Journal of Machine Learning for Biomedical Imaging 2024:006 vol. 2, pp. 717–760 Submitted 8/2023
Published 4/2024

Deep Monte-Carlo EM for Semantic Segmentation using Weakly-and-Semi-Supervised Learning Using Very Few Expert Segmentations

Akshay V. Gaikwad akshayg@cse.iitb.ac.in
Computer Science and Engineering (CSE) Department, Indian Institute of Technology (IIT) Bombay, Mumbai, India.

Suyash P. Awate suyash@cse.iitb.ac.in
Computer Science and Engineering (CSE) Department, Indian Institute of Technology (IIT) Bombay, Mumbai, India.

Abstract

Typical methods for semantic image segmentation rely on large training sets comprising per-pixel semantic segmentations. In medical-imaging applications, obtaining a large number of expert segmentations can be difficult because of the underlying demands on the experts' time and the budget. However, in many such applications, it is much easier to obtain image-level information indicating the class labels of the objects of interest present in the image. We propose a novel deep-neural-network (DNN) framework for the *semantic segmentation* of images relying on *weakly-and-semi-supervised* learning from a training set comprising (i) *very few images having per-pixel semantic segmentations* and (ii) all images having class labels for the objects of interest present within. To enable weakly-and-semi-supervised learning, our framework proposes to couple the tasks of semantic segmentation and image classification by incorporating a semantic-segmenter DNN followed by a translator DNN with end-to-end learning. We propose variational learning relying on *Monte-Carlo expectation maximization*, infering a posterior distribution on the hidden variable that models the segmenter-DNN's latent space. We propose a Metropolis-Hastings *sampler* for the posterior distribution, along with sample *reparametrizations* to enable end-to-end back-propagation. Results on three publicly available real-world microscopy datasets show the benefits of our framework over existing methods, along with empirical insights into the workings of various approaches.

Keywords: Semantic segmentation , Monte-Carlo EM , variational learning , Metropolis-Hastings sampling , weakly-and-semi-supervised learning.

1. Introduction

In the field of medical image analysis, semantic image segmentation is an important task with widespread applications in digital pathology, radiology, and endoscopy. Such applications include diagnosis, radiotherapy, and image-guided intervention. Semantic image segmentation involves predicting class-label probabilities for every pixel in the image. In the broader field of image analysis, key approaches to semantic image segmentation have relied on several principles including (i) automatic feature extraction from images (Yao et al., 2012; Mostajabi et al., 2015), (ii) hierarchical conditional random fields (CRFs) (Ladicky et al., 2009; Lempitsky et al., 2011; Vemulapalli et al., 2016; Chandra and Kokkinos, 2016) that hierarchically model the local and global dependencies between the pixel labels, (iii) region proposals (Arbelaez et al., 2012; Li et al., 2013; Ren et al., 2015; He et al., 2017; Hariharan

https://doi.org/10.59275/j.melba.2024-2fgd

et al., 2014, 2015) that are ranked based on the target class labels, (iv) spatial pyramid pooling (He et al., 2015; Zhao et al., 2017; Chen et al., 2018), (v) scale-aware attention (Chen et al., 2016), and (vii) hierarchical feature learning including encoder-decoder based fully convolutional deep neural networks (DNNs) like SegNet (Badrinarayanan et al., 2017), U-Net (Ronneberger et al., 2015), FCN (Long et al., 2015), DeepLabV3 (Chen et al., 2018), and FRRN (Pohlen et al., 2017). Within the field of medical image analysis, key approaches to semantic or binary image segmentation include (i) the one by Ciresan et al. (2012), (ii) U-Net (Ronneberger et al., 2015; Ö Çiçek et al., 2016) and its extensions (Xu et al., 2021; Fu et al., 2019) based on feature fusion, and (iii) VA MaskR-CNN (Wang et al., 2019) that extends MaskR-CNN (He et al., 2017) to 3D volumes.

Closely related to the image-segmentation task are the tasks of image classification and object detection/recognition. Image classification involves predicting one or more class labels associated with the image. Popular methods for image classification include AlexNet (Krizhevsky et al., 2012), VGG-Net (Simonyan and Zisserman, 2015), ResNet (He et al., 2016), and DenseNet (Huang et al., 2017). Object detection/recognition often involves predicting a bounding box as well as a class label for every object of interest present in the image. Popular methods for object detection/recognition include Faster R-CNN (Ren et al., 2015), YOLO9000 (Redmon and AFarhadi, 2017), SSD (Liu et al., 2016), and M2Det (Zhao et al., 2019). The knowledge of the image class, or the classes of objects in the image, can help semantic image segmentation.

Typical methods for semantic image segmentation rely on a training set with per-pixel semantic segmentations associated with all images. For instance, learning a MaskR-CNN (He et al., 2017) requires a large training set of images along with bounding-box segmentations, per-pixel object segmentations within each bounding box, and object-class labels for each bounding box. Similarly, typical methods using fully convolutional DNNs, e.g., SegNet (Badrinarayanan et al., 2017), U-Net (Ronneberger et al., 2015), FCN (Long et al., 2015), DeepLabV3 (Chen et al., 2018), and FRRN (Pohlen et al., 2017), rely on large expert-segmented training sets and full supervision.

For medical applications, it is often infeasible to obtain (i) a large training set, because of unavailability of data that can be made available for studies, and (ii) a sufficiently large number of expert segmentations, because of limitations on experts' time and budgets. A typical DNN learned from a small training set and even fewer expert-segmented images is unable to learn well, leading to large errors in segmentation as well as object detection/recognition. On the other hand, many medical applications involve images with objects of interest (e.g., a cell or an organ) where the image-level information about the types/classes of objects is readily available within typical clinical workflows/protocols, and at a small fraction of expert time. More specifically, in the case of blood-tissue miscroscopy images the class-label information of the form of (i) type of WBCs, (ii) classes of nuclei or (iii) level of infection in RBCs is readily available.

This paper makes novel contributions. We propose a novel DNN framework for *semantic segmentation* of blood-tissue microscopy images relying on *weakly-and-semi-supervised* learning from small training sets comprising (i) *very few images having per-pixel semantic segmentations* and (ii) all images having class labels for the objects of interest present within. To enable weakly-and-semi-supervised learning, our framework proposes to couple the tasks of semantic segmentation with image classification, with end-to-end learning; this coupled

framework first outputs a (probabilistic) semantic segmentation that is then concatenated with the observed image and passed as input to a translator DNN to produce (probabilistic) class labels indicating the classes of objects present in the image. One of the important goals of our method is to infer the uncertainty during inference, in both the segmentation and recognition. Capturing uncertainty is especially important in challenging training scenarios, e.g., where models train using very small amounts of data and/or supervision. This requires the modeling of a distribution on output segmentations/probability-maps and output classifications/probability-vectors during inference/deployment; such a distribution gives a family of outputs rather than a single output. Furthermore, Bayesian/variational modeling can, in addition to modeling uncertainty, also improve model learning itself (Wells et al., 1996; Allassonniere et al., 2010). We propose a variational learning framework relying on *Monte-Carlo expectation maximization* (MCEM), infering a posterior distribution on the hidden variable modeling the segmenter DNN's latent space. During the learning phase, we design a *Metropolis-Hastings* (MH) sampler for the posterior distribution, along with sample *reparametrizations* to enable end-to-end backpropagation. By the design of this posterior distribution, the segmenter receives additional information about the objects present in the image based on the class-label data associated with the image, thereby enabling weak supervision. During inference on test images, our variational framework can inform about the uncertainty associated with the probabilistic per-pixel semantic segmentation and the probabilistic image-level classification. We employ three publicly available real-world microscopy datasets to show the benefits of our framework over existing approaches, along with empirical insights into the workings of various methods.

The rest of the paper is organized as follows. Section 2 describes the related works including the recent methods that rely on weakly-and-semi-supervised learning. Section 3 describes our novel DNN framework (Section 3.1) for semantic segmentation of images, which relies on coupling semantic segmentation with image classification to leverage the weakly-labeled images; Section 3.2 describes the segmentation framework and Section 3.3 describes the classification framework. Section 3.4 describes the novel MCEM-based variational-learning framework. Section 3.5 describes the MH sampling algorithm and the reparametrization scheme. Section 3.6 describes the inference strategy on test data. Section 4 describes three publicly available real-world datasets for evaluation, with results and empirical insights comparing our method with existing methods. Section 5 concludes the paper.

2. Related Work

This section describes related works in semantic image segmentation relying on semi-supervised learning, weakly-supervised learning, weakly-and-semi-supervised learning, and variational learning.

Semi-supervised learning refers to learning instances where the training set has a subset for which all types of ground-truth annotations are available (e.g., in our context, image-level class labels as well as per-pixel semantic segmentations) and the remaining subset that is devoid of any kind of annotation. Typical methods for semi-supervised learning for semantic segmentation (Mittal et al., 2019; Hung et al., 2018; Souly et al., 2017) rely on a segmenter and a discriminator, where the latter discriminates between the distribution of predicted segmentations and the distribution of ground-truth segmentations available in

the training set. The discriminator then aids in learning to segment images for which the ground-truth segmentation is unavailable. However, learning such a discriminator usually requires large training sets of images with per-pixel segmentations (e.g., tens of thousands of segmented images), which may be unavailable in many clinical/scientific application scenarios. Some methods (Mittal et al., 2019; Hung et al., 2018) may counter the limited size of the training subset having per-pixel segmentations by assuming the availability of a pre-trained discriminator, but pre-training a reliable discriminator itself will need a large training set. In contrast, our framework focuses on learning semantic segmentation from significantly smaller training sets (e.g., a few tens of segmented images) by leveraging weakly-supervised learning relying on image-level class labels indicating the classes of the objects of interest present in the image.

Weakly-supervised learning for image segmentation refers to learning from ground-truth information that is devoid of per-pixel segmentations, but comprises other annotations that typically have lesser information. For instance, many weakly-supervised methods for semantic segmentation (Li et al., 2021; Lee et al., 2021; Wang et al., 2020; Zhang et al., 2020a; Chang et al., 2020; Fan et al., 2020; Chen et al., 2020; Zhang et al., 2019; Ahn and Kwak, 2018) use image-level class labels along with the class-activation-map (CAM) approach of Zhou et al. (2016) to estimate semantic segmentations for the images. Our framework assumes that (i) a fraction of the training set has images with per-pixel segmentations available, whereas (ii) the rest of the training set has images with only class labels available but no per-pixel segmentations available. In addition, our framework does *not* rely on CAM which itself requires a classifier to be pre-trained using a large training set.

This section, so far, has described (i) learning in the *purely semi-supervised* scheme: where a subset of the training-set images have per-pixel segmentations and the remaining images are devoid of any kind of annotations, and (ii) learning in the *purely weakly-supervised* setting: where none of the training-set images have per-pixel segmentations and all of the images have the class labels available. Learning in the *weakly-and-semi-supervised* setting is essentially a combination of these two aforementioned cases, i.e., where a subset of the training-set images have per-pixel segmentations and all of the images in the training set have class labels. Curating such a training set takes limited expert time, and such datasets are readily available through typical clinical workflows. There are many instances of such methods, e.g., some using CAMs (Lee et al., 2021; Yao and Gong, 2020; Ahn et al., 2019; Hong et al., 2015; Lee et al., 2021; Ouali et al., 2020; Lee et al., 2019; Wei et al., 2018), and others (Pan et al., 2022; Zhou et al., 2019; Xu et al., 2015; Papandreou et al., 2015).

For weakly-and-semi-supervised segmentation, one group of methods (Lee et al., 2021; Yao and Gong, 2020; Hong et al., 2015; Ouali et al., 2020; Ahn et al., 2019; Lee et al., 2019; Wei et al., 2018) leverage class labels through a pre-trained classifier/discrminator, to estimate missing segmentations. For instance, the methods in Yao and Gong (2020) and Hong et al. (2015) use a DNN pair, where there is (i) a primary DNN that predicts class labels, followed by (ii) another branched DNN that outputs a segmentation by leveraging the CAMs (corresponding to the predicted class labels) to restrict the search space during semi-supervised segmentation. These methods evaluate on datasets where the object class informs on the object shape that helps in object segmentation, unlike some of the medical datasets in this paper where the shape of objects is similar across classes. Unlike these methods, our end-to-end-learned DNN first produces a probabilistic semantic segmentation

that is then combined with the observed image to produce class labels for the underlying objects of interest. The other methods (Lee et al., 2021; Ouali et al., 2020; Ahn et al., 2019; Lee et al., 2019; Wei et al., 2018) incorporate CAMs to simulate a segmentation for those training-set images that are devoid of ground-truth segmentations, where the CAMs are derived from an independently learned DNN for image classification. However, it is well known that CAMs often differ significantly from the ground-truth segmentations (Kang et al., 2021; Dunnmon et al., 2019; Wei et al., 2018; Lu et al., 2020), leading to many false positives and false negatives. Although some methods like Wei et al. (2018); Lu et al. (2020) attempt to enhance the CAMs using prior information (e.g., on smoothness and size), they are typically unable to estimate high-quality segmentations. Unlike these CAM-based methods, (i) we leverage class labels to infer a posterior distribution on the semantic segmentations using a joint framework for segmentation and image classification, (ii) we design a MH sampler for the posterior distribution for use within MCEM-based learning, and (iii) we employ end-to-end variational learning. A group of methods (Zhou et al., 2019) propose to learn an adversarial DNN to discriminate between the distribution of predicted segmentations and the distribution of expert segmentations. However, effective learning of the discriminator in the high-dimensional space of image segmentations typically requires a large set of expert segmentations. In contrast, our scheme proposes a novel variational MCEM method to leverage a small training set along with a tiny set of expert segmentations.

For weakly-and-semi-supervised segmentation, other methods rely on a variety of schemes to estimate pseudo-ground-truth segmentations for images devoid of expert segmentations. Xu et al. (2015) employ a max-margin clustering framework to learn from several types of weak labels for semantic segmentation. Unlike Yao and Gong (2020); Ouali et al. (2020); Lee et al. (2019); Ahn et al. (2019); Zhou et al. (2019); Wei et al. (2018); Hong et al. (2015); Xu et al. (2015), our framework (i) involves weakly-and-semi-supervised learning, (ii) involves variational modeling and MCEM inference, and (iii) infers a posterior distribution on the missing segmentations by leveraging the training-set images, their available segmentations, and their class labels. Papandreou et al. (2015) propose weakly-and-semi-supervised learning for semantic segmentation by using a CRF model to couple the image-level labels with the prediced segmentation. Unlike Papandreou et al. (2015), our approach (i) couples the image-level labels to both the predicted segmentation and the observed/input image, (ii) employs a DNN for this coupling, and (iii) proposes MCEM learning by sampling the missing segmentations from their posterior distribution. Unlike our MCEM framework, their scheme for learning relies on a two-stage algorithm that first uses their CRF to construct a pseudo-ground-truth segmentation and then uses the pseudo-ground-truth to optimize DNN parameters; our empirical analysis shows that the generated pseudo-ground-truth segmentations can have serious flaws, in which case their algorithm causes the pseudo-ground-truths to misled their DNN optimization. Pan et al. (2022) propose weakly-and-semi-supervised learning for semantic segmentation using self-supervised low-rank network with fixed multi-view transforms of the image along with a classification loss. They employ multi-view mask calibration along with a refinement module to construct pseudo-ground-truths for the missing segmentations. It uses matrix decomposition based approach for low-rank representation. Unlike Pan et al. (2022), our approach (i) proposes MCEM learning by sampling the missing segmentations from their posterior distribution and (ii) employs a (translator)

DNN for learning class probabilities. Our empirical analysis in Section 4.11 shows that these pseudo-ground-truths have several flaws that can mislead the DNN learning.

Weakly-supervised learning for object detection typically refers to learning from ground-truth information of the form of, say, points or object class labels, which are not only devoid of per-pixel segmentations but also devoid of bounding boxes of the objects. Weakly-and-semi-supervised learning for object detection typically relies on a subset of the training set with bounding boxes, and an absence of per-pixel segmentations available for any image in the training set. Typical methods like Zhang et al. (2022b,c); Chen et al. (2021); Yan et al. (2017); Zhang et al. (2020b), and methods discussed in Zhang et al. (2022a) rely on such weakly-and-semi-supervised learning mechanisms where their outputs lack per-pixel semantic segmentation of objects of interest in the image. Thus, we do not consider these methods for evaluation against weakly-and-semi-supervised learning for semantic segmentations as they would probably underperform severely when we construct per-pixel semantic segmentations from their outputs. Also, it is unfair/irrelevant to compare the network architectures for semantic segmentation (e.g., U-Net (Ronneberger et al., 2015), FCN (Long et al., 2015), DeepLabV3 (Chen et al., 2018)) with architectures for bounding-box prediction (e.g., FasterR-CNN (Ren et al., 2015), YOLO9000 (Redmon and AFarhadi, 2017), SSD (Liu et al., 2016), M2Det (Zhao et al., 2019)).

The class of methods relying on Bayesian or variational inference for semantic segmentation rely on outputing a distribution on the segmentations. For instance, some DNN methods (Kendall and Gal, 2017; Lakshminarayanan et al., 2017; Rupprecht et al., 2017) model a mean segmentation map along with a model for the per-pixel variability that is factored across pixels. Other schemes (Batra et al., 2012; Kirillov et al., 2015, 2016) generate a set of diverse (by design) semantic segmentations; such segmentations can be insightful when they are considered separately, but cannot be used as a group for deriving a single semantic segmentation from them. Some methods (Kohl et al., 2018) for semantic segmentation employ principles in variational DNN learning (Kingma and Welling, 2014) to model a distribution in latent space and, thereby, are able to model the inter-pixel spatial dependencies in the distribution on segmentations. Unlike these variational DNN methods, our method (i) focuses on weakly-and-semi-supervised learning for semantic segmentation and (ii) proposes a novel statistical framework for variational learning relying on MCEM.

Our preliminary work (Gaikwad and Awate, 2021) using weakly-and-semi-supervised learning focuses on the joint tasks of (i) binary segmentation, to separate the object/foreground from the background, and (ii) object recognition, unlike the work in this manuscript that focuses on semantic segmentation. Specifically, (Gaikwad and Awate, 2021) models each image to have either (i) a single connected region of interest or (ii) multiple regions of interest of the same class. Unlike Gaikwad and Awate (2021), our framework in this manuscript outputs a per-pixel semantic segmentation, indicating the type of class (including the background) at every pixel. To enable weakly-and-semi-supervised learning with limited availability of expert segmentations, it leverages a translator DNN that maps the segmenter DNN's output (concatenated with the input image) to the ground-truth image-level class labels, and thereby implicitly infuses the information contained in the image-level class labels into the segmenter DNN. Unlike Gaikwad and Awate (2021), this manuscript models each image to allow multiple objects/regions of interest, where each can belong to one of

many classes. In these ways, our theoretical framework in this manuscript significantly extends, and differs from, the framework in Gaikwad and Awate (2021).

