Semi-Supervised Semantic Segmentation via Marginal Contextual Information

Anonymous authors
Paper under double-blind review

Abstract

We present a novel confidence refinement scheme that enhances pseudo-labels in semi-supervised semantic segmentation. Unlike existing methods, which filter pixels with low-confidence 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 1.39 mIoU improvement over the prior art on PASCAL VOC 12 with 366 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, there is no shortage of available data, but 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., 2019; 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, 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 & Valpola, 2017). Additionally, the teacher may undergo weaker augmentations than the student (Berthelot et al., 2019), simplifying the teacher's task.

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.

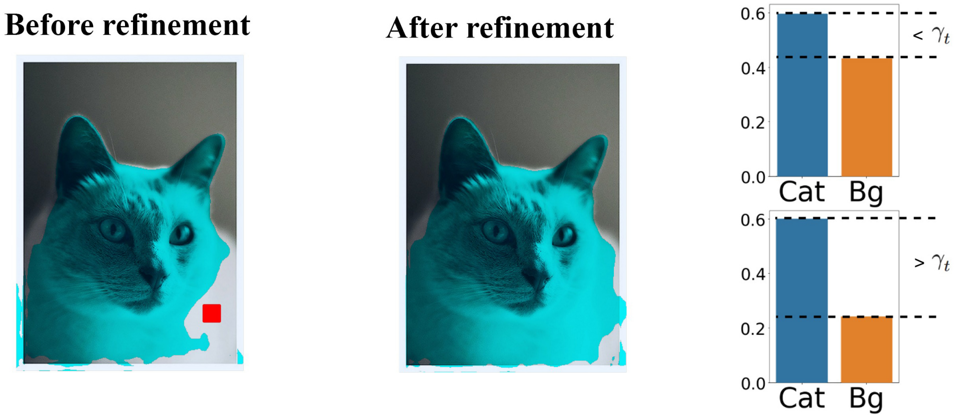


Figure 1: **Confidence refinement. Left:** pseudo-labels generated without refinement. **Middle:** pseudo-labels obtained from the same model after refinement with marginal contextual information. **Right Top:** predicted probabilities of the top two classes of the pixel highlighted by the red square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

Confidence-based filtering is a popular technique to address this issue (Sohn et al., 2020a). This approach assigns pseudo-labels only when the model's confidence surpasses a specified threshold, 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 (Sohn et al., 2020b; Liu et al., 2021; 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 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 than 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 benchmarks. S4MC achieves significant improvements in performance over previous state-of-the-art methods. In particular, we observed an increase of **+1.39 mIoU** on PASCAL VOC 12 (Everingham et al., 2010) using 366 annotated images, **+1.01 mIoU** on Cityscapes (Cordts et al., 2016) using only 186 annotated images, and increase **+1.5 mIoU** on COCO (Lin et al., 2014) using 463 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 an 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 & Aila, 2016; Berthelot et al., 2019; 2020; Xie et al., 2020). This can be done through consistency regularization (Laine & Aila, 2016; Tarvainen & Valpola, 2017; Miyato et al., 2018), which ensures consistent predictions between different views or different models' prediction of the unlabeled data. To ensure that the pseudo-labels are helpful, the temperature of the prediction (soft pseudo-labels; Berthelot et al., 2019) 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).

2.2 Semi-Supervised Semantic Segmentation

In semantic segmentation, most SSL methods rely on consistency regularization and developing augmentation strategies compatible with segmentation tasks (French et al., 2020; Ke et al., 2020; Chen et al., 2021; Zhong et al., 2021; Xu et al., 2022). 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 (FP) on unlabeled data (Ouali et al., 2020; Zou et al., 2021; Liu et al., 2022b; Yang et al., 2023) are also used to enhance consistency and the 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) included unreliable pseudo-labels into training by employing contrastive loss with the least confident classes predicted by the model. Unimatch (Yang et al., 2023) combined SSL (Sohn et al., 2020a) with several self-supervision signals, i.e., two strong augmentations and one more with FP, obtained good results without complex losses or class-level heuristics. 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 & Lancaster, 1990). Incorporating spatial context explicitly has been proven beneficial in segmentation tasks, for example, by encouraging smoothness like in the Conditional Random Fields 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 helpful by Nekrasov et al. (2021). This work uses the context from neighboring pixel predictions to enhance pseudo-label propagation.

3 Method

This section describes the proposed method using the teacher–student paradigm with teacher averaging (Tarvainen & Valpola, 2017). Adjustments for image-level consistency are described in Appendix F.

3.1 Overview

In semi-supervised semantic segmentation, we are given a labeled training set $\mathcal{D}_\ell = \left\{(\mathbf{x}_i^\ell, \mathbf{y}_i)\right\}_{i=1}^{N_\ell}$, 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-averaging approach 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

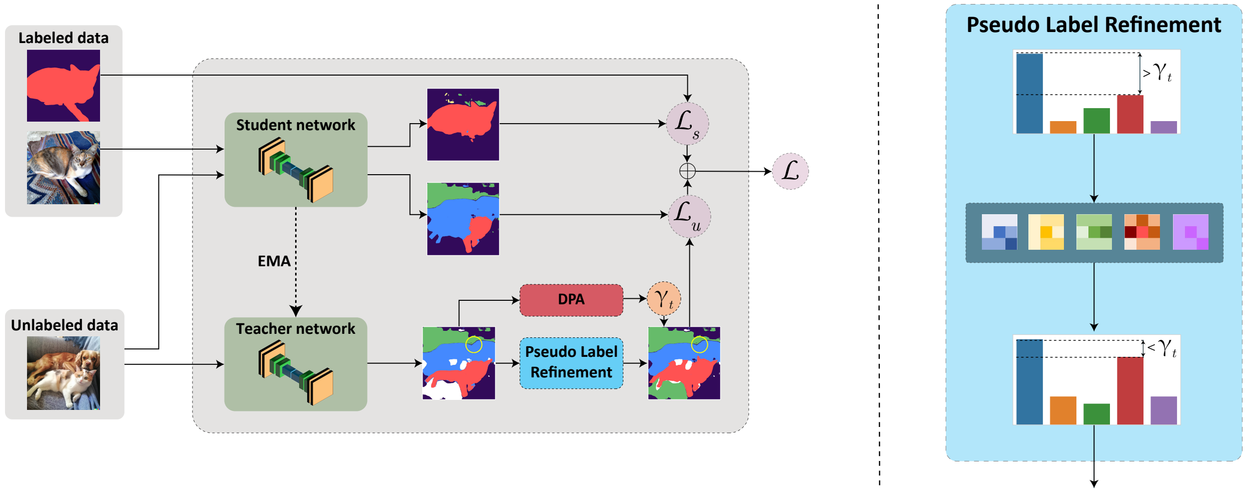


Figure 2: **Left:** S4MC employs a teacher–student paradigm for semi-supervised segmentation. Labeled images are used to supervise the student network directly; both networks process unlabeled images. Teacher predictions are refined and used to evaluate the margin value, which is then thresholded to 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. After refinement, the margin reduces, thereby preventing error propagation.

