Semi-Supervised Semantic Segmentation via Marginal Contextual Information

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present a novel confidence refinement scheme that enhances pseudo-labels in semi-supervised semantic segmentation. Unlike current leading methods, which filter pixels with low-confidence teacher predictions in isolation, our approach leverages the spatial correlation of labels in segmentation maps by grouping neighboring pixels and considering their pseudo-labels collectively. With this contextual information, our method, named S4MC, increases the amount of unlabeled data used during training while maintaining the quality of the pseudo-labels, all with negligible computational overhead. Through extensive experiments on standard benchmarks, we demonstrate that S4MC outperforms existing state-of-the-art semi-supervised learning approaches, offering a promising solution for reducing the cost of acquiring dense annotations. For example, S4MC achieves a substantial 6.34 mIoU improvement over the prior state-of-the-art method on PASCAL VOC 12 with 92 annotated images. The code to reproduce our experiments is available at https://s4mcontext.github.io/.

1 Introduction

Supervised learning has been the driving force behind advancements in modern computer vision, including classification (Krizhevsky et al., 2012; Dai et al., 2021), object detection (Girshick, 2015; Zong et al., 2022), and segmentation (Zagoruyko et al., 2016; Chen et al., 2018a; Li et al., 2022; Kirillov et al., 2023). However, it requires extensive amounts of labeled data, which can be costly and time-consuming to obtain. In many practical scenarios, while there is no shortage of available data, only a fraction can be labeled due to resource constraints. This challenge has led to the development of semi-supervised learning (SSL) (Rasmus et al., 2015; Berthelot et al., 2019b; Sohn et al., 2020a; Yang et al., 2022a), a methodology that leverages both labeled and unlabeled data for model training.

This paper focuses on applying SSL to semantic segmentation, which has applications in various areas such as perception for autonomous vehicles (Bartolomei et al., 2020), mapping (Van Etten et al., 2018) and agriculture (Milioto et al., 2018). SSL is particularly appealing for segmentation tasks, as manual labeling can be prohibitively expensive.

A widely adopted approach for SSL is pseudo-labeling (Lee, 2013; Arazo et al., 2020). This technique dynamically assigns supervision targets to unlabeled data during training based on the model's predictions. To generate a meaningful training signal, it is essential to adapt the predictions before integrating them into the learning process. Several techniques have been proposed for that, such as using a teacher network to generate supervision to a student network (Hinton et al., 2015). The teacher network can be made more powerful during training by applying a moving average to the student network's weights (Tarvainen and Valpola, 2017). Additionally, the teacher may undergo weaker augmentations than the student (Berthelot et al., 2019b), simplifying the teacher's task.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Figure 1: **Confidence refinement. Left:** pseudo-labels generated by the teacher network without refinement. **Middle:** pseudo-labels obtained from the same model after refinement with marginal contextual information. **Right Top:** predicted probabilities of the top two classes of the pixel highlighted by the red square before, and **Bottom:** after refinement. S4MC allows additional correct pseudo labels to propagate.

However, pseudo-labeling is intrinsically susceptible to confirmation bias, which tends to reinforce the model predictions instead of improving the student model. Mitigating confirmation bias becomes particularly important when dealing with erroneous predictions made by the teacher network.

One popular technique to address this issue is confidence-based filtering (Sohn et al., 2020a). This approach assigns pseudo-labels only when the model's confidence surpasses a specified threshold, thereby reducing the number of incorrect pseudo-labels. Though simple, this strategy was proven effective and inspired multiple improvements in semi-supervised classification (Zhang et al., 2021; Rizve et al., 2021), segmentation (Wang et al., 2022), and object detection in images (Sohn et al., 2020b; Liu et al., 2021) and 3D scenes (Zhao et al., 2020; Wang et al., 2021). However, the strict filtering of the supervision signal leads to extended training periods and, potentially, to overfitting when the labeled instances used are insufficient to represent the entire sample distribution. Lowering the threshold would allow for higher training volumes at the cost of reduced quality, further hindering the performance (Sohn et al., 2020a).

In response to these challenges, we introduce a novel confidence refinement scheme for the teacher network predictions in segmentation tasks, designed to increase the availability of pseudo-labels without sacrificing their accuracy. Drawing on the observation that labels in segmentation maps exhibit strong spatial correlation, we propose to group neighboring pixels and collectively consider their pseudo-labels. When considering pixels in spatial groups, we asses the event-union probability, which is the probability that at least one pixel belongs to a given class. We assign a pseudo-label if this probability is sufficiently larger that the event-union probability of any other class. By taking context into account, our approach *Semi-Supervised Semantic Segmentation via Marginal Contextual Information* (S4MC), enables a relaxed filtering criterion which increases the number of unlabeled pixels utilized for learning while maintaining high-quality labeling, as demonstrated in Fig. 1.

We evaluated S4MC on multiple semi-supervised segmentation benchmarks. S4MC achieves significant improvements in performance over previous state-of-the-art methods. In particular, we observed a remarkable increase of **+6.34 mIoU** on PASCAL VOC 12 (Everingham et al., 2010) using only 92 annotated images and an increase of **+1.85 mIoU** on Cityscapes (Cordts et al., 2016) using only 186 annotated images. These findings highlight the effectiveness of S4MC in producing high-quality segmentation results with minimal labeled data.

2 Related Work

2.1 Semi-Supervised Learning

Pseudo-labeling (Lee, 2013) is a popular and effective technique in SSL, where labels are assigned to unlabeled data based on model predictions. To make the most of these labels during training, it is essential to refine them (Laine and Aila, 2016; Berthelot et al., 2019b,a; Xie et al., 2020). One way to achieve this is through consistency regularization (Laine and Aila, 2016; Tarvainen and Valpola, 2017; Miyato et al., 2018), which ensures consistent predictions between different views of the unlabeled

data. Alternatively, a teacher model can be used to obtain pseudo-labels, which are then used to train a student model. To ensure that the pseudo-labels are useful, the temperature of the prediction (soft pseudo-labels; Berthelot et al., 2019b) can be increased, or the label can be assigned to samples with high confidence (hard pseudo-labels; Xie et al., 2020; Sohn et al., 2020a; Zhang et al., 2021). In this work we follow the hard pseudo-label assignment approach and improve upon previous methods by proposing a confidence refinement scheme.