- This manuscript extends the single-level variational modeling of the latent space in Gaikwad and Awate (2021) to a multi-level variational modeling of the latent space.
- Second, this paper presents new empirical results using the new theoretical framework.
- Third, this paper also presents empirical analysis with six additional baseline methods, i.e., four additional baselines for weakly-and-semi-supervised learning, i.e., Wei et al. (2018), Ouali et al. (2020), Lee et al. (2021), Pan et al. (2022), one additional baseline for weakly-supervised learning proposed in Lee et al. (2021), and one additional baseline for semi-supervised learning proposed in Ouali et al. (2020).
- Fourth, the empirical analysis in this paper sheds key insights into the model designs and capabilities underlying various methods (ours and baselines) by visualizing intermediate outputs of the DNNs underlying various methods.
- Fifth, this paper shows uncertainty maps underlying the segmentation produced by our variational framework.

3. Methods

We describe our novel framework for the task of semantic image segmentation using weakly-and-semi-supervised variational DNN learning based on MCEM. The framework assumes that expert-provided image-level class-label information is present for every image in the training set, but that expert-provided per-pixel semantic segmentations are available for only a subset of images in the training set. Our variational framework jointly learns (i) a variational segmenter DNN and (ii) a translator DNN. The variational segmenter DNN models a distribution over a multiscale latent-space variable (which is a natural result of our variational modeling strategy on UNet-style architectures that involve skip connections across multiple spatial scales between the encoder and the decoder), and outputs a distribution on the semantic probability maps. The translator DNN outputs image-level class probabilities corresponding to a pair of an input image and a segmentation probability map output by the segmenter, enabling weak supervision. During training, the framework infers a posterior distribution on the latent-space variable of the variational segmenter.

3.1 Joint Statistical Modeling of Semantic Segmenter and Translator

Let random field X model the *input image* with V pixels/voxels. The input image may contain zero or multiple objects of interest, where any object of interest is known to belong to one of K classes. The pixels *not* belonging to any object of interest may be termed as "background" pixels, i.e., the $(K+1)$-th class; note that the background region may indeed comprise other objects *not* belonging to the K classes of interest. Examples of objects of interest include cells, structures, organs, infected regions, etc., each of which can belong to different medical types (e.g., classes within normal and abnormal cases).

Let random field Z model the *semantic segmentation* of image X. Within random field Z, each pixel v has an associated categorical random variable that indicates whether

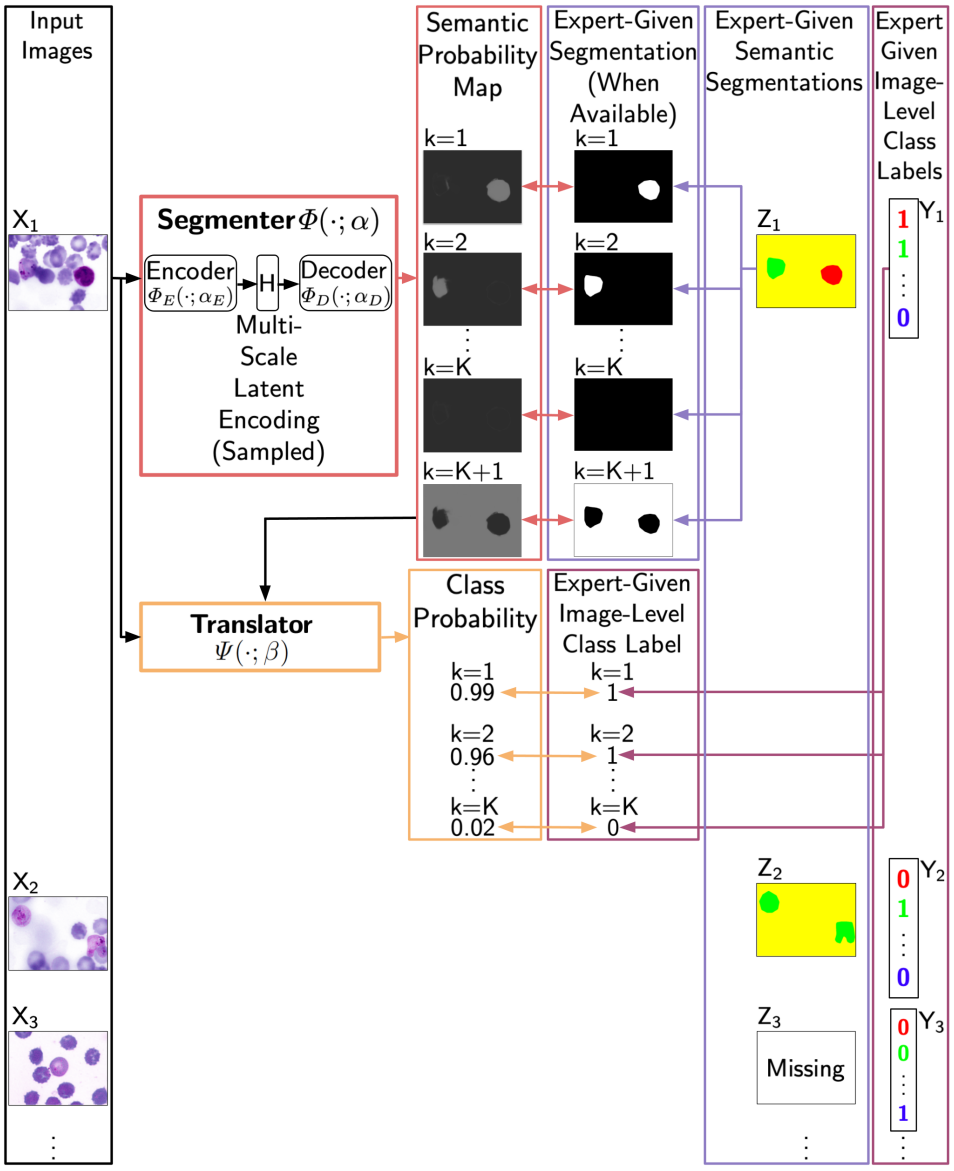


Figure 1: **Our Variational DNN framework for Semantic Segmentation using Weakly-and-Semi-Supervised Learning with MCEM.** The training data comprises the set of images $\{X_1, X_2, \cdots\}$, the set of expert-given semantic segmentations $\{Z_1, Z_2, \cdots\}$ (with some segmentations missing), and the set of expert-given class-label vectors $\{Y_1, Y_2, \cdots\}$. H is the hidden random variable modeling the multi-scale latent space within the segmenter DNN. The details of the semantic-segmenter DNN appear later in Figure 2. The details of the translator DNN, outputing image-level class labels, appear later in Figure 3.

pixel v belongs (i) to the background or (ii) to one of the K classes of the objects of interest. We represent every pixel of semantic segmentation Z by a one-hot binary random

vector of size $K+1$. Let $Z[v]$ be the one-hot-encoded vector at pixel v, with components $[Z[v][1], Z[v][2], \cdots, Z[v][K+1]]$. Thus, if pixel v belongs to an object of class k, then $Z[v][k] = 1$ and other values $Z[v][j]$, $\forall j \neq k$, equal 0. If pixel v does not belong to any object class of interest, then $Z[v][K+1] = 1$ and other values $Z[v][j]$, $\forall j \leq K$, equal 0.

Let the binary random vector Y, of size K, model the *image-level class labels* indicating the classes of the objects of interest present in the image X. In general, Y is a multi-hot vector. Y does not provide any information about the objects' locations within the image. If the image contains at least one object of class k, then $Y[k] = 1$, otherwise $Y[k] = 0$.

Let the joint distribution $P(X, Y, Z)$ model the statistical dependencies between the input image X, its semantic segmentation Z, and its class-label vector Y. During learning, we model the joint distribution $P(X, Y, Z)$ through the conditional distribution of the semantic segmentation $P(Z|X)$ and the conditional distribution of the class-label vector $P(Y|X, Z)$. While the choice of factors follows from the standard probability chain rule, the factors are also motivated by a generative model of the triplet (X, Z, Y) as follows. Starting with the input-datum image X, we can generate (using the segmenter that models $P(Z|X)$) the semantic segmentation Z which gives per-pixel information of the labels over the entire spatial domain of the input image X. Then, given X and the segmenter (and thereby Z that is a produced by the segmenter using X), the translator aggregates the information across the spatial domain to generate a single probability vector that denotes the class probabilities at the image level. During inference, for input image x', the variational framework allows us to infer a distribution over semantic probability maps $P(Z|x')$; and also infer a distribution over class-label probability vectors $P(Y|x')$. Figure 1 shows the joint DNN model for semantic segmentation and image-level classification.

In the fully-supervised learning mode, a DNN would rely on a training set of triples (X, Y, Z), where X is an input image, Z is the associated semantic segmentation, and Y the associated class-label vector. However, for image X, obtaining a per-pixel expert semantic segmentations Z is laborious and expensive. Thus, in practice, a large part of the training set will be devoid of the semantic segmentations Z and only comprise pairs (X, Y). This paper focuses on weakly-and-semi-supervised learning, where the training set $\mathcal{T} := \{(x_s, y_s, z_s)\}_{s=1}^{S} \cup \{(x_u, y_u)\}_{u=1}^{U}$, with $S \ll (S+U)$ (Figure 1).

3.2 Variational Segmenter DNN's Statistical Model

We propose a DNN-based variational model $P(Z|X)$, corresponding to the segmenter component in Figure 1, to generate a semantic segmentation for the input image X. We propose to model $P(Z|X)$ relying on two components: (i) an encoder mapping $\Phi_E(\cdot; \alpha_E)$ parameterized by α_E, and (ii) a decoder mapping $\Phi_D(\cdot; \alpha_D)$ parameterized by α_D. Let $\alpha := \alpha_E \cup \alpha_D$. Let a multiscale random vector H model the *latent space* of the segmenter (Figure 2).

First, to model the latent-space distribution, the encoder $\Phi_E(\cdot; \alpha_E)$ maps the image X to a factored multivariate Gaussian distribution over the latent-space, with factor means $\Phi_E^{\text{mean}}(X; \alpha_E)$ and factor log-variances $\Phi_E^{\text{log-var}}(X; \alpha_E)$, i.e.,

$$P(H|X; \alpha_E) := G(H; \Phi_E^{\text{mean}}(X; \alpha_E), \Phi_E^{\text{log-var}}(X; \alpha_E)), \tag{1}$$

where $G(\cdot; \mu, \lambda)$ is a factored multivariate Gaussian with means in the vector μ and the variances as the exponentials of the components in the vector λ. Such a factored model

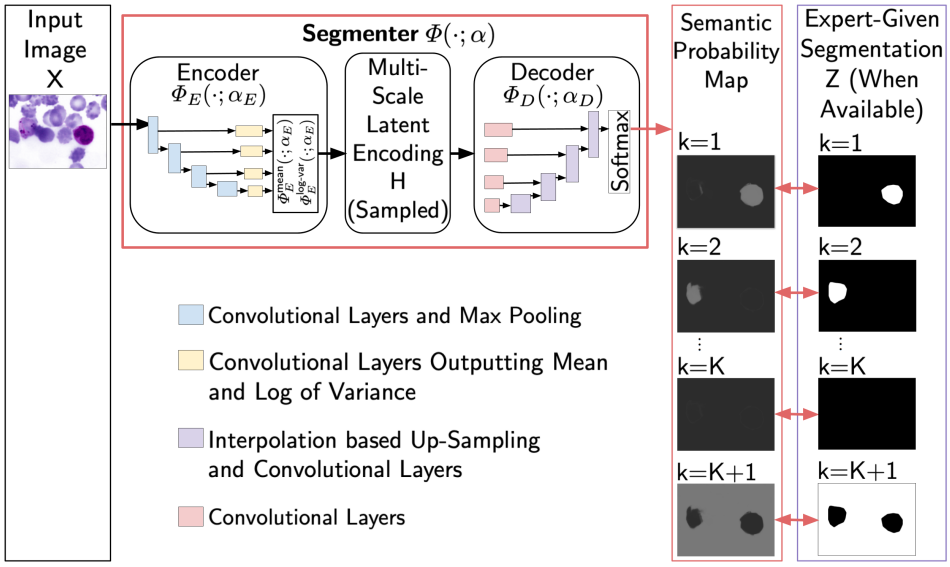


Figure 2: **Our Variational Segmenter DNN Component** within the framework shown earlier in Figure 1. For each input image X, the multiscale latent-space hidden random variable H is modeled by a input-image-specific Gaussian distribution. Variational learning entails sampling from this distribution coupled with sample reparametrization to enable end-to-end backpropagation based optimization. The variational-segmenter output is a distribution over semantic probability maps that are represented as $(K+1)$ scalar-valued images, where the k-th image represents the per-pixel probability map indicating the presence of objects of class k in the input image X. When the expert segmentation is available, the DNN learning relies on matching the semantic probability maps to the expert segmentation.

is a valid modeling approach because the encoder is designed to be highly nonlinear (as in typical DNNs) and can easily learn to map the distribution of input images to an axis-aligned (un-rotated) zero-mean multivariate Gaussian in the latent space (this is equivalent to a factored multivariate Gaussian). In fact, if at all some hypothetical encoder mapped the input distribution to a general multivariate Gaussian, then simple linear transformations of a translation and a rotation can transform any arbitrary multivariate Gaussian to a factored multivariate Gaussian. Thus, the actual DNN encoder can easily learn to include this linear transformation (implicitly) into its nonlinear mapping to output a factored multivariate Gaussian in latent space.

Second, the decoder $\Phi_D(\cdot;\alpha_D)$ maps latent-space random-vector instances h to K *probability maps* underlying the semantic segmentation, where the probability of the v-th pixel in image X belonging to an object of class k is given by $\Phi_D(h;\alpha_D)[v][k]$. Thus, at each pixel v of the segmenter output, $\Phi_D(h;\alpha_D)[v]$ is a $(K+1)$-length vector of class probabilities such that $\sum_{k=1}^{K+1} \Phi_D(h;\alpha_D)[v][k] = 1$. Also, for a fixed class k, the map $\Phi_D(h;\alpha_D)[\cdot][k]$ gives the probability of each pixel belonging to an object of class k. We model the distribution on the semantic segmentation image Z at every pixel using a categorical/Multinoulli distribution

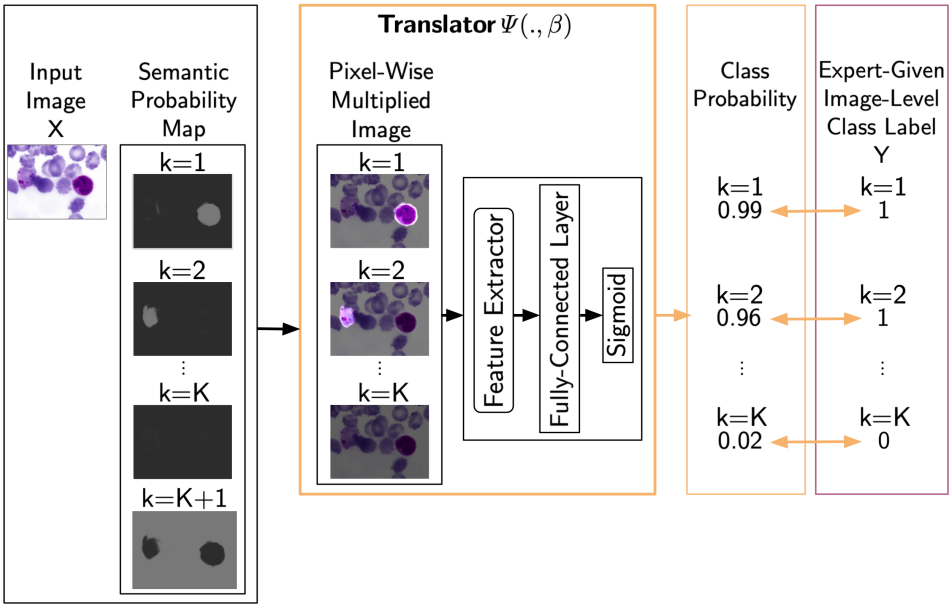


Figure 3: **Our Translator DNN Component** within the framework shown earlier in Figure 1. The DNN model for image classification takes as input a pair of images comprising (i) the image X and (ii) the semantic-segmentation probability maps. For each class k, the translator DNN takes the input as the image X pixel-wise multiplied with the semantic-segmentation probability map, and subsequently maps it to the image-level class-label probability that is matched with the expert class-label $Y[k]$.

with parameters given by the probability maps $\Phi_D(h;\alpha_D)$. Thus,

$$P(Z|H;\alpha_D) := \prod_{v=1}^{V}\prod_{k=1}^{K+1}(\Phi_D(H;\alpha_D)[v][k])^{Z[v][k]}. \tag{2}$$

Finally, the conditional distribution of the semantic segmentation Z is

$$P(Z|X;\alpha) = \int_h P(Z|h;\alpha_D)P(h|X;\alpha_E)dh \approx \frac{1}{R}\sum_{r=1}^{R} P(Z|h^r;\alpha_D), \tag{3}$$

where h^r are sampled independently in the multiscale latent space from the Gaussian $P(H|X;\alpha_E)$. In this way, for each input X, the Gaussian model $P(H|X;\alpha_E)$ over the latent space along with the nonlinear transformation $\Phi_D(\cdot;\alpha_D)$ on H together implicitly model a distribution over the semantic probability maps $P(Z|X;\alpha)$, incorporating both inter-pixel and intra-pixel dependencies of the class probabilities associated with the semantic segmentation.

3.3 Translator DNN's Statistical Model

We propose a DNN-based statistical model $P(Y|X)$ (Figure 3) to relate the image-level class labels Y for image X; while this enables weakly-and-semi-supervised learning, it also

aids during supervised learning. Given X and the probability maps $\Phi_D(H;\alpha_D)$, we assume conditional independence of the set of image-level class labels $\{Y[k]\}_{k=1}^{K}$. So, we model

$$P(Y|X,H;\alpha_D,\beta) := \prod_{k=1}^{K} P(Y[k]|X,\Phi_D(H;\alpha_D)[\cdot][k];\beta). \tag{4}$$

Let a DNN model a mapping $\Psi(\cdot;\beta)$, parameterized by β, from each image pair $(X,\Phi_D(H;\alpha_D)[\cdot][k])$ to an associated parameter $\Psi(X,\Phi_D(H;\alpha)_D[\cdot][k];\beta)$ underlying a Bernoulli random variable associated with image-level class label $Y[k]$. Thus,

$$\begin{aligned} &P(Y[k]|X,\Phi_D(H,X;\alpha_D)[\cdot][k];\beta) := \\ &(\Psi(X,\Phi_D(H,X;\alpha_D)[\cdot][k];\beta))^{Y[k]}(1-\Psi(X,\Phi_D(H,X;\alpha_D)[\cdot][k];\beta))^{1-Y[k]}. \end{aligned} \tag{5}$$

Finally,

$$P(Y|X;\theta) := \int_h P(Y|X,h;\alpha_D,\beta)P(h|X;\alpha_E)dh \approx \frac{1}{Q}\sum_{q=1}^{Q} P(Y|X,h^q;\alpha_D,\beta), \tag{6}$$

where h^q is sampled in latent space from the Gaussian distribution $P(H|X;\alpha_E)$.

