Handling Noisy Annotation for Remote Sensing Semantic Segmentation via Boundary-aware Knowledge Distillation

Yue Sun[1], Dong Liang[1*] , Shaoyuan Li[1], Songcan Chen[1], and Sheng-Jun Huang[1]

***Abstract*—In recent years, image segmentation has made significant progress, but acquiring annotated data is still a considerable challenge, especially in remote sensing imagery (RSI). The complex structure and inter-category confusion of RSI increase the time-consuming and cost of pixel-level annotation, and noisy annotations inevitably appear. This paper proposes a boundary-aware knowledge distillation method (BAKD) to handle noisy annotations by evaluating their uncertainty. BAKD consists of two core strategies: Predictive Confidence Evaluation (PCE) and Boundary-annotated Reliability Evaluation (BRE). The predictive confidence jointly decided by the teacher and student networks reflects the annotation's uncertainty. The boundary-annotated reliability directly measures the annotation's uncertainty based on the distance from the annotation to the semantic boundary. Leveraging these two types of uncertainty information, BAKD assigns each sample a comprehensive boundary-aware weight to identify samples with potential noisy annotations. This alleviates the impact of noisy annotation on the model's training and improves its generalization performance. Experimental results show that BAKD achieves competitive semantic segmentation performance on the Potsdam and Vaihingen benchmarks compared with the state-of-the-art KD methods. In addition, BAKD can be easily integrated into semantic segmentation methods based on KD, extending their applicability in handling noisy annotations.**
Codes are available at https://github.com/sunyueue/BAKD.git.

***Index Terms*—Semantic Segmentation, Knowledge Distillation, Sample Weighting, Noisy Annotation.**

I. Introduction

SEMANTIC segmentation of remote sensing imagery (RSI) has applied in many fields such as hazard assessment [1], [2], urban planning [3], [4], farmland detection [5], [6] and natural disaster detection [7]. The current deep neural networks,e.g., DeepLab [8]–[11], PSPNet [12], HRNet [13], have achieved remarkable success and are widely used in semantic segmentation. However, these methods' superiority mainly relies on supervised learning, which requires a large amount of manually annotated training data with high quality. Compared with visual tasks such as image classification or object detection, pixel-level annotation for semantic segmentation tasks takes a long time and requires experts with domain knowledge to implement [14], [15]. According to [16], it takes nearly 90 minutes to annotate a high-resolution natural urban landscape image at the pixel level. Since RSI generally produces higher resolution and more complex structures than other image types, the annotation cost would be much higher [17]. Therefore, obtaining high-quality manual annotations has become a significant obstacle to developing deep learning models for RSI semantic segmentation tasks [18].

(a) Image (b) Fine annotation (c) Coarse annotation

Fig. 1. The challenges in RSI semantic segmentation: (a) Factors that affect the cost and quality of annotation in RSI. (b) Fine manual annotation is expensive and laborious. (c) Coarse annotation reduces costs but limits the segmentation model's performance.

Due to the annotator's subjective cognitive bias, the labeling results will inevitably contain some errors. Complex RSI is affected by factors such as occlusion and inter-category confusion, and annotation errors are inevitable [19]. As shown in the green box in Fig. 1 (a), the tall building and its shadow occlude the low vegetation, rendering the occluded area invisible in RSI and introducing annotation uncertainty. In complex scenes with intricate backgrounds or overlapping objects, the pixel features of objects from different categories can be easily confused. As shown in the red box in Fig. 1, the similarities between trees and low vegetation often lead to annotation errors. Additionally, small targets like the car illustrated in the yellow box in Fig. 1 (a) pose challenges in obtaining precise location information, often resulting in misaligned boundaries during annotation. These issues complicate manual annotation and increase the risk of noisy annotations in the semantic boundary. Given that manual annotation is time-consuming and prone to many annotation errors in the semantic boundary, we leverage coarse annotations to train our semantic segmentation model. While the coarse annotations in Fig. 1 (c) may overlook finer details along semantic boundaries in Fig. 1 (b), leading to pixel-by-pixel boundary misalignment, they offer a significant reduction in annotation costs.

Noisy annotations within coarse annotations make supervised learning prone to overfitting bias, hindering the segmen-

All authors are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing 211106, China.
All authors:
1 All authors are {sun.yue,liangdong,lisy,s.chen,huangsj}@nuaa.edu.cn.
* Corresponding author.

tation model's ability to grasp correct knowledge and limiting its performance upper bound [20]. Co-learning methods [21]–[23] employ dual models to select low-loss samples for learning to improve the model's performance on noisy annotations. However, these techniques predominantly focus on image-level sample selection and have limited applicability to pixel-level semantic segmentation tasks. Knowledge Distillation (KD) [24] leverages soft labels from the teacher model in place of single hard labels because soft labels contain richer category relationship information. Soft labels enhance the student model's performance compared to rigid hard labels, which may be noisy in coarse annotations. Studies [25], [26] also affirm that label smoothing and KD strategies enhance the model's performance under different degrees of noisy annotations. Semantic segmentation methods based on KD [27]–[31] leverage diverse supervision information from the teacher network to effectively enhance the student network's robustness on noisy annotations. Nonetheless, these methods overlook the teacher network's potential to guide the student network in screening noisy annotations and learning correct ones. Additionally, they fail to fully consider leveraging boundary information from coarse annotations to guide model training. Noisy annotations in coarse annotations will cause the model to learn incorrect annotation information, particularly at the semantic boundary where this impact is more pronounced. Therefore, investigating how the teacher model can guide in handling noisy annotations and effectively utilizing boundary information from coarse annotations is crucial for extending the applicability of KD in coarsely annotated RSI.

To address the challenge posed by coarse annotations in complex scenarios, we propose a new boundary-aware knowledge distillation framework (BAKD) based on annotation uncertainty. We devise two uncertainty evaluation strategies to derive the comprehensive boundary-aware weight: Predictive Confidence Evaluation (PCE) and Boundary-annotated Reliability Evaluation (BRE). PCE leverages the teacher and student networks' prediction discrepancy to collaboratively evaluate the predictive confidence for each sample. Given that coarse annotations may contain noisy annotations at the semantic boundary, we directly evaluate the annotations' reliability based on the distance from the annotation to the semantic boundary. During training, BAKD allocates a comprehensive boundary-aware weight to each sample by fusing these two types of uncertainty information. Subsequently, BAKD screens training samples based on the boundary-aware weight and identifies samples with potentially noisy annotations. For these samples, BAKD allocates lower weights to mitigate their negative impact on the student network training.

The main contributions of this paper are therefore:

1) We propose an innovative boundary-aware knowledge distillation (BAKD) framework to address noisy annotation. It can be seamlessly integrated with mainstream semantic segmentation methods to extend their applicability in handling noisy annotations.

2) Considering the characteristics of coarse annotations, we devise two uncertainty evaluation strategies: PCE suppresses noisy annotations by combining the predictive confidence of teacher and student networks. BRE leverages the distance from annotation to the semantic boundary to obtain annotated reliability and suppress the negative impact of noisy annotations.

3) Extensive experiment on the ISPRS Potsdam and ISPRS Vaihingen datasets validates the effectiveness and practicability of BAKD. Compared with mainstream methods, BAKD facilitates more stable model training in noisy annotations and significantly enhances the model's segmentation performance.

Subsequently, in Section II, we provide a concise overview of the related work in the domain of semantic segmentation based on KD and inaccurate supervision. Section III provides a detailed description of our proposed BAKD method. Section IV presents extensive experiments on two classic RSI semantic segmentation datasets, demonstrating the superiority of BAKD. Conclusions are articulated in Section V.

II. Related Work

A. Semantic Segmentation based on Knowledge Distillation

Hinton *et al.* [24] first introduced the concept of knowledge distillation (KD). Most previous studies on KD, such as [30], [32], focused on image classification. However, image-level KD does not take the locally structured information for semantic segmentation into account, so it has natural defects for pixel-level semantic segmentation. Most efforts have focused on defining the knowledge for the segmentation task to solve this problem. Liu *et al.* [31]extracted structured knowledge from the teacher network to the student network by using two structured distillation schemes. Wang *et al.* [33] put forward a new intra-class feature variation distillation (IFVD), which transformed the cumbersome teacher model into a compact student model. Shu *et al.* [29] introduced a new channel-wise KD method that minimized differences between teacher and student networks by utilizing asymmetric KL divergence. Feng *et al.* [28] improved the classification accuracy of existing compact networks by capturing similar knowledge in the pixel and category dimensions, respectively. To solve the problem that the previous techniques ignore the global semantic relationship between pixels in different images, Yang *et al.* [27] attempted to model pixel-pixel and pixel-region comparison relationships in semantic segmentation tasks as knowledge and transfer global pixel correlation from teachers to students for semantic segmentation.

However, these methods overlook the teacher network's capability to guide the student network to screen noisy annotations and learn correct annotations in noisy annotation scenarios. To address this problem, BAKD allocates lower weights to samples with noisy annotations by jointly evaluating the annotation's uncertainty from teacher and student networks. Additionally, previous methods fail to consider how to leverage boundary information from coarse annotations to guide model training. BAKD leverages the distance from the annotation to the semantic boundary to evaluate the annotation's uncertainty and suppresses the noisy annotation's negative impact.

B. Classification from Noisy Annotations with Co-Learning

Co-learning was originally proposed as a strategy for image classification tasks with noisy labels. Han *et al.* [22] demon-

strated that collaborative training effectively deals with label noise. Co-learning improved the model's performance on noisy data by utilizing two models and selecting small loss samples from the samples for learning. Among them, Decouple [21] proposed how to measure the disagreement between the two models and decide when and how to update the model parameters based on the disagreement. Co-teaching [22]and Co-teaching+ [23] maintained two independent DNN networks, each selected samples that it believed to have more minor losses and passed them to the other network for further training. Each network back-propagated the small batch selected by its peer network to update itself. Compared with Co-teaching [22], Co-teaching+ [23] also introduced a decoupled divergence strategy, which first screened out samples with inconsistent predictions and then selected small loss samples from these samples for training. However, these image-level sample selection methods have certain limitations. In the presence of significant noisy annotations in the dataset, simply discarding entire images inevitably leads to overfitting, a phenomenon confirmed in subsequent experiments. Therefore, our BAKD utilizes two models to evaluate the predictive confidence of each pixel from a pixel-level perspective and screen out correct annotations to train the model.

C. Semantic Segmentation from Noisy Annotations

Image segmentation from noisy annotations is a critical problem. Recent studies addressed this problem by explicitly considering systematic human labeling errors [34] and modifying the segmentation loss to increase robustness [35], [36]. Other works proposed utilizing two interconnected networks to learn together to discover noisy gradient information [37]. Alternatively, they learned high-level spatial structures of images and used them as supervisory signals to mitigate the impact of incorrect annotations [38]. However, the disadvantage of these methods is that some samples with completely clean annotations are required. Liu *et al.* [39] proposed adaptively triggered online object-wise label correction (AIO2) to address label noise arising from incomplete label sets. However, AIO2 is only applicable to binary segmentation tasks.

In recent research, Liu *et al.* [40] proposed an adaptive early learning correction (ADELE) method, which monitors the IoU curves of each class to detect the onset of the memory stage. However, ADELE [40] requires the recording of the IoU values for each pixel in every training iteration, which imposes significant demands on memory and computational resources, diverging from practical application requirements. In addition, the ADELE paper also points out that the success of ADELE depends on the quality of the initial annotations. When the initial annotation quality is poor, it is difficult to achieve the correction conditions and the errors cannot be completely corrected. In contrast, BAKD's evaluation strategy can still effectively identify these noisy annotations in the case of severe noisy annotations. Fang *et al.* [41] drew inspiration from the Co-teaching concept in image classification to design a reliable mutual distillation (RMD) method, which leverages the collaboration of two segmentation models to filter out label noise from coarse annotations. However, the confidence of small targets is often low in RSI segmentation, and RMD's filtering strategy may excessively eliminate small target classes, resulting in insufficient training samples and overfitting. Furthermore, the mutual training between the two models can lead to interference, particularly in the early stages, where a poorly performing model may adversely affect the learning performance of the other model. In contrast, BAKD incorporates a warm-up phase that utilizes a pre-trained teacher model to convey reliable knowledge, thereby enhancing the stability and efficiency of the learning process.

D. Prediction Uncertainty in Remote Sensing Imagery

The complex structure of remote sensing imagery, confusion between categories, and noisy annotations in coarse annotations make it unavoidable that the current segmentation network would misjudge a certain category, which leads to uncertainty in model prediction. Therefore, solving the problem of prediction uncertainty is the key to further improving the segmentation performance of the model.

To this end, Dong *et al.* [42] estimated the uncertainty of the prediction by using entropy measurement to identify the pixels that need to be updated. Chen *et al.* [43] employed the class probability values predicted by the model to mine high-confidence samples from images with coarse class annotations as pseudo-labels. Cao *et al.* [44] used the absolute difference of probability maps as an uncertainty-aware analysis tool to obtain more reliable pseudo-labels. In the latest research, Lyu *et al.* [45] introduced an uncertainty analysis method to improve the accuracy of semantic segmentation by improving the utilization of remote sensing image features. Li *et al.* [46] proposed an uncertainty-aware network (UANet) that gradually guides attention to uncertain pixels during feature interaction. Li *et al.* [47] also proposed an uncertainty-aware detail-preserving network (UADPNet), which introduced an arbitrary uncertainty estimator at the data level to obtain an assessment of uncertainty and highlighted uncertain pixels through an uncertainty-aware fusion module (UAFM).

However, these methods about prediction uncertainty rely on the model's prediction to measure uncertainty. For instance, UANet [46] assessed the uncertainty of foreground and background pixels by calculating the pixel's probability, depending entirely on the model's predictions. Similarly, the UAFM module in UADPNet [47] also utilized the model's prediction to measure uncertainty. The student network has not yet converged in the early stages of model training, and the predictions may not be sufficiently reliable, leading to errors in uncertainty evaluation that subsequently affect the model's training effectiveness. In contrast, BAKD combines KD to jointly guide the student model's training using teacher and student networks' predictions, providing a more reliable uncertainty assessment in the early training phase.

III. THE METHODOLOGY

This section presents our method's complete workflow. We start by introducing the loss function paradigm for semantic segmentation with knowledge distillation (KD) in Subsection III-A. In Section III-B, we define Boundary-aware Knowledge

Distillation (BAKD), which combines two evaluation strategies: Predictive Confidence Evaluation (PCE) and Boundary-annotated Reliability Evaluation (BRE). Subsequently, we elaborate on the specific implementation details of the PCE and BRE strategies in Sections III-C and III-D, respectively. Finally, we discuss integrating BAKD into other semantic segmentation methods based on KD in Subsection III-E.