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 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 $\mathbf{x}_{j,k}^i$ the pixel in the j-th row and k-th column. We adopt a thresholding-based criterion inspired by FixMatch (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:

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

where $p_c(x_{j,k}^i)$ is the pixel probability of class c. A commonly used score is given by $\kappa(x_{j,k}^i; \theta_t) = \max_c\{p_c(x_{j,k}^i)\}$. However, using a pixel-wise margin (Scheffer et al., 2001; Shin et al., 2021), produces more stable results. Denoting by max2 the second-highest value, the margin is given by the difference between

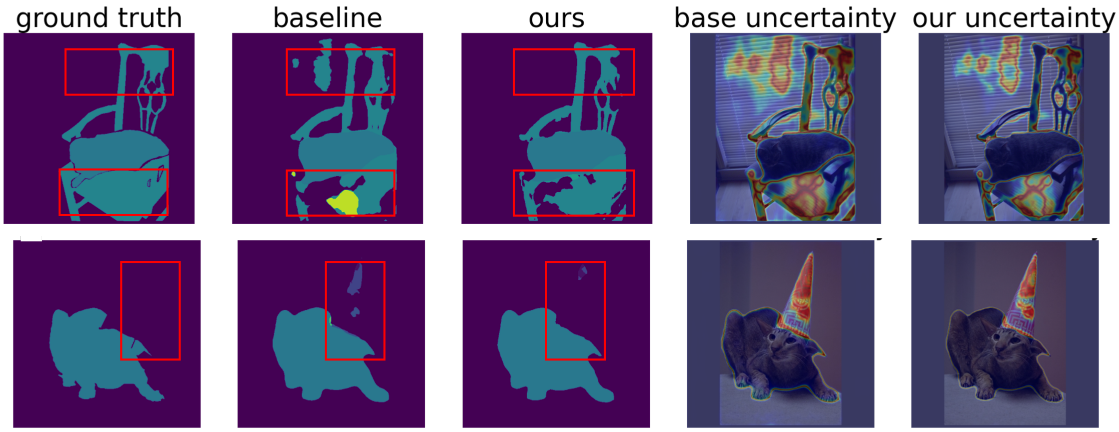


Figure 3: **Qualitative results.** The segmentation map predicted by S4MC (*ours*) is compared to using no refinement module (*baseline*) and to the ground truth. *Heat map* represents the uncertainty of the model (κ^{-1}), showing more confident predictions in certain areas and smoother segmentation maps (marked by the red boxes). Additional examples are shown in Appendix A.

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 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 linearly decreases from α_0 to zero during the training. As the model predictions improve with each iteration, gradually lowering the threshold increases the number of propagated pseudo-labels without compromising 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 each pixel's predicted pseudo-label by considering its neighboring pixels' predictions. 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:

$$\tilde{p}_c(x^i_{j,k}) = p_c(x^i_{j,k}) + p_c(x^i_{\ell,m}) - p_c(x^i_{j,k}, x^i_{\ell,m}), \tag{6}$$

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, assuming a non-negative correlation between the probabilities of neighboring pixels is reasonable. This is mainly due to the nature of segmentation maps, which are typically piecewise constant.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. (6), we obtain an upper bound for the event union probability:

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

For each class c, we select the neighbor with the maximal information utilization using Eq. (7):

$$\tilde{p}^{\mathbf{N}}_c(x^i_{j,k}) = \max_{\ell,m} \tilde{p}_c(x^i_{j,k}). \tag{8}$$

Computing the event union over all classes employs neighboring predictions to amplify differences in ambiguous cases. Similarly, this prediction refinement prevents the creation of over-confident predictions not supported by additional spatial evidence and helps reduce 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 6. For larger neighborhoods, we decrease the probability contribution of the neighboring pixels with a distance-dependent factor:

$$\tilde{p}_c(x^i_{j,k}) = p_c(x^i_{j,k}) + \beta_{\ell,m}\big[p_c(x^i_{\ell,m}) - p_c(x^i_{j,k}, x^i_{\ell,m})\big], \tag{9}$$

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).

We explored alternatives at the pixel level, such as the maximum class probability ($\kappa_{\max}$) and entropy (κ_{ent}). Table F.1 in the Appendix studies the impact of different confidence functions on pseudo-label refinement.

Considering multiple neighbors, we can use the formulation for three or more events. In practice, we calculate it iteratively, starting with two event-union defined by Eq. (9), using it as $p_c(x^i_{j,k})$, finding the next desired event using Eq. (8) with the remaining neighbors, and repeating the process.

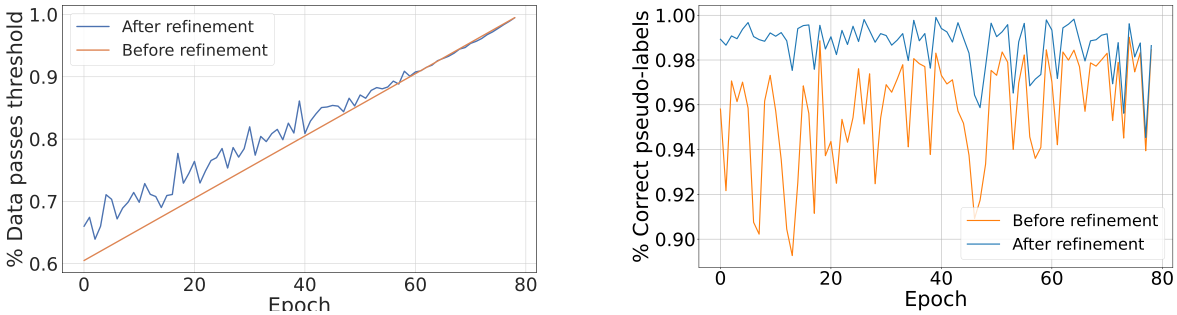
3.2.2 Threshold Setting

A high threshold can prevent transferring the teacher model's wrong "beliefs" to the student model. However, this comes at the expense of learning from fewer examples, resulting in a less comprehensive model. To determine the DPA threshold, we use 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 (unchanged) threshold. We tuned α_0 value in Table 7 and set $\alpha_0 = 0.4$, i.e., 60% of raw predictions pass the threshold at $t = 0$.

