# *Supplementary Material* for AEFT with GPP Layer for WSSS

Sung-Hoon Yoon*, Hyeokjun Kweon*, Jegyeong Cho, Shinjeong Kim, and Kuk-Jin Yoon

Korea Advanced Institute of Science and Technology
{yoon307,0327june,j2k0618,aakseen,kjyoon}@kaist.ac.kr

## 1 Additional Discussions

By replacing the GAP layer with the proposed Gated Pyramid Pooling (GPP) layer, the performance is increased to 54.2% (+5.8%). By incorporating the proposed Adversarial Erasing Framework via Triplet (AEFT) with the GPP layer, the mIoU performance is further increased to 56.0% (+7.6%). Though the proposed method achieves state-of-the-art in segmentation level, ours show comparable or even inferior performance in seed level than other methods (Refer Table 4). Although the performance of seed itself is important, the quality of pseudo-labels is much more important in WSSS. Since the seed with high mIoU does not ensure the good quality of pseudo-labels (Refer Table 4), it is also important to generate CAMs that can benefit from the post-processing techniques (CRF/RW/IRN), which are commonly used in WSSS. In this section, we provide additional experimental results (ablation) on PASCAL VOC 2012.

### 1.1 GPP: Compatibility with other state-of-the-art

To verify whether the proposed GPP layer can be applied to other state-of-the-art methods or not, we incorporate the GPP layer with SEAM [50]. The reported performance of CAMs of SEAM is 55.4% with a classifier based on the GAP layer. By replacing the GAP layer with the proposed GPP layer, the performance of resulting CAMs is increased to 57.5% (+2.1%) in mIoU. Since the GPP layer does not require the specific architecture or learning framework, we expect that it can be employed in subsequent WSSS studies as the baseline.

### 1.2 GPP: Eq.4 and Eq.5

Activation functions of which ranges are positive (*e.g.* ReLU, Sigmoid) usually have been applied to focus on positive activation when generating CAMs. Previous WSSS works for spatially varying pooling method [3,10] also regarded the positive values of CAMs only by using softmax function on CAMs. However, in the perspective of generating precise CAMs, it is important to consider the

* Equal contribution.

regions with negative activation (*e.g.* background or regions of the other classes) as well as the positively activated regions. To effectively incorporate the negatively activated regions in the learning process, we devise the proposed method based on the GPP layer to amplify the activation of CAMs in both positive and negative directions (refer Eq. 5 in the main paper). In the cases of using only positive ($ReLU(\hat{P}_{16}) \odot ReLU(f)$) or negative ($ReLU(-\hat{P}_{16}) \odot ReLU(-f)$) when training the classifier, the performance was lower than the baseline (48.4%) in both cases. Furthermore, when $\beta$ in Eq.4 in the main paper is set to 1 (without gating negative activation), the mIoU performance is 53.6% which is 0.6% lower than the best model (54.2%). This ablation study for the GPP layer supports that simultaneously utilizing the positive and negative activation can be helpful for the network to generate precise CAMs.

### 1.3 AEFT: Without GPP

When we train the network with the AEFT only (without GPP), we can also achieve meaningful performance gain (↑ 4.1%, 52.5% mIoU) compared to the baseline. However, as shown in the third row of Table 2 (*Direct*), directly maximizing or minimizing the distance of embedding obtained from CAMs degrade the mIoU. Thus, we can not fully decouple the GPP layer from the AEFT.

### 1.4 AEFT: Grid Search for Hyperparameters

To verify whether our framework is sensitive to the threshold, we evaluate the proposed method from GT as shown in Table A1 and find our framework is quite robust to threshold. We set $t_L$ to not much remain the object, while $t_H$ is set to perfectly erase high-confident regions. Since the role of $t_H$ is to erase the high confidence region, we observe that the performance is not degraded unless $t_H$ is too low (e.g. >0.50). We observed too large $t_L$ makes it easy for the model to increase the distance between the anchor ($e_{AL}$) and negative ($e_N$) embeddings (since the object has overly remained). However, when $t_L$ is lower than 0.2, the deviation of mIoU is within 1% range. For the main paper, we select the best setting ($t_H = 0.60$ and $t_L = 0.20$).