2.2 Semi-Supervised Semantic Segmentation

In semantic segmentation, most SSL methods rely on a combination of consistency regularization and the development of augmentation strategies compatible with segmentation tasks. Given the uneven distribution of labels typically encountered in segmentation maps, techniques such as adaptive sampling, augmentation, and loss re-weighting are commonly employed (Hu et al., 2021). Feature perturbations on unlabeled data (Ouali et al., 2020; Zou et al., 2021; Liu et al., 2022b) are also used to enhance consistency, along with the application of virtual adversarial training (Liu et al., 2022b). Curriculum learning strategies that incrementally increase the proportion of data used over time are beneficial in exploiting more unlabeled data (Yang et al., 2022b; Wang et al., 2022). A recent approach introduced by (Wang et al., 2022) cleverly utilizes *unreliable* predictions by employing contrastive loss with the least confident classes predicted by the model. However, most existing works primarily focus on individual pixel label predictions. In contrast, we delve into the contextual information offered by spatial predictions on unlabeled data.

2.3 Contextual Information

Contextual information encompasses environmental cues that assist in interpreting and extracting meaningful insights from visual perception (Toussaint, 1978; Elliman and Lancaster, 1990). Incorporating spatial context explicitly has been proven beneficial in segmentation tasks, for example, by encouraging smoothness like in the Conditional Random Fields (CRF) method (Chen et al., 2018a) and attention mechanisms (Vaswani et al., 2017; Dosovitskiy et al., 2021; Wang et al., 2020). Combating dependence on context has shown to be useful by Nekrasov et al. (2021). In this work, we leverage the context from neighboring pixel predictions to enhance pseudo-label propagation.

3 Method

3.1 Overview

In semi-supervised semantic segmentation, we are given a labeled training set of images $\mathcal{D}_\ell = \left\{(\mathbf{x}_i^\ell, \mathbf{y}_i)\right\}_{i=1}^{N_\ell}$, and an unlabeled set $\mathcal{D}_u = \{\mathbf{x}_i^u\}_{i=1}^{N_u}$ sampled from the same distribution, i.e., $\left\{\mathbf{x}_i^\ell, \mathbf{x}_i^u\right\} \sim D_x$. Here, $\mathbf{y}$ are 2D tensors of shape $H \times W$, assigning a semantic label to each pixel of $\mathbf{x}$. We aim to train a neural network f_θ to predict the semantic segmentation of unseen images sampled from D_x.

We follow a teacher–student approach (Tarvainen and Valpola, 2017) and train two networks f_{θ_s} and f_{θ_t} that share the same architecture but update their parameters separately. The student network f_{θ_s} is trained using supervision from the labeled samples and pseudo-labels created by the teacher's predictions for unlabeled ones. The teacher model f_{θ_t} is updated as an exponential moving average (EMA) of the student weights. $f_{\theta_s}(\mathbf{x}_i)$ and $f_{\theta_t}(\mathbf{x}_i)$ denote the predictions of the student and teacher models for the $\mathbf{x}_i$ sample, respectively. At each training step, a batch of $\mathcal{B}_\ell$ and $\mathcal{B}_u$ images is sampled from $\mathcal{D}_\ell$ and $\mathcal{D}_u$, respectively. The optimization objective can be written as the following loss:

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

$$\mathcal{L}_s = \frac{1}{M_l} \sum_{\mathbf{x}_i^\ell, \mathbf{y}_i \in \mathcal{B}_l} \ell_{CE}(f_{\theta_s}(\mathbf{x}_i^\ell), \mathbf{y}_i) \tag{2}$$

$$\mathcal{L}_u = \frac{1}{M_u} \sum_{\mathbf{x}_i^u \in \mathcal{B}_u} \ell_{CE}(f_{\theta_s}(\mathbf{x}_i^u), \hat{\mathbf{y}}_i), \tag{3}$$

where $\mathcal{L}_s$ and $\mathcal{L}_u$ are the losses over the labeled and unlabeled data correspondingly, λ is a hyperparameter controlling their relative weight, and $\hat{\mathbf{y}}_i$ is the pseudo-label for the i-th unlabeled

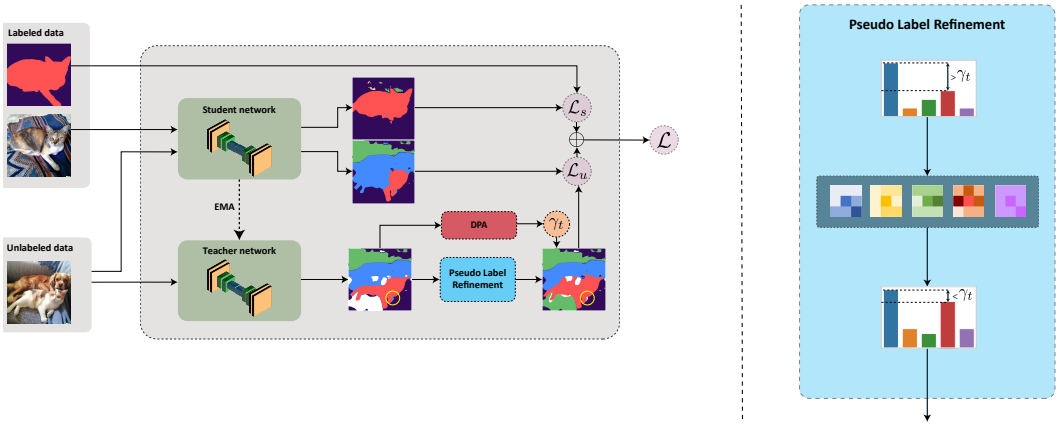


Figure 2: **Left:** S4MC employs a teacher-student paradigm for semi-supervised segmentation. Labeled images are used to supervise the student network directly. Both teacher and student networks process unlabeled images. Predictions from the teacher network are refined and used to evaluate the margin value, which is then thresholded to produce pseudo-labels that guide the student network. The threshold, denoted as γ_t, is dynamically adjusted based on the teacher network's predictions. **Right:** Our confidence refinement module exploits neighboring pixels to adjust per-class predictions, as detailed in Section 3.2.1. The class distribution of the pixel marked by the yellow circle on the left is changed. Before refinement, the margin surpasses the threshold and erroneously assigns the blue class (dog) as a pseudo-label. However, after refinement, the margin significantly reduces, thereby preventing the propagation of this error.

image. Not every pixel of $\mathbf{x}_i$ has a corresponding label or pseudo-label, and M_l and M_u denote the number of pixels with label and assigned pseudo-label in the image batch, respectively.

