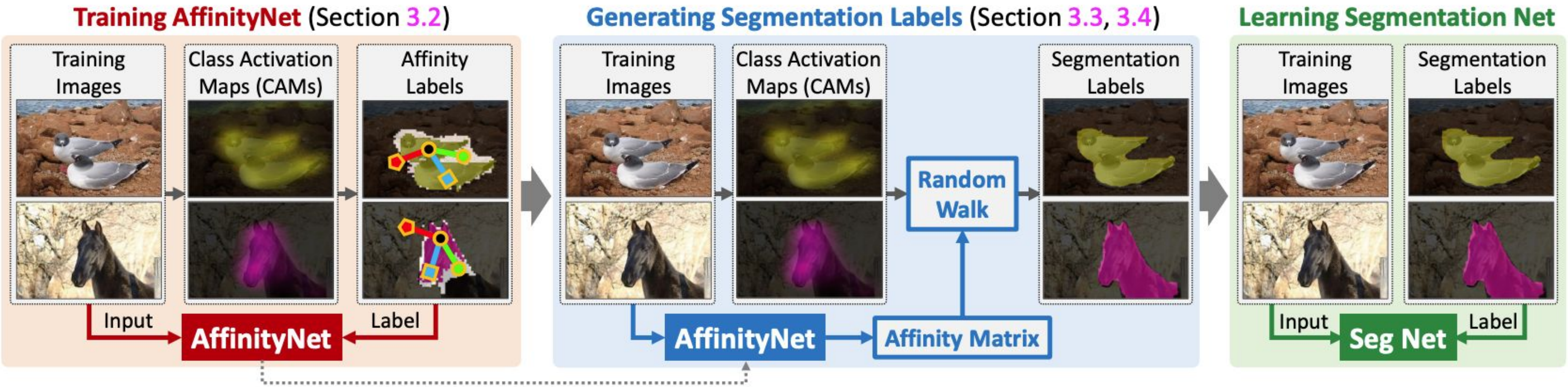
Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation

CVPR’18

Junsoo Lee

19.09.30

Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation, CVPR'18



- Image-level label setting -> weakly supervised for semantic segmentation.
- Learning pixel-level affinities which encourage random work to propagate the activations to nearby and semantically identical areas, and penalize propagation to area of the other classes.

Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation, CVPR'18

1. Computing CAMs

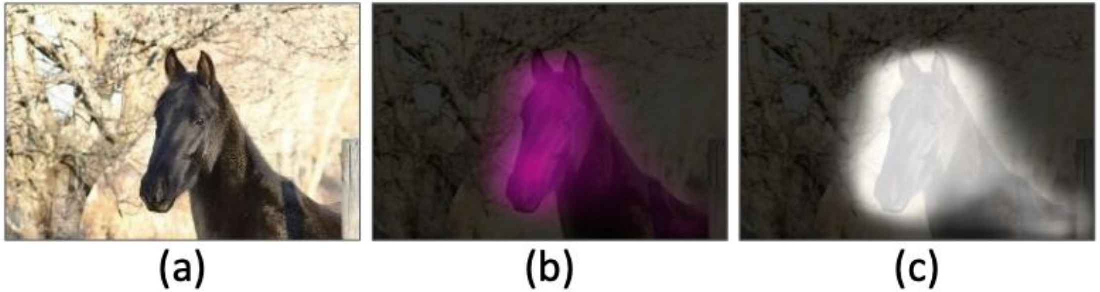


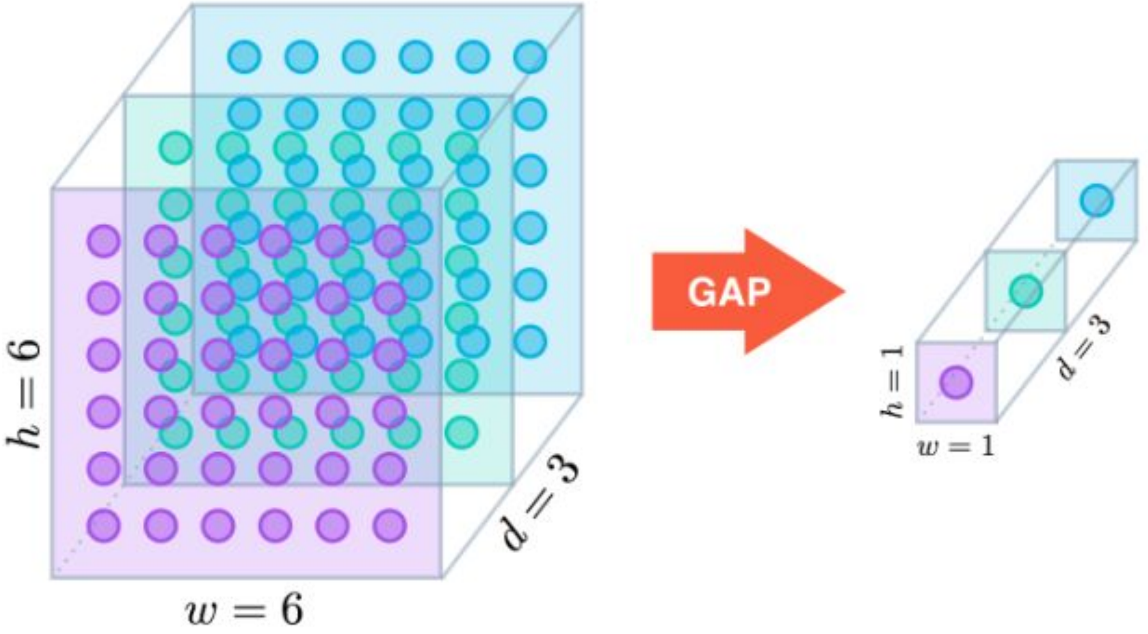
Figure 2. Visualization of CAMs obtained by our approach. (a) Input image. (b) CAMs of object classes: Brighter means more confident object region. (c) CAMs of background: Darker means more confident background region.

