This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Uncertainty and Energy based Loss Guided Semi-Supervised Semantic Segmentation

Rini Smita Thakur, Vinod K Kurmi
Indian Institute of Science Education and Research Bhopal, India
{rinithakur, vinodkk}@iiserb.ac.in

Abstract

Semi-supervised (SS) semantic segmentation exploits both labeled and unlabeled images to overcome tedious and costly pixel-level annotation problems. Pseudolabel supervision is one of the core approaches of training networks with both pseudo labels and ground-truth labels. This work uses aleatoric or data uncertainty and energy based modeling in intersection-union pseudo supervised network.The aleatoric uncertainty is modeling the inherent noise variations of the data in a network with two predictive branches. The per-pixel variance parameter obtained from the network gives a quantitative idea about the data uncertainty. Moreover, energy-based loss realizes the potential of generative modeling on the downstream SS segmentation task. The aleatoric and energy loss are applied in conjunction with pseudo-intersection labels, pseudo-union labels, and ground-truth on the respective network branch. The comparative analysis with state-of-the-art methods has shown improvement in performance metrics. The code is availaible at https://visdomlab.github.io/DUEB/.

1. Introduction

Semantic segmentation (SS) is an important branch of computer vision, which finds application in medical imaging, autonomous driving, and other intelligent industrial systems. There is significant performance improvement with the advent of supervised deep learning methodologies with extensive labeled data [30]. However, in many practical scenarios labeled data scarcity is the major problem. The constraint for supervised image segmentation is the requirement of enormous pixel-label annotated data, whose acquisition is tedious and time-consuming. It takes around 1.5 hours to annotate a single high-resolution urban street scene (Cityscapes dataset) [7], and sixty minutes for camouflaged object image [10] (COD10K dataset). So, there is a gradual shift from fully supervised approaches to weakly supervised and SS approaches. The weakly supervised methods leverage image-level labels with bounding boxes or scribbles, whereas SS methods leverage both labeled and unlabeled images. The SS semantic methods are categorized into adversarial methods, consistency regularization, contrastive learning, pseudo-labeling, and hybrid methods [33,44]. Pseudo-labeling is one of the most useful, simple and effective methods in SS segmentation. Pseudo-labeling implies the generation of pseudolabels from the supervised trained network and re-training with the new dataset composed of ground truth and pseudolabels.

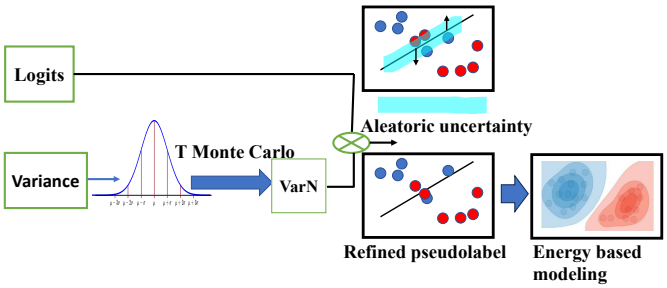


Figure 1. Refinement of pseudolabel with aleatoric uncertainty and energy based modeling

There is always a quest for the refinement of pseudo labels as the network is prone to confirmation bias. Uncertainty quantification is the preferred approach for refining pseudolabels, relying on two fundamental types of uncertainty: epistemic (model) uncertainty, which reflects the variability in model parameters, and aleatoric (data) uncertainty, which captures the inherent noise present in the input data. There are various test-time epistemic uncertainty (model) estimation techniques for pseudolabel refinement [3]. However, the inclusion of aleatoric (data) uncertainty in the Bayesian framework during training has not been explored in the SS segmentation setup. The non-inclusion of training time aleatoric uncertainty can be attributed to the fact that it requires significant architectural change for per-pixel variance calculation, which is not always feasible [27]. Secondly, appropriate change in the training process is required for variance optimization ac-

cording to the specific computer vision problem. On the other hand, epistemic uncertainty can be modeled at the inference time even in pre-trained models with a more generalized approach without cost modification. Therefore, there is a huge surge in epistemic methods such as thresholding, entropy minimization, ensembling and Monte-Carlo dropout with Bayesian approximation [34], but aleatoric inclusion is a neglected domain. So, we hypothesize to include the effect of object boundaries, inter and intra-class variation, and low-contrast regions with training-time aleatoric uncertainty quantification.

The other constraint with the SS segmentation methods is the discriminative and deterministic framework, which fails to capture the generative distribution. The consistency regularized pseudo-supervised methods [5, 11, 18, 46, 47] leverage supervised cross-entropy loss with ground-truths, unsupervised cross-entropy loss with the pseudolabels and consistency loss to capture perturbations within discriminative and deterministic framework. However, the discriminative networks can be reinterpreted as energy based model (EBM) to include the advantages of generative networks, i.e., implicit modeling of input data $p(x)$ to avoid overfitting with limited labeled training data, strong generalizability for out-of-detection sample, adversarial robustness, and improved uncertainty calibrations [24, 45].

In this study, we employ an EBM with a joint distribution of input images and labels of data to obtain generative functionality using a simple pseudo-labeled classifier. The discriminative classifier can be re-interpreted as the energy-based model for the joint distribution of input and ground truth [13, 22]. The EBMs parameterize any multi-dimensional input into the scalar value.

The joint energy based modeling and inclusion of certain uncertainty measures such as variance, entropy, dropout, etc., either at training or test time, can improve classification performance [2, 6, 19]. Therefore, it has motivated us to design variance-based loss for the segmentation network in order to capture inherent noise variations of the input. Furthermore, additional energy-based loss can help the model in learning the joint distribution of data $p(x, y)$ leading to a more enhanced refinement. It favours aleatoric uncertainty quantification by implicitly modeling $p(x)$.

