Out-of-Candidate Rectification for Weakly Supervised Semantic Segmentation

Zesen Cheng[1,2*] Pengchong Qiao[1,2,3*] Kehan Li[1,2] Siheng Li[4] Pengxu Wei[5]
Xiangyang Ji[4] Li Yuan[1,3] Chang Liu[4] ✉ Jie Chen[1,2,3] ✉

[1] School of Electronic and Computer Engineering, Peking University, Shenzhen, China
[2] AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China
[3] Peng Cheng Laboratory, Shenzhen, China [4] Tsinghua University [5] Sun Yat-Sen University

{cyanlaser,pcqiao,kehanli}@stu.pku.edu.cn weipx3@mail.sysu.edu.cn
{xyji,liuchang2022}@tsinghua.edu.cn {yuanli-ece,jiechen2019}@pku.edu.cn

Abstract

Weakly supervised semantic segmentation is typically inspired by class activation maps, which serve as pseudo masks with class-discriminative regions highlighted. Although tremendous efforts have been made to recall precise and complete locations for each class, existing methods still commonly suffer from the unsolicited ***Out-of-Candidate*** *(OC) error predictions that do not belong to the label candidates, which could be avoidable since the contradiction with image-level class tags is easy to be detected. In this paper, we develop a group ranking-based* ***O****ut-of-****C****andidate* ***R****ectification (OCR) mechanism in a plug-and-play fashion. Firstly, we adaptively split the semantic categories into* ***In-Candidate*** *(IC) and OC groups for each OC pixel according to their prior annotation correlation and posterior prediction correlation. Then, we derive a differentiable rectification loss to force OC pixels to shift to the IC group. Incorporating OCR with seminal baselines (e.g., AffinityNet, SEAM, MCTformer), we can achieve remarkable performance gains on both Pascal VOC (+3.2%, +3.3%, +0.8% mIoU) and MS COCO (+1.0%, +1.3%, +0.5% mIoU) datasets with negligible extra training overhead, which justifies the effectiveness and generality of OCR.* †

1. Introduction

Due to the development of deep learning, significant progress has been made in deep learning-based semantic segmentation [42, 47]. However, its effectiveness requires huge amounts of data with precise pixel-level labels. Collecting precise pixel-level labels is very time-consuming and labor-intensive, thus much research shifts attention to training effective semantic segmentation models with relatively low manual annotation cost, i.e., Weakly Supervised Semantic Segmentation (WSSS). There exist various types of weak supervision for semantic segmentation such as image-level tag labels [1, 2, 26, 61, 77], bounding boxes [14, 30, 34], scribbles [40, 57] and points [5]. In this work, we focus on WSSS based on image-level tag labels since image-level tags demand the least annotation cost, which just needs the information on the existence of the target object categories.

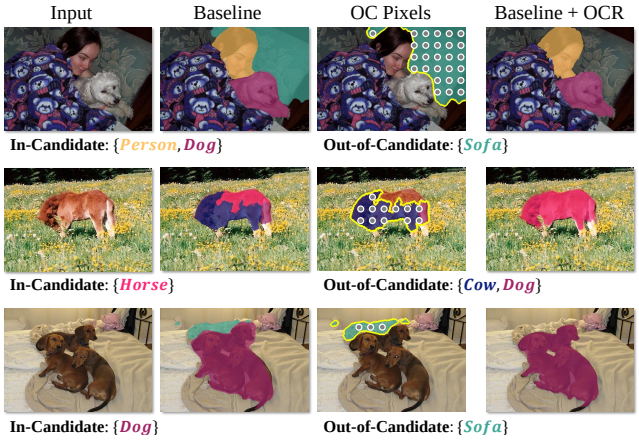


Figure 1. **Motivation of our OCR**. We visualize the segmentation results from the baseline method (e.g. SEAM) and the baseline with our proposed OCR. The predictions from baseline methods are easily disturbed by OC pixels, that is, pixels whose semantic categories are in contradiction with label candidate set (inner of the Yellow contour). Our proposed OCR can rectify these OC pixels and suppress this unreasonable phenomenon.

Most of the previous WSSS methods follow such a standard workflow [2]: 1). generating high-quality Class Activation Maps (CAM) [61, 70]; 2). generating pseudo labels from CAMs [2, 78]; 3). training segmentation networks from pseudo labels. Previous works mainly fo-

* Equal contribution ✉ Corresponding Author
† github.com/sennnnn/Out-of-Candidate-Rectification

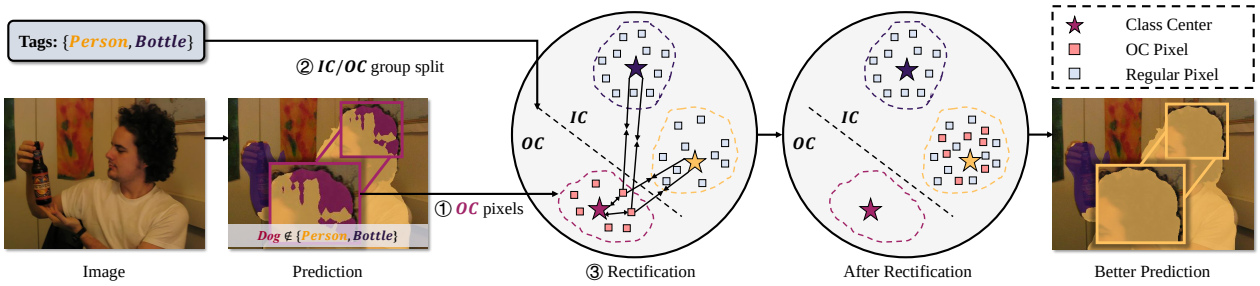


Figure 2. **Conceptual workflow of our OCR**. The OC pixels are selected out by checking if the semantic categories are in contradiction with image-level candidate tags. Then we adaptively split the categories into IC group and OC group. Finally, we utilize rectification loss for group ranking and let OC pixels escape from OC group to IC group.

cus on the first and second procedures. However, training segmentation network from pseudo labels is also vital because neural network can exploit shared patterns between pseudo labels [4] and largely improve final segmentation results [39]. But the pseudo label generation relies on high-quality CAM, while the pseudo labels usually are incomplete and imprecise because CAM only focuses on discriminative object parts and can not fully exploit object regions [7]. The existence of noise in pseudo labels provides confused knowledge to segmentation networks and results in error predictions. According to the observations in Fig. 1, the segmentation networks trained by noisy pseudo labels usually output pixels with semantic categories that do not belong to the candidate label set, i.e., image-level tag labels. This special type of prediction errors are defined as ***Out-of-Candidate*** (OC). These errors can be easily detected by checking if the semantic category of pixel is in contradiction with image-level tag labels, which is seldom considered before. For better identifying this phenomenon, we extra name these error pixels as OC pixels and name the illegal categories as OC categories. In contrast, the potentially correct categories for OC pixels are defined as ***In-Candidate*** (IC) categories.

