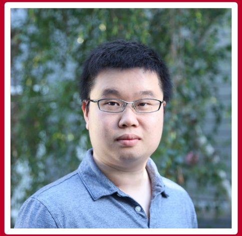
Confidence Regularized Self-Training


Yang Zou


Zhiding Yu

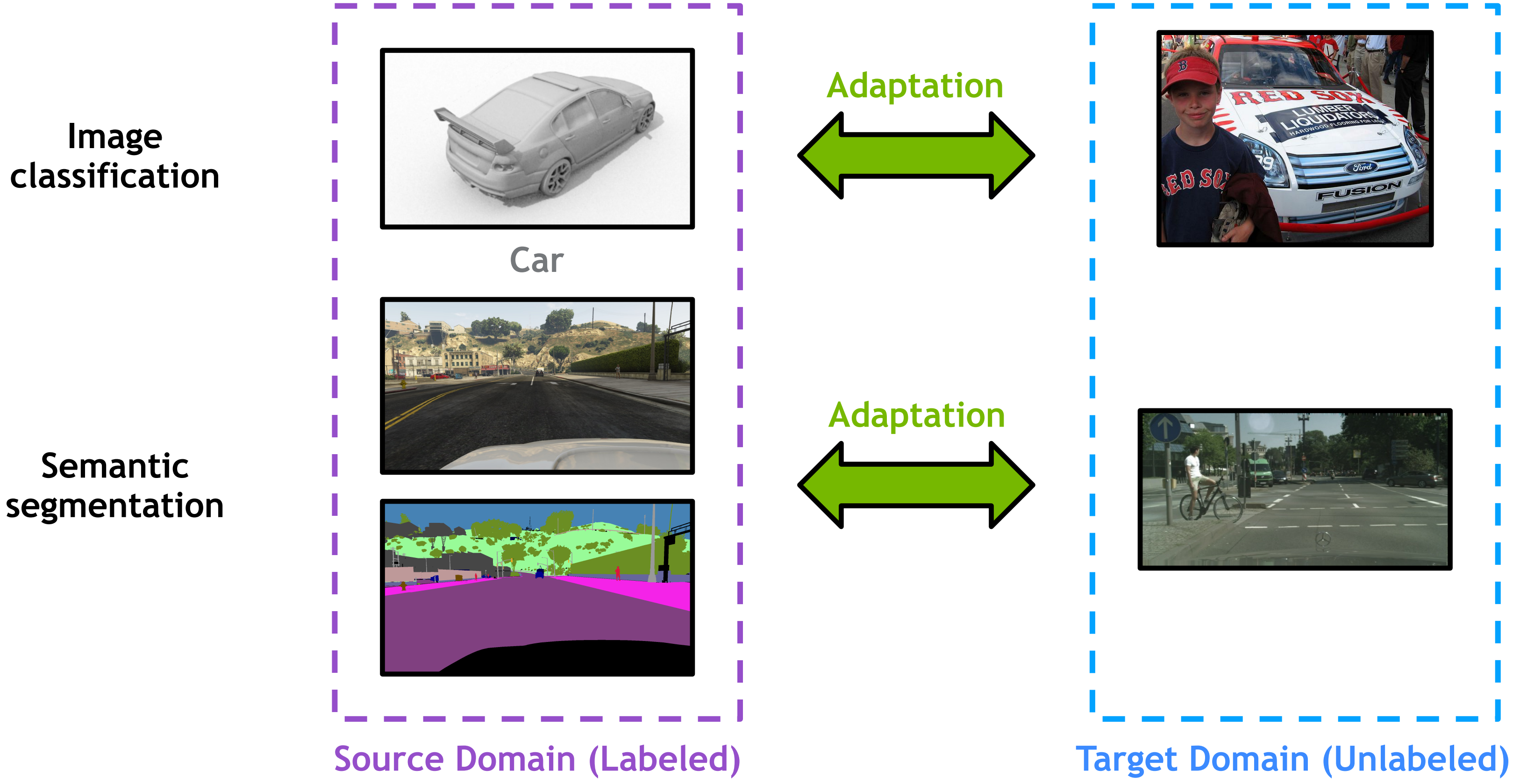

Xiaofeng Liu


Vijayakumar Bhagavatula


Jinsong Wang

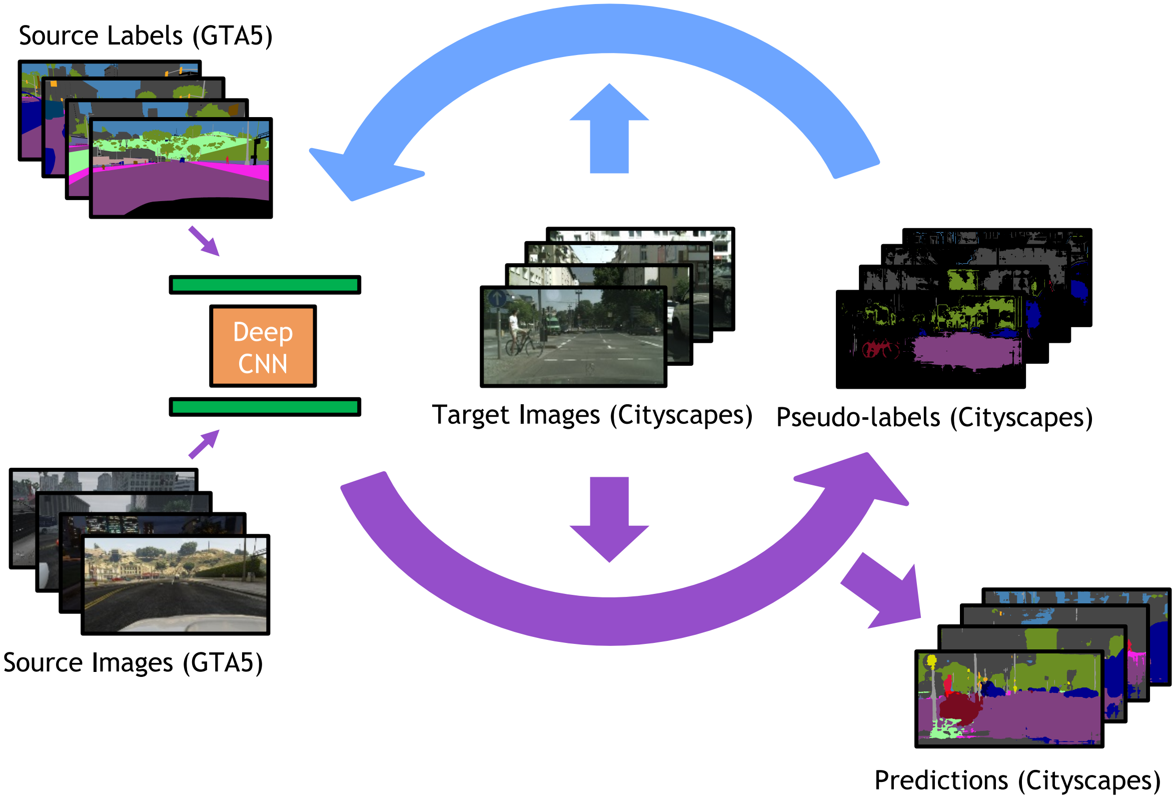
Carnegie Mellon University NVIDIA General Motors

Unsupervised Domain Adaptation (UDA)



UDA through Iterative Deep Self-Training

GTA5 → Cityscapes



Yang Zou, Zhiding Yu et al., **Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training**, ECCV18

