Uncertainty Estimation via Response Scaling for Pseudo-Mask Noise Mitigation in Weakly-Supervised Semantic Segmentation

Yi Li[1,2], **Yiqun Duan**[3], **Zhanghui Kuang**[2], **Yimin Chen**[2], **Wayne Zhang**[2], **Xiaomeng Li**[1,4,*]

[1] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology
[2] SenseTime Research
[3] Department of Computer Science, University of Technology Sydney
[4] The Hong Kong University of Science and Technology Shenzhen Research Institute
ylini@connect.ust.hk, eexmli@ust.hk

Abstract

Weakly-Supervised Semantic Segmentation (WSSS) segments objects without a heavy burden of dense annotation. While as a price, generated pseudo-masks exist obvious noisy pixels, which result in sub-optimal segmentation models trained over these pseudo-masks. But rare studies notice or work on this problem, even these noisy pixels are inevitable after their improvements on pseudo-mask. So we try to improve WSSS in the aspect of noise mitigation. And we observe that many noisy pixels are of high confidence, especially when the response range is too wide or narrow, presenting an uncertain status. Thus, in this paper, we simulate noisy variations of response by scaling the prediction map multiple times for uncertainty estimation. The uncertainty is then used to weight the segmentation loss to mitigate noisy supervision signals. We call this method URN, abbreviated from Uncertainty estimation via Response scaling for Noise mitigation. Experiments validate the benefits of URN, and our method achieves state-of-the-art results at 71.2% and 41.5% on PASCAL VOC 2012 and MS COCO 2014 respectively, without extra models like saliency detection. Code is available at https://github.com/XMed-Lab/URN.

Introduction

Semantic segmentation is the fundamental task in computer vision. One of the main challenges is the prohibitive cost of obtaining dense pixel-level annotations to supervise model training. To reduce the annotation cost, researchers replace the dense pixel-level annotations by weak annotations, such as scribble (Lin et al. 2016; Tang et al. 2018), bounding boxes (Xu, Schwing, and Urtasun 2015; Dai, He, and Sun 2015; Khoreva et al. 2017), points (Bearman et al. 2016) and image-level labels (Kolesnikov and Lampert 2016; Pathak, Krahenbuhl, and Darrell 2015; Pinheiro and Collobert 2015; Ahn and Kwak 2018). In this paper, our goal is to develop a novel method for Weakly-Supervised Semantic Segmentation (WSSS) based on image-level class labels only.

Most existing methods (Wang et al. 2020; Zhang et al. 2020; Li et al. 2021) in WSSS follow the common pipeline that generates pseudo-masks firstly and then trains the segmentation model with the pseudo-masks in a fully-supervised manner. These methods improve the quality of pseudo-mask by solving the under activation problem of Class Activation Map (Zhou et al. 2016) as these papers (Kolesnikov and Lampert 2016; Jiang et al. 2019; Chang et al. 2020), also distinguishing foreground from background with AffinityNet (Ahn and Kwak 2018) or saliency (Yao and Gong 2020; Fan et al. 2020). But the noisy pixels are inevitable after applying these methods. As far as we know, only PMM (Li et al. 2021) tries to relieve the influence of noise. While it does not work on refined pseudo-mask and COCO dataset, without location of noisy pixels. So in this paper, we mitigate the noise in a more effective way.

We observe that the noisy pixels are mainly divided into missing positives and false positives. To be specific, if the response range (scope of activated area) is narrow, the missing positives take the lead, otherwise false positives become the majority as shown in Fig.2. Because of the variation of the response scale, there are many false pixels (noise) with high confidence. And this phenomenon is exactly belongs to uncertainty. Thus, in this paper, we locate the noise via uncertainty estimation.

Uncertainty estimation requires multiple predictions for each image, so how to generate these predictions and reflect the noise is important in the setting of WSSS. By observation of false pixels whose confidences are high in Figure 2, the uncertain pixels are highly related to the response scale. So we simulate response maps in different activation scales by probability scaling. Besides, we add dense-CRF (Zheng et al. 2015) to response maps after scaling scheme as the formation of pseudo-mask. After above operations, we estimate the uncertainty with these scaled responses, and the affinity is built between uncertainty and noise.

After uncertainty estimation via response scaling, we normalize the uncertainty and transform it to loss weight in optimization to mitigate the damage of noise. We observe that the pixel whose uncertainty is higher is more likely to be noise as shown in Figure 1. Thus, a lower loss weight is applied to relieve the adverse effect of it. We multiply the weight map and loss map with a mean operation to get the final loss in optimization.

The overall process is call URN abbreviated from Uncertainty estimation via Response scaling for Noise mitigation.

Figure 1: Uncertainty estimated by response scaling is positively correlated with false pixels (noise) of pseudo-mask. Left: ground truth. Middle: noisy pseudo-mask. Right: uncertain map after normalization.
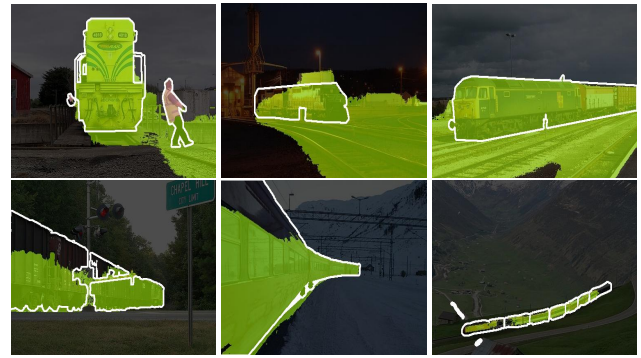

Figure 2: False pixels (noise) in pseudo-masks of the state-of-the-art WSSS method (Li et al. 2021). Top: wide response scales cause false positives. Bottom: narrow response scales make missing positives. The white line is boundary of ground-truth.