To suppress the occurrence of OC phenomenon, we propose group ranking-based **O**ut-of-**C**andidate **R**ectification (OCR) to rectify OC pixels from OC categories to IC categories by solving a group ranking problem (i.e., the prediction score of IC group needs to be larger than the prediction score of OC group). In Fig. 2, OCR is illustrated as three procedures: OC pixels selection, IC/OC categories group split and rectification. Firstly, we find out OC pixels whose classification result is in contradiction with image-level tag labels. Secondly, we adaptively split the classes into IC classes group and OC classes group for each OC pixel by considering prior label correlation information from the image-level tag labels and posterior label correlation information from the network prediction. Finally, rectification loss is used to modulate the distance between OC pixels and class centers of IC group and OC group. It constraints that the OC pixels and OC class centers are pushed away and the OC pixels and IC class centers are pulled closer so that those OC pixels are rectified to correct classes.

Out-of-Candidate Rectification (OCR) is designed in a plug-and-play style to provide reasonable supervision signals with trivial training costs and to improve evaluation results with no extra cost for inference. To fairly show the effectiveness and generality, we adopt the same settings of several previous methods (AffinityNet [2], SEAM [61], MCTformer [70]) and evaluate our proposed OCR on the PASCAL VOC 2012 and MS COCO 2014 datasets. Experiments demonstrate that our OCR improves the performance of final segmentation results. Specifically, our module improves AffinityNet, SEAM and MCTformer by 3.2%, 3.3% and 0.8% mIoU on PASCAL VOC 2012 dataset and 1.0%, 1.3% and 0.5% mIoU on MS COCO 2014 dataset.

2. Related works

Most current WSSS approaches are built upon such a paradigm: 1). Generating high-quality CAMs; 2). Refining CAM for pseudo label generation; 3). Training segmentation network from pseudo labels. The core technology of this paradigm is CAMs [79]. However, the raw CAMs can only cover the discriminative part of object regions so it is challenge to provide clean supervision for segmentation networks. Previous methods focus on improving these three procedures of the paradigm for easing the limitations:

Generating high-quality CAMs. How to generate high-quality CAMs is the key research topic of WSSS because the improvement of CAMs can boost the whole workflow from the source. The key is to let CAMs cover object regions as precisely and completely as possible. A few methods design heuristic strategies, like "Hide & Seek" [52] and Erasing [62], adopted on images [36, 73, 76] or feature maps [12, 27, 32] to force the network to exploit novel regions rather than only discriminative regions. Besides, Other strategies utilize sub-categories [7], self-supervised learning [11, 51, 61], contrastive learning [17, 29, 68, 80] and cross-image information [21, 37, 54] to generate pre-

cise and complete CAMs. Recently, Vision Transformer (ViT) [16] is proposed as a new generation of visual neural networks. Because of the long-range context nature, vision transformer can better capture semantic context so some recent works begin to utilize ViT as the classification network for generating high-quality CAMs [22, 48, 49, 70]. What's more, some of previous methods try to improve the local receptive field of classification network [28, 63, 71]. Improving CAMs by modifying the standard classification loss is also researched by previous works, e.g., [64, 77].

Generating pseudo labels from CAMs. After acquiring high-quality CAMs, we decode them into pseudo labels. Originated from AffinityNet [2], some works focus on exploiting pixel-level affinity learning which can facilitate the generation of higher-quality pseudo labels from CAMs [1, 21, 60, 69, 78]. Besides, prior other works design diverse mechanisms for pseudo label generation, for instance, densecrf [74], texture exploiting [3] and multi-estimations [20] for better pseudo label generation.

Training from pseudo labels. Although dozens of advanced CAM generation or refinement technology are proposed, the existence of noise in pseudo labels is inevitable. In this context, some works explore about how to better acquire clean segmentation supervision from pseudo labels. For example, URN [38] uses uncertainty estimation to mitigate noise of pseudo labels when training segmentation networks. Different from previous researches, our works mainly focus on solving OC pixels which are a special type of prediction error firstly defined by us.

3. Method

3.1. Preliminaries

We first define WSSS using the following setup. Let $\mathcal{X}$ be the input image space, and $\mathcal{Y} = \{1, 2, ..., C\}$ be the output image-level label space. After sampling from the input image space and output label space, we can define a training dataset $\mathcal{D} = \{(x_i, S_i)\}_{i=1}^{n}$, where each tuple comprises an image $x_i \in \mathcal{X}$ and an image-level label set $S_i \subset \mathcal{Y}$. Equivalent to the supervised semantic segmentation setup, the goal of WSSS is to obtain a pixel classifier that can identify the real class of each pixel. The main difference is that the WSSS setup can not be accessible to exact pixel-level labels. This limitation lets segmentation algorithms struggle to acquire correct pixel-level supervision signals without interference. A basic assumption of WSSS is that the ground truth label of pixels $y_i^{(x,y)}$ belongs to image x_i is concealed in label set S_i, i.e., $y_i^{(x,y)} \in S_i$, but it's invisible to learning models. According to this assumption, previous methods for WSSS leverage class activation map (CAM) [22, 79] to generate pseudo segmentation labels $\hat{y}_i = \{\hat{y}_i^{(x,y)} | \hat{y}_i^{(x,y)} \in S_i\}_{x=1,y=1}^{H \times W}$ which are a type of noisy estimation of real labels y_i. With these pseudo segmentation labels, we can define an extra segmentation noisy dataset $D_s = \{(x_i, \hat{y}_i)\}_{i=1}^{n}$. Then a segmentation network f_θ is used to fit segmentation noisy dataset D_s for generating final segmentation results z_i.

3.2. Overall Pipeline

To prove the generality of our approach, we build our pipeline upon several prominent and representative baseline methods: AffinityNet [2], SEAM [61] and MCTformer [70]. All of these methods follow a standard workflow: 1). Generating CAMs; 2). Refining CAM with affinity learning; 3). Training segmentation network.

