Domain Adaptation for Semantic Segmentation with Maximum Squares Loss

Minghao Chen, Hongyang Xue, Deng Cai*

State Key Lab of CAD&CG, College of Computer Science, Zhejiang University, Hangzhou, China
Fabu Inc., Hangzhou, China
Alibaba-Zhejiang University Joint Institute of Frontier Technologies, Hangzhou, China
minghaochen01@gmail.com, hyxue@outlook.com, dengcai@cad.zju.edu.cn

ICCV 2019

Introduction

- Real-world datasets with pixel-wise semantic labels demand an enormous amount of manual annotation work.
- Because of this "**curse of dataset annotation**", real-world datasets for semantic segmentation often contain only a small number of samples.
- Utilize synthetic datasets - the model trained on the synthetic dataset cannot generalize well to real-world examples via direct transfer.
- Unsupervised domain adaptation (UDA):
  - labeled synthetic dataset - source domain
  - unlabeled real-world dataset - target domain
  - utilize the unlabeled data from the target domain to help minimize the performance gap between these two domains.

UDA

$$\mathcal{D}_S = \{(x_s, y_s) | x_s \in \mathbb{R}^{H \times W \times 3}, y_s \in \mathbb{R}^{H \times W}\} \qquad \mathcal{D}_T = \{x_t | x_t \in \mathbb{R}^{H \times W \times 3}\}$$

Objective: $\mathcal{L}(x_s, x_t) = \mathcal{L}_{CE}(p_s, y_s) + \lambda_T \mathcal{L}_T(x_t)$

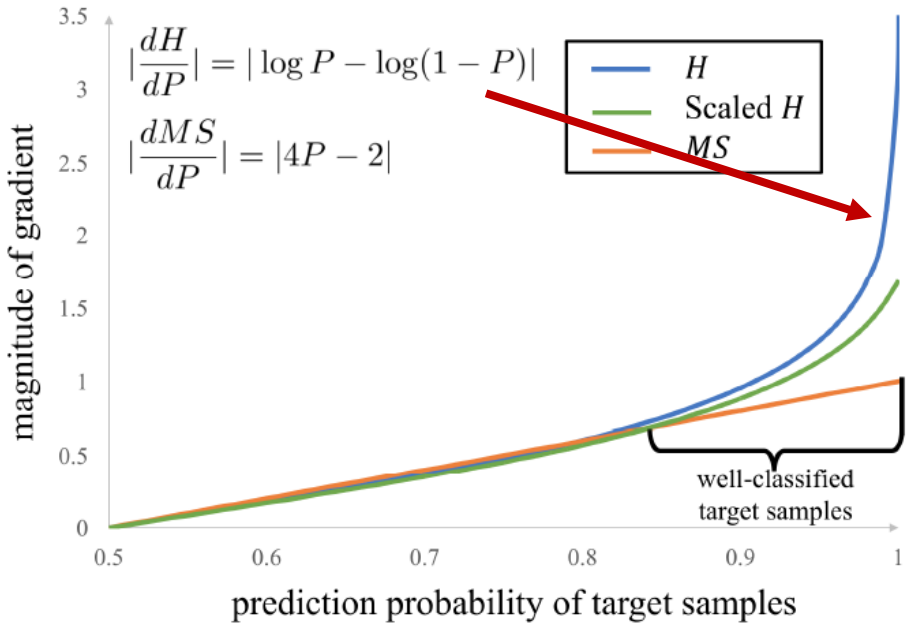
$$\mathcal{L}_{CE}(p_s, y_s) = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} y_s^{n,c} \log(p_s^{n,c})$$

