathrm{L}}|\mathbf{X}^{\mathrm{L}}), \mathbf{Y}^{\mathrm{L}}). \quad (5)$$

To this end, the joint loss function for semisupervised building footprint segmentation is formulated as

$$L = L_{\mathrm{supervised}} + L_{\mathrm{contrast}} + L_{\mathrm{consistency}}. \quad (6)$$

The associated optimization scheme of PiCoCo is described in Algorithm 1.

IV. Experiments

A. Experimental Setup

1) Dataset Configuration: In our experiments, we exploit two building footprint extraction benchmark datasets: 1) Inria Aerial Image Labeling Dataset [5] and 2) NZ32km2 [9].

1) *Inria*: This dataset contains 180 aerial orthorectified tiles with a spatial resolution of 0.3 m covered 405-km^2 areas over Austin, Chicago, Kitsap County, Vienna, and West Tyrol in the training set and 180 tiles over Bellingham, San Francisco, Bloomington, Innsbruck, and East Tyrol in the enclosed test set. For our experiments, we split the original training set into training and test sets according to [56]. From the training set, we only randomly crop an area with the size of 600 × 600 pixels covered 0.0324 km^2 for each tile to construct labeled training set (as shown in Fig. 3), and the others belong to unlabeled training set. Thus, we have a training set with 153 labeled patches with the size of 600 × 600 pixels and 15 147 unlabeled patches.
2) *NZ32km2*: This dataset covers a 32-km^2 area in Christchurch, New Zealand, which is composed of eight tiles (four tiles for training and test) with a spatial resolution of 0.075 m. Similar to the above setting, we randomly crop 122 patches with the size of 600 × 600 to be the labeled patches, which only cover areas of 0.247 km^2. Thus, we finally have a training set with 122 and 12 166 labeled and unlabeled image patches, as summarized in Table I.

TABLE I
Number of Labeled and Unlabeled Patches in the Training Sets of Inria and NZ32km2

	Inria	NZ32km2
w/ labels	153	122
w/o labels	15147	12166

2) Implementation Details: For the segmentation architecture f, we adopt DeepLabV3+ [53] since it is one of the most popular CNN architectures for segmentation. ResNet50 is selected as the CNN backbone for extracting hierarchical deep features. The classifier head ϕ_c and the projection head ϕ_p are both composed of two stacked convolutional layers to extract the class predictions and dense vectors, which are constructed by Conv[3 × 3]-BN-ReLu and Conv[1 × 1] layers. The spatial size of the input images is 600 × 600, and we feed the images into the segmentation model after the augmentation by: 1) RandomCrop with the size of 512 × 512 pixels; 2) RandomFlip; and 3) RandomRotate. The number M_q of query vectors is 256 and the number M_k of negative key vectors is 512. The threshold δ is set as 0.97 and τ is defined as 0.1. For the consistency learning strategy, the heavy augmentation is composed of: 1) RandomResizedCrop with a scale from 0.2 to 1; 2) RandomBrightnessContrast; 3) HueSaturation; and 4) ToGray. We utilize the stochastic gradient descent optimizer to train the segmentation model with an initial learning rate of 5×10^{-3} and the polynomial scheduler. We train the networks

for a total of 200 epochs with mini-batches of the size 12. All the experiments are implemented in PyTorch [57] and carried out on an NVIDIA RTX3090 GPU. To validate the effectiveness of the proposed method, we compare it to several state-of-the-art methods from both the perspectives of building extraction and semisupervised semantic segmentation.

1) *UNet [58]:* By fusing multilevel feature maps to simultaneously capture hierarchical semantics and preserving fine-grained shapes of building masks, UNet architecture and its modified versions have been widely served as strong baselines for building footprint segmentation [59], [60].
2) *MFRN [33]:* The MFRN architecture is a kind of the multifeature reuse network, where each layer is connected to all the subsequent layers with the same size.
3) *Multitask [61]:* Besides the CE loss, a multitask learning framework with the integration of truncated distance loss is proposed to learn the accurate contours of buildings.
4) *BF [42]:* A boundary regularization loss is proposed for sufficiently penalizing the misalignment of boundaries. By combining the boundary loss with CE and Dice losses, the state-of-the-art building extraction performance is achieved.
5) *ReCo [52]:* A pixelwise contrast loss is introduced for semantic segmentation, which achieves the state-of-the-art performances on both supervised and semisupervised semantic segmentation.
6) *CPS [51]:* A novel consistency regularization approach is proposed for semisupervised semantic segmentation, which imposes the consistency on two segmentation models perturbed with different initialization of the same input image.

It is worth noting that the CNN backbones for extracting deep features of the considered methods are the same, i.e., ResNet50.

