– Supplementary Material –
MIC: Masked Image Consistency for Context-Enhanced Domain Adaptation

Lukas Hoyer [1] Dengxin Dai [2] Haoran Wang [2] Luc Van Gool [1,3]
[1] ETH Zurich [2] Max Planck Institute for Informatics, Saarland Informatics Campus [3] KU Leuven
{lhoyer,vangool}@vision.ee.ethz.ch, {ddai,hawang}@mpi-inf.mpg.de

A. Overview

In the supplementary material for MIC, we provide the source code (Sec. B), study the influence of further MIC parameters (Sec. C), analyze additional aspects of the behavior of MIC (Sec. D), extend the state-of-the-art comparison for semantic segmentation (Sec. E), provide a comprehensive qualitative comparison with previous works (Sec. F), and discuss potential limitations (Sec. G).

B. Source Code

The source code to train MIC is available at `https://github.com/lhoyer/MIC`. For further information on the environment setup and experiment execution, please refer to `README.md`. The implementation of MIC is based on the source code of HRDA [10] and mmsegmentation [4] for semantic segmentation, SDAT [19] for image classification, and SADA [3] for object detection.

C. Influence of Further MIC Parameters

C.1. MIC Prediction Region

To gain a better understanding of the working principles of MIC, we additionally study how MIC behaves if only masked or unmasked regions of the image are included in the MIC loss (i.e. $\mathcal{L}^M$ is only calculated for regions, where $\mathcal{M}_{ij}$ is 0 or 1). Tab. S1 shows that both MIC with a loss for masked patches and MIC with a loss for unmasked patches gain about +1.5 mIoU over DAFormer without MIC. When the MIC loss is calculated for both regions (default setting), the performance further improves by about +0.8 mIoU.

The improved performance for predicting masked patches shows that MIC profits from predicting regions with missing local information from the context. This task enhances the use of context relations for local predictions.

The improved performance for predicting unmasked patches shows that MIC profits from predicting regions with local information but without their complete context information. As not all context relations are available due to the masking, the network learns to exploit different combinations of context relations. This task enhances the robustness of the network towards missing context relations. During inference, this is particularly helpful to correctly predict partly-occluded objects (see Sec. F).

Both capabilities are complementary and can be successfully combined when applying the MIC loss to all image patches.

Table S1. Study of the MIC loss applied to specific image regions with DAFormer [9] on GTA→CS.

MIC Loss Region	mIoU
–	68.3
Masked Patches	69.8
Unmasked Patches	69.7
All Patches	70.6

Table S2. Parameter study of the MIC loss weight λ^M with DAFormer [9] on GTA→CS.

MIC Loss Weight λ^M	mIoU
0.0	68.3
0.1	68.9
0.5	69.5
1.0	70.6
2.0	70.1
10.0	67.9

C.2. MIC Loss Weight λ^M

Further, we study the influence of the MIC loss weight λ^M with DAFormer on GTA→CS. Tab. S2 shows that equal weighting of MIC loss ($\lambda^M = 1$) and the other loss terms achieves the best performance. A smaller weight gradually degrades the performance up to the point where no MIC is used. Also, a larger loss weight results in a decreased performance. If it is too large such as $\lambda^M = 10$, the performance can drop below the baseline. In that case, the MIC loss term dominates the total loss so that the other terms such as the source and adaptation loss cannot work effectively.

Table S3. Parameter study of the MIC teacher momentum α with DAFormer [9] on GTA→CS and with SDAT [19] on VisDA-2017.

Teacher Momentum α	$\text{mIoU}_{GTA\to CS}$	mAcc_{VisDA}
0.9	70.0	92.8
0.99	70.3	92.7
0.999	70.6	80.5
0.9999	69.3	79.5

Table S4. Ablation study of color augmentation for MIC with DAFormer [9] on GTA→CS and CS→ACDC.

MIC Domain	$\text{mIoU}_{GTA\to CS}$	$\text{mIoU}_{CS\to ACDC(Val)}$
–	68.3	55.1
w/o Color Augmentation	70.3	59.8
w/ Color Augmentation	70.6	58.7

C.3. Teacher Momentum α

Tab. S3 shows the influence of the MIC teacher network momentum α on the UDA performance for GTA→Cityscapes (semantic segmentation) and VisDA-2017 (image classification). For GTA→CS, it can be seen that the default value of $\alpha = 0.999$ from DAFormer [9] achieves the best performance. A smaller α (faster teacher update) gradually decreases the performance. Similarly, a higher teacher α also results in a performance drop. Probably, a too large α (slow teacher update) results in outdated pseudo-labels, which hamper the consistency training. For VisDA-2017, $\alpha = 0.9$ achieves the best performance, showing that a faster update of the teacher is useful for successful adaptation in this case.

C.4. Data Augmentation on Different Datasets

Tab. S4 compares MIC without and with color augmentation (brightness, contrast, saturation, hue, and blur following the parameters of [9, 10, 23]) on GTA→CS and CS→ACDC. It can be seen that color augmentation improves MIC for GTA→CS while it decreases the performance on CS→ACDC. We assume that the color augmentation can corrupt the content of dark nighttime images due to the locally already low brightness and contrast. If the color augmentation corrupts the content of the unmasked patches of the image, the masked image consistency loss can be rendered meaningless. Therefore, we forgo color augmentation for target domains with nighttime images (DarkZurich and ACDC).

D. Extended Analysis of MIC

D.1. Influence of Pseudo-Label Quality

To analyze the influence of the pseudo-label (PL) quality on the performance of MIC through the training, Fig. S1 plots the validation mIoU of MIC(DAFormer) with respect to the PL mIoU on the train set at several training iterations. Both are clearly correlated, which is expected given that better PL improve the model and a better model improves PL by the EMA update. To simulate worse PL, we add PL noise by randomly swapping the classes of PL segments. Even with 20% PL noise, which reduces the PL mIoU by -38, the final val. mIoU only decreases by -4. This shows that MIC is relatively robust to PL noise during the training.

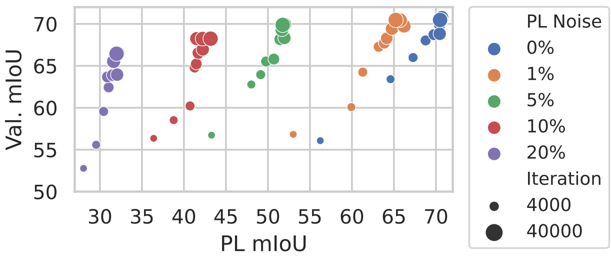

Figure S1. Influence of pseudo-label (PL) noise on the performance of MIC with DAFormer [9] on GTA→CS through the training process at different training iterations.

D.2. MIC as Standalone

MIC is designed as an orthogonal plug-in to enhance existing adaptation methods on various UDA benchmarks (see Tab. 1-5 in the main paper). Therefore, MIC requires an adaptation loss $\mathcal{L}^T$ from a host UDA method in order to work well. Without $\mathcal{L}^T$, the performance expectedly drops to 62.5 mIoU on GTA→CS with a DAFormer network. However, this is still +10.5 mIoU better than the host method in this case (DAFormer w/o $\mathcal{L}^T$).

D.3. Why Selecting Random Patches?

We have chosen random patch selection to promote a simple design that can be easily integrated into various UDA methods for different vision tasks and does not require a specific architecture or data priors. Despite its simplicity, we show that it is a very powerful strategy (see Tab. 1-5 in the main paper). As an image contains many objects with different context relations, it is hard to know in advance which relations are important. Random masking provides diverse context combinations through the training so that the network can identify different relevant relations.

D.4. Why does MIC Work for Image Classification?

Even though there is only one prediction per image for the image classification task, the network internally maintains spatial intermediate features. MIC can help to model context relations of object parts in these features, which can improve the distinction of ambiguous object parts on the target domain and reduce the effect of ill-adapted parts.

Table S5. Semantic segmentation UDA on CS→FoggyZurich

Method	Training with Simulated Fog	mIoU
CMAda2+ [6]	✓	43.4
CMAda3+ [6]	✓	46.8
FIFO [14]	✓	48.4
CuDA-Net+ [18]	✓	49.1
DAFormer [9]		40.8
MIC (DAFormer)		43.5
HRDA [10]		46.0
MIC (HRDA)		**49.7**

E. Extended Comparison for UDA Semantic Segmentation

Cityscapes→Foggy Zurich Supplementing the four semantic segmentation UDA benchmarks of the main paper, Tab. S5 further provides the semantic segmentation performance of MIC on Cityscapes [5] to Foggy Zurich [21]. MIC was trained using the annotated Cityscapes training set as source domain and the unlabeled Foggy Zurich medium fog set as target domain. For validation, the model was tested on the Foggy Zurich test v2 set. Tab. S5 shows that MIC(HRDA) significantly improves HRDA by +3.7 mIoU while MIC(DAFormer) gains +2.7 mIoU over DAFormer. MIC(HRDA) also outperforms specialized fog domain adaptation methods, which additionally utilize annotated Cityscapes images with simulated fog (Foggy Cityscapes DBF [21]) during training.

