Progressive Learning With Cross-Window Consistency for Semi-Supervised Semantic Segmentation

Bo Dang, Yansheng Li, *Senior Member, IEEE*, Yongjun Zhang, *Member, IEEE*, and Jiayi Ma, *Senior Member, IEEE*

***Abstract*— Semi-supervised semantic segmentation focuses on the exploration of a small amount of labeled data and a large amount of unlabeled data, which is more in line with the demands of real-world image understanding applications. However, it is still hindered by the inability to fully and effectively leverage unlabeled images. In this paper, we reveal that cross-window consistency (CWC) is helpful in comprehensively extracting auxiliary supervision from unlabeled data. Additionally, we propose a novel CWC-driven progressive learning framework to optimize the deep network by mining weak-to-strong constraints from massive unlabeled data. More specifically, this paper presents a biased cross-window consistency (BCC) loss with an importance factor, which helps the deep network explicitly constrain confidence maps from overlapping regions in different windows to maintain semantic consistency with larger contexts. In addition, we propose a dynamic pseudo-label memory bank (DPM) to provide high-consistency and high-reliability pseudo-labels to further optimize the network. Extensive experiments on three representative datasets of urban views, medical scenarios, and satellite scenes with consistent performance gain demonstrate the superiority of our framework. Our code is released at https://jack-bo1220.github.io/project/CWC.html.**

***Index Terms*— Semi-supervised semantic segmentation, consistency loss, pseudo-label supervision.**

I. Introduction

SEMANTIC segmentation, as a fundamental and essential task, is widely employed in a wide range of situations, such as automated driving, medical pathology diagnosis, and land cover survey [2], [3], [4], [5], [6], [7]. The brilliant performance of data-driven deep learning algorithms largely depends on huge volumes of annotated data. In practice, massive unlabeled images are collected, but it is hard to acquire the corresponding pixel-level annotations. Despite the availability of advanced semi-automatic labeling algorithms [8], [9], the process of generating annotated data is still tremendously labor-intensive and time-consuming, particularly the annotating process of remote sensing and medical images requires the participation of experts with domain knowledge. To alleviate this issue, numerous semi-supervised learning methods [10], [11], [12], [13], [14], [15], [16], [17] have been developed and achieve promising performance.

Although lots of achievements have been obtained in semi-supervised semantic segmentation, many tricky challenges still remain. The first challenge is how to generate or select pseudo-labels with high-reliability for preventing catastrophic performance degradation. Minimizing the adverse impact of the noise of pseudo-labels is a longstanding but unsolved issue in self-training pipelines [13]. The second challenge is that heterogeneous consistency traits are not fully utilized. As shown in Figure 1, the prediction of the model for overlapping regions of image patches across diverse contextual windows should exhibit semantic consistency, which we refer to as **c**ross-**w**indow **c**onsistency (CWC). It can be viewed as a unique form of data augmentation (*i.e.* , contextual augmentation) and applied to unlabeled data [1]. Similar ideas are also involved in self-supervised learning [18] and image-to-image translation [19], which shows that CWC is promising. However, it is still insufficient for existing works to fully exploit the merit of CWC. For instance, Directional Contrastive loss from [1] requires manual setting of some key parameters (such as the positive filtering threshold) that must be tuned depending on the datasets. As a whole, CWC is preliminarily explored in the consistency loss modeling, but unfortunately ignored in the selection of high-quality pseudo-labels.

Similar to the peripheral vision system in human vision [20], [21], human visual reasoning processes need to rely on multiple contour regions that cover different contextual information. In life, when humans view images from cross windows, the visual center of the cerebral cortex often produces the same response on overlapping regions. These facts guide us to leverage CWC to exploit unlabeled data.

In light of the aforementioned challenges and the inspiration of human vision, we propose a progressive learning framework guided by the philosophy of CWC to systematically exploit the benefits of this inherent consistency. Our framework

Manuscript received 10 September 2023; revised 21 February 2024 and 26 August 2024; accepted 30 August 2024. Date of publication 17 September 2024; date of current version 23 September 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42371321 and in part by the State Key Project of the National Natural Science Foundation of China under Grant 42030102. The associate editor coordinating the review of this article and approving it for publication was Prof. Yulan Guo. *(Corresponding author: Yansheng Li.)*

Bo Dang, Yansheng Li, and Yongjun Zhang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: bodang@whu.edu.cn; yansheng.li@whu.edu.cn; zhangyj@whu.edu.cn).

Jiayi Ma is with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: jyma2010@gmail.com).

Digital Object Identifier 10.1109/TIP.2024.3458854

progressively optimizes deep network by mining weak-to-strong constraints from unlabeled data. Specifically, in the first stage, we introduce a general and effective **b**iased **c**ross-window **c**onsistency (BCC) loss that measures the semantic consistency of overlapping regions based on the segmentation confidence maps. In the second stage, we further extend this fundamental concept by designing a unique pseudo-label reliability evaluating method and establishing a highly dynamic and rewarding **d**ynamic **p**seudo-label **m**emory bank (DPM) to assist in exposing the model to strong pseudo-label constraints. Benefiting from our proposed pseudo-label reliability evaluation algorithm guided by the inherent cross-window dependencies of images and a well-designed DPM, our approach calculates the contextual prediction consistency of overlapping regions across various windows to ensure the constant and dynamic update of information in the DPM.

Our framework is generalized and can be adapted simply to semi-supervised semantic segmentation applications (*e.g.* , urban street scenes segmentation in computer vision, medical nuclear segmentation in pathological analysis, and land cover classification in remote sensing). Extensive experiments on the Cityscapes [22], MoNuSeg [23], and DeepGlobe [24] datasets demonstrate a considerable performance improvement over the state-of-the-art methods. By systematically exploring CWC, our main contributions are summarized as follows:

- The BCC loss with the importance factor is designed to maintain larger contextual semantic consistency among overlapping confidence maps.
- We propose a DPM using a novel pseudo-label reliability evaluation method to minimize the adverse effects of ill-posed pseudo-labels.
- Our framework outperforms previous methods on extensive datasets from different fields, which demonstrates the strong generalization and competitiveness of our work.

II. Related Work

Semi-supervised semantic segmentation's crux and core is how to properly utilize unlabeled data and be able to further enhance the generalization of the model with less labeled data. With the rapid improvement of semi-supervised learning (SSL) methods [25], [26], [27], [28], [29], solutions based on different paradigms have made progress in semi-supervised semantic segmentation tasks. The current semi-supervised semantic segmentation approach consists of three typical pipelines: GAN-based, self-training, and consistency regularization methods.

A. GAN-Based Methods

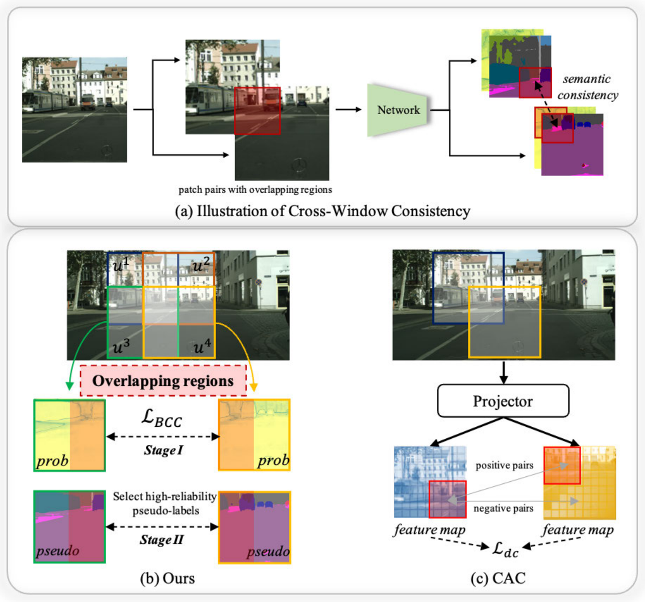
In semi-supervised semantic segmentation, GANs [30], [31] are used as discriminative tools or supervised signals alone or in conjunction with other methods. For example, Hung et al. [32] use discriminators of GAN networks to find pseudolabeled plausible regions. Previous works [31], [33] add the GAN branch as an auxiliary supervision in natural and medical images, respectively. Zhai et al. [34] employ a framework consisting of two generators and a discriminator for adversarial learning, where each generator produces segmentation masks that mutually supervise each other. Souly et al. [10] posit that the incorporation of synthetic visual data can induce real samples to approximate the feature space, thereby facilitating the enhancement of segmentation outcomes. Consequently, they employ the GAN to fabricate non-authentic samples, while leveraging weakly annotated information to enhance the quality of the GAN-generated samples. In contrast to the aforementioned approaches that utilize GANs as discriminative tools or supervised signals, we address semi-supervised semantic segmentation from another perspective.

B. Consistency Regularization-Based Methods

Consistency regularization-based methods make features of samples from the same category more compact in the feature space, while keeping features of samples from different categories as far as possible. The benefit comes in the implement's flexibility, which includes the design and metrics of consistency traits. Specifically, CutMix [35], ClassMix [36], and various other data augmentations [37] are federated in the consistency regularization framework in order to transform or perturb the input data to satisfy the constraints of the consistency measure, just as [11], [38], [39], [40] do. Similarly, further broader disturbances and different initialization model are published to achieve a gain [12], [14], respectively. Contrastive loss that performs well on other tasks is relocated to the consistency regularization paradigm in owing to the rapid advancement of contrastive learning and self-supervised learning [41], [42], [43], [44]. For instance, InfoNCE [45], which attempts to bring positive pairs closer and push negative pairs apart and shines in self-supervised learning, has been extensively modified and adapted to many previous methods [1], [15], [46], [47], [48]. In addition, CCT [11] emphasizes the validity of the mean square error (MSE) as an elegant consistency loss. In this paper, we focus on minimizing differences among unlabeled data across diverse windows to mitigate cross-window bias during the first stage.

