UNSUPERVISED DOMAIN ADAPTATION FOR URBAN SCENES SEGMENTATION

Biasetton M., Michieli U., Agresti G., Zanuttigh P. - University of Padova
{biasetto, michieli, agrestig, zanuttigh}@dei.unipd.it

Università degli Studi di Padova

Abstract

The semantic understanding of urban scenes is one of the key components for autonomous driving systems. Deep neural networks require huge amount of labeled data, which is difficult and expensive to acquire. A recent workaround is to exploit synthetic data but differences between real and synthetic scenes limit the performance. We propose an unsupervised domain adaptation strategy from a synthetic supervised training to real data.
Experimental results demonstrate that the proposed approach is able to adapt a network trained on synthetic datasets to a real one.

Dataset

SOURCE (SYNTHETIC)

GTA

~25k images
High quality
Car viewpoints

SYNTHIA

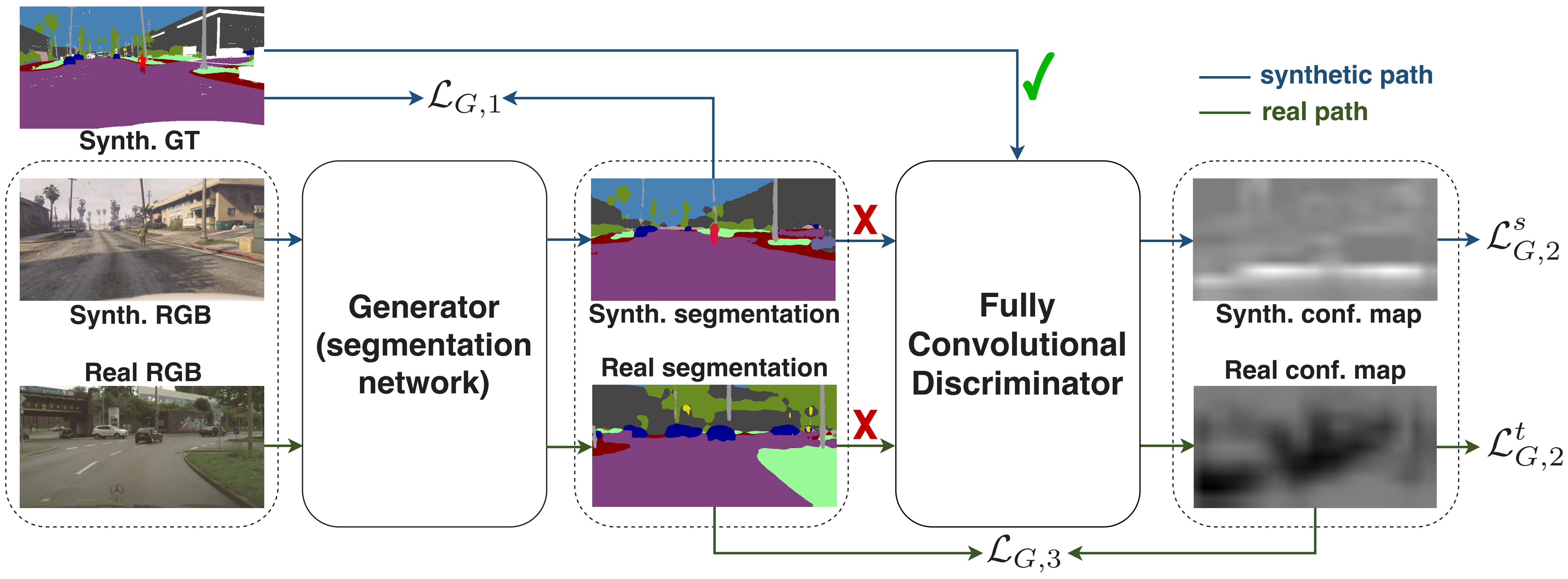
~9k images
Medium quality
Different viewpoints

TARGET (REAL WORLD)

CITYSCAPES

~3k images
Car viewpoints

Proposed Approach



- $\mathcal{L}_{G,1}$: standard cross-entropy loss (on source dataset)
- Trained end-to-end minimizing $\mathcal{L}_{full} = \mathcal{L}_{G,1} + w^{s,t}\mathcal{L}^{s,t}_{G,2} + w'\mathcal{L}_{G,3}$

