COINet: Adaptive Segmentation with Co-Interactive Network for Autonomous Driving

Jie Liu[1], Xiaoqing Guo[1], Baopu Li[2], Yixuan Yuan[1†]

1 Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China
2 Baidu Research, USA
† Corresponding Author

Overview

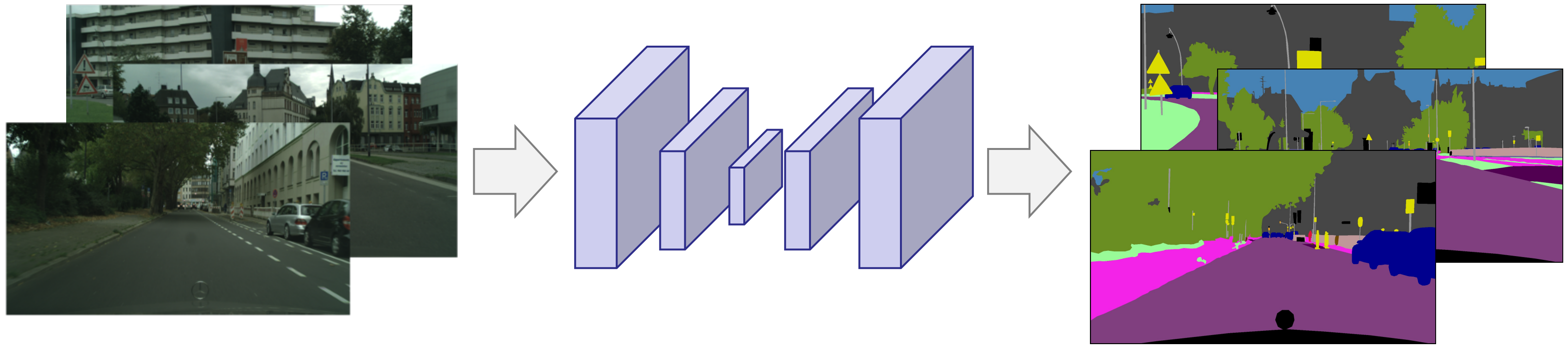
Department of Electrical Engineering
香港城市大學
City University of Hong Kong