Generating CAMs. For acquiring CAMs, we first pretrain a multi-label classifier f'_θ. Then we use a multi-label classification loss $\mathcal{L}_{cls}$ to supervise the classifier:

$$\mathcal{L}_{cls} = E_{(x_i, S_i) \sim D} \Big[\sum_{j=1}^{C} - S_i^j \log(\sigma(h_i^j)) - (1 - S_i^j) \log(1 - \sigma(h_i^j)) \Big], \tag{1}$$

where $\sigma(\cdot)$ is sigmoid function, h_i^j is the classification logits of j-th category and i-th sample, i.e., $h_i^j = \mathrm{GAP}(f'_\theta(x_i))$. $f'_\theta(x_i) \in \mathbb{R}^{C \times H \times W}$. GAP is the global average pooling operation. After pretraining, we need to normalize the initial CAMs to get "seed CAM" $\mathcal{P}$:

$$\mathcal{P}_i^j = \frac{f'_\theta(x_i)^j}{\max(f'_\theta(x_i)^j)}. \tag{2}$$

Refining CAMs with affinity learning. Ahn et al. [2] propose AffinityNet to learn the affinities between adjacent pixels from the reliable seeds of "seed CAM". Then the AffinityNet is used to predict an affinity matrix to refine "seed CAM" to pseudo mask labels by random walk. Following the settings of baselines, we adopt this refinement.

Training segmentation network. Following the settings of baseline methods [2, 61, 70] and some previous works [69, 74, 76], we choose DeepLabv1 [9] based on ResNet38 [66] as the segmentation network f_θ. The segmentation network is used to fit pseudo segmentation labels $\hat{y}$ and the training objective is below:

$$\mathcal{L}_{seg} = E_{(x_i, \hat{y}_i) \sim D_s} \left[\sum_{j=1}^{C} -\hat{y}_i^j log(\frac{exp(z_i^j)}{\sum_{k=1}^{C} exp(z_i^k)}) \right], \tag{3}$$

where z_i is its logits output, i.e., $f_\theta(x_i) = z_i$. In order to make the formula more clear, we ignore the pixel location index (x, y) and sample index i. Because of the label noise, OC errors easily occurs during this phrase. To resist OC errors, our method modify the standard training objective and introduce an extra loss:

$$\mathcal{L} = \mathcal{L}_{seg} + \alpha \mathcal{L}_{rec}, \tag{4}$$

where $\mathcal{L}_{rec}$ is the rectification loss which will be introduced below and α is the loss modulation coefficient.

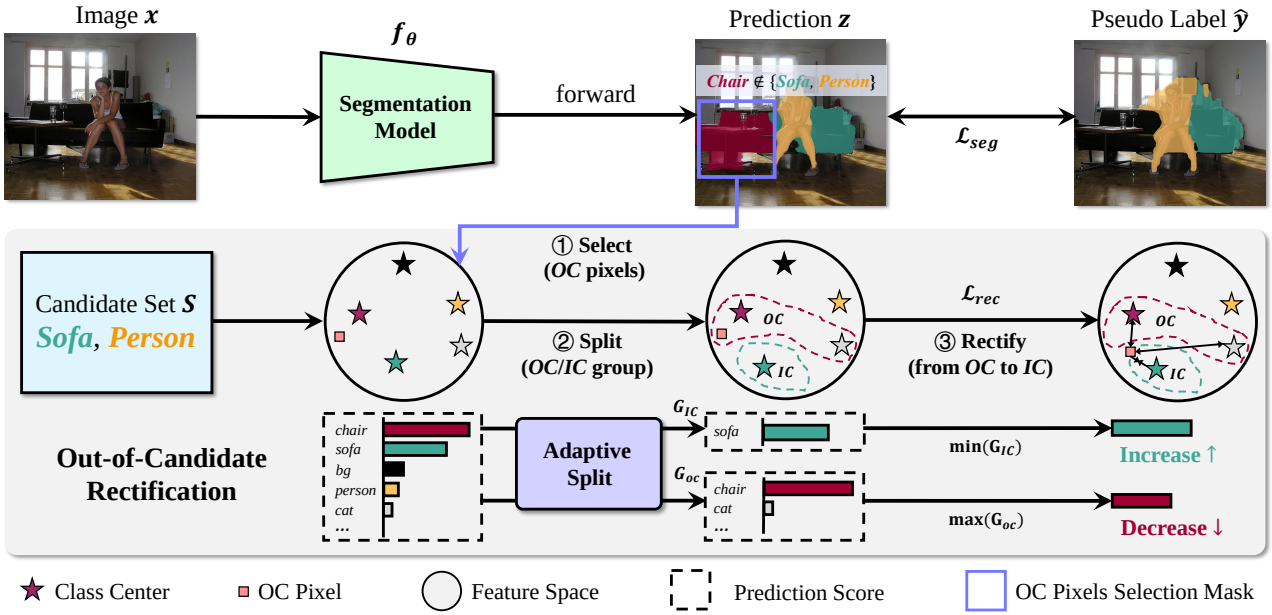


Figure 3. **Detailed workflow of our OCR** which illustrates the behavior about how to rectify an OC pixel from OC classes group to IC classes group. Note that IC classes group is just the candidate set of correct class for single pixel. So, we adopt an adaptive strategy to filter out useless classes in IC classes group when splitting classes into IC classes group and OC classes group, which can reduce the probability of containing incorrect classes in IC classes group. For instance, both "*sofa*" and "*person*" belong to tag labels. However, the "*sofa*" is the correct class rather than "*person*" and involving "*person*" in IC classes group is not beneficial to rectify OC pixels to correct class. The adaptive split strategy filter out "*person*" class according to the correlation information for correct rectification.

3.3. Out-of-Candidate Rectification

To our best knowledge, previous methods ignore the OC problem during training segmentation network from pseudo segmentation labels. We are the first to design relative mechanism (OCR) to suppress the occurrence of OC pixels. The proposed OCR is comprised of three parts: OC pixels selection, IC and OC group split and rectification loss. The three steps are introduced below.

OC Pixels Selection. When training segmentation network, previous methods only utilize the pseudo segmentation labels $\hat{y}$. We find that the prior information provided by candidate label set S is critical. With the prior information, those OC pixels can be easily detected. We implement this detection by designing a mask m_{oc} (for a single pixel):

$$m_{oc} = \begin{cases} 1, & \underset{k}{\text{argmax}}(z^k) \in \overline{S}, \\ 0, & \underset{k}{\text{argmax}}(z^k) \in S \cup \{bg\}, \end{cases} \tag{5}$$

where z^k is the segmentation logits of i-th class and bg is the background class.