C. Self-Training-Based and Pseudo Labeling-Based Methods

Self-training-based and Pseudo labeling-based methods commonly leverage student-teacher models to produce and re-train pseudo-labels. Chronic challenges include how to generate or select pseudo-labels with high-confidence to optimize the model and tackle the class-imbalanced issue. In response to the first problem outlined, ST++ [13] proposes a straightforward yet effective pipeline that boosts model stability through strong and weak data transformations and by gradually utilizing all pseudo-labels. Instead of ignoring the doubtful pixels of pseudo-labels, U^{2}PL [49] treats them as negative samples to be compared with the matching positive samples. ELN [50] and Yuan et al. [51] create the ELN module to correct pseudo labels and self-correction loss to prevent overfiting to the noise of low-confidence pseudo-labels, respectively. Yi et al. [52] use graph attention network to correct noisy labels. Specifically, Hu et al. [53] enhance the quality of pseudo-labels by utilizing LIDAR pseudo-labeling to estimate



(a) Illustration of Cross-Window Consistency

(b) Ours

(c) CAC

Fig. 1. The main concept underlying our work is that (a) the prediction of the model for overlapping regions of image patches across diverse contextual windows should exhibit semantic consistency, which we refer to as cross-window consistency (CWC). In contrast to (c) CAC [1], we employ the stable confidence maps rather than feature maps to encourage consistency of overlapping regions from wider contextual windows, and further apply CWC traits to select high-reliability pseudo-labels.

the depth of street-view images. Nevertheless, the applicability of this pipeline to medical and remote sensing scenarios poses a challenge. The class-imbalance bias of pseudo-labels undermine the generalization of the model, particularly when there are very few unlabeled samples or when the sample contains a significant long-tail effect. Numerous solutions [54], [55], [56] recognize this issue and align class distributions to rectify the imbalance. Note that the existing methods do not perfectly address the mentioned challenges. This work pursues selecting rewarding pseudo-labels to avoid the misleading of ill-posed pseudo-labels and overfitting of fixed pseudo-labels, based on the idea that overlapping regions on image patches from diverse contextual windows exhibit semantic consistency.

III. Methodology

A. Problem Definition

The goal of semi-supervised semantic segmentation is to employ a small set of labeled data $\mathcal{B}_l = \{(x_i, y_i)\}_{i=1}^{M}$ with unlabeled data $\mathcal{B}_u = \{u_i\}_{i=1}^{N}$ to train a model $\mathcal{F}$ that can provide accurate results on test data. In general, the overall optimization loss can be formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_s + \lambda \mathcal{L}_u, \tag{1}$$

where λ is a trade-off weight between labeled and unlabeled data supervision. Typically, the labeled supervised loss $\mathcal{L}_s$ is the cross-entropy loss or correlation variant (*e.g.* , OHEM [57]) of the inferences and annotated labels. The unsupervised loss $\mathcal{L}_u$ can be defined flexibly as consistency loss, pseudo-label loss, entropy minimum loss, thereby encouraging the model to fit the unlabeled data.

Algorithm 1 Our Framework Pseudocode

```
Input: Labeled images and corresponding labels $\mathcal{B}_l = \{(x_i, y_i)\}_{i=1}^{M}$
       Unlabeled images $\mathcal{B}_u = \{u_i\}_{i=1}^{N}$
       Validation set $\mathcal{B}_v = \{(x_i, y_i)\}_{i=1}^{V}$
Output: Trained model $\mathcal{F}$
1:  #Stage I: Biased Cross-Window consistency supervision (weak constraint)
2:  Train $\mathcal{F}$ on $\mathcal{B}_l$ with cross-entropy loss and $\mathcal{B}_u$ based on Eq. (4)
3:  Initialize previous best $\mathcal{S} \leftarrow 0$
4:  Calculate previous best $\mathcal{S} \leftarrow \mathrm{meanIOU}(\mathcal{F}(\mathcal{B}_v))$
5:  Initialize count $count \leftarrow 0$
6:  #Stage II: Dynamic pseudo-label memory bank (strong constraint)
7:  for $u_i \in \mathcal{B}_u$ do
8:      Get reliability score $\mathcal{R}_i$ based on Eqs. (7) and (8)
9:  end for
10: Select Top-K% scored unlabeled images and generate corresponding pseudo-labels $y_i^*$ into the dynamic pseudo-label memory bank $\mathcal{Q} = \{(u_i, y_i^*)\}_{i=1}^{N \times K\%}$
11: while epoch<maximum number of epochs do
12:     Train $\mathcal{F}$ on $\mathcal{B}_l \cup \mathcal{Q}$ with cross-entropy loss
13:     $count \leftarrow count + 1$
14:     Calculate current $\mathcal{S}' \leftarrow \mathrm{meanIOU}(\mathcal{F}(\mathcal{B}_v))$
15:     if $\mathcal{S}' > \mathcal{S} + \triangle$ or $count > \mathcal{K}$ then
16:         Update previous best $\mathcal{S} \leftarrow \mathcal{S}'$
17:         Re-initialize count $count \leftarrow 0$
18:         Update the $\mathcal{Q}$ (Jump to Line 7)
19:     end if
20: end while
Return: $\mathcal{F}$
```

B. Motivation and Overview

The majority of existing consistency regularization-based approaches [11], [46], [47], [48] focus on learning feature consistency following perturbation or data augmentation, whereas CAC [1] introduces a context-aware consistency loss that compares high-level feature consistency. In contrast to it, (1) we focus on the fact that the feature representation computation is unstable (even with the participation of the feature projection [58]), and (2) we further apply the philosophy of CWC to select high-reliability pseudo-labels, and subsequent experiments validate its effectiveness, as highlighted in Figure 1.

Motivated by the aforesaid discussions, our overall optimization objective function can be defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_s + \lambda_{BCC}\mathcal{L}_{BCC} + \lambda_{DPM}\mathcal{L}_{DPM}, \tag{2}$$

$$\mathcal{L}_s = \frac{1}{|\mathcal{B}_l|}\sum_{\mathbf{X}\in\mathcal{B}_l}\frac{1}{W \times H}\sum_{i=0}^{W \times H}\left(\ell_{ce}\left(\mathbf{p}_i, y_i\right)\right), \tag{3}$$

where W and H represent the width and height of images. For the labeled dataset $\mathcal{B}_l$, the semantic segmentation model $\mathcal{F}$ is employed to generate its confidence map $\mathbf{p}_i$ (after $Softmax$ normalization), which is supervised by the ground

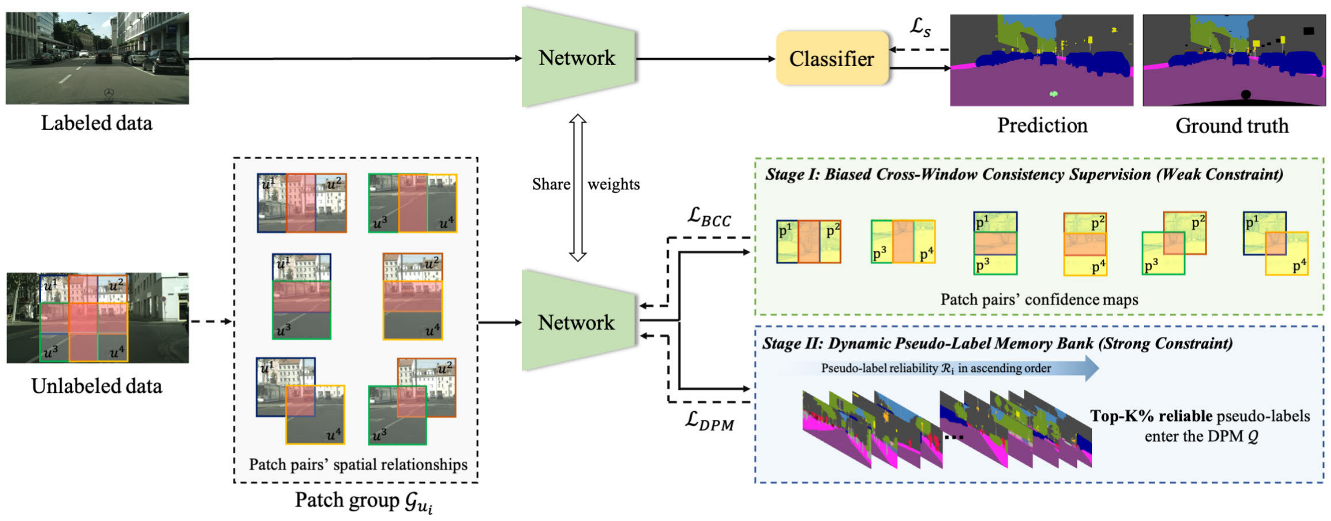

Labeled data
Network
Classifier
$\mathcal{L}_s$
Prediction
Ground truth
Share weights
Unlabeled data
u^1
u^2
u^3
u^4
Patch pairs' spatial relationships
Patch group $\mathcal{G}_{u_i}$
Network
$\mathcal{L}_{BCC}$
$\mathcal{L}_{DPM}$
Stage I: Biased Cross-Window Consistency Supervision (Weak Constraint)
p^1
p^2
p^3
p^4
Patch pairs' confidence maps
Stage II: Dynamic Pseudo-Label Memory Bank (Strong Constraint)
Pseudo-label reliability $\mathcal{R}_i$ in ascending order
Top-K% reliable pseudo-labels enter the DPM Q

Fig. 2. An overview of our proposed framework. labeled data is employed for supervised training the model $\mathcal{F}$. For unlabeled data, in stage I, we present a novel BCC loss with an importance factor (Eq. (4)) to encourages the model to maintain consistency across overlapping confidence maps in different windows but does not restrict specific class attributes. In addition, we propose the DPM to rank the pseudo-label reliability (Eqs. (7) and (8)) in light of CWC and dynamically update and manage rewarding pseudo-labels.

truth y_i using the cross-entropy loss ℓ_{ce}. As for the unlabeled dataset $\mathcal{B}_u$, our BCC loss $\mathcal{L}_{BCC}$ reflects the weak constraint, as described in Section III-C. The pseudo-label supervised loss $\mathcal{L}_{DPM}$ is the strong constraint, and Section III-D describes the pseudo-label filtering and usage.

