

---

# Semi-Supervised Semantic Segmentation via Marginal Contextual Information

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

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

## 15 1 Introduction

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

24 This paper focuses on applying SSL to semantic segmentation, which has applications in various  
25 areas such as perception for autonomous vehicles (Bartolomei et al., 2020), mapping (Van Etten et al.,  
26 2018) and agriculture (Milioto et al., 2018). SSL is particularly appealing for segmentation tasks, as  
27 manual labeling can be prohibitively expensive.

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



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

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

39 One popular technique to address this issue is confidence-based filtering (Sohn et al., 2020a). This  
 40 approach assigns pseudo-labels only when the model’s confidence surpasses a specified threshold,  
 41 thereby reducing the number of incorrect pseudo-labels. Though simple, this strategy was proven  
 42 effective and inspired multiple improvements in semi-supervised classification (Zhang et al., 2021;  
 43 Rizve et al., 2021), segmentation (Wang et al., 2022), and object detection in images (Sohn et al.,  
 44 2020b; Liu et al., 2021) and 3D scenes (Zhao et al., 2020; Wang et al., 2021). However, the strict  
 45 filtering of the supervision signal leads to extended training periods and, potentially, to overfitting  
 46 when the labeled instances used are insufficient to represent the entire sample distribution. Lowering  
 47 the threshold would allow for higher training volumes at the cost of reduced quality, further hindering  
 48 the performance (Sohn et al., 2020a).

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

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

## 65 2 Related Work

### 66 2.1 Semi-Supervised Learning

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

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

## 2.2 Semi-Supervised Semantic Segmentation

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

## 2.3 Contextual Information

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

# 3 Method

## 3.1 Overview

In semi-supervised semantic segmentation, we are given a labeled training set of images  $\mathcal{D}_\ell = \{(\mathbf{x}_i^\ell, \mathbf{y}_i)\}_{i=1}^{N_\ell}$ , and an unlabeled set  $\mathcal{D}_u = \{\mathbf{x}_i^u\}_{i=1}^{N_u}$  sampled from the same distribution, i.e.,  $\{\mathbf{x}_i^\ell, \mathbf{x}_i^u\} \sim D_x$ . 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  $f_\theta$  to predict the semantic segmentation of unseen images sampled from  $D_x$ .

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

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_u \quad (1)$$

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

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

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

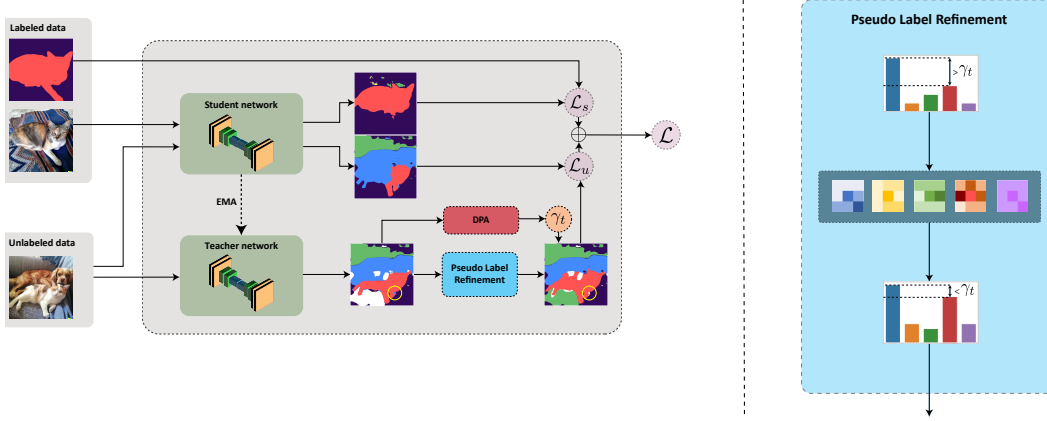


Figure 2: **Left:** S4MC employs a teacher-student paradigm for semi-supervised segmentation. Labeled images are used to supervise the student network directly. Both teacher and student networks process unlabeled images. Predictions from the teacher network are refined and used to evaluate the margin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

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

### 3.1.1 Pseudo-label Propagation

For a given image  $\mathbf{x}_i$ , we denote by  $\mathbf{x}_{j,k}^i$  the pixel in the  $j$ -th row and  $k$ -th column. We adopt a thresholding-based criterion inspired by (Sohn et al., 2020a). By establishing a score, denoted as  $\kappa$ , 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  $\gamma_t$ :