URN is practical and effective for the segmentation with noisy pseudo-mask from weakly supervision. We validate the proposed URN with baseline method PMM (Li et al. 2021) on a prevailing dataset PASCAL VOC 2012 (Everingham et al. 2010) and a challenging dataset MS COCO 2014 with image-level supervision. And we get new state-of-the-art mIoU at 71.2% on VOC and 41.5% on COCO.

We summarize the contributions as follows :

- We observe that false pixels (noise) with high confidences are related to the scale of response range, and we locate these uncertain pixels via response scaling.
- We originally mitigate the damage of noise in optimization of Weakly-Supervised Semantic Segmentation via loss weight from uncertainty.
- We achieve new state-of-the-art results on main datasets in Weakly-Supervised Semantic Segmentation.

Related Work

Weakly-Supervised Semantic Segmentation. In recent years, considerable efforts have been devoted to developing label-efficient semantic segmentation methods (Lin et al. 2016; Tang et al. 2018; Xu, Schwing, and Urtasun 2015; Dai, He, and Sun 2015; Khoreva et al. 2017; Kolesnikov and Lampert 2016; Pathak, Krahenbuhl, and Darrell 2015; Pinheiro and Collobert 2015; Ahn and Kwak 2018; Wang et al. 2020; Yu et al. 2019a; Li et al. 2020a, 2018). Among these methods, image-level class labels require the least annotation cost. Existing WSSS methods mainly adopt two-stage procedures. The first stage optimizes class activation maps (CAMs) (Zhou et al. 2016) produced by a multi-label classification network to generate pseudo ground-truth. The second stage is to train a fully supervised semantic segmentation network via pseudo ground-truth. To improve WSSS, most algorithms focus on improving the quality of CAMs

through expanding partial response area to whole foreground region (Kolesnikov and Lampert 2016; Jiang et al. 2019; Chang et al. 2020). Some methods distinguish foreground from background to improve the quality of CAMs by self-supervised method (Ahn and Kwak 2018; Ahn, Cho, and Kwak 2019) or introducing extra information liking saliency detection model (Lee et al. 2019; Yao and Gong 2020; Jiang et al. 2019; Yao and Gong 2020; Li et al. 2020b; Fan et al. 2020) and object proposals (Liu et al. 2020).

In WSSS, researchers mainly focus on generating high-quality pseudo labels to improve the performance, less attention has been paid on *how to improve the segmentation model with these imperfect labels*. In this paper, We realize that since noisy pixels are inevitable, the appropriate way is to mitigate the damage of uncertain pixels.

Uncertainty in Segmentation. Deep neural networks provide a probability with each prediction, occasionally, it is false even the probability is high. In paper (Kendall and Gal 2017), this phenomenon is called epistemic uncertainty resulting from the model itself. Early works use Monte-Carlo Dropout (Kendall, Badrinarayanan, and Cipolla 2015; Gal and Ghahramani 2016a; Blundell et al. 2015; Gal and Ghahramani 2016b) to approximate the posterior distribution for uncertainty estimation, also, Ensemble (Lakshminarayanan, Pritzel, and Blundell 2017) and multi-head (Lee et al. 2015, 2016; Rupprecht et al. 2017) are also applied. Follow-up researchers majorly concentrate on proposing improved approaches to acquire better output uncertainty maps. Here, inference-based (Kendall and Gal 2017; Tanno et al. 2017; Jungo, Balsiger, and Reyes 2020) mostly utilize Bayesian Inference or Stochastic Inference to improve the quality of uncertainty map, meanwhile auto-encoder-based (Kingma and Welling 2014; Sohn, Lee, and Yan 2015) methods introduce reconstruction to acquire better spatial precision of the uncertainty map. These papers mostly predict uncertainty masks to facilitate human's judgement, such as it in medical image segmentation scenarios. However, previous methods mostly emphasize on improving the qual-

ity of the uncertainty map but rarely explore how to improve the segmentation precision through uncertainty. Instead, they assemble the predictions for better results.

Different from previous works that estimate uncertainty for visualization or annotation, we utilize uncertainty to mitigate noise in optimization of segmentation for better performance. And the uncertainty is estimated based on the observation of noise's characteristics in Weakly-supervised Semantic Segmentation.

Learning with Noisy Labels. Our focus in this paper is to mitigate noisy labels in WSSS. Here, we discuss some related work in learning with noisy labels. Existing methods of learning with noisy labels could be classified into four categories. 1) Label correction methods propose to improve the quality of the labels by applying label corrections via using complex noise prediction models such as graphical models (Xiao et al. 2015), conditional random fields (Vahdat 2017), and neural networks (Lee et al. 2017; Veit et al. 2017). 2) Label-weighted loss methods modify loss functions during training, based on label-dependent weights (Natarajan et al. 2013; Goldberger and Ben-Reuven 2017; Sukhbaatar et al. 2014) or estimated noise transition matrix that defines the probability of mislabeling one class with another (Han et al. 2018a; Patrini et al. 2017; Reed et al. 2014; Pereyra et al. 2017). For example, Xu et al. (2019) introduces a Mutual Information (MI) loss for robust fine-tuning of a CE pre-trained model. 3) Fine-grained label correction in training designs adaptive training strategies that are more robust to noisy labels. These methods mostly correct labels by teacher-student distillation (Jiang et al. 2018; Yu et al. 2019b; Kumar and Ithapu 2019) and apply jointly training or learning-mutually mechanism between parallel models (Han et al. 2018b; Tanaka et al. 2018; Kim et al. 2019). 4) Robust loss functions also are proposed as more generic solution for robust learning by proposing and proving new loss function, such as Mean Absolute Error (MAE) losses (Ghosh, Kumar, and Sastry 2017), gradient saturation MAE (Zhang and Sabuncu 2018), Generalized Cross-Entropy (GCE) loss (Zhang and Sabuncu 2018), and Symmetric Cross-Entropy (SCE) Wang et al. (2019c). Empirically justified approaches that directly modify the magnitude of the loss gradients are also an active line of research (Wang et al. 2019a,b).

