Weak-shot Semantic Segmentation by Transferring Semantic Affinity and Boundary

Siyuan Zhou[1]
ssluvble@sjtu.edu.cn
Li Niu[1*]
ustcnewly@sjtu.edu.cn
Jianlou Si[2]
sijianlou@sensetime.com
Chen Qian[2]
qianchen@sensetime.com
Liqing Zhang[1]
zhang-lq@cs.sjtu.edu.cn

[1] MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University Shanghai, China

[2] SenseTime Research, SenseTime Beijing, China

Abstract

Weakly-supervised semantic segmentation (WSSS) with image-level labels has been widely studied to relieve the annotation burden of the traditional segmentation task. In this paper, we show that existing fully-annotated base categories can help segment objects of novel categories with only image-level labels, even if base categories and novel categories have no overlap. We refer to this task as weak-shot semantic segmentation, which could also be treated as WSSS with auxiliary fully-annotated categories. Recent advanced WSSS methods usually obtain class activation maps (CAMs) and refine them by affinity propagation. Based on the observation that semantic affinity and boundary are class-agnostic, we propose a method under the WSSS framework to transfer semantic affinity and boundary from base to novel categories. As a result, we find that pixel-level annotation of base categories can facilitate affinity learning and propagation, leading to higher-quality CAMs of novel categories. Extensive experiments on PASCAL VOC 2012 dataset prove that our method significantly outperforms WSSS baselines on novel categories.

1 Introduction

Semantic segmentation [4, 10, 23, 35, 37, 39, 42, 55] is fundamental in computer vision and has been greatly advanced through the rapid development of deep learning techniques. Traditional fully-supervised segmentation heavily relies on expensive pixel-level (full) annotations. To solve this issue, segmentation paradigms requiring fewer or weaker annotations have gradually attracted research attention, like weakly-/semi-supervised segmentation [22, 41, 47, 53] and one-/few-shot segmentation [15, 24, 43, 44, 50]. However, these paradigms have limitations in practical applications. Semi-supervised/few-shot segmentation can not handle new categories with no pixel-level annotations. Weakly-supervised semantic segmentation (WSSS)

*Corresponding author.

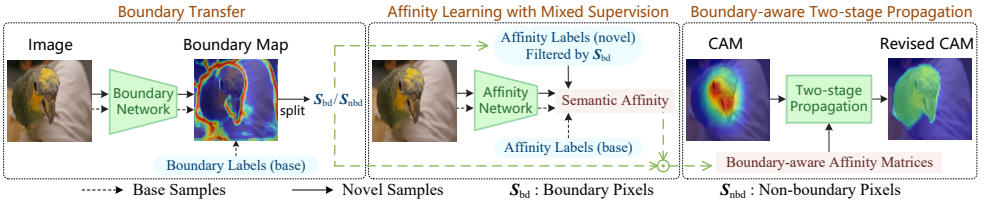

Figure 1: Overview of our RETAB: (i) train a boundary network to transfer semantic boundary from base to novel categories, (ii) train an affinity network to learn semantic affinity in a mixed-supervised manner, and (iii) perform boundary-aware two-stage propagation to revise CAMs based on the learned semantic boundary and affinity. Details can be found in Section 3.2.

leverages more accessible weak annotations, but the performance gap between WSSS and fully-supervised segmentation is still non-negligible.

In this work, we propose a new learning paradigm called weak-shot semantic segmentation. Assume that we have a fully-annotated segmentation dataset containing training samples of only base categories (marked as base samples). We are also provided with extra weakly-annotated training samples containing objects of base or novel categories (marked as novel samples). Therefore, our problem can be considered as weakly-supervised segmentation with an auxiliary fully-annotated dataset, but the set of well-labeled categories is limited. At test time, our objective is to segment images where both base and novel categories may exist. The critical problem in our task is transferring class-agnostic knowledge from base categories to novel ones to enhance the segmentation performance of novel categories. Considering different types of weak annotation, we focus on image-level labels in this work since it is a popular research direction in WSSS [3, 17, 17, 52, 53, 56]. That is, base samples have pixel-level labels while novel samples only have image-level labels in our problem.

We draw inspiration from classical WSSS methods to start our work. Until recently, some advanced WSSS methods are based on Class Activation Map (CAM) [62] which can effectively localize discriminative parts of objects by training a classification network with image-level labels. The typical WSSS framework usually obtains CAM as the initial response, expands the response region to acquire pseudo labels, and uses pseudo labels to train a segmentation network. Among WSSS methods under this framework, PSA [1] is a representative one with two main steps: learn semantic affinities between pair-wise pixels within local neighborhoods on the feature map (*i.e.*, affinity learning), and generate a transition matrix to perform random walk [38] on CAMs (*i.e.*, affinity-based propagation).