A. Preliminary

Semantic segmentation is a dense pixel-level prediction task that assigns a specific class to each pixel in an image. Given an input image I with dimensions $W \times H \times 3$, the feature extractor of a segmenter first extracts a feature map F, where H and W represent the height and width of the input image and feature map, respectively. The categorical logit map $\mathbf{Z}$ is generated from feature map F by applying a classifier. In image semantic segmentation tasks, the cross-entropy (CE) loss is commonly used to measure the difference between the predicted probability distribution $\sigma\left(\mathbf{Z}_{h,w}\right)$ and the ground truth label $\mathbf{y}_{h,w}$. Optimization using CE as task loss:

$$L_{task} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} CE\left(\sigma\left(\mathbf{Z}_{h,w}\right), \mathbf{y}_{h,w}\right) \tag{1}$$

Specifically, for each pixel at position (h, w), the CE loss is computed as:

$$CE\left(\sigma\left(\mathbf{Z}_{h,w}\right), \mathbf{y}_{h,w}\right) = -\sum_{c=1}^{C} \mathbf{y}_{h,w}^{(c)} \log\left(\sigma\left(\mathbf{Z}_{h,w}^{(c)}\right)\right) \tag{2}$$

where $\mathbf{y}_{h,w}$ ($\mathbf{y}_{h,w} \in \{0,1\}^C$) is the one-hot encoded ground truth label at the (h, w)-th pixel, $\mathbf{y}_{h,w}^{(c)}$ denotes the value corresponding to the c-th class of the one-hot encoded label at the (h, w)-th pixel, and $\mathbf{Z}_{h,w}$ denotes the output logits for the (h, w)-th pixel. The softmax function σ generates the category probability.

The existing KD methods usually employ a pixel-wise alignment among class probabilities between a cumbersome teacher network t and a lightweight student network s to obtain a distillation loss, which can be formulated as follows:

$$L_{kd} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} KL\left[\sigma\left(\frac{\mathbf{Z}_{h,w}^{t}}{T}\right) \| \sigma\left(\frac{\mathbf{Z}_{h,w}^{s}}{T}\right)\right] \tag{3}$$

where $\mathbf{Z}_{h,w}^{t}$ and $\mathbf{Z}_{h,w}^{s}$ represent the output logits for the (h, w)-th pixel produced from the teacher and the student network, respectively. σ function calculates the category probability of the (h, w)-th pixel generated by the student and teacher networks, respectively. The parameter T represents the temperature taken by distillation and reflects the label's softening degree. For a fair comparison with previous works [27], [31], we set T =1 in our experiments. KL denotes the Kullback-Leibler divergence, which measures the difference between two probability distributions. Since the KD methods we compared employ Kullback-Leibler (KL) divergence, we choose the most commonly used KL divergence as the distillation loss, which can be formulated as follows:

$$D_{KL}\left(P \| Q\right) = \sum_{c=1}^{C} P_c \log\left(\frac{P_c}{Q_c}\right) \tag{4}$$

where $P = \sigma\left(\frac{\mathbf{Z}_{h,w}^{t}}{T}\right)$, represents the probability distribution of the c-th class at position (h, w) from the teacher network, and $Q = \sigma\left(\frac{\mathbf{Z}_{h,w}^{s}}{T}\right)$, represents the probability distribution of the same class at the same position from the student network.

B. Define the Boundary-aware Knowledge Distillation

Learning the semantic boundary is a significant challenge in segmentation tasks, especially given that most noisy annotations tend to occur at these boundaries. We plan to generate a pixel-level loss weight to suppress the noise of the coarse annotations, enabling the model to learn from the correct annotations. Based on the above inspiration, we propose a boundary-aware knowledge distillation (BAKD) method that captures the annotations' uncertainty through two strategies: **Predictive Confidence Evaluation (PCE)** and **Boundary-annotated Reliability Evaluation (BRE)**.

As shown in Fig. 2, we initially compute the prediction discrepancy between the teacher and the student networks' predictions and corresponding annotations during the training process. Then, we collaboratively evaluate to obtain a dynamic predictive confidence score PC based on this discrepancy. The lower PC score indicates the higher annotation's uncertainty and the more likely it is a noisy annotation. Given that noise typically intensifies near the semantic boundary in coarse annotations, we calculate the distance D_{dis} from each annotation to these boundaries. Subsequently, employing our devised exponential normalization approach, we derive the Boundary-annotated reliability score BR. Annotations closer to boundaries exhibit lower reliability, potentially indicating noisy annotations. Consequently, a comprehensive boundary-aware weight W_{BAW} is assigned to each sample by combining the PC and BR, where BR complements PC's boundary considerations. By comprehensively considering these two types of information, BAKD effectively identifies samples with potentially noisy annotations. We incorporate the weight W_{BAW} into the loss function using Eq. 5 to update the student network parameters. Details on generating W_{BAW} are elaborated below.

$$L_{BAKD} = W_{BAW} \cdot L_{task} + L_{kd} \tag{5}$$

C. Predictive Confidence Evaluation (PCE) Strategy

Screening samples with noisy annotations requires evaluating the annotation's uncertainty as a starting point. Traditionally, predictive confidence from the network serves as a pivotal metric for evaluating the annotation's uncertainty. Low confidence typically indicates classification difficulty, while it often signifies mislabeling in noisy supervised datasets. After introducing KD, the most straightforward approach involves using the prediction discrepancy between the reliable teacher network's prediction and corresponding annotation as

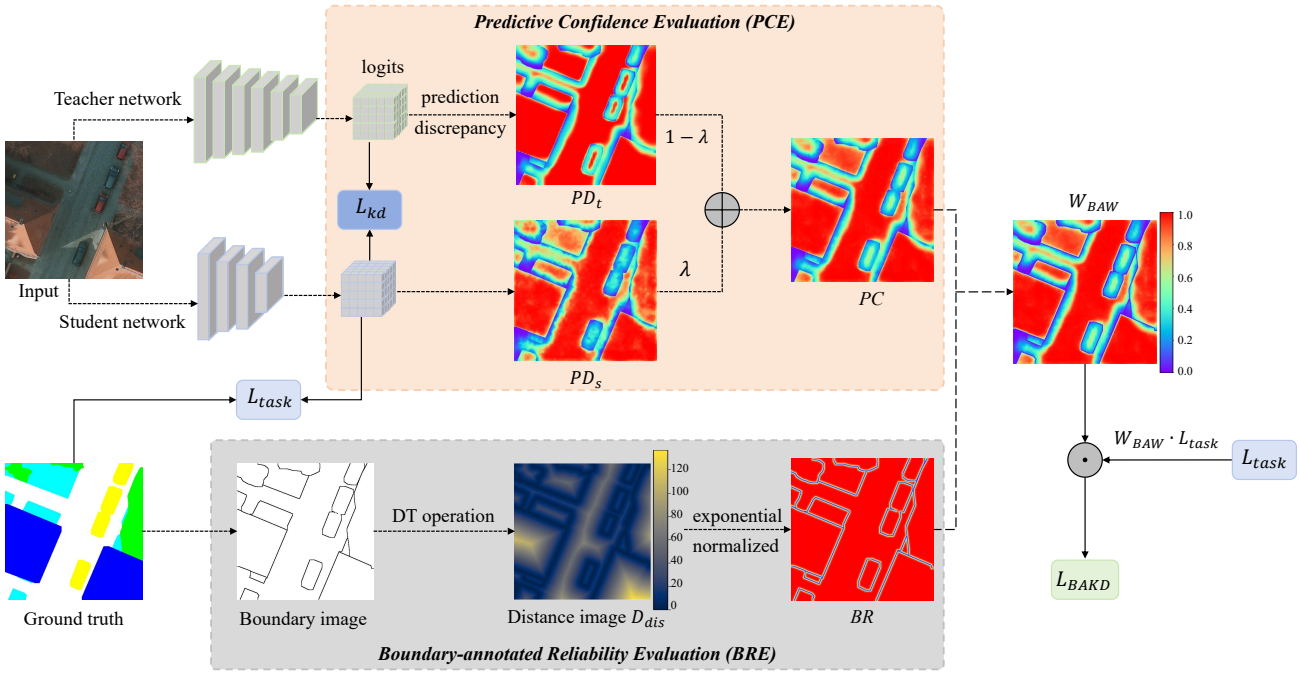


Fig. 2. Overview of our BAKD method. The PCE strategy collaboratively evaluates the prediction discrepancy between the teacher and student networks to obtain the predictive confidence score PC for each sample. The BRE strategy employs the distance transformation (DT) operation to calculate the distance D_{dis} from each annotation to the semantic boundary. Subsequently, our designed parameterized exponential normalization process determines the boundary-annotated reliability score BR. Finally, the combination of PC and BR yields the comprehensive boundary-aware weight W_{BAW}.

a confidence indicator. Higher prediction discrepancies usually suggest potentially noisy annotations. However, relying solely on the teacher network's prediction discrepancy to evaluate annotation's uncertainty overlooks the student network's perspective and cognitive processes. Just as a question deemed easy by a teacher can be challenging for some students. Therefore, incorporating the student's predictions to refine the annotation's evaluation criteria becomes essential.

Based on this motivation, we propose a teacher-student cooperation method to evaluate the annotation's uncertainty and dynamically adjust the collaboration weights between the teacher and student networks as the student network's cognitive levels increase. Initially, for each pixel in the image, the cross entropy is used to calculate the prediction discrepancy between predictions $\mathbf{Z}$ and the corresponding annotation $\mathbf{y}$ of the teacher and the student networks, respectively:

$$PD_s = -\sum_{c=1}^{C} \mathbf{y}_{h,w}^{(c)} \cdot \log\left(\sigma\left(\mathbf{Z}_{s(h,w)}^{(c)}\right)\right) \tag{6}$$

$$PD_t = -\sum_{c=1}^{C} \mathbf{y}_{h,w}^{(c)} \cdot \log\left(\sigma\left(\mathbf{Z}_{t(h,w)}^{(c)}\right)\right) \tag{7}$$

where C denotes the total number of classes. $\sigma\left(\mathbf{Z}_{s(h,w)}^{(c)}\right)$ and $\sigma\left(\mathbf{Z}_{t(h,w)}^{(c)}\right)$ are probabilities that the (h,w)-th pixel belongs to the c-th class for the student and teacher networks respectively.

Subsequently, the predictive confidence score PD for each annotation is derived by dynamically weighting the prediction discrepancy between the teacher and student networks:

$$PD = \lambda \cdot PD_s + (1-\lambda) \cdot PD_t \tag{8}$$

where λ is a weight adjustment parameter. Selecting an appropriate λ value is critical. Intuitively, the student network has not yet converged in the early stages of model training. Relying on the student network to evaluate predictive confidence may transmit and amplify errors, posing challenges to correcting the student network [48]. Conversely, the teacher network's supervisory information tends to be more precise and dependable. Therefore, we introduce a warm-up strategy to ensure the stability and reliability of the confidence evaluation. In practice, we set the initial λ value to 0. As training progresses, the student network's performance has dramatically improved, enabling it to accurately grasp its needs and play a more significant role in predictive confidence evaluation. Therefore, λ gradually increases, indicating that the student network is constantly improving. This also means that the teacher network will continue to play a role in the later training stage even if the student network performs well.

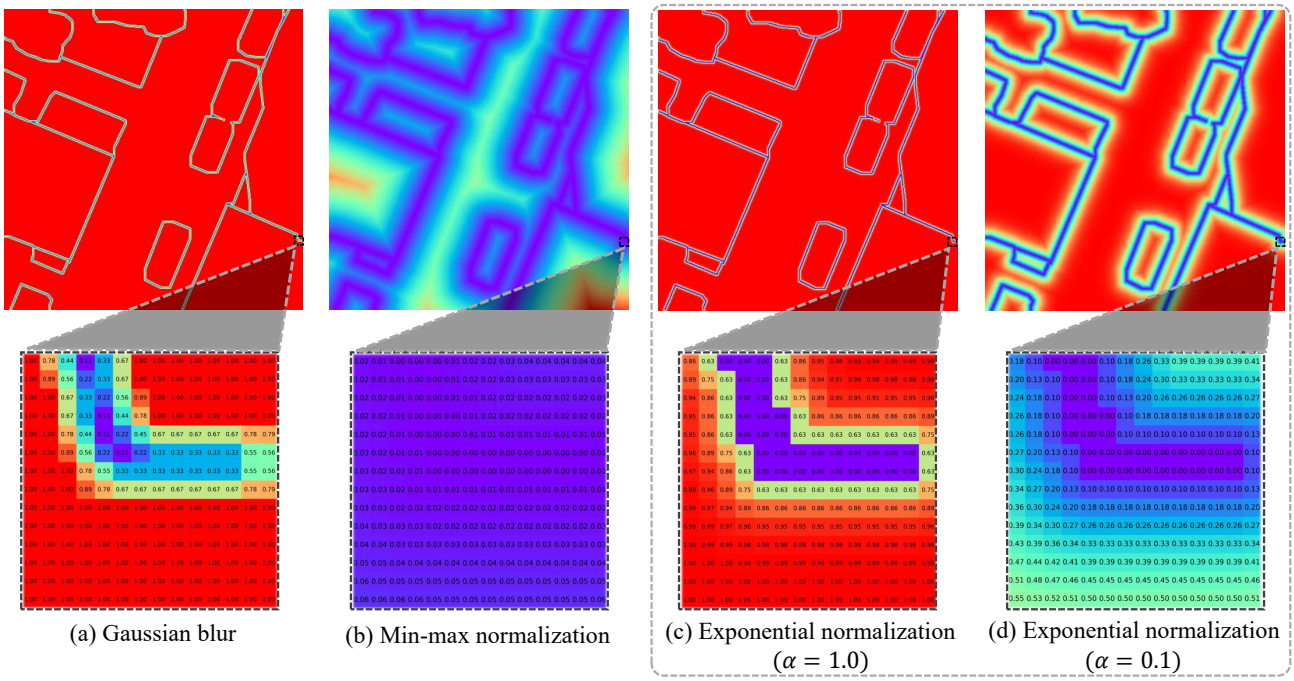
Subsequently, normalizing the confidence scores within the range of [0,1] yields a predictive confidence score PC for each sample:

$$PC = \exp\{-PD\} \tag{9}$$

The PC map in Fig. 2 reveals that low-confidence pixels cluster near the semantic boundary, illustrating PC's sensitivity to annotation uncertainty in boundary regions. Algorithm 1 provides the pseudo-code illustrating PCE strategy to generate confidence score PC in the overall training pipeline.

