SEMI-SUPERVISED SEMANTIC SEGMENTATION VIA BOOSTING UNCERTAINTY ON UNLABELED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

We bring a new perspective to semi-supervised semantic segmentation by providing an analysis on the labeled and unlabeled distributions in training datasets. We first figure out that the distribution gap between labeled and unlabeled datasets cannot be ignored, even though the two datasets are sampled from the same distribution. To address this issue, we theoretically analyze and experimentally prove that appropriately boosting uncertainty on unlabeled data can help minimize the distribution gap, which benefits the generalization of the model. We propose two strategies and design an uncertainty booster algorithm, specially for semi-supervised semantic segmentation. Extensive experiments are carried out based on these theories, and the results confirm the efficacy of the algorithm and strategies. Our plug-and-play uncertainty booster is tiny, efficient, and robust to hyperparameters but can significantly promote performance. Our approach achieves state-of-the-art performance in our experiments compared to the current semi-supervised semantic segmentation methods on the popular benchmarks: Cityscapes and PASCAL VOC 2012 with different train settings.

1 INTRODUCTION

Semantic segmentation has been a fundamental tool for various downstream applications. When deep learning methods are adopted in this area, the lack of fine-grained annotations is gradually prominent. Our paper focuses on semi-supervised semantic segmentation(Chen et al., 2021; Luo et al., 2022; 2021a;b), which learns a model with a few labeled data and excess unlabeled data. Under these settings, how to appropriately utilize unlabeled data to improve generalization becomes critical.

Notice that, even if all labeled and unlabeled data are sampled from the same distribution, there is still a non-negligible distribution gap between the two clusters of data. This is the key question we deal with in this paper. Some recent approaches have attempted to tackle this question by designing consistency regularization(Chen et al., 2021; Lee et al., 2021; Luo et al., 2021a) or evaluating unlabeled *o.o.d* (out of distribution) data via uncertainty approaches(Wang et al., 2022; Kwon & Kwak, 2022), when traditional methods always reduce the output uncertainty to get an improvement. However, we want to argue that, due to the distribution gap, boosting uncertainty on the logits of unlabeled *o.o.d* data can benefit the generalization of the model in semi-supervised semantic segmentation.

We theoretically prove that elaborately designing an uncertainty booster for the model and applying it to unlabeled data can reduce the distribution gap, which can improve the generalization of the model. After that, we propose the requirements and strategies to design a suitable uncertainty booster for segmentation. The core principle is that we should consider the original distribution of the unlabeled images. Specifically, we demonstrate two strategies of selecting proper distribution and proper *o.o.d* data.

Based on the proposed strategies, we design an uncertainty booster for semi-supervised semantic segmentation to alleviate the distribution gap between the labeled and unlabeled datasets. Our newly designed module is benefited from the following advantages:

(1) **Plug-and-play** The uncertainty booster can be used in all semi-supervised segmentation methods that require retraining pseudo labels.

(2) **Green and Efficient** There are no trainable parameters and only a few fixed parameters in the booster, which means our module imposes nearly no impact on the training speed and only takes up very little memory and training sources.

(3) **Robustness** Due to our ablation study in Table 4, our module is very robust to hyperparameters. Different hyperparameters show little influence on the promotion.

Experiments are carried out on the chosen baseline in (Chen et al., 2021). Our strategy achieves state-of-the-art performance compared to current methods on the Cityscapes(Cordts et al., 2016) and PASCAL VOC 2012 (Everingham et al.) benchmarks under various data partition protocols.

1.1 Related Work

Semi-supervised learning Semi-supervised learning has two typical paradigms: consistency regularization(Bachman et al., 2014; French et al., 2019; Sajjadi et al., 2016; Xu et al., 2021) and self-training(Lee et al., 2013; Zou et al., 2020). The derived methods focus on data augmented self-training which utilizes strong augmentation such as CutMix(Yun et al., 2019), CutOut(DeVries & Taylor, 2017), ClassMix(Olsson et al., 2021). Recent approaches pay attention to how to better release the potential of unlabeled data(Mendel et al., 2020; Ke et al., 2020; Kwon & Kwak, 2022; Wang et al., 2022), which, for example, aim to improve the quality of pseudo labels via distinguishing reliable and unreliable pseudo label(Wang et al., 2022). However, these methods do not theoretically analyze the difference between the distributions of unlabeled and labeled datasets, which is the essence of making full use of unlabeled data. In contrast, our method gives a complete analysis of this question and designs an algorithm for semi-supervised semantic segmentation.