In this paper, we design our method under the typical WSSS framework and focus on better expanding the initial response. With the assumption that semantic affinities and boundaries are class-agnostic, we attempt to transfer these two types of information from base to novel categories. Our method is called **R**esponse **E**xpansion by **T**ransferring semantic **A**ffinity and **B**oundary (RETAB), which contains an affinity learning step and an affinity-based propagation step. In the affinity learning step, our goal is to design an affinity network that learns semantic affinities from ground-truth labels of base samples and CAMs of novel samples. Inspired by prior works on instance segmentation [18, 28], boundary knowledge could function as an effective tool to assist with affinity learning. However, CAMs are noisy on semantic boundaries, leading to imprecise semantic affinities for novel categories. To solve this problem, we first consider boundary (*resp.*, non-boundary) pixels as unconfident (*resp.*, confident) pixels [12]. Then we propose to train a boundary network supervised by

base samples. The predicted boundaries of novel samples are used to split apart CAMs into boundary regions and non-boundary regions. To filter out noisy supervisions in CAMs of novel samples, we only consider non-boundary regions when we train the affinity network.

In the affinity-based propagation step, we propose a two-stage propagation strategy to revise CAMs. In the first stage, random walk is restricted within the non-boundary regions. Pixels with dominant category labels in the confident regions will be propagated to fit object shapes. In the second stage, unconfident pixels in the boundary regions are propagated under the guidance of confident pixels in the non-boundary regions. Hopefully, confident pixels can regulate the random walk of unconfident ones and facilitate propagation on object boundaries. After propagation, pseudo segmentation labels are obtained based on the revised CAMs (revised responses). Finally, we train a segmentation network under the mixed supervisions from ground-truth labels of base samples and pseudo labels of novel samples, which is the only network used in the inference stage. In summary, our main contributions are:

- We study a novel paradigm called weak-shot semantic segmentation that utilizes full annotations of base categories to benefit segmenting objects of novel categories with only image-level labels.
- We propose a simple yet effective method called RETAB to transfer class-agnostic semantic affinity and semantic boundary from base to novel categories under the typical WSSS framework, together with *a novel boundary-aware two-stage propagation strategy*. Our method can be integrated into any WSSS method under this framework.
- The effectiveness of RETAB is verified on PASCAL VOC 2012 dataset [16]. RETAB significantly outperforms WSSS baselines and naive weak-shot segmentation baselines on novel categories.

2 Related Works

Weakly-supervised Semantic Segmentation: Weakly-supervised semantic segmentation (WSSS) [19, 31, 46, 48] has attracted considerable interest because weak annotations are conveniently available. Most advanced WSSS methods with image-level labels [1, 8, 33, 51, 57, 58] are based on the class activation map (CAM) [62] obtained from a classification network. "Seed, expand, and constrain", three principles proposed by SEC [29], are followed by many WSSS works. Some of them work on improving the seed or initial response [8, 32, 51]. The other works followed the *coarse-to-fine* strategy to expand or propagate responses [1, 25, 50, 59]. Compared with WSSS, our proposed weak-shot segmentation takes advantage of full annotations in existing datasets, and can be considered as WSSS with an auxiliary fully-annotated dataset containing limited well-labeled categories. In this work, our RETAB is realized under the typical WSSS framework and focuses on the response expansion problem.

Few-/Zero-shot Semantic Segmentation: Few-shot semantic segmentation [15, 24, 44, 60] and zero-shot semantic segmentation [7, 20] have been studied by plenty of works in recent years. They both divide the whole category set into base categories and novel categories. Few-shot semantic segmentation assumes that only a few training images are available for each novel category, but pixel-level annotations are still required for novel categories. Zero-shot semantic segmentation relies on category-level semantic representations that are often weak and ambiguous, so the performance is far from satisfactory. Likewise, our proposed weak-shot semantic segmentation has a split of base categories and novel categories. Differently, we

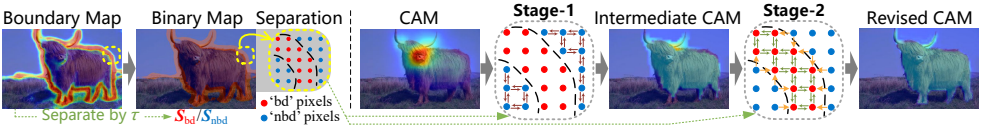

Figure 2: Illustration of boundary-aware two-stage propagation. The predicted boundary map is separated into boundary ('bd') pixels $\mathcal{S}_{\text{bd}}$ and non-boundary ('nbd') pixels $\mathcal{S}_{\text{nbd}}$. In the first stage only 'nbd' pixels are propagated, while in the second stage 'nbd' pixels guide the propagation of 'bd' pixels. See Section 3.2 for more details.

provide novel categories with weak annotations that are easily accessible and useful for learning segmentation models.

Semi-supervised Semantic Segmentation: Typically, semi-supervised semantic segmentation [22, 26, 27, 40] addresses the issue of utilizing a set of well-labeled images to enhance the segmentation quality for another set of unlabeled images [26, 40] or weakly-labeled images [22, 27]. In this work, our proposed weak-shot semantic segmentation follows a similar idea to separate training samples into two sets with different annotation levels. Differently, our task further splits categories into base ones and novel ones, which involves cross-category knowledge transfer.