D. Boundary-annotated Reliability Evaluate (BRE) Strategy

Discussion: In annotating remote sensing images, identifying noisy annotations near the semantic boundary poses a



(a) Gaussian blur (b) Min-max normalization (c) Exponential normalization ($\alpha = 1.0$) (d) Exponential normalization ($\alpha = 0.1$)

Fig. 3. Utilizing the exponential normalization method on the boundary map of annotated images to generate annotated reliability scores offers distinct advantages. Adjusting the parameter α based on the noise level present at the dataset's boundary makes it possible to control the distribution of scores in the boundary regions flexibly. Specifically, when α =1.0, the reliability scores for smaller regions near the boundary decrease (c). When α = 0.1, the reliability scores also decrease for larger regions near the boundary (d).

Algorithm 1 Predictive Confidence Evaluation (PCE)

Input: Input images x, labels y, the parameter of teacher network θ_t, the iterations of warm-up iteration $iter_{warm-up}$, cross-entropy loss function CE, Kullback-Leibler divergence function KL, Softmax function σ, weight adjustment parameter λ.
Output: Predictive confidence score PC.

if $iter \leq iter_{warm-up}$ **then**
  $PD_t = -\sum_{c=1}^{C} \mathbf{y}_{h,w}^{(c)} \cdot \log\left(\sigma\left(\mathbf{Z}_{t(h,w)}^{(c)}\right)\right)$
  $PC = \exp\{-PD_t\}$
else
  $PD_s = -\sum_{c=1}^{C} \mathbf{y}_{h,w}^{(c)} \cdot \log\left(\sigma\left(\mathbf{Z}_{s(h,w)}^{(c)}\right)\right)$
  $PD_t = -\sum_{c=1}^{C} \mathbf{y}_{h,w}^{(c)} \cdot \log\left(\sigma\left(\mathbf{Z}_{t(h,w)}^{(c)}\right)\right)$
  $PD = \lambda \cdot PD_s + (1-\lambda) \cdot PD_t$
  $PC = \exp\{-PD\}$
end if
return PC

significant challenge. The closer to the semantic boundary, the lower the annotated reliability. We explore three distinct methods to evaluate annotated reliability based on the boundary information of the annotation map:

(1) Gaussian blur: Initially, we apply Gaussian blurring to the binary boundary map to derive an annotated reliability map. For the boundary map, only the pixel value at the boundary is 0, and the rest of the pixel values are 1. After Gaussian smoothing, only the pixels closest to the boundary will be averaged to a smaller value. This method may not be ideal for datasets with significant noisy annotations. In addition, Gaussian smoothing will blur the noise and inadequately capture the annotated reliability at the boundary position well, as shown in Fig. 3 (a). (2) Min-Max Normalization: We calculate the distance from each pixel to the boundary using the distance transform method to create a distance map. After normalization with min-max scaling, pixels near the boundary receive lower weights, as shown in Fig. 3 (b). This approach may not be suitable for datasets with light boundary noise. In addition, too many pixels are assigned lower weights, leading to insufficient training samples and causing overfitting problems. (3) Exponential Normalization: Another more flexible method is to utilize the parameterized exponential function $1 - \exp\{-\alpha \cdot d_{h,w}\}$ to normalize the distance map, thereby obtaining the annotated reliability map. By adjusting the parameter α according to the boundary noise level, we can flexibly regulate the speed of weight decrease in the boundary area, as shown in Fig. 3 (c) and Fig. 3 (d).

After the experimental comparison, the parameterized exponential normalization method can better reflect the degree to which each annotation belongs to the semantic boundary (confirmed in subsequent experiments). Subsequently, we integrate this strategy into the knowledge distillation process. Specifically, as shown in Fig. 2, we first obtain the boundary image of the coarse annotation map, which can be directly obtained through a straightforward traversal algorithm that examines the pixel's neighborhood. Any pixel with a category annotation conflicting with its neighbors is flagged as a boundary pixel. Subsequently, the distance map D_{dis} is generated

Algorithm 2 Boundary-annotated Reliability Evaluation (BRE)

Input: Input images x, labels y, parameter α.
Output: Boundary-annotated reliability score BR.

Step 1: Generate Boundary Image
Initialize an empty boundary image e with the same size as y
for each pixel (h, w) in y **do**
  if any neighboring pixel of (h, w) has a different label **then**
    Set $e(h, w) = 1$ {Mark as boundary pixel}
  else
    Set $e(h, w) = 0$ {Mark as non-boundary pixel}
  end if
end for
Step 2: Calculate Distance Image
$d_{h,w}$ = DistanceTransform(e) {Compute distance from the nearest boundary for the h,w-pixel on label y}
Step 3: Calculate Boundary-annotated Reliability
$BR = 1 - \exp\{-\alpha \cdot d_{h,w}\}$
return BR

by calculating the Euclidean distance from each pixel to the nearest boundary pixel using the distance transformation formula. D_{dis} effectively represents the distance from each annotation to the nearest semantic boundary.

Following this, we introduce the boundary-annotated reliability evaluation function BR based on the exponential normalization to characterize this relationship:

$$BR = 1 - \exp\{-\alpha \cdot d_{h,w}\} \tag{10}$$

where $d_{h,w}$ represents the Euclidean distance from pixel h, w to the nearest boundary, with α serving as a parameter dictating the function's shape. The design of this function follows the following principles: When pixel h, w is distant from the boundary, $d_{h,w}$ is large, and W_{dis} approaches 1, signifying high annotated reliability. When pixel h, w is near the boundary, $d_{h,w}$ diminishes, and W_{dis} approaches 0, indicating reduced annotated reliability. α controls the diffusion range of boundary uncertainty and can be adjusted according to actual conditions. In Fig. 2, the boundary-annotated reliability score BR of pixels near the boundary is small, denoting pronounced uncertainty in their annotations. Algorithm 2 provides the pseudo-code illustrating BRE strategy to generate boundary-annotated reliability score BR.

The PCE strategy combines the prediction confidence scores of the student and teacher networks. As the student's cognitive level improves, the collaborative weight of the student and teacher networks is dynamically adjusted to obtain a more accurate prediction confidence score PC. The BRE strategy uses the edge information of the coarse annotation and designs a parameterized exponential function normalization method, which can dynamically adjust the boundary reliability score BR of each sample according to the noisy degree of the semantic boundary. In our method, we combine the predictive confidence score PC and the boundary-annotated reliability score BR to obtain the comprehensive boundary-aware weight W_{BAW}. Notably, BR exclusively influences the weights of pixels near the boundary, with those closer to the object's center maintaining weights close to 1. If PC and BR are combined using average or proportional weighting, the BR will weaken the effect of the PC for pixels close to the object's center. We opt for the minimum value between the two as the final boundary-aware weight W_{BAW}:

$$W_{BAW} = \min(PC, BR) \tag{11}$$

W_{BAW} can ensures that BR predominates near the semantic boundary while enabling PC to have a stronger impact elsewhere. By selecting the minimum value, we can give full play to the advantages of the two uncertainty evaluation strategies and effectively mitigate the negative impact of noise annotations on model training.

Subsequently, we adopt the method of assigning low weights to mitigate the negative impact of these samples on the student network and apply the weight W_{BAW} to the loss function L_{task} of the semantic segmentation task:

$$L_{weight} = W_{BAW} \cdot L_{task} \tag{12}$$

Finally, we employ the BAKD strategy to calculate the weighted segmentation loss $L_{weighted}$ for each pixel, guiding the student network's parameter updates by minimizing L_{BAKD}:

$$L_{BAKD} = L_{weighted} + L_{kd} = W_{BAW} \cdot L_{task} + L_{kd} \tag{13}$$

E. Integrating with Other Approaches

Since BAKD only affects L_{task} loss, it can be integrated with other KD methods without introducing additional optimization goals. In experiments, we integrate BAKD with AT [30], CWD [29], DSD [28] and CIRKD [27] methods. Taking the AT [30] and CIRKD [27] methods as examples, we will demonstrate how to integrate BAKD into AT and CIRKD and derive the corresponding distillation loss formula.

1) The distillation loss of AT method after integrating BAKD.

The total loss defined by AT [30]:

$$L_{AT} = L_{task} + \frac{\beta}{2} \sum_{j \epsilon I} || \frac{Q_S^j}{||Q_S^j||_2} - \frac{Q_T^j}{||Q_T^j||_2} ||_2 \tag{14}$$

where Q_S^j and Q_T^j are respectively the j-th pair of student and teacher attention maps in vectorized form, $\frac{Q_S^j}{||Q_S^j||_2}$ and $\frac{Q_T^j}{||Q_T^j||_2}$ are the result of using l_2-normalization attention maps. The calculation details of $\frac{\beta}{2} \sum_{j \epsilon I} || \frac{\mathbf{Q}_S^j}{||\mathbf{Q}_S^j||_2} - \frac{\mathbf{Q}_T^j}{||\mathbf{Q}_T^j||_2} ||_2$ are described in AT [30]. L_{task} is the semantic segmentation task loss function mentioned in Section III-A, represented by the cross entropy function.

The distillation loss of BAKD :

$$L_{BAKD} = W_{BAW} \cdot L_{task} + L_{kd} \tag{15}$$

where L_{kd} is the pixel-level distillation loss mentioned in Section III-A, represented by Kullback-Leibler (KL) divergence.

The distillation loss of AT [30] method after integrating BAKD can be derived:

$$L_{AT\_BAKD} = W_{BAW} \cdot L_{task} + \frac{\beta}{2} \sum_{j \epsilon I} || \frac{Q_S^j}{||Q_S^j||_2} - \frac{Q_T^j}{||Q_T^j||_2} ||_2 \tag{16}$$

In the new loss term described above, we follow the default parameter settings in AT [30] and set the weighting parameter β to 10^3 divided by the number of elements in the attention map and batch size for each layer. We only modify the weights for each pixel by incorporating the relative difficulty factors obtained through BAKD in front of the L_{task}.

2)The distillation loss of CIRKD method after integrating BAKD.

The total loss defined by CIRKD [27]:

$$L_{CIRKD} = L_{task} + L_{kd} + \alpha L_{batch\_p2p} + \beta L_{memory\_p2p} + \gamma L_{memory\_p2r} \tag{17}$$

where L_{batch_p2p} represents distillation loss of mini-batch-based pixel-to-pixel, L_{memory_p2p} denotes distillation loss of memory-based pixel-to-pixel, L_{memory_p2r} denotes distillation loss of memory-based pixel-to-region. α, β and γ are the weight balance parameters. The further calculation details of L_{batch_p2p}, L_{memory_p2p} and L_{memory_p2r} are described in [27].

The distillation loss of BAKD :

$$L_{BAKD} = W_{BAW} \cdot L_{task} + L_{kd} \tag{18}$$

After integrating BAKD, the modified distillation loss of the CIRKD method is derived as follows:

$$L_{CIRKD\_BAKD} = W_{BAW} \cdot L_{task} + L_{kd} + \alpha L_{batch\_p2p} + \beta L_{memory\_p2p} + \gamma L_{memory\_p2r} \tag{19}$$

In this new loss term, we follow the default parameter settings in CIRKD [27], setting the weighting parameter α to 1, β to 0.1, and γ to 0.1. Moreover, BAKD can seamlessly integrate with other semantic segmentation methods based on KD. This integration allows us to enhance the performance of the student network further from existing approaches.

IV. Experiments and Results Analysis

To verify BAKD's effectiveness, we conduct extensive experiments on RSI semantic segmentation datasets ISPRS Potsdam and ISPRS Vaihingen. Subsequently, we compare BAKD with other methods and summarize our experiment.

A. Experimental Setup

1) Datasets: We validate the proposed method by performing experiments on high-resolution aerial images of two German cities, Vaihingen and Potsdam, acquired through flight missions provided by the 2-D Semantic Annotation Challenge organized by Working Group II/4 of the International Society for Photogrammetry and Remote Sensing (ISPRS) [49]. The ISPRS Vaihingen dataset consists of 33 images, and each image has approximately 2100×2100 pixels and a spatial resolution of 9 cm. Each image has three bands, corresponding to near-infrared (NIR), red (R) and green (G) wavelengths. Each pixel in the image is classified into one of 6 land cover categories (buildings, cars, low vegetation, impervious surfaces, trees, and clutter/background). We select 16 images for training, 8 images for validation, and 9 images for testing. The ISPRS Potsdam dataset consists of 38 high-resolution aerial images, each with a pixel size of 6000×6000 and a spatial resolution of 5 cm. All images are annotated with the same 6-category pixel-level labels as the Vaihingen dataset. We select 24 images for training, 7 images for validation, and 7 images for testing our model. In our experiments, we use a 512×512 patch size, which fits our memory budget. Since CNN-based semantic segmentation models are prone to boundary effects, we use 256-pixel overlapping patches to reduce this effect.

2) Simulation of Annotation Errors: We simulate the types and quantities of coarse annotations in the training data to analyze the impact of specific errors. The baseline model is based on the original training labels from the dataset, assuming they are correct. This allows us to compare our baseline with current state-of-the-art results and further investigate the effects of simulated annotation errors on model performance. Our goal is to generate labels that closely resemble human annotation errors [34], which are shown in Fig. 4, and below are the coarse annotations obtained through three different methods:

A. Mask Dilation (Subsequent experiments were mainly based on this noise type):

1. Randomly Generate Dilation Iterations: Randomly generate dilation iterations suited to the characteristics of different classes. In the noisy supervised Vaihingen and Potsdam datasets, the dilation iterations for large classes (such as Building, Low vegetation, Impervious surface, and Tree) are randomly assigned an integer range of 3 to 10. For small object classes (such as Car and Clutter), the iterations are randomly assigned an integer range of 2 to 5. In the severely noisy supervised Vaihingen and Potsdam datasets, the dilation iterations for large classes are randomly assigned an integer range of 7 to 15. For small object classes, the iterations are randomly assigned an integer range of 4 to 6.

2. Perform Dilation Operation: Apply the specified iteration counts to the foreground masks of each class to expand the class boundaries.

B. Mask Erosion:

1. Randomly Generate Erosion Iterations: Randomly generate erosion iterations based on the characteristics of different classes.The randomly set erosion iterations are consistent with the dilation in **A. Mask Dilation.**

2. Perform Erosion Operation: Apply the specified iteration counts to the foreground masks of each class to reduce the class boundaries.