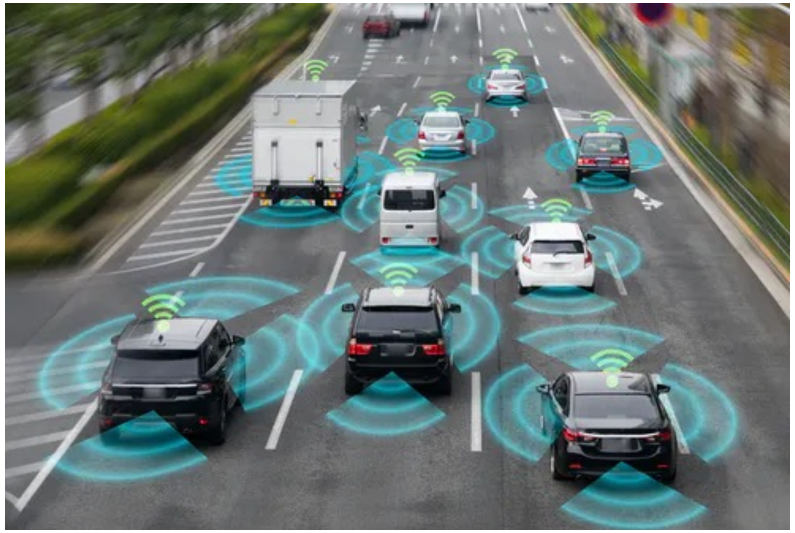
Overview

Department of Electrical Engineering
香港城市大學
City University of Hong Kong

Background: Segmentation

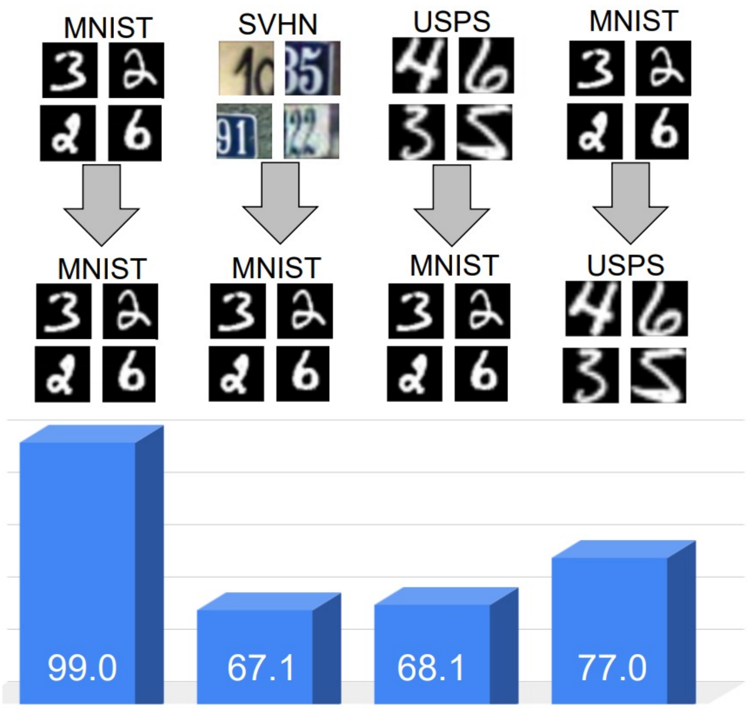
- One of the fundamental computer vision problems
- Assign semantic label for each pixel in the images
- Practical real-world application: autonomous driving





Background: Unsupervised Domain Adaptation

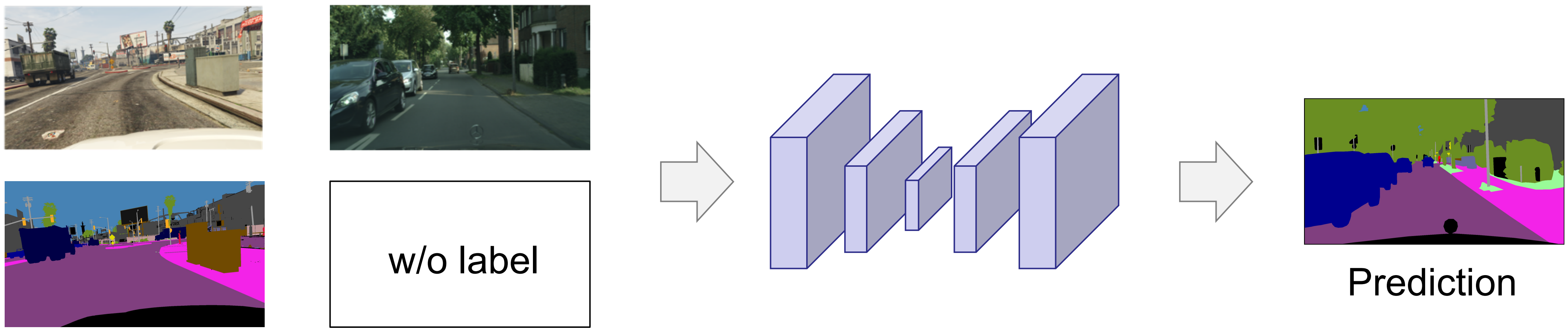
- Challenge1: When applying model to a new domain, the performance will drop
- Challenge2: The annotation is labor-intensive and expensive, especially for pixel-level label



~60 min per image

Background: Unsupervised Domain Adaptation

- Given: Source data w/ annotations + Target data w/o annotations(new domain)
- Object: Transfer the knowledge to a new domain without annotations.