Weak-shot Learning: Actually, the weak-shot learning paradigm, *i.e.*, full annotations for base categories and weak annotations for novel categories, has been studied in image classification [9], object detection [13, 34, 36, 61], and instance segmentation [5, 18, 23, 30, 63]. Weak-shot classification [9] supposes that base categories have clean image labels and novel categories only have noisy ones. Weak-shot detection [36], also called mixed-supervised [34] or cross-supervised [13] detection, requires that base categories have box-level annotations while novel ones only have image-level labels. Weak-shot instance segmentation, usually called partially-supervised instance segmentation, utilizes mask annotations of base categories and only bounding boxes of novel ones. The abovementioned methods generally transfer class-agnostic target (*e.g.*, similarity, objectness) or learn the mapping from weak annotation to full annotation. To the best of our knowledge, weak-shot semantic segmentation has not been explored. Compared with weak-shot classification/detection, we focus on the more challenging segmentation task. Compared with weak-shot instance segmentation, we only utilize higher-available image-level labels instead of stronger boxes.

3 Methodology

In weak-shot semantic segmentation, we have base categories $\mathcal{C}^b$ and novel categories $\mathcal{C}^n$, satisfying $\mathcal{C} = \mathcal{C}^b \cup \mathcal{C}^n$ and $\mathcal{C}^b \cap \mathcal{C}^n = \emptyset$. We assume that background bg belongs to base categories, *i.e.*, $bg \in \mathcal{C}^b$. Training data contain $\mathcal{N}^b$ base samples and $\mathcal{N}^n$ novel samples with no intersection. Base samples only contain base categories, whereas novel samples may contain base or novel categories. We provide pixel-level labels for base samples and only image-level labels for novel samples. We use c_i^* to denote the ground-truth segmentation label of the i-th pixel on the feature map for any base sample. Next, we will first take a glance at PSA [1], which represents the typical WSSS framework. Then, we describe our RETAB to transfer class-agnostic semantic affinity and boundary under this framework, which is outlined in Figure 1. Finally, we discuss the training of the segmentation network.

3.1 Review of CAM and PSA

As mentioned in Section 1, the typical WSSS framework [1, 8, 51] adopts CAM as the initial response, which is obtained by training a classification network with image-level labels. CAM M_l of a typical category $l \in \mathcal{C}$ highlights the discriminative regions of this category. A group of WSSS methods design models to augment CAM [8, 51], which could also function as the initial response. Another group of WSSS methods attempts to design effective algorithms to expand responses [1, 25, 50], in which PSA [1] is a representative one. Specifically, PSA designs an AffinityNet to learn feature map f^{aff}, based on which pair-wise semantic affinity $\hat{a}_{ij}$ between pixel i and j is calculated by $\hat{a}_{ij} = \exp\left\{-\left\|f^{\text{aff}}(x_i, y_i) - f^{\text{aff}}(x_j, y_j)\right\|_1\right\}$. In the training stage, PSA only considers pixel pairs in neighbor set $\mathcal{P}$:

$$\mathcal{P} = \left\{(i,j) \mid d\big((x_i,y_i),(x_j,y_j)\big) < \gamma,\ i \neq j\right\}. \tag{1}$$

where γ is a search radius and $d(\cdot,\cdot)$ represents Euclidean distance. The network is supervised by affinity labels a_{ij} obtained from CAMs (refer to [1] for details), where $a_{ij} = 1$ for intra-category pairs and $a_{ij} = 0$ for inter-category pairs. The trained AffinityNet enforces the predicted semantic affinity $\hat{a}_{ij}$ to be close to 1 if pixel i and j belong to the same category, and close to 0 otherwise. Predicted $\hat{a}_{ij}$ form the affinity matrix $\hat{A}$, with which random walk is performed on each M_l to obtain the revised response of category l. Next, we describe our RETAB that adapts response expansion for weak-shot semantic segmentation.

3.2 Pipeline of RETAB

Boundary Transfer. We first introduce transferring semantic boundaries from base categories to novel ones because this will facilitate our two major steps to expand responses. As shown in Figure 1, we train a boundary network to predict a boundary map from the input image. We use $\hat{p}_i$ to represent the predicted boundary probability (after Sigmoid) for pixel i on the feature map. During training, the boundary network is only supervised with base samples. Boundary labels b_i^* for base samples are derived from segmentation labels, and we denote $b_i^* = 1/0$ for boundary/non-boundary pixels. In training, the model tends to suppress responses in boundary regions, which is caused by the unbalanced boundary, foreground, and background pixels in the training samples. To address it, we split pixels into three subsets: boundary pixels $\mathcal{S}_{\text{bd}} = \{i \mid b_i^* = 1\}$, non-boundary foreground pixels $\mathcal{S}_{\text{fg}} = \{i \mid b_i^* = 0, c_i^* \in \mathcal{C}^b \backslash \{bg\}\}$, and non-boundary background pixels $\mathcal{S}_{\text{bg}} = \{i \mid b_i^* = 0, c_i^* = bg\}$. Cross-entropy classification losses are applied to three sets and merged to form the total loss of the boundary network:

$$\mathcal{L}_B = -\sum_{i\in\mathcal{S}_{\text{bd}}} \frac{\log(\hat{p}_i)}{|\mathcal{S}_{\text{bd}}|} - \frac{1}{2}\sum_{i\in\mathcal{S}_{\text{fg}}} \frac{\log(1-\hat{p}_i)}{|\mathcal{S}_{\text{fg}}|} - \frac{1}{2}\sum_{i\in\mathcal{S}_{\text{bg}}} \frac{\log(1-\hat{p}_i)}{|\mathcal{S}_{\text{bg}}|}.$$

We expect the class-agnostic boundary knowledge embedded in the trained boundary model to be transferred to novel categories. Therefore, we perform boundary prediction on novel samples. To facilitate affinity learning and affinity-based propagation, we pre-set a threshold τ to divide the boundary prediction for each novel sample into two parts: a boundary region $\hat{\mathcal{S}}_{\text{bd}} = \{i \mid \hat{p}_i \geqslant \tau\}$ and a non-boundary region $\hat{\mathcal{S}}_{\text{nbd}} = \{i \mid \hat{p}_i < \tau\}$, as shown in Figure 2.

Affinity Learning with Mixed Supervision. In PSA, AffinityNet is supervised by coarse affinity labels a_{ij} obtained from CAMs, which are imprecise because CAMs only highlight discriminative parts. We similarly implement an affinity network to predict pair-wise semantic

affinities $\hat{a}_{ij}$ within local neighborhoods. Differently, we adopt more robust affinity labels a^*_{ij} by utilizing fully-annotated base samples and weakly-annotated novel samples to train the network in a mixed-supervised manner (see Figure 1). The improvement comes from two aspects. Firstly, affinity labels of base samples are purified through pixel-level annotations. This helps the model learn better semantic affinities and facilitates novel categories by assuming that semantic affinities are class-agnostic. For base samples, affinity labels a^*_{ij} in $\mathcal{P}$ are obtained from the segmentation label, *i.e.*, $a^*_{ij} = 1$ when $c^*_i = c^*_j$, and $a^*_{ij} = 0$ otherwise. Secondly, pixel-level labels for novel samples are not given, so we can only obtain coarse affinity labels from CAMs. Our idea is to leverage predicted boundaries to purify the supervision. Since inaccurate pixels usually occur on CAM boundaries [12], we only use affinity labels in non-boundary regions. Specifically, we narrow the set $\mathcal{P}$ in (1) to a cleaner set $\mathcal{P}_{\text{nbd}}$ by filtering out pixel pairs in boundary regions: $\mathcal{P}_{\text{nbd}} = \big\{(i,j) \mid d\big((x_i,y_i),(x_j,y_j)\big) < \gamma,\ i \neq j \text{ and } i,j \notin \hat{\mathcal{S}}_{\text{bd}}\big\}$. For novel samples, our affinity labels a^*_{ij} in $\mathcal{P}_{\text{nbd}}$ are obtained in the same way as a_{ij} (see Section 3.1). By filtering out noisy affinity labels, the model can learn more accurate affinities from novel samples. Details of the cross-entropy loss to optimize the affinity network can be found in PSA. The loss form remains unchanged and is suitable for all samples by replacing a_{ij} with a^*_{ij} for base samples and replacing $\mathcal{P}$ with $\mathcal{P}_{\text{nbd}}$ for novel ones.

Boundary-aware Two-stage Propagation. After training the affinity network, we predict the semantic affinity $\hat{a}_{ij}$ between pixel i and j within the entire neighbor set $\mathcal{P}$. Similar to [1], we perform random walk on CAMs based on $\hat{a}_{ij}$ which represents the transition probability that pixel i should be propagated to pixel j. Distinctive from propagating on the whole CAM for a single stage in the classical random walk, we adopt a boundary-aware two-stage propagation strategy that concurrently utilizes semantic affinities and boundaries (see Figure 2). In the first stage, the random walk is restricted within non-boundary regions to prevent the disturbance from unconfident pixels on the boundaries, in which case the boundary regions are analogous to isolation belts. Under this design, the pixels with dominant category labels will be propagated to other pixels to fit object shapes and complete the masks. Restricting the propagation area can be implemented by setting certain entries in the affinity matrix to zero, leading to a sparse affinity matrix $\hat{A}^{(1)}$ as follows:

$$\hat{A}^{(1)}_{ij} = \begin{cases} \hat{a}_{ij}, & \forall\, i \neq j \text{ s.t. } i,j \in \hat{\mathcal{S}}_{\text{nbd}},\ (i,j) \in \mathcal{P}, \\ 1\ , & \forall\, i = j, \\ 0\ , & \text{otherwise}. \end{cases} \tag{2}$$

