Adversarial Erasing Framework via Triplet with Gated Pyramid Pooling Layer for Weakly Supervised Semantic Segmentation

Sung-Hoon Yoon*, Hyeokjun Kweon*, Jegyeong Cho, Shinjeong Kim, and Kuk-Jin Yoon (*equal contribution)
Visual intelligence Lab., Dept. Mechanical Eng., KAIST, Korea

EUROPEAN CONFERENCE ON COMPUTER VISION
TEL AVIV 2022

Motivation

- WSSS aims to learn SS with **weak yet inexpensive labels** (*e.g.* class tags) only.
- Most of WSSS methods exploits **Class Activation Maps (CAMs)** to localize the object in the image.
- However, from the perspective of SS, there are two main issues in CAMs: Impreciseness & Sparseness.

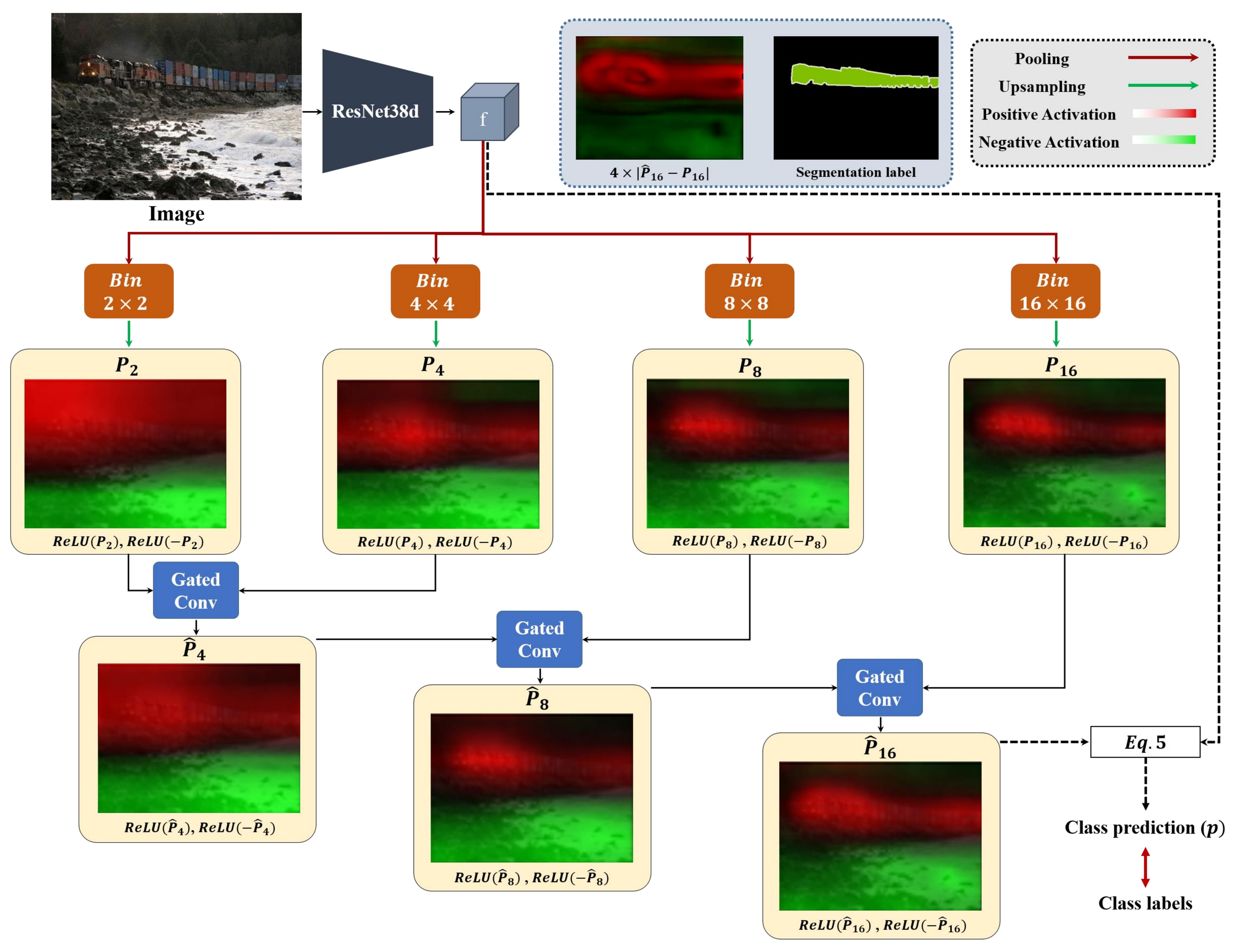
Impreciseness (Not fit with the object boundary)

- Most WSSS methods use Global Average Pooling (GAP) to aggregate the region-level feature maps into the image-level class prediction.
- However, the GAP layer **aggressively averages all the features** even on the object-irrelevant regions.

