

AIO2: Online Correction of Object Labels for Deep Learning With Incomplete Annotation in Remote Sensing Image Segmentation

Chenyang Liu¹, Student Member, IEEE, Conrad M. Albrecht², Member, IEEE, Yi Wang³, Student Member, IEEE, Qingyu Li, and Xiao Xiang Zhu⁴, Fellow, IEEE

Abstract—While the volume of remote sensing (RS) data is increasing daily, deep learning in Earth observation (EO) faces lack of accurate annotations for supervised optimization. Crowdsourcing projects such as openstreetmap (OSM) distribute the annotation load to their community. However, such annotation inevitably generates noise due to insufficient control of the label quality, lack of annotators, frequent changes of the Earth’s surface as a result of natural disasters, and urban development, among many other factors. We present *adaptively triggered online object-wise label correction* (AIO2) to address annotation noise induced by incomplete label sets. AIO2 features an *adaptive correction trigger* (ACT) module that avoids label correction when the model training under- or overfits, and an *online object-wise label correction* (O2C) methodology that employs spatial information for automated label modification. AIO2 utilizes a mean teacher model to enhance training robustness with noisy labels to both stabilize the training accuracy curve for fitting in ACT and provide pseudo labels for correction in O2C. Moreover, O2C is implemented *online* without the need to store updated labels every training epoch. We validate our approach on two building footprint segmentation datasets with different spatial resolutions. Experimental results with varying degrees of building label noise demonstrate the robustness of AIO2. Source code will be available at <https://github.com/zhu-xlab/AIO2.git>.

Index Terms—Building detection, curriculum learning, deep learning, early learning, label correction, memorization effects, noisy labels, remote sensing (RS), semantic segmentation.

I. INTRODUCTION

DEEP learning has become a powerful tool of big data mining in Earth observation (EO) [1]. However, supervised deep learning methods are notorious data-hungry, requiring large amounts of high-quality labeled data to avoid overfitting. Despite the abundance of remote sensing (RS) images, obtaining accurately annotated labels poses a significant challenge due to the expensive, laborious, and time-consuming nature of the annotation process, which often involves domain experts and field surveys.

Nevertheless, there are many sources of labels from which we can easily obtain large amounts of labeled data with minimal efforts. For instance, Volunteered Geographic Information sources like openstreetmap (OSM) collect label information from individuals in a volunteer capacity and make it freely available [2]. Another approach is to design automatic labeling tools, such as *AutoGeoLabel* [3], to generate labels rapidly for RS images from high-quality data sources, e.g., light detection and ranging (LiDAR) data. Additionally, various land use land cover products, including Google’s Dynamic World [4], ESA’s World Cover [5], and Esri’s Land Cover [6], offer rich information for EO. Nevertheless, these label sources often result in unreliable labels, e.g., noisy labels due to insufficient human annotation. For example, [7] documents human uncertainty for the classification of local climate zones. As reported in [8], deep learning models are known for their large number of parameters and capability of learning complex functions, yet vulnerability to label noise. This also applies to segmentation tasks [9]. Therefore, these readily available labels require special considerations when applied to real-world scenarios. Beyond model training, noisy labels may significantly affect the evaluation of methodologies as well [10].

While learning from noisy labels (LNLs) has been extensively studied for image classification tasks, few approaches have been developed for image segmentation tasks. Existing LNL methods for segmentation tasks mainly borrow ideas from LNL for classification and semi-supervised segmentation

Manuscript received 7 September 2023; revised 14 January 2024; accepted 29 February 2024. Date of publication 5 March 2024; date of current version 21 March 2024. This work was supported by the Helmholtz Association’s Initiative and Networking Fund on the HAICORE@FZJ Partition. The work of Chenyang Liu, Conrad M. Albrecht, and Yi Wang was supported by the Helmholtz Association through the Framework of HelmholtzAI, Local Unit Munich Unit @Aeronautics, Space and Transport (MASTr), under Grant ZT-I-PF-5-01. The work of Qingyu Li and Xiao Xiang Zhu was supported in part by the Excellence Strategy of the Federal Government and the Länder through the Technical University of Munich Innovation Network EarthCare; and in part by the German Federal Ministry of Education and Research (BMBF) in the Framework of the International Future AI Laboratory “AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond” under Grant 01DD20001. (Corresponding author: Chenyang Liu.)

Chenyang Liu and Yi Wang are with the Chair of Data Science in Earth Observation, Technical University of Munich (TUM), 80333 Munich, Germany, and also with the Remote Sensing Technology Institute, German Aerospace Center (DLR), 82234 Weßling, Germany (e-mail: chenyang.liu@dlr.de; chenyang.liu@tum.de; Yi.Wang@dlr.de; yi4.wang@tum.de).

Conrad M. Albrecht is with the Remote Sensing Technology Institute, German Aerospace Center (DLR), 82234 Weßling, Germany (e-mail: Conrad.Albrecht@dlr.de).

Qingyu Li and Xiao Xiang Zhu are with the Chair of Data Science in Earth Observation, Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: qingyu.li@tum.de; xiaoxiang.zhu@tum.de).

Digital Object Identifier 10.1109/TGRS.2024.3373908

methods. In the former case from classification tasks, a set of regularization techniques such as consistency regularization [11] or entropy minimization [12] is used to constrain the optimization space. Nevertheless, label noise behaves differently in these two types of tasks. In classification tasks, the entire image is treated as a single sample unit and can be considered to have approximately similar levels of uncertainty. Thus, random flipping can be used to simulate label noise for classification tasks. In contrast, the sample unit in segmentation tasks is a pixel, and neighboring pixels are interconnected through spatial dependencies [13]. As a result, pixels located near boundaries are more difficult to define. From this perspective, we can classify pixel-wise label noise into two categories: assignment noise and shape noise. Assignment noise occurs when objects are labeled incorrectly, while shape noise refers to inexact object delineation caused by such phenomena as coarse annotations. In practice, inaccurate co-registration of image-mask pairs is another common source of label noise, mainly leading to shape noise with misaligned boundaries [14]. Generally, assignment noise incurs more severe damage on model training than shape noise does. This difference is illustrated in Section III-B. Moreover, LNL for natural image segmentation is usually studied in the context of weakly supervised learning, where pixel-wise noisy labels are derived with image-level annotations by Grad-CAM and its variants from object-centric images [15], [16]. Thus, the primary label noise is the shape noise, while RS applications usually face more complex noise types due to different image characteristics and more diverse noisy label sources.

As regards the ideas borrowed from the semi-supervised learning domain, self-training methods are naturally related to the noisy label problem, where pseudo labels generated by the classifier itself inevitably incur some inaccurate assignments [17]. Following this paradigm, [18], [19], [20] correct possibly wrong labels in the training set by means of high-confidence or low-uncertainty predictions. To make these methods effective, the questions of when and how to correct the labels should be considered. In semi-supervised scenarios, only a small part of the accurately labeled patches is available at the beginning. The training set is gradually expanded via adding pseudo labels as training continues, during which the impact of bad pseudo labels can be offset to some extent by the advantages brought by training size expansion. LNL settings do not confer this advantage since the classifier originally has access to a large number of labels that are not as accurate as expected. Therefore, manually setting the warm-up length as in [19] can easily lead to an early or late start of correction, risking correction effectiveness degradation when model predictions are not reliable enough. Liu et al. [21] propose an adaptive method for label correction initialization in which the training accuracy curve is fit on an exponential function and the change in its gradients is monitored. While promising, this method has a sensitive threshold setting to noise rates, and the fluctuation of accuracy curves makes the detection results unstable. In terms of how to correct, current correction criteria usually take softmax/sigmoid outputs as confidence indicators [20]. The threshold is either predefined by users or

flexibly adjusted in an image-wise fashion [19], more or less ignoring the spatial dependencies among pixels. One further possibility is to determine data and model uncertainty, e.g., via Bayesian Neural Networks, for sample selection [22]. Yet, a major challenge is the lack of ground truth to evaluate the estimated data uncertainty when developing such methods to address real-world problems.

In this work, we study building footprint identification from aerial imagery to develop a novel methodology to handle, among other types of noise, incomplete label noise, in which a given set of building outlines is known to miss a subset of existing buildings (false negative), but annotated buildings are assumed to have accurate outlines. Existing research on quality assessment of OSM building data has found that in most areas position accuracy is comparable to cadastral maps, while completeness is relatively low, and that it varies worldwide [23], [24], [25]. As mentioned above, assignment noise potentially imposes a more significant negative impact on model training than shape noise. Thus, we focus our study on incomplete label noise as a first step toward a systematic solution to using OSM labels for model training. Our approach significantly differs from semi-supervised scenarios, where all the labeled patches are carefully annotated, with all the objects marked. For us, all the patches are labeled with objects dropped from the ground truth by annotation as background. Thus, we call it “incomplete label noise.”

Based on the aforementioned issues, we propose a new method called adaptively triggered online object-wise label correction (AIO2). This approach consists of two main components, an adaptive correction trigger (ACT) module and an online object-wise label correction (O2C) module, to address the “when” and “how” questions in the self-cleansing process without human interference. In short, AIO2 adopts ACT to automatically trigger the O2C by monitoring the dynamics of training accuracy curves measured by numerical gradients. Specifically, our framework incorporates a mean teacher model [26] originally designed for semi-supervised learning. The teacher model is updated by exponentially averaging historical model weights, thus leading to a minimal extra computational burden without backpropagation. In turn, the training accuracy curves by the teacher model are smoother, which can reduce the negative effects of fluctuations on early learning detection results. Moreover, we partially decouple the online label correction process and the model training by utilizing the predictions from teacher models as pseudo labels, with which we design an object-wise correction module for label cleansing. The main contributions of our work are as follows.

- 1) We introduce a new label correction method termed AIO2 for segmentation tasks with incomplete label noise, which is less sensitive to parameter settings and more compatible with spatial characteristics of pixels.
- 2) We analyze in detail the memorization effects in segmentation tasks as a basis for our methodology design. The resulting insights have served as valuable input for future extensions of noisy label training programs.
- 3) We present two new modules, namely, ACT and O2C, which are particularly designed for segmentation tasks

to solve the “when” and “how” problems in the self-cleansing process without human interference.

In a nutshell, our methodology exploits the spatial context of pixels to explore memorization effects in pixel-level segmentation tasks. The high-level objective of semantic segmentation from RS modalities is the generation of map data. *Vectorizing* rasterized segmentation maps involves grouping pixels into single identities such as buildings, and other geospatial “objects.” Our work devises strategies for the analysis and adjustment of geospatial image semantic segmentation tasks at the object level, such as evaluation of training performance, label correction, and uncertainty assignment of pixels based on relative position (object “boundary” versus “bulk”).

The article is organized as follows: Section II summarizes related studies on LNL with deep learning models. Next, the memorization effects of noisy labels in segmentation tasks along with technical details of the proposed AIO2 method are described in Section III. We elaborate the experimental results in Section IV, and conclude this work with some discussions and future lines in Section V.

II. RELATED WORKS

In this section, we review some recent advances in LNL with deep learning models for image classification and segmentation tasks, with a special focus on the RS domain.

A. LNL for Image Classification Tasks

The problem of noisy labels is well-investigated in the image classification field. One promising label source is web crawling: it is easy and cheap to obtain a large amount of labeled data, although it is somewhat unreliable [27]. To reduce the negative effects of label noise on model training, existing studies seek to find solutions using three main approaches: robust architecture modification [28], [29], [30], label cleansing [31], [32], [33], [34], [35], [36], [37], [38], and robust loss function design [26], [39], [40], [41], [42], [43], [44]. A comprehensive review of LNL methods for classification is provided in [45]. Each of the three methods mentioned above is described briefly below.

A key concept of robust architecture modification is to add a label transition layer on top of a softmax layer of base deep neural networks in order to explicitly transfer the hidden “true” labels to their noisy versions for training [28]. In the test phase, the transition layer is removed to enclose “clean” predictions. It can be modeled in a feature-independent fashion [29] or a feature-conditional way [30]. Label cleansing is more straightforward than the other two method types, employing sample selection or correction. Incorrectly assigned labels can be recognized according to high uncertainty quantified by loss [31] or softmax outputs [32]. Some works also leverage the discrepancy between simultaneously trained deep neural networks such as Decouple [33], MentorNet [34], Co-teaching [35], [36], and DivideMix [37]. Another reframes the noisy label problem as a domain shift one, and separates clean and noisy labels with the aid of data augmentation [38]. Robust loss function design is probably the most popular of the LNL methods, partially due to its flexibility and its theoretical basis

in risk minimization [46], [47]. Some state-of-the-art methods include a consistency constraint between teacher and student model predictions in a self-ensembling framework [26], early-learning regularization by integrating historical predictions to combat the memorization phenomenon [39], compatibility loss between corrected label distributions and original noisy labels to avoid crazy deviation of corrections from original labels [40], bootstrapping using the convex combination of original noisy labels and predictions to prevent direct fitting on the noise distribution [41], and so on [42], [43], [44].

