Supplementary Material: Attention-based Class-Conditioned Alignment for Multi-Source Domain Adaptation of Object Detectors

Atif Belal[1], Akhil Meethal[1], Francisco Perdigon Romero[2], Marco Pedersoli[1], Eric Granger[1]
[1] LIVIA, ILLS, Dept. of Systems Engineering, ETS Montreal, Canada
[2] GAIA Montreal, Ericsson Canada
atif.belal.1@ens.etsmtl.ca, akhilpm135@gmail.com, francisco.perdigon.romero@ericsson.com,
{marco.pedersoli, eric.granger}@etsmtl.ca

1. Datasets

In our paper, we used five datasets to create the multi-source domain adaptation settings. These datasets are listed below and summarized in the Tab. 1:

1. BDD100k - The BDD100K [16] is a large-scale diverse driving dataset. It contains 70,000 training and 10,000 testing images captured across various times, like Daytime, Night, and Dusk/Dawn. This variation makes it a good choice for the DA problem.

2. Cityscapes - Cityscapes [3] is an autonomous driving research dataset with images captured from urban street scenes. It contains 2,975 training and 500 testing images.

3. Kitty - The KITTI [5] dataset is a self-driving dataset that comprises a collection of images and associated sensor data captured from a moving vehicle in urban environments. It consists of 7,481 training images RGB images.

4. MS COCO - MS COCO [7] is one of the most widely used benchmark datasets in computer vision. It is a complex dataset having large scale and appearance variations for the instances. It has around 330,000 images containing 80 object categories.

5. Synscapes - Synscapes [12] is a synthetic autonomous driving dataset that provides more variability for testing our method. It consists of 25,000 training images.

2. Study on the Class-Embedding layer

In this section, we visualize the information learned by the class-embedding layer. Fig. 1 shows the activation of each class-embedding layer (corresponding to each object category) with ROI-Pooled features that contain the object category. For this, we crop the region of the object category from the image and find activation with each class-embedding layer. After that, we use Softmax to normalize the values. It can be observed from Fig. 1 that each embedding layer is activated most when it matches the corresponding object category. This shows that the class-embedding is successfully learning class information. Also, the under-represented object categories (bike, truck, truck) are highly activated with their corresponding embedding layer. It shows that our attention-based instance-level alignment mechanism is helping in datasets with class imbalance by focusing on under-represented classes.

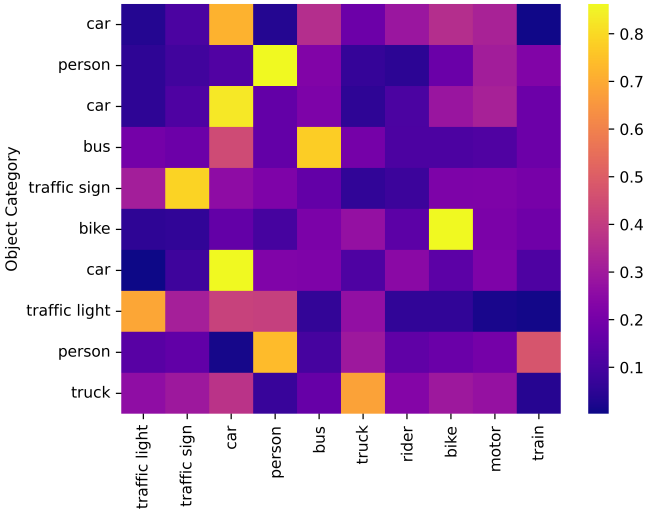


Figure 1. Heatmap showing the activation of each class embedding with instance features of some object categories. The X-axis represents the list of all the class embedding corresponding to all the object categories, and the Y-axis represents the object category present in the ROI-Pooled feature.

3. Increase in number of parameters with source domains

Earlier works in the MSDA for OD [13, 15] relied on learning domain-specific parameters. It rapidly increases training parameters with each source domain, as summarized in Tab. 2. Later, PMT [1] mitigated this by using prototypes instead of domain-specific subnets. However, in PMT the number of parameters slightly increases with each source domain. In contrast, our ACIA doesn't have any domain-specific parameters, so there is no change in the number of parameters with source domains.

Setting	Src.1	# Img.	Src.2	# Img.	Src.3	# Img.	Target	# Img.
Cross-Time	Day	36,728	Night	27,971	-	-	Dawn	5,027
Cross-Camera	Cityscapes	2,831	Kitty	6,684	-	-	Day	36,728
Mixed Domain	Cityscapes	2,975	COCO	71,745	Synscapes	25,000	Day	36,728

Table 1. Summary of the different MSDA settings used in our study.