During learning, promoting $\log P(Y|X,H;\alpha_D,\beta)$ penalizes (for all classes k) the cross entropy between the ground-truth Bernoulli distribution with probabilities $[Y[k],1-Y[k]]$ and the DNN-output Bernoulli distribution with probabilities $[\Psi(.;\beta),1-\Psi(.;\beta)]$. The translator DNN $\Psi(.;\beta)$ plays a crucial role in implicitly infusing the information within the ground-truth class-label vector Y into the learning of the segmenter DNN, irrespective of whether the expert segmentation Z is present or absent in the training set. First, the translator DNN informs the learning of the segmenter-DNN's decoder by functionally connecting the segmenter output to the image-level class-labels Y. Furthermore, our MCEM learning framework (Section 3.4) leverages the translator-based model $P(Y|X,H;\alpha_D,\beta)$ to inform the sampling of the hidden random variable H from its posterior distribution using a Markov-Chain Monte Carlo (MCMC) algorithm (Section 3.5); the hidden H relates to the segmenter-encoder's output. Thus, the translator DNN allows us to leverage the information in the image-level class-labels for backpropagation-based optimization (end-to-end learning) of the segmenter-DNN's decoder as well as the encoder (after sample reparametrization, described later in Section 3.5).

3.4 Monte-Carlo EM for Weakly-and-Semi-Supervised Learning

Given training-set $\mathcal{T}$, we propose a MCEM framework for joint DNN learning of the segmenter and the translator. Let the DNN *parameters* be $\theta := \alpha \cup \beta$. For images X_s and X_u, we model the hidden latent-space representations as H_s and H_u, respectively. The *complete data* $\mathcal{T}^{\text{complete}} := \{(x_s,y_s,z_s,H_s)\}_{s=1}^{S} \cup \{(x_u,y_u,H_u)\}_{u=1}^{U}$. The complete-data likelihood is

$$P(\mathcal{T}^{\text{complete}};\theta) := \prod_{u=1}^{U} P(x_u,y_u,H_u;\theta) \prod_{s=1}^{S} P(x_s,y_s,z_s,H_s;\theta). \tag{7}$$

E Step. At iteration i, with parameter estimates θ^i, EM designs a minorized version of the log-likelihood function as the expectation of the complete-data log-likelihood over the

posterior distribution $\prod_{u=1}^{U} P(H_u|x_u, y_u; \theta^i) \prod_{s=1}^{S} P(H_s|x_s, y_s, z_s; \theta^i)$ of the missing latent-space encodings H_u and H_s, i.e.,

$$Q(\theta; \theta^i) := E_{\prod_{u=1}^{U} P(H_u|x_u,y_u;\theta^i) \prod_{s=1}^{S} P(H_s|x_s,y_s,z_s;\theta^i)} \Big[\sum_{u=1}^{U} \log P(x_u, y_u, H_u; \theta) + \sum_{s=1}^{S} \log P(x_s, y_s, z_s, H_s; \theta) \Big]. \tag{8}$$

MCEM approximates the analytically intractable expectation as a Monte-Carlo average using an independent and identically distributed (i.i.d.) sample $\{(h_u^t) \sim P(H_u|x_u, y_u; \theta^i)\}_{t=1}^{T} \cup \{h_s^t \sim P(H_s|x_s, y_s, z_s; \theta^i)\}_{t=1}^{T}$, to give

$$\begin{aligned}
&\widehat{Q}(\theta; \theta^i) \\
&:= \sum_{u=1}^{U} \frac{1}{T} \sum_{t=1}^{T} \log P(x_u, y_u, h_u^t; \theta) + \sum_{s=1}^{S} \frac{1}{T} \sum_{t=1}^{T} \log P(x_s, y_s, z_s, h_s^t; \theta) \\
&:= \sum_{u=1}^{U} \frac{1}{T} \sum_{t=1}^{T} \log \big(P(y_u|x_u, h_u^t; \beta \cup \alpha_D) P(h_u^t|x_u; \alpha_E) P(x_u) \big) \\
&+ \sum_{s=1}^{S} \frac{1}{T} \sum_{t=1}^{T} \log \big(P(y_s|x_s, h_s^t; \beta \cup \alpha_D) P(z_s|h_s^t; \alpha_D) P(h_s^t|x_s; \alpha_E) P(x_s) \big) .
\end{aligned} \tag{9}$$

We sample the latent-space encodings h_u^t and h_s^t, from $P(H_u|x_u, y_u; \theta^i)$ and $P(H_s|x_s, y_s, z_s; \theta^i)$, using a MH sampler incorporating an efficient proposal distribution, as described later in Section 3.5.

M Step. At iteration i, EM maximizes $\widehat{Q}(\theta; \theta^i)$ over parameters θ leading to the updated parameters θ^{i+1}. Selecting a sufficiently large Monte-Carlo sample, MCEM inherits the behaviour of EM that leads to convergence to a stationary point of the log-likelihood function of the observed training set $\mathcal{T}$. The terms $P(x_u)$ and $P(x_s)$, involving the prior model on input images, are independent of the DNN parameters θ, and thereby can be ignored during DNN optimization.

3.5 Efficient MH Sampling of Hidden Latent-Space Encodings

Within iteration i of MCEM, we propose a MH sampling algorithm to (i) sample the encodings h_u from the posterior distribution $P(H_u|x_u, y_u; \theta^i)$, and (ii) sample the encodings h_s from the posterior distribution $P(H_s|x_s, y_s, z_s; \theta^i)$.

The MH sampling strategy is a MCMC method that, given a current state, first generates a candidate by sampling from a proposal distribution and then, based on some probability, either updates the state to the candidate or retains the current state. For sampling H, we now consider several strategies to generate candidates and propose one that is computationally straightforward and leads to a good probability of acceptance of the generated candidate. First, it is well known that the naive strategy of sampling from a simple (fixed parametric) proposal distribution over latent space $P(H)$ is inefficient, because it would fail to capture (i) the true distribution of H and (ii) the dependencies between the latent-space encoding H and the datum X that are present in the posterior distribution that we desire

to sample from. Second, modeling a realistic proposal distribution is difficult because of the unavailability of the required observations for the hidden variables H; moreover, sampling from such a distribution would continue to lead the sample being independent of the input datum X. Third, some schemes aim to improve the aforementioned strategy by adapting the covariance of the proposal distribution to the local structure of the (posterior) distribution that we desire to sample from. However, such schemes often add significant complexity in reliably modeling the local covariance using the Hessian of the distribution. Rather, we design an improved proposal distribution that leverages the DNN model learned at iteration i, as follows.

For the term $P(H_u|x_u, y_u; \theta^i)$, Bayes rule factors it, upto a normalizing constant, into the form $P(y_u|x_u, H_u; \alpha_D^i, \beta^i)P(H_u|x_u; \alpha_E^i)$, where (i) the first factor $P(y_u|x_u, H_u; \alpha_D^i, \beta^i)$ is modeled using the segmenter's decoder $\Phi_D(\cdot; \alpha_D^i)$ and the translator $\Psi(\cdot; \beta^i)$, and (ii) the second factor $P(H_u|x_u; \alpha_E^i)$ is modeled using the segmenter's encoder $\Phi_E(\cdot; \alpha_E^i)$ as a factored Gaussian. Within the Markov chain in the MH sampler, when the current latent-space encoding (state) is h, we propose to use the candidate-proposal distribution $P(H_u|x_u; \alpha_E^i)$ to draw a candidate latent-space encoding (state) h'. Sampling the encoding h' from the factored Gaussian distribution $P(H_u|x_u; \alpha_E^i)$ on latent space is computationally efficient because it needs only a single forward pass through the segmenter's encoder. Also, this proposal distribution is effective because it is (i) informed by the training dataset through the usage of the current parameter estimates θ^i, (ii) informed by the specific input datum x_u, and (iii) is closely related to the posterior distribution $P(H_u|x_u, y_u; \theta^i)$ that we desire to sample from. In this way, the proposal candidates tend to have a good acceptance probability within the MH sampler. The MH sampler's acceptance probability for candidate h' is

$$\min\left(1, \frac{P(h'|x_u, y_u; \theta^i)}{P(h|x_u, y_u; \theta^i)} \frac{P(h|x_u; \alpha_E^i)}{P(h'|x_u; \alpha_E^i)}\right) = \min\left(1, \frac{P(y_u|x_u, h'; \alpha_D^i, \beta^i)}{P(y_u|x_u, h; \alpha_D^i, \beta^i)}\right). \tag{10}$$

Thus, at iteration i, while the candidate generation relies on the learned segmenter's encoder $\Phi_E(\cdot; \alpha_E^i)$, the candidate acceptance relies on, both, the learned segmenter's decoder $\Phi_D(\cdot; \alpha_D^i)$ and the learned translator $\Psi(\cdot; \beta^i)$. In this way, for training-set images x_u devoid of expert segmentations, the corresponding observed class-labels y_u help the MCEM to select an informative sample $\{h_u^t\}_{t=1}^T$ that, in turn, improves the learning of the translator and the segmenter.

We use an analogous strategy for sampling from $P(H_s|x_s, y_s, z_s; \theta^i)$, where latent-space hidden variable H_s is associated with those images x_s for which an expert segmentation z_s is available. Here, Bayes rule refactors $P(H_s|x_s, y_s, z_s; \theta^i)$, upto a normalizing constant, into the form $P(y_s|x_s, H_s; \alpha_D^i, \beta^i)P(H_s|x_s, z_s; \theta^i) \propto P(y_s|x_s, H_s; \theta^i)P(z_s|H_s; \alpha_D^i)P(H_s|x_s; \alpha_E^i)$, where, at iteration i, (i) the first factor $P(y_s|x_s, H_s; \alpha_D^i, \beta^i)$ is modeled by the mapping $\Phi_D(\cdot; \alpha_D^i)$ underlying the segmenter's decoder and the mapping $\Psi(\cdot; \beta^i)$ underlying the translator, (ii) the second factor $P(z_s|H_s; \alpha_D^i)$ is modeled by the mapping $\Phi_D(\cdot; \alpha_D^i)$ underlying the segmenter's decoder, and (iii) the third factor $P(H_s|x_s; \alpha_E^i)$ is modeled by the mapping $\Phi_E(\cdot; \alpha_E^i)$ underlying the segmenter's encoder. In this case, within the Markov chain in the MH sampler, when the current latent-space encoding (state) is h, we propose to use the candidate-proposal distribution $P(H_s|x_s; \alpha_E^i)$ to draw a candidate latent-space

encoding (state) h'. The MH sampler's acceptance probability for the candidate (h') is

$$\min\left(1, \frac{P(h'|x_s, y_s, z_s; \theta^i)}{P(h|x_s, y_s, z_s; \theta^i)} \frac{P(h|x_s; \alpha_E^i)}{P(h'|x_s; \alpha_E^i)}\right) = \tag{11}$$

$$\min\left(1, \frac{P(y_s|x_s, h'; \alpha_D^i, \beta^i) P(z_s|h'; \alpha_D^i)}{P(y_s|x_s, h; \alpha_D^i, \beta^i) P(z_s|h; \alpha_D^i)}\right) \tag{12}$$

Thus, our candidate generation relies on the learned segmenter's encoder model $\Phi_E(\cdot; \alpha_E^i)$ at iteration i, and our candidate acceptance relies on the segmenter's decoder model $\Phi_D(\cdot; \alpha_D^i)$ and the translator model $\Psi(\cdot; \beta^i)$ at iteration i.

In this paper, we sample each encoding h^t by (i) initializing the Markov chain to a state from the previous iteration $t-1$ and (ii) running the Markov chain through a burn-in period that uses a different pseudo-random number sequence for each t. In this paper, we find that a burn-in of 10 iterations and a sample size $T = 20$ works reasonably with the stochastic optimization.

To enable end-to-end learning, while sampling $h \sim P(H|x; \alpha_E^i)$, we reparameterize its d-th component as

$$h[d] := \Phi_E^{\text{mean}}(X; \alpha_E)[d] + \eta \exp(0.5 \Phi_E^{\text{log-var}}(X; \alpha_E)[d]), \tag{13}$$

where η is a random draw from a standard normal distribution. Such a reparameterization of h, in terms of the DNN parameters α_E, allows backpropagating the loss-function gradients through h to α_E.

Algorithm 1 summarizes the training algorithm.

3.6 Deep-MCEM based Inference for Test Images

We perform the inference on test images using the optimal DNN model parameters $\theta^* = \alpha^* \cup \beta^*$ obtained using the weakly-and-semi-supervised learning.

For a test image x', we propose to infer the pixel-wise distribution of probability values $P(Z[v][k]|H; \alpha_D^*)$, resulting from the underlying distribution over H that depends on x'. As described in Equation 3, we obtain the mean $\mu_{\text{seg}}[v][k]$ of the distribution of probability values $P(Z[v][k]|H; \alpha_D^*)$ using Monte-Carlo sampling of the latent-space encoding h from the Gaussian $P(H|x'; \alpha_E^*)$, i.e.,

$$\mu_{\text{seg}}[v][k] := P(Z[v][k]|x'; \alpha^*) := \frac{1}{T'} \sum_{t=1}^{T'} P(Z[v][k]|h^t; \alpha_D^*), \tag{14}$$

with independently sampled $h^t \sim P(H|x'; \alpha_E^*)$. We obtain a hard semantic segmentation, by computing, at each pixel v, a class label k^* for which the mean probability $\mu_{\text{seg}}[v][k]$ is the highest. We obtain the variance $\left(\sigma_{\text{seg}}[v][k]\right)^2$ of the distribution of the probability values $P(Z[v][k]|H; \alpha_D^*)$ as,

$$\left(\sigma_{\text{seg}}[v][k]\right)^2 := \frac{1}{T'} \sum_{t=1}^{T'} (P(Z[v][k]|h^t; \alpha_D^*) - \mu_{\text{seg}}[v][k])^2, \tag{15}$$

Input Training Set $\mathcal{T}^{\text{complete}} := \{(x_s, y_s, z_s, H_s)\}_{s=1}^{S} \cup \{(x_u, y_u, H_u)\}_{u=1}^{U}$
DNN parameters to optimize $\theta := \alpha \cup \beta$.
Random initialization of DNN parameters θ as θ^0.
$bestVal = 0$
for $i = 1 : I$ **do**
  % E step for iteration i.
  % MH sampling within E step, as described in Section 3.5.
  for $t = 1 : T$ **do**
    % Sample latent-space vectors h_s for images x_s.
    for $s = 1 : S$ **do**
      $h_s^t \leftarrow MHSampler(x_s, y_s, z_s; \theta^{i-1})$
    end
    % Sample latent-space vectors h_u for images x_u.
    for $u = 1 : U$ **do**
      $h_u^t \leftarrow MHSampler(x_u, y_u; \theta^{i-1})$
    end
  end
  % This implicitly gives us the function $\widehat{Q}(\theta; \theta^{i-1})$, as described in Equation 9.
  % M step for iteration i.
  Optimize $\widehat{Q}(\theta; \theta^{i-1})$ over DNN parameters θ using stochastic gradient descent to give θ^i.
end

Algorithm 1: Pseudo-code of Deep MCEM based Variational Learning for Semantic Segmenation.

with independently sampled $h^t \sim P(H|x'; \alpha_E^*)$. We propose to treat the standard deviation $\sigma_{\text{seg}}[v][k]$ as the uncertainty in the prediction of the segmentation probability $P(Z[v][k]|H; \alpha_D^*)$. In this paper, we use a sample size of $T' = 128$.

For the test image x', we propose to infer the image-level distribution of class probability values $P(Y[k]|H; \alpha_D^*, \beta^*)$, resulting from the underlying distribution over H that depends on x'. As described in Equation 6, we get the mean $\mu_{\text{class}}[k]$ of the distribution of probability values $P(Y[k]|H; \alpha_D^*, \beta^*)$ using Monte-Carlo sampling of the latent-space encoding h from the Gaussian $P(H|x'; \alpha_E^*)$, i.e.,

$$\mu_{\text{class}}[k] := P(Y[k]|x'; \theta^*) := \frac{1}{T'} \sum_{t=1}^{T'} P(Y[k]|h^t; \alpha_D^*, \beta^*). \tag{16}$$

Equation 5 and Equation 16 indicate that $\mu_{\text{class}}[k]$ relates to the probability of the image having an object of class k. Thus, we get the hard image-level class labels by selecting those classes for which $\mu_{\text{class}}[k] \geq 0.5$. We get the variance $\big(\sigma_{\text{class}}[k]\big)^2$ of the distribution of probability values $P(Y[k]|H; \alpha_D^*, \beta^*)$ as,

$$\big(\sigma_{\text{class}}[k]\big)^2 := \frac{1}{T'} \sum_{t=1}^{T'} (P(Y[k]|h^t; \alpha_D^*, \beta^*) - \mu_{\text{class}}[k])^2, \tag{17}$$

using independently sampled $h^t \sim P(H|x'; \alpha_E^*)$. We propose to treat the standard deviation $\sigma_{\text{class}}[k]$ as the uncertainty in the prediction of the segmentation probability $P(Y[k]|H; \alpha_D^*, \beta^*)$. In this paper, we use a sample size of $T' = 128$.

During inference on a test image x', sampling the latent space variable H from the Gaussian distribution $P(H|x'; \alpha_E^*)$ is computationally efficient as it needs a single forward-pass through the segmenter's encoder to get the distribution $P(H|x'; \alpha_E^*)$, followed by draws from a standard normal to sample h.

4. Results and Discussion

We evaluate our method described in Section 3 for weakly-and-semi-supervised semantic segmentation *WeakS-Ours* that leverages (i) the fully-supervised S images from the training set having both expert segmentations and image-level class labels as well as (ii) the weakly-supervised U images the training set having only image-level class labels without any expert segmentations. For WeakS-Ours,

(i) the segmenter $\Phi(\cdot; \alpha)$ relies on the standard U-Net architecture for semantic segmentation (Ronneberger et al., 2015) and (ii) the translator $\Psi(\cdot; \beta)$ relies on the standard ResNet architecture for classification (He et al., 2016).

The U-Net architecture of the DNN for semantic segmentation (Figure 2) takes as input an RGB image and produces as output a $K + 1$-channel semantic probability map. The U-Net architecture comprises (i) the encoder, (ii) the multiscale latent-space encoding H, and (iii) the decoder. The encoder has four blocks (colored blue in Figure 2) performing convolutions and max-pooling, followed by convolutional layers (colored yellow in Figure 2) that outputs two vectors ϕ_E^{mean} and $\phi_E^{log-var}$ (at each of the four scales) to parametrize the Gaussian distribution of the multiscale latent-space variable H. The multiscale latent-space variable H combines four output vectors, one at each of the four scales, of lengths 2048, 1024, 512, and 256. The decoder has four convolutional layers that take the sampled latent-space encodings as input (colored pink in Figure 2), followed by blocks (colored purple in Figure 2) performing interpolation-based up-sampling and convolutions. The segmenter finally applies a softmax function to get the probability maps for each of the $K + 1$ classes.

The input to the translator is a pair of images comprising (i) the original RGB image and (ii) its semantic probability map output by the segmenter. We propose to leverage the per-class semantic probability maps as attention maps (for spatial per-pixel rescaling/reweighting) for the original RGB image (Figure 3). Each of the K attention-weighted images forms an input to the translator DNN that relies on the standard ResNet50 architecture. The translator outputs a K-length probability vector, where the value of the k-th component indicates the probability of the original image containing an object of class k.

