# Supplementary for 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

In this document, we provide additional materials to support our main submission. In Section 1, we introduce the basic splits of PASCAL VOC 2012 dataset [4] used in the main paper. All the experiments in Section 2, 3, 4, 5, 6, 7, 8, 9 and 10 depend on these basic splits. In Section 2, we analyze the hyper-parameters in our model. In Section 3, 4 and 5, we perform ablation studies to verify the functionalities of different parts in our method. In Section 6, we analyze the source of improvement for the final segmentation results. In Section 7, 8, 9 and 10, we supplement experiments in the main paper with more quantitative and qualitative results of pseudo segmentation labels and final segmentation results. In Section 11 and 12, we then extend the basic splits to more general ones to deal with additional problems in weak-shot semantic segmentation. Finally, we discuss the limitations and future works in Section 13.

## 1 Basic Splits of PASCAL VOC 2012 Dataset

As mentioned in Section 4.1 in the main paper, we follow the category splits in PASCAL-$5^i$ [6] to form four basic folds, *i.e.*, fold 0, fold 1, fold 2, and fold 3. In fold $i$ ($0 \leqslant i \leqslant 3$), categories with indices $5i+1, 5i+2, 5i+3, 5i+4, 5i+5$ are regarded as 5 novel categories and the remaining categories (including "background") are regarded as 16 base categories. The detailed split of base categories and novel categories for each fold is shown in Table 3.

We use the augmented training set (*trainaug* set) with 10582 images from SBD [5] as the complete training set for our experiments conducted on PASCAL VOC 2012 dataset. Based on the abovementioned category splits, we further divide 10582 training samples into base samples and novel samples for each fold. The images containing only base categories are included as base samples, whereas those containing at least one novel category are deemed

*Corresponding author.
© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

| | *trainaug* | | *train* | |
|---|---|---|---|---|
| **fold** | $\mathcal{N}^b$ | $\mathcal{N}^n$ | $\mathcal{N}^b$ | $\mathcal{N}^n$ |
| **0** | 7746 | 2836 | 1051 | 413 |
| **1** | 6978 | 3604 | 950 | 514 |
| **2** | 5040 | 5542 | 800 | 664 |
| **3** | 8437 | 2145 | 1091 | 373 |

Table 1: The number of base samples and novel samples in fold $i$ ($0 \leqslant i \leqslant 3$). $\mathcal{N}^b$ and $\mathcal{N}^n$ represent the number of base samples and novel samples, respectively.

| **fold** | sample | acc. | prec. | recall | $F_1$-score |
|---|---|---|---|---|---|
| **0** | base | 0.831 | 0.594 | 0.777 | 0.612 |
| **0** | novel | 0.854 | 0.597 | 0.774 | 0.622 |
| **1** | base | 0.846 | 0.593 | 0.777 | 0.615 |
| **1** | novel | 0.836 | 0.594 | 0.743 | 0.614 |

Table 2: Analysis of boundary prediction results on PASCAL VOC 2012 *train* set. Base samples and novel samples in fold 0 and fold 1 are evaluated separately.

novel samples. The total count of base samples and novel samples in each fold are listed in Table 1 for both the *trainaug* set and the official *train* set.

Following recent WSSS researches [1, 3, 8, 11], we adopt *trainaug* set for model training, *train* set for evaluating CAMs or pseudo labels, and *val* set or *test* set for evaluating segmentation results.

## 2 Analysis of Hyper-parameter

Recall that we use a threshold $\tau$ to divide the boundary prediction results into boundary regions and non-boundary regions, as discussed in Section 3.2 in the main paper. To investigate the impact of $\tau$, we conduct experiments by trying different $\tau$ in the range of $[0.1, 0.9]$ on all four folds. Table 4 summarizes the *train* set mIoUs of revised responses, which utilize CAM [12] as the initial response and our proposed RETAB as the propagation strategy. We can observe that most high mIoUs occur when $\tau$ lies in the range of $[0.3, 0.7]$. In our implementation, we consistently set $\tau = 0.5$ in all experiments.

## 3 Analysis of Boundary Transfer

To demonstrate the effectiveness of boundary transfer, we evaluate boundary predictions on PASCAL VOC 2012 *train* set. Figure 1 visualizes some boundary maps of base samples and novel samples predicted by our boundary network on fold 0 and fold 1. As discussed in Section 3.2 in the main paper, the predicted boundary map can be transformed into a binary prediction map, where 1 and 0 respectively denote the pixels in boundary regions and non-boundary regions. Ground-truth binary boundary labels for evaluation can be obtained by transforming the semantic boundary labels in SBD [5] into class-agnostic ones. Then we can evaluate the performance of the boundary network in a binary classification manner. As shown in Table 2, we respectively calculate the accuracy, precision, recall, and $F_1$-score

| **class** | bg | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbk | person | plant | sheep | sofa | train | tv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **index** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| **fold 0** | *b* | *n* | *n* | *n* | *n* | *n* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* |
| **fold 1** | *b* | *b* | *b* | *b* | *b* | *b* | *n* | *n* | *n* | *n* | *n* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* |
| **fold 2** | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *n* | *n* | *n* | *n* | *n* | *b* | *b* | *b* | *b* | *b* |
| **fold 3** | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *n* | *n* | *n* | *n* | *n* |