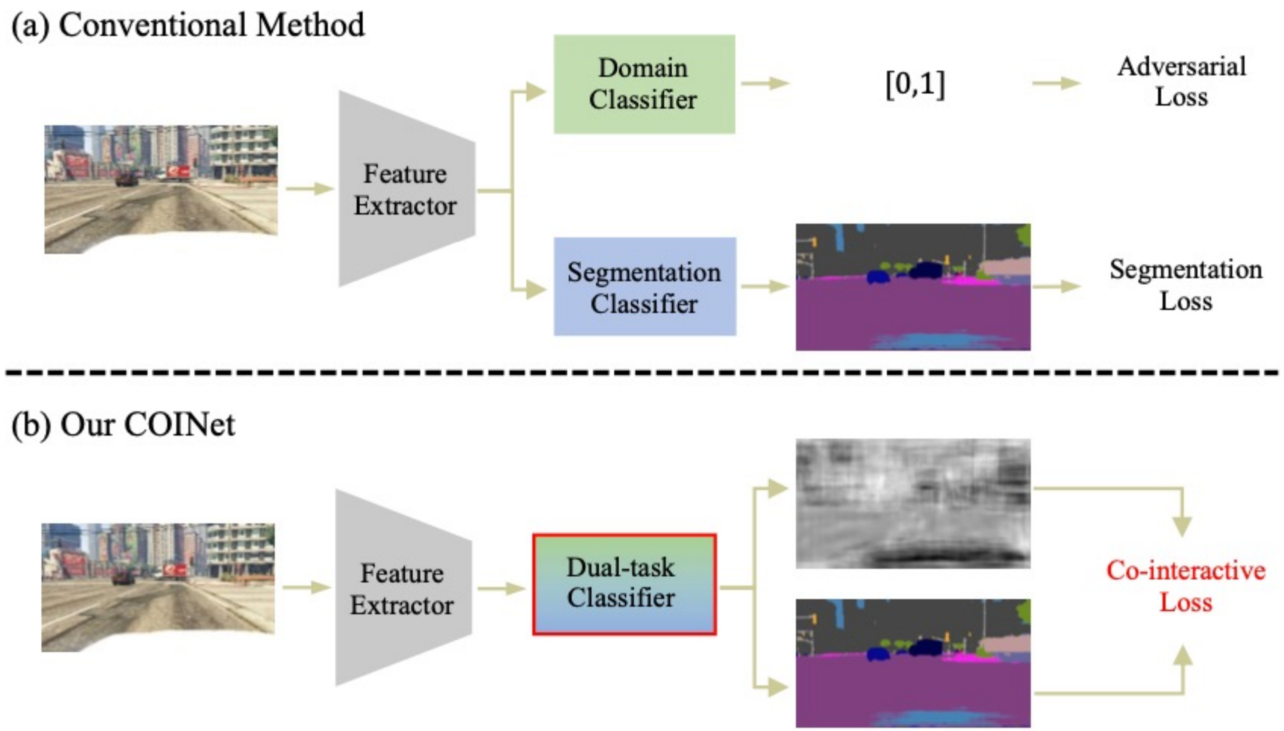
Source Domain　　Target Domain

Overview

Department of Electrical Engineering
香港城市大學
City University of Hong Kong

Motivation

- Ignore the interactive relationship between segmentation task and domain task.
- Not consider the semantic gap among different feature maps.