$$M_c(x, y) = \mathbf{w}_c^\top f^{\text{cam}}(x, y),$$

$$M_{\text{bg}}(x, y) = \left\{1 - \max_{c \in C} M_c(x, y)\right\}^\alpha$$

Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation, CVPR'18

1. Computing CAMs

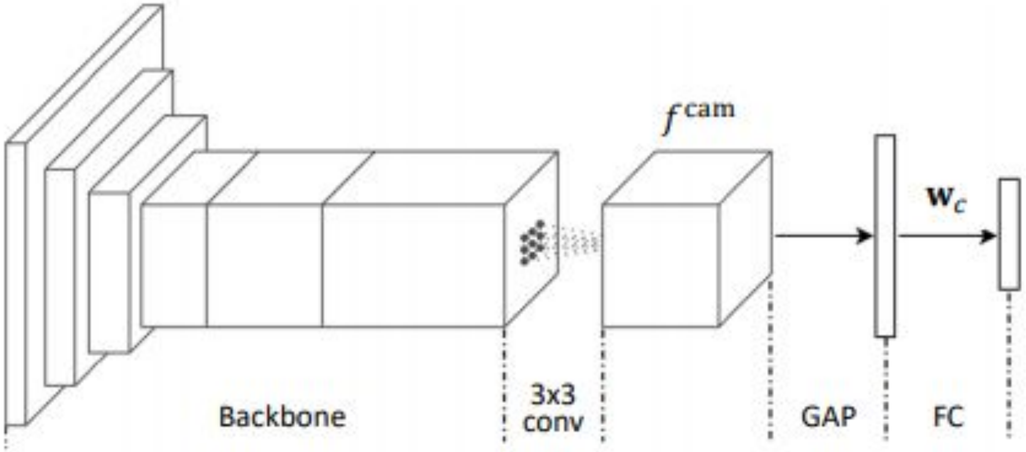


$$M_c(x, y) = \mathbf{w}_c^\top f^{\text{cam}}(x, y),$$

$$M_{\text{bg}}(x, y) = \left\{1 - \max_{c \in C} M_c(x, y)\right\}^\alpha$$

Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation, CVPR'18

1. Computing CAMs



(c) Our network for computing CAMs

$$M_c(x, y) = \mathbf{w}_c^\top f^{\text{cam}}(x, y),$$

$$M_{\text{bg}}(x, y) = \left\{1 - \max_{c \in C} M_c(x, y)\right\}^\alpha$$

Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation, CVPR'18

2. Generating Semantic Affinity Labels

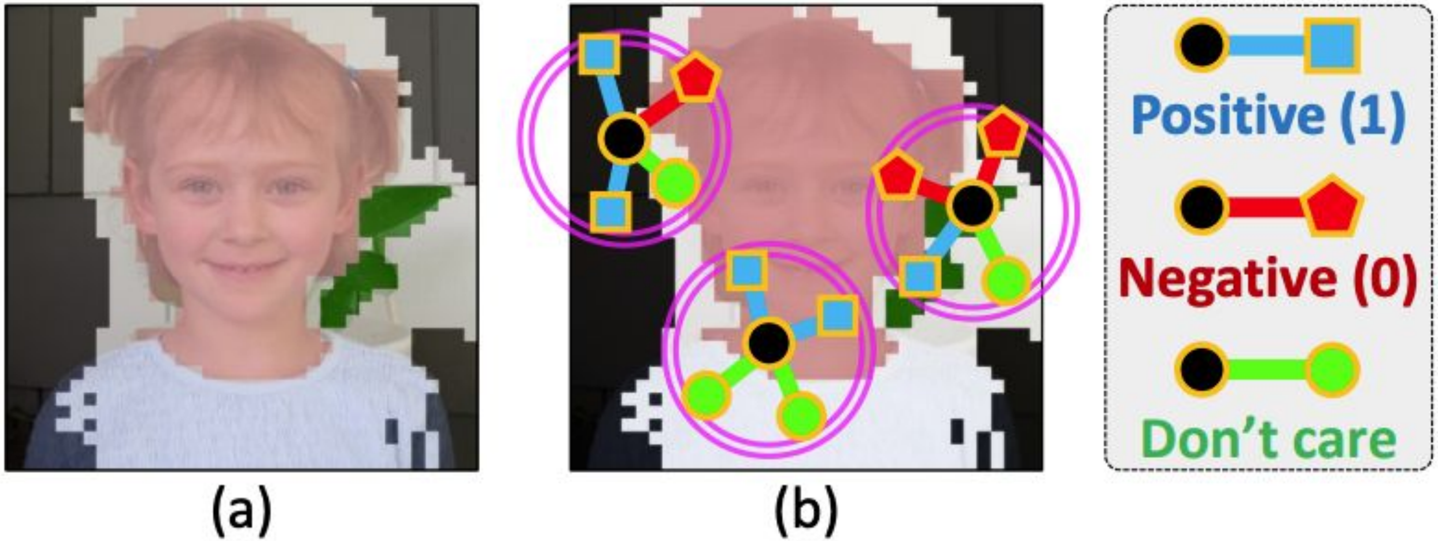


Figure 4. Conceptual illustration of generating semantic affinity labels. (a) Confident areas of object classes and background:

Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation, CVPR'18

3. AffinityNet Training

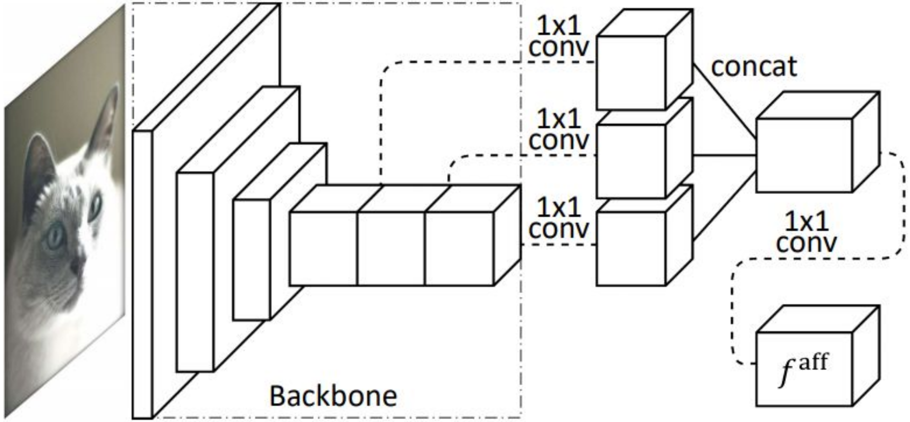


Figure 3. Overall architecture of AffinityNet. The output feature map f^{aff} is obtained by aggregating feature maps from multiple levels of a backbone network so that f^{aff} can take semantic information at various field-of-views. Specifically, we first apply 1×1

$$W_{ij} = \exp\Big\{ - \big\| f^{\text{aff}}(x_i, y_i) - f^{\text{aff}}(x_j, y_j) \big\|_1 \Big\},$$

$$\mathcal{P} = \big\{ (i,j) \mid \mathrm{d}\big((x_i, y_i), (x_j, y_j)\big) < \gamma, \forall i \neq j \big\}$$

$$\begin{aligned} \mathcal{P}^+ &= \big\{ (i,j) \mid (i,j) \in \mathcal{P}, W^*_{ij} = 1 \big\}, \\ \mathcal{P}^- &= \big\{ (i,j) \mid (i,j) \in \mathcal{P}, W^*_{ij} = 0 \big\}, \end{aligned}$$

her break $\mathcal{P}^+$ into $\mathcal{P}^+_{\text{fg}}$ and $\mathcal{P}^+_{\text{bg}}$ for objects a

Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation, CVPR’18

4. Revising CAMs using AffinityNet

$$T = D^{-1} W^{\circ\beta}, \;\; \text{where} \;\; D_{ii} = \sum_j W_{ij}^{\beta}.$$

$$\text{vec}(M_c^*) = T^t \cdot \text{vec}(M_c) \quad \forall c \in C \cup \{\text{bg}\},$$