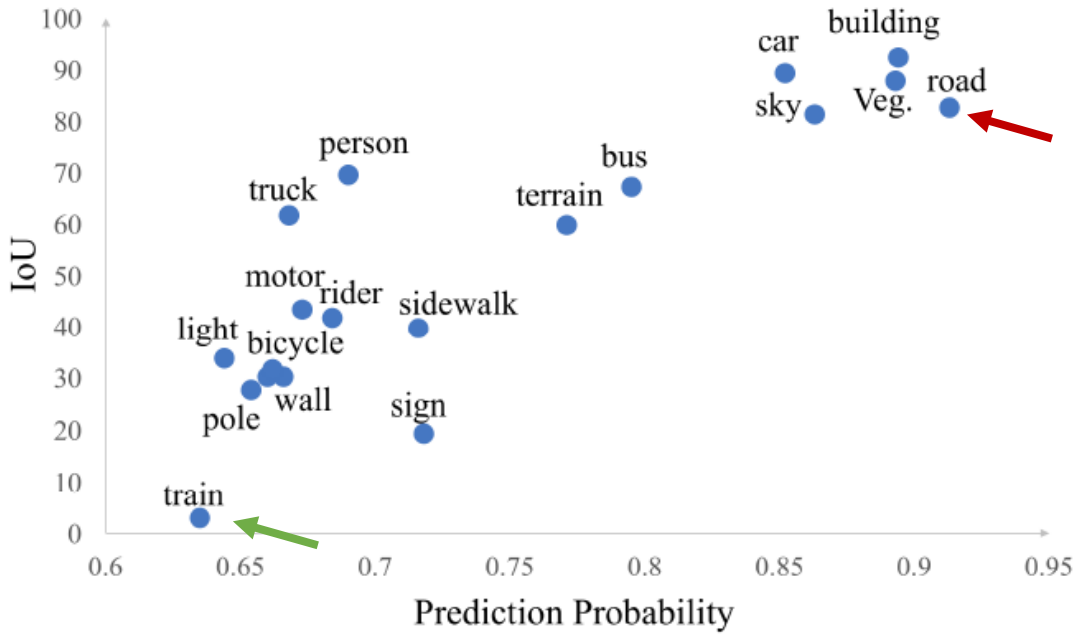
$$\mathcal{L}_T(x_t) = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} p_t^{n,c} \log(p_t^{n,c})$$

Entropy Minimization

$$H(p|x_t) = -p \log p - (1-p) \log(1-p),$$
$$|\frac{dH}{dp}| = |\log p - \log(1-p)|.$$

Probability Imbalance Problem





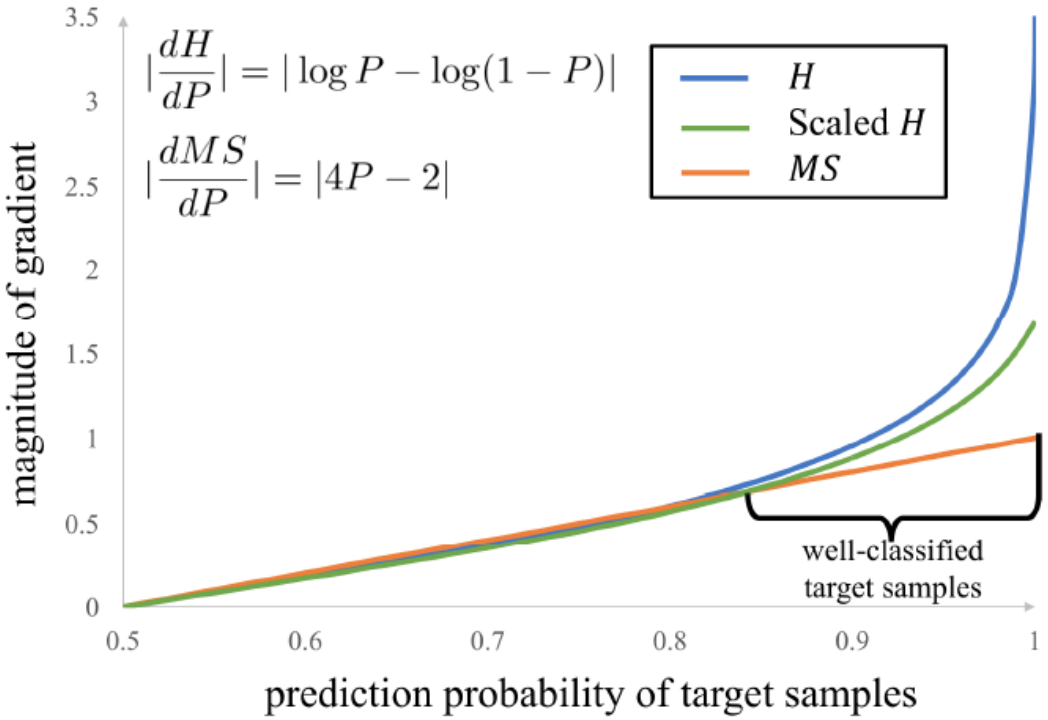
the gradient of the entropy minimization method (H) is focused on well-classified samples in the target domain

From GTA5 to Cityscapes, the mean of prediction probability v.s. Intersection over Union(IoU) for each target class.

