Supplementary Material: mDALU: Multi-Source Domain Adaptation and Label Unification with Partial Datasets

Rui Gong [1], Dengxin Dai [1,4], Yuhua Chen [1], Wen Li [3], Luc Van Gool [1,2]
[1] Computer Vision Lab, ETH Zurich, [2] VISICS, KU Leuven, [3] UESTC, [4] MPI for Informatics
{gongr, dai, yuhua.chen, vangool}@vision.ee.ethz.ch, liwenbnu@gmail.com

In this supplementary, we provide additional information for,

S1 detailed framework structure and implementation of our approach,

S2 more detailed information about the datasets involved in experiments,

S3 experimental results when having more than two source domains,

S4 more experimental results and additional visualization results for semantic segmentation.

S1. Framework Structure and Implementation

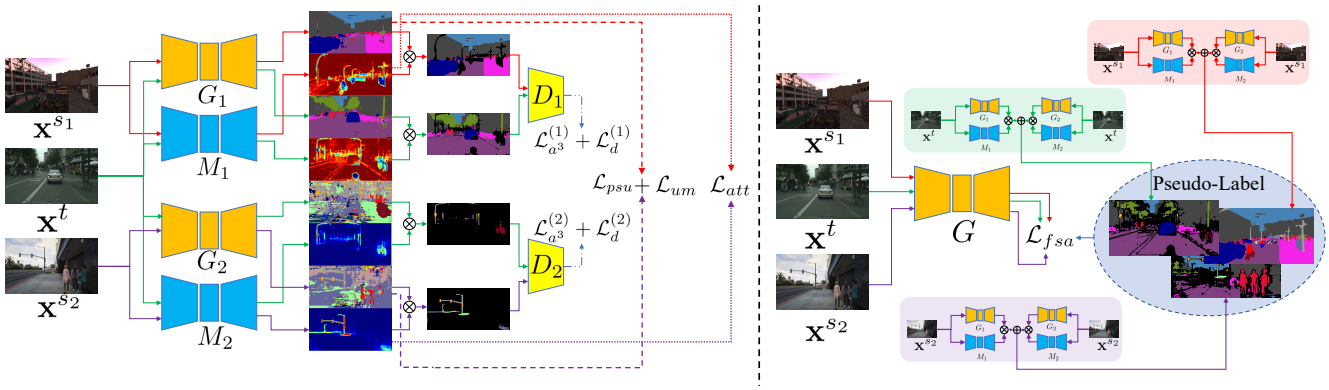
In Sec. 3 and Fig. 2 of the main paper, we introduce our approach to mDALU problem, and here we provide more detailed structure and implementation of our approach. The overview of our approach is shown in Fig. S1. In the image classification experiment, the hyperparameter λ in Eq. (10) of the main paper is set as 1.0, and δ in Eq. (12) and Eq. (13) of the main paper is set as 0.5. The images are resized to 32×32. We use the the Adam optimizer [8] with $\beta_1 = 0.9, \beta_2 = 0.999$ and the weight decay as 5×10^{-4}. The learning rate is set as 2×10^{-4}. We adopt the same network architecture as that of the digits classification experiments in [12]. In the 2D semantic image segmentation experiments, the hyperparameter λ in Eq. (10) of the main paper is set as 0.001, and δ in Eq. (12) and Eq. (13) of the main paper is set as 0.2, 0.5 and 0.4 for SYNTHIA, GTA5 and Cityscapes dataset, respectively. The images are resized to 1024×512. We use the SGD optimizer for training the semantic segmentation network, whose momentum is 0.9, weight decay is 5×10^{-4} and learning rate is 2.5×10^{-4} with polynomial decay of power 0.9. Meanwhile, the Adam optimizer is used for training the discriminator network, whose momentum is $\beta_1 = 0.9, \beta_2 = 0.99$, weight decay is 5×10^{-4} and learning rate is 1×10^{-4} with polynomial decay of power 0.9. We adopt the same semantic segmentation and discriminator network architecture as that of [16]. In the cross-modal semantic segmentation experiments, we follow the exactly same data augmentation and preprocess procedure as that of [7]. The hyperparameter δ in Eq. (12) and Eq. (13) of the main paper is set as 0.2. We use the Adam optimizer for training the 2D and 3D semantic segmentation network, with $\beta_1 = 0.9, \beta_2 = 0.999$. The learning rate is set as 1×10^{-3}.

S2. Datasets Overview of mDALU Benchmark

In Sec. 4 of the main paper, we introduce the benchmark setup of the mDALU problem. Here we provide more details about the datasets involved in the benchmark.

