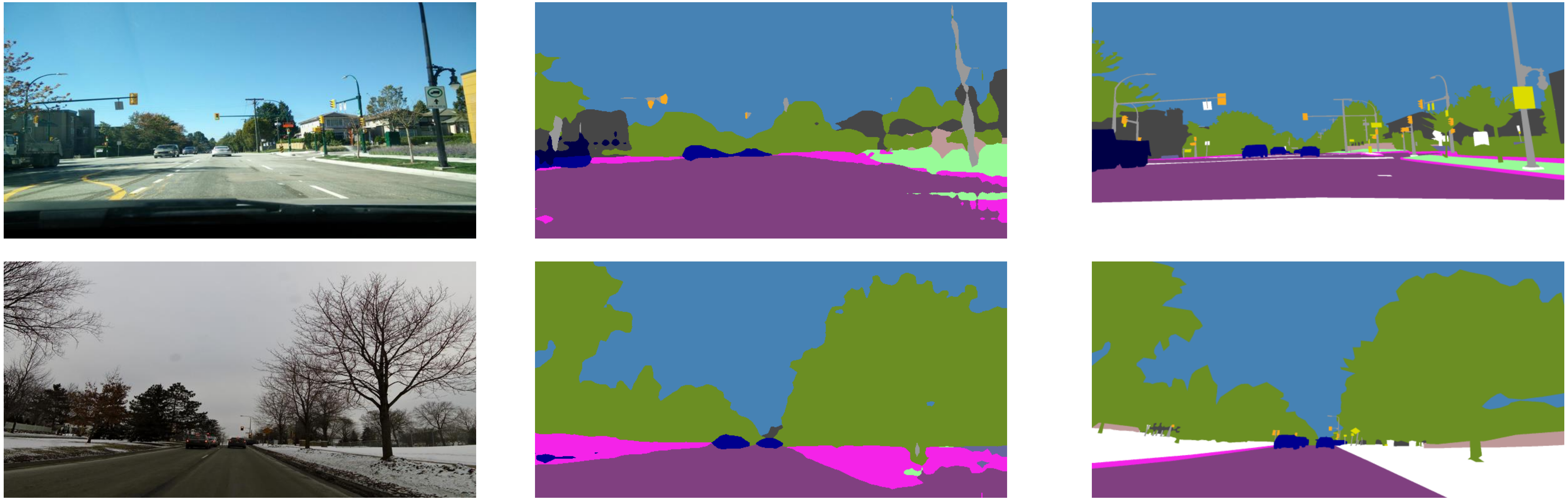
Unsupervised Domain Adaptation with Multiple Domain Discriminators and Adaptive Self-Training

