Adaptive Affinity Fields for Semantic Segmentation

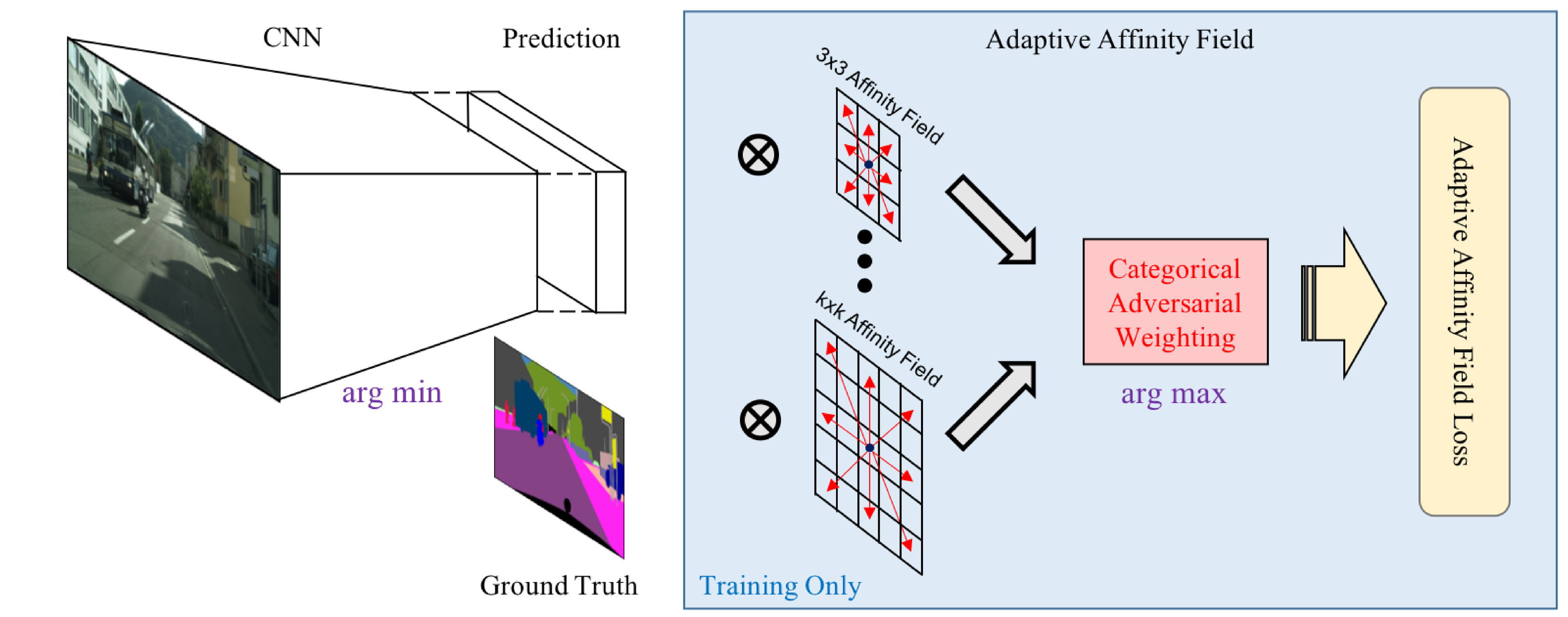
Tsung-Wei Ke*, Jyh-Jing Hwang*, Ziwei Liu, Stella X. Yu

Berkeley UNIVERSITY OF CALIFORNIA

ICSI

Code

Overview



- Existing methods often use per-pixel supervision and ignore label correlations among pixels.
- Our method captures and matches semantic relations between pixels in the label space.
- Effective representation: Encode spatial structures in distributed, pixel-centric relations.
- Efficient computation: 2 hyper-parameters only, zero overhead during inference.
- Accurate segmentation: Get details right and generalize to visual appearance changes.