ere $\text{vec}(\cdot)$ means vectorization of a matrix, and

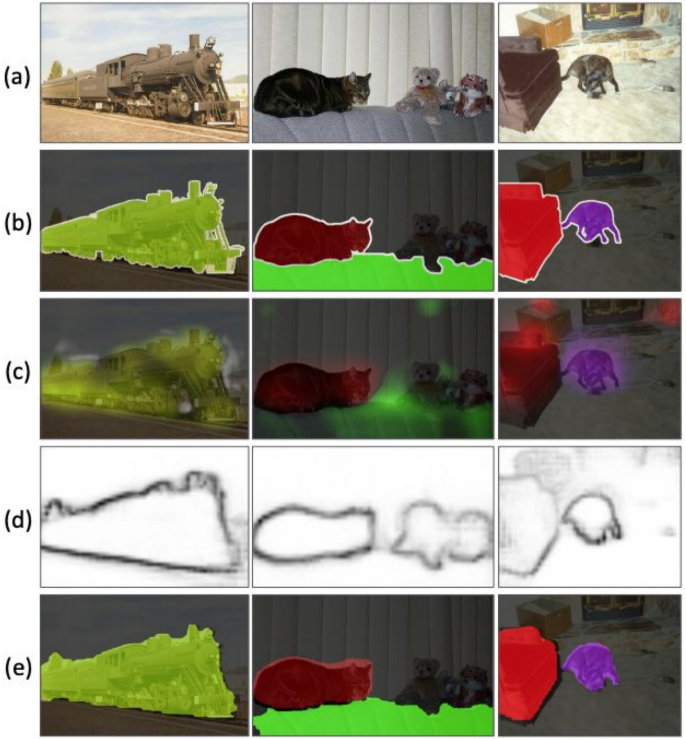
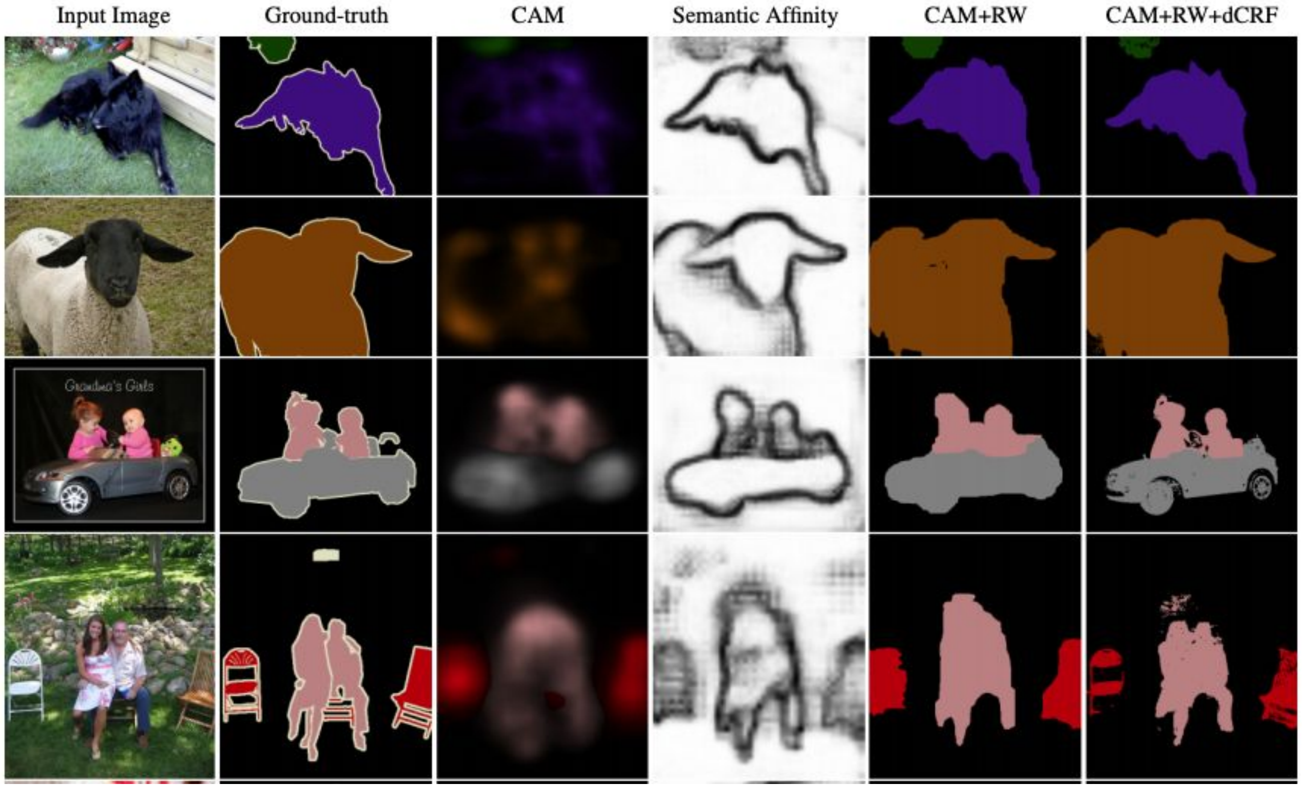


Figure 5. Qualitative examples of synthesized segmentation labels of training images in the PASCAL VOC 2012 benchmark. (a) Input images. (b) Groundtruth segmentation labels. (c) CAMs of object classes. (d) Visualization of the predicted semantic affinities. (e) Synthesized segmentation annotations.