C. Relative Shift of Object Boundaries:

1. Generate Random Displacement: Generate random displacements in both horizontal and vertical directions to de-

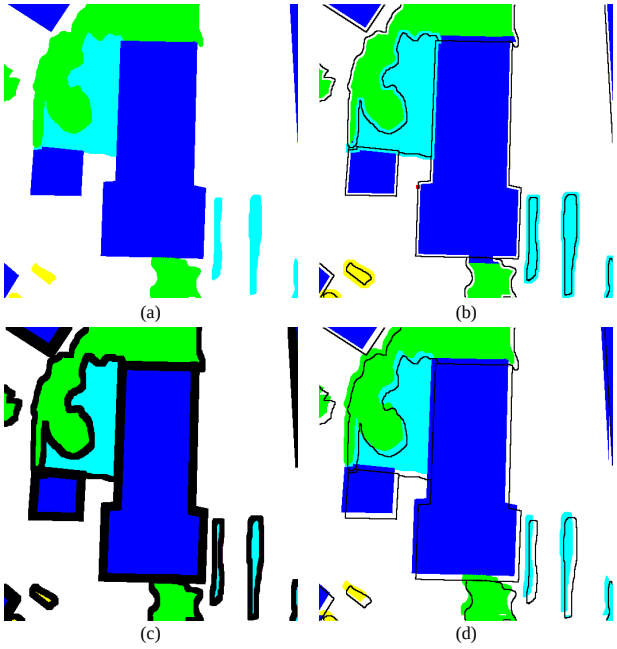


Fig. 4. Annotations before and after introducing different annotation noise. (a) Original annotation, (b) Annotation corrupted with random dilation, (c) Annotation corrupted with random erosion, (d) Annotation corrupted with shift.

termine the new position of each pixel. In the experiment, we randomly select two integers within the ranges of -10 to -5 and 5 to 10, respectively, as the horizontal and vertical displacement amounts (we exclude the values from -5 to 0 and 0 to 5 due to their relatively small offsets).

2. Apply Displacement and Construct New Image: Calculate the new coordinates after displacement and copy the original pixel values to the new image; if new coordinates exceed the boundaries, retain the original pixel values.

3) Network architecture: We employ DeepLabV3 [10] with ResNet-101 backbone [50] as the teacher network for all experiments. For student networks, we employ various segmentation architectures to verify the effectiveness of BAKD. Specifically, DeepLabV3 with ResNet-18 backbone, PSPNet [12] with ResNet-18 backbone and DeepLabV3 with MobileNetV2 backbone [51] are adopted. ResNet-101, ResNet-18, and MobileNetV2 backbone networks were all pre-trained on the ImageNet [52] dataset.

4) Training strategy: Our framework is implemented in PyTorch on two RTX 3090 GPUs. The networks are trained using mini-batch stochastic gradient descent (SGD) with a momentum of 0.9 and weight decay of 0.0005. We set the number of iterations to 40,000. The learning rate is initialized at 0.02 and is multiplied by $(1 - \frac{iter}{iter_{total}})^{0.9}$ during training. Normal data augmentation techniques such as random flipping and scaling in the $[0.5, 2]$ range are applied during training. The temperature T is set to be 1. The batch size is 16, and all experiments are conducted using mixed-precision training. For the noisy supervised datasets, we conduct ablation studies to determine the optimal values for the parameter α. Specifically, we set $\alpha = 0.8$ for the noisy supervised Potsdam dataset and $\alpha = 0.8$ for the noisy supervised Vaihingen dataset. For original annotated datasets and seriously noisy supervised datasets, we set α to 1.0, as we do not conduct an ablation study for these cases. Additionally, the parameter λ is initialized at 0 for the first 4,000 iterations and linearly increased to 0.5 over the subsequent iterations.

5) Evaluation metrics: Following the standard setting, we adopt two commonly used metrics to evaluate the performance of different methods, including the mIoU (mean intersection over union) and mF_1 (mean F_1).

1. Mean Intersection Over Union: mIoU is one of the main indicators of semantic segmentation model performance, which is defined as the average of the intersection over union (IoU) of each category. Assuming that true positives (TP), false negatives (FP), and false negatives (FN) represent the number of pixels correctly classified, misclassified, and missed, respectively, mIoU is defined as:

$$\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FP_c + FN_c} \tag{20}$$

where C is the number of classes.

2. mF_1: mF_1 also considers TP_c, FP_c, and FN_c, but it focuses on the harmonic mean of precision and recall to mitigate the effects of imbalanced class distributions. mF_1 is calculated as:

$$mF1 = \frac{1}{C}\sum_{c=1}^{C}\frac{2 \times TP_c}{2 \times TP_c + FP_c + FN_c} \tag{21}$$

B. Performance

1) Performance Comparisons with Existing Methods on Noisy Supervised Datasets: To verify the performance of BAKD on a noisy supervised dataset, we compare it with three related methods: 1. Mainstream semantic segmentation methods based on KD (SSKD), including KD [24], AT [30], CWD [29], DSD [28], CIRKD [27] and RDD [48]. In experiments, we adopt DeepLabV3 with ResNet-101 backbone as the teacher network, denoted as "Teacher", and DeepLabV3 with ResNet-18 backbone as the student network, denoted as "Student". 2. Classification from Noisy Annotations with Co-Learning (CNACL) method, including decoupling [21], co-teaching [22] and co-teaching+ [23]. For a fair comparison, in the experiment, we had the same teacher and student network architecture as SSKD. 3. Semantic segmentation from noisy annotations (SSNA), the typical ADELE [40] and RMD [41] are employed as comparison methods. In the experiments, we employ DeepLabV3 with ResNet-18 backbone as the segmentation model.

For noisy supervised Vaihingen dataset. To validate the performance of mIoU and mF_1, we evaluate BAKD on the noisy supervised Vaihingen dataset, as illustrated in Table I. For classification from noisy annotations with the Co-Learning (CNACL) method, the mutual supervision between two models during training helps mitigate the impact of noisy annotations to a certain extent. However, CNACL relies on image-level noise filtering, and in noisy annotation semantic segmentation, noisy annotations usually appear in local areas without

TABLE I
QUANTITATIVE SEGMENTATION RESULTS FOR OUR PROPOSED BAKD AND OTHER EXISTING METHODS ON THE NOISY SUPERVISED VAIHINGEN DATASET, SHOWCASING IOU SCORES FOR THE FIVE CLASSES, ALONG WITH THE MIOU SCORE AND mF_1 SCORE. THE BEST SCORE IN EACH COLUMN IS HIGHLIGHTED IN **BOLD**, WHEREAS THE SECOND-BEST SCORE IS IN UNDERLINE. VALUES WITHIN PARENTHESES INDICATE THE PERFORMANCE VARIANCE RELATIVE TO THE BASELINE MODEL (STUDENT).

Method	IoU					mIoU (%)	mF_1 (%)
	Imp.Surf.	Building	Low veg	Tree	Car		
Classification from Noisy Annotations with Co-Learning (CNACL)							
+Decoupling [21]	73.27	76.45	57.47	69.60	39.03	63.16 (↓ 1.83)	71.20 (↓ 2.69)
+Co-teaching [22]	**74.56**	77.84	**58.42**	**70.05**	45.07	65.19 (↑ 0.20)	72.78 (↓ 1.11)
+Co-teaching+ [23]	73.02	76.40	52.79	68.65	36.58	61.48 (↓ 3.51)	70.12 (↓ 3.77)
Semantic segmentation from noisy annotations (SSNA)							
ADELE [40]	74.37	**81.47**	57.00	69.39	54.64	67.37 (↑ 2.38)	75.98 (↑ 2.09)
RMD [41]	73.01	79.27	56.27	69.64	55.13	66.66 (↑ 1.67)	75.23 (↑ 1.34)
Semantic segmentation based on KD (SSKD)							
Teacher	74.18	79.46	57.47	69.34	63.52	68.79	77.04
Student	71.16	77.03	52.03	66.01	58.76	64.99	73.89
+KD [24]	72.30	77.96	53.40	66.71	61.31	66.34 (↑ 1.35)	74.96 (↑ 1.07)
+AT [30]	72.39	77.50	55.12	68.48	**62.14**	67.13 (↑ 2.14)	75.63 (↑ 1.74)
+CWD [29]	72.59	77.35	54.63	67.17	58.96	66.14 (↑ 1.15)	75.03 (↑ 1.14)
+DSD [28]	72.17	78.16	54.28	67.01	61.73	66.67 (↑ 1.68)	75.11 (↑ 1.22)
+CIRKD [27]	73.41	78.62	55.47	68.05	60.61	67.23 (↑ 2.24)	75.83 (↑ 1.94)
+RDD [48]	73.50	78.53	55.00	67.87	61.78	67.34 (↑ 2.35)	75.85 (↑ 1.96)
+Ours	73.85	79.81	58.34	69.80	58.91	**68.14 (↑ 3.15)**	**76.37 (↑ 2.48)**

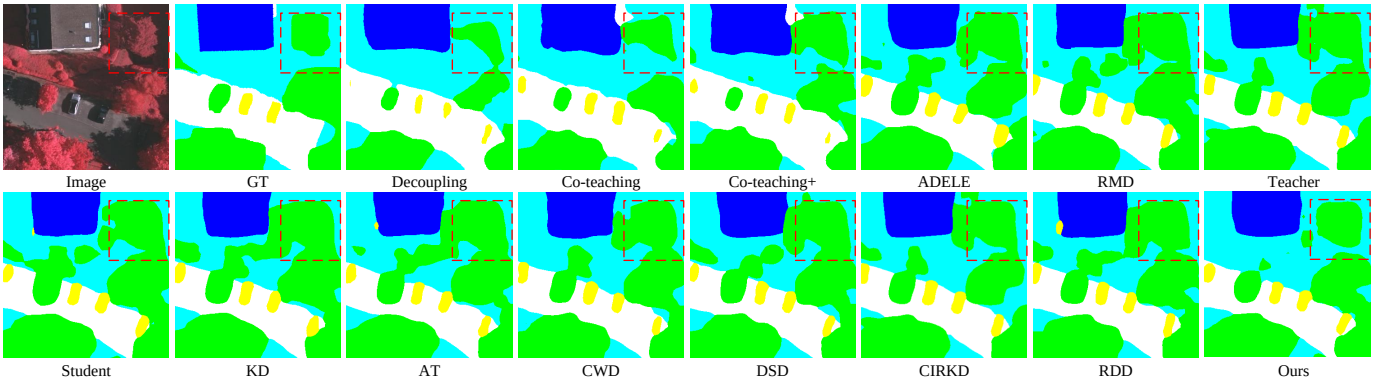


Fig. 5. Qualitative segmentation results for our proposed BAKD and other existing methods on the noisy supervised Vaihingen dataset. Legend—white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, yellow: cars, red: clutter/background. The red dashed boxes mark some areas where the semantic segmentation result is obviously optimized. The semantic labels produced by BAKD are more consistent with the ground truth (GT). (Best viewed in color.)

affecting the entire image. Consequently, discarding entire images due to noise can lead to a scarcity of training samples, potentially resulting in overfitting and diminished model performance, notably in the segmentation of minority categories. For instance, considering categories with a limited number of instances like "Car", excluding these samples during training results in fewer instances of the "Car" category. This reduction notably decreases the mIoU for the "Car" category across all three CNACL methods compared to the baseline. The ADELE method significantly enhances segmentation performance by pixel-level correction of noisy annotations. Although ADELE has also made great progress, our BAKD performs better. Compared with ADELE, our BAKD improves the mIoU and mF_1 indicators by 0.77% and 0.39%, respectively. The RMD method uses the collaboration of two segmentation models to filter out label noise from coarse annotations and improve the segmentation performance. Compared with RMD, the BAKD method improves the mIoU and mF_1 indicators by 1.48% and 1.14% respectively. For Mainstream semantic segmentation methods based on KD (SSKD) method, all structured KD methods improve the student network's segmentation performance compared to training without KD. Our BAKD outperforms other KD methods regarding mIoU and mF_1 with significant advantages. Compared with the basic student network method without KD, BAKD improves the mIoU and mF_1 by 3.15%and 2.48% respectively. Compared with the CIRKD method, BAKD improves the mIoU and mF_1 by 0.91% and 0.54%, respectively. Compared with the state-of-the-art RDD method, BAKD improves the mIoU and mF_1 by 0.8% and 0.52%, respectively. In addition, the qualitative results are shown in Fig. 5. The validity of our BAKD is intuitively demonstrated, and the semantic labels produced by BAKD are more consistent with the ground truth.

For noisy supervised Potsdam dataset. We also evaluate BAKD on the noisy supervised Potsdam dataset, with the experimental outcomes are detailed in Table II. Due to the ample images in the Potsdam dataset, filtering image-level samples does not lead to severe overfitting. Consequently, all three CNACL methods alleviate the impact of noisy annota-

TABLE II
QUANTITATIVE SEGMENTATION RESULTS FOR OUR PROPOSED BAKD AND OTHER EXISTING METHODS ON THE NOISY SUPERVISED POTSDAM DATASET, SHOWCASING IOU SCORES FOR THE FIVE CLASSES, ALONG WITH THE MIOU SCORE AND mF_1 SCORE. THE BEST SCORE IN EACH COLUMN IS HIGHLIGHTED IN **BOLD**, WHEREAS THE SECOND-BEST SCORE IS IN UNDERLINE. VALUES WITHIN PARENTHESES INDICATE THE PERFORMANCE VARIANCE RELATIVE TO THE BASELINE MODEL (STUDENT).