3.3 Putting it All Together

We perform semi-supervised learning 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 (2). Unlabeled images are fed into the student and teacher models. We sort the values of κ_{margin} (5) 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. (9). Pixels with top class matching original label and margin values higher than γ_t are assigned pseudo-labels as described in Eq. (4), for the unsupervised loss (3). The entire pipeline is visualized in Fig. 2.

The impact of S4MC is shown in Fig. 4, comparing the fraction of pixels that pass the threshold with and without refinement. S4MC uses more unlabeled data during most of the training (a), while the refinement ensures high-quality pseudo-labels (b). We further study true positive (TP) and false positive (FP) rates, as shown in Fig. E.1 in the Appendix. We show qualitative results in Fig. 3, including both the confidence heatmap and the pseudo-labels with and without the impact of S4MC.



(a) **Data fraction that passes the threshold**. Our method increases the number of pseudo-labeled pixels, mostly in the early stage of the training.

(b) **Accuracy of the pseudo-labels**. 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. The training was performed using S4MC; metrics with and without S4MC were calculated.

Table 1: Comparison between our method and prior art on the PASCAL VOC 2012 `val` (total of 1,464 training images) under different partition protocols. The caption describes the share of the training set used as labeled data and the actual number of labeled images. * denotes reproduced results using official implementation. ± denotes the standard deviation over three runs.

Method	**1/16 (92)**	**1/8 (183)**	**1/4 (366)**	**1/2 (732)**	**Full (1464)**
CutMix-Seg (French et al., 2020)	52.16	63.47	69.46	73.73	76.54
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^{2}PL (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
PCR (Xu et al., 2022)	70.06	74.71	77.16	78.49	80.65
FixMatch* (Yang et al., 2023)	68.07	73.72	76.38	77.97	79.97
UniMatch* (Yang et al., 2023)	73.75	75.05	77.7	79.9	80.43
CutMix-Seg + S4MC	70.96	71.69	75.41	77.73	80.58
FixMatch + S4MC	73.13	74.72	77.27	79.07	79.6
UniMatch$^{\psi}$ + S4MC	**74.72** ± **0.283**	**75.21** ± **0.244**	**79.09** ± **0.183**	**80.12** ± **0.120**	**81.56** ± **0.103**

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. Implementation details are given in Appendix C.

4.1 Setup

Datasets In our experiments, we use PASCAL VOC 2012 (Everingham et al., 2010), Cityscapes (Cordts et al., 2016), and MS COCO (Lin et al., 2014) datasets.

PASCAL comprises 20 object classes (+ background). 2,913 annotated images are divided into training and validation sets of 1,464 and 1,449 images, respectively. Zoph et al. (2020) shown that joint training of PASCAL with training images with augmented annotations (Hariharan et al., 2011) outperforms joint training with COCO (Lin et al., 2014) or ImageNet (Russakovsky et al., 2015). Based on this finding, we use extended PASCAL VOC 2012 (Hariharan et al., 2011), which includes 9,118 augmented training images,

Table 2: Comparison between our method and prior art on the PASCAL VOC 2012 `val` (total of 1,464 training images) under different partition protocols using ResNet-50 as backbone model. The caption describes the share of the training set used as labeled data.

Method	**1/16**	**1/8**	**1/4**	**1/2**	**Full**
Supervised Baseline	44.0	52.3	61.7	66.7	72.9
PseudoSeg (Zou et al., 2021)	54.89	61.88	64.85	70.42	71.00
PC^{2}Seg (Zhong et al., 2021)	56.9	64.6	67.6	70.9	72.3
UniMatch (Yang et al., 2023)	71.9	72.5	76.0	77.4	78.7
UniMatch$^{\psi}$ + S4MC	**72.62**	**72.83**	**76.44**	**77.83**	**79.41**

Table 3: Comparison between our method and prior art on the augmented PASCAL VOC 2012 `val` dataset under different partitions, using additional unlabeled data from Hariharan et al. (2011) (total of 10,582 training images). We included the number of labeled images in parentheses for each partition ratio. * denotes reproduced results using official implementation.

Method	**1/16**	**1/8**	**1/4**	**1/2**
CutMix-Seg (French et al., 2020)	71.66	75.51	77.33	78.21
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^{2}PL (Wang et al., 2022)	77.21	79.01	79.3	80.50
PCR (Xu et al., 2022)	78.6	**80.71**	**80.78**	80.91
FixMatch* (Yang et al., 2023)	74.35	76.33	76.87	77.46
UniMatch* (Yang et al., 2023)	76.6	77.0	77.32	77.9
CutMix-Seg + S4MC	**78.84**	79.67	79.85	**81.11**
FixMatch + S4MC	75.19	76.56	77.11	78.07
UniMatch$^{\psi}$ + S4MC	76.95	77.54	77.62	78.08

Table 4: Comparison between our method and prior art on the Cityscapes `val` dataset (total of 2,976 training images) 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 and the number of labeled images. * denotes reproduced results using official implementation.

Method	**1/16**	**1/8**	**1/4**	**1/2**
CutMix-Seg (French et al., 2020)	69.03	72.06	74.20	78.15
AEL (Hu et al., 2021)	74.45	75.55	77.48	79.01
U^{2}PL (Wang et al., 2022)	70.30	74.37	76.47	79.05
PS-MT (Liu et al., 2022b)	-	76.89	77.6	79.09
PCR (Xu et al., 2022)	73.41	76.31	78.4	79.11
FixMatch* (Yang et al., 2023)	74.17	76.2	77.14	78.43
UniMatch* (Yang et al., 2023)	75.99	77.55	78.54	79.22
CutMix-Seg + S4MC	75.03	77.02	78.78	78.86
FixMatch + S4MC	75.2	77.61	79.04	79.74
UniMatch$^{\psi}$ + S4MC	**77.0**	**77.78**	**79.52**	**79.76**

wherein only a subset of pixels are labeled. Following prior art, we conducted two sets of experiments: in the first, we used only the original set, while in the second, "augmented" setup, we also used augmented data.

