UNSUPERVISED DOMAIN ADAPTATION FOR SEMANTIC SEGMENTATION

Anurag Singh

Zeynep Gerem

Outline

- Motivation
- Our Method
- Results

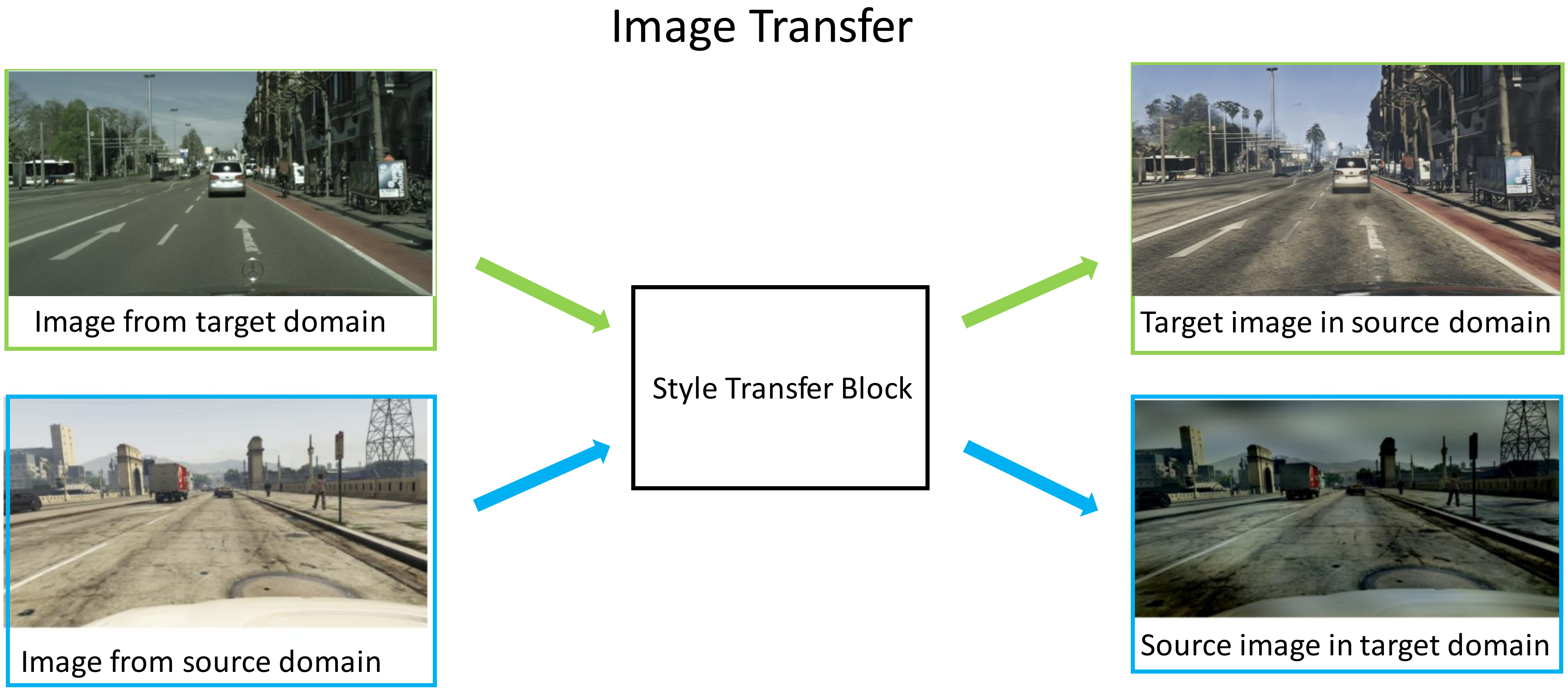


Figure 1: Style transfer of source and target images

Fourier Domain Adaptation for Semantic Segmentation

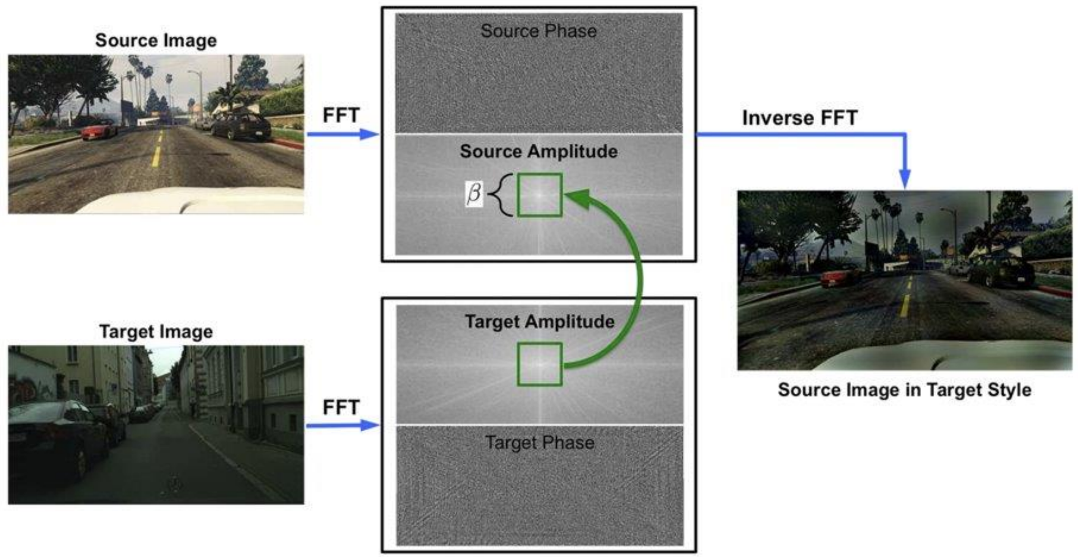