Teo Spadotto, Marco Toldo, Umberto Michieli, Pietro Zanuttigh

January 15th, 2021

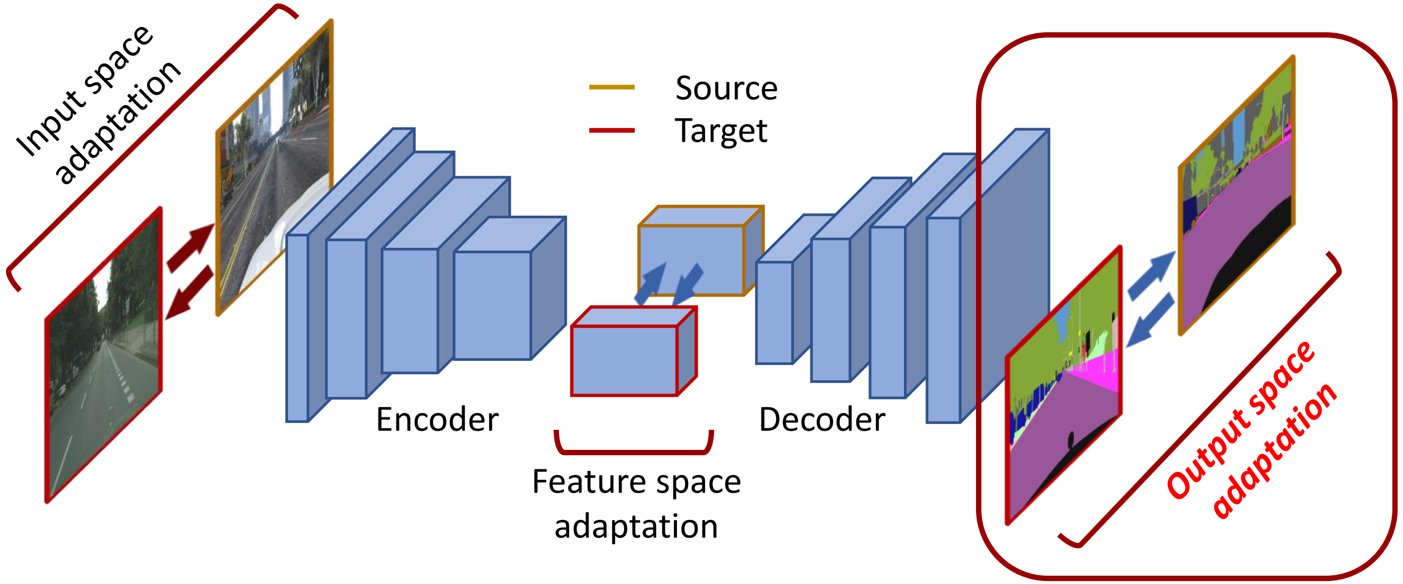
Semantic Segmentation



RGB Prediction Ground Truth

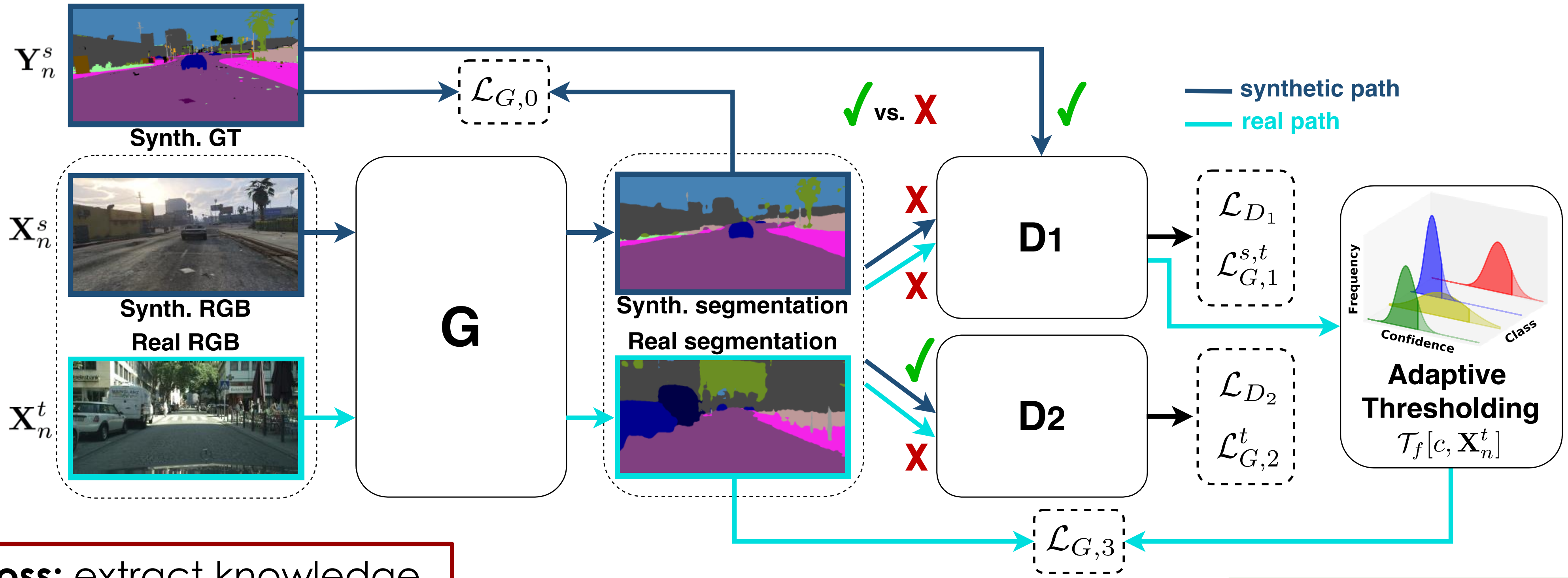
- Dense labeling task: assign a class label to each single pixel in an image
- Nowadays solved with deep learning, typically auto-encoder CNNs
- Large generic datasets for training are available but it is challenging to get data specific to the task