Cityscapes dataset includes urban scenes from 50 cities with 30 classes, of which only 19 are typically used for evaluation (Chen et al., 2018a;b).

MS COCO dataset is a challenging segmentation benchmark with 80 object classes (+ background). 123k images are split into 118k and 5k for training and validation.

Implementation details We implement S4MC with teacher–student paradigm of consistency regularization, both with teacher averaging (Tarvainen & Valpola, 2017; French et al., 2020) and augmentation variation (Sohn et al., 2020a; Yang et al., 2023) frameworks. All variations use DeepLabv3+ (Chen et al., 2018b), while for feature extraction, we use ResNet-101 (He et al., 2016) for PASCAL VOC and Cityscapes, and Xception-65 (Chollet, 2016) for MS COCO. For the teacher averaging setup, the teacher parameters θ_t are updated via an exponential moving average (EMA) of the student parameters: $\theta_t^{\eta} = \tau\theta_t^{\eta-1} + (1-\tau)\theta_s^{\eta}$, where $0 \leq \tau \leq 1$ defines how close the teacher is to the student and η denotes the training iteration. We used $\tau = 0.99$. In the augmentation variation approach, pseudo-labels are generated through weak augmentations, and optimization is performed using strong augmentations. Additional details are provided in Appendix C.

Evaluation We compare S4MC with state-of-the-art methods and baselines under the standard partition protocols – using 1/2, 1/4, 1/8, and 1/16 of the training set as labeled data. For the "classic" setting of the PASCAL experiment, we additionally use 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 comparison. For the ablation studies, we use PASCAL VOC 2012 `val` with 1/4 partition.

4.2 Results

PASCAL VOC 2012. Table 1 compares our method with state-of-the-art baselines on the PASCAL VOC 2012 dataset. While Tables 2 and 3 shows the comparison results on PASCAL with additional unlabeled data from SBD (Hariharan et al., 2011) using ResNet-50 and ResNet-101, respectively. S4MC outperforms all compared methods in standard partition protocols using the PASCAL VOC 12 dataset and some partitions when using SBD annotations. More significant improvement can be observed for partitions of extremely low annotated data, where other methods suffer from starvation due to poor teacher generalization. Qualitative results are shown in Fig. 3. Our refinement procedure aids in adding falsely filtered pseudo-labels and removing erroneous ones.

Cityscapes. Table 4 presents the comparison with state-of-the-art methods on the Cityscapes `val` (Cordts et al., 2016) dataset under various partition protocols. S4MC outperforms the compared methods in most partitions, and combined with the FixMatch scheme, S4MC outperforms compared approaches across all partitions.

MS COCO. Table 5 presents the comparison with state-of-the-art methods on the MS COCO `val` (Lin et al., 2014) dataset. S4MC outperforms the compared state-of-the-art methods in most regimes, using the data splits published in (Yang et al., 2023). In this experimental setting, the model sees a small fraction of the data, which could be hard to generalize over all classes. Yet, the mutual information using neighboring predictions seems to compensate somewhat as more supervision signals propagate from the unlabeled data.

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 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 Fig. A.1.

4.3 Ablation Study

Neighborhood size and neighbor selection criterion. Our prediction refinement scheme employs event-union probability with neighboring pixels. We examine varying neighborhood sizes ($N = 3, 5, 7$), number of neighbors ($k = 1, 2$), and selection criteria for neighbors. We compare the following methods for

Table 5: Comparison between our method and prior art on COCO (Lin et al., 2014) `val` (total of 118,336 training images) on different partition protocols. For each partition protocol, the caption gives the share of the training set used as labeled data and the number of labeled images. * denotes reproduced results using official implementation.

Method	**1/512**	**1/256**	**1/128**	**1/64**	**1/32**
Supervised Baseline	22.9	28.0	33.6	37.8	42.2
PseudoSeg (Zou et al., 2021)	29.8	37.1	39.1	41.8	43.6
PC2Seg (Zhong et al., 2021)	29.9	37.5	40.1	43.7	46.1
UniMatch* (Yang et al., 2023)	31.9	38.9	**43.86**	47.8	49.8
UniMatch$^{\psi}$ + S4MC	**32.9**	**40.4**	43.78	**47.98**	**50.58**

Table 6: The effect of neighborhood size and neighbor selection criterion on the Pascal VOC 12 with 1/4 labeled data. We denote the number of neighbors as k.

Selection criterion	**Neighborhood size N**			
	1×1	**3×3**	**5×5**	**7×7**
Random neighbor (k=1)	77.7	77.03	76.85	76.87
Cosine similarity (k=1)		78.02	78.05	77.99
Max-prob (k=1)		**79.09**	78.23	77.76
Max-prob (k=2)		77.77	77.82	78.03
Min-prob (k=1)		75.62	75.11	73.95

Table 7: The effect of α_0, the initial proportion of confidence pixels for the Pascal VOC 12 with 1/4 labeled data.

α_0	20%	30%	40%	50%	60%
mIoU	78.13	77.53	**79.1**	78.24	77.99

choosing the neighboring predictions: (a) Random selection, (b) Cosine similarity, (c) Max probability, and (d) Minimal probability from a complete neighborhood. Note that for the cosine similarity, we choose a single most similar prediction vector for all classes, thus mostly enhancing confidence and not overturning predictions.We also compare with $N = 1$ neighborhood, corresponding to not using S4MC. As seen from Table 6, $N = 3$ neighborhood with one neighboring pixel of the highest class probability proved most efficient in our experiments. Aggregating the probability of a randomly selected neighbor has negligible influence on model performance. Using multiple neighbors demonstrates improved performance as the neighborhood expands. Notably, the inclusion of minimal class probability pixels adversely affects model performance, primarily attributed to the contribution of neighbors that exhibit high certainty in belonging to a distinct class.

We also examine the contribution of the proposed pseudo-label refinement (PLR) and DPA. Results in Table 8 show that the PLR improves the mask mIoU by 1.09%, while DPA alone harms the performance. This indicates that PLR helps semi-supervised learning mainly because it enforces more spatial dependence on the pseudo-labels.

Threshold parameter tuning We utilize a dynamic threshold that depends on an initial value, α_0. In Table 7, we examine the effect of different initial values to establish this threshold. A smaller α_0 propagates too many errors, leading to significant confirmation bias. In contrast, a larger α_0 would mask most of the data, rendering the semi-supervised learning process lengthy and inefficient.

Table 8: Ablation study on the different components of S4MC on top of FixMatch for the augmented Pascal VOC 12 with 1/2 labeled data. **PLR** is the pseudo-label refinement module and **DPA** is dynamic partition adjustment.

