Progressive Bayesian Inference for Scribble-Supervised Semantic Segmentation

Chuanwei Zhou, Chunyan Xu*, Zhen Cui

PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education,
Jiangsu Key Lab of Image and Video Understanding for Social Security,
School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.
{cwzhou, cyx, zhen.cui}@njust.edu.cn

Abstract

The scribble-supervised semantic segmentation is an important yet challenging task in the field of computer vision. To deal with the pixel-wise sparse annotation problem, we propose a Progressive Bayesian Inference (PBI) framework to boost the performance of the scribble-supervised semantic segmentation, which can effectively infer the semantic distribution of these unlabeled pixels to guide the optimization of the segmentation network. The PBI dynamically improves the model learning from two aspects: the Bayesian inference module (i.e., semantic distribution learning) and the pixel-wise segmenter (i.e., model updating). Specifically, we effectively infer the semantic probability distribution of these unlabeled pixels with our designed Bayesian inference module, where its guidance is estimated through the Bayesian expectation maximization under the situation of partially observed data. The segmenter can be progressively improved under the joint guidance of the original scribble information and the learned semantic distribution. The segmenter optimization and semantic distribution promotion are encapsulated into a unified architecture where they could improve each other with mutual evolution in a progressive fashion. Comprehensive evaluations of several benchmark datasets demonstrate the effectiveness and superiority of our proposed PBI when compared with other state-of-the-art methods applied to the scribble-supervised semantic segmentation task.

Introduction

Semantic segmentation, which refers to achieving accurate dense pixel-wise class prediction of an image, is a fundamental computer vision task (Krähenbühl and Koltun 2011; Chen et al. 2017). It serves many other tasks such as multi-task learning (Cui et al. 2022; Zhou et al. 2020), intelligent diagnostic (Falk et al. 2019), video segmentation (Xu et al. 2021a; Zhou et al. 2021) etc. For now, great progress has been achieved due to the rapid development of deep segmentation networks. However, sufficient training of deep networks requires a great amount of fully annotated segmentation masks, which suffer great labeling burdens for pixel-wise annotations. Towards relieving the heavy reliance on highly costly annotations, scribble-supervised semantic segmentation resorts to arbitrarily drawn lines to train a satisfactory segmentation model. The scribbles are much easier to obtain and provide sparse annotations which could indicate the rough location of the semantic regions to guide the segmenter learning.

The scribble-supervised semantic segmentation task is formed as interactive segmentation in the early stage, and it is usually solved by utilizing graphical models to build inter-pixel or inter-region relationships to expand the scribbles towards those unknown regions (Rother, Kolmogorov, and Blake 2004; Grady 2006). When it comes to the neural network era, some methods attempt to combine graphical models and deep neural networks to produce better segmentation. Pan et al. perform a random walk process in the deep features to decrease the representation uncertainty to produce more confident segmentation. NormalCut (Tang et al. 2018a) and KernelCut (Tang et al. 2018b) utilize the graph cuts to build extra regularization terms to constrain the segmenter learning for more stable segmentation results. The graphical methods demand specific designs for certain scenarios and they are hard to deploy in practical applications. In addition, BPG (Wang et al. 2019) introduces a pre-trained edge detection network to provide complementary boundary supervision to regulate the segmenter learning, but it will inevitably introduce extra information from other datasets. Despite directly relying on the original scribbles to supervise the network update, other works resort to inferring pseudo labels in the unlabeled regions to mine more supervision signals. ScribbleSup (Lin et al. 2016) generates pseudo labels in the whole image by jointly considering the scribbles and prediction probabilities to solve a CRF (Lin et al. 2016) optimization objective. RAWKS (Vernaza and Chandraker 2017) develops a random walk process to generate pseudo labels in the unlabeled regions to train a better segmenter by utilizing the hitting probabilities as label transmission matrices. A2GNN (Zhang et al. 2021b) introduces a graph neural network to learn to expand the supervision signals from the scribbles to unlabeled regions. PSI (Xu et al. 2021b) learns adaptive thresholds with a network module imposed on the features and probabilities to generate the pseudo labels. All existing pseudo-label methods only utilize the prediction uncertainties of those unlabeled pixels to generate hard pseudo labels, but they give insufficient consideration to the feature distribution correlations of the labeled and unlabeled regions

*Chunyan Xu is the corresponding author

during the label inference process.

In this paper, we propose a progressive Bayesian inference (PBI) framework which infers the semantic distribution of these unlabeled data to provide auxiliary supervision. The segmentation network could be progressively optimized with the semantic distributions which are inferred from the scribbles by mining the feature distribution correlations between the labeled and unlabeled regions to boost the segmentation performance. Specifically, we design a Bayesian inference module to effectively learn the semantic distribution of these unlabeled/unobserved data in the feature space, which formulates it as learning under the situation of the partially observed data. In particular, to optimize the network parameters of the Bayesian inference module, we use the Bayesian expectation maximization strategy to estimate the posterior probability distribution of these unlabeled data by performing the maximum likelihood estimation in an iterative solution. To achieve more robust semantic distributions of these unlabeled data, the Bayesian inference module is then used to perform distribution learning with the guide of the estimated probability information. Subsequently, we utilize both the scarce scribble annotations as well as the learned semantic distributions to guide the segmenter learning. The proposed PBI builds an iterative optimization process by exploiting the inferred semantic distribution, then dynamically improves the model learning from two aspects: the Bayesian inference module (i.e., semantic distribution mining) and the pixel-wise segmenter (i.e., model updating). The segmenter optimization and semantic label distribution promotion are encapsulated into a unified architecture where the two parts could improve each other with mutual evolution in a progressive fashion. Extensive experiments have demonstrated that our proposed PBI could boost the performance of the scribble-supervised semantic segmentation and state-of-the-art segmentation performances have been achieved in standard benchmarks.