Uncertainty in Deep learning Uncertainties can be divided into *aleatoric uncertainty* and *epistemic uncertainty*(Gal et al., 2016). The aleatoric uncertainty is also referred to *data uncertainty*, which means some of the ground truth may be incorrect. The epistemic uncertainty, referred to as model uncertainty, represents the uncertainty of the model, including whether the model parameters best explain the observed data and whether the structure best fits the data. Some classical approaches that qualify uncertainty include bayesian epistemic uncertainty estimation via dropout(Gal & Ghahramani, 2016), aleatoric uncertainty estimation via multi-network outputs(Kendall & Gal, 2017), epistemic uncertainty estimation via ensembling(Lakshminarayanan et al., 2017). While for self-training in semi-supervised semantic segmentation, the pseudo labels of unlabeled data play the role of ground truth to finetune the model, which involves both aleatoric and epistemic uncertainty. We can connect both of the uncertainties by presenting high-quality pseudo labels. Thus, in this paper, we only consider the uncertainty of unlabeled data to analyze both of the uncertainties.

2 Preliminaries

2.1 Boosting Uncertainty

In this subsection, we will briefly introduce boosting uncertainty. Many common methods adopt minimizing uncertainty as an effective strategy to reduce overfitting, yielding better performances. A simple way to minimize uncertainty is adding l_2 regularization, forcing the model to produce a convincing result. While boosting uncertainty aims to let the model output a slightly fuzzy result and change the distribution, take a simple example, if the original output is a $[0.9, 0.05, 0.05]$ for a classification model, we may modify the model to yield $[0.85, 0.075, 0.075]$ instead of the original one via boosting uncertainty. In this case, the gap between the distribution of the model output after boosting uncertainty and the original distribution is getting wider hinges on the boosting strategy we choose.

2.2 Settings

Before we describe our findings, we shall clarify the symbols and settings used in this paper. To simplify the statement, we divide the ideal training dataset distribution D into two subsets, D_L and D_U, respectively denoted as sampled distributions of the labeled and unlabeled datasets. We then denote the $i.i.d.$ sampled elements of distribution D_L as $L = \{(x_d, y_d)\}_{d=1}^{D} \sim D_L$ as the empirical distribution of D_L, in which $x \in \mathbb{R}^{k \times k}$ is from a $k \times k$-dimensional input space, $y = \{1, 2, \cdots, K\}$

where K is the number of classes. So it is with D_U. The loss we use is *0-1 loss*, defined as: $\mathcal{L}(\cdot) = 1\{h(x) \neq y\}$.

We first denote the vanilla model, which is trained only on the labeled dataset. Then we denote a posterior distribution P on hypotheses that depends on the real distribution L of the labeled dataset, parameterized by $\mathbf{W}_l$ with bounded induced norm: $P = \prod_{i=1}^{d} \mathcal{N}(\mathbf{W}_l \bar{x_l}, \mathbf{I})$, whereas $\mathbf{W}_l$ is the weights and biases of vanilla model h. In this settings, $\bar{x_l}$ is the input representations from L.

Secondly, we denote another posterior distribution Q on hypotheses that depends on dataset D_U **and** distribution P, parameterized by $\mathbf{W}_{lu}$: $Q = \prod_{i=1}^{d} \mathcal{N}(\mathbf{W}_{lu} \bar{x}_u, \mathbf{I})$, whereas $\mathbf{W}_{lu}$ is the weights and biases of h trained on L and U, $\bar{x_u}$ is the input representations from U.

We finally denote an uncertainty boosted posterior distribution on hypotheses that also depends on dataset D_U **and** distribution P: $Q_m = \prod_{i=1}^{d} \mathcal{N}((\mathbf{W}_{lu} + \mathbf{b})\bar{x}_u, \mathbf{I})$, parameterized by $\mathbf{W}_{lu}$, which is the weights and biases of h', whereas $\mathbf{b}$ indicates the distribution of uncertainty booster. By this, we define $R^E_{D_U}(h)$ as the *Expected Risk* of h applying on D_U and $R^G_{D_U}(h)$ as the *Empirical Risk* of h applying on D_U.

3 Theoretical Motivation For Boosting Uncertainty on Unlabeled Data

This section aims to figure out whether boosting uncertainty on unlabeled data can help the model improve the generalization in semi-supervised semantic segmentation. Even though the labeled and unlabeled datasets are sampled from the same distribution, there is still a non-negligible distribution gap between these two sub-distributions, which is harmful to the model to yield pseudo labels for the unlabeled dataset. This section aims to explore and give an effective solution to this question.

3.1 Boosting Uncertainty Helps Remitting Distribution Gap Between Labeled and Unlabeled Data Distributions

We will begin our derivation by considering the vanilla semi-supervised semantic model and its variant of using an uncertainty booster. Given a trained model on labeled data, we will first explore the difference between the expected and empirical risks of the two models(vanilla and the variant of using uncertainty booster) on unlabeled data.

Our theory aims to find a model that can perform well on both labeled and unlabeled datasets, which can generate a better pseudo label for the unlabeled dataset for further training. This means the model can reveal the distance between labeled and unlabeled distributions and thus have a good generalization. On this basis, the optimization goal for minimizing the distribution gaps is defined as:

$$F_2(h, h', D_L, D_U) = \min_{h'} |R^E_{D_U}(h') - R^E_{D_L}(h)| \tag{1}$$

where h' is the model that utilizes the uncertainty booster. As $R^E_{D_L}(h)$ depends on the model and labeled dataset selection, we can regard the $R^E_{D_L}(h)$ as a constant. Thus, we mainly focus on how the uncertainty booster influence $R^E_{D_U}(h')$.