3) Evaluation Metrics: For all the tiles in the two test sets, we calculate intersection over union (IoU), Dice, precision, recall, and overall accuracy (OA) scores for each tile and obtain the associated mean values, where these metrics are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{TN}} \tag{9}$$

$$\text{Dice} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{TN}} \tag{10}$$

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \tag{11}$$

TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

B. Experimental Results

1) Comparison to State-of-the-Art Approaches: Fig. 4 displays the learning curves of all the considered methods on the validation sets of the two benchmark datasets. It can be observed that the Dice scores of UNet, MFRN, and Multitask are lower than the other methods. One plausible reason is that the number of labeled training images is limited for sufficiently training the associated segmentation models from scratch. As a comparison, the DeepLabV3+ architecture is constructed upon the pretrained CNN backbones on ImageNet, e.g., ResNet50, and can be well adapted to the novel segmentation task even with limited training images. Tables II and III demonstrate the mean values of all the metrics evaluated on the compared methods of Inria and NZ32km2 datasets, respectively. In addition, we also present the results based on the DeepLabV3+ architecture trained by the images with full labels as the performance upper bound. Consistently with the above observation, the methods with DeepLabV3+ architectures perform better than the others on the tests sets. For Inria, `PiCoCo` can improve the mean IoU score with more than 2% than the recently proposed CPS method and 3% than the contrast learning-based ReCo method. `PiCoCo` achieves all the best performances on the five metrics among the compared methods. For NZ32km2, `PiCoCo` outperforms the CPS method with a mean IoU value more than 1% and reaches the best performances of three metrics out of five. Compared to CPS, `PiCoCo` only exploits one segmentation model and achieves the out-performance. CPS simultaneously trains two segmentation models and exploits cross-consistency loss to optimize each other, which requires more GPU memory and computational cost than `PiCoCo`. Compared with another contrast learning-based method, ReCo, `PiCoCo` can perform better since it uses the consistency to regularize the learning in order to well adapt the model on the unlabeled images. In addition, within ReCo, the dense vectors from both labeled and unlabeled images are made contrast to discover the latent representation space. Due to some unreliable pseudo labels from the unlabeled images, they may influence the accurate construction of the latent representation space. As a comparison, we only make contrast of the dense vectors on the labeled images to avoid such issue. Besides the above quantitative comparison, we also display some visual results in Fig. 5. Compared to some methods, such as BF, `PiCoCo` has less false positive predictions. Moreover, the building boundaries obtained by the proposed method are more accurate than the other methods, even the method BF that integrates a boundary loss. As shown in the following subsection, we enforce to make contrast of the hard query vectors with respect to the negative ones. The locations of those informative samples usually lie on the building boundaries so that the representations of the boundaries can be more accurately learned.

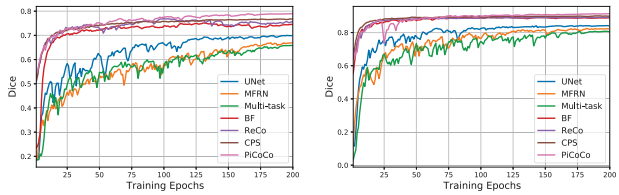

Fig. 4. Learning curves of all the considered methods on the validation sets of the two datasets: (a) Inria and (b) NZ32km2.

TABLE II
EVALUATION METRICS OF ALL THE CONSIDERED METHODS ON THE INRIA TEST SET

No.	Method	mIoU	mDice	mPrecision	mRecall	mOA
1	UNet [58]	65.38	78.89	82.00	76.35	94.03
2	MFRN [33]	59.94	74.72	71.72	78.87	92.48
3	Multi-task [61]	58.77	73.65	78.16	70.69	93.03
4	BF [42]	71.15	83.04	84.50	81.75	94.80
5	ReCo [52]	70.03	82.23	85.95	79.23	94.75
6	CPS [51]	71.67	83.41	85.05	81.98	94.97
7	PiCoCo	**73.78**	**84.82**	**86.17**	**83.64**	**95.38**
8	DeepLabv3+[full label] [53]	81.76	89.90	90.82	89.05	97.02

TABLE III
EVALUATION METRICS OF ALL THE CONSIDERED METHODS ON THE NZ32KM2 TEST SET

No.	Method	mIoU	mDice	mPrecision	mRecall	mOA
1	UNet [58]	72.11	81.51	80.04	86.63	94.81
2	MFRN [33]	69.16	79.44	73.46	**90.08**	93.77
3	Multi-task [61]	67.54	78.07	75.98	85.01	93.38
4	BF [42]	80.78	87.66	88.09	89.21	97.04
5	ReCo [52]	81.63	88.73	89.55	88.79	96.94
6	CPS [51]	82.08	**90.07**	90.18	88.73	97.13
7	PiCoCo	**83.19**	89.38	**90.57**	89.70	**97.33**
8	DeepLabv3+[full label] [53]	87.18	91.88	92.47	92.28	98.10