Unlike the previous reweight methods based on probability, there are many pixels with high confidences of pseudo-masks at uncertain status. Thus, for the first time in WSSS, we propose a new method to estimate the uncertainty to locate the possible noise for mitigation. Besides, we present a distillation scheme based on pseudo-masks in WSSS.

Methodology

In this section, we firstly introduce the pipeline of Weakly-supervised Semantic Segmentation and where our method is applied. Then we elaboration the proposed Uncertainty Estimation via Response Scaling. In the last, we describe the transformation from uncertainty to loss weight. We depict the whole process as Fig.3 for better understanding.

Let's define the segmentation function with image $I \in \mathbb{R}^{H \times W}$ as $\mathcal{F}_S(I)$ which trained from pseudo-mask

$M \in \mathbb{R}^{H \times W}$, and its prediction $X \in \mathbb{R}^{C \times H \times W}$ is the input of the proposed method Uncertainty estimation via Response scaling for Noise mitigation (URN) $Urn(X)$ whose output is loss weight $Y \in \mathbb{R}^{C \times H \times W}$ for noise mitigation in segmentation optimization.

URN in Weakly-supervised Semantic Segmentation

Latest Weakly-supervised Semantic Segmentation has three main components, namely classification, pseudo-mask refinement and segmentation. We name them as $\mathcal{F}_C(I, L)$, $\mathcal{F}_R(M, I)$, $\mathcal{F}_S(I, M)$ respectively. Note that the first variable is input of inference and the second variable is ground-truth in training. $\mathcal{F}_C$ is trained with image-level annotation A and converts image I to Class Activation Map $CAM \in \mathbb{R}^{C \times H \times W}$ via operation Cam as Eq.1

$$CAM = Cam(\mathcal{F}_C(I)) \qquad (1)$$

Then CAM is refined by $\mathcal{F}_R$ to obtain pseudo mask M. In most methods $\mathcal{F}_R$ is dense-CRF and in our baseline it is PPMG (Li et al. 2021). In paper (Ahn and Kwak 2018) it is AffinityNet with Random Walk to learn the foreground and background. We collectively referred them to $\mathcal{F}_R$ as:

$$M = \mathcal{F}_R(CAM) \qquad (2)$$

In this paper, we focus on the last stage $\mathcal{F}_S$ which is trained from M and inference with image. Our improvement is an additive loss weight Y from Uncertainty Estimation via Response Scaling with a transformation operation. We conclude these operations as Urn:

$$Y = Urn(\mathcal{F}_S(I)) \qquad (3)$$

After Urn calculating the weight mask, $\bar{\mathcal{F}}_S$ is trained with Cyclic Pseudo-mask M_c in PMM (Li et al. 2021), image I and weight mask Y for the final deployment.

Uncertainty Estimation via Response Scaling

Bayesian neural networks (Denker and LeCun 1990; MacKay 1992) set deep model into a probabilistic model by learning the distribution of prediction. While Bayesian neural network's prediction is hard to obtain. Thus, they apply varied inference to approximate the posterior of the model via dropout, such as Monte Carlo Dropout (Kendall and Gal 2017) or ensemble.

In WSSS, our goal is to measure the variance of pseudo-masks due to the inevitable noise on pseudo-masks. But predictions from dropout or different models cannot directly obtain pseudo-masks in the behavior of noise. Activation ranges are likely too narrow or too wide simultaneously, because the operation target is the network weight instead of the response itself. We observe that the noise in pseudo-mask of segmentation is always related to the response range of foregrounds. Motivated by this observation, we propose the Uncertainty Estimation via Response Scaling to mitigate the noise. Scaling operation with CRF could simulate the true behavior of noisy pseudo-masks. There, the variance of this noisy pseudo-masks in real distribution is the uncertainty that reflect the variance of real noise in pseudo-masks.

Figure 3: Process of URN. Outputs are weight mask Y and cyclic mask M_c with purple borders. Blue arrows indicate the training inputs of CNNs, and orange arrows belong to the operations of URN. $\mathcal{F}_R$, CRF and PPMG are optional post-processes. The weight mask looks like the intermediate between pseudo-mask and cyclic mask in this image.

We start the estimation from segmentation model $\mathcal{F}_S$ which is trained from pseudo-masks, we have its prediction feature map as:

$$X = \mathcal{F}_S(I) \qquad (4)$$

We firstly record the existed labels in pseudo-mask M, and these labels are corresponded to the indexes $\mathbb{R}^{\bar{C}}$ of X in channel dimension. We only keep the positive channels for next processing. we have the selected feature map $\bar{X}$ as:

$$\bar{X} = X_{(\mathbb{R}^{\bar{C}},:,:)}, \text{s.t.}, \mathbb{R}^{\bar{C}} \in M \qquad (5)$$

Then we apply the scaling scheme to $\bar{X}$ via exponential function with varied power factors $S \in \mathbb{R}^N$ on the target channel to generate varied response maps, and in this process, other channels keep the same. When the scale factor is upper than 1, the response scale of target class declines and the number of false positive reduces too. If it's lower than 1 the scale expands and there are less missing positive pixels. For each label in $\mathbb{R}^{\bar{C}}$, we adjust the scale of one channel and keep others the same for category specifically variation. We have scaled predictions $P \in \mathbb{R}^{\bar{C} \times N \times \bar{C} \times H \times W}$ for each class and scale as:

$$P_{(\bar{c},n,\bar{c},:,:)} = \bar{X}^{S_n}_{(\bar{c},:,:)}, \forall n, \bar{c} \qquad (6)$$

As we want to estimate the uncertainty of pseudo-mask predicted from segmentation, we introduce dense-CRF function Crf and argmax Arg to generate pseudo-masks in varied scales as follows:

$$\bar{P}_{(\bar{c},n,:,:,:)} = Crf(\bar{P}_{(\bar{c},n,:,:,:)}, I), \forall n, \bar{c} \qquad (7)$$