Theorem 1 (McAllester, 2003) and (Germain et al., 2016) provide an upper bound of the difference of expected risk $R^E_{D_U}(h')$ and empirical risk $R^G_{D_U}(h')$ with probability of at least $1 - \delta$:

$$R^E_{D_U}(h') - R^G_{D_U}(h') \leq \sqrt{\frac{\mathbf{KL}[Q_m \| P] + \log \frac{2\sqrt{N}}{\delta}}{2N}} \tag{2}$$

where $\mathbf{KL}$ is the K-L divergence, N is the number of the samples in U.

For a semi-supervised model that utilizes the uncertainty booster, the K-L divergence in the upper bound can be calculated as:

$$\begin{aligned} \mathbf{KL}[Q_m \| P] &= \sum_{i=1}^{d} \mathbf{KL}(\mathcal{N}((\mathbf{W}_{lu} + \mathbf{b})\bar{x}_u, \mathbf{I}) \| \mathcal{N}(\mathbf{W}_l \bar{x_l}, \mathbf{I})) \\ &= d \| \mathbf{W}_l \bar{x_l} - (\mathbf{W}_{lu} + \mathbf{b})\bar{x}_u \|_2^2 \end{aligned} \tag{3}$$

Therefore, from Eq. 2 and Eq. 3, we have that:

$$R^E_{D_U}(h') \leq R^G_{D_U}(h') + \sqrt{\frac{d\|\mathbf{W}_l\bar{x}_l - (\mathbf{W}_{lu}+\mathbf{b})\bar{x}_u\|_2^2}{2N}} + \sqrt{\frac{\log\frac{2\sqrt{N}}{\delta}}{2N}} \tag{4}$$

Moreover, for the vanilla model h, the form of Eq. 4 is written as:

$$R^E_{D_U}(h) \leq R^G_{D_U}(h) + \sqrt{\frac{d\|\mathbf{W}_l\bar{x}_l - \mathbf{W}_{lu}\bar{x}_u\|_2^2}{2N}} + \sqrt{\frac{\log\frac{2\sqrt{N}}{\delta}}{2N}} \tag{5}$$

The detailed proof is presented in the appendix.

In Eq. 1, after training on labeled dataset L, the model can learn a sub-distribution on the distribution of the labeled dataset. As unlabeled data has no ground-truth labels, the *Expected Risk*: $R^E_{D_U}(h)$ for predictor h on unlabeled dataset distribution always surpass than that of $R^E_{D_L}(h)$ on labeled dataset. We will first discuss this most common situation.

Situation 1: $R^E_{D_U}(h')$ is larger than $R^E_{D_L}(h)$ In this situation, Eq. 1 turns out to minimize the upper bound of $R^E_{D_U}(h')$. This means we should minimize the RHS of Eq. 4 compared to the RHS of Eq. 5. As the input data is the same and the influence of uncertainty booster is tiny in an iteration, we can suppose $R^G_{D_U}(h)$ and $R^G_{D_U}(h')$ remain the same. So the comparison of the scales of the upper bounds in the RHS of Eq. 4 and Eq. 5 turns out to focus on the scales of $\|\mathbf{W}_l\bar{x}_l - (\mathbf{W}_{lu}+\mathbf{b})\bar{x}_u\|_2^2$

We can see that the aim of minimizing $\|\mathbf{W}_l\bar{x}_l - (\mathbf{W}_{lu}+\mathbf{b})\bar{x}_u\|_2^2$ is to keep $(\mathbf{W}_{lu}+\mathbf{b})\bar{x}_u$ closer to $\mathbf{W}_l\bar{x}_l$ from labeled dataset. While for a uniform input $\bar{x}_u$ sampled from U, as the model h has already trained on labeled dataset, the weights $\mathbf{W}_{lu}$ have a strong affinity for the distribution of labeled distribution L, thus when input *o.o.d* data in unlabeled data, $\mathbf{W}_{lu}\bar{x}_u$ will be dragged far away from original labeled dataset distribution, which incurs a much higher upper bound of *Expected Risk*, so we may push back $(\mathbf{W}_{lu}+\mathbf{b})\bar{x}_u$ to $\mathbf{W}_l\bar{x}_l$ via the appending of $\mathbf{b}$ which is an uncertainty booster that can slightly alter the distribution. Thus, if $\mathbf{b}$ is carefully designed and applied on possible unlabeled *o.o.d* data, the $R^E_{D_U}(h')$ can have a lower upper bound than the original $R^E_{D_U}(h)$. We then further analyze how to design the proposed booster.

Situation 2: $R^E_{D_U}(h)$ is smaller than $R^E_{D_L}(h)$ In this rare case, Eq. 1 turns out to be:

$$F_1(h, h', D_L, D_U) = \min_h(R^E_{D_L}(h) - R^E_{D_U}(h')) \tag{6}$$

Thus, we focus on maximizing the upper bound of $R^E_{D_U}(h)$, so in Eq. 4, we just simply introduce some random noise to increase the difference between $\mathbf{W}_l\bar{x}_l$ and $(\mathbf{W}_{lu}+\mathbf{b})\bar{x}_u$ via $\mathbf{b}$ and hence, we can minimize function F.