Adversarial Training

$$\mathcal{L}^{s,t}_{G,2} = -\log(D(G(\mathbf{X}^{s,t}_n)))$$

$$\mathcal{L}_D = -\log(1 - D(G(\mathbf{X}^{s,t}_n))) + \log(D(\mathbf{Y}^s_n))$$

s: source dataset
t: target dataset

Self-Taught Loss

Predictions of G are more reliable where D marks them as GT with high accuracy

$$\mathcal{L}_{G,3} = -I_{T_u} \cdot W^t_c \cdot \hat{\mathbf{Y}}_n[c] \cdot \log\left(G(\mathbf{X}^t_n)[c]\right)$$

c: classes

class weighting

threshold on confidence maps from D

Results

From GTA	road	sidewalk	building	wall	fence	pole	t light	t sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
Ours ($\mathcal{L}_{G,1}$ only)	45.3	20.6	50.1	9.3	12.7	19.5	4.3	0.7	81.9	21.1	63.3	52.0	1.7	77.9	26.0	39.8	0.1	4.7	0.0	27.9
Ours ($\mathcal{L}_{full}$) [1]	54.9	23.8	50.9	16.2	11.2	20.0	3.2	0.0	79.7	31.6	64.9	52.5	7.9	79.5	27.2	41.8	0.5	10.7	1.3	**30.4**
Hung et al. [2]	81.7	0.3	68.4	4.5	2.7	8.5	0.6	0.0	82.7	21.5	67.9	40.0	3.3	80.7	34.2	45.9	0.2	8.7	0.0	29.0

From SYNTHIA	road	sidewalk	building	wall	fence	pole	t light	t sign	veg	sky	person	rider	car	bus	mbike	bike	mIoU
Ours ($\mathcal{L}_{G,1}$ only)	10.3	20.5	35.5	1.5	0.0	28.9	0.0	1.2	83.1	74.8	53.5	7.5	65.8	18.1	4.7	1.0	25.4
Ours ($\mathcal{L}_{full}$) [1]	78.4	0.1	73.2	0.0	0.0	16.9	0.0	0.2	84.3	78.8	46.0	0.3	74.9	30.8	0.0	0.1	**30.2**
Hung et al. [2]	72.5	0.0	63.8	0.0	0.0	16.3	0.0	0.5	84.7	76.9	45.3	1.5	77.6	31.3	0.0	0.1	29.4

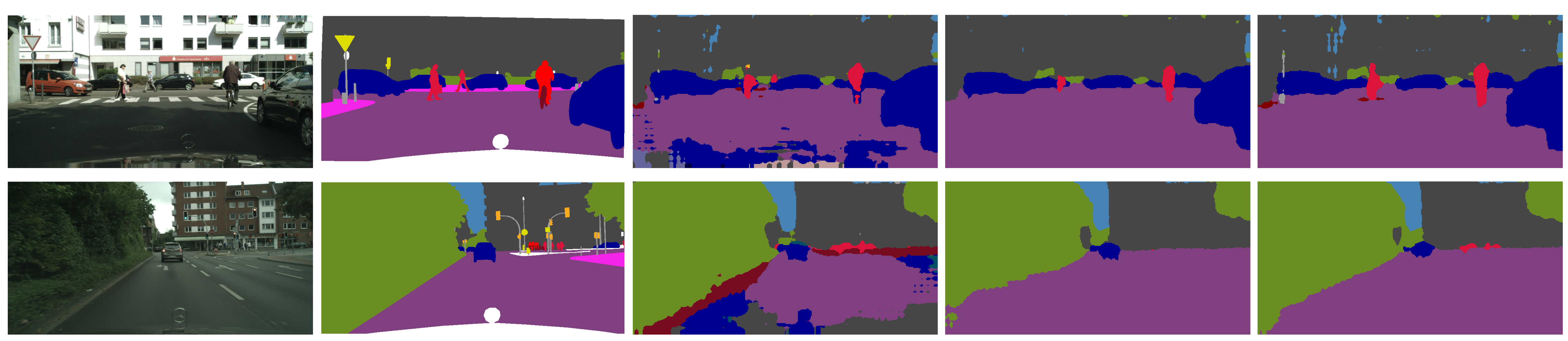


image | annotation | baseline ($\mathcal{L}_{G,1}$) | Hung et al. [2] | $\mathcal{L}_{full}$ [1]

References

[1] Biasetton M., Michieli U., Agresti G., Zanuttigh P., "Unsupervised Domain Adaptation for Semantic Segmentation of Urban Scenes", CVPR Workshop on Autonomous Driving (WAD), 2019.

[2] Hung W., Tsai Y., Liou Y., Lin Y., Yang M., "Adversarial Learning for Semi-Supervised Semantic Segmentation", BMVC, 2018.