Class-Balanced Self-Training (CBST)

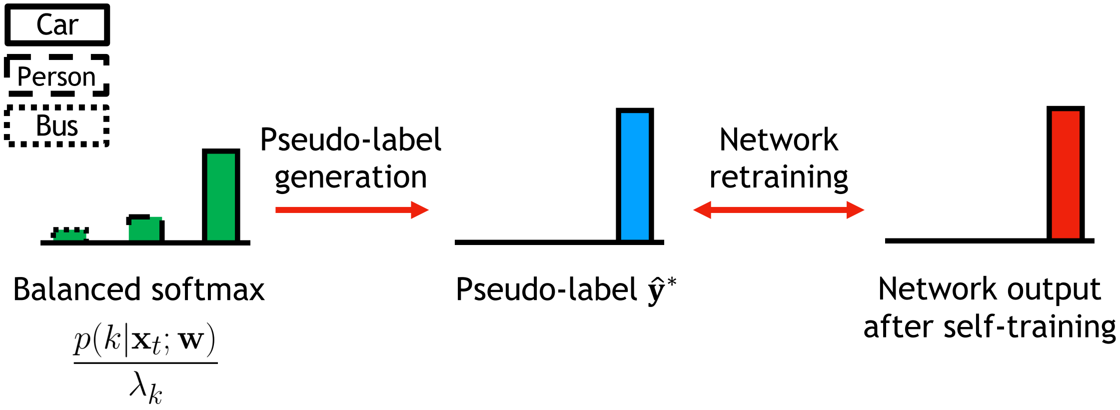
$$\min_{\mathbf{w},\hat{\mathbf{Y}}_T} \mathcal{L}_{CB}(\mathbf{w},\hat{\mathbf{Y}}_T) = -\sum_{s\in S}\sum_{k=1}^{K} y_s^{(k)} \log p(k|\mathbf{x}_s;\mathbf{w}) - \sum_{t\in T}\sum_{k=1}^{K} \hat{y}_t^{(k)} \log \frac{p(k|\mathbf{x}_t;\mathbf{w})}{\lambda_k}$$

$$s.t.\ \hat{\mathbf{y}}_t = (\hat{y}_t^{(1)},...,\hat{y}_t^{(K)}) \in \Delta^{K-1} \cup \{\mathbf{0}\},\ \forall t$$