Ideally, all items of the Eq. (2) are optimized together, but it is extremely GPU memory-consuming. To operate pervasively on the overwhelming majority of devices and considering the adaptability of our method to networks with a larger number of parameters, we propose a progressive learning strategy to achieve the ultimate optimization goal. Algorithm 1 and Figure 2 offer a comprehensive pseudocode description and an intuitive overivew of our whole framework, respectively.

C. Stage I: Biased Cross-Window Consistency Supervision (Weak Constraint)

For each unlabeled image u_i, four adjacent and overlapping patches, having a default minimum overlap size of either $\frac{H}{2}$ or $\frac{W}{2}$, are randomly cropped and defined as a patch group denoted by $\mathcal{G}_{u_i} = \left\{\left(u_i^1, u_i^2, u_i^3, u_i^4\right)\right\}_{i=1}^{N}$. The spatial relationships among these patches are illustrated in Figure 2. Any patch pairs $\left(u_i^k, u_i^l\right)$ containing overlapping regions u_{oi} are processed by the encoder $\mathcal{E}$, decoder $\mathcal{D}$, and classifier $\mathcal{C}$ to obtain a $softmax$ normalized confidence map $\mathbf{p}_i$, which is then used to calculate the BCC loss.

BCC loss encourages overlapping regions within the pair groups to have semantically consistent representations but does not restrict specific class attributes. In particular, to encourage the model to focus on **prominent** differences with different semantic classes in overlapping regions, we propose the importance factor $\mathcal{M}_{imp}$ to eliminate insignificant differences (pixel positions with different confidence maps but the same semantic classes), as illustrated in Figure 3. Intuitively, the importance factor $\mathcal{M}_{imp}$ will amplify the difference between the feature information of pixels that are more valuable to the model. Table IV experimental results demonstrate its benefits. Our BCC loss can be written as

$$\mathcal{L}_{BCC} = \frac{1}{|B_u|} \sum_{\mathbf{X} \in B_u} \frac{1}{W_o \times H_o} \sum_{i=0}^{W_o \times H_o} \sum_{1 \leqslant k < l \leqslant 4} \ell_2\left(\mathbf{p}_{oi}^k, \mathbf{p}_{oi}^l\right) \cdot \mathcal{M}_{imp}, \tag{4}$$

$$\mathbf{p}_{oi}^k = \operatorname{Softmax}\left(\mathcal{C}\left(u_{oi}^k\right)\right), \tag{5}$$

$$\mathcal{M}_{imp} = \mathbf{1}\left\{\operatorname{argmax}\left(\mathbf{p}_{oi}^k\right) \neq \operatorname{argmax}\left(\mathbf{p}_{oi}^l\right)\right\}, \tag{6}$$

where $\ell_2\left(\mathbf{p}_{oi}^k, \mathbf{p}_{oi}^l\right) = \left\|\mathbf{p}_{oi}^k - \mathbf{p}_{oi}^l\right\|_2^2$ calculates the square of the euclidean distance of the confidence maps of the overlapping regions. W_o and H_o represent the width and height of overlapping regions. k and l represent the sequence number of spatial relationships among the patch group $\mathcal{G}_{u_i}$.

1) Discussion: $\mathcal{L}_{BCC}$ uses an elegant ℓ_2 for the distance measure between the anchor and positive samples, which ensures the high efficiency of our entire framework. It achieves a 1.7% and 8.98% performance improvement on Cityscapes and MoNuSeg after the joint $\mathcal{M}_{imp}$, respectively.

D. Stage II: Dynamic Pseudo-Label Memory Bank (Strong Constraint)

To increase the quality of training pseudo-labels and prevent noise from adversely affecting the model. Supported by CWC, we come up with the concept of the DPM $\mathcal{Q}$ for selecting reliable pseudo-labels and dynamically maintaining a high confidence and high consistency pseudo-label repository to undertake pseudo-label supervision.

How to dynamically update? Previous attempts [25] to evaluate the reliability of pseudo-labels mainly focus on pixel-level filtering methods, with the common strategy being

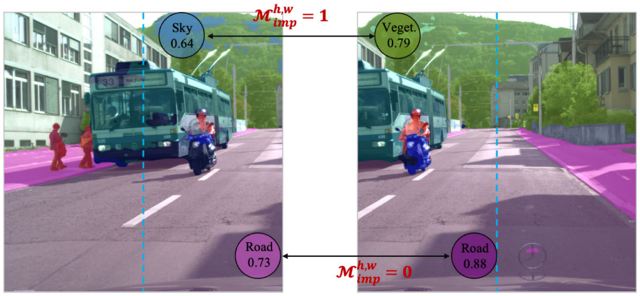


Fig. 3. Illustration of the importance factor $\mathcal{M}_{imp}$. It eliminates insignificant differences (pixel positions with different confidence maps but the same semantic classes).

to filter out low-confidence pixel information via manual or adaptive thresholding. However, filtered pixels often include complex and vital information, which may increase the negative consequences of the long-tail effect. ST++ [13] presents a method for image-level selection by evaluating the stability of pseudo-labels at various training phases. It needs to calculate the model's classification differences for numerous phases, which is time-consuming. In our framework, we investigate an efficient algorithm for evaluating the reliability of pseudo-labels in order to promote rewarding updating of DPM $\mathcal{Q}$. Specifically, we reckon that **the greater the semantic consistency of pseudo-labels in overlapping regions including patches from different contexts, the greater the reliability of pseudo-labels**. Thus, for each unlabeled image u_i, we randomly crop and form a patch group $\mathcal{G}_{u_i} = \left\{\left(u_i^1, u_i^2, u_i^3, u_i^4\right)\right\}_{i=1}^{N}$ having four adjacent and overlapping patches of size $H_o \times W_o$. We utilize the model developed currently to evaluate the meanIOU of the predictions of the four patches' overlapping regions as the reliability score $\mathcal{R}_i$.

$$\mathbb{C}_{oi} = \sum_{1 \leqslant k < l \leqslant 4} \text{ConfusionMatrix}\left(y_{oi}^{k*}, y_{oi}^{l*}\right), \tag{7}$$

$$\mathcal{R}_i = \text{meanIOU}\left(\mathbb{C}_{oi}\right), \tag{8}$$

where $y_{oi}^{k*} = \text{argmax}\left(\mathcal{C}\left(u_{oi}^{k}\right)\right)$ is pseudo masks of overlapping regions u_{oi}. The ConfusionMatrix is a confusion matrix with a size of *class number* × *class number*. The rows and columns respectively correspond to the pseudo masks from the overlapping regions of different windows. It provides a representation of the disparities between two classifications generated by a model [59]. After getting the reliability scores of all unlabeled images, we sort the entire set of unlabeled images based on these scores and select the Top-K% reliable pseudo-labels and corresponding images to entry into the DPM $\mathcal{Q}$. (The original pseudo-labels in the DPM will be totally replaced.)

When will an update be made? To maintain the optimal pseudo-labels in the DPM at all times, the DPM will be automatically updated when the model achieves a gain of $\triangle$ on the validation set or when it reaches a predefined $\mathcal{K}$ epochs of training. This ensures that the model always obtains more rewarding information from the DPM.

We employ pseudo-labels y_{ui}^* from the DPM to rigorously supervise the confidence map $\mathbf{p}_{ui}$ of the corresponding model output using cross-entropy loss:

$$\mathcal{L}_{DPM} = \frac{1}{|\mathcal{B}_u|} \sum_{\mathbf{X} \in \mathcal{B}_u} \frac{1}{W \times H} \sum_{i=0}^{W \times H} \left(\ell_{ce}\left(\mathbf{p}_{ui}, y_{ui}^*\right)\right), \tag{9}$$

where W and H represent the width and height of unlabeled images, respectively.

1) Discussion: Instead of always utilizing all unlabeled images and corresponding pseudo-labels [60], our DPM dynamically filters out the less-reliable pseudo-labels based on CWC. In addition, our DPM updates information more frequently and flexibly in order to assist the model in learning more rewarding unlabeled data. The experimental results listed in Table IV and Figure 7 demonstrate its benefits.