4.1 Baseline Methods for Comparison

We compare our method with seven baseline methods including: (A) five baselines for weakly-and-semi-supervised semantic segmentation, i.e., (i) *WeakS-CRF* proposed in Papandreou et al. (2015), (ii) *WeakS-CAM-Dil* proposed in Wei et al. (2018), (iii) *WeakS-CAM-CCT* proposed in Ouali et al. (2020), (iv) *WeakS-AdvCAM-CCT* proposed in Lee et al. (2021), and (v) *WeakS-SLRNet* proposed in Pan et al. (2022); (B) one baseline for

weakly-supervised learning *Weakly-AdvCAM* proposed in Lee et al. (2021); (C) one baseline for semi-supervised learning *Semi-CCT* proposed in Ouali et al. (2020). All the baselines for weakly-and-semi-supervised learning train using (i) S images having both expert segmentations and image-level class labels and (ii) U images with image-level labels and without expert segmentations. The baseline Weakly-AdvCAM for weakly-supervised learning trains using (ii) $S+U$ images with image-level labels and without expert segmentations. The baseline Weakly-AdvCAM for weakly-supervised learning trains using (i) S images having both expert segmentations and image-level class labels and (ii) U images without any supervision that is without image-level labels and without expert segmentations.

Across all the baseline methods, as well as our method, we ensure consistent designs of the DNN architectures for fair comparisons. All the baseline methods employ a DNN for semantic segmentation which uses the same (standard) backbone U-Net architecture in our framework. The weakly-and-semi-supervised learning methods WeakS-CAM-Dil, WeakS-AdvCAM-CCT, and WeakS-CAM-CCT; and the weakly-supervised learning method WeakS-AdvCAM rely on a pre-trained DNN classifier to produce CAMs, where the classifier DNNs rely on the same backbone ResNet architecture as in our framework. WeakS-CRF enables weakly-and-semi-supervised learning by coupling the semantic segmentation to the image-class labels using a CRF instead of a DNN.

4.2 Real-World Microscopy Datasets

The BCCD histopathology dataset (`github.com/Shenggan/BCCD_Dataset`) shows blood tissue including white blood cells (WBCs) of four classes, i.e., eosinophil, lymphocyte, neutrophil, and monocyte. We have a set of 410 image tiles of size 480×640 pixels. Because of very limited examples (only around 20) of monocytes for reliable training and evaluation, we remove the instances of the monocyte class from the dataset. Thus, the curated dataset has $K = 3$ classes. From this set, we randomly select (i) $S + U = 150$ tiles for training, (ii) 50 tiles for validation (to tune free parameters for all methods), and (iii) 190 tiles for testing. Because of the small size of this dataset, during training, we augment the dataset by including a randomly flipped (horizontally or vertically) version of each image tile, effectively making the training-set size as $S + U = 300$.

The Malaria histopathology dataset (`data.broadinstitute.org/bbbc/BBBC041`) has image tiles showing Giemsa-stained microscopy images of blood tissue infected by the malarial parasite plasmodium vivax. The dataset consists of WBCs, non-infected red blood cells (RBCs), and infected RBCs. The infected RBCs are differentiated into different classes that indicate the life-cycle stage of the malarial parasite within the RBC. The goal for image analysis is to detect the infected RBCs and indicate their classes. We have data from $K = 3$ such infected-RBC classes, i.e., ring, trophozoite, and schizont. We generate multiple cropped image tiles from the original image. While generating such image tiles we ensure the dataset property of have multiple types of cells is maintained. We have a set of 1273 image tiles of size 600×450. From this set, we randomly select (i) $S + U = 500$ image for training, (ii) 200 image for validation (to tune free parameters for all methods), and (iii) 573 tiles for testing. Each tile can contain zero or more number of infected RBCs. For tiles containing multiple infected RBCs, each infected RBC can belong to a different class.

The Lizard histopathology dataset provided by Graham et al. (2021) (`https://warwick.ac.uk/fac/cross_fac/tia/data/lizard`) has histopathology microscopy images of blood tissues. The dataset consists of colon nuclei semantic segmentation where, the nuclei's are classified into the 6 categories as (i) epithelial, (ii) lymphocyte, (iii) plasma, (iv) neutrophil, (v) eosinophil, and (vi) connective. The goal for image analysis is to segment the foreground nuclei regions in the image and indicate their classes. We have data from $K = 3$ such classes, i.e., epithelial, lymphocyte, and connective. We generate multiple cropped image tiles from the original image. We have a set of 1349 image tiles of size 256×256. From this set, we randomly select (i) $S + U = 800$ images for training, (ii) 200 images for validation (to tune free parameters for all methods), and (iii) 349 tiles for testing. Each tile can contain zero or more nuclei of different classes.

For all the datasets, we obtain manual expert segmentations of the cells of interest. These data originate from multiple clinical sites and, thereby, despite standard staining protocols, are susceptible to natural staining variation. Thus, we perform stain normalization Reinhard et al. (2001) during pre-processing. All the input images shown in this paper are stain normalized.

4.3 Training Strategy for All Methods

We ensure consistency in DNN-training schemes across all three baselines as well as the methods within our framework. For all methods, we train using stochastic gradient descent (Robbins and Monro, 1951; LeCun et al., 1998) with the learning-rate parameter set to 0.01 and the momentum parameter set to 0.9; we find that the results of all methods are insensitive to small changes in these values. To reduce overfitting for all methods, we follow the following strategy: (i) during training, after every few iterations, we save the model parameters and (ii) after the training finishes, we pick the model parameters that perform the best on the validation set.

4.4 Evaluation Strategy for All Methods

We evaluate performance using the mean intersection-over-union (mean-IoU; equivalently mean-Jaccard) between (i) manual segmentations provided by the expert and (ii) the hard segmentation image produced by the methods (e.g., as described in Section 3.6 for our methods). We evaluate each method by varying the number of the training-set images having expert segmentations, i.e., we change the level of supervision $\gamma := 100 \cdot S/(S+U)\%$ from 10% (indicating a very low level of supervision; $S \ll U$) to 100% (indicating full supervision; $S \gg U = 0$). We also evaluate our two methods for their accuracy in predicting the image-level class labels. We evaluate the variability in the performance of each method by performing multiple repeated experiments involving a random selection of the training set, the validation set, and the test set.

In addition to the comparative analysis across methods, we present empirical insights into some of the key aspects of the weakly-and-semi-supervised learning mechanisms of all methods, i.e., (i) the posterior sampling for methods within our framework, (ii) efficacy of CAM-based approaches in estimating missing segmentations, (iii) the role of the CRF model within Papandreou et al. (2015) to couple their DNN segmentation with the image-

level class labels and , (iv) the method in Pan et al. (2022) that uses multi-view based self-supervised learning.

4.5 Scheme for Qualitative Visualization of Results

This section describes our schemes for visualizing (i) probabilistic semantic segmentations, (ii) hard semantic segmentations, and (iii) uncertainty estimates underlying semantic segmentations. All our datasets involve $K = 3$ classes of the objects of interest. First, to visualize a probabilistic semantic segmentation, in our case of $K = 3$, at each pixel, we embed the information in the $(K + 1)$-length vector of class probabilities into the three channels of a RGB image. Specifically, we use the first three values in this vector to assign values to the color components corresponding to red, green, and blue. Second, to visualize a hard semantic segmentation, at each pixel, we embed the information in the $(K+1)$-length one-hot vector of class labels into the three channels of a color/RGB image. Specifically, we indicate each of the $(K + 1) = 4$ classes, by the colors red, green, blue, and yellow. Third, to visualize the uncertainty maps underlying the K object classes of interest, in our case of $K = 3$, at each pixel, we embed the information in the K-length uncertainty vector into the three channels of a RGB image. Specifically, we use the three values in this uncertainty vector to assign values to the color components corresponding to red, green, and blue.

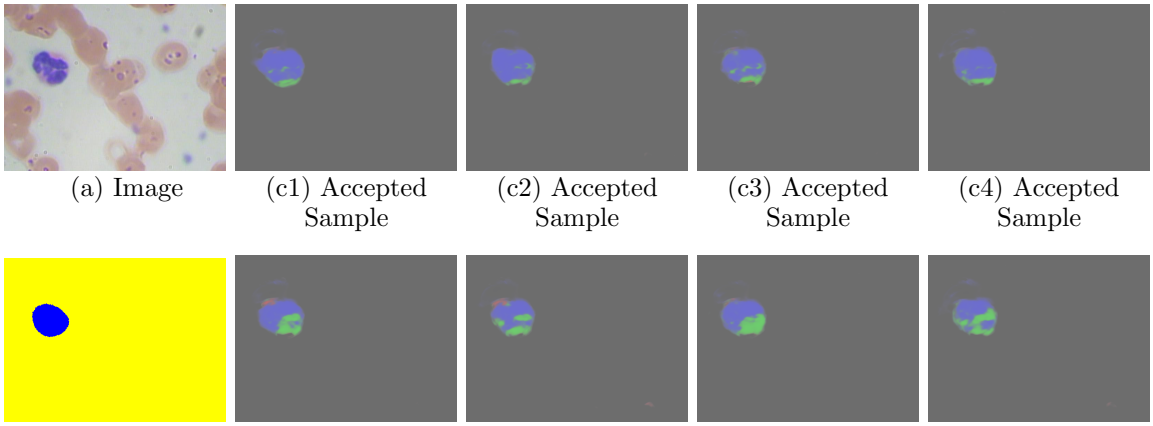
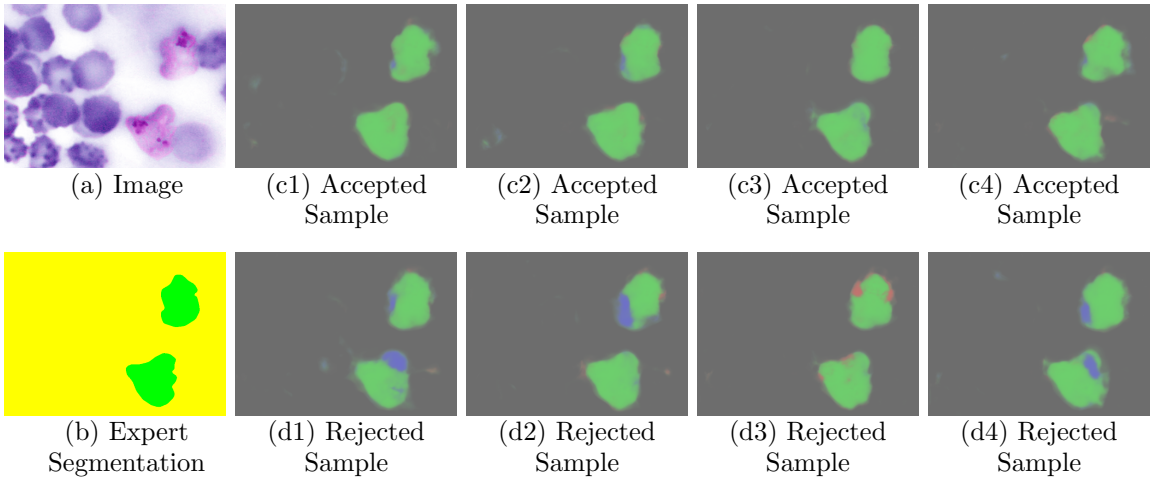


Figure 4: **Empirical Insights Into the MH-Sampling Within Our Framework: BCCD Dataset (a)** Input RGB Image from the training set, and **(b)** its expert-given semantic segmentation. **(c1)–(c4)** Examples of semantic probability maps resulting from the candidates h, in latent space, *accepted* by the MH sampler described in Section 3.5. **(d1)–(d4)** Examples of semantic probability maps resulting from the candidates h, in latent space, *rejected* by the MH sampler described in Section 3.5. The visualization scheme for the semantic probability maps in (c1)–(c4) and (d1)–(d4) is described in Section 4.5.



(a) Image (c1) Accepted Sample (c2) Accepted Sample (c3) Accepted Sample (c4) Accepted Sample

(b) Expert Segmentation (d1) Rejected Sample (d2) Rejected Sample (d3) Rejected Sample (d4) Rejected Sample

Figure 5: **Empirical Insights Into the MH-Sampling Within Our Framework: Malaria Dataset** The description is the same as in Figure 4.

4.6 Empirical Insights Into the MH-Sampling Within Our Framework

We provide some insights into the MH-sampling scheme described in Section 3.5; qualitatively (Figure 4, Figure 5 and Figure 6) and quantitatively (Table 1). For a typical image x in the training set, Figure 4 and Figure 5 show examples of semantic probability maps, i.e., $\Phi_D(h, x; \alpha_D^i)$, produced by the decoder corresponding to the MH-sampler-accepted and MH-sampler-rejected candidates of the latent-space encoding h drawn from the posterior distribution ($P(h_u|x_u, y_u; \theta^i)$ or $P(h_s|x_s, y_s, z_s; \theta^i)$). The accepted sampled latent-space encodings h typically lead to semantic probability maps that are more similar to the expert-given semantic segmentation, and, thereby, also more consistent with the expert-given image-level class labels. In other words, the MH sampler tends to reject those candidates h that lead to (i) a semantic probability map having larger discrepancy with respect to the expert segmentation Z and/or (ii) image-level class-label probabilities that are lower for the expert-given image-level class labels Y. For instance, in Figure 4 and Figure 5, the accepted cases had an h that lead to the probabilities $P(y|x, h; \alpha_D, \beta)$ of the image-level multi-hot class-label vector y in the range $[0.979, 0.989]$, whereas the same probabilities for the rejected cases were in the lower range of $[0.921, 0.949]$. In this way, during training, the MH sampler produces the set of accepted latent-space encodings $\{h^t\}_{t=1}^T$ that are informed by the expert-given image-level class labels as well as the expert-given semantic segmentation (when one is available). The accepted sampled set, in turn, informs and improves the encoder model through the subsequent sample reparametrization and backpropagation.

We evaluate the MH sampler (quantitatively in Table 1 and qualitatively in Figure 6) by using three different strategies of generating the proposal distribution: (i) when the distribution is modeled by the currently-learned encoder (our MH-sampler, where the encoder is being dynamically updated during data-driven learning); (ii) when the distribution is an isotropic Gaussian with the mean set to the current state and the variances fixed to large values (this risks producing many candidates in regions where the posterior distribution

Table 1: **Empirical Insights Into the MH-Sampling: Quantitative analysis.** Acceptance rate of the MH sampler, in DNNs trained at different supervision levels, using different proposal distributions (manually tuned versus learned).

Fraction of training-set with expert segmentations	Acceptance rate for isotropic-Gaussian proposal with manually-tuned large variance	Acceptance rate for isotropic-Gaussian proposal with manually-tuned small variance	Acceptance rate for our factored-Gaussian proposal with learned variances
10%	72.73%	95.14%	90.69%
50%	70.11%	94.89%	85.02%
100%	69.81%	94.62%	83.34%

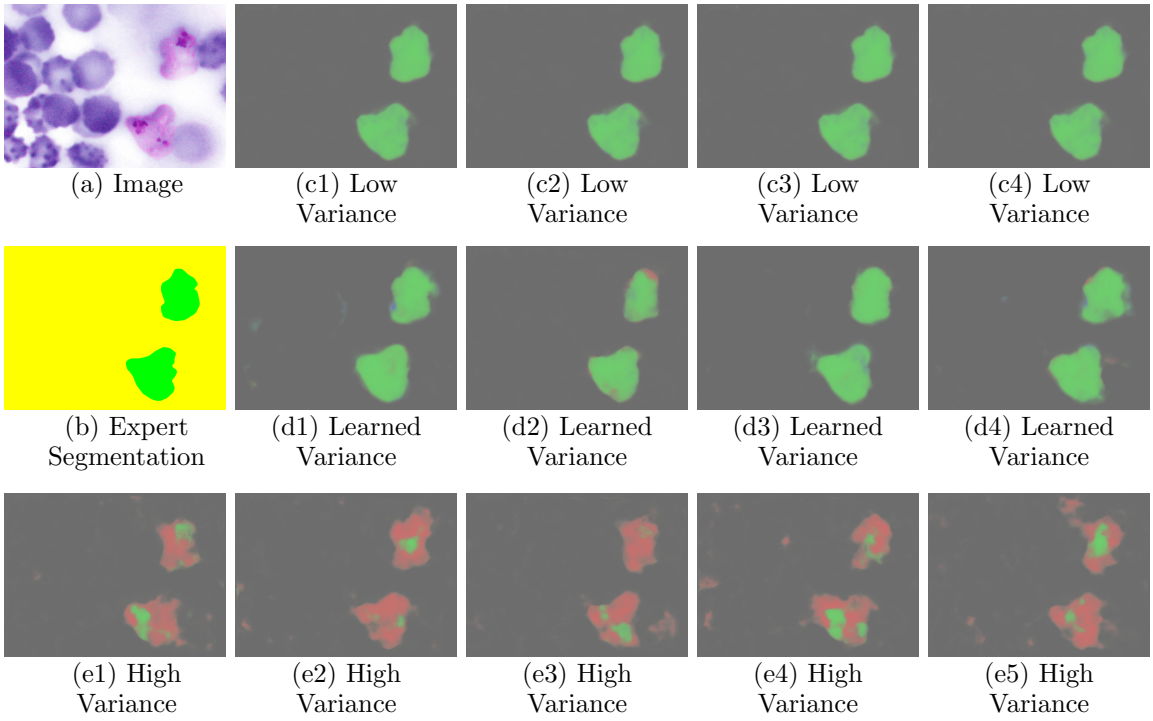


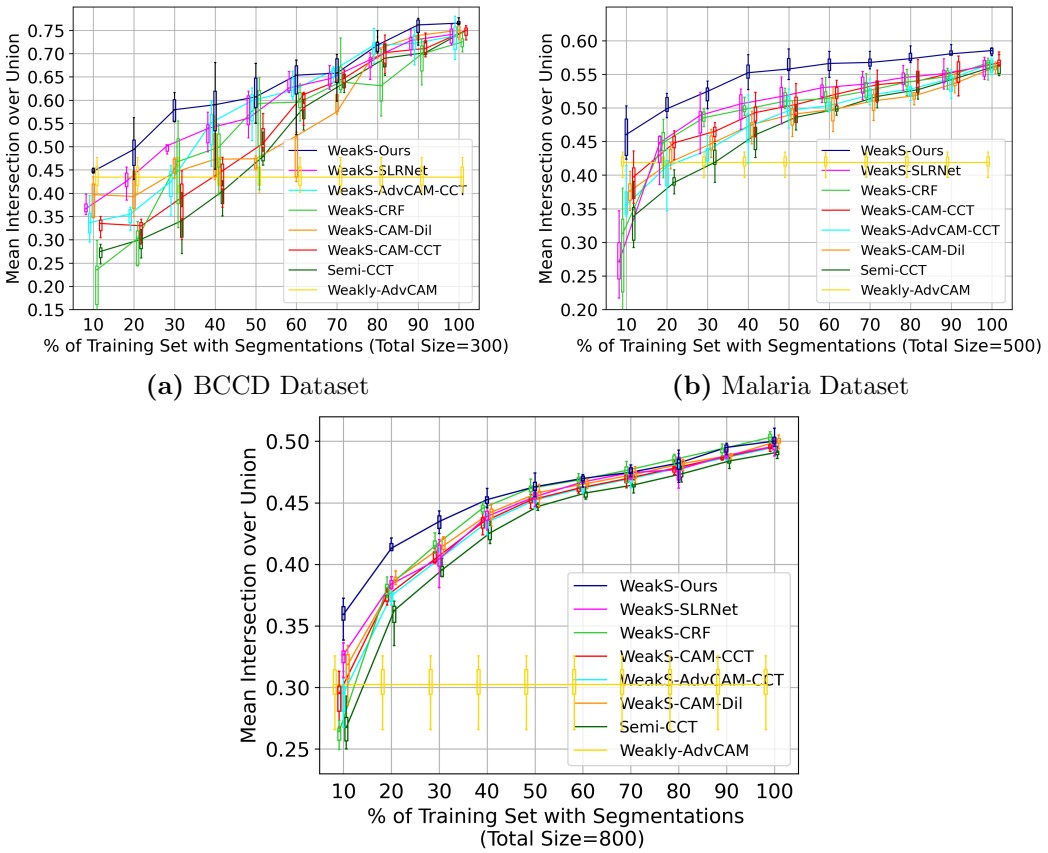
Figure 6: **Empirical Insights Into the MH-Sampling: Qualitative analysis.** Accepted latent-space proposals from the MH sampler, visualized as semantic-probability maps after decoder transformation, using final trained DNNs (at 50% supervision) using different proposal distributions (as described in Table 1).

has low probability mass); and (iii) when the distribution is an isotropic Gaussian with the mean set to the current state and the variances fixed to small values (this risks limiting the exploration of the state space). Table 1 shows the acceptance rates of samples generated using the final trained DNNs at different supervision levels. In general, it is very difficult

to manually tune the per-dimension variances of the proposal distribution in latent space, with the risks involving (i) low acceptance rate of the proposal candidates, (ii) poor exploration by the sampler of the high-probability regions, and (iii) having the sampler stuck in low-probability regions. Figure 6 shows the semantic-probability maps corresponding to the accepted samples generated using the final trained DNNs at the supervision level of 50%. The accepted samples of the MH sampler using proposals from a isotropic Gaussian with small variance (Figure 6(c1)-(c4)) shows imperceptible variability (poor exploration) across the semantic-probability maps. The accepted samples of the MH sampler using proposals from a isotropic Gaussian with large variance (Figure 6(e1)-(e5)) indicates the sampler leading to low-probability regions (semantic-probability maps far from ground-truth) and getting stuck there. The accepted samples of the MH sampler using proposals from a factored-Gaussian with variances learned from our framework (Figure 6(d1)-(d4)) show a high acceptance rate and good variability in the resulting semantic-probability maps that remain close to the ground-truth. Thus, our MH-sampler is able to sample from a proposal distribution that: (i) is computationally efficient and straightforward to sample from, (ii) leads to a high rate of acceptance of the proposed candidates, and (iii) leads to accepted candidates that are good representatives of the variability modeled by the distribution.