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

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

$$\kappa_{\text{margin}}(\mathbf{x}_{j,k}^i) = \max_c \{p_c(\mathbf{x}_{j,k}^i)\} - \max_{c \neq \arg \max_c \{p_c(\mathbf{x}_{j,k}^i)\}} \{p_c(\mathbf{x}_{j,k}^i)\}. \quad (5)$$

### 3.1.2 Dynamic Partition Adjustment (DPA)

Following U<sup>2</sup>PL (Wang et al., 2022), we use a decaying threshold  $\gamma_t$ . DPA replaces the fixed threshold with a quantile-based threshold that decreases with time. At each iteration, we set  $\gamma_t$  as the  $\alpha_t$ -th quantile of  $\kappa_{\text{margin}}$  over all pixels of all images in the batch.  $\alpha_t$  is defined as follows:

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

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

## 3.2 Marginal Contextual Information

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

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

### 3.2.1 Confidence Margin Refinement

We refine the predicted pseudo-label of each pixel by considering the predictions of its neighboring pixels. Given a pixel  $x_{j,k}^i$  with a corresponding per-class prediction  $p_c(x_{j,k}^i)$ , we examine neighboring pixels  $x_{\ell,m}^i$  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  $c$ :

$$\tilde{p}_c(x_{j,k}^i) = p_c(x_{j,k}^i) + p_c(x_{\ell,m}^i) - p_c(x_{j,k}^i, x_{\ell,m}^i), \quad (7)$$

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

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

$$\tilde{p}_c(x_{j,k}^i) \leq p_c(x_{j,k}^i) + p_c(x_{\ell,m}^i) - p_c(x_{j,k}^i) \cdot p_c(x_{\ell,m}^i). \quad (8)$$

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

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

$$\tilde{p}_c^{\mathbf{N}}(x_{j,k}^i) = \max_{\ell,m} \tilde{p}_c(x_{j,k}^i). \quad (9)$$

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

$$\tilde{p}_c(x_{j,k}^i) = p_c(x_{j,k}^i) + \beta_{\ell,m} [p_c(x_{\ell,m}^i) - p_c(x_{j,k}^i, x_{\ell,m}^i)], \quad (10)$$

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

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

### 3.2.2 Threshold Setting

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

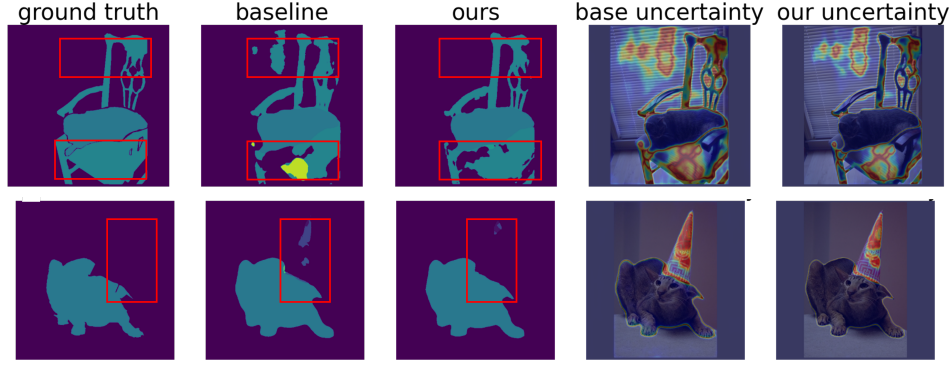


Figure 3: **Qualitative results of S4MC.** The outputs of two trained models and the annotated ground truth. The segmentation map predicted by S4MC (*Ours*) compared to the segmentation map using no refinement module (*Baseline*) and to the ground truth. *Heat map* represents the uncertainty of the model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

### 180 3.3 Putting it All Together

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

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

## 194 4 Experiments

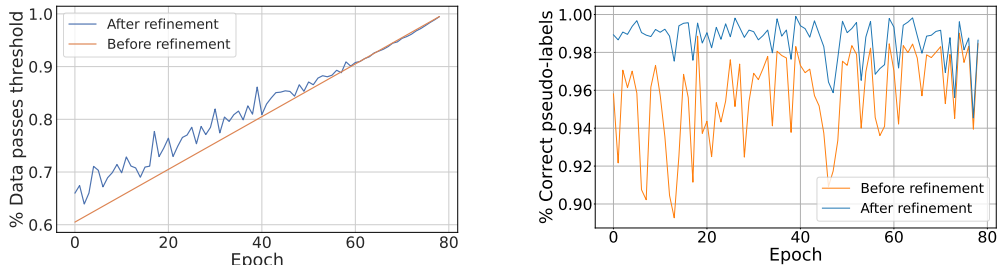
195 This section presents our experimental results. The setup for the different datasets and partition  
 196 protocols is detailed in Section 4.1. Section 4.2 compares our method against existing approaches and  
 197 Section 4.3 provides the ablation study. Further implementation details are given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