The SS segmentation network provides per-pixel variance for heteroscedastic uncertainty calculation. A normal distribution is placed over the variance, and this variance is added to the model's prediction. The heteroscedastic aleatoric loss incorporates this variance parameter along with the energy-based loss to further refine the results. Figure 1 depicts the pseudolabel refinement with aleatoric uncertainty using both logits and variance.
The proposed framework, data uncertainty and energy loss (DUEB) is the pseudolabeling based SS segmentation network built upon the Conservative Progressive Collaborative Learning (CPCL) [11] framework. CPCL is a hybrid network that leverages consistency regularization and pseudo-labeling in a discriminative framework with two parallel branches. It generates union-intersection pseudo labels based on agreement and disagreement indicators on the outputs of two branches. These union and intersection help to generate better pseudo-labels. The main contributions of the DUEB are:

- We propose the aleatoric data uncertainty-based loss framework with DeepLabv3+ [4]. We incorporate the distorted logit loss based on data-dependent variance and optimize the difference between undistorted logit and distorted logit loss.
- Enhance the SS prediction by energy-based loss to incorporate generative modeling using the discriminative function.
- Usage of union-intersection pseudo-labels for data uncertainty and energy loss calculation in an unsupervised setting and usage of ground truth in a supervised setting.

2. Related Work

2.1. Semi-supervised segmentation

The combination of consistency regularized network along with pseudo-supervision is widely used for the SS segmentation [5, 28, 36, 46, 47]. The general methodology involves multiple identical networks or multiple networks with the same structure but different initialization with two or more predictive branches. Perturbations (input/feature/network) are applied, and corresponding pseudolabels are generated for all the branches. Cutout, CutMix, ClassMix and ComplexMix are commonly used image augmentation methods for input perturbations [23]. Fixmatch [36], Pseudoseg [47], and Unimatch [41] impose consistency on strongly and weakly augmented images. Fixmatch generates pseudolabels with thresholding, whereas PseudoSeg utilizes a calibrated fusion strategy. In PseudoSeg [47], the first network outputs decoder prediction and self-attention GradCam with weakly augmented image input. Thereafter, pseudolabels generated with calibrated fusion strategy supervise another network with strongly augmented image input. Unimatch [41] is built upon Fixmatch, and it uses additional feature perturbations and two strong views perturbations, unlike Fixmatch. Cross Pseudo Supervision network (CPS) [5] and S3MPL [46] network forces consistency between two differently initialized networks ($f_{\theta 1}$, $f_{\theta 2}$) with same augmented input. In CPS, the pseudo-hot label map of the first network $f_{\theta 1}$ is used to supervise another network $f_{\theta 2}$ and vice-versa. However, in S3MPL mixed pseudolabel

map based on higher prediction confidence is used to supervise both ($f_{\theta 1}$, $f_{\theta 2}$). The SS medical segmentation uses uncertainty aware pseudolabels in a consistency regularized student-teacher framework [28]. The noisy pseudolabels are refined with uncertainty estimation by Kullback–Leibler variance. U^2PL [40] network proposes a methodology to sort reliable and unreliable pseudolabels based on entropy.
Recently, SS segmentation has progressed towards transformer-based architectures (e.g. Semi-CVT [15], AllSpark [39]). The extension of CPS is done in transformer-based architectures using ViT backbone in [25, 26].

2.2. Uncertainty Estimation

Uncertainty quantification procedures are pivotal in uncertainty reduction during optimization and decision-making with two major approaches: Bayesian and Ensemble learning [1]. The uncertainties in computer vision are classified as aleatoric (data) uncertainty and epistemic (model). Aleatoric uncertainty captures the inherent noise distributions in the input image pixels, whereas epistemic uncertainty captures the network uncertainty. The epistemic uncertainty is incorporated in SS segmentation works, whereas irreducible aleatoric uncertainty is a neglected domain. It is captured by putting prior distribution over the weights followed by variational Bayesian approximation over the weights posterior distribution. The epistemic uncertainty is incorporated in the segmentation network with Monte-Carlo dropout. The Monte Carlo dropout involves multiple forward passes turning off certain nodes as per dropout probability. The 3D SS left atrium student-teacher framework network exploits epistemic uncertainty information with predictive entropy metric to evaluate uncertainty aware consistency loss [43]. The other medical segmentation network identifies certain and uncertain areas based on the conservative-radical module with multiple decoders [35]. Kullback–Leibler variance is used for uncertainty estimation to rectify the noisy pseudolabels for SS medical image segmentation [28]. There are several methods to evaluate epistemic uncertainty with a single pass to reduce the computational burden with Monte-Carlo passes [31].

Aleatoric uncertainty captures image noise, occlusions, blur, uncertain visual cues, etc. Heteroscedastic aleatoric uncertainty is modeled by Gaussian output models or Dirichlet models [29]. The Gaussian output network models the noise at the logits into the Gaussian distribution. The mean of Gaussian distribution is equal to the output of the corresponding class, whereas variance is learned from the input image. The associated likelihood is obtained from the expectation of this probability by applying Monte Carlo to these distributions, drawing a certain number of samples from the Gaussian modeling of the logits. The resulting likelihood is formulated into the cross entropy loss to embed the aletoric uncertainty component. Aleatoric uncertainty is also applied during test-time augmentation, where the distribution of predictions is calculated with Monte-Carlo simulations with the prior distribution of image transformations and noise [21, 38].

3. Method

In this section, we first define the problem statement, followed by a description of network architecture, data uncertainty and energy based loss in a union-intersection pseudolabel supervised framework. We have incorporated the variance layer in DeepLabv3+ [4] architecture and formulated aleatoric uncertainty and energy based loss in union-intersection pseudolabels supervised network.