- the Hadamard power of the original affinity matrix, ignores immaterial affinities in W.
- Using this technique makes out random walk propagation more conservative.

Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation, CVPR’18

5. Qualitative Results



Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation, CVPR'18

6. Quantitative Results

Method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mean
EM-Adapt [28]	67.2	29.2	17.6	28.6	22.2	29.6	47.0	44.0	44.2	14.6	35.1	24.9	41.0	34.8	41.6	32.1	24.8	37.4	24.0	38.1	31.6	33.8
CCNN [29]	68.5	25.5	18.0	25.4	20.2	36.3	46.8	47.1	48.0	15.8	37.9	21.0	44.5	34.5	46.2	40.7	30.4	36.3	22.2	38.8	36.9	35.3
MIL+seg [30]	79.6	50.2	21.6	40.9	34.9	40.5	45.9	51.5	60.6	12.6	51.2	11.6	56.8	52.9	44.8	42.7	31.2	55.4	21.5	38.8	36.9	42.0
SEC [14]	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	**75.2**	22.1	53.5	28.3	65.8	57.8	62.3	52.5	32.5	62.6	32.1	45.4	45.3	50.7
AdvErasing [37]	83.4	**71.1**	30.5	72.9	41.6	55.9	63.1	60.2	74.0	18.0	**66.5**	32.4	71.7	56.3	64.8	52.4	37.4	69.1	31.4	58.9	43.9	55.0
Ours-DeepLab	87.2	57.4	25.6	69.8	45.7	53.3	76.6	**70.4**	74.1	28.3	63.2	**44.8**	75.6	**66.1**	65.1	71.1	40.5	66.7	37.2	58.4	49.1	58.4
Ours-ResNet38	**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**

Table 2. Performance on the PASCAL VOC 2012 *val* set, compared to weakly supervised approaches based only on image-level labels.

Method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mean
EM-Adapt [28]	76.3	37.1	21.9	41.6	26.1	38.5	50.8	44.9	48.9	16.7	40.8	29.4	47.1	45.8	54.8	28.2	30.0	44.0	29.2	34.3	46.0	39.6
CCNN [29]	70.1	24.2	19.9	26.3	18.6	38.1	51.7	42.9	48.2	15.6	37.2	18.3	43.0	38.2	52.2	40.0	33.8	36.0	21.6	33.4	38.3	35.6
MIL+seg [30]	78.7	48.0	21.2	31.1	28.4	35.1	51.4	55.5	52.8	7.8	56.2	19.9	53.8	50.3	40.0	38.6	27.8	51.8	24.7	33.3	46.3	40.6
SEC [14]	83.5	56.4	28.5	64.1	23.6	46.5	70.6	58.5	71.3	23.2	54.0	28.0	68.1	62.1	70.0	55.0	38.4	58.0	39.9	38.4	48.3	51.7
AdvErasing [37]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	55.7
Ours-DeepLab	88.0	61.1	29.2	73.0	40.5	54.1	75.2	**70.4**	**75.1**	27.8	62.5	51.4	78.4	**68.3**	76.2	71.8	40.7	**74.9**	**49.2**	55.0	48.3	60.5
Ours-ResNet38	**89.1**	**70.6**	**31.6**	**77.2**	**42.2**	**68.9**	**79.1**	66.5	74.9	**29.6**	**68.7**	**56.1**	**82.1**	64.8	**78.6**	**73.5**	**50.8**	70.7	47.7	**63.9**	**51.1**	**63.7**

Table 3. Performance on the PASCAL VOC 2012 *test* set, compared to weakly supervised approaches based only on image-level labels.

Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation, CVPR'18

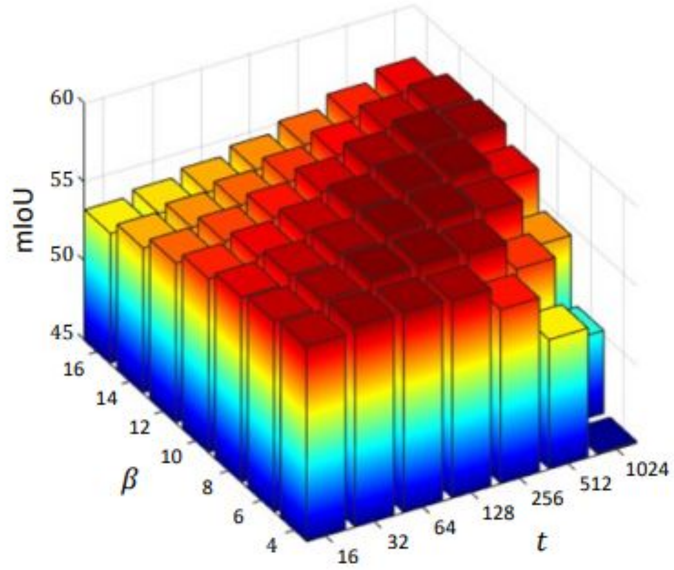
6. Quantitative Results

Method	Sup.	Extra Data	*val*	*test*
TransferNet [10]	$\mathcal{I}$	MS-COCO [20]	52.1	51.2
Saliency [26]	$\mathcal{I}$	MSRA [21], BSDS [24]	55.7	56.7
MCNN [35]	$\mathcal{I}$	YouTube-Object [31]	38.1	39.8
CrawlSeg [11]	$\mathcal{I}$	YouTube Videos	58.1	58.7
What'sPoint [1]	$\mathcal{P}$	-	46.0	43.6
RAWK [36]	$\mathcal{S}$	-	61.4	-
ScribbleSup [18]	$\mathcal{S}$	-	63.1	-
WSSL [28]	$\mathcal{B}$	-	60.6	62.2
BoxSup [6]	$\mathcal{B}$	-	62.0	64.6
SDI [12]	$\mathcal{B}$	BSDS [24]	65.7	67.5
FCN [22]	$\mathcal{F}$	-	-	62.2
DeepLab [3]	$\mathcal{F}$	-	67.6	70.3
ResNet38 [38]	$\mathcal{F}$	-	80.8	82.5
Ours-DeepLab	$\mathcal{I}$	-	58.4	60.5
Ours-ResNet38	$\mathcal{I}$	-	61.7	63.7

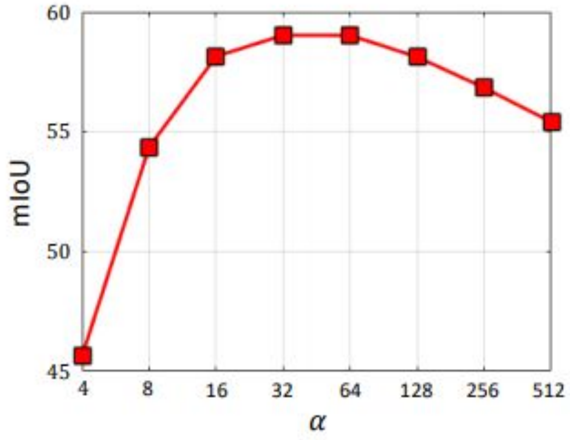
Table 4. Performance on the PASCAL VOC 2012 *val* and *test* sets. The supervision types (Sup.) indicate: $\mathcal{P}$–point, $\mathcal{S}$–scribble, $\mathcal{B}$–bounding box, $\mathcal{I}$–image-level label, and $\mathcal{F}$–segmentation label.

Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation, CVPR'18

7. Analysis on Effects of the Hyper-parameters



(a) Accuracy versus β and t



(b) Accuracy versus α

Figure 8. Accuracy (mIoU) of segmentation labels synthesized by CAM+RW for different hyper-parameter values on the VOC 2012 *train*.