$$\lambda_k > 0$$

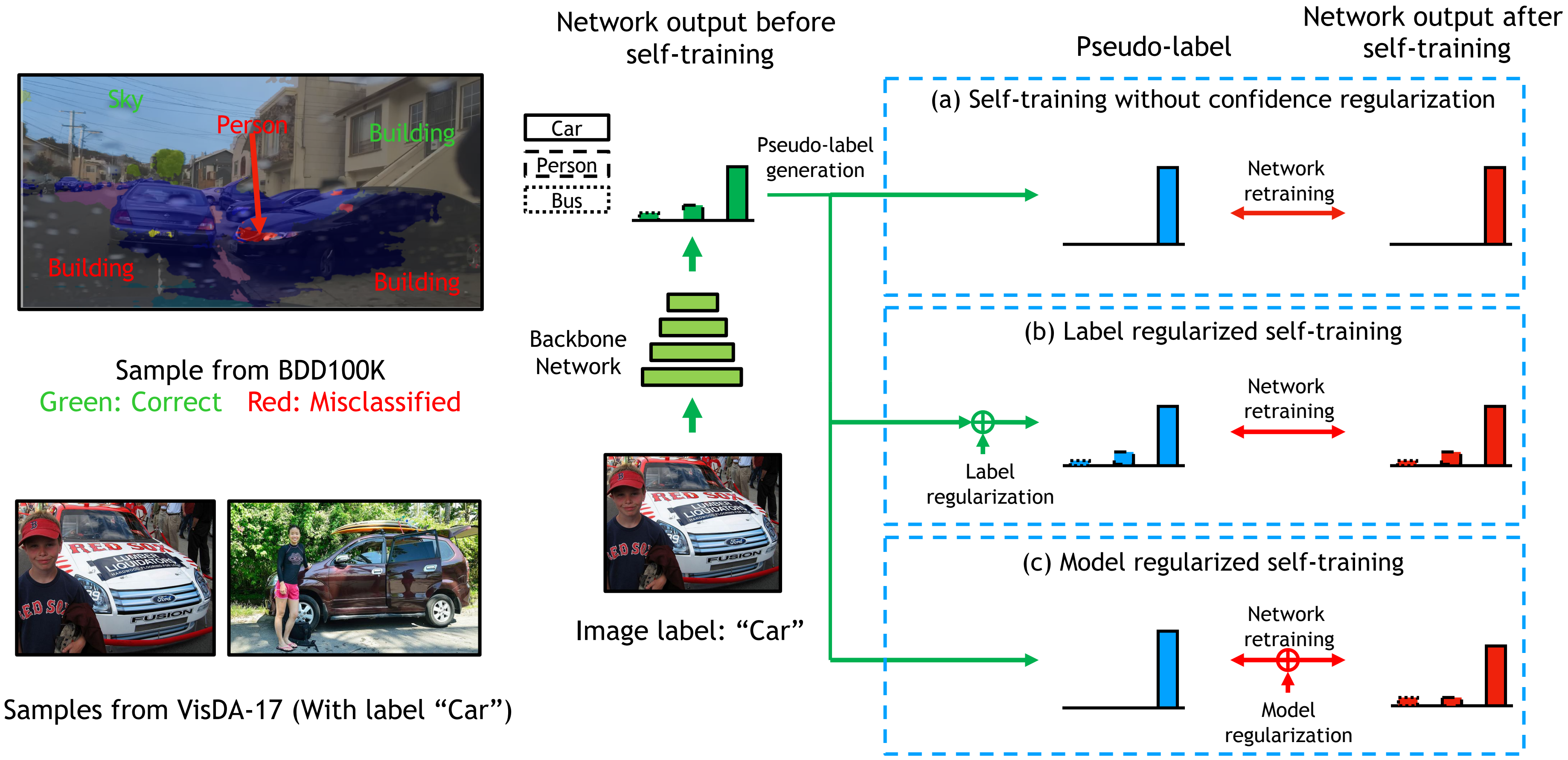
where: $\mathbf{x}$: input sample $\mathbf{p}$: class predication vector $\mathbf{y}$: label vector $\hat{\mathbf{y}}$: pseudo-label vector
$\mathbf{w}$: network parameters s: source sample index t: target sample index Δ^{K-1}: probability simplex

$$\hat{y}_t^{(k)*} = \begin{cases} 1, \text{if } k = \arg\max_k \{\frac{p(k|\mathbf{x}_t;\mathbf{w})}{\lambda_k}\} \\ \quad \text{and} \quad p(k|\mathbf{x}_t;\mathbf{w}) > \lambda_k \\ 0, \text{otherwise} \end{cases}$$



Yang Zou, Zhiding Yu et al., **Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training**, ECCV18