Method	IoU					mIoU (%)	mF_1 (%)
	Imp.Surf.	Building	Low veg	Tree	Car		
Classification from Noisy Annotations with Co-Learning (CNACL)							
+Decoupling [21]	79.63	87.14	70.26	70.17	69.02	75.24 (↑ 1.35)	79.64 (↑ 0.80)
+Co-teaching [22]	78.72	86.09	69.31	69.32	67.85	74.26 (↑ 0.37)	79.21 (↑ 0.37)
+Co-teaching+ [23]	79.76	87.52	70.65	70.89	70.01	75.77 (↑ 1.88)	79.81(↑ 0.97)
Semantic segmentation from noisy annotations (SSNA)							
ADELE [40]	82.38	**91.11**	72.67	73.12	65.35	76.73 (↑ 2.84)	81.02 (↑ 2.18)
RMD [41]	**82.74**	90.72	71.58	73.04	64.37	76.49 (↑ 2.50)	80.82 (↑ 1.98)
Semantic segmentation based on KD (SSKD)							
Teacher	80.75	89.04	72.52	74.44	69.30	77.21	81.27
Student	79.81	88.36	69.90	70.33	61.54	73.99	78.84
+KD [24]	81.16	89.38	71.77	71.97	67.99	76.45 (↑ 2.56)	80.86 (↑ 2.02)
+AT [30]	81.38	88.94	72.07	72.38	70.10	76.97 (↑ 3.08)	80.98 (↑ 2.14)
+CWD [29]	80.16	88.65	71.02	71.80	70.45	76.42 (↑ 2.53)	80.84 (↑ 2.00)
+DSD [28]	80.02	88.62	71.00	72.08	67.23	75.79 (↑ 1.90)	79.89 (↑ 1.05)
+CIRKD [27]	81.76	89.55	72.50	72.45	70.13	77.28 (↑ 3.39)	81.24 (↑ 2.40)
+RDD [48]	81.24	89.18	72.25	73.64	69.43	77.15 (↑ 3.16)	81.11 (↑ 2.27)
+Ours	81.70	89.28	**72.73**	**73.91**	**70.76**	**77.68 (↑ 3.79)**	**81.67 (↑ 2.83)**

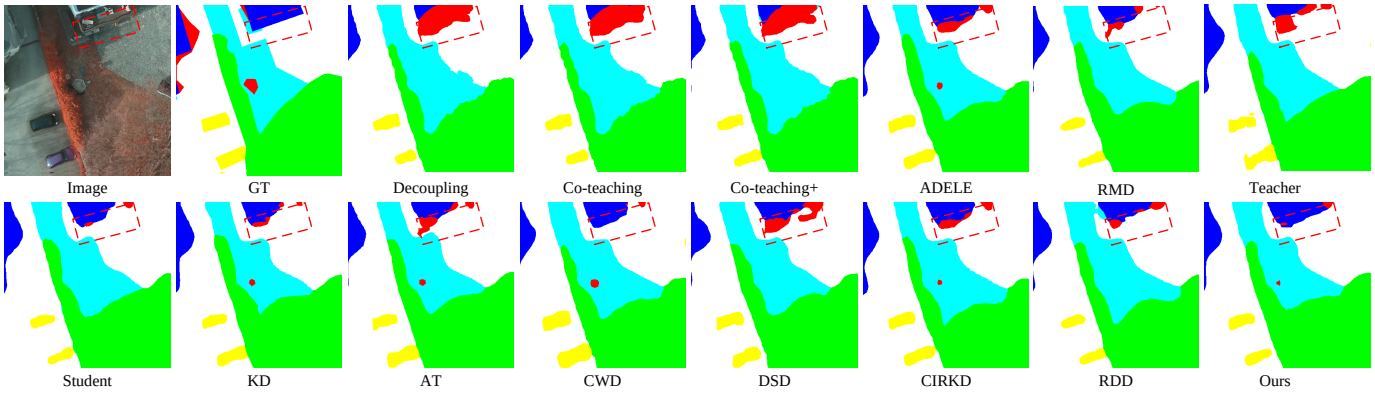


Fig. 6. Qualitative segmentation results for our proposed BAKD and other existing methods on the noisy supervised Potsdam dataset. The red dashed boxes mark some areas where the semantic segmentation result is obviously optimized. The semantic labels produced by BAKD are more consistent with the ground truth (GT). (Best viewed in color.)

tions to some extent, thereby enhancing model performance. However, their performance falls short of semantic segmentation methods based on KD (SSKD). Although the two SSNA methods achieved significant progress compared to the basic student network, they are still inferior to our proposed BAKD. Compared with ADELE, BAKD improves the mIoU and mF_1 by 0.95% and 0.65% respectively. Compared with RMD, BAKD improves the mIoU and mF_1 by 1.19% and 0.85% respectively. For the SSKD method, BAKD outperforms other KD methods, further validating its effectiveness in semantic segmentation tasks. Specifically, compared with the basic student network method without KD, BAKD improves the mIoU and mF_1 by 3.69% and 2.83%, respectively. Compared with CIRKD method, BAKD improves the mIoU and mF_1 by 0.4% and 0.43%, respectively. Compared with the state-of-the-art RDD method, BAKD improves the mIoU and mF_1 by 0.53% and 0.56%, respectively. In addition, the qualitative results are shown in Fig. 6, highlighting semantic labels that align more closely with ground truth.

2) Performance Comparisons on seriously noisy supervised datasets: To further verify the robustness of BAKD under noisy annotations, we perform more severe erosion and dilation operations on the Vaihingen and Potsdam datasets to generate datasets with seriously noisy annotations and conducted experimental verification on these datasets.

For seriously noisy supervised Vaihingen dataset. We evaluated our BAKD on the seriously noisy, supervised Vaihingen dataset. The experimental results are shown in Table III. In the CNACL method, due to the serious noise in the annotation, screening the training data excludes a large number of images. Despite filtering out more noisy samples, severe overfitting in the CNACL method led to only marginal improvements compared to the baseline method, with performance even decreasing in some categories. Specifically, during the training process of Co-teaching, approximately half of the data was discarded, resulting in an 8.62% decrease in the IoU of the car category compared to the baseline method (Student). Compared to Decoupling, BAKD improved the IoU for the car category by 8.96%. Compared to Co-teaching, BAKD enhanced the IoU for the car category by 18.41%. Additionally,

TABLE III
QUANTITATIVE SEGMENTATION RESULTS FOR OUR PROPOSED BAKD AND OTHER EXISTING METHODS ON THE SERIOUSLY NOISY SUPERVISED VAIHINGEN DATASET, SHOWCASING IOU SCORES FOR THE FIVE CLASSES, ALONG WITH THE MIOU SCORE AND mF_1 SCORE. THE BEST SCORE IN EACH COLUMN IS HIGHLIGHTED IN **BOLD**, WHEREAS THE SECOND-BEST SCORE IS IN UNDERLINE. VALUES WITHIN PARENTHESES INDICATE THE PERFORMANCE VARIANCE RELATIVE TO THE BASELINE MODEL (STUDENT).

Method	IoU					mIoU (%)	mF_1 (%)
	Imp.Surf.	Building	Low veg	Tree	Car		
Classification from Noisy Annotations with Co-Learning (CNACL)							
+Decoupling [21]	68.84	74.80	55.32	63.80	29.90	58.53 (↑ 0.86)	67.60 (↑ 0.67)
+Co-teaching [22]	67.63	70.02	46.45	55.91	20.45	52.09 (↓ 5.58)	62.89 (↓ 4.04)
+Co-teaching+ [23]	67.93	75.52	**56.85**	64.04	25.90	58.05 (↑ 0.38)	67.00 (↑ 0.07)
Semantic segmentation from noisy annotations (SSNA)							
ADELE [40]	66.43	75.56	49.17	62.59	40.13	58.78 (↑ 1.11)	68.03 (↑ 1.10)
RMD [41]	68.50	74.37	53.46	62.89	38.29	59.50 (↑ 1.83)	68.21 (↑ 1.28)
Semantic segmentation based on KD (SSKD)							
Teacher	70.45	77.01	57.86	64.25	41.56	62.23	70.42
Student	68.84	75.44	52.69	62.31	29.07	57.67	66.93
+KD [24]	68.32	75.58	53.82	62.33	35.97	59.20 (↑ 1.53)	68.18 (↑ 1.25)
+AT [30]	69.00	75.77	52.99	62.28	39.62	59.93 (↑ 2.26)	69.19 (↑ 2.26)
+CWD [29]	**69.63**	76.00	54.51	63.39	36.84	60.07 (↑ 2.40)	68.92 (↑ 1.99)
+DSD [28]	67.93	76.23	53.23	61.89	40.10	59.84 (↑ 2.17)	68.76 (↑ 1.83)
+CIRKD [27]	69.46	76.11	52.82	63.68	37.28	59.87 (↑ 2.20)	69.07 (↑ 2.14)
+RDD [48]	68.81	75.00	54.65	63.59	**43.19**	61.05 (↑ 3.38)	69.93 (↑ 3.00)
+Ours	69.06	**76.55**	56.58	**64.49**	38.86	**61.11 (↑ 3.44)**	**69.95 (↑ 3.02)**

TABLE IV
QUANTITATIVE SEGMENTATION RESULTS FOR OUR PROPOSED BAKD AND OTHER EXISTING METHODS ON THE SERIOUSLY NOISY SUPERVISED POTSDAM DATASET, SHOWCASING IOU SCORES FOR THE FIVE CLASSES, ALONG WITH THE MIOU SCORE AND mF_1 SCORE. THE BEST SCORE IN EACH COLUMN IS HIGHLIGHTED IN **BOLD**, WHEREAS THE SECOND-BEST SCORE IS IN UNDERLINE. VALUES WITHIN PARENTHESES INDICATE THE PERFORMANCE VARIANCE RELATIVE TO THE BASELINE MODEL (STUDENT).

Method	IoU					mIoU (%)	mF_1 (%)
	Imp.Surf.	Building	Low veg	Tree	Car		
Classification from Noisy Annotations with Co-Learning (CNACL)							
+Decoupling [21]	73.52	82.10	66.98	66.75	57.66	69.40 (↑ 0.76)	74.06 (↑ 0.38)
+Co-teaching [22]	74.91	83.02	63.46	63.55	57.31	68.45 (↓ 0.19)	73.49 (↓ 0.19)
+Co-teaching+ [23]	73.75	83.02	67.14	66.42	58.69	69.80 (↑ 1.16)	74.15 (↑ 0.47)
Semantic segmentation from noisy annotations (SSNA)							
ADELE [40]	76.01	86.54	67.22	72.91	61.18	72.77 (↑ 4.13)	77.27 (↑ 3.59)
RMD [41]	77.90	86.96	70.78	73.41	55.35	72.88 (↑ 4.24)	77.35 (↑ 3.67)
Semantic segmentation based on KD (SSKD)							
Teacher	76.11	86.06	69.64	73.40	61.33	73.91	77.91
Student	75.54	85.97	68.84	70.23	42.64	68.64	73.68
+KD [24]	78.46	**87.80**	70.74	71.38	57.97	73.27 (↑ 4.63)	77.43 (↑ 3.75)
+AT [30]	77.85	86.70	70.36	72.37	59.74	73.40 (↑ 4.76)	77.81 (↑ 4.13)
+CWD [29]	77.92	87.42	70.56	70.78	57.91	72.92 (↑ 4.28)	77.38 (↑ 3.70)
+DSD [28]	76.21	86.31	68.85	71.21	55.96	71.71 (↑ 2.57)	77.00 (↑ 3.32)
+CIRKD [27]	78.30	86.75	**71.48**	**73.44**	59.49	73.89 (↑ 5.25)	78.44 (↑ 4.76)
+RDD [48]	78.17	87.62	71.10	73.00	60.35	74.05 (↑ 5.41)	78.59 (↑ 4.91)
+Ours	**78.62**	87.08	71.24	73.28	**62.22**	**74.49 (↑ 5.85)**	**78.78 (↑ 5.10)**

BAKD achieved a 12.96% improvement in the IoU for the car category compared to Co-teaching+. ADELE significantly improves segmentation performance by correcting incorrect annotations at the pixel level. When the annotation noise is more serious, ADELE's early learning may not occur because there may not be enough information in the noisy annotations to correct the errors. As seen from Table III, in the case of severe noise, ADELE is even worse than the semantic segmentation based on the KD method. The confidence screening strategy in the RMD method may excessively remove small target categories with low confidence, such as Car. In the face of severe noise, it is not as good as the semantic segmentation method based on knowledge distillation. BAKD increased IoU in the car category by 0.57% compared to RMD. For Mainstream semantic segmentation methods based on KD (SSKD) method, all structured KD methods improve the segmentation performance of the student network. Our BAKD outperforms other KD methods regarding mIoU and mF_1 with significant advantages. Especially in the category of Car, due to the presence of boundary noise annotations, the IoU of Car is only 29.07%. However, by optimizing our BAKD method, the IoU of the Car category is improved by up to 9.79%. This result highlights the effectiveness of the BAKD method in solving the challenges posed by noisy annotations. This implies that BAKD's segmentation performance improvement would be more significant if the annotations for all categories were seriously noisier.

For seriously noisy supervised Potsdam dataset. We also evaluate our BAKD on the seriously noisy supervised Potsdam dataset. The experimental results are shown in Table IV. Since the Potsdam dataset is large in scale, filter-based methods (CNACL) are less likely to suffer from overfitting. CNACL

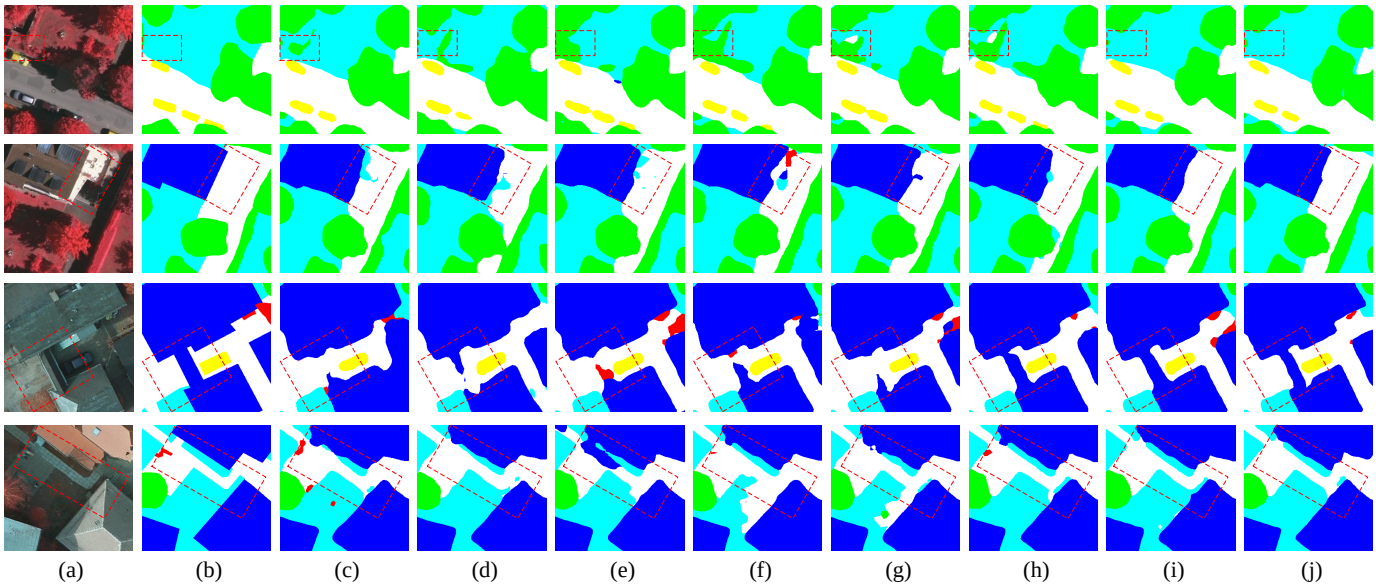


Fig. 7. Qualitative comparison for our proposed BAKD and other existing semantic segmentation methods based KD on the original annotated Vaihingen dataset (upper two rows) and Potsdam dataset (lower two rows). From left to right are (a) Input image, (b) Ground Truth, the results of (c) Student, (d) KD, (e) AT, (f) CWD, (g) DSD, (h) CIRKD, (i) BAKD (Ours), (j) Teacher, respectively. The relatively obvious areas of improvement are marked with red rectangles. BAKD performs better on the boundary areas.

can improve segmentation performance to a certain extent, but the improvement effect still has restrictions. For example, in the small target car category, BAKD improved the IoU of the car category by 4.56% compared to Decoupling. Compared to Co-teaching, BAKD enhanced the IoU of the car category by 4.91%. Additionally, BAKD achieved a 3.53% improvement in the IoU of the car category compared to Co-teaching+. Furthermore, due to RMD's exclusion of low-confidence small target classes during the filtering process, its improvement in the car category is not as significant as that of BAKD. BAKD achieved a 6.87% increase in the IoU of the car category compared to RMD. For Mainstream semantic segmentation methods based on KD (SSKD) method, all structured KD methods improve the segmentation performance of the student network. Our BAKD outperforms other KD methods regarding mIoU and mF_1 with significant advantages. For the Car category significantly affected by boundary noise, its mIoU was only 42.64%. However, by optimizing our BAKD method, the IoU of the Car category increased by 16.58%. This result again highlighted BAKD's excellent performance in dealing with noise labeling challenges.