TABLE IV
ABLATION STUDY OF THE PROPOSED METHOD

	Inria				NZ32km2			
CE-Dice	✓	✓	✓	✓	✓	✓	✓	✓
Contrast		✓		✓		✓		✓
Consistency			✓	✓			✓	✓
mIoU	67.49	73.26	70.22	**73.78**	79.60	83.17	81.55	**83.19**
mDice	80.47	84.47	82.42	**84.82**	87.28	**90.25**	89.00	89.38
mPrecision	83.91	85.92	83.38	**86.17**	88.42	**90.87**	88.88	90.57
mRecall	77.51	83.20	81.82	**83.64**	87.53	89.27	89.38	**89.70**
mOA	94.22	95.32	94.58	**95.38**	96.63	97.29	97.03	**97.33**

2) Ablation Study: To sufficiently understand the two main loss terms in PiCoCo, we conduct extensive ablation experiments to separately analyze the effects of these terms. First, we train the segmentation models with $L_{\text{supervised}}$, $L_{\text{supervised}} + L_{\text{contrast}}$, and $L_{\text{supervised}} + L_{\text{consistency}}$ and evaluate their results in Table IV. It can be seen that both the two loss terms are necessary for segmenting the building regions with limited training images with annotations. By involving L_{contrast} and $L_{\text{consistency}}$, the mean IoU performances can be boosted by more than 6% and 3% for Inria and NZ32km2 datasets, respectively. Compared to $L_{\text{consistency}}$, the performance improvement of L_{contrast} is larger, which indicates that the latent representation space constructed by the labeled images plays a vital role in the generalization capability of the trained model on the unseen images. As shown in Fig. 6, we also present the probabilities of the predicted building areas in some examples. Without the contrast loss, some background areas cannot be well distinguished from the building regions nearby. For example, in the first row of Fig. 6(a), some predicted probabilities of the parking lot areas are larger in the results of $L_{\text{supervised}}$ and $L_{\text{supervised}} + L_{\text{consistency}}$ than the results with the contrast loss involved. This indicates that the contrast learning on the representations in the latent space can assist the improvement of building extraction, especially for some confused areas. To clearly analyze the effect of the contrast loss, we demonstrate the class confidences of some examples obtained by PiCoCo in Fig. 7. Brighter color denotes the higher confidence, while the darker color indicates the lower confidence. Within the strategy of contrast learning, we focus on selecting hard query vectors from the locations with low confidences. It shows that those locations most likely belong to the building boundaries, from where the representations are paid more attention to be discriminative with respect to the background representations. In this way, more accurate building boundaries can be obtained, and less false positive predictions exist in the results. In addition, we also calculate the similarities between $\mathcal{R}_q^c$ and $\mathbf{f}_k^{c,+}$, and the ones between $\mathcal{R}_q^c$ and $\mathbf{f}_k^{c,-}$ for building pixels. Fig. 8(a) and (c) shows the histograms of the similarities between the building representations and their mean