4.7 Evaluation of All Methods on Datasets

We evaluate and compare all the methods quantitatively (Figure 7; Table 2, Table 3, Table 4) and several methods qualitatively (Figure 8, Figure 9, Figure 10, Figure 11, Figure 12, Figure 13). For instance, for all the datasets, in the case where very few expert segmentations are available, i.e., when $S \ll (S + U)$, WeakS-Ours outperforms all other methods, quantitatively and qualitatively, in weakly-and-semi-supervised learning. Qualitatively, WeakS-Ours produces higher-quality segmentations compared to all other methods, at virtually all the levels of supervision, by reducing errors in class-label predictions for objects as well as the background, as seen in Figure 8, Figure 9, Figure 10, Figure 11, Figure 12, Figure 13. WeakS-Ours performs better than all baselines (which rely on either CAM-based schemes or a CRF-based scheme for weakly-and-semi-supervised learning) for theoretical reasons detailed in Section 2. Nevertheless, detailed empirical insights into those theoretical arguments appear later in Section 4.10 (for CAM-based schemes), Section 4.9 (for the CRF-based scheme) and Section 4.11 (for WeakS-SLRNet Pan et al. (2022)). Figure 7 (along with Table 2, Table 3, Table 4) shows that WeakS-SLRNet is the best performing method among the baselines for most of the supervision levels and for all the datasets. WeakS-CRF also shows promising results quantitatively and qualitatively despite being an older method as compared to the other baseline methods. WeakS-AdvCAM-CCT performs better than WeakS-CAM-CCT for BCCD dataset, but performs worse for the Malaria dataset, mainly because of the inclusion of a lot of false positives for Malaria dataset during anti-adversarial learning for WeakS-AdvCAM-CCT. The weakly-and-semi-supervised methods WeakS-CAM-CCT, WeakS-AdvCAM-CCT almost always perform better than the semi-supervised method Semi-CCT, by leveraging the weakly labelled data. The weakly-supervised learning method Weakly-AdvCAM is unable to utilize the semantic segmentations available for a part of the training set, and thus shows a constant performance across



(a) BCCD Dataset

(b) Malaria Dataset

(c) Lizard Dataset

Figure 7: **Results: Semantic Segmentation for BCCD, Malaria and Lizard Datasets; Quantitative Analysis.** Performance of all the baseline methods compared with WeakS-Ours, measuring the mean-IoU between (i) the (hard) semantic segmentation output by the method and (ii) the expert semantic segmentation, employing models trained across varying levels of supervision γ. Box plots show variability in test-set performance over randomly selected sets (with fixed cardinality) for training, validation, and testing.

the levels of supervision; it performs better than the other baseline methods only at very low supervision level (10%), but shows lower performance as compared to WeakS-Ours.

We visualize sampled semantic-probability maps and the uncertainty maps (as detailed in Section 4.5) for the test images in Figure 8; Figure 9; Figure 10; Figure 11, Figure 12, Figure 13. The uncertainty map models the per-pixel ambiguity/variability in the estimation of the semantic probability map. For the BCCD dataset, the uncertainty maps (Figure 8(d); Figure 9(d)) show very less uncertainty overall, with some uncertainty near the WBC boundary. In contrast, for the Malaria dataset, the uncertainty maps (Figure 10(d); Figure 11(d)) show much higher values; this is consistent with our qualitative perception of the semantic segmentation task in the Malaria dataset being more difficult than the

Table 2: **Results: Semantic Segmentation for BCCD; Quantitative Analysis.** The median values for the boxplots shown in Figure 7(a).

Supervision Levels →	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
WeakS-Ours	**44.7**	**49.4**	**57.9**	**59.0**	**60.7**	**65.4**	65.8	**71.8**	**76.2**	**76.5**
WeakS-SLRNet	36.7	42.6	50.2	53.9	56.1	62.6	64.6	68.7	72.8	74.1
WeakS-CRF	23.6	30.3	46.8	49.9	59.3	59.6	63.9	63.1	69.9	72.5
WeakS-CAM-CCT	33.5	33.0	39.1	44.8	50.7	61.2	64.7	70.2	71.1	75.0
WeakS-AdvCAM-CCT	33.7	35.5	42.5	55.2	59.9	62.1	**66.4**	**71.8**	72.2	73.8
WeakS-CAM-Dil	39.7	39.3	44.5	47.3	47.4	52.3	57.5	71.0	73.9	75.0
Semi-CCT	27.5	30.1	34.1	40.4	47.9	58.4	63.7	69.0	70.2	75.0
Weakly-AdvCAM	43.5	43.5	43.5	43.5	43.5	43.5	43.5	43.5	43.5	43.5

Table 3: **Results: Semantic Segmentation for Malaria; Quantitative Analysis.** The median values for the boxplots shown in Figure 7(b).

Supervision Levels →	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
WeakS-Ours	**46.0**	**49.9**	**52.5**	**55.2**	**55.8**	**56.6**	**56.8**	**57.3**	**58.0**	**58.5**
WeakS-SLRNet	27.1	44.8	48.9	50.6	51.7	53.0	53.5	54.5	55.1	56.2
WeakS-CRF	31.3	44.8	48.5	49.8	51.0	51.4	52.5	53.8	54.7	56.7
WeakS-CAM-CCT	39.8	44.8	46.5	49.3	50.5	52.0	53.4	54.0	55.5	56.7
WeakS-AdvCAM-CCT	35.6	41.3	43.7	47.2	49.6	50.4	51.9	52.8	54.4	56.6
WeakS-CAM-Dil	37.6	41.9	44.7	47.5	49.0	49.8	51.0	51.8	53.6	56.1
Semi-CCT	33.9	38.9	41.9	46.0	48.6	49.8	51.7	52.7	54.5	56.4
Weakly-AdvCAM	41.9	41.9	41.9	41.9	41.9	41.9	41.9	41.9	41.9	41.9

task in the BCCD dataset, which also reflects in the quantitative analysis in Figure 7 and Table 2, Table 3. To give some insights, for the two infected RBCs in Figure 10(a), the uncertainty in segmentation, as seen in Figure 10(d), is significantly higher for the RBC that has very subtle low-contrast textures without a clear separation between the RBC and its surrounding. Indeed, the semantic segmentation maps (in the second row and third row in Figure 11), across all methods, clearly indicate more errors in pixels corresponding to this particular RBC. We give another example from Figure 11(a), where the three infected RBCs belong to the same class (as seen in the expert segmentation in Figure 11(b)). However, while two of those three RBCs have an appearance that is typical of their class, one

Table 4: **Results: Semantic Segmentation for Lizard; Quantitative Analysis.** The median values for the boxplots shown in Figure 7(c).

Supervision Levels →	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
WeakS-Ours	**35.9**	**41.3**	**43.5**	**45.3**	**46.3**	**47.0**	**47.5**	48.2	**49.5**	50.0
WeakS-SLRNet	32.6	38.4	40.5	43.9	45.5	46.7	47.4	47.7	48.9	49.6
WeakS-CRF	26.4	38.1	41.5	44.6	46.2	46.8	47.6	**48.5**	49.4	**50.3**
WeakS-CAM-CCT	29.5	37.5	40.4	43.4	45.2	46.2	47.0	47.8	48.7	49.5
WeakS-AdvCAM-CCT	29.5	37.5	40.4	43.4	45.2	46.2	47.0	47.8	48.7	49.5
WeakS-CAM-Dil	31.9	38.8	41.5	44.2	45.8	46.6	47.3	48.2	48.9	50.0
Semi-CCT	26.8	36.2	39.6	42.5	44.7	45.8	46.5	47.3	48.4	49.1
Weakly-AdvCAM	30.2	30.2	30.2	30.2	30.2	30.2	30.2	30.2	30.2	30.2

of the RBCs has a darker and denser texture that leads to relatively higher uncertainty for pixels within that RBC (Figure 11(d)). Indeed, the semantic segmentation maps (in the second row and third row in Figure 11), across all methods, clearly indicate more errors in pixels corresponding to this particular RBC. In case of Lizard dataset, the uncertainty values (Figure 12(d); Figure 13(d)) are higher and distributed across the image; this is consistent with our qualitative perception of the semantic segmentation task in the Lizard dataset being even more difficult than the task in the Malaria dataset, which also reflects in the quantitative analysis in Figure 7 and Table 4. To give some insights, the uncertainty in Figure 12(d); Figure 13(d) is observed to be higher in value in and around the nuclei. Indeed, the semantic segmentation maps (in the second row and third row in Figure 12 and Figure 13), across all the methods, show that WeakS-Ours shows better classification

Table 5: **Results: Computational Costs.** Time taken (in minutes) for training DNNs over the Lizard dataset for all methods at 20% and 100% supervision levels. We ran the optimizer for each method for 50 epochs that was sufficient for the optimizers to converge for all methods.

Supervision Level	**WeakS-Ours**	WeakS-SLRNet	WeakS-CRF	WeakS-CAM-CCT	WeakS-AdvCAM-CCT	WeakS-CAM-Dil	Semi-CCT	Weakly-AdvCAM
20%	125	25	12	140	140	12	138	9
100%	96	25	9	83	83	12	82	9

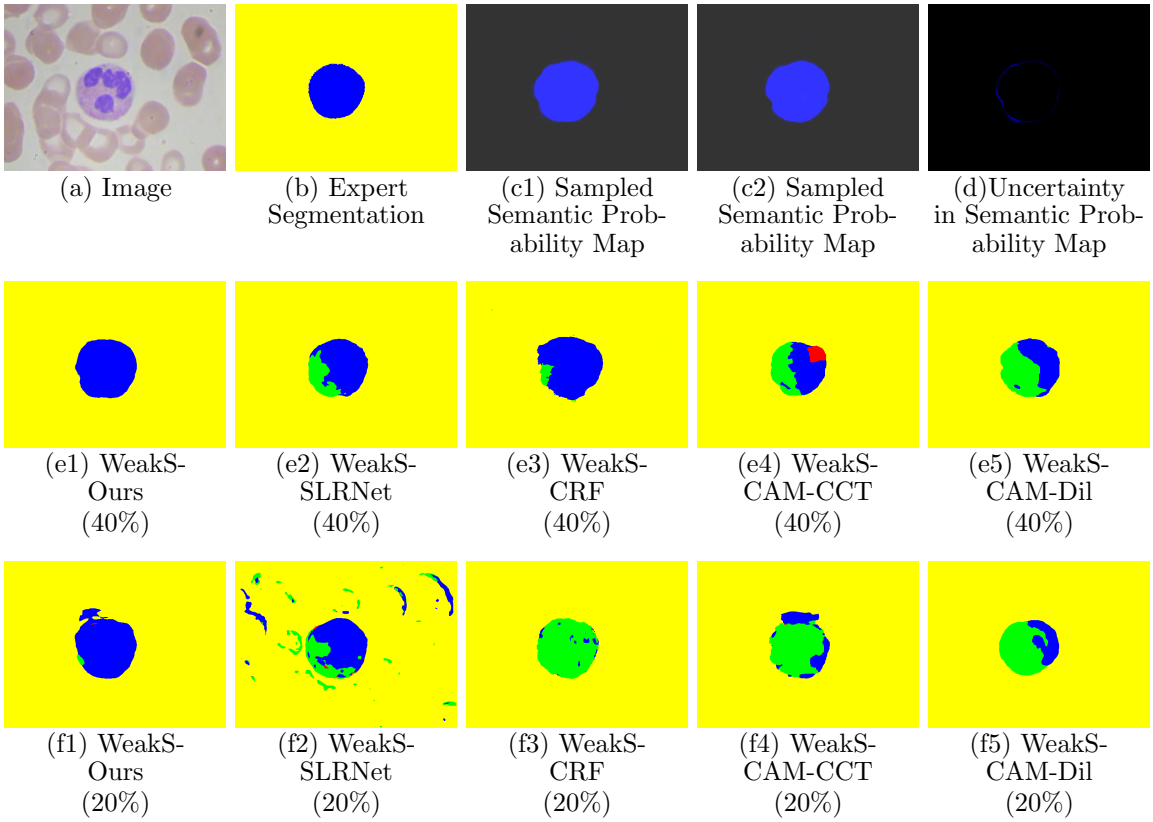
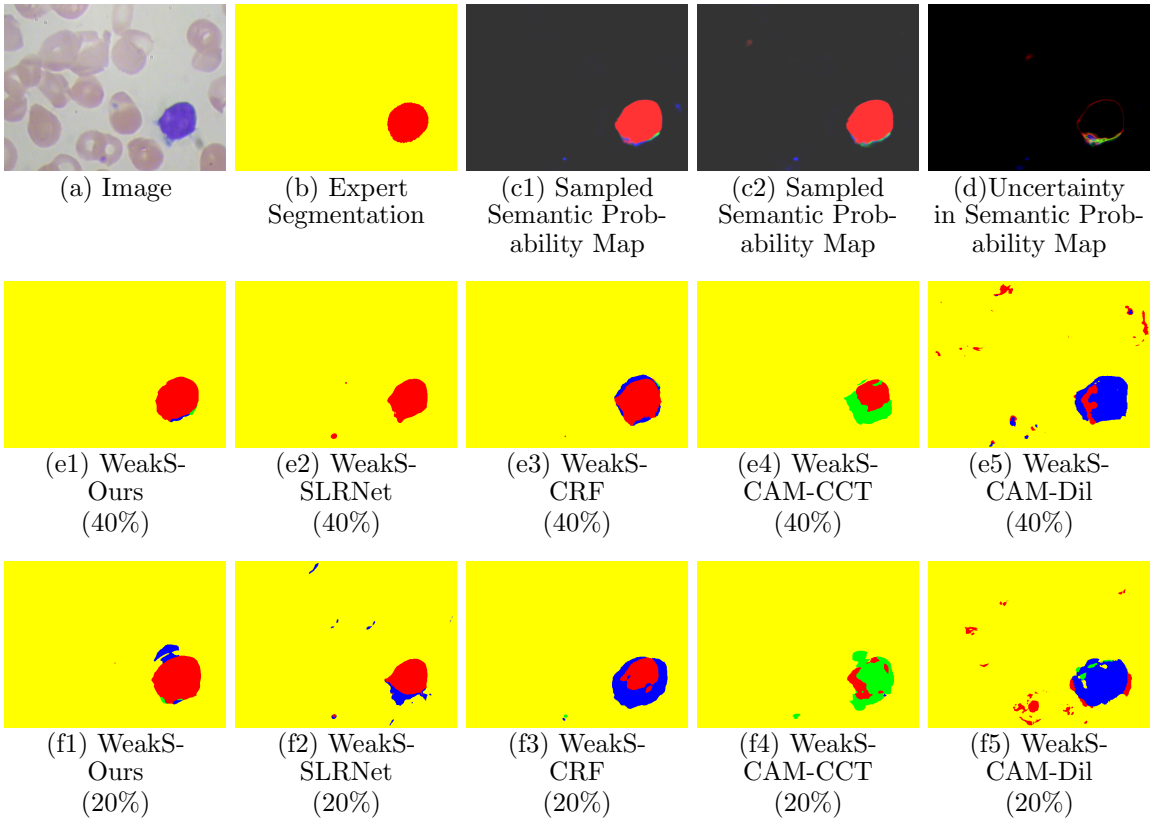


Figure 8: **Results on BCCD Dataset: Qualitative Analysis**. **(a)** Input RGB image and **(b)** its expert-given semantic segmentation. **(c1)-(c2)** Semantic probability maps resulting from sampled latent-space encodings h at $\gamma = 40\%$ supervision for WeakS-Ours. **(d)** Uncertainty map at $\gamma = 40\%$ supervision for WeakS-Ours. **(e1)**–**(e5)** show the results across all methods at $\gamma = 40\%$ supervision. **(f1)**–**(f5)** show the results across all methods at $\gamma = 20\%$ supervision. The visualization scheme for all semantic segmentations and semantic probability maps is described in Section 4.5.

of foreground cells at lower lower supervision levels (speficially for the red class label) and better precision in locating the boundaries of nucleis'.

We also compare *WeakS-Ours* with all the baseline methods with respect to the time taken for training and inference on the Lizard dataset (Table 5). For inference, for the entire test set, WeakS-Ours takes 116 seconds, whereas all the baseline methods take around 20 to 30 seconds. The methods involving sampling, i.e., WeakS-Ours and all CCT-based methods (WeakS-CAM-CCT, WeakS-AdvCAM-CCT, and Semi-CCT), have higher training time. However, the training-time for WeakS-Ours is lower as compared to the CCT based methods, i.e., WeakS-CAM-CCT, WeakS-AdvCAM-CCT, and Semi-CCT. For inference, WeakS-Ours takes more time as compared to the baseline methods owing to sampling (Section 3.6).