### 198 4.1 Setup

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

201 The PASCAL VOC dataset comprises 20 object classes (plus background). The dataset includes  
 202 2,913 annotated images, divided into a training set of 1,464 images and a validation set of 1,449  
 203 images. In addition, the dataset includes 9,118 coarsely annotated training images (Hariharan et al.,  
 204 2011), in which only a subset of the pixels are labeled. Following previous research, we conduct two  
 205 sets of experiments. The ‘classic’ experiment utilizes only the original training set (Wang et al., 2022;  
 206 Zou et al., 2021), while the ‘coarse’ experiment uses all available data (Wang et al., 2022; Chen et al.,  
 207 2021; Hu et al., 2021).

208 The Cityscapes (Cordts et al., 2016) dataset includes urban scenes from 50 different cities with  
 209 30 classes, of which only 19 are typically used for evaluation (Chen et al., 2018a,b). Similarly  
 210 to PASCAL, in addition to 2,975 training and 500 validation images, the dataset includes 19,998  
 211 coarsely annotated images, which we do not use in our experiment.



(a) **Data fraction that passes the threshold.** The baseline model has a fixed percentage, as it is based on DPA. Our method increases the number of pixels assigned pseudo-label, mostly in the early stage of the training when the model is under-confident. (b) **Fraction of correct pseudo-labels** the assigned pseudo-labels with the correct class divided by the total assigned pseudo-label. S4MC produces more quality pseudo-labels during the training process, most notably at the early stages.

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

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

| Method                                   | 1/16 (92)    | 1/8 (183)    | 1/4 (366)    | 1/2 (732)    | Full (1464)  |
|------------------------------------------|--------------|--------------|--------------|--------------|--------------|
| Supervised Only                          | 45.77        | 54.92        | 65.88        | 71.69        | 72.50        |
| CutMix-Seg (French et al., 2020)         | 52.16        | 63.47        | 69.46        | 73.73        | 76.54        |
| PseudoSeg (Zou et al., 2021)             | 57.60        | 65.50        | 69.14        | 72.41        | 73.23        |
| PC <sup>2</sup> Seg (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        | <u>72.0</u>  | 73.10        | 74.70        | -            |
| ST++ (Yang et al., 2022b)                | 65.2         | 71.0         | 74.6         | 77.3         | 79.1         |
| U <sup>2</sup> 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         | <u>76.6</u>  | <u>78.4</u>  | 80.0         |
| S4MC + CutMix-Seg (Ours)                 | <u>70.96</u> | 71.69        | 75.41        | 77.73        | <u>80.58</u> |
| S4MC + FixMatch (Ours)                   | <b>74.32</b> | <b>75.62</b> | <b>77.84</b> | <b>79.72</b> | <b>81.51</b> |

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

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

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

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

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

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

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

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

## 4.2 Results

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

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

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

| (a) Neighborhood choice. |                       |              |              | (b) $\alpha_0$ in Eq. (6), which controls the initial proportion of confidence pixels |       |              |       |       |
|--------------------------|-----------------------|--------------|--------------|---------------------------------------------------------------------------------------|-------|--------------|-------|-------|
| Selection criterion      | Neighborhood size $N$ |              |              |                                                                                       |       |              |       |       |
|                          | $3 \times 3$          | $5 \times 5$ | $7 \times 7$ | 20%                                                                                   | 30%   | 40%          | 50%   | 60%   |
| Random neighbor          | 73.25                 | 71.1         | 70.41        | 74.45                                                                                 | 73.85 | <b>75.41</b> | 74.56 | 74.31 |
| Max neighbor             | <b>75.41</b>          | 75.18        | 74.89        |                                                                                       |       |              |       |       |
| Min neighbor             | 74.54                 | 74.11        | 70.28        |                                                                                       |       |              |       |       |
| Two max neighbors        | 74.14                 | 75.15        | 74.36        |                                                                                       |       |              |       |       |

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

### 4.3 Ablation Study

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

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

**Threshold parameter tuning** As outlined in Section 3.1.2, we utilize a dynamic threshold that depends on an initial value,  $\alpha_0$ . In Table 4b, we examine the effect of different initial quantiles to establish this threshold. A smaller  $\alpha_0$  would propagate too many errors, leading to significant confirmation bias. In contrast, a larger  $\alpha_0$  would mask most of the data, resulting in insufficient label propagation, rendering the semi-supervised learning process lengthy and inefficient. We found that an  $\alpha_0$  of 40% yields the best performance.