PLR	**DPA**	**mIoU**
		79.51
✓		78.25
	✓	79.94
✓	✓	**80.13**

Table 9: Evaluation of Boundary IoU (Cheng et al., 2021) comparing models trained with FixMatch+S4MC and with FixMatch using 183 annotated images on COCO. The model is based on Xception-65 as in Table 5.

FixMatch	**FixMatch+S4MC**
31.1	29.9

Mask boundaries Table 9 demonstrates the limitation of our method in terms of boundary IoU (Cheng et al., 2021). Contrary to the improvement S4MC provides to IoU, the boundary IoU is reduced. That aligns with the qualitative results, as our model predictions masks are smoother in regions far from the boundaries and less confident around the boundaries.

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 more unlabeled data. S4MC outperforms existing approaches and achieves state-of-the-art results on multiple popular benchmarks under various data partition protocols, such as MS COCO, Cityscapes, and Pascal VOC 12. Despite its effectiveness in lowering the annotation requirement, there are several limitations to using S4MC. First, its reliance on event-union relaxation is applicable only in cases involving spatial coherency. As a result, using our framework for other dense prediction tasks would require an examination of this relaxation's applicability. Furthermore, our method uses a fixed-shape neighborhood without considering the object's structure. It would be interesting to investigate the use of segmented regions to define new neighborhoods; this is a future direction we plan to explore.

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*, pp. 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)*, pp. 5808–5815, 2020. doi: 10.1109/IROS45743.2020.9341347. (cited on p. 1)

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 (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/1cd138d0499a68f4bb72bee04bbec2d7-Abstract.html`. (cited on pp. 1, 2, and 3)

David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin A. Raffel. ReMixMatch: semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HklkeR4KPB`. (cited on p. 2)

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 8)

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. 8 and 9)

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)*, pp. 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 p. 3)

Bowen Cheng, Ross Girshick, Piotr Dollar, Alexander C. Berg, and Alexander Kirillov. Boundary IoU: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15334–15342, June 2021. URL `https://openaccess.thecvf.com/content/CVPR2021/html/Cheng_Boundary_IoU_Improving_Object-Centric_Image_Segmentation_Evaluation_CVPR_2021_paper.html`. (cited on p. 11)

François Chollet. Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807, 2016. (cited on p. 9)

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, 7, and 9)

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 (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 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. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy. (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, 7, and 22)

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. 3, 7, 8, and 9)

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*, pp. 991–998, 2011. doi: 10.1109/ICCV.2011.6126343. (cited on pp. 7, 8, and 9)

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. 9)

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 (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 22106–22118. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/b98249b38337c5088bbc660d8f872d6a-Paper.pdf. (cited on pp. 3 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 (eds.), *European Conference on Computer Vision*, pp. 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 p. 3)

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 (eds.), *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)

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *European Conference on Computer Vision*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. URL https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48. (cited on pp. 2, 7, 9, and 10)

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)*, pp. 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)*, pp. 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 p. 3)

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 (eds.), *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 (eds.), *Advances in Intelligent Data Analysis*, pp. 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 22)

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*, pp. 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 22)

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 (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 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, 9, 19, 21, and 22)

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. URL https://arxiv.org/abs/2005.04757. (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 (eds.), *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 9)

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, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_ECA-Net_Efficient_Channel_Attention_for_Deep_Convolutional_Neural_Networks_CVPR_2020_paper.html. (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)*, pp. 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 (CVPR)*, June 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_ECA-Net_Efficient_Channel_Attention_for_Deep_Convolutional_Neural_Networks_CVPR_2020_paper.html. (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 pp. 5 and 21)

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, 5, 7, 8, and 9)

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 (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6256–6268. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/44feb0096faa8326192570788b38c1d1-Abstract.html. (cited on pp. 2 and 3)

Haiming Xu, Lingqiao Liu, Qiuchen Bian, and Zhen Yang. Semi-supervised semantic segmentation with prototype-based consistency regularization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26007–26020. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/a70ee7ea485e4fd36abbfc4adf591c28-Abstract-Conference.html. (cited on pp. 3, 7, and 8)

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)*, pp. 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)*, pp. 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 and 7)

Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7236–7246, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yang_Revisiting_Weak-to-Strong_Consistency_in_Semi-Supervised_Semantic_Segmentation_CVPR_2023_paper.html. (cited on pp. 3, 7, 8, 9, 10, 19, 22, and 23)

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. 18)

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 (eds.), *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 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 (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18408–18419. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/995693c15f439e3d189b06e89d145dd5-Abstract.html. (cited on pp. 2, 3, and 21)

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)*, pp. 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 pp. 3, 8, and 10)

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)

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3833–3845. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/27e9661e033a73a6ad8cefcde965c54d-Paper.pdf. (cited on p. 7)

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, 8, and 10)

Algorithm 1: Pseudocode: Pseudo label refinement of S4MC, PyTorch-like style.

```
# X: predict prob of unlabeled data B x C x H x W
# k: number of neigbors

#create neighborhood tensor
neigborhood=[]
X = X.unsqueeze(1)
X = torch.nn.functional.pad(X, (1, 1, 1, 1, 0, 0, 0, 0))
for i,j in [(None,-2),(1,-1),(2,None)]:
    for k,l in [(None,-2),(1,-1),(2,None)]:
        if i==k and i==1:
            continue
        neighborhood.append(X[:,:,i:j, k:l])
neighborhood = torch.stack(neighborhood)

#pick k neighbors for union event
ktop_neighbors,neigbor_idx=torch.topk(neighborhood, k=k,axis=0)
for nbr in ktop_neighbors:
    beta = torch.exp((-1/2) * neigbor_idx)
    X = X + beta*nbr - (X*nbr*beta)
```

A Visual results

We present in Figs. A.1 and A.2 an extension of Fig. 3, showing more instances from the unlabeled data and the corresponding pseudo-labeled with the baseline model and S4MC.

Our method can achieve more accurate predictions during the inference phase without refinements. This results in more seamless and continuous predictions, which accurately depict objects' spatial configuration.

B Computational cost

Let us denote the image size by $H \times W$ and the number of classes by C.

First, the predicted map of dimension $H \times W \times C$ is stacked with the padded-shifted versions, creating a tensor of shape [n,H,W,C]. K top neighbors are picked via top-k operation and calculate the union event as presented in Eq. (9). (The pseudo label refinement pytorch-like pseudo-code can be obtained in Algorithm 1 for $N = 4$ and k max neighbors.)