IV. Experiments

A. Dataset Description

We evaluate our approach on three publicly available benchmarks to encompass various application scenarios such as urban street scenes semantic segmentation, histopathological tissue detection, and land cover classification, represented by Cityscapes [22], MoNuSeg [23], and DeepGlobe [24]. Cityscapes contains 2975 training images with fine-annotated labels of 19 semantic classes, 500 validation images, and 1525 test images. We compare our method with state-of-the-art methods under 1/30, 1/16, 1/8, and 1/4 partition protocols following ST++ [13] and CPS [12]. MoNuSeg is published by the multi-organ nuclei segmentation challenge [23] and consists of 30, 7, and 14 histopathologic images (1000 × 1000 pixels) for training, validation, and testing, respectively. Our and other methods are implemented under 1/30, 1/6, 1/3, and full supervision partition protocols. DeepGlobe contains 803 satellite images (2448 × 2448 pixels) that are applied for land cover classification analysis in the field of remote sensing. Following [61], we divide the images into the training set, validation set, and test set with 454, 207, and 142 images, respectively. Similarly, our and other methods are implemented under 1/16, 1/8, 1/4, and full supervision partition protocols.

B. Evaluation Protocol

We employ DeepLabv3+ [62] with ResNet-50 [63] that has been pre-trained on ImageNet [64] as our segmentation model to ensure a fair comparison with prior work. We use the mean intersection-over-union (mIOU) metric to evaluate the segmentation performance of all datasets, following previous work. For a comprehensive evaluation of MoNuSeg, we also used Dice coefficient (DC) and Jaccard coefficient (JC), which are commonly used in biomedical segmentation. In line with established conventions, we report results on the 500 Cityscapes `val` set, the 14 MoNuSeg `test` set, and the 142 DeepGlobe `test` set using only single-scale testing and without any post-processing techniques.

C. Implementation Details

The batch-size is set to 2 in Stage I and 4 in Stage II. The initial learning rate of Stage I of the backbone is 0.005, 0.004, and 0.005 for Cityscapes, MoNuSeg, and DeepGlobe, respectively, whereas the learning rate of the segmentation

TABLE I

Comparison With State-of-the-Art Methods on **Cityscapes** val Set Under Different Partition Protocols. We Use DeepLabv3+ as the Segmentation Network and ResNet-50 as the Backbone. “SupOnly” Means Supervised Training Without Using Any Unlabeled Data. † Means We Reproduce the Approach and Other Results Are Collected From [12], [38], [54]. * Implies That the Image Resolution Is 800px and CutMix Is Applied. The Best Results Are Marked in **red bold** and the Second Best Ones Are Marked in blue

Method	Publication	1/30 (100)	1/16 (186)	1/8 (372)	1/4 (744)
SupOnly†	-	55.1	61.8	66.2	72.3
CPS [12]	CVPR’21	-	69.8	74.4	76.9
CAC [1]	CVPR’21	60.9	69.4	74.0	-
DARS [56]	ICCV’21	-	66.9	73.7	-
ST++† [13]	CVPR’22	61.4	70.1	73.2	74.7
U^{2}PL [49]	CVPR’22	59.8	70.6	73.0	76.3
USRN [54]	CVPR’22	-	71.2	75.0	-
PS-MT [38]	CVPR’22	-	-	74.4	75.2
CPCL [66]	TIP’23	-	69.9	74.6	77.0
Ours | 720	-	**67.3**	**72.8**	**76.6**	**77.6**
CPS* [12]	CVPR’21	-	74.5	76.6	77.8
n-CPS* [67]	Arxiv’21	-	**76.1**	77.6	78.4
PS-MT* [38]	CVPR’22	-	-	77.1	78.4
UniMatch* [68]	CVPR’23	64.5	75.0	76.8	77.5
LaserMix* [69]	CVPR’23	-	75.5	77.1	78.3
Ours* | 800	-	**68.6**	75.5	**77.7**	**78.7**

TABLE II

Comparison With State-of-the-Art Methods on **MoNuSeg** test Set Under Different Partition Protocols. All Methods Are Reproduced by Us Via DeepLabv3+ With ResNet-50 for a Fair Comparison. The Best Results Are Marked in **red bold** and the Second Best Ones Are Marked in blue

Partition	Method	DC (%)	mIOU (%)	JC (%)
1/30 (1)	SupOnly	61.87	60.13	45.54
	CutMix [35]	65.35	64.89	48.75
	CAC [1]	63.29	60.88	46.84
	ST++ [13]	68.68	67.05	52.49
	UniMatch [68]	63.77	63.51	47.15
	Ours	**75.51**	**74.46**	**60.89**
1/6 (5)	SupOnly	73.16	71.95	58.52
	CutMix [35]	74.71	73.77	59.91
	CAC [1]	74.10	73.40	59.21
	ST++ [13]	77.85	76.46	63.83
	UniMatch [68]	74.96	75.32	60.43
	Ours	**79.62**	**78.62**	**66.28**
1/3 (10)	SupOnly	74.53	72.72	60.07
	CutMix [35]	76.72	76.07	62.73
	CAC [1]	74.43	74.80	59.63
	ST++ [13]	78.41	77.27	64.62
	UniMatch [68]	79.15	77.84	65.63
	Ours	**80.37**	**78.48**	**67.27**
Full (30)	SupOnly	77.82	76.82	64.20
	CutMix [35]	78.13	77.03	64.50
	CAC [1]	80.71	79.35	67.73
	ST++ [13]	79.98	78.80	66.79
	UniMatch [68]	78.15	77.31	64.33
	Ours	**81.00**	**79.73**	**68.18**

head is 10 times that of the backbone. The initial learning rate of Stage II is reduced to 0.003, 0.003, and 0.004. We use the SGD optimizer to train Cityscapes, MoNuSeg, and DeepGlobe for 280, 200, and 150 epochs under a poly learning rate scheduler, respectively. Following ST++ [13], the labeled data is randomly flipped and resized within a range from 0.5 to 2.0. Meanwhile, unlabeled images are augmented using color jitter, grayscale, and blur. The training image resolution (i.e., the window size) is set to 720/800 for Cityscapes and 512 for MoNuSeg and DeepGlobe. Images of the window size are randomly cropped from the original images. To get a fair comparison result for Cityscapes, we employ OHEM loss with the same parameters as previous work. For MoNuSeg and DeepGlobe, we train all models using only standard cross-entropy loss, without Sync-BN [65] and auxiliary loss. Moreover, the trade-off weights λ_{BCC} and λ_{DPM} are set to 0.16 and 1.0, respectively. And the selection of reliable pseudo-labels in Stage II selects, by default, the top 50% of data for storage in the DPM. When the validation set metric achieves a 2% gain or *count* reaches 25, the DPM is automatically updated with more trustworthy pseudo-labels.

D. Comparison With State-of-the-Art Methods

An innovative framework based on the CWC traits of images is proposed. In this section, we implement advanced methods on various datasets with the same segmentation network and setting to ensure the fairness of comparison.

1) Performance Comparison on Cityscapes: Table I shows the results of our method compared with other state-of-the-art methods on the Cityscapes dataset. We reproduce the representative methods within the same network and setting according to their publicly available codes or use the results reported in the original papers. Our framework achieves a stable improvement under different partition protocols. Specifically, our framework outperforms the supervised baseline (SupOnly) by +12.2%, +11.0%, +10.4%, and +5.3% under 1/30, 1/16, 1/8, and 1/4 partition protocols, respectively. Besides, ours outperforms state-of-the-art methods with larger margins by 4.1%, 1.6%, and 1.6% in mIOU for 1/30, 1/16, and 1/8 split of Cityscapes. We present some qualitative results under 1/8 protocol with a training resolution of 720 in Figure 4. In comparison to the previous state-of-the-art method, our method displays more accurate segmentation results thanks to the proposed progressive learning strategy.

2) Performance Comparison on MoNuSeg: Table II shows the comparison results on the MoNuSeg dataset. Compared with latest and advanced methods, ours surpasses them with large margins under all partition protocols (especially when there are few labeled samples participating in training). The DC gap between the SupOnly on full set (77.82%) and our 1/30 labeled setting result (75.51%) is only 2.31%. Under the 1/6 labeled setting, ours outperforms the supervised baseline on the full set (79.62% vs. 77.82%). We also observe that even under full supervision, our framework still obtains a +3.18% gain. Qualitative results under 1/3 protocol are displayed in Figure 5 on the MoNuSeg test set.

3) Performance Comparison on DeepGlobe: We show the comparison results for the DeepGlobe dataset in Table III. Ours brings significant and stable improvements compared to the SupOnly and other popular advanced methods under

TABLE III

COMPARISON WITH STATE-OF-THE-ART METHODS ON **DEEPGLOBE** TEST SET UNDER DIFFERENT PARTITION PROTOCOLS. ALL METHODS ARE REPRODUCED BY US VIA DEEPLABV3+ WITH RESNET-50 FOR A FAIR COMPARISON. * MEANS ONLY USING THE BCC SUPERVISION (STAGE I), AND DISCUSSIONS ABOUT THE FULL DATA SETTING ARE PRESENTED IN SECTION IV-K. THE BEST RESULTS ARE MARKED IN **RED BOLD** AND THE SECOND BEST ONES ARE MARKED IN BLUE

Method	1/16 (28)	1/8 (56)	1/4 (113)	Full* (454)
SupOnly	55.47	62.19	66.82	68.64
CutMix [35]	53.33	61.46	66.07	66.59
CAC [1]	56.47	62.23	66.26	69.45
ST++ [13]	55.61	62.21	65.67	69.25
Ours	**58.36**	**65.17**	**68.51**	**70.01**

all partition protocols. Besides, the mIOU gap between the SupOnly on full set (68.64%) and our 1/4 labeled setting result (68.51%) is only 0.13%. We display some qualitative results under 1/4 protocol in Figure 6 on the DeepGlobe test set.