Sparseness (Highlight most discriminative regions only)

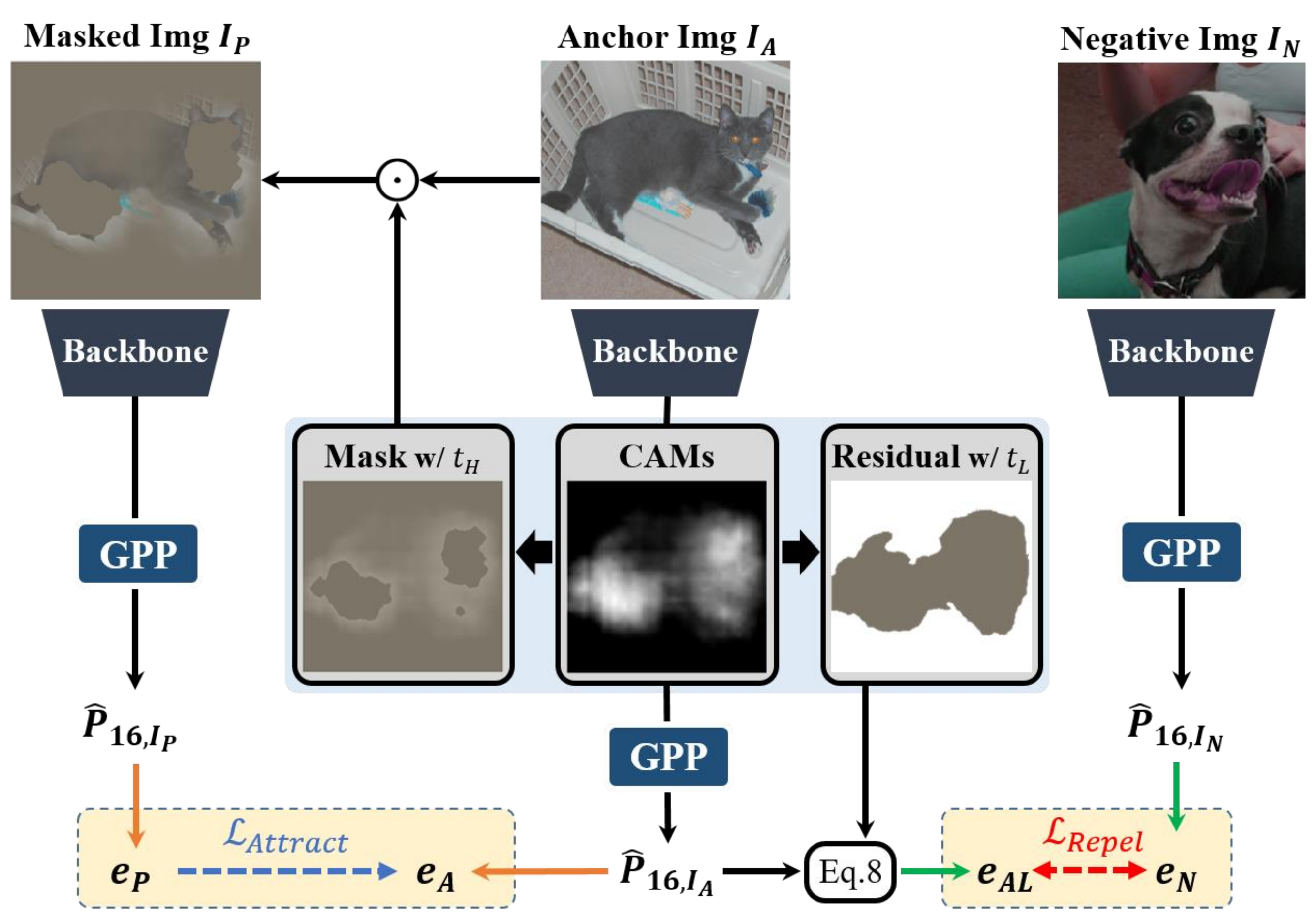
- Adversarial Erasing (AE) methods effectively address the sparseness problem by iteratively erasing-and-finding the most discriminative regions.
- However, AE methods suffer from **over-expansion problem due to the rigid classification loss.**