C. Ablation Study

1) Ablation study about performance comparisons with mainstream KD methods on original annotated datasets: To verify the performance of BAKD on the original annotated datasets, we compare it with mainstream semantic segmentation methods based on KD, including KD [24], AT [30], CWD [29], DSD [28] and CIRKD [27] on the above two representative datasets. In the experiment, we employ DeepLabV3 with ResNet-101 backbone as the teacher network, denoted as "Teache", and DeepLabV3 with ResNet-18 backbone as the student model, denoted as "Student". The experimental outcomes are detailed in Table V. On both original annotated datasets, all structured KD methods improve the segmentation performance of the student network compared to training without KD. Notably, our BAKD exhibits superior performance in terms of mIoU and mF_1 metrics, showcasing significant advantages over other KD methods. On the original annotated Vaihingen dataset, compared with the CIRKD method, BAKD improves the mIoU and mF_1 by 0.45% and 0.19%, respectively. On the original annotated Potsdam dataset, compared with the CIRKD method, BAKD improves the mIoU and mF_1 by 0.29% and 0.15%, respectively. This further proves that our BAKD still performs well on fine manual annotated datasets. Furthermore, qualitative results are depicted in Fig. 7, illustrating the visual impact of our BAKD method.

TABLE V
ABLATION STUDY ABOUT THE PERFORMANCE COMPARISONS WITH MAINSTREAM KD METHODS ON ORIGINAL ANNOTATED DATASETS. THE BEST SCORE IN EACH COLUMN IS HIGHLIGHTED IN **BOLD**. OUR BAKD OUTPERFORMS OTHER KD METHODS REGARDING MIOU AND mF_1 WITH SIGNIFICANT ADVANTAGES.

Method	Original annotated Vaihingen dataset		Original annotated Potsdam dataset	
	mIoU (%)	mF_1 (%)	mIoU (%)	mF_1 (%)
Teacher	77.42	83.73	81.79	84.83
Student	74.66	81.56	78.62	81.95
+KD [24]	75.34 (↑ 0.68)	82.11	79.05 (↑ 0.43)	82.67
+AT [30]	75.50 (↑ 0.84)	82.14	80.01 (↑ 1.39)	83.40
+CWD [29]	75.28 (↑ 0.62)	82.04	79.99 (↑ 1.37)	83.84
+DSD [28]	75.44 (↑ 0.78)	82.34	80.14 (↑ 1.52)	83.65
+CIRKD [27]	75.55 (↑ 0.89)	82.34	80.41 (↑ 1.79)	83.91
+Ours	**76.00 (↑ 1.34)**	**82.53**	**80.70 (↑ 2.08)**	**84.06**

2) Ablation study about the effect of α: We conduct ablation experiments to analyze the impact of the parameter α in the BRE strategy. The parameter α determines the degree of

TABLE VI
ABLATION STUDY ABOUT THE EFFECT OF α ON THE NOISY SUPERVISED DATASET. THE BEST SCORE IN EACH COLUMN IS HIGHLIGHTED IN **BOLD**.

α	Noisy supervised Vaihingen dataset		Noisy supervised Potsdam dataset	
	mIoU (%)	mF_1 (%)	mIoU (%)	mF_1 (%)
0.2	67.73	76.27	77.47	81.48
0.4	67.87	76.12	77.33	81.20
0.6	68.11	76.33	77.55	81.64
0.8	68.12	76.35	**77.68**	**81.67**
1.0	**68.14**	**76.37**	77.40	81.64

diffusion of uncertainty around the boundary. For larger α values, the BR score decreases in smaller regions near the boundary. Conversely, with smaller α values, the BR score decreases in larger regions near the boundary. Fine-tuning the optimal α value according to the degree of boundary noise in different datasets is crucial for enhancing segmentation performance. Consequently, we conduct ablation experiments on two noisy supervised datasets, with the results detailed in Table VI. Notably, on the noisy supervised Vaihingen dataset, an α value of 1.0 achieves the best segmentation performance. Similarly, on the noisy supervised Potsdam dataset, an α value of 0.8 achieves the optimal segmentation results.

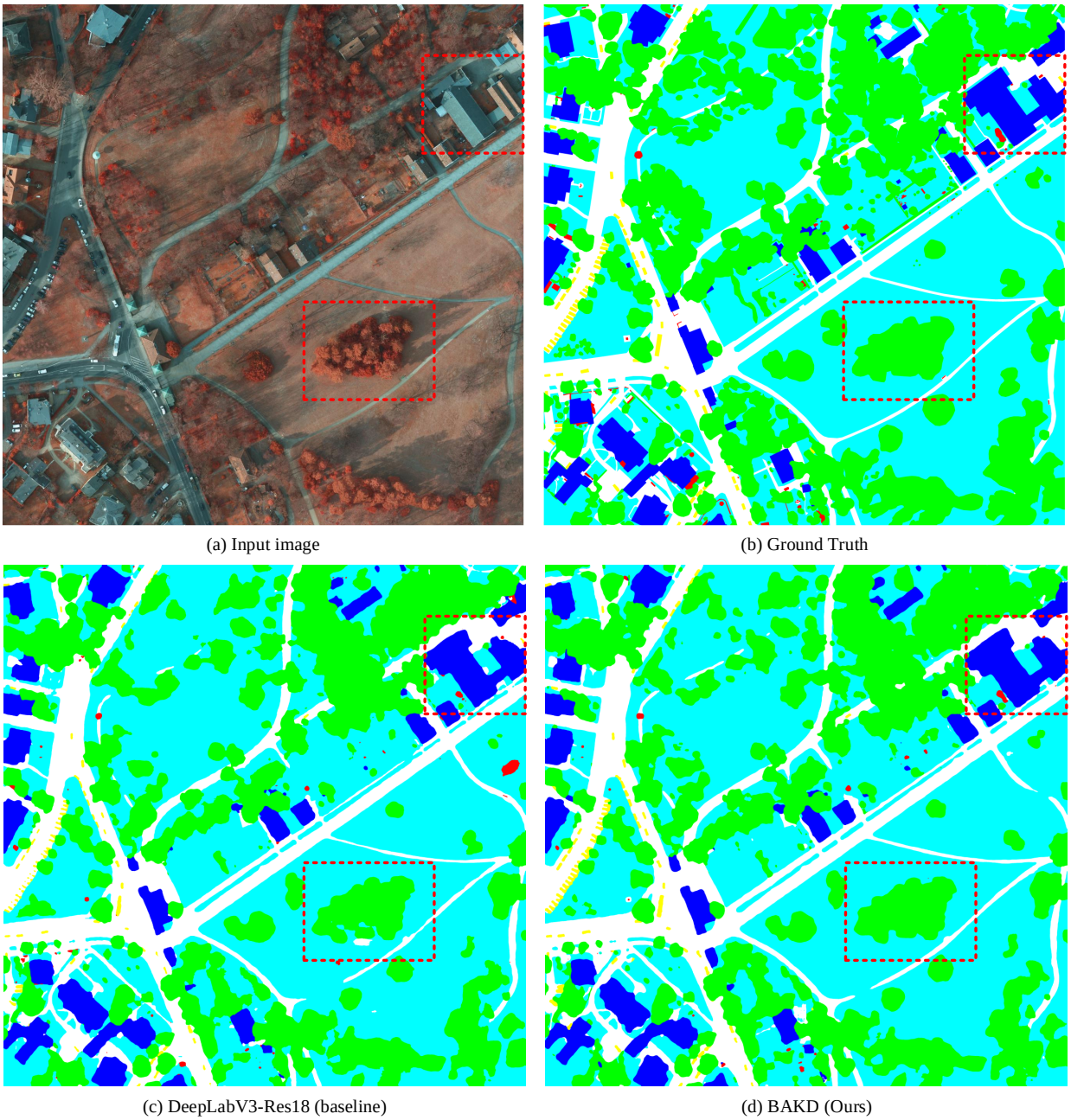
3) Ablation study about the different student networks: To verify the effectiveness and robustness of our BAKD, we evaluate various student networks on the noisy, supervised Potsdam dataset. In experiments, we adopt DeepLabV3 with ResNet-101 backbone (DeepLabV3-Res101) as the teacher network. DeepLabV3 with ResNet-18 backbone (DeepLabV3-Res18), DeepLabV3 with MobileNetV2 backbone (DeepLabV3-MBV2) and PSPNet with ResNet-18 backbone (PSPNet-Res18) as student networks. The experimental results are presented in Table VII. Notably, BAKD achieves considerable results in all student networks, proving the robustness of BAKD to changes in student network architecture. When using a more robust backbone network, performance improves and even exceeds the performance of the teacher network. Like other semantic segmentation methods based on knowledge distillation, BAKD does not modify the model architecture. Therefore, the parameters (Params) of the model and the floating point operations (FLOPs) resulting from inference remain consistent with the underlying backbone network. Specifically, for all methods where the student model is DeepLabV3-Res18, the FLOPs are 85.98G and the Params are 13.61M. Similarly, for all methods where the student model is PSPNet-Res18, the FLOPs are 67.51G, and the Params are 12.92M. It should be noted that we use lightweight MobileNetV2 as the backbone network, aiming to verify BAKD's performance on lightweight networks to explore the possibility of deploying BAKD on mobile devices. As can be seen from the table, although the model based on MobileNetV2 has fewer parameters and fewer calculations (22.62G FLOPs, 3.23M Params), its mIoU reaches 76.80%. This shows that our method is also suitable for more lightweight models, such as SqueezeNet [53] series and ShuffleNet [54], [55] series. BAKD can achieve high segmentation accuracy while using as few computing resources and parameters as possible, thereby better deploying in resource-constrained RSI environments such as embedded and mobile devices. We qualitatively evaluate the student network utilizing DeepLabV3 with ResNet-18 as the backbone network on the unprocessed large-size Potsdam test set. As illustrated in Fig. 8, the segmentation results generated by the student model trained with the BAKD approach exhibit closer alignment with the ground truth.

TABLE VII
ABLATION STUDY ABOUT DIFFERENT STUDENT NETWORKS ON THE NOISY SUPERVISED POTSDAM DATASET. FLOPS ARE MEASURED BASED ON THE TEST SIZE OF 512×512. * DENOTES THAT NOT INITIALIZE THE BACKBONE WITH IMAGENET [52] PRE-TRAINED WEIGHTS. THE BEST SCORE IN EACH COLUMN IS HIGHLIGHTED IN **BOLD**. BAKD ACHIEVED CONSIDERABLE RESULTS IN ALL STUDENT NETWORKS.

Method	mIoU(%)	mF_1 (%)	FLOPs(G)	Params(M)
T: DeepLabV3-Res101	77.21	81.27	384.41	61.11
S: DeepLabV3-Res18	73.99	78.84	85.98	13.61
+BAKD	**77.68 (↑ 3.69)**	**81.67**		
S: DeepLabV3-Res18*	72.56	78.13	85.98	13.61
+BAKD	**77.41 (↑ 4.85)**	**81.50**		
S: DeepLabV3-MBV2	73.55	78.38	22.62	3.23
+BAKD	**76.80 (↑ 3.25)**	**81.08**		
S: PSPNet-Res18	73.63	78.33	67.51	12.92
+BAKD	**77.49 (↑ 3.86)**	**81.28**		

4) Ablation study of components in BAKD: We conduct ablation experiments on the noisy supervised Potsdam dataset to evaluate the effectiveness of two uncertainty evaluation strategies. The results are summarized in Table VIII. Among them: Experimental group (a) means that only the predictive confidence score (PC) is used to participate in the loss calculation of the segmentation task. The results indicate that the PCE strategy can improve the segmentation performance of the student network to a certain extent. The experimental group (b) employs only the boundary-annotated reliability score (BR) in the loss calculation, which can also improve the student network's segmentation performance. However, the effect is slightly weaker than the BR score. The experimental group (c) computes the final boundary-aware weight W_{BAW} by averaging the PC and BR scores for participation in the loss calculation. The experimental group (d) means that the minimum weight between PC and BR for each pixel is selected as the final boundary-aware weight W_{BAW} for inclusion in the loss calculation. The results suggest that considering these two types of uncertainty information can better identify samples with potentially noisy annotations, thereby mitigating the impact of noisy annotations on knowledge distillation. Given that the BR score near the center tends to approach 1, the averaging method might diminish the impact of PC, while selecting the minimum weight can better emphasize the combined effect of both scores.