To summarize, our major contributions are as follows: i) We propose a novel progressive Bayesian inference (PBI) framework to boost the performance of the scribble-supervised semantic segmentation, where the semantic distribution is effectively inferred under the situation of partially observed data, and then adopted to guide the segmenter learning in an iterative manner. ii) We specifically design a Bayesian inference module for learning robust semantic distribution with a probabilistic inference net, which is trained by the estimated Bayesian posterior distribution of class-wise features. iii) We conduct extensive experiments to validate the effectiveness of the proposed method on the scribble-supervised semantic segmentation task and report state-of-the-art performances on the PASCAL VOC 2012 dataset (Everingham et al. 2010) and the PASCAL Context dataset (Hariharan et al. 2011).

Related Works

Scribble-supervised semantic segmentation The scribble-supervised semantic segmentation task arises from the urgent demand of effective deep segmentation network training with little annotation cost. Some methods designed specific loss functions to constrain the network training. For example, NormalCut (Tang et al. 2018a) and KernelCut (Tang et al. 2018b) resolved to the CRF (Krähenbühl and Koltun 2011) criterion to derive topology-constrained optimization objectives. URNE (Pan et al. 2021) resorted to a siamese structure to design a self-supervised constraint for the network optimization to encourage consistent segmentation. Different from them, the proposed PBI does not rely on any specific regularization terms but the normal ones. BPG (Wang et al. 2019) brought in extra information of an auxiliary edge detection network to provide boundary supervision for more accurate segmentation prediction. In contrast, our proposed PBI framework resorts to no extra sources except for a clustering assumption which is usually utilized for the segmentation tasks to produce complementary information. Other methods produced pseudo labels to mine more supervision signals in the unlabeled regions for segmenter learning. ScribbleSup (Lin et al. 2016) built up a deep framework that generated a hard pseudo map over the whole image region according to the original scribble as well as a CRF (Krähenbühl and Koltun 2011) model to train the segmentation network. RAWKS (Vernaza and Chandraker 2017) utilized the random-walk model to propagate the scarce scribbles to those unlabeled regions by treating the hitting probabilities as label transmission matrices. A2GNN (Zhang et al. 2021b) learned an affinity attention graph neural network to infer hard pseudo labels in the unlabeled regions. PSI (Xu et al. 2021b) proposed to learn dynamic thresholds via an auxiliary module to generate hard one-hot pseudo labels for the segmentation network training. All the above methods utilized the hard pseudo labels which solely rely on the pixel prediction uncertainties of the unlabeled pixels to train the segmenter. In contrast, our PBI induces semantic distributions for those unannotated data by building a Bayesian inference module to explore the feature distribution correlations between the labeled and unlabeled regions in the feature space.

Pseudo label learning The pseudo label is commonly adopted in tasks where enough supervision is lacking. Some methods directly generated hard pseudo-labels to make up for the loss of sufficient training supervision. Lee et al. (Lee et al. 2013) first introduced the pseudo label method into the deep network by selecting the classes which achieved the maximum probability for the unlabeled data to produce fake supervisions to relieve the supervision deficiency. FixMatch (Sohn et al. 2020) regulated the model training by generating pseudo labels based on the weakly augmented samples to supervise the strongly augmented counterparts. FlexMatch (Zhang et al. 2021a) generated adaptive pseudo-labels by calculating one customized threshold for each class with a curriculum learning framework. Some other methods generated probabilistic distributions as supervisions for the unlabeled samples to facilitate sufficient network training. For example, DLDL (Gao et al. 2017) generated the label distributions by minimizing its distance to the rough ground-truth labels to overcome the label ambiguity. PENCIL (Yi and Wu 2019) produced label distributions by regarding them as learnable parameters to defeat the label noises. Although they also produce the label distributions, the proposed PBI differs greatly from them as they update the label distribution probabilities as learnable variables while our semantic label distributions are inferred from a specifically introduced

probabilistic inference net.

The Proposed Method

Overview As shown in Fig. 1, we propose a progressive Bayesian inference (PBI) framework to perform the scribble-supervised semantic segmentation, where the segmenter could be progressively optimized by the learned semantic distribution. The whole architecture is mainly composed of two parts: the segmenter learning and the Bayesian inference module, which are implemented in an alternating manner. Given an image X as the input sample, we can predict the pixel-wise segmentation probability P by employing a segmenter with the encoder-decoder architecture: $P = f_{\text{seg}}(X, \Phi)$, and Φ is the segmenter parameters. The convolutional features F of the segmenter are fed into the probabilistic inference net Ψ to induce the semantic label distribution Q, i.e., $Q = f_{\text{infer}}(F, \Psi)$. The network parameters of the segmenter can be learned under the joint supervisions provided by the original scribble information S as well as the inferred semantic label distributions Q, i.e., $\mathcal{L}_{ce}(P, S) + \alpha\mathcal{L}_{ld}(P, Q)$. In the Bayesian inference module, we use the Bayesian expectation maximization strategy to devise the semantic probability distribution $\widehat{Q}$, which is then adopted to facilitate the effective learning of the probabilistic inference net Ψ by using the supervision function $\mathcal{L}_{bi}(Q, \widehat{Q})$. The Bayesian expectation-maximization process can comprehensively consider the segmentation prediction P, the convolutional feature F, as well as the scribble information S to derive the auxiliary supervision $\widehat{Q}$ in a Bayesian manner. The segmenter Φ and probabilistic inference net Ψ are optimized in an alternating way to form a close-looping learning framework, among which the two nets could be promoted each other for mutual evolution.