Figure 2: An example of spectral transfer. A source image from GTA5 dataset transferred to target style [1]

- Mask with beta hyper-parameter to filter some range of frequencies

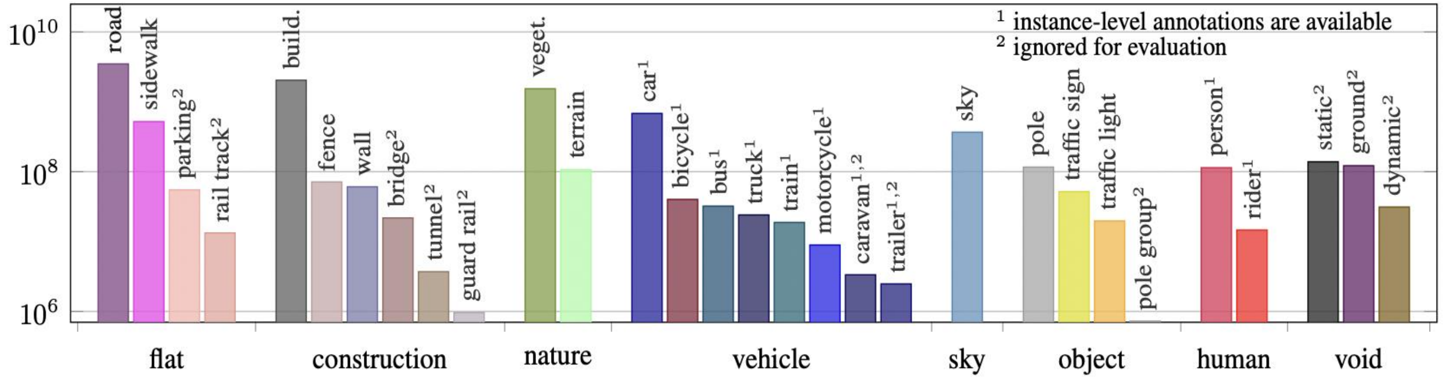
$$M_\beta(h,w) = \mathbb{1}_{(h,w)\in[-\beta H:\beta H, -\beta W:\beta W]}$$