S2.1. Image Classification

In the image classification benchmark of the main paper, we adopt three digits datasets, including MNIST [9], Synthetic Digits [3], and SVHN [11] dataset. MNIST is a hand-written numbers image dataset, SVHN is a street view house numbers image dataset and Synthetic Digits is a synthetic numbers image dataset. In the image classification benchmark of the main paper, we adopt these three different style digits images, to introduce larger domain gap between different source domains to effectively evaluate the validity of different methods for mDALU problem. In Sec. S3, we introduce two more datasets, MNIST-M [3] and USPS [6] , to evaluate the effectiveness of our approach when dealing with more than two source domains. MNIST-M is a synthetic numbers image dataset, and USPS is a hand-written numbers image dataset. We follow the setup of splitting the dataset in [12, 13]. In each of MNIST, MNIST-M, SVHN and Synthetic Digits, 25000 images for training are sampled from the training subset, and 9000 images for testing are sampled from the testing subset. And for the USPS dataset, due to there are only 9298 images in total are available, the whole training set covering 7438 images is used for training, while the whole testing set with 1860 images is adopted for testing. MNIST, MNIST-M, SVHN, Synthetic Digits, USPS are abbreviated as MT, MM, SVHN, SYN, and UP, respectively. The detailed label space of different source



(a) Partially-Supervised Adaptation (b) Fully-Supervised Adaptation

Figure S1: Overview of our approach to mDALU problem. Our approach is composed of two stages: (a) partially-supervised adaptation stage, and (b) fully-supervised adaptation stage. In the partially-supervised adaptation stage, there are three modules involved, the domain attention (DAT) module, the uncertainty maximization (UM) module, and the attention-guided adversarial alignment (A^3) module. Besides the supervised semantic segmentation loss $\mathcal{L}_{psu}$ on the source domain, the DAT module is trained in the supervised way with $\mathcal{L}_{att}$, the UM module is trained in the supervised way with $\mathcal{L}_{um}$ and the A^3 module is trained in the adversarial way with $\mathcal{L}_{a^3} + \mathcal{L}_d$. In the fully-supervised adaptation stage, in order to complete the label space, the pseudo-label, for all the samples $\mathbf{x}^{s_1}$, $\mathbf{x}^{s_2}$, $\mathbf{x}^t$ from all related domains, is generated by fusing the probability map weighted by attention map from different branches, G_1, M_1 and G_2, M_2. Then the semantic segmentation network G is trained in the complete and unified label space with the generated pseudo-label and the supervised loss $\mathcal{L}_{fsa}$. In the implementation, G_1, G_2, M_1, M_2 share the same encoder and adopt different label predictors.

(a) MT (b) SYN (c) SVHN (d) MM (d) UP

Figure S2: Example images of different datasets in mDALU image classification benchmark.

domains and the target domain under different experiments setup is listed in Table S1 and Table S2. The example images of different datasets are shown in Fig. S2.

S2.2. 2D Semantic Image Segmentation

In the 2D semantic image segmentation benchmark of the main paper, we adopt the synthetic image datasets, GTA5 [14] and SYNTHIA [15] and the real image dataset, Cityscapes [2]. We introduce the label space of different datasets in the main paper. Here we provide more additional information about the datasets.

Cityscapes. Cityscapes is a dataset composed of the street scene images collected from different European cities. We use the training set of Cityscapes covering 2993 images, without the label information, as the target domain during the training stage. And we adopt the validation set of Cityscapes, which are composed of 500 images and densely labeled with 19 classes, to evaluate the semantic segmentation performance of the model on the target domain.

GTA5. GTA5 is a synthetic urban scene image dataset, whose images are rendered from the game engine. The scene of the images is based on the city of Los Angeles. In our 2D semantic image segmentation benchmark, we use 24966 densely labeled images in the GTA5 dataset as one of our source domains, whose annotation is compatible with that of Cityscapes.

SYNTHIA. SYNTHIA is a synthetic dataset, containing photo-realistic images rendered from a virtual city. We use the SYNTHIA-RAND-Cityscapes subset, which contains 9400 densely labeled images and the 16 class annotation of which is compatible with that of Cityscapes. In our 2D semantic image segmentation benchmark, the labeled SYNTHIA dataset serves as one of our source domains.

S2.3. Cross-Modal Semantic Segmentation

In the cross-modal semantic segmentation benchmark of the main paper, three datasets are involved, Cityscapes [2], Nuscenes [1] and A2D2 [5]. We introduce the label space of different datasets in the main paper. Here we provide more information on the datasets and the mapping between our label space and the annotated class label in different datasets.