IC and OC Group Split. We define two groups: IC group G_{ic} and OC group G_{oc}. We require IC group and OC group to satisfy such a group ranking relationship:

$$z^k > z^l \quad \text{s.t. } \forall k \in G_{ic}, \forall l \in G_{oc}. \tag{6}$$

Intuitively, we can assign those classes which belong to label candidate set with background, i.e., $k \in S \cup \{bg\}$, as IC group and those classes which do not belong to label candidate set, i.e., $l \notin S \cup \{bg\}$, as OC group. But the latent correlation between categories is not considered. We first count the prior correlation matrix $\mathcal{M}$ according to the co-occurrence of different classes:

$$\mathcal{M}_{k,l} = \frac{\sum_{i=1}^{L} \mathbb{1}_{k \in S_i, l \in S_i}}{L}, \tag{7}$$

where L is the number of samples in whole dataset, S_i denotes image-level tag set of i-th sample and $\mathcal{M}_{k,l}$ means the correlation score between k-th class and l-th class. Then we define anchor class:

$$\mathcal{A} = \underset{j \in S_i \cup \{bg\}}{\text{argmax}} \left(z^j\right), \tag{8}$$

where $\mathcal{A}$ denotes the class index of anchor class. For OC classes group, it can be easily identified by the contradiction with image-level tags. For IC classes group, we hope to contain as few useless classes as possible while keeping the correct class. Note that anchor class has a high probability of being the ground truth so that we use anchor class as a ruler to filter out useless classes in image-level tags for building IC classes group. According to the analysis above, we develop an adaptive strategy to split IC and OC Group:

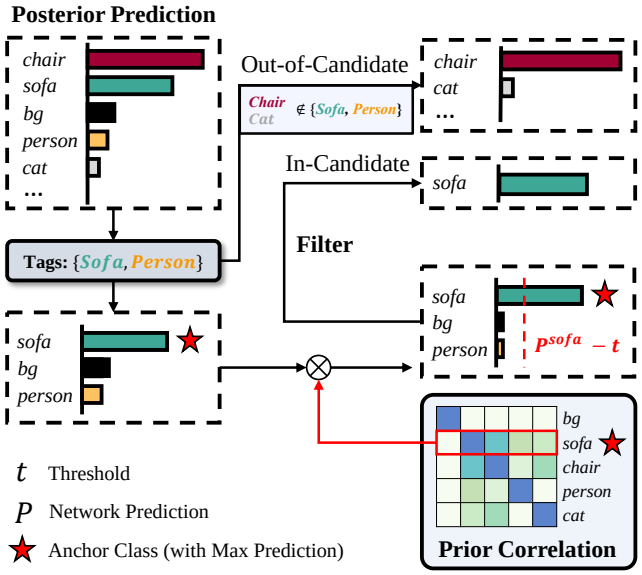


Figure 4. **Adaptive Split.** The OC classes group are selected out according to the contradiction with the tag labels. For IC classes group, we firstly select anchor class $\mathcal{A}$ (i.e, max prediction class). Specifically, the anchor class is "*sofa*". Then we use prior correlation of anchor class $\mathcal{M}_{\mathcal{A}}$ to modulate the prediction score of classes in tags S. Finally, those classes in tags whose prediction score is lower than threshold ($P^{\mathcal{A}} - t$) are filtered.

$$G_{ic} = \left\{k | k \in S \cup \{bg\}, P^{A} - P^{k} \times \mathcal{M}_{\mathcal{A},k} < t\right\}, \tag{9}$$

$$G_{oc} = \{l \,|\, l \notin S \cup \{bg\}\}, \tag{10}$$

where P is the posterior probability prediction from network (i.e., segmentation logits after softmax operation) and t is a threshold for filtering useless classes for IC classes group. To better understand the adaptive split strategy, the split process is illustrated in Fig. 4.

Rectification Loss. The core idea of OCR is to rectify those OC pixels and let these pixels have higher activation for IC classes group than OC classes group. To achieve this, we formulate a group ranking problem and expect Eq. 6 to hold. The optimization objective can be defined as:

$$\max_{l \in G_{oc}} z^{l} < \min_{k \in G_{ic}} z^{k}, \tag{11}$$

where z_i^k denotes the classes in the IC group and z_i^l denotes the classes in the OC group, which can be also written as $\max_{l \in G_{oc}} z^{l} - \min_{k \in G_{ic}} z^{k} < 0$. The Eq. 11 means we force the minimum of the OC pixels logits for the categories in IC group G_{ic} to be larger than the maximum of the OC pixels logits for the categories in OC group G_{oc}. To this purpose, we refer to some wonderful works in metric learning [25, 50, 55, 58, 59] and semi-supervised learning [45] for making Eq. 11 as a loss function:

$$\mathcal{L}_{rec} = \max_{k \in G_{oc}} z^{l} - \min_{k \in G_{ic}} z^{k} + \Delta, \tag{12}$$

where Δ is a value margin that can make the loss function more scalable. But this loss function can not directly attend the gradient backward process of gradient descent optimization since the max and min functions are globally non-differentiable. To make equation $\mathcal{L}_{rec}$ differentiable, we adopt smooth approximation from previous research [44]:

$$\max(z^1, z^2, ..., z^n) \approx \log(\sum_{i=1}^{n} e^{z^i}). \tag{13}$$

Based on the above functional approximation, we derive our rectification loss $\mathcal{L}_{rec}$ for single pixel as:

$$\mathcal{L}_{rec} = \max_{k \in G_{oc}} z^{k} - \min_{l \in G_{ic}} z^{l} + \Delta \tag{14}$$

$$= \max_{k \in G_{oc}} z^{k} + \max_{l \in G_{ic}} (-z^{l}) + \Delta \tag{15}$$

$$\approx \log\left[\sum_{l \in G_{oc}} e^{z^{l}+\Delta} \times \sum_{k \in G_{ic}} e^{-z^{k}}\right] \tag{16}$$

To avoid excessive optimization, we use *ReLU* function to rectify loss:

$$\mathcal{L}_{rec} = ReLU\left[\log\left(\sum_{k \in G_{ic}} e^{-z^{k}} \times \sum_{l \in G_{oc}} e^{z^{l}+\Delta}\right)\right] \tag{17}$$

We convert *ReLU* to its smooth approximation [23] for acquiring a gradient friendly loss:

$$ReLU(z) = \max(z, 0) \approx \log(1 + e^{z}) \tag{18}$$

The final formula of rectification loss is

$$\mathcal{L}_{rec} = m_{oc} \log\left[1 + \sum_{k \in G_{ic}} e^{-z^{k}} \times \sum_{l \in G_{oc}} e^{z^{l}+\Delta}\right] \tag{19}$$

where m_{oc} is the OC pixel selection mask.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate our method on two datasets, *i.e.*, PASCAL VOC 2012 [18] and MS COCO 2014 [41]. **PASCAL VOC** has 1,464, 1,449, and 1,456 images for training (train), validation (val) and test sets, respectively. It has 20 object classes and one background class. Following the common practice of prior works [7, 33, 53, 61, 69, 76], an augmented set of 10,582 images, with additional data from [24], was used for training. Furthermore, **MS COCO** consists of 80 object classes and one background class whose training and validation sets contain 82,081 and 40,137 images, respectively. Following [13, 35], we remove images without target classes and adopt the ground-truth labels of COCO stuff [6].