Table 3: The basic splits of 21 categories for PASCAL VOC 2012 dataset. In each fold, '*b*' denotes a base category and '*n*' denotes a novel category. *bg* is the background category.

| | fold 0 | | | fold 1 | | | fold 2 | | | fold 3 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau$ | $\mathcal{C}$ | $\mathcal{C}^b$ | $\mathcal{C}^n$ | $\mathcal{C}$ | $\mathcal{C}^b$ | $\mathcal{C}^n$ | $\mathcal{C}$ | $\mathcal{C}^b$ | $\mathcal{C}^n$ | $\mathcal{C}$ | $\mathcal{C}^b$ | $\mathcal{C}^n$ |
| 0.1 | 69.0 | 71.7 | 60.1 | 68.6 | 68.3 | 69.7 | 68.6 | 68.1 | 70.3 | 69.0 | 70.8 | 63.3 |
| 0.2 | 70.3 | 72.8 | 62.3 | 70.3 | 70.0 | 71.0 | 70.1 | 69.4 | 72.3 | 70.1 | 71.9 | **64.1** |
| 0.3 | 70.9 | 73.4 | **62.8** | 70.8 | 70.5 | **71.9** | 70.7 | 70.0 | 72.9 | **70.3** | 72.3 | **64.1** |
| 0.4 | 71.0 | 73.7 | 62.6 | 71.1 | 70.9 | 71.6 | 70.8 | 70.1 | 73.0 | 70.2 | 72.4 | 63.4 |
| 0.5 | **71.2** | 74.0 | 62.5 | 71.3 | 71.2 | 71.6 | **70.9** | 70.2 | **73.3** | 70.1 | 72.4 | 62.8 |
| 0.6 | 71.1 | 74.0 | 62.1 | 71.3 | 71.3 | 71.4 | **70.9** | 70.2 | 73.2 | 70.1 | **72.5** | 62.3 |
| 0.7 | **71.2** | **74.2** | 61.8 | **71.4** | **71.5** | 71.2 | **70.9** | **70.4** | 72.7 | 69.9 | 72.4 | 61.9 |
| 0.8 | 71.1 | 74.1 | 61.7 | 71.3 | 71.4 | 71.0 | **70.9** | **70.4** | 72.6 | 69.8 | 72.4 | 61.5 |
| 0.9 | 71.1 | 74.1 | 61.5 | 71.3 | 71.4 | 70.8 | 70.8 | 70.3 | 72.5 | 69.4 | 72.3 | 61.2 |

Table 4: Different choices of hyper-parameter $\tau$ on four folds. The mIoU evaluation of revised responses for "CAM+RETAB" is conducted on PASCAL VOC 2012 *train* set.

for base samples and novel samples in the *train* set on fold 0 and fold 1. As illustrated, although the boundary network is trained on only base samples, it also performs well in predicting boundaries for novel samples, verifying the effectiveness of knowledge transfer across categories.

# 4 Analysis of Semantic Affinity

In this subsection, we compare the semantic affinity learned by PSA [1] and our RETAB on PASCAL VOC 2012 *train* set. The predicted semantic affinities $\hat{a}_{ij}$ are transformed into the binary values $\hat{a}'_{ij}$ for both base samples and novel samples: $\hat{a}'_{ij} = 1$ if $\hat{a}_{ij} \geqslant 0.5$, and $\hat{a}'_{ij} = 0$ otherwise. Ground-truth semantic affinities $a^*_{ij}$ could be obtained by the method described in Section 3.2 in the main paper. Then we can evaluate the predicted semantic affinities within neighbor set $\mathcal{P}$ (see Eqn.(1) in the main paper) in a binary classification manner. We respectively calculate the accuracy, precision, recall, and $F_1$-score for base samples and novel samples in the *train* set on fold 0 and fold 1. Table 5 displays the quantitative results comparing PSA and our RETAB. As illustrated, our method learns more satisfying semantic affinity for both base samples and novel samples than PSA, verifying RETAB's transferability of semantic affinity from base categories to novel categories.