Within the RS community, the typical type of image classification task is scene classification, in which the noisy label problem is studied from the single-label and multi-label aspects. In single-label cases, some ideas are borrowed directly from the computer vision domain, such as using co-teaching for sample selection [48], and smooth loss to constrain the optimization process [49]. Specifically for RS data, Damodaran et al. [50] propose an entropic optimal transport loss inherently exploiting the geometric structure of the underlying data. Other methods are designed from the feature learning perspective, either doing sample cleansing [51], [52] or modifying the loss function to be noise-robust [53], [54]. In terms of multiclass scene classification with noisy labels, only a few works address the problem mainly employing loss design [55], [56] and sample selection [57].

B. LNL for Image Segmentation Tasks

Unlike the extensive research in image classification, related LNL works for segmentation are relatively rare in the computer vision domain, which is generally studied along with weakly supervised methods using pixel-wise labels derived from activation maps guided by image-level annotations as noisy labels [15], [16]. These labels for object-centric images are primarily contaminated by shape noise. On the contrary, noisy labels are more frequently encountered in RS image segmentation tasks, where pixel-wise annotations are more difficult and ambiguous in combination with the clustered background, and thus require more expertise. Furthermore, there are more noisy label sources for RS image segmentation, such as OSM and various land use land cover products. To combat noisy labels, some works have been designed under the assumption that a small number of clean labels are available during training [58], [59], [60]. However, this is not the case in many real scenarios. We thus concentrate on methods without usage of clean labels in the following review.

Inspired by LNL methods for classification, [61], [62] appended a probabilistic model to ordinary deep neural networks in order to capture the relationship between noisy labels and their latent true counterparts for road and building extraction. Nevertheless, more common and effective solutions employ robust loss functions such as bootstrapping [63], consistency constraints [11], [64], and loss reweighting with weights of each sample estimated by an attention mechanism [65] or reliability [16]. These methods, though effective to some extent, are sensitive to parameter setting, and sometimes unable to generalize well due to the long-tail distribution problem in segmentation tasks. In addition, Albrecht et al. [66]

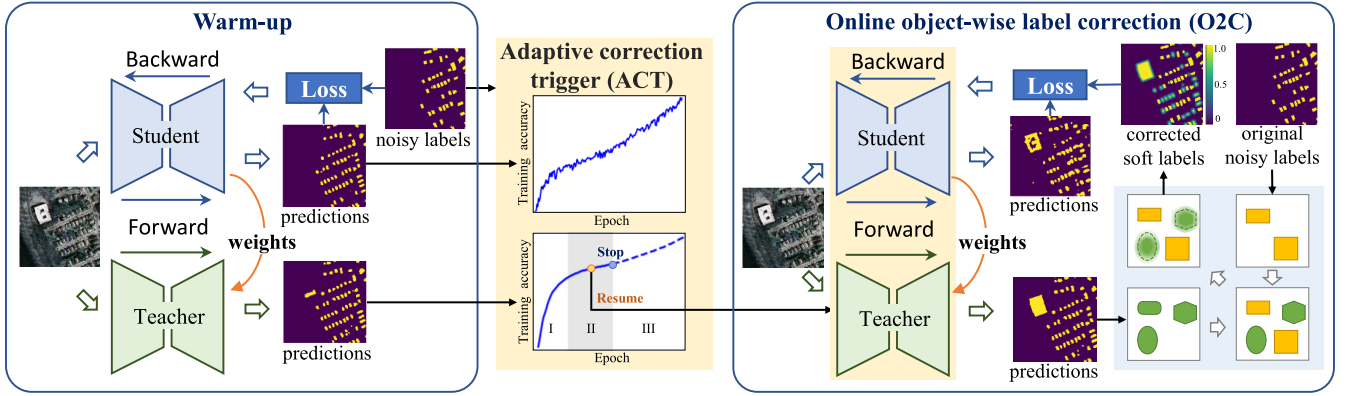


Fig. 1. Flowchart of the proposed two-stage AIO2 method for object-level incomplete label sets. Model training is initially conducted using the given noisy labels, where ACT actively monitors the training dynamics to determine when to trigger O2C for label correction.

exploit a CycleGAN for iterative addition of missing labels from style-translation of aerial images into rasterized OSM scenes. This approach incorporates spatial correlations and geographic context of human infrastructure such as roads, buildings, and parks. On the other hand, encouraged by the connection of the noisy label problem with semi-supervised learning, confidence/uncertainty-based pixel-wise sample selection or correction is widely used, in which model performance is highly dependent on the predefined threshold setting [20]. To alleviate this drawback, Dong et al. [19] involve a patch-based threshold adjustment technique that can partially release the dependency on manual threshold setting. They also combine it with regularization constraining on original noisy labels, a popular strategy aiming to reduce the negative effects of mistaken corrected labels [18]. Besides, Sun et al. [67] utilized mutual teaching with two structurally identical models to update noisy pseudo labels for hyperspectral image change detection. The aforementioned approaches for RS image segmentation with noisy labels primarily rely on pixel-wise correction, with some adjustments to enhance adaptability to RS images. However, this pixel-wise constraint neglects crucial spatial information shared by neighboring pixels, a key factor in segmentation tasks. Additionally, these methods require a manually set warm-up stage, introducing instability. In a recent work, Liu et al. [21] proposed an adaptive early learning detection for medical image segmentation with noisy labels. While promising, its application to RS images proved unstable due to sensitive hyperparameter settings and susceptibility to accuracy curve fluctuations. In response, our proposed method, AIO2, is developed to achieve more robust early learning detection and enhance the effectiveness of sample correction with spatial information.

III. METHODOLOGY

Fig. 1 presents an overview of the proposed AIO2 method. Analogous to other correction-based methods, AIO2 is initialized from a warm-up stage, taking given noisy labels as reference data to train the network. The ACT module at the same time ceases the training when the model starts overfitting to noisy labels. Both the student and teacher models are then reloaded from a previous checkpoint according to

the refined detection result. Thereafter, training is resumed with O2C coming into force. In this procedure, a teacher model is introduced whose weights are updated by exponential moving average (EMA) on historical weights of the student model. The teacher model on the one hand provides more smooth training accuracy curves for the ACT module to automatically terminate the warm-up phase and simultaneously trigger O2C for label cleansing, and on the other hand provide pseudo labels for O2C, and thus is able to partly decouple the label correction process from model training. In the following, Section III-A first gives a brief description of the mean teacher model. Some insights on memorization effects are discussed in Section III-B, followed by technical details of the ACT and O2C modules in Sections III-C and III-D, respectively.

A. Mean Teacher Model

Temporal ensembling was first introduced into semi-supervised learning domain by implementing an EMA of historical successive predictions on each training example [68]. Encouraged by its success, mean teacher modeling was developed by applying EMA on model weights instead of predictions. The effectiveness of mean teacher modeling to combat label noise was evaluated on classification tasks [26]. Doing so conferred two obvious advantages: better model stability, and a decreased storage and computational burden. In our work, we found that the stability of the mean teacher model is beneficial to both of our newly designed modules.

Let $\theta_n^{(s)}$ denote the parameters of the student model at the n th iteration updated in a regular model training approach through backpropagation

$$\theta_n^{(s)} = \theta_{n-1}^{(s)} - \eta \nabla \mathcal{L}(\theta_{n-1}^{(s)}) \quad (1)$$

with η as the learning rate, and $\nabla \mathcal{L}(\cdot)$ the gradients of loss function with respect to each parameter. After the update of $\theta_n^{(s)}$, the counterpart of teacher model $\theta_n^{(t)}$ can be derived via EMA by

$$\theta_n^{(t)} = \begin{cases} \theta_n^{(s)}, & n = 0 \\ \alpha \theta_{n-1}^{(t)} + (1 - \alpha) \theta_n^{(s)}, & n > 0 \end{cases} \quad (2)$$

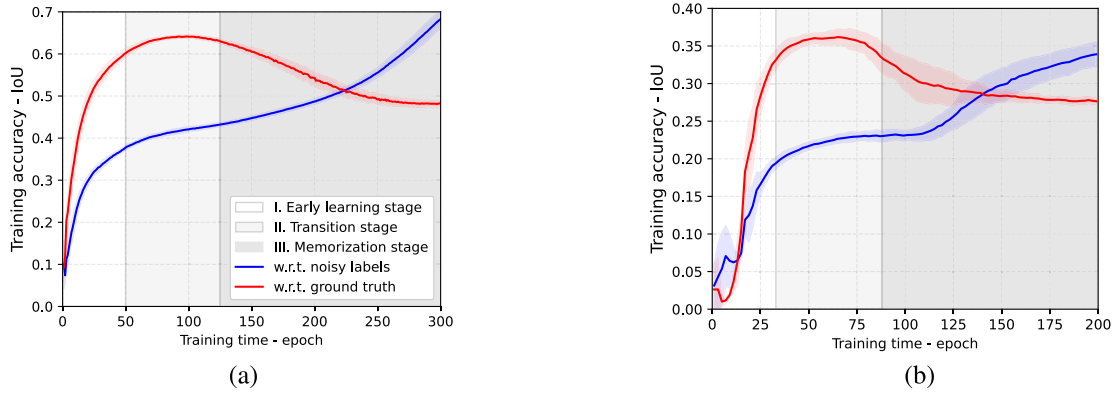


Fig. 2. Three-stage training without special considerations for label noise (colors online). Training accuracies of teacher models obtained with incomplete noisy labels of a drop rate of 0.5 on (a) Massachusetts dataset and (b) Germany dataset. Note: For real-world scenarios, training accuracies (blue) needs to be based on noisy labels. Ground-truth (GT) label accuracies (red) are presented for reference only. For this figure and all those that follow, statistically fluctuating accuracy curves have been smoothed, with the solid line indicating the mean value and the shaded, semitransparent region marking the 1σ -area.

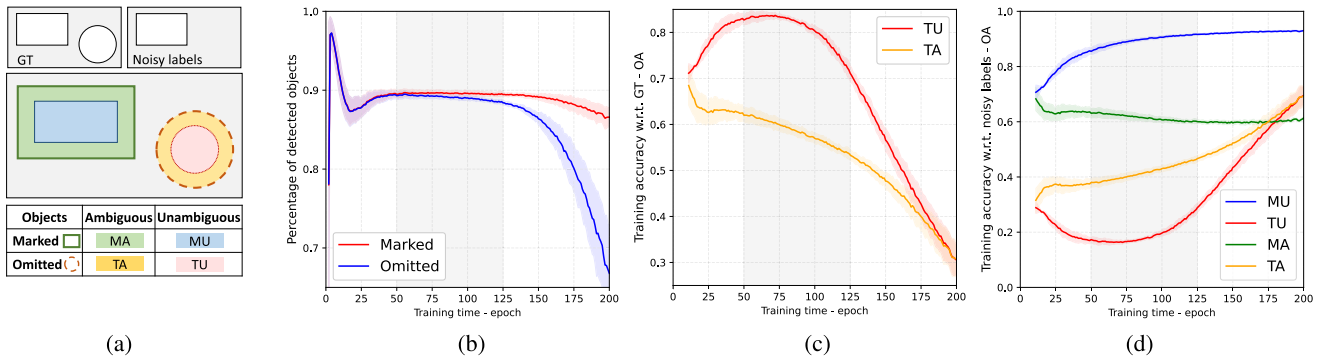


Fig. 3. Numerical exploration of memorization effects on noisy labels for segmentation tasks (colors online). We statistically analyze the model training of the (teacher) model on the Massachusetts dataset at a drop rate of 0.5. From the object-wise perspective, we divide all the objects in GT masks into Marked (solid green rectangle) and Omitted (dashed orange circle), and report their (b) Detection rates. An object is rated *detected* when it is at least partially predicted. From the pixel-wise perspective, we split all the object pixels into four groups as shown in (a), and calculate their overall accuracies (OAs) with respect to (c) GT labels and (d) noisy labels. Gray background shadows highlight the transition phase.

where $\theta_0^{(s)}$ are the randomly initialized models, and α is the smoothing coefficient hyperparameter empirically set as 0.999 [26].