Additional Baselines In the main paper, we have shown a selection of the most relevant methods for domain-adaptive semantic segmentation. In the extended comparison in Tab. S6, we supplement the selection of previous works. It can be observed that also in the extended comparison, MIC(HRDA) outperforms all previous methods by a large margin. There are a few cases, where another method achieves a better performance for a specific class (e.g. DAP [11] for vegetation on Synthia→Cityscapes) but their performance falls behind MIC for other classes, resulting in a lower mIoU.

MIC with DAFormer Further, we provide MIC with DAFormer on all four benchmarks in Tab. S6. Compared to DAFormer, MIC(DAFormer) achieves significant performance improvements across the different datasets. The performance of MIC(DAFormer) can be further improved by utilizing sliding window inference as suggested in HRDA [10] to use the same inference input size as the training crop, which works better for the learned positional embedding of the Transformer encoder. MIC(DAFormer)$_{\text{slide}}$ improves the performance on all four benchmarks, especially for day-to-nighttime and clear-to-adverse-weather adaptation. Similar to MIC(HRDA), major improvements come from the classes *sidewalk*, *fence*, *pole*, *traffic sign*, *terrain*, and *rider*.

MIC with DeepLabV2 For a more fair comparison with ResNet-based UDA methods, we further provide detailed results of MIC(HRDA$_{\text{DLv2}}$), which uses a DeepLabV2 [2] network architecture with a ResNet-101 [8] backbone, in Tab. S6. It can be seen that MIC(HRDA$_{\text{DLv2}}$) significantly outperforms recent ResNet-based methods such as DecoupleNet [13], DAP [11], CPSL [15], and HRDA$_{\text{DLv2}}$ [10] on synthetic-to-real adaptation as well as CCDistill [7], DANIA [29], and HRDA$_{\text{DLv2}}$ [10] on day-to-nighttime/clear-to-adverse-weather adaptation.

F. Further Example Predictions

Supplementing the example predictions in the main paper, we show further representative examples of the strength and weaknesses of MIC in comparison with strong state-of-the-art methods.

Synthetic-to-Real Segmentation: On GTA→CS semantic segmentation, MIC(HRDA) achieves considerable performance improvements for the classes *sidewalk*, *fence*, *bus*, and *rider* (see Tab. S6). This is also reflected in the example predictions in Fig. S2-S5. In these examples, it can be observed that previous methods often recognize only parts of ambiguous regions while other parts of the same region are misclassified. As MIC was trained to utilize context relations, it has learned to reason more holistically about context relations in the images. Therefore, MIC can probably utilize the correctly recognized object parts to resolve the semantics of ambiguous image regions. More specifically, for *sidewalk* (Fig. S2), MIC is able to segment *sidewalk* more completely and even recognizes segments that previous methods failed to identify. For *fence* (Fig. S3), MIC reduces the segmentation of objects behind the fence instead of the fence. For *bus* (Fig. S4), MIC better segments ambiguous textures inside the bus and better recognizes partly-occluded busses. For *rider* (Fig. S5), MIC better segments the upper body and head of close riders and is able to recognize distant riders, probably by utilizing the bicycles as a context clue.

However, there are also some difficult examples, where UDA methods including MIC fail to correctly segment the image (Fig. S6). For example, MIC still struggles to differentiate vehicles with rare appearances, sidewalk that merges with the road, sidewalk under parking cars, and pedestrians standing close to bicycles.

Clear-to-Adverse-Weather Segmentation: On CS→ ACDC semantic segmentation, the same observations as for GTA→CS apply for the classes *sidewalk* (Fig. S7), *fence* (Fig. S8), and *bus/train* (Fig. S9). However, there are some

Table S6. Extended comparison of the semantic segmentation performance (IoU in %) on four different UDA benchmarks.