Method	Number of source domains				
	1	2	3	4	5
DMSN	45.994	75.426	104.858	134.290	163.722
TRKP	45.994	59.942	73.890	87.838	101.786
PMT	46.586	46.587	46.588	46.589	46.590
ACIA (ours)	45.659	45.659	45.659	45.659	45.659

Table 2. The number of parameters vs the number of source domains. No parameter growth in our method.

4. Equalization Loss v2 for class-imbalance

Focal loss was introduced as an improvement over cross-entropy loss, it helps in dealing with foreground-background imbalance. But, the MSDA setting (cross-time) used in our paper have foreground-foreground class imbalance. In this section, we replace the focal loss for object detection with equalization loss v2 [11] to tackle the problem of class imbalance. EQLv2 loss was proposed for long-tail object detection and have demonstrated significant improvements in detecting under-represented objects without sacrificing performance on more frequent classes. Equalization Loss v2 adjusts the gradients during backpropagation to address class imbalance in object detection. It applies a class-specific weighting factor to the gradient of the classification loss. The frequent classes are down-weighted and rare classes are up-weighted based on their frequency, thereby balancing the learning process for all classes. In Tab. 3, we compare the performance of EQL v2 loss with focal loss on the cross-time setting. We also trained PMT with this loss for comparison. It can be observed that in both the lower and upper bound, EQL v2 is outperforming focal loss. This shows the effectiveness of EQL v2 loss in a class-imbalance problem. In case of PMT, EQL v2 is improving their performance by a slight margin. This shows that PMT is not very effective in imbalance dataset scenarios. For ACIA, there is no improvement when focal loss is replaced by EQLv2 loss. This further proves that our method is every effective when dealing with class imbalance problem.

5. Additional Experiments on the Cross-Time Setting

In this section, we provide some additional results on the cross-time settings. Here, we present a setting where source domain contains images which are mostly in bright environment while target domain contains dark/dull environment. For this, we used Daytime and Dusk/Dawn domain of BDD100K as source domains, while Night domain of BDD100K is used as target domain. This setting is challenging due to large domain shift between source and target domain. The result for this setting is reported in Tab. 4, we also trained PMT and reported their performance for comparison. It can be observed that, we are outperforming PMT by 2.3 mAP. This shows that our instance-level alignment performs better when the domain shift is large between source and target domain.

Setting	Method	Focal Loss	EQLv2
Lower Bound	Source Only	28.9	31.6
MSDA	PMT	45.3	45.7
	ACIA (Ours)	**47.9**	**47.5**
Upper Bound	Target-Only	26.6	28.3
	All-Combined	45.6	45.8
	Fine-Tuning	50.9	51.7

Table 3. Comparison of performance when using focal loss and equalization loss v2 for OD on the cross-time setting.

Setting	Method	mAP
Lower Bound	Source Only	24.2
MSDA	PMT	34.9
	ACIA (Ours)	**37.2**
Upper Bound	Target-Only	37.8
	All-Combined	42.1
	Fine-Tuning	46.8

Table 4. mAP performance of ACIA and baseline methods for cross-time adaptation on BDD100k, when the two sources are Daytime and Dusk/Dawn subsets, and the target is the *Night* subset.

6. Additional Experiments on the Mixed Domain Adaptation Setting

In this section, we provide some additional results on the mixed domain adaptation settings. In the main paper, we only showed the results when $C + M$ was employed as the two source domains (because the previous papers followed that setting only). In Tab. 5 we compare our results with PMT[1] when $C + S$ and $M + S$ are employed as the source domain. It can be observed that our method outperforms PMT for both settings. This further shows that our attention-based instance alignment performs better than prototype-based instance alignment with complex domain shifts.

Method	C+S	M+S
PMT	30.1	34.9
ACIA (ours)	**33.7**	**35.8**

Table 5. Additional results on the Mixed Domain Adaptation Setting.

7. Effect of Class-Alignment in UDA methods

In this section, we show the effect of class-wise alignment on UDA methods [6, 9]. Tab. 6 shows the results. We studied three different cases: (1) No class alignment - we used their proposed method only. (2) Class-Alignment with target domain - here we incorporated our class alignment component with their method, aligning the source domains and the target domain. (3) Class-Alignment without target domain - our class alignment component with their method, but this time only the source domains are considered. The results clearly show that our class alignment is effective with both methods. Thus, we can conclude that the proposed class-conditional alignment is effective for both UDA and MSDA methods. It can be observed that similar to our method, removing target data for the instance-level alignment further enhances the model performance.