In Table 6, we conduct ablation studies on PASCAL VOC 2012 *train* set to investigate the semantic affinities learned by affinity networks under different training strategies. We show the performance of revised responses for different strategies using the same initial response and the same propagation method. As shown in Table 6, using pseudo affinity labels of all samples to train the affinity network (row 1) results in low mIoU scores. When we only utilize ground-truth affinity labels of base samples for training (row 2), the model witnesses a

Figure 1: Visualizations on boundary prediction results for base samples and novel samples in fold 0 and fold 1 of PASCAL VOC 2012 *train* set. We display several predicted boundary maps together with their corresponding ground-truth boundary maps.

| | | PSA | | | | RETAB | | | |
|---|---|---|---|---|---|---|---|---|---|
| **fold** | sample | acc. | prec. | recall | $F_1$ | acc. | prec. | recall | $F_1$ |
| **0** | base | 0.811 | 0.583 | 0.664 | 0.597 | **0.841** | **0.633** | **0.763** | **0.662** |
| **0** | novel | 0.823 | 0.577 | 0.643 | 0.591 | **0.862** | **0.639** | **0.745** | **0.669** |
| **1** | base | 0.821 | 0.578 | 0.650 | 0.591 | **0.852** | **0.634** | **0.762** | **0.664** |
| **1** | novel | 0.802 | 0.587 | 0.667 | 0.600 | **0.836** | **0.628** | **0.731** | **0.653** |

Table 5: Analysis on semantic affinities for base samples and novel samples on PASCAL VOC 2012 *train* set. Semantic affinities predicted by PSA and our RETAB are evaluated on fold 0 and fold 1.

significant performance improvement because more precise semantic affinities can be learned, especially on boundary regions. This phenomenon implies that even if our RETAB (boundary network and affinity network) is only supervised on base samples, affinities could also be transferred well to novel samples. This advantage makes our RETAB more practical in real-world applications because we can directly use a RETAB pretrained on base samples to perform inference on extensive novel samples without further fine-tuning. In row 3 of Table 6, we further use pseudo affinities of novel samples to train the affinity network in a mixed-supervised manner, which slightly enhances the performance on novel categories. After filtering out noisy supervisions in boundary regions of CAM for novel samples (row 4), we finally obtain the best strategy used in our RETAB.

## 5 Analysis of Boundary-aware Two-stage Propagation

By taking fold 0 and fold 1 as examples, we analyze the functionality of boundary-aware two-stage propagation on Pascal VOC 2012 *train* set. We compare three propagation methods: 1) "one-stage": classical one-stage random walk [1], 2) "nbd+bd": separately propagate in

| Supervision for base samples | Supervision for novel samples | fold 0 all | fold 0 base | fold 0 novel | fold 1 all | fold 1 base | fold 1 novel |
|---|---|---|---|---|---|---|---|
| pseudo | pseudo | 63.6 | 66.4 | 54.8 | 63.6 | 63.1 | 65.2 |
| GT | - | 70.8 | 73.8 | 61.0 | 70.6 | 70.8 | 69.8 |
| GT | pseudo | 70.7 | 73.7 | 61.2 | 70.9 | 71.0 | 70.5 |
| GT | pseudo+filter | **71.2** | **74.0** | **62.5** | **71.3** | **71.2** | **71.6** |

Table 6: The ablation study of semantic affinities learned by affinity networks with different training strategies. Utilizing "CAM" as the initial response and our two-stage propagation as the expansion algorithm, we show the mIoU evaluation for revised responses on PASCAL VOC 2012 *train* set. **GT**: use ground-truth affinity labels obtained from pixel-level segmentation labels. **pseudo**: use pseudo affinity labels generated from "CAM". **filter**: filter out coordinate pairs in boundary regions.

| Propagation strategy | fold 0 all | fold 0 base | fold 0 novel | fold 1 all | fold 1 base | fold 1 novel |
|---|---|---|---|---|---|---|
| one-stage | 70.7 | 73.7 | 61.1 | 70.3 | 71.1 | 67.9 |
| nbd+bd | 66.0 | 68.4 | 58.3 | 65.7 | 65.4 | 66.8 |
| BTP | **71.2** | **74.0** | **62.5** | **71.3** | **71.2** | **71.6** |

Table 7: The ablation study of different propagation strategies on fold 0 and fold 1 of Pascal VOC 2012 *train* set. We show the mIoU evaluation for revised responses, using "CAM" as the initial response and the best semantic affinities learned by our RETAB. **one-stage**: classical random walk. **nbd+bd**: separately propagate in the non-boundary regions and in the boundary regions. **BTP**: our proposed boundary-aware two-stage propagation strategy.

the non-boundary region and in the boundary region, where confident pixels can not guide the propagation of unconfident ones, 3) "BTP": our boundary-aware two-stage propagation as described in Section 3.2 in the main paper. We adopt CAM [12] as the initial response and use the best semantic affinities learned by our RETAB. The results are summarized in Table 7. As illustrated, our proposed "BTP" outperforms both "one-stage" and "nbd+bd" especially on novel categories, verifying the importance of utilizing confident pixels to guide the propagation of unconfident ones.