In the second stage, we hope that confident non-boundary pixels propagated in the first stage can regulate the random walk process of boundary pixels, whereas unconfident boundary pixels are not allowed to affect confident ones, so the sparse affinity matrix $\hat{A}^{(2)}$ should be

$$\hat{A}^{(2)}_{ij} = \begin{cases} \hat{a}_{ij}, & \forall\, i \neq j \text{ s.t. } j \in \hat{\mathcal{S}}_{\text{bd}},\ (i,j) \in \mathcal{P}, \\ 1\ , & \forall\, i = j, \\ 0\ , & \text{otherwise}. \end{cases} \tag{3}$$

We can observe that $\hat{A}^{(1)}$ is bidirectional, but $\hat{A}^{(2)}$ contains unidirectional components, indicating the critical difference between two-stage propagation and the original random walk with only one bidirectional $\hat{A}$. Our two-stage propagation is applied on CAM M_l by firstly using $\hat{A}^{(1)}$ and secondly using $\hat{A}^{(2)}$ as the affinity matrix in the random walk process. We

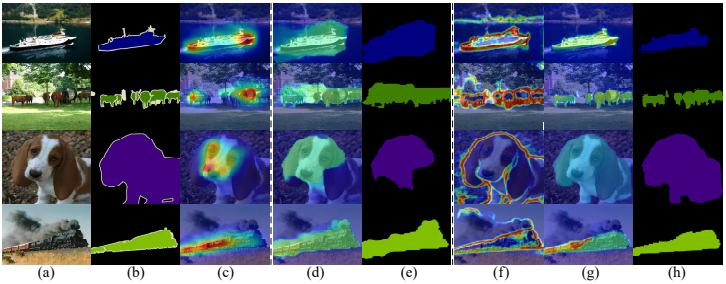

Figure 3: Visualized pseudo labels on VOC12 *train* set. Examples from top to bottom belong to novel samples in fold 0,1,2,3, respectively. (a) image. (b) GT. (c) CAM. (d)(e) CAM+RW and pesudo labels. (f) boundary prediction. (g)(h) CAM+RETAB and pesudo labels.

strictly follow [1] to perform random walk with affinity matrix without tuning parameters. In detail, we first generate transition probability matrices corresponding to $\hat{A}^{(1)}$ and $\hat{A}^{(2)}$. Then, the random walk in each stage is accomplished by iteratively multiplying the corresponding transition matrix to M_l until a predefined number of iterations is reached. We refer to the CAMs after boundary-aware two-stage propagation as revised CAMs (or revised responses).

3.3 Mixed-supervised Segmentation

Since revised responses have smaller resolutions than input images, we first up-sample them to the original sizes using bilinear interpolation. Then, we apply argmax on the category channel of the concatenated revised responses to obtain pseudo segmentation labels of novel samples. Finally, we can train any segmentation network in a fully-supervised manner using mixed supervisions from ground-truth labels of base samples and pseudo labels of novel samples. During inference, the segmentation network takes in a test image to predict categories within $\mathcal{C}^b \cup \mathcal{C}^n$ for each pixel because test images may contain either base or novel categories.

4 Experiments

4.1 Experimental Setting

Datasets and Evaluation Metrics. Following most WSSS works, we conduct experiments on PASCAL VOC 2012 dataset [16] with 21 classes, including 20 foreground object classes and a background class. The official dataset includes 1464 training images, 1449 validation images, and 1456 test images. Following common practice in semantic segmentation, we adopt an augmented training set with 10582 images from SBD [21]. Following the category split rule in PASCAL-5^i [44], which is commonly used in few-shot segmentation [15, 24, 49, 50], we evenly divide the 20 foreground categories into four folds. Categories in each fold are regarded as 5 novel categories, and the remaining categories (including background) are regarded as 16 base categories. We further divide 10582 training samples into base samples and novel samples for each fold. Images containing only base categories are included as base samples, whereas those containing at least one novel category are deemed novel samples. More separation details for categories and samples are left to supplementary. For each fold, we retain full annotations of base samples and only image-level labels for novel samples. During inference,

we use the official validation set and test set to verify the segmentation performance. We adopt Intersection-over-Union (IoU) as the evaluation metric. Due to the specific characteristics of our task, we calculate mean IoU on all categories, base categories, and novel categories, which are referred to as all-mIoU, base-mIoU, and novel-mIoU, respectively.

Implementation Details. The boundary network is based on ResNet38 [54]. We concatenate feature maps in shallow layers and deep layers from three stages with 1×1 convolutions. The concatenated feature map is followed by another 1×1 convolution with the output dimension being one and a Sigmoid layer. The network structure of our affinity network is the same as AffinityNet in PSA [1], which is also based on ResNet38. For both boundary network and affinity network, we augment training images with color jittering, random cropping (448×448), and horizontal flip. Both networks are trained with batch size 8 for eight epochs. SGD [6] optimizer is adopted with weight decay of $5e^{-4}$. The learning rate is initialized as 0.001 (*resp.*, 0.01) for the boundary network (*resp.*, affinity network) and decreases following the polynomial policy $lr_{\text{iter}} = lr_{\text{initial}}(1 - \frac{\text{iter}}{\text{max\_iter}})^{\alpha}$ with $\alpha = 0.9$. We strictly follow WSSS baselines [1, 51, 58] to adopt DeepLab [10] with ResNet38 [54] backbone pretrained on ImageNet [14] as the segmentation network. The training schemes (data augmentation, batch size, training epoch, optimizer, learning rate, and other parameters) remain unchanged for segmentation. All experiments are conducted on 2 NVIDIA RTX 2080Ti GPUs with PyTorch.