Method	Structure Guidance	Training	Run-time Inference	Performance
CRF [15]	input image	medium	yes	76.53
GAN [12]	ground-truth labels	hard	no	76.20
Our AAF	label affinity	easy	no	**79.24**

Table 1. Performance (% mIoU) is reported with PSPNet on Cityscapes validation set.

- Pixel-wise labeling loss:

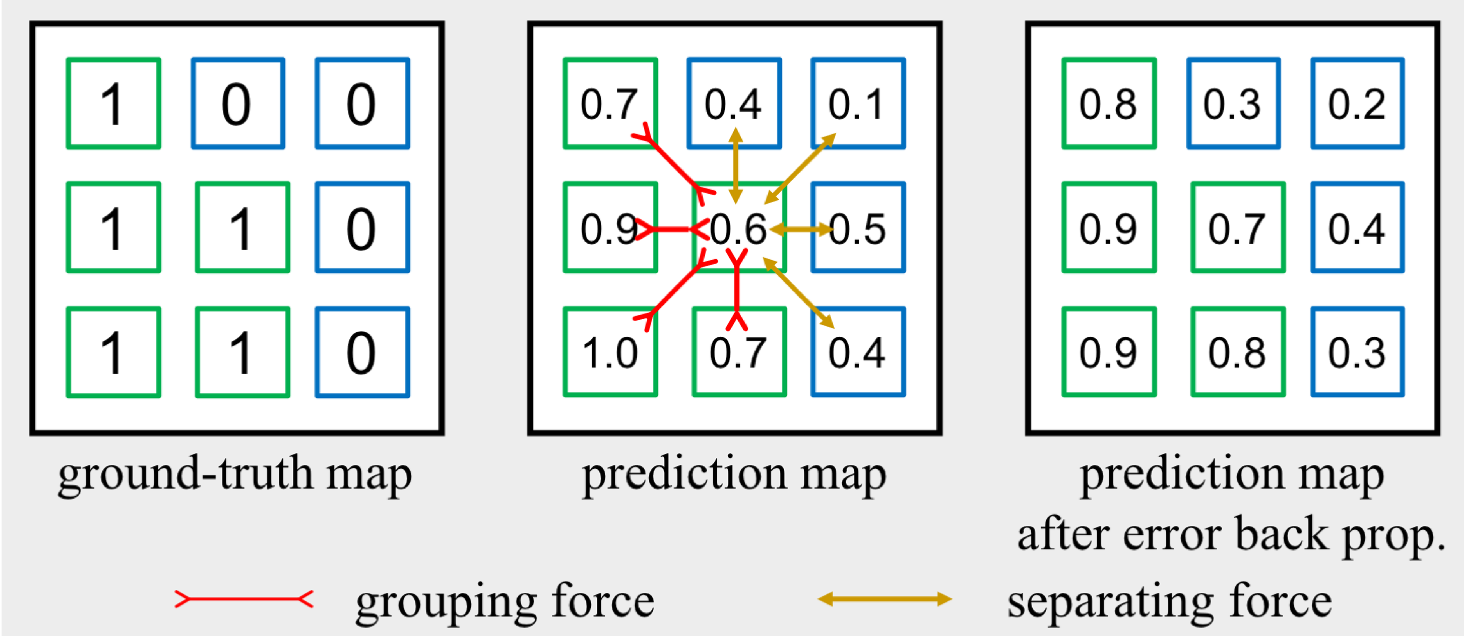
$$\mathcal{L}^i_{\text{unary}} = \mathcal{L}^i_{\text{cross-entropy}} = -\log \hat{y}_i(l).$$

- Region-wise labeling loss:

$$\mathcal{L}^i_{\text{unary}}(\hat{y}_i, y_i) + \lambda \mathcal{L}^i_{\text{region}}(\mathcal{N}(\hat{y}_i), \mathcal{N}(y_i))$$

