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 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. As a result, our method increases the amount of unlabeled data used during training while maintaining the quality of the pseudo-labels 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 to 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://github.com/s4mcontext/s4mc.

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 large amounts of labeled data, which can be costly and time-consuming to obtain. In practical scenarios, often a substantial amount of data is available, but only a portion of it can be labeled due to limited resources. 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), which offers a way to utilize both labeled and unlabeled data in model training.

This paper focuses on applying SSL to the task of semantic segmentation, which involves assigning each pixel of an image a semantic label corresponding to an object category. Semantic segmentation has applications in various areas such as perception for autonomous vehicles (Bartolomei et al., 2020), mapping (Van Etten et al., 2018), agriculture (Milioto et al., 2018), and medicine (Asgari Taghanaki et al., 2021). SSL is particularly appealing for segmentation tasks (Hu et al., 2021; Wang et al., 2022), 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 meaningful training signal, it is essential to modify the predictions in some way before incorporating them into the learning process. Several techniques can be used 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, teacher's may undergo weaker augmentations than student's ones (Berthelot et al., 2019b), making the teacher's task easier.

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

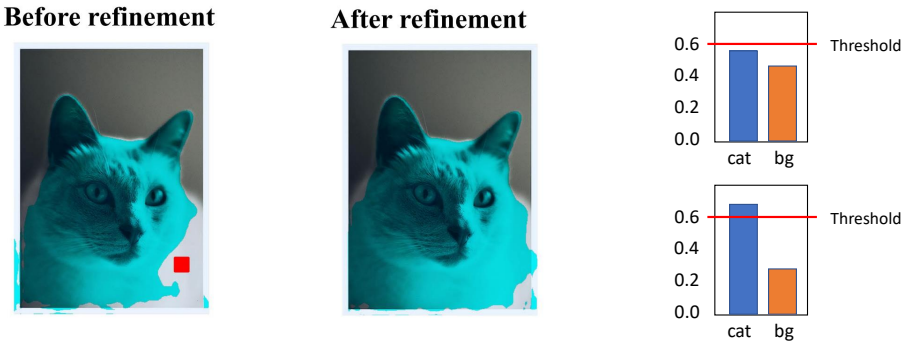


Figure 1: **Pseudo-label refinement. Left:** pseudo-labels generated by the model. **Middle:** pseudo-labels obtained from the same model after refinement with marginal contextual information. **Right top:** the predicted probabilities of the top two classes of the pixel in the middle of the red square. **Right bottom:** the probabilities obtained after refinement with marginal contextual information (7).

Moreover, pseudo-labeling is intrinsically prone to confirmation bias, which leads to reinforcement of the model's predictions rather than student model improvement. Reducing confirmation bias is especially important for erroneous predictions made by the teacher network.

One popular technique for that is confidence-based filtering (Sohn et al., 2020a). To reduce the number of incorrect pseudo-label, a threshold is set to assign pseudo-labels only when the model's confidence is high. Though simple, this idea was proven quite powerful 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 while using fewer labeled instances that may not represent the entire sample distribution leads to overfitting. Lowering the threshold would allow for higher training volumes at the cost of reduced quality, further hindering the performance (Sohn et al., 2020a).

The primary contribution of this work is a novel confidence refinement scheme of the teacher network predictions in segmentation tasks that increases the number of pseudo-labels available without sacrificing 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 look at the probability that at least one pixel is associated with a given class, and assign pseudo-label if this probability is higher then any other class union-event probability. By taking marginal 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) and entropy minimization (Grandvalet and Bengio, 2004) are two of the most popular and effective basic approaches in modern SSL. Entropy minimization aims to minimize

the entropy of unseen examples, causing confidant models without supervision. Pseudo-labeling, the more common setup, is the process of automatically assigning labels to unlabeled data based on the model's predictions. However, to utilize these labels in the training process, it is crucial to refine them to generate additional training signals (Laine and Aila, 2016; Berthelot et al., 2019b,a; Xie et al., 2020). Consistency regularization (Laine and Aila, 2016; Tarvainen and Valpola, 2017; Miyato et al., 2018) is one way to achieve this by adding a loss term that ensures consistency of model predictions between different views of the unlabeled data. Alternatively, we can use a stronger model (teacher) to obtain the pseudo-labels and train a weaker one (student). The teacher model can be a temporal ensemble of the student model (Tarvainen and Valpola, 2017). Furthermore, we can enhance the quality of pseudo-labels by increasing the temperature of the prediction (Berthelot et al., 2019b) (soft pseudo-labels) or assigning the exact label to samples whose confidence exceeds a certain threshold (Xie et al., 2020; Sohn et al., 2020a; Zhang et al., 2021) (hard pseudo-labels).