E. Ablation Studies

The notable contributions of our framework are condensed into 1) BCC loss with the importance factor, 2) an efficient method for evaluating reliability of pseudo-labels, and 3) the DPM. We conduct our ablation studies with DeepLabv3+ and ResNet-50 on the 1/8 split of Cityscapes (training image resolution: 720) to verify the effectiveness of them.

1) Effectiveness of the BCC Loss: In Table IV, we show the outcomes of the naive consistency loss (Exp. I) and our proposed BCC loss $\mathcal{L}_{BCC}$ with the importance factor $\mathcal{M}_{imp}$ (Exp. II), demonstrating that the importance factor $\mathcal{M}_{imp}$ can provide a +1.7% gain over the supervised baseline method (Exp. SupOnly) with only labeled data. And the performance of the model declines (−1.5%) when $\mathcal{L}_{BCC}$ is not used in the first stage of training (Exp. VII). They indicate that BCC loss encourages overlapping regions to maintain prominent semantic consistency.

2) Impact of Parameters of the BCC Loss: For each unlabeled image u_i, four adjacent and overlapping patches with a window size of $H \times W$ are randomly cropped and designated as a patch group. And the default minimum overlap size is defined as $\frac{H}{2}$ or $\frac{W}{2}$. The size of the overlapping area determines the difference in the context information contained between adjacent patches. We conduct an ablation study on different minimum overlap sizes, as shown in Table V, which demonstrates that most suitable size for $\mathcal{L}_{BCC}$ is $\frac{H}{2}$ or $\frac{W}{2}$.

we further consider n patch pairs with overlapping areas in each patch group. In Table VI, we find that when n reaches the maximum value of 6, the result outperforms other counterparts, which proves that the wider cross-window information involved is beneficial for mining unlabeled data.

3) Effectiveness of the Pseudo-Label Reliability Evaluation: A perfect model maintains a consistent and robust self-awareness across different contextual environments, as evidenced by its ability to generate consistent overlapping region predictions. In contrast, a suboptimal model (e.g., a randomly initialized model) yields confused predictions when presented with images from different contextual windows. The performance of models trained at different stages differs for each unlabeled image. Our method for evaluating reliability is based on the above philosophy and design to identify the Top-K unlabeled data that are best suited to the current model.

We verify its effectiveness from two perspectives: (1) as shown in Experiments III and IV in Table IV, the mIOU gap between random selection (50%) and reliable selection (50%) based on Eq. (7) and (8) is 1.3%, indicating that our pseudo-label reliability evaluation is effective. (2) We directly train the model with all unlabeled images and their pseudo-labels, and its performance is nearly identical to that of a model trained with only 50% high-reliable pseudo-labels. This indicates our approach's capability to reduce noise interference in pseudo-labels, as we mentioned in Section III-D.

In addition, to further illustrate the advantages of this image-level reliability selection, we implement a comparison with other candidates: 1) pixel-level filtering (like FixMatch), 2) image-level filtering based on softmax scores, and 3) model stability at multiple stages as reliability (like ST++). Specifically, the pixel-level filtering method ignores the pseudo-label information of the pixels with a maximum class confidence of less than 0.75 during the training phase. It is noteworthy to mention that filtering low-confidence pixel information often discards critical information, which can lead to negative impacts from the long-tail effect. As shown in Table VII, our image-level reliability selection is superior to others.

4) Efficiency of the Pseudo-Label Reliability Evaluation: We employ the same settings and device to make a fair comparison with ST++ [13] on the efficiency of evaluating the reliability of pseudo-labels. ST++ needs to calculate the model's classification differences for numerous phases. In contrast to ST++ which needs to calculate model classification differences in multiple stages, our method only needs to calculate model differences in overlapping areas in a single stage. In Figure 7(a), the efficiency of the pseudo-label evaluation method based on CWC proposed by us **(2.29 FPS)** is twice that of ST++ **(1.12 FPS)**, which provides support for the high dynamic performance of DPM.

5) Effectiveness of the DPM: In Table IV, the comparison between our method's result and that of Experiment IV demonstrates that the DPM can bring a significant improvement by +3.6%. In addition, the difference between Experiment III and Experiment VI demonstrates the advantage of the DPM even when random selection is employed.

Furthermore, Figure 7(b) illustrates the model performance of the DPM at each automatic update, as well as the renewal ratio of the images corresponding to the memory bank's pseudo-labels. Specifically, approximately 35% of unlabeled images are changed at each update, demonstrating the high dynamics of the DPM. With the continual update of DPM, our framework's performance is also continuously enhanced, indicating that it is wise to gradually utilize high-reliability pseudo-labels instead of all pseudo-labels to optimize the model.

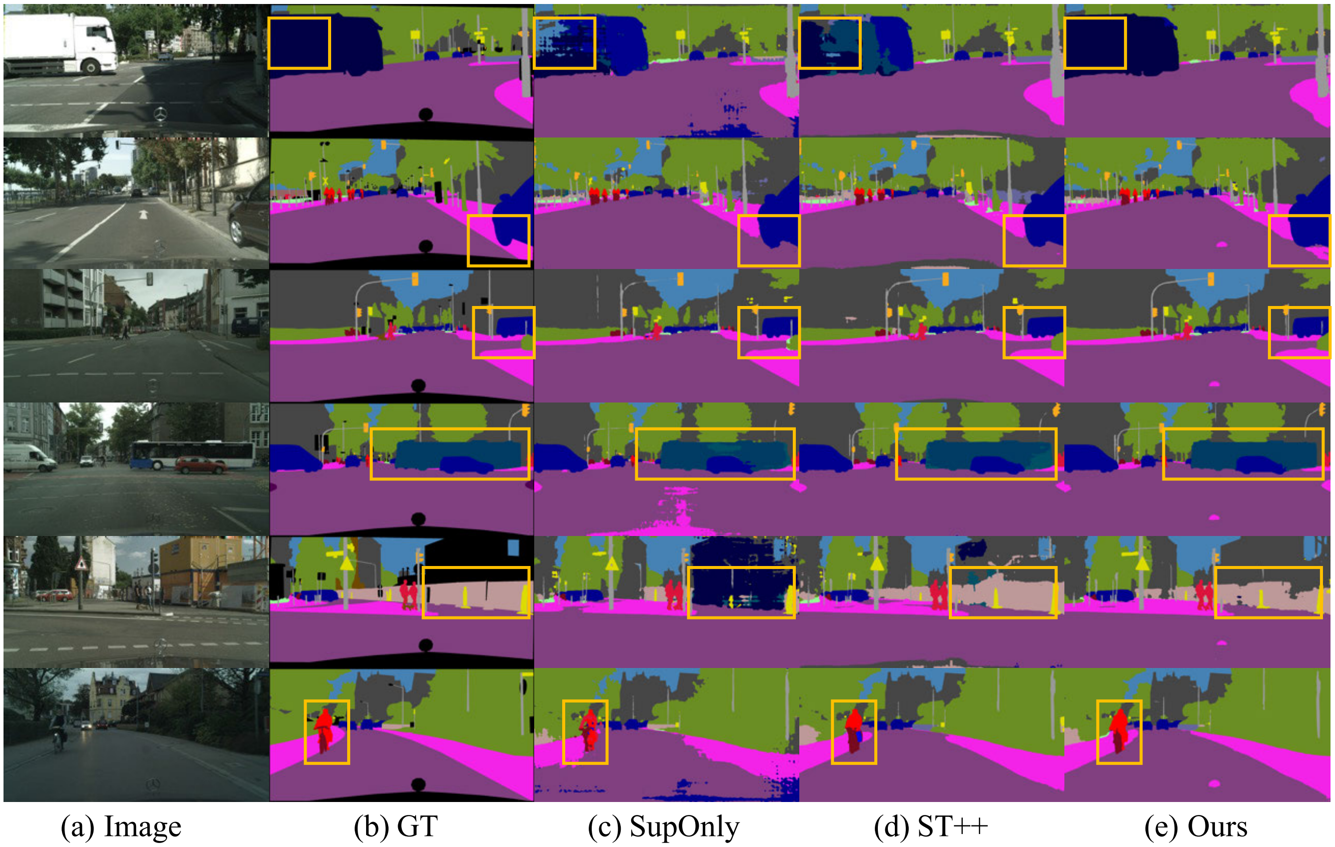


Fig. 4. Qualitative results on the Cityscapes val set.(a) and (b) are corresponding to images and Ground Truth(GT), (c) represents the results of supervised baseline(SupOnly), (d) is the results of ST++ [13], and (e) is the results of our framework. Orange rectangles highlight the difference among of them.