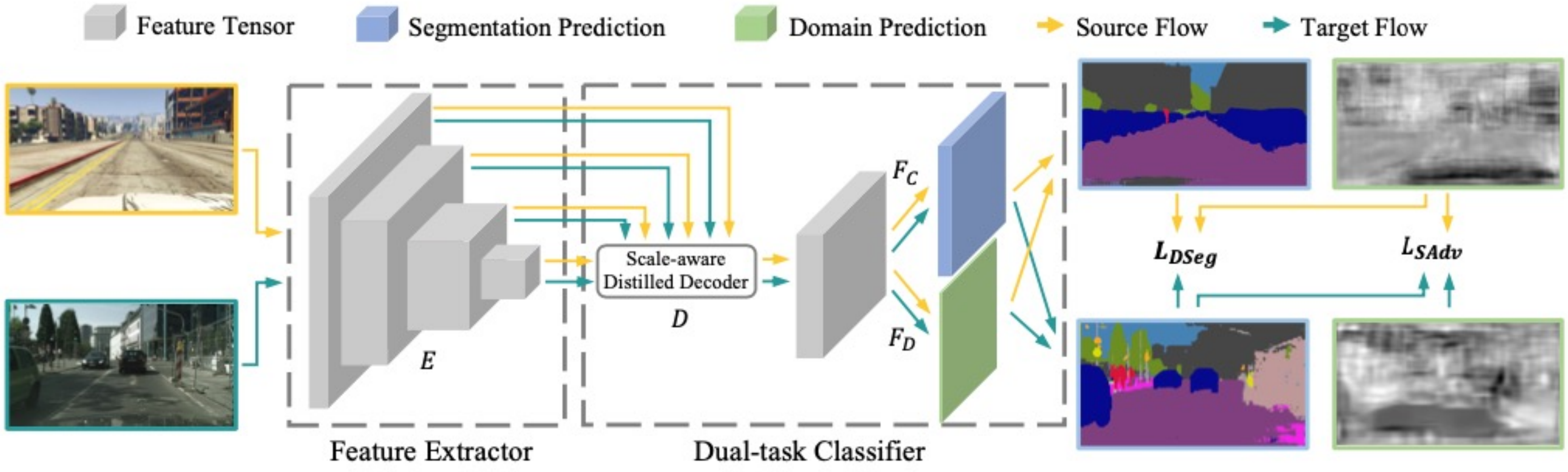


Overview

Methodology：COINet

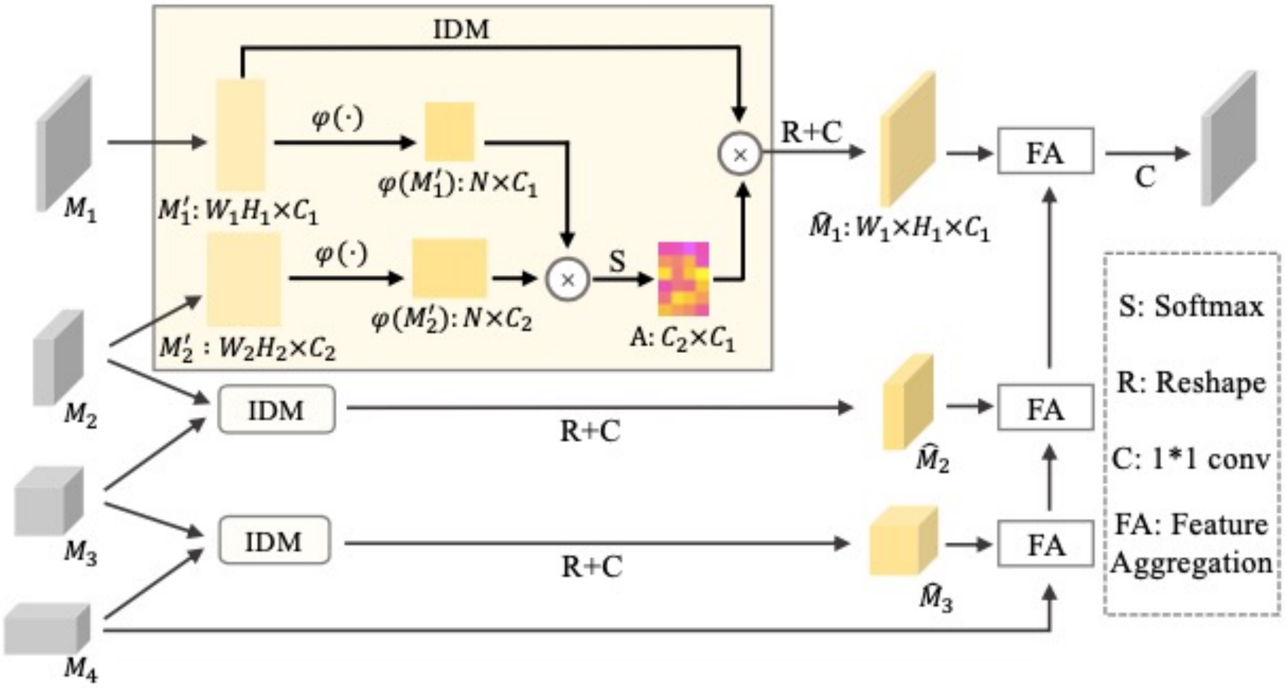
Department of Electrical Engineering
香港城市大學
City University of Hong Kong

- Feature Extractor E: employ DeepLabv2 to extract image feature.
- Scale-aware Distilled Decoder D: eliminate the domain gap among multi-scale feature maps and fuse them.
- Domain Prediction Branch F_D: predict the domain results
- Segmentation Branch F_C: predict the semantic results
- Co-interactive Loss L_{DSeg} and L_{SAdv}: align the feature distribution and refine the segmentation classifier decision boundary.



