W2P: Switching from Weak Supervision to Partial Supervision for Semantic Segmentation

Fangyuan Zhang[1,2] **, Tianxiang Pan**[1,2]**, Junhai Yong**[1,2]**, Bin Wang**[1,2*]

[1]School of Software, Tsinghua University, China
[2]Beijing National Research Center for Information Science and Technology (BNRist), China
zhangfy19@mails.tsinghua.edu.cn, ptx9363@gmail.com, yongjh@tsinghua.edu.cn, wangbins@tsinghua.edu.cn

Abstract

Current weakly-supervised semantic segmentation (WSSS) techniques concentrate on enhancing class activation maps (CAMs) with image-level annotations. Yet, the emphasis on producing these pseudo-labels often overshadows the pivotal role of training the segmentation model itself. This paper underscores the significant influence of noisy pseudo-labels on segmentation network performance, particularly in boundary region. To address above issues, we introduce a novel paradigm: Weak to Partial Supervision (W2P). At its core, W2P categorizes the pseudo-labels from WSSS into two unique supervisions: trustworthy clean labels and uncertain noisy labels. Next, our proposed partially-supervised framework adeptly employs these clean labels to rectify the noisy ones, thereby promoting the continuous enhancement of the segmentation model. To further optimize boundary segmentation, we incorporate a noise detection mechanism that specifically preserves boundary regions while eliminating noise. During the noise refinement phase, we adopt a boundary-conscious noise correction technique to extract comprehensive boundaries from noisy areas. Furthermore, we devise a boundary generation approach that assists in predicting intricate boundary zones. Evaluations on the PASCAL VOC 2012 and MS COCO 2014 datasets confirm our method's impressive segmentation capabilities across various pseudo-labels.

Introduction

Weakly-supervised semantic segmentation (WSSS) (Lee et al. 2019; Wang et al. 2020; Lee, Kim, and Yoon 2021; Chen et al. 2022; Lee et al. 2021b; Li et al. 2022; Xu et al. 2022) achieves segmentation by using image-level labels instead of precise pixel-wise annotations in fully-supervised semantic segmentation (FSSS) (Xie et al. 2021; Chen et al. 2018; Long, Shelhamer, and Darrell 2015). It follows the two-stage paradigm: a classification model generates class activation maps (CAMs) (Zhou et al. 2016) as pseudo-labels, which are then used to train a segmentation network.

Contemporary methods in WSSS primarily focus on enhancing CAMs during the initial stage. Despite progressive improvements in the metrics for pseudo-labels, these enhancements do not translate into improved segmentation performance. Table 1 demonstrates that although

	SEAM	IRN	Δ	EDAM	AMN	Δ
T-mIoU	63.6	66.3	2.7↑	68.0	72.2	4.2↑
T-mAcc	80.2	74.2	6.0↓	80.4	77.3	3.1↓
V-mIoU	64.5	63.5	1.0↓	70.9	69.5	1.4↓

Table 1: We analyze the quality of pseudo-labels and segmentation performance of various WSSS methods on the PASCAL VOC 2012 $train$ (T) and val (V) set using the mean Intersection-over-Union (mIoU) and Accuracy (mAcc).

AMN (Lee, Kim, and Shim 2022) achieves a higher mIoU compared to EDAM (Wu et al. 2021), it does not lead to improved segmentation performance in the *val* set. A similar occurrence can be observed between SEAM (Wang et al. 2020) and IRN (Ahn, Cho, and Kwak 2019). This inconsistency can be attributed to the presence of inherent noise during the initial stage, resulting in lower accuracy as shown in Table 1. Figure 1(b) depicts how training the FSSS network with low-quality pixels in the pseudo ground-truths causes it to converge towards sub-optimal solutions.

In this study, we contend that it is crucial to give greater emphasis to robust learning with noisy labels for the second stage rather than solely focusing on optimizing the CAMs in the first stage. Unlike image classification, the process of pixel-wise learning in noisy environments is more intricate. This complexity arises from the difficulty of accurately predicting the main supervisory signals in the boundary regions of WSSS (Rong et al. 2023; Wang et al. 2022). To tackle these challenges, as illustrated in Figure 1(c), we propose a novel framework named *Weak to Partial Supervision* (W2P), which consists of two modules: the *boundary-preserving noise detection* (BPND) module and the *Partially-supervised Learning* (PSL) module.

The BPND module trains a segmentation model using a few iterations with pseudo-labels generated by established WSSS methods. Motivated by the early-learning theory (Liu et al. 2020, 2022), which indicates a significant discrepancy between the noisy pseudo-labels and model predictions in the early stages of training, the BPND module employs pixel-wise "small-loss" metric to distinguish between clean and noisy pseudo-labels. While "small-loss"criterion effectively selects trustworthy clean pseudo-labels, it fails

*Corresponding author

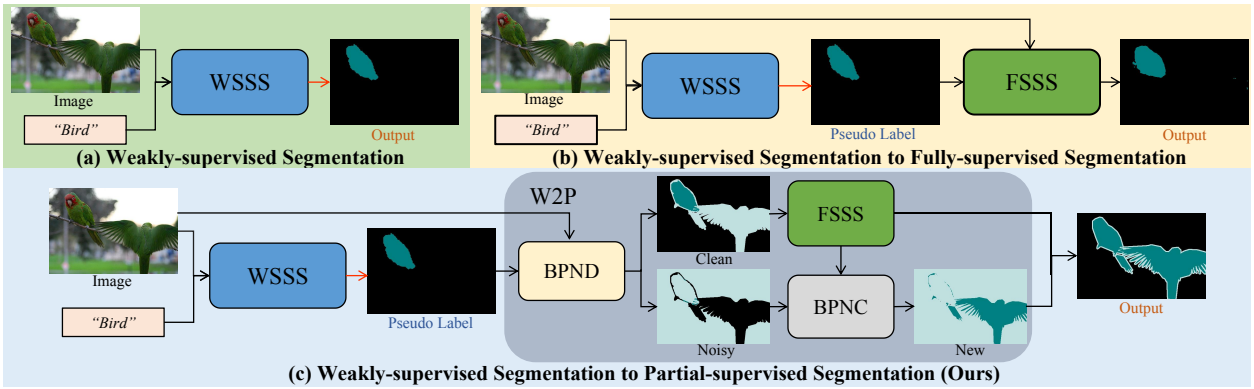

Figure 1: An overview of various training paradigms for WSSS: (a) Basic weakly-supervised semantic segmentation. (b) Weakly-supervised to fully-supervised semantic segmentation. (c) Our proposed W2P framework.