TABLE IV

ABLATION STUDY ON THE EFFECTIVENESS OF VARIOUS COMPONENTS IN OUR FRAMEWORK. $\mathcal{L}_{BCC}$: BIASED CROSS-WINDOW CONSISTENCY LOSS, $\mathcal{M}_{imp}$: IMPORTANCE FACTOR, RANDOM SELECT (50%) MEANS SELECTING 50% PSEUDO-LABELS TO RETRAIN RANDOMLY. RELIABLE SELECT (50%) MEANS SELECTING TOP-50% RELIABLE PSEUDO-LABELS TO RETRAIN BASED ON EQS. (7) AND (8). ALL UNLABELED DATA MEANS USING ALL PSEUDO-LABELS TO RETRAIN DIRECTLY. DPM: DYNAMIC PSEUDO-LABEL MEMORY BANK $\mathcal{Q}$

ID	$\mathcal{L}_{BCC}$ (w/o $\mathcal{M}_{imp}$)	$\mathcal{L}_{BCC}$ (w/ $\mathcal{M}_{imp}$)	Random select (50%)	Reliable select (50%)	All unlabeled data	DPM $\mathcal{Q}$	mIOU(%)
SupOnly							66.2
I	✔						66.5
II		✔					67.9
III		✔	✔				71.7
IV		✔		✔			73.0
V		✔			✔		73.1
VI		✔	✔			✔	73.8
VII				✔		✔	75.1
Ours		✔		✔		✔	**76.6**

TABLE V

ABLATION STUDY ON DIFFERENT MINIMUM OVERLAP SIZES. H AND W REPRESENT THE HEIGHT AND WIDTH OF THE INPUT IMAGE

Minimum overlap size	$\frac{H}{3}$ or $\frac{W}{3}$	$\frac{H}{2}$ or $\frac{W}{2}$ (default)	$\frac{2H}{3}$ or $\frac{2W}{3}$
mIOU (%)	67.1	**67.9**	67.6

TABLE VI

ABLATION STUDY ON DIFFERENT NUMBER OF PATCH PAIRS IN EACH PATCH GROUP

n	0	1	3	6 (default)
mIOU (%)	66.2	66.6	66.5	**67.9**

We also conduct ablation experiments on the ratio of selected reliable pseudo-labels. The default setting 50% is effective enough, as demonstrated in Table VIII.

F. Experimental Evidence and Analysis of Unstable Feature Representation Computation

We conduct experiments to compare the performance of computing consistency using confidence maps and feature

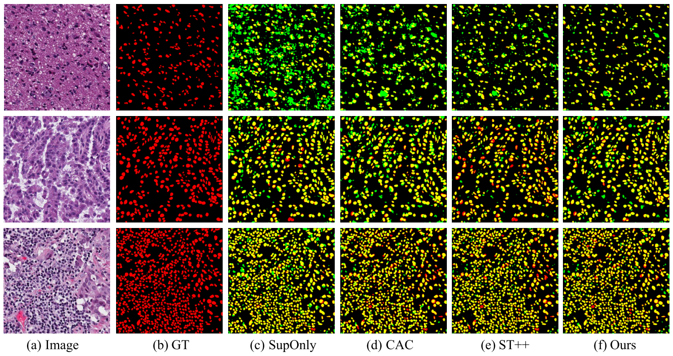


Fig. 5. Qualitative results on the MoNuSeg test set.(a) and (b) are corresponding to images and Ground Truth(GT), (c) represents the results of supervised baseline(SupOnly), (d) and (e) is the results of CAC [1] and ST++ [13], and (f) is the results of our framework. Green and red present the predictions and ground truth respectively, while yellow indicates their overlap regions.

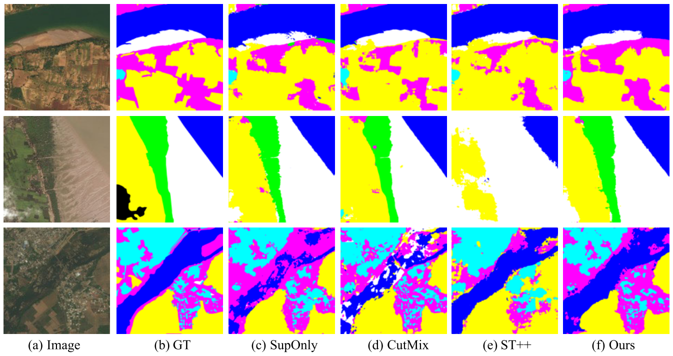


Fig. 6. Qualitative results on the DeepGlobe test set.(a) and (b) are corresponding to images and Ground Truth(GT), (c) represents the results of supervised baseline(SupOnly), (d) and (e) is the results of CutMix [35] and ST++ [13], and (f) is the results of our framework. Cyan represents "urban", yellow represents "agriculture", magenta represents "rangeland", green represents "forest", blue represents "water", white represents "barren" and black represents "unknown".

TABLE VII

EFFECTIVENESS OF THE PSEUDO-LABEL RELIABILITY EVALUATION. PERFORMANCE ANALYSIS OVER PIXEL-LEVEL FILTERING STRATEGY AND OUR IMAGE-LEVEL RELIABILITY SELECTION STRATEGY

Method	mIOU(%)
Pixel-level filtering	72.5
image-level filtering based on softmax scores	71.9
ST++	73.2
Ours	**76.6**

TABLE VIII

ABLATION STUDY ON THE RATIO OF RELIABLE PSEUDO-LABELS

Ratio	20%	50%(default)	80%
mIOU (%)	74.5	76.6	76.3

TABLE IX

COMPARISON OF DIFFERENT INPUTS IN $\mathcal{L}_{BCC}$

	Feature map	Feature map + Φ (proj_dim=64)	Feature map + Φ (proj_dim=128)	Ours (stage I)
mIOU (%)	64.5	65.2	63.9	**67.9**

TABLE X

COMPARISON WITH STATE-OF-THE-ART METHODS ON **CITYSCAPES** `VAL` SET UNDER DIFFERENT PARTITION PROTOCOLS. WE USE DEEPLABV3+ AS THE SEGMENTATION NETWORK AND **RESNET-101** AS THE BACKBONE. † MEANS WE REPRODUCE THE APPROACH AND OTHER RESULTS ARE COLLECTED FROM [12], [38]. * IMPLIES THAT THE IMAGE RESOLUTION IS GREATER THAN 720PX AND CUTMIX IS APPLIED. THE BEST RESULTS ARE MARKED IN **RED BOLD** AND THE SECOND BEST ONES ARE MARKED IN BLUE

Method	Publication	1/16 (186)	1/8 (372)	1/4 (744)
SupOnly†	-	62.2	69.1	72.3
CCT [11]	CVPR'20	69.6	74.5	76.4
GCT [70]	ECCV'20	66.9	73.0	76.5
AEL [55]	NIPS'21	**74.5**	75.6	77.5
PS-MT [38]	CVPR'22	-	76.9	77.6
ST++† [13]	CVPR'22	70.3	73.9	76.8
PCR [47]	NIPS'22	73.4	76.3	78.4
Ours | 720	-	**74.5**	**77.0**	**78.6**
U^{2}PL* [49]	CVPR'22	70.3	74.4	76.5
UniMatch* [68]	CVPR'23	**76.6**	77.9	**79.2**
ESL* [71]	ICCV'23	75.1	77.2	78.9
Ours* | 800	-	75.8	**78.0**	78.9

maps. Table IX show that using feature maps before the classifier to calculate CWC significantly reduces performance, although the performance can be improved to some extent when nonlinear projectors Φ are added. The larger channel dimension of feature maps consumes more memory and cannot be extended to wider contextual windows, which contrasts with confidence maps whose dimension is limited to the number of semantic classes. Moreover, we posit that confidence maps contain direct information about semantic classes, while feature maps represent broader and more vague information, which is a crucial factor in explaining the performance gap between them. Furthermore, our experiments indicate that the $\mathcal{L}_{BCC}$ outperforms CAC in both `DeepGlobe` and `MoNuSeg`. This finding, combined with the observation that the performance of CAC drops by approximately 10% without Φ, corroborates the claim that feature representation calculation is unstable.

G. Comparison of Different Backbones

Similar to previous methods, we also adopt DeepLabv3+ [62] with **ResNet-101** [63] as the segmentation network and conducted experiments on Cityscapes dataset. The results presented in Table X suggest that in most cases, the accuracy of our method is generally comparable to that of previous methods. Moreover, the improvement in accuracy compared to the supervised baseline (SupOnly) demonstrates that our framework remains effective even without relying on a specific backbone.

H. Evaluation on the Larger Segmentation Dataset

To further substantiate the efficacy of our methodology and its applicability across diverse segmentation models,

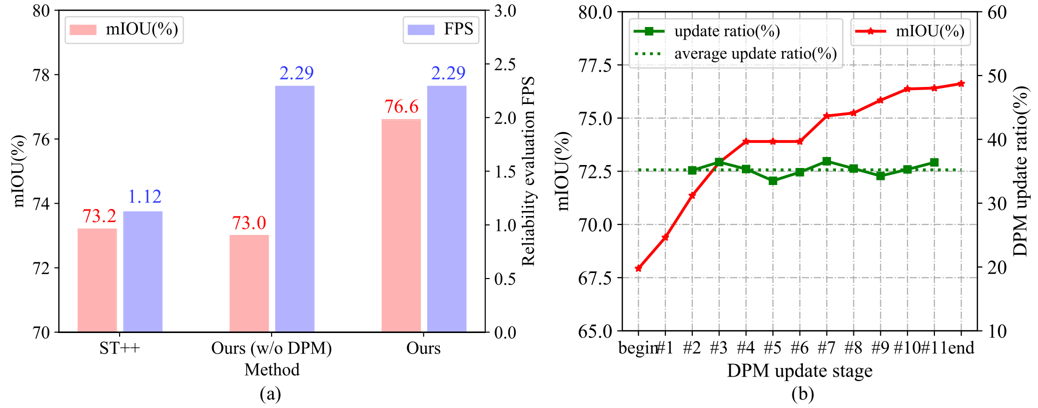


Fig. 7. (a) Comparison between the efficiency and accuracy of evaluating the reliability of pseudo-labels. (b) The model performance and the update ratio of the DPM at each automatic update stage. #*k*: the *k*-th update and training. `begin`: the beginning of Stage II. `end`: the end of training.