In all, if better selected and designed, boosting uncertainty on unlabeled *o.o.d* data may reduce the difference between the *Expected Risks* of labeled and unlabeled distributions, indicating that this strategy helps minimize the potential distribution gap between labeled distribution and unlabeled distribution.

3.2 A Theory of Designing Uncertainty Booster for Segmentation

In the last subsection, we figure out that boosting uncertainty in segmentation may help reduce the distribution gap for labeled and unlabeled distributions. At the same time, there is still a disturbing risk when rethinking Eq. 1. Since boosting uncertainty may reduce the upper bound of $R^E_{D_U}(h')$, there are questions about how much we reduce the upper bound suitable for the model. If a lousy booster is chosen, was there a catastrophic influence on the distribution? We will further discuss these several problems.

3.2.1 Conditions of Compliance when Boosting Uncertainty

In this subsection, to better understand the dense segmentation task, we will focus on the pixel distribution in the image. We will begin with how to generate an excellent pseudo label for a single image.

Theorem 2 (McAllester, 2003) observed that, let $\mathcal{H}$ be a hypothesis space, $h \in \mathcal{H}$, D_L be the labeled dataset distribution, I_U be the distribution of a single unlabeled image. $\mathrm{R}_X(h)$ is the *Expected Risk* on I_U, $\mathrm{R}_D(h)$ is the *Expected Risk* on D_L. We have:

$$\forall h \in \mathcal{H}, \mathrm{R}_X(h) \leq \mathrm{R}_D(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}\left(D_L, I_U\right) + \mu_{h^*} \tag{7}$$

whereas the $d_{\mathcal{H}\Delta\mathcal{H}}$ is denoted as empirical discrepancy distance:

$$d_{\mathcal{H}\Delta\mathcal{H}}\left(D_L, I_U\right) = 2 \sup_{(h_1,h_2)\in\mathcal{H}^2} \left| \mathop{\mathbf{E}}_{\mathbf{x}\sim D_L} Pr\left[h_1(\mathbf{x}) \neq h_2(\mathbf{x})\right] - \mathop{\mathbf{E}}_{\mathbf{x}\sim I_U} Pr\left[h_1(\mathbf{x}) \neq h_2(\mathbf{x})\right] \right| \tag{8}$$

and

$$\mu_{h^*} = \mathrm{R}_{\mathcal{S}}\left(h^*\right) + \mathrm{R}_{\mathcal{T}}\left(h^*\right), h^* = \mathrm{argmin}_{h\in\mathcal{H}}\left(\mathrm{R}_{\mathcal{S}}(h) + \mathrm{R}_{\mathcal{T}}(h)\right) \tag{9}$$

and x is the pixels in the images.

As the unlabeled dataset has no ground-truth labels, the μ_{h^*} is inaccessible, and both the labeled dataset and the unlabeled image are sampled from the same distribution D. The μ_{h^*} is assumed to be low and trivial. As the target is to minimize the discrepancy between $\mathrm{R}_X(h)$ and $\mathrm{R}_D(h)$ in Eq. 7. The aim turns out to minimize $d_{\mathcal{H}\Delta\mathcal{H}}\left(D_L, I_U\right)$. Based on (Mansour et al., 2009), we can modify Eq. 8 into:

$$d_{\mathcal{H}\Delta\mathcal{H}}\left(D_L, I_U\right) = 2 \sup_{(h_1,h_2)\in\mathcal{H}^2} \left| \mathop{\mathbf{E}}_{\mathbf{x}\sim D_L} L_r\left(h_1(\mathbf{x}), h_2(\mathbf{x})\right) - \mathop{\mathbf{E}}_{\mathbf{x}\sim I_U} L_r\left(h_1(\mathbf{x}), h_2(\mathbf{x})\right) \right| \tag{10}$$

Where L_r can be a general real-valued loss. If L_r is a L_2 loss:

$$\begin{aligned} d_{\mathcal{H}\Delta\mathcal{H}}\left(D_L, I_U\right) &= 2 \sup_{(h_1,h_2)\in\mathcal{H}^2} \left| \mathop{\mathbf{E}}_{\mathbf{x}\sim D_L} \left[\left(h_1(x) - h_2(x)\right)^2\right] - \mathop{\mathbf{E}}_{\mathbf{x}\sim I_U} \left[\left(h_1(x) - h_2(x)\right)^2\right] \right| \\ &= 2 \sup_{(h_1,h_2)\in\mathcal{H}^2} \left| \sum_{\mathbf{x}\in\left(D_L \cup I_U\right)} \left[D_L(\mathbf{x}) - I_U(\mathbf{x})\right] \left[\left(h_1(x) - h_2(x)\right)^2\right] \right| \end{aligned} \tag{11}$$

In which we can see, $\left[\left(h_1(x) - h_2(x)\right)^2\right]$ depends on $\left(D_L \cup I_U\right)$ which is sampled from the support of D which we could hardly control and is greater than or equal to 0, so minimizing difference of $\mathrm{R}_X(h)$ and $\mathrm{R}_D(h)$ depends on minimizing $\left[D_L(\mathbf{x}) - I_U(\mathbf{x})\right]$. This means for each pixel in unlabeled images, the difference between the pixel distribution of the output of the unlabeled images and the prior distribution of the unlabeled images should be as close as possible.