Cityscapes. Cityscapes [2] is a 2D urban scene image dataset, and has been introduced in the Sec. S2.2. In the cross-modal semantic segmentation benchmark, we adopt the training set of Cityscapes, covering 2975 images, as the

Experiment	Label Space									
Non-Overlapping(Table 2 in main paper)	Domain	Source1	Source2	Target	Source1	Source2	Target	Source1	Source2	Target
	Dataset	SVHN	SYN	MT	MT	SVHN	SYN	MNIST	SYN	SVHN
	Class	0∼4	5∼9	0∼9	0∼4	5∼9	0∼9	0∼4	5∼9	0∼9
Partially-Overlapping(Table 4 in main paper)	Domain	Source1	Source2	Target	Source1	Source2	Target	Source1	Source2	Target
	Dataset	SVHN	SYN	MT	MT	SVHN	SYN	MNIST	SYN	SVHN
	Class	0∼6	3∼9	0∼9	0∼6	3∼9	0∼9	0∼6	3∼9	0∼9

Table S1: The label space of different source domains and the target domain in the mDALU image classification benchmark of the main paper.

More Source Domains Experiments (Table S5 in supplementary)					
Domain	Source1	Source2	Source3	Source4	Target
Dataset	SVHN	SYN	MM	UP	MT
Class	0∼2	2∼4	4∼6	7∼9	0∼9
Dataset	MT	SYN	MM	UP	SVHN
Class	0∼2	2∼4	4∼6	7∼9	0∼9
Dataset	MT	SVHN	MM	UP	SYN
Class	0∼2	2∼4	4∼6	7∼9	0∼9
Dataset	MT	SVHN	SYN	UP	MM
Class	0∼2	2∼4	4∼6	7∼9	0∼9
Dataset	MT	SVHN	SYN	MM	UP
Class	0∼2	2∼4	4∼6	7∼9	0∼9

Table S2: The label space of different source domains and the target domain in the mDALU image classification benchmark of the more source domains experiments in the supplementary.

2D source domain. Unlike the Sec. S2.2 does not use the label information of Cityscapes training images, we use the ground truth label of Cityscapes training images, but the label space of Cityscapes in our experiments only covers 6 classes, road, sidewalk, building, pole, sign and nature. The mapping from the original Cityscapes annotated classes and our label space is listed in Table S4.

Nuscenes. Nuscenes [1] is an autonomous driving dataset covering 1000 driving scenes, which are collected from the Boston and Singapore. Each scene, of 20-second length, is sampled and annotated at 2HZ, resulting in 40K well-annotated keyframes for 3D bounding boxes of the objects. In our cross-modal semantic segmentation benchmark, we adopt the training set of the Nuscenes, including 28130 keyframes 3D LiDAR points, as the 3D source domain. Then as done in [7], we generate the 3D point-wise semantic labels from the 3D bounding boxes, by assigning the object label to the points inside the bounding box and taking the points outside the bounding box as unlabeled points. The label space of the 3D source domain includes 4 classes, person, car, truck and bike. The mapping between the object label annotation in Nuscenes and our label space is reported in Table S4.

A2D2. A2D2 [5] is an autonomous driving dataset, including simultaneously recorded paired 2D images and 3D LiDAR points. The A2D2 covers 20 scenes, which are corresponding to 28637 frames for training. And the scene 20180807_145028 is used for validation. The 2D images are densely labeled with 38 semantic classes. Following [7], the 3D point-wise semantic labels are generated by the re-projection to the 2D images. In our cross-modal semantic segmentation benchmark, the A2D2 serves as the target domain. We use the training set of A2D2 without the label information during training, including the paired 2D images and 3D LiDAR points. And we use the validation set 20180807_145028 with the ground truth label for evaluat-

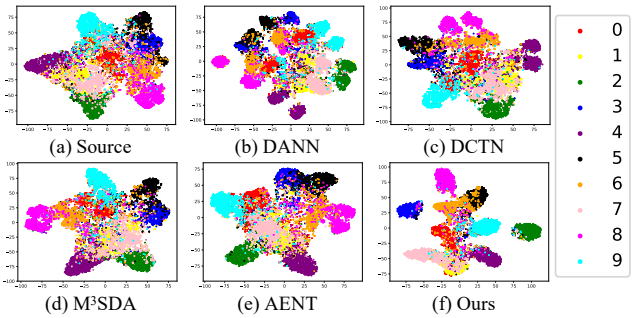


Figure S3: t-SNE Visualization of the feature embedding on the mDALU image classification benchmark, MT, SVHN, MM, UP → SYN. We adopt the same t-SNE parameters for all visualization.