Method	Seg	mIoU (%)	
		val	*test*
‡BES [ECCV20] [8]	V2-Res101	65.7	66.6
*MCIS [ECCV20] [54]	V2-Res101	66.2	66.9
*ICD [CVPR20] [19]	V2-Res101	67.8	68.0
‡AdvCAM [CVPR21] [33]	V2-Res101	68.1	68.0
*NSROM [CVPR21] [72]	V2-Res101	68.3	68.5
*GroupWSSS [AAAI21] [37]	V2-Res101	68.7	69.0
†EDAM [CVPR21] [65]	V2-Res101	70.9	70.6
‡AMR [AAAI22] [46]	V2-Res101	68.8	69.1
AFA [CVPR22] [49]	MiT-B1 [67]	66.0	66.3
†AffinityNet [CVPR18] [2]	V1-Res38	61.7	63.7
†SSDD [ICCV19] [51]	V1-Res38	64.9	65.5
†SEAM [CVPR20] [61]	V1-Res38	64.5	65.7
†CONTA [NeurIPS20] [75]	V1-Res38	66.1	66.7
‡CDA [ICCV21] [53]	V1-Res38	66.1	66.8
†CPN [ICCV21] [76]	V1-Res38	67.8	68.5
‡OC-CSE [ICCV21] [31]	V1-Res38	68.4	64.2
MCTformer [CVPR22] [70]	V1-Res38	71.9	71.6
*AffinityNet	V1-Res38	61.7	63.7
*OCR+AffinityNet	V1-Res38	**64.9** ↑ **3.2**	**65.2** ↑ **1.5**
†SEAM	V1-Res38	64.5	65.7
†OCR+SEAM	V1-Res38	**67.8** ↑ **3.3**	**68.4** ↑ **2.7**
MCTformer	V1-Res38	71.9	71.6
OCR+MCTformer	V1-Res38	**72.7** ↑ **0.8**	**72.0** ↑ **0.4**

Table 1. **Main Results on Pascal VOC** [18] *val* and *test* split. Seg denotes the segmentation networks used by models. For instance, V1-Res38 denotes Deeplabv1 [9] based on ResNet38 and V2-Res101 is DeeplabV2 [10] based on ResNet101. *, † and ‡ denote models using VGG16, ResNet38 or ResNet50 as the classification network backbone. Besides, segmentation and classification of SS-WSSS [3], URN [38] and AFA [49] share same network. The classification network of MCTformer is DeiT-S [56].

Evaluation protocol. To be consistent with previous works [33, 70], we adopt the mean Intersection-over-Union (mIoU) to evaluate the semantic segmentation performance on the *val* set of two datasets. The semantic segmentation results on the PASCAL VOC *test* set are acquired from the official PASCAL VOC online evaluation server.

Implementation details. In line with our baseline methods (*i.e.*, AffinityNet [2], SEAM [61], MCTformer [70]), we choose DeepLab-LargeFOV (V1) [9] based on ResNet38 [66] backbone network whose output stride is 8 as our segmentation network. All of the backbone networks are pre-trained on ImageNet [15]. Note that we calculate a prior correlation score matrix $\mathcal{M}$ by counting the co-occurrence between different classes, the detailed scores are available in the appendix. The threshold t is set to 0.2. At train time, we use SGD whose momentum and weight decay are 0.9 and 5e-4 as our optimizer. The initial learning rate is 1e-3 which is multiplied by an exponential decay factor during training process. We train our models for 30 epochs with a batch size of 16. For data augmentation, the training images are randomly rescaled with a scale ratio from 0.7 to 1.3 and then are cropped to 321×321. At test time, we use test-time augmentation and DenseCRFs with the hyper-parameters suggested in [9] for post-processing.

Method	Cls	Seg	mIoU (%)
			val
Luo *et al.* [AAAI20] [43]	Res101	V2-VGG16	29.9
GroupWSSS [AAAI21] [37]	VGG16	V2-VGG16	28.4
URN [AAAI22] [38]	Res101	V2-Res101	40.7
AFA [CVPR22] [49]	MiT-B1	MiT-B1	38.9
*AffinityNet [CVPR18] [2]	Res38	V1-Res38	29.5
SEAM [CVPR20] [61]	Res38	V1-Res38	31.9
CONTA [NeurIPS20] [75]	Res38	V1-Res38	32.8
OC-CSE [ICCV21] [31]	Res38	V1-Res38	36.4
CDA [ICCV21] [53]	Res38	V1-Res38	33.2
MCTformer [CVPR22] [70]	Res38	V1-Res38	42.0
*AffinityNet	Res38	V1-Res38	29.5
OCR+AffinityNet	Res38	V1-Res38	**30.5** ↑ **1.0**
SEAM	Res38	V1-Res38	31.9
OCR+SEAM	Res38	V1-Res38	**33.2** ↑ **1.3**
MCTformer	DeiT-S	V1-Res38	42.0
OCR+MCTformer	DeiT-S	V1-Res38	**42.5** ↑ **0.5**

Table 2. **Main Results on MS COCO** [41] *val* split. Cls and Seg denote the classification backbone and segmentation network used by models, respectively. Note that AffinityNet doesn't provide official evaluation results on MS COCO dataset, the results of AffinityNet [2](*) are implemented by us.

4.2. Main Results

For calibrating the training settings of segmentation network, we mainly adopt our OCR into baseline methods (e.g., AffinityNet, SEAM and MCTformer) which use ResNet38 based deeplabv1 as segmentation network.