Method		Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU
Synthetic-to-Real: GTA→Cityscapes (Val.)																					
AdaptSeg [24]	ResNet-Based	86.5	25.9	79.8	22.1	20.0	23.6	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5	41.4
ADVENT [25]		89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
CBST [32]		91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
BDL [17]		91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
FADA [26]		91.0	50.6	86.0	43.4	29.8	36.8	43.4	25.0	86.8	38.3	87.4	64.0	38.0	85.2	31.6	46.1	6.5	25.4	37.1	50.1
DACS [23]		89.9	39.7	87.9	30.7	39.5	38.5	46.4	52.8	88.0	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.1
SAC [1]		90.4	53.9	86.6	42.4	27.3	45.1	48.5	42.7	87.4	40.1	86.1	67.5	29.7	88.5	49.1	54.6	9.8	26.6	45.3	53.8
CorDA [27]		94.7	63.1	87.6	30.7	40.6	40.2	47.8	51.6	87.6	47.0	89.7	66.7	35.9	90.2	48.9	57.5	0.0	39.8	56.0	56.6
ProDA [31]		87.8	56.0	79.7	46.3	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5
ProCA [12]		91.9	48.4	87.3	41.5	31.8	41.9	47.9	36.7	86.5	42.3	84.7	68.4	43.1	88.1	39.6	48.8	40.6	43.6	56.9	56.3
DecoupleNet [13]		87.6	49.3	87.2	42.5	41.6	46.6	57.4	44.0	89.0	43.9	90.6	73.0	43.8	88.1	32.9	53.7	44.3	49.8	57.2	59.1
DAP [11]		94.5	63.1	89.1	29.8	47.5	50.4	56.7	58.7	89.5	50.2	87.0	73.6	38.6	91.3	50.2	52.9	0.0	50.2	63.5	59.8
CPSL [15]		92.3	59.9	84.9	45.7	29.7	52.8	61.5	59.5	87.9	41.6	85.0	73.0	35.5	90.4	48.7	73.9	26.3	53.8	53.9	60.8
HRDA$_{DLv2}$ [10]		96.2	73.1	89.7	43.2	39.9	47.5	60.0	60.0	89.9	47.1	90.2	75.9	49.0	91.8	61.9	59.3	10.2	47.0	65.3	63.0
MIC (HRDA$_{DLv2}$)		96.5	74.3	90.4	47.1	42.8	50.3	61.7	62.3	90.3	49.2	90.7	77.8	53.2	93.0	66.2	68.0	6.8	38.0	60.6	64.2
DAFormer [9]	DAFormer	95.7	70.2	89.4	53.5	48.1	49.6	55.8	59.4	89.9	47.9	92.5	72.2	44.7	92.3	74.5	78.2	65.1	55.9	61.8	68.3
MIC (DAFormer)		96.7	75.0	90.0	58.2	50.4	51.1	56.7	62.1	90.2	51.3	92.9	72.4	47.1	92.8	78.9	83.4	75.6	54.2	62.6	70.6
MIC (DAFormer)$_{slide}$		96.9	76.5	90.1	57.6	52.2	51.2	56.7	61.8	90.3	**51.7**	92.9	72.5	47.9	92.9	79.5	85.5	76.8	53.6	62.9	71.0
HRDA [10]		96.4	74.4	91.0	**61.6**	51.5	57.1	63.9	69.3	91.3	48.4	94.2	79.0	52.9	93.9	84.1	85.7	75.9	63.9	67.5	73.8
MIC (HRDA)		**97.4**	**80.1**	**91.7**	61.2	**56.9**	**59.7**	**66.0**	**71.3**	**91.7**	51.4	**94.3**	**79.8**	**56.1**	**94.6**	**85.4**	**90.3**	**80.4**	**64.5**	**68.5**	**75.9**
Synthetic-to-Real: Synthia→Cityscapes (Val.)																					
ADVENT [25]	ResNet-Based	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	–	84.1	57.9	23.8	73.3	–	36.4	–	14.2	33.0	41.2
CBST [32]		68.0	29.9	76.3	10.8	1.4	33.9	22.8	29.5	77.6	–	78.3	60.6	28.3	81.6	–	23.5	–	18.8	39.8	42.6
FADA [26]		84.5	40.1	83.1	4.8	0.0	34.3	20.1	27.2	84.8	–	84.0	53.5	22.6	85.4	–	43.7	–	26.8	27.8	45.2
DACS [23]		80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	–	90.8	67.6	38.3	82.9	–	38.9	–	28.5	47.6	48.3
SAC [1]		89.3	47.2	85.5	26.5	1.3	43.0	45.5	32.0	87.1	–	89.3	63.6	25.4	86.9	–	35.6	–	30.4	53.0	52.6
CorDA [27]		**93.3**	**61.6**	85.3	19.6	5.1	37.8	36.6	42.8	84.9	–	90.4	69.7	41.8	85.6	–	38.4	–	32.6	53.9	55.0
ProDA [31]		87.8	45.7	84.6	37.1	0.6	44.0	54.6	37.0	88.1	–	84.4	74.2	24.3	88.2	–	51.1	–	40.5	45.6	55.5
ProCA [12]		90.5	52.1	84.6	29.2	3.3	40.3	37.4	27.3	86.4	–	85.9	69.8	28.7	88.7	–	53.7	–	14.8	54.8	53.0
DecoupleNet [13]		77.8	48.6	75.6	32.0	1.9	44.4	52.9	38.5	87.8	–	88.1	71.1	34.3	88.7	–	58.8	–	50.2	61.4	57.0
DAP [11]		84.2	46.5	82.5	35.1	0.2	46.7	53.6	45.7	**89.3**	–	87.5	75.7	34.6	**91.7**	–	**73.5**	–	49.4	60.5	59.8
CPSL [15]		87.2	43.9	85.5	33.6	0.3	47.7	57.4	37.2	87.8	–	88.5	79.0	32.0	90.6	–	49.4	–	50.8	59.8	57.9
HRDA$_{DLv2}$ [10]		85.8	47.3	87.3	27.3	1.4	50.5	57.8	61.0	87.4	–	89.1	76.2	48.5	87.3	–	49.3	–	55.0	68.2	61.2
MIC (HRDA$_{DLv2}$)		84.7	45.7	88.3	29.9	2.8	53.3	61.0	59.5	86.9	–	88.8	78.2	53.3	89.4	–	58.8	–	56.0	**68.3**	62.8
DAFormer [9]	DAFormer	84.5	40.7	88.4	41.5	6.5	50.0	55.0	54.6	86.0	–	89.8	73.2	48.2	87.2	–	53.2	–	53.9	61.7	60.9
MIC (DAFormer)		83.0	40.9	88.2	37.6	**9.0**	52.4	56.0	56.5	87.6	–	93.4	74.2	51.4	87.1	–	59.6	–	57.9	61.2	62.2
MIC (DAFormer)$_{slide}$		82.6	40.7	88.3	40.2	**9.0**	52.4	55.7	56.6	87.6	–	93.4	74.1	52.5	87.2	–	62.2	–	57.4	61.1	62.6
HRDA [10]		85.2	47.7	88.8	**49.5**	4.8	57.2	65.7	60.9	85.3	–	92.9	79.4	52.8	89.0	–	64.7	–	63.9	64.9	65.8
MIC (HRDA)		86.6	50.5	**89.3**	47.9	7.8	**59.4**	**66.7**	**63.4**	87.1	–	**94.6**	**81.0**	**58.9**	90.1	–	61.9	–	**67.1**	64.3	**67.3**
Day-to-Nighttime: Cityscapes→DarkZurich (Test)																					
ADVENT [25]	ResNet-Based	85.8	37.9	55.5	27.7	14.5	23.1	14.0	21.1	32.1	8.7	2.0	39.9	16.6	64.0	13.8	0.0	58.8	28.5	20.7	29.7
AdaptSeg [24]		86.1	44.2	55.1	22.2	4.8	21.1	5.6	16.7	37.2	8.4	1.2	35.9	26.7	68.2	45.1	0.0	50.1	33.9	15.6	30.4
BDL [17]		85.3	41.1	61.9	32.7	17.4	20.6	11.4	21.3	29.4	8.9	1.1	37.4	22.1	63.2	28.2	0.0	47.7	39.4	15.7	30.8
GCMA[†] [20]		81.7	46.9	58.8	22.0	20.0	41.2	40.5	41.6	64.8	31.0	32.1	53.5	47.5	75.5	39.2	0.0	49.6	30.7	21.0	42.0
MGCDA[†] [22]		80.3	49.3	66.2	7.8	11.0	41.4	38.9	39.0	64.1	18.0	55.8	52.1	53.5	74.7	66.0	0.0	37.5	29.1	22.7	42.5
DANNet[†] [28]		90.0	54.0	74.8	41.0	21.1	25.0	26.8	30.2	**72.0**	26.2	**84.0**	47.0	33.9	68.2	19.0	0.3	66.4	38.3	23.6	44.3
CDAda[†] [30]		90.5	60.6	67.9	37.0	19.3	42.9	36.4	35.3	66.9	24.4	79.8	45.4	42.9	70.8	51.7	0.0	29.7	27.7	26.2	45.0
CCDistill[†] [7]		89.6	58.1	70.6	36.6	22.5	33.0	27.0	30.5	68.3	33.0	80.9	42.3	40.1	69.4	58.1	0.1	72.6	47.7	21.3	47.5
HRDA$_{DLv2}$ [10]		88.7	65.5	68.3	41.9	18.1	50.6	6.0	39.6	33.3	34.4	0.3	57.6	51.7	75.0	70.9	8.5	63.6	41.0	38.8	44.9
MIC (HRDA$_{DLv2}$)		82.8	69.6	75.5	44.0	21.0	51.1	43.4	48.3	39.3	37.1	0.0	59.4	53.6	73.6	74.2	9.2	78.7	40.0	37.2	49.4
DAFormer [9]	DAFormer	93.5	65.5	73.3	39.4	19.2	53.3	44.1	44.0	59.5	34.5	66.6	53.4	52.7	82.1	52.7	9.5	89.3	50.5	38.5	53.8
MIC (DAFormer)		88.2	60.5	73.5	53.5	23.8	52.3	44.6	43.8	68.6	34.0	58.1	57.8	48.2	78.7	58.0	13.3	91.2	46.1	42.9	54.6
MIC (DAFormer)$_{slide}$		89.9	65.0	75.9	54.9	25.5	53.3	44.6	44.0	70.0	39.2	62.0	58.4	48.7	79.8	59.6	**21.0**	**91.3**	**53.4**	**44.7**	56.9
HRDA [10]		90.4	56.3	72.0	39.5	19.5	57.8	**52.7**	43.1	59.3	29.1	70.5	60.0	58.6	**84.0**	**75.5**	11.2	90.5	51.6	40.9	55.9
MIC (HRDA)		**94.8**	**75.0**	**84.0**	**55.1**	**28.4**	**62.0**	35.5	**52.6**	59.2	**46.8**	70.0	**65.2**	**61.7**	82.1	64.2	18.5	**91.3**	52.6	44.0	**60.2**
Clear-to-Adverse-Weather: Cityscapes→ACDC (Test)																					
ADVENT [25]	ResNet-Based	72.9	14.3	40.5	16.6	21.2	9.3	17.4	21.2	63.8	23.8	18.3	32.6	19.5	69.5	36.2	34.5	46.2	26.9	36.1	32.7
AdaptSegNet [24]		69.4	34.0	52.8	13.5	18.0	4.3	14.9	9.7	64.0	23.1	38.2	38.6	20.1	59.3	35.6	30.6	53.9	19.8	33.9	33.4
BDL [17]		56.0	32.5	68.1	20.1	17.4	15.8	30.2	28.7	59.9	25.3	37.7	28.7	25.5	70.2	39.6	40.5	52.7	29.2	38.4	37.7
GCMA[†] [20]		79.7	48.7	71.5	21.6	29.9	42.5	56.7	57.7	75.8	39.5	87.2	57.4	29.7	80.6	44.9	46.2	62.0	37.2	46.5	53.4
MGCDA[†] [22]		73.4	28.7	69.9	19.3	26.3	36.8	53.0	53.3	75.4	32.0	84.6	51.0	26.1	77.6	43.2	45.9	53.9	32.7	41.5	48.7
DANNet[†] [28]		84.3	54.2	77.6	38.0	30.0	18.9	41.6	35.2	71.3	39.4	86.6	48.7	29.2	76.2	41.6	43.0	58.6	32.6	43.9	50.0
DANIA[†] [29]		88.4	60.6	81.1	37.1	32.8	28.4	43.2	42.6	**77.7**	50.5	**90.5**	51.5	31.1	76.0	37.4	44.9	64.0	31.8	46.3	53.5
HRDA$_{DLv2}$ [10]		84.9	63.2	83.1	33.1	32.3	46.0	42.7	55.4	69.2	52.8	83.1	63.2	37.8	78.1	48.5	58.5	62.4	42.8	57.2	57.6
MIC (HRDA$_{DLv2}$)		88.7	63.9	84.1	38.4	35.7	45.7	51.5	60.3	72.7	52.3	85.8	62.5	39.8	84.7	37.7	68.7	71.9	46.0	56.5	60.4
DAFormer [9]	DAFormer	58.4	51.3	84.0	42.7	35.1	50.7	30.0	57.0	74.8	52.8	51.3	58.3	32.6	82.7	58.3	54.9	82.4	44.1	50.7	55.4
MIC (DAFormer)		58.5	51.6	84.9	48.1	39.8	50.8	39.7	59.9	77.1	54.9	51.9	63.9	40.7	84.1	63.1	66.2	85.5	46.3	57.1	59.2
MIC (DAFormer)$_{slide}$		60.5	60.5	86.1	54.7	**42.0**	51.4	41.2	61.2	77.6	57.4	53.6	64.6	40.2	85.9	68.7	73.8	87.0	50.1	58.8	61.9
HRDA [10]		88.3	57.9	88.1	**55.2**	36.7	56.3	**62.9**	65.3	74.2	57.7	85.9	68.8	45.7	88.5	**76.4**	82.4	87.7	52.7	60.4	68.0
MIC (HRDA)		**90.8**	**67.1**	**89.2**	54.5	40.5	**57.2**	62.0	**68.4**	76.3	**61.8**	87.0	**71.3**	**49.4**	**89.7**	75.7	**86.8**	**89.1**	**56.9**	**63.0**	**70.4**