Bayesian Expectation Maximization For $\widehat{Q}$ In the case of sparse annotation, the key is how to effectively mine information from unlabeled data to guide the network learning process. Different from existing work that generates hard pseudo-labels of unlabeled samples based on the prediction uncertainties, we utilize the feature distribution correlations of the labeled and unlabeled regions to mine the semantic distribution of unlabeled data in the learning process. In particular, given an image X as the input sample, we use D and U to represent the labeled and unlabeled pixels, i.e., $X = \langle D, U\rangle$. We try to find the approximating distribution $\widehat{Q}$ of unlabeled data by constructing with a known parametric distribution $\mathcal{P}$ from the labeled regions. To deal with this problem, we minimize the relative entropy (also known as Kullback-Leibler divergence, KL) between two probability distributions to solve $\widehat{Q}$:

$$\arg\min_{\widehat{Q}} KL(\widehat{Q}||\mathcal{P}). \tag{1}$$

The relative entropy formulation can be also expanded as the expectation of the logarithmic difference between the two probabilities $\widehat{Q}$ and $\mathcal{P}$:

$$\begin{aligned} KL(\widehat{Q}||\mathcal{P}) &= \boldsymbol{E}_{\widehat{Q}}\left[\log \widehat{Q}/\mathcal{P}\right] \\ &= \boldsymbol{E}_{\widehat{Q}}[\log \widehat{Q}] - \boldsymbol{E}_{\widehat{Q}}[\log \mathcal{P}]. \end{aligned} \tag{2}$$

We can then use the Bayesian posterior estimation method to get the distribution $\mathcal{P}$ over these unlabeled pixels by applying this to the case of learning from partially observed data (i.e., scribble annotations):

$$\mathcal{P} = \widetilde{\mathcal{P}}/\mathcal{Z}. \tag{3}$$

Here the distribution $\mathcal{P}$ is built over these unlabeled pixels U (i.e., $\mathcal{P}(U|D, \Theta)$), where the observed data D and the corresponding parameters Θ are fixed now. $\mathcal{Z}$ is the distribution over these labeled data $\mathcal{P}(D|\Theta)$ and $\widetilde{\mathcal{P}}$ is the joint probability $\mathcal{P}(U, D|\Theta)$. Let $l(\Theta : \langle D, U\rangle)$ denote the log-likelihood of the parameters Θ with respect to all completed pixels. The logarithm of the joint probability can be rewritten as $\log\widetilde{\mathcal{P}} = l(\Theta : \langle D, U\rangle)$. By using the entropy definition $\boldsymbol{E}_{\widehat{Q}}[\log \widehat{Q}] = -\boldsymbol{H}_{\widehat{Q}}(U)$, the relative entropy can be rewritten as:

$$\begin{aligned} KL(\widehat{Q}||\mathcal{P}) &= -\boldsymbol{H}_{\widehat{Q}}(U) - E_{\widehat{Q}}[\log\widetilde{\mathcal{P}}] + E_{\widehat{Q}}[\log\mathcal{Z}] \\ &= -\boldsymbol{H}_{\widehat{Q}}(U) - E_{\widehat{Q}}[l(\Theta : \langle D, U\rangle)] + \log\mathcal{Z}. \end{aligned} \tag{4}$$

Note that the term $\log\mathcal{Z}$ does not depend on $\widehat{Q}$. Hence, minimizing the relative entropy $KL(\widehat{Q}||\mathcal{P})$ is equivalent to maximizing the following energy functional:

$$\arg\max_{\widehat{Q},\Theta} \boldsymbol{H}_{\widehat{Q}}(U) + \boldsymbol{E}_{\widehat{Q}}[l(\Theta : \langle D, U\rangle)], \tag{5}$$

where the first term $\boldsymbol{H}_{\widehat{Q}}(U)$ is the entropy value of the semantic distribution $\widehat{Q}$ over these unlabeled pixels, and the second term $\boldsymbol{E}_{\widehat{Q}}[l(\Theta : \langle D, U\rangle)]$ is the expected log-likelihood relative to $\widehat{Q}$. Here we can use the common Gaussian mixture model to model the distribution of these labeled data, i.e., $\Theta = \{(\mu_{c,k}, \Sigma_{c,k})|c = 0, 1, \cdots, C; k = 1, 2, \cdots, K\}$, and C and K separately represent the total number of the semantic classes and the number of mixtures for each class. $\mu_{c,k}$ and $\Sigma_{c,k}$ are the mean and diagonal covariance matrix for the k-th mixture of class c. The solution of the objective (i.e., Eqn. (5)) leads to an expectation-maximization iteration which alternates between optimizing the parameters Θ and semantic distribution $\widehat{Q}$.