## 5 Conclusion

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

## References

- Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International Joint Conference on Neural Networks*, pages 1–8, 2020. doi: 10.1109/IJCNN48605.2020.9207304. URL <https://arxiv.org/abs/1908.02983>. (cited on p. 1)
- Luca Bartolomei, Lucas Teixeira, and Margarita Chli. Perception-aware path planning for UAVs using semantic segmentation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5808–5815, 2020. doi: 10.1109/IROS45743.2020.9341347. (cited on p. 1)
- David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin A. Raffel. ReMixMatch: semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint*, November 2019a. URL <https://arxiv.org/abs/1911.09785>. (cited on p. 2)
- David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A. Raffel. MixMatch: a holistic approach to semi-supervised learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL <https://proceedings.neurips.cc/paper/2019/hash/1cd138d0499a68f4bb72bee04bbec2d7-Abstract.html>. (cited on pp. 1, 2, and 3)
- Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018a. doi: 10.1109/TPAMI.2017.2699184. URL <https://arxiv.org/abs/1412.7062>. (cited on pp. 1, 3, and 6)
- Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, September 2018b. URL [https://openaccess.thecvf.com/content\\_ECCV\\_2018/html/Liang-Chieh\\_Chen\\_Encoder-Decoder\\_with\\_Atrous\\_ECCV\\_2018\\_paper.html](https://openaccess.thecvf.com/content_ECCV_2018/html/Liang-Chieh_Chen_Encoder-Decoder_with_Atrous_ECCV_2018_paper.html). (cited on pp. 6 and 7)
- Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2613–2622, June 2021. URL [https://openaccess.thecvf.com/content/CVPR2021/html/Chen\\_Semi-Supervised\\_Semantic\\_Segmentation\\_With\\_Cross\\_Pseudo\\_Supervision\\_CVPR\\_2021\\_paper.html](https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Semi-Supervised_Semantic_Segmentation_With_Cross_Pseudo_Supervision_CVPR_2021_paper.html). (cited on pp. 6, 7, and 8)
- Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. URL [https://www.cv-foundation.org/openaccess/content\\_cvpr\\_2016/html/Cordts\\_The\\_Cityscapes\\_Dataset\\_CVPR\\_2016\\_paper.html](https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Cordts_The_Cityscapes_Dataset_CVPR_2016_paper.html). (cited on pp. 2, 6, and 8)
- Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. CoAtNet: marrying convolution and attention for all data sizes. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3965–3977. Curran Associates, Inc., 2021. URL <https://proceedings.neurips.cc/paper/2021/hash/20568692db622456cc42a2e853ca21f8-Abstract.html>. (cited on p. 1)
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. (cited on p. 3)
- Dave G. Elliman and Ian T. Lancaster. A review of segmentation and contextual analysis techniques for text recognition. *Pattern Recognition*, 23(3):337–346, 1990. ISSN 0031-3203. doi: [https://doi.org/10.1016/0031-3203\(90\)90021-C](https://doi.org/10.1016/0031-3203(90)90021-C). URL <https://www.sciencedirect.com/science/article/pii/003132039090021C>. (cited on p. 3)
- Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. doi: 10.1007/s11263-009-0275-4. URL <https://doi.org/10.1007/s11263-009-0275-4>. (cited on pp. 2, 6, 7, 15, and 16)
- Geoffrey French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham D. Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *British Machine Vision Conference*. BMVA Press, 2020. URL <https://www.bmvc2020-conference.com/assets/papers/0680.pdf>. (cited on pp. 7 and 8)

333 Ross Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, December  
334 2015. URL [https://openaccess.thecvf.com/content\\_iccv\\_2015/html/Girshick\\_Fast\\_R-CNN\\_](https://openaccess.thecvf.com/content_iccv_2015/html/Girshick_Fast_R-CNN_ICCV_2015_paper.html)  
335 [ICCV\\_2015\\_paper.html](https://openaccess.thecvf.com/content_iccv_2015/html/Girshick_Fast_R-CNN_ICCV_2015_paper.html). (cited on p. 1)