Method	No Class Align.	Class-Align. w/ Target	Class-Align. w/o Target
Strong-Weak [9]	29.9	33.7	34.2
Adaptive Teacher [6]	34.6	36.8	37.6

Table 6. Importance of Class-alignment in UDA methods. The proposed attention-based class-conditioned aligner is effective for UDA methods as well.

8. Architecture of the Discriminator Networks

We use two domain classifiers as the discriminator networks for adversarial training: an image-level domain classifier and an instance-level domain classifier. Their architecture is summarized in Fig. 2. The image-level domain classifier receives its input from the final layer of the backbone network used for feature extraction. This classifier is fully convolutional with a final N+1 class prediction layer (corresponds to the number of source domains plus the target domain). The instance-level classifier receives its input from the attention layer. This classifier consists of only linear layers with a final N-way prediction layer (the target domain is not used here because the GT boxes from the target domain are not available).

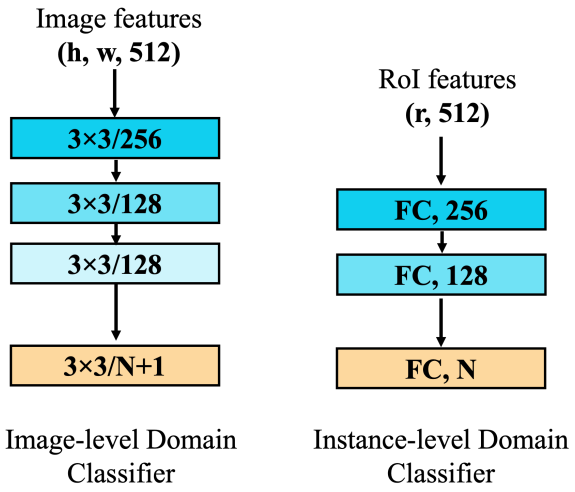


Figure 2. Detailed architecture of the networks used for image-level and instance-level domain classifiers. The activation functions used in image-level and instance-level classifiers are leaky ReLU and GELU respectively. Additionally, Layernorm is used in the instance-level domain classifier. (r= no of GT boxes in the image, h,w = height and width of the feature map, FC = Fully connected layer)

9. Class-wise AP

We also report the detailed class-wise AP of ACIA on the Cross-Time and Mixed Adaptation settings in Tab. 7 and Tab. 8 respectively. Note that, for the Cross-camera Adaptation settings, there is only one class, so this is not applicable.

In both settings, our method outperforms the others for all classes. The class-wise AP shows improvement both in the majority and minority classes as well - eg: car (majority) and traffic sign (minority) in the cross-time settings. In the mixed adaptation settings, we outperform others for all the classes, in both the case of two and three source domains.

[1] We trained for this setting using the code provided by them

Setting	Source	Method	Bike	Bus	Car	Motor	Person	Rider	Light	Sign	Train	Truck	mAP
Lower Bound	D	Source Only	35.1	51.7	52.6	9.9	31.9	17.8	21.6	36.3	-	47.1	30.4
	N		27.9	32.5	49.4	15.0	28.7	21.8	14.0	30.5	-	30.7	25.0
	D+N		31.5	46.9	52.9	8.4	29.5	21.6	21.7	34.3	-	42.2	28.9
UDA	D+N	Strong-Weak [10]	29.7	50.0	52.9	11.0	31.4	21.1	23.3	35.1	-	44.9	29.9
		Graph Prototype [2]	31.7	48.8	53.9	20.8	32.0	21.6	20.5	33.7	-	43.1	30.6
		Cat. Regularization [14]	25.3	51.3	52.1	17.0	33.4	18.9	20.7	34.8	-	47.9	30.2
		UMT [4]	42.3	48.1	56.4	13.5	35.3	26.9	31.1	41.7	-	40.1	33.5
		Adaptive Teacher [6]	43.1	48.9	56.9	14.7	36.0	27.1	32.7	43.8	-	42.7	34.6
MSDA	D+N	MDAN [17]	37.1	29.9	52.8	15.8	35.1	21.6	24.7	38.8	-	20.1	27.6
		M^3SDA [8]	36.9	25.9	51.9	15.1	35.7	20.5	24.7	38.1	-	15.9	26.5
		DMSN [15]	36.5	54.3	55.5	20.4	36.9	27.7	26.4	41.6	-	50.8	35.0
		TRKP [13]	48.4	56.3	61.4	22.5	41.5	27.0	41.1	47.9	-	51.9	39.8
		PMT [1]	55.3	59.8	67.6	29.9	47.6	32.7	46.3	56.0	-	57.7	45.3
		ACIA(Ours)	**56.1**	**61.0**	**69.2**	**31.9**	**51.8**	**39.8**	**49.2**	**59.0**	-	**61.0**	**47.9**
Upper Bound	D+N	Target Only	27.2	39.6	51.9	12.7	29.0	15.2	20.0	33.1	-	37.5	26.6
		All-Combined	56.4	59.9	67.3	30.8	47.9	33.9	47.2	57.8	-	54.8	45.3
		Fine-Tuning	63.3	68.1	72.5	39.3	52.2	37.2	54.1	63.1	-	59.1	50.9