**Table A1.** Ablation regarding threshold $(t_H, t_L)$ in AEFT. The column-index represents $t_H$ and the row-index represents $t_L$. The performance (mIoU, %) is evaluated with the PASCAL VOC 2012 *train* set.

| $t_L$ \ $t_H$ | 0.50 | 0.55 | 0.60 | 0.65 | 0.70 |
|---|---|---|---|---|---|
| 0.10 | 55.0 | 55.0 | 55.1 | 55.0 | 55.0 |
| 0.15 | 55.1 | 55.2 | 55.3 | 55.4 | 55.5 |
| 0.20 | 55.3 | 55.5 | **56.0** | 55.5 | 55.5 |
| 0.25 | 54.2 | 54.2 | 54.5 | 54.5 | 54.3 |

### 1.5 Comparison with softmaxed-CAMs

We implemented the softmaxed-CAMs as in [3] (background threshold = 0.2) and achieves 49.4%, which is 4.8% lower than the GPP. Since the class prediction in [3] is invariant to the size of CAMs, it suffers from inferior recall than using GAP. Qualitative comparison results of CAMs is shown in Fig. A1. As shown in figure, though the softmaxed-CAMs are sharper than CAMs (w/ GPP) but only localize the partial region.

## 2 Details Regarding MS-COCO Experiments

To show the superiority of the proposed method, we conducted an experiment on MS-COCO 2014 dataset. However, due to the page limit of the main paper, we provide additional experimental results regarding MS-COCO 2014 dataset in this *supplementary material*. For the experiment, the optimal $t_H$ and $t_L$ is set to 0.70 and 0.10, respectively. Since there exist more classes in MS-COCO 2014 dataset (80) than PASCAL VOC 2012 dataset (20), we let AEFT have more relaxed criteria for triplet loss with a smaller margin ($\epsilon$=0.1). The other settings (*e.g.* batch size and learning rate) are the same as the PASCAL VOC 2012 experiments. With the proposed Gated Pyramid Pooling (GPP) layer and Adversarial Erasing Framework via Triplet (AEFT), the mIoU quality of the generated seeds (CAMs) is increased to 38.5% (+7.2%) compared with the baseline (31.3%).

For the fair comparison with other state-of-the-art methods, we also apply IRN for refinement. The generated pseudo labels achieve 46.7% mIoU on MS-COCO 80k *train* set. With the help of the GPP layer that effectively captures information from the global context to fine-details and AEFT that expands the CAMs while preventing over-erasing, we also achieved state-of-the-art (44.8%) on MS-COCO 2014 dataset.

## 3 Additional Results

Table A2 shows the class-wise IoU of the semantic segmentation model on PASCAL VOC 2012 *val* set. In addition, we provide additional qualitative results for both PASCAL VOC 2012 and MS-COCO 2014 datasets. Since the results of CAMs generated by our proposed method were not provided in the main paper, we provide the qualitative comparison of our CAMs with the baseline CAMs in Fig. A2. Qualitative comparison results of the semantic segmentation is also shown in Fig A3. With the proposed GPP and AEFT, the generated CAMs are precise as well as activate the entire object regions. This is more evident in the MS-COCO 2014 dataset that contains more small and diverse objects, as shown in Fig A4. Here, Fig. A5 shows the segmentation results with single class while Fig. A6 shows the results with multiple classes in MS-COCO 2014. Since the segmentation model is trained with high-quality pseudo labels generated by the proposed framework, the model is not only good at localize the entire object but also can capture fine-details (even very small objects: refer second row in Fig A5 and fourth row in Fig A6).