Pascal VOC. Tab. 1 provides the comparison results of our OCR against representative methods on Pascal VOC *val* split and *test* split. As shown in Tab. 1, our OCR consistently increases the performance of baseline methods (e.g., AffinityNet, SEAM and MCTformer). Specifically, our OCR can not only improve AffintityNet, SEAM and MCTformer by 3.2%, 3.3% and 0.8% on Pascal VOC *val* split but also improve AffinityNet, SEAM and MCTformer by 1.5%, 2.7% and 0.4% on Pascal VOC *test* split. Besides, OCR with MCTformer sets a new State-Of-The-Art.

MS COCO. Tab. 2 reports the results of our OCR compared to previous methods on MS COCO *val* split. MS COCO is a more challenging dataset than Pascal VOC because it has more semantic categories and the OC phenomenon is more likely to occur, where our OCR still improves some previous representative methods (AffinityNet, SEAM and MCTformer) by 1.0%, 1.3% and 0.5%. Same as Pascal VOC, we also build a new State-Of-The-Art by adopting OCR into MCTformer.

4.3. Quantitative Analysis

Extra Computation Cost. Our OCR is used to boost the training progress of segmentation network and can be removed when evaluating segmentation network. In order to check if it introduces heavy computation overhead when training segmentation network, we evaluate OCR based on SEAM with different image crop size. In Tab. 3,

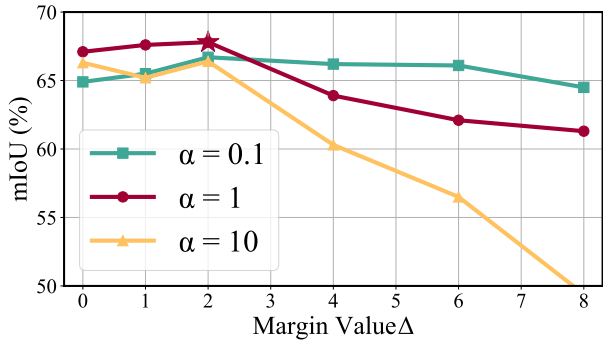


Figure 5. The performance (mIoU (%)) influence of the margin value Δ and the loss coefficient α for penalizing loss $L_{penalize}$. This study adopts SEAM as baseline and are evaluated on Pascal VOC *val* split. The best parameter selection of α and Δ are 1 and 2.

Method	Size	mIoU (%)	Train Speed
SEAM (**w/ OCR**)	256^2	63.6 (↑ **3.2**)	14.38 (↑ **0.56**)
	321^2	64.5 (↑ **3.3**)	23.32 (↑ **0.83**)
	448^2	64.7 (↑ **3.7**)	42.29 (↑ **1.18**)

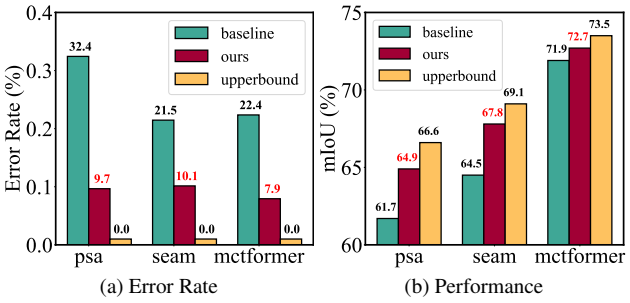
Table 3. Inference performance (mIoU (%)) and training speed (min./epoch) of OCR based on SEAM. The size is the crop size when preprocessing images. All of the results are based on SEAM and are evaluated on Pascal VOC *val* split.

OCR improve performance (mIoU (%)) by 3.2~3.7% with only 0.56~1.18 min./epoch additional training time, which shows the effectiveness and efficiency of our OCR.

Different Value Margin. The margin value Δ can control prediction differences between IC and OC groups. In other words, it can locally control the intensity of rectification. In order to check the influence of rectification intensity, we conduct extensive experiments on margin Δ. As shown in Fig. 5, the overall best choice of Δ is 2. When placing a larger value for the margin Δ, excessive rectification degrades the performance of segmentation network. When placing a smaller value for margin Δ, insufficient rectification provides suboptimal performance.

Loss Coefficient. We use a loss coefficient α to globally control the rectification intensity of rectification loss L_{rec}. In Fig. 5, we conduct empirical experiments to check the influence of the loss coefficient. When we adjust the loss coefficient α to evaluate segmentation network, we find that the overall optimal choice of α is 1.0.

Different Group Split Strategy. As mentioned in Sec 3.3, IC classes group is just the candidate set of correct classes and is not equivalent to the ground truth of OC pixels. So we have to filter out some useless classes in IC classes group. We compare this adaptive strategy *ada*(·) to two brute force strategies: 1). all (·) which means selecting all tag labels into IC classes group; 2). max (·) which means only selecting anchor class (i.e., the class with max prediction score) into IC classes group. In Tab. 4a, we can find that the OCR with all (·) strategy decreases the baseline performance by 1.1% mIoU which shows that useless classes in IC classes group cause negative effect. In contrast, we find that ada (·) and max (·) strategies can improve the performance of baseline by 2.7% and 3.3% mIoU. The comparison above tells us that it is necessary to remove useless classes in IC classes group. Besides, the ada (·) group split strategy is more effective than the max (·) strategy, which means the anchor class is not always the correct class of OC pixels and our adaptive strategy can keep the correct class of OC pixels better than the max (·) strategy.



(a) Error Rate (b) Performance

Figure 6. Effectiveness analysis of our proposed method. (a). The rate of images with Out-Of-Candidate error on Pascal VOC 2012 *val* split. (b). The mIoU metric on Pascal VOC 2012 *val* split. We choose SEAM as our baseline in this experiment.

G_{ic}	G_{oc}	mIoU
None	None	64.5
all ($S \cup \{bg\}$)	$\overline{S}$	63.4 (↓ 1.1)
max ($S \cup \{bg\}$)	$\overline{S}$	67.2 (↑ **2.7**)
ada ($S \cup \{bg\}$)	$\overline{S}$	67.8 (↑ **3.3**)

(a) IC/OC group split.

Rec. Pixels	mIoU
None	64.5
IC	64.7 (↑ **0.2**)
OC	67.8 (↑ **3.3**)
ALL	66.9 (↑ **2.4**)

(b) Rectified pixels select.

Table 4. Ablation study of (a). IC/OC group split strategy; (b). Rectified pixels select strategy. "Rec. Pixels" means those pixels which will be rectified by our OCR. max (·) denotes select the anchor class which has max prediction score in IC classes group. ada (·) denotes selecting class by considering prior and posterior correlation information. These experiments are based on SEAM.