5) Ablation study about three ways of generating different BR scores: We conduct ablation experiments on the three methods of using boundary maps to generate different BR scores discussed in the BRE strategy on the noisy supervised Potsdam dataset. The experimental results are shown in Table IX. Comparing not using the BRE strategy, the BR score generated by the three methods enhances the model's segmentation performance. Specifically, the Gaussian blur only averages pixels closest to the boundary to a smaller value



(a) Input image (b) Ground Truth

(c) DeepLabV3-Res18 (baseline) (d) BAKD (Ours)

Fig. 8. Qualitative comparison for our proposed BAKD and the baseline student network on the unprocessed large-size Potsdam test set. The relatively obvious areas of improvement are marked with red rectangles.

TABLE VIII
THE EFFECT OF COMPONENTS IN THE PROPOSED METHOD ON THE NOISY SUPERVISED POTSDAM DATASET. THE BEST SCORE IN EACH COLUMN IS HIGHLIGHTED IN **BOLD**. MIN COMBINATION FULLY PLAYS A ROLE IN THE ADVANTAGES OF THE TWO EVALUATION STRATEGIES.

Method	PCE	BRE	Combination mode	mIoU (%)	mF_1 (%)
T: DeepLabV3-Res101				77.21	81.27
S: DeepLabV3-Res18				73.99	78.84
(a)	✓			77.11	81.13
(b)		✓		76.98	80.85
(c)	✓	✓	Average combination	77.29	81.31
(d)	✓	✓	Min combination	**77.68**	**81.67**

after the Gaussian smoothing operation. This method is not ideal for data with seriously noisy annotations. In contrast, the min-max normalization assigns smaller weights to a broader range of pixels near the boundary, potentially leading to insufficient training samples and overfitting. The exponential normalization can flexibly use the noise level at the boundary of the dataset to adjust the BR of each pixel, thereby flexibly controlling the distribution of BR scores in the boundary regions. Therefore, this method achieves the best performance.

6) Ablation study about the effect of warm-up iteration ratio: To enable the student network to quickly have a certain ability to judge the annotation's uncertainty, in the early stage of training, we adopted a method similar to "Warm-up" training.

TABLE IX
ABLATION EXPERIMENT ABOUT THREE WAYS OF GENERATING DIFFERENT BR. THE BEST SCORE IN EACH COLUMN IS HIGHLIGHTED IN **BOLD**. THE PARAMETERIZED EXPONENTIAL FUNCTION NORMALIZATION METHOD OUTPERFORMED OTHER BR GENERATION METHODS.

Way of generate BR	mIoU(%)	mF_1 (%)
Without BR	77.11	81.13
Gaussian blur	77.56	81.39
Min-max normalization	77.43	81.44
Exponential normalization (Ours)	**77.68**	**81.67**

TABLE X
THE EFFECT OF WARM-UP ITERATION RATIO IN THE PROPOSED PCW STRATEGY ON THE NOISY SUPERVISED POTSDAM DATASET. THE BEST SCORE IN EACH COLUMN IS HIGHLIGHTED IN **BOLD**.

Method	λ	Warm-up	mIoU (%)	mF_1 (%)
T: DeepLabV3-Res101			77.21	81.27
S: DeepLabV3-Res18 (without KD)			73.99	78.84
(a)	$[0, 0.5]$	0%	77.35	81.32
(b)	$[0, 0.5]$	10%	**77.68**	**81.67**
(c)	$[0, 0.5]$	20%	77.43	81.52
(d)	$[0, 0.5]$	30%	77.25	81.29

Provide supervision information through the teacher network to guide the student network in learning relatively correct supervised samples to achieve rapid convergence. The training at this stage should be short to avoid a long warm-up phase that cannot fully utilize the potential of the student model itself. To determine the optimal scale parameter, we try different scales from 0% to 30% to evaluate the performance of the student model at different scales, as shown in Table X. The results demonstrate that using a warm-up strategy in the initial 10% of training iterations leads to the best segmentation performance.

7) Ablation study about the effect of λ: We conducted ablation experiments on the parameter λ to explore the influence of Knowledge Distillation (KD) on BAKD within the PUW strategy. Here, λ adjusts the weight ratio between the student and teacher networks in this strategy. In Table XI, "λ=0" indicates λ set to 0, utilizing only the teacher network's evaluation score as PC. "λ=1" indicates λ set to 1, and using only the student model's evaluation score as PC. "λ=[0,1]" indicates λ gradually increases from 0 to 1 during training iterations. The evaluation score of the student network gradually participates in calculating PC. Similarly, "λ=[0,0.3]" indicates a gradual increase from 0 to 0.3 in λ during training. "λ=[0,0.5]" indicates a gradual increase from 0 to 0.5 in λ during training. "λ=[0,0.7]" indicates a gradual increase from 0 to 0.7 in λ during training. In Fig. 9, we present the confidence score maps for groups (i), (ii), and (iv), as well as the baseline strategy that solely uses the student network for training. When iteration= 0, the student network has not yet been trained, resulting in low prediction confidence scores for each pixel. At this stage, relying entirely on the immature student network to provide sample weights can amplify errors, posing challenges for the student network's correction. In contrast, the pre-trained teacher network offers more reliable confidence scores. Therefore, in the early training stage, we utilize the supervisory information provided by the teacher network to assist the student network in converging rapidly.

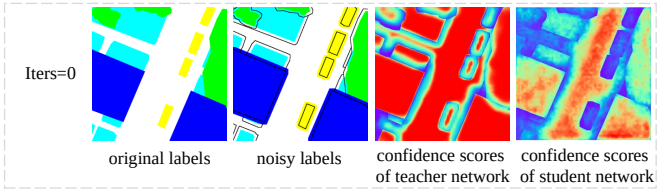
TABLE XI
THE EFFECT OF λ IN THE PROPOSED PCW STRATEGY ON THE NOISY SUPERVISED POTSDAM DATASET. THE BEST SCORE IN EACH COLUMN IS HIGHLIGHTED IN **BOLD**.

Method	λ	Warm-up	mIoU (%)	mF_1 (%)
T: DeepLabV3-Res101			77.21	81.27
S: DeepLabV3-Res18 (without KD)			73.99	78.84
(i)	0	10%	77.02	81.09
(ii)	1	10%	76.86	80.91
(iii)	$[0, 0.3]$	10%	76.93	81.01
(iv)	$[0, 0.5]$	10%	**77.68**	**81.67**
(v)	$[0, 0.7]$	10%	77.32	81.27
(vi)	$[0, 1.0]$	10%	77.11	81.18

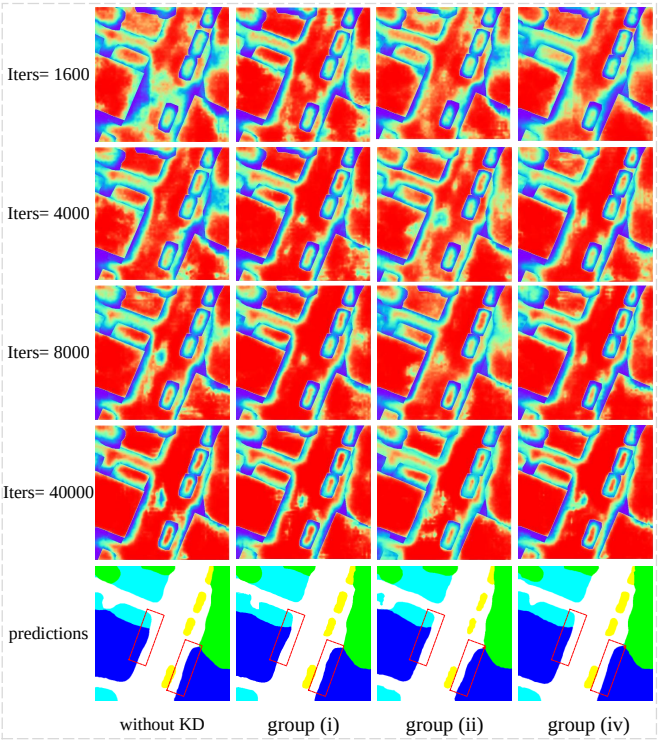
Upon examining the confidence maps for groups (i), (ii), and (iv), we observe the following finding: Group (i) relies entirely on the teacher network to provide sample weights. Although its predicted confidence is relatively reliable in the early iterations, its influence gradually diminishes as training progresses. In contrast, group (ii) depends solely on the immature student network for sample weights, resulting in less reliable predicted confidence in the early iterations compared to groups (ii) and (iv). This reliance can amplify errors and increase the difficulty of corrections for the student network. In contrast, the prediction results of group (iv) show higher consistency with the semantic boundaries of the labels.

Additionally, we conducted a sensitivity analysis on the λ to investigate the impact of different final λ values on model performance. The experimental results demonstrate that the model performs optimally when the final λ = 0.5. Further analysis reveals that when the final λ = 0.3, the student network's participation in the later training stages is insufficient, leading to limited contributions to the evaluation of predictive confidence. This prevents the model from fully leveraging the student network's gradually improving cognitive capabilities during training. Conversely, when the final λ = 0.7, the teacher network's supervisory role is relatively weakened in the later training stages, making the model more susceptible to the student network's inaccurate predictions, thereby affecting the overall convergence speed and model's stability. Therefore, final λ = 0.5 achieves a well-balanced collaboration between the teacher and student networks. It ensures the teacher network's reliable supervision in evaluating predictive confidence during the early training stages while fully leveraging the student network's contributions to predictive confidence evaluation in the later training stages.

8) Ablation study about different noise types: To thoroughly investigate the impact of different types of noise annotations on model performance, we conducted experiments with three types of noise annotations on the Vaihingen datasets. By comparing the experimental results of different noise types, we can comprehensively evaluate the model's performance under relatively realistic noise data. As shown in Table XII, the experimental results demonstrate that our BAKD method effectively improves both mIoU and BF scores under various noise conditions. These findings validate the robustness and effectiveness of our approach, further confirming its applicability in handling different types of annotation noise.



(a) Confidence score maps of the teacher and student networks at iteration = 0



(b) Confidence score maps and prediction results of the student network at different iterations

Fig. 9. The influence of different λ on confidence score maps at different training iterations (Iters).

TABLE XII
ABLATION STUDY ABOUT DIFFERENT NOISE TYPES ON THE NOISY SUPERVISED VAIHINGEN DATASET.

Noisy type	Method	mIoU (%)	mF_1 (%)
Dilation Noise	Student	64.99	73.89
	BAKD	68.14 (↑ 3.15)	76.37 (↑ 2.48)
Erosion Noise	Student	61.38	70.02
	BAKD	66.51 (↑ 5.03)	75.13 (↑ 5.11)
Shift Noise	Student	66.02	74.94
	BAKD	69.63 (↑ 3.61)	77.27 (↑ 2.33)

D. Integrating with Existing KD methods and training time

To show that our method is complementary to other semantic segmentation based on KD methods, we evaluate the performance impact of integrating BAKD into existing KD methods on the noisy supervised Vaihingen and Postdam datasets. We use DeepLabV3-Res18 as the student network and explore the effectiveness of BAKD in the simplest integrated way under the experimental settings of the original method without changing any hyperparameters. The experimental results are shown in Table XIII. Among the five baseline methods, BAKD effectively improves all methods' performance and

TABLE XIII
INTEGRATING WITH EXISTING KD METHODS ON NOISY SUPERVISED VAIHINGEN AND POTSDAM DATASETS. "+" DENOTES IMPLEMENTING THE CORRESPONDING SCHEMES. VALUES WITHIN PARENTHESES INDICATE THE PERFORMANCE VARIANCE RELATIVE TO EACH KD METHOD. BAKD EFFECTIVELY IMPROVES THE PERFORMANCE OF ALL KD METHODS.

Method	Noisy supervised Vaihingen dataset		Noisy supervised Potsdam dataset	
	mIoU (%)	mF_1 (%)	mIoU (%)	mF_1 (%)
Teacher	68.79	77.04	77.21	81.27
Student	64.99	73.89	73.99	78.84
AT [30]	67.13	75.63	76.97	80.98
+BAKD	67.47(↑ 0.34)	75.75	**77.53 (↑ 0.56)**	**81.42**
CWD [29]	66.14	75.03	76.42	80.84
BAKD	66.76(↑ 0.62)	75.40	**77.48 (↑ 1.06)**	**81.35**
DSD [28]	66.67	75.11	75.79	79.89
+BAKD	67.03(↑ 0.36)	75.47	**77.65 (↑ 0.67)**	**81.56**
CIRKD [27]	67.23	75.83	77.28	81.24
+BAKD	67.90(↑ 0.67)	76.12	**77.89 (↑ 0.61)**	**81.52**
KD [24]	66.34	74.96	76.45	80.86
+BAKD (Ours)	**68.14 (↑ 1.80)**	76.37	**77.68 (↑ 1.23)**	**81.67**

increases the student model's robustness to noisy annotations. Specifically, for the noisy supervised Vaihingen dataset, after integrating BAKD, the mIoU of the AT method increased by 0.34%, the mIoU of the CWD method increased by 0.62%, the mIoU of the DSD method increased by 0.36%, the mIoU of the CIRKD method increased by 0.67%, and the mIoU of the KD method increased by 1.80%. For the noisy supervised Postdam dataset, after integrating BAKD, the mIoU of the AT method increased by 0.56%, the mIoU of the CWD method increased by 1.06%, the mIoU of the DSD method increased by 1.86%, the mIoU of the CIRKD method increased by 0.61%, and the mIoU of the KD method increased by 1.23%. BAKD extends these KD methods' applicability in handling noisy annotations. Fig. 10 compares prediction maps for methods with and without BAKD integration on the noisy supervised Postdam dataset. The visualization results of the methods integrated with BAKD generally perform better on complex regions such as edges and obscured objects.

