Contrastive Model Adaptation for Cross-Condition Robustness in Semantic Segmentation

Supplementary Material

David Bruggemann Christos Sakaridis Tim Brödermann Luc Van Gool
ETH Zurich, Switzerland
{brdavid, csakarid, timbr, vangool}@vision.ee.ethz.ch

A. Training Details

We provide additional training details in this section. All models were trained using Automatic Mixed Precision on a single consumer TITAN RTX GPU.

A.1. Optimization

We train for 10k iterations for all datasets. During the first 2.5k iterations, gradient backpropagation from the projection head to the backbone is stopped, to avoid noisy weight updates on the pretrained backbone weights. We use an AdamW [4] optimizer with weight decay 0.01 and a linear learning rate decay with linear warm-up for the first 1500 iterations. The chosen learning rates for CMA are 1×10^{-5} (SegFormer-based) and 2×10^{-8} (DeepLabv2-based), using 1 (SegFormer-based) or 2 (DeepLabv2-based) adverse-reference image pairs per batch. For the weights of the projection head, the learning rate is multiplied by a factor of 10, as it is initialized randomly. The individual loss weights are $\lambda_{\text{ent}} = 0.01$ and $\lambda_{\text{cdc}} = 1.0$ for SegFormer-based CMA and $\lambda_{\text{ent}} = 1.0$ and $\lambda_{\text{cdc}} = 1.0$ for DeepLabv2-based CMA.

A.2. CDC Loss Hyperparameters

For partitioning the dense feature map into patches, a 7×7 grid is used. Positives and negatives are encoded with an exponential moving average network using a momentum of 0.9999. Negatives are then stored in a queue of size 65536. The temperature τ of the InfoNCE loss varies depending on the dataset, we use 0.3 for ACDC, 0.03 for Dark Zurich, 0.3 for RobotCar, and 0.1 for CMU. For eliminating unreliable patches in the confidence modulation, a threshold of 0.2 is used throughout.

A.3. Data Handling

Training data augmentation consists of random cropping to square shape—such that the crop size coincides with the shorter sidelength of the input—and random horizontal flipping. Note that no resizing is applied.

Test predictions are generated through a sliding window approach. The windows are square, with a sidelength equal to the shorter input sidelength. Consecutive windows overlap for between 0% and 50% of their sidelength, depending on the input aspect ratio.

A.4. Baselines

We reimplemented the baselines TENT [9], HCL [3], and URMA [8] for a fair comparison, carefully following their published code for reference. The learning rate, loss weights, as well as method-specific hyperparameters were separately tuned for each method. For SegFormer-based HCL, we had to introduce random subsampling of anchors for the contrastive loss, due to prohibitive memory demands otherwise.

B. ACG Benchmark

The purpose of ACG is to provide a generalization benchmark estimating a model's adverse-condition robustness to diverse inputs, whereby the model is trained on another dataset such as ACDC [7], Mapillary Vistas [5], *etc*. The evaluation benchmark consists of training, validation, and test images from the public datasets Wilddash2 [12], BDD100K [11], Foggy Zurich [2], and Foggy Driving [6]. Models trained on these four datasets can therefore not be evaluated on ACG.

B.1. Construction

We constructed the ACG benchmark as follows:

1. For each of Wilddash2, BDD100K, Foggy Driving, and Foggy Zurich, we inspected all images with public semantic segmentation annotations and extracted images depicting fog, night, rain, or snow—or a combination thereof. For Wilddash2 we only considered

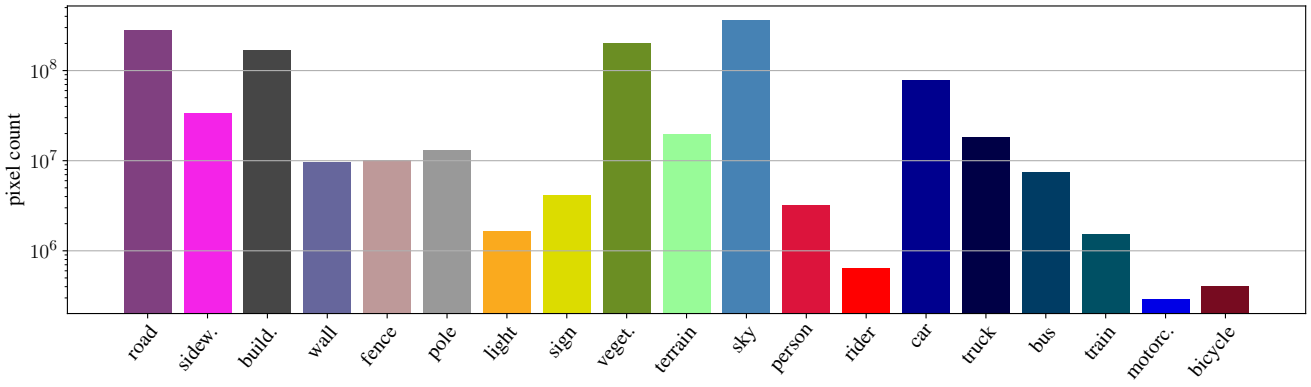

Figure B-1. Number of annotated pixels per class in the ACG benchmark.

images taken in Europe and North America, to confine the geographical domain shift.

2. For every selected image, we checked the quality of the semantic segmentation labels. Images with unreliable ground-truth were eliminated. For BDD100K, this step eliminated a majority of images.

3. For selected images from Foggy Driving or Foggy Zurich, we checked for potential geographical overlap with ACDC, since all three datasets were recorded in the greater area of Zurich. Images with geographical overlap were eliminated.

Through these three steps, we selected 919 images from a pool of 15173. However, upon closer inspection we observed that there were no rainy images containing the "train" class, which would prevent condition-wise evaluation (see Sec. B.2). We therefore collected 3 copyright-free images from the web depicting rainy street scenes with trains or trams and finely annotated the pixels of class "train". In total, ACG consists of 922 adverse-condition images with high-quality ground-truth annotations.

The ground-truth annotations follow the labeling convention of Cityscapes [1], consisting of 19 classes. For Wilddash2, semantic classes were mapped back to Cityscapes classes according to the mapping given by [12].

B.2. Data Splits

We divide the 922 images into 4 subsets, classified by condition, to enable condition-wise evaluation. Each image depicting a nighttime scene was assigned to ACG-night, regardless of the weather condition. For daytime images, each image was assigned to either ACG-fog, ACG-rain, or ACG-snow, depending on the dominant weather condition. The resulting subset sizes are 121 for ACG-fog, 225 for ACG-rain, 276 for ACG-snow, and 300 for ACG-night.

Table C-1. Effect of individual training losses on ACDC validation performance.

	CMA	w/o $\mathcal{L}_{\text{ent}}$	w/o $\mathcal{L}_{\text{st}}$	w/o $\mathcal{L}_{\text{cdc}}$
ACDC val mIoU	67.2	66.7	57.7	60.1

Table C-2. Sensitivity of CMA to the confidence threshold (default value of 0.2).

confidence threshold	0	0.1	0.2	0.3	0.4	0.5
ACDC val mIoU	66.8	67.1	67.2	67.0	67.0	66.7

B.3. Class Distribution

The numbers of annotated pixels per class are shown in Fig. B-1. Importantly, each class is also represented within each condition-subset.

C. Additional Ablations

Effect of Entropy and Self-Training Losses. We show in Table C-1 the effect of the individual training losses on ACDC validation performance. Omitting either the self-training or our CDC loss leads to a large performance drop, while omitting the entropy loss has a more minor effect.

Sensitivity to Patch Confidence Threshold. Table C-2 shows the sensitivity of CMA to the confidence threshold value, which is set to 0.2 in Eq. (3).

D. Condition-Wise ACDC Performances

In Tables D-1, D-2, D-3, D-4 we report the test set results for the condition-wise evaluations on ACDC-fog, ACDC-night, ACDC-rain, and ACDC-snow. For all the methods, Cityscapes is used as the source dataset and the full ACDC training set as the target dataset. On ACDC-night, ACDC-rain, and ACDC-snow, CMA outperforms all other methods, while being second best on ACDC-fog.

E. Source Model Predictions

Fig. E-1 shows SegFormer source model predictions on corresponding reference (normal condition, left) and target (adverse condition, right) images of the ACDC dataset. Overall, the Cityscapes-trained source model produces more accurate predictions on the reference images.

F. Qualitative Results

We provide more qualitative segmentation results on randomly selected ACDC validation images in Fig. F-1.

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2

[2] Dengxin Dai, Christos Sakaridis, Simon Hecker, and Luc Van Gool. Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. *IJCV*, 128(5):1182–1204, 2020. 1

[3] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. In *NeurIPS*, 2021. 1, 4, 5, 6

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[5] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The Mapillary Vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 1

[6] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, 2018. 1

[7] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The Adverse Conditions Dataset with Correspondences for semantic driving scene understanding. In *ICCV*, 2021. 1

[8] Prabhu Teja S and Francois Fleuret. Uncertainty reduction for model adaptation in semantic segmentation. In *CVPR*, 2021. 1, 4, 5, 6

[9] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 1, 4, 5, 6

[10] 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 *NeurIPS*, 2021. 4, 5

[11] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 1

[12] Oliver Zendel, Matthias Schörghuber, Bernhard Rainer, Markus Murschitz, and Csaba Beleznai. Unifying panoptic segmentation for autonomous driving. In *CVPR*, 2022. 1, 2

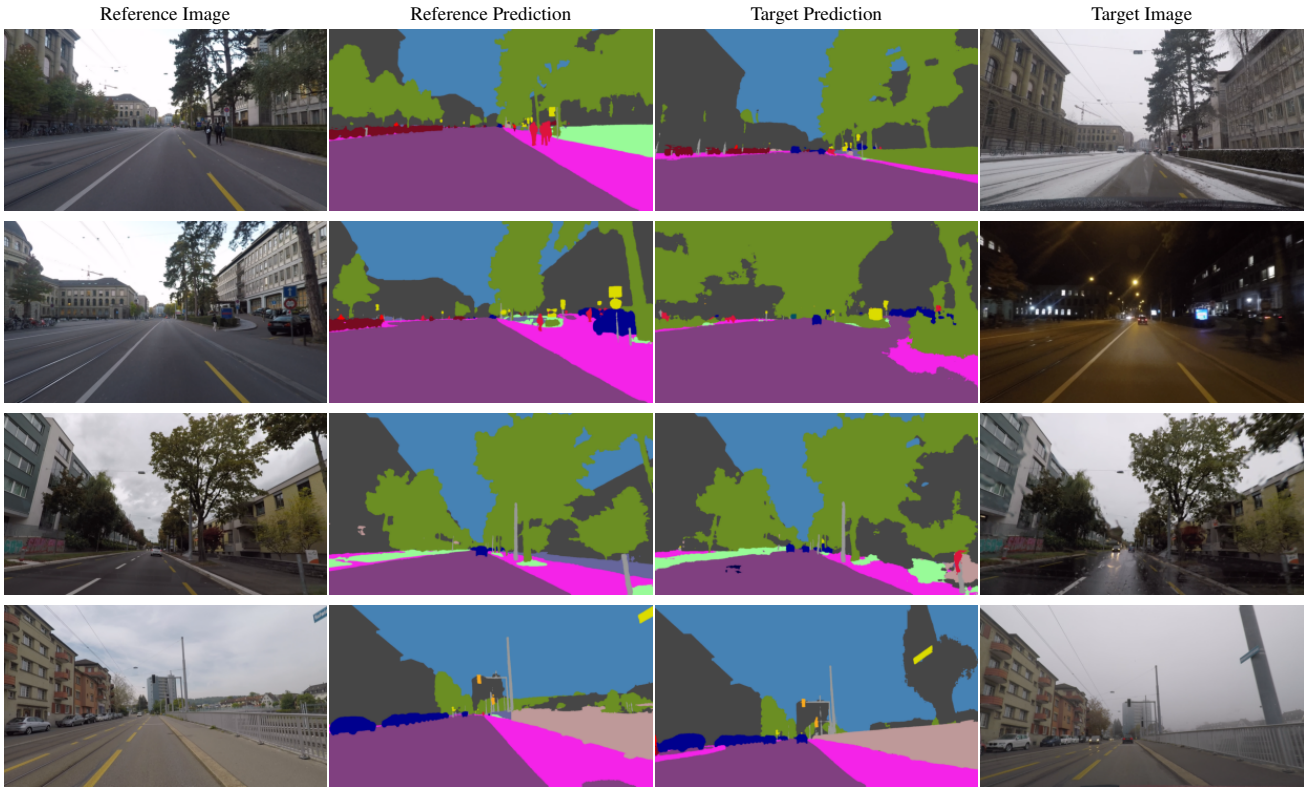

Figure E-1. Comparison of SegFormer predictions on pairs of reference and target images.

Table D-1. Comparison to the state of the art in model adaptation on Cityscapes→ACDC, with reported performance on the ACDC-fog test set.

Method		ACDC-fog IoU ↑																			
		road	sidew.	build.	wall	fence	pole	light	sign	veget.	terrain	sky	person	rider	car	truck	bus	train	motorc.	bicycle	**mean**
Source model	SegFormer [10]	87.8	60.7	73.1	44.5	30.1	42.1	52.3	64.4	81.4	68.8	93.4	51.1	53.2	78.4	66.0	39.7	75.1	43.2	**47.4**	60.7
TENT [9]		83.0	61.1	68.2	44.1	30.4	44.1	52.1	64.4	81.1	69.3	89.9	50.9	54.7	78.6	67.1	39.5	75.4	45.9	47.1	60.4
HCL [3]		88.5	63.2	79.8	45.3	30.6	44.7	53.7	65.9	81.8	69.6	95.5	52.5	55.0	79.4	68.0	40.7	74.0	40.7	46.9	61.9
URMA [8]		89.3	61.8	87.9	51.4	**36.3**	**52.3**	58.1	**67.9**	85.7	**71.8**	97.2	54.5	**62.5**	**82.3**	**70.6**	**62.0**	82.0	**52.9**	36.2	**66.5**
CMA		**93.5**	**75.3**	**88.6**	**53.4**	33.0	52.2	**58.2**	67.0	**86.9**	71.5	**97.8**	**55.6**	42.0	80.4	70.0	54.8	**83.3**	43.0	37.4	65.5

Table D-2. Comparison to the state of the art in model adaptation on Cityscapes→ACDC, with reported performance on the ACDC-night test set.

Method		ACDC-night IoU ↑																			
		road	sidew.	build.	wall	fence	pole	light	sign	veget.	terrain	sky	person	rider	car	truck	bus	train	motorc.	bicycle	**mean**
Source model	SegFormer [10]	87.9	52.7	64.1	34.0	20.2	37.2	34.5	40.2	51.8	32.4	6.6	54.5	31.4	72.8	49.6	65.2	54.1	34.0	41.4	45.5
TENT [9]		85.9	53.3	64.3	34.4	20.2	37.8	35.2	40.3	52.3	33.9	2.9	53.8	31.9	72.5	46.2	63.8	53.8	34.0	40.9	45.1
HCL [3]		88.2	54.3	64.4	35.3	20.7	39.1	36.8	40.4	52.0	32.1	2.8	55.2	33.7	73.5	49.2	66.5	58.1	35.4	41.7	46.3
URMA [8]		90.6	60.1	71.9	42.6	26.7	47.5	47.5	47.4	46.7	42.9	0.4	54.4	34.6	76.8	42.1	65.6	71.0	**38.0**	37.2	49.7
CMA		**95.2**	**77.5**	**84.3**	**43.9**	**30.9**	**49.4**	**52.0**	**49.6**	**74.2**	**51.2**	**78.4**	**61.4**	**41.2**	**79.2**	**63.6**	**75.1**	**75.8**	34.6	**47.3**	**61.3**

Table D-3. Comparison to the state of the art in model adaptation on Cityscapes→ACDC, with reported performance on the ACDC-rain test set.

Method		ACDC-rain IoU ↑																			
		road	sidew.	build.	wall	fence	pole	light	sign	veget.	terrain	sky	person	rider	car	truck	bus	train	motorc.	bicycle	**mean**
Source model	SegFormer [10]	83.1	46.7	89.5	40.5	47.2	54.0	67.0	66.9	92.6	40.2	97.6	63.5	24.6	87.8	65.1	72.7	81.0	42.8	58.0	64.3
TENT [9]		83.1	47.2	89.2	40.9	47.6	54.5	66.9	67.3	92.7	41.4	97.1	63.7	25.4	87.9	65.3	74.8	82.2	43.1	57.4	64.6
HCL [3]		84.2	50.5	90.1	42.7	48.9	57.0	68.5	69.0	93.0	40.9	97.8	65.4	26.1	88.7	68.1	74.4	80.4	43.8	58.0	65.6
URMA [8]		87.2	61.0	92.4	52.0	51.9	57.2	**72.0**	**73.1**	**93.8**	**46.1**	**98.1**	**68.8**	31.8	**90.6**	**73.2**	85.9	**86.9**	**51.7**	51.9	69.8
CMA		**93.3**	**76.3**	**92.8**	**58.1**	**58.2**	**61.2**	70.4	71.8	**93.8**	45.0	97.9	67.4	**36.8**	89.7	72.2	**88.5**	86.4	50.5	**66.7**	**72.5**

Table D-4. Comparison to the state of the art in model adaptation on Cityscapes→ACDC, with reported performance on the ACDC-snow test set.

Method		ACDC-snow IoU ↑																			
		road	sidew.	build.	wall	fence	pole	light	sign	veget.	terrain	sky	person	rider	car	truck	bus	train	motorc.	bicycle	**mean**
Source model	SegFormer [10]	82.0	44.9	80.5	30.4	45.4	46.8	65.6	63.1	86.8	5.2	93.6	67.8	40.8	87.1	56.4	76.7	83.1	32.8	60.3	60.5
TENT [9]		81.8	45.6	79.1	31.3	45.4	48.0	65.5	63.3	86.9	4.6	91.8	67.4	43.1	87.0	53.3	76.6	83.2	33.6	61.9	60.5
HCL [3]		82.9	47.4	83.2	35.4	46.8	50.1	67.8	64.9	87.7	5.3	95.6	69.8	43.9	87.6	60.1	76.9	83.2	35.3	63.4	62.5
URMA [8]		88.0	58.9	87.2	**52.0**	51.7	**57.8**	**75.6**	70.3	88.8	5.8	**97.1**	75.0	**63.6**	89.0	**69.6**	79.0	**89.8**	**50.1**	65.4	69.2
CMA		**92.4**	**70.5**	**88.3**	50.4	**55.6**	56.3	74.8	**71.1**	**90.8**	**29.4**	96.9	**77.4**	63.5	**90.1**	63.5	**79.6**	89.0	45.6	**73.9**	**71.5**

Figure F-1. Qualitative segmentation results of SegFormer-based adaptation methods on ACDC validation images.