## 6 Source of Improvement

Table 8 gives an ablation study of each part in our approach that contributes to the improvement of the final segmentation results on PASCAL VOC 2012 *val* set. Experiments are conducted on fold 0. In Table 8, row 1 is the WSSS baseline PSA (CAM+RW) [1]. Row 2 is our created naive weak-shot segmentation baseline "seggt" and row 3 is the augmented weak-shot segmentation baseline "affgt+seggt". Except row 1, all experiments adopt ground-truth segmentation masks of base samples so as to train the segmentation network in a mixed-supervised manner. Comparing row 3 with row 2, we find that applying ground-truth affinity labels of base samples to train the affinity network can improve 0.3 base-mIoU (%) and 7.3 novel-mIoU (%). After filtering out noisy affinity labels in novel samples, the model in row 4 gains another 1.9 improvement on novel-mIoU (%). Comparing row 5 with row 4, replacing the original random walk with our proposed boundary-aware two-stage propagation further

| WSSS | affgt | FILTER | BTP | seggt | all-mIoU | base-mIoU | novel-mIoU |
|---|---|---|---|---|---|---|---|
| ✓ | | | | | 61.7 | 62.8 | 58.1 |
| ✓ | | | | ✓ | 71.9 | 76.3 | 57.8 |
| ✓ | ✓ | | | ✓ | 73.9 | 76.6 | 65.1 |
| ✓ | ✓ | ✓ | | ✓ | 74.3 | 76.6 | 67.0 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **75.0** | **76.8** | **69.2** |

Table 8: The ablation study for each source of improvement in RETAB. Models are trained on fold 0 and the final segmentation results are evaluated on VOC12 *val* set. **WSSS**: weakly-supervised semantic segmentation baseline PSA [1]. **affgt**: use ground-truth affinity labels for base samples when training the affinity network. **FILTER**: filter out coordinate pairs in boundary regions for novel samples. **BTP**: boundary-aware two-stage propagation. **seggt**: use ground-truth segmentation labels for base samples when training the segmentation network.

| | CAM | | | CAM+RW | | | CAM+RW(affgt) | | | CAM+RETAB | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **fold** | all | base | novel | all | base | novel | all | base | novel | all | base | novel |
| **0** | 48.0 | 51.4 | 37.4 | 61.0 | 63.8 | 52.3 | 69.7 | 73.1 | 58.7 | **71.2** | **74.0** | **62.5** |
| **1** | 48.0 | 47.8 | 48.8 | 61.0 | 61.0 | 61.2 | 70.0 | 71.0 | 66.6 | **71.3** | **71.2** | **71.6** |
| **2** | 48.0 | 47.2 | 50.7 | 61.0 | 58.9 | 67.9 | 69.7 | 69.6 | 70.0 | **70.9** | **70.2** | **73.3** |
| **3** | 48.0 | 47.6 | 49.4 | 61.0 | 62.0 | 57.9 | 69.0 | 71.9 | 59.8 | **70.1** | **72.4** | **62.8** |
| | SEAM | | | SEAM+RW | | | SEAM+RW(affgt) | | | SEAM+RETAB | | |
| **fold** | all | base | novel | all | base | novel | all | base | novel | all | base | novel |
| **0** | 55.4 | 58.8 | 44.6 | 63.6 | 66.4 | 54.8 | 67.5 | 71.1 | 56.1 | **68.2** | **71.2** | **58.4** |
| **1** | 55.4 | 55.3 | 55.7 | 63.6 | 63.1 | 65.2 | 67.2 | 67.1 | 67.7 | **68.1** | **67.6** | **69.7** |
| **2** | 55.4 | 53.2 | 62.3 | 63.6 | 60.9 | 72.4 | 67.0 | 65.1 | 73.2 | **67.9** | **65.7** | **74.8** |
| **3** | 55.4 | 56.0 | 53.6 | 63.6 | 65.5 | 57.6 | 66.7 | 69.4 | 58.2 | **67.5** | **69.7** | **60.7** |
| | CPN | | | CPN+RW | | | CPN+RW(affgt) | | | CPN+RETAB | | |
| **fold** | all | base | novel | all | base | novel | all | base | novel | all | base | novel |
| **0** | 57.4 | 60.5 | 47.5 | 67.8 | 70.5 | 59.3 | 70.5 | 73.0 | 62.3 | **71.5** | **73.3** | **65.9** |
| **1** | 57.4 | 55.9 | 62.3 | 67.8 | 66.6 | 71.6 | 71.2 | 70.7 | 72.9 | **72.1** | **71.3** | **74.6** |
| **2** | 57.4 | 56.3 | 61.0 | 67.8 | 66.3 | 72.7 | 71.6 | 70.9 | 74.0 | **72.3** | **71.5** | **75.0** |
| **3** | 57.4 | 58.6 | 53.8 | 67.8 | 69.3 | 62.9 | 70.2 | 72.2 | 64.0 | **70.7** | **72.4** | **65.1** |