3.2.2 Two Strategies for Designing an Uncertainty Booster for Segmentation

Strategy 1: The Criteria for Distribution Imitation Based on Theorem 2, the uncertainty boosted output should have a similar distribution to the prior distribution of the image. From an intuitive perspective, the pixel distributions are various in different images, so we shall focus on image-wise distribution for each pixel when boosting uncertainty for a segmentation model. In a nutshell, *the distribution of uncertainty booster that we apply to each image in the segmentation model shall be subject to the distribution of the image itself.*

Strategy 2: The Criteria for Data Selection In addition to considering the selection criteria of the distribution, we shall also consider the scale that we boost uncertainty. We should focus on the unlabeled *o.o.d* data relative to the prior output distribution. We first proposed that if a particular model h is trained on a known sampled distribution, we test it on a sampled data point. If the trained model yields a higher uncertainty on the sampled data point, the more likely this data point is out of distribution from the known distribution; thus, we need to give this data point a more significant

disturbance. Based on Theorem 2, we raise the second template for designing the uncertainty booster: *The scale of boosting uncertainty for data depends on the scale of the uncertainty that the model generates on the data.* This means that the model should pay more attention to the uncertain data points. The more the model is unconfident, the more we boost the uncertainty.

4 UNCERTAINTY BOOSTER MODULE (UBM)

In this section, based on the two strategies proposed in section 3.2.2, we design a plug-and-play uncertainty booster module (UBM) specialized for semi-supervised semantic segmentation. In the meantime, our proposed module requires negligible extra memory or computation while achieving noticeable performance gain for segmentation models.

4.1 REGIONAL UNCERTAINTY VOTER (RUV)

We aim to find a proper distribution to boost pixel-level uncertainty according to Strategy 1. As mentioned above, commonly used Gaussian or Uniform Distributions applied to the whole dataset may fail in semantic segmentation because the distribution of a specific image varies from one another. The nature of semantic consistency renders a pixel closely related to its adjacent pixels in an image. Thus, imposing non-regional perturbation on pixels prohibits the model from learning the actual distribution. As such, we design the uncertainty booster on an image-wise case-by-case basis.

We consider taking regional information into account and propose the *Regional Uncertainty Voter* (RUV) to produce a customized artificial distribution. Given a one-hot pseudo label $p^{oh} \in \mathbb{R}^{h\times w\times K}$ of an image $x^{h\times w}$, where K is the number of classes.

We count the number of pixels belonging to each class in the $h_v \times w_v$ vicinity V of every individual pixel x_i, $i \in [0, hw-1]$, by which pixels can The module can perceive and aggregate unique regional information in the image, which is done by a specifically defined kernel.

Then we divide the counting result map by the cardinality of V to yield a probability map $C \in \mathbb{R}^{h\times w\times K}$.

Compared with the universal Gaussian or Uniform Distribution, our region-aware distribution is more natural and sensible to impose. We formulate the calculation of C in Eq. 12:

$$C_j = \frac{\sum_{k=0}^{K-1} \text{weight}(j,k) \star p_k^{oh}}{|V|}, \quad \text{where weight}(j,k) = \begin{cases} \mathbf{1}^{h_v\times w_v} & \text{if } j=k \\ \mathbf{0}^{h_v\times w_v} & \text{else} \end{cases} \text{ and } C_j \in C \quad (12)$$

where $\star$ is the valid 2D cross-correlation operator, $C_j \in \mathbb{R}^{h\times w}$ is the probability map of class j, weight(j,k) maps the k th layer (k th class) of p^{oh} to C_j. Note that weight(j,k) is a constant kernel for gathering neighbor predictions. The voter in Eq. 12 can be efficiently computed by `Conv2d` with our pre-designed kernel weight and is free from back-propagation, rendering neglectable computation cost.