$\bar{P}$ indicates the scaled predictions after dense-CRF, then we get the pseudo-masks in varied scales and classes $\bar{M} \in \mathbb{R}^{\bar{C} \times N \times H \times W}$ by reducing the third category channel (ranking from 0) via argmax:

$$\bar{M} = Arg(\bar{P}, dim = 2) \qquad (8)$$

After get scales pseudo-masks we start to estimate the uncertainty, we select variance $\sigma^2 \in \mathbb{R}^{\bar{C} \times H \times W}$ as the metric to approximate the uncertainty for its simplicity and versatility. We calculate Var in second dimension (scaling factors) N for each foreground categories:

$$Var(\bar{M}) = \frac{1}{N-1} \sum_{n=0}^{N} (\bar{M}_{(\bar{c},n,:,:)} - \mathbb{E}(\bar{M}_{\bar{c},:,:,:}))^2, \forall \bar{c} \quad (9)$$

Finally, we apply Max operation in category dimension as Eq.10 to get class agnostic uncertainty map $U \in \mathbb{R}^{H \times W}$ for the later loss reweight with min-max normalization $Norm$.

$$U = Norm(Max(Var(\bar{M}), dim = 0)) \qquad (10)$$

Noise Mitigation in Optimization

Reweight scheme is widely usage in computer vision tasks like classification and detection to balance category or adjusting the importance of certain kind of samples. Following the mechanism, we deploy a lower loss weight for the possible noise pixels based on the uncertainty estimated by response scaling to mitigate the damage of noise in training.

Since pixels with high uncertainty deserve low loss weight to limit its influence. We set initial weight mask $W \in \mathbb{R}^{H \times W}$ as:

$$W = 1 - U \qquad (11)$$

To make the weight mask more controllable, we assign certain pixels and uncertain pixels via threshold t guided by W. And we mitigate the impact of possible noise by a lower loss weight. We define the final weight mask Y as:

$$Y_{(h,w)} = \begin{cases} 1 & W_{(h,w)} >= t \\ t & W_{(h,w)} < t \end{cases} \forall h, w \qquad (12)$$

In the end, we multiply the weight mask Y to the loss mask $L \in \mathbb{R}^{H \times W}$ from cross entropy loss with target cyclic pseudo-mask M_c and prediction map X. We have the segmentation loss in as Eq.(13):

$$L(X, M_c)_{(h,w)} = -Y_{(h,w)} \cdot log \frac{exp(X_{(M_{c(h,w)},h,w)})}{\sum_{c=0}^{C} exp(X_{c,h,w})}, \forall h, w \tag{13}$$

Experiments

Datasets

PASCAL VOC 2012: It is the most prevalent dataset in Weakly-supervised Semantic Segmentation, because of the moderate difficulty and quantity. To be specific, it is consisted of 20 foreground categories and one background class, divided into train set, validation set and test set at the quantities of 1464 images, 1449 images, 1456 images respectively. Besides the original data, additional images and annotations from SBD (Hariharan et al. 2011) are used which is called trainaug set at number 10582. In WSSS, all the pixel-level annotations are converted to image-level multi-label annotations in classification phase.

MS COCO 2012: MS COCO is the main dataset in object detection, also some works in WSSS report their results on the version of 2012. It is a challenging dataset which provides more categories than PASCAL VOC, and it is smaller in average object size. MS COCO 14 dataset ranges from 0 to 90, among them, 80 categories are valid foreground with one background, and other 10 categories are not evaluated. The train set contains 82081 images and the number of validation set is 40137. Save to PASCAL VOC 2012, the evaluation metric is Mean intersection over union (mIoU).

Implementation Details

Baseline: We apply our method on PMM (Li et al. 2021). It improves the generation and utilization of pseudo-masks based on SEAM (Wang et al. 2020). The backbone is ResNet-38 (Wu, Shen, and Van Den Hengel 2019), besides PMM uses Res2Net-101 (Gao et al. 2019) and ScaleNet-101 (Li et al. 2019) and achieve state-of-the-art results on both VOC and COCO. In this paper, we verify our methods on these three backbones and ResNet-101(He et al. 2016). Our settings are same to PMM. The different is that, we add our URN in this segmentation codebase. Specifically, the codebase is MMSegmentation (MMlab 2020), and PMM uses PSPnet (Zhao et al. 2017) to get same results reported in the paper of SEAM. For VOC, the batch size is 16 on 8 GPUs at learning rate 0.005 for 20000 iterations in ploy policy. And training COCO requres 32 GPUSs at batch size 64 and learning rate 0.02. The iteration number is 40000. For the augmentation, the resized image is limited between 512 to 2048 with crop size 512×512 for training. Besides, random flip and distortion are applied as transformation. The crop size of test phase is same as train phase, with dense-CRF as post-processing. For the classification part, use the original code, because we only focus on the segmentation phase. Note that, all the models are pretrained from imagenet.

URN: The weight mask is saved offline in PNG format to save storage space. For easily implementation, we concatenate pseudo-mask and weight mask into one image to aviod multiple ground-truths and multiple preprocessing. We split the weight mask and restore its range to 1 from 0 during rewight. The scale factors S in Eq.(7) are set to {0.15,0.2,0.25,4,5,6} in VOC and {0.4,0.5,0.6,2,3,4} in COCO. The specific numbers are set by visualization without strict requirements. The minimum loss value t which divide uncertain and uncertain pixels is determined by experiment at 0.05.

Pseudo-mask Distillation: As the segmentation model $\mathcal{F}_S$ is not the final deploy model, which returns prediction feature map and cyclic mask. Thus, we propose a practical and effective distillation scheme based on the cyclic pseudo-mask. If we have a teacher segmentation model $\mathcal{F}_S$ whose prediction M_c is in high quality, we train student models via this pseudo-mask for better guidance. In this paper, we set Res2Net-101 as the teacher model in VOC, and ScaleNet-101 is the teacher model in COCO. We also analyze the gains of Pseudo-mask Distillation in Tab.4.