2.2 Semi-Supervised Semantic Segmentation

SSL semantic segmentation schemes rely mostly on consistency regularization, segmentation-compatible augmentation design, and pseudo-labeling. Since segmentation benchmarks turn out to have unevenly distributed labels, some methods (Zhang et al., 2021; Hu et al., 2021) aim to achieve balanced predictions via adaptive sampling, augmenting, and loss re-weighting. A common approach (Ouali et al., 2020; Zou et al., 2021; Liu et al., 2022b) utilizes feature perturbations on unlabeled data to improve consistency prediction. PS-MT (Liu et al., 2022b) introduces virtual adversarial training perturbations, U^{2}PL (Wang et al., 2022) utilize unreliable pixels by contrastive loss with the least confident classes predicted by the model. Certain methods (Yang et al., 2022c; Wang et al., 2022) use a curriculum learning approach, increasing the portion of usage of data over time. (Fan et al., 2022; Chen et al., 2021) revise co-teaching (Han et al., 2018) for semantic segmentation.

The methods above have introduced significant improvements in the training scheme. Nevertheless, most of these do not address the use of unreliable pixels. To the best of our knowledge, no prior art look at the contextual information gained from the spatial predictions on unlabeled data.

2.3 Contextual Information

Contextual information refers to environmental cues that aid in interpreting and extracting meaningful insights from visual perception (Toussaint, 1978; Elliman and Lancaster, 1990). It also enable to incorporate prior knowledge, like continues structure of objects for example, can help models transcend pixel-level analysis. (**Moshe**: please read the above) Recent developments in computer vision self-supervision utilize the ability of large models to improve contextual understanding, and solve related problems. Contrastive learning methods (He et al., 2019; Chen et al., 2020b; Hénaff et al., 2020; Chen et al., 2020a) reduce the class ambiguity by forcing the learned features to disentangle. Other methods, such as Masked Autoencoders (He et al., 2021; Noroozi and Favaro, 2016), are forced to fill in generated gaps in the image. These methods implicitly improve the contextual perception of the model but require additional computational power and a large amount of data. Under semi-supervision, where the annotated data is limited, these advantages become limited or vanish.

3 Method

In this section, we first give an overview of the problem definition and existing approaches, followed by a general description of the pseudo-labeling regime for unlabeled data in Section 3.1. Our proposed method to generate more reliable pseudo-labels is introduced in Section 3.2.

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 $\mathbf{x}$. In practical scenarios, $N_u \gg N_l$. We aim to train a neural network f_θ to predict the semantic segmentation of unseen images sampled from D_x.

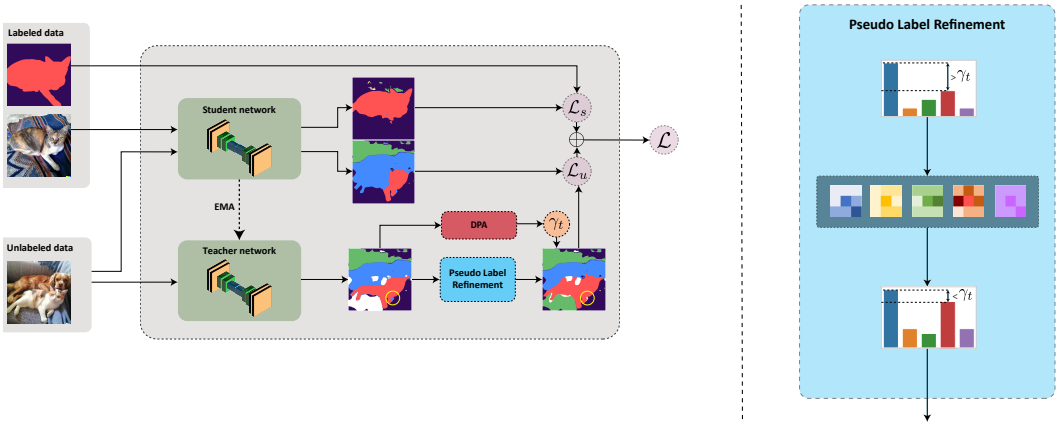


Figure 2: **Left:** The S4MC pipeline employs a teacher-student paradigm for semi-supervised segmentation. Labeled images are used to directly supervise the student network. Both teacher and student networks process unlabeled images. Predictions from the teacher network are refined and used to establish a margin criterion, 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 pseudo-label refinement module exploits neighboring pixels to adjust per-class predictions, as detailed in Section 3.2.1. Pay attention to the transformation in the class distribution of the pixel marked by the yellow circle. Before refinement, the margin criterion 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.(**OrL**: rewrote the thing – please check.)