Table 9: Mean IoU evaluation of initial responses and revised responses on VOC12 *train* set. We adopt three initial responses, *i.e.*, "CAM" [12], "SEAM" [8], and "CPN" [10]. We compare three response propagation methods: "RW", "RW(affgt)", and "RETAB". "RW" is introduced in PSA [1]. "RW(affgt)" is based on "RW" and further uses ground-truth affinity labels of base samples when training the affinity network. "RETAB" is our proposed method.

enhances 2.2 novel-mIoU (%), resulting in RETAB with the best mIoUs.

# 7 Evaluation on Pseudo Segmentation Labels

We evaluate pseudo segmentation labels by assessing CAMs before/after propagation on PASCAL VOC 2012 *train* set. To verify the transferability of our method, we adopt three

| 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 [7] | 64.9 | 67.2 | 57.5 | 64.9 | 62.9 | 71.2 | 64.9 | 63.2 | 70.4 | 64.9 | 67.8 | 55.7 |
| BES [3] | 65.7 | 67.0 | 61.7 | 65.7 | 63.8 | 71.6 | 65.7 | 64.7 | 68.8 | 65.7 | 68.7 | 56.0 |
| SvM [10] | 66.6 | 66.3 | 67.8 | 66.6 | 65.9 | 68.9 | 66.6 | 66.2 | 68.1 | 66.6 | 69.6 | 57.0 |
| CAM+RW [1] | 61.7 | 62.8 | 58.1 | 61.7 | 61.3 | 62.8 | 61.7 | 60.5 | 65.3 | 61.7 | 63.7 | 55.1 |
| CAM+RW(seggt) | 71.9 | 76.3 | 57.8 | 73.7 | 74.9 | 69.9 | 72.9 | 74.1 | 69.0 | 72.9 | 78.3 | 55.8 |
| CAM+RW(affgt+seggt) | 73.9 | 76.6 | 65.1 | 74.6 | 75.0 | 73.3 | 73.6 | 74.3 | 71.4 | 73.2 | **78.4** | 56.6 |
| CAM+RETAB | **75.0** | **76.8** | **69.2** | **75.7** | **75.6** | **76.1** | **74.2** | **74.9** | **72.0** | **73.7** | **78.4** | **58.5** |
| SEAM+RW [8] | 64.5 | 66.2 | 59.0 | 64.5 | 63.3 | 68.3 | 64.5 | 62.9 | 69.7 | 64.5 | 67.1 | 56.2 |
| SEAM+RW(seggt) | 72.4 | 76.2 | 60.2 | 74.1 | 74.5 | 72.6 | 72.6 | 72.9 | 71.7 | 73.2 | 78.2 | 57.4 |
| SEAM+RW(affgt+seggt) | 73.3 | 76.5 | 63.1 | 74.6 | 74.7 | 74.3 | 73.1 | 73.3 | 72.5 | 73.8 | 78.7 | 58.1 |
| SEAM+RETAB | **74.0** | **76.6** | **65.4** | **74.9** | **75.0** | **74.5** | **73.6** | **73.7** | **73.0** | **74.3** | **79.1** | **58.9** |
| CPN+RW [10] | 67.8 | 69.1 | 63.8 | 67.8 | 65.9 | 73.7 | 67.8 | 66.8 | 70.9 | 67.8 | 70.8 | 58.1 |
| CPN+RW(seggt) | 73.7 | 76.4 | 65.0 | 74.9 | 75.0 | 74.6 | 74.0 | 74.9 | 71.0 | 73.5 | 78.1 | 58.8 |
| CPN+RW(affgt+seggt) | 74.7 | **76.8** | 68.1 | 75.4 | 75.5 | 75.2 | 74.3 | 75.0 | 71.9 | 73.9 | 78.6 | 59.0 |
| CPN+RETAB | **75.1** | 76.7 | **70.0** | **75.9** | **75.7** | **76.5** | **74.8** | **75.3** | **73.3** | **74.5** | **79.0** | **60.3** |
| Fully Oracle | 76.6 | 76.8 | 76.1 | 76.6 | 76.2 | 78.1 | 76.6 | 75.4 | 80.6 | 76.6 | 79.2 | 68.2 |