Method	wall	fence	pole	person	rider	motorcycle	bicycle	mIoU
Ours (PSF w/o relabeling)	11.5	17.8	**33.6**	47.7	13.2	7.2	**43.4**	38.4
Ours (PSF w/ relabeling)	**13.3**	**17.9**	30.6	**53.7**	**18.2**	**19.8**	43.2	**40.0**

Table S3: Quantitative comparison of w/ and w/o relabeling inconsistent taxonomies in PSF module. The detailed performance on inconsistent taxonomies classes is also shown. The mIoU is reported for 19 classes. The best results are denoted in bold.

ing the performance. The label space of the target domain for evaluation includes 10 classes, road, sidewalk, building, pole, sign, nature, person, car, truck and bike. The mapping between the label space and the annotated 38 semantic classses in A2D2 is shown in Table S4.

S3. Experiments with More Source Domains

In this section, we evaluate the effectiveness of our approach when dealing with more than two source domains. Based on the classification benchmark of the main paper, we here introduce two more datasets, MNIST-M [3] and USPS [6], which are abbreviated as "MM" and "UP" respectively. Then as done in the main paper, each time, one of "MT", "SYN", "SVHN", "'MM' and "UP" is taken as the target domain, while the other four are used as source domains. The label space of different source domains in the experiments is listed in Table S2.

Experimental results. In Table S5, we report the quantitative experimental results of the classification benchmark, after introducing two more datasets, MM and UP. It can be seen that our approach with the "partially-supervised adaptation" stage highly outperforms the source-only baseline, the adaptation-based methods DANN, DCTN, and M^{3}SDA, and the label-unification based method AENT. It achieves an average accuracy of 80.83% on the target domain. Then by exploiting the "fully-supervised adaptation" stage, the performance is further improved to 82.88%. It proves the effectiveness and the robustness of our approach for addressing the mDALU problem when more than two source domains are given. In Fig. S3, the qualitative comparison of feature embedding, t-SNE visualization [10], between our approach and other methods is shown. It shows that our approach is able to learn more discriminative features than other methods. It further verifies the good performance of our approach to mDALU problem.

S4. More Experimental Results for Semantic Segmentation

Detailed experimental results for semantic segmentation. In Table 5a and Table 8 of the main paper, we show the quantitative comparison, through the mIoU, between our approach and other methods, on the 2D and cross-modal semantic segmentation benchmark. Correspondingly, we here provide more detailed experimental results in Table S6 and Table S7, covering the per-class IoU results.

Attention visualization for semantic segmentation. During the "partially-supervised adaptation" stage, we introduce the attention map in the domain attention (DAT) module, the attention-guided adversarial alignment (A^3) module and the inference via attention-guided fusion. In order to verify the effectiveness of our attention map prediction, we show the qualitative visualization of the attention map on the target domain images in Fig. S4. Corresponding to the Sec. 3.2.1 of the main paper, the attention map $\tilde{\mathbf{a}}_1^t$ and $\tilde{\mathbf{a}}_2^t$, are generated by feeding the target domain image $\mathbf{x}^t$ into the attention network M_1 and M_2. It is shown that our predicted attention map $\tilde{\mathbf{a}}_1^t$, corresponding to the source domain $\mathcal{S}_1$, has higher attention value, for the objects belonging to the partial label space $\mathcal{C}_1$, such as the road, sidewalk, building, vegetation, sky and car. And the predicted attention map $\tilde{\mathbf{a}}_2^t$, corresponding to the source domain $\mathcal{S}_2$, has higher attention value, for the objects belonging to the partial label space $\mathcal{C}_2$, such as the fence, pole, light, sign, bus, motorcycle and bicycle. It proves the validity of our attention map prediction.

Additional qualitative results for semantic segmentation. In Fig. 4 of the main paper, we show the qualitative comparison results between our approach and other methods on the 2D semantic image segmentation benchmark, and the source domain images are not translated with CycleGAN [20], *i.e.,* the "NT" setting. Here we provide additional qualitative comparison results between our approach and other methods on the 2D semantic image segmentation benchmark, and the source images are translated with CycleGAN [20], *i.e.,* the "T" setting. As shown in Fig. S5, it can be seen that our approach obviously outperforms other methods on the 2D semantic image segmentation benchmark. It further verifies the effectiveness of our approach to mDALU problem.