Furthermore, we compare using the same student network to ensure consistent computational load and parameters, thereby assessing the time resource consumption differences between BAKD and existing related methods. Table XIV presents the experimental results comparing training times. Since Decoupling, Co-teaching, and RMD methods require training two networks simultaneously, their training time is relatively long, even exceeding that of a complex teacher network. In contrast, the KD method based on the offline teacher network only needs to train one student network, and the training time is relatively low. The ADELE method necessitates calculating and recording the IoU values for each pixel during every training iteration, leading to significant computational overhead. Since BAKD can be integrated with KD methods, we also obtained the training times before and after combining various KD methods with BAKD. Table XIV shows that among the numerous KD methods, BAKD only adds 13 minutes to the baseline KD method. In contrast, the training times for the AT, CWD, DSD, and CIRKD methods are all longer than those for BAKD. Notably, the CIRKD method employs a memory bank

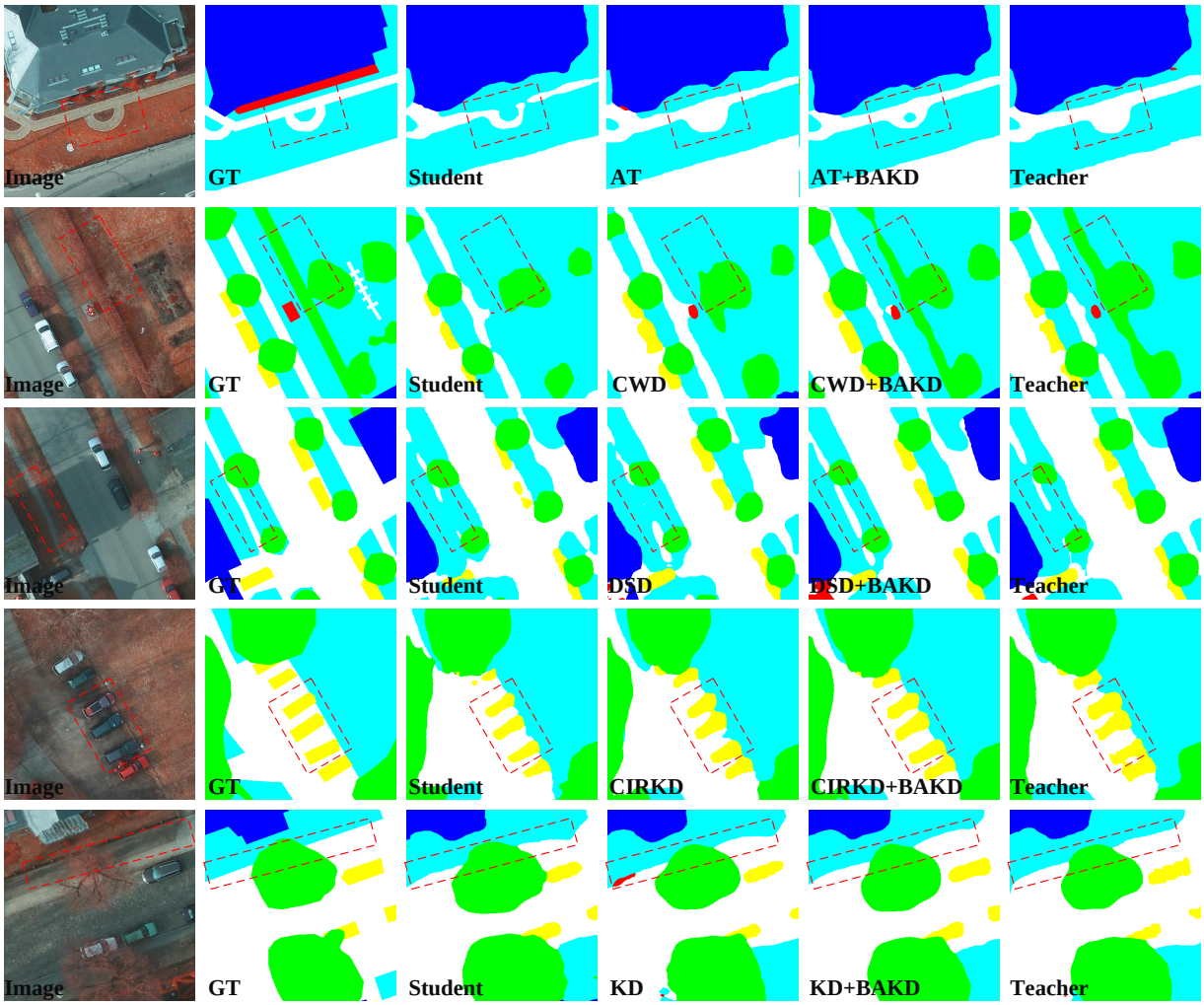


Fig. 10. Predictions of KD methods with and without integrating BAKD. BAKD performs better on complex regions such as edges and obscured objects.

for contrastive learning. It requires optimizing five loss terms, resulting in 6 hours and 10 minutes of training, almost twice as long as other methods. This also verifies the problem that adding multiple optimization objectives will increase training difficulty and time. This comparison concludes that integrating RDD does not incur excessive computational overhead or training time. It only incurs an average increase of less than 8% in training time across different methods.

TABLE XIV
COMPARISON OF TRAINING TIME FOR EXPERIMENTS WITH AND WITHOUT BAKD INTEGRATION.

Method	Cost
T: DeepLabV3-Res101	5 h 36 m
S: DeepLabV3-Res18	2 h 43 m
Decoupling [21]	6 h 33 m
Co-teaching [22]	5 h 34 m
Co-teaching+ [23]	5 h 04 m
ADELE [40]	6 h 45 m
RMD [41]	6 h 34 m
AT [30]	3 h 34 m
AT [30] with BAKD	3 h 50 m (↑ 7.84%)
CWD [29]	4 h 44 m
CWD [29] with BAKD	4 h 56 m (↑ 4.23%)
DSD [28]	3 h 40 m
DSD [28] with BAKD	3 h 54 m (↑ 6.36%)
CIRKD [27]	6 h 10 m
CIRKD [27] with BAKD	6 h 24 m (↑ 3.78%)
KD [24]	2 h 58 m
KD [24] with BAKD (Ours)	3 h 11 m (↑ 7.30%)

V. LIMITATIONS

The BAKD method focuses on effectively handling the ubiquitous semantic boundary noise. The effectiveness of the boundary-annotated reliability evaluate (BRE) strategy in BAKD is largely affected by the quality of the initial annotations. When the annotation quality at the semantic boundary is low, the BRE strategy can play a positive role. However, in the case where there is less noise on semantic boundaries, the performance gain of the BRE strategy is not significant.

VI. CONCLUSIONS

We propose the Boundary-aware Knowledge Distillation method (BAKD) for the RSI semantic segmentation task with noisy annotations. BAKD includes two evaluation strategies, which leverage the predictive confidence of the teacher and

student networks, along with the boundary-annotated reliability, to identify samples with potentially noisy annotations and adjust the learning process accordingly. BAKD has good scalability and is easy to integrate with existing distillation methods to further improve the robustness of noisy annotations. Experimental results show that BAKD outperforms mainstream knowledge distillation methods.

In this study, we focused on the semantic segmentation task of remote-sensing images with noisy annotations. The BAKD method effectively utilizes relatively easy-to-obtain rough annotations for training, thereby reducing manual annotation costs. However, the BAKD method's applicability is limited under certain noise conditions. In future research, we will explore more suitable evaluation methods for annotations with different noise types (such as Gaussian or more complex mixed noise) and varying noise levels, aiming to enhance the model's robustness and generalization ability across various practical application scenarios. At the same time, we plan to further explore BAKD's application potential in other visual tasks (such as classification and detection) to promote the development of training technology under noisy supervised datasets. Moreover, although this study has examined the segmentation performance of BAKD across different student architectures, we will also focus on the performance of BAKD on more lightweight student models (such as SqueezeNet series and ShuffleNet series) in resource-constrained RSI environments in future work.

REFERENCES

[1] H. N. Pham, K. B. Dang, T. V. Nguyen, N. C. Tran, X. Q. Ngo, D. A. Nguyen, T. T. H. Phan, T. T. Nguyen, W. Guo, and H. H. Ngo, "A new deep learning approach based on bilateral semantic segmentation models for sustainable estuarine wetland ecosystem management," *Science of The Total Environment*, vol. 838, p. 155826, 2022.

[2] L. Bragagnolo, L. Rezende, R. Da Silva, and J. Grzybowski, "Convolutional neural networks applied to semantic segmentation of landslide scars," *Catena*, vol. 201, p. 105189, 2021.

[3] G. Can, D. Mantegazza, G. Abbate, S. Chappuis, and A. Giusti, "Semantic segmentation on swiss3dcities: A benchmark study on aerial photogrammetric 3d pointcloud dataset," *Pattern Recognition Letters*, vol. 150, pp. 108–114, 2021.

[4] B. Trenčanová, V. Proença, and A. Bernardino, "Development of semantic maps of vegetation cover from uav images to support planning and management in fine-grained fire-prone landscapes," *Remote Sensing*, vol. 14, no. 5, p. 1262, 2022.

[5] H. Sheng, X. Chen, J. Su, R. Rajagopal, and A. Ng, "Effective data fusion with generalized vegetation index: Evidence from land cover segmentation in agriculture," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 60–61.

[6] J. K. Jadhav, A. P. Sonavale, and R. Singh, "Segmentation analysis using particle swarm optimization-self organizing map algorithm and classification of remote sensing data for agriculture," in *Intelligent Data Communication Technologies and Internet of Things: ICICI 2019*. Springer, 2020, pp. 659–668.

[7] C. Ji, W. Zhou, J. Lei, and L. Ye, "Infrared and visible image fusion via multiscale receptive field amplification fusion network," *IEEE Signal Processing Letters*, 2023.

[8] C. Liang-Chieh, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *ICLR*, 2015.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *TPAMI*, vol. 40, no. 4, pp. 834–848, 2017.

[10] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2017.

[13] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *TPAMI*, vol. 43, no. 10, pp. 3349–3364, 2020.

[14] D. Liang, B. Kang, X. Liu, P. Gao, X. Tan, and S. Kaneko, "Cross-scene foreground segmentation with supervised and unsupervised model communication," *Pattern Recognition*, vol. 117, p. 107995, 2021.

[15] D. Liang, Y. Du, H. Sun, L. Zhang, N. Liu, and M. Wei, "Nlkd: Using coarse annotations for semantic segmentation based on knowledge distillation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2335–2339.

[16] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[17] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li, "Treeunet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 156, pp. 1–13, 2019.

[18] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Systems with Applications*, vol. 169, p. 114417, 2021.

[19] G. Xu, M. Deng, G. Sun, Y. Guo, and J. Chen, "Improving building extraction by using knowledge distillation to reduce the impact of label noise," *Remote Sensing*, vol. 14, no. 22, p. 5645, 2022.

[20] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical image analysis*, vol. 65, p. 101759, 2020.

[21] E. Malach and S. Shalev-Shwartz, "Decoupling" when to update" from" how to update"," *Advances in neural information processing systems*, vol. 30, 2017.

[22] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[23] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, and M. Sugiyama, "How does disagreement help generalization against label corruption?" in *International conference on machine learning*. PMLR, 2019, pp. 7164–7173.

[24] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NeurIPS*, 2015.

[25] M. Lukasik, S. Bhojanapalli, A. Menon, and S. Kumar, "Does label smoothing mitigate label noise?" in *International Conference on Machine Learning*. PMLR, 2020, pp. 6448–6458.

[26] F. Sarfraz, E. Arani, and B. Zonooz, "Knowledge distillation beyond model compression," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 6136–6143.

[27] C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, and Q. Zhang, "Cross-image relational knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[28] Y. Feng, X. Sun, W. Diao, J. Li, and X. Gao, "Double similarity distillation for semantic image segmentation," *TIP*, vol. 30, pp. 5363–5376, 2021.

[29] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen, "Channel-wise knowledge distillation for dense prediction," in *ICCV*, 2021.

[30] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *ICLR*, 2017.

[31] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

[32] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in *ICCV*, 2019.

[33] Y. Wang, W. Zhou, T. Jiang, X. Bai, and Y. Xu, "Intra-class feature variation distillation for semantic segmentation," in *ECCV*, 2020.

[34] J. Jacob, O. Ciccarelli, F. Barkhof, and D. C. Alexander, "Disentangling human error from the ground truth in segmentation of medical

images," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 750–15 762, 2021.

[35] Y. Shu, X. Wu, and W. Li, "Lvc-net: Medical image segmentation with noisy label based on local visual cues," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer, 2019, pp. 558–566.

[36] G. Wang, X. Liu, C. Li, Z. Xu, J. Ruan, H. Zhu, T. Meng, K. Li, N. Huang, and S. Zhang, "A noise-robust framework for automatic segmentation of covid-19 pneumonia lesions from ct images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2653–2663, 2020.

[37] S. Min, X. Chen, Z.-J. Zha, F. Wu, and Y. Zhang, "A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 4578–4585.

[38] Y. Luo, G. Liu, W. Li, Y. Guo, and G. Yang, "Deep neural networks learn meta-structures to segment fluorescence microscopy images," *arXiv preprint arXiv:2103.11594*, 2021.

[39] C. Liu, C. M. Albrecht, Y. Wang, Q. Li, and X. X. Zhu, "Aio2: Online correction of object labels for deep learning with incomplete annotation in remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[40] S. Liu, K. Liu, W. Zhu, Y. Shen, and C. Fernandez-Granda, "Adaptive early-learning correction for segmentation from noisy annotations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2606–2616.

[41] C. Fang, Q. Wang, L. Cheng, Z. Gao, C. Pan, Z. Cao, Z. Zheng, and D. Zhang, "Reliable mutual distillation for medical image segmentation under imperfect annotations," *IEEE Transactions on Medical Imaging*, vol. 42, no. 6, pp. 1720–1734, 2023.

[42] R. Dong, W. Fang, H. Fu, L. Gan, J. Wang, and P. Gong, "High-resolution land cover mapping through learning with noise correction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[43] H. Chen, W. Yang, L. Liu, and G.-S. Xia, "Coarse-to-fine semantic segmentation of satellite images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 217, pp. 1–17, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271624002958

[44] 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 Journal of Photogrammetry and Remote Sensing*, vol. 188, pp. 157–176, 2022.

[45] X. Lyu, R. Zheng, and L. Zhang, "Semantic segmentation of weakly annotated remote sensing images based on feature adversary and uncertainty perception," *IEEE Geoscience and Remote Sensing Letters*, 2024.

[46] J. Li, W. He, W. Cao, L. Zhang, and H. Zhang, "Uanet: An uncertainty-aware network for building extraction from remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[47] J. Li, H. Huang, W. He, H. Zhang, and L. Zhang, "Overcoming the uncertainty challenges in flood rapid mapping with multi-source optical data," in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2024, pp. 780–784.

[48] D. Liang, Y. Sun, Y. Du, S. Chen, and S.-J. Huang, "Relative difficulty distillation for semantic segmentation," *Science China Information Sciences*, vol. 67, no. 9, p. 192105, 2024.

[49] F. Rottensteiner, G. Sohn, M. Gerke, and J. D. Wegner, "Isprs semantic labeling contest," *ISPRS: Leopoldshöhe, Germany*, vol. 1, no. 4, p. 4, 2014.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2016.

[51] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[52] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[53] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[54] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[55] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.