- CRF: label consistency btw visually similar pixels
- GAN: structure priors in label predictions
- Our AAF: pixel-centric label pattern similarity

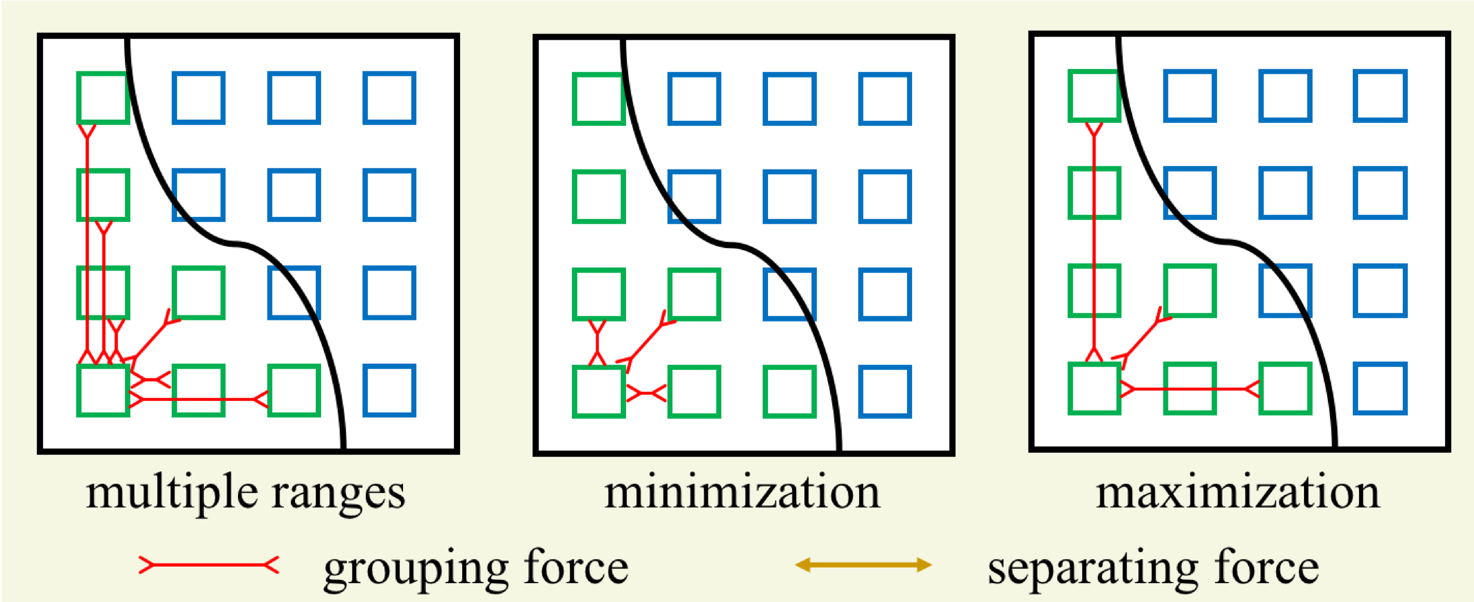
Our Model



- Pixels of same / diff. gt labels desire same / diff predictions regardless of actual label values.
- Our affinity field loss:

$$\mathcal{L}^{ic}_{\text{affinity}} = \begin{cases} \mathcal{L}^{i\bar{b}c}_{\text{affinity}} = D_{KL}(\hat{y}_j(c) \| \hat{y}_i(c)) & \text{if } y_i(c) = y_j(c) \\ \mathcal{L}^{ibc}_{\text{affinity}} = \max\{0, m - D_{KL}(\hat{y}_j(c) \| \hat{y}_i(c))\} & \text{otherwise} \end{cases}$$

- No message passing or iterative refinement.