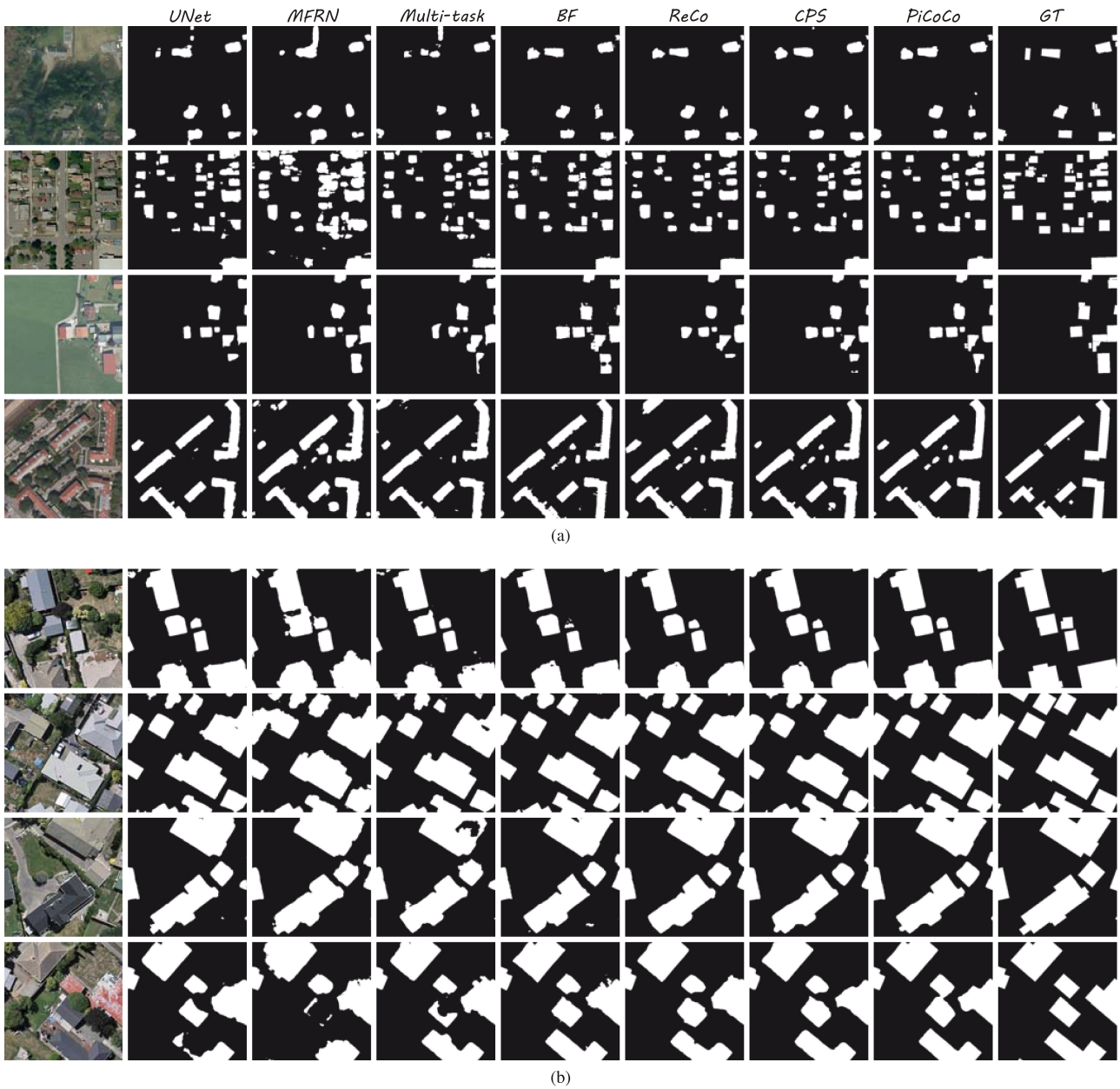

Fig. 5. Some visual comparison examples among all the considered methods: (a) Inria and (b) NZ32km2. It can be observed that PiCoCo has less false positive predictions. Moreover, the building boundaries obtained by the proposed method are more accurate than the others.

representation for Inria and NZ32km2, respectively. Fig. 8(b) and (d) shows the histograms of the similarities between the building representations and the mean representation of background. Obviously, the contrast loss can pull the intraclass representations together and push the ones from different classes away in the latent space. Thus, the learned classifier can well distinguish the building and background pixels. As demonstrated in Fig. 9, we display two predicted building footprints via PiCoCo for the test tiles in the Inria (a) and NZ32km2 (b) datasets. It can be seen that most buildings can be accurately segmented based on the proposed method.

3) Hyperparameter Analysis: We analyze the sensitivities of the parameters τ and (M_q, M_k) on the building extraction performance. τ controls the radius of the hypersphere on which the normalized representations are projected. From another perspective, it also determines the smoothness of the contrastive distribution in the contrast loss term. With larger τ, the distribution becomes smoother and smaller τ will sharpen the distribution [62]. To increase the contrastive capability of each pixel in the feature space. we make τ small and it lies in the range from 0.05 to 1 [63], and we calculate five metrics and display them in Table V. It can be observed that the building extraction accuracies do not vary much for different τ, which indicates that the choice of τ can lie in a relatively wide range, and it does not require to be carefully tuned. The parameter pair (M_q, M_k) controls the number of query and key vectors to be compared. In general, as studied in [52], larger numbers of M_q and M_k indicate that more vectors are made contrast,

TABLE V
SENSITIVITY ANALYSIS OF THE TEMPERATURE PARAMETER τ

	Inria				NZ32km2			
τ	0.05	0.1	0.5	1	0.05	0.1	0.5	1
mIoU	73.93	73.78	73.69	73.17	82.71	83.19	83.24	83.92
mDice	84.94	84.82	84.77	84.43	89.58	89.38	89.46	90.69
mPrecision	84.78	86.17	85.99	85.01	90.09	90.57	90.44	91.27
mRecall	85.20	83.64	83.70	83.95	89.54	89.70	89.69	89.62
mOA	95.31	95.38	95.37	95.18	97.22	97.33	97.32	97.41

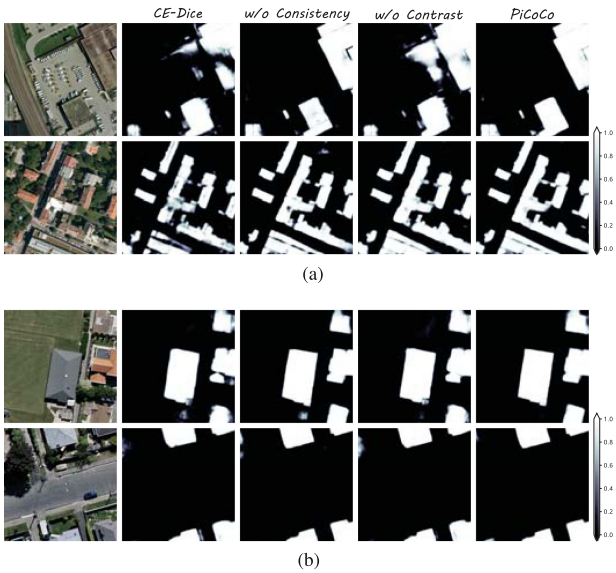

Fig. 6. Predicted probabilities of building regions based on $L_{\text{supervised}}$, $L_{\text{supervised}} + L_{\text{contrast}}$, $L_{\text{supervised}} + L_{\text{consistency}}$, and PiCoCo: (a) Inria and (b) NZ32km2.