336 Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours  
337 from inverse detectors. In *International Conference on Computer Vision*, pages 991–998, 2011. doi:  
338 10.1109/ICCV.2011.6126343. (cited on pp. 6 and 8)

339 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image  
340 recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June  
341 2016. URL [https://www.cv-foundation.org/openaccess/content\\_cvpr\\_2016/html/He\\_Deep\\_](https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html)  
342 [Residual\\_Learning\\_CVPR\\_2016\\_paper.html](https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html). (cited on p. 7)

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

345 Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic  
346 segmentation via adaptive equalization learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang,  
347 and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages  
348 22106–22118. Curran Associates, Inc., 2021. URL [https://proceedings.neurips.cc/paper/2021/](https://proceedings.neurips.cc/paper/2021/file/b98249b38337c5088bbc660d8f872d6a-Paper.pdf)  
349 [file/b98249b38337c5088bbc660d8f872d6a-Paper.pdf](https://proceedings.neurips.cc/paper/2021/file/b98249b38337c5088bbc660d8f872d6a-Paper.pdf). (cited on pp. 3, 6, 7, and 8)

350 Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson W. H. Lau. Guided collaborative training for pixel-  
351 wise semi-supervised learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm,  
352 editors, *European Conference on Computer Vision*, pages 429–445, Cham, 2020. Springer International  
353 Publishing. ISBN 978-3-030-58601-0. URL [https://www.ecva.net/papers/eccv\\_2020/papers\\_](https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/1932_ECCV_2020_paper.php)  
354 [ECCV/html/1932\\_ECCV\\_2020\\_paper.php](https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/1932_ECCV_2020_paper.php). (cited on pp. 7 and 8)

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

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

362 Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference*  
363 *on Learning Representations*, 2016. URL <https://openreview.net/forum?id=BJ6o0fqge>. (cited on p.  
364 2)

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

369 Feng Li, Hao Zhang, Huaizhe xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask DINO:  
370 towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint*, June  
371 2022. URL <https://arxiv.org/abs/2206.02777>. (cited on p. 1)

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

374 <https://openreview.net/forum?id=6u6N8WWwYSM>. (cited on p. 7)

375 Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira,  
376 and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *International Conference on*  
377 *Learning Representations*, 2021. URL [https://openreview.net/forum?id=MJIve1zgr\\_](https://openreview.net/forum?id=MJIve1zgr_). (cited on p.  
378 2)

379 Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro.  
380 Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *IEEE/CVF*  
381 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4258–4267, June 2022b.  
382 URL [https://openaccess.thecvf.com/content/CVPR2022/html/Liu\\_Perturbed\\_and\\_Strict\\_](https://openaccess.thecvf.com/content/CVPR2022/html/Liu_Perturbed_and_Strict_Mean_Teachers_for_Semi-Supervised_Semantic_Segmentation_CVPR_2022_paper.html)  
383 [Mean\\_Teachers\\_for\\_Semi-Supervised\\_Semantic\\_Segmentation\\_CVPR\\_2022\\_paper.html](https://openaccess.thecvf.com/content/CVPR2022/html/Liu_Perturbed_and_Strict_Mean_Teachers_for_Semi-Supervised_Semantic_Segmentation_CVPR_2022_paper.html). (cited  
384 on pp. 3, 7, and 8)

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

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

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

Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. URL [https://openaccess.thecvf.com/content\\_CVPR\\_2020/html/Ouali\\_Semi-Supervised\\_Semantic\\_Segmentation\\_With\\_Cross-Consistency\\_Training\\_CVPR\\_2020\\_paper.html](https://openaccess.thecvf.com/content_CVPR_2020/html/Ouali_Semi-Supervised_Semantic_Segmentation_With_Cross-Consistency_Training_CVPR_2020_paper.html). (cited on pp. 3, 7, and 8)

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

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

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

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

Gyungin Shin, Weidi Xie, and Samuel Albanie. All you need are a few pixels: Semantic segmentation with pixelpick. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1687–1697, October 2021. URL [https://openaccess.thecvf.com/content/ICCV2021W/ILDAV/html/Shin\\_All\\_You\\_Need\\_Are\\_a\\_Few\\_Pixels\\_Semantic\\_Segmentation\\_With\\_ICCVW\\_2021\\_paper.html](https://openaccess.thecvf.com/content/ICCV2021W/ILDAV/html/Shin_All_You_Need_Are_a_Few_Pixels_Semantic_Segmentation_With_ICCVW_2021_paper.html). (cited on pp. 4 and 15)