The overall space (memory) complexity of the calculation is $O(n \times H \times W \times C)$, which is negligible considering all parameters and gradients of the model. Time complexity adds three tensor operations (stack, topk, and multiplication) over the $H \times W \times C$ tensor, where the multiplication operates k times, which means $O(k \times H \times W \times C)$. This is again negligible for any reasonable number of classes compared to tens of convolutional layers with hundreds of channels.

To verify that, we conducted a training time analysis comparing FixMatch and FixMatch + S4MC over PASCAL with 366 labeled examples, using distributed training with 8 Nvidia RTX 3090 GPUs. FixMatch average epoch takes 28:15 minutes, and FixMatch + S4MC average epoch takes 28:18 minutes, an increase of about 0.2% in runtime.

C Implementation Details

All experiments were conducted for 80 training epochs with the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and learning rate policy of $lr = lr_{\text{base}} \cdot \left(1 - \frac{\text{iter}}{\text{total iter}}\right)^{\text{power}}$.

For the teacher averaging consistency, we apply resize, crop, horizontal flip, GaussianBlur, and with a probability of 0.5, we use Cutmix (Yun et al., 2019) on the unlabeled data.

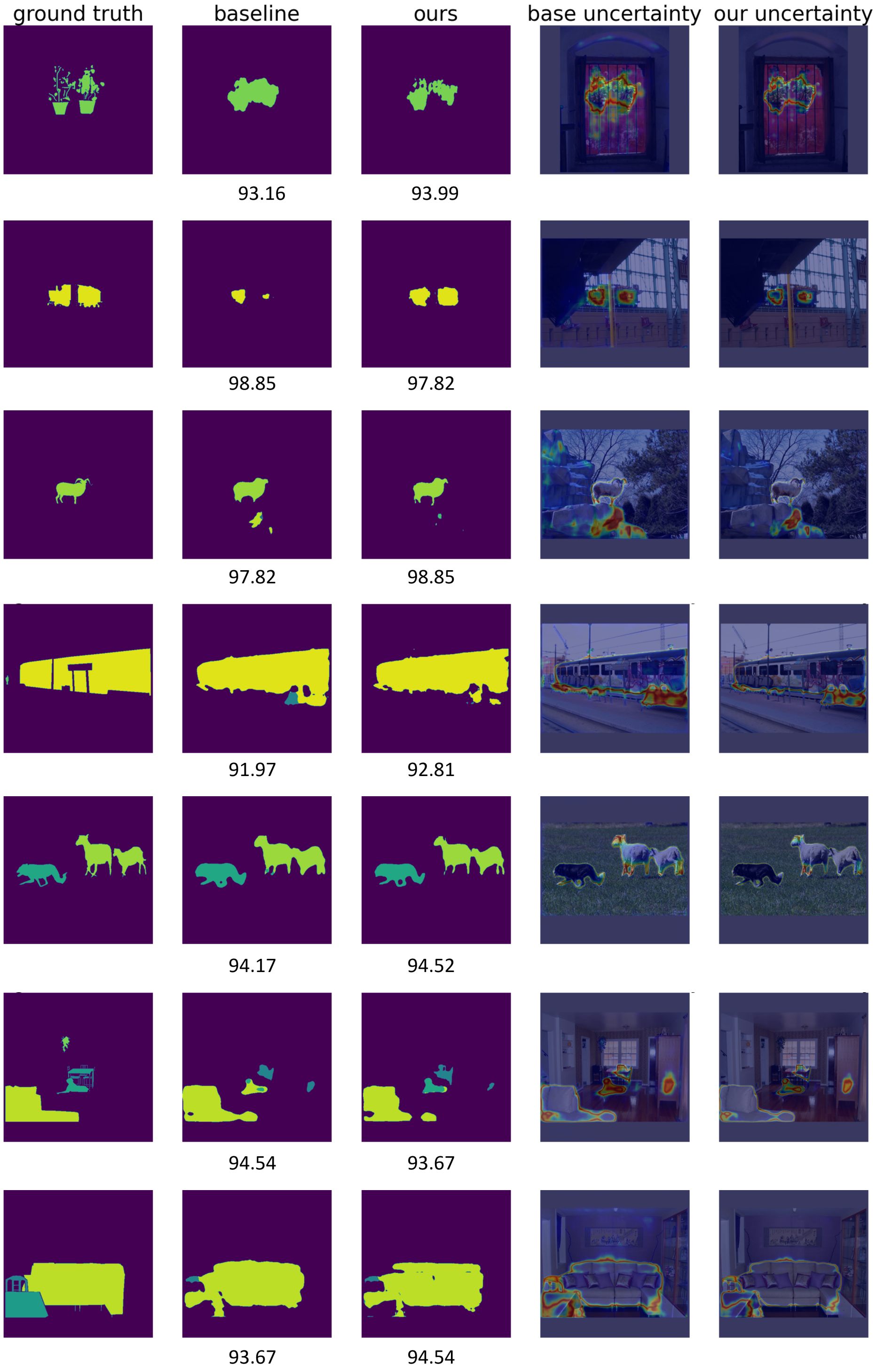


Figure A.1: **Example of refined pseudo-labels**, the structure is as in Fig. 3, and the numbers under the predictions show the pixel-wise accuracy of the prediction map.

For the augmentation variation consistency (Sohn et al., 2020a; Yang et al., 2023), we apply resize, crop, and horizontal flip for weak and strong augmentations as well as ColorJitter, RandomGrayscale, and Cutmix for strong augmentations.