(a) Image (b) Expert Segmentation (c1) Sampled Semantic Probability Map (c2) Sampled Semantic Probability Map (d)Uncertainty in Semantic Probability Map

(e1) WeakS-Ours (40%) (e2) WeakS-SLRNet (40%) (e3) WeakS-CRF (40%) (e4) WeakS-CAM-CCT (40%) (e5) WeakS-CAM-Dil (40%)

(f1) WeakS-Ours (20%) (f2) WeakS-SLRNet (20%) (f3) WeakS-CRF (20%) (f4) WeakS-CAM-CCT (20%) (f5) WeakS-CAM-Dil (20%)

Figure 9: **Results on BCCD Dataset: Qualitative Analysis** on another example. The description is the same as in Figure 8.

We also provide a preliminary glimpse into the calibration error of various methods by evaluating their expected calibration error (ECE) (Guo et al., 2017) over the Malaria dataset. We evaluate ECE using 10 bins. The observed ECE values at 20% supervision level for the different methods are: Semi-CCT 0.142, WeakS-AdvCAM-CCT 0.147, WeakS-CAM-CCT 0.152, Weakly-AdvCAM 0.157, WeakS-Ours 0.161, WeakS-SLRNet 0.173, WeakS-CRF 0.207 and WeakS-CAM-Dil 0.223. Lower the ECE, better the calibration. Thus, WeakS-Ours tends to be a bit better calibrated as compared to WeakS-SLRNet, WeakS-CRF, and WeakS-CAM-Dil. The CCT-based methods give a bit lower error ECEs than WeakS-Ours. We also observed that ECE typically decreases with increasing supervision.

4.8 Evaluation of Ablated Versions of Our Method

We construct three ablated versions of our method *WeakS-Ours* as follows: (i) *WeakS-Ours-SingleLevelVariational* restricting the variational modeling only for the coarsest level of the latent-space encoding (as it was there in our preliminary work in), but including modeling each missing segmentation as a random variable within MCEM; (ii) *WeakS-Ours-NonVariational* that removes the ability from WeakS-Ours to learn a variational model, i.e., removes modeling each missing segmentation as a random variable and also removes

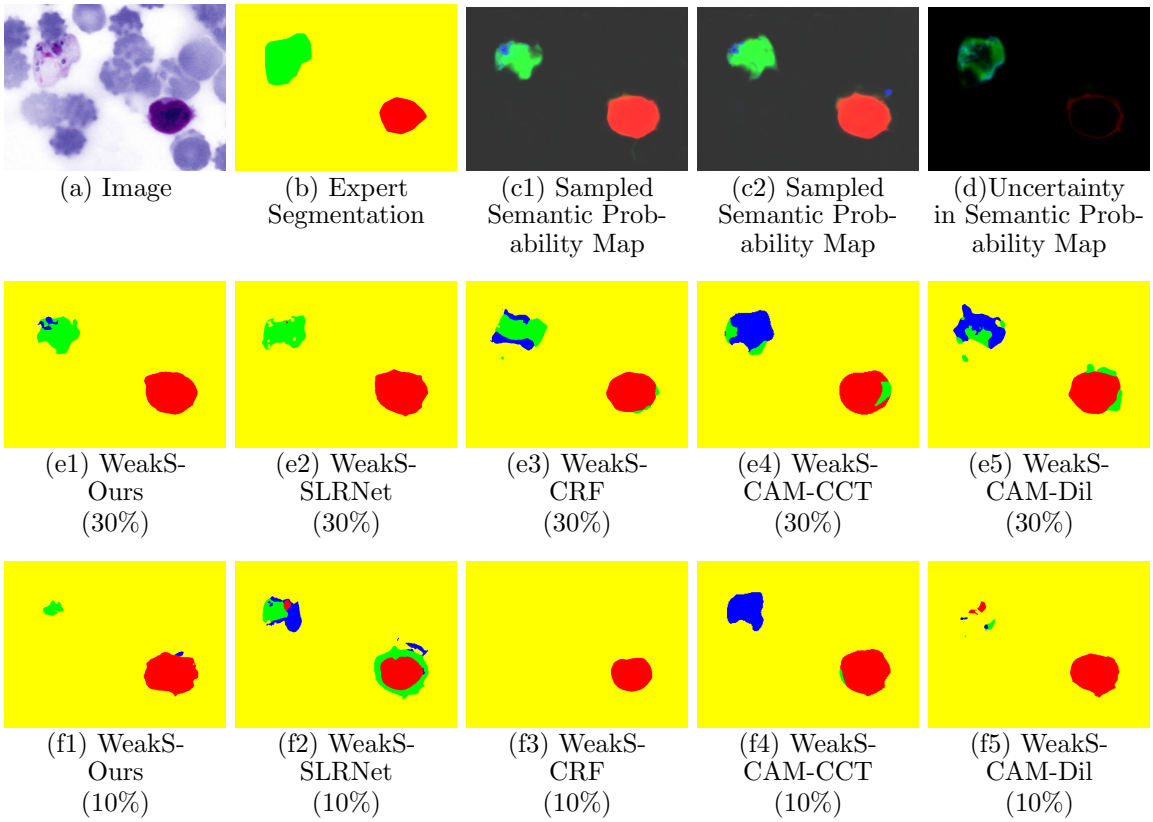
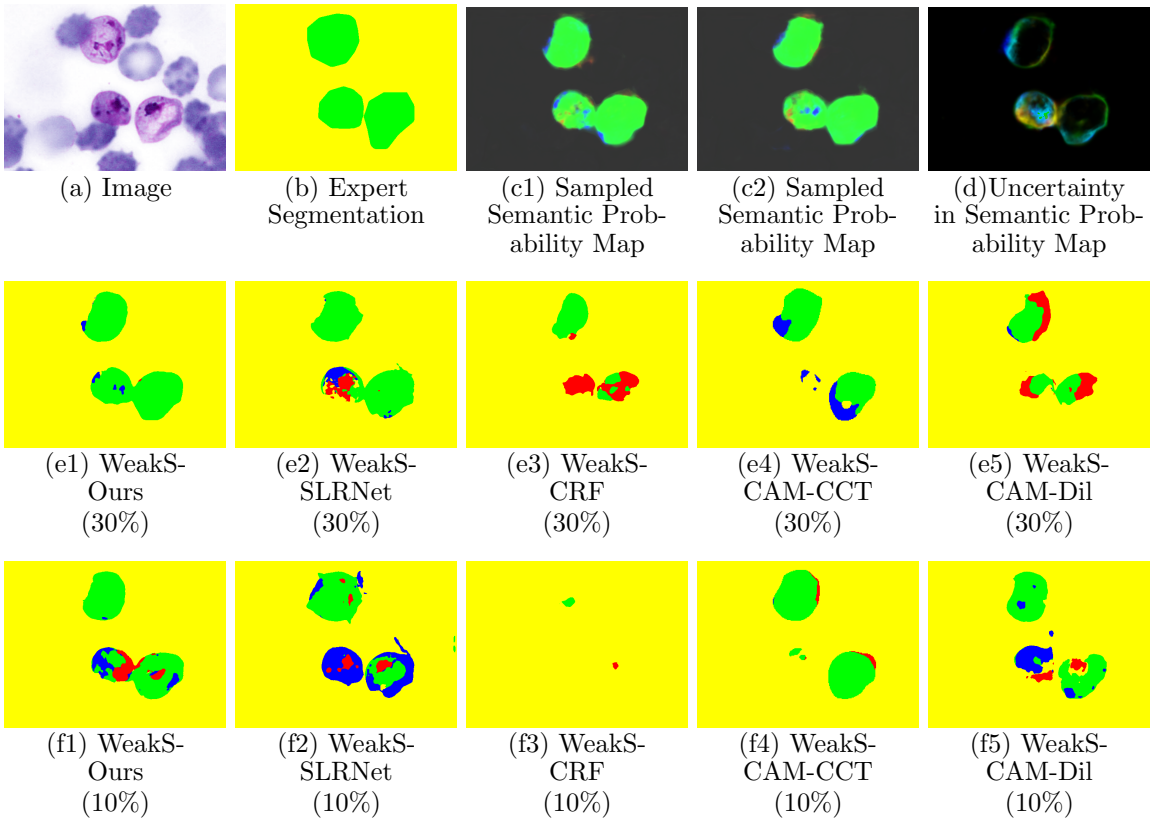


Figure 10: **Results on Malaria Dataset: Qualitative Analysis**. **(a)** Input RGB image and **(b)** its expert-given semantic segmentation. **(c1)-(c2)** Semantic probability maps resulting from sampled latent-space encodings h at $\gamma = 30\%$ supervision for WeakS-Ours. **(d)** Uncertainty map at $\gamma = 30\%$ supervision for WeakS-Ours. **(e1)**–**(e5)** show the results across all methods at $\gamma = 30\%$ supervision. **(f1)**–**(f5)** show the results across all methods at $\gamma = 10\%$ supervision. The visualization scheme for all semantic segmentations and semantic probability maps is described in Section 4.5.

modeling latent-space of the segmenter as a random variable; (iii) *S-Ours-NonVariational* further removes the ability from WeakS-Ours-NonVariational to leverage weak supervision in the form of the image-level class labels in the absence of the per-pixel segmentation.

The quantitative results over the ablated versions of our method in Figure 14, show that, compared to WeakS-Ours, there is: (i) a large drop in performance in WeakS-Ours-NonVariational, indicating the importance of the variational learning within our framework; (ii) a huge drop in performance in S-Ours-NonVariational, indicating the utility of weak supervision in the form of image-level labels; and (iii) a statistically significant drop, at low levels of supervision, in performance in WeakS-Ours-SingleLevelVariational, indicating the benefits of multi-level variational modeling in latent space.



(a) Image | (b) Expert Segmentation | (c1) Sampled Semantic Probability Map | (c2) Sampled Semantic Probability Map | (d)Uncertainty in Semantic Probability Map

(e1) WeakS-Ours (30%) | (e2) WeakS-SLRNet (30%) | (e3) WeakS-CRF (30%) | (e4) WeakS-CAM-CCT (30%) | (e5) WeakS-CAM-Dil (30%)

(f1) WeakS-Ours (10%) | (f2) WeakS-SLRNet (10%) | (f3) WeakS-CRF (10%) | (f4) WeakS-CAM-CCT (10%) | (f5) WeakS-CAM-Dil (10%)

Figure 11: **Results on Malaria Dataset: Qualitative Analysis** on another example. The description is the same as in Figure 10.

Although the key focus of the proposed approach is on semantic segmentation, to gain further insights into the working of translator, we provide some empirical analysis for the image-level classification task. Figure 15 shows the evaluation of the accuracy of image-level class-label outputs produced by the translator within our framework; for all the datasets, WeakS-Ours improves over S-Ours for the task of classification by leveraging weakly-and-semi-supervised learning through the training-set images x_u paired with their image-level class labels y_u (without the associated expert segmentations). We also analyzed the image-level class-probability vectors along with their uncertainty vectors, associated with the test images in Figure 8, Figure 9 for BCCD dataset; and Figure 10, Figure 11 for Malaria dataset. For the BCCD-dataset example in Figure 8, the image-level class-probability vector and the associated uncertainty values output by WeakS-Ours were: $[0.008, 0.005, 0.992]$ and $[0.003, 0.001, 0.003]$. For another example of BCCD-dataset in Figure 9, the values were: $[0.991, 0.011, 0.014]$ and $[0.003, 0.006, 0.008]$ respectively. For the Malaria-dataset example in Figure 10, the image-level class-probability vector and the associated uncertainty values output by WeakS-Ours were: $[0.994, 0.962, 0.021]$ and $[0.001, 0.092, 0.041]$. For another example of Malaria-dataset in Figure 11, the image-level class-probability vector and the associated uncertainty values were: $[0.026, 0.999, 0.049]$ and $[0.018, 0.0007, 0.08]$ respectively.

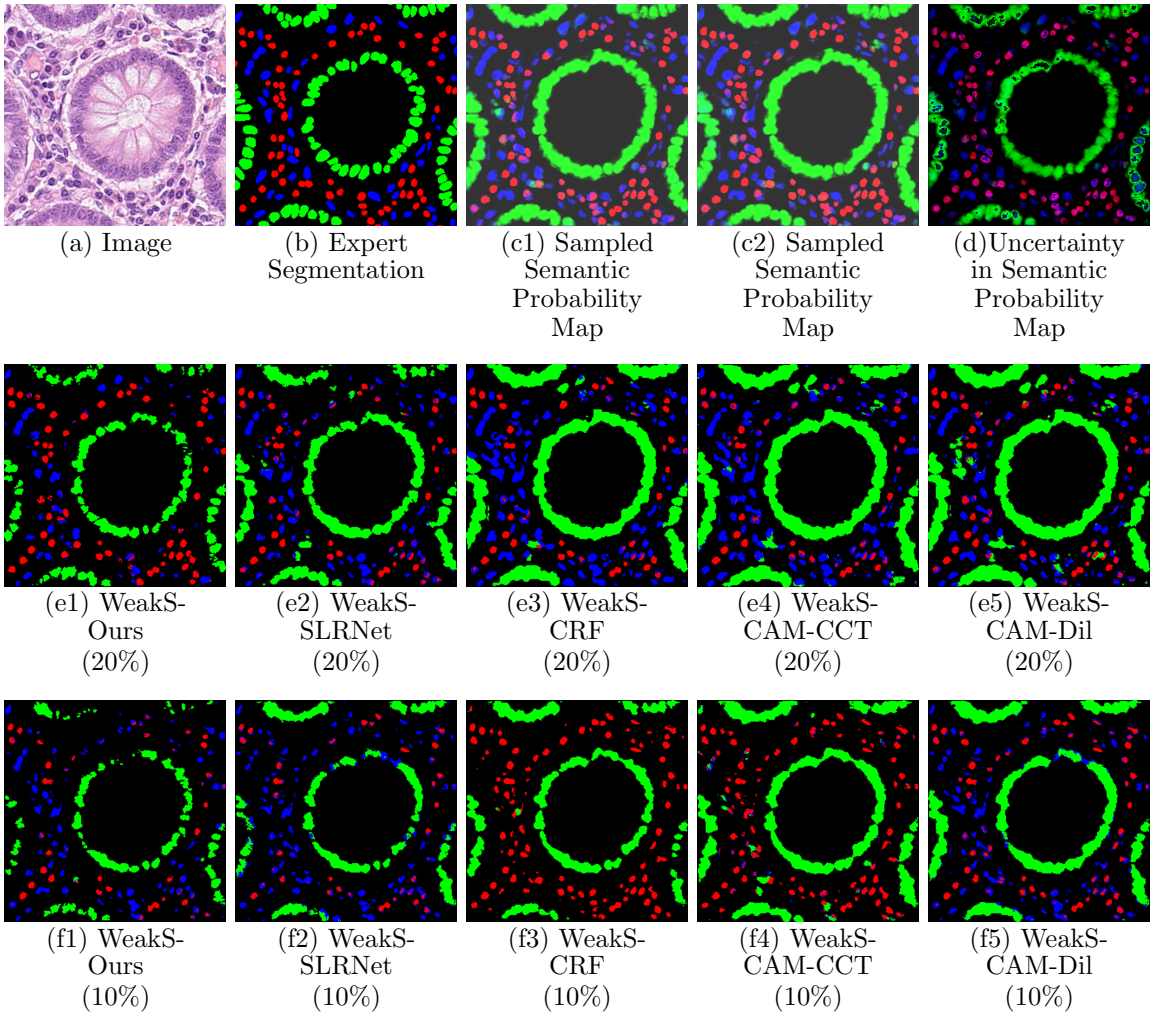


Figure 12: **Results on Lizard Dataset: Qualitative Analysis**. **(a)** Input RGB image and **(b)** its expert-given semantic segmentation. **(c1)-(c2)** Semantic probability maps resulting from sampled latent-space encodings h at $\gamma = 20\%$ supervision for WeakS-Ours. **(d)** Uncertainty map at $\gamma = 20\%$ supervision for WeakS-Ours. **(e1)**–**(e5)** show the results across all methods at $\gamma = 20\%$ supervision. **(f1)**–**(f5)** show the results across all methods at $\gamma = 10\%$ supervision. The visualization scheme for all semantic segmentations and semantic probability maps is described in Section 4.5.

As compared to the Malaria dataset examples, the examples from the BCCD dataset show a higher confidence in the image-level class predictions, as indicated by (i) higher probabilities output for the classes of objects present in the image, (ii) lower probabilities output for the classes of objects absent in the image, and (iii) lower uncertainty values associated with all the classes. For the Malaria-dataset example in Figure 10, compared to the other classes, there is larger uncertainty in the trophozoite class associated with the