- The stylization of source to target is done according to

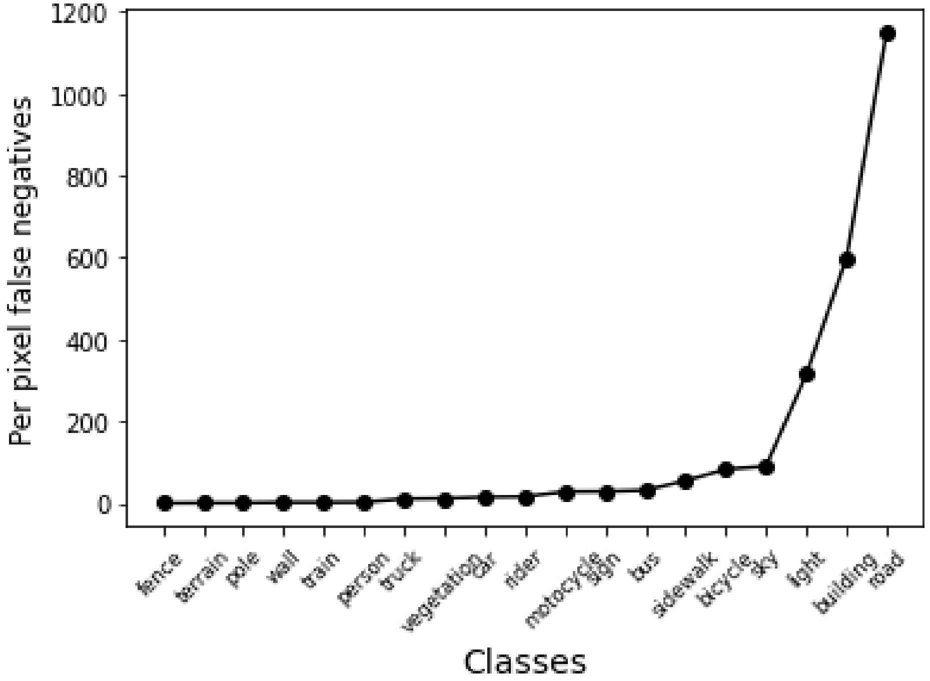
$$x^{s\to t} = \mathcal{F}^{-1}([M_\beta \circ \mathcal{F}^A(x^t) + (1-M_\beta)\circ \mathcal{F}^A(x^s), \mathcal{F}^P(x^s)])$$

[1] Yanchao et al. Fda: Fourier domain adaptation for semantic segmentation. In CVPR, 2020.