Comparison with State-of-the-Art

We conduct experiments on PASCAL VOC 2012 and MS COCO 2014 as Tab.1. The results in the top part are all trained from image-level annotations. And the works in the middle part introduce extra models or datasets. Except (Khoreva et al. 2017) we don't list the methods supervised by bounding box, as we focus on image-level supervision methods. We organize the results by datasets and compare our method to the previous state-of-the-arts as follows.

PASCAL VOC 2012: We list the results on both validation set and test set on VOC. The most used backbone is ResNet-101, so we divide these results from others by color. For ResNet-101 our URN surpass the works in 2020 more than 3% at 69.5% on validation set. Another prevalent backbone is ResNet-38 which is wider than ResNet-101. Our result on this backbone is 70.6% on test set. It is higher than our baseline PMM by 1.6% and is about 4% higher than CONTA which iterates 3 times on both classification and segmentation. Also, it is 4.9% higher than SEAM. For the two stronger backbones ScaleNet-101 and Res2Net-101, we achieve 70.1% and 71.2% respectively.

MS COCO 2014: Compare to VOC, COCO is more challenging. There are 4 times categories and the training images are 8 times more, also the object size is smaller than VOC. Due to these difficulties, a few of works report the results on it, and there are results from validation set only before. For ResNet-38, SEAM's result is 31.7% and PMM is at 36.7%, while our URN achieve SOTA at 40.5%. As well as other backbones, our results are higher than 40.5%, among them the best mIoU is 41.5%.

Comparison with Methods with Extra Information: In the middle part of Tab.1, we list some latest methods based on saliency, object proposals or extra datasets. We can see that these works are higher than the methods without extra information in average. But our results still beyond them at every dataset and backbone. Especially, URN is 7.1% higher

Method	Backbone	Supervision	VOC12 val	VOC12 test	COCO14 val
BFBP (Saleh et al. 2016)	VGG16	$\mathcal{I}$	46.6†	48.0†	20.4†
SEC (Kolesnikov and Lampert 2016)	VGG16	$\mathcal{I}$	50.7†	51.7†	22.4†
AffinityNet (Ahn and Kwak 2018)	ResNet-38	$\mathcal{I}$	61.7	63.7	-
IRNet (Ahn, Cho, and Kwak 2019)	ResNet-50	$\mathcal{I}$	63.5	64.8	-
OAA (Jiang et al. 2019)	ResNet-101	$\mathcal{I}$	63.9	65.6	-
ICD (Fan et al. 2020)	ResNet-101	$\mathcal{I}$	64.1	64.3	-
SEAM (Wang et al. 2020)	ResNet-38	$\mathcal{I}$	64.5	65.7	31.7 -
SSDD (Shimoda and Yanai 2019)	ResNet-101	$\mathcal{I}$	64.9	65.5	-
CONTA (Zhang et al. 2020)	ResNet-38	$\mathcal{I}$	66.1	66.7	32.8
SC-CAM (Chang et al. 2020)	ResNet-101	$\mathcal{I}$	66.1	65.9	-
Sun et al. (Sun et al. 2020)	ResNet-101	$\mathcal{I}$	66.2	66.9	-
PMM(Li et al. 2021)	ResNet-38	$\mathcal{I}$	<u>68.5</u>	<u>69.0</u>	<u>36.7</u>
PMM(Li et al. 2021)	ScaleNet-101	$\mathcal{I}$	<u>67.1</u>	<u>67.7</u>	<u>40.2</u>
PMM(Li et al. 2021)	Res2Net-101	$\mathcal{I}$	<u>70.0</u>	<u>70.5</u>	<u>35.7</u>
DSRG (Huang et al. 2018)	ResNet-101	$\mathcal{I}$+ESnE+MSRA-B	61.4	63.2	26.0†
FickleNet (Lee et al. 2019)	ResNet-101	$\mathcal{I}$+$\mathcal{S}$	64.9	65.3	-
SDI (Khoreva et al. 2017)	ResNet-101	$\mathcal{I}$+$\mathcal{D}$+BSDS	65.7	67.5	-
OAA˙ (Jiang et al. 2019)	ResNet-101	$\mathcal{I}$+$\mathcal{S}$	65.2	66.4	
SGAN (Yao and Gong 2020)	ResNet-101	$\mathcal{I}$+$\mathcal{S}$	67.1	67.2	<u>33.6</u>
ICD (Fan et al. 2020)	ResNet-101	$\mathcal{I}$+$\mathcal{S}$	67.8	68.0	-
Li et al. (Li et al. 2020b)	ResNet-101	$\mathcal{I}$+$\mathcal{S}$	<u>68.2</u>	<u>68.5</u>	28.4†
LIID (Liu et al. 2020)	ResNet-101	$\mathcal{I}$+SOP	66.5	67.5	-
LIID (Liu et al. 2020)	Res2Net-101	$\mathcal{I}$+SOP	69.4	70.4	-
URN	ResNet-38	$\mathcal{I}$	**69.4**$_{0.9}$	**70.6**$_{1.6}$	**40.5**$_{3.8}$
URN	ResNet-101	$\mathcal{I}$	**69.5**$_{1.3}$	**69.7**$_{1.2}$	**40.7**$_{7.1}$
URN	ScaleNet-101	$\mathcal{I}$	**70.1**$_{3.0}$	**70.8**$_{3.1}$	**40.8**$_{0.6}$
URN	Res2Net-101	$\mathcal{I}$	**71.2**$_{1.2}$	**71.5**$_{1.0}$	**41.5**$_{5.8}$

Table 1: Performance companion with state-of-the-art WSSS methods on VOC 2012 and COCO 2014. The middle part lists the methods with extra supervision. $\mathcal{I}$, $\mathcal{S}$, $\mathcal{D}$ indicate supervisions of image-level tag, saliency, detection respectively. SOP is segment-based object proposals. Other extra information is about data. † indicates backbone of VGG. For each backbone, the second result has an underline and the best is bold with gain in lower right corner.

than SGAN on ResNet-101 on COCO, which demonstrates the strong effectiveness of our method.