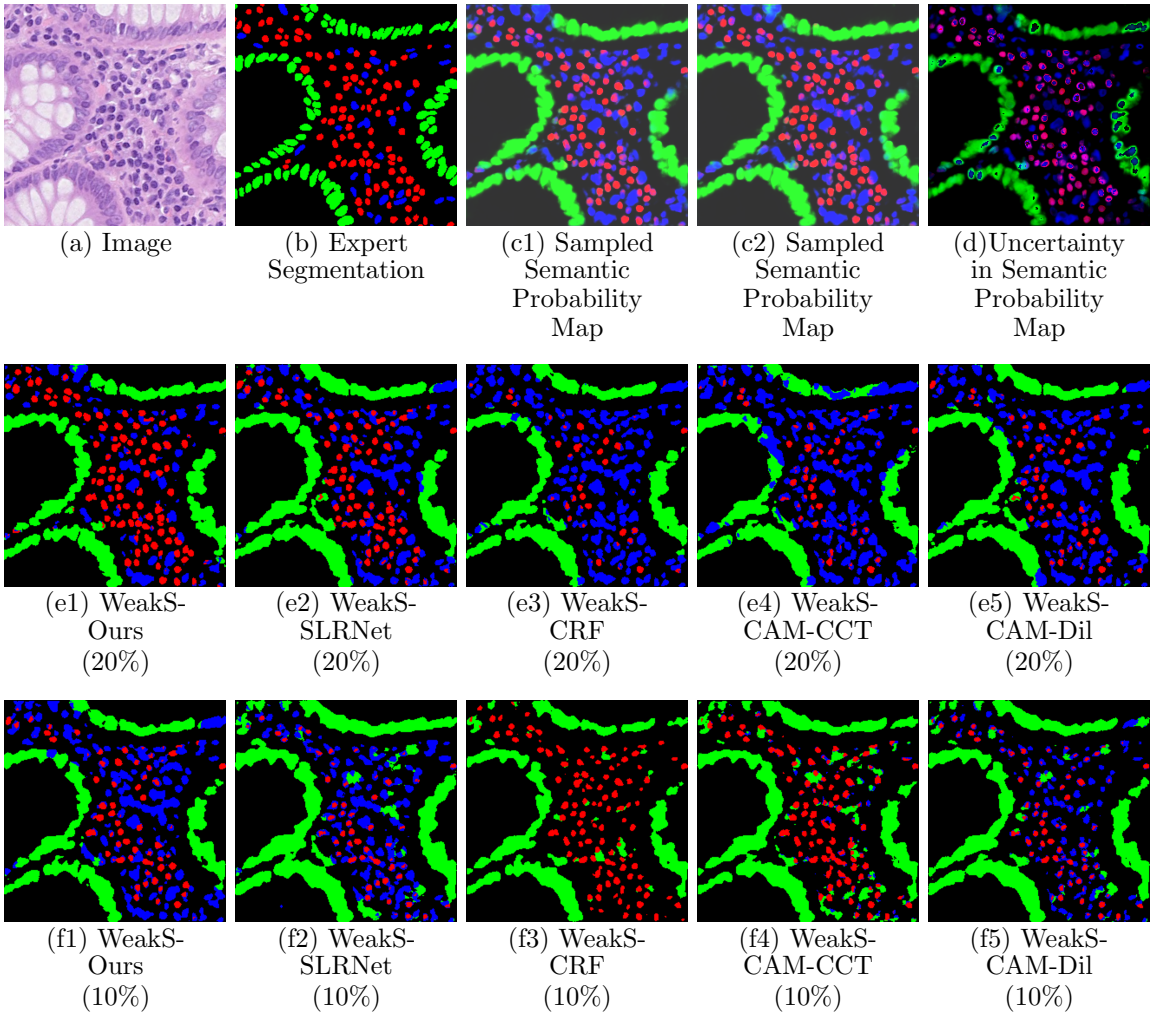
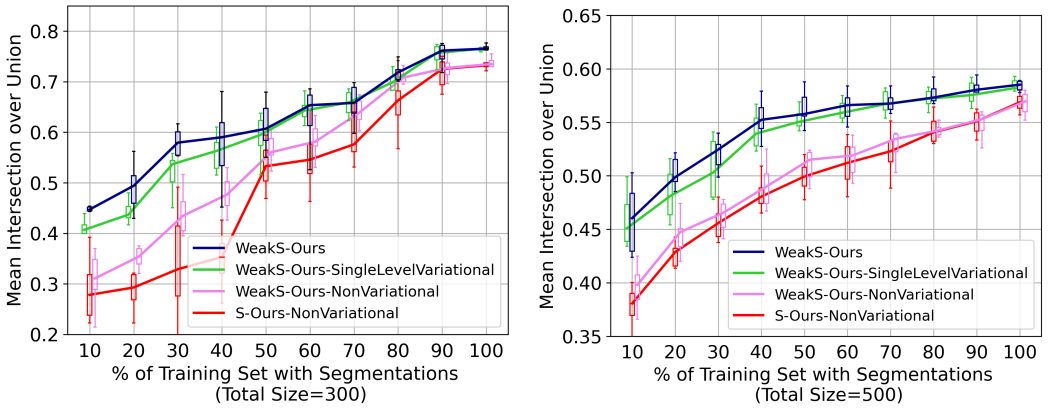


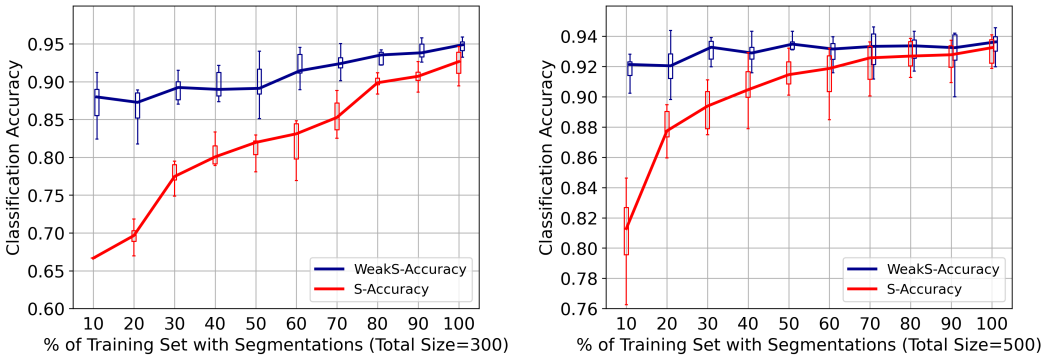
Figure 13: **Results on Lizard Dataset: Qualitative Analysis** on another example. The description is the same as in Figure 12.

RBC at the top of the image (labeled green in the expert segmentation in Figure 10(b)), which stems from the semantic probability maps for that RBC being more imprecise (Figure 10(c1)-(c2),(e1)) and more variable (Figure 10(d)), which in turn is because of the subtle and low-contrast textures in that RBC (Figure 10(a)). For the Malaria-dataset example in Figure 11, despite the higher uncertainty in the per-pixel semantic-segmentation probabilities for one of the three trophozoite RBCs, the uncertainty in the image-level probability vector for the trophozoite class remains low because, for the other two trophozoite RBCs, the per-pixel semantic probabilities are accurate (Figure 11(c1)-(c2)) consistently (Figure 11(d)).



(a) BCCD Dataset **(b)** Malaria Dataset

Figure 14: **Results: Results on Ablated Versions of Our Method.** Performance of the ablated versions of WeakS-Ours, i.e., WeakS-Ours-SingleLevelVariational, WeakS-Ours-NonVariational, and S-Ours-NonVariational, measuring the mean-IoU at increasing levels of supervision γ.



(a) BCCD Dataset **(b)** Malaria Dataset

Figure 15: **Results: Classification for BCCD and Malaria Datasets, Quantitative Analysis.** Performance of our methods, measuring the accuracy in predicting the image-level class labels, employing models trained across varying levels of supervision γ. The box plots show the variability in the test-set performance over randomly selected sets (with fixed cardinality) for training, validation, and testing.

4.9 Empirical Insights into CRF-Based Baseline WeakS-CRF (Papandreou et al., 2015)

WeakS-CRF (Papandreou et al., 2015) leverages the image-level class label for weak supervision and semi supervision by designing a bias function using a CRF to estimate the missing expert segmentation. However, the CRF-based scheme, while having the advantage of simplicity relative to other DNN-based models, has some major limitations as shown in the

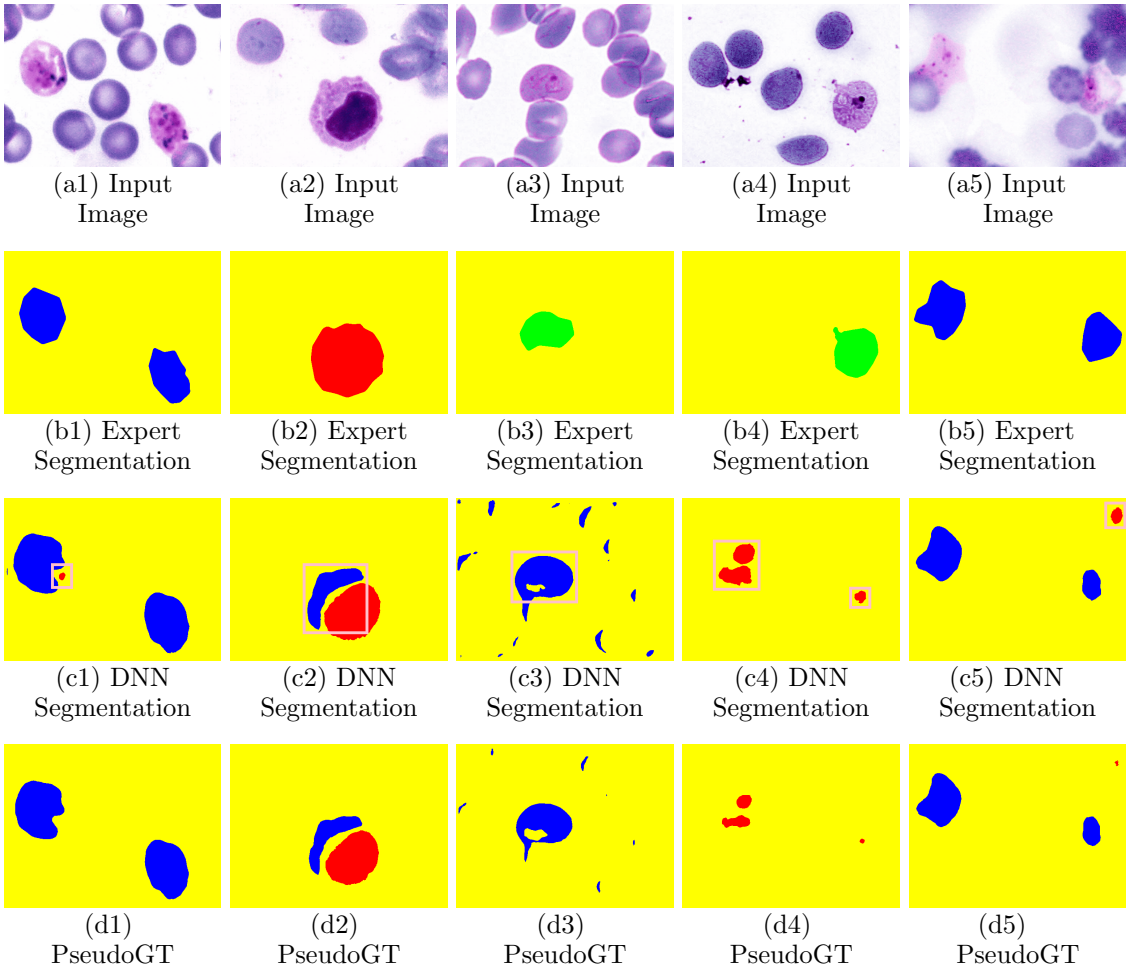


Figure 16: **Empirical Insights into CRF-Based Baseline WeakS-CRF (Papandreou et al., 2015). (a1)–(a5)** Input RGB images, and **(b1)–(b5)** their expert-given semantic segmentations. **(c1)–(c5)** Semantic segmentations produced by the DNN component in the WeakS-CRF model trained at the level of supervision $\gamma = 30\%$. **(d1)–(d5)** Pseudo-ground-truth semantic segmentations estimated using the CRF component, through its bias-introduction scheme, in the WeakS-CRF model trained at the level of supervision $\gamma = 30\%$.

several representative examples in Figure 16. First, for the images in Figure 16(a1) and Figure 16(a5), the expert segmentations (Figure 16(b1),(b5)) indicate the presence of the classes depicted by the colors blue (for the object) and yellow (for the background). Thus, the CRF-introduced bias (desirably) modifies the predicted segmentations (Figure 16(c1),(c5)) to promote the blue and yellow classes and, thereby, removing/reducing the erroneous red region; WeakS-CRF then employs the resulting modified discrete segmentation (a "pseudo ground-truth"; (Figure 16(d1),(d5))) for further training of the DNN. However, the same strategy fails to work well in the examples in Figures 16(a2)–(d2), Figures 16(a3)–(d3),

and Figures 16(a4)–(d4) because, first, the DNN segmentation detects a large object of a class that is absent in the image and, subsequently, the bias modification scheme only marginally shrinks the object size in the pseudo-ground-truth without being able to remove it completely. Now that the pseudo-ground-truth indicates an object of a class that is contradictory to that indicate by the expert, this pseudo-ground-truth can seriously mislead the subsequent DNN training. In addition, in the examples in Figures 16(a3)–(d3) and Figures 16(a4)–(d4), the DNN first fails to detect an object (green) that is present in the image, and, subsequently, the bias modification is unable to insert any pixel corresponding to the missing object (green) on the pseudo-ground-truth. Now that the pseudo-ground-truth fails to indicate any pixel belonging to an object class (green) that should be present in the image (as per the expert), this pseudo-ground-truth can seriously mislead the subsequent DNN training. The methods within our framework avoid creating such pseudo-ground-truth estimation that are contradictory to the image-level class labels, and leverages the translator to continuously maintain a differentiable functional connection between the semantic probability maps and the image-level class label for reliable end-to-end learning.

4.10 Empirical Insights into CAM-Based Baselines

The CAM based methods of WeakS-CAM-Dil (Wei et al., 2018), WeakS-CAM-CCT (Ouali et al., 2020), WeakS-AdvCAM-CCT (Lee et al., 2021), and Weakly-AdvCAM (Lee et al., 2021) rely on CAMs, during training, to estimate the missing segmentations by first computing the CAMs for all classes (using a pre-trained classifier DNN and the known image-level class labels) and subsequently using the CAMs to generate a discrete pseudo-ground-truth semantic segmentation. Figure 17 shows some examples of CAM-based pseudo-ground-truth segmentations. For WeakS-CAM-Dil and WeakS-CAM-CCT, the generation of CAMs relies on the same pre-trained ResNet classifier employing dilated convolutions. For WeakS-AvdCAM-CCT and WeakS-AdvCAM, the generation of CAMs relies on the pre-trained ResNet classifier employing adversarial climbing using the CAM outputs by the DNN.

For the images in Figure 17(a1)-(a2), compared to their expert segmentations (Figure 17(b1)-(b2)), the generated pseudo-ground-truth segmentations in Figure 17(c1)-(c2) show regions in the background mistakenly labeled as parts of objects. Such errors in the pseudo-ground-truth segmentations can seriously mislead the DNN training. On the other hand, for the images in Figure 17(a3)-(a4), compared to their expert segmentations (Figure 17(b3)-(b4)), the generated pseudo-ground-truth segmentations in Figure 17(c3)-(c4) show parts of objects mistakenly labeled as background, effectively shrinking the object size in the pseudo-ground-truth segmentation. Such errors in the pseudo-ground-truth segmentation also mislead the DNN training. Figure 17(a5)-(c5) show another kind of error in generating the CAM-based pseudo-ground-truth segmentation when the object of interest is misclassified. Because all such pseudo-ground-truth segmentations, across various examples in Figure 17, are pre-computed before DNN training, the resulting mis-training of the DNNs can outweigh the benefits of incorporating the weak/semi supervision.

For training images with missing segmentations, in addition to the CAM-based pseudo-ground-truth segmentation, WeakS-CAM-Dil employs a heuristic to iteratively recreate another hard pseudo-ground-truth segmentation from the DNN-output probability maps by eliminating pixel labels outside the set of labels in expert-given image-level class labels. Such

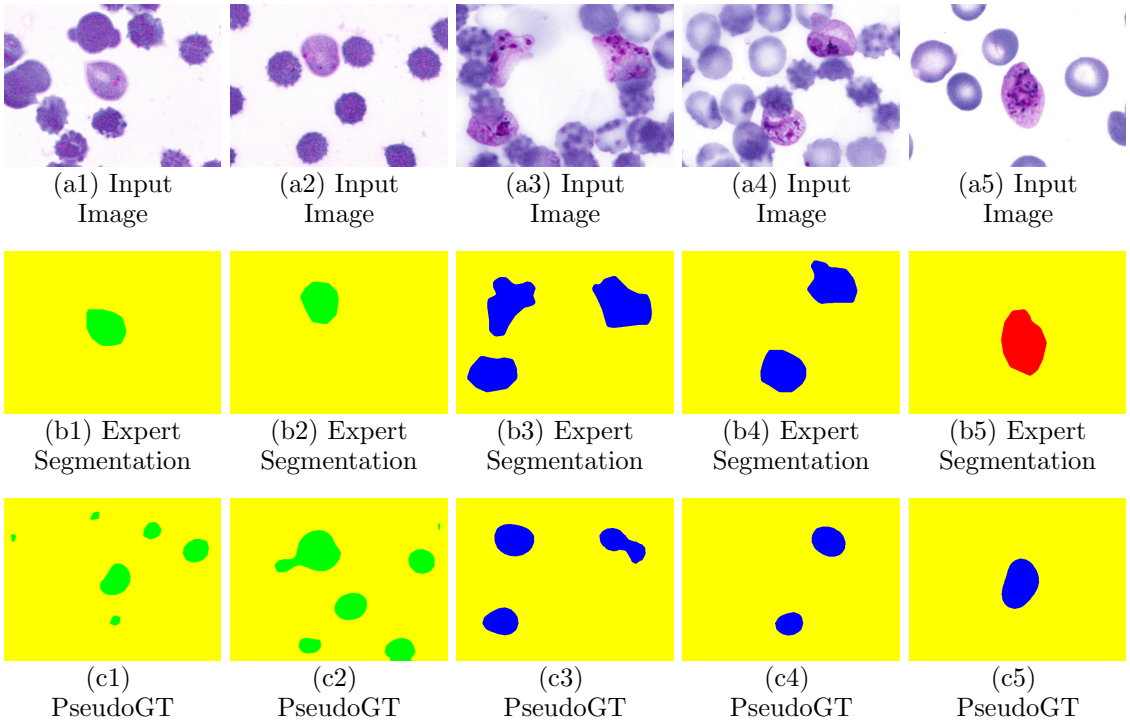


Figure 17: **Empirical Insights into CAM-Based Baselines. (a1)–(a5)** Input RGB images, and **(b1)–(b5)** their expert-given semantic segmentations. **(c1)–(c5)** Pseudo-ground-truth semantic segmentations estimated from the CAMs obtained using a pre-trained classifier.

an operation fails to lead to a differentiable objective function for backpropagation based optimization. In contrast, our framework employs a translator DNN to connect the semantic probability maps with the image-level class labels producing a differentiable objective function (using MH sampling and reparametrization) for end-to-end learning.

WeakS-CAM-CCT, in addition to the CAM-based pseudo-ground-truth segmentation, employs a heuristic in their cross-consistency training strategy that relies on applying a random perturbation using adhoc schemes. WeakS-AdvCAM-CCT follows WeakS-CAM-CCT for weakly-and-semi-supervised learning, but constructs the CAM-based pseudo-ground-truth segmentation after employing adversarial climbing on the CAM outputs of the DNN. In contrast, our variational framework relies on random sampling from the posterior distribution of the latent-space hidden variable, which follows naturally from our design of the EM optimization framework. Moreover, unlike our framework, WeakS-CAM-CCT and WeakS-AdvCAM-CCT rely on many auxiliary decoders within their segmenter, thereby increasing the number of DNN parameters in their model.