Contrastive Learning with Class Imbalance



Distribution of pixels among classes in Cityscapes dataset



Avg False negative count over (30k) for 2k subset size

Unsupervised contrastive learning

- ➢ High number of false negatives

Supervised contrastive learning

- ➢ Need for labels

OUR METHOD

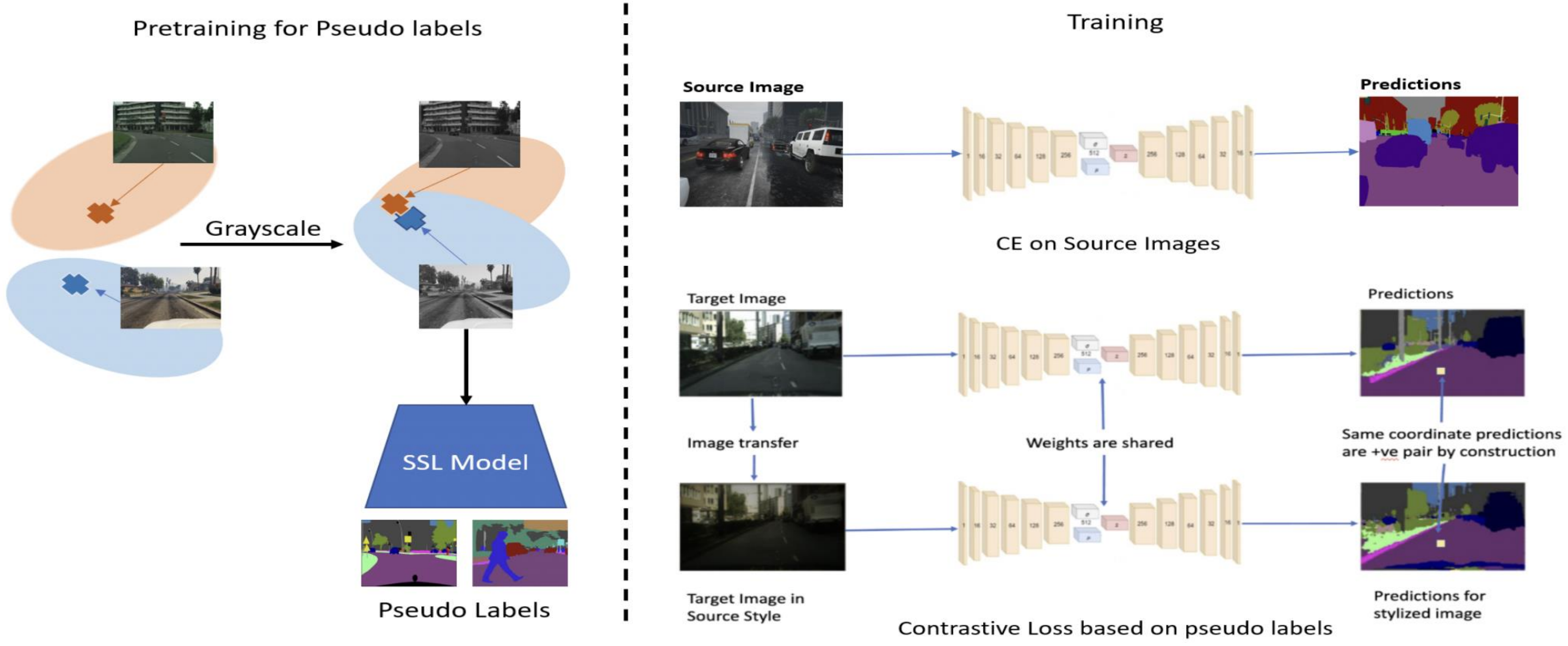
PROPOSED METHOD

Model 1: Training to get pseudo labels

- Cross entropy on source images
- Entropy minimization on target images
- Get predictions for source stylized target images -> pseudo labels

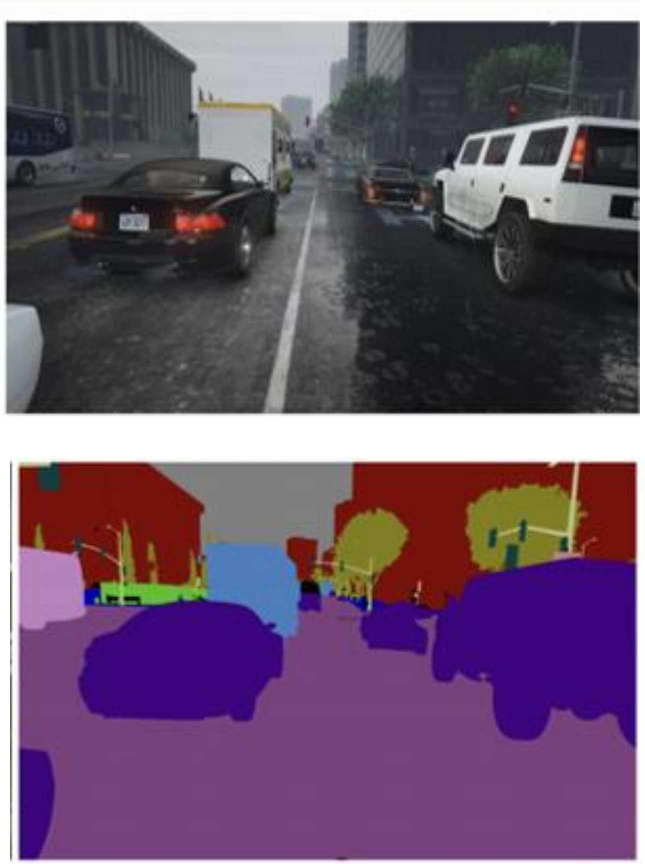
Model 2: Training with pseudo labels

- Cross entropy on target stylized source images
- Entropy minimization on target images
- Supervised contrastive loss on target and source stylized target image pair using pseudo labels