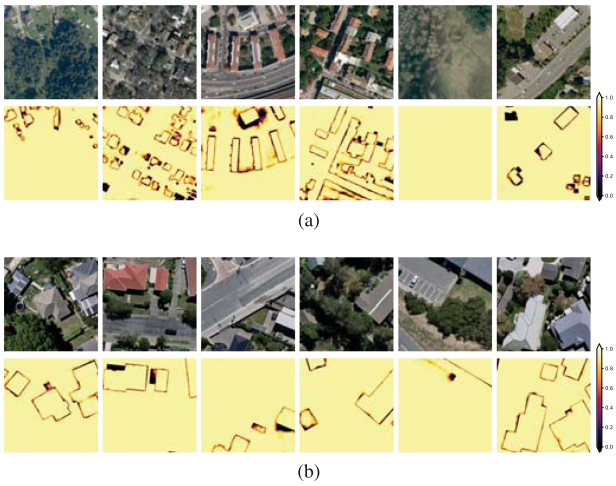

Fig. 7. Class confidences of some examples in (a) Inria and (b) NZ32km2 datasets. The brighter color denotes higher confidence, and the darker color indicates lower confidence.

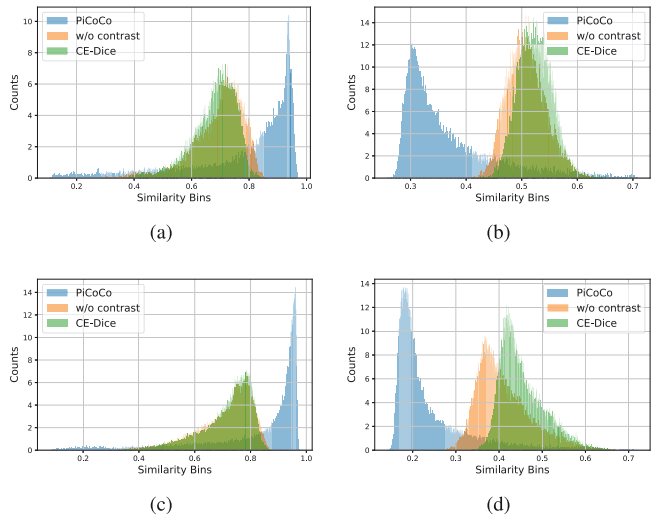

Fig. 8. Histograms of the similarities between $\mathcal{R}_q^c$ and $\mathbf{f}_k^{c,+}$ [(a) Inria and (c) NZ32km2] and the ones between $\mathcal{R}_q^c$ and $\mathbf{f}_k^{c,-}$ [(b) Inria and (d) NZ32km2] for building pixels.

which can improve the sufficiency of contrast learning. As shown in Table VI, the building segmentation accuracy is relatively stable when M_q and M_k are varied from 128 to 512 and from 256 to 1024, respectively. Different from multiclass semantic segmentation, building footprint segmentation belongs to the binary segmentation category. Thus, a query or key number from 128 to 1024 is sufficient for the contrast learning. Moreover, as discussed above, the active query vectors are most likely located around the building boundaries that do not contain large amounts of vectors. Thus, we can choose the numbers of query and key vectors starting from 128.

4) Discussion: As analyzed above, we carried out extensive experiments to validate the performance of PiCoCo when the segmentation model was trained on limited images with building annotations. Compared to the other state of the art, PiCoCo achieved the best performance especially on the Inria dataset. Differently from NZ32km2, Inria contains the RS images from different cities over the world, where the building and background pixels are more diverse. This requires that the trained segmentation model should have strong generalization capability on unseen images. With around 1% annotated pixels on the whole dataset, DeepLabv3+ trained on the fully labeled dataset only leads the performance with a mean IoU margin of 8% compared to PiCoCo. Based on the ablation study, both the contrast and consistency loss terms take effects on the performance gain for the building extraction. This may also give insights for other RS image interpretation tasks, such as scene classification, to simultaneously exploit these two terms when limited annotations in the training set are available. From the sensitivity analysis

(a) (b)

Fig. 9. Building footprint segmentation results of two test tiles based on PiCoCo: (a) Inria and (b) NZ32km2.