Comparison between w/ and w/o relabeling inconsistent taxonomies in the source domain. In Sec. 3.2.6 of the main paper, we introduce the extension of our method for inconsistent taxonomies. In the PSF module, besides the unlabeled samples in the source domain being completed with the predicted pseudo-label as in Eq. (12), we add Eq. (17) to relabel the conflict part $\mathbf{c}_p^q \cap \mathbf{c}_m^n$ in the source domain $\mathcal{S}_m$. In Table 7 of the main paper, we show the performance of our extended method 40.0% under inconsistent taxonomies setting, which outperforms other competing methods significantly and proves the effectiveness of our extended method for inconsistent taxonomies. Here, we compare the ablation of our extended method, w/o relabeling inconsistent taxonomies in the source domain, against our full extended method, to further verify the effectiveness

label space	A2D2	Cityscapes	Nuscenes
road	'rd normal street', 'zebra crossing', 'solid line', 'rd restricted area', 'slow drive area', 'drivable cobblestone', 'dashed line', 'painted driv. instr.'	'road'	–
sidewalk	'sidewalk', 'curbstone'	'sidewalk'	–
building	'buildings'	'building'	–
pole	'poles'	'pole'	–
sign	'traffic sign 1', 'traffic sign 2', 'traffic sign 3'	'traffic sign'	–
nature	'nature object'	'vegetation', 'terrain'	–
person	'pedestrian 1', 'pedestrian 2', 'pedestrian 3'	–	'pedestrian'
car	'car 1', 'car 2', 'car 3', 'car 4', 'ego car'	–	'car'
truck	'truck 1', 'truck 2', 'truck 3'	–	'truck'
bike	'bicycle 1', 'bicycle 2', 'bicycle 3', 'bicycle 4', 'small vehicles 1', 'small vehicles 2', 'small vehicles 3'	–	'motorcycle', 'bicycle'

Table S4: Class mapping between the label space and the annotated classes in different datasets.

Method	MT	SYN	SVHN	MM	UP	Avg
Source	86.90 ± 0.40	63.80 ± 0.15	51.84 ± 2.13	52.09 ± 0.69	91.83 ± 0.78	69.29 ± 0.83
DANN[4]	86.38 ± 1.44	63.76 ± 0.88	51.58 ± 2.27	52.14 ± 0.61	89.98 ± 1.42	68.77 ± 1.32
DCTN [18]	63.87 ± 0.10	53.33 ± 1.15	43.57 ± 0.98	40.23 ± 0.48	59.78 ± 1.19	52.16 ± 0.78
M^3SDA [12]	87.26 ± 1.54	63.40 ± 0.32	48.96 ± 0.92	52.28 ± 1.60	90.20 ± 0.97	68.42 ± 1.07
AENT[19]	79.55 ± 2.40	63.22 ± 0.41	52.58 ± 2.27	48.65 ± 0.31	87.62 ± 1.36	66.32 ± 1.35
Ours w/o PSF	**94.90±0.23**	**78.37±0.58**	**72.18±0.44**	**63.01±0.74**	**95.70±0.44**	**80.83±0.49**
Ours	**96.60±0.07**	**80.68±0.30**	**73.82±0.35**	**66.62±0.62**	**96.70 ± 0.22**	**82.88±0.31**

Table S5: Quantitative comparison between our method and other SOTA methods, under mDALU image classification benchmark with 4 source domains. "MT", "SYN", "SVHN", "MM", and "UP" represent the target domain. We implement AENT on classification by utilizing the ambiguity cross entropy loss proposed in [19]. The best results are denoted in bold.

of relabeling inconsistent taxonomies as in Eq. (17). From Table S3, it is shown that the performance is improved by 1.6% from 38.4% to 40.0%, by relabeling inconsistent taxonomies in the source domain. And the detailed performance comparison on the inconsistent taxonomies classes in Table S3 also proves the effectiveness of relabeling inconsistent taxonomies in the source domain.

References

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 2, 3

[2] 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

[3] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 1, 4

[4] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 5

[5] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020. 2, 3

[6] Jonathan J. Hull. A database for handwritten text recognition research. *TPAMI*, 16(5):550–554, 1994. 1, 4

[7] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Perez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *CVPR*, 2020. 1, 3

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[9] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010. 1