[†] Method uses additional daytime/clear-weather geographically-aligned reference images.

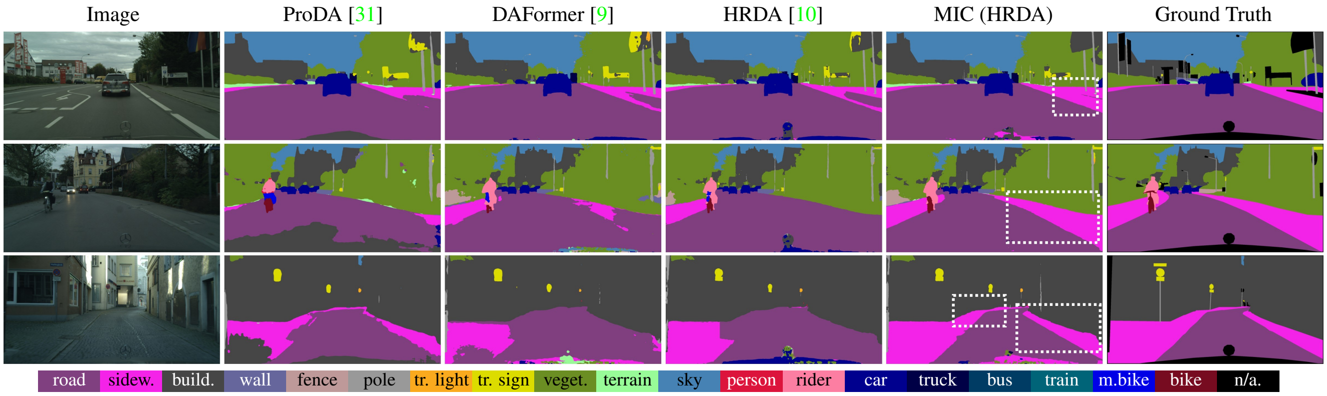

Figure S2. Example predictions showing a better segmentation of *sidewalk* by MIC on GTA→Cityscapes.

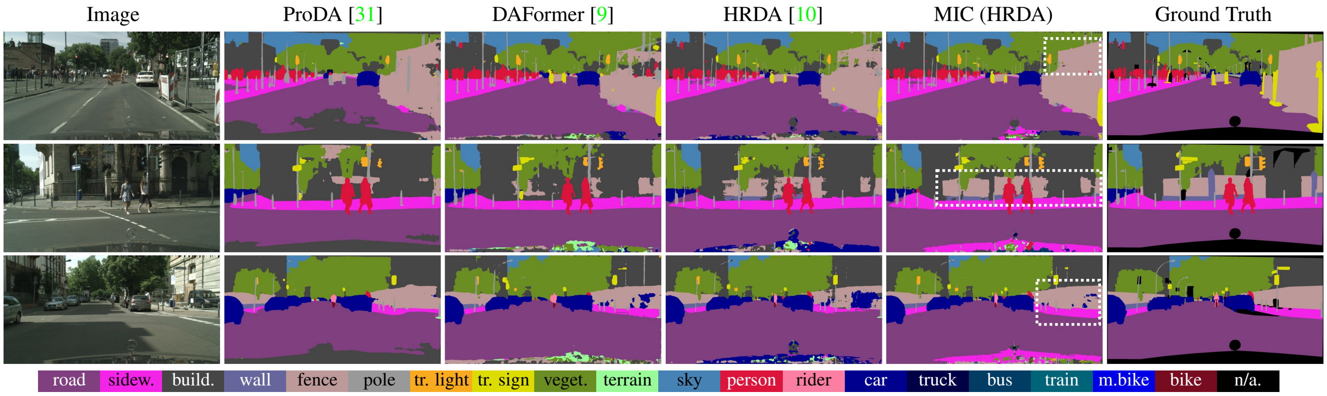

Figure S3. Example predictions showing a better segmentation of *fence* by MIC on GTA→Cityscapes.

distinct failure cases. In particular, UDA methods including MIC fail to segment snow-covered *sidewalk*, distinguish *sky/vegetation/building* in dark image ares, and struggle with motion blur of dynamic objects.

Clear-to-Foggy-Weather Detection: On CS→Foggy CS object detection, MIC(SADA) is able to detect objects that previous methods failed to recognize. For example, MIC better detects the classes *bus* and *truck* (Fig. S11) as well as *rider*, *motorcycle*, and *bicycle* (Fig. S12). Typical failure cases (Fig. S13) include multiple detections for a single object, missed detections, and the confusion of semantically similar objects.

Synthetic-to-Real Classification: For VisDA-2017 image classification UDA, we provide a random selection of examples, where MIC(SDAT) performs better than SDAT in Fig. S14. It can be seen that MIC can better distinguish semantically similar classes such as *train* vs. *bus*, *bus* vs *truck*, and *truck* vs *car*. Further, we show a random selection of failure cases of MIC(SDAT) in Fig. S15. MIC mostly confuses semantically similar vehicle classes, especially if instances are at the decision boundary between two classes or different classes are present in an image.

Supervised Segmentation: Fig. S16 compares DAFormer and MIC(DAFormer) when trained in a supervised fashion on Cityscapes. It shows improvements for regions that are difficult to identify such as instances of *terrain*, *sidewalk*, *bus*, and *rider*. Generally, the supervised DAFormer without MIC already performs very well, so that the potential for improvement is smaller, which is also reflected in the quantitative results in the main paper.

G. Potential Limitations

Even though UDA methods achieve evolvingly higher performances for synthetic-to-real and clear-to-adverse weather adaptation, the current methods are still not reliable enough to be safely deployed in real-world autonomous driving as can be seen in the failure cases in Fig. S6, S10, and S13. For these cases, it is still necessary to collect annotations on the target domain to achieve safe operation. We hope that this gap to supervised learning can be gradually narrowed in the future, but we assume that, for some corner cases, a few annotations might still be necessary to reliably guide the adaptation.