Ablation Studies

Effectiveness of URN: To verify the effectiveness of our method, we set experiments based on our baseline PMM which is the current state-of-the-art. We want to show that, our method works well even on a high baseline. The evaluated dataset is PASCAL VOC 2012 on the validation set with the metric of mIoU. We compare the results between baseline and ours on four backbones in Tab.3. It shows that our method achieves significant improvements on the all backbones. We push the previous state-of-the-art to 71.2% from 70.0%, and the gains on ResNet-101, ResNet-38 are 1.4% and 0.9% respectively. Among them, ScaleNet-101 raise most with a growth of 2.9%. These results suggest that our method is effective and it works well on varied backbones.

Weight Threshould: There is an important hyper paramter in URN which controls the intensity of reweight. To select the optimal value of t in Eq.(12), we search it via experiments as shown in Tab.3. The backbone is Res2Net on PASCAL VOC 2012 validation set. If t is set to 1, the weights

Backbone	Baseline	Ours
ResNet-38	68.5	69.4
ResNet-101	68.1	69.5
ScaleNet-101	67.2	70.1
Res2Net-101	70.0	71.2

Table 2: Comparison with baselines on PASCAL VOC 2O12 validation set.

are same to original weights, and 0 means droping all the uncertain pixels. It suggests that the pixels under uncertain area are useful, thus it is not wise to drop all of them. 0.05 is the optimal value, it means the weights of certain pixels is 20 times than the uncertain pixels, and it works best.

Pseudo-mask Distillation: Distillation is a widely used technique. In this paper we apply the mechanism by a very simple but practical way. We select the pseudo-masks whose quality are best as teacher and train other student models via them. Since student models are the deployed models, and they are able to learn the knowledge from teacher masks without increasing inference time. We verify the gains of Pseudo-mask Distillation in Tab. 4. We can see that the

t	mIoU
1	70.0
0.5	70.3
0.1	70.8
0.05	**71.2**
0	70.2

Table 3: Selection of threshold t in Eq.(12).

Backbone	Basline	without PD	with PD
ScaleNet-101	67.1	68.3	70.1
Res2Net-101	70.0	71.2	-

Table 4: Results of Pseudo-mask Distillation (PD). Note: Res2Net is the teacher model without PD.

ScaleNet-101 and Res2Net-101 both raise 1.2% after applying URN without PD, and ScaleNet-101 raises 1.8% from Pseudo-mask Distillation. Note that Res2Net is the teacher in VOC, thus there is no result with PD.

Probability vs. Uncertainty: As the uncertainty estimation replies on the prediction of segmentation, and some reweight methods use probability as metric in tasks like classification. We try to use the probability as loss weight for comparison. Specifically, we detach the prediction feature map and apply softmax to make the probability as loss weight. As shown in Tab.5, in the same settings, the mIoU on VOC validation is 70.5%, while our URN is 71.2%. So probability has positive correlation to noise, but URN works better. Because in WSSS there are many false pixels with high probability.

Visual Comparison

To verifies the effectiveness of our methods, we compare our results with SEAM and PMM on VOC12 validation set in Fig.4, and the visualizations of COCO are shown in Fig.5. In these two figures, we can see that our method performs better than these two baselines.

Conclusion

In summary, the noise is closely connected with the response scale, and we estimate the uncertainty by response scaling to simulate the various response scale. Then we mitigate the noise in segmentation optimization with the uncertainty from scaling. Besides, we propose Pseudo-mask Distillation for the weaker backbones in implementation. Experimentally, we verify the improvements in obligation studies and compare our results to the previous state-of-the-art methods, which demonstrates the effectiveness of our method.

Acknowledgements

Method	Backbone	VOC val
Probability	Res2Net-101	70.5
URN	Res2Net-101	71.2

Table 5: Comparison of two reweight schemes.
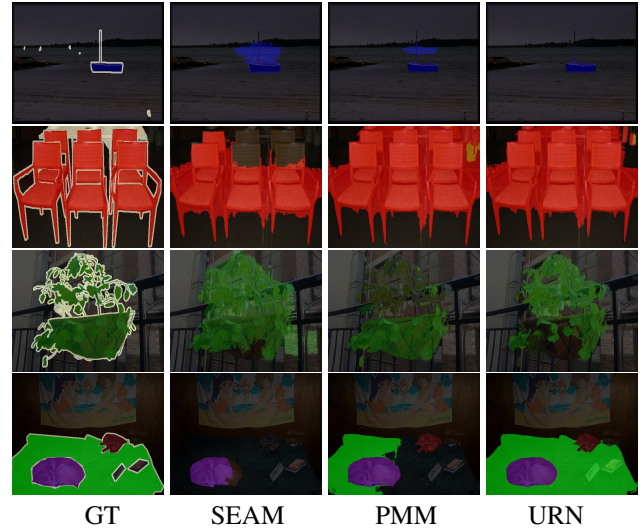

GT　　SEAM　　PMM　　URN

Figure 4: Visual comparison on PASCAL VOC 2012 validation set.