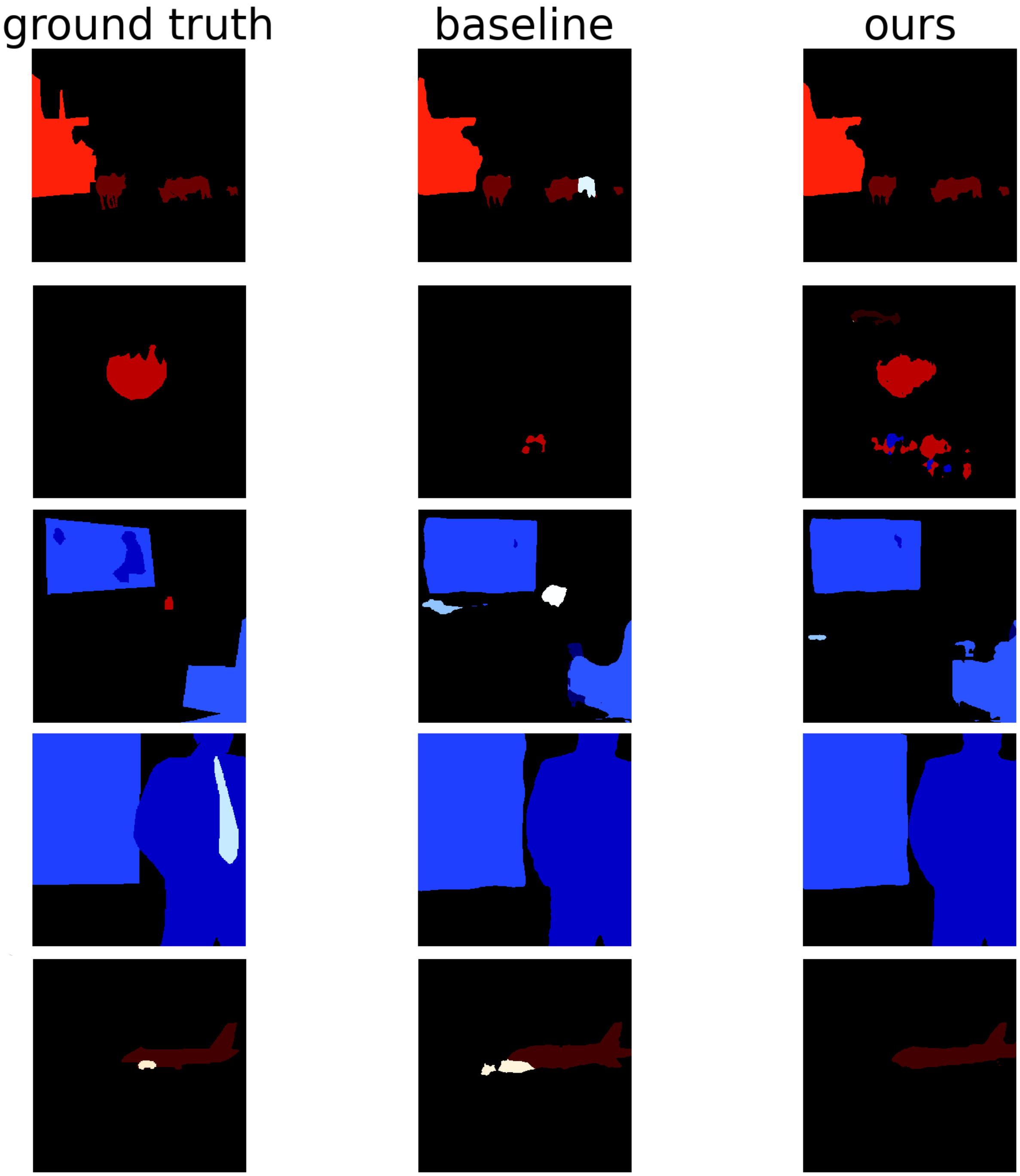


Figure A.2: Qualative results of our method in comparison to UniMatch baseline over COCO with 1/32 of the labeled examples

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 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.

D Limitations and Potential Negative Social Impacts

Limitations. The constraint imposed by the spatial coherence assumption also restricts the applicability of this work to dense prediction tasks. Improving pseudo-labels' quality for overarching tasks such as classification might necessitate reliance on data distribution and the exploitation of inter-sample relationships. We are currently exploring this avenue of research.

Societal impact. Similar to most semi-supervised models, we utilize a small subset of annotated data, which can potentially introduce biases from the data into the model. Further, our PLR module assumes spatial coherence. While that holds for natural images, it may yield adverse effects in other domains, such as medical imaging. It is important to consider these potential impacts before choosing to use our proposed method.

E Pseudo-labels quality analysis

The quality improvement and the quantity increase of pseudo-labels are shown in Fig. 4. Further analysis of the quality improvement of our method is demonstrated in Fig. E.1 by separating the *true positive* and *false positive*.

Within the initial phase of the learning process, the enhancement in the quality of pseudo-labels can be primarily attributed to the advancement in true positive labels. In our method, the refinement not only facilitates the inclusion of a larger number of pixels surpassing the threshold but also ensures that a significant majority of these pixels are high quality.

As the learning process progresses, most improvements are obtained from a decrease in false positives pseudo-labels. This analysis shows that our method effectively minimizes the occurrence of incorrect pseudo-labeled, particularly when the threshold is set to a lower value. In other words, our approach reduces confirmation bias from decaying the threshold as the learning process progresses.

F weak–strong consistency

We need to redefine the supervision branch to adjust the method to augmentation level consistency framework (Sohn et al., 2020a; Zhang et al., 2021; Wang et al., 2023). Recall that within the teacher averaging framework, we denote $f_{\theta_s}(\mathbf{x}_i)$ and $f_{\theta_t}(\mathbf{x}_i)$ as the predictions made by the student and teacher models for input $\mathbf{x}_i$, where the teacher serves as the source for generating confidence-based pseudo-labels. In the context of image-level consistency, both branches differ by augmented versions $\mathbf{x}_i^w$, $\mathbf{x}_i^s$ and share identical weights f_θ. Here, $\mathbf{x}_i^w$ and $\mathbf{x}_i^s$ represent the weak and strong augmented renditions of the input $\mathbf{x}_i$, respectively. Following the framework above, the branch associated with weak augmentation generates the pseudo-labels.

F.1 Confidence function alternatives

In this paper, we introduce a confidence function to determine pseudo-label 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,

$$\kappa_{\max}(x^i_{j,k}) = \max_c p_c(x^i_{j,k}), \tag{F.1}$$

which is commonly used to generate pseudo-labels.