Datasets

Source Domain Dataset



GTA5

Target Domain Dataset

Cityscapes

RESULTS

Training for pseudo labels

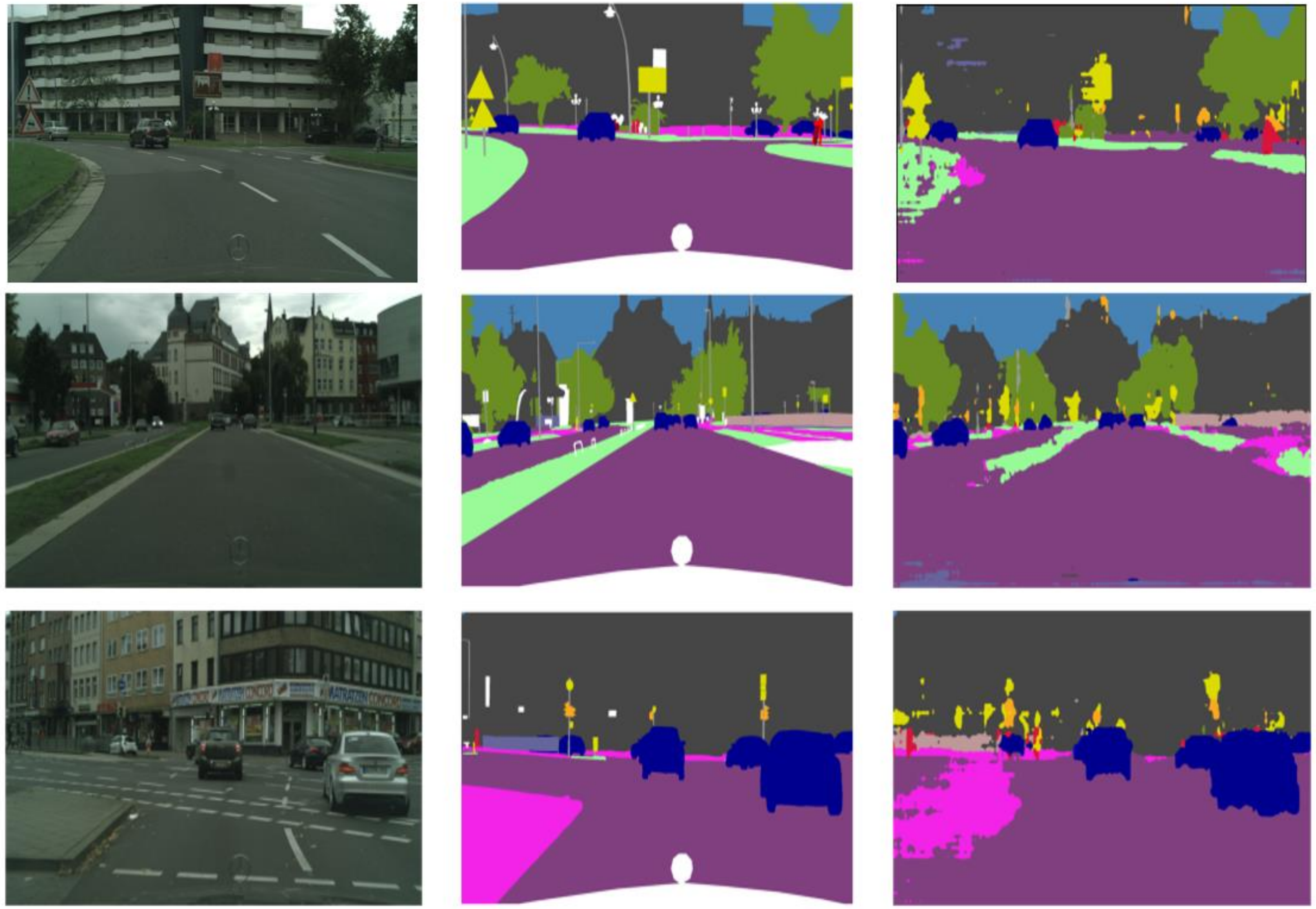
Train on grayscale source and target images
DeepLabV2 trained for 100k iterations
No image translation while training the model

Method	*road*	*sidewalk*	*building*	*wall*	*fence*	*pole*	*light*	*sign*	*vegetation*	*terrain*	*sky*	*person*	*rider*	*car*	*truck*	*bus*	*train*	*motocycle*	*bicycle*	mIoU
Trg in Trg	78.98	41.19	65.04	13.87	16.73	22.16	22.32	34.33	68.18	25.79	58.69	53.01	30.18	77.76	12.88	31.3	4.7	12.34	45.46	37.63
Trg in Src	89.42	40.66	76.9	13.96	22.18	24.03	21.51	39.23	80.78	24.08	81.87	48.06	33.16	80.88	34.43	42.23	26.33	15.59	45.99	44.28
Pseudo Labels	89.12	44.14	78.71	13.21	27.79	24.34	25.42	40.08	79.84	27.79	85.43	52.06	40.39	81.18	28.33	37.59	29.78	25.99	48.09	46.28