Baselines. Our RETAB can be incorporated into any WSSS baseline under the typical framework with three steps: 1) obtain CAM as the initial response, 2) propagate response to acquire pseudo labels, and 3) train a segmentation network. We incorporate our method into PSA [1], SEAM [51], and CPN [58]. They all use random walk with AffinityNet (*abbr.* RW [1, 2]) as the second step, and the same segmentation network in the third step. The difference lie in the first step. PSA directly uses "CAM" proposed in [62] as the initial response, while SEAM and CPN design augmented responses based on their proposed architectures. For ease of representation, we use "CAM+RW", "SEAM+RW", and "CPN+RW" to denote the overall pipelines of baselines. For fairness, when integrating our method into each baseline, we use the same initial response and segmentation network as this baseline, resulting in our "CAM+RETAB", "SEAM+RETAB", and "CPN+RETAB".

Hyper-parameters. During boundary prediction, we set the threshold τ as 0.5 via cross-validation (see supplementary for detailed discussion). For the other hyper-parameters (search radius $\gamma = 5$, the parameters to generate affinity labels a^*_{ij} for novel samples and the parameters for random walk in each stage of our two-stage propagation), we use the default values in WSSS baseline PSA [1], SEAM [51], and CPN [58] without further tuning.

4.2 Evaluation on Pseudo Segmentation Labels

We evaluate pseudo segmentation labels by assessing CAMs before/after propagation on VOC12 *train* set. Figure 3 visualizes the pseudo segmentation labels of some novel samples by comparing "CAM+RW" with "CAM+RETAB". As shown, pseudo labels generated by RETAB better adapt to object boundaries than the classical random walk, demonstrating RETAB's transferability of semantic affinity and boundary from base categories to novel ones. More quantitative analyses and visualizations results can be found in the supplementary.

Method	fold 0			fold 1			fold 2			fold 3		
	$\mathcal{C}$	$\mathcal{C}^b$	$\mathcal{C}^n$	$\mathcal{C}$	$\mathcal{C}^b$	$\mathcal{C}^n$	$\mathcal{C}$	$\mathcal{C}^b$	$\mathcal{C}^n$	$\mathcal{C}$	$\mathcal{C}^b$	$\mathcal{C}^n$
SSDD [45]	65.5	67.6	58.8	65.5	64.5	68.7	65.5	63.5	72.1	65.5	68.0	57.7
BES [12]	66.6	68.8	59.6	66.6	64.9	71.9	66.6	64.7	72.6	66.6	69.3	57.8
SvM [59]	66.7	67.5	64.2	66.7	65.8	69.7	66.7	65.6	70.6	66.7	69.6	57.7
CAM+RW [1]	63.7	65.4	58.1	63.7	63.7	63.8	63.7	61.4	71.0	63.7	65.8	56.8
CAM+RW(seggt)	73.8	78.5	58.6	74.8	**76.5**	69.5	73.7	74.4	71.5	73.9	79.2	56.9
CAM+RW(affgt+seggt)	75.2	78.7	64.0	75.3	**76.5**	71.5	74.6	75.2	72.7	74.1	**79.3**	57.5
CAM+RETAB	**76.3**	**78.8**	**68.0**	**76.0**	76.1	**75.9**	**75.4**	**75.4**	**75.6**	**74.8**	79.2	**60.8**
SEAM+RW [51]	65.7	67.8	59.0	65.7	64.7	68.9	65.7	63.7	72.3	65.7	68.2	57.9
SEAM+RW(seggt)	74.0	78.7	59.1	74.5	75.6	71.1	73.5	73.3	74.0	73.7	78.1	59.6
SEAM+RW(affgt+seggt)	74.9	**78.9**	62.1	75.2	76.4	71.4	74.3	74.3	74.3	74.2	78.7	59.8
SEAM+RETAB	**75.5**	**78.9**	**64.6**	**76.0**	**76.6**	**74.0**	**75.1**	**75.0**	**75.6**	**74.8**	**79.0**	**61.5**
CPN+RW [58]	68.5	70.7	61.5	68.5	66.8	73.8	68.5	66.6	74.5	68.5	71.2	59.7
CPN+RW(seggt)	74.7	78.6	62.4	75.5	76.0	74.0	75.5	75.7	74.8	74.4	78.8	60.2
CPN+RW(affgt+seggt)	76.1	79.0	66.8	76.5	**76.8**	75.7	75.6	75.6	75.5	74.9	**79.3**	60.7
CPN+RETAB	**76.6**	**79.1**	**68.8**	**76.7**	76.7	**76.7**	**75.9**	**75.8**	**76.2**	**75.3**	**79.3**	**62.4**
Fully Oracle	77.9	79.1	74.4	77.9	77.7	78.7	77.9	76.4	82.9	77.9	79.6	72.6