to preserve boundary regions that present challenges during early learning. To address this, we incorporate low-level semantics and introduce a boundary-preserving noise detection strategy based on the superpixel structure. This strategy utilizes structural constraints to shift regions near boundaries with small losses toward the interior of the object, thereby preserving an accurate boundary representation.

After boundary-preserving noise detection, we train the partially-supervised learning framework using clean partial supervisions, with model predictions as reliable supervision for noisy regions. To reduce errors and ensure robust training, we propose a class-wise adaptive threshold paradigm to filter out unreliable predictions. To identify challenging boundaries, we suggest a boundary-preserving noise correction algorithm. Additionally, we propose a boundary generation strategy to enhance boundary predictions by duplicating and transferring high-confidence object boundaries between images, creating artificial boundary pixels.

In summary, the main contributions are as follows:

- We propose a new WSSS paradigm that converts weak to partial supervision by redefining the second stage of WSSS as the segmentation problem with noisy labels.
- We present a boundary-preserving noise detection module for pseudo-labels selection, while preserving complete boundary structure.
- Upon selecting reliable supervision, the partially-supervised learning module offers complementary signals for the unreliable parts. We further introduce a boundary preserving noise correction and boundary generation strategy to enhance boundary segmentation.

Related Work

Weakly-Supervised Semantic Segmentation

Existing image-level WSSS methods commonly use CAMs (Zhou et al. 2016) as initial seeds to generate pseudo segmentation labels. Due to the inherent discrepancies between semantic labels and pixel-wise annotations, it is challenging to achieve complete coverage of the object region. To address this issue, current solutions target at enhancing the quality of CAMs by utilizing the transformer's attention module (AFA (Ru et al. 2022), MCTformer (Xu et al. 2022), ViT-PCM (Rossetti et al. 2022)), separating the foreground regions with contrastive learning (PPC (Du et al. 2022), ToCo (Ru et al. 2023)), iteratively erasing (OC-CSE (Kweon et al. 2021), ECS (Sun et al. 2021), AEFT (Yoon et al. 2022)), changing the optimization target (RIB (Lee et al. 2021a), PMM (Li et al. 2021)), generating more accurate seeds (Su et al. 2021) and incorporating additional signals such as saliency maps (ICD (Fan et al. 2020), EPS (Lee et al. 2021b), DRS (Kim, Han, and Kim 2021), AuxSeg (Xu et al. 2021) and SANCE (Li, Fan, and Zhang 2022)).

However, these methods primarily focus on generating pseudo-labels, with little attention given to training with these labels. URN (Li et al. 2022) proposes a method to identify noisy labels by estimating uncertainty across different scales. ADELE (Liu et al. 2022) adaptively corrects pseudo-labels for different categories. BECO (Rong et al. 2023) introduces a co-training paradigm for correcting noise. These methods neglect the importance of boundary regions in the learning process. In contrast to these methods, we propose a new Weak-to-Partial framework that shifts the focus in WSSS from relying on CAMs to robustly handle noisy labels. W2P generates partial supervision and progressively refine noisy parts, with boundary-preserving segmentation.

Robust Learning with Noisy Labels

Learning with noisy labels for classification tasks has recently received significant attention. Existing solutions can be categorized into two groups: approaches that aim to reduce the negative impact of noisy labels, and techniques that focus on fixing inaccurate annotations. The former group includes methods that reduce the negative impact of noisy labels through improved robust optimization (Zhang and Sabuncu 2018; Ma et al. 2020; Wang et al. 2019), designed regularization techniques (Liu et al. 2020; Tanaka et al. 2018), robust architecture (Chen and Gupta 2015; Goldberger and Ben-Reuven 2017; Han et al. 2018a), and sample selection (Han et al. 2018b; Jiang et al. 2018). These algorithms emphasize the role of clean data during the training process, neglecting the information in the noisy labeled data. To address this issue, recent studies (Li, Socher, and

Hoi 2020; Yi et al. 2023; Xia et al. 2022) propose the progressive correction of noisy supervision through model predictions, achieving state-of-the-art performance.

Despite the significant progress in noisy learning for classification tasks, there has been limited research on the more prevalent problem of robust learning in segmentation tasks. Only a few studies (Shu, Wu, and Li 2019; Guo and Yuan 2022) have developed noise-robust architectures for medical image segmentation. ADELE (Liu et al. 2022) exploits the phenomenon of early learning in semantic segmentation to adaptively correct noisy labels with various categories. BECO (Rong et al. 2023) proposes a co-training paradigm for noise correction. These methods mainly tackle noisy segmentation as a pixel-wise classification task, disregarding the challenging boundary regions in WSSS tasks. In contrast, we propose the W2P paradigm, which performs boundary-preserving noise detection and correction.

Method

Overview

Weakly-Supervised Semantic Segmentation (WSSS) aims to train a segmentation network using a weakly annotated dataset denoted as $X = \{(I, y)\}$, where I is the image and $y = [y_1, y_2, ..., y_C]^T$ denotes the corresponding image-level label, with C denoting the number of object categories.

