Cross-Domain Transfer Learning with CoRTe: Consistent and Reliable Transfer from Black-Box to Lightweight Segmentation Model
Supplementary Material

Claudia Cuttano Antonio Tavera Fabio Cermelli Giuseppe Averta
Barbara Caputo
name.surname@polito.it
Politecnico di Torino, Corso Duca degli Abruzzi, 24 — 10129 Torino, ITALIA

1. Source Model Generation

To simulate the black-box source model, we train a complex semantic segmentation model on the source domain. Specifically, the source model $\mathcal{S}$ is based on DAFormer [1]. It consists of a MiT-B5 encoder [5] and a context-aware feature fusion decoder [1]. The network is trained with pairs of $\{x_s, y_s\}$, where $x_s \in \mathcal{X}_S$ corresponds to a source image and $y_s \in \mathcal{Y}_S$ is its respective ground truth. A standard cross-entropy loss is used to train the model for 40k iterations, with batches of 2 images each. We adopt two strategies introduced in [1] to limit overfitting and stabilize the training: (i) Rare Class Sampling (RCS) and (ii) Thing-Class ImageNet Feature Distance. The first attempts at mitigating the class unbalance by sampling more frequently images containing rare classes, while the latter regularizes the distance between the bottleneck features of the segmentation network and the bottleneck features of the ImageNet model. We refer to the companion paper for the remaining training hyperparameters.

2. Networks Comparison

Tab. 1 reports the details of the networks employed as source [1] and target [5] models[1]. More specifically, it is worth noticing that the first is considerably computationally complex, consisting of an encoder with 81.4M parameters and a decoder with 3.7M parameters. In contrast our target architecture, which corresponds to SegFormer-B0, only has 3.8M parameters. This difference in size and complexity, with the source model being 22.4x larger than SegFormer-B0, has an impact on the throughput for inference. Indeed, SegFormer-B0 requires a significantly lower number of flops (18 GFlops vs. 274, $\approx$ 15 times less), thus boosting the inference throughput. In our experiments, carried out on a single NVIDIA TITAN RTX with 24 GB memory, DAFormer processes 5.6 images per second, whereas the target model has a throughput of 30.3 images per second. The combination of limited inference time and low complexity makes the target model a suitable option for real-time applications on low resources hardware.

[1]The code will be released upon acceptance.

Method	Encoder	Decoder	#Params encoder	#Params decoder	Speed (img/s)	Flops (G)
Source model	MIT-B5 [5]	DAFormer [1]	81.4M	3.7M	5.6	274
Target model	MIT-B0 [5]	SegFormer [5]	3.4M	0.4M	30.3	18

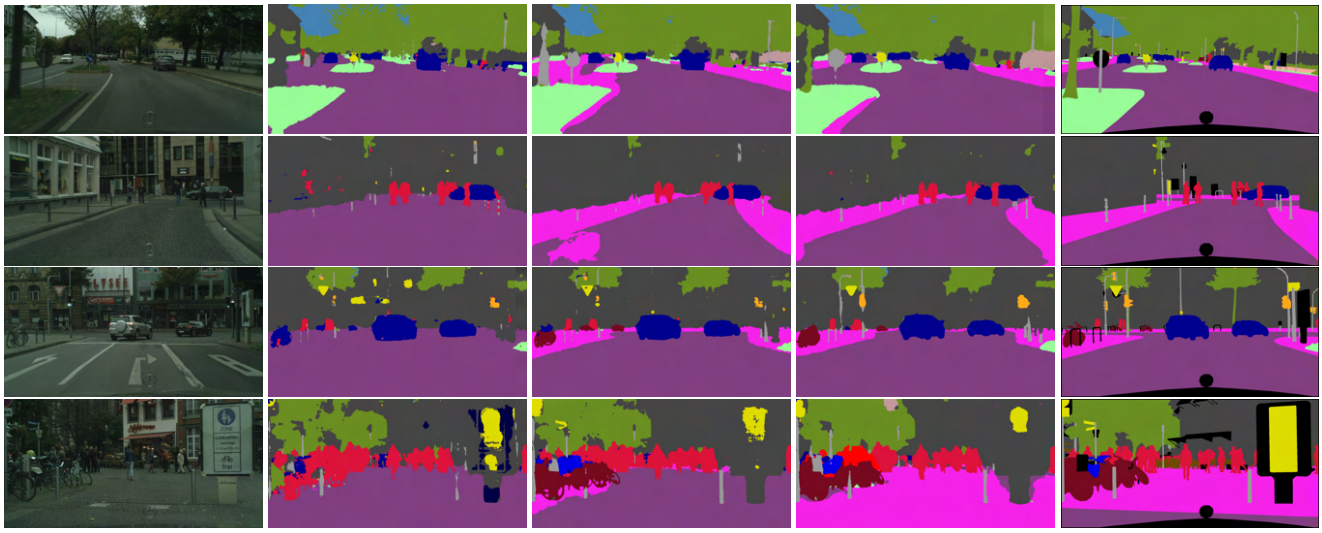
Table 1: #Parameters, Speed (img/s) and #Flops (G) for source (DAFormer) and target (SegFormer-B0) networks. #Flops computed on Cityscapes data resized to 1024×512.