3.1.1 Pseudo-label Propagation

For a given image $\mathbf{x}_i$, we denote by $\mathbf{x}^i_{j,k}$ the pixel in the j-th row and k-th column. We adopt a thresholding-based criterion inspired by (Sohn et al., 2020a). By establishing a score, denoted as κ, which is based on the class distribution predicted by the teacher network, we assign a pseudo-label to a pixel if its score exceeds a threshold γ_t:

$$\hat{\mathbf{y}}^i_{j,k} = \begin{cases} \arg\max_c \{p_c(x^i_{j,k})\} & \text{if } \kappa(x^i_{j,k}; \theta_t) > \gamma_t, \\ \text{ignore} & \text{otherwise,} \end{cases} \tag{4}$$

where $p_c(x^i_{j,k})$ is the pixel probability of class c. A commonly used score is given by $\kappa(x^i_{j,k}; \theta_t) = \max_c\{p_c(x^i_{j,k})\}$. However, we found that using a pixel-wise margin, inspired by the work of Scheffer et al. (2001) and Shin et al. (2021), produces more stable results. This approach calculates the margin as the difference between the highest and the second-highest values of the probability vector:

$$\kappa_{\text{margin}}(x^i_{j,k}) = \max_c\{p_c(x^i_{j,k})\} - \max_c 2\{p_c(x^i_{j,k})\}. \tag{5}$$

3.1.2 Dynamic Partition Adjustment (DPA)

Following U^{2}PL (Wang et al., 2022), we use a decaying threshold γ_t. DPA replaces the fixed threshold with a quantile-based threshold that decreases with time. At each iteration, we set γ_t as the α_t-th quantile of κ_{margin} over all pixels of all images in the batch. α_t is defined as follows:

$$\alpha_t = \alpha_0 \cdot (1 - {}^t/_{\text{iterations}}). \tag{6}$$

As the model predictions improve with each iteration, gradually lowering the threshold increases the number of propagated pseudo-labels without compromising their quality.

3.2 Marginal Contextual Information

Utilizing contextual information (Section 2.3), we look at surrounding predictions (predictions on neighboring pixels) to refine the semantic map at each pixel. We introduce the concept of "Marginal

Contextual Information," which involves integrating additional information to enhance predictions across all classes. At the same time, reliability-based pseudo-label methods focus on the dominant class only (Sohn et al., 2020a; Wang et al., 2023). Section 3.2.1 describes our confidence refinement, followed by our thresholding strategy and a description of S4MC methodology.

3.2.1 Confidence Margin Refinement

We refine the predicted pseudo-label of each pixel by considering the predictions of its neighboring pixels. Given a pixel $x^i_{j,k}$ with a corresponding per-class prediction $p_c(x^i_{j,k})$, we examine neighboring pixels $x^i_{\ell,m}$ within an $N \times N$ pixel neighborhood surrounding it. We then calculate the probability that at least one of the two pixels belongs to class c:

$$\tilde{p}_c(x^i_{j,k}) = p_c(x^i_{j,k}) + p_c(x^i_{\ell,m}) - p_c(x^i_{j,k}, x^i_{\ell,m}), \tag{7}$$

where $p_c(x^i_{j,k}, x^i_{\ell,m})$ denote the joint probability of both $x^i_{j,k}$ and $x^i_{\ell,m}$ belonging to the same class c.

While the model does not predict joint probabilities, it is reasonable to assume a non-negative correlation between the probabilities of neighboring pixels. This is largely due to the nature of segmentation maps, which are typically piecewise constant. Consequently, any information regarding the model's prediction of neighboring pixels belonging to a specific class should not lead to a reduction in the posterior probability of the given pixel also falling into that class. The joint probability can thus be bounded from below by assuming independence: $p_c(x^i_{j,k}, x^i_{\ell,m}) \geqslant p_c(x^i_{j,k}) \cdot p_c(x^i_{\ell,m})$. By substituting this into Eq. (7), we obtain an upper bound for the event union probability:

$$\tilde{p}_c(x^i_{j,k}) \leq p_c(x^i_{j,k}) + p_c(x^i_{\ell,m}) - p_c(x^i_{j,k}) \cdot p_c(x^i_{\ell,m}). \tag{8}$$

This formulation allows us to filter out confidence margins that do not exceed the threshold.

For each class c, we select the neighbor with the maximal information gain using Eq. (8):

$$\tilde{p}^{\mathbf{N}}_c(x^i_{j,k}) = \max_{\ell,m} \tilde{p}_c(x^i_{j,k}). \tag{9}$$

Computing the event union over all classes employs neighboring predictions to amplify differences in ambiguous cases. Consider, for instance, an uncertain pixel prediction with a 0.5 probability of belonging to one of two classes. If a neighboring pixel has a 0.7 probability of belonging to the first class and only a 0.3 probability of belonging to the second, this results in a significant event union probabilities margin of 0.2. Similarly, this prediction refinement prevents the creation of over-confident predictions that is not supported by additional spatial evidence and helps in reducing confirmation bias. The refinement is visualized in Fig. 1. In our experiments, we used a neighborhood size of 3×3. To determine whether the incorporation of contextual information could be enhanced with larger neighborhoods, we conducted an ablation study focusing on the neighborhood size and the neighbor selection criterion, as detailed in Table 4a. For larger neighborhoods, we decrease the probability contribution of the neighboring pixels with a distance-dependent factor:

$$\tilde{p}_c(x^i_{j,k}) = p_c(x^i_{j,k}) + \beta_{\ell,m}\big[p_c(x^i_{\ell,m}) - p_c(x^i_{j,k}, x^i_{\ell,m})\big], \tag{10}$$

where $\beta_{\ell,m} = \exp\big(-\frac{1}{2}(|\ell - j| + |m - k|)\big)$ is a spatial weighting function. Empirically, contextual information refinement affects mainly the most probable one or two classes. This aligns well with our choice to use the margin confidence (5).

Considering more than two events (more than one neighbor), one can use the formulation for three or four event-union. In practice, we find two event-union using Eq. (10), assign it as $p_c(x^i_{j,k})$, find the next desired event using Eq. (9) with the remaining neighbors, and repeat the process.