TABLE XI

Comparison With State-of-the-Art Methods on **PASCAL VOC2012** *augmented* Set Under Different Partition Protocols. We Use the Same Split as U^{2}PL [49]. The Best Results Are Marked in **Red Bold** and the Second Best Ones Are Marked in Blue

Method	Publication	1/16 (662)	1/8 (1323)	1/4 (2646)
U^{2}PL [49]	CVPR'22	77.2	79.0	79.3
GTA-Seg [16]	NIPS'22	77.8	80.5	80.6
SemiCVT [74]	CVPR'23	78.2	80.0	80.2
UniMatch [68]	CVPR'23	80.9	81.9	80.4
LogicDiag [17]	ICCV'23	79.7	80.2	80.6
Ours	-	**82.2**	**83.5**	**83.6**

TABLE XII

The Training Burden of Our Method on Cityscapes Dataset With 1/8 Labeled Data

	Training time (each epoch) (s)	GPU Memory (M)	Params (M)
Stage I	309.81	15483	40.475
Stage II	713.85	13827	40.475

TABLE XIII

Results (IoU) of Different Classes on the **DeepGlobe** test Set Under Different Partition Protocols. We Use DeepLabv3+ as the Segmentation Network and ResNet-50 as the Backbone. *Gain*: The mIOU Gain of Between SupOnly and Our Approach

	Urban	Agriculture	Rangeland	Forest	Water	Barren	mIOU(%)
Partition Protocol : 1/16 (28)							
SupOnly	74.49	77.23	28.11	61.88	62.57	28.53	55.47
Ours	75.50	75.52	23.96	63.47	75.36	36.36	58.36
Gain	+1.01	-1.71	-4.15	+1.59	+12.79	+7.83	+2.89
Partition Protocol : 1/8 (56)							
SupOnly	75.44	82.77	29.21	67.32	67.67	50.73	62.19
Ours	75.47	82.18	32.73	69.22	77.13	54.32	65.17
Gain	+0.03	-0.59	+3.52	+1.90	+9.46	+3.59	+2.98
Partition Protocol : 1/4 (113)							
SupOnly	72.56	85.09	33.17	76.99	74.15	58.94	66.82
Ours	77.37	84.80	35.36	76.36	77.18	60.00	68.51
Gain	+4.81	-0.29	+2.19	-0.63	+3.03	+1.06	+1.69
Partition Protocol : Full (454)							
SupOnly	76.52	85.13	37.73	75.80	75.99	60.68	68.64
Ours	78.19	86.08	40.23	75.50	78.86	61.16	70.01
Gain	+1.67	+0.95	+2.50	-0.30	+2.87	+0.48	+1.37

we conduct comparative experiments on PASCAL VOC2012 *augmented* dataset [72]. The PASCAL VOC2012 dataset is a large benchmark for semantic segmentation tasks. The *original* set comprises approximately 4000 samples, which have been meticulously partitioned into three subsets: train, val, and test, containing 1464, 1449, and 1456 images, respectively. This dataset offers pixel-level annotations for 21 categories, including the background class. In line with established conventions [12], [49], we augment the training data by incorporating 9118 coarsely labeled images from the SBD dataset [73]. Recently, Transformer-based models have made remarkable advancements in the field of semantic segmentation. Recently, there has been a surge of interest in exploring CNN-Transformer-based approaches in the realm of semi-supervised semantic segmentation, as evidenced by the emergence of methods such as SemiCVT [74] and others [75], [76]. In our study, we employ the Swin Transformer-Base [77] architecture as the encoder, while relying on the widely adopted UPerNet [78] as the decoder. The other experimental configurations remain consistent with previous work [68].

The performance improvements observed in all partition protocols, as depicted in Table XI, showcase the adaptability of our approach when applied to advanced segmentation networks. Furthermore, these results highlight the generalizability of our method when dealing with relatively large-scale unlabeled data.

I. Training Burden

Table XII shows the training burden of our approach at different stages. And the time consumed for each update of DPM is 20.6 minutes on Cityscapes dataset with 1/8 labeled data.

J. Per-Class Results

In Tables XIII and XIV, we present in detail the IOU performance of our results and some other methods for **per-class** on DeepGlobe and Cityscapes datasets, respectively. It is

TABLE XIV

Results (IoU) of Different Classes on the **Cityscapes** `val` Set Under Different Partition Protocols. We Use DeepLabv3+ as the Segmentation Network and ResNet-101 as the Backbone. *Gain*: The mIOU Gain of Between SupOnly and Our Approach. Our Framework Achieves the Most Significant Improvement on the Tailed Classes (*e.g.*, '`wall`', '`rider`', '`truck`', '`bus`', and '`train`'), Indicating That Our Method Alleviates the Class Imbalance Issue to a Certain Extent

	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	mIOU(%)
Partition Protocol : 1/16 (186)																				
SupOnly	96.0	73.9	89.4	29.2	41.3	54.9	62.9	70.8	90.6	56.6	92.3	74.2	44.5	91.5	36.7	42.0	23.6	39.8	70.7	62.2
Ours	97.7	82.6	91.2	49.0	55.8	59.7	68.5	77.7	92.0	60.0	94.0	80.2	57.5	94.6	73.1	76.9	68.5	60.7	75.1	74.5
Gain	+1.7	+8.7	+1.8	+19.8	+14.5	+4.8	+5.6	+6.9	+1.4	+3.4	+1.7	+6.0	+13.0	+3.1	+36.4	+34.9	+44.9	+20.9	+4.4	+12.3
Partition Protocol : 1/8 (372)																				
SupOnly	96.5	77.1	90.7	37.6	51.5	60.1	64.9	74.8	91.5	55.8	93.4	76.6	51.5	93.1	52.0	67.2	48.5	55.5	73.6	69.1
Ours	97.9	83.9	92.3	58.7	60.4	63.5	70.8	79.2	92.3	60.9	94.7	81.9	62.1	95.1	76.1	79.2	71.8	64.8	76.9	77.0
Gain	+1.4	+6.8	+1.6	+21.1	+8.9	+3.4	+5.9	+4.4	+0.8	+5.1	+1.3	+5.3	+10.6	+2.0	+24.1	+12.0	+23.3	+9.3	+3.3	+7.9
Partition Protocol : 1/4 (744)																				
SupOnly	97.5	81.4	91.2	37.6	55.8	63.4	68.7	77.1	91.6	57.8	93.8	79.1	57.1	93.4	60.7	73.7	55.6	63.0	75.0	72.3
Ours	97.7	83.3	92.8	61.9	63.1	64.9	71.2	79.7	92.5	60.0	94.4	82.8	64.4	95.3	76.8	87.1	78.8	68.7	77.8	78.6
Gain	+0.2	+1.9	+1.6	+24.3	+7.3	+1.5	+2.5	+2.6	+0.9	+2.2	+0.6	+3.7	+7.3	+1.9	+16.1	+13.4	+23.2	+5.7	+2.8	+6.3

TABLE XV

Results (mIOU%) About the Varying Scale of Unlabeled Data in the Full Data Setting on DeepGlobe `test` Set

Scale	1/16 (28)	1/8 (56)	1/4 (113)	1/2 (227)	Full (454)
SupOnly	-	-	-	-	68.64
Stage I	68.29	68.31	68.63	69.03	**70.01 (+1.37)**
Stage II	68.13	67.37	66.86	66.11	65.47

worth noting that our framework achieves the most significant improvement on the tailed classes (*e.g.* , '`wall`', '`rider`', '`truck`', '`bus`', and '`train`'), indicating that our method alleviates the class imbalance issue to a certain extent.

K. Discussion About the Full Data Setting

In the full data setting on DeepGlobe dataset, images fed to the unsupervised branch are collected from the labeled training set, as CAC [1] does. Here we add an ablation study on the effect of the varying scale of unlabeled data in the full data setting. Table XV shows that BCC supervision (Stage I) is also beneficial in the full data setting, and with the increase in the scale of unlabeled data, performance gradually improves. It is recommended to only use Stage I leads to the optimal performance in the full data setting.

V. Conclusion

We propose a progressive learning framework for developing CWC systematically via mining weak-to-strong constraints. At the early stage, we propose a BCC loss with the importance factor to encourage the model to maintain consistency across overlapping confidence maps in different windows but does not restrict specific class attributes. We conceptualize the DPM to dynamically update and manage high-reliability pseudo-labels to strongly constrain the model in the latter period. The evaluation strategy of pseudo-label reliability based on CWC is the key of DPM. Our framework achieves the state-of-the-art performance across three representative datasets from various fields.

References

[1] X. Lai et al., "Semi-supervised semantic segmentation with directional context-aware consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1205–1214.

[2] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[4] Y. Li, B. Dang, W. Li, and Y. Zhang, "GLH-water: A large-scale dataset for global surface water detection in large-size very-high-resolution satellite imagery," in *Proc. AAAI*, 2024, vol. 38, no. 20, pp. 22213–22221.

[5] Y. Li, T. Shi, Y. Zhang, and J. Ma, "SPGAN-DA: Semantic-preserved generative adversarial network for domain adaptive remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5406717, doi: 10.1109/TGRS.2023.3313883.

[6] X. Guo et al., "SkySense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery," in *Proc. CVPR*, Jun. 2024, pp. 27672–27683.