GT　　SEAM　　PMM　　URN
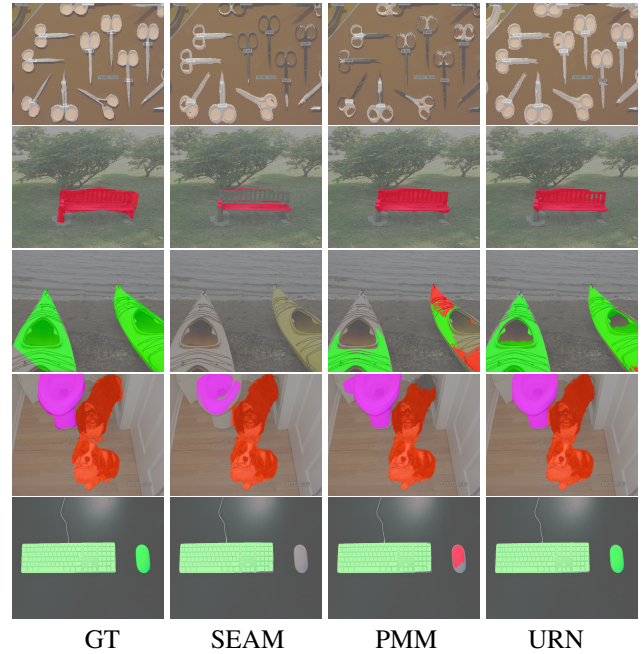
Figure 5: Visual comparison on MS COCO 2014 validation set.

References

Ahn, J.; Cho, S.; and Kwak, S. 2019. Weakly Supervised Learning of Instance Segmentation With Inter-Pixel Relations. In *CVPR*, 2204–2213.

Ahn, J.; and Kwak, S. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 4981–4990.

Bearman, A.; Russakovsky, O.; Ferrari, V.; and Fei-Fei, L. 2016. What the point: Semantic segmentation with point supervision. In *ECCV*, 549–565. Springer.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight Uncertainty in Neural Network. In *ICML*, volume 37 of *Proceedings of Machine Learning Research*, 1613–1622. PMLR.

Chang, Y.-T.; Wang, Q.; Hung, W.-C.; Piramuthu, R.; Tsai, Y.-H.; and Yang, M.-H. 2020. Weakly-supervised semantic segmentation via sub-category exploration. In *CVPR*, 8991–9000.

Dai, J.; He, K.; and Sun, J. 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 1635–1643.

Denker, J. S.; and LeCun, Y. 1990. Transforming neural-net output levels to probability distributions. In *NeurIPS*, 853–859.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

Fan, J.; Zhang, Z.; Song, C.; and Tan, T. 2020. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *CVPR*, 4283–4292.

Gal, Y.; and Ghahramani, Z. 2016a. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 1050–1059. PMLR.

Gal, Y.; and Ghahramani, Z. 2016b. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *ICML*, volume 48 of *Proceedings of Machine Learning Research*, 1050–1059. PMLR.

Gao, S.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; and Torr, P. H. 2019. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*.

Ghosh, A.; Kumar, H.; and Sastry, P. 2017. Robust Loss Functions under Label Noise for Deep Neural Networks. In *AAAI*.

Goldberger, J.; and Ben-Reuven, E. 2017. Training deep neural-networks using a noise adaptation layer. In *ICLR*.

Han, B.; Yao, J.; Niu, G.; Zhou, M.; Tsang, I.; Zhang, Y.; and Sugiyama, M. 2018a. Masking: A New Perspective of Noisy Supervision. In *NeurIPS*.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018b. Co-teaching: robust training deep neural networks with extremely noisy labels. In *NeurIPS*.

Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; and Malik, J. 2011. Semantic contours from inverse detectors. In *ICCV*, 991–998. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Huang, Z.; Wang, X.; Wang, J.; Liu, W.; and Wang, J. 2018. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, 7014–7023.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.-J.; and Fei-Fei, L. 2018. MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels. In *ICML*.

Jiang, P.-T.; Hou, Q.; Cao, Y.; Cheng, M.-M.; Wei, Y.; and Xiong, H.-K. 2019. Integral object mining via online attention accumulation. In *ICCV*, 2070–2079.

Jungo, A.; Balsiger, F.; and Reyes, M. 2020. Analyzing the Quality and Challenges of Uncertainty Estimations for Brain Tumor Segmentation. *Frontiers in Neuroscience*, 14: 282.

Kendall, A.; Badrinarayanan, V.; and Cipolla, R. 2015. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*.

Khoreva, A.; Benenson, R.; Hosang, J.; Hein, M.; and Schiele, B. 2017. Simple Does It: Weakly Supervised Instance and Semantic Segmentation. In *CVPR*.

Kim, Y.; Yim, J.; Yun, J.; and Kim, J. 2019. Nlnl: Negative learning for noisy labels. In *CVPR*, 101–110.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

Kolesnikov, A.; and Lampert, C. H. 2016. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 695–711. Springer.

Kumar, A.; and Ithapu, V. K. 2019. SeCoST: Sequential Co-Supervision for Weakly Labeled Audio Event Detection. *arXiv preprint arXiv:1910.11789*.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *NeurIPS*, 6402–6413.

Lee, J.; Kim, E.; Lee, S.; Lee, J.; and Yoon, S. 2019. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *CVPR*, 5267–5276.

Lee, K.-H.; He, X.; Zhang, L.; and Yang, L. 2017. CleanNet: Transfer Learning for Scalable Image Classifier Training with Label Noise. *arXiv preprint arXiv:1711.07131*.

Lee, S.; Prakash, S. P. S.; Cogswell, M.; Ranjan, V.; Crandall, D.; and Batra, D. 2016. Stochastic Multiple Choice Learning for Training Diverse Deep Ensembles. In *NeurIPS*, 2119–2127.

Lee, S.; Purushwalkam, S.; Cogswell, M.; Crandall, D.; and Batra, D. 2015. Why M heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*.

Li, X.; Yu, L.; Chen, H.; Fu, C.-W.; and Heng, P.-A. 2018. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. *BMVC*.

Li, X.; Yu, L.; Chen, H.; Fu, C.-W.; Xing, L.; and Heng, P.-A. 2020a. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2): 523–534.