Table 1. Results of the first model on GTA-5→Cityscapes with DeepLab backbone

Pseudo label Visualization

Cityscapes Images | Ground Truth | Pseudo Labels

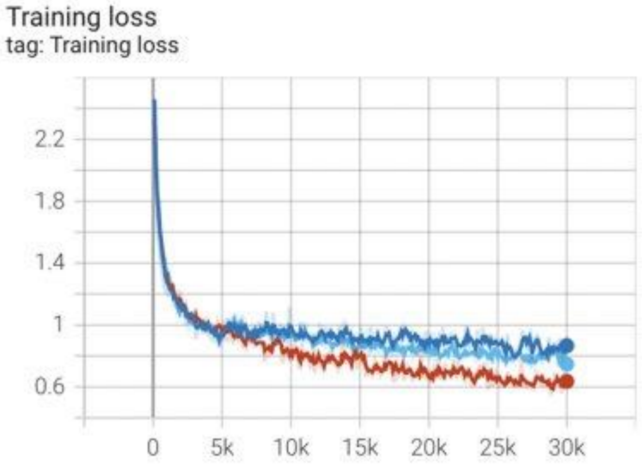


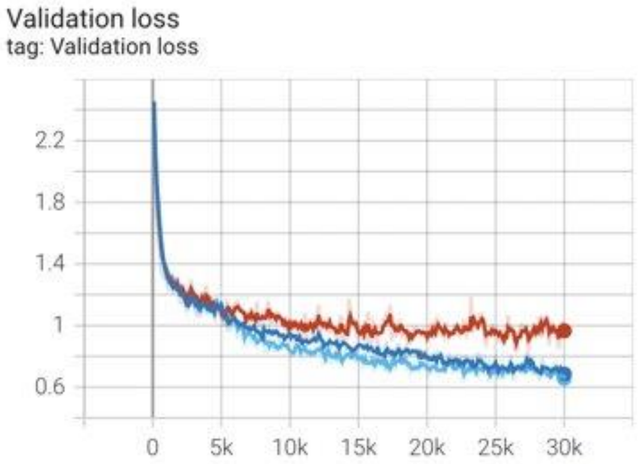
UDA on ENet: Supervised Contrastive Learning

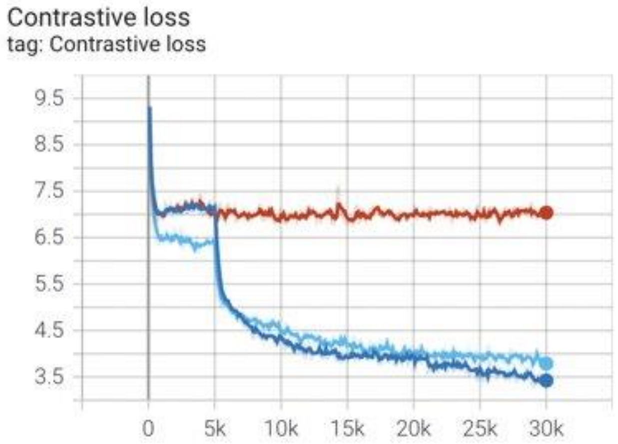
Cross entropy on source images

Contrastive loss with pseudo labels

Contrastive loss with ground truth labels







Method	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motocycle	bicycle	mIoU
CE on Src	29.45	14.82	40.5	0.17	0.0	0.11	0.0	0.0	16.47	0.48	65.48	0.0	0.0	12.97	0.28	0.0	0.0	0.0	0.0	9.51
CL using PL	85.98	23.25	65.05	0.31	0.0	0.0	0.0	0.0	70.56	3.22	77.69	0.0	0.0	57.37	0.0	0.0	0.0	0.0	0.0	20.18
CL using GT	90.53	48.83	62.93	0.01	0.0	0.0	0.0	0.0	69.3	0.0	77.53	0.0	0.0	65.06	0.0	0.0	0.0	0.0	0.0	21.8

Table 3. Quantative comparison on GTA-5→Cityscapes with ENet backbone

Main Results:

Method	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motocycle	bicycle	mIoU
VGG16 backbone																				
CBST[23]	66.7	26.8	73.7	14.8	9.5	28.3	25.9	10.1	75.5	15.7	51.6	47.2	6.2	71.9	3.7	2.2	5.4	18.9	32.4	30.9
SIBAN[10]	83.4	13.0	77.8	20.4	17.5	24.6	22.8	9.6	81.3	29.6	77.3	42.7	10.9	76.0	22.8	17.9	5.7	14.2	2.0	34.2
Cycada[6]	85.2	37.2	76.5	21.8	15.0	23.8	22.9	21.5	80.5	31.3	60.7	50.5	9.0	76.9	17.1	28.2	4.5	9.8	0	35.4
AdvEnt[19]	86.9	28.7	78.7	28.5	25.2	17.1	20.3	10.9	80.0	26.4	70.2	47.1	8.4	81.5	26.0	17.2	18.9	11.7	1.6	36.1
DCAN[20]	82.3	26.7	77.4	23.7	20.5	20.4	30.3	15.9	80.9	25.4	69.5	52.6	11.1	79.6	24.9	21.2	1.30	17.0	6.70	36.2
CLAN[11]	88.0	30.6	79.2	23.4	20.5	26.1	23.0	14.8	81.6	34.5	72.0	45.8	7.9	80.5	26.6	29.9	0.0	10.7	0.0	36.6
LSD[15]	88.0	30.5	78.6	25.2	23.5	16.7	23.5	11.6	78.7	27.2	71.9	51.3	19.5	80.4	19.8	18.3	0.9	20.8	18.4	37.1
BDL[8]	89.2	40.9	81.2	29.1	19.2	14.2	29.0	19.6	83.7	35.9	80.7	54.7	23.3	82.7	25.8	28.0	2.3	25.7	19.9	41.3
FDA-MBT[21]	86.1	35.1	80.6	30.8	20.4	27.5	30.0	26.0	82.1	30.3	73.6	52.5	21.7	81.7	24.0	30.5	29.9	14.6	24.0	42.2
Ours	92.79	55.48	74.99	7.17	21.35	11.39	25.65	39.98	80.51	26.35	84.28	38.34	32.72	70.8	18.95	25.74	28.76	37.91	37.03	**42.64**
ResNet101 backbone																				
AdaStruct[17]	86.5	25.9	79.8	22.1	20.0	23.6	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5	41.4
DCAN[20]	85.0	30.8	81.3	25.8	21.2	22.2	25.4	26.6	83.4	36.7	76.2	58.9	24.9	80.7	29.5	42.9	2.5	26.9	11.6	41.7
DLOW[5]	87.1	33.5	80.5	24.5	13.2	29.8	29.5	26.6	82.6	26.7	81.8	55.9	25.3	78.0	33.5	38.7	0.0	22.9	34.5	42.3
Cycada[6]	86.7	35.6	80.1	19.8	17.5	38.0	39.9	41.5	82.7	27.9	73.6	64.9	19	65.0	12.0	28.6	4.5	31.1	42.0	42.7
CLAN[11]	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
ABStruct[2]	91.5	47.5	82.5	31.3	25.6	33.0	33.7	25.8	82.7	28.8	82.7	62.4	30.8	85.2	27.7	34.5	6.4	25.2	24.4	45.4
AdvEnt[19]	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
BDL[8]	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
FDA	90.0	40.5	79.4	25.3	26.7	30.6	31.9	29.3	79.4	28.8	76.5	56.4	27.5	81.7	27.7	45.1	17.0	23.8	29.6	44.6
FDA-MBT[21]	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	**50.45**
Ours	92.68	53.3	78.22	23.44	17.81	25.29	15.67	36.52	79.85	34.42	82.8	45.88	31.82	83.67	47.95	42.76	30.61	17.59	41.66	46.42

Table 1. Quantative comparison on GTA-5→Cityscapes with VGG16 and ResNet 101 backbone where FDA-BMT refers to the entire method proposed in [21] and FDA refers to single setting training.

Ablations

Method	*road*	*sidewalk*	*building*	*wall*	*fence*	*pole*	*light*	*sign*	*vegetation*	*terrain*	*sky*	*person*	*rider*	*car*	*truck*	*bus*	*train*	*motocycle*	*bicycle*	mIoU
FDA ($\beta = 0.01, \lambda_{ent} = 0$)	29.45	14.82	40.5	0.17	0.0	0.11	0.0	0.0	16.47	0.48	65.48	0.0	0.0	12.97	0.28	0.0	0.0	0.0	0.0	9.51
FDA ($\beta = 0.01$)	82.42	2.96	59.78	0.04	0.0	0.02	0.0	0.0	64.96	18.46	61.65	0.0	0.0	48.11	0.87	0.0	0.0	0.0	0.0	17.86
FDA ($\beta = 0.05$)	82.41	13.73	56.1	0.01	0.0	0.77	0.0	0.0	65.06	5.01	62.1	0.0	0.0	33.86	2.25	0.0	0.0	0.0	0.0	16.91
FDA ($\beta = 0.09$)	72.07	22.07	56.32	0.66	0.0	2.17	0.0	0.0	63.37	12.83	57.97	0.0	0.0	30.11	1.46	0.0	0.0	0.0	0.0	16.79
CL using PL (Ours)	80.11	27.72	72.31	1.61	0.0	0.0	0.0	0.0	71.43	0.38	77.47	0.0	0.0	61.33	0.0	0.0	0.0	0.0	0.0	20.65
CL using GT	90.53	48.83	62.93	0.01	0.0	0.0	0.0	0.0	69.3	0.0	77.53	0.0	0.0	65.06	0.0	0.0	0.0	0.0	0.0	21.8