Methodology: Scale-aware Distilled Decoder

- Multi-scale feature maps: high level semantic information, shallow detailed texture information
- Semantic gaps among different scale feature maps
- Inter-Distilled Module (IDM): utilize the deep feature map to guide the semantic distillation of adjacent shallow feature map



- Calculate Channel affinity

$$A^{(i,j)} = \frac{\exp\left(\varphi(M_k')^i \cdot \varphi(M_{k-1}')^j\right)}{\sum_{j=1}^{C_{k-1}} \exp\left(\varphi(M_k')^i \cdot \varphi(M_{k-1}')^j\right)}$$

- Distill feature map

$$\hat{M}_{k-1} = M_{k-1}' A^T$$

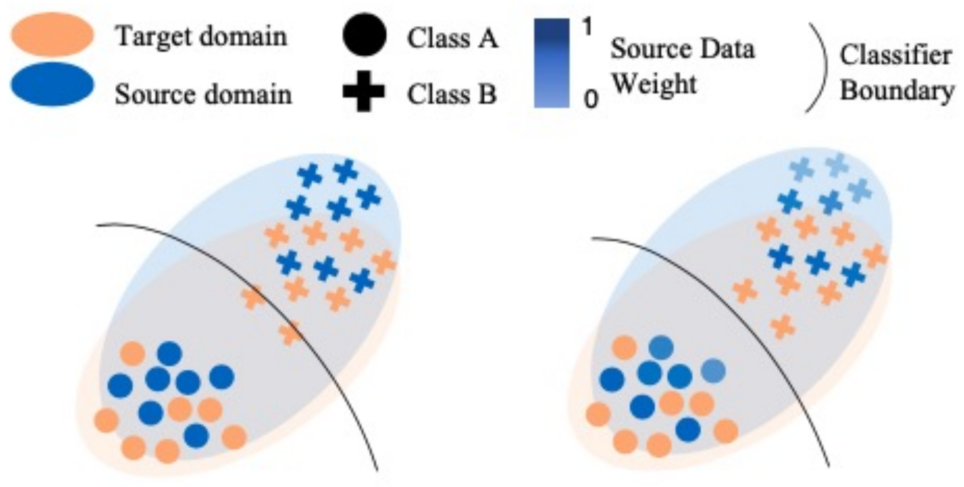
Methodology：Co-interactive Loss

- Domain promote segmentation: Enlarge the weight of source features which are regarded as target domain.

$$\mathcal{L}_{DSeg}(E, F_C) = \sum_{i=1}^{WH} -(1 + \rho_{seg} p_i^D) y_i \log p_i^C$$

- Segmentation promote domain: Reduce the adversarial weight for target features with high confidence.

$$\mathcal{L}_{SAdv}(E, F_D) = \sum_{i=1}^{WH} \left[-\left(1 + \frac{\rho_{adv}}{p^C_{t(i,\hat{y}_t)}}\right) z \log p^D_{t(i)} - (1-z) \log\left(1 - p^D_{s(i)}\right)\right]$$



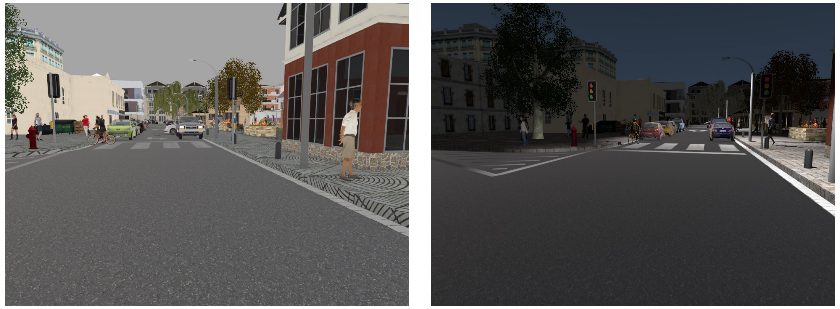
(a) Uniform Weight (b) Confidence Weight