4.2 UNCERTAINTY ADAPTIVE STRATEGY (UAS)

Our uncertainty booster is required to be careful and smart. Intuitively, boosting uncertainty wildly would negatively impact performance because correct and certain distribution that already learned is likely to be deviated by the booster. Therefore, the selection criterion is crucial and should be tailored for each pixel of every image at every single state of training. To address this issue, we propose calculating the confidence value, `Conf`, based on the entropy of each pixel, which decides how strong the booster should be for the corresponding pixel. Conventionally, we regard pixels with great `Conf` value as well-classified ones, where extra uncertainty is unnecessary. While those with small `Conf` values are expected to be unconfident *o.o.d* pixels, thus, a strong booster is needed. To sum up, assume $pred \in \mathbb{R}^{h\times w\times K}$ to be the prediction probabilistic map of the model, we define `Conf` of pixel $x_i \in x$ in Eq. 13 and the normalized adaptive weight $W_{x_i} \in W_x$ of pixel x_i for the pseudo label p^{oh} in Eq. 14:

$$\texttt{Conf}(x_i) = \sum_{k=0}^{K-1} pred_{i,k} \ \log(pred_{i,k}) \quad (13)$$

$$W_{x_i} = \frac{\texttt{Conf}(x_i) - \min \texttt{Conf}(x)}{\max \texttt{Conf}(x) - \min \texttt{Conf}(x)},\ W_{x_i} \in W_x \tag{14}$$

In all, given a vanilla pseudo label p^{oh}, based on the RUV and UAS, we can define the uncertainty boosted pseudo label $\hat{p}_i$ in Eq. 15:

$$\hat{p} = \text{UBM}(p^{oh}) = W_x * p^{oh} + (1 - W_x) * C,\ p^{oh} = Onehot(Pred) \tag{15}$$

The proposed pipeline is shown in Fig. 1; the proposed UBM module is simple, low-parameters, and efficient. After the UBM module finishes processing the vanilla probability map $Pred$, we use the output pseudo labels $\hat{p}$ for further training.

4.3 Overall Strategy

We follow CPS (Chen et al., 2021) as the baseline, which consists of two independent models, namely $f(x;\theta_1)$ and $f(x;\theta_2)$. The two models have the same network structure and loss definition but different initialization. For labeled data L, both $f(x;\theta_1)$ and $f(x;\theta_2)$ are trained by CrossEntropy (CE) loss with ground truth; For unlabeled data U, the models generate pseudo labels p_1^{U} and p_2^{U} for each other as the ground truth of CE loss.

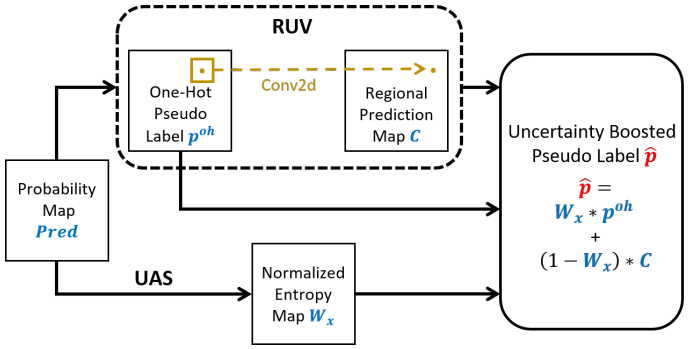


Figure 1: The pipline of UBM.

We only apply the proposed module UBM on the two pseudo labels p_1^{U} and p_2^{U}. Finally, the overall loss function for $f(x;\theta_1)$ is defined in Eq. 16, vice versa for $f(x;\theta_2)$.

$$\mathcal{L}_1 = \frac{1}{|L|}\sum_{x\in L} \text{CE}(p, y) + \lambda \frac{1}{|U|}\sum_{x\in U} \text{CE}(p_1^{\text{U}}, \text{UBM}(p_2^{\text{U}})) \tag{16}$$

where p is the probabilistic output of labeled data, y is the ground truth of labeled data, p_1^{U} is the probabilistic output of the unlabeled data of this model, p_2^{U} is the one-hot output of the same unlabeled data of the other model, UBM is the proposed uncertainty booster module. λ is the trade-off weight between the two CE losses.

5 Experiments

Datasets & Evaluation PASCAL VOC 2012 dataset is a standard object-centric semantic segmentation dataset, which contains 20 foreground classes and one background class. We follow previous works and adopt the augmented set (Hariharan et al., 2011) with 10,582 training images and 1449 validation images. Cityscapes (Cordts et al., 2016) is a dataset for urban scene understanding, consisting of 2,975 training images with fine-annotated labels and 500 validation images. For both datasets, training images are split under label ratios of 1/16, 1/8, 1/4, and 1/2, respectively. We directly adopt all the split partitions provided by CPS (Chen et al., 2021).

We use the mean Intersection-over-Union (mIoU) metric to evaluate the segmentation performance. We report results on PASCAL VOC 2012 val set and Cityscapes val set for all label ratios. Following (Wang et al., 2022), for PASCAL VOC 2012, we center crop images to a fixed resolution; as for Cityscapes, we use sliding-window evaluation.

Implementation We use ResNet-101(He et al., 2016) pretrained on ImageNet(Deng et al., 2009) as our backbone and DeepLabv3+(Chen et al., 2018) as the segmentation head. Following CPS (Chen et al., 2021), we add a deep stem block to our backbone, remove the last down-sampling operations, and employ dilated convolutions in the subsequent convolution layers. In addition, we use mini-batch

SGD with the momentum of 0.9 and a weight decay of 0.0001 to train our model. As in CPS (Chen et al., 2021), we train PASCAL VOC 2012 for 60 epochs with $\lambda = 1.5$ and Cityscapes with OHEM loss for 240 epochs with $\lambda = 6$, where the learning rates are 0.0025 and 0.005, respectively. For simplicity, the size of the vicinity V in UBM is set to 5×5 in all experiments.