3.1. Problem Statement

In SS semantic segmentation tasks, both labeled and unlabeled data are available, with the condition that the amount of unlabeled data is significantly greater than the labeled data. The labeled data with P images and corresponding ground-truth are given by $D_l = \{(X_l^1, G_l^1), ..., (X_l^P, G_l^P)\}$. The Q unlabeled images is given by $D_u = \{X_u^1, ..., X_u^Q\}$. A large amount of unlabeled data yields good segmentation results on test data. In this work, we follow similar architecture details presented in [11]. The main framework consists of two segmentation networks, $f(;\theta_c)$ and $f(;\theta_p)$ with the same structure but different initialization. It utilizes labeled data in a traditional supervised manner and unlabeled data with union-intersection pseudo labels. Both networks predict the pixel-wise semantics segmentation class labels and pixel-wise variance (uncertainty) values.

3.2. Network Architecture

We propose a novel SS segmentation framework that incorporates aleatoric or data uncertainty-based loss and energy-based loss as shown in Figure 2. The framework consists of two networks $f_{\theta_c}(;\theta_c)$ and $f_{\theta_p}(;\theta_p)$, termed as the conservative branch and progressive branch, respectively. θ_c and θ_p are the parameters of these networks respectively. Each network produces logits and variance at the output. The supervised learning is done with traditional categorical cross-entropy in which the same labeled images are given as the input in both networks. In an unsupervised setting, two unlabeled images X_1 and X_2 are given as input to both $f_{\theta_c}(;\theta_c)$ and $f_{\theta_p}(;\theta_p)$.The CutMix method is adopted for generating the strongly augmented image from individual images a and b to generate the output pixel based on the random mask m given in Eq. 1:

$$mix(a, b, m) = (1 - m) \odot a + m \odot b \tag{1}$$

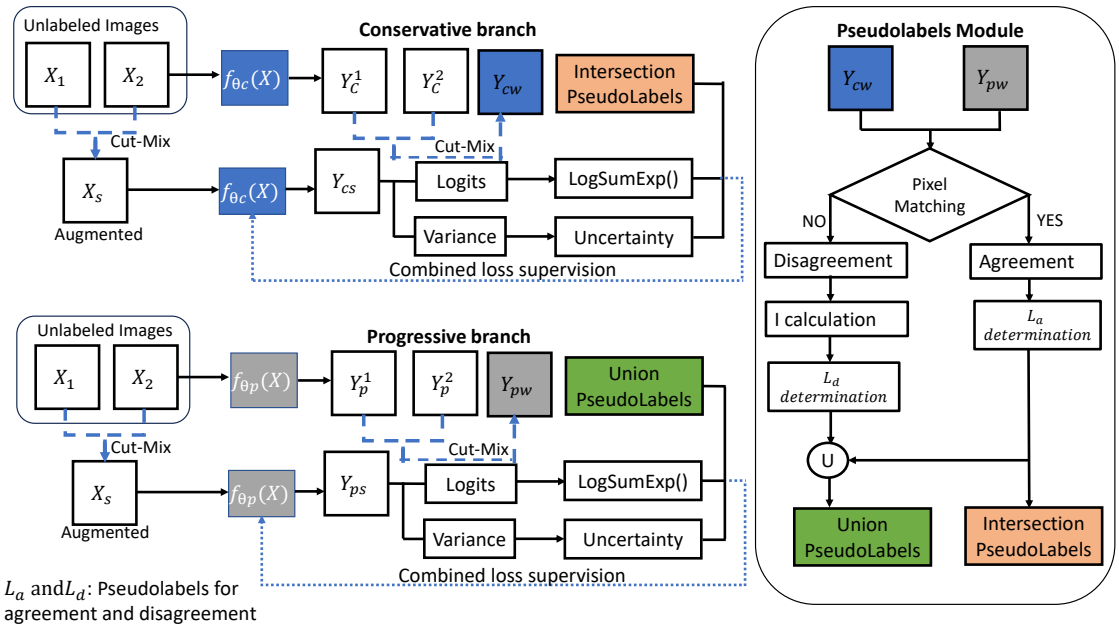


Figure 2. Block Diagram of proposed uncertainty and energy loss based framework with pseudolabels module for semi-supervised segmentation. Two branches: conservative and progressive are trained in parallel with intersection/union pseudolabels, data uncertainty loss and energy based loss.

The strongly augmented image X_s is obtained from Eq. 2.

$$X_s = mix(X_1, X_2, m) \tag{2}$$

The images X_1, X_2 and X_s are given as inputs in the conservative branch to generate Y_c^1, Y_c^2 and Y_{cs} given by Eq. 3, 4 and 5

$$Y_c^1 \leftarrow \arg\max_y f_{\theta_c}(y|X_1) \tag{3}$$

$$Y_c^2 \leftarrow \arg\max_y f_{\theta_c}(y|X_2) \tag{4}$$

$$Y_{cs} \leftarrow \arg\max_y f_{\theta_c}(y|X_s) \tag{5}$$

The output Y_c^1, Y_c^2 are combined at the output using the same CutMix procedure as that of input to give Y_{cw} using the same mask m, i.e. $mix(Y_c^1, Y_c^2, m)$. Similarly, we can get the output of the progressive branch as Y_{ps} and Y_{pw} $(mix(Y_p^1, Y_p^2, m))$ with the network $f(\theta_p)$.

The mixed output of both networks (Y_{cw}, Y_{pw}) obtained by mixing two individual images output is used for the generation of pseudo labels. These pseudo labels supervise the branches whose input is the Cut-Mix version of two images. The pseudo labels of the conservative branch are obtained by the intersection between individual pixels of Y_{cw} and Y_{pw}. Whereas the pseudo labels for the progressive branch are formed by the union operation of Y_{cw} and Y_{pw}. The union operation combines pixels using both agreement and disagreement indicators.

The intersection pseudo labels y^i_{inter}, are obtained by per-pixel comparison of Y_{cw} and Y_{pw}, as both outputs should be consistent with each other. The union pseudolabels y^i_{union}, consider both agreement and disagreement of pixels in Y_{cw} and Y_{pw}.