In the E-step, we fix the semantic distribution $\widehat{Q}$ to maximize the log-likelihood w.r.t the parameters Θ. The class mixture parameters are separately estimated in the corresponding class regions $V_q^c = \{(x, y)|\hat{q}_{x,y} > \tau^c\}$ where τ^c is a confidence threshold and we set it to 0.97 for the foreground classes (i.e., $c = 1, 2, \cdots, C$) and 0.6 for the background class (i.e., $c = 0$) as in the previous work (Kolesnikov and Lampert 2016). Considering the feature representation capability of the trained segmenter, we also determine a sampling region $V = \{V^c|c = 0, 1, \cdots, C\}$ to sample stable features, and $V_f^c = \{(x, y)|\,\|x - x^c\|_1 \le r, \|y - y^c\|_1 \le r, \forall(x^c, y^c) \in D^c\}$, where D^c denotes the scribble regions for the c-th class with r being a sampling radius. Hence the sampling region V^c is an intersection of the two regions (i.e., $V^c = V_q^c \cap V_f^c$), and the mixture model parameters $\mu_{c,k}$ and $\Sigma_{c,k}$ are then solved following the conventional Gaussian mixture model solution with a k-means clustering initialization. In the M-step, we fix the mixture model parameters to

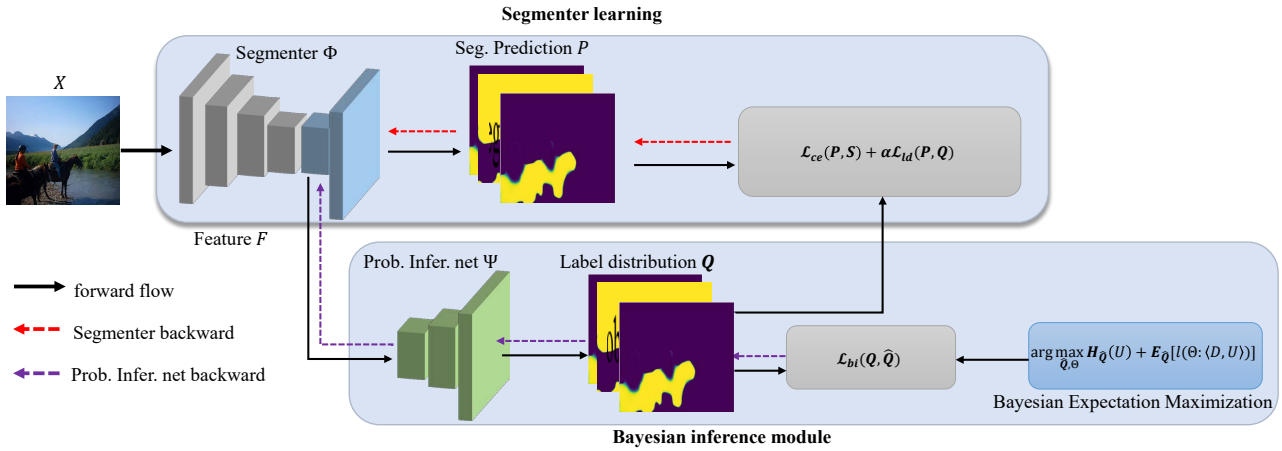

Figure 1: An overview of the proposed PBI. The whole architecture is mainly composed of two complementary components: the segmenter learning and the Bayesian inference module, and the two parts are implemented in an alternating way for co-evolution. For more details, please refer to the manuscript.

solve the semantic distribution $\widehat{Q}$. It is exactly to estimate the class portion in each pixel p with the currently estimated mixture parameters. The distribution value q_p^c of $\widehat{Q}$ in pixel p for the class c are calculated via:

$$\hat{q}_p^c = \frac{\max_k \mathcal{N}(f_p|\mu_{c,k},\Sigma_{c,k})}{\sum_{c=0}^{C}\max_k \mathcal{N}(f_p|\mu_{c,k},\Sigma_{c,k})}, \tag{6}$$

where $\mathcal{N}(f|\mu,\Sigma)$ is the probability density of the corresponding feature f w.r.t the Gaussian distribution with mean μ and covariance matrix Σ.

The above EM-iteration requires a good initialization, and thus we directly use the segmenter estimated probability as an initialization for $\widehat{Q}$. Hence the produced distribution $\widehat{Q}$ is a Bayesian posterior of the semantic distributions, and it effectively mines the correlation between the labeled and unlabeled feature distributions with the built EM iterations. The estimated posterior distribution $\widehat{Q}$ is then utilized to optimize the probabilistic inference net Ψ. To embrace the details as well as the class semantics, we concatenate the low and high level features to form the feature F. Considering the evolved representation capability of the gradually trained segmenter, we progressively enlarge the sampling radius r by 20 every 20 epochs. As a consequence, a larger sampling region could envelop more feature variations, and in turn, the resulting semantic distribution could produce more stable feature representations. Our experiment results have demonstrated that the progressive Bayesian posterior estimation process could generate useful semantic label distribution for better learning the probabilistic inference net, and then improve the prediction capability of the segmenter to boost the segmentation performance.