Preventing Overconfidence in Self-Training



Label Regularized Self-Training (LR)

$$\min_{\mathbf{w},\hat{\mathbf{Y}}_T} \mathcal{L}_{LR}(\mathbf{w},\hat{\mathbf{Y}}_T) = -\sum_{s\in S}\sum_{k=1}^{K} y_s^{(k)} \log p(k|\mathbf{x}_s;\mathbf{w}) - \sum_{t\in T}\left[\sum_{k=1}^{K} \hat{y}_t^{(k)} \log \frac{p(k|\mathbf{x}_t;\mathbf{w})}{\lambda_k} \boxed{- \alpha r_c(\hat{\mathbf{y}}_t)}\right]$$

$$s.t.\ \hat{\mathbf{y}}_t = (\hat{y}_t^{(1)},...,\hat{y}_t^{(K)}) \in \Delta^{K-1} \cup \{\mathbf{0}\},\ \forall t$$

$$\lambda_k > 0$$

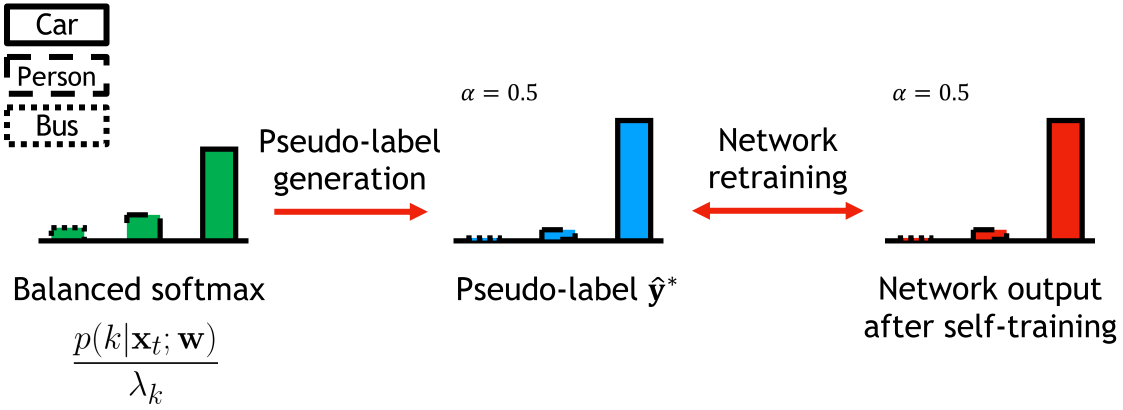
where: α: regularizer weight

$$\mathcal{C}(\hat{\mathbf{y}}_t) = -\hat{y}_t^{(k)} \sum_{k=1}^{K} \log \frac{p(k|\mathbf{x}_t;\mathbf{w})}{\lambda_k} + \alpha r_c(\hat{\mathbf{y}}_t)$$

$$\hat{\mathbf{y}}_t^{\dagger} = \arg\min_{\hat{\mathbf{y}}_t} \mathcal{C}(\hat{\mathbf{y}}_t)$$

$$s.t.\ \hat{y}_t \in \Delta^{(K-1)},\ \forall t$$

$$\hat{\mathbf{y}}_t^{*} = \begin{cases} \hat{\mathbf{y}}_t^{\dagger}, & \text{if } \mathcal{C}(\hat{\mathbf{y}}_t^{\dagger}) < \mathcal{C}(\mathbf{0}) \\ \mathbf{0}, & \text{otherwise} \end{cases}$$



Model Regularized Self-Training (MR)

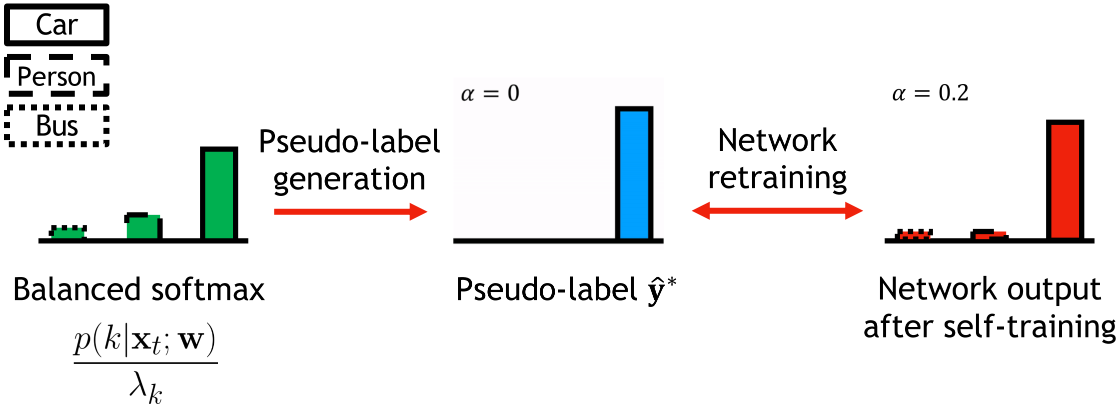
$$\min_{\mathbf{w},\hat{\mathbf{Y}}_T} \mathcal{L}_{MR}(\mathbf{w},\hat{\mathbf{Y}}_T) = -\sum_{s\in S}\sum_{k=1}^{K} y_s^{(k)} \log p(k|\mathbf{x}_s;\mathbf{w}) - \sum_{t\in T}\Big[\sum_{k=1}^{K} \hat{y}_t^{(k)} \log \frac{p(k|\mathbf{x}_t;\mathbf{w})}{\lambda_k} \boxed{- \alpha r_c(p(\mathbf{x}_t;\mathbf{w}))}\Big]$$