Table 10: Comparison of segmentation performance on PASCAL VOC 2012 *val* set. Columns marked by $\mathcal{C}$, $\mathcal{C}^b$ and $\mathcal{C}^n$ represents all-mIoU, base-mIoU and novel-mIoU, respectively. We have five groups of methods. The first group includes some popular WSSS methods. The second / third / fourth group represents our WSSS baseline PSA / SEAM / CPN, together with our created weak-shot segmentation baselines (seggt and affgt+seggt), and our implemented RETAB with the same initial response. The fifth group is a fully-supervised method. Except the first group, all experiments use DeepLab [2] with ResNet38 [9] backbone.

initial responses: "CAM", "SEAM" and "CPN". Table 9 compares the performance of three response expanding methods, *i.e.*, "RW", "RW(affgt)" and our proposed "RETAB", where "RW(affgt)" is based on "RW" and further uses ground-truth affinity labels of base samples when training the affinity network. As illustrated, "RW(affgt)" greatly advances base-mIoUs since more accurate semantic affinities in base samples are learned. Still, this method can not achieve satisfying results on novel categories. By comparison, our "RETAB" achieves consistently higher mIoU results than baselines on all four folds. In detail, "RETAB" improves 9.1 $\sim$ 10.3 all-mIoU (%), 10.2 $\sim$ 11.3 base-mIoU (%) and 4.9 $\sim$ 10.4 novel-mIoU (%) when compared with "RW" in PSA[1], improves 3.9 $\sim$ 4.6 all-mIoU (%), 4.2 $\sim$ 4.8 base-mIoU (%) and 2.4 $\sim$ 4.5 novel-mIoU (%) when compared with "RW" in SEAM, and improves 4.3 $\sim$ 6.6 all-mIoU (%), 4.0 $\sim$ 7.8 base-mIoU (%) and 3.2 $\sim$ 11.1 novel-mIoU (%) when compared with "RW" in CPN. These improvements come from two aspects. On the one hand, we utilize more confident semantic affinities by incorporating ground-truth labels and boundary predictions in affinity learning. On the other hand, two-stage propagation makes it possible that confident pixels guide the propagation of unconfident ones. Interestingly, although the initial response "SEAM" outperforms "CAM", "CAM+RETAB" works better than "SEAM+RETAB" after propagation. It shows that the usefulness of RETAB is related to the chosen initial response.

[1]The official *train* set all-mIoU of revised responses is 58.1 in PSA [1], and our reproduced result is 61.0.

# 8 Evaluation on Segmentation Performance

This section supplements Section 4.3 in the main paper. Table 1 in the main paper shows the segmentation performance on the *test* set. In this section, Table 10 further provides results on the *val* set. By comparison, our method still outperforms baselines on all folds, especially for novel categories, which is consistent with the performance on the *test* set.

Figure 2: Visualized pseudo segmentation labels of novel samples on PASCAL VOC 2012 *train* set. (a) image. (b) GT label. (c) "CAM+RW". (d) "CAM+RETAB". (e) "SEAM+RW". (f) "SEAM+RETAB". (g) "CPN+RW". (h) "CPN+RETAB". Names of novel categories presented in the images are marked as blue on the left side.

Figure 3: Qualitative semantic segmentation results for novel categories on PASCAL VOC 2012 *val* set. (a) image. (b) GT label. (c) "CAM+RW". (d) "CAM+RETAB". (e) "SEAM+RW". (f) "SEAM+RETAB". (g) "CPN+RW". (h) "CPN+RETAB". Names of novel categories presented in the images are marked as blue on the left side.

# 9 Qualitative Results on Pseudo Segmentation Labels

In Figure 2, we show more visualizations of pseudo segmentation labels for novel samples generated by different methods in fold 0, 1, 2, and 3 on PASCAL VOC 2012 *train* set, which supplements Figure 3 in the main paper. In Figure 2, column (c), (e), and (g) represent the

WSSS baseline PSA [1], SEAM [8], and CPN [10]. Column (d), (f), and (h) denote our implemented RETAB with the same initial responses as these three baselines, respectively. By comparing (c) with (d), comparing (e) with (f), and comparing (g) with (h), we find that our method outperforms all baselines in recovering object shapes on the boundary regions of pseudo labels. Boundary-aware two-stage propagation restricts the random walk process within the target object regions and allows confident pixels to guide unconfident ones, resulting in pseudo segmentation labels with higher quality.

## 10 Qualitative Results on Semantic Segmentation

We show more visualizations in Figure 3 for the segmentation results on PASCAL VOC 2012 *val* set predicted by different methods in fold 0, 1, 2 and 3, which supplements Figure 4 in the main paper. In Figure 3, column (c), (e), and (g) respectively displays the WSSS baseline PSA [1], SEAM [8], and CPN [10]. It is obvious that object boundaries predicted by these baselines are not precise. Also, some adjacent objects are wrongly classified as inaccurate categories. Column (d), (f), and (h) in Figure 3 represent our implemented RETAB with the same initial responses and the same segmentation network as three baselines, respectively. For a fair comparison, we compare (c) with (d), compare (e) with (f), and compare (g) and (h) to verify the effectiveness of our method. As illustrated, RETAB performs considerably better than three baselines in filling object regions while not overstepping the object boundaries. Also, our method can output more delicate details, like animal legs and plant leaves. Consequently, we can safely claim that RETAB successfully transfers pixel-level knowledge from base categories to novel ones and narrows the gap between weakly-supervised and fully-supervised segmentation.