We present a new framework called *Weak-to-Partial* (W2P) that focuses on the second stage of the WSSS. Using the provided dataset X, we train a WSSS model and generate initial pseudo-labels for training images. Now we obtain a new dataset: $X_s = \{(I, Y)\}$, where $Y \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}}$ denotes the inaccurate segmentation map with the spatial size H, W. To train a segmentation model on this noisy dataset, the W2P framework incorporates a *boundary preserving noisy detection* (BPND) module to focus on identifying precise partial supervision, providing dependable signals for the *partially-supervised learning* (PSL) module. The PSL module leverages high-quality partial supervisions from Y_{clean} to enhance the quality of the Y_{noisy} using model predictions.

Noisy Label Generation

The generation of noisy labels follows the first stage of the existing WSSS methods. It extracts the initial segmentation from CAMs extracted from the classification model trained with X. As the first and second stages of WSSS are independent, our W2P can be effortlessly applied to existing WSSS solutions. To showcase the generalization capability of W2P, we choose three WSSS baseline methods as generators: IRN(Ahn, Cho, and Kwak 2019), ReCAM(Chen et al. 2022), and AMN(Lee, Kim, and Shim 2022).

Boundary Preserving Noise Detection

Noisy labels are incorporated in the W2P training framework to improve the segmentation network. As in previous studies (Huang et al. 2022; Sui, Zhang, and Wu 2022; Li, Socher, and Hoi 2020), the quality of clean pseudo labels Y_{clean} significantly impacts the model performance in the context of noisy robust learning. In our W2P, the introduction of erroneous annotations adversely affects the performance of the subsequent PSL module, making it unable to provide reliable labels for the noisy region. Consequently, this impacts the overall performance.

Therefore, designing a strategy to distinguish between noisy and clean pseudo-labels is crucial. The prevalent technique for noise separation is the small-loss criterion, where labels misaligned with predictions during early-learning stage are considered noisy, while the rest are deemed clean. Based on this theory, we propose a noise detection strategy using the loss of the predictions and pseudo-labels as an indicator of inconsistency. The loss is calculated as:

$$P = f_a(I, \theta_a), \quad L = \mathrm{CE}(P, Y), \tag{1}$$

where f_a represents the segmentation model trained over t epochs with parameters θ_a. The model prediction is denoted as P, and the Cross-Entropy loss function is denoted as CE. Noisy labels are identified as pseudo-labels with losses exceeding a designated threshold.

While the *small-loss* efficiently identifies clean pseudo-labels, it faces difficulty in differentiating between boundary and noisy pixels. Precisely learning pixel-wise annotations in the boundary area is challenging due to semantic confusion. Consequently, the small-loss metric naturally results in a loss of boundary supervision. To tackle this issue, we propose **B**oundary **P**reserving **N**oise **D**etection (BPND), which retains boundaries while eliminating noisy pseudo-labels.

BPND aims to leverage the structural prior and spatial correlation within pixels to preserve object boundaries. This is achieved by employing superpixels (Achanta et al. 2010; den Bergh et al. 2012), which are clusters of low-level features commonly used in visual and shape analysis tasks. Figure 3 illustrates that pixels in both the boundary and inner regions often have the same categorical semantics within each superpixel. Therefore, the preservation of the boundary region can be accomplished by relying on the semantics of the internal region. Next, we present our BPND strategy, along with the use of superpixel representation.

Concretely, given an image I, we calculate the loss L between its prediction P and pseudo-labels Y using Equation.1. The superpixel representation of I is $SP = \{SP_i\}_{i=1}^{H\times W}$, where $SP_i \in 1, 2, ..., K$ and K is the number of superpixels. Pixel j belongs to superpixel k if $SP_j = k$. The pixels within the same superpixel are expected to have consistently labeled semantic labels. Therefore, we propose reducing high loss in boundary regions by exploiting low loss in the interior regions. In practice, we employ a averaging operation to compute the smoothed loss $L_k = \frac{1}{N_k}\sum_{j:SP_j=k} l_j$.

where $N_k = |j : SP_j = k|$ is the size of the superpixel and l_j is the loss of pixel j. In practice, as shown in Figure 2, we calculate the smoothed loss for each superpixel as follows:

$$L = \mathrm{R}(L, K), SP = \mathrm{OH}(SP), \tag{2}$$

$$L_s = SP \otimes \mathrm{N}(L \odot SP), \tag{3}$$

where $\mathrm{R}(, b)$ represents the repeat operation along the channel dimension for b times and OH refers to the one-hot operation for superpixel indexes. $\otimes$ and $\odot$ denote the matrix