Table 1: Comparison of segmentation performance on VOC12 *test* set. Columns marked by $\mathcal{C}/\mathcal{C}^b/\mathcal{C}^n$ represents all-/base-/novel-mIoU. We have five groups of methods. The first group includes some popular WSSS methods. The second / third / fourth group represents the WSSS baseline PSA / SEAM / CPN, together with our created weak-shot segmentation baselines (seggt and affgt+seggt), and our RETAB. The fifth group is a fully-supervised method. Except the first group, all experiments use DeepLab [10] with ResNet38 [54] backbone.

4.3 Evaluation on Segmentation Performance

Naturally, the fully-supervised (*resp.,* weakly-supervised) setting serves as the upper (*resp.,* lower) bound of our weak-shot setting. In this subsection, we compare the final segmentation results of our method with the upper/lower bound. Quantitative results on VOC12 *test* set are summarized in Table 1 (see supplementary for results on the *val* set). "CAM+RW", "SEAM+RW" and "CPN+RW" represent three WSSS baselines we use. We correspondingly create three naive weak-shot segmentation baselines, denoted by "(seggt)". The simple modification is to train the final segmentation step with the mixed supervisions of ground-truth segmentation labels for base samples and pseudo segmentation labels generated by WSSS baselines for novel samples. Also, we create three augmented weak-shot segmentation baselines, denoted by "(affgt+ seggt)". Based on naive baselines, these augmented baselines further utilize ground-truth affinity labels of base samples when training the affinity network. Our methods are represented by "CAM+RETAB", "SEAM+RETAB" and "CPN+RETAB". "Fully Oracle" is a reproduced fully-supervised baseline. As illustrated in Table 1, our RETAB consistently and significantly outperforms WSSS baselines and weak-shot baselines for novel categories on all four folds, verifying the effectiveness of knowledge transfer across categories. Compared with "Fully Oracle", our method can recover 97.7% $\sim$ 100% of its bound for base categories and 83.7% $\sim$ 97.5% of its bound for novel categories, implying that RETAB successfully narrows the gap between weakly-supervised and fully-supervised segmentation.

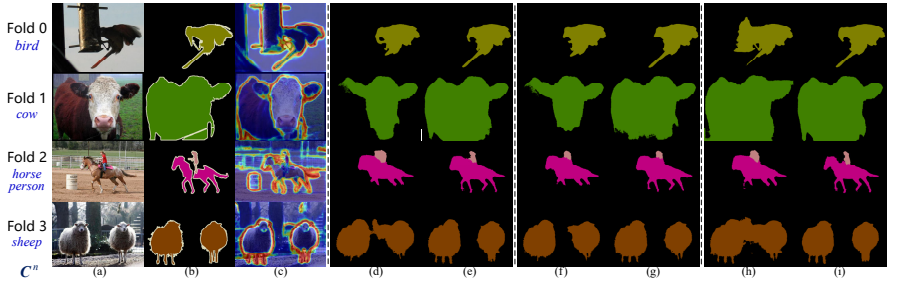

Figure 4: Visualized segmentation results of novel categories on VOC12 *val* set. Examples from top to bottom contain objects of novel categories (marked as blue on the left side) in fold 0,1,2,3, respectively. (a) image. (b) GT label. (c) boundary prediction. (d) CAM+RW. (e) CAM+RETAB. (f) SEAM+RW. (g) SEAM+RETAB. (h) CPN+RW. (i) CPN+RETAB.

Figure 4 shows some visualizations on segmenting objects of novel categories with different methods (see supplementary for more visualizations). Our RETAB works better in recovering object shapes, especially in boundary regions.

4.4 Ablation Studies and Generalization to Other Settings

We carefully analyze the functionality of boundary transfer, semantic affinity, and boundary-aware two-stage propagation, respectively. Different parts in our method are verified to demonstrate the source of improvement for the segmentation results. A significance test is also included. For more practical application, we generalize our weak-shot semantic segmentation task to two settings: 1) generalization to potential novel categories in the background, and 2) generalization to fewer fully-annotated training data. We observe that our RETAB could generalize to discover potential cues of novel categories in the background of base samples with a simple follow-up self-training step. We also observe that RETAB can use a small proportion of base samples to facilitate a large number of novel samples. These detailed results can be found in the supplementary.

5 Conclusion

This paper has proposed a novel paradigm called weak-shot semantic segmentation that utilizes pixel-level annotations of base categories to improve the segmentation performance on novel categories with only image-level labels. Under the typical WSSS framework, a simple yet effective method called RETAB is developed to expand response regions by transferring class-agnostic semantic affinity and boundary. Our work provides a simple yet effective baseline to promote future research on weak-shot semantic segmentation.