Unsupervised Domain Adaptation



- Labeling data is available only for the source dataset
- Goal: achieve good results on a different (but related) target dataset
- Domain shift limits performance: need for adaptation
- The adaptation can be performed at input, feature or ***output space***

Output Level Adaptation



Task Loss: extract knowledge from source supervision

$$\mathcal{L}_{full} = \mathcal{L}_{G,0} + w_1^{s,t}\mathcal{L}_{G,1}^{s,t} + w_2^{t}\mathcal{L}_{G,2}^{t} + w_3\mathcal{L}_{G,3}$$

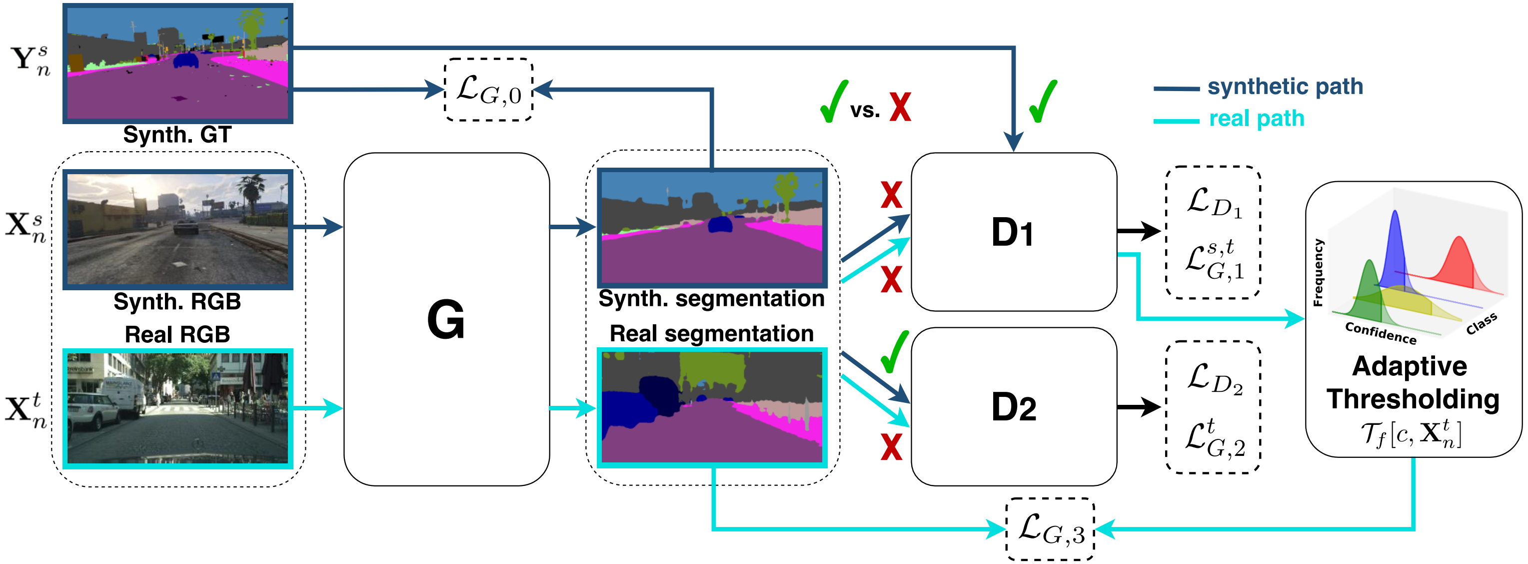
Self-Training: use confident target pseudo labels for self-taught supervision

Output Level Adversarial Modules: align prediction maps between source and target domains w.r.t. source GT maps

Adversarial Adaptation

Two Adversarial Adaptations

- D1: GT vs Prediction
  ⇨ indirect domain alignment
  ⇨ both source and target predictions can be used
- D2: Source vs Target
  ⇨ direct domain alignment