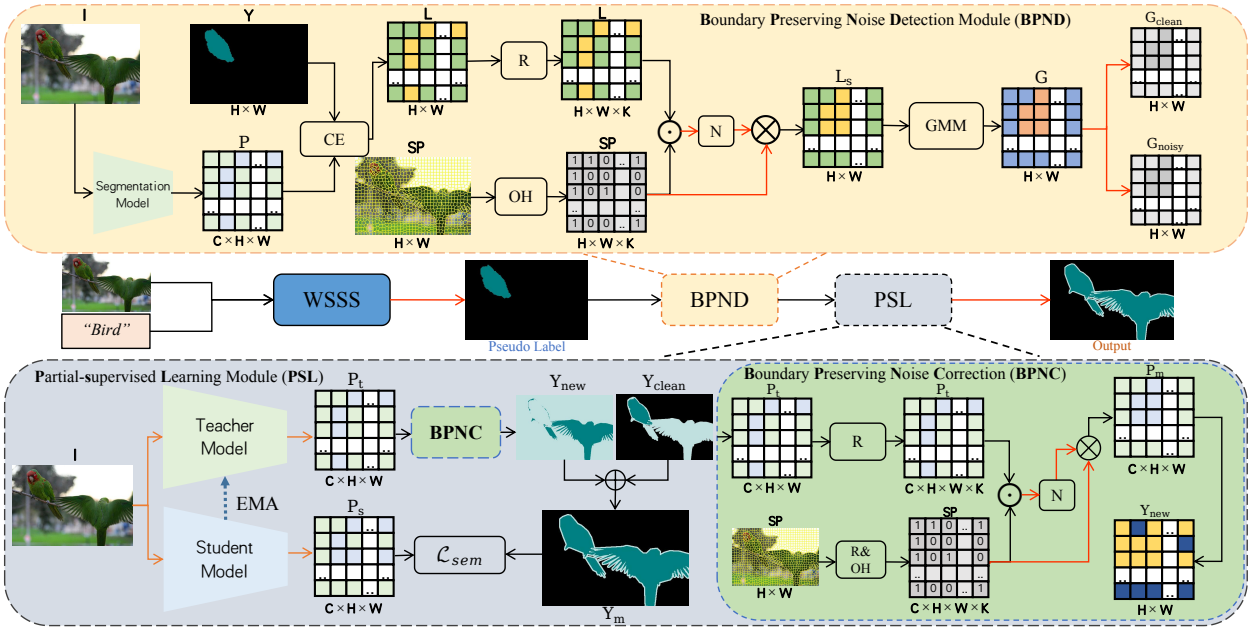

Figure 2: Overview of our Weak-to-Partial (W2P) method: Firstly, we generate initial noisy labels using images and image-level labels with the WSSS module. Next, we utilize the proposed Boundary Preserving Noise Detection (BPND) module to separate the noisy labels and generate clean supervisions with two masks G_{clean} and G_{noisy}. Then, W2P incorporates a Partially-supervised Learning (PSL) module that generates complementary supervisions (Y_{new}) to refine segmentation within the region defined by G_{noisy}. Finally, we combine supervisions from Y_{new} and Y_{clean} to create complete supervisions (Y_m), which are crucial for training the segmentation model and enhancing performance.

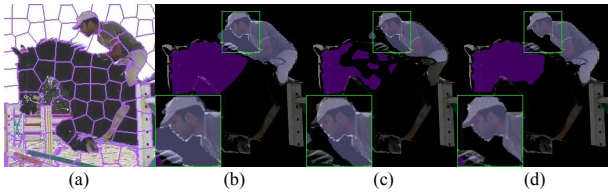

Figure 3: Visual representation of the noisy detection: (a) presents the image and its corresponding superpixels. (b) displays the original noisy label. (c) and (d) exhibit the selected reliable supervision using the "small-loss" method and the BPND module, respectively.

and pixel-wise multiplication operation, respectively. N is the normalization operation along the spatial dimensions.

After obtaining the smoothed loss L_s, several issues arise. The non-normalized cross-entropy loss poses a challenge in establishing a suitable classification threshold for clean and noisy data. Moreover, the variation in pseudo-labels across categories requires laborious task of determining appropriate thresholds for each category. To address these challenges, we propose an adaptive threshold strategy that autonomously differentiates between clean and noisy pseudo-labels across categories. Specifically, we fit a two-component Gaussian Mixture Model (GMM) (Permuter, Francos, and Jermyn 2006) with the loss L_b for category b, using the Expectation-Maximization algorithm. With the sharpness of distribution, GMM is efficiently in distinguishing the two-modality (clean and noisy) data. Consequently, the clean probability w_i for each pixel in category b is calculated using the posterior probability $p(g_b|l_i)$, where g_b corresponds to the Gaussian component with the smaller mean (smaller loss) for category b.

The problem of threshold selection, originally based on L_b, now relies on the clean probability w_i. Here, a threshold τ_1 is used to create masks G_{noisy} and G_{clean} for each category, providing partial supervision. Despite using the same threshold, the category-specific GMM effectively captures the noise distribution, enabling the generation of adaptive thresholds for different categories. The masks for clean and noisy regions, G_{clean} and G_{noisy}, are produced:

$$G^i_{clean} = \begin{cases} \text{True}, & p(g|l_i) > \tau_1 \\ \text{False}, & p(g|l_i) \leq \tau_1 \end{cases}, G_{noisy} = \sim G_{clean}. \tag{4}$$

Subsequently, partial supervisions Y_{clean} and Y_{noisy} are generated with G_{clean} and G_{noisy}, respectively. This process mitigates the presence of low-quality pseudo-labels and enhances the performance of the PSL.

Partially-supervised Learning

In this module, we propose a partially-supervised training (PSL) algorithm that utilizes the provided Y_{clean} and Y_{noisy} splits from the BPND module. Initially, we train a segmentation framework using the reliable Y_{clean}. Then, we utilize this framework to generate more reliable labels denoted as

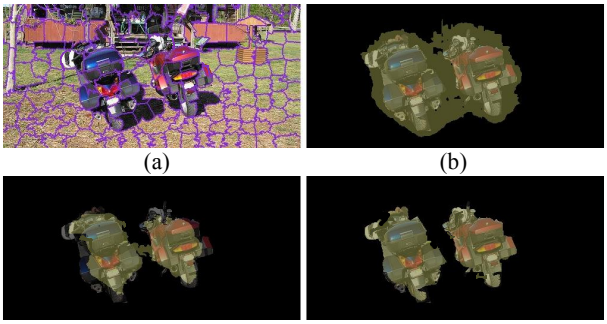

Figure 4: Visual representation of the new supervision: (a) presents the image with corresponding superpixels. (b) displays the original prediction. (c) and (d) exhibit the selected reliable supervision using the "high-confidence" metric and the BPNC module, respectively.