**Table A2.** Class-wise IoU comparison on PASCAL VOC 2012 *val* set with only image-level supervision.

| Method | bg | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbk | person | plant | sheep | sofa | train | tv | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AffinityNet [1] | 88.2 | 68.2 | 30.6 | 81.1 | 49.6 | 61.0 | 77.8 | 66.1 | 75.1 | 29.0 | 66.0 | 40.2 | 80.4 | 62.0 | 70.4 | 73.7 | 42.5 | 70.7 | 42.6 | 68.1 | 51.6 | 61.7 |
| SEAM [50] | 88.8 | 68.5 | 33.3 | 85.7 | 40.4 | 67.3 | 78.9 | 76.3 | 81.9 | 29.1 | 75.5 | 48.1 | 79.9 | 73.8 | 71.4 | 75.2 | 48.9 | 79.8 | 40.9 | 58.2 | 53.0 | 64.5 |
| BES [10] | 88.9 | 74.1 | 29.8 | 81.3 | 53.3 | 69.9 | **89.4** | 79.8 | 84.2 | 27.9 | 76.9 | 46.6 | 78.8 | 75.9 | 72.2 | 70.4 | **50.8** | 79.4 | 39.9 | 65.3 | 44.8 | 65.7 |
| OC-CSE [27] | 90.2 | **82.9** | 35.1 | 86.8 | 59.4 | 70.6 | 82.5 | 78.1 | **87.4** | 30.1 | 79.4 | 45.9 | 83.1 | **83.4** | 75.7 | 73.4 | 48.1 | **89.3** | 42.7 | 60.4 | 52.3 | 68.4 |
| Ours | **91.9** | 77.6 | **37.8** | **88.9** | **64.5** | **73.8** | 87.8 | **81.2** | 87.1 | **34.6** | **83.9** | **52.9** | **85.3** | 82.0 | **77.0** | **79.7** | 38.9 | 88.5 | **44.4** | **74.4** | **56.0** | **70.9** |

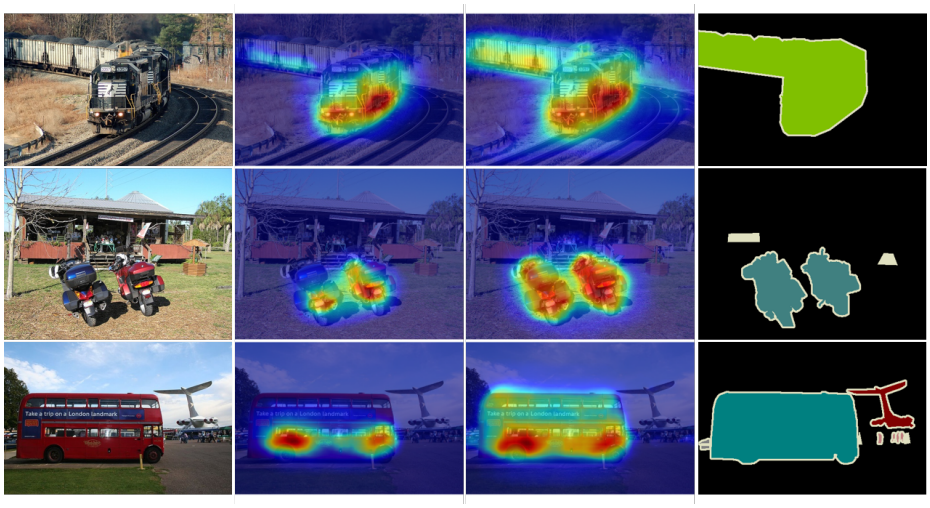

**Fig. A1.** Qualitative comparison results of CAMs. From left to right: Images, softmaxed-CAMs (similar to Araslanov *et al.* [3]), CAMs from Ours (only GPP), Ground truth labels. The images are from PASCAL VOC 2012 *train* set.