✗ **Fake:** Source&Target pred. ✓ **Real:** Source GT

$$\mathcal{L}_{G,1}^{s,t} = -\sum_{p \in \mathbf{X}_n^{s,t}} \log(D_1(G(\mathbf{X}_n^{s,t}))^{(p)}) \qquad \mathcal{L}_{D_1} = -\sum_{p \in \mathbf{X}_n^{s,t}} \log(1 - D_1(G(\mathbf{X}_n^{s,t}))^{(p)}) + \log(D_1(\mathbf{Y}_n^s)^{(p)})$$

$$\mathcal{L}_{G,2}^{t} = -\sum_{p \in \mathbf{X}_n^{t}} \log(D_2(G(\mathbf{X}_n^{t}))^{(p)}) \qquad \mathcal{L}_{D_2} = -\sum_{p \in \mathbf{X}_n^{s,t}} \log(1 - D_2(G(\mathbf{X}_n^{t}))^{(p)}) + \log(D_2(G(\mathbf{X}_n^{s}))^{(p)})$$

✗ **Fake:** Target prediction ✓ **Real:** Source prediction

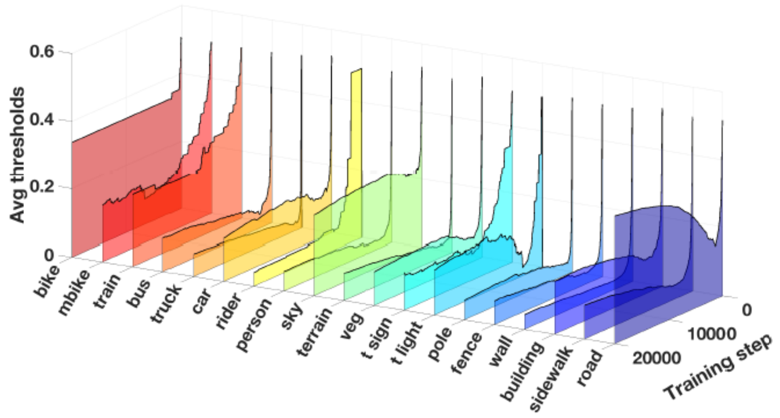
Self-Training

Loss: Weighted cross-entropy with pseudo labels

$$\mathcal{L}_{G,3} = -\sum_{p\in \mathbf{X}_n^t}\sum_{c\in\mathcal{C}} \mathcal{M}_f^{(p)} \cdot W_c^s \cdot \hat{\mathbf{Y}}_n^{(p)}[c] \cdot \log\left(G(\mathbf{X}_n^t)^{(p)}[c]\right)$$

Adaptive Threshold based on class-wise distribution

$$\mathcal{T}_f[c,\mathbf{X}_n^t] = \mathcal{Q}_f(D_1(G(\mathbf{X}_n^t)[c]))$$



$$\mathcal{M}_f^{(p)} = \begin{cases}1 & \text{if } \left(D_1(G(\mathbf{X}_n^t))^{(p)} > \mathcal{T}_f[c,\mathbf{X}_n^t]\right) \wedge \left(\hat{\mathbf{Y}}_n^{(p)}[c]=1\right)\\ 0 & \text{otherwise}\end{cases}$$

Confidence based mask computed from discriminator's output

- Use highly confident network predictions for self-teaching on target dataset
- Use discriminator’s output as a confidence measure
- *Class* and *step* adaptive thresholding dynamically updated during training