Maximum Squares Loss

$$\mathcal{L}_T(x_t) = -\frac{1}{2N}\sum_{n=1}^{N}\sum_{c=1}^{C}(p_t^{n,c})^2$$

$$MS(p|x_t) = -p^2 - (1-p)^2,$$
$$|\frac{dMS}{dp}| = |4p-2|.$$



Interpretation from f-divergence View

- The target part loss $\mathcal{L}_T(x_t)$ can be treated as the distance between the model prediction distribution $p^{n,c}$ and uniform distribution: $\mathcal{U} = \frac{1}{C}$

Pearson χ^2 divergence: $f(t) = t^2 - 1$

max $$D_{\chi^2}(p^{n,c} \| \mathcal{U}) = C \sum_c (p^{n,c})^2 - 1$$

min $$\mathcal{L}_T(x_t) = -\frac{1}{2N} \sum_{n=1}^{N} \sum_{c=1}^{C} (p_t^{n,c})^2$$

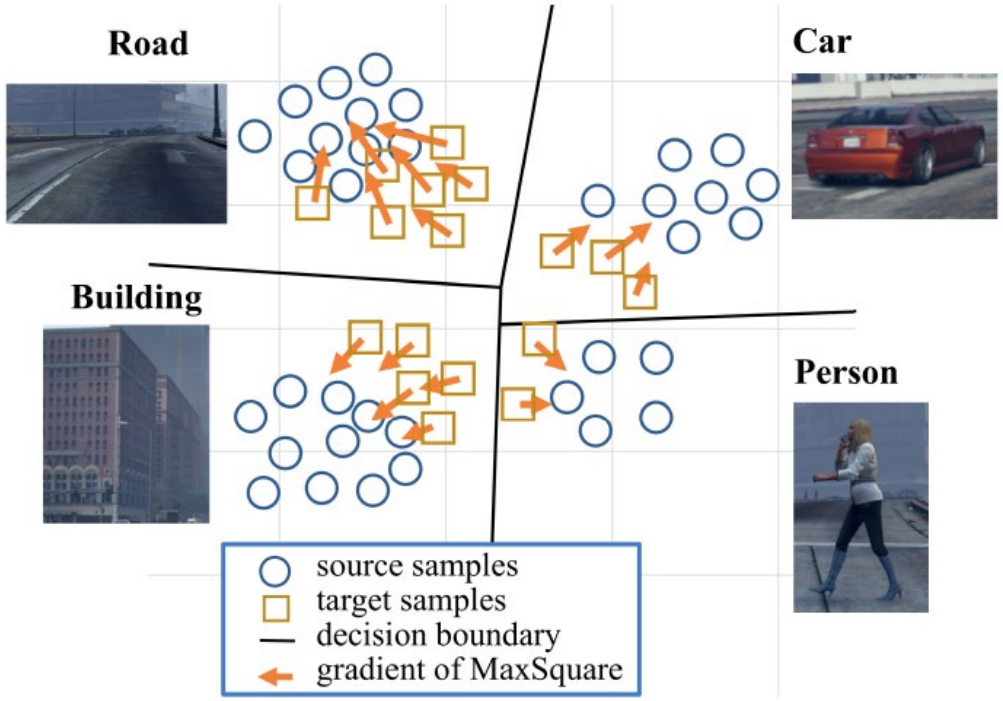


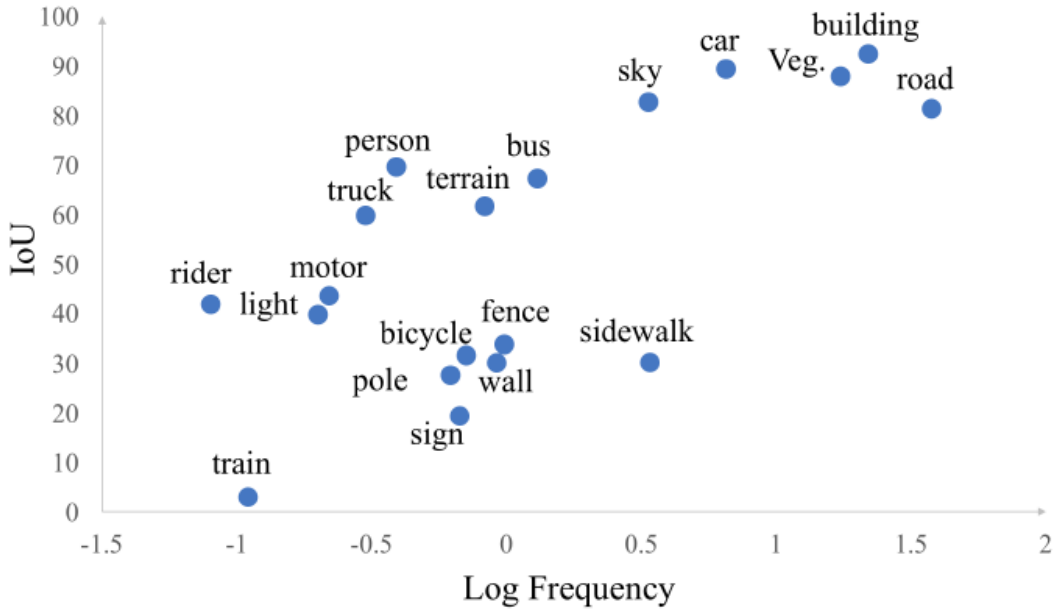