Y_{new}, which outperform the noisy pseudo-labels Y_{noisy}. Finally, we merge the Y_{clean} and Y_{new} labels to supervise and enhance the performance of semantic segmentation model.

Figure 2 illustrates the partially-supervised learning module consisting of a student model f_s and a teacher model f_t with parameters θ_s and θ_t, respectively. The student model f_s learns from the combined supervision Y_m derived from Y_{new} and Y_{clean}, while the teacher model f_t generates reliable supervisions Y_{new} to substitute the original noisy labels Y_{noisy}. To reduce the computational complexity and improve the reliability of the teacher model, we optimize the student model using gradient backward while updating the teacher model through exponential moving average (EMA) (Tarvainen and Valpola 2017):

$$\theta_t = \lambda * \theta_t + (1 - \lambda) * \theta_s, \tag{5}$$

where λ is the EMA coefficient. Utilizing the teacher model, we generate the merged supervision Y_m as:

$$P_t = f_t(I, \theta_t), \tag{6}$$

$$Y_{new} = \arg\max P_t, S_{new} = \max(\text{softmax}(P_t)), \tag{7}$$

$$Y_m = Y_{new} \oplus Y_{clean}, \tag{8}$$

where $\oplus$ merges the two partial supervisions, and S_{new} represents the confidence associated with Y_{new}.

The performance of the model heavily depends on the quality of Y_{new}. Accumulated errors caused by incorrect modifications in the noisy regions negatively impact PSL training. Hence, we utilize the commonly-used "high-confidence" metric to filter out predictions with low confidences that are deemed unreliable. To avoid the manual adjustment of thresholds for different categories, we present an algorithm that generates adaptive thresholds. This algorithm employs a two-component Gaussian Mixture Model (GMM) to fit the prediction confidences, denoted as S_{new}, for each category. Accurate predictions are selected based on the reliable probability $p(g_b|s_i)$, where g_b denotes the Gaussian component with a higher mean (indicating higher confidences) for category b, and s_i represents the prediction confidence of the i-th pixel in S_{new}. By utilizing adaptive thresholds, we produce the reliable supervision Y_{new} using the following method:

$$Y^i_{new} = \begin{cases} Y^i_{new}, & p(g|s_i) > \tau_2 \\ 255, & p(g|s_i) \le \tau_2 \end{cases}. \tag{9}$$

The value 255 is the "ignore" indicator during training.

While the high-threshold strategy effectively reduces noise in Figure 4, it also decreases the recall for pixels in the boundary regions. This limitation, in turn, results in a lack of crucial supervisory signal in the boundary area, thereby negatively impacting the overall performance of W2P.

To address this issue, we propose a paradigm called **b**oundary **p**reserving **n**oise **c**orrection (BPNC) and an algorithm for generating boundaries to enhance boundary prediction. The first approach extracts boundary pixels using structural constraints, while the second approach constructs boundary pixels through the copying and pasting of high-confidence areas from one image to another.

Boundary Preserving Noise Correction In BPNC, superpixels identify low confidence boundaries. Figure 4 shows that same-category pixels are usually grouped together within superpixels. Boundary regions have lower confidence, while the interior regions have higher confidence. Thus, high confidence in interior regions can identify low confidence boundaries. Similar to BPND, the smoothed prediction P_m is calculated for each superpixel as follows:

$$P_t = \text{R}(P_t, K), SP = \text{OH}(\text{R}(SP, C)), \tag{10}$$

$$P_m = SP \otimes \text{N}(P_t \odot SP). \tag{11}$$

We employ two techniques in the implementation to accelerate computation. Firstly, we only smooth the logits of the most dominant class in prediction for each superpixel. Secondly, the overall operation is performed on the predictions without up-sampling. These operations can be efficiently computed in parallel on GPUs, resulting in negligible additional computational time. With smoothed prediction, the labels Y_{new} and the confidences S_{new} are updated by:

$$Y_{new} = \arg\max P_m, S_{new} = \max(\text{softmax}(P_m)). \tag{12}$$

Boundary Generation The proposed BPNC algorithm improves boundary segmentation reliability. To capitalize on the enhanced boundaries, we suggest creating artificial boundary pixels by copying and pasting high-confidence areas between images. This allows for the generation of diverse boundary scenes by pasting accurate object segmentations onto various image backgrounds.

For the input images I_1 and I_2, along with their corresponding smoothed labels Y^1_{new}, Y^2_{new}, we aim to formulate the mixed image I_{mix} and the mixed target Y_{mix} as:

$$I_{mix} = ((1 - M) \odot I_1) \oplus (M \odot I_2), \tag{13}$$

$$Y_{mix} = ((1 - M) \odot Y^1_{new}) \oplus (M \odot Y^2_{new}). \tag{14}$$

The copy-paste mask M is formulated as:

$$M_i = \begin{cases} 1, & p(g|s^i_2) > \tau_2 \\ 0, & p(g|s^i_2) \le \tau_2 \end{cases}, \tag{15}$$

where s^i_2 represents the confidence of i-th pixel in I_2.

Overall Loss Function The overall Weak-to-Partial (W2P) loss is computed using two pairs: a mixed image I_{mix} and mixed target Y_{mix}, and an original image I and combined supervision Y_m. The loss is defined as follows:

$$P_s^{mix} = f_s(I_{mix}, \theta_s), P_s = f_s(I, \theta_s), \tag{16}$$

$$\mathcal{L}_{sem} = \mathrm{CE}(P_s^{mix}, Y_{mix}) + \mathrm{CE}(P_s, Y_m). \tag{17}$$

Since the teacher model exhibits superior robustness to noise, it is exclusively retained for the inference phase.