3.2.2 Threshold Setting

Setting a high threshold can mitigate confirmation bias from the teacher model's "beliefs" transferring to the student model. However, this comes at the expense of learning from fewer examples, potentially resulting in a less comprehensive model. *Dynamic Partition Adjustment* (DPA; Wang et al., 2022) attempt to address this issue by setting a threshold that decreases over time. We adopt this method in determining the threshold from the teacher predictions pre-refinement $p_c(x^i_{j,k})$, but we filter values based on $\tilde{p}_c(x^i_{j,k})$. Consequently, more pixels pass the threshold that remains unaffected. We set $\alpha_0 = 0.4$, i.e., 60% of raw predictions pass the threshold at $t = 0$, as this value demonstrated superior performance in our experiments. An ablation study for α_0 is provided in Table 4b.

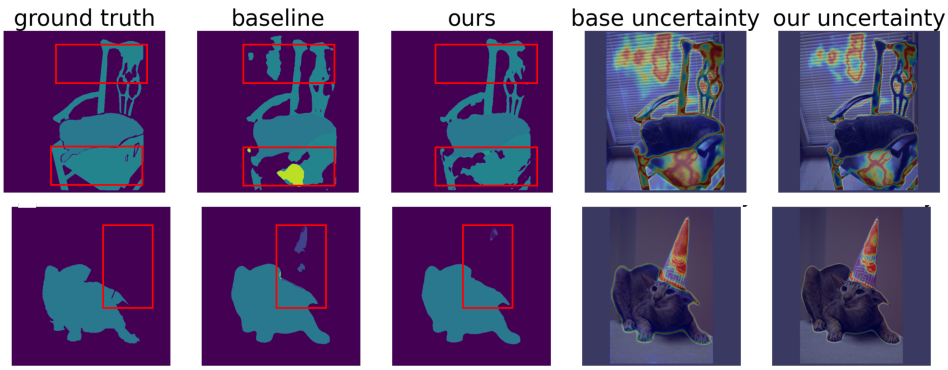


Figure 3: **Qualitative results of S4MC.** The outputs of two trained models and the annotated ground truth. The segmentation map predicted by S4MC (*Ours*) compared to the segmentation map using no refinement module (*Baseline*) and to the ground truth. *Heat map* represents the uncertainty of the model as κ^{-1}, showing a more confident prediction over certain areas, yielding to a smother segmentation maps (compared in the red boxes).

3.3 Putting it All Together

We perform semi-supervised for semantic segmentation by pseudo-labeling pixels using their neighbors' contextual information. Labeled images are only fed into the student model, producing the supervised loss (Eq. (2)). Unlabeled images are fed into the student and teacher models. We sort the margin based κ_{margin} (Eq. (5)) values of teacher predictions and set γ_t as described in Section 3.2.2. The per-class teacher predictions are refined using the *weighted union event* relaxation, as defined in Eq. (10). Pixels with higher margin values than γ_t are assigned with pseudo-labels as described in Eq. (4), producing the unsupervised loss (Eq. (3)). A visualization of the entire pipeline is depicted in Fig. 2.

The impact of S4MC is demonstrated in Fig. 4, which compares the fraction of pixels that pass the threshold with and without refinement. (a) Our method makes greater use of unlabeled data during most of the training process, (b) while the refinement ensures high-quality pseudo-labels. Qualitative results are presented in Fig. 3, where one can see both the confidence heatmap and the pseudo-labels with and without the impact of S4MC.

4 Experiments

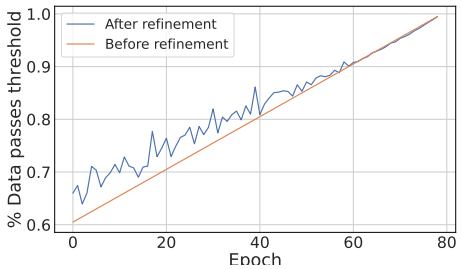
This section presents our experimental results. The setup for the different datasets and partition protocols is detailed in Section 4.1. Section 4.2 compares our method against existing approaches and Section 4.3 provides the ablation study. Further implementation details are given in the Appendix.

4.1 Setup

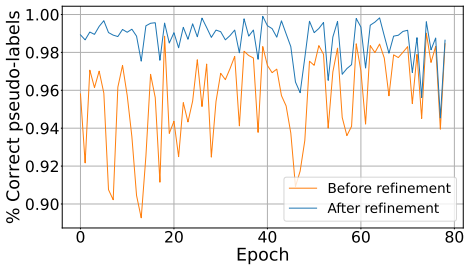
Datasets In our experiments, we use PASCAL VOC 2012 (Everingham et al., 2010) and Cityscapes (Cordts et al., 2016) datasets.

The PASCAL VOC dataset comprises 20 object classes (plus background). The dataset includes 2,913 annotated images, divided into a training set of 1,464 images and a validation set of 1,449 images. In addition, the dataset includes 9,118 coarsely annotated training images (Hariharan et al., 2011), in which only a subset of the pixels are labeled. Following previous research, we conduct two sets of experiments. The 'classic' experiment utilizes only the original training set (Wang et al., 2022; Zou et al., 2021), while the 'coarse' experiment uses all available data (Wang et al., 2022; Chen et al., 2021; Hu et al., 2021).

The Cityscapes (Cordts et al., 2016) dataset includes urban scenes from 50 different cities with 30 classes, of which only 19 are typically used for evaluation (Chen et al., 2018a,b). Similarly to PASCAL, in addition to 2,975 training and 500 validation images, the dataset includes 19,998 coarsely annotated images, which we do not use in our experiment.



(a) **Data fraction that passes the threshold**. The baseline model has a fixed percentage, as it is based on DPA. Our method increases the number of pixels assigned pseudo-label, mostly in the early stage of the training when the model is under-confident.



(b) **Fraction of correct pseudo-labels** the assigned pseudo-labels with the correct class divided by the total assigned pseudo-label. S4MC produces more quality pseudo-labels during the training process, most notably at the early stages.

Figure 4: Pseudo-label quantity and quality on PASCAL VOC 2012 (Everingham et al., 2010) with 366 labeled images using our margin (5) confidence function.

Table 1: Comparison between our method and prior art on the PASCAL VOC 2012 `val` on different partition protocols. the caption describes the share of the training set used as labeled data and, in parentheses, the actual number of labeled images. Larger improvement can be observed for partitions of extremely low annotated data, where other methods suffer from starvation due to poor teacher generalization.