$$s.t.\ \hat{\mathbf{y}}_t = (\hat{y}_t^{(1)},...,\hat{y}_t^{(K)}) \in \Delta^{K-1} \cup \{\mathbf{0}\},\ \forall t$$

$$\lambda_k > 0$$

where: α: regularizer weight

$$\min_{\mathbf{w}} -\sum_{t\in T}[\sum_{k=1}^{K} \hat{y}_t^{(k)} \log p(k|\mathbf{x}_t;\mathbf{w}) - \alpha r_c(p(\mathbf{x}_t;\mathbf{w}))]$$



Proposed Confidence Regularizers

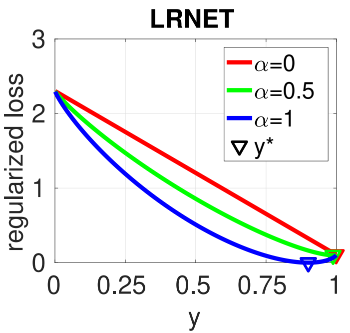
LR-Entropy (LRENT) $r_c(\hat{\mathbf{y}}_t) = \sum_{k=1}^{K} \hat{y}_t^{(k)} \log\left(\hat{y}_t^{(k)}\right)$

Pseudo-label solver $$\hat{y}_t^{(i)\dagger} = \frac{\left(\frac{p(i|\mathbf{x}_t)}{\lambda_k}\right)^{\frac{1}{\alpha}}}{\sum_{k=1}^{K}\left(\frac{p(k|\mathbf{x}_t)}{\lambda_k}\right)^{\frac{1}{\alpha}}}$$

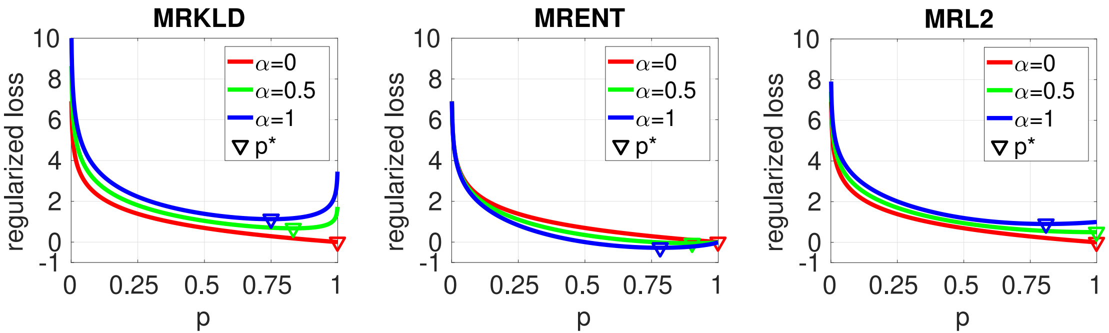
MR-KLDiv (MRKLD) $r_c(p(\mathbf{x}_t;\mathbf{w})) = -\sum_{k=1}^{K} \frac{1}{K} \log p(k|\mathbf{x}_t)$

MR-Entropy (MRENT) $r_c(p(\mathbf{x}_t;\mathbf{w})) = \sum_{k=1}^{K} p(k|\mathbf{x}_t) \log p(k|\mathbf{x}_t)$

MR-L2 (MRL2) $r_c(p(\mathbf{x}_t;\mathbf{w})) = \sum_{k=1}^{K} p(k|\mathbf{x}_t)^2$



Pseudo-label generation loss
v.s. probability



Regularized retraining loss
v.s. probability

Theoretical Analysis

Probabilistic Explanation

Proposition 1. CRST can be modeled as a regularized maximum likelihood for classification (RCML) problem optimized via classification expectation maximization.

Convergence Analysis

Proposition 2. Given pre-determined λ_k's, CRST is convergent with gradient descent for network retraining optimization.

Softlabel and Softmax with temperature

Proposition 3. If λ_k are equal for all k, the soft pseudo-label of *LRENT* given is exactly the same as softmax with temperature

Label smoothing

Proposition 4. Self-training with MRKLD is equivalent to self-training with pseudo-label uniformly smoothed by $\epsilon = (K\alpha - \alpha)/(K + K\alpha)$, where α is the regularizer weight.