Image-wise Class-balanced Weighting Factor

classes with higher accuracy always have more pixels on the label map, which leads to an imbalance in quantity

$$m^{n,c^*} = \begin{cases} 1 & \text{if } c^* = \arg\max_c p^{n,c} \\ 0 & \text{otherwise}, \end{cases}$$

$$N^c = \sum_n m^{n,c}.$$

$$\mathcal{L}_T(x_t) = -\sum_{n=1}^{N}\sum_{c=1}^{C} \frac{1}{2(N^c)^{\alpha} \times N^{(1-\alpha)}} (p_t^{n,c})^2$$



Experiment

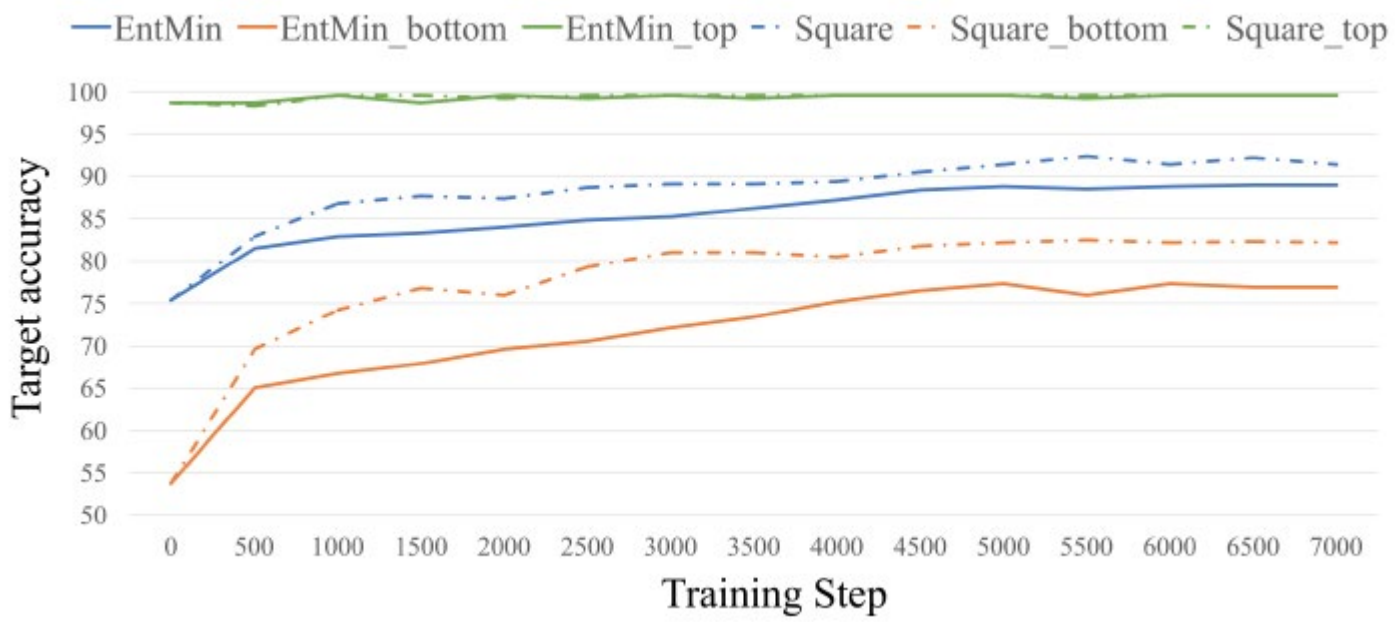
- Classification:
  - Datasets: Amazon (A), Webcam (W) and DSLR (D)
  - Transfer tasks: A->W, D->W, W->D, A->D, D->A and W->A.
  - Model: ResNet50 pre-trained on ImageNet