Usually, the disagreement of the pixels denotes that the pixel is uncertain as both networks give different predictions.There is always a trade-off between selecting high-quality pseudolabels and utilizing all the pseudolabels. The prediction-confidence based approach used in some of the recent works avoids performance degradation at the expense of wastage of unlabeled data. In [16], only $27-36\%$ pixels of the unlabeled Cityscapes dataset are used. Therefore, in the case of disagreement, we chose the class with a higher disagreement indicator as it is a more difficult class to predict. The class-wise disagreement indicator (I) given in CPCL [11] evaluates the pseudo labels of the disagreement regions. The final union pseudo labels are the union of the agreement and disagreement pixels between two predictions, thereby leveraging a high quantity of pseudolabels. The intersection pseudolabels supervise $f_{\theta_c}(;\theta_c)$, whereas union pseudolabels supervise $f_{\theta_p}(;\theta_p)$ with Cut-Mix strongly augmented image as the input X_s. A detailed explanation of class-wise disagreement indicator and dynamic confidence-based weight w_u^i is given in Supplementary. The loss function with intersection and union pseudolabels is given as:

$$\mathcal{L}_{int} = \frac{1}{N}\sum_{i}^{N} w_u^i \,\mathrm{CE}\left(f_{\theta_c}^i\left(X_s\right); y_{\text{inter}}^i\right) \tag{6}$$

$$\mathcal{L}_{uni} = \frac{1}{N}\sum_{i}^{N} w_u^i \,\mathrm{CE}\left(f_{\theta_p}^i\left(X_s\right); y_{\text{union}}^i\right) \tag{7}$$

$f_{\theta_c}^i(X_s)$ and $f_{\theta_p}^i(X_s)$ are the i^{th} pixel output of given input image X. For the supervised data:

$$\mathcal{L}_{sup} = \frac{1}{N}\sum_{i}^{N} \mathrm{CE}\left(f_{\theta}^i\left(X_l\right); G_l^i\right) \tag{8}$$

These losses can be combined as:

$$\mathcal{L}_{det} = \mathcal{L}_{sup} + \gamma_{int}\mathcal{L}_{int} + \gamma_{uni}\mathcal{L}_{uni} \tag{9}$$

γ_{int} and γ_{uni} are the hyperparameter for union and intersection supervision based loss function. These pseudo labels obtained using the union and interaction from two networks do not consider the uncertainty associated with the data. Thus we propose the following loss function to further enhance the pseudo-label supervision.

3.3. Data Uncertainty Estimation

Both the networks share the same segmentation structure as DeepLabv3+ [4] with different weights initialization. The network's last layer is modified to give a variance layer in the output of the network along with the predicted logits. The variance layer is the convolutional layer which is applied over the features of the last layer. Thus, the model outputs per-pixel variance along with the logits. The Bayesian aleatoric uncertainty loss function trains the variance layer's parameter. We follow similar to [17, 19] to obtain the variance and data uncertainty in the deep model. For any given input X, corresponding label Y and model $f(;\theta)$, the data uncertainty can be obtained from following:

$$\hat{Y}, \sigma_x = f(X;\theta) \tag{10}$$

$$diff = CE(\hat{Y}, Y) - CE\left(\hat{Y} + \epsilon_{\mathrm{t}}, Y\right); \epsilon_{\mathrm{t}} \sim \mathcal{N}(0, \sigma_x) \tag{11}$$

Here σ_x is the pixels-wise variance for input X and $\hat{Y}$ is the logit output for given input X and model$f(;\theta)$. $CE(,)$is the cross-entropy loss function. $\mathcal{N}(0, \sigma_x)$ is the normal distribution with mean 0 and σ_x is data dependent variance.
In the aleatoric loss, variance-based components are added to the original undistorted cross entropy loss l_u which is obtained from the model without variance layer. Apart from l_u, the other component reduces the difference between cross-entropy loss on the original logits and cross-entropy loss on the distorted logits. The distorted logits have Gaussian noise sampled according to the predicted variance. Gaussian noise is sampled from the normal distribution with mean zero and standard deviation equivalent to the square root of the predicted variance. Monte Carlo (MC) sampling generates multiple distorted versions of the logits by adding Gaussian noise. The difference tends to reduce the data uncertainty as it tends to optimize the variance with respect to undistorted logits. This difference is passed through the exponential linear unit for negative value handling. The expression e^{σ^2} exponentially scales the variance. The exponential function is chosen because it amplifies larger values of variance more significantly while maintaining smooth behaviour for smaller values. This means the model will penalize larger uncertainties more heavily, encouraging it to reduce uncertainty wherever possible. The overall data uncertainty loss l_{du} is given by Eq. 12:

$$\mathcal{L}_{\text{ale}} = \frac{1}{T}\sum_{T}\left[(-ELU * diff) * l_u + l_u + \left(e^{\sigma^2} - 1\right)\right] \tag{12}$$

where l_u is the undistorted loss, i.e., cross-entropy loss (CE) between ground-truth (Y) and $\hat{Y}$ i.e. $l_u = CE(\hat{Y}, Y)$, T are Monte-Carlo samples, ELU is an exponential linear unit. For the conservative and progressive branches, we can define the uncertainty loss as $\mathcal{L}_{ale}^c$ and $\mathcal{L}_{ale}^p$, respectively. Note that for calculating the $\mathcal{L}_{ale}^c$, we use y_{inter} as the pixel label. Similarly, y_{union} are used as the pixel label for obtaining $\mathcal{L}_{ale}^p$.

3.4. Energy Modeling

The objective of energy-based models is to incorporate generative modeling along with discriminative modeling. The energy-based model is also very well suited for uncertainty calibration for SS learning [45]. The generative model learns the joint distribution of the data and label $p(x, y)$, whereas the discriminative model learns conditional distribution $p(y|x)$. The log-likelihood distribution of joint distribution can be obtained from the discriminative function and energy of data distribution [13].

$$log\ p_\theta(x, y) = log\ p_\theta(x) + log\ p_\theta(y|x) \tag{13}$$

In the above equation, conditional class distribution $log\ p_\theta(y|x)$ is learned by cross-entropy loss, and $log\ p_\theta(x)$ can be learned from the energy-based models.