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:

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

where $\mathcal{L}_s$ is the supervised loss over the labeled data, $\mathcal{L}_u$ is the loss over the unlabeled data, and λ_u is a hyperparameter controlling the weight of the unlabeled data loss. The exact definition of the two losses depends on the algorithm being used; in our case, both supervised and unsupervised losses are the standard cross-entropy loss. The difference is the use of hard pseudo-labels in $\mathcal{L}_u$:

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

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

where $\hat{\mathbf{y}}_i$ is the pseudo-label for the i-th unlabeled image. Note that not every pixel of $\mathbf{x}_i$ has a corresponding $\hat{\mathbf{y}}_i$. (**OrL**: But we divide by the batch size, so we ignore that actual number of pseudo-labeled pixels?) We set λ_u to be the ratio of labeled and unlabeled batch sizes $\lambda_u = |\mathcal{B}_u|/|\mathcal{B}_\ell|$ to match the weight of labeled and unlabeled images even when batches are sampled in an unbalanced manner. The choice of the pseudo-label generation algorithm is critical for the whole setup. Our novel approach is described in Section 3.2.1.

Pseudo-Labeling Propagation: For a given image $\mathbf{x}_i$, we denote by $\mathbf{x}_{j,k}^i$ the pixel in the j-th column and k-th row. We follow FixMatch (**OrL**: add citation) and assign a pseudo-label to a pixel if the confidence of some prediction criterion is high enough:

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

where we denote by $p(x^i_{j,k})$ the vector of teacher prediction probabilities, i.e., $f_{\theta_t}(\mathbf{x_i})_{j,k}$. $p_c(x^i_{j,k})$ is the probability score of class c and κ quantifies the teacher model confidence. In the simplest case, $\kappa(x^i_{j,k}; \theta_t) = \max_c p_c(x^i_{j,k})$.

To produce high-quality pseudo-labels, we need to quantify the teacher's confidence, denote by κ.We define the pixel-wise margin, inspired by Scheffer et al. (2001) and Shin et al. (2021), as the difference between the first and second maximal values of the probability vector:

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

where $\max 2$ denotes the vector's second maximum value.

Dynamic Partition Adjustment: Following U^{2}PL (Wang et al., 2022), we use a decaying threshold, namely, Dynamic Partition Adjustment (DPA). DPA replaces the fixed threshold with a quantile-based threshold γ_t that decreases with time. At each iteration, we evaluate the α_t quantile of Eq. (5) over all pixels of all examples and set γ_t to its value. α_t is defined as follows:

$$\alpha_t = \alpha_0 \cdot \left(1 - \frac{i}{\text{iterations}}\right). \tag{6}$$

Since the model predictions become more precise with time, decreasing the threshold over time increases the number of unlabeled samples fed into the model 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 while 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 pseudo-labeling refinement, followed by our thresholding strategy and a description of S4MC methodology. (**Moshe**: please read this section)

3.2.1 Pseudo-label refinement

We refine the predicted pseudo-label of each pixel based on its spatial neighborhood. Given a pixel $x^i_{j,k}$ located at (j, k) with an associated per-class prediction $p_c(x^i_{j,k})$, we look at neighboring pixels $x^i_{\ell,m}$ inside an $N \times N$ pixel neighborhood around it and look at the probability that at least one of two pixels belong 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.

The joint probabilities for different pixels are unknown. However, it is reasonable to assume that the correlation between the probabilities that neighboring pixels belong to the same class c is non-negative. In other words, the information about the prediction of the model on neighboring pixels belonging to the same class cannot decrease the posterior on the given pixel.