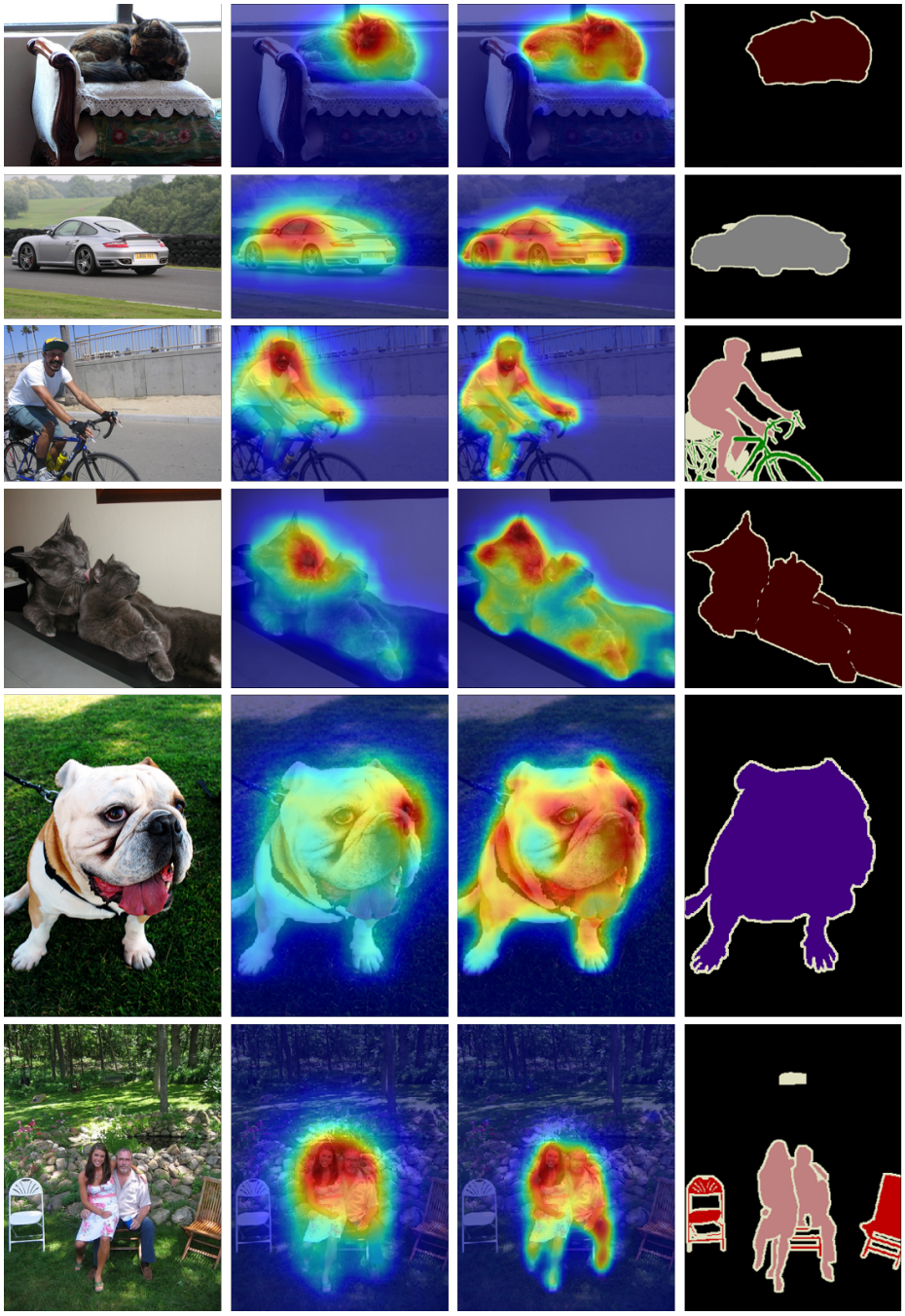

**Fig. A2.** Qualitative comparison results of CAMs between the baseline and the proposed method. From left to right: Images, Baseline CAMs [1], CAMs from Ours, Ground truth labels. The images are from PASCAL VOC 2012 *train* set.