Experiments

Experimental Settings

Our experiments use the benchmarks PASCAL VOC 2012 and MS COCO 2014 for WSSS. PASCAL VOC 2012 has 10,582 training images, 1,449 validation images, and 1,456 testing images across 21 categories. MS COCO 2014 has 81 categories with 82,783 training images and 40,504 validation images. The mean Intersection-over-Union (mIoU) is the evaluation metric we use. We use the SLIC (Achanta et al. 2010) algorithm for superpixel generation. To certify the effectiveness of W2P, we present extensive ablation studies on PASCAL VOC 2012 val dataset.

In the first stage of WSSS, we use IRN to generate pseudo-labels, unless specified. The W2P framework in the second stage utilizes DeeplabV3+ with a ResNet101 backbone pretrained on ImageNet as the segmentation network, following prior studies. We exclude general tricks like an output stride of 8 and COCO pretrained models. During inference, we employ established practices and use multi-scale techniques along with dense CRF.

The W2P framework's hyperparameters need minimal tuning, with threshold values τ_1 and τ_2 set to 0.9 and λ set to 0.99. The BPND stage is trained for 8 epochs on VOC and 4 epochs on COCO, while the PSL stage takes 72 epochs on VOC and 36 epochs on COCO. A batch size of 16 is used for all experiments.

Comparison with State-of-the-arts

PASCAL VOC 2012. Table 2 shows W2P's performance compared to state-of-the-art WSSS methods on PASCAL VOC 2012. W2P achieves exceptional results with mIoU of 74.0 and 73.9 using an ImageNet pretrained backbone, setting new state-of-the-art in Image-level WSSS. W2P outperforms IRN by 10.7 and 9.3, as well as ReCAM (5.5 and 5.5) and AMN (4.5 and 4.3), other IRN-based methods. Compared to methods using saliency maps from saliency detection models like SANCE and DRS, our method demonstrates significant superiority. Moreover, our method outperforms MCTformer and other transformer-based methods.

MS COCO 2014. We also report the performance of our method on the challenging MS COCO 2014 dataset to showcase its superiority. Table 3 presents the comparison results on the MS COCO 2014 validation set. Our W2P achieves a new state-of-the-art mIoU of 46.4, demonstrating its effectiveness on a large-scale dataset.

Method	BackBone	Val	Test
ICD(CVPR20)	ResNet101	67.8	68.0
EPS(CVPR21)	ResNet101	71.0	71.8
EDAM(CVPR21)	ResNet101	70.9	70.6
AuxSeg(ICCV21)	ResNet38	69.0	68.6
DRS(AAAI21)	ResNet101	71.2	71.4
SANCE(CVPR22)	ResNet101	72.0	72.9
IRN(CVPR19)	ResNet101	63.5	64.8
SEAM(CVPR20)	ResNet38	64.5	65.7
URN(AAAI22)	ResNet101	69.5	69.7
ReCAM(CVPR22)	ResNet101	68.5	68.4
ADELE(CVPR22)	ResNet101	69.3	68.8
PPC(CVPR22)	ResNet101	67.7	67.4
AMN(CVPR22)	ResNet101	69.5	69.6
AEFT(ECCV22)	ResNet101	70.9	71.7
BECO(CVPR23)	ResNet101	72.1	71.8
W2P (Ours)	ResNet101	**74.0**	**73.9**
AFA(CVPR22)	MiT-B1	69.3	68.8
MCTformer(CVPR22)	ResNet38	70.9	71.7
ViT-PCM(ECCV22)	ResNet101	72.1	71.8
ToCo(CVPR23)	ViT-B	69.8	70.5
W2P (Ours)	MiT-B2	**76.0**	**75.7**

Table 2: Performance of WSSS methods in mIoU on PASCAL VOC 2012 val and test.

Method	BackBone	Sup.	Val
IRN(CVPR19)	ResNet101	I	41.4
CDA(ICCV21)	ResNet38	I	33.2
RIB(NeurIPS21)	ResNet101	I	43.8
MCTformer(CVPR22)	ResNet38	I	42.0
URN(CVPR22)	ResNet101	I	40.7
BECO(CVPR23)	ResNet101	I	45.1
W2P (Ours)	ResNet101	I	**46.4**

Table 3: Performance comparison of WSSS methods in terms of mIoU on the COCO val set.

Improvement on boundary segmentation. To validate the predictions of W2P on boundary areas, we present qualitative comparisons from the VOC val set in Figure 5. Compared with IRN and BECO, our W2P improves predictions on challenging boundary areas and enhances object segmentation. Additionally, we provide quantitative results of boundary improvement in Table 4.

Ablation Study

Analysis of the proposed components. We evaluate the proposed components' effectiveness on different pseudo-labels generated by IRN, ReCAM, and AMN in Table 5. Using BPND alone trains FSSS directly with selected clean labels. Employing PSL solely means training the segmentation network with all pseudo-labels, allowing updates in all regions. Equipped with BPND, the W2P framework significantly improves performance from 65.1 to 74.0 with IRN. With M+CRF, performance is further improved from 73.5 to

IRN	ReCAM	AMN	BECO	W2P (ours)
24.7	28.0	29.2	33.4	**36.0**

Table 4: Performance of different methods in terms of boundary mIoU (Cheng et al. 2021) on VOC 2012 *val* set.

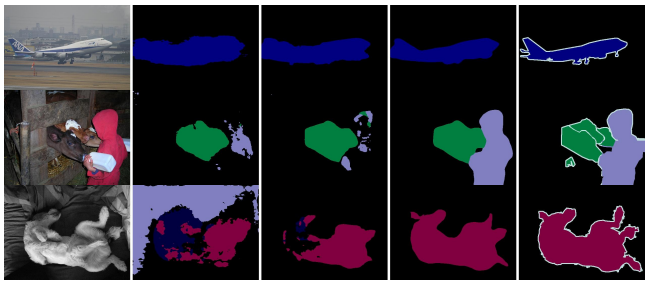

Figure 5: Visualization of segmentation results on PASCAL VOC 2012 *val* set.