Method	**1/16 (92)**	**1/8 (183)**	**1/4 (366)**	**1/2 (732)**	**Full (1464)**
Supervised Only	45.77	54.92	65.88	71.69	72.50
CutMix-Seg (French et al., 2020)	52.16	63.47	69.46	73.73	76.54
PseudoSeg (Zou et al., 2021)	57.60	65.50	69.14	72.41	73.23
PC^2Seg (Zhong et al., 2021)	57.00	66.28	69.78	73.05	74.15
CPS (Chen et al., 2021)	64.10	67.40	71.70	75.90	-
ReCo (Liu et al., 2022a)	64.80	72.0	73.10	74.70	-
ST++ (Yang et al., 2022b)	65.2	71.0	74.6	77.3	79.1
U^2PL (Wang et al., 2022)	67.98	69.15	73.66	76.16	79.49
PS-MT (Liu et al., 2022b)	65.8	69.6	76.6	78.4	80.0
S4MC + CutMix-Seg (Ours)	70.96	71.69	75.41	77.73	80.58
S4MC + FixMatch (Ours)	**74.32**	**75.62**	**77.84**	**79.72**	**81.51**

Implementation details We implement S4MC on top of two framework variants: CutMix-Seg (French et al., 2020) and FixMatch (Sohn et al., 2020a). Both use DeepLabv3+ (Chen et al., 2018b) with a Imagenet-pre-trained (Russakovsky et al., 2015) ResNet-101 (He et al., 2016). The teacher parameters θ_t are updated via an exponential moving average (EMA) of the student parameters Tarvainen and Valpola (2017): $\theta_t^\eta = \tau\theta_t^{\eta-1} + (1-\tau)\theta_s^\eta$, where $0 \leq \tau \leq 1$ defines how close the teacher is to the student and η denotes the training iteration. We used $\tau = 0.99$. Additional details are provided in Appendix D.

Evaluation We compare S4MC with baselines under the common partition protocols – using $1/2$, $1/4$, $1/8$, and $1/16$ of the training data as labeled data. For the 'classic' setting of the PASCAL experiment, we additionally compare using all the finely annotated images. We follow standard protocols and use mean Intersection over Union (mIoU) as our evaluation metric. We use the data split published by Wang et al. (2022) when available to ensure a fair comparisons. For the ablation studies, we use PASCAL VOC 2012 `val` with $1/4$ partition.

Methods in comparison We compare against popular SSL segmentation methods: CutMix-Seg (French et al., 2020), CCT (Ouali et al., 2020), GCT (Ke et al., 2020), PseudoSeg (Zou et al., 2021), CPS (Chen et al., 2021), PC^2Seg (Zhong et al., 2021), AEL (Hu et al., 2021), U^2PL (Wang et al.,

Table 2: Comparison between our method and prior art on the 'coarse' PASCAL VOC 2012 `val` dataset under different partition protocols, using additional unlabeled data from (Hariharan et al., 2011). For each partition ratio we included the number of labeled images in parentheses. As in 1, larger improvements are observed for partitions with less annotated data.

Method	**1/16 (662)**	**1/8 (1323)**	**1/4 (2646)**	**1/2 (5291)**
Supervised Only	67.87	71.55	75.80	77.13
CutMix-Seg (French et al., 2020)	71.66	75.51	77.33	78.21
CCT (Ouali et al., 2020)	71.86	73.68	76.51	77.40
GCT (Ke et al., 2020)	70.90	73.29	76.66	77.98
CPS (Chen et al., 2021)	74.48	76.44	77.68	78.64
AEL (Hu et al., 2021)	77.20	77.57	78.06	80.29
PS-MT (Liu et al., 2022b)	75.5	78.2	78.7	-
U^2PL (Wang et al., 2022)	77.21	79.01	79.3	80.50
S4MC + CutMix-Seg (Ours)	78.49	79.67	79.85	81.11
S4MC + FixMatch (Ours)	**80.77**	**81.9**	**82.3**	**83.3**

Table 3: Comparison between our method and prior art on the Cityscapes `val` dataset under different partition protocols. Labeled and unlabeled images are selected from the Cityscapes `training` dataset.For each partition protocol, the caption gives the share of the training set used as labeled data, in parentheses, the number of labeled images.

Method	**1/16 (186)**	**1/8 (372)**	**1/4 (744)**	**1/2 (1488)**
Supervised Only	62.96	69.81	74.08	77.46
CutMix-Seg (French et al., 2020)	69.03	72.06	74.20	78.15
CCT (Ouali et al., 2020)	69.32	74.12	75.99	78.10
GCT (Ke et al., 2020)	66.75	72.66	76.11	78.34
CPS (Chen et al., 2021)	69.78	74.31	74.58	76.81
AEL (Hu et al., 2021)	74.45	75.55	77.48	79.01
U^2PL (Wang et al., 2022)	70.30	74.37	76.47	79.05
PS-MT (Liu et al., 2022b)	-	76.89	77.6	79.09
S4MC + CutMix-Seg (Ours)	75.03	77.02	78.78	78.86
S4MC + FixMatch (Ours)	**76.3**	**78.25**	**78.95**	**79.13**

2022), PS-MT (Liu et al., 2022b), and ST++ (Yang et al., 2022b). "Supervised Only" stands for supervised training without using any unlabeled data. As a baseline, we use CutMix-Seg (French et al., 2020).

4.2 Results

PASCAL VOC 2012. Table 1 compares our method with state-of-the-art baselines on the PASCAL VOC 2012 dataset. While Table 2 shows the comparison results on the PASCAL VOC 2012 dataset with additional coarsely annotated data from SBD (Hariharan et al., 2011). In both setups, S4MC outperform all the compared methods in standard partition protocols, both when using labels only for the original PASCAL VOC 12 dataset and when using SBD annotations as well. Qualitative results are shown in Fig. 3. As can be seen our refinement procedure aids in both adding falsely filtered pseduo-labels as well as removing erroneous ones.

Cityscapes. Table 3 Presents the comparison results on the Cityscapes `val` dataset. Table 3 compares our method with other state-of-the-art methods on the Cityscapes (Cordts et al., 2016) dataset under various partition protocols. S4MC outperforms the compared methods in most partitions, except for the $1/2$ setting, and combined with Fixmatch scheme, S4MC outperforms compared approaches across all partitions.

Table 4: The effect of neighborhood size and neighbor selection criterion.

(a) Neighborhood choice.