In this case the joint probability can be bounded from below by independent assumption: $p_c(x^i_{j,k}, x^i_{\ell,m}) \geqslant p_c(x^i_{j,k}) \cdot p_c(x^i_{\ell,m}))$, While this boundary could be use as a naive approximation, we additionally investigate other ways to have a closer bounded estimation of the joint probability, as presented in the Appendix C in the supplementary material. For each class c, we use $\tilde{p}_c(x^i_{j,k})$ from Eq. (7) and chose the neighbor with the maximal information gain on the same class, i.e.,

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

(**OrL**: we don't use eq7 thought, we use 7 and 8, no? should we actually plug in 8 into 7 to have the correct expression being used?)(**Moshe**: rephrased it, it's assigning 7 into 8, is that not clear or not smart to do?) In this way, an unreliable label predictions can gain information from the evidence the model acquired from the neighborhood to not create over-confident predictions (**OrL**: the example is not clear enough. its a good idea to give an example or intuition but make sure to connect it to

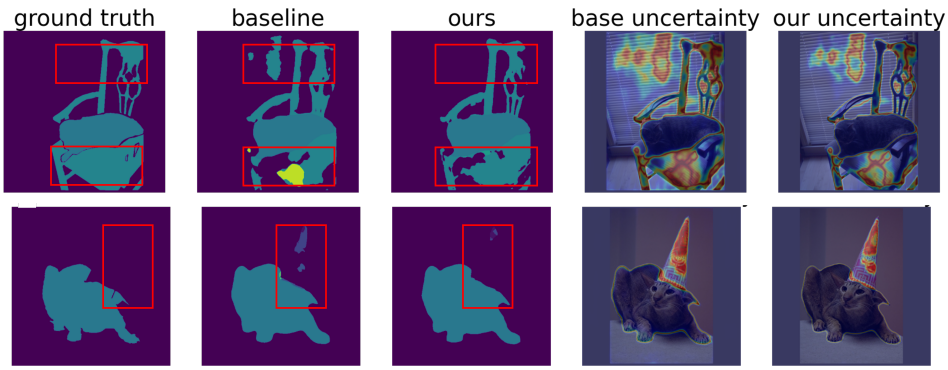


Figure 3: **Example of refined pseudo-labels.** The outputs of two trained models and the annotated ground truth. **Ours**: Prediction map after refinement as described in Section 3.2. **Baseline**: Prediction map without pseudo-label refinement. **Heat map**: Represent the uncertainty of the model as ${\kappa_{margin}}^{-1}$, based on Eq. (5). The baseline model predicts pixels as unassociated with classes of adjacent pixels and classes that do not occur in the image and generally have regions with higher values.

what you just derived in math)(**Moshe**: reffering to the cat example from figure 1 would be good? or describe a new example?), thus reducing the confirmation bias. The refinement is visualized in Fig. 1. The spatial addition to the contextual information can benefit from more than one neighbor. However, union events saturate rapidly and get a value close to 1 when there are more than two events. An ablation study on the neighborhood size and how to choose the neighbor pixels is provided in Table 4.

For larger neighborhoods, we decrease the probability contribution of the neighboring pixels with a distance-dependent factor :

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

where $\beta_{\ell,m} = \exp\big(-\frac{1}{2}(|\ell - j| + |m - k|)\big)$ is a spatial weighting function.

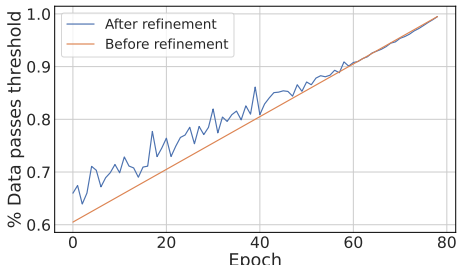
Empirically, contextual information refinement affects mainly the most probable class or two. This aligns well with our choice to use the margin confidence (5), which only considers the two most probable classes.

3.2.2 Threshold setting

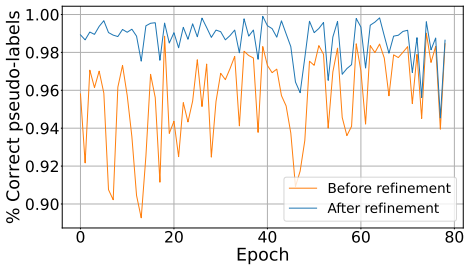
Setting a high threshold can help to reduce the confirmation bias from the teacher model "beliefs" to the student model at the cost of learning from fewer examples, that in turn leads to a less comprehensive model. DPA aim to close the gap by setting a temporally lower threshold that increase with time. We modify the DPA to account for the spatial information. Instead of using the same value for determining the threshold and filtering the labels, we choose the threshold based on raw teacher predictions $p_c(x^i_{j,k})$, but filter values based on $\tilde{p}_c(x^i_{j,k})$. As a result, more pixels pass the threshold, but the threshold itself is not affected directly. We set $\alpha_0 = 0.4$, i.e., 60% of raw predictions pass the threshold at $t = 0$, as this value performed the best in our experiments. An ablation study for α_0 is provided in Table 5.