BPND	PSL	M+CRF	IRN	ReCAM	AMN
			65.1	67.1	67.9
✓			67.5	69.2	70.0
	✓		70.7	70.9	71.8
✓	✓		73.1	73.5	74.5
✓	✓	✓	74.0	74.4	75.6

Table 5: Performance of different pseudo-labels in terms of mIoU on VOC 2012 *val* set. M+CRF denotes the label refinement with multi-scale test and dense-CRF.

BPNC	AT	BG	IRN	ReCAM	AMN
			70.5	70.7	71.6
		✓	71.9	71.8	73.5
✓		✓	73.6	73.6	75.0
✓	✓	✓	74.0	74.4	75.6

Table 6: Analysis of PSL module. AT: Adaptive Threshold. BG: Boundary Generation.

Module \ Method	S_{20}	S_{50}	S_{70}	AT
BPND	61.0	70.9	69.5	74.0
PSL	72.8	73.6	73.0	74.0

Table 7: Analysis of clean label selection in BPND and PSL module. AT: Adaptive Threshold. S_n: Select the n% data with the lowest loss values for BPND or select the n% data with the highest confidences for PSL.

τ_1	0.5	0.7	0.9	0.95	0.99
Segm.	72.2	73.8	74.0	73.2	67.6

Table 8: Value of τ_1.

τ_2	0.5	0.7	0.9	0.95	0.99
Segm.	73.4	73.1	74.0	73.6	73.0

Table 9: Value of τ_2.

λ	0.5	0.7	0.9	0.99	0.999
Segm.	70.3	71.0	73.7	74.0	73.9

Table 10: Value of λ.

	ADELE	BECO	W2P (Ours)
Time	1.7 h	32 m	11 m

Table 11: Training time for one epoch. h: hour. m: minute.

74.4 with ReCAM. These results validate the effectiveness of our W2P framework for different CAMs. For an in-depth analysis of the PSL module, please refer to Table 6.

Analysis of the adaptive thresholds. Our study compares the adaptive thresholds for BPND and PSL with the common strategy (Huang et al. 2022; Han et al. 2018b) of dividing data into clean and noisy. Table 7 shows our AT strategy's superior performance.

Impact with hyper-parameters. Tables 8 and 9 show the importance of setting moderate values for both τ_1 and τ_2. This ensures accurate noise removal and the utilization of reliable predictions for correction. In Table 10, the EMA framework performs well with a threshold above 0.9. We conduct an analysis of the duration required for training each epoch in Table 11, demonstrating substantial advantages of our algorithm compared to prior research.

Conclusion

In this study, we present W2P, a new weakly supervised segmentation method that focuses on the second stage of WSSS for robust learning despite noisy labels, specifically targeting boundary segmentation. By using a class-wise GMM paradigm, our noise detection module selects reliable pseudo-labels while preserving the boundary annotations accurately. Our partially-supervised learning module utilizes separate partial-supervision to learn from clean supervision and generate accurate signals for noisy parts. Additionally, we propose a boundary correction strategy and a boundary generation method to improve boundary segmentation with only image-level supervision. Extensive experiments on multiple benchmarks show that our method surpasses other state-of-the-art WSSS methods.

Acknowledgements

This work was supported by the NSFC under Grant 62072271. Jun-Hai Yong was supported by the NSFC under Grant 62021002.

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; and Süsstrunk, S. 2010. SLIC Superpixels. *epfl*.

Ahn, J.; Cho, S.; and Kwak, S. 2019. Weakly Supervised Learning of Instance Segmentation With Inter-Pixel Relations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, L.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *European Conference on Computer Vision (ECCV)*.

Chen, X.; and Gupta, A. 2015. Webly Supervised Learning of Convolutional Networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chen, Z.; Wang, T.; Wu, X.; Hua, X.; Zhang, H.; and Sun, Q. 2022. Class Re-Activation Maps for Weakly-Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cheng, B.; Girshick, R.; Dollár, P.; Berg, A. C.; and Kirillov, A. 2021. Boundary IoU: Improving Object-Centric Image Segmentation Evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

den Bergh, M. V.; Boix, X.; Roig, G.; de Capitani, B.; and Gool, L. V. 2012. SEEDS: Superpixels Extracted via Energy-Driven Sampling. In *European Conference on Computer Vision (ECCV)*.

Du, Y.; Fu, Z.; Liu, Q.; and Wang, Y. 2022. Weakly Supervised Semantic Segmentation by Pixel-to-Prototype Contrast. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fan, J.; Zhang, Z.; Song, C.; and Tan, T. 2020. Learning Integral Objects With Intra-Class Discriminator for Weakly-Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Goldberger, J.; and Ben-Reuven, E. 2017. Training deep neural-networks using a noise adaptation layer. In *International Conference on Learning Representations (ICLR)*.

Guo, X.; and Yuan, Y. 2022. Joint Class-Affinity Loss Correction for Robust Medical Image Segmentation with Noisy Labels. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part IV*.

Han, B.; Yao, J.; Niu, G.; Zhou, M.; Tsang, I. W.; Zhang, Y.; and Sugiyama, M. 2018a. Masking: A New Perspective of Noisy Supervision. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I. W.; and Sugiyama, M. 2018b. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Huang, Z.; Bao, Y.; Dong, B.; Zhou, E.; and Zuo, W. 2022. W2N: Switching from Weak Supervision to Noisy Supervision for Object Detection. In *European Conference on Computer Vision (ECCV)*.

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

Kim, B.; Han, S.; and Kim, J. 2021. Discriminative Region Suppression for Weakly-Supervised Semantic Segmentation. In *AAAI*.