Selection criterion	Neighborhood size N		
	3×3	5×5	7×7
Random neighbor	73.25	71.1	70.41
Max neighbor	**75.41**	75.18	74.89
Min neighbor	74.54	74.11	70.28
Two max neighbors	74.14	75.15	74.36

(b) α_0 in Eq. (6), which controls the initial proportion of confidence pixels

20%	30%	40%	50%	60%
74.45	73.85	**75.41**	74.56	74.31

Contextual information at inference. Given that our margin refinement scheme operates through prediction adjustments, we explored whether it could be employed at inference time to further enhance performance. The results reveal a negligible improvement in the DeepLab-V3-plus model, from an 85.7 mIOU to 85.71. This underlines that the performance advantage of S4MC primarily derives from the adjusted margin, as the most confident class is rarely swapped. A heatmap of the prediction over several samples is presented in Fig. 3 and Appendix E.

4.3 Ablation Study

We ablate different components of our method using the CutMix-Seg framework variant, and evaluated using the Pascal VOC 12 dataset with a partition protocol of $1/4$ labeled images.

Neighborhood size and neighbor selection criterion. Our prediction refinement scheme employs event-union probability with neighboring pixels, which depends on the chosen neighbor to pair with the current pixel. To assess this, we tested varying neighborhood sizes ($N = 3, 5, 7$) and criteria for selecting the neighboring pixel: (a) random, (b) maximal class probability, (c) minimal class probability, and (d) two neighbors, as described in Section 3.2.1. As shown in Table 4a, a small 3×3 neighborhood with one neighboring pixel of the highest class probability proved most efficient in our experiments.

Threshold parameter tuning As outlined in Section 3.1.2, we utilize a dynamic threshold that depends on an initial value, α_0. In Table 4b, we examine the effect of different initial quantiles to establish this threshold. A smaller α_0 would propagate too many errors, leading to significant confirmation bias. In contrast, a larger α_0 would mask most of the data, resulting in insufficient label propagation, rendering the semi-supervised learning process lengthy and inefficient. We found that an α_0 of 40% yields the best performance.

5 Conclusion

In this paper, we introduce S4MC, a novel approach for incorporating spatial contextual information in semi-supervised segmentation. This strategy refines confidence levels and enables us to leverage a larger portion of unlabeled data. S4MC outperforms existing approaches and achieves state-of-the-art results on multiple popular benchmarks under various data partition protocols, such as Cityscapes and Pascal VOC 12. While we believe S4MC offers a good solution to lowering the annotation requirement, it has several limitations. First, the event-union relaxation is relevant in problems where spatial coherency is expected. The generalization of our framework to other dense prediction tasks would necessitate an assessment of whether this relaxation is applicable. Furthermore, our method employs a fixed shape neighborhood without considering the structure of objects. It would be intriguing to investigate the use of segmented regions to define new neighborhoods, and this is a direction we plan to explore in the future.

References

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International Joint Conference on Neural Networks*, pages 1–8, 2020. doi: 10.1109/IJCNN48605.2020.9207304. URL https://arxiv.org/abs/1908.02983. (cited on p. 1)

Luca Bartolomei, Lucas Teixeira, and Margarita Chli. Perception-aware path planning for UAVs using semantic segmentation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5808–5815, 2020. doi: 10.1109/IROS45743.2020.9341347. (cited on p. 1)

David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin A. Raffel. ReMixMatch: semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint*, November 2019a. URL https://arxiv.org/abs/1911.09785. (cited on p. 2)

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A. Raffel. MixMatch: a holistic approach to semi-supervised learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL https://proceedings.neurips.cc/paper/2019/hash/1cd138d0499a68f4bb72bee04bbec2d7-Abstract.html. (cited on pp. 1, 2, and 3)

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018a. doi: 10.1109/TPAMI.2017.2699184. URL https://arxiv.org/abs/1412.7062. (cited on pp. 1, 3, and 6)

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, September 2018b. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Liang-Chieh_Chen_Encoder-Decoder_with_Atrous_ECCV_2018_paper.html. (cited on pp. 6 and 7)

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2613–2622, June 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Semi-Supervised_Semantic_Segmentation_With_Cross_Pseudo_Supervision_CVPR_2021_paper.html. (cited on pp. 6, 7, and 8)

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Cordts_The_Cityscapes_Dataset_CVPR_2016_paper.html. (cited on pp. 2, 6, and 8)

Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. CoAtNet: marrying convolution and attention for all data sizes. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3965–3977. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc//paper/2021/hash/20568692db622456cc42a2e853ca21f8-Abstract.html. (cited on p. 1)

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. (cited on p. 3)

Dave G. Elliman and Ian T. Lancaster. A review of segmentation and contextual analysis techniques for text recognition. *Pattern Recognition*, 23(3):337–346, 1990. ISSN 0031-3203. doi: https://doi.org/10.1016/0031-3203(90)90021-C. URL https://www.sciencedirect.com/science/article/pii/003132039090021C. (cited on p. 3)

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. doi: 10.1007/s11263-009-0275-4. URL https://doi.org/10.1007/s11263-009-0275-4. (cited on pp. 2, 6, 7, 15, and 16)

Geoffrey French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham D. Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *British Machine Vision Conference*. BMVA Press, 2020. URL https://www.bmvc2020-conference.com/assets/papers/0680.pdf. (cited on pp. 7 and 8)

Ross Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, December 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Girshick_Fast_R-CNN_ICCV_2015_paper.html. (cited on p. 1)

Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *International Conference on Computer Vision*, pages 991–998, 2011. doi: 10.1109/ICCV.2011.6126343. (cited on pp. 6 and 8)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html. (cited on p. 7)

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. (cited on p. 1)

Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22106–22118. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/b98249b38337c5088bbc660d8f872d6a-Paper.pdf. (cited on pp. 3, 6, 7, and 8)

Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson W. H. Lau. Guided collaborative training for pixel-wise semi-supervised learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision*, pages 429–445, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58601-0. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/1932_ECCV_2020_paper.php. (cited on pp. 7 and 8)

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. (cited on p. 1)

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html. (cited on p. 1)

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2016. URL https://openreview.net/forum?id=BJ6oOfqge. (cited on p. 2)

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop: Challenges in Representation Learning (WREPL)*, July 2013. URL http://deeplearning.net/wp-content/uploads/2013/03/pseudo_label_final.pdf. (cited on pp. 1 and 2)

Feng Li, Hao Zhang, Huaizhe xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask DINO: towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint*, June 2022. URL https://arxiv.org/abs/2206.02777. (cited on p. 1)