5.1 Comparison with Existing Alternatives

We compare our method with the current state-of-the-art methods, including DCC(Lai et al., 2021), ST++(Yang et al., 2022), U^{2}PL(Wang et al., 2022) etc. Experiments are carried out via the same network architecture. Notice that we don't use CutMix(Yun et al., 2019), which is a powerful augmentation, and we still achieve state-of-the-art performance.

Results on PASCAL VOC 2012 Table 1 compares our proposed method with state-of-the-art methods on PASCAL VOC 2012 dataset. Compared to the baseline, our uncertainty booster steadily promotes the performance, achieving impressive improvements of **+1.41, +1.42, +1.75** and **+2.00**, respectively under 1/16, 1/8, 1/4, 1/2 partition protocols. Compared to state-of-the-art methods, our method outperforms nearly all the methods in all settings.

Results on Cityscapes Table 2 presents the comparison with state-of-the-art methods on the Cityscapes dataset. The UBM brings about **+2.03, +0.58, +2.78**, and **+2.60** of improvements under 1/16, 1/8, 1/4, 1/2 partition protocols compared to baseline. Also, our method outperforms all the state-of-the-art methods.

Table 1: Comparison with state-of-the-art methods on PASCAL VOC 2012. All the methods are based on DeepLabv3+ and ResNet-101 Backbone. All other results are referred from ST++ (Yang et al., 2022)

Method	CutMix	1/16(662)	1/8 (1323)	1/4(2646)	1/2(5291)
Supervised		66.30	70.60	73.10	77.21
CutMix (French et al., 2019)	✓	71.66	75.51	77.33	78.21
GCT (Ke et al., 2020)	✗	67.20	72.50	75.10	77.40
DCC (Lai et al., 2021)	✗	72.40	74.60	76.30	–
ST (Yang et al., 2022)	✓	72.90	75.70	76.40	–
ST++ (Yang et al., 2022)	✓	**74.50**	76.30	76.60	–
CPS (Chen et al., 2021)	✗	72.18	75.83	77.55	78.64
CPS+UBM	✗	**73.59**$^{+1.41}$	**77.25**$^{+1.42}$	**79.30**$^{+1.75}$	**80.64**$^{+2.00}$

Table 2: Comparison with state-of-the-art methods on Cityscapes. All methods are based on DeepLabv3+ and ResNet-101 Backbone. All other results are referred from U^{2}PL (Wang et al., 2022).

Method	CutMix	1/16(186)	1/8(372)	1/4(744)	1/2(1488)
Supervised		65.74	72.53	74.43	77.83
CutMix(French et al., 2019)	✓	67.06	71.83	76.36	78.25
GCT(Ke et al., 2020)	✗	66.75	72.66	76.11	78.34
DCC (Lai et al., 2021)	✗	–	69.70	72.70	77.50
RCC (Zhang et al., 2022)	✓	–	74.04	76.47	–
U^{2}PL(Wang et al., 2022)	✓	70.30	74.37	76.47	79.05
CPS(Chen et al., 2021)	✗	69.78	74.31	74.58	76.81
CPS+UBM	✗	**71.81**$^{+2.03}$	**74.89**$^{+0.58}$	**77.36**$^{+2.78}$	**79.41**$^{+2.60}$

5.2 Ablation Studies

All the ablation studies are carried out on PASCAL VOC 2012 dataset with labeled ratio 1/4, with DeepLabv3+ and ResNet-101 backbone, the size of the vicinity is set to 15×15.

The Effectiveness of Components in Uncertainty Booster To further analyze the effective portions of our methods, we separately introduce Uniform Distribution instead of RUV and remove UAS. The results are shown in Table 3 (Van. indicates Vanilla; UD indicates Uniform Distribution booster). We can see that if we add UAS for each unlabeled image, there will be an increase of **+1.49**. But if we add a Uniform Distribution booster and UAS to boost uncertainty, there will be a remarkable decrease in mIoU with **-6.39**. This proves that the distribution of pixels in different images is remarkably different. When we add both the UAS and RUV, we achieve the highest performance of **+1.93**. That is because RUV catches non-local distributions of the pixel, thus can better generate a more similar distribution to the input unlabeled image distribution.

Ablation Study on The Size of The Vicinity We also ablate the vicinity size of the 2D cross-correlation operator. As table 4 shows, the results remain almost the same, which proves that if we focus on image-wise distribution for each pixel when boosting uncertainty for the model, the hyperparameters count for little influence on the model.