Overview

Experiment

- Source Dataset
  - GTAV: 24996 images collected from computer game with pixel-level labels
  - SYNTHIA: 9400 synthetic images with pixel-level labels
- Target Dataset
  - Cityscapes: 2975 training images and 500 validation images

GTAV



SYNTHIA

Cityscapes

Experiment

- Achieve superior results comparing with other SOTA methods

TABLE I: Unsupervised Adaptation Model performance from GTAV [32] to Cityscapes [7].

Method	road	side.	build.	wall	fence	pole	light	sign	vege.	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU	Gain
Source only	75.8	16.8	72.2	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	-
SSF-DAN (19′) [17]	90.3	38.9	81.7	24.8	22.9	30.5	37.0	21.2	**84.8**	**38.8**	76.9	58.8	30.7	85.7	30.6	38.1	5.9	28.3	36.9	45.4	8.8
CrCDA (20′) [29]	**92.4**	**55.3**	82.3	31.2	29.1	32.5	33.2	35.6	83.5	34.8	84.2	58.9	32.2	84.7	**40.6**	46.1	2.1	31.1	32.7	48.6	12.0
UIDA (20′) [33]	90.6	37.1	82.6	30.1	19.1	29.5	32.4	20.6	85.7	40.5	79.7	58.7	31.1	86.3	31.5	48.3	0.0	30.2	35.8	46.3	9.7
LSE (20′) [34]	90.2	40.0	83.5	31.9	26.4	32.6	38.7	37.5	81.0	34.2	84.6	**61.6**	33.4	82.5	32.8	45.9	6.7	29.1	30.6	47.5	10.9
WeakDA (20′) [35]	91.6	47.4	84.0	30.4	28.3	31.4	37.4	35.4	83.9	38.3	83.9	61.2	28.2	83.7	28.8	41.3	8.8	24.7	**46.4**	48.2	11.6
BCDM (21′) [30]	90.5	37.3	83.7	**39.2**	22.2	28.5	36.0	17.0	84.2	35.9	**85.9**	59.1	35.5	85.2	31.1	39.3	**21.1**	26.7	27.5	46.6	10.0
COINet	91.8	47.3	**85.1**	34.2	**29.1**	**35.2**	**40.7**	**40.9**	80.8	36.4	81.2	59.3	**36.5**	**87.3**	33.4	47.5	5.6	29.9	32.1	**49.2**	**12.6**

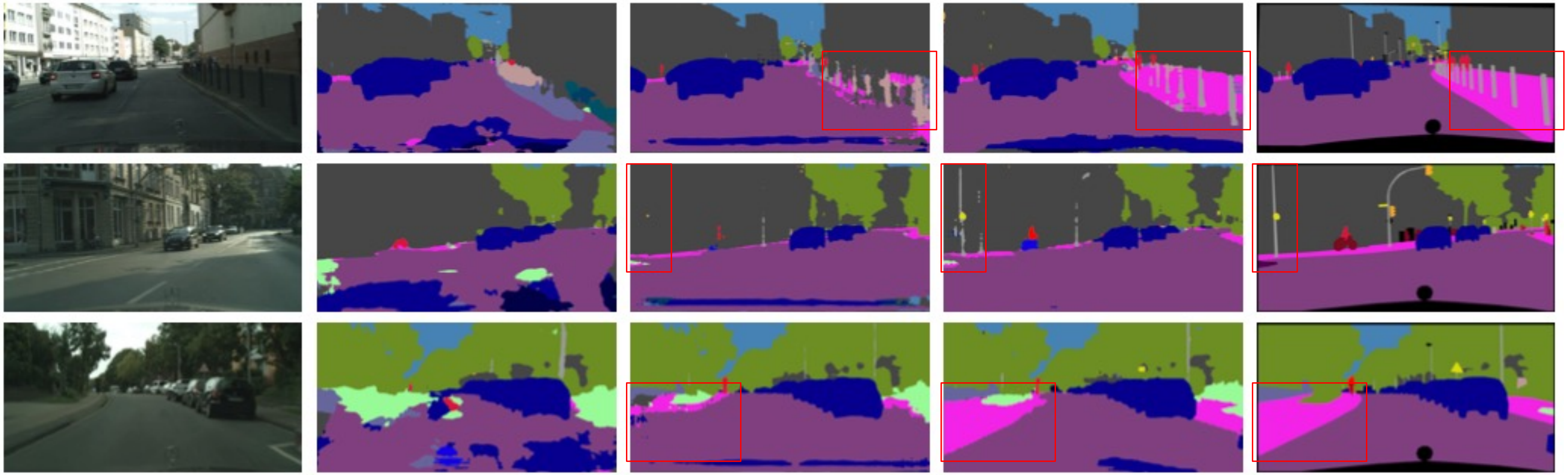
TABLE II: Unsupervised Adaptation Model performance from SYNTHIA [36] to Cityscapes [7].