3.2.3 Putting it all together

We perform SSL by pseudo-labeling pixels using their neighbors' contextual information. First, a batch of labeled images is fed into the student model only, producing the supervised loss as defined in Eq. (2). Then, a batch of unlabeled images is fed into the student and teacher models. We sort the confidence scores for all pixels from the teacher model and find γ_t. The teacher predictions are refined using the *weighted union event* distribution for each class, as defined in Eq. (7). Then pixels with the new confidence values that pass the temporal threshold are assigned pseudo-labels as described in Eq. (4), producing the unsupervised loss as defined in Eq. (3).Fig. 4 shows the additional number of pixels at each iteration and the confidence distribution before and after using the spatial extensions. A scheme of our method can be viewed in Fig. 2.



(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 classic PASCAL VOC 2012 `val` dataset with different amounts of labeled data. Labeled images are selected from the PASCAL VOC 2012 `train` dataset without the additional augmented set. For each partition protocol, the caption gives the share of the training set used as labeled data and, in parentheses, the actual number of labeled images. "Supervised Only" stands for supervised training without using any unlabeled data.

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	72.00	73.10	74.70	-
ST++ (Yang et al., 2022c)	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 (Ours)	**70.96**	71.69	75.41	77.73	**80.58**
S4MC ψ (Ours)	**74.32**	**75.62**	**77.84**	**79.72**	**81.51**

The effect of S4MC visualized in Fig. 4 that compares the fraction of pixels that pass the threshold between the baseline and our method. One can see that our method utilizes more unlabeled data for the training procedure in most of the training process and reduce it by a little when the threshold is low, aiming to use the teacher predictions in a manner that prevent propagating errors. Another evidence of that is the percentage of correct pseudo-labels, where one can see that our method passes more correct pseudo-labels.

4 Experiments

In this section, we provide our experimental results. Section 4.1 describes the experimental setup for the various datasets and partition protocols. Implementation details are provided in the appendix. We compare our method with existing methods for semi-supervised semantic segmentation in Section 4.2. Finally, we offer an ablation study on our choices in Section 5.

4.1 Setup

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

Table 2: Comparison between our method and prior art on the coarse PASCAL VOC 2012 `val` dataset under different partition protocols. Labeled images are selected from the PASCAL VOC 2012 `train` dataset with additional data from (Hariharan et al., 2011), for a total of 10,582 annotated images. For each partition protocol, the caption gives the share of the training set used as labeled data and, in parentheses, the actual number of labeled images. "Supervised Only" stands for supervised training without using any unlabeled data. As a baseline, we use CutMix-Seg (French et al., 2020).

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^{2}PL (Wang et al., 2022)	77.21	79.01	79.3	80.50
S4MC (Ours)	**78.49**	**79.67**	**79.85**	**81.11**
S4MC ψ (Ours)	**80.77**	**81.9**	**82.3**	**83.3**

The PASCAL VOC dataset includes 20 object classes (plus background). The dataset includes 2,913 annotated images, separated into a training set of 1,464 images and a validation set of 1,449 images. In addition, the dataset includes 9,118 coarse-annotated training images (Hariharan et al., 2011), in which all the annotated pixels can be used as labeled pixels, and the rest are ignored. Following prior research, we conduct two sets of experiments. One (classic) uses only the original training set (Wang et al., 2022; Zou et al., 2021), and the second (coarse) 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 usually 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. For both datasets, we compare S4MC with others under the common partition protocols – using $1/2$, $1/4$, $1/8$, and $1/16$ of the training data as labeled data. For classic PASCAL, we also compare using all the finely annotated images. We use RandAugment (Cubuk et al., 2020) or CutMix (Yun et al., 2019) for training augmentations.

Network Architecture We use DeepLabv3+ (Chen et al., 2018b) in all our experiments. As an encoder backbone, we use an 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': $\theta_t^\eta = \tau\theta_t^{\eta-1} + (1-\tau)\theta_s^\eta$, where $0 \leq \tau \leq 1$ defines how close the teacher is to the student and η denotes the time. For the teacher update, following Tarvainen and Valpola (2017), we set $\tau = 0.99$.