B. Memorization Effects

Memorization effects were first reported on image classification tasks [8], [69], implying a two-stage training with noisy labels. More precisely, in the first *early-learning* stage, model performance is continuously improved by dominant learning from most of the accurately labeled samples, while in the later *memorization* stage, model performance begins to be degraded for overfitting to label noise information. A similar phenomenon has also been observed in segmentation tasks [9], [21]. Unlike the original elucidation of memorization effects, we reinterpret this phenomenon as a three-stage training with noisy labels, adding a transition stage between the early-learning and memorization stages [see Fig. 2]. In the following, we inspect this phenomenon at both pixel and object levels.

We first quantify the memorization effects from the object-wise perspective. We classify all objects/building footprints in the training set as either.

- 1) *Marked (M)*: Identified as per the incomplete set, true positive [rectangle in Fig. 3(a)].
- 2) *Omitted (T)*: Set of labels to complete the set, false negatives from the perspective of the incomplete set [circles in Fig. 3(a)].

Note: As per the definition of *incomplete labeling* false positives are not considered. Fig. 3(b) presents the detection rates. As the plot indicates, the transformation from early learning to memorization is smooth. After an initial unstable, oscillating warm-up phase, the model performance plateaus (transition stage), before the detection rate starts to sharply drop (memorization stage).

At the pixel level, we visualize the memorization phenomenon inspired by [21], additionally considering the spatial correlation of pixels. Our grouping strategy is illustrated in Fig. 3(a). The pixels P of a single object O ($P \in O$) are broken down into two categories.

- 1) *Ambiguous (A)*, P sufficiently close to the boundary ∂O of O , $d(P, \partial O) < D$ for some maximum distance D and distance function d such as the Hausdorff metric [70].
- 2) *Unambiguous (U)*, otherwise.

Given the previous object-level definitions of *marked* and *omitted*, all object pixels are categorized into one of the

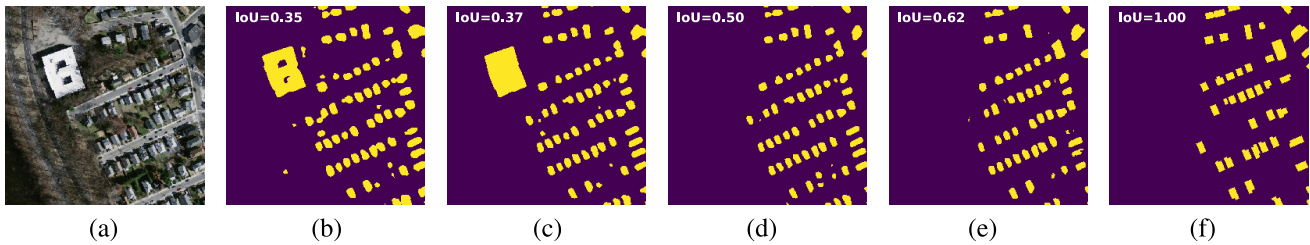


Fig. 4. Example demonstrating the three-stage training. (a) Optical image. (b)–(e) Predictions from the (b) early-learning, (c) transition, and (d) and (e) memorization stages obtained after 30, 75, 150, and 300 training epochs, respectively. We list the IoU with respect to (f) noisy labels at the upper left corner.

four groups: *marked-ambiguous* (MA), *marked-unambiguous* (MU), *omitted-ambiguous* (TA), and *omitted-unambiguous* (TU). Fig. 3(c) and (d) presents the evolution of the OA over the course of model training. Note that the OAs of MA and MU are the same no matter what reference data are used. Therefore, they are only presented in Fig. 3(d). Due to memorization effects, we observe a notable bias as training progresses, i.e., while the OAs with respect to GT labels of TU and TA decrease in Fig. 3(c), their counterpart with respect to noisy labels increase in Fig. 3(d). Yet, the memorization of TU and TA is out of sync. Noise memorization for TA takes place immediately upon the start of model training. In contrast, TU pixels stay unaffected to about epoch 75, as shown in Fig. 3(c), before overfitting reduces the OA with respect to the GT. As illustrated in Fig. 4, model training is mainly about structure information in the first early learning stage [see Fig. 4(b)]. As a result, the model can extract most of the instances in the scene at the second transition stage [see Fig. 4(c)]. However, in the memorization stage, the model overfits to label noise at a rapid pace, and learns to drop GT building footprints missing in the incomplete labels due to dominant learning of TU samples [see Fig. 4(d) and (e)]. Consequently, the OA with respect to GT drops while the OA with respect to noisy labels ramps up. On the other hand, it seems that the boundary regions of building footprints are more vulnerable to overfitting. Indeed, properly defining the exact outline of a building from aerial imagery is a challenging task, even for humans. For example, do annotations follow subtle details of the building’s facade, or is the whole footprint approximated by a simplified rectangle? Are open courtyards (with green areas) considered part of the building footprint? We observe that overfitting reflects most notably in TA pixels, resulting in coarse approximation of building footprint outlines before memorization affects the inner core of objects semantically segmented [see Fig. 4(b) and (c)].

To summarize, as illustrated in Fig. 2, on both datasets, in the first early-learning stage, dominant learning of MU leads to rapid increase of training accuracies with respect to both GT and noisy labels. Then, in the second transition stage, the learning of MU is close to an end—it is mainly TA samples that are being memorized. The training arrives at a plateau where both kinds of training accuracies keep relatively stable. In the final memorization stage, the training accuracies with respect to noisy labels again start to increase rapidly while training accuracies with respect to GT drop severely, largely due to the ultimate memorization of TU samples. In practice, the number of epochs representing different stages would vary

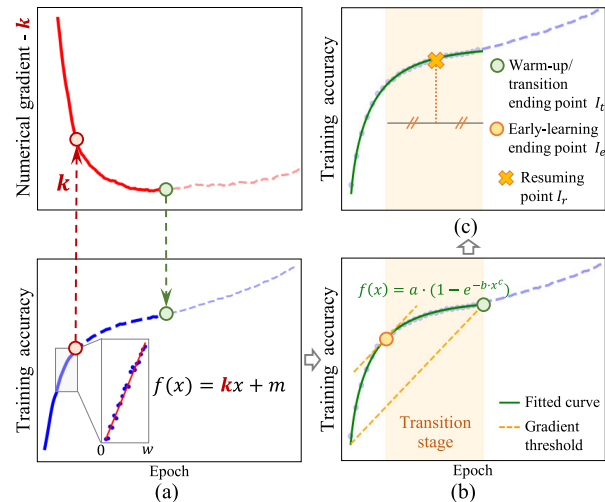


Fig. 5. ACT module. Three-stage strategy to identify “when” to start label correction (orange \times), where the nonnegative number w in (a) determines the window size of epochs to numerically estimate k , the blue faded dashed line (—) indicates trends in the training accuracy obtained without the application of our ACT module. (a) Numerical gradient estimation, (b) transition stage detection, and (c) resuming.

when the degree of incompleteness, the image resolution, as well as the tasks (e.g., road detection and urban green space detection) change as indicated in Fig. 2(a) and (b).

In Sections III-C and III-D, based on the analysis of the three-stage training, we introduce the ACT and O2C modules to solve the “when” and “how” questions in our label correction pipeline.

C. ACT Module

As stated above, the model reaches the highest potential in the transition stage when directly trained with noisy labels. A natural idea to solve the “when” question is to initiate the correction procedure in this stage, using the most reliable predictions. As shown in Fig. 2, the training accuracy (with respect to noisy labels) increases much faster both before and after the transition stage, which provides us the potential to monitor the growth rate of the training accuracy curve for detection.

Fig. 5 showcases the detection procedure of the ACT module following this idea. After an initial warm-up and slow-down in training accuracy, a saddle point in training accuracy toward reacceleration is a hallmark of data overfitting/memorization taking effect (green \circ), which marks the ending of the transition stage (yellow shaded area) of learning structural information for semantic segmentation. We define

the beginning of the transition stage (orange \circ) as the epoch I_e , where the rate of training accuracy increase matches the overall rate of training accuracy increase from epoch zero to I_t . The resuming point from which the label correction is triggered is finally taken as the middle point of the transition stage.

To numerically determine the gradient of the training accuracy over epochs, we apply local linear fitting over w epochs based on the outputs of the teacher network. This approach reduces random fluctuations in the estimation of the training accuracy's derivative of the student model. Given the sliding window size w , the numerical gradient at the i th epoch is estimated by fitting the $(i - w)$ th to the i th training accuracy data points $\{(x, y)\}_w = \{(1, f_{i-w}), \dots, (w, f_i)\}$ to a linear function

$$f(x) = k_i x + m_i = y \quad (3)$$

where the slope parameter k_i represents the growth rate of training accuracy at each epoch i .

The second step is to determine the transition stage. We expect k to continuously decrease in the first two stages, while it ramps up again when entering the last memorization stage. So, we terminate the warm-up phase at the end of the transition stage when k starts to increase. Let z be the *look-ahead buffer zone size* to determine whether k_i has hit its lowest numerical value. Then, the ending point of warm-up/transition stage I_t is detected as

$$I_t = j \quad \text{when} \quad k_j = \min(k_j, \dots, k_{j+z}). \quad (4)$$

To improve the robustness of detection, we perform the analysis, (4), for a set of sliding window sizes $W = \{w\}$, i.e., we obtain the set $I = \{I_t^{(w)}\}_W$ over which we average accordingly

$$\langle I_t \rangle = \frac{1}{|W|} \left[\sum_{w \in W} I_t^{(w)} \right] \quad (5)$$

with $|W|$ the number of window sizes picked. Details on the choice of w values are documented in Section IV-A. The buffer parameter z is simply set to the mean $z = \langle w \rangle = \sum_{w \in W} w / |W|$.

Thereafter, to detect the ending point of the early-learning stage or the starting point of the transition stage I_e , we need a threshold to tell when the curve becomes sufficiently flat. A manually set threshold depends heavily on the quality of noisy labels, which is hard to fix in most scenarios. Alternatively, we propose that the slope between the first and I_t th epochs serves as the adaptive threshold (the orange lines in Fig. 5), that is, we compute the slope of the arc as

$$\sigma = \frac{f_{I_t} - f_1}{I_t}. \quad (6)$$

At this step, we turn to a different evaluation of the gradients compared to (3) above, now fitting a multiparameter curve, because:

- 1) We have sufficient data points from the warm-up phase to globally fit the exponentially saturating function (7), and
- 2) Its analytic gradients (8) are monotonically decreasing.

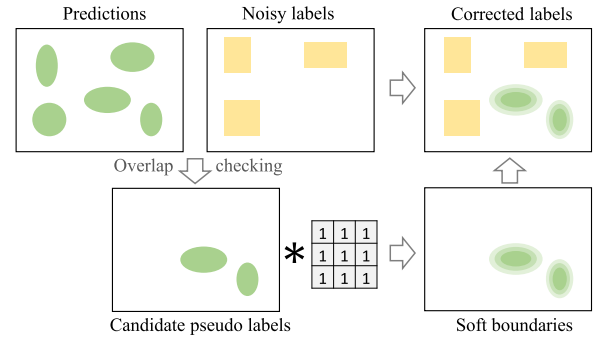


Fig. 6. O2C module. Spatial constraint is used to solve “how” to correct labels, where candidate pseudo labels are selected by filtering predictions (solid green ellipse) against noisy labels (yellow rectangles), and $*$ is the convolution operator to introduce uncertainty around boundary areas.

In this sense, we fit I_t training accuracy points $\{(x, y)\}_{I_t} = \{(1, f_1), \dots, (I_t, f_{I_t})\}$ to

$$f(x) = a(1 - \exp(-bx^c)) = y \quad (7)$$

with a , b , and c fitting parameters. Accordingly, we obtain

$$f'(x) = abc x^{c-1} \exp(-bx^c). \quad (8)$$

Specifically, $0 < a < 1$ corresponds to the magnitude of (7). We constrain $b > 0$ and $0 < c < 1$ to restrict (8) to a monotonically decreasing function. Correspondingly, employing the threshold, (6), we can explicitly count to the starting point of the transition phase as follows:

$$I_e = \sum_{i=1}^{I_t} \text{sgn}(f'(i) - \sigma) \quad (9)$$

where $\text{sgn}(\cdot)$ is the sign function with $\text{sgn}(x) = 1$ if $x > 0$, otherwise 0, and σ is the adaptive threshold defined in (6).