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

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

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

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

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

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

He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J. Guibas. 3D IoU Match: 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](https://openaccess.thecvf.com/content/CVPR2021/html/Wang_3DIOUMatch_Leveraging_IoU_Prediction_for_Semi-Supervised_3D_Object_Detection_CVPR_2021_paper.html). (cited on p. 2)

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

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. FreeMatch: self-adaptive thresholding for semi-supervised learning. In *International Conference on Learning Representations*, 2023. URL [https://openreview.net/forum?id=PDrUPTXJI\\_A](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](https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Semi-Supervised_Semantic_Segmentation_Using_Unreliable_Pseudo-Labels_CVPR_2022_paper.html). (cited on pp. 2, 3, 4, 5, 6, 7, and 8)

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

Fan Yang, Kai Wu, Shuyi Zhang, Guannan Jiang, Yong Liu, Feng Zheng, Wei Zhang, Chengjie Wang, and Long Zeng. Class-aware contrastive semi-supervised learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14421–14430, June 2022a. URL [https://openaccess.thecvf.com/content/CVPR2022/html/Yang\\_Class-Aware\\_Contrastive\\_Semi-Supervised\\_Learning\\_CVPR\\_2022\\_paper.html](https://openaccess.thecvf.com/content/CVPR2022/html/Yang_Class-Aware_Contrastive_Semi-Supervised_Learning_CVPR_2022_paper.html). (cited on p. 1)

Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. ST++: make self-training work better for semi-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4268–4277, June 2022b. URL [https://openaccess.thecvf.com/content/CVPR2022/html/Yang\\_ST\\_Make\\_Self-Training\\_Work\\_Better\\_for\\_Semi-Supervised\\_Semantic\\_Segmentation\\_CVPR\\_2022\\_paper.html](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](https://openaccess.thecvf.com/content_ICCV_2019/html/Yun_CutMix_Regularization_Strategy_to_Train_Strong_Classifiers_With_Localizable_Features_ICCV_2019_paper.html). (cited on p. 16)

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

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

- 495 Na Zhao, Tat-Seng Chua, and Gim Hee Lee. SESS: self-ensembling semi-supervised 3D object detection.  
 496 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*,  
 497 June 2020. URL [https://openaccess.thecvf.com/content\\_CVPR\\_2020/html/Zhao\\_SESS\\_](https://openaccess.thecvf.com/content_CVPR_2020/html/Zhao_SESS_Self-Ensembling_Semi-Supervised_3D_Object_Detection_CVPR_2020_paper.html)  
 498 [Self-Ensembling\\_Semi-Supervised\\_3D\\_Object\\_Detection\\_CVPR\\_2020\\_paper.html](https://openaccess.thecvf.com/content_CVPR_2020/html/Zhao_SESS_Self-Ensembling_Semi-Supervised_3D_Object_Detection_CVPR_2020_paper.html). (cited on p.  
 499 2)
- 500 Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent  
 501 semi-supervised semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*,  
 502 pages 7273–7282, October 2021. URL [https://openaccess.thecvf.com/content\\_ICCV2021/html/](https://openaccess.thecvf.com/content_ICCV2021/html/Zhong_Pixel_Contrastive-Consistent_Semi-Supervised_Semantic_Segmentation_ICCV_2021_paper.html)  
 503 [Zhong\\_Pixel\\_Contrastive-Consistent\\_Semi-Supervised\\_Semantic\\_Segmentation\\_ICCV\\_](https://openaccess.thecvf.com/content_ICCV2021/html/Zhong_Pixel_Contrastive-Consistent_Semi-Supervised_Semantic_Segmentation_ICCV_2021_paper.html)  
 504 [2021\\_paper.html](https://openaccess.thecvf.com/content_ICCV2021/html/Zhong_Pixel_Contrastive-Consistent_Semi-Supervised_Semantic_Segmentation_ICCV_2021_paper.html). (cited on p. 7)
- 505 Zhuofan Zong, Guanglu Song, and Yu Liu. DETRs with collaborative hybrid assignments training. *arXiv*  
 506 *preprint*, November 2022. URL <https://arxiv.org/abs/2211.12860>. (cited on p. 1)
- 507 Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister.  
 508 PseudoSeg: designing pseudo labels for semantic segmentation. In *International Conference on Learning*  
 509 *Representations*, 2021. URL <https://openreview.net/forum?id=-Tw099rbVRu>. (cited on pp. 3, 6,  
 510 and 7)