Acknowledgements

The work was supported by the Shanghai Municipal Science and Technology Major/Key Project, China (2021SHZDZX0102, 20511100300) and National Natural Science Foundation of China (Grant No. 61902247).

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 2018.

[2] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019.

[3] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, 2020.

[4] Gedas Bertasius, Lorenzo Torresani, Stella X Yu, and Jianbo Shi. Convolutional random walk networks for semantic image segmentation. In *CVPR*, 2017.

[5] David Biertimpel, Sindi Shkodrani, Anil S Baslamisli, and Nóra Baka. Prior to segment: Foreground cues for weakly annotated classes in partially supervised instance segmentation. In *ICCV*, 2021.

[6] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[7] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32:468–479, 2019.

[8] 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 *CVPR*, 2020.

[9] Junjie Chen, Li Niu, Liu Liu, and Liqing Zhang. Weak-shot fine-grained classification via similarity transfer. In *NeurIPS*, 2021.

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[11] 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.

[12] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *ECCV*, 2020.

[13] Zitian Chen, Zhiqiang Shen, Jiahui Yu, and Erik Learned-Miller. Cross-supervised object detection. *arXiv preprint arXiv:2006.15056*, 2020.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[15] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, 2018.

[16] 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.

[17] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *CVPR*, 2020.

[18] Qi Fan, Lei Ke, Wenjie Pei, Chi-Keung Tang, and Yu-Wing Tai. Commonality-parsing network across shape and appearance for partially supervised instance segmentation. In *ECCV*, 2020.

[19] Ruochen Fan, Qibin Hou, Ming-Ming Cheng, Gang Yu, Ralph R Martin, and Shi-Min Hu. Associating inter-image salient instances for weakly supervised semantic segmentation. In *ECCV*, 2018.

[20] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. Context-aware feature generation for zero-shot semantic segmentation. In *ACM Multimedia*, 2020.

[21] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.

[22] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. *Advances in neural information processing systems*, 28, 2015.

[23] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *CVPR*, 2018.

[24] Tao Hu, Pengwan Yang, Chiliang Zhang, Gang Yu, Yadong Mu, and Cees GM Snoek. Attention-based multi-context guiding for few-shot semantic segmentation. In *AAAI*, 2019.

[25] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, 2018.

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

[27] Mostafa S Ibrahim, Arash Vahdat, Mani Ranjbar, and William G Macready. Semi-supervised semantic image segmentation with self-correcting networks. In *CVPR*, 2020.

[28] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *CVPR*, 2021.

[29] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016.

[30] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. In *ICCV*, 2019.

[31] Suha Kwak, Seunghoon Hong, and Bohyung Han. Weakly supervised semantic segmentation using superpixel pooling network. In *AAAI*, 2017.

[32] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *CVPR*, 2019.

[33] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *CVPR*, 2021.

[34] Yan Li, Junge Zhang, Kaiqi Huang, and Jianguo Zhang. Mixed supervised object detection with robust objectness transfer. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):639–653, 2018.

[35] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016.

[36] Yan Liu, Zhijie Zhang, Li Niu, Junjie Chen, and Liqing Zhang. Mixed supervised object detection by transferring mask prior and semantic similarity. In *NeurIPS*, 2021.

[37] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[38] László Lovász et al. Random walks on graphs: A survey. *Combinatorics, Paul erdos is eighty*, 2(1):1–46, 1993.

[39] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015.

[40] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, 2020.

[41] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015.

[42] Guo-Jun Qi. Hierarchically gated deep networks for semantic segmentation. In *CVPR*, 2016.

[43] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alexei A Efros, and Sergey Levine. Few-shot segmentation propagation with guided networks. *arXiv preprint arXiv:1806.07373*, 2018.

[44] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017.

[45] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *ICCV*, 2019.

[46] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017.

[47] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi and weakly supervised semantic segmentation using generative adversarial network. *arXiv preprint arXiv:1703.09695*, 2017.

[48] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *ECCV*, 2020.

[49] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *ICCV*, 2019.

[50] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*, 2018.

[51] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR*, 2020.

[52] 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 *CVPR*, 2017.

[53] 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 *CVPR*, 2018.

[54] 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.

[55] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[56] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *AAAI*, 2020.

[57] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. In *NeurIPS*, 2020.

[58] Fei Zhang, Chaochen Gu, Chenyue Zhang, and Yuchao Dai. Complementary patch for weakly supervised semantic segmentation. In *ICCV*, 2021.

[59] 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 *ECCV*, 2020.

[60] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics*, 50(9):3855–3865, 2020.

[61] Yuanyi Zhong, Jianfeng Wang, Jian Peng, and Lei Zhang. Boosting weakly supervised object detection with progressive knowledge transfer. In *ECCV*, 2020.

[62] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

[63] Yanzhao Zhou, Xin Wang, Jianbin Jiao, Trevor Darrell, and Fisher Yu. Learning saliency propagation for semi-supervised instance segmentation. In *CVPR*, 2020.