[10] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008. 4

[11] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS workshops*, 2011. 1

[12] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. 1, 5

[13] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. In *ICLR*, 2020. 1

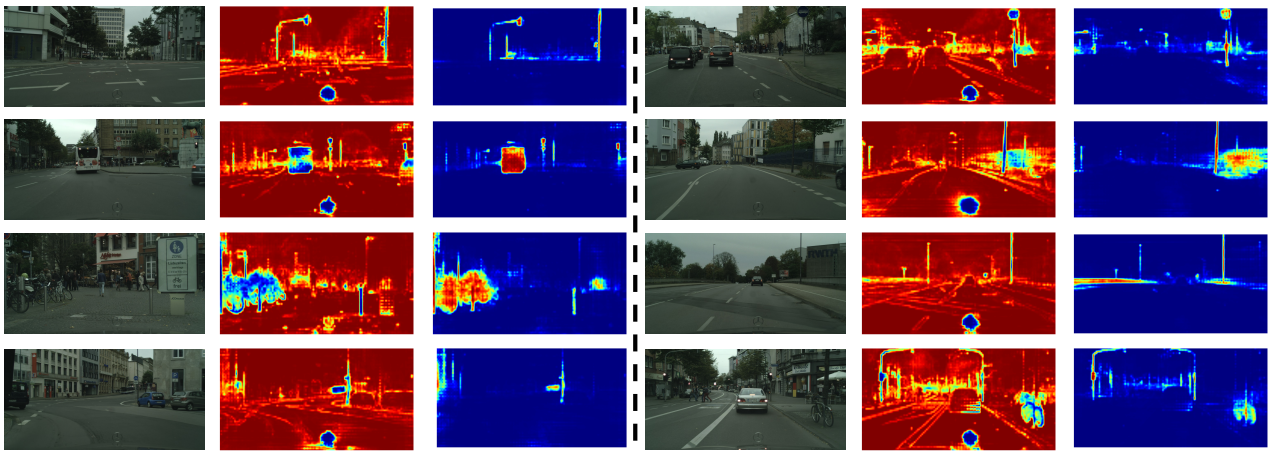
[14] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 2

GTA5+SYNTHIA→Cityscapes																					
Setting	Method	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrian	sky	person	rider	car	truck	bus	train	motorbike	bicycle	mIoU
NT	Source	3.0	10.1	42.1	7.3	6.6	10.6	18.2	**31.2**	61.3	3.9	73.2	27.5	16.2	9.9	1.4	1.6	0.0	8.6	3.8	17.7
	AdaptSegNet[16]	0.1	0.0	1.4	4.0	6.6	5.4	14.6	22.8	5.9	1.9	35.9	1.3	18.0	0.6	3.0	1.8	0.7	13.0	9.4	7.7
	MinEnt[17]	32.0	10.0	73.0	15.4	18.1	20.5	29.5	19.9	75.3	3.9	**79.6**	51.3	18.7	18.5	4.3	4.8	**9.2**	**20.3**	10.3	27.1
	Advent[17]	6.3	1.0	27.7	4.5	6.3	6.5	16.9	19.3	16.7	2.0	40.6	6.8	17.1	7.7	3.7	6.6	1.2	15.0	18.5	11.8
	Ours (w/o PSF)	**82.8**	30.8	78.9	17.5	15.8	28.0	34.8	18.9	79.1	10.5	78.4	52.0	18.2	71.4	16.8	34.3	2.0	11.0	8.0	36.3
	Ours (ADV)	82.1	**35.2**	78.1	**27.3**	18.8	**29.6**	33.0	21.1	78.3	**36.9**	75.3	**58.9**	**25.0**	69.6	19.3	33.8	0.0	15.6	**22.9**	40.1
	Ours (PSF)	77.8	31.9	**79.5**	17.9	18.1	29.0	**34.9**	20.9	80.2	9.0	**79.6**	55.6	20.9	**74.4**	16.9	25.5	0.0	17.0	18.9	37.3
	Ours (ADV+PSF)	81.7	34.1	**79.5**	26.7	**19.4**	29.0	32.0	23.2	**82.3**	31.4	79.5	57.5	22.3	66.6	**26.8**	**40.2**	0.0	19.4	20.4	**40.6**
T	Source	28.7	9.5	52.3	11.1	10.0	9.5	16.4	**30.6**	55.9	2.7	67.5	40.8	21.1	38.7	6.9	4.3	**6.4**	**22.1**	20.6	24.0
	AdaptSegNet[16]	78.3	34.5	75.7	16.2	15.6	11.5	19.0	10.8	78.0	16.5	76.3	42.6	8.4	59.6	10.9	8.8	0.5	14.2	8.7	30.8
	MinEnt[17]	58.5	20.6	70.5	12.0	17.9	18.3	19.9	27.1	74.3	8.0	79.1	46.5	20.5	37.7	9.1	20.4	2.8	18.9	10.6	30.1
	Advent[17]	78.0	34.3	75.9	14.5	5.8	9.8	17.2	10.2	76.4	15.0	76.9	40.6	3.1	61.3	19.3	14.5	0.0	9.9	12.5	30.3
	Ours(w/o PSF)	86.0	40.8	79.1	13.2	22.7	33.5	33.3	18.9	79.9	33.2	72.0	49.7	19.1	63.3	20.6	10.1	0.0	13.4	34.0	38.1
	Ours (ADV)	86.2	41.3	81.6	21.1	**23.3**	33.4	32.0	20.6	81.0	32.1	**79.8**	57.5	**26.4**	70.5	**24.8**	**31.4**	0.2	18.3	27.1	41.5
	Ours (PSF)	**87.8**	**42.9**	81.2	17.3	22.0	34.1	**36.9**	17.9	82.2	34.2	73.6	**58.9**	25.1	**76.5**	24.4	28.9	0.1	19.8	**41.9**	42.4
	Ours (PSF+ADV)	86.8	42.5	**82.5**	**23.0**	23.1	**34.4**	36.3	29.1	**82.9**	**34.3**	76.5	56.5	24.1	75.5	23.6	17.3	0.3	22.0	41.6	**42.8**

Table S6: Per-Class IoU on the mDALU 2D semantic image segmentation benchmark. "NT" means source domain images are not translated with CycleGAN, and "T" means source domain images are translated with CycleGAN. The mIoU results are reported over 19 classes. The best results are denoted in bold.