Experiment: UDA for Image Classification

Results on VisDA-17

Method	Aero	Bike	Bus	Car	Horse	Knife	Motor	Person	Plant	Skateboard	Train	Truck	Mean
Source [50]	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
MMD [33]	87.1	63.0	76.5	42.0	90.3	42.9	85.9	53.1	49.7	36.3	85.8	20.7	61.1
DANN [15]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
ENT [18]	80.3	75.5	75.8	48.3	77.9	27.3	69.7	40.2	46.5	46.6	79.3	16.0	57.0
MCD [51]	87.0	60.9	**83.7**	64.0	88.9	79.6	84.7	**76.9**	**88.6**	40.3	83.0	25.8	71.9
ADR [50]	87.8	79.5	**83.7**	65.3	**92.3**	61.8	**88.9**	73.2	87.8	60.0	**85.5**	32.3	74.8
SimNet-Res152 [44]	**94.3**	82.3	73.5	47.2	87.9	49.2	75.1	79.7	85.3	68.5	81.1	50.3	72.9
GTA-Res152 [53]	-	-	-	-	-	-	-	-	-	-	-	-	77.1
Source-Res101	68.7	36.7	61.3	**70.4**	67.9	5.9	82.6	25.5	75.6	29.4	83.8	10.9	51.6
CBST	87.2±2.4	78.8±1.0	56.5±2.2	55.4±3.6	85.1±1.4	79.2±10.3	83.8±0.4	77.7±4.0	82.8±2.8	88.8±3.2	69.0±2.9	72.0±3.8	76.4±0.9
MRL2	87.0±2.9	79.5±1.9	57.1±3.2	54.7±2.9	85.5±1.1	78.1±11.7	83.0±1.5	77.7±3.7	82.4±1.7	88.6±2.7	69.1±2.2	71.8±3.0	76.2±1.0
MRENT	87.1±2.7	78.3±0.7	56.1±4.0	54.4±2.7	84.4±2.3	79.9±10.6	83.7±1.1	77.9±4.4	82.7±2.4	87.4±2.8	70.0±1.4	72.8±3.3	76.2±0.8
MRKLD	87.3±2.5	79.4±1.9	60.5±2.4	59.7±2.5	87.6±1.4	**82.4±4.4**	86.5±1.1	78.4±2.6	84.6±1.7	86.4±2.8	72.5±2.4	69.8±2.5	77.9±0.5
LRENT	87.7±2.4	78.7±0.8	57.3±3.3	54.5±4.0	84.8±1.7	79.7±10.3	84.2±1.4	77.4±3.7	83.1±1.5	**88.3±2.6**	70.9±2.1	**72.6±2.4**	76.6±0.9
MRKLD+LRENT	88.0±0.6	79.2±2.2	61.0±3.1	60.0±1.0	87.5±1.2	81.4±5.6	86.3±1.5	78.8±2.1	85.6±0.9	86.6±2.5	73.9±1.3	68.8±2.3	**78.1±0.2**

Results on Office-31

Method	A→W	D→W	W→D	A→D	D→A	W→A	Mean
ResNet-50 [21]	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
DAN [33]	80.5±0.4	97.1±0.2	99.6±0.1	78.6±0.2	63.6±0.3	62.8±0.2	80.4
RTN [35]	84.5±0.2	96.8±0.1	99.4±0.1	77.5±0.3	66.2±0.2	64.8±0.3	81.6
DANN [15]	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
ADDA [61]	86.2±0.5	96.2±0.3	98.4±0.3	77.8±0.3	69.5±0.4	68.9±0.5	82.9
JAN [36]	85.4±0.3	97.4±0.2	99.8±0.2	84.7±0.3	68.6±0.3	70.0±0.4	84.3
GTA [53]	**89.5±0.5**	97.9±0.3	99.8±0.4	87.7±0.5	72.8±0.3	71.4±0.4	86.5
CBST	87.8±0.8	98.5±0.1	**100±0.0**	86.5±1.0	71.2±0.4	70.9±0.7	85.8
MRL2	88.4±0.2	98.6±0.1	**100±0.0**	87.7±0.9	71.8±0.2	**72.1±0.2**	86.4
MRENT	88.0±0.4	98.6±0.1	**100±0.0**	87.4±0.8	**72.7±0.2**	71.0±0.4	86.4
MRKLD	88.4±0.9	98.7±0.1	**100±0.0**	88.0±0.9	71.7±0.8	70.9±0.4	86.3
LRENT	88.6±0.4	98.7±0.1	**100±0.0**	**89.0±0.8**	72.0±0.6	71.0±0.3	86.6
MRKLD+LRENT	89.4±0.7	**98.9±0.4**	**100±0.0**	88.7±0.8	72.6±0.7	70.9±0.5	**86.8**