After determining the values of I_e and I_t , we finally resume the models at the middle of the transition stage that trigger O2C from the I_r th epoch with

$$I_r = \left\lfloor \frac{I_e + I_t}{2} \right\rfloor \quad (10)$$

which is expected to be close to the best-performed model in the warm-up phase. However, saving every checkpoint file consumes a lot of storage space. A trick here is to save a selection of checkpoints, e.g., every five checkpoints, and resume from the one closest to I_r .

D. Online Object-Wise Label Correction

As discussed in Section III-B and shown in Fig. 3, the memorization of noisy labels mainly takes place on TA samples around boundaries during the transition stage with a high object detection rate of model. Based on this observation, we design an O2C module as a substitute for the commonly used pixel-wise correction strategies.

Fig. 6 presents the workflow of O2C. One major improvement is the selection of pseudo label candidates in an object-wise fashion by checking the overlap between predictions and given noisy labels. We reserve the marked objects,

and do label correction only for those that are omitted. In addition, considering the memorization effects on TA samples, we apply a smooth filter to generate soft boundaries for candidate pseudo objects. For simplicity, we use an all-one filter in Fig. 6, which can also be replaced by other kinds of smooth filters, such as a Gaussian filter.

Additionally, “online” in the O2C name includes one-off label correction at each iteration without saving historical correction results. This is another major difference from commonly used pixel-wise correction strategies, which correct labels incrementally.

Algorithm 1 AIO2 Framework

Input: training set with noisy labels $D_T(\mathbf{x}, y)$, look-ahead buffer zone size z and a set of sliding window sizes $W = w$ for sample correction detection

Output: predicted segmentation masks

```

1: Initialization: randomly initialized student model  $\theta_0^{(s)}$ 
   and teacher model  $\theta_0^{(t)} = \theta_0^{(s)}$  with  $\theta_0^{(t)}$ .requires_grad=False,
   empty accuracy list of teacher model  $A$ , empty numerical
   gradient lists  $K^{(w)}$ , indicator for sample correction
    $S=False$ 
2: for  $i=1$  to Epoch do
3:   for  $j=1$  to Batch do
4:     //O2C for label correction
5:     if  $S$  then
6:        $\tilde{y} = \text{O2C}(y, \theta_{i-1}^{(t)}(\mathbf{x}))$ , cf. Section III-D
7:     else
8:        $\tilde{y} = y$ 
9:     end if  $S$ 
10:    //model updating
11:    update student model  $\theta_{i(j)}^{(s)}$  with  $(\mathbf{x}, \tilde{y})$ , cf. Eqs. (1),
      (11)
12:    update teacher model  $\theta_{i(j)}^{(t)}$  with  $\theta_{i(j)}^{(s)}$ , cf. Eq. (2)
13:  end for  $j$ 
14:  //ACT for label correction detection
15:  if not  $S$  then
16:    update  $A = A + [\text{IoU}(D_T, \theta_i^{(t)})]$ 
17:    update  $K^{(w)} = K^{(w)} + [\text{FIT}(w, i, A)]$  cf. Eq. (3)
18:    if  $k_{i-z}^{(w)}$  for  $w$  in  $W$  all meet Eq. (4) then
19:      get the resuming point  $I_r$ , cf. Eqs. (5)-(10)
20:      //resume from the  $I_r$ th epoch
      and trigger label correction
21:       $i \leftarrow i-z, \theta_i^{(s)} \leftarrow \theta_{i-z}^{(s)}, \theta_i^{(t)} \leftarrow \theta_{i-z}^{(t)}, C \leftarrow \text{True}$ 
22:    end if  $k_{i-z}^{(w)}$ 
23:  end if not  $S$ 
24: end for  $i$ 

```

E. AIO2 Framework

In summary, the proposed AIO2 framework employs a two-stage pipeline to train segmentation networks with incomplete noisy labels. In the initial warm-up stage, models are trained using the given noisy labels. Subsequently, in the second stage, O2C is applied for label correction before optimization. The transition between the two stages is automatically managed by ACT. It is crucial to notice that the

teacher model plays a predominant role in both the ACT and O2C modules, while solely the student model is directly trained with training data and gradient descent. We summarize the implementation details in Algorithm 1 for a more comprehensive understanding of AIO2. Specifically, within each epoch, the student model undergoes the optimization with the training set, using a combined loss of the distribution-based cross-entropy loss and the region-based dice loss [71], as follows:

$$L = L_{ce} + L_{dice} \quad (11)$$

$$L_{ce} = - \sum_{i=1}^N \sum_{j=1}^C y_{i,j} \log(p_{i,j}) \quad (12)$$

$$L_{dice} = 1 - \frac{2 \sum_{j=1}^C \sum_{i=1}^N y_{i,j} p_{i,j}}{\sum_{j=1}^C \sum_{i=1}^N (y_{i,j} + p_{i,j})} \quad (13)$$

where $y_{i,j}$ represents the one-hot label for the i th sample at class j , $p_{i,j}$ denotes the corresponding prediction from the softmax layer, N and C are the numbers of samples and classes. After that, the teacher model is updated by EMA with the new weights of the student model, as opposed to traditional gradient descent.

IV. EXPERIMENTS

In this section, we evaluate the effectiveness of our proposed AIO2 method on two building footprint extraction datasets with different spatial resolutions. We first give an overview of the two datasets and label noise injection strategy along with other settings in Section IV-A, followed by detailed experimental results, including ablation studies and parameter sensitivity analysis, and related discussions.

A. Datasets and Settings

1) *Datasets:* The **Massachusetts Dataset**¹ [72] is composed of 151 RGB aerial images collected over the City of Boston with a size of 1500×1500 pixels and a spatial resolution of 1 m. The corresponding building masks are also provided for each image. The whole dataset was randomly split into three subsets comprising a training set of 137 images, a test set of ten images, and a validation set of four images. We keep the original split, and further crop each image into a series of 256×256 small patches. After some images without labels are removed, the final datasets comprise 3065 patches for training, 250 for test, and 100 for validation.

The **Germany Dataset** [73] consists of 2052 image-label pairs with a size of 320×320 pixels generated across ten Germany cities including Bielefeld, Bochum, Bonn, Cologne, Dortmund, Duesseldorf, Duisburg, Essen, Muenster, and Wuppertal. The image data were from Planet basemap images with a lower spatial resolution of 3 m and three bands (RGB). The building masks were rasterized from vector cadastral data² to 3 m GSD to pair with image data. In our experiments, we randomly select 200 patches for test and 50 for validation.

¹Downloaded from <https://www.cs.toronto.edu/vmnh/data/>.

²Accessed via GEOportal.NRW (<https://www.geoportal>) on August 5, 2021.

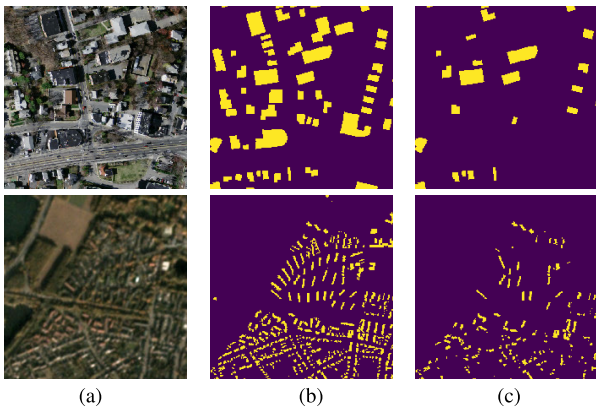


Fig. 7. Example of data triples for two datasets. (a) Optical images. (b) Accurate GT labels. (c) Simulated noisy labels. The first and second rows correspond to the Massachusetts (1 m) and Germany (3 m) datasets, respectively. Buildings are highlighted in yellow in (b) and (c).

TABLE I

QUALITY ASSESSMENT OF SYNTHETIC NOISY LABELS GIVEN DIFFERENT DISCARDING PERCENTAGES α_0 ON TWO DATASETS

α_0		0.3	0.5	0.7
Massachusetts	OR	29.48±0.29	49.40±0.86	69.46±0.29
	IoU	71.35±0.46	51.58±0.85	31.71±0.37
	OA	96.17±0.06	93.52±0.11	90.86±0.05
Germany	OR	29.75±0.58	49.31±0.60	69.58±0.36
	IoU	70.35±0.43	50.69±0.56	30.29±0.36
	OA	96.78±0.02	94.62±0.07	92.40±0.03

Note: The standard deviations were calculated from three replays with different random seeds. The three assessment metrics used here are Omission Rate (OR) denoting the actual instance discarding percentages of noisy labels versus GT, Intersection over Union (IoU) of building class, and Overall Accuracy (OA).

2) *Label Noise Injection Strategy*: We inject incomplete label noise by randomly discarding a certain proportion of instances from GT building masks, as illustrated in Fig. 7. Let α denote the discarding percentage. To simulate the inconsistency in sample quality among different local areas, we perform uniform sampling from the range centered on the given discarding percentage for the whole dataset α_0 to obtain the actual α for each patch. The range is defined as $[\alpha - r, \alpha + r]$ with $r = \min(1 - \alpha, \alpha)$. For example, we sample α from $[0, 0.6]$, $[0, 1]$, $[0.4, 1]$ given $\alpha_0 = 0.3, 0.5, 0.7$, respectively. We implement three replays under each α_0 in our experiments to show the statistical significance of results. The quality assessment of our synthetic noisy labels is presented in Table I. As can be observed, the overall omission rates are close to α_0 though α differs on each patch.

3) *Models and Compared Methods*: We utilize U-Nets [74] as our building extraction models with vanilla U-Net encoder and EfficientNet B5 [75] as backbones for the Massachusetts and Germany datasets, respectively. We have seven methods to compare in total, including two baselines (U-Nets directly trained with GT and noisy labels) and five other methods, two based on pixel-wise label correction and three using regularization techniques. The baseline results of training with GT can be taken as the potential upper limit of these sample correction methods.

For the pixel-wise label correction, a common approach is to apply a fixed threshold K on confidence values (softmax outputs) to select correction candidates [20], [21]. We refer to this approach as **pixel-wise**, and choose $K = 0.6$ as the threshold value after a parameter tuning. In addition, we compare our proposed approach with an **adaptive pixel-wise** version, which automatically sets the thresholds for each patch as the minimum between K and its averaged confidence value on this patch [19]. Furthermore, we also upgrade it by incorporating class-wise thresholds for better performance. To ensure the effectiveness of pixel-wise correction methods, we combine them with our designed ACT module, and take teacher model outputs as corrected pseudo labels.

In addition, we employ three regularization techniques: **consistency constraint** [11], [26], which enforces consistency between teacher and student model outputs; **bootstrapping** [41], [63], which combines original noisy labels and model predictions as soft reference data for loss calculation; and **noisy label regularization** [18], [19], which adds a weighted loss with respect to the original noisy labels in the adaptive pixel-wise training scheme. The consistency constraint is formulated in the form of mean squared error (MSE) with an adaptive weight gradually ramping up to 0.7 in the first 80 epochs. Bootstrapping combines soft pseudo labels (softmax outputs) \hat{p} and original hard noisy labels y by $y' = \beta \cdot y + (1 - \beta) \cdot \hat{p}$, with β exponentially decreasing from 1 to 0.3 in the first 80 epochs. The weight for noisy label regularization is set as 0.25 via parameter searching.

For the proposed AIO2 method, we set the sliding window size w group in ACT as $[10, 20, 30, 40]$ for early learning detection. Taking into account the spatial resolution of the two datasets, we select 5 and 3 as the filter size to generate soft boundaries of pseudo labels in O2C on the Massachusetts (1 m) and Germany (3 m) datasets, respectively. Adam serves as the optimizer in all methods with a learning rate of $1e-3$ and $5e-3$ for the Massachusetts and Germany datasets, respectively. Since we assume no clean labels to be available during training in the noisy label settings, we only use the validation set to select models when trained with GT. The final results reported for the other seven LNL methods derive from three replays after 325 and 300 epochs on the Massachusetts and Germany datasets, respectively. Our evaluation measures include intersection over union (IoU), precision (the positive fraction among predictions), recall (the positive fraction among GT labels/those supposed to be positive), and $F1$ score (the harmonic mean of the precision and recall) of the building class, along with OA.