Li, X.; Zhou, T.; Li, J.; Zhou, Y.; and Zhang, Z. 2020b. Group-Wise Semantic Mining for Weakly Supervised Semantic Segmentation. *arXiv preprint arXiv:2012.05007*.

Li, Y.; Kuang, Z.; Chen, Y.; and Zhang, W. 2019. Data-driven neuron allocation for scale aggregation networks. In *CVPR*, 11526–11534.

Li, Y.; Kuang, Z.; Liu, L.; Chen, Y.; and Zhang, W. 2021. Pseudo-mask Matters inWeakly-supervised Semantic Segmentation. *arXiv preprint arXiv:2108.12995*.

Lin, D.; Dai, J.; Jia, J.; He, K.; and Sun, J. 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 3159–3167.

Liu, Y.; Wu, Y.-H.; Wen, P.-S.; Shi, Y.-J.; Qiu, Y.; and Cheng, M.-M. 2020. Leveraging instance-, image-and dataset-level information for weakly supervised instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

MacKay, D. J. 1992. A practical Bayesian framework for back-propagation networks. *Neural computation*, 4(3): 448–472.

MMlab. 2020. MMSegmentation, an Open Source Semantic Segmentation Toolbox. https://github.com/open-mmlab/mmsegmentation. Accessed: 2021-05-04.

Natarajan, N.; Dhillon, I. S.; Ravikumar, P. K.; and Tewari, A. 2013. Learning with noisy labels. In *NeurIPS*, 1196–1204.

Pathak, D.; Krahenbuhl, P.; and Darrell, T. 2015. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 1796–1804.

Patrini, G.; Rozza, A.; Menon, A.; Nock, R.; and Qu, L. 2017. Making neural networks robust to label noise: a loss correction approach. In *CVPR*.

Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, L.; and Hinton, G. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Pinheiro, P. O.; and Collobert, R. 2015. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 1713–1721.

Reed, S.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; and Rabinovich, A. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.

Rupprecht, C.; Laina, I.; DiPietro, R.; Baust, M.; Tombari, F.; Navab, N.; and Hager, G. D. 2017. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *ICCV*, 3591–3600. IEEE.

Saleh, F.; Aliakbarian, M. S.; Salzmann, M.; Petersson, L.; Gould, S.; and Alvarez, J. M. 2016. Built-in foreground/background prior for weakly-supervised semantic segmentation. In *ECCV*, 413–432. Springer.

Shimoda, W.; and Yanai, K. 2019. Self-supervised difference detection for weakly-supervised semantic segmentation. In *ICCV*, 5208–5217.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning Structured Output Representation using Deep Conditional Generative Models. In *NeurIPS*, 3483–3491.

Sukhbaatar, S.; Bruna, J.; Paluri, M.; Bourdev, L.; and Fergus, R. 2014. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*.

Sun, G.; Wang, W.; Dai, J.; and Van Gool, L. 2020. Mining cross-image semantics for weakly supervised semantic segmentation. In *ECCV*, 347–365. Springer.

Tanaka, D.; Ikami, D.; Yamasaki, T.; and Aizawa, K. 2018. Joint optimization framework for learning with noisy labels. In *CVPR*.

Tang, M.; Djelouah, A.; Perazzi, F.; Boykov, Y.; and Schroers, C. 2018. Normalized Cut Loss for Weakly-Supervised CNN Segmentation. In *CVPR*.

Tanno, R.; Worrall, D. E.; Ghosh, A.; Kaden, E.; Sotiropoulos, S. N.; Criminisi, A.; and Alexander, D. C. 2017. Bayesian image quality transfer with CNNs: exploring uncertainty in dMRI super-resolution. In *MICCAI*, 611–619. Springer.

Vahdat, A. 2017. Toward Robustness against Label Noise in Training Deep Discriminative Neural Networks. In *NeurIPS*.

Veit, A.; Alldrin, N.; Chechik, G.; Krasin, I.; Gupta, A.; and Belongie, S. 2017. Learning From Noisy Large-Scale Datasets With Minimal Supervision. In *CVPR*.

Wang, X.; Hua, Y.; Kodirov, E.; and Robertson, N. M. 2019a. IMAE for Noise-Robust Learning: Mean Absolute Error Does Not Treat Examples Equally and Gradient Magnitude's Variance Matters. *arXiv preprint arXiv:1903.12141*.

Wang, X.; Kodirov, E.; Hua, Y.; and Robertson, N. M. 2019b. Derivative Manipulation For Adjusting Emphasis Density Function: A General Example Weighting Framework. *arXiv preprint arXiv:1905.11233*.

Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; and Bailey, J. 2019c. Symmetric Cross Entropy for Robust Learning with Noisy Labels. In *ICCV*, 322–330.

Wang, Y.; Zhang, J.; Kan, M.; Shan, S.; and Chen, X. 2020. Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation. In *CVPR*.

Wu, Z.; Shen, C.; and Van Den Hengel, A. 2019. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90: 119–133.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *CVPR*.

Xu, J.; Schwing, A. G.; and Urtasun, R. 2015. Learning to Segment Under Various Forms of Weak Supervision. In *CVPR*.

Xu, Y.; Cao, P.; Kong, Y.; and Wang, Y. 2019. L_dmi: An information-theoretic noise-robust loss function. In *NeurIPS*.

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

Yu, L.; Wang, S.; Li, X.; Fu, C.-W.; and Heng, P.-A. 2019a. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *MICCAI*, 605–613. Springer.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019b. How does Disagreement Help Generalization against Label Corruption? In *ICML*.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.; and Sun, Q. 2020. Causal intervention for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2009.12547*.

Zhang, Z.; and Sabuncu, M. R. 2018. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In *NeurIPS*.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *CVPR*, 2881–2890.

Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; and Torr, P. H. S. 2015. Conditional Random Fields as Recurrent Neural Networks. In *2015 IEEE International Conference on Computer Vision*, 1529–1537.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*, 2921–2929.