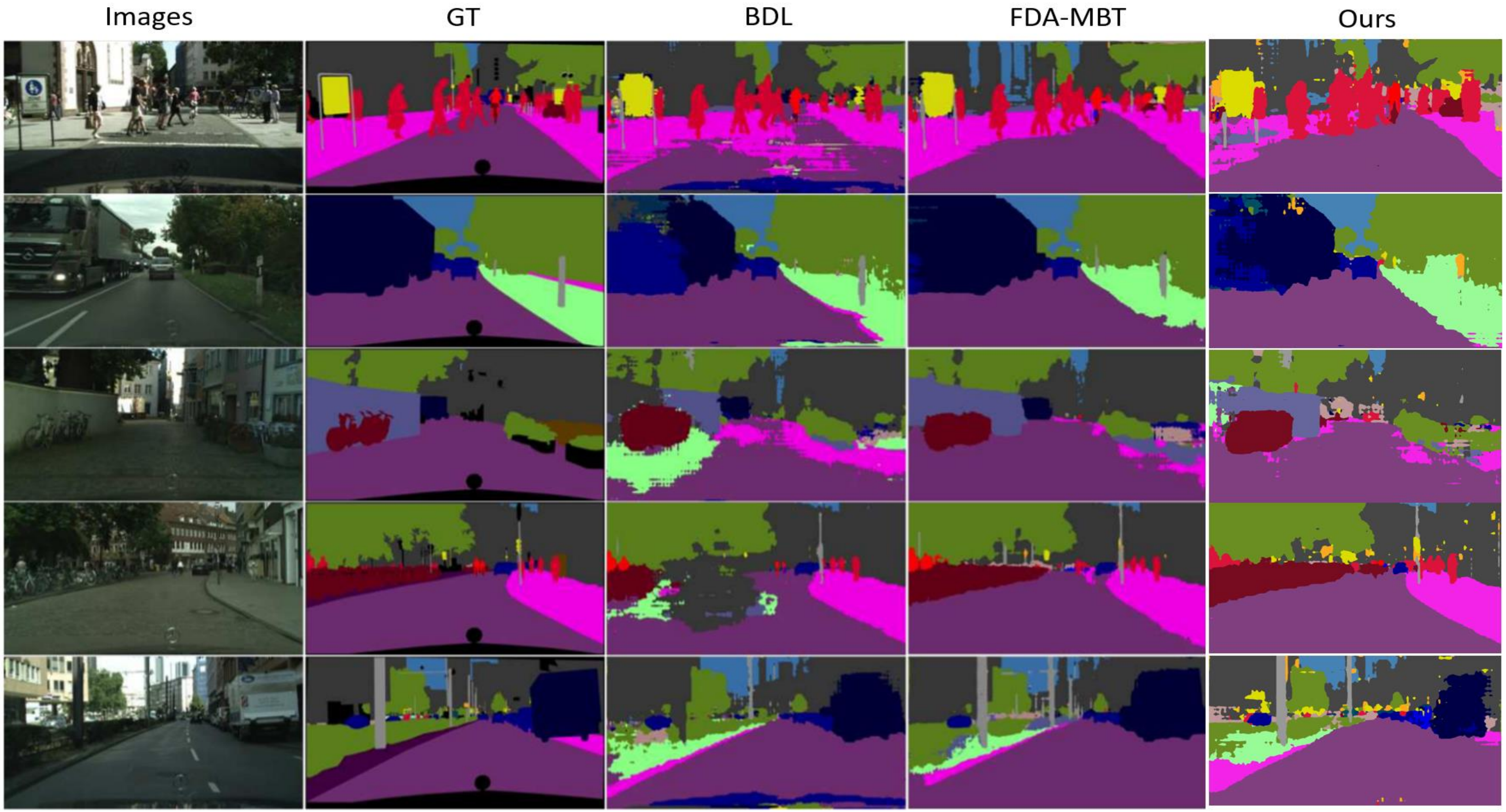
Ablation on ENet backbone

Table 2. Ablation of the first and the second model for GTA-5→Cityscapes

Components			mIoU
CE	CL	$\lambda_{ent} = 0$	
	✓		46.42
✓			45.42
			45.01
		✓	44.64

Table 3. Ablation on Deeplab backbone

Qualitative Results on DeepLab



Takeaway

- Contrastive Learning(CL) > psuedo label CE in unsupervised setting
- Contrastive Learning(CL) can have architecture & dataset challenges

Thank You

REPO LINK