Evaluation The primary evaluation metric is mean Intersection over Union (mIoU). We use the data split published by Wang et al. (2022) when available to ensure consistency in our comparisons. For the ablation studies, we use PASCAL VOC 2012 `val` with $1/2$ partition.

4.2 Comparison with Existing Methods

We compare our approach with popular and state-of-the-art semi-supervised semantic 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^{2}Seg (Zhong et al., 2021), AEL (Hu et al., 2021), U^{2}PL (Wang et al., 2022), PS-MT (Liu et al., 2022b), and ST++ (Yang et al., 2022c).

Results on the PASCAL VOC 2012 Dataset Table 1 compares our method with the other state-of-the-art methods on the classic PASCAL VOC 2012 dataset, while Table 2 shows the comparison results on the PASCAL VOC 2012 dataset with SBD (Hariharan et al., 2011) additional coarse annotated data. For both setups, S4MC outperform all the compared methods in standard partition

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 and, in parentheses, the actual number of labeled images. "Supervised Only" stands for supervised training without using any unlabeled data.

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 (Ours)	**75.03**	**77.02**	**78.78**	78.86
S4MC ψ (Ours)	**76.3**	**78.25**	**78.95**	**79.13**

Table 4: Ablation study on the neighborhood, over Pascal VOC 12 with partition protocol of $1/4$ (366 labeled images).

Method	**Neighborhood size N**		
	$\mathbf{3 \times 3}$	$\mathbf{5 \times 5}$	$\mathbf{7 \times 7}$
Max neighbor	75.41	75.18	74.89
Random neighbor	73.25	71.1	70.41
Two max neighbors	74.14	75.15	74.36
Min neighbor	74.54	74.11	70.28

protocols, both when using labels only for the original PASCAL VOC 12 dataset and when using SBD annotations as well.

Results on the Cityscapes Dataset Table 3 Presents the comparison results on the Cityscapes `val` dataset. Table 3 compares our method with the other state-of-the-art methods on the Cityscapes (Cordts et al., 2016) dataset under various partition protocols. S4MC outperforms the current state-of-the-art methods in most partitions, except for $1/2$ setting. Our method combined with a more powerful Fixmatch with feature perturbation (Yang et al., 2022b) approach outperforms state-of-the-art approaches across all partitions.

Contextual spatial information at inference time The work in this paper raises a question about our usage of contextual information at inference time. Accordingly, we investigate the use of our method to see if the prediction map of a pre-trained model improves. The results show that the DeepLab-V3-plus model improves by minor improvement from 85.7 mIOU to 85.71. Sometimes we even saw a degradation in the performance of our models; nevertheless, this was still very insignificant. The reason for the insignificance of impact is that the refinement does change the model confidence based on spatial information, but not to the extent of swapping the maximal class in most cases. A heat map of the prediction over some samples is shown in the appendix.

5 Ablation Study

Effectiveness of using neighbor pixels To explore the use of spatial contextual information using neighbor pixels, we conduct experiments that ask what neighboring pixels can elevate the reliability of the predictions. Table 4 examine using a neighbor with maximal probability, a neighbor with minimal probability, or a random neighbor, as well as two neighbors, for different neighborhood sizes. Using a small 3×3 neighborhood with one neighboring pixel was the most efficient in our experiments.

Table 5: **Ablation study on α_0 in Eq. (6)**, which controls the initial proportion of confidence pixels over the PASCAL dataset with the partition of 1/4 (366) annotated.

20%	30%	40%	50%	60%
74.45	73.85	**75.18**	74.56	74.31

Hyperparameter tuning We ablate several important parameters for S4MC. Table 4 demonstrates the effect of both the neighborhood size and the number of neighbor pixels, with different sampling strategies on the mIoU results on PASCAL VOC 2012 `val` set.

Table 5 studies the impact of different initial quantiles that help set a threshold. We find $\alpha_0 = 40\%$ achieves the best performance. Small α_0 will propagate too many mistakes, leading to severe conformation bias. Large α_0 means we mask most of the data, will propagate very little data, and make the semi-supervision learning process long and inefficient.