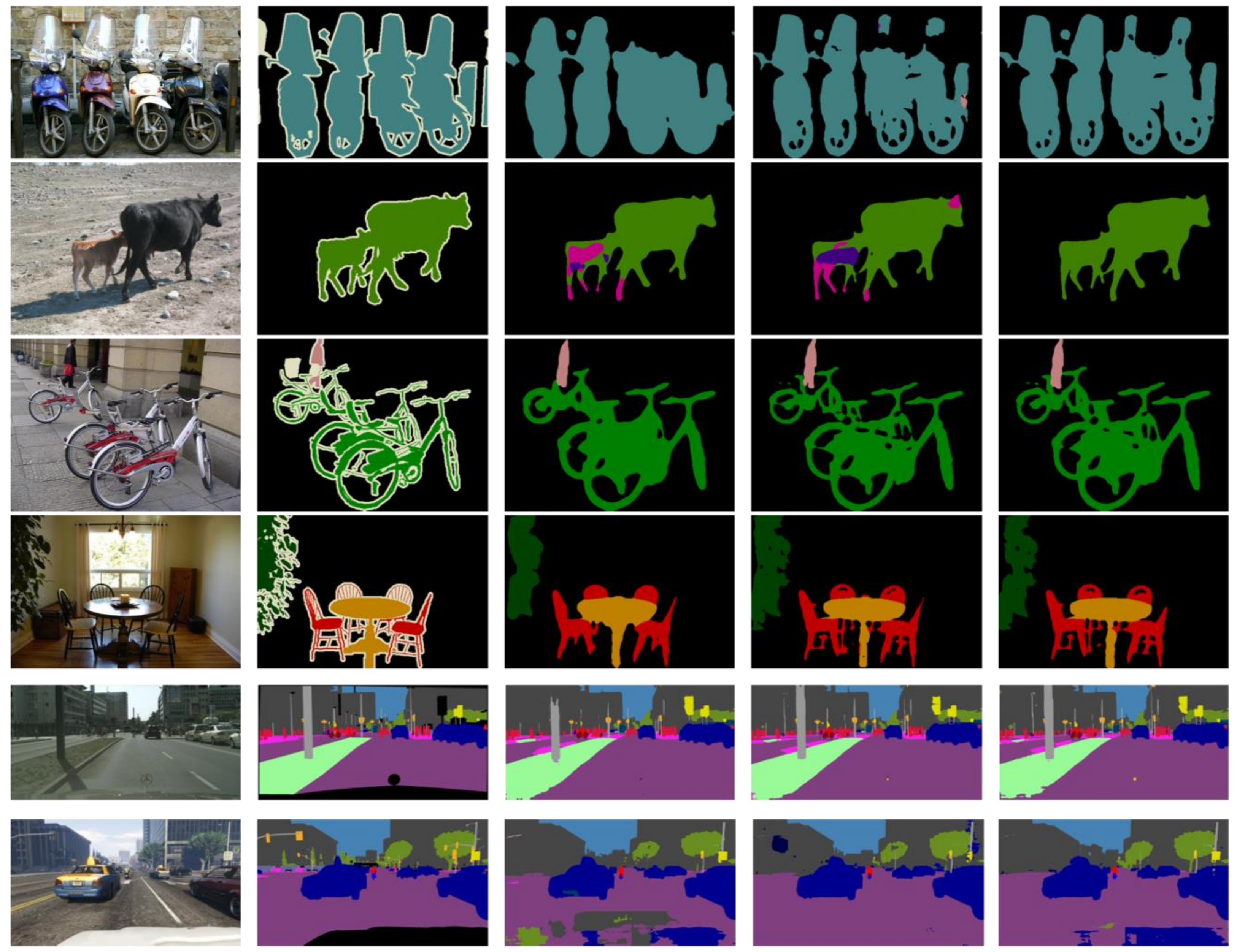


- Affinity fields in small / large neighborhoods encode near / long range structural relations.
- One size does not fit all classes; picking the one with minimal affinity loss results in trivial solutions.
- Select the right size by pushing the affinity field matching to the hard negative cases.
- Our adversarial learning for adaptive kernel sizes:

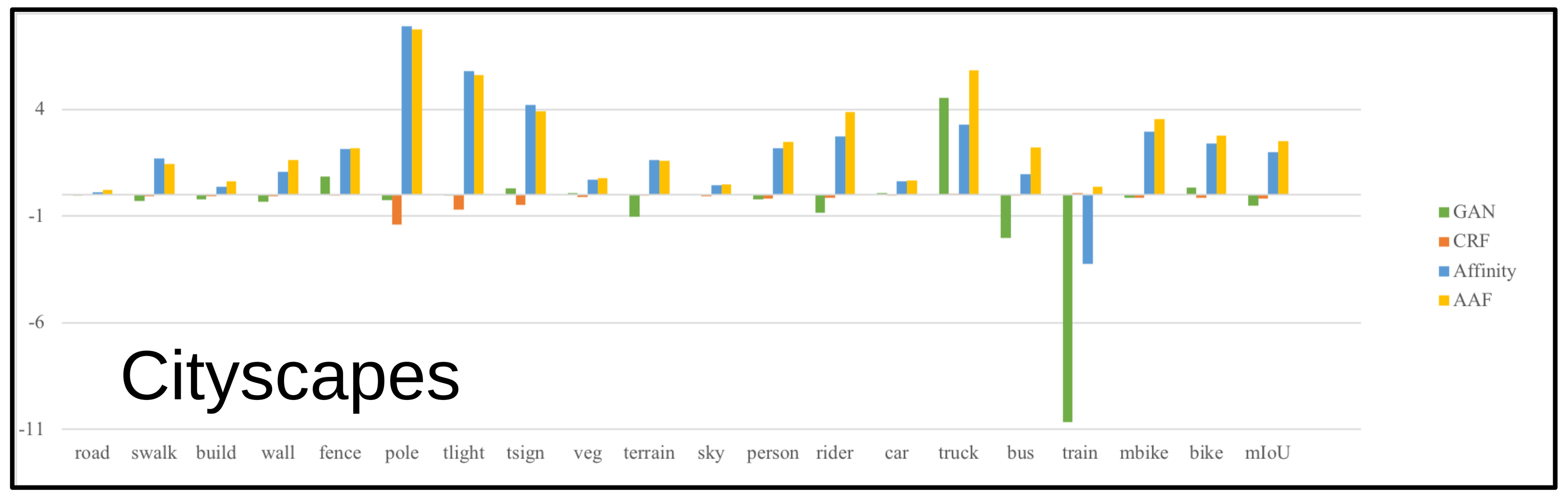
$$\mathcal{L}_{\text{multiscale}} = \sum_c \sum_k w_{ck} \mathcal{L}^{ck}_{\text{region}} \quad \text{s.t.} \sum_k w_{ck} = 1$$

$$S^* = \underset{S}{\operatorname{argmin}} \max_{w} \mathcal{L}_{\text{unary}} + \mathcal{L}_{\text{multiscale}}.$$