Table 3: Ablations on Two Strategies

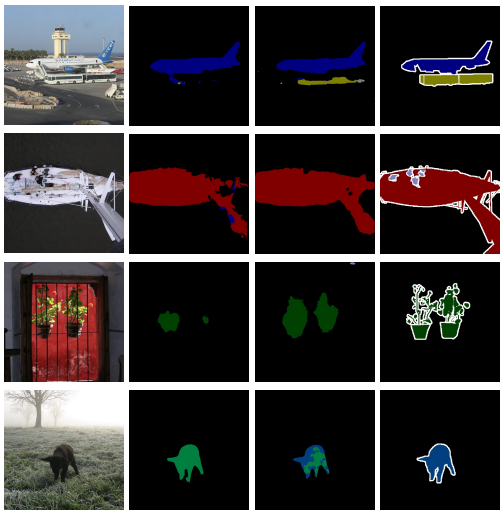
	Van.	UAS	UD/UAS	RUV/UAS
mIoU	77.55	79.04	72.91	**79.48**

Table 4: Ablations on Different Vicinity Sizes

Size	3×3	5×5	9×9	15×15
mIoU	79.45	79.30	79.40	**79.48**

5.3 Qualitative Results

The qualitative results tested on 1/8 labeled data of PASCAL VOC are presented in Figure 2. Our method outperforms the baseline in many scenarios. We get a more complete and accurate segmentation result rather than baseline method, and incorrect semantics can be corrected by nearby semantic information. More results are presented in the supplementary material.



(a) Input (b) Baseline (c) UBM (d) GT

Figure 2: Qualitative results on PASCAL VOC 2012 val set.

6 Conclusion and Future Outlook

In this paper, we theoretically and experimentally propose that boosting uncertainty on unlabeled data helps with the generalization of the model in semi-supervised semantic segmentation. We demonstrate two advanced strategies to design a novel uncertainty booster. The first strategy aims to map the uncertainty-boosted output closer to the prior labeled output of the model. The second strategy proposes that the model should pay more attention to the uncertain data points, which means the more the model is unconfident, the more we boost the uncertainty of the data points. Following the theoretical strategies, we design a plug-and-play module that does not need any training. Our module makes the old baseline method outperform the current methods on PASCAL VOC 2012 and Cityscapes via different partition protocols without increasing too much training cost. Our work can trigger the research interest in the distribution gap and inspire more work on developing uncertainty methods in semi-supervised learning.

REFERENCES

Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2613–2622, 2021.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Yarin Gal et al. Uncertainty in deep learning. 2016.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A new pac-bayesian perspective on domain adaptation. In *International conference on machine learning*, pp. 859–868. PMLR, 2016.

Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pp. 991–998. IEEE, 2011.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *European conference on computer vision*, pp. 429–445. Springer, 2020.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Donghyeon Kwon and Suha Kwak. Semi-supervised semantic segmentation with error localization network. *arXiv preprint arXiv:2204.02078*, 2022.

Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1205–1214, 2021.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896, 2013.

Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4071–4080, 2021.

Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. pp. 8801–8809, 2021a.

Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting Zhang. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 318–329, 2021b.

Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. 2022.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

David McAllester. Simplified pac-bayesian margin bounds. In *Learning theory and Kernel machines*, pp. 203–215. Springer, 2003.

Robert Mendel, Luis Antonio de Souza, David Rauber, João Paulo Papa, and Christoph Palm. Semi-supervised segmentation based on error-correcting supervision. In *European Conference on Computer Vision*, pp. 141–157. Springer, 2020.

Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1369–1378, 2021.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4248–4257, 2022.

Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, pp. 11525–11536. PMLR, 2021.

Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4268–4277, 2022.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Jianrong Zhang, Tianyi Wu, Chuanghao Ding, Hongwei Zhao, and Guodong Guo. Region-level contrastive and consistency learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:2204.13314*, 2022.

Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020.

A APPENDIX

For the vanilla model h, the upper bound of $R^E_{D_U}(h) - R^G_{D_U}(h)$ turns out to be:

$$R^E_{D_U}(h) - R^G_{D_U}(h) \leq \sqrt{\frac{\mathbf{KL}[Q\|P] + \log\frac{2\sqrt{N}}{\delta}}{2N}} \tag{17}$$

Thus, for the upper bound of Eq. 17, the K-L divergence is:

$$\begin{aligned} \mathbf{KL}[Q\|P] &= \sum_{i=1}^{d} \mathbf{KL}(\mathcal{N}(\mathbf{W}_{lu}\bar{x}_u, \mathbf{I})\|\mathcal{N}(\mathbf{W}_l\bar{x}_l, \mathbf{I})) \\ &= d\|\mathbf{W}_l\bar{x}_l - \mathbf{W}_{lu}\bar{x}_u\|_2^2 \end{aligned} \tag{18}$$

Thus, for the vanilla model h, the form of Eq. 17 is written as:

$$R^E_{D_U}(h) \leq R^G_{D_U}(h) + \sqrt{\frac{d\|\mathbf{W}_l\bar{x}_l - \mathbf{W}_{lu}\bar{x}_u\|_2^2}{2N}} + \sqrt{\frac{\log\frac{2\sqrt{N}}{\delta}}{2N}} \tag{19}$$

which is shown in the main body of the paper.