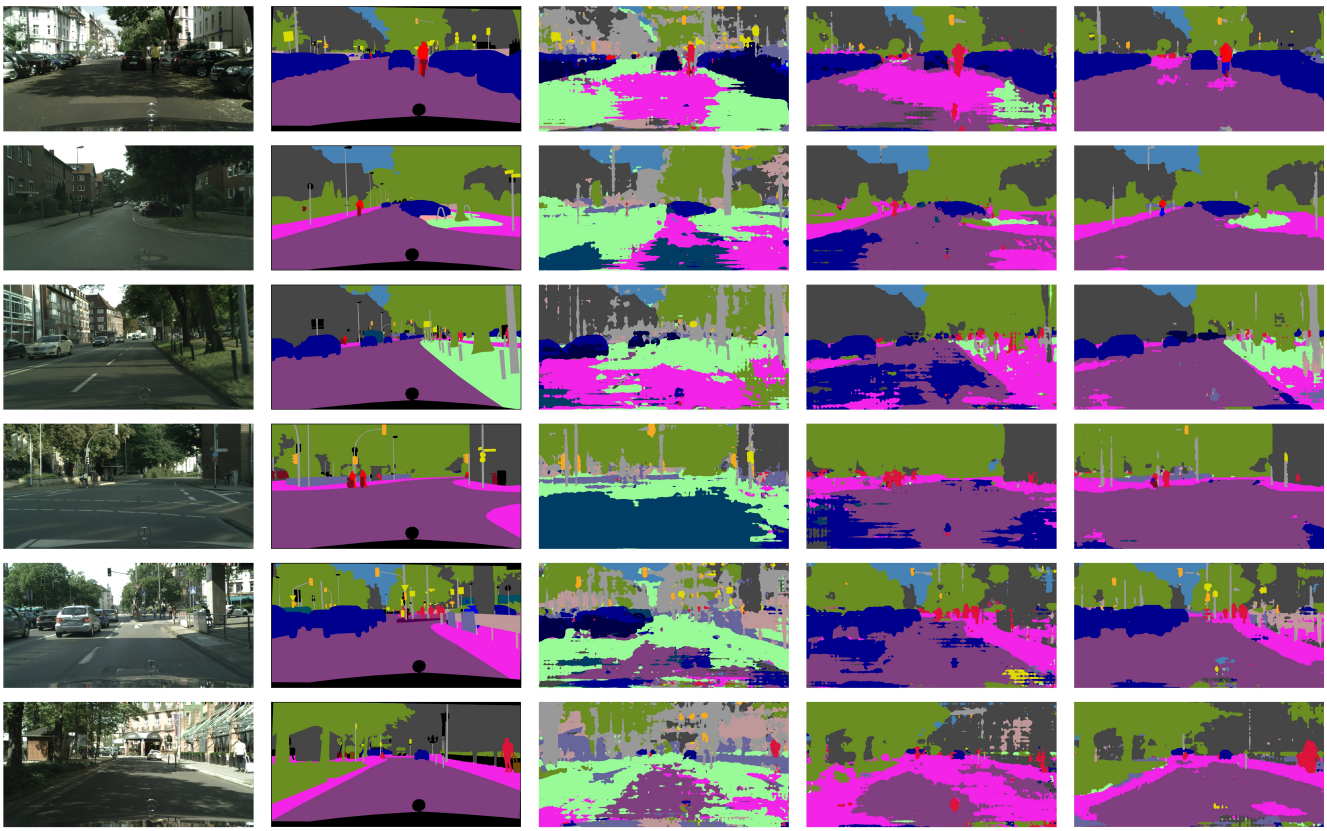
Cityscapes+Nuscenes→A2D2												
Modality	Method	road	sidewalk	building	pole	sign	nature	person	car	truck	bike	mIoU
2D	Sources	83.1	48.7	85.0	**34.8**	36.1	87.0	0.0	0.0	0.0	0.0	37.5
	xMUDA	68.2	13.8	22.3	22.1	15.7	3.6	0.1	15.9	1.6	0.0	16.3
	ES + MinEnt	55.7	15.6	64.9	19.8	21.7	45.2	0.0	0.0	0.0	0.0	22.3
	ES + KL	14.4	20.0	74.2	15.3	36.6	46.2	0.0	9.3	1.5	0.0	21.7
	xMUDA + AKL	44.8	29.7	46.5	36.2	33.6	61.4	0.0	21.0	1.8	0.0	27.5
	xMUDA + AKL + COMP	70.3	38.1	76.4	25.0	30.5	80.8	0.0	0.0	0.0	0.0	32.1
	Ours (w/o PSF)	85.8	54.3	81.8	34.1	40.8	81.4	0.0	0.0	2.8	0.0	38.1
	Ours	**92.8**	**59.9**	**90.0**	30.4	**60.7**	**90.6**	**13.8**	**71.6**	**39.1**	**0.4**	**54.9**
3D	Source	0.0	0.0	0.0	0.0	0.0	0.0	2.1	16.1	1.4	0.0	2.0
	xMUDA	0.0	0.0	0.0	0.0	0.0	0.0	1.2	14.6	1.5	0.0	1.7
	ES + MinEnt	0.0	0.0	0.0	0.0	0.0	0.0	1.9	12.0	1.5	0.0	1.5
	ES + KL	0.0	0.0	0.0	0.0	0.0	0.0	1.9	9.8	1.8	1.2	1.5
	xMUDA + AKL	0.0	0.0	0.0	0.0	0.0	0.0	2.4	18.5	1.3	0.4	2.3
	xMUDA + AKL + COMP	6.1	1.8	0.0	0.0	0.0	0.0	2.1	17.9	1.3	0.0	2.9
	Ours (w/o PSF)	0.6	0.7	0.3	0.0	0.0	2.4	1.3	16.1	2.3	0.0	2.4
	Ours	**82.0**	**27.7**	**80.3**	**1.4**	**7.5**	**80.8**	**7.2**	**54.9**	**25.6**	**3.5**	**37.1**
Fuse	Source	85.5	51.8	83.8	**41.8**	40.2	83.8	6.3	23.0	8.8	0.0	42.5
	xMUDA	55.8	2.2	2.8	3.5	2.8	0.2	2.7	19.4	1.7	0.0	9.1
	ES + MinEnt	63.1	7.5	69.7	9.0	13.8	30.2	2.6	11.0	1.2	0.0	20.8
	ES + KL	10.6	21.2	65.0	18.2	26.8	34.7	5.4	12.0	2.4	0.3	19.7
	xMUDA + AKL	13.2	36.9	20.1	34.1	31.1	44.5	4.6	24.8	1.7	0.1	21.1
	xMUDA + AKL + COMP	74.1	43.5	74.4	35.2	35.5	71.0	4.1	34.7	5.0	0.0	37.7
	Ours(w/o PSF)	91.1	57.3	85.7	39.7	47.4	85.9	8.6	57.8	25.3	0.4	49.9
	Ours	**91.7**	**58.6**	**90.1**	34.5	**58.8**	**90.3**	**15.4**	**72.4**	**43.6**	**1.3**	**55.7**