Quantitative Results

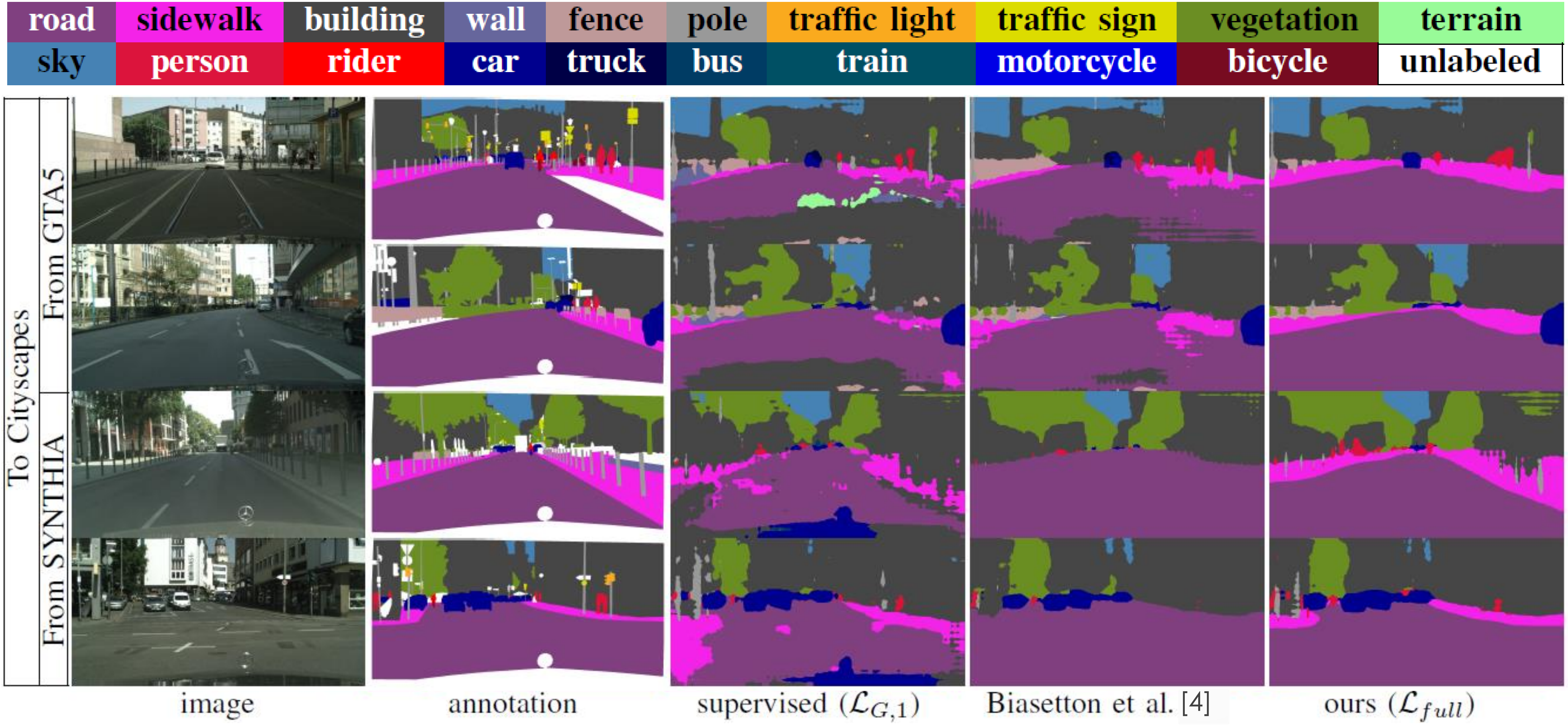
Method	mIoU	
	GTA5→CS	SYNTHIA→CS
Supervised (baseline)	31.9	28.8
Hoffman et al. [1]	27.1	20.1
Hung et al. [2]	29.0	29.4
Zhang et al. [3]	28.9	29.0
Biasetton et al. [4]	30.4	30.2
Michieli et al. [5]	33.3	31.3
Ours	**35.1**	**34.6**

Method	mIoU	
	GTA5→MAP	SYNTHIA→MAP
Supervised (baseline)	37.8	31.1
Hung et al. [2]	34.4	27.0
Biasetton et al. [4]	35.2	28.2
Michieli et al. [5]	38.5	32.0
Ours	**41.9**	**34.9**