B. Experimental Results

1) *Massachusetts Dataset*: We first show the performance of various compared methods during the training in Fig. 8. It can be seen that our proposed AIO2 method achieves the best results no matter how severely the labels are corrupted. In situations with low label noise rates, AIO2 is able to attain results comparable to those when training with GT, while in situations with high label noise rates, AIO2 has a more significant advantage over other LNL methods. Additionally, the use of the ACT module can appropriately trigger the

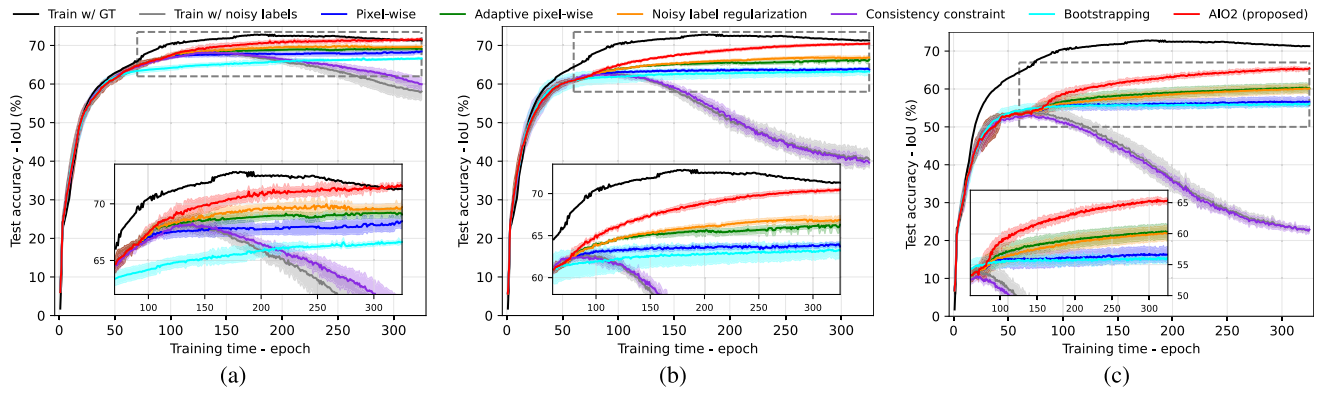


Fig. 8. Test accuracy (IoU) versus training time (epoch) obtained by considered methods trained with GT or incomplete noisy labels under different given discarding percentages (α_0) on the Massachusetts dataset. (a) $\alpha_0 = 0.3$. (b) $\alpha_0 = 0.5$. (c) $\alpha_0 = 0.7$.

TABLE II
IOUS OBTAINED BY CONSIDERED METHODS AFTER 325 EPOCHS ON THE MASSACHUSETTS DATASET

IoU (%)		Final			Maximum		
α_0		0.3	0.5	0.7	0.3	0.5	0.7
Baseline	Train w/ GT	72.47			72.83		
	Train w/ noisy labels	58.06±2.13	40.43±2.17	22.84±0.48	68.51±0.50	62.82±0.60	54.22±0.82
Regularization	Consistency constraint	59.94±1.69	39.59±0.66	22.90±0.86	68.31±0.38	62.85±0.28	53.35±0.54
	Bootstrapping	66.63±0.19	63.23±1.01	56.11±0.58	66.83±0.26	63.52±0.96	56.33±0.75
	Noisy label regularization	<u>69.71±0.58</u>	<u>66.88±0.51</u>	60.09±1.14	<u>70.10±0.52</u>	<u>67.10±0.44</u>	60.15±1.19
Correction	Pixel-wise	68.45±0.60	63.71±0.42	56.63±1.18	68.46±0.60	64.00±0.28	56.85±1.22
	Adaptive pixel-wise	69.06±0.31	66.29±0.35	60.25±1.21	69.41±0.20	66.37±0.36	60.42±1.24
	AIO2 (proposed)	71.56±0.29	70.47±0.17	65.45±0.27	71.72±0.27	70.50±0.19	65.62±0.39

Note: the best and second best results among LNL methods are highlighted in bold and underlined, respectively.

TABLE III

OAS, PRECISIONS, RECALLS, AND $F1$ SCORES OBTAINED BY CONSIDERED METHODS AFTER 325 EPOCHS ON THE MASSACHUSETTS DATASET

Accuracy (%)		OA			Precision			Recall			F1		
α_0		0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7
Baseline	Train w/ GT	94.86			85.23			83.19			84.20		
	Train w/ noisy labels	92.14	88.64	85.07	<u>86.60</u>	<u>86.94</u>	<u>87.10</u>	64.39	43.51	23.88	73.84	57.97	37.49
Regularization	Consistency constraint	92.40	88.58	85.09	87.50	87.78	88.56	66.24	42.33	23.85	75.37	57.11	37.57
	Bootstrapping	93.35	92.65	91.05	80.42	81.78	82.29	80.14	73.97	64.02	80.27	77.66	72.00
	Noisy label regularization	94.09	<u>93.31</u>	<u>92.08</u>	80.02	81.26	84.42	<u>84.93</u>	79.47	67.55	<u>82.40</u>	<u>80.35</u>	75.04
Correction	Pixel-wise	93.82	92.81	91.42	83.68	82.75	84.41	79.18	73.58	63.29	81.36	77.89	72.32
	Adaptive pixel-wise	93.84	93.06	92.02	78.51	79.88	82.48	85.36	<u>80.02</u>	<u>69.33</u>	81.79	79.94	<u>75.32</u>
	AIO2 (proposed)	94.58	94.08	93.01	84.71	83.31	81.57	82.41	82.17	77.11	83.55	82.74	79.28

Note: the best and second best results among LNL methods are highlighted in bold and underlined, respectively.

correction program before the model overfits to noisy labels. Note that the model showcases the memorization effects, that is, the performance is first improved and then degraded on all the training data of different label noise rates. It indicates that memorization effects can be generally observed in the LNLs, thus making it possible for the ACT module to work in common cases. Furthermore, the methods related to label correction strategies, especially those using adaptive thresholds, typically perform better than regularizations. Bootstrapping can partially reduce the influence of label noise in the training process. However, the consistency constraint has a limited impact, particularly when the labels are highly contaminated. Similarly, noisy label regularization provides only marginal improvements, and in some cases, can even worsen the

performance of the pixel-wise label correction method when the quality of training labels is extremely poor.

Some statistical results are shown in Tables II and III. In general, AIO2 performs the best among all LNL methods, achieving the highest IoUs, OAs, and $F1$ scores across different levels of label noise. Furthermore, AIO2 exhibits the most stable performance, with the small standard deviations as well as the final results very close to the best ones that the model can ever achieve during training. Moreover, from Table III, we can observe that the recall is still quite low by consistency constraint, although the precision is improved slightly, indicating that the consistency constraint can help the model learn details, but fails to combat missing instance annotations.

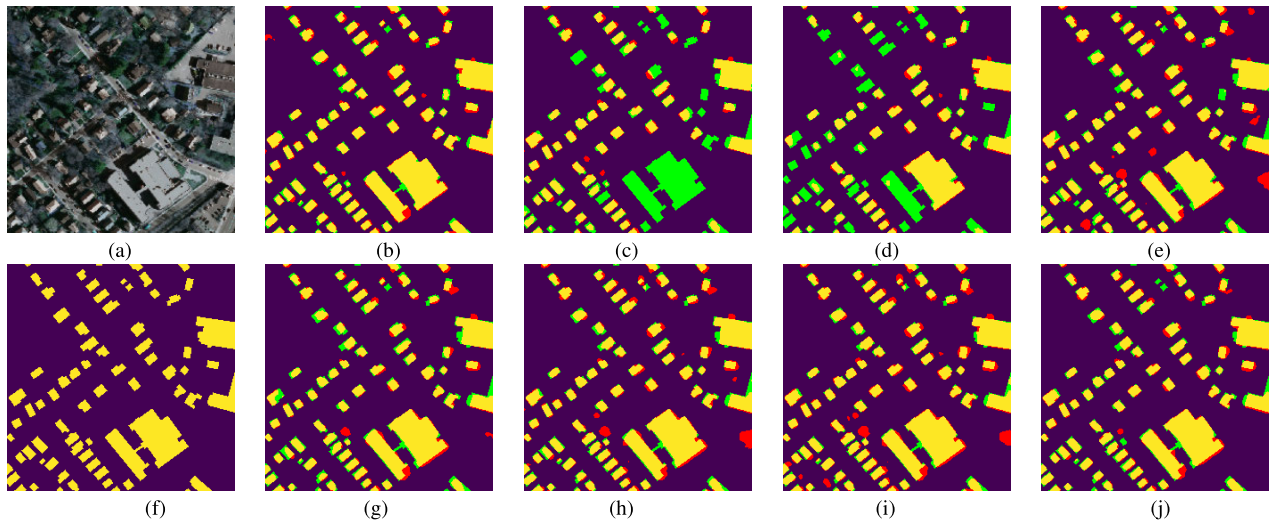


Fig. 9. Segmentation maps obtained by considered methods after 325 epoch training on noisy labels with $\alpha_0 = 0.5$ for the Massachusetts dataset, where false positive and false negative are highlighted with red and green, respectively. (a) Optical image. (b) Train w/GT. (c) Train w/noisy labels. (d) Consistency constraint. (e) Bootstrapping. (f) GT. (g) Pixel-wise. (h) Adaptive pixel-wise. (i) Noisy label regularization. (j) AIO2 (proposed).

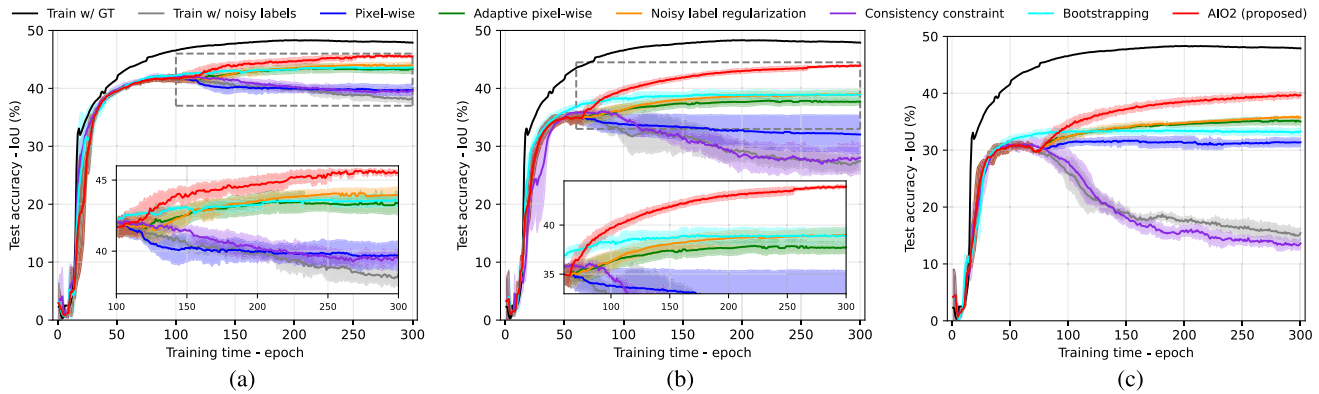


Fig. 10. Test accuracy (IoU) versus training time (epoch) obtained by considered methods trained with GT or incomplete noisy labels under different given discarding percentages (α_0) on the Germany dataset. (a) $\alpha_0=0.3$. (b) $\alpha_0=0.5$. (c) $\alpha_0=0.7$.

Finally, we present some visual results in Fig. 9, from which we can draw the same conclusions as before. The proposed AIO2 can generate better segmentation maps than other considered LNL methods, and is able to detect almost all the instances in the scene and also depict the shapes better than others. On the other hand, the model trained purely on noisy labels (Train w/ noisy labels) overfits to incomplete label noise, thereby omitting a number of houses in the segmentation map. While the consistency constraint improves this a bit, some instances are still excluded, as shown in Fig. 9(d).