Our Results



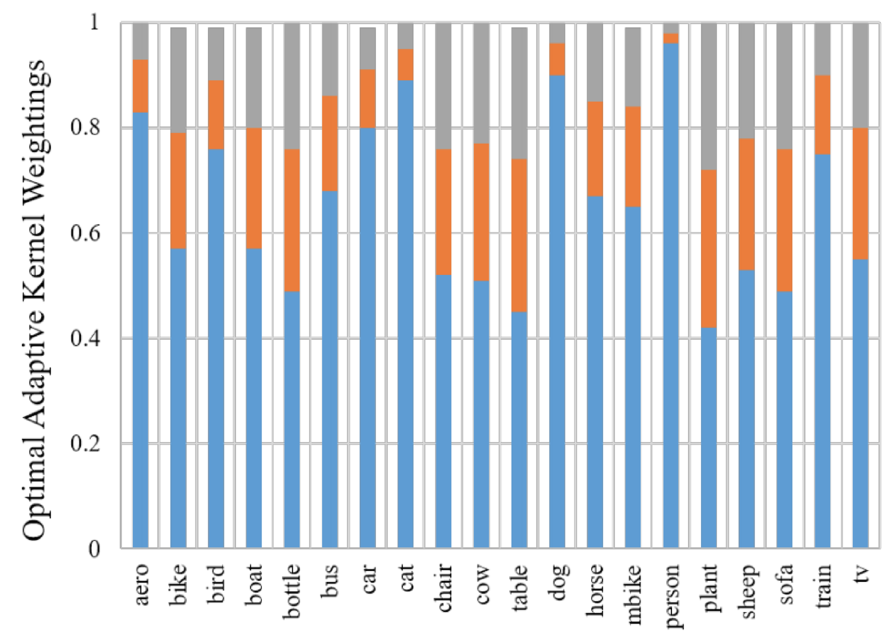
(a) Image (b) Ground Truth (c) PSPNet (d) Affinity (e) AAF

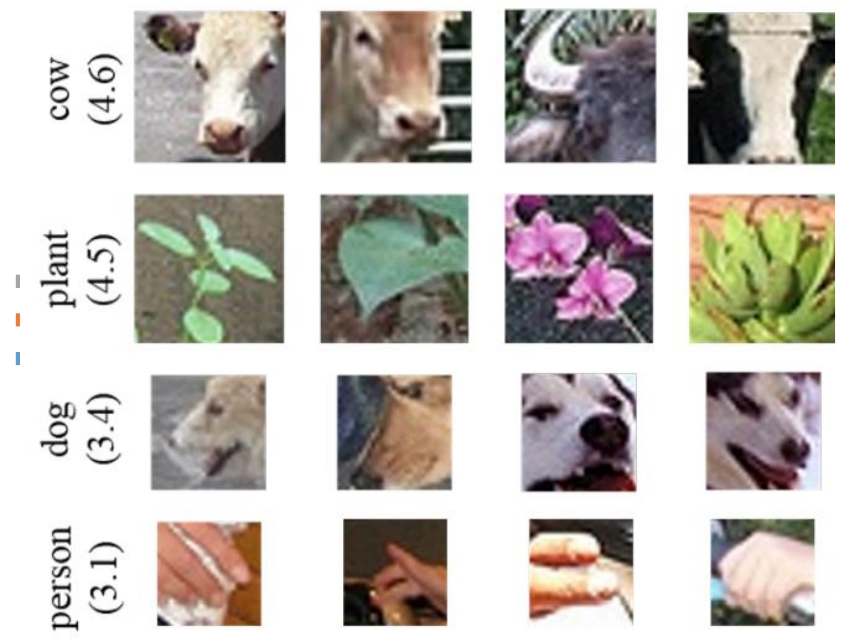


- More accurate segmentation on PASCAL VOC and Cityscapes

Method	road	swalk	build.	wall	fence	pole	tlight	tsign	veg.	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU	pix. acc
PSPNet	61.79	**34.26**	37.30	**13.31**	18.52	26.51	31.64	17.51	55.00	8.57	**82.47**	42.73	**49.78**	**69.25**	**34.31**	18.21	**25.00**	33.14	6.86	35.06	68.78
Affinity	75.26	30.34	44.10	12.91	**20.19**	**29.78**	31.50	23.98	**64.25**	11.83	74.32	48.28	49.12	67.39	25.76	**23.82**	20.29	**41.48**	5.63	**36.86**	75.13
AAF	**83.07**	27.82	**51.16**	10.41	18.76	28.58	**31.74**	**24.98**	61.38	**12.25**	70.65	**50.53**	48.06	53.35	26.80	20.97	24.50	39.56	**9.37**	36.52	**78.28**

- Better generalization: Trained on Cityscapes, tested on GTA5





- Learned kernel sizes for different classes