Different Pixel Selection Strategy. Pixel selection strategy controls the operation domain of OCR. In Tab. 4b, we find that if we only rectify IC pixels, the OCR only brings a negligible 0.2% mIoU improvement for baseline. If we rectify all of the pixels in an image, the OCR improves the performance of the baseline by 2.4% which is lower than only rectifying the OC pixels. So using OCR to extra rectify IC pixels is unnecessary and may cause a suboptimal effect for rectifying OC pixels. Only using OCR to rectify OC pixels achieve optimal improvement for the baseline.

Error Rate Analysis. In order to verify if the OCR can suppress the occurrence of OC pixels, we count the proportion of OC error predictions of baseline and our method on Pascal VOC *val* split. As shown in Fig. 6a, our proposed OCR significantly reduces the proportion of OC error for the baseline methods (i.e., AffinityNet, SEAM and MCTformer) from 32.4%, 21.5% and 22.4% to 9.7%, 10.1% and

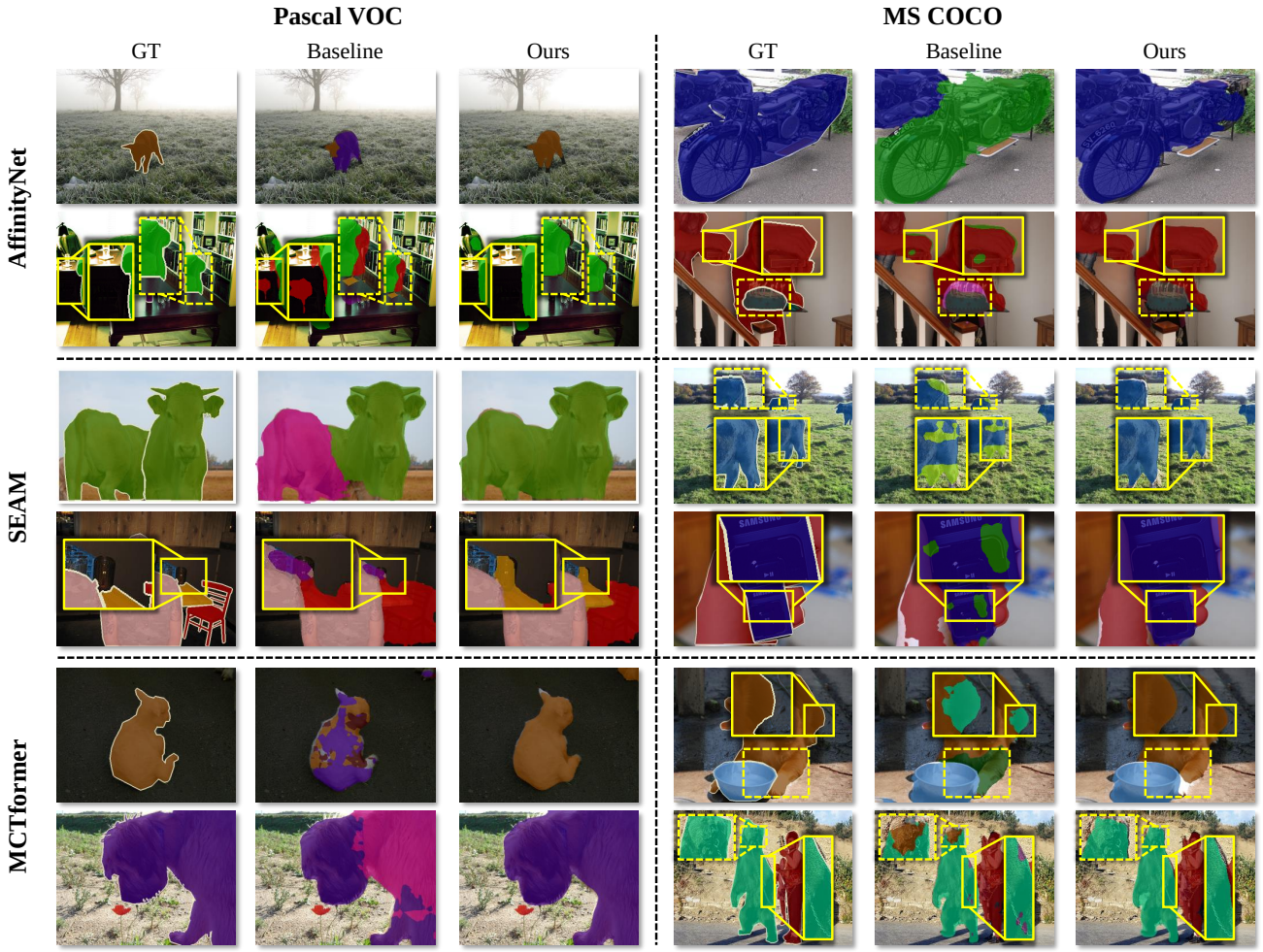


Figure 7. **Qualitative segmentation results comparison** between baseline methods and our methods. Prediction results from baseline methods usually contain Out-Of-Candidate pixels on both Pascal VOC and MS COCO dataset (2nd and 5th columns). Our methods (OCR) can fix these pixels to correct class (3rd and 6th columns). Yellow boxes are used to highlight the effect of our method.

7.9%. With the reduction of OC error, the performance of baseline models is synchronously improved from 61.7%, 64.5% and 71.9% mIoU to 64.9%, 67.8% and 72.9% mIoU.

4.4. Qualitative Analysis

Fig. 7 depicts qualitative comparisons of our method with baselines against vanilla baselines (i.e., AffinityNet, SEAM and MCTformer) over representative examples on both VOC and COCO datasets. It clearly shows that those OC pixels occurring on baseline predictions are rectified after adopting our OCR whether it is a simple scenario that only contains a single object (e.g., the first row and the last three columns) or a complex scenario that contains multiple objects (e.g., the second row and the first three columns).

5. Conclusion

In this paper, we observe that previous WSSS methods usually output pixels whose semantic categories are in contradiction with image-level candidate tags. Then we propose some new concepts (OC/IC) to describe this special type of segmentation error. To tackle these errors, we propose Out-of-Candidate Rectification (OCR). The OCR first defines IC and OC classes group. Then we formulate the relationship between IC and OC groups by a group ranking problem. Finally, we derive a differentiable rectification loss to solve the group ranking problem for suppressing the OC phenomenon. We incorporate OCR with several representative baseline methods for evaluation. The experiments show that our OCR can consistently improve baseline methods on both Pascal VOC and MS COCO datasets, which can demonstrate the effectiveness and generality of our OCR.