PBI Learning Φ And Ψ The goal of the proposed PBI framework is to explore the semantic distribution of unlabeled pixels for learning a better segmenter. The overall PBI training process is composed of two alternating optimization steps: the segmenter learning and the Bayesian inference module. During segmenter learning, the input image X is first forwarded into the segmenter Φ to produce the segmentation prediction P. Meanwhile, the convolutional features of the input image F are also fed into the probabilistic inference net Ψ to learn a more robust semantic distribution Q. Then the semantic distribution Q along with the original scribbles S are jointly utilized as the supervision signal to guide the segmenter update:

$$\begin{aligned}\Phi &\leftarrow \Omega(\Phi, \mathcal{L}_{seg}(P,Q,S)),\\ &= \Omega(\Phi, \mathcal{L}_{seg}(f_{\text{seg}}(X,\Phi),Q,S)),\end{aligned} \tag{7}$$

where $\mathcal{L}_{seg}$ is the segmentation loss function, Ω is an optimizer. The segmentation loss function $\mathcal{L}_{seg}$ considers both the original scribble annotation S and the inferred semantic label distribution Q as follows:

$$\mathcal{L}_{\text{seg}}(P,Q,S) = \mathcal{L}_{ce}(P,S) + \alpha\mathcal{L}_{ld}(P,Q), \tag{8}$$

where $\mathcal{L}_{ce}$ is the commonly used cross-entropy between P and S for these labeled pixels. $\mathcal{L}_{ld}$ refers to the KL-divergence $KL(Q\|P)$ for making P approximate Q in the unlabeled areas, and it also includes the entropy loss as a constraint. α is a factor to balance the effects from scribbles S and the explored semantic distributions Q, and we linearly grow it from 0.0 to 0.5 in the beginning 100 epochs and it is fixed in the following epochs.

The Bayesian inference module is to train the probabilistic inference net Ψ to produce a robust global approximation of the semantic distribution by fixing the segmenter parameters Φ. The probabilistic inference net Ψ consists of three cascaded convolutional layers as well as a softmax activation layer, and it takes the intermediate segmentation features F as input to produce the semantic distribution Q. We utilize the Bayesian posterior $\widehat{Q}$ induced by the developed Bayesian posterior estimator to update the net Ψ following:

$$\begin{aligned}\Psi &\leftarrow \Omega(\Psi, \mathcal{L}_{bi}(Q,\widehat{Q})),\\ &= \Omega(\Psi, \mathcal{L}_{bi}(f_{\text{infer}}(F,\Psi),\widehat{Q})),\end{aligned} \tag{9}$$

where $\mathcal{L}_{bi}$ is the Bayesian inference loss to make a robust approximation of the estimated posterior distribution $\widehat{Q}$. Since

both Q and $\widehat{Q}$ are probability distributions, we utilize a robust variant of the KL divergence as:

$$\mathcal{L}_{bi} = KL_{robust}(Q\|\widehat{Q}), \tag{10}$$

$$= \sum_{u\in U}\sum_{c=0}^{C} \widehat{q}_u^c \log\left(a + b\cdot q_u^c\right), \tag{11}$$

where q_u^c and $\widehat{q}_u^c$ are the values of Q and $\widehat{Q}$ for the class c at pixel u, $a = \frac{0.4}{C-1}$ and $b = 1 - C\cdot a$ following (Larsen et al. 1998). It is also possible to directly utilize the estimated posterior distribution $\widehat{Q}$, instead of the learned probability Q, as the auxiliary supervision. However, $\widehat{Q}$ is closely relevant to P due to the EM iteration, and directly using $\widehat{Q}$ as the pseudo-labels is prone to introduce accumulated errors. Hence, the probabilistic inference net Ψ is specifically designed to learn a more robust semantic distribution Q as the global approximation for the segmenter optimization. Our experimental results have also demonstrated the effectiveness of the introduced probabilistic inference net.

The segmenter learning and Bayesian inference module are optimized alternately to mine better segmentation feature representations and semantic distributions. Our experimental results have demonstrated superior performances of the proposed method over the hard pseudo-label methods, verifying the effectiveness of the proposed progressive Bayesian inference framework. In the testing stage, we directly input the testing images into the trained segmenter Φ to produce the segmentation predictions, and the Bayesian expectation maximization, as well as the probabilistic inference net, are no longer required for the segmentation mask inference.

Experiment Results

Experiment Settings

Following the standard protocol (Zhang et al. 2021b; Xu et al. 2021b; Pan et al. 2021), we utilize the PASCAL VOC 2012 semantic segmentation dataset (Everingham et al. 2010) and the PASCAL Context dataset (Mottaghi et al. 2014) to evaluate our proposed PBI framework. We adopt the DeepLabV3+ (Chen et al. 2018) with backbone ResNet101 as the segmenter Φ. We follow the conventional model initialization protocol as previous work (such as ScribbleSup (Lin et al. 2016), PSI (Xu et al. 2021b)). The segmenter is initialized from the ImageNet pre-trained ResNet and the Prob. Infer. net is initialized from scratch following the 'Kaiming_normal' strategy. The SGD optimizer with momentum and weight decay being 0.9 and 5e-4 is adopted as the optimizer Ω to train the networks. The learning rate is initially set to 1e-4 and then slowly decayed with a 'poly' schedule, and the whole framework is trained for 200 epochs with a batch size of 8. For the first 100 epochs, the sampling region V remains in the original scribbles S. Starting from the 100-th epoch, we progressively expand the sampling region V with a radius r of 21 every 20 epochs. The Gaussian mixture number K is set to 3 empirically. In the optimization process, we implement random augmentations including scaling ($[0.5, 2.0]$), flipping (p=0.5), rotation ($[-10, 10]$) and cropping (512×512). The multi-scale, as well as the flipping strategies, are adopted during the testing phase, but no CRF (Adams, Baek, and Davis 2010) post-processing is utilized which is the same as in other methods (Wang et al. 2019; Xu et al. 2021b). All the experiments are implemented with the PyTorch framework (Paszke et al. 2019). Following the previous literature, we adopt the mean Intersection-over-Union (mIoU) score as our evaluation metric.