Energy-based models originate from the fact that any probability distribution $p_\theta(x)$ can be expressed in terms of energy function with the Boltzmann distribution given by Eq. 14:

$$p_\theta(x) = \frac{exp(-E_\theta(x))}{Z(\theta)} \tag{14}$$

where $E_{\theta(x)} : \mathbb{R}^D \to \mathbb{R}$ is the energy function which maps D dimension input into a scalar, and $Z_\theta = \int_x exp(-E_\theta(x))$ is the normalizing constant (partition function). The energy

Partition	Methods	mIoU	Animal	Vehicle	Indoor	Person	Background
1/2	Supervised	74.05	84.45	76.90	55.76	81.95	93.57
	CPCL	75.30	86.63	77.23	57.11	84.23	94.12
	DUEB	**75.94**	**87.08**	**78.25**	**57.46**	**85.61**	**94.17**
1/4	Supervised	71.66	78.98	74.31	55.91	82.05	93.23
	CPCL	74.58	85.39	76.62	56.69	83.65	93.70
	DUEB	**75.85**	**86.69**	**78.6**	**57.13**	**85.48**	**94.22**
1/8	Supervised	67.16	73.47	71.94	48.66	81.98	92.28
	CPCL	73.74	84.24	76.53	54.91	84.36	93.60
	DUEB	**74.89**	**85.47**	**78.75**	**54.91**	**85.2**	**94.14**
1/16	Supervised	62.00	67.65	67.88	41.56	79.99	91.63
	CPCL	71.66	82.73	76.24	**49.70**	83.57	93.02
	DUEB	**72.41**	**84.66**	**77.09**	49.60	**84.09**	**93.32**

Table 1. Quantitative Performance of PASCAL VOC 2012 in percent over the Supervised and CPCL baselines with ResNet-50 backbone.(ANIMAL: Bird, Cat, Cow, Dog, Horse, Sheep; VEHICLE: Aeroplane, Bicycle, Boat, Bus, Car, Motorbike and Train; INDOOR: Bottle, Chair, Dining Table, Potted Plant, Sofa, and monitor; PERSON:Person; BACKGROUND: Background.

Partition		1/4		1/8		1/16	
Loss		Datasets		Datasets		Datasets	
Uncertainty	Energy	PASCALVOC	Cityscapes	PASCALVOC	Cityscapes	PASCALVOC	Cityscapes
		74.58	76.98	73.74	74.60	71.66	69.92
√		75.23	77.28	74.10	75.98	71.90	72.07
	√	73.85	77.19	73.00	75.80	71.14	72.19
√	√	**75.85**	**77.85**	**74.89**	**76.16**	**72.41**	**72.38**

Table 2. mIoU of different loss components in addition to cross-entropy loss at different partitions.

function value is low for the samples drawn from the data distribution and high otherwise [42]. The energy function can be defined in terms of the LogSumExp(.) of the logits of a classifier given by Eq. 15:

$$E_\theta(x) = -LogSumExp_y(f_\theta(x)|y) \tag{15}$$

where $f_\theta(x)|y$ denotes the logits corresponding to y class-label.

The energy-based loss $\mathcal{L}_e$, which tends to maximize the logits over model distribution, is given by this equation:

$$\mathcal{L}_e = LogSumExp_y(f_\theta(x)|y) \tag{16}$$

From Eq. 16, we can obtain the energy-based loss function for the conservative and progressive branch as $\mathcal{L}_e^c$ and $\mathcal{L}_e^p$ respectively. In our network, the energy-based loss is applied over conservative and progressive branches using intersection and union pseudo labels. It is also applied as a part of supervised loss using the ground-truth label.

3.5. Total loss

The total loss is the combination of loss using the union and intersection label as pesdueo label defined in Eq.

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \gamma_{ale}(\mathcal{L}_{ale}^c + \mathcal{L}_{ale}^p) + \gamma_e(\mathcal{L}_e^c + \mathcal{L}_e^p) \tag{17}$$

γ_{ale} and γ_e are the hyperparameter for data uncertainty and energy based loss functions. In our case we use both the value as 1.

4. Experiments

We evaluate the performance of the proposed method on two standard benchmark datasets: Cityscapes [8] and PASCAL VOC 2012 [9]. The experiments are performed using the PyTorch deep learning framework on a server with an NVIDIA A100. The detailed algorithm and training details are presented in the Supplementary materials.

4.1. Dataset and Implementation Details

4.1.1 Datasets

The Cityscapes dataset is formed explicitly for the urban driving scene. It comprises 2975 training images and 500 validation images. The resolution of each image is 2048×1024, which is randomly cropped to 800×800, keeping the original resolution intact. There are 19 classes for pixel annotation.

PASCAL VOC 2012 dataset is the benchmark semantic segmentation dataset of the common objects. It has 1464 training images and 1449 validation images, with twenty foreground classes and one background class. We have created the augmented version of the dataset using the Segmentation Boundaries Dataset following the convention of the previous work [11]. Therefore, the entire training set now contains 10582 images with random cropping of 512×512, keeping the original resolution intact.