4.11 Empirical Insights into WeakS-SLRNet (Pan et al., 2022)

WeakS-SLRNet (Pan et al., 2022) leverages image-level class labels for weak-and-semi-supervision by having an image-level classification loss on the class-wise prediction of segmentation masks summed across pixels. For the images with missing expert segmentations, WeakS-SLRNet uses a refinement module on top of multi-view masks calibration to generate a pseudo-ground-truth segmentation. WeakS-SLRNet employs a low-rank module that maps the high-dimensional latent vector to a lower dimension by decomposing it into lower dimensional dictionary matrix and encoding matrix. The image-level class probabilities are constructed using per-pixel segmentation outputs of the DNN which are prone to errors when learned using a training set with low-supervision level. In contrast, our framework relies on a classifier DNN that is learned using the entire training set and takes as input the combination of original image and the semantic probability maps to provide concrete class probabilities that are also indicative of the quality of segmentation. WeakS-SLRNet uses a heuristic approach to generate pseudo-ground-truths using a cross-view based refinement module. Moreover, unlike our framework that has straightforward neural-network layers at each level for mapping features to a low-dimensional latent vector, WeakS-SLRNet uses a complex cross-view based low-rank-factorization module for reducing the dimension of the

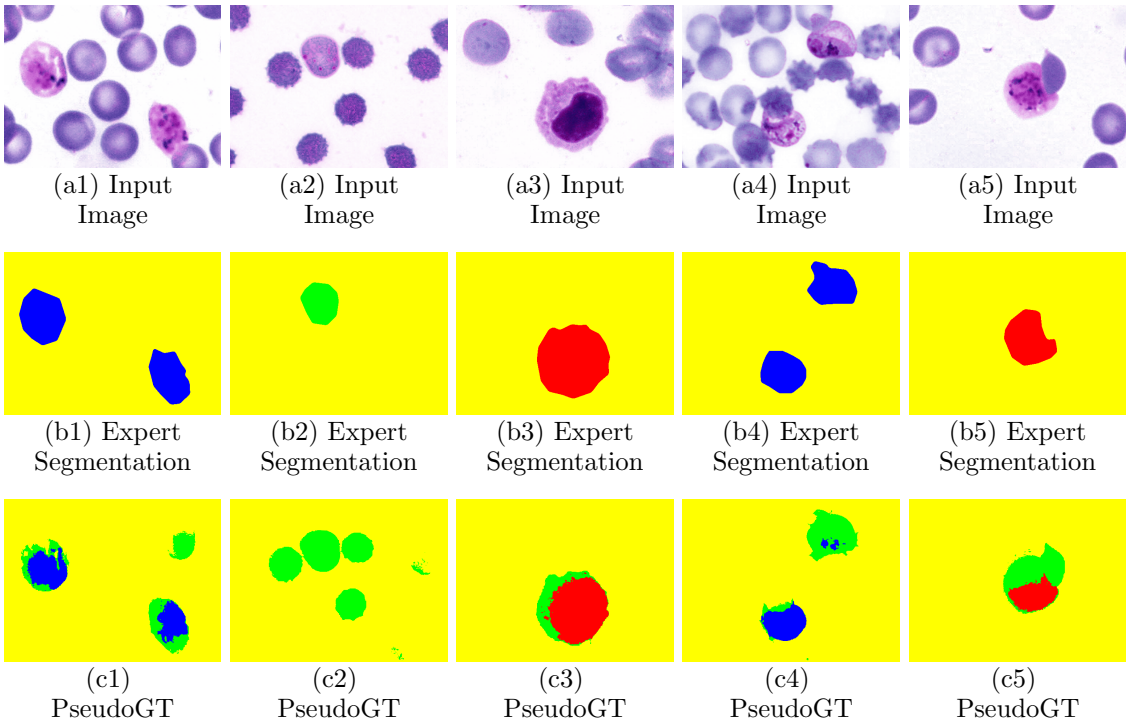


Figure 18: **Empirical Insights into WeakS-SLRNet (Pan et al., 2022)**. **(a1)–(a5)** Input RGB images, and **(b1)–(b5)** their expert-given semantic segmentations. **(c1)–(c5)** Pseudo-ground-truth semantic segmentations estimated by WeakS-SLRNet.

latent vector. WeakS-SLRNet relies on fixed multi-view approach for self-supervision. In contrast, our variational framework has a theoretically sound way of handling missing segmentations through expectation over the samples generated from the posterior distribution of the latent-space hidden variable. WeakS-SLRNet does a fairly good job in estimating the pseudo-ground-truth segmentations, but it has some limitations as shown in the several representative examples in Figure 18. For the images in Figure 18(a1)-(a2), compared to their expert segmentations (Figure 18(b1)-(b2)), the generated pseudo-ground-truth segmentations in Figure 18(c1)-(c2) show background regions mislabeled as object parts. For the images in Figure 18(a3)-(a5), compared to their expert segmentations (Figure 18(b3)-(b5)), the generated pseudo-ground-truth segmentations in Figure 18(c3)-(c5) show that a part of the object is incorrectly labelled. Figure 18(a5)-(c5) show another kind of error in generating pseudo-ground-truth segmentations, when one of the objects of interest is misclassified. All such errors in the pseudo-ground-truth segmentations can seriously mislead the DNN training. Even though the class labels are known for all the images, the classification loss has hardly any impact for large misclassifications because the classifier aggregates the probabilities of all the pixels class-wise; thus, the weakly-supervised examples are unable to guide the DNN towards much better learning. Our method uses the translator DNN to effectively reject the sampled segmentations depicting incorrect regions and misclassifications, as well as the false prediction of foreground object in-place of background.

5. Conclusion

We propose a novel variational DNN framework for the semantic segmentation of blood-tissue microscopy images relying on weakly-and-semi-supervised learning from a training set comprising (i) a few images having per-pixel semantic segmentations and (ii) all images having class labels for the objects of interest present within. To enable weakly-and-semi-supervised learning, our framework couples semantic segmentation with image classification, with end-to-end learning. We propose a novel variational framework relying on MCEM based learning, infering a posterior distribution on the hidden variable modeling the segmenter-DNN's multiscale latent space. We propose a MH sampler for the posterior distribution, along with sample reparametrizations to enable end-to-end backpropagation. During inference on test images, our variational framework can inform about the uncertainty associated with the probabilistic per-pixel semantic segmentation and the probabilistic image-level classification. We provide empirical analysis that sheds key insights into the model designs and capabilities underlying various weakly-and-semi-supervised learning schemes by visualizing intermediate outputs/images of the DNNs underlying various methods. Results on three publicly available real-world datasets show the benefits of our framework.

Acknowledgments

The authors are grateful for support from the Infrastructure Facility for Advanced Research and Education in Diagnostics grant funded by Department of Biotechnology (DBT), Government of India (BT/INF/22/SP23026/2017).

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

Authors do not have any conflicts of interest.

References

J Ahn and S Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *IEEE Comp. Vis. Pattern Recog.*, pages 4981–4990, 2018.

J Ahn, S Cho, and S Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Comp. Vis. Pattern Recog.*, page 2209, 2019.

Stephanie Allassonniere, Estelle Kuhn, and Alain Trouve. Construction of Bayesian deformable models via a stochastic approximation algorithm: A convergence study. *Bernoulli*, 16(3):641–78, 2010.

P Arbelaez, B Hariharan, C Gu, S Gupta, L Bourdev, and J Malik. Semantic segmentation using regions and parts. In *IEEE Comp. Vis. Pattern Recog.*, pages 3378–3385, 2012.

V Badrinarayanan, A Kendall, and R Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Tran. Pattern Anal. Mach. Intell.*, 39(12):2481–2495, 2017.

D Batra, P Yadollahpour, A Guzman-Rivera, and G Shakhnarovich. Diverse m-best solutions in markov random fields. In *Euro. Conf. Comp. Vis.*, pages 1–16, 2012.

S Chandra and I Kokkinos. Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In *Euro. Conf. Comp. Vis.*, page 402, 2016.

Y Chang, Q Wang, W Hung, R Piramuthu, Y Tsai, and M Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *IEEE Comp. Vis. Pattern Recog.*, pages 8991–9000, 2020.

L Chen, Y Yang, J Wang, W Xu, and A Yuille. Attention to scale: Scale-aware semantic image segmentation. In *IEEE Comp. Vis. Pattern Recog.*, pages 3640–3649, 2016.

L Chen, Y Zhu, G Papandreou, F Schroff, and H Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Euro. Conf. Comp. Vis.*, pages 833–851, 2018.

L Chen, W Wu, C Fu, X Han, and Y Zhang. Weakly supervised semantic segmentation with boundary exploration. In *Euro. Conf. Comp. Vis.*, pages 347–362, 2020.

L Chen, T Yang, X Zhang, W Zhang, and J Sun. Points as queries: Weakly semi-supervised object detection by points. In *IEEE/CVF Comp. Vis. Pattern Recog.*, pages 8823–8832, 2021.

D Ciresan, A Giusti, L Gambardella, and J Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Adv. Neural. Info. Proc. Sys.*, pages 2843–2851, 2012.

J Dunnmon, D Yi, C Langlotz, R Christopher, D Rubin, and M Lungren. Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology*, 290(2):537–544, 2019.

J Fan, Z Zhang, C Song, and T Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *IEEE Comp. Vis. Pattern Recog.*, pages 4283–4292, 2020.

X Fu, N Cai, K Huang, H Wang, P Wang, C Liu, and H Wang. M-Net: A novel u-net with multi-stream feature fusion and multi-scale dilated convolutions for bile ducts and hepatolith segmentation. *IEEE Access*, 7:148645, 2019.

A Gaikwad and S Awate. Deep mcem for weakly-supervised learning to jointly segment and recognize objects using very few expert segmentations. In *Conf. Inf. Proc. Med. Imag.*, pages 624–636, 2021.

S Graham, M Jahanifar, A Azam, M Nimir, Y Tsang, K Dodd, E Hero, H Sahota, A Tank, K Benes, et al. Lizard: A large-scale dataset for colonic nuclear instance segmentation and classification. In *Int. Conf. Comp. Vis.*, pages 684–693, 2021.

C Guo, G Pleiss, Y Sun, and K Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330, 2017.

B Hariharan, P Arbelaez, R Girshick, and J Malik. Simultaneous detection and segmentation. In *Euro. Conf. Comp. Vis.*, pages 297–312, 2014.

B Hariharan, P Arbelaez, R Girshick, and J Malik. Hypercolumns for object segmentation and fine-grained localization. In *IEEE Comp. Vis. Pattern Recog.*, pages 447–456, 2015.

K He, X Zhang, S Ren, and J Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Tran. Pattern Anal. Mach. Intell.*, 37(9):1904–1916, 2015.

K He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition. In *IEEE Comp. Vis. Pattern Recog.*, pages 770–778, 2016.

K He, G Gkioxari, P Dollar, and R Girshick. Mask R-CNN. In *Int. Conf. Comp. Vis.*, pages 2980–2988, 2017.

S Hong, H Noh, and B Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *Neural. Info. Proc. Sys.*, page 1495, 2015.

G Huang, Z Liu, L van der Maaten, and K Weinberger. Densely connected convolutional networks. In *Comp. Vis. Pattern Recog.*, page 4700, July 2017.

W Hung, Y Tsai, Y Liou, Y Linand, and M Yang. Adversarial learning for semi-supervised semantic segmentation. In *Brit. Mach. Vis. Conf.*, page 65, 2018.

H Kang, H Park, Y Ahn, A Messem, and W Neve. Towards a quantitative analysis of class activation mapping for deep learning-based computer-aided diagnosis. In *Medical Imaging*, pages 119–131, 2021.

A Kendall and Y Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Adv. Neural. Info. Proc. Sys.*, pages 5574–5584, 2017.

D Kingma and M Welling. Auto-encoding variational bayes. In *Int. Conf. Learn. Repr.*, 2014.

A Kirillov, D Shlezinger, D Vetrov, C Rother, and B Savchynskyy. M-best-diverse labelings for submodular energies and beyond. In *Adv. Neural. Info. Proc. Sys.*, pages 613–621, 2015.

A Kirillov, A Shekhovtsov, C Rother, and B Savchynskyy. Joint m-best-diverse labelings as a parametric submodular minimization. In *Adv. Neural. Info. Proc. Sys.*, pages 334–342, 2016.

S Kohl, B Romera-Paredes, C Meyer, J De Fauw, J Ledsam, K Maier-Hein, S Eslami, D Rezende, and O Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Adv. Neural. Info. Proc. Sys.*, pages 6965–6975, 2018.

A Krizhevsky, I Sutskever, and G Hinton. Imagenet classification with deep convolutional neural networks. In *Adv. Neural. Info. Proc. Sys.*, pages 1097–1105, 2012.

L Ladicky, C Russell, P Kohli, and P Torr. Associative hierarchical crfs for object class image segmentation. In *Int. Conf. Comp. Vis.*, page 739, 2009.

B Lakshminarayanan, A Pritzel, and C Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Adv. Neural. Info. Proc. Sys.*, pages 6402–6413, 2017.

Y LeCun, L Bottou, G Orr, and K-R Muller. Efficient backprop. In *Neural Networks: Tricks of the Trade*, pages 9–48, 1998.

J Lee, E Kim, S Lee, J Lee, and S Yoon. FickleNet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *IEEE Comp. Vis. Pattern Recog.*, pages 5262–5271, 2019.

J Lee, E Kim, and S Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *IEEE Comp. Vis. Pattern Recog.*, pages 4071–4080, 2021.

V Lempitsky, A Vedaldi, and A Zisserman. Pylon model for semantic segmentation. In *Adv. Neural. Info. Proc. Sys.*, pages 1485–1493, 2011.

Y Li, Z Kuang, L Liu, Y Chen, and W Zhang. Pseudo-mask matters in weakly-supervised semantic segmentation. In *IEEE/CVF Int. Conf. Comp. Vis.*, pages 6964–6973, 2021.

Z Li, E Gavves, K van de Sande, C Snoek, and A Smeulders. Codemaps - segment, classify and search objects locally. In *Int. Conf. Comp. Vis.*, page 2136, 2013.

W Liu, D Anguelov, D Erhan, C Szegedy, S Reed, C Fu, and A Berg. Ssd: Single shot multibox detector. In *Euro. Conf. Comp. Vis.*, page 21, 2016.

J Long, E Shelhamer, and T Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Comp. Vis. Pattern Recog.*, pages 3431–3440, 2015.

W Lu, X Jia, W Xie, L Shen, Y Zhou, and J Duan. Geometry constrained weakly supervised object localization. In *Euro. Conf. Comp. Vis.*, pages 481–496, 2020.

S Mittal, M Tatarchenko, and T Brox. Semi-supervised semantic segmentation with high- and low-level consistency. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(4):1369–79, 2019.

M Mostajabi, P Yadollahpour, and G Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *IEEE Comp. Vis. Pattern Recog.*, pages 3376–3385, 2015.

Ö Çiçek, A Abdulkadir, S Lienkamp, T Brox, and O Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In *Med. Imag. Comput. Comp.-Assist. Interv.*, pages 424–432, 2016.

Y Ouali, C Hudelot, and M Tami. Semi-supervised semantic segmentation with cross-consistency training. In *IEEE Comp. Vis. Pattern Recog.*, page 12671, 2020.

J Pan, P Zhu, K Zhang, B Cao, Y Wang, D Zhang, J Han, and Q Hu. Learning self-supervised low-rank network for single-stage weakly and semi-supervised semantic segmentation. In *Int. Journal Com. Vis.*, pages 1181–1195, 2022.

G Papandreou, L Chen, K Murphy, and A Yuille. Learning to segment under various forms of weak supervision. In *Int. Conf. Comp. Vis.*, page 1742, 2015.

T Pohlen, A Hermans, M Mathias, and B Leibe. Full-resolution residual networks for semantic segmentation in street scenes. In *IEEE Comp. Vis. Pattern Recog.*, pages 4151–4160, 2017.

J Redmon and AFarhadi. YOLO9000: Better, faster, stronger. In *IEEE Comp. Vis. Pattern Recog.*, pages 6517–6525, 2017.

E Reinhard, M Ashikhmin, B Gooch, and P Shirley. Color transfer between images. *IEEE Comp. Graph. App.*, 21(5):34–41, 2001.

S Ren, K He, R Girshick, and J Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Neural Info. Proc. Sys.*, pages 91–99, 2015.

H Robbins and S Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, pages 400–407, 1951.

O Ronneberger, P Fischer, and T Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Med. Imag. Comput. Comp.-Assist. Interv.*, volume 9351, pages 234–241, 2015.

C Rupprecht, I Laina, R DiPietro, M Baust, F Tombari, N Navab, and G Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Int. Conf. Comp. Vis.*, pages 3611–3620, 2017.

K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *Int. Conf. Learn. Repr.*, 2015.

N Souly, C Spampinato, and M Shah. Semi supervised semantic segmentation using generative adversarial network. In *Int. Conf. Comp. Vis.*, pages 5688–5696, 2017.

R Vemulapalli, O Tuzel, M Liu, and R Chellapa. Gaussian conditional random field network for semantic segmentation. In *IEEE Comp. Vis. Pattern Recog.*, pages 3224–3233, 2016.

X Wang, S Han, Y Chen, D Gao, and N Vasconcelos. Volumetric attention for 3d medical image segmentation and detection. In *Med. Imag. Comput. Comp.-Assist. Interv.*, pages 175–184, 2019.

Y Wang, J Zhang, M Kan, S Shan, and X Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *IEEE Comp. Vis. Pattern Recog.*, pages 12275–12284, 2020.

Y Wei, H Xiao, H Shi, Z Jie, Feng, and T Huang. Revisiting Dilated Convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In *IEEE Comp. Vis. Pattern Recog.*, pages 7268–7277, 2018.

W. M. Wells, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz. Adaptive segmentation of MRI data. *IEEE Tr. Med. Imaging*, 15(4):429–443, 1996.

J Xu, A Schwing, and R Urtasun. Learning to segment under various forms of weak supervision. In *Comp. Vis. Pattern Recog.*, pages 3781–3790, 2015.

Y Xu, Z Zhou, X Li, N Zhang, M Zhang, and P Wei. FFU-Net: Feature fusion u-net for lesion segmentation of diabetic retinopathy. *BioMed Res. Int.*, 2021.

Z Yan, J Liang, W Pan, J Li, and C Zhang. Weakly-and semi-supervised object detection with expectation-maximization algorithm. In *arXiv*, 2017.

J Yao, S Fidler, and R Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *IEEE Comp. Vis. Pattern Recog.*, pages 702–709, 2012.

Q Yao and X Gong. Saliency guided self-attention network for weakly and semi-supervised semantic segmentation. *IEEE Access*, 8:14413–14423, 2020.

B Zhang, J Xiao, Y Wei, M Sun, and K Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. *Conf. Assoc. Advanc. Artif. Intell.*, 34(7):12765–12772, 2020a.

D Zhang, J Han, L Zhao, and T Zhao. From discriminant to complete: Reinforcement searching-agent learning for weakly supervised object detection. *IEEE Trans. Neur. Net. Learning Sys.*, 31(12):5549–5560, 2020b.

D Zhang, J Han, G Cheng, and M Yang. Weakly supervised object localization and detection: A survey. *IEEE Tran. Pattern Anal. Mach. Intell.*, 44(9):5866–5885, 2022a.

D Zhang, W Zeng, J Yao, and J Han. Weakly supervised object detection using proposal- and semantic-level relationships. *IEEE Tran. Pattern Anal. Mach. Intell.*, 44(6):3349–3363, 2022b.

S Zhang, Z Yu, L Liu, X Wang, A Zhou, and K Chen. Group r-cnn for weakly semi-supervised object detection with points. In *IEEE/CVF Comp. Vis. Pattern Recog.*, pages 9417–9426, 2022c.

T Zhang, G Lin, J Cai, T Shen, C Shen, and A Kot. Decoupled spatial neural attention for weakly supervised semantic segmentation. *IEEE Tran. Mult.*, 21(11):2930–2941, 2019.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Comp. Vis. Pattern Recog.*, pages 2881–2890, 2017.

Q Zhao, T Sheng, Y Wang, Z Tang, Y Chen, L Cai, and Haibin Ling. M2Det: A single-shot object detector based on multi-level feature pyramid network. In *Conf. Assoc. Advanc. Artif. Intell.*, pages 9259–9266, 2019.

B Zhou, A Khosla, A Lapedriza, A Oliva, and A Torralba. Learning deep features for discriminative localization. In *IEEE Comp. Vis. Pattern Recog.*, pages 2921–2929, 2016.

Y Zhou, X He, L Huang, L Liu, F Zhu, S Cui, and L Shao. Collaborative learning of semi-supervised segmentation and classification for medical images. In *IEEE Comp. Vis. Pattern Recog.*, pages 2079–2088, 2019.