2) *Germany Dataset*: Fig. 10 first illustrates the real-time performance of the model as the training proceeds. AIO2 can promptly trigger the ACT module before the model starts to overfit to label noise, partially leading to the best performance among all compared LNL methods. However, pixel-wise correction-based methods do not perform better than bootstrapping to the extent they do on the Massachusetts dataset. This is possibly caused by the lower spatial resolution of planet images, which amplifies the uncertainty of individual pixel samples.

The corresponding IoU statistics and other accuracy results are presented in Tables IV and V, respectively. The superiority

of the proposed AIO2 method is evident from its exceptional performance in terms of IoU, OA, and $F1$ scores, as well as its enhanced stability. In addition, although pixel-wise correction strategies exhibit higher recall rates than AIO2 on the Massachusetts dataset, AIO2 demonstrates higher precision rates on this Germany dataset. Nevertheless, the harmonic mean of precision and recall—that is, the $F1$ scores—suggest that AIO2 still outperforms other compared LNL methods on both datasets. These results highlight the effectiveness of the O2C module at utilizing spatial information to balance the number of samples to correct when applied to different datasets.

To visually evaluate the effectiveness of the proposed method, Fig. 11 presents a series of segmentation maps for a specific test image by various compared methods. These segmentation maps demonstrate that AIO2 produces the top-performing outcome among all LNL methods, with a lower false positive rate and a clearer portrayal of details, followed by bootstrapping. However, there tends to be more green parts, that is, the false negative parts, in the map by AIO2 than those by the noisy label regularization method, when compared to the results on the Massachusetts dataset. We attribute this

TABLE IV
IoUs OBTAINED BY CONSIDERED METHODS AFTER 300 EPOCHS ON THE GERMANY DATASET

IoU (%)		Final			Maximum		
α_0		0.3	0.5	0.7	0.3	0.5	0.7
Baseline	Train w/ GT	48.10			48.32		
	Train w/ noisy labels	38.42±0.59	27.46±1.72	15.42±1.05	42.03±0.62	35.84±0.59	31.18±0.65
Regularization	Consistency constraint	39.60±0.66	28.03±2.80	13.61±0.97	42.38±0.26	36.32±0.39	31.51±0.82
	Bootstrapping	43.59±0.46	38.96±0.85	33.34±0.67	43.75±0.25	39.15±0.80	33.78±0.63
	Noisy label regularization	43.97±0.54	<u>39.01±0.75</u>	<u>35.76±0.14</u>	44.13±0.37	39.05±0.73	35.91±0.20
Correction	Pixel-wise	39.73±0.95	32.03±3.35	31.38±0.76	42.02±0.62	35.64±0.64	31.86±0.71
	Adaptive pixel-wise	43.40±0.79	37.73±0.63	35.06±0.67	43.65±0.74	37.97±0.71	35.33±0.69
	AIO2 (proposed)	45.52±0.21	43.91±0.12	39.66±0.52	45.80±0.11	43.98±0.17	39.83±0.49

Note: the best and second best results among LNL methods are highlighted in bold and underlined, respectively.

TABLE V
OAs, PRECISIONS, RECALLS, AND $F1$ SCORES OBTAINED BY CONSIDERED METHODS AFTER 300 EPOCHS ON THE GERMANY DATASET

Accuracy (%)		OA			Precision			Recall			F1		
α_0		0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7
Baseline	Train w/ GT	93.37			69.48			61.36			65.18		
	Train w/ noisy labels	92.74	91.99	90.70	<u>71.84</u>	<u>76.97</u>	<u>77.66</u>	45.87	31.45	16.93	55.98	44.60	27.78
Regularization	Consistency constraint	<u>92.83</u>	92.05	90.57	72.94	78.17	81.75	46.95	32.30	14.77	57.11	45.58	24.99
	Bootstrapping	92.60	<u>92.53</u>	<u>92.13</u>	63.13	64.85	65.45	58.49	49.87	41.12	60.70	56.36	50.49
	Noisy label regularization	92.35	91.54	91.77	59.51	53.86	57.96	<u>62.91</u>	<u>60.35</u>	49.81	61.09	<u>56.89</u>	<u>53.51</u>
Correction	Pixel-wise	92.57	91.41	91.37	66.69	57.25	58.78	50.04	46.32	43.18	56.97	50.04	49.56
	Adaptive pixel-wise	91.98	91.06	91.14	57.70	50.90	53.84	64.36	61.16	52.31	60.75	56.89	53.51
	AIO2 (proposed)	93.22	93.07	92.59	67.68	65.58	65.45	58.34	57.19	50.52	62.65	61.09	57.02

Note: the best and second best results among LNL methods are highlighted in bold and underlined, respectively.

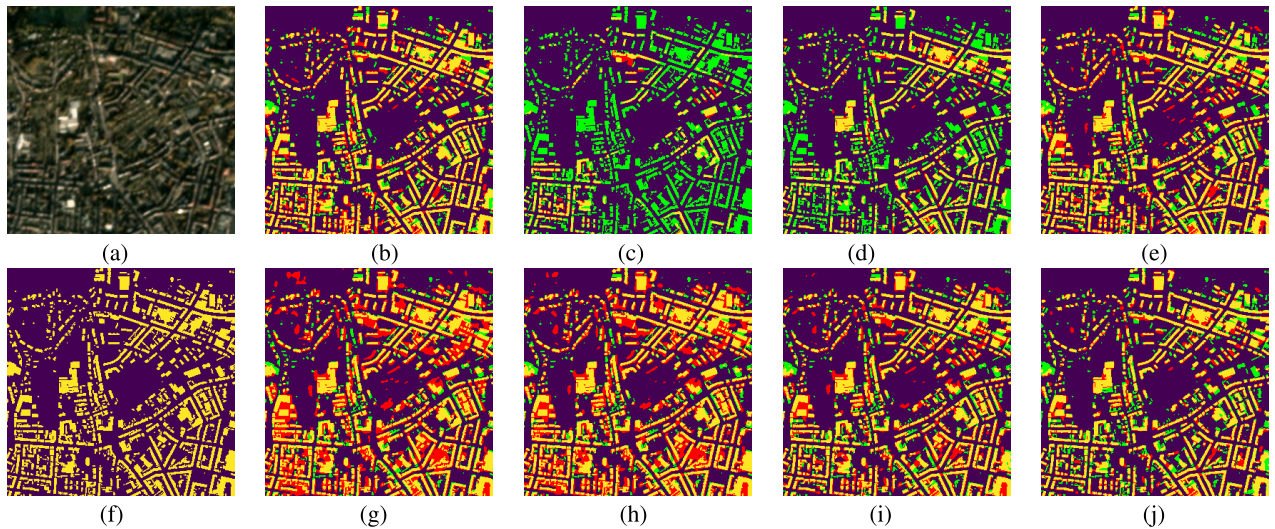


Fig. 11. Segmentation maps obtained by considered methods after 300 epoch training on noisy labels with $\alpha_0 = 0.5$ for the Germany dataset, where false positive and false negative are highlighted with red and green, respectively. (a) Optical image. (b) Train w/GT. (c) Train w/noisy labels. (d) Consistency constraint. (e) Bootstrapping. (f) GT. (g) Pixel-wise. (h) Adaptive pixel-wise. (i) Noisy label regularization. (j) AIO2 (proposed).

phenomenon to the lower spatial resolution of planet images. Many predicted objects tend to be connected to each other, thereby discarded by the overlap check in O2C. Pixel-wise correction methods typically result in over-segmentation maps. This problem can be partially alleviated by employing noisy label regularization. In contrast, the maps generated by training with noisy labels and the consistency constraint are clearly impaired by the incompleteness issue as a consequence of overfitting to noisy labels.

C. Ablation Studies

In addition to the experiments discussed above, we conducted ablation studies to test the roles of the newly designed ACT module, the teacher model, and the soft boundary trick in the O2C module.

1) *ACT Module*: To evaluate the necessity of using ACT in label correction methods, we conducted tests on both object-wise and pixel-wise label correction strategies triggered at different numbers of epochs. The results are presented

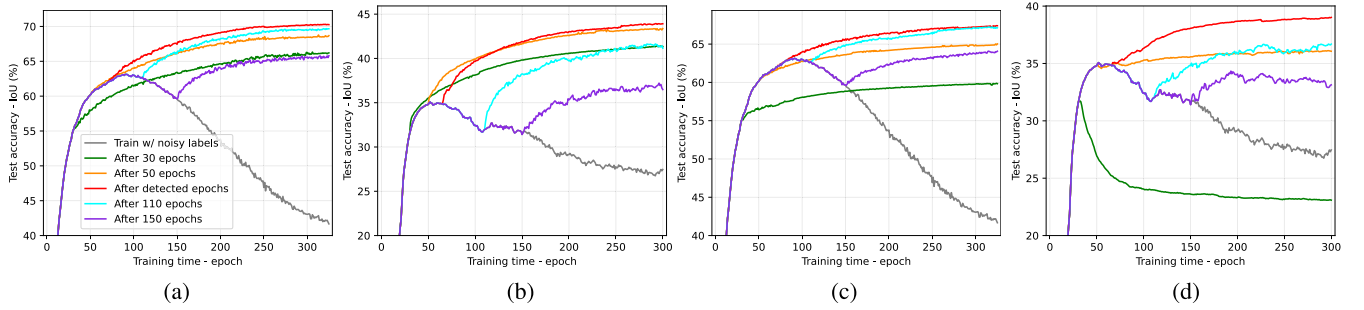


Fig. 12. Test accuracy (IoU) versus training time (epoch) obtained by combining object-wise (O2C) and pixel-wise label correction strategies with different numbers of warm-up epochs on the Massachusetts and Germany datasets, of which the labels are corrupted with a discarding percentage $\alpha_0 = 0.5$. Specifically, the results with respect to pixel-wise correction were generated by class-wise adaptive threshold with noisy label regularization. (a) O2C+Massachusetts. (b) O2C+Germany. (c) Pixel-wise+Massachusetts. (d) Pixel-wise+Germany.

TABLE VI

EARLY-LEARNING DETECTION RESULTS OF THE NUMBERS OF EPOCHS WHERE THE LABEL CORRECTION IS TRIGGERED BY THE ACT MODULE

	α_0	Replay 1	Replay 2	Replay 3	avg.
Massachusetts	0.3	93 (67.66/70.36)	89 (67.32/70.67)	101 (67.08/70.31)	94 (67.35/70.45)
	0.5	82 (62.39/63.04)	70 (62.11/62.90)	73 (61.79/62.76)	75 (62.10/62.90)
	0.7	68 (52.10/53.02)	80 (53.08/53.33)	87 (54.24/54.37)	78 (53.14/53.58)
Germany	0.3	115 (41.22/41.68)	107 (41.93/41.93)	119 (41.32/41.61)	114 (41.49/41.74)
	0.5	67 (35.23/35.41)	76 (32.90/33.38)	67 (33.64/33.82)	70 (33.92/34.20)
	0.7	77 (26.26/26.50)	73 (26.16/26.51)	80 (25.71/26.87)	77 (26.04/26.63)

Note: the training accuracies wrt GT (those at the detected epoch/maximum during the training) are shown in brackets.

in Fig. 12. These results suggest that the timing of the label correction process has a significant impact on the final performance. Starting the label correction procedure too early (when the model is still underfitting) or too late (when the model starts to overfit to noisy labels) can both lead to suboptimal results. In this context, the proposed ACT can effectively mitigate these negative effects by replacing the manual warm-up stage setting with adaptive early learning detection.

Additionally, Table VI presents detailed results of early learning detection by ACT. It displays the model's performance on the training set at detected epochs in comparison with the best counterpart achieved by models when directly trained with noisy labels. Two pieces of information can be gleaned from this table. First, the detected models are very close to the best-performing ones, which partly explains why ACT can help models obtain the promising results shown in Fig. 12. Second, the repeated implementation of ACT shows stable detection performance. Note that while on the Massachusetts dataset with $\alpha_0 = 0.3$ there is little difference in model performance between the detected and the best models, the ACT module can still guarantee the quality of pseudo labels in the O2C module [see Fig. 8]. This is partly due to the high resolution of images (lower pixel uncertainties) and a less severe corruption of labels. In this case, models directly trained on noisy labels would experience less significant damage, leading to an elongated and flattened transition phase where the peak performance deviates from the center point.