3. Qualitative Analysis

Fig. 1 and Fig. 2 provide a qualitative comparison between our solution and some of the methods considered in the main paper. More specifically, Fig. 1 confirms the CoRTe superiority when applied to the GTA→Cityscapes scenario, providing good predictions across all classes, especially stuff, and being overall the closest to the ground truth. The only exception is for the *traffic sign*, on which both HRDA [2] and DACS [4], as also confirmed by the experiments in Tab.1 of the main paper, provide higher-level predictions. Similar remarks can be made for the results obtained in the SYNTHIA→Cityscapes scenario. Indeed, Fig. 2 confirms the numerical results obtained in Tab.2, showing the superior ability of CoRTe to segment fine objects and to recognize rare classes.

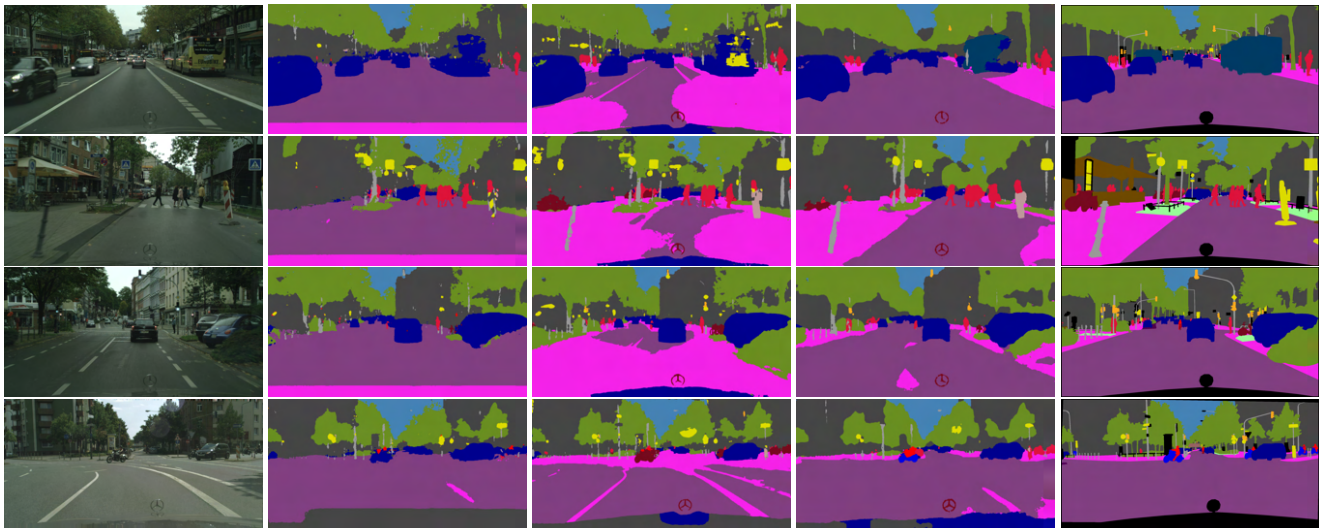
4. Real-to-Real

In addition to the traditional UDA synthetic-to-real settings, we evaluate CoRTe's performance also on the real-to-real Cityscapes→ACDC task, where the Adverse Condition Dataset (ACDC) [3] serves as the unlabeled target domain. We report the results in tab 2. With CoRTe (47.7%), we outperform all the baselines by a clear margin, confirming the



(a) Image (b) DACS [4] (c) HRDA [2] (d) CoRTe (ours) (e) Ground Truth

Figure 1: *Visual comparison for GTA → Cityscapes.*



(a) Image (b) DACS[4] (c) HRDA [2] (d) CoRTe (ours) (e) Ground Truth

Figure 2: *Visual comparison for SYNTHIA → Cityscapes.*

potential of CoRTe for real-world applications.

References

[1] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9924–9935, June 2022. 1

[2] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 372–391. Springer, 2022. 1, 2, 3

Method	SF	T→S	Road	Sidewalk	Building	Wall	Fence	Pole	T.Light	T.sight	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorbike	Bicycle	mIoU
Source-only	✗	✗	67.1	8.1	60.0	28.2	20.8	32.3	54.7	20.5	59.7	19.2	76.3	50.0	22.2	71.1	38.4	23.0	27.6	30.4	2.2	37.5
No adapt	✗	✗	65.1	15.8	55.1	12.9	17.0	28.1	44.2	36.5	68.7	21.1	76.7	29.6	5.3	60.0	29.2	16.1	34.5	12.8	15.3	33.9
DACS[4]	✗	✗	70.8	42.7	66.0	25.1	13.4	40.9	51.6	37.5	70.9	31.1	81.7	30.1	9.1	65.1	45.6	67.6	31.7	18.3	16.8	42.9
HRDA[2]	✗	✗	81.9	40.0	63.9	27.7	15.9	44.0	58.5	44.2	71.3	34.4	81.7	26.6	15.6	64.1	48.4	51.3	54.9	21.1	11.2	45.1
KL-DIV	✓	✓	69.1	33.7	74.1	29.9	15.8	42.6	55.2	42.2	72.6	32.8	71.2	40.1	9.3	73.5	60.9	76.5	53.7	12.6	20.4	46.6
CoRTe	✓	✓	71.1	41.5	76.9	31.5	17.8	45.4	38.1	42.1	70.9	33.5	69.8	42.0	10.8	70.8	53.0	81.2	69.7	16.1	23.3	**47.7**

Table 2: mIoU for Cityscapes→ACDC. **SF** denotes the *Source-Free* methods, whereas **T→S** refers to the methods that leverage the black-box model for training the target network. All methods use B0 as encoder, while *Source-only* uses B5.

[3] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021. 1

[4] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021. 1, 2, 3

[5] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12077–12090. Curran Associates, Inc., 2021. 1