4.1.2 Implementation Details

We follow a similar experimental protocol defined in the CPCL model for a fair comparison. Both the con-

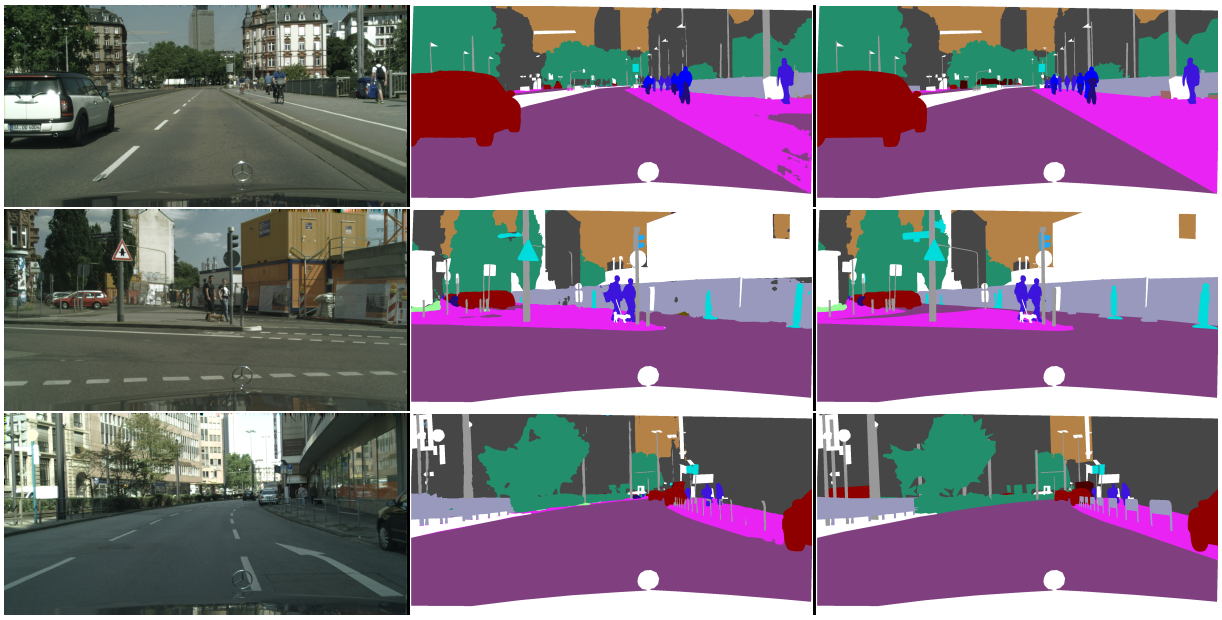


Figure 3. Segmentation Results of Cityscapes dataset (partition protocol:1/8) Left:Input Middle:Predictions of DUEB Right:Ground-truth.

Partition	Methods	mIoU	Flat	Human	Vehicle	Construction	Object	Nature	Sky
1/2	Supervised	75.36	89.87	71.71	74.21	67.35	69.91	78.52	94.52
	CPCL	**78.17**	91.44	**73.88**	**79.29**	69.15	**72.44**	**79.68**	**94.72**
	DUEB	77.58	**91.58**	71.97	79.07	**69.23**	70.55	79.24	94.70
1/4	Supervised	73.22	88.88	71.53	70.4	64.07	69.8	75.98	94.34
	CPCL	76.98	90.85	72.77	77.4	68.24	71.89	78.06	94.50
	DUEB	**77.85**	**91.60**	**73.45**	**78.63**	**70.16**	**71.99**	**78.12**	**94.72**
1/8	Supervised	68.63	87.89	69.81	60.02	60.68	67.85	74.43	94.01
	CPCL	74.6	89.48	72.88	72.89	65.05	70.49	77.13	94.4
	DUEB	**76.16**	**90.61**	**72.91**	**75.92**	**65.53**	**72.11**	**78.42**	**94.65**
1/16	Supervised	61.67	85.27	65.66	46.88	53.49	63.03	72.71	93.51
	CPCL	69.92	88.41	70.15	61.96	62.69	68.5	75.97	94.02
	DUEB	**72.38**	**89.76**	**71.67**	**66.65**	**64.38**	**70.34**	**77.10**	**94.24**

Table 3. Quantitative Performance of CityScapes dataset in percent over Supervised and CPCL baseline with ResNet-50 backbone (FLAT:Road and Sidewalk; HUMAN:Person and Rider; VEHICLE: Car, Truck, Bus, Train, MotorCycle, and Bicycle; CONSTRUCTION: Building Wall and Fence; OBJECT: Pole, Traffic Lights, and Traffic Sign, NATURE: Vegetation and Terrain; SKY: Sky)

servative and progressive branch share the same structure (Deeplabv3+) with randomly initialized segmentation heads. We use SGD with momentum with an initial learning rate of 10^{-4} and 5×10^{-3} with Cityscapes and PASCAL VOC, respectively. The learning rate is multiplied by $(1 - \frac{iter}{max_iter})^{0.9}$, momentum is set to 0.9, and the weight decay is 10^{-4}. The data partition protocol divides the data into labeled and unlabeled sets. It randomly extracts 1/2, 1/4, 1/8, and 1/16 of the whole data labeled part and considers the remaining data as unlabeled part. We follow the data partition protocol of previous works, [11, 18]. The Cut-Mix strategy is used for the generation of strongly augmented images using three rectangle regions of random ratio (in the range of [0.25, 0.5]). The Monte Carlo sample value T is set to 10 for both datasets. The implementation code will be made public.