Experiment: UDA for Semantic Segmentation

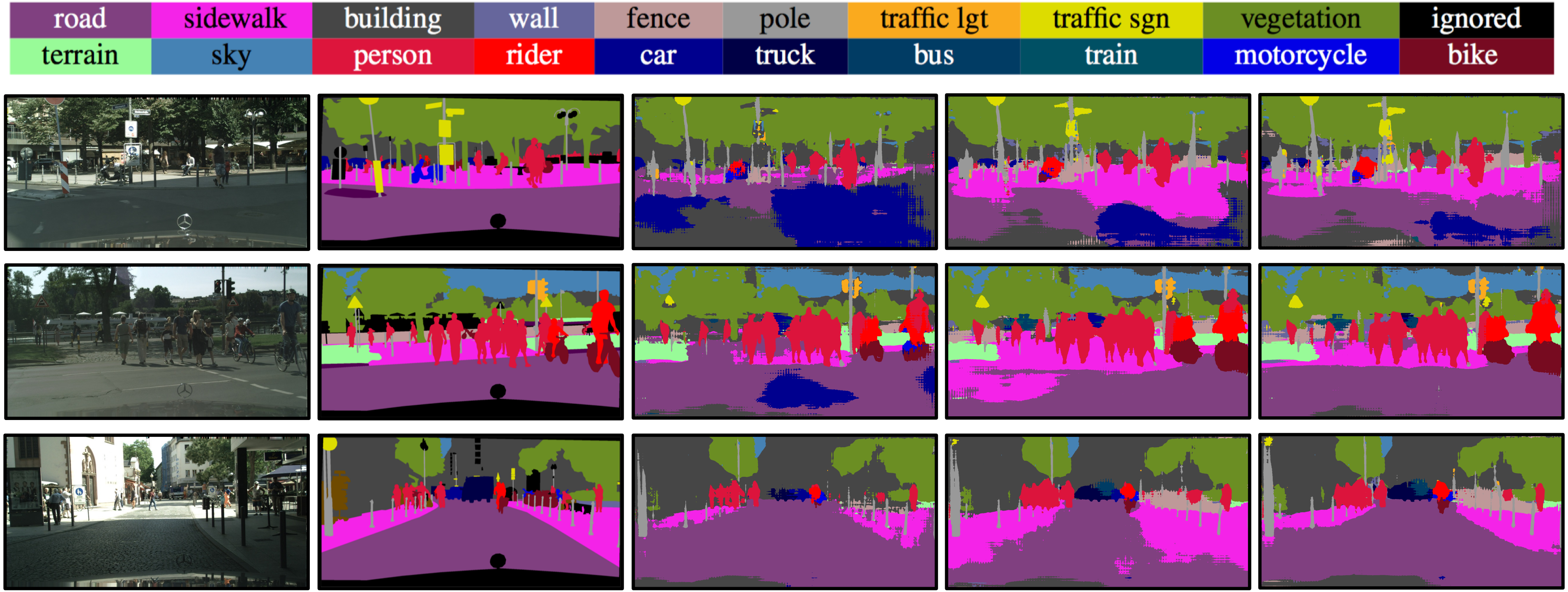
Results on SYNTHIA -> Cityscapes (mIoU* - 13 class)

Method	Backbone	Road	SW	Build	Wall*	Fence*	Pole*	TL	TS	Veg.	Sky	PR	Rider	Car	Bus	Motor	Bike	mIoU	mIoU*
Source	DRN-105	14.9	11.4	58.7	1.9	0.0	24.1	1.2	6.0	68.8	76.0	54.3	7.1	34.2	15.0	0.8	0.0	23.4	26.8
MCD [51]		84.8	**43.6**	79.0	3.9	0.2	29.1	7.2	5.5	83.8	83.1	51.0	11.7	79.9	27.2	6.2	0.0	37.3	43.5
Source	DeepLabv2	55.6	23.8	74.6	–	–	–	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	–	38.6
AdaptSegNet [60]		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
AdvEnt [63]	DeepLabv2	**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	ResNet-38	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 [69]		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
Source	DeepLabv2	64.3	21.3	73.1	2.4	1.1	31.4	7.0	27.7	63.1	67.6	42.2	19.9	73.1	15.3	10.5	38.9	34.9	40.3
CBST		68.0	29.9	76.3	10.8	1.4	33.9	22.8	**29.5**	77.6	78.3	60.6	28.3	81.6	23.5	18.8	39.8	42.6	48.9
MRL2		63.4	27.1	76.4	14.2	1.4	35.2	**23.6**	29.4	78.5	77.8	61.4	**29.5**	82.2	22.8	18.9	42.3	42.8	48.7
MRENT		69.6	32.6	75.8	12.2	**1.8**	35.3	23.3	**29.5**	77.7	78.9	60.0	28.5	81.5	25.9	19.6	41.8	43.4	49.6
MRKLD		67.7	32.2	73.9	10.7	1.6	**37.4**	22.2	31.2	80.8	80.5	60.8	29.1	82.8	25.0	19.4	45.3	**43.8**	**50.1**
LRENT		65.6	30.3	74.6	**13.8**	1.5	35.8	23.1	29.1	77.0	77.5	60.1	28.5	82.2	22.6	**20.1**	41.9	42.7	48.7