6 Conclusion

In this paper, we present a novel way of looking at contextual information. Based on the spatial contextual information, we propose a framework (S4MC) for semi-supervised semantic segmentation using this perspective and insights. Our work leverages spatial information and helps the model in scenarios where the annotated data fail to build quality pseudo-labels. Our method can be easily incorporated into any framework that performs dense prediction and utilizes thresholding. It does not need extra trainable parameters or annotated data and requires negligible additional compute. S4MC outperforms existing approaches and achieves state-of-the-art results in multiple popular benchmarks under various data partition protocols, such as Cityscapes and Pascal VOC 12. Exploiting other aspects of the contextual information, such as spectral contextual information, coherency for instance, remains open for future work.

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)

Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 54(1): 137–178, 2021. doi: 10.1007/s10462-020-09854-1. URL https://arxiv.org/abs/1910.07655. (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. 3)

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 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 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 p. 8)

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, July 2020a. URL https://proceedings.mlr.press/v119/chen20j.html. (cited on p. 3)

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. 3, 7, 8, and 9)

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint*, March 2020b. URL https://arxiv.org/abs/2003.04297. (cited on p. 3)

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, 8, and 9)

Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. RandAugment: practical automated data augmentation with a reduced search space. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18613–18624. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html. (cited on p. 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)

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, 16, and 17)

Jiashuo Fan, Bin Gao, Huan Jin, and Lihui Jiang. UCC: uncertainty guided cross-head co-training for semi-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9947–9956, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Fan_UCC_Uncertainty_Guided_Cross-Head_Co-Training_for_Semi-Supervised_Semantic_Segmentation_CVPR_2022_paper.html. (cited on p. 3)

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, 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)

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://papers.nips.cc/paper/2004/hash/96f2b50b5d3613adf9c27049b2a888c7-Abstract.html. (cited on p. 2)

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/a19744e268754fb0148b017647355b7b-Abstract.html. (cited on p. 3)

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 p. 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. 8)

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2019. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html. (cited on p. 3)

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2021. (cited on p. 3)

Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aäron van den Oord. Data-efficient image recognition with contrastive predictive coding. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4182–4192. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/henaff20a.html. (cited on p. 3)

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. 1, 3, 8, and 9)

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. 8 and 9)

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. 3)

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, 8, and 9)

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. 3)

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 69–84. Springer, 2016. (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, 8, and 9)

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. 8)

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. 5 and 16)

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. 5 and 16)

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, 5, and 17)

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, 3, 4, and 8)

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)

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)

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. 1, 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, 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 p. 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, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. *ArXiv*, abs/2208.09910, 2022b. (cited on pp. 9 and 17)

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 2022c. 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 pp. 8 and 17)

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 pp. 7 and 8)

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, 7, and 8)

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.

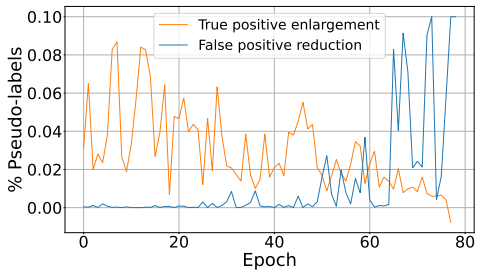


Figure 5: **Quality of pseudo-labels**, on PASCAL VOC 2012 (Everingham et al., 2010) over time. At the early stage of the learning process, the increase in pseudo-labels' quality is mostly due to true positive improvement. In other words, the refinement not only helps more pixels pass the threshold, 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,

$$\kappa_{\max}(x^i_{j,k}) = \max_c p_c(x^i_{j,k}), \tag{10}$$

which is commonly used to generate pseudo-labels.

The second alternative is negative entropy, defined as

$$\kappa_{\text{ent}}(x^i_{j,k}) = \sum_{c \in C} p_c(x^i_{j,k}) \log\left(p^c_{i,j}\right). \tag{11}$$

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:

$$\kappa_{\text{margin}}(x_{i,j}) = \max_c(p_c(x^i_{j,k})) - \max 2_c(p_c(x^i_{j,k})), \tag{12}$$

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 6.

Table 6: 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 6 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.