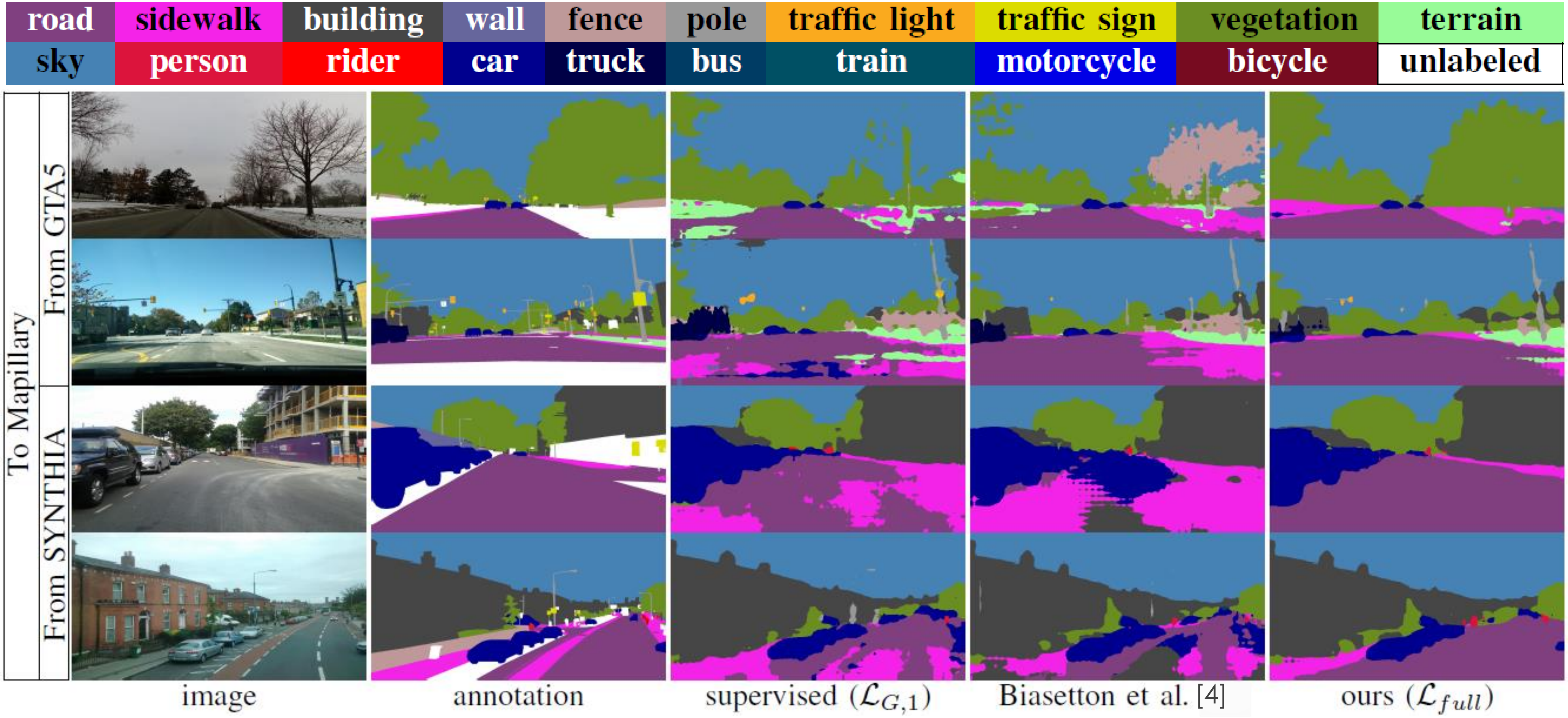
[1] J. Hoffman et al., "FCNs in the wild: Pixel-level adversarial and constraint-based adaptation," arXiv, 2016
[2] W.-C. Hung et al., "Adversarial learning for semi-supervised semantic segmentation,", BMVC, 2018
[3] Y. Zhang et al., "Curriculum domain adaptation for semantic segmentation of urban scenes," ICCV, 2017
[4] M. Biasetton et al., "Unsupervised Domain Adaptation for Semantic Segmentation of Urban Scenes," CVPRW, 2019
[5] U. Michieli et al., "Adversarial learning and self-teaching techniques for domain adaptation in semantic segmentation," IEEE Transaction on Intelligent Vehicles, 2020

- 2 Source synthetic datasets (GTA5 or SYNTHIA)
- 2 Target real-world datasets (Cityscapes and Mapillary)
- Results computed using a DeepLab-v2 network with Resnet-101 as encoder

Visual Results (Cityscapes)



Visual Results (Mapillary)



Conclusions

- We presented a novel adversarial learning and self-teaching scheme for unsupervised domain adaptation
- Domain discriminators capture both source vs target and ground truth vs prediction statistics
- Adaptive self-training strategy
- Experimental results on synthetic to real adaptation show that the approach outperforms competing schemes using output-level adaptation

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZION

Paper webpage:
https://lttm.dei.unipd.it/paper_data/semanticDA