Results on GTA5 -> Cityscapes

Method	Backbone	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg.	Terrain	Sky	PR	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU
Source	DRN-26	42.7	26.3	51.7	5.5	6.8	13.8	23.6	6.9	75.5	11.5	36.8	49.3	0.9	46.7	3.4	5.0	0.0	5.0	1.4	21.7
CyCADA [23]		79.1	33.1	77.9	23.4	17.3	32.1	33.3	31.8	81.5	26.7	69.0	62.8	14.7	74.5	20.9	25.6	6.9	18.8	20.4	39.5
Source	DRN-105	36.4	14.2	67.4	16.4	12.0	20.1	8.7	0.7	69.8	13.3	56.9	37.0	0.4	53.6	10.6	3.2	0.2	0.9	0.0	22.2
MCD [51]		90.3	31.0	78.5	19.7	17.3	28.6	30.9	16.1	83.7	30.0	69.1	58.5	19.6	81.5	23.8	30.0	5.7	25.7	14.3	39.7
Source	DeepLabv2	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
AdaptSegNet [60]		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
AdvEnt [63]	DeepLabv2	89.4	33.1	**81.0**	26.6	26.8	27.2	33.5	24.7	83.9	**36.7**	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
Source	DeepLabv2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	29.2
FCAN [67]		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.6
Source	DeepLabv2	71.3	19.2	69.1	18.4	10.0	35.7	27.3	6.8	79.6	24.8	72.1	57.6	19.5	55.5	15.5	15.1	11.7	21.1	12.0	33.8
CBST		91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
MRL2		**91.9**	55.2	80.9	32.1	21.5	36.7	30.0	19.0	84.8	34.9	80.1	56.1	23.8	83.9	28.0	29.4	20.5	24.0	40.3	46.0
MRENT		91.8	53.4	80.6	32.6	20.8	34.3	29.7	21.0	84.0	34.1	80.6	53.9	24.6	82.8	30.8	34.9	16.6	26.4	42.6	46.1
MRKLD		91.0	**55.4**	80.0	33.7	21.4	37.3	32.9	24.5	**85.0**	34.1	80.8	57.7	24.6	84.1	27.8	30.1	26.9	26.0	42.3	47.1
LRENT		91.8	53.5	80.5	32.7	21.0	34.0	29.0	20.3	83.9	34.2	**80.9**	53.1	23.9	82.7	30.2	35.6	16.3	25.9	42.8	45.9
Source	ResNet-38	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 [69]		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
MRL2		84.4	52.7	74.7	38.0	**32.2**	43.7	**53.7**	38.6	73.9	24.4	64.4	45.6	24.6	63.2	3.22	31.9	**45.9**	44.2	34.8	46.0
MRENT		84.6	49.5	73.9	**35.8**	25.1	**46.2**	53.3	43.3	75.2	24.2	63.8	48.2	**33.8**	65.7	2.89	32.6	39.2	**50.0**	34.7	46.4
MRKLD		84.5	47.7	74.1	27.9	22.1	43.8	46.5	37.8	83.7	22.7	56.1	56.8	26.8	81.7	22.5	46.2	27.5	32.3	**47.9**	46.8
LRENT		80.3	40.8	65.8	24.6	30.5	43.1	49.5	40.3	82.1	26.0	54.6	59.4	32.1	68.0	31.9	30.0	21.9	44.8	46.7	45.9
CBST-SP	ResNet-38	85.6	55.1	76.9	26.8	23.4	38.9	47.1	**46.9**	83.4	25.5	68.7	45.6	15.7	79.7	27.7	50.3	38.2	33.4	44.6	48.1
MRKLD-SP		90.8	46.0	79.9	27.4	23.3	42.3	46.2	40.9	83.5	19.2	59.1	63.5	30.8	83.5	36.8	**52.0**	28.0	36.8	46.4	49.2
MRKLD-SP-MST		91.7	45.1	80.9	29.0	23.4	43.8	47.1	40.9	84.0	20.0	60.6	**64.0**	31.9	**85.8**	**39.5**	48.7	25.0	38.0	47.0	**49.8**

Experiment: Qualitative Results (GTA5 -> Cityscapes)



Original Image | Ground Truth | Source Model | CBST | CRST (MRKLD)

Conclusions and Future Works

Conclusions

- Compared with supervised learning, self-training is an under-determined problem (EM with latent variables).
- Our work shows the importance of confidence regularizations as inductive biases to help under-constrained problems such as unsupervised domain adaptation and semi-supervised learning.
- CRST is still aligned with entropy minimization. The proposed confidence regularization only serves as a safety measure to prevent over self-training/entropy minimization.
- MR-KLD is most recommended in practice for its efficiency and good performance.

Future Works

- This work could potentially inspire many other meaningful regularizations/inductive biases for similar problems.