[7] Y. Li et al., "MFVNet: A deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation," *Sci. China Inf. Sci.*, vol. 66, no. 4, Apr. 2023, Art. no. 140305.

[8] H. Ling, J. Gao, A. Kar, W. Chen, and S. Fidler, "Fast interactive object annotation with curve-GCN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5257–5266.

[9] Y. Hao et al., "EdgeFlow: Achieving practical interactive segmentation with edge-guided flow," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1551–1560.

[10] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5688–5696.

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

[12] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2613–2622.

[13] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "ST++: Make self-training work better for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 4268–4277.

[14] J. Fan, B. Gao, H. Jin, and L. Jiang, "UCC: Uncertainty guided cross-head cotraining for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9937–9946.

[15] H. Wu, Z. Wang, Y. Song, L. Yang, and J. Qin, "Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11656–11665.

[16] Y. Jin, J. Wang, and D. Lin, "Semi-supervised semantic segmentation via gentle teaching assistant," in *Proc. NeurIPS*, vol. 35, 2022, pp. 2803–2816.

[17] C. Liang, W. Wang, J. Miao, and Y. Yang, "Logic-induced diagnostic reasoning for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 16197–16208.

[18] K. Chen, L. Hong, H. Xu, Z. Li, and D.-Y. Yeung, "MultiSiam: Self-supervised multi-instance Siamese representation learning for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7526–7534.

[19] M. Ko et al., "Self-supervised dense consistency regularization for image-to-image translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18301–18310.

[20] J. Y. Lettvin, "On seeing sidelong," *Science*, vol. 16, no. 4, pp. 10–20, Jul. 1976.

[21] J. Min, Y. Zhao, C. Luo, and M. Cho, "Peripheral vision transformer," in *Proc. NeurIPS*, vol. 35, 2022, pp. 32097–32111.

[22] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[23] N. Kumar et al., "A multi-organ nucleus segmentation challenge," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1380–1391, May 2020.

[24] I. Demir et al., "DeepGlobe 2018: A challenge to parse the earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.

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

[26] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. NeurIPS*, vol. 17, 2004, pp. 529–536.

[27] B. Zoph et al., "Rethinking pre-training and self-training," in *Proc. NeurIPS*, vol. 33, 2020, pp. 3833–3845.

[28] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2018.

[29] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "MixMatch: A holistic approach to semi-supervised learning," in *Proc. NeurIPS*, vol. 32, 2019, pp. 5049–5059.

[30] I. J. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[31] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high-and low-level consistency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1369–1379, Apr. 2019.

[32] W. C. Hung, Y. H. Tsai, Y. T. Liou, Y.-Y. Lin, and M. H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *Proc. BMVC*, 2018, pp. 1–17.

[33] J. Hou, X. Ding, and J. D. Deng, "Semi-supervised semantic segmentation of vessel images using leaking perturbations," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 2625–2634.

[34] D. Zhai, B. Hu, X. Gong, H. Zou, and J. Luo, "ASS-GAN: Asymmetric semi-supervised GAN for breast ultrasound image segmentation," *Neurocomputing*, vol. 493, pp. 204–216, Jul. 2022.

[35] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6023–6032.

[36] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "ClassMix: Segmentation-based data augmentation for semi-supervised learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2021, pp. 1369–1378.

[37] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Jul. 2019.

[38] Y. Liu, Y. Tian, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro, "Perturbed and strict mean teachers for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4258–4267.

[39] Z. Zhao, L. Yang, S. Long, J. Pi, L. Zhou, and J. Wang, "Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11350–11359.

[40] Z. Zhao, S. Long, J. Pi, J. Wang, and L. Zhou, "Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23705–23714.

[41] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6210–6219.

[42] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020.

[43] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[44] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 15750–15758.

[45] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[46] Y. Zhong, B. Yuan, H. Wu, Z. Yuan, J. Peng, and Y.-X. Wang, "Pixel contrastive-consistent semi-supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7273–7282.

[47] H.-M. Xu, L. Liu, Q. Bian, and Z. Yang, "Semi-supervised semantic segmentation with prototype-based consistency regularization," in *Proc. NeurIPS*, vol. 35, 2022, pp. 26007–26020.

[48] H. Xiao et al., "Semi-supervised semantic segmentation with cross teacher training," *Neurocomputing*, vol. 508, pp. 36–46, Oct. 2022.

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

[50] D. Kwon and S. Kwak, "Semi-supervised semantic segmentation with error localization network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9957–9967.

[51] J. Yuan, Y. Liu, C. Shen, Z. Wang, and H. Li, "A simple baseline for semi-supervised semantic segmentation with strong data augmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8229–8238.

[52] R. Yi, Y. Huang, Q. Guan, M. Pu, and R. Zhang, "Learning from pixel-level label noise: A new perspective for semi-supervised semantic segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 623–635, 2022.

[53] P. Hu, S. Sclaroff, and K. Saenko, "Leveraging geometric structure for label-efficient semi-supervised scene segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 6320–6330, 2022.

[54] D. Guan, J. Huang, A. Xiao, and S. Lu, "Unbiased subclass regularization for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9968–9978.

[55] H. Hu, F. Wei, H. Hu, Q. Ye, J. Cui, and L. Wang, "Semi-supervised semantic segmentation via adaptive equalization learning," in *Proc. NeurIPS*, vol. 34, 2021, pp. 22106–22118.

[56] R. He, J. Yang, and X. Qi, "Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6910–6920.

[57] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 761–769.

[58] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, vol. 119, 2020, pp. 1597–1607.

[59] O. Caelen, "A Bayesian interpretation of the confusion matrix," *Ann. Math. Artif. Intell.*, vol. 81, nos. 3–4, pp. 429–450, 2017.

[60] R. Dupre, J. Fajtl, V. Argyriou, and P. Remagnino, "Improving dataset volumes and model accuracy with semi-supervised iterative self-learning," *IEEE Trans. Image Process.*, vol. 29, pp. 4337–4348, 2020.

[61] W. Chen, Z. Jiang, Z. Wang, K. Cui, and X. Qian, "Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8916–8925.

[62] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with Atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[64] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[65] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[66] S. Fan, F. Zhu, Z. Feng, Y. Lv, M. Song, and F.-Y. Wang, "Conservative-progressive collaborative learning for semi-supervised semantic segmentation," *IEEE Trans. Image Process.*, vol. 32, pp. 6183–6194, 2023, doi: 10.1109/TIP.2023.3242819.

[67] D. Filipiak, P. Tempczyk, and M. Cygan, "N-CPS: Generalising cross pseudo supervision to N networks for semi-supervised semantic segmentation," 2021, *arXiv:2112.07528*.

[68] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7236–7246.

[69] L. Kong, J. Ren, L. Pan, and Z. Liu, "LaserMix for semi-supervised LiDAR semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21705–21715.

[70] Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. Lau, "Guided collaborative training for pixel-wise semi-supervised learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 429–445.

[71] J. Ma, C. Wang, Y. Liu, L. Lin, and G. Li, "Enhanced soft label for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1185–1195.

[72] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[73] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 991–998.

[74] H. Huang et al., "SemiCVT: Semi-supervised convolutional vision transformer for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11340–11349.

[75] P. Li et al., "Semi-supervised semantic segmentation under label noise via diverse learning groups," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, vol. 30, Oct. 2023, pp. 1229–1238.

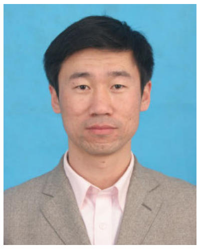
[76] Y. Li, X. Wang, L. Yang, L. Feng, W. Zhang, and Y. Gao, "Diverse cotraining makes strong semi-supervised segmentor," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 16055–16067.

[77] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[78] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 418–434.

Bo Dang received the B.S. degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2022, where he is currently pursuing the Ph.D. degree with the School of Remote Sensing and Information Engineering. He has published several papers in CVPR, AAAI, and ISPRS JPRS. His research interests include remote sensing semantic segmentation and remote sensing foundation model.

Yansheng Li (Senior Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent system from the Huazhong University of Science and Technology, Wuhan, China, in 2015. From 2017 to 2018, he was a Visiting Assistant Professor with the Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. He is currently a Full Professor and the Vice Dean of the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan. He has authored more than 100 peer-reviewed papers, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, CVPR, ECCV, and AAAI. His research interests include knowledge graph, deep learning, and their applications in remote sensing big data mining. He was awarded the Young Surveying and Mapping Science and Technology Innovation Talent Award of the Chinese Society for Geodesy, Photogrammetry and Cartography, in 2022. He received the recognition of the Best Reviewer of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING in 2021. He is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and a Junior Editorial Member of *The Innovation*.



Yongjun Zhang (Member, IEEE) received the B.S. degree in geodesy, the M.S. degree in geodesy and surveying engineering, and the Ph.D. degree in geodesy and photography from Wuhan University, Wuhan, China, in 1997, 2000, and 2002, respectively. He is currently a Full Professor and the Dean of the School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 150 research articles and one book. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multisource datasets, artificial intelligence-driven remote sensing image interpretation, and 3D city reconstruction.

Jiayi Ma (Senior Member, IEEE) received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively. He is currently a Professor with the Electronic Information School, Wuhan University. His research interests include computer vision and machine learning. He is an Area Editor of *Information Fusion* and an Associate Editor of IEEE/CAA JOURNAL OF AUTOMATICA SINICA, *Neurocomputing*, *Image and Vision Computing*, and *Geo-Spatial Information Science*.