Kweon, H.; Yoon, S.; Kim, H.; Park, D.; and Yoon, K. 2021. Unlocking the Potential of Ordinary Classifier: Class-specific Adversarial Erasing Framework for Weakly Supervised Semantic Segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Lee, J.; Choi, J.; Mok, J.; and Yoon, S. 2021a. Reducing Information Bottleneck for Weakly Supervised Semantic Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Lee, J.; Kim, E.; Lee, S.; Lee, J.; and Yoon, S. 2019. FickleNet: Weakly and Semi-Supervised Semantic Image Segmentation Using Stochastic Inference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lee, J.; Kim, E.; and Yoon, S. 2021. Anti-Adversarially Manipulated Attributions for Weakly and Semi-Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lee, M.; Kim, D.; and Shim, H. 2022. Threshold Matters in WSSS: Manipulating the Activation for the Robust and Accurate Segmentation Model Against Thresholds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lee, S.; Lee, M.; Lee, J.; and Shim, H. 2021b. Railroad Is Not a Train: Saliency As Pseudo-Pixel Supervision for Weakly Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, J.; Fan, J.; and Zhang, Z. 2022. Towards Noiseless Object Contours for Weakly Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, J.; Socher, R.; and Hoi, S. C. H. 2020. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *International Conference on Learning Representations (ICLR)*.

Li, Y.; Duan, Y.; Kuang, Z.; Chen, Y.; Zhang, W.; and Li, X. 2022. Uncertainty Estimation via Response Scaling for Pseudo-Mask Noise Mitigation in Weakly-Supervised Semantic Segmentation. In *AAAI*.

Li, Y.; Kuang, Z.; Liu, L.; Chen, Y.; and Zhang, W. 2021. Pseudo-mask Matters in Weakly-supervised Semantic Segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Liu, S.; Liu, K.; Zhu, W.; Shen, Y.; and Fernandez-Granda, C. 2022. Adaptive Early-Learning Correction for Segmentation from Noisy Annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-Learning Regularization Prevents Memorization of Noisy Labels. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ma, X.; Huang, H.; Wang, Y.; Romano, S.; Erfani, S. M.; and Bailey, J. 2020. Normalized Loss Functions for Deep Learning with Noisy Labels. In *International Conference on Machine Learning (ICML)*.

Permuter, H. H.; Francos, J. M.; and Jermyn, I. 2006. A study of Gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognit.*

Rong, S.; Tu, B.; Wang, Z.; and Li, J. 2023. Boundary-enhanced Co-training for Weakly Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rossetti, S.; Zappia, D.; Sanzari, M.; Schaerf, M.; and Pirri, F. 2022. Max Pooling with Vision Transformers Reconciles Class and Shape in Weakly Supervised Semantic Segmentation. In *European Conference on Computer Vision (ECCV)*.

Ru, L.; Zhan, Y.; Yu, B.; and Du, B. 2022. Learning Affinity from Attention: End-to-End Weakly-Supervised Semantic Segmentation with Transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ru, L.; Zheng, H.; Zhan, Y.; and Du, B. 2023. Token Contrast for Weakly-Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shu, Y.; Wu, X.; and Li, W. 2019. LVC-Net: Medical Image Segmentation with Noisy Label Based on Local Visual Cues. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part VI*.

Su, Y.; Sun, R.; Lin, G.; and Wu, Q. 2021. Context Decoupling Augmentation for Weakly Supervised Semantic Segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Sui, L.; Zhang, C.; and Wu, J. 2022. Salvage of Supervision in Weakly Supervised Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sun, K.; Shi, H.; Zhang, Z.; and Huang, Y. 2021. ECS-Net: Improving Weakly Supervised Semantic Segmentation by Using Connections Between Class Activation Maps. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Tanaka, D.; Ikami, D.; Yamasaki, T.; and Aizawa, K. 2018. Joint Optimization Framework for Learning With Noisy Labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, C.; Zhang, Y.; Cui, M.; Ren, P.; Yang, Y.; Xie, X.; Hua, X.; Bao, H.; and Xu, W. 2022. Active Boundary Loss for Semantic Segmentation. In *AAAI*. AAAI Press.

Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; and Bailey, J. 2019. Symmetric Cross Entropy for Robust Learning With Noisy Labels. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Wang, Y.; Zhang, J.; Kan, M.; Shan, S.; and Chen, X. 2020. Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, T.; Huang, J.; Gao, G.; Wei, X.; Wei, X.; Luo, X.; and Liu, C. H. 2021. Embedded Discriminative Attention Mechanism for Weakly Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xia, X.; Liu, T.; Han, B.; Gong, M.; Yu, J.; Niu, G.; and Sugiyama, M. 2022. Sample Selection with Uncertainty of Losses for Learning with Noisy Labels. In *International Conference on Learning Representations (ICLR)*.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Álvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xu, L.; Ouyang, W.; Bennamoun, M.; Boussaïd, F.; Sohel, F.; and Xu, D. 2021. Leveraging Auxiliary Tasks with Affinity Learning for Weakly Supervised Semantic Segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Xu, L.; Ouyang, W.; Bennamoun, M.; Boussaïd, F.; and Xu, D. 2022. Multi-class Token Transformer for Weakly Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yi, R.; Guan, D.; Huang, Y.; and Lu, S. 2023. Class-Independent Regularization for Learning with Noisy Labels. In *AAAI*.

Yoon, S.; Kweon, H.; Cho, J.; Kim, S.; and Yoon, K. 2022. Adversarial Erasing Framework via Triplet with Gated Pyramid Pooling Layer for Weakly Supervised Semantic Segmentation. In *European Conference on Computer Vision (ECCV)*.

Zhang, Z.; and Sabuncu, M. R. 2018. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhou, B.; Khosla, A.; Lapedriza, À.; Oliva, A.; and Torralba, A. 2016. Learning Deep Features for Discriminative Localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.