4.2. Performance Analysis

We evaluate our framework DUEB on PASCAL VOC and Cityscapes dataset at all the partition protocols (ratio of labeled data). It has been observed that there is a significant improvement in mIoU at all partition protocols for the PASCALVOC dataset, as given in Table 1. The performance analysis is done with CPCL and supervised CPCL baseline [11]. It has been observed that there is a significant improvement by DUEB over CPCL supervised baseline as the ratio of labeled data decreases. The performance gains with respect to the supervised CPCL baseline is **1.89%**, **4.19%**, **7.73%**, and **10.41%** under partition protocol of 1/2, 1/4, 1/8 and 1/16, respectively on PASCAL VOC using ResNet-50 backbone. The best performance is over the animal class of

Methods	Network	1/16	1/8	1/4	1/2
Supervised	ResNet-50	62.00	67.16	71.66	74.05
MT [37]$_{NeurIPS'17}$	ResNet-50	66.77	70.78	73.22	75.41
	ResNet-101	70.59	73.20	76.62	77.61
GCT [18]$_{ECCV'20}$	ResNet-50	64.05	70.47	73.45	75.20
	ResNet-101	69.77	73.30	75.25	77.14
CCT [32]$_{CVPR'20}$	ResNet-50	65.22	70.87	73.43	74.75
	ResNet-101	67.94	73.00	76.17	77.56
Cut-Mix [12] $_{BMVC'20}$	ResNet-50	68.9	70.70	72.46	74.49
	ResNet-101	72.56	72.69	74.25	75.89
CPS [5]$_{CVPR'21}$	ResNet-50	68.21	73.20	74.24	75.91
CPCL [11]$_{TIP'23}$	ResNet-50	71.66	73.74	74.58	75.30
	ResNet-101	73.44	76.40	77.16	77.67
Unimatch [41]	ResNet-50	74.5	75.8	76.1	–
	ResNet-101	76.5	77.0	77.2	–
DUEB	ResNet-50	72.41	74.89	75.85	75.94
DUEB	ResNet-101	74.13	76.70	78.00	78.13

Table 4. Quantitative Performance (mIoU in percentage) of PASCAL-VOC dataset with other methods.

Methods	Network	1/16	1/8	1/4	1/2
Supervised	ResNet-50	61.67	68.63	73.22	75.36
MT [37]$_{NeurIPS'17}$	ResNet-50	66.14	72.03	74.47	77.43
	ResNet-101	68.08	73.71	76.53	78.59
GCT [18]$_{ECCV'20}$	ResNet-50	65.81	71.33	75.30	77.09
	ResNet-101	66.90	72.96	76.45	78.58
CCT [32] $_{CVPR'20}$	ResNet-50	66.35	72.46	75.68	76.78
	ResNet-101	69.64	74.48	76.35	78.29
CPS [5]$_{CVPR'21}$	ResNet-50	69.79	74.39	76.85	78.64
PGC [20]$_{WACV'23}$	ResNet-50	–	71.20	73.90	76.80
CPCL [11]$_{TIP'23}$	ResNet-50	69.92	74.60	76.98	78.17
Unimatch [41]	ResNet-50	75.0	76.8	77.5	78.6
	ResNet-101	76.6	77.9	79.2	79.5
DUEB	ResNet-50	72.38	76.16	77.85	77.58
DUEB	ResNet-101	74.16	77.36	78.80	79.47

Table 5. Quantitative Performance (mIoU in percentage) of CityScapes dataset with other methods.

PASCAL VOC, as there is an increment of about 17 % over the supervised baseline and 2% over the CPCL network with ResNet-50 backbone. However, the performance is stagnant when it comes to background class as higher mIoU is already achieved by supervised baseline.

The CityScapes dataset also has significant mIoU improvement over partition protocols 1/4, 1/8 and 1/16, as given in Table 3. However, at 1/2 data partition, it outperformed the supervised CPCL baseline but showed a reduction slight reduction with respect to CPCL with ResNet-50. In Cityscapes, too, there is an improvement in performance as the ratio of labeled images decreases. The best performance is achieved by the vehicle group, which is 19.77% over the supervised CPCL baseline and 4.69% over CPCL. However, the performance is stagnant when it comes to sky class as higher mIoU is already achieved by supervised baseline. The visualization of the segmentation map is provided in Figure 3. It has been observed that DUEB with ResNet-101 backbone outperforms the state-of-the-art methods for all partition protocols as given in Table 1 and Table 3. More details are available in the Supplementary material.

4.3. Ablation Study

We analyze the network to evaluate mIoU on separate loss components. In the entire network, standard cross entropy is loss with ground truth for the supervised part and with pseudo-intersection-union labels for the unsupervised part. Data uncertainty loss and energy-based loss have been added in addition to cross-entropy loss. From Table 2 it has been observed that data uncertainty loss is more robust than the energy-based loss when used individually. However, a combination of both uncertainty-based loss and energy-based loss performs better in terms of mIoU than the individual components. The reason is energy-based modeling helps to quantify the aleatoric uncertainty because it is robust to noisy data and aims to minimize energy for plausible data points.

4.4. Comparative Analysis

Table 4 and Table 5 gives a comparative analysis of DUEB on PASCAL-VOC and CityScapes with state-of-the art methods: GCT [18], CCT [32], Cut-Mix Seg, [12], MT [37], CPS [5], PGC [20], CPCL [11], Unimatch [41]. All compared methods are based on DeepLabv3+ with ResNet-50 or ResNet-101 [14] backbone. We report the results of some previous methods in Table 4 and Table 5 from [5] and [11]. It has outperformed some previous methods on both ResNet-50/101 backbone in terms of mIoU. The performance improvement is more evident when a few number of labeled images are used.

5. Conclusion

We propose data uncertainty and energy-based modeling for SS semantic segmentation. The per-pixel variance parameter, along with the energy, have resulted in performance improvement. Moreover, the usage of union-intersection pseudo labels for unsupervised parts renders usage of both large quantity and high quality. The proposed methodology is generic, i.e., it can be applied to existing SS semantic segmentation networks. The modification in the architecture of CNN to incorporate the output variance layer can yield improved results. However, there is further scope for improvement in the areas of pseudo-label generation, uncertainty loss incorporation in transformers, and model coupling.