Method	Sup.	Backbone	mIoU(%)
DeepLab	F	ResNet101	76.8
TreeFCN	F	ResNet101	80.9
MCOF	I	ResNet101	60.3
AffinityNet	I	ResNet38	58.4
ICD	I	VGG16	64.0
IAL	I	VGG16	62.0
BoxSup	B	VGG16	62.0
SDI	B	VGG16	65.7
ScribbleSup*	S	VGG16	63.1
RAWKS	S	ResNet101	59.5
NormalCut	S	ResNet101	72.8
KernelCut	S	ResNet101	73.0
BPG	S	ResNet101	73.2
PSI	S	ResNet101	74.9
URNE	S	ResNet101	76.1
A2GNN	S	TreeFCN	76.2
PBI (Ours)	S	ResNet101	**77.2**

Table 1: Comparison with state-of-the-art methods on the PASCAL VOC 2012 validation set. 'F','I','B' and 'S' separately mean the full supervision, the image-level tags, the boxes and the scribbles. The symbol '*' means CRF post-processing.

Method	Sup.	mIoU(%)
PSPNet	F	47.8
DANet	F	52.6
OCR	F	56.2
BoxSup	Semi	40.5
ScribbleSup*	S	36.1
RAWKS	S	36.0
DeepLabV3+	S	37.1
PSI	S	43.1
PBI (Ours)	S	**43.7**

Table 2: Comparison with state-of-the-art methods on the PASCAL Context dataset. 'F' and 'S' separately mean the full supervisions and the scribbles. The symbol '*' means the CRF post-processing.

Comparison With State-of-the-art Methods

PASCAL VOC 2012: We first compare our proposed PBI with other state-of-the-art methods on the PASCAL VOC 2012 dataset (Everingham et al. 2010). The detailed results have been listed in Table 1. Our proposed PBI achieves the best performance of 77.2% among all the scribble-supervised

segmentation methods. Our PBI exceeds those regularization based methods including NormalCut (Tang et al. 2018a), KernelCut (Tang et al. 2018b) and URNE (Pan et al. 2021) separately by 4.4%, 4.2% and 1.1%. It may be that our mined semantic distribution could implicitly provide better segmentation learning regularization. Our method outperforms BPG (Wang et al. 2019) which introduces extra boundary supervisions by 4.0%, and it indicates that our proposed method is able to make up for the loss of accurate edge information. It may be that the proposed multi-Gaussian-based Bayesian posterior estimation could embrace a great number of class variations including those edge pixels. Finally, when compared with PSI (Xu et al. 2021b) or A2GNN (Zhang et al. 2021b) which produce hard pseudo labels, our method achieves performance improvements of 2.3% and 1.0%. It verifies the effectiveness of our mined semantic distributions in training better segmentation networks. It's worth noting that our PBI could bridge the performance gap between scribble-supervised methods and fully supervised methods (Wu, Shen, and Van Den Hengel 2019; Song et al. 2019) by a large margin, where it even outperforms the WiderResnet (Wu, Shen, and Van Den Hengel 2019) by 0.4%. It further validates the effectiveness of our proposed PBI in training better segmentation networks.

PASCAL Context: We additionally conduct experiments on the more challenging PASCAL Context dataset (Hariharan et al. 2011) to further evaluate the generalization capability of our PBI. The detailed results have been listed in Table 2. All the compared scribble-supervised methods including ours obey the pure scribble-supervision principle. But the Box-sup method has adopted the fully annotated masks in the PASCAL Context dataset (Hariharan et al. 2011) along with all the bounding-box annotated images from the PASCAL 2007 (Everingham et al. 2010) dataset. It could be observed that among all the compared scribble-supervised methods, our proposed PBI achieves the best performance of 43.7%, excelling the second-best method PSI (Xu et al. 2021b) which generates hard pseudo labels by 0.6%. It is obviously a non-trivial improvement and shows that our method could generalize to more complicated datasets, which verifies the effectiveness of the semantic distributions mined by our PBI. It is also worth noting that our proposed PBI could further narrow the performance gap between the fully supervised and the weakly supervised methods, further validating the effectiveness of the PBI framework.

Ablation Studies

All ablation studies are conducted on the PASCAL VOC 2012 dataset.

We first conduct experiments to evaluate the effectiveness of our adopted parts and the experiment results are listed in Table 3. We first train a baseline solely by the original scribbles, and it achieves a mIoU of 70.1%. When the hard label is generated as extra supervision signals (i.e. Hard One-hot Label), the performance is boosted by 3.0% to 73.1%. Afterward, the probabilistic inference net is introduced to generate the semantic distributions. We first compute the Bayesian posterior based on the original scribbles and only use the induced class posterior distribution to finetune the segmenter.