TABLE VI
SENSITIVITY ANALYSIS OF THE NUMBER OF QUERIES AND KEY VECTORS (M_q, M_k)

	Inria			NZ32km2		
(M_q, M_k)	(128,256)	(256,512)	(512,1024)	(128,256)	(256,512)	(512,1024)
mIoU	73.77	73.78	73.65	83.16	83.19	83.77
mDice	84.83	84.82	84.75	89.93	89.38	89.86
mPrecision	85.79	86.17	85.06	90.03	90.57	91.28
mRecall	84.04	83.64	84.56	90.15	89.70	89.59
mOA	95.38	95.38	95.33	97.29	97.33	97.40

of hyperparameters, the main parameters of PiCoCo are not required to be tediously tuned. Thus, PiCoCo can be easily adapted to different building extraction datasets. According to the segmentation results over larger areas, PiCoCo has the potential to be exploited in real scenarios, since it can accurately predict the building segments without large amounts of ground truth.

V. CONCLUSION

In this article, we propose a novel framework for semisupervised learning building segments from training data with limited annotations. To achieve this, we simultaneously make contrast between the latent representations of building and background pixels and impose the consistency between the predictions obtained from the unlabeled images and their perturbed versions. Based on the extensive experiments on two RS building datasets, our results validate the effectiveness of the newly proposed PiCoCo and demonstrate its performance compared to other state-of-the-art methods. As future work, we plan to further investigate contrast and consistency learning strategies for generating vectorized building polygons in a semisupervised manner.

REFERENCES

[1] J. Kang, M. Körner, Y. Wang, H. Taubenböck, and X. X. Zhu, "Building instance classification using street view images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 44–59, 2018.
[2] 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, May 2021.
[3] B. Adriano *et al.*, "Learning from multimodal and multitemporal earth observation data for building damage mapping," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 132–143, 2021.
[4] K. Steinnocher, A. De Bono, B. Chatenoux, D. Tiede, and L. Wendt, "Estimating urban population patterns from stereo-satellite imagery," *Eur. J. Remote Sens.*, vol. 52, no. sup2, pp. 12–25, 2019.
[5] 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.
[6] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, "Learning aerial image segmentation from online maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6054–6068, Nov. 2017.
[7] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Graduate Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2013.
[8] 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.
[9] G. Wu *et al.*, "A boundary regulated network for accurate roof segmentation and outline extraction," *Remote Sens.*, vol. 10, no. 8, 2018, Art. no. 1195.

[10] Q. Chen, L. Wang, Y. Wu, G. Wu, Z. Guo, and S. L. Waslander, "Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings," *ISPRS J. Photogrammetry Remote Sens.*, vol. 147, pp. 42–55, 2019.

[11] H. Jing, X. Sun, Z. Wang, K. Chen, W. Diao, and K. Fu, "Fine building segmentation in high-resolution SAR images via selective pyramid dilated network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6608–6623, 2021.

[12] J. Du *et al.*, "A novel framework for 2.5-D building contouring from large-scale residential scenes," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 4121–4145, Jun. 2019.

[13] Z. Li, W. Shi, Q. Wang, and Z. Miao, "Extracting man-made objects from high spatial resolution remote sensing images via fast level set evolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 883–899, Feb. 2015.

[14] N. L. Gavankar and S. K. Ghosh, "Automatic building footprint extraction from high-resolution satellite image using mathematical morphology," *Eur. J. Remote Sens.*, vol. 51, no. 1, pp. 182–193, 2018.

[15] L. Sahar, S. Muthukumar, and S. P. French, "Using aerial imagery and GIS in automated building footprint extraction and shape recognition for earthquake risk assessment of urban inventories," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3511–3520, Sep. 2010.

[16] M. Brédif, O. Tournaire, B. Vallet, and N. Champion, "Extracting polygonal building footprints from digital surface models: A fully-automatic global optimization framework," *ISPRS J. Photogrammetry Remote Sens.*, vol. 77, pp. 57–65, 2013.

[17] M. Awrangjeb, C. Zhang, and C. S. Fraser, "Automatic extraction of building roofs using LiDAR data and multispectral imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 83, pp. 1–18, 2013.

[18] S. Du, Y. Zhang, Z. Zou, S. Xu, X. He, and S. Chen, "Automatic building extraction from LiDAR data fusion of point and grid-based features," *ISPRS J. Photogrammetry Remote Sens.*, vol. 130, pp. 294–307, 2017.

[19] Z. Gharibbafghi, J. Tian, and P. Reinartz, "Modified superpixel segmentation for digital surface model refinement and building extraction from satellite stereo imagery," *Remote Sens.*, vol. 10, no. 11, 2018, Art. no. 1824.

[20] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[21] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[22] D. Hong *et al.*, "Endmember-guided unmixing network (EGU-Net): A general deep learning framework for self-supervised hyperspectral unmixing," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2021.3082289.

[23] S. Xian, W. Zhirui, S. Yuanrui, D. Wenhui, Z. Yue, and F. Kun, "AIR-SARShip-1.0: High-resolution SAR ship detection dataset," *J. Radars*, vol. 8, no. 6, pp. 852–862, 2019.

[24] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. D. Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 130, pp. 139–149, 2017.

[25] Z. Huang, G. Cheng, H. Wang, H. Li, L. Shi, and C. Pan, "Building extraction from multi-source remote sensing images via deep deconvolution neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 1835–1838.