Method	road	side.	build.	light	sign	vege.	sky	person	rider	car	bus	motor	bike	mIoU	Gain
Source only	55.6	22.6	63.8	5.8	13.4	72.9	78.4	51.3	15.1	33.6	21.2	13.9	22.9	36.2	-
CLAN (19′) [16]	81.3	37.0	80.1	16.1	13.7	78.2	81.5	53.4	21.2	73.0	32.9	22.6	30.7	47.8	11.6
SSF-DAN (19′) [17]	**84.6**	41.7	**80.8**	11.5	14.7	80.8	85.3	57.5	21.6	**82.0**	36.0	19.3	**34.5**	50.0	13.8
CrCDA (20′) [29]	86.2	44.9	79.5	9.4	11.8	78.6	86.5	57.2	26.1	76.8	39.9	21.5	32.1	50.0	13.8
UIDA (20′) [33]	84.3	37.7	79.5	9.2	8.4	80.0	84.1	57.2	23.0	78.0	**38.1**	20.3	36.5	48.9	12.7
LSE (20′) [34]	82.9	**43.1**	78.1	9.1	14.4	77.0	83.5	58.1	**25.9**	71.9	38.0	**29.4**	31.2	49.4	13.2
COINet	83.1	42.3	79.2	**19.8**	**25.7**	**82.1**	**85.6**	**59.2**	24.5	81.3	33.7	28.3	26.8	**51.6**	**15.4**

Experiment

Department of Electrical Engineering
香港城市大學
City University of Hong Kong

- Perform well in small objects.
- Preserve high performance for well-aligned categories.



Target image | Source only | AdaptSeg | Ours | Ground Truth

Experiment

- Cluster center distance measures the degree of alignment.
- Our method achieves lower distance, indicating better feature distribution alignment.
- Ablation study validates the effectiveness of each key component.