Hard One-hot Label	✓					
Bayesian posterior estimation		✓	✓			
Prob. Infer. net. learning				✓		✓
Progressive sampling			✓			✓
mIoU(%)	70.1	73.1	73.6	76.4	74.5	77.2

Table 3: The performances of our method with different parts enabled on the PASCAL 2012 validation set.

We then utilize the estimated Bayesian posterior $\widehat{Q}$ to finetune the segmenter (Bayesian posterior estimation), and it obtains a performance of 73.6%. It surpasses the hard pseudo label variant by 0.5% and verifies the effectiveness of our induced Bayesian posterior estimation. The probabilistic inference net (Prob. infer. net learning) is further introduced to learn robust semantic distribution, and it boosts the performance by 0.9%. It verifies that the introduced probabilistic inference net could produce more robust semantic distributions for segmenter learning. We further conduct an experiment based on the above two semantic distribution variants by implementing the progressive sampling region growing strategy for the posterior estimation (i.e., Progressive sampling), and the performances further grow by 2.8% and 2.7%. The non-trivial performance increment clearly shows that the proposed progressive Bayesian posterior inference procedure could effectively promote semantic label distribution mining to train better segmentation networks.

We have also plotted the utilized supervision signals including the hard pseudo labels as well as the learned semantic distributions (including the maximum classes and probabilities) in Fig. 2. Our learned semantic label distributions are able to better capture the semantic object occupations than the hard labels. It also demonstrates that the mined semantic distributions could decrease the label uncertainties when compared with the hard labels for instance the sheep in the first row and the plant in the second row. It vividly shows the superiority of the semantic distributions learned in PBI.

We conduct experiments to analyze the adopted low and high level features and the Gaussian mixture number K, and the results are listed in Table 4. When only adopting the low-level features, the model struggles to achieve satisfactory performances since the class feature clusters highly overlap and much harmful information will be introduced. If only use the high-level features, the performances will be greatly boosted because the highly abstracted deep semantic features could provide accurate feature clustering for most pixels. When both the low and high level features are jointly adopted, it could further boost the segmentation and achieve satisfactory performances. It is in that the hybrid feature could provide complementary features for accurate class clusters and boundaries. The performance saturates at a moderate mixture number $K = 3$ which could effectively balance the feature variations and noises.

We evaluate the effectiveness of the progressive progress of the Bayesian posterior estimation, and the results are listed in Table 5 with growing sampling radii and changing iteration numbers. When only sampling feature vectors in the

a)image b)GT c)scribble d)psuedo label e)learned label distri. f)max-prob map

Figure 2: Visual comparison between the hard pseudo label and the learned semantic distribution.

Feature	K=1	K=2	K=3	K=4	K=5
low	68.6	70.1	70.3	70.5	70.4
high	74.7	75.8	76.4	76.3	76.0
low+high	76.1	76.6	77.2	77.1	77.0

Table 4: The mIoU(%) on the PASCAL VOC 2012 (Everingham et al. 2010) validation set with different level features as well as different mixture numbers K for Bayesian posterior estimation.

Radius r	0	21	41	61	81	101	121	141
Iteration	80	100	120	140	160	180	200	220
mIoU(%)	75.4	76.1	76.6	76.9	77.0	77.2	77.1	76.8

Table 5: The performances on the PASCAL VOC 2012 (Everingham et al. 2010) validate set with growing radius r for Bayesian posterior estimation.

original scribble regions (iteration=80) to learn the semantic distributions, the proposed method could achieve a mIoU of 75.4%. The performance keeps increasing when the radius grows from 21 to 101 and it saturates at 'r=101' since a growing sampling region could embrace more class feature variations. When the radius further grows from 101 to 141, the performance drops a little which may be that the accumulated noises interfere with the segmenter training.

We conduct experiments to analyze the time consumption for each training stage and the inference time, and the time is averaged over all the images. The detailed time consumption comparison results have been listed in Table 6. In the training process, the majority of time consumption is on the gradient backward process, where the segmenter backward requires updating 116 layers with 76.9ms while the Prob. infer. net backward only needs to update 9 layers with 2.5ms. The optimization for the Prob. infer. net only takes about 6% of the whole training time (5.3ms *vs* 86.7 ms) for one training sample in each iteration, and the Bayesian EM iteration for posterior estimation consumes 13.5ms per sample. Hence

Stages	SF	SB	EM	PF	PB	TF
Time (ms)	4.5	76.9	13.5	2.8	2.5	1.7

Table 6: The consumed time for different stages of PBI. 'SF', 'SB' 'EM', 'PF', 'PB', and 'TF' separately represent the segmenter forward, segmenter backward, Bayesian EM iteration, Prob. infer. net forward, Prob. infer. net backward and testing forward stages.

our introduced Bayesian inference module will not bring too many extra computation burdens. In the testing stage, we abandon the Bayesian inference module and segmenter back-propagation, so that the segmentation prediction is very fast with 1.7 ms/sample, which is as efficient as other methods.