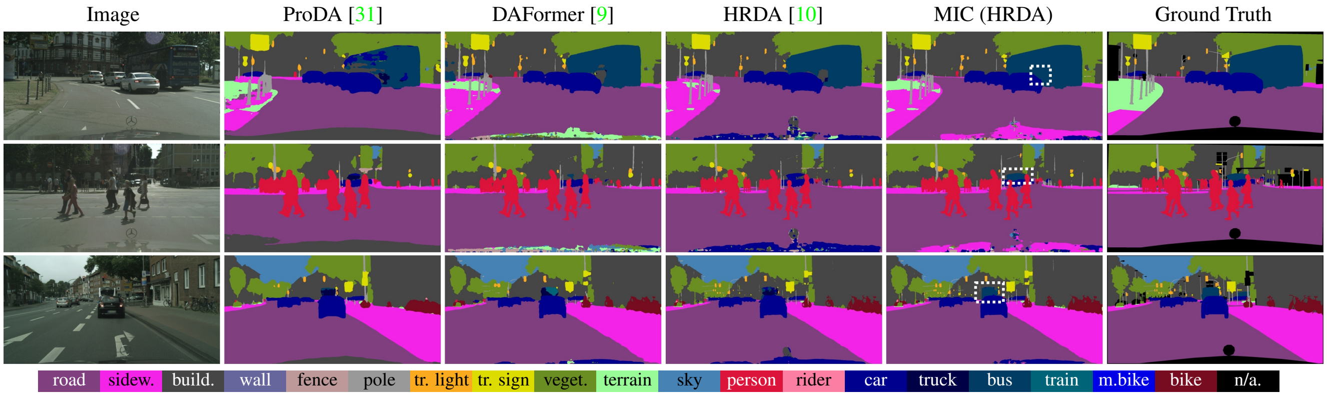

Figure S4. Example predictions showing a better segmentation of *bus* by MIC on GTA→Cityscapes.

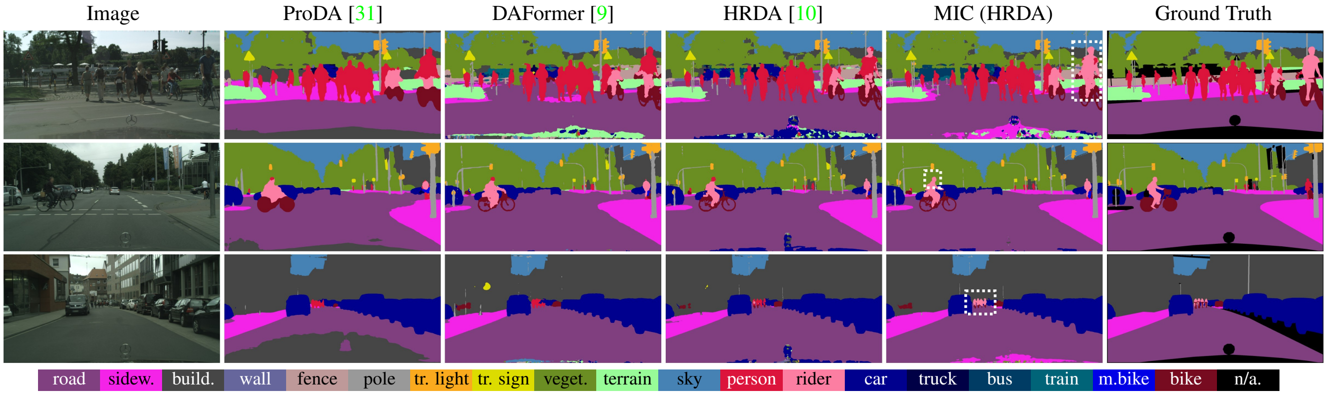

Figure S5. Example predictions showing a better segmentation of *rider* by MIC on GTA→Cityscapes.

As MIC is specifically exploiting context relations for domain adaptation, it is based on two assumptions. First, MIC assumes that context information is a relevant factor for recognition. For classes, where context is less important, such as *building* or *vegetation* for synthetic-to-real adaptation, MIC has a limited potential for improvement. And second, MIC assumes that the relevant context relations are captured by the training data. If objects appear out-of-context during inference, MIC might be more susceptible to these corner cases. In the experimental analysis, it is shown that these assumptions mostly hold on a wide range of practically-relevant UDA benchmarks and MIC outperforms previous methods by a significant margin.

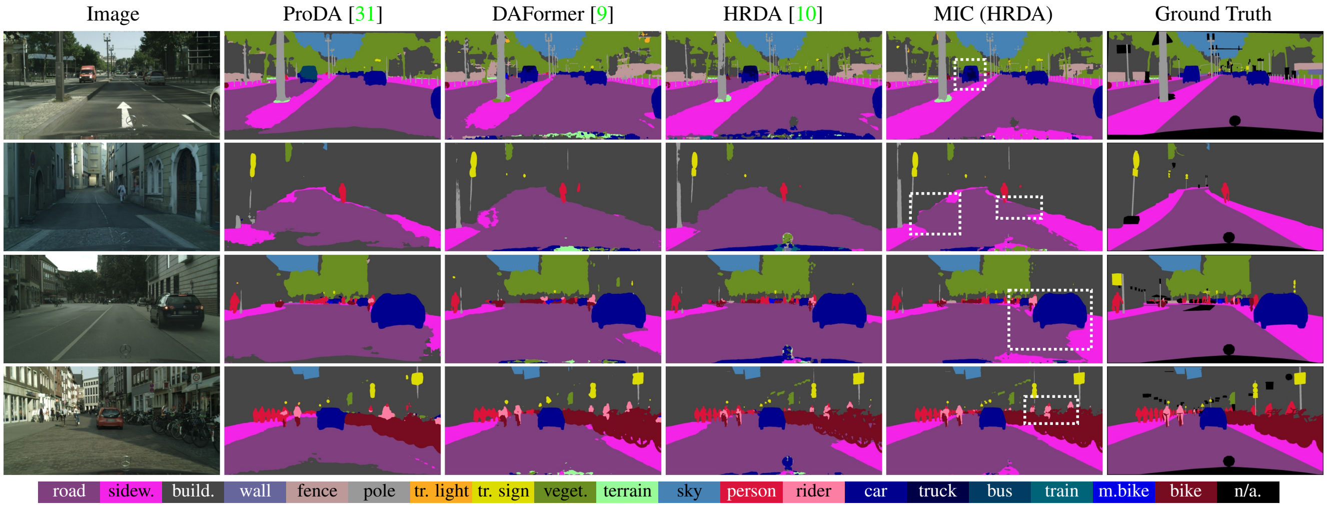

Figure S6. Failure cases of MIC on GTA→Cityscapes.

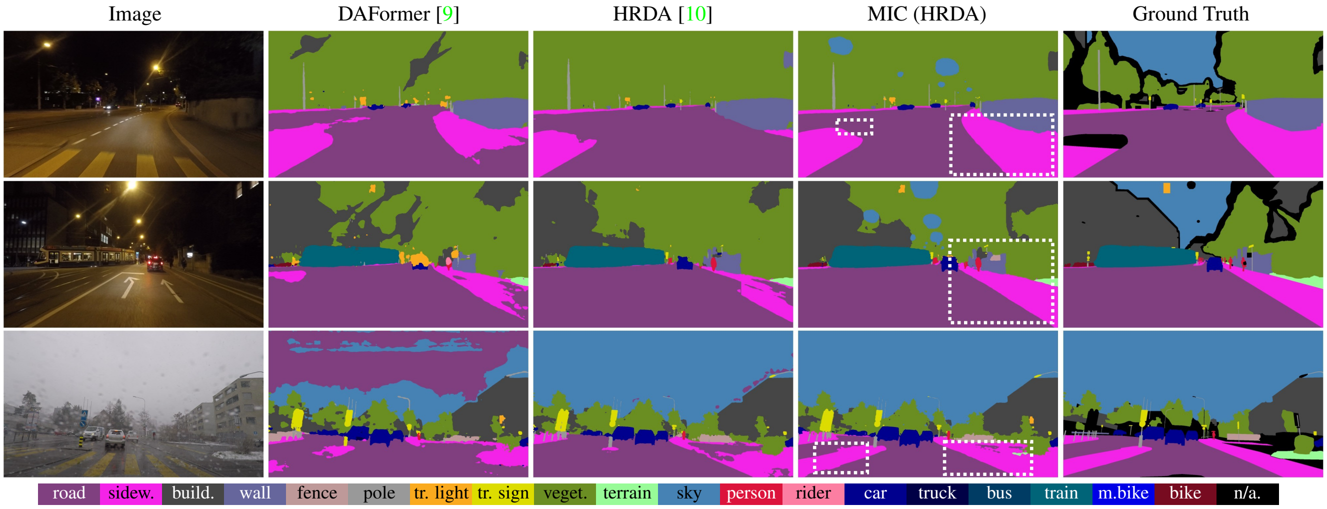

Figure S7. Example predictions showing a better segmentation of *sidewalk* by MIC on Cityscapes→ACDC.