Acknowledgments Rini acknowledges the financial support provided by the Department of Science and Technology, Government of India, through the WISE Post-Doctoral Fellowship Program (Reference No. DST/WISE-PDF/ET-33/2023), which facilitated the successful completion of this work.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. 3

[2] Ismail Alarab and Simant Prakoonwit. Uncertainty estimation based adversarial attack in multi-class classification. *Multimedia Tools and Applications*, 82(1):1519–1536, 2023. 2

[3] Cheng Chen, Quande Liu, Yueming Jin, Qi Dou, and Pheng-Ann Heng. Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 225–235, Cham, 2021. Springer International Publishing. 1

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2, 3, 5

[5] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. 2, 8

[6] Marc Combalia, Ferran Hueto, Susana Puig, Josep Malvehy, and Veronica Vilaplana. Uncertainty estimation in deep neural networks for dermoscopic image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 744–745, 2020. 2

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 1

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6

[9] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 6

[10] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2774–2784, 2020. 1

[11] Siqi Fan, Fenghua Zhu, Zunlei Feng, Yisheng Lv, Mingli Song, and Fei-Yue Wang. Conservative-progressive collaborative learning for semi-supervised semantic segmentation. *IEEE Transactions on Image Processing*, 2023. 2, 3, 4, 6, 7, 8

[12] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *BMVC*, 2020. 8

[13] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *ICLR*, 2019. 2, 5

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 8

[15] Huimin Huang, Shiao Xie, Lanfen Lin, Ruofeng Tong, Yen-Wei Chen, Yuexiang Li, Hong Wang, Yawen Huang, and Yefeng Zheng. Semicvt: Semi-supervised convolutional vision transformer for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11340–11349, 2023. 3

[16] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018. 4

[17] Indu Joshi, Ayush Utkarsh, Riya Kothari, Vinod K Kurmi, Antitza Dantcheva, Sumantra Dutta Roy, and Prem Kumar Kalra. Data uncertainty guided noise-aware preprocessing of fingerprints. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 5

[18] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 429–445. Springer, 2020. 2, 7, 8

[19] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 2, 5

[20] Heejo Kong, Gun-Hee Lee, Suneung Kim, and Seong-Whan Lee. Pruning-guided curriculum learning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5914–5923, 2023. 8

[21] Shanu Kumar, Vinod Kumar Kurmi, Praphul Singh, and Vinay P Namboodiri. Mitigating uncertainty of classifier for unsupervised domain adaptation. *arXiv preprint arXiv:2107.00727*, 2021. 3

[22] Vinod K. Kurmi, Venkatesh K. Subramanian, and Vinay P. Namboodiri. Domain impression: A source data free domain adaptation method. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 615–625, January 2021. 2

[23] Dominik Lewy and Jacek Mańdziuk. An overview of mixing augmentation methods and augmentation strategies. *Artificial Intelligence Review*, 56(3):2111–2169, 2023. 2

[24] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative

models: Semi-supervised learning and strong out-of-domain generalization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8296–8307, 2021. 2

[25] Peixia Li, Pulak Purkait, Thalaiyasingam Ajanthan, Majid Abdolshah, Ravi Garg, Hisham Husain, Chenchen Xu, Stephen Gould, Wanli Ouyang, and Anton Van Den Hengel. Semi-supervised semantic segmentation under label noise via diverse learning groups. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1229–1238, 2023. 3

[26] Yijiang Li, Xinjiang Wang, Lihe Yang, Litong Feng, Wayne Zhang, and Ying Gao. Diverse cotraining makes strong semi-supervised segmentor. *arXiv preprint arXiv:2308.09281*, 2023. 3

[27] Antonio Loquercio, Mattia Segu, and Davide Scaramuzza. A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters*, 5(2):3153–3160, 2020. 1

[28] Liyun Lu, Mengxiao Yin, Liyao Fu, and Feng Yang. Uncertainty-aware pseudo-label and consistency for semi-supervised medical image segmentation. *Biomedical Signal Processing and Control*, 79:104203, 2023. 2, 3

[29] José Mena, Oriol Pujol, and Jordi Vitria. A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective. *ACM Computing Surveys (CSUR)*, 54(9):1–35, 2021. 3

[30] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022. 1

[31] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. *arXiv preprint arXiv:2102.11582*, 2021. 3

[32] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12671–12681, 2020. 8

[33] Adrian Peláez-Vegas, Pablo Mesejo, and Julián Luengo. A survey on semi-supervised semantic segmentation. *arxiv*, 2023. 1

[34] Yinghuan Shi, Jian Zhang, Tong Ling, Jiwen Lu, Yefeng Zheng, Qian Yu, Lei Qi, and Yang Gao. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(3):608–620, 2022. 2

[35] Yinghuan Shi, Jian Zhang, Tong Ling, Jiwen Lu, Yefeng Zheng, Qian Yu, Lei Qi, and Yang Gao. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(3):608–620, 2022. 3

[36] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 2

[37] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 8

[38] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019. 3

[39] Haonan Wang, Qixiang Zhang, Yi Li, and Xiaomeng Li. Allspark: Reborn labeled features from unlabeled in transformer for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3627–3636, 2024. 3

[40] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022. 3

[41] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7236–7246, 2023. 2, 8

[42] Xiulong Yang, Qing Su, and Shihao Ji. Towards bridging the performance gaps of joint energy-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15732–15741, 2023. 6

[43] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 605–613. Springer, 2019. 3

[44] Man Zhang, Yong Zhou, Jiaqi Zhao, Yiyun Man, Bing Liu, and Rui Yao. A survey of semi-and weakly supervised semantic segmentation of images. *Artificial Intelligence Review*, 53:4259–4288, 2020. 1

[45] Stephen Zhao, Jorn-Henrik Jacobsen, and Will Grathwohl. Joint energy-based models for semi-supervised classification. In *ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning*, volume 1, 2020. 2, 5

[46] Yan Zhou, Ruyi Jiao, Dongli Wang, and Haibin Zhou. S³mpl:semi-supervised semantic segmentation with mixed pseudo label. In *2022 37th Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pages 1158–1163, 2022. 2

[47] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg:

Designing pseudo labels for semantic segmentation. In *International Conference on Learning Representations*, 2021. 2