[26] K. Bittner, S. Cui, and P. Reinartz, "Building extraction from remote sensing data using fully convolutional networks," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 42, no. W1, pp. 481–486, 2017.

[27] G. Sun, H. Huang, A. Zhang, F. Li, H. Zhao, and H. Fu, "Fusion of multiscale convolutional neural networks for building extraction in very high-resolution images," *Remote Sens.*, vol. 11, no. 3, 2019, Art. no. 227.

[28] J. Vargas, S. Srivastava, D. Tuia, and A. Falcao, "OpenStreetMap: Challenges and opportunities in machine learning and remote sensing," *IEEE Geosci. Remote Sens. Magn.*, vo. 9, no. 1, pp. 184–199, 2020.

[29] J. Kang, R. Fernandez-Beltran, X. Sun, J. Ni, and A. Plaza, "Deep learning-based building footprint extraction with missing annotations," *IEEE Geosci. Remote Sens. Lett.*, to be published. doi: 10.1109/LGRS.2021.3072589.

[30] 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.

[31] P. Liu *et al.*, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 830.

[32] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020.

[33] L. Li, J. Liang, M. Weng, and H. Zhu, "A multiple-feature reuse network to extract buildings from remote sensing imagery," *Remote Sens.*, vol. 10, no. 9, 2018, Art. no. 1350.

[34] Q. Li, Y. Shi, X. Huang, and X. X. Zhu, "Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF)," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7502–7519, Nov. 2020.

[35] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2021.

[36] 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.*, 2017, pp. 1125–1134.

[37] A. Ziaee, R. Dehbozorgi, and M. Döller, "A novel adaptive deep network for building footprint segmentation," 2021, *arXiv:2103.00286*.

[38] Z. Li, J. D. Wegner, and A. Lucchi, "Topological map extraction from overhead images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1715–1724.

[39] Q. Chen, L. Wang, S. L. Waslander, and X. Liu, "An end-to-end shape modeling framework for vectorized building outline generation from aerial images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 170, pp. 114–126, 2020.

[40] N. Girard, D. Smirnov, J. Solomon, and Y. Tarabalka, "Polygonal building segmentation by frame field learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021.

[41] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, Nov. 2018.

[42] A. Bokhovkin and E. Burnaev, "Boundary loss for remote sensing imagery semantic segmentation," in *Proc. Int. Symp. Neural Netw*. 2019, pp. 388–401.

[43] X. Sun, A. Shi, H. Huang, and H. Mayer, "BAS4Net: 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.

[44] Y. Zhang, X. Zheng, G. Liu, X. Sun, H. Wang, and K. Fu, "Semi-supervised manifold learning based multigraph fusion for high-resolution remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 2, pp. 464–468, Feb. 2014.

[45] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. X. Zhu, "Joint and progressive subspace analysis (JPSA) with spatial-spectral manifold alignment for semisupervised hyperspectral dimensionality reduction," *IEEE Trans. Cybern.*, vol. 51, no. 7, pp. 3602–3615, Jul. 2021.

[46] G. French, T. Aila, S. Laine, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, high-dimensional perturbations," 2019, *arXiv:1906.01916*.

[47] J. Kim, J. Jang, and H. Park, "Structured consistency loss for semi-supervised semantic segmentation," 2020, *arXiv:2001.04647*.

[48] 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.

[49] J. Peng, G. Estrada, M. Pedersoli, and C. Desrosiers, "Deep co-training for semi-supervised image segmentation," *Pattern Recognit.*, vol. 107, 2020, Art. no. 107269.

[50] Y. Zou *et al.*, "Pseudoseg: Designing pseudo labels for semantic segmentation," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[51] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021.

[52] S. Liu, S. Zhi, E. Johns, and A. J. Davison, "Bootstrapping semantic segmentation with regional contrast," 2021, *arXiv:2104.04465*.

[53] 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.*, 2018, pp. 801–818.

[54] P. Khosla *et al.*, "Supervised contrastive learning," *Adv. Neural Informat. Process. Syst.*, vol. 33, 2020.

[55] L. Melas-Kyriazi and A. K. Manrai, "PixMatch: Unsupervised domain adaptation via pixelwise consistency training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021.

[56] W. Chen, Z. Jiang, Z. Wang, K. Cui, and X. Qian, "Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8924–8933.
[57] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 8026–8037, 2019.
[58] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2015, pp. 234–241.
[59] G. Pasquali, G. C. Iannelli, and F. Dell'Acqua, "Building footprint extraction from multispectral, spaceborne earth observation datasets using a structurally optimized U-Net convolutional neural network," *Remote Sens.*, vol. 11, no. 23, 2019, Art. no. 2803.
[60] N. He, L. Fang, and A. Plaza, "Hybrid first and second order attention Unet for building segmentation in remote sensing images," *Sci. China Inf. Sci.*, vol. 63, no. 4, 2020, Art. no. 140305.
[61] B. Bischke, P. Helber, J. Folz, D. Borth and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1480–1484.
[62] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[63] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, "Deep metric learning based on scalable neighborhood components for remote sensing scene characterization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8905–8918, Dec. 2020.