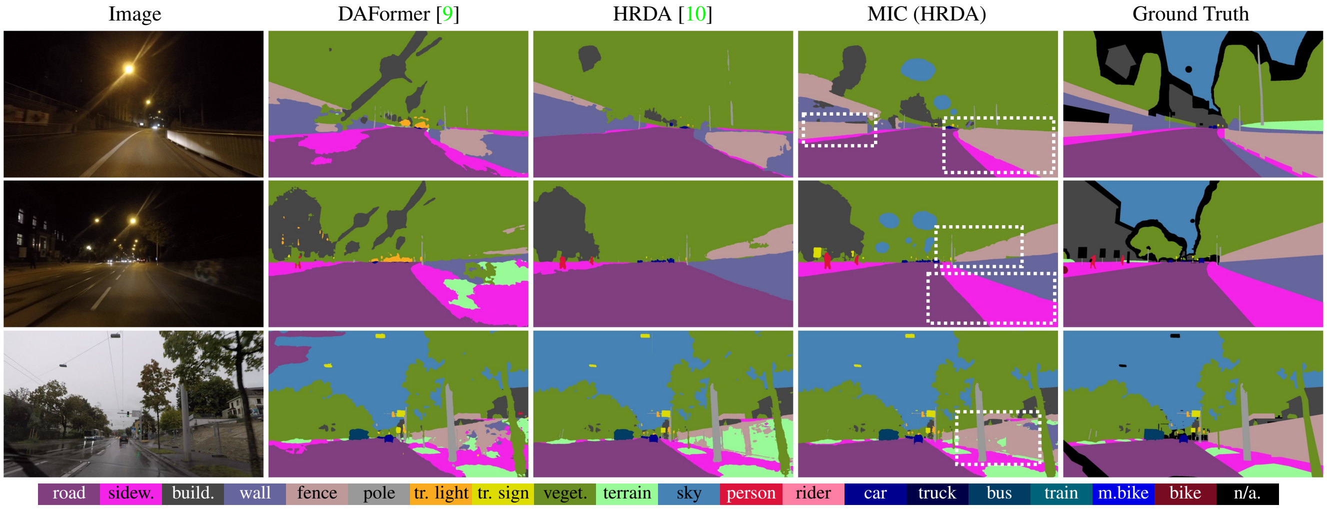

Figure S8. Example predictions showing a better segmentation of *fence* by MIC on Cityscapes→ACDC.

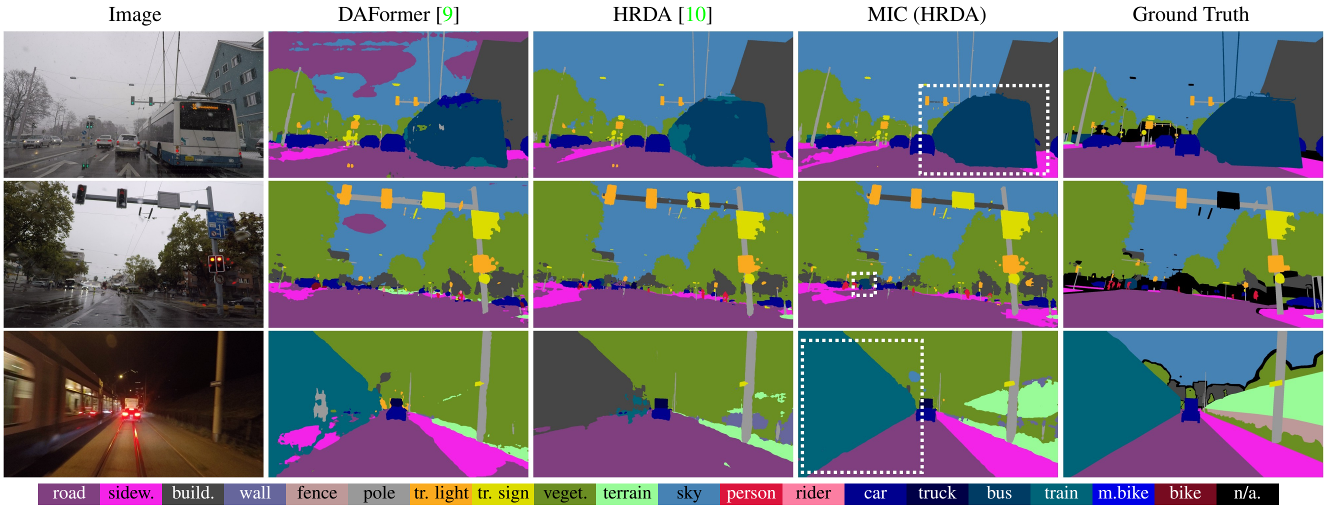
Image
DAFormer [9]
HRDA [10]
MIC (HRDA)
Ground Truth

Figure S9. Example predictions showing a better segmentation of *bus* and *train* by MIC on Cityscapes→ACDC.

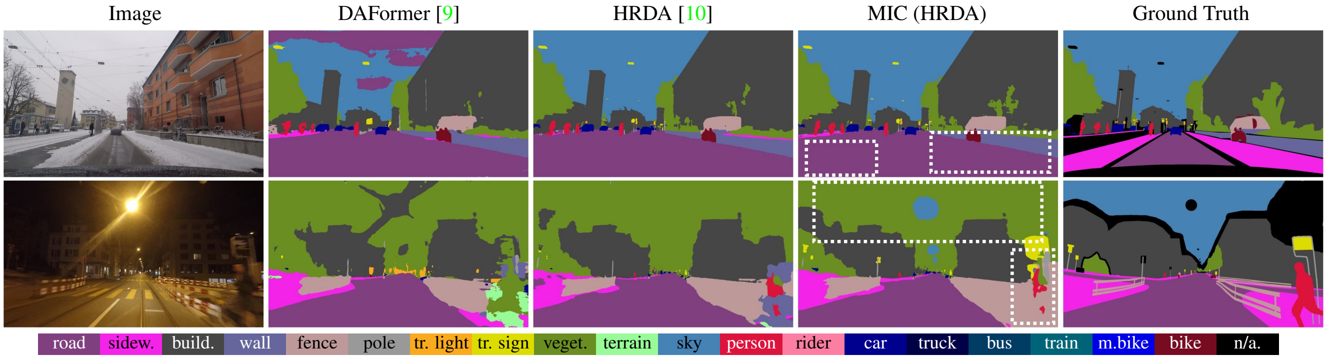
Image
DAFormer [9]
HRDA [10]
MIC (HRDA)
Ground Truth

Figure S10. Failure cases of MIC on Cityscapes→ACDC.

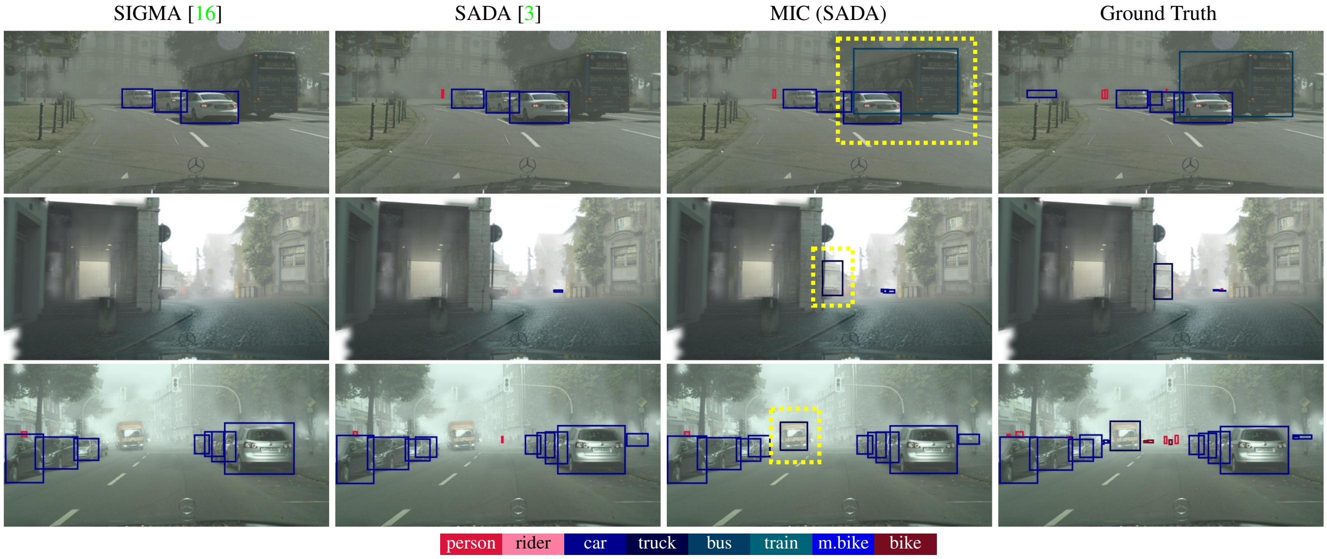
SIGMA [16]
SADA [3]
MIC (SADA)
Ground Truth

Figure S11. Example predictions showing a better detection of *bus* and *truck* by MIC on Cityscapes→Foggy Cityscapes.