Shikun Liu, Shuaifeng Zhi, Edward Johns, and Andrew J. Davison. Bootstrapping semantic segmentation with regional contrast. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=6u6N8WWwYSM. (cited on p. 7)

Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=MJIve1zgR_. (cited on p. 2)

Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4258–4267, June 2022b. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_Perturbed_and_Strict_Mean_Teachers_for_Semi-Supervised_Semantic_Segmentation_CVPR_2022_paper.html. (cited on pp. 3, 7, and 8)

Andres Milioto, Philipp Lottes, and Cyrill Stachniss. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2229–2235, 2018. doi: 10.1109/ICRA.2018.8460962. URL https://arxiv.org/abs/1709.06764. (cited on p. 1)

Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018. doi: 10.1109/TPAMI.2018.2858821. URL https://ieeexplore.ieee.org/abstract/document/8417973. (cited on p. 2)

Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. Mix3d: Out-of-context data augmentation for 3d scenes. *3DV 2021*, 2021. (cited on p. 3)

Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Ouali_Semi-Supervised_Semantic_Segmentation_With_Cross-Consistency_Training_CVPR_2020_paper.html. (cited on pp. 3, 7, and 8)

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://papers.nips.cc/paper/2015/hash/378a063b8fdb1db941e34f4bde584c7d-Abstract.html. (cited on p. 1)

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S. Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=-ODN6SbiUU. (cited on p. 2)

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. (cited on p. 7)

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden Markov models for information extraction. In Frank Hoffmann, David J. Hand, Niall Adams, Douglas Fisher, and Gabriela Guimaraes, editors, *Advances in Intelligent Data Analysis*, pages 309–318, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-44816-7. URL https://link.springer.com/chapter/10.1007/3-540-44816-0_31. (cited on pp. 4 and 15)

Gyungin Shin, Weidi Xie, and Samuel Albanie. All you need are a few pixels: Semantic segmentation with pixelpick. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1687–1697, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021W/ILDAV/html/Shin_All_You_Need_Are_a_Few_Pixels_Semantic_Segmentation_With_ICCVW_2021_paper.html. (cited on pp. 4 and 15)

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A. Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: simplifying semi-supervised learning with consistency and confidence. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 596–608. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html. (cited on pp. 1, 2, 3, 4, 5, and 7)

Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020b. (cited on p. 2)

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html. (cited on pp. 1, 2, 3, and 7)

Godfried T. Toussaint. The use of context in pattern recognition. *Pattern Recognition*, 10(3):189–204, 1978. ISSN 0031-3203. doi: https://doi.org/10.1016/0031-3203(78)90027-4. URL https://www.sciencedirect.com/science/article/pii/0031320378900274. The Proceedings of the IEEE Computer Society Conference. (cited on p. 3)

Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. SpaceNet: a remote sensing dataset and challenge series. *arXiv preprint*, June 2018. URL https://arxiv.org/abs/1807.01232. (cited on p. 1)

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. (cited on p. 3)

He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J. Guibas. 3DIoUMatch: leveraging IoU prediction for semi-supervised 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14615–14624, June 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Wang_3DIoUMatch_Leveraging_IoU_Prediction_for_Semi-Supervised_3D_Object_Detection_CVPR_2021_paper.html. (cited on p. 2)

Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020. (cited on p. 3)

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. FreeMatch: self-adaptive thresholding for semi-supervised learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PDrUPTXJI_A. (cited on p. 5)

Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo labels. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Semi-Supervised_Semantic_Segmentation_Using_Unreliable_Pseudo-Labels_CVPR_2022_paper.html. (cited on pp. 2, 3, 4, 5, 6, 7, and 8)

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/44feb0096faa8326192570788b38c1d1-Abstract.html. (cited on pp. 2 and 3)

Fan Yang, Kai Wu, Shuyi Zhang, Guannan Jiang, Yong Liu, Feng Zheng, Wei Zhang, Chengjie Wang, and Long Zeng. Class-aware contrastive semi-supervised learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14421–14430, June 2022a. URL https://openaccess.thecvf.com/content/CVPR2022/html/Yang_Class-Aware_Contrastive_Semi-Supervised_Learning_CVPR_2022_paper.html. (cited on p. 1)

Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. ST++: make self-training work better for semi-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4268–4277, June 2022b. URL https://openaccess.thecvf.com/content/CVPR2022/html/Yang_ST_Make_Self-Training_Work_Better_for_Semi-Supervised_Semantic_Segmentation_CVPR_2022_paper.html. (cited on pp. 3, 7, and 8)

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Yun_CutMix_Regularization_Strategy_to_Train_Strong_Classifiers_With_Localizable_Features_ICCV_2019_paper.html. (cited on p. 16)

Sergey Zagoruyko, Adam Lerer, Tsung-Yi Lin, Pedro O. Pinheiro, Sam Gross, Soumith Chintala, and Piotr Dollár. A MultiPath network for object detection. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 15.1–15.12. BMVA Press, September 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.15. URL https://dx.doi.org/10.5244/C.30.15. (cited on p. 1)

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18408–18419. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/995693c15f439e3d189b06e89d145dd5-Abstract.html. (cited on pp. 2 and 3)

Na Zhao, Tat-Seng Chua, and Gim Hee Lee. SESS: self-ensembling semi-supervised 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Zhao_SESS_Self-Ensembling_Semi-Supervised_3D_Object_Detection_CVPR_2020_paper.html. (cited on p. 2)

Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7273–7282, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Zhong_Pixel_Contrastive-Consistent_Semi-Supervised_Semantic_Segmentation_ICCV_2021_paper.html. (cited on p. 7)

Zhuofan Zong, Guanglu Song, and Yu Liu. DETRs with collaborative hybrid assignments training. *arXiv preprint*, November 2022. URL https://arxiv.org/abs/2211.12860. (cited on p. 1)

Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. PseudoSeg: designing pseudo labels for semantic segmentation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=-TwO99rbVRu. (cited on pp. 3, 6, and 7)

A Pseudo-labels quality analysis

The quality improvement and the quantity increase of pseudo-labels are shown in figure Fig. 4. Further analysis of the quality improvement of our method can be shown in Fig. 5 by separating the true positive increase and the false positive decrease.