Jian Kang (Member, IEEE) received the B.S. and M.E. degrees in electronic engineering from the Harbin Institute of Technology, Harbin, China, in 2013 and 2015, respectively, and the Dr.-Ing. degree in signal processing in earth observation from Technical University of Munich, Munich, Germany, in 2019.

In August 2018, he was a Guest Researcher with the Institute of Computer Graphics and Vision, Graz University of Technology, Graz, Austria. From 2019 to 2020, he was with the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Berlin, Germany. He is currently with the School of Electronic and Information Engineering, Soochow University, Suzhou, China. His research interests include signal processing and machine learning techniques and their applications in remote sensing. In particular, he is interested in multidimensional data analysis, geophysical parameter estimation based on interferometric synthetic aperture radar data, synthetic aperture radar denoising, and deep-learning-based techniques for remote sensing image analysis.

Dr. Kang received the first place of the Best Student Paper Award at 2018 European Conference on Synthetic Aperture Radar, Aachen, Germany. His joint work was selected as one of the ten Student Paper Competition Finalists at 2020 IEEE International Symposium on Geoscience and Remote Sensing.

Zhirui Wang received the B.Sc. degree from the Harbin Institute of Technology, Harbin, China, in 2013, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2018, both in electronic engineering.

He is currently an Assistant Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. His research interests include synthetic aperture radar terrain classification and synthetic aperture radar target detection and recognition.

Ruoxin Zhu received the B.Sc. and M.Sc. degrees in photogrammetry and remote sensing from the Zhengzhou Institute of Surveying and Mapping, Zhengzhou, China, in 2013 and 2016, respectively.

He is currently with the State Key Laboratory of Geo-Information Engineering, Xi'an Research Institute of Surveying and Mapping, Xi'an, China.

Xian Sun (Senior Member, IEEE) received the B.Sc. degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2009, all in electronic engineering.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

Ruben Fernandez-Beltran (Senior Member, IEEE) received the B.Sc. degree in computer science, the M.Sc. degree in intelligent systems, and the Ph.D. degree in computer science from Universitat Jaume I, Castellón de la Plana, Spain, in 2007, 2011, and 2016, respectively.

He is currently a Postdoctoral Researcher with the Computer Vision Group, Universitat Jaume I, where he is a member of the Institute of New Imaging Technologies. He is a Visiting Researcher with the University of Bristol, Bristol, U.K., University of Cáceres, Cáceres, Spain, and Technische Universität Berlin, Berlin, Germany. His research interests include multimedia retrieval, spatiospectral image analysis, and pattern recognition techniques applied to image processing and remote sensing.

Dr. Fernandez-Beltran is a Member of the Spanish Association for Pattern Recognition and Image Analysis, which is part of the International Association for Pattern Recognition. He was awarded with the Outstanding Ph.D. Dissertation Award at Universitat Jaume I in 2017.

Antonio Plaza (Fellow, IEEE) received the M.Sc. and Ph.D. degrees in computer engineering from the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, University of Extremadura, Cáceres, Spain, in 1999 and 2002, respectively.

He is currently the Head of the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, University of Extremadura. He has authored more than 600 publications, including around 300 JCR journal articles (over 170 in IEEE journals), 23 book chapters, and around 300 peer-reviewed conference proceeding papers. He has reviewed more than 500 manuscripts for more than 50 different journals. His research interests include hyperspectral data processing and parallel computing of remote sensing data.

Dr. Plaza was a Member of the Editorial Board of IEEE GEOSCIENCE AND REMOTE SENSING NEWSLETTER from 2011 to 2012 and *IEEE Geoscience and Remote Sensing Magazine* in 2013. He was also a Member of the Steering Committee of IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING. He received the recognition as a Best Reviewer of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2009 and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING in 2010. He was also a recipient of the Most Highly Cited Paper (2005–2010) in *Journal of Parallel and Distributed Computing*, the 2013 Best Paper Award of IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and the Best Column Award of *IEEE Signal Processing Magazine* in 2015. He received best paper awards at IEEE International Conference on Space Technology and IEEE Symposium on Signal Processing and Information Technology. He was the Director of Education Activities for the IEEE Geoscience and Remote Sensing Society (GRSS) from 2011 to 2012 and the President of the Spanish Chapter of the IEEE GRSS from 2012 to 2016. He was as an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING from 2007 to 2012. He was the Editor-in-Chief for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING from 2013 to 2017. He has guest edited ten special issues on hyperspectral remote sensing for different journals. He is an Associate Editor for IEEE ACCESS (received the recognition as an Outstanding Associate Editor of the journal in 2017). He is the Editor-in-Chief for the IEEE JOURNAL ON MINIATURIZATION FOR AIR AND SPACE SYSTEMS. He has been included in the Highly Cited Researchers list from Clarivate Analytics in 2018–2020.