2) *Teacher Model*: As mentioned in Section III-D, we decouple model training from the label correction process by utilizing the predictions of teacher models as pseudo labels. By way of comparison, we employ student models as a

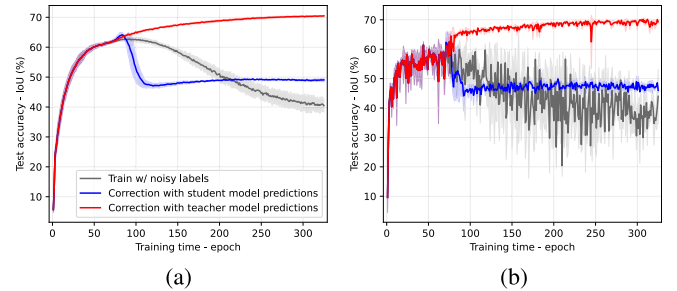


Fig. 13. Test accuracies of (a) teacher models and (b) student models by using teacher (red) or student (blue) model predictions as pseudo labels in O2C on the Massachusetts dataset with noisy labels generated under $\alpha_0 = 0.5$.

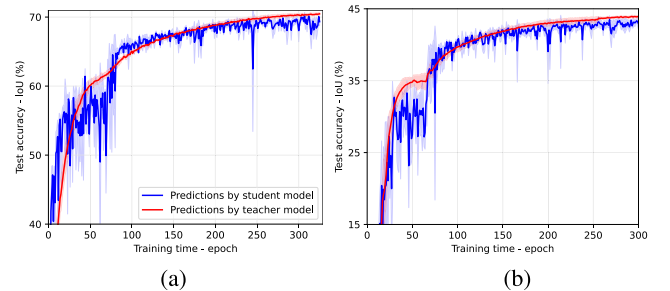


Fig. 14. Test accuracies of teacher (red) and student (blue) models trained by the proposed AIO2 on (a) Massachusetts and (b) Germany datasets with noisy labels generated under $\alpha_0 = 0.5$.

substitute for the teacher models to act as the corrected label source in the O2C module. Fig. 13 illustrates the results of two cases, demonstrating that model training collapses to some degree when label correction is initiated with student model predictions as pseudo labels.

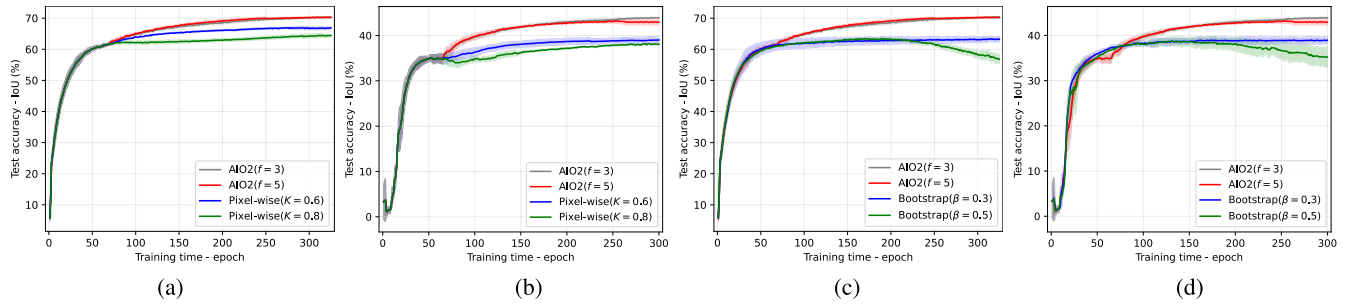


Fig. 15. Parameter sensitivity analysis of AIO2 with noisy labels generated under $\alpha_0 = 0.5$. Versus (a) (pixel-wise) noisy label regularization on the Massachusetts dataset, (b) (pixel-wise) noisy label regularization on the Germany dataset, (c) bootstrapping on the Massachusetts dataset, and (d) bootstrapping on the Germany dataset.

TABLE VII

FINAL TEST ACCURACIES OBTAINED BY AIO2 USING DIFFERENT SOFT FILTER SIZES ON THE MASSACHUSETTS AND GERMANY DATASETS WITH NOISY LABELS GENERATED UNDER $\alpha_0 = 0.5$

Filter size	Massachusetts	Germany
0	68.67 \pm 0.25	42.02 \pm 0.89
3	70.29 \pm 0.26	43.91\pm0.12
5	70.26\pm0.24	42.93 \pm 0.62
7	70.31 \pm 0.32	40.35 \pm 0.76

Note: the results by default settings for each dataset in the previous sections are highlighted in bold.

In addition, we plot test accuracies obtained by teacher and student models in AIO2 as a function of training time, as shown in Fig. 14. Clearly, teacher models can achieve both better and more stable results than student models, thereby enhancing the overall robustness of the proposed method.

3) *Soft Boundary Trick*: Rather than using hard labels, O2C applies a soft filter on correction candidates in order to generate soft labels around boundaries. The reason behind this is that the boundary samples naturally enjoy a higher level of uncertainty and are more likely to be misclassified. We conduct experiments on two datasets by removing the soft boundary trick and trying different filter sizes. The results shown in Table VII demonstrate the effectiveness of the soft boundary trick, which improves the model performance by approximately 2% points on both datasets. Moreover, the model is not overly sensitive to the filter size, except for a slight drop in performance when using a filter size of 7 on the Germany dataset. This is reasonable, since the spatial resolution of this dataset (3 m) is relatively low. A 7×7 filter covers an area of roughly 441 m², which is big enough to mix up everything for building extraction. This is also why we chose a smaller soft filter size (3) empirically for the Germany dataset than for the Massachusetts dataset. Thus, we take the filter size as a fine-tuning hyperparameter, and recommend that readers empirically adjust the filter size based on the target application and the spatial resolution of images. For instance, segmentation of large objects such as urban green spaces, or a higher spatial resolution (e.g., better than 1 m) would be better served by a bigger filter size, while a smaller filter size better suits tiny objects like roads, or a lower spatial resolution. An unrealistic filter size is probably harmful to model performance.

D. Parameter Sensitivity Analysis

In this section, we compare the parameter sensitivity of the proposed AIO2 with that of noisy label regularization and bootstrapping. Recall that noisy label regularization is based on adaptive pixel-wise label correction using a predefined confidence threshold K , while bootstrapping requires a manually set weight β between noisy labels and predictions. Therefore, we plot the results with different filter sizes f of AIO2 in Fig. 15, in contrast to those obtained by noisy label regularization with different K and by bootstrapping with different β . Similar to the results presented in Table VII, AIO2 with different values of f performs on a par with each another, whereas changes in K and β can significantly affect the performance of their corresponding methods. In fact, an inappropriate setting of β in bootstrapping can even lead to a drop in accuracy. These findings demonstrate that the proposed method is less sensitive to parameter settings, making it a promising choice for practical applications.

V. CONCLUSION AND PERSPECTIVES

In this work, we introduced and evaluated a novel mechanism to efficiently train binary semantic segmentation models on incomplete labels by means of AIO2 correction. AIO2 is a fully automatic, iterative label correction framework comprising two key components: the ACT module and the O2C module. Both modules interact without explicitly setting a predefined warm-up phase. While ACT exploits the characteristics of the training accuracy curve over training epochs, O2C features an object-level correction strategy instead of the widely-used pixel-level algorithms. This way, AIO2 automates the addition of pseudo labels to the training dataset, and exploits spatial information to assist with sample correction for segmentation. Besides, O2C operates *on-line* with little extra storage required due to the exploitation of a mean teacher model where the EMA also partially decouples the label correction process from the student model training.

Experimental results obtained on two geographically distinct datasets (Germany and the United States) with spatial resolutions varying by about one order of magnitude indicate the effectiveness of the proposed method. For example, when dropping about 30% of building labels in 1 m-resolution overhead imagery, AIO2 yields accuracy improvements of about 10 percentage points compared to naive supervised training

with noisy labels. When the spatial resolution decreases by an order of magnitude for the Germany dataset, we still observe improvements of about 5% points showcasing the robustness of AIO2. However, we can still observe a larger IoU gap between AIO2 and training with GT labels on the Germany dataset than that on the Massachusetts dataset, which indicates the limitation of the proposed AIO2 method when coping with RS images of a lower resolution.

This work is our initial step toward a systematic solution for training deep neural network models from noisy labels for geospatial semantic segmentation. In future works, we will devote ourselves to improving the effectiveness of AIO2 on handling contiguous objects on the low-resolution RS images. Furthermore, we will test AIO2 on the combined noisy labels additionally with shape label noise, and expand AIO2's application to multiclass segmentation tasks such as land cover mapping. Exploring the potential of AIO2 in a multiround fashion is another interesting topic to investigate.

ACKNOWLEDGMENT

The authors thank Nikolai Skuppin for sharing the Germany dataset [73] with them, and Nassim Ait Ali Braham for discussions during the group meetings.

REFERENCES

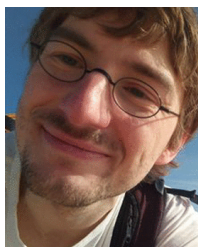
- [1] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.
- [2] J. E. Vargas-Munoz, S. Srivastava, D. Tuia, and A. X. Falcão, "OpenStreetMap: Challenges and opportunities in machine learning and remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 184–199, Mar. 2021.
- [3] C. M. Albrecht, F. Marianno, and L. J. Klein, "AutoGeoLabel: Automated label generation for geospatial machine learning," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2021, pp. 1779–1786.
- [4] C. F. Brown et al., "Dynamic world, near real-time global 10 m land use land cover mapping," *Sci. Data*, vol. 9, no. 1, p. 251, 2022.
- [5] D. Zanaga et al., "ESA WorldCover 10 m 2020 v100," Zenodo, v100, Oct. 2021, doi: [10.5281/zenodo.5571936](https://doi.org/10.5281/zenodo.5571936).
- [6] K. Karra, C. Kontgis, Z. Statman-Weil, J. C. Mazzariello, M. Mathis, and S. P. Brumby, "Global land use/land cover with sentinel 2 and deep learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2021, pp. 4704–4707.
- [7] X. X. Zhu et al., "So2Sat LCZ42: A benchmark data set for the classification of global local climate zones [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 3, pp. 76–89, Sep. 2020.
- [8] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Commun. ACM*, vol. 64, no. 3, pp. 107–115, 2021.
- [9] C. Liu, C. M. Albrecht, Y. Wang, and X. X. Zhu, "Peaks fusion assisted early-stopping strategy for overhead imagery segmentation with noisy labels," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2022, pp. 4842–4847.
- [10] R. Hänsch and O. Hellwich, "The truth about ground truth: Label noise in human-generated reference data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 5594–5597.
- [11] G. Wang et al., "A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2653–2663, Aug. 2020.
- [12] P. Li et al., "Exploring label probability sequence to robustly learn deep convolutional neural networks for road extraction with noisy datasets," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5614018.
- [13] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral-spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2018.
- [14] A. Maiti, S. J. O. Elberink, and G. Vosselman, "Effect of label noise in semantic segmentation of high resolution aerial images and height data," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 2, pp. 275–282, May 2022.
- [15] W. Shen et al., "A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9284–9305, Aug. 2023.
- [16] B. Zhang, J. Xiao, Y. Wei, M. Sun, and K. Huang, "Reliability does matter: An end-to-end weakly supervised semantic segmentation approach," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12765–12772.
- [17] F. Zhang, Y. Shi, Z. Xiong, W. Huang, and X. X. Zhu, "Pseudo features-guided self-training for domain adaptive semantic segmentation of satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5612414.
- [18] Y. Cao and X. Huang, "A coarse-to-fine weakly supervised learning method for green plastic cover segmentation using high-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 188, pp. 157–176, Jun. 2022.
- [19] R. Dong, W. Fang, H. Fu, L. Gan, J. Wang, and P. Gong, "High-resolution land cover mapping through learning with noise correction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4402013.
- [20] Y. Cao and X. Huang, "A full-level fused cross-task transfer learning method for building change detection using noise-robust pretrained networks on crowdsourced labels," *Remote Sens. Environ.*, vol. 284, Jan. 2023, Art. no. 113371.
- [21] S. Liu, K. Liu, W. Zhu, Y. Shen, and C. Fernandez-Granda, "Adaptive early-learning correction for segmentation from noisy annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 2606–2616.
- [22] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 1513–1589, Oct. 2023, doi: [10.1007/s10462-023-10562-9](https://doi.org/10.1007/s10462-023-10562-9).
- [23] M. Brovelli and G. Zamboni, "A new method for the assessment of spatial accuracy and completeness of OpenStreetMap building footprints," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 8, p. 289, Jul. 2018.
- [24] Y. Zhang, Q. Zhou, M. A. Brovelli, and W. Li, "Assessing OSM building completeness using population data," *Int. J. Geographical Inf. Sci.*, vol. 36, no. 7, pp. 1443–1466, Jul. 2022.
- [25] B. Herfort, S. Lautenbach, J. P. de Albuquerque, J. Anderson, and A. Zipf, "A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap," *Nature Commun.*, vol. 14, no. 1, p. 3985, Jul. 2023.
- [26] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.
- [27] J. Wei, Z. Zhu, H. Cheng, T. Liu, G. Niu, and Y. Liu, "Learning with noisy labels revisited: A study using real-world human annotations," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–23.
- [28] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," 2014, *arXiv:1406.2080*.
- [29] A. J. Bekker and J. Goldberger, "Training deep neural-networks based on unreliable labels," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Feb. 2016, pp. 2682–2686.
- [30] J. Goldberger and E. Ben-Reuven, "Training deep neural-networks using a noise adaptation layer," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–9.
- [31] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa, "Joint optimization framework for learning with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5552–5560.
- [32] G. Pleiss, T. Zhang, E. Elenberg, and K. Q. Weinberger, "Identifying mislabeled data using the area under the margin ranking," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 17044–17056.
- [33] E. Malach and S. Shalev-Shwartz, "Decoupling 'when to update' from 'how to update,'" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
- [34] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Proc. ICML*, 2018, pp. 2304–2313.
- [35] B. Han et al., "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.
- [36] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, and M. Sugiyama, "How does disagreement help generalization against label corruption?" in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7164–7173.