Table S7: Per-Class IoU on the mDALU cross-modal semantic segmentation benchmark. The mIoU results are reported over 10 classes. The best results are denoted in bold.



(a) Cityscapes (b) Attention1 (c) Attention2 (a) Cityscapes (b) Attention1 (c) Attention2

Figure S4: Visualization of the attention map $\tilde{\mathbf{a}}_1^t$ and $\tilde{\mathbf{a}}_2^t$ of the target domain images. (a) is the Cityscapes image $\mathbf{x}^t$. (b) is the attention map $\tilde{\mathbf{a}}_1^t$, generated by feeding the $\mathbf{x}^t$ into the attention network M_1. (c) is the attention map $\tilde{\mathbf{a}}_2^t$, generated by feeding the $\mathbf{x}^t$ into the attention network M_2. Red parts are the parts with higher attention value, while the blue parts with lower attention value.



(a) Image (b) Ground Truth (c) Source Only (d) AdaptSegNet (e) Ours

Figure S5: Qualitative comparison of semantic segmentation results, under the mDALU 2D semantic image segmentation benchmark, SYNTHIA + GTA5 $\rightarrow$ Cityscapes. The source images are translated with CycleGAN, *i.e.,* setting "T".

[15] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 2

[16] 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*, 2018. 1, 6

[17] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Mathieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 6

[18] Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *CVPR*, 2018. 5

[19] Xiangyun Zhao, Samuel Schulter, Gaurav Sharma, Yi-Hsuan Tsai, Manmohan Chandraker, and Ying Wu. Object detection with a unified label space from multiple datasets. In *ECCV*, 2020. 5

[20] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *ICCV*, 2017. 4