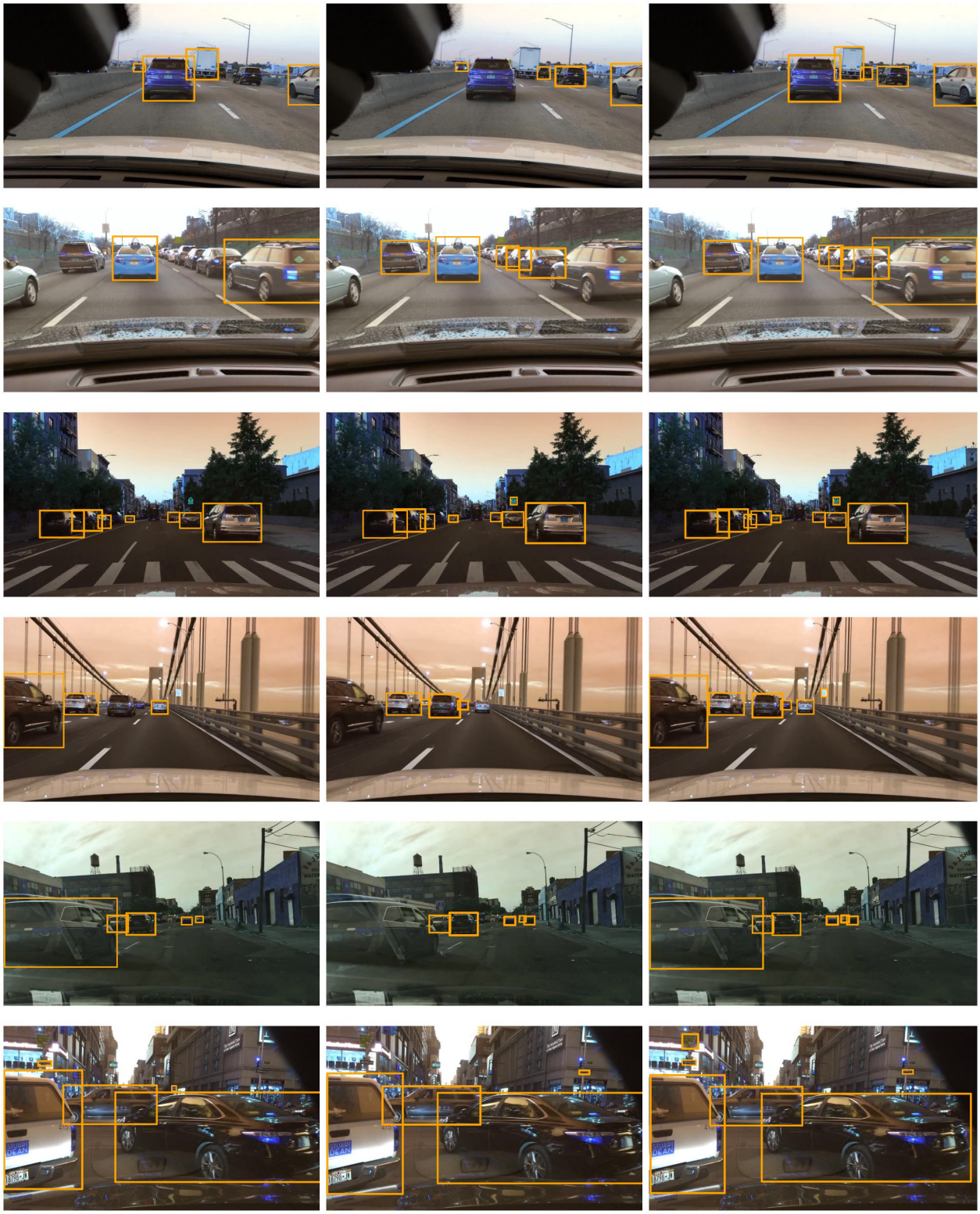
Table 7. Class-wise AP of ACIA compared against the baseline lower bound, UDA, MSDA, and upper bound methods in the cross-time settings. Source domains are daytime (D) and night (N) subsets and the target is always Dusk/Dawn of BDD100K.