- Segmentation:
  - synthetic datasets to real-world:
    GTA5->Cityscapes, SYNTHIA->Cityscapes
  - cross-city adaptation: Cityscapes->NTHU
  - Model: Deeplabv2 with ResNet-101 backbones pre-trained on ImageNet

Classification

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
ResNet-50 [12]	68.4±0.2	96.7±0.1	99.3±0.1	68.9±0.2	62.5±0.3	60.7±0.3	76.1
DANN [10]	82.0±0.4	96.9±0.2	99.1±0.1	79.7±0.4	**68.2±0.4**	**67.4±0.5**	82.2
EntMin	89.0±0.1	99.0±0.1	**100.0±.0**	86.3±0.3	67.5±0.2	63.0±0.1	84.1
MaxSquare	**92.4±0.5**	**99.1±0.1**	**100.0±.0**	**90.0±0.2**	68.1±0.4	64.2±0.2	**85.6**

A->W



Segmentation

GTA5→Cityscapes																				
Method	Backbone	road	sidewalk	building	wall	fence	pole	light	sign	veg.	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU (%)
Source only [36]	Wider	70.0	23.7	67.8	15.4	18.1	40.2	41.9	25.3	78.8	11.7	31.4	62.9	29.8	60.1	21.5	26.8	7.7	28.1	12.0	35.4
CBST [36]	ResNet-38	86.8	46.7	76.9	26.3	24.8	42.0	46.0	38.6	80.7	15.7	48.0	57.3	27.9	78.2	24.5	49.6	17.7	25.5	45.1	45.2
CBST-SP [36]	[32]	88.0	**56.2**	77.0	27.4	22.4	**40.7**	**47.3**	**40.9**	82.4	21.6	60.3	50.2	20.4	83.8	35.0	**51.0**	**15.2**	20.6	37.0	46.2
AdaptSegNet [28]		86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
MinEnt [31]	ResNet101	86.2	18.6	80.3	27.2	24.0	23.4	33.5	24.7	83.3	31.0	75.6	54.6	25.6	**85.2**	30.0	10.9	0.1	21.9	**37.1**	42.3
AdvEnt+MinEnt [31]		87.6	21.4	82.0	**34.8**	**26.2**	28.5	35.6	23.0	84.5	35.1	76.2	58.6	**30.7**	84.8	34.2	43.4	0.4	28.4	35.3	44.8
Source only		71.4	15.3	74.0	21.1	14.4	22.8	33.9	18.6	80.7	20.9	68.5	56.6	27.1	67.4	32.8	5.6	7.7	28.4	33.8	36.9
MinEnt†		84.2	34.4	80.7	27.0	15.7	25.8	32.6	18.0	83.4	29.4	76.9	58.7	24.0	78.7	35.9	29.9	6.5	28.3	31.4	42.2
MaxSquare	ResNet101	88.1	27.7	80.8	28.7	19.8	24.9	34.0	17.8	83.6	34.7	76.0	58.6	28.6	84.1	37.8	43.1	7.2	32.2	34.2	44.3
MaxSquare+IW		89.3	40.5	81.2	29.0	20.4	25.6	34.4	19.0	83.6	34.4	76.5	59.2	27.4	83.8	**38.4**	43.6	7.1	32.2	32.5	45.2
MaxSquare+IW+Multi		**89.4**	43.0	**82.1**	30.5	21.3	30.3	34.7	24.0	**85.3**	**39.4**	**78.2**	**63.0**	22.9	84.6	36.4	43.0	5.5	**34.7**	33.5	**46.4**