Gated Pyramid Pooling (GPP) layer



- GPP sets **different pooling weight** for each feature.
- The weighting factors are acquired at **multi-scale**, which is crucial to handle the impreciseness.
- We incorporate the **gating mechanism** to aggregate the feature pyramid into a single multi-scale-aware prediction, in a **coarse-to-fine** manner.
- Further, we devise a **sign-preserving attention** and make GPP to amplify both pos. and neg. directions.
  - Positive prediction: Existence of the class
  - Negative prediction: Non-existence of the class

Adversarial Erasing Framework via Triplet (AEFT)



- AEFT aims **understand the concept of erasing**, in a more **flexible** manner compared to rigid classification.
- We define the triplet as follows:
  - Anchor image I_A (original input image)
  - Positive image I_P (masked image)
  - Negative image I_N (image with no overlapped class)
- For anchor-negative pair, we erase the highly activated region and **make the repelling more difficult**.
- AEFT could use the CAMs as its metric space directly, but **using GPP feature space** is much more effective.

Ablation Studies

- The left table shows that the **GPP layer outperforms the GAP layer**.
  - The pyramid pooling of the features at multiple scales provides complementary benefits.
  - Our coarse-to-fine fusion approach is much better than naïve averaging or inverse direction.
- The right table shows that the **AEFT effectively handles the over-expansion problem**.
  - Compared to the *rigid* one, our *soft* approach enables higher recall without losing precision.
  - Also, using negative samples (*repel*) further increases the precision by a large margin.

2x2	4x4	8x8	16x16	Aggregation	mIoU (%)
✓				-	49.9
	✓			-	51.6
		✓		-	52.9
			✓	-	53.1
✓	✓	✓	✓	$\mathcal{A}$	53.3
✓	✓	✓	✓	$\mathcal{G}_I$	51.3
✓	✓	✓	✓	$\mathcal{G}$	**54.2**

	Precision(%)	Recall(%)	mIoU(%)
GPP only	66.5	75.6	54.2
Attract (Rigid)	65.1 (-1.4)	76.1 (+0.5)	53.4 (-0.8)
Attract (Soft)	66.6 (+0.1)	77.2 (+1.6)	55.0 (+0.8)
Atrract (Soft)+Repel	68.4 (+1.9)	76.3 (+0.7)	56.0 (+1.8)

* *Attract (Rigid)*: uses rigid classification labels for the masked image.
* *Attract (Soft)*: minimizes the distance between the anchor and positive
* *Attract (Soft)+Repel* : uses repelling loss along with *Attract (Soft)*

Results

Methods	*Backbone*	VOC *val*	VOC *test*	COCO *val*
AffinityNet [2]$_{CVPR18}$	ResNet38	61.7	63.7	-
ICD [15]$_{CVPR20}$	ResNet101	64.1	64.3	-
IRNet [1]$_{CVPR19}$	ResNet50	63.5	64.8	32.6
SSDD [43]$_{ICCV19}$	ResNet38	64.9	65.5	-
SEAM [50]$_{CVPR20}$	ResNet38	64.5	65.7	31.9
Sub-category [7]$_{CVPR20}$	ResNet101	66.1	65.9	-
CONTA [57]$_{NIPS20}$	ResNet38	66.1	66.7	33.4
RRM [56]$_{AAAI20}$	ResNet101	66.3	66.5	-
BES [10]$_{ECCV20}$	ResNet101	65.7	66.6	-
CDA [44]$_{ICCV21}$	ResNet38	66.1	66.8	-
ECS [46]$_{ICCV21}$	ResNet38	66.6	67.6	-
AdvCAM [30]$_{CVPR21}$	ResNet101	68.1	68.0	-
OC-CSE[27]$_{ICCV21}$	ResNet38	68.4	68.2	36.4
CPN [58]$_{ICCV21}$	ResNet38	67.8	68.5	-
RIB [28]$_{NeurIPS21}$	ResNet101	68.3	68.6	43.8
PMM [35]$_{ICCV21}$	ResNet38	68.5	69.0	36.7
Ours	ResNet38	**70.9**	**71.7**	**44.8**