The second alternative is negative entropy, defined as

$$\kappa_{\text{ent}}(x^i_{j,k}) = \sum_{c \in C} p_c(x^i_{j,k}) \log\bigl(p^c_{i,j}\bigr). \tag{F.2}$$

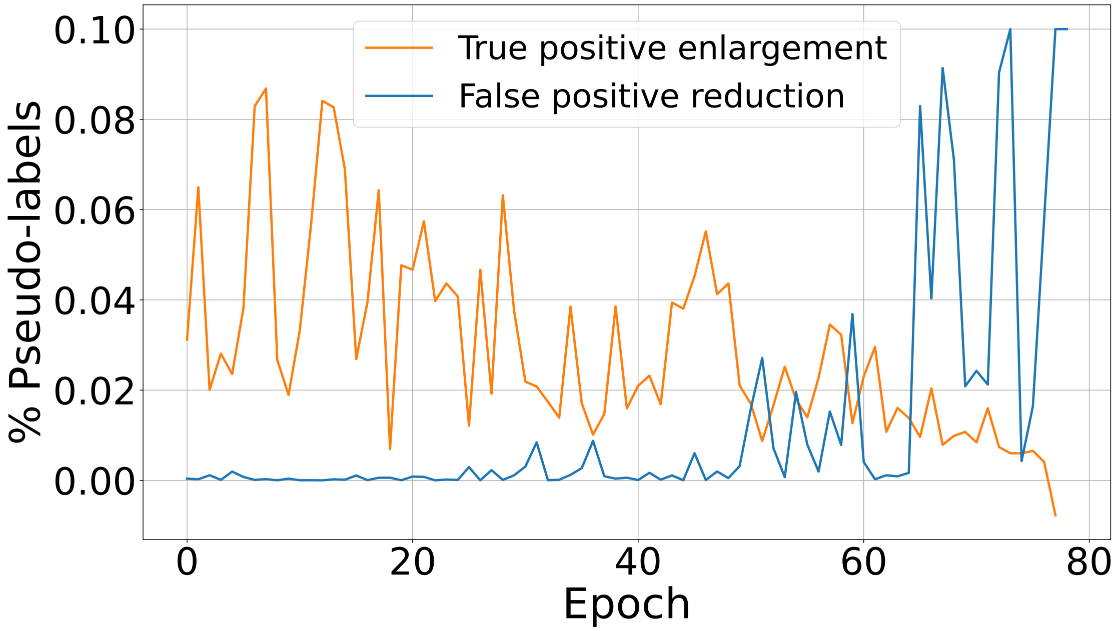


Figure E.1: **Quality of pseudo-labels**, on PASCAL VOC 2012 (Everingham et al., 2010) over training iterations. Fig. 4 separated to *True positive* and *False positive* analysis. *True positive* are the bigger part of improvement at the early stage of the training process, while reduction of *false positive* is the main contribution late in the training process

Table F.1: Ablation study on the confidence function κ, over Pascal VOC 12 with partition protocols

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

Note that this is indeed a confidence function since high entropy corresponds to high uncertainty, and low entropy corresponds to high confidence.

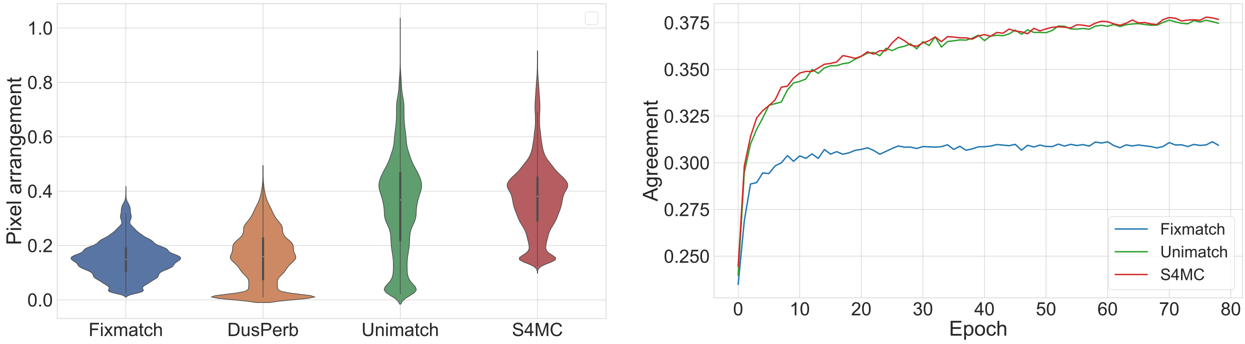
The third option is for us to define the margin function (Scheffer et al., 2001; Shin et al., 2021) as the difference between the first and second maximal values of the probability vector and also described in the main paper:

$$\kappa_{\text{margin}}(x_{i,j}) = \max_{c}(p_c(x^i_{j,k})) - \max2_c(p_c(x^i_{j,k})), \tag{F.3}$$

where max2 denotes the vector's second maximum value. All alternatives are compared in Table F.1.

F.2 Decomposition and analysis of Unimatch

Unimatch (Yang et al., 2023) investigating the consistency and suggest using FixMatch (Sohn et al., 2020a) and a strong baseline for semi-supervised semantic segmentation. Moreover, they provide analysis that shows that combining three students for each supervision signal, one feature level augmentation, feature perturbation, denoted by FP, and two strong augmentations, denoted by S1 and S2. Fusing Unimatch and our method did not provide significant improvements, and we examined the contribution of different components of Unimatch. We measured the pixel agreement as described in Eq. (9) and showed that the



(a) The spatial agreement as we define in in 9 compared between different variations of Unimatch and S4MC, on PASCAL VOC 12 dataset.

(b) The spatial agreement, compared between different variations of Unimatch (Yang et al., 2023) and S4MC, on PASCAL over time.

Figure F.1: Spatial agreement analysis

feature perturbation branch has the same effect on pixel agreement as S4MC. Fig. F.1 present the distribution of agreement using FixMatch (S1), DusPerb (S1,S2), Unimatch (S1, S2, FP) and S4MC (S1, S2).

G Bounding the joint probability

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

$$p_c^1(x^i_{j,k}, x^i_{\ell,m}) \approx p_c(x^i_{j,k}) \cdot p_c(x^i_{\ell,m}) \tag{G.1}$$

In addition to the independence approximation, it is possible to estimate the unconditional expectation of two neighboring pixels belonging to the same class based on labeled data:

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

To avoid overestimating that could lead to overconfidence, we set

$$p_c(x^i_{j,k}, x^i_{\ell,m}) = \max(p_c^1(x^i_{j,k}, x^i_{\ell,m}), p_c^2(x^i_{j,k}, x^i_{\ell,m})) \tag{G.3}$$

That upper bound of joint probability ensures that the independence assumption does not underestimate the joint probability, preventing overestimating the union event probability. Using Eq. (G.3) increase the mIoU by **0.22** on average, compared to non use of S4MC refinement, using 366 annotated images from PASCAL VOC 12 Using only Eq. (G.2) reduced the mIoU by **-14.11** compared to the non-use of S4MC refinement and harmed the model capabilities to produce quality pseudo-labels.