**Fig. A3.** Qualitative segmentation results of the proposed method on the PASCAL VOC 2012 *validation* set. From top to bottom: Images, Ours, Ground truth labels.

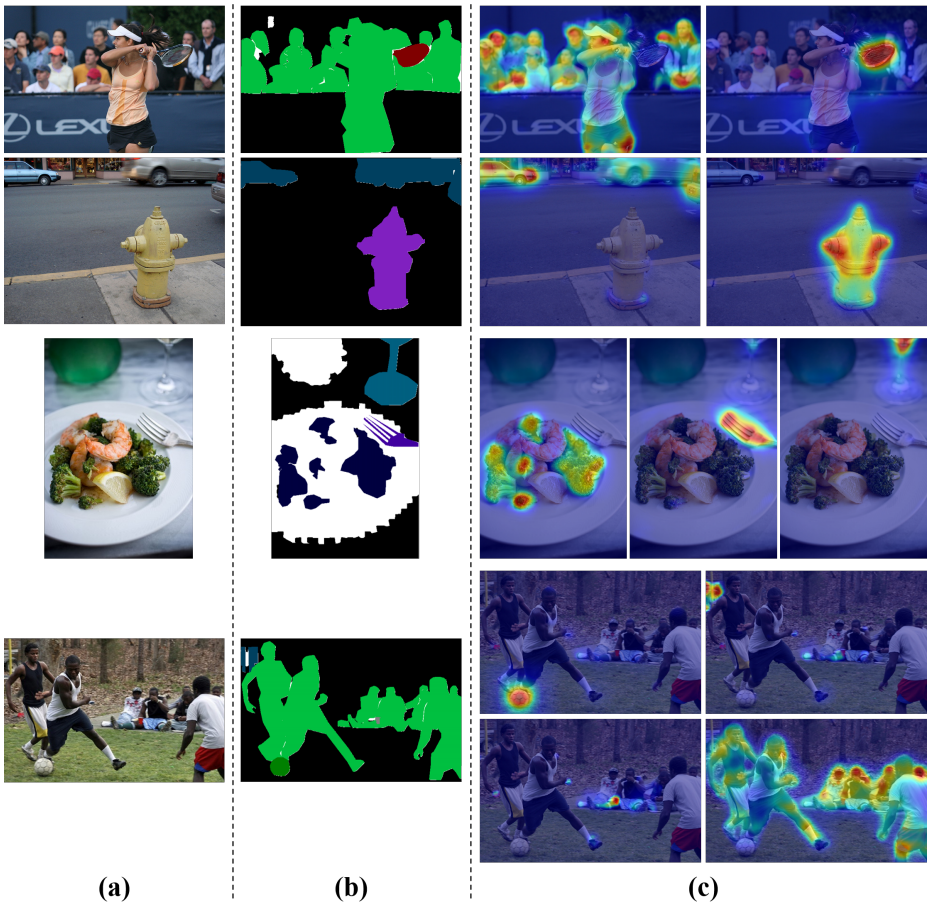


**Fig. A4.** Qualitative CAMs results of the proposed method. From (a) to (c): Images, Ground truth labels, and CAMs from Ours. The images are from MS-COCO 2014 *train* set.

**Fig. A5.** Qualitative segmentation results of the proposed method on the MS-COCO 2014 *val* set. Each image contains only one class. From left to right: Images, Ours, Ground truth labels.


**Fig. A6.** Qualitative segmentation results of the proposed method on the MS-COCO 2014 *val* set. Each image contains multiple classes and the legend located at the top represents each class. From left to right: Images, Ours, Ground truth labels.