## 11 Generalization to Potential Novel Categories in the Background

Recall that the basic goal of weak-shot semantic segmentation is to improve the segmentation performance on weakly-annotated images (novel samples) with an auxiliary fully-annotated dataset (base samples). In practical scenarios, base samples might contain potential novel categories in some regions labeled "background". Specifically, when labeling the fully-annotated dataset, certain categories (base categories) are selected to annotate precisely, and all the other pixels are marked as "background". Actually, these background pixels may contain novel categories we aim at. If we can discover these potential novel categories in base samples, we might achieve better segmentation performances on novel categories for the newly-coming weakly-annotated images.

To simulate the above situation, we propose two novel folds for PASCAL VOC 2012 dataset: fold $0^*$ and fold $1^*$. They follow the same category split rule as fold 0 and fold 1, respectively. Furthermore, based on the split of base samples and novel samples in fold 0 and fold 1, we further randomly choose approximately 50% novel samples and move them to base samples by changing their pixel labels of novel categories into the "background" label with index 0. We refer to these new labels as revised segmentation labels. Consequently, base samples now contain potential novel categories in the background. After this operation, the number of base samples increases, and the number of novel samples decreases. In our experiments, fold $0^*$ contains 9184 base / 1398 novel samples, and fold $1^*$ contains 8755 base

| Method | fold 0* | | | fold 1* | | |
|---|---|---|---|---|---|---|
| | $\mathcal{C}$ | $\mathcal{C}^b$ | $\mathcal{C}^n$ | $\mathcal{C}$ | $\mathcal{C}^b$ | $\mathcal{C}^n$ |
| CAM+RW(seggt) | 66.3 | 76.0 | 35.1 | 63.0 | 74.8 | 25.1 |
| CAM+RW(affgt+seggt) | 67.5 | 76.1 | 39.9 | 64.6 | 75.0 | 31.3 |
| CAM+RETAB | 67.8 | 74.4 | 46.5 | 65.3 | 74.4 | 36.3 |
| CAM+RETAB+ST | **71.9** | **76.2** | **58.0** | **74.5** | **75.8** | **70.3** |
| Fully Oracle | 76.6 | 76.8 | 76.1 | 76.6 | 76.2 | 78.1 |

Table 11: Comparison of segmentation performance on PASCAL VOC 2012 *val* set for fold 0* and fold 1*. "CAM+RW(seggt)" and "CAM+RW(affgt+seggt)" are weak-shot segmentation baselines. "CAM+RETAB" denotes our implemented RETAB. "CAM+RETAB+ST" denotes our method followed by a self-training step. "Fully Oracle" is the fully-supervised upper bound. All experiments use DeepLab [2] with ResNet38 [9] backbone for segmentation.

/ 1827 novel samples in the *trainaug* set. Next, we discuss the training strategy to deal with this typical problem, which is also included in our weak-shot semantic segmentation task.

Based on the new split and the revised segmentation labels, the training process of our RETAB and the segmentation network remains unchanged. Furthermore, we adopt a self-training strategy (*abbr.* ST) after the segmentation step to discover potential novel categories in the background of base samples. ST can be regarded as a self-supervised fine-tuning step for the segmentation network with mixed supervisions from base samples and novel samples. We introduce each iteration of ST in detail as follows, which includes three steps. First, we use the current segmentation network to perform segmentation predictions on the current batch of training samples. Second, for the background pixels classified as novel categories, we flip the pixel label from "background" to the predicted novel category for each of these pixels. Third, the revised segmentation labels are used to fine-tune the segmentation network via back-propagation. By gradually discovering potential novel categories in the background, the segmentation performance of novel categories could be iteratively improved. The above procedure is repeated until a predefined number of iteration is reached. In our implementation, we apply ST for 10000 iterations. Except that the initial learning rate is adjusted to half, all other parameters are not changed when we fine-tune the segmentation network.