- [37] J. Li, R. Socher, and S. C. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–14.
- [38] P.-F. Zhang, Z. Huang, G. Bai, and X.-S. Xu, "IDEAL: High-order-ensemble adaptation network for learning with noisy labels," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 325–333.
- [39] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda, "Early-learning regularization prevents memorization of noisy labels," in *Proc. NIPS*, vol. 33, 2020, pp. 20331–20342.
- [40] K. Yi and J. Wu, "Probabilistic end-to-end noise correction for learning with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 7017–7025.
- [41] S. E. Reed and H. Lee, "Training deep neural networks on noisy labels with bootstrapping," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–11.
- [42] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.
- [43] Y. Lyu and I. W. Tsang, "Curriculum loss: Robust learning and generalization against label corruption," 2019, *arXiv:1905.10045*.
- [44] Y. Liu and H. Guo, "Peer loss functions: Learning from noisy labels without knowing noise rates," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6226–6236.
- [45] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8135–8153, Nov. 2023.
- [46] N. Manwani and P. S. Sastry, "Noise tolerance under risk minimization," *IEEE Trans. Cybern.*, vol. 43, no. 3, pp. 1146–1151, Jun. 2013.
- [47] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 1–7.
- [48] X. Tai, G. Wang, C. Grecos, and P. Ren, "Coastal image classification under noisy labels," *J. Coastal Res.*, vol. 102, no. sp1, pp. 151–156, Dec. 2020.
- [49] Z. Huang, C. O. Dumitru, Z. Pan, B. Lei, and M. Datcu, "Classification of large-scale high-resolution SAR images with deep transfer learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 1, pp. 107–111, Jan. 2021.
- [50] B. B. Damodaran, R. Flamary, V. Seguy, and N. Courty, "An entropic optimal transport loss for learning deep neural networks under label noise in remote sensing images," *Comput. Vis. Image Understand.*, vol. 191, Feb. 2020, Art. no. 102863.
- [51] Y. Li, Y. Zhang, and Z. Zhu, "Error-tolerant deep learning for remote sensing image scene classification," *IEEE Trans. Cybern.*, vol. 51, no. 4, pp. 1756–1768, Apr. 2020.
- [52] B. Tu, W. Kuang, W. He, G. Zhang, and Y. Peng, "Robust learning of mislabeled training samples for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5623–5639, 2020.
- [53] J. Kang, R. Fernandez-Beltran, X. Kang, J. Ni, and A. Plaza, "Noise-tolerant deep neighborhood embedding for remotely sensed images with label noise," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2551–2562, 2021.
- [54] J. Kang, R. Fernandez-Beltran, P. Duan, X. Kang, and A. J. Plaza, "Robust normalized softmax loss for deep metric learning-based characterization of remote sensing images with label noise," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8798–8811, Oct. 2021.
- [55] T. Burgert, M. Ravanbakhsh, and B. Demir, "On the effects of different types of label noise in multi-label remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5413713.
- [56] A. K. Aksoy, M. Ravanbakhsh, and B. Demir, "Multi-label noise robust collaborative learning for remote sensing image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, early acces, Oct. 20, 2024, doi: [10.1109/TNNLS.2022.3209992](https://doi.org/10.1109/TNNLS.2022.3209992).
- [57] G. Sumbul and B. Demir, "Generative reasoning integrated label noise robust deep image representation learning," *IEEE Trans. Image Process.*, vol. 32, pp. 4529–4542, 2023.
- [58] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.
- [59] 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.
- [60] N. Ahmed, R. M. Rahman, M. S. G. Adnan, and B. Ahmed, "Dense prediction of label noise for learning building extraction from aerial drone imagery," *Int. J. Remote Sens.*, vol. 42, no. 23, pp. 8906–8929, Dec. 2021.
- [61] P. Li et al., "Robust deep neural networks for road extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6182–6197, Jul. 2021.
- [62] Z. Zhang, W. Guo, M. Li, and W. Yu, "GIS-supervised building extraction with label noise-adaptive fully convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 12, pp. 2135–2139, Dec. 2020.
- [63] C. Henry, F. Fraundorfer, and E. Vig, "Aerial road segmentation in the presence of topological label noise," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 2336–2343.
- [64] K. Malkin et al., "Label super-resolution networks," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–22.
- [65] C. Lin et al., "Deep learning network intensification for preventing noisy-labeled samples for remote sensing classification," *Remote Sens.*, vol. 13, no. 9, p. 1689, Apr. 2021.
- [66] C. M. Albrecht et al., "Change detection from remote sensing to guide OpenStreetMap labeling," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 7, p. 427, Jul. 2020.
- [67] J. Sun, J. Liu, L. Hu, Z. Wei, and L. Xiao, "A mutual teaching framework with momentum correction for unsupervised hyperspectral image change detection," *Remote Sens.*, vol. 14, no. 4, p. 1000, Feb. 2022.
- [68] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–13.
- [69] D. Arpit et al., "A closer look at memorization in deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 233–242.
- [70] M.-P. Dubuisson and A. K. Jain, "A modified Hausdorff distance for object matching," in *Proc. 12th Int. Conf. Pattern Recognit.*, vol. 1, 1994, pp. 566–568.
- [71] S. Jadon, "A survey of loss functions for semantic segmentation," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Oct. 2020, pp. 1–7.
- [72] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Univ. Toronto, Comput. Center Toronto, Toronto, ON, Canada, 2013.
- [73] N. Skuppim, E. J. Hoffmann, Y. Shi, and X. X. Zhu, "Building type classification with incomplete labels," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 5844–5847.
- [74] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241.
- [75] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, K. Chaudhuri and R. Salakhutdinov, Eds. Jun. 2019, pp. 6105–6114.



Chenying Liu (Student Member, IEEE) received the B.S. and M.S. degrees (Hons.) from Sun Yat-sen University, Guangzhou, China, in 2017 and 2020, respectively. She is currently pursuing the Ph.D. degree with the German Aerospace Center (DLR), Weßling, Germany, and the Technical University of Munich, Munich, Germany.

From 2020 to 2021, she was a Research Associate with Sun Yat-sen University. Her research interests include remote sensing and Earth observation, weakly supervised learning, multimodal learning, and signal processing.



Conrad M. Albrecht (Member, IEEE) received the Ph.D. (Dr. rer. nat.) degree in physics from Heidelberg University, Heidelberg, Germany, in 2014, with a focus on distributed computing to study physics at low temperatures.

After Post-Doctoral Researcher at the Institute for Theoretical Physics, Heidelberg University in 2014, for more than six years, he was a Research Scientist with the Physical Sciences Department, IBM T. J. Watson Research Center, Yorktown, NY, USA. He has been a PI of the HelmholtzAI Young

Investigator Group (YIG) “Large-Scale Data Mining in Earth Observation” (DM4EO), Munich, Germany, since April 2021. In July 2023, he was appointed as a Visiting Associate Professor with the Institute of Nasca, Yamagata University, Yamagata, Japan, contributing to research in machine learning for the UNESCO World Heritage of the Nasca Culture in Peru. He is currently a Researcher with the Earth Observation Center, German Aerospace Center, Weßling, Germany. His research agenda interconnects physical models and numerical analysis, employing big data technologies, and machine learning through open-science research.

Dr. Albrecht co-organized workshops at the IEEE BigData Conference, IGARSS conferences, and the AAAS Annual Meeting. At home in the European Union and the United States, some of his transatlantic initiatives include: in 2022, he started a collaboration on physics-aware deep learning with the Earth and Environmental Engineering department of Columbia University, New York City, USA. In 2023, he was invited by the Alexander-von-Humboldt Foundation to present at the German-American Frontiers of Engineering Symposium. Beginning in 2024, initiated by the Helmholtz Imaging and Data Science, Munich, he is in close touch with the German Department, Princeton University, Princeton, NJ, USA, for cross-Atlantic undergraduate internships.



Yi Wang (Student Member, IEEE) received the B.E. degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2018, and the M.Sc. degree in geomatics engineering from the University of Stuttgart, Stuttgart, Germany, in 2021. He is currently pursuing the Ph.D. degree with the German Aerospace Center, Weßling, Germany, and the Technical University of Munich, Munich, Germany.

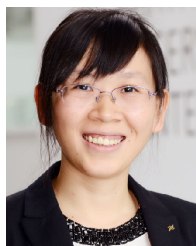
In 2020, he spent three months at the Perception System Group, Sony Corporation, Stuttgart,

Germany. His research interests include remote sensing, deep learning, computer vision, and self-supervised learning.



Qingyu Li received the bachelor’s degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2015, the master’s degree in Earth oriented space science and technology (ESPACE) from Technische Universität München (TUM), Munich, Germany, in 2018, the master’s degree in photogrammetry and remote sensing from Wuhan University in 2019, and the Doctor of Engineering (Dr.-Ing.) degree from TUM, in 2022.

From 2019 to 2022, she was a Research Associate with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany. She is currently a Post-Doctoral Researcher with the Chair of Data Science in Earth Observation, TUM. Her research interests include remote sensing data processing and interpretation, artificial intelligence, remote sensing applications, and urban analysis.



Xiao Xiang Zhu (Fellow, IEEE) received the Master (M.Sc.) degree, her doctor of engineering (Dr.-Ing.) degree and her “Habilitation” in the field of signal processing from Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is the Chair Professor for Data Science in Earth Observation at Technical University of Munich (TUM) and was the founding Head of the Department “EO Data Science” at the Remote Sensing Technology Institute, German Aerospace Center

(DLR). Since May 2020, she is the PI and director of the international future AI lab “AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond”, Munich, Germany. Since October 2020, she also serves as a Director of the Munich Data Science Institute (MDSI), TUM. From 2019 to 2022, She has been a co-coordinator of the Munich Data Science Research School (www.mu-ds.de) and the head of the Helmholtz Artificial Intelligence – Research Field “Aeronautics, Space and Transport”. She was a guest scientist or visiting professor at the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, the University of Tokyo, Tokyo, Japan, and University of California, Los Angeles, United States in 2009, 2014, 2015, and 2016, respectively. She is currently a visiting AI professor at ESA’s Phi-lab, Frascati, Italy. Her main research interests are remote sensing and Earth observation, signal processing, machine learning and data science, with their applications in tackling societal grand challenges, e.g. Global Urbanization, UN’s SDGs, and Climate Change.

Dr. Zhu has been a member of young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She serves in the scientific advisory board in several research organizations, among others the German Research Center for Geosciences (GFZ, 2020–2023) and Potsdam Institute for Climate Impact Research (PIK). She is an associate Editor of IEEE Transactions on Geoscience and Remote Sensing, Pattern Recognition and served as the area editor responsible for special issues of IEEE Signal Processing Magazine (2021–2023). She is a Fellow of IEEE.