Acknowledgements. This work was supported in part by the National Key R&D Program of China (No. 2022ZD0118201), the Natural Science Foundation of China (No. 61972217, 32071459, 62176249, 62006133, 62271465), the Natural Science Foundation of Guangdong Province in China (No. 2019B1515120049).

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019. 1, 3

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 3, 6

[3] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3, 6

[4] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pages 233–242. PMLR, 2017. 2

[5] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European Conference on Computer Vision*, 2016. 1

[6] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 5

[7] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2, 5

[8] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *European Conference on Computer Vision*, 2020. 6

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *International Conference on Learning Representations*, 2015. 3, 6

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 6

[11] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4288–4298, 2022. 2

[12] Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, and Qianru Sun. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 969–978, 2022. 2

[13] Junsuk Choe, Seungho Lee, and Hyunjung Shim. Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4256–4271, 2020. 5

[14] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015. 1

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009. 6

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[17] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4329, 2022. 2

[18] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 5, 6

[19] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 6

[20] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Employing multi-estimations for weakly-supervised semantic segmentation. In *European Conference on Computer Vision*, pages 332–348. Springer, 2020. 3

[21] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 2, 3

[22] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[23] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011. 5

[24] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2011. 5

[25] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5

[26] Seunghoon Hong, Donghun Yeo, Suha Kwak, Honglak Lee, and Bohyung Han. Weakly supervised semantic segmentation using web-crawled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 1

[27] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems*, 2018. 2

[28] Peng-Tao Jiang, Yuqi Yang, Qibin Hou, and Yunchao Wei. L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16886–16896, 2022. 3

[29] Tsung-Wei Ke, Jyh-Jing Hwang, and Stella X Yu. Universal weakly supervised segmentation by pixel-to-segment contrastive learning. In *International Conference on Learning Representations*, 2021. 2

[30] Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 1

[31] Hyeokjun Kweon, Sung-Hoon Yoon, Hyeonseong Kim, Daehee Park, and Kuk-Jin Yoon. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 6

[32] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised segmentation using stochastic inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[33] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 5, 6

[34] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2643–2652, 2021. 1

[35] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 5

[36] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 2

[37] Xueyi Li, Tianfei Zhou, Jianwu Li, Yi Zhou, and Zhaoxiang Zhang. Group-wise semantic mining for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 2, 6

[38] Yi Li, Yiqun Duan, Zhanghui Kuang, Yimin Chen, Wayne Zhang, and Xiaomeng Li. Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1447–1455, 2022. 3, 6

[39] Yi Li, Zhanghui Kuang, Liyang Liu, Yimin Chen, and Wayne Zhang. Pseudo-mask matters in weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6964–6973, 2021. 2

[40] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 1

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 5, 6

[42] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 1

[43] Wenfeng Luo and Meng Yang. Learning saliency-free model with generic features for weakly-supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 6

[44] Frank Nielsen and Ke Sun. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. *Entropy*, 18(12):442, 2016. 5

[45] Pengchong Qiao, Zhidan Wei, Yu Wang, Chang Liu, Zhennan Wang, Guoli Song, and Jie Chen. Pcr: Pessimistic consistency regularization for semi-supervised segmentation. *arXiv preprint arXiv:2210.08519*, 2022. 5

[46] Jie Qin, Jie Wu, Xuefeng Xiao, Lujun Li, and Xingang Wang. Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2117–2125, 2022. 6

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[48] Simone Rossetti, Damiano Zappia, Marta Sanzari, Marco Schaerf, and Fiora Pirri. Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation. In *European Conference on Computer Vision*, pages 446–463. Springer, 2022. 3

[49] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16846–16855, 2022. 3, 6

[50] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 5

[51] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2, 6

[52] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017. 2

[53] Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context decoupling augmentation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 5, 6

[54] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *European Conference on Computer Vision*, 2020. 2, 6

[55] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020. 5

[56] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 2021. 6

[57] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7158–7166, 2017. 1

[58] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017. 5

[59] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 5

[60] Xiang Wang, Sifei Liu, Huimin Ma, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation by iterative affinity learning. *International Journal of Computer Vision*, 128(6):1736–1749, 2020. 3

[61] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 3, 5, 6

[62] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 2

[63] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 3

[64] Tong Wu, Guangyu Gao, Junshi Huang, Xiaolin Wei, Xiaoming Wei, and Chi Harold Liu. Adaptive spatial-bce loss for weakly supervised semantic segmentation. In *European Conference on Computer Vision*, pages 199–216. Springer, 2022. 3

[65] Tong Wu, Junshi Huang, Guangyu Gao, Xiaoming Wei, Xiaolin Wei, Xuan Luo, and Chi Harold Liu. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 6

[66] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019. 3, 6

[67] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 6

[68] Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. *arXiv preprint arXiv:2203.13505*, 2022. 2

[69] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, Ferdous Sohel, and Dan Xu. Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3, 5

[70] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4310–4319, 2022. 1, 2, 3, 6

[71] Lian Xu, Hao Xue, Mohammed Bennamoun, Farid Boussaid, and Ferdous Sohel. Atrous convolutional feature network for weakly supervised semantic segmentation. *Neurocomputing*, 421:115–126, 2021. 3

[72] Yazhou Yao, Tao Chen, Guo-Sen Xie, Chuanyi Zhang, Fumin Shen, Qi Wu, Zhenmin Tang, and Jian Zhang. Non-salient region object mining for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 6

[73] Sung-Hoon Yoon, Hyeokjun Kweon, Jegyeong Cho, Shinjeong Kim, and Kuk-Jin Yoon. Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation. In *European Conference on Computer Vision*, pages 326–344. Springer, 2022. 2

[74] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 3

[75] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. In *Advances in Neural Information Processing Systems*, 2020. 6

[76] Fei Zhang, Chaochen Gu, Chenyue Zhang, and Yuchao Dai. Complementary patch for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2, 3, 5, 6

[77] Tianyi Zhang, Guosheng Lin, Weide Liu, Jianfei Cai, and Alex Kot. Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In *European Conference on Computer Vision*, 2020. 1, 3

[78] Xiangrong Zhang, Zelin Peng, Peng Zhu, Tianyang Zhang, Chen Li, Huiyu Zhou, and Licheng Jiao. Adaptive affinity loss and erroneous pseudo-label refinement for weakly supervised semantic segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5463–5472, 2021. 1, 3

[79] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 2, 3

[80] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4299–4309, 2022. 2