Setting	Source	Method	Person	Car	Rider	Truck	Motor	Bicycle	Bike	mAP
Lower Bound	C	Source Only	26.9	44.7	22.1	17.4	17.1	18.8	16.7	23.4
Lower Bound	C+M	Source Only	35.2	49.5	26.1	25.8	18.9	26.1	26.5	29.7
UDA		UMT [4]	30.7	28.0	3.9	11.2	19.2	17.8	18.7	18.5
UDA		Adaptive Teacher [6]	31.2	31.7	15.1	16.4	17.1	20.9	27.9	22.9
MSDA		TRKP [13]	39.2	53.2	32.4	28.7	25.5	31.1	37.4	35.3
MSDA		PMT [1]	41.1	53.5	31.2	31.9	33.7	34.9	44.6	38.7
MSDA		**ACIA(ours)**	**43.3**	**58.1**	**33.3**	**35.1**	**33.7**	**38.6**	**45.2**	**41.0**
Upper Bound		All-Combined	40.2	60.1	47.1	60.0	29.2	36.3	56.9	47.1
Upper Bound		Fine-Tuning	44.1	61.4	49.0	61.1	30.8	39.2	58.8	49.2
Lower Bound	C+M+S	Source Only	36.6	49.0	22.8	24.9	26.9	28.4	27.7	30.9
UDA		UMT [4]	32.7	39.6	6.6	21.2	21.3	25.7	28.5	25.1
UDA		Adaptive Teacher [6]	36.3	42.6	19.7	23.4	24.8	27.1	33.2	29.6
MSDA		TRKP [13]	40.2	53.9	31.0	30.8	30.4	34.0	39.3	37.1
MSDA		PMT [1]	43.3	54.1	32.0	32.6	35.1	36.1	44.8	39.7
MSDA		**ACIA(ours)**	**44.9**	**59.2**	**33.8**	**33.5**	**38.3**	**39.9**	**46.5**	**42.3**
Upper Bound		All-Combined	41.7	63.9	49.5	58.1	31.6	39.1	53.5	48.2
Upper Bound		Fine-Tuning	49.2	63.5	56.1	62.6	35.1	43.7	57.2	52.5
Upper Bound	C+M+S	Target Only	35.3	53.9	33.2	46.3	25.6	29.3	46.7	38.6

Table 8. Class-wise AP of ACIA compared against the baselines in the mixed adaptation settings. Source domains are Cityscapes(C), MS COCO(M), and Synscapes(S) datasets while the Daytime domain of BDD100K is the target domain.

10. More Detection Visualization

In Fig. 3 we present more visualization of the detections on BDD100k cross-time for the three multi-source adaptation approaches presented in the paper: no class-wise adaptation (similar to [13, 15]), prototype-based class-conditional adaptation [1] and our attention-based class-conditional adaptation. From the visualizations, it can be observed that our method is performing better detection compared to the other approaches highlighting the impact of an efficient class-conditioned alignment.



(a) Methods w/o class alignment. (b) Prototype-based method. (c) Our ACIA method.

Figure 3. Comparison of instance-level adaptation detection on BDD100k cross-time setting. **(a)** MSDA without class-conditioned instance adaptation as in [13, 15]. **(b)** With the prototype-based class-conditional adaptation [1]. **(c)** ODs with our ACIA approach.

References

[1] Belal, A., Meethal, A., Romero, F.P., Pedersoli, M., Granger, E.: Multi-source domain adaptation for object detection with prototype-based mean teacher. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1277–1286 (January 2024) 1, 4, 5

[2] Cai, Q., Pan, Y., Ngo, C.W., Tian, X., Duan, L., Yao, T.: Exploring object relation in mean teacher for cross-domain detection (2019). https://doi.org/10.48550/ARXIV.1904.11245, https://arxiv.org/abs/1904.11245 4

[3] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding (2016) 1

[4] Deng, J., Li, W., Chen, Y., Duan, L.: Unbiased mean teacher for cross-domain object detection (2021) 4

[5] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361 (2012). https://doi.org/10.1109/CVPR.2012.6248074 1

[6] Li, Y.J., Dai, X., Ma, C.Y., Liu, Y.C., Chen