Conclusion

In this paper, a novel progressive Bayesian inference (PBI) framework is proposed to boost the scribble-supervised semantic segmentation performances by inferring the semantic probability distribution of the unlabeled data. The semantic label distribution is induced for segmenter optimization by a specifically introduced probabilistic inference network which could be well learned in the Bayesian inference module. We derive the Bayesian posterior by mining the feature correlations of the labeled and unlabeled data so that effective optimization of the inference network could be realized. The PBI dynamically improves the model learning from two aspects: the pixel-wise segmenter (i.e., model updating) and Bayesian inference module (i.e., semantic label distribution mining), where the two parts are encapsulated into a close-looping optimization process for mutual promotion. Our experiments have demonstrated the effectiveness of the proposed PBI framework and state-of-the-art results have been reported in the PASCAL VOC 2012 dataset and the PASCAL Context dataset. In the future, we might consider applying the proposed PBI framework to other weakly-supervised tasks.

Acknowledgments

The authors would like to thank all reviewers for their instructive comments. This work was supported by the National Science Fund of China (Grant Nos. 61972204 and 62072244), the Natural Science Foundation of Shandong Province (Grant No. ZR2020LZH008), and in part by State Key Laboratory of High-end Server & Storage Technology.

References

Adams, A.; Baek, J.; and Davis, M. A. 2010. Fast high-dimensional filtering using the permutohedral lattice. In *Computer Graphics Forum*, 753–762.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. 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.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, 801–818.

Cui, Z.; Zhou, L.; Wang, C.; Xu, C.; and Yang, J. 2022. Visual Micro-Pattern Propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 303–338.

Falk, T.; Mai, D.; Bensch, R.; Çiçek, Ö.; Abdulkadir, A.; Marrakchi, Y.; Böhm, A.; Deubner, J.; Jäckel, Z.; Seiwald, K.; et al. 2019. U-Net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 67–70.

Gao, B.-B.; Xing, C.; Xie, C.-W.; Wu, J.; and Geng, X. 2017. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6): 2825–2838.

Grady, L. 2006. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11): 1768–1783.

Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; and Malik, J. 2011. Semantic contours from inverse detectors. In *Proceedings of the IEEE International Conference on Computer Vision*, 991–998.

Kolesnikov, A.; and Lampert, C. H. 2016. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Proceedings of the European Conference on Computer Vision*, 695–711.

Krähenbühl, P.; and Koltun, V. 2011. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*, 109–117.

Larsen, J.; Nonboe, L.; Hintz-Madsen, M.; and Hansen, L. K. 1998. Design of robust neural network classifiers. In *ICASSP*, 1205–1208.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 896.

Lin, D.; Dai, J.; Jia, J.; He, K.; and Sun, J. 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3159–3167.

Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.-G.; Lee, S.-W.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 891–898.

Pan, Z.; Jiang, P.; Wang, Y.; Tu, C.; and Cohn, A. G. 2021. Scribble-Supervised Semantic Segmentation by Uncertainty Reduction on Neural Representation and Self-Supervision on Neural Eigenspace. In *Proceedings of the IEEE International Conference on Computer Vision*, 7416–7425.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 8026–8037.

Rother, C.; Kolmogorov, V.; and Blake, A. 2004. " GrabCut" interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 309–314.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, 596–608.

Song, L.; Li, Y.; Li, Z.; Yu, G.; Sun, H.; Sun, J.; and Zheng, N. 2019. Learnable tree filter for structure-preserving feature transform. In *Advances in Neural Information Processing Systems*, volume 32.

Tang, M.; Djelouah, A.; Perazzi, F.; Boykov, Y.; and Schroers, C. 2018a. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1818–1827.

Tang, M.; Perazzi, F.; Djelouah, A.; Ben Ayed, I.; Schroers, C.; and Boykov, Y. 2018b. On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European Conference on Computer Vision*, 507–522.

Vernaza, P.; and Chandraker, M. 2017. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7158–7166.

Wang, B.; Qi, G.; Tang, S.; Zhang, T.; Wei, Y.; Li, L.; and Zhang, Y. 2019. Boundary Perception Guidance: A Scribble-Supervised Semantic Segmentation Approach. In *International Joint Conference on Artificial Intelligence*, 3663–3669.

Wu, Z.; Shen, C.; and Van Den Hengel, A. 2019. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 119–133.

Xu, C.; Wei, L.; Cui, Z.; Zhang, T.; and Yang, J. 2021a. Metavos: Learning to adapt online target-specific segmentation. *IEEE Transactions on Image Processing*, 4760–4772.

Xu, J.; Zhou, C.; Cui, Z.; Xu, C.; Huang, Y.; Shen, P.; Li, S.; and Yang, J. 2021b. Scribble-Supervised Semantic Segmentation Inference. In *Proceedings of the IEEE International Conference on Computer Vision*, 15354–15363.

Yi, K.; and Wu, J. 2019. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7017–7025.

Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; and Shinozaki, T. 2021a. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems*, volume 34.

Zhang, B.; Xiao, J.; Jiao, J.; Wei, Y.; and Zhao, Y. 2021b. Affinity attention graph neural network for weakly supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, C.; Xu, C.; Cui, Z.; Zhang, T.; and Yang, J. 2021. Self-Teaching Video Object Segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 1623–1637.

Zhou, L.; Cui, Z.; Xu, C.; Zhang, Z.; Wang, C.; Zhang, T.; and Yang, J. 2020. Pattern-structure diffusion for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4514–4523.