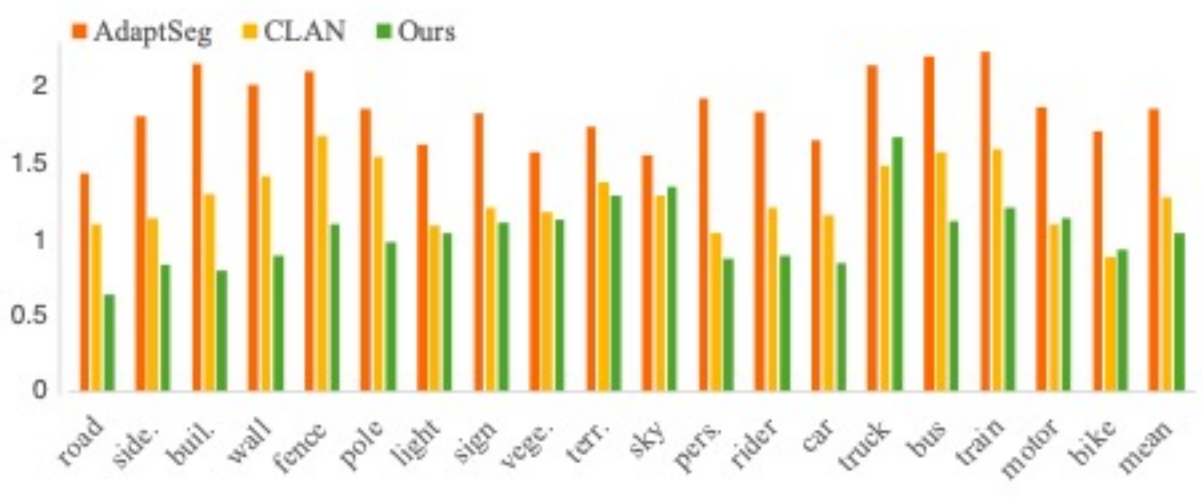


Fig. 6: Quantitative analysis for the feature alignment. We show each Cluster Center Distance of three approaches.

TABLE III: Ablation Studies of each component.

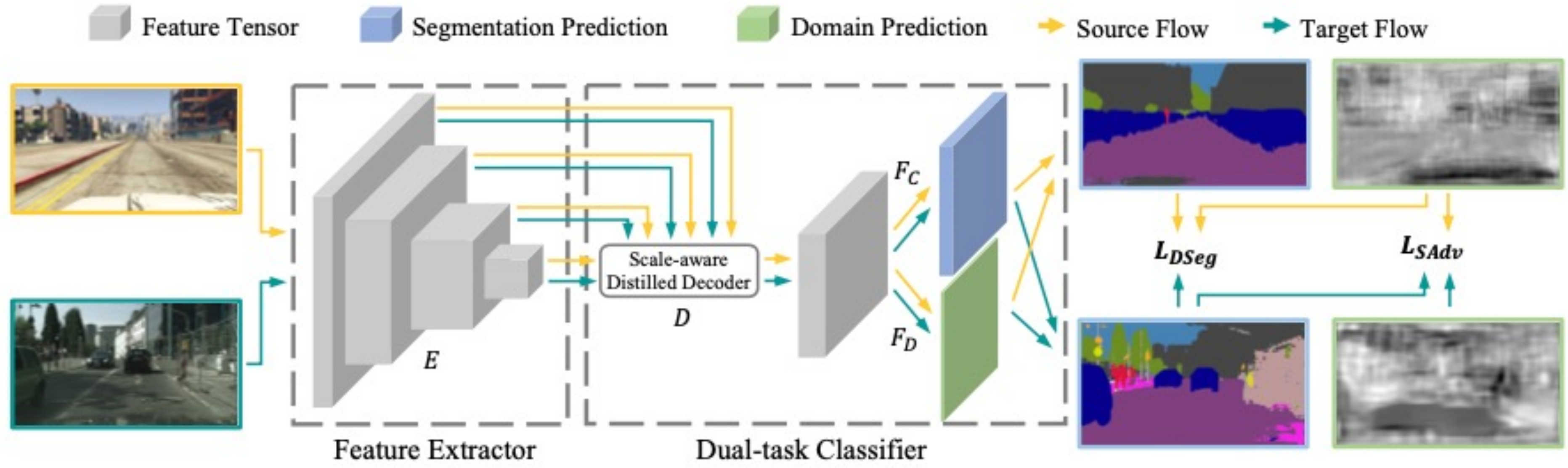
DTC	$\mathcal{L}_{DSeg}$	$\mathcal{L}_{SAdv}$	mIoU	Gain(%)
			45.5	-
✓			46.9	1.4
✓	✓		48.2	2.7
✓		✓	47.7	2.2
✓	✓	✓	49.2	3.7

Overview

Department of Electrical Engineering
香港城市大學
City University of Hong Kong

Conclusion

- Propose a co-interactive network (COINet) addressing unsupervised domain adaptation problem.
- Scale-aware Distilled Decoder fuses multi-scale feature maps smoothly.
- Co-interactive loss promotes two tasks with each other.
- Comprehensive experiments demonstrate the effectiveness of these modules.



Thank you very much for your attention!