Table 11 summarizes segmentation results on the *val* set for fold 0* and fold 1*. We list four methods: 1) our created naive weak-shot baseline "CAM+RW(seggt)" which uses revised labels of base samples and pseudo labels of novel samples generated by PSA [1] for segmentation, 2) our created augmented weak-shot baseline "CAM+RW(affgt+seggt)" which is based on the naive baseline and further uses ground-truth affinity labels of base samples for affinity training, 3) our method "CAM+RETAB" which uses revised labels of base samples and pseudo labels of novel samples generated by our RETAB for segmentation, and 4) our method with self-training "CAM+RETAB+ST". As shown in Table 11, our method consistently outperforms all the baselines on novel categories. After adopting the self-training strategy, more potential novel categories in the base samples are discovered. As a result, the segmentation network witnesses a huge improvement on novel categories. In conclusion, 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.

| **class** | bg | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbk | person | plant | sheep | sofa | train | tv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **index** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| **fold 4** | *b* | *n* | *n* | *n* | *n* | *n* | *n* | *n* | *n* | *n* | *n* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* | *b* |
| **fold 5** | *b* | *n* | *n* | *n* | *n* | *n* | *n* | *n* | *n* | *n* | *n* | *n* | *n* | *n* | *n* | *n* | *b* | *b* | *b* | *b* | *b* |

Table 12: The category splits of fold 4 and fold 5 for PASCAL VOC 2012 dateset. In each fold, '*b*' denotes a base category and '*n*' denotes a novel category. *bg* represents for the background category.

| Method | fold 4 | | | fold 5 | | |
|---|---|---|---|---|---|---|
| | $\mathcal{C}$ | $\mathcal{C}^b$ | $\mathcal{C}^n$ | $\mathcal{C}$ | $\mathcal{C}^b$ | $\mathcal{C}^n$ |
| CAM+RW [1] | 61.7 | 62.8 | 60.5 | 61.7 | 60.6 | 62.1 |
| CAM+RW(seggt) | 67.7 | 72.6 | 62.4 | 64.2 | 67.8 | 62.8 |
| CAM+RW(affgt+seggt) | 70.6 | 74.8 | 66.9 | 66.7 | 70.9 | 65.0 |
| CAM+RETAB | **74.3** | **75.6** | **72.8** | **69.3** | **71.5** | **68.4** |
| Fully Oracle | 76.6 | 76.2 | 77.1 | 76.6 | 72.6 | 78.2 |

Table 13: Comparison of segmentation performance on fold 4 and fold 5 of PASCAL VOC 2012 *val* set. "CAM+RW" represents our WSSS baseline PSA. "CAM+RW(seggt)" and "CAM+RW(affgt+seggt)" are weak-shot segmentation baselines. "CAM+RETAB" denotes our implemented RETAB. "Fully Oracle" is the fully-supervised upper bound. All experiments use DeepLab [2] with ResNet38 [9] backbone for segmentation.

## 12 Generalization to Fewer Fully-annotated Training Data

To prove that RETAB can use a small proportion of base samples to facilitate a large number of novel samples, we create fold 4 and fold 5 to include more novel categories and fewer base categories. As shown in Table 12, there are 10 novel categories in fold 4 and 15 novel categories in fold 5. Fold 4 contains 4443 base / 6139 novel samples, and fold 5 contains 1014 base / 9568 novel samples in the *trainaug* set. Following the evaluation setting in Section 4.3 in the main paper, we compare our method "CAM+RETAB" with the WSSS baseline "CAM+RW" and our created weak-shot segmentation baselines: "CAM+RW(seggt)" and "CAM+RW(affgt+seggt)". The segmentation results on the *val* set are listed in Table 13. Our method surpasses all baseline methods, especially on novel categories. Also, it can recover a considerable proportion of the fully-supervised upper bound. These evaluation results prove that our proposed RETAB can generalize well even if weakly-annotated novel samples are much more than fully-annotated base samples.

## 13 Limitations and Future Works

In Figure 4, we list some failure cases of segmentation results of our method on the *val* set. Although RETAB successfully narrows the gap between weakly-supervised and fully-supervised segmentation, it could also probably make unsatisfying predictions on some hard novel categories, like "bottle" in fold 0 and "person" in fold 2. "bottle" objects tend to have paster on the body, which could easily be treated as part of the boundaries, thus misleading the propagation to undesirable results. "person" objects take up a considerable proportion

Figure 4: Failure cases of semantic segmentation results on VOC12 *val* set. Names of novel categories presented in the images are marked as blue on both sides of the figure.

in the dataset. This category has spatial relationships with many other categories, making "person" hard to segment when it is split into the set of novel categories.

In this work, we focus on the second step under the typical WSSS framework to improve the response expansion algorithm. However, the first step, *i.e.*, the way to obtain the initial response, is also crucial. The quality of the initial response directly affects the performance of downstream operations. In the future, we will focus on how to transfer pixel-level knowledge from base categories to novel categories to improve the performance of the initial response for novel samples. Since our proposed RETAB can work on any initial response, the improved initial response could also use RETAB as the second step for response expansion. We hope that RETAB can function as a simple yet effective baseline to facilitate future researches on weak-shot semantic segmentation.

# 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] 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.

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

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

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

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

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

[9] 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.

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

[11] 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.

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