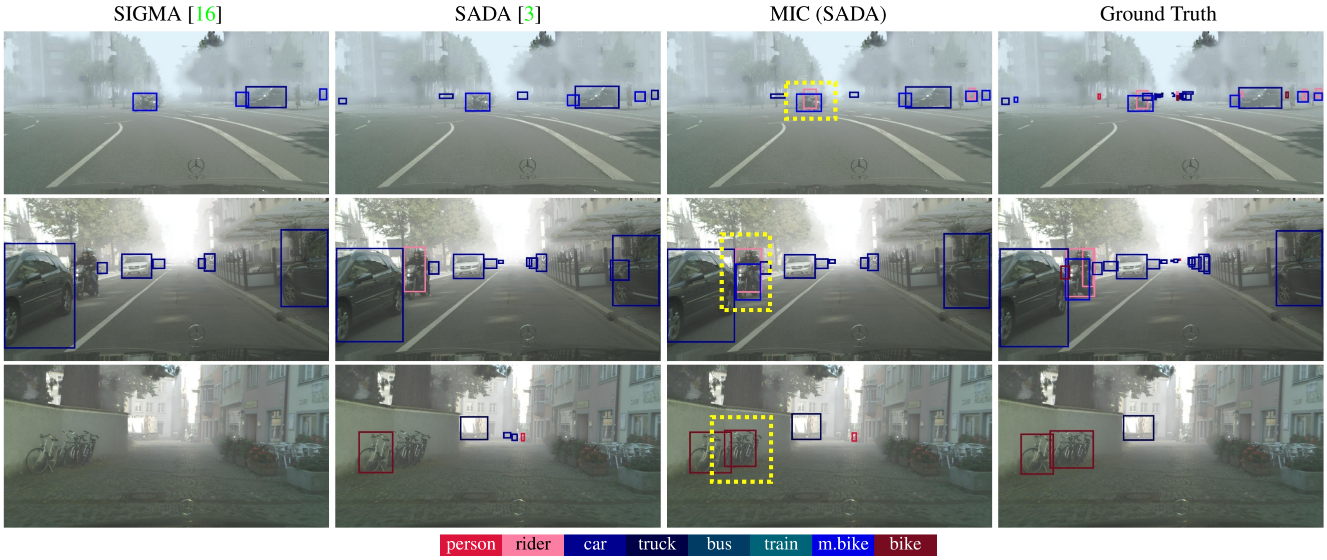

Figure S12. Example predictions showing a better detection of *rider*, *motorcycle*, and *bicycle* by MIC on Cityscapes→Foggy Cityscapes.

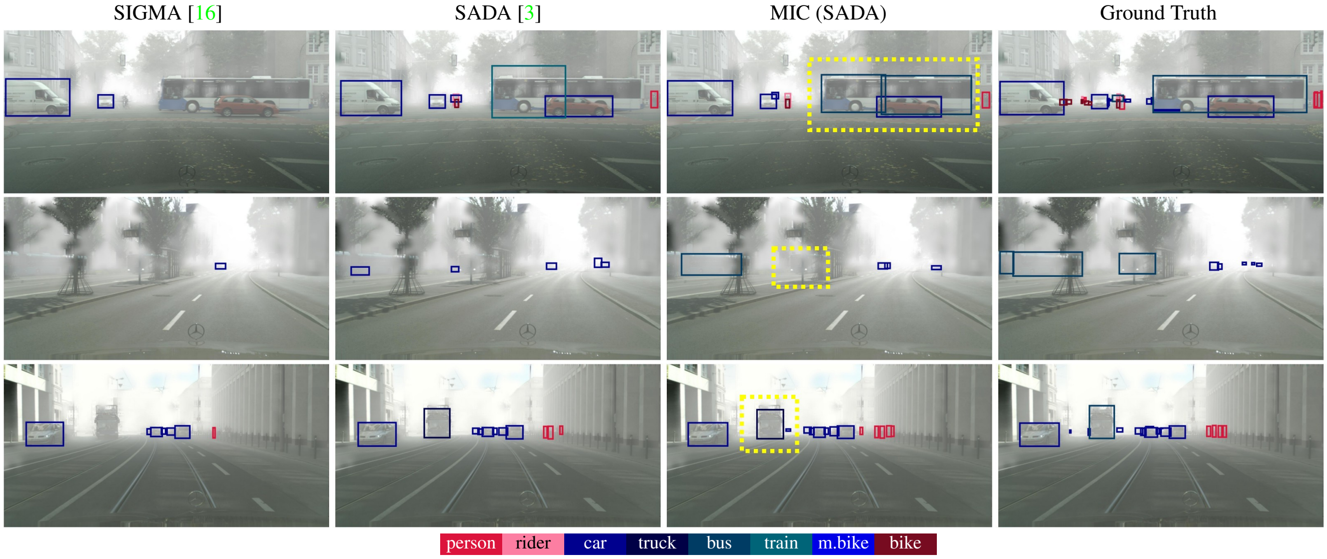

Figure S13. Failure cases of MIC on Cityscapes→Foggy Cityscapes.

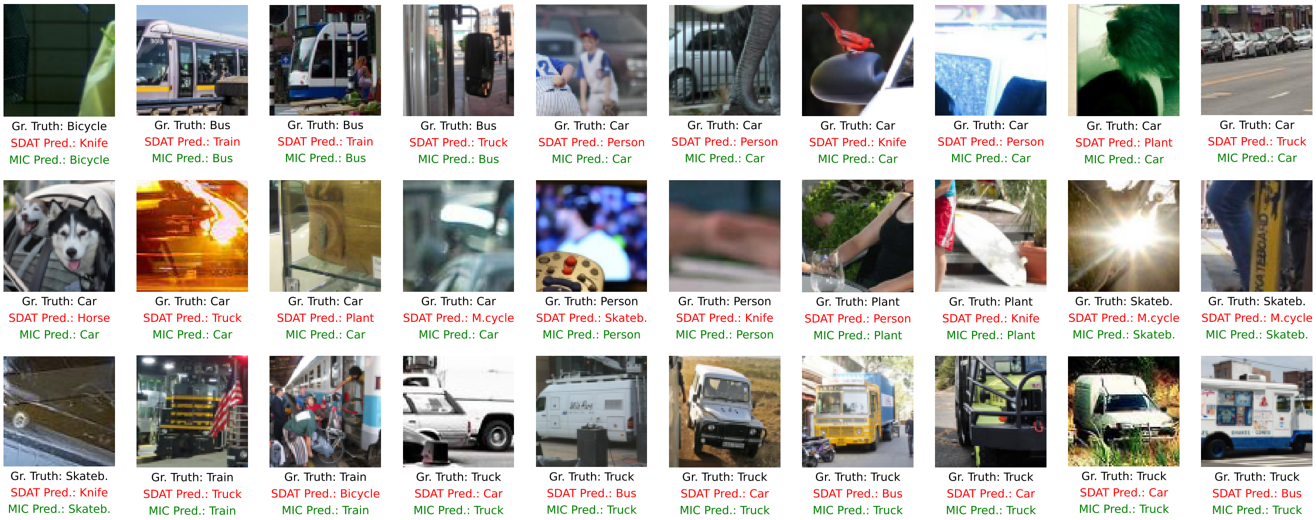

Figure S14. Example predictions showing a better recognition performance of MIC on VisDA.