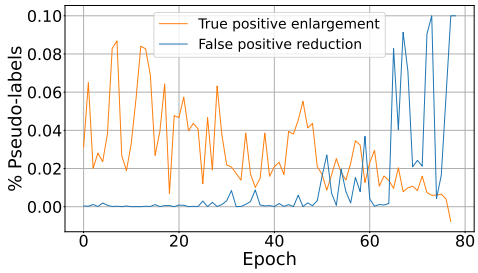


Figure 5: **Quality of pseudo-labels**, on PASCAL VOC 2012 (Everingham et al., 2010) over time. At the early stage of the learning process, the increase in pseudo-labels' quality is mostly due to true positive improvement. In other words, the refinement not only helps more pixels pass the threshold, but most of these are of good quality. Later in the learning process, most of the improvement comes from a reduction in false-positive. This means that our method reduces the number of pseudo-labeled images assigned with the wrong label when the threshold is low.

B Confidence function

In this paper we introduce a confidence function to determine pseudo-labels propagation. We introduced $\kappa_{\text{margin}}(x_{i,j})$ and mentioned other alternatives have been examined.

Here we define several options for the confidence function.

The simplest option is to look at the probability of the dominant class,

$$\kappa_{\max}(x^i_{j,k}) = \max_c p_c(x^i_{j,k}), \tag{11}$$

which is commonly used to generate pseudo-labels.

The second alternative is negative entropy, defined as

$$\kappa_{\text{ent}}(x^i_{j,k}) = \sum_{c \in C} p_c(x^i_{j,k}) \log\left(p^c_{i,j}\right). \tag{12}$$

Since high entropy corresponds to high uncertainty, low entropy corresponds to high confidence, as required from the certainty.

The third option is for us to define the margin function (Scheffer et al., 2001; Shin et al., 2021) as the difference between the first and second maximal values of the probability vector and also described in the main paper:

$$\kappa_{\text{margin}}(x_{i,j}) = \max_c(p_c(x^i_{j,k})) - \max 2_c(p_c(x^i_{j,k})), \tag{13}$$

where $\max 2$ denotes the vector's second maximum value. Figure Fig. 6 shows the change of distribution of the margin function over time. All alternatives are compared in Table 5.

Table 5: Ablation study on the confidence function κ, over Pascal VOC 12 with partition protocols

Function	1/4 (366)	1/2 (732)	Full (1464)
$\kappa_{\max}$	74.29	76.16	79.49
κ_{ent}	75.18	77.55	79.89
κ_{margin}	75.41	77.73	80.58

Table 5 studies the impact of different confidence functions on pseudo-label refinement. We found that using a margin to describe confidence is a suitable way when there is a contradiction in smooth regions.

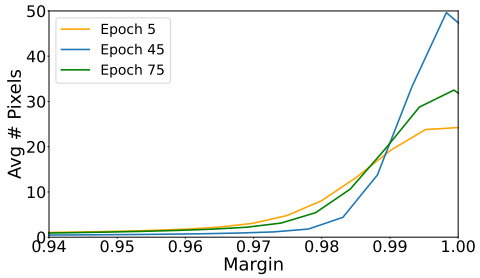


Figure 6: **Distribution of confidence estimation**, an average over a batch of 256 unlabeled images from PASCAL VOC 2012 (Everingham et al., 2010) over time, using margin confidence estimation. At the early stage of the learning process, the model is less confident over the unlabeled images. at the intermediate stage, the model yields more confident predictions. Late in the learning process, the method tries to compensate for conformation bias; since the basic method propagates almost all pixels, our model tries to reduce confidence and supervision on bad examples and reduce the confirmation bias.

C Bounding the joint probability

In this paper, we had the union event estimation with the independent assumption, defined as

$$p_c^1(x_{j,k}^i, x_{\ell,m}^i) \approx p_c(x_{j,k}^i) \cdot p_c(x_{\ell,m}^i) \tag{14}$$

Besides independent even, another estimation is the unconditional empirical expectation of two neighboring pixels belonging to the same class, i.e.,

$$p_c^2(x_{j,k}^i, x_{\ell,m}^i) = \frac{1}{|\mathcal{N}_l| \cdot H \cdot W \cdot |\mathbf{N}|} \sum_{y,i,j} \sum_{k,\ell \in \mathbf{N}_{i,j}} \mathbb{1}\{y_{i,j} = y_{k,\ell}\} \tag{15}$$

Since we want to improve the estimation and avoid overestimating the union event that could lead to overconfidence, we set

$$p_c(x_{j,k}^i, x_{\ell,m}^i) = \max(p_c^1(x_{j,k}^i, x_{\ell,m}^i), p_c^2(x_{j,k}^i, x_{\ell,m}^i)) \tag{16}$$

That upper bound of joint probability ensures that the independence assumption does not underestimate the joint probability, which in turn, prevents overestimation of the union event probability. Using Eq. (16) increase the mIOU by **0.22** on average, for using 366 annotated images from PASCAL VOC 12. Using only $p_c^2(x_{j,k}^i, x_{\ell,m}^i)$ as a joint probability assumption led to diverged model, reducing the mIOU by **-14.11**.

D Implementation Details

All experiments are conducted for 80 training epochs with the simple stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and learning rate policy of $lr = lr_{\text{base}} \cdot \left(1 - \frac{\text{iter}}{\text{total iter}}\right)^{\text{power}}$. With the probability of 0.5, we apply CutMix (Yun et al., 2019) augmentation on the unlabeled data.

For PASCAL VOC 2012 $lr_{\text{base}} = 0.001$ and the decoder only $lr_{\text{base}} = 0.01$, the weight decay is set to 0.0001 and all images are cropped to 513×513 and $\mathcal{B}_l = \mathcal{B}_u = 3$.

For Cityscapes, all parameters use $\text{lr}_{\text{base}} = 0.01$, and the weight decay is set to 0.0005. The learning rate decay parameter is set to $\text{power} = 0.9$. Due to memory constraints, all images are cropped to 769×769 and $\mathcal{B}_\ell = \mathcal{B}_u = 2$. All experiments are conducted on a machine with 8 Nvidia RTX A5000 GPUs.

E More visual results

An extension of Fig. 3, shows more instances from the unlabeled data and their pseudo-labeled with the baseline model and S4MC.

Figure 7: **Example of refined pseudo-labels**, the structure is as in Fig. 3

, the numbers under the predictions show the pixel-wise accuracy of the prediction map. Our method obtains more continuous predictions, with higher certainty for such pixels, thus tending to apply lower certainty for small separated object predictions which allows the model to refrain from learning such mistakes and obtains higher confidence near the borders of predicted objects, allowing the model to learn from these pixels.