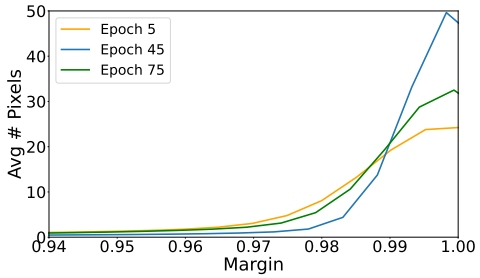


Figure 6: **Distribution of confidence estimation**, an average over a batch of 256 unlabeled images from PASCAL VOC 2012 (Everingham et al., 2010) over time, using margin confidence estimation. At the early stage of the learning process, the model is less confident over the unlabeled images. at the 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

$$p_c^1(x_{j,k}^i, x_{\ell,m}^i) \approx p_c(x_{j,k}^i) \cdot p_c(x_{\ell,m}^i) \tag{13}$$

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

$$p_c^2(x_{j,k}^i, x_{\ell,m}^i) = \frac{1}{|\mathcal{N}_l| \cdot H \cdot W \cdot |\mathbf{N}|} \sum_{y,i,j} \sum_{k,\ell \in \mathbf{N}_{i,j}} \mathbb{1}\{y_{i,j} = y_{k,\ell}\} \tag{14}$$

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

$$p_c(x_{j,k}^i, x_{\ell,m}^i) = \max(p_c^1(x_{j,k}^i, x_{\ell,m}^i), p_c^2(x_{j,k}^i, x_{\ell,m}^i)) \tag{15}$$

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. (15) 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.

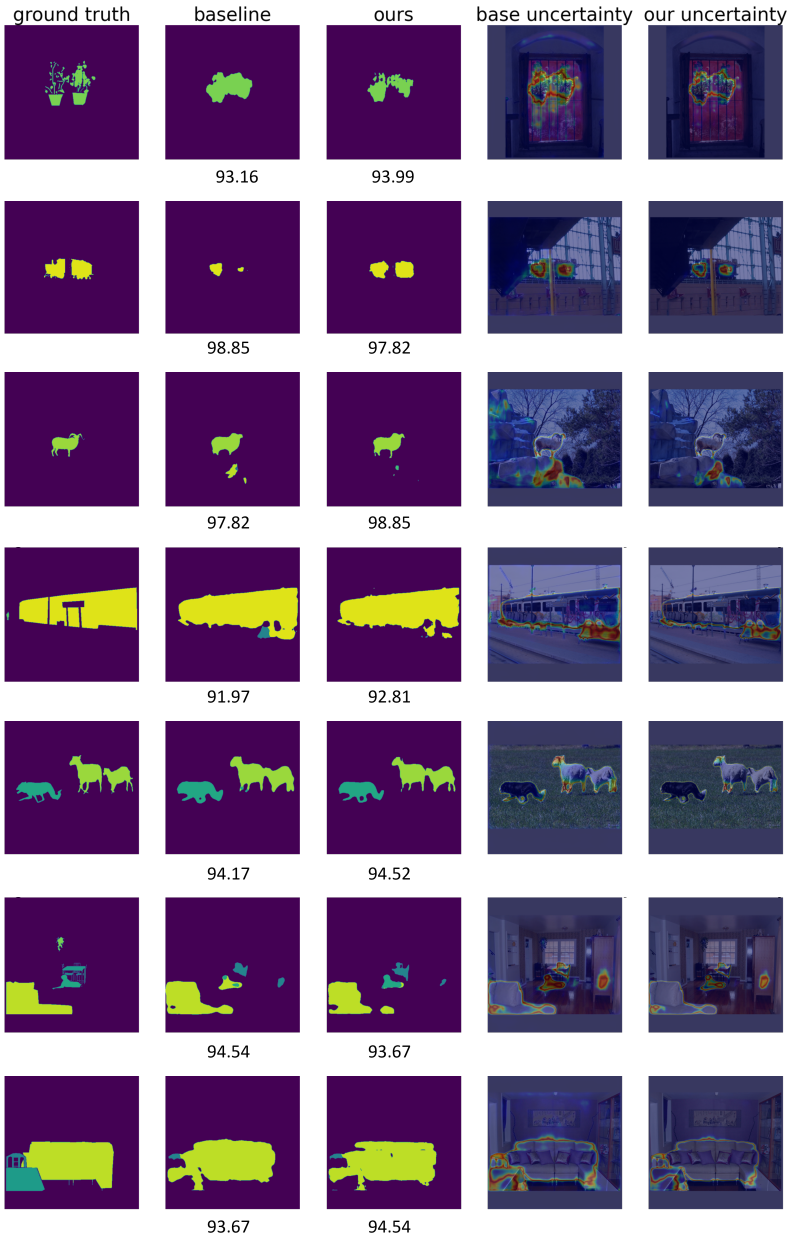
Experiments annotated with ψ uses the Fixmatch framework (Sohn et al., 2020a) with feature perturbation (Yang et al., 2022b).

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.

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.



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.