Table 2: Results for GTA5-to-Cityscapes experiments. "MaxSquare" denotes our maximum squares loss method and "MaxSquare+IW" is the maximum squares loss combined with our image-wise weighting factor (Eq. 13). " Multi" denotes combining the multi-level self-guided method in Section 3.4. For comparison, we reproduce the result of entropy minimization method [31], which is denoted as "MinEnt†". CBST [36] adopts a wider ResNet model [32], which is more powerful than the original ResNet [12] that we adopt.

Segmentation

SYNTHIA→Cityscapes																			
Method	Backbone	road	sidewalk	building	wall*	fence*	pole*	light	sign	veg.	sky	person	rider	car	bus	motor	bike	mIoU (%)	mIoU* (%)
Source only [36]	Wider	32.6	21.5	46.5	4.8	0.1	26.5	14.8	13.1	70.8	60.3	56.6	3.5	74.1	20.4	8.9	13.1	29.2	33.6
CBST [36]	ResNet-38	53.6	23.7	75.0	**12.5**	0.3	**36.4**	**23.5**	**26.3**	**84.8**	74.7	**67.2**	17.5	**84.5**	28.4	15.2	**55.8**	**42.5**	**48.4**
AdaptSegNet [28]		84.3	**42.7**	77.5	-	-	-	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	**18.9**	32.3	-	46.7
MinEnt [31]	ResNet101	73.5	29.2	77.1	7.7	0.2	27.0	7.1	11.4	76.7	82.1	57.2	21.3	69.4	29.2	12.9	27.9	38.1	44.2
AdvEnt+MinEnt [31]		**85.6**	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	57.9	**23.8**	73.3	**36.4**	14.2	33.0	41.2	48.0
Source only		17.7	15.0	74.3	10.1	0.1	25.5	6.3	10.2	75.5	77.9	57.1	19.2	31.2	31.2	10.0	20.1	30.1	34.3
MinEnt†		67.8	28.3	79.0	4.8	0.1	24.7	4.0	7.3	81.7	84.1	58.9	19.4	75.9	36.2	10.4	26.1	38.0	44.5
MaxSquare	ResNet101	77.4	34.0	78.7	5.6	0.2	27.7	5.8	9.8	80.7	83.2	58.5	20.5	74.1	32.1	11.0	29.9	39.3	45.8
MaxSquare+IW		78.5	34.7	76.3	6.5	0.1	30.4	12.4	12.2	82.2	**84.3**	59.9	17.9	80.6	24.1	15.2	31.2	40.4	46.9
MaxSquare+IW+Multi		82.9	40.7	**80.3**	10.2	**0.8**	25.8	12.8	18.2	82.5	82.2	53.1	18.0	79.0	31.4	10.4	35.6	**41.4**	**48.2**

Table 3: Results for SYNTHIA-to-Cityscapes experiments.

Segmentation

Cross-City Adaptation															
City	Method	road	sidewalk	building	light	sign	veg.	sky	person	rider	car	bus	motor	bike	mIoU (%)
Rome	Cross city [5]	79.5	29.3	84.5	0.0	22.2	80.6	82.8	29.5	13.0	71.7	37.5	25.9	1.0	42.9
	CBST [36]	**87.1**	**43.9**	**89.7**	14.8	**47.7**	85.4	90.3	45.4	**26.6**	**85.4**	20.5	49.8	**10.3**	53.6
	AdaptSegNet [28]	83.9	34.2	88.3	18.8	40.2	**86.2**	**93.1**	**47.8**	21.7	80.9	47.8	48.3	8.6	53.8
	Source only	85.0	34.7	86.4	17.5	39.0	84.9	85.4	43.8	15.5	81.8	46.3	38.4	4.8	51.0
	MaxSquare	80.0	27.6	87.0	**20.8**	42.5	85.1	92.4	46.7	22.9	82.1	53.5	**50.8**	8.8	53.9
	MaxSquare+IW	82.9	32.6	86.7	20.7	41.6	85.0	93.0	47.2	22.5	82.2	**53.8**	50.5	9.9	**54.5**
Rio	Cross city [5]	74.2	43.9	79.0	2.4	7.5	77.8	69.5	39.3	10.3	67.9	**41.2**	27.9	10.9	42.5
	CBST [36]	**84.3**	**55.2**	85.4	**19.6**	**30.1**	80.5	77.9	55.2	28.6	**79.7**	33.2	37.6	11.5	52.2
	AdaptSegNet [28]	76.2	44.7	84.6	9.3	25.5	81.8	87.3	**55.3**	32.7	74.3	28.9	**43.0**	27.6	51.6
	Source only	74.2	42.2	84.0	12.1	20.4	78.3	87.9	50.1	25.6	76.6	40.0	27.6	17.0	48.9
	MaxSquare	70.9	39.2	**85.6**	14.5	19.7	**81.8**	**88.1**	55.2	31.5	77.2	39.3	43.1	**30.1**	52.0
	MaxSquare+IW	76.9	48.8	85.2	13.8	18.9	81.7	**88.1**	54.9	**34.0**	76.8	39.8	44.1	29.7	**53.3**
Tokyo	Cross city [5]	83.4	**35.4**	72.8	12.3	12.7	77.4	64.3	42.7	21.5	64.1	**20.8**	8.9	40.3	42.8
	CBST [36]	**85.2**	33.6	**80.4**	8.3	**31.1**	**83.9**	78.2	53.2	28.9	**72.7**	4.4	27.0	47.0	48.8
	AdaptSegNet [28]	81.5	26.0	77.8	**17.8**	26.8	82.7	**90.9**	55.8	**38.0**	72.1	4.2	24.5	**50.8**	49.9
	Source only	81.4	28.4	78.1	14.5	19.6	81.4	86.5	51.9	22.0	70.4	18.2	22.3	46.4	47.8
	MaxSquare	79.3	28.5	78.3	14.5	27.9	82.8	89.6	57.3	31.9	71.9	6.0	29.1	49.2	49.7
	MaxSquare+IW	81.2	30.1	77.0	12.3	27.3	82.8	89.5	**58.2**	32.7	71.5	5.5	**37.4**	48.9	**50.5**
Taipei	Cross city [5]	78.6	28.6	80.0	13.1	7.6	68.2	82.1	16.8	9.4	60.4	34.0	26.5	9.9	39.6
	CBST [36]	**86.1**	**35.2**	84.2	15.0	**22.2**	75.6	74.9	22.7	**33.1**	**78.0**	37.6	**58.0**	30.9	50.3
	AdaptSegNet [28]	81.7	29.5	85.2	26.4	15.6	76.7	91.7	31.0	12.5	71.5	41.1	47.3	27.7	49.1
	Source only	82.6	33.0	**86.3**	16.0	16.5	78.3	83.3	26.5	8.4	70.7	36.1	47.9	15.7	46.3
	MaxSquare	81.2	32.8	85.4	**31.9**	14.7	**78.3**	**92.7**	28.3	8.6	68.2	42.2	51.3	32.4	49.8
	MaxSquare+IW	80.7	32.5	85.5	32.7	15.1	78.1	91.3	**32.9**	7.6	69.5	**44.8**	52.4	**34.9**	**50.6**

Table 6: Results for Cross-City experiments.

Conclusion

- Demonstrate the **probability imbalance** problem when applying the entropy minimization method to UDA for semantic segmentation.
- Propose **maximum squares loss:** prevent easy-to-transfer classes from dominating the training on the target domain.
- For class imbalance in the target domain, propose **class weighting factor**, based on the prediction quantity of each class.