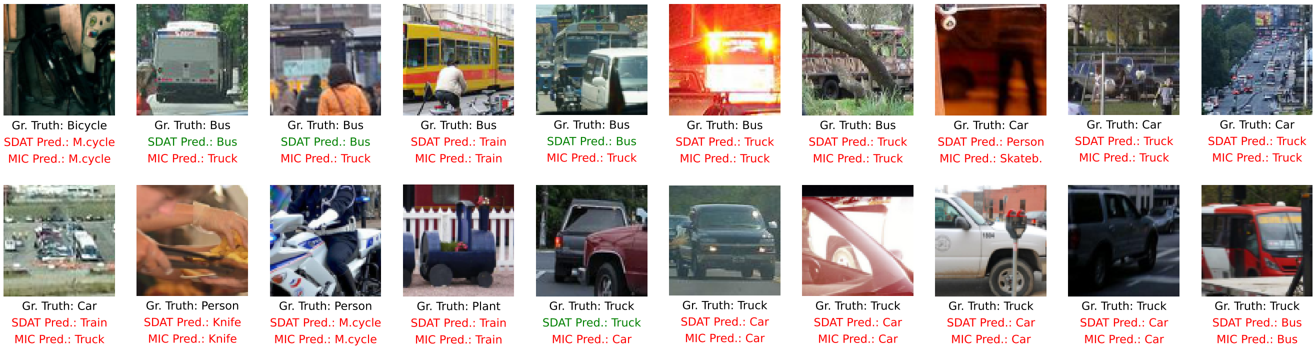

Figure S15. Failure cases of MIC on VisDA.

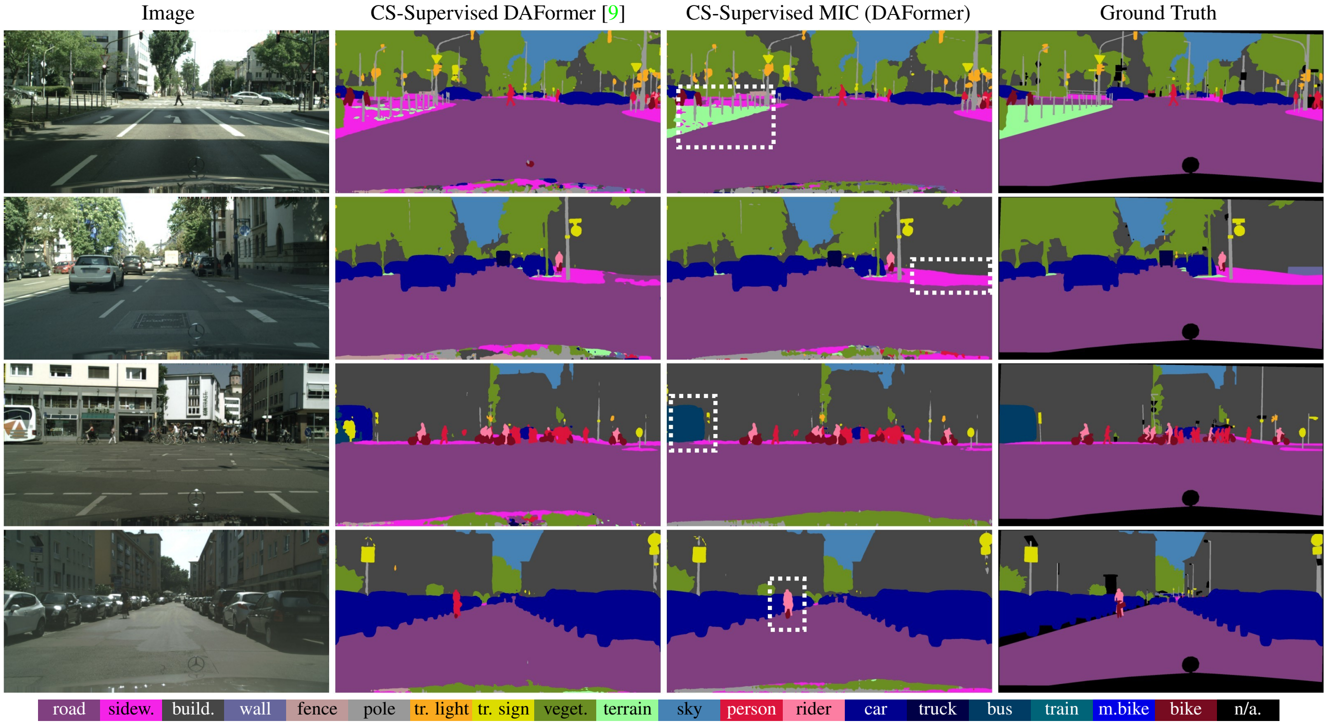

Figure S16. Example predictions showing a better segmentation of difficult classes such as *terrain*, *sidewalk*, *bus*, and *rider* by MIC in a **supervised** training setup on Cityscapes.

References

[1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *CVPR*, pages 15384–15394, 2021. 4

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*, 40(4):834–848, 2017. 3

[3] Yuhua Chen, Haoran Wang, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Scale-aware domain adaptive faster r-cnn. *IJCV*, 129(7):2223–2243, 2021. 1, 8, 9

[4] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 1

[5] 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*, pages 3213–3223, 2016. Dataset URL: https://www.cityscapes-dataset.com/, Dataset License: https://www.cityscapes-dataset.com/license/. 3

[6] 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. 3

[7] Huan Gao, Jichang Guo, Guoli Wang, and Qian Zhang. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. In *CVPR*, pages 9913–9923, 2022. 3, 4

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[9] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 11

[10] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In *ECCV*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[11] Xinyue Huo, Lingxi Xie, Hengtong Hu, Wengang Zhou, Houqiang Li, and Qi Tian. Domain-agnostic prior for transfer semantic segmentation. In *CVPR*, pages 7075–7085, 2022. 3, 4

[12] Zhengkai Jiang, Yuxi Li, Ceyuan Yang, Peng Gao, Yabiao Wang, Ying Tai, and Chengjie Wang. Prototypical contrast adaptation for domain adaptive semantic segmentation. In *ECCV*, pages 36–54, 2022. 4

[13] Xin Lai, Zhuotao Tian, Xiaogang Xu, Yingcong Chen, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Decouplenet: Decoupled network for domain adaptive semantic segmentation. In *ECCV*, pages 369–387, 2022. 3, 4

[14] Sohyun Lee, Taeyoung Son, and Suha Kwak. Fifo: Learning fog-invariant features for foggy scene segmentation. In *CVPR*, pages 18911–18921, 2022. 3

[15] Ruihuang Li, Shuai Li, Chenhang He, Yabin Zhang, Xu Jia, and Lei Zhang. Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation. In *CVPR*, pages 11593–11603, 2022. 3, 4

[16] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *CVPR*, pages 5291–5300, 2022. 8, 9

[17] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, pages 6936–6945, 2019. 4

[18] Xianzheng Ma, Zhixiang Wang, Yacheng Zhan, Yinqiang Zheng, Zheng Wang, Dengxin Dai, and Chia-Wen Lin. Both style and fog matter: Cumulative domain adaptation for semantic foggy scene understanding. In *CVPR*, pages 18922–18931, 2022. 3

[19] Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, Arihant Jain, and Venkatesh Babu Radhakrishnan. A closer look at smoothness in domain adversarial training. In *ICML*, pages 18378–18399, 2022. 1, 2

[20] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *ICCV*, pages 7374–7383, 2019. Dataset URL: https://www.trace.ethz.ch/publications/2019/GCMA_UIoU/. 4

[21] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *ECCV*, pages 687–704, 2018. Dataset URL: https://people.ee.ethz.ch/~csakarid/Model_adaptation_SFSU_dense/. 3

[22] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *PAMI*, 2020. Dataset URL: https://www.trace.ethz.ch/publications/2019/GCMA_UIoU/. 4

[23] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. DACS: Domain Adaptation via Cross-domain Mixed Sampling. In *WACV*, pages 1379–1389, 2021. 2, 4

[24] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, pages 7472–7481, 2018. 4

[25] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, pages 2517–2526, 2019. 4

[26] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *ECCV*, pages 642–659, 2020. 4

[27] Qin Wang, Dengxin Dai, Lukas Hoyer, Olga Fink, and Luc Van Gool. Domain adaptive semantic segmentation with self-supervised depth estimation. In *ICCV*, pages 8515–8525, 2021. 4

[28] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *CVPR*, pages 15769–15778, 2021. 4

[29] Xinyi Wu, Zhenyao Wu, Lili Ju, and Song Wang. A one-stage domain adaptation network with image alignment for unsupervised nighttime semantic segmentation. *PAMI*, 2021. 3, 4

[30] Qi Xu, Yinan Ma, Jing Wu, Chengnian Long, and Xiaolin Huang. Cdada: A curriculum domain adaptation for nighttime semantic segmentation. In *ICCVW*, pages 2962–2971, 2021. 4

[31] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *CVPR*, pages 12414–12424, 2021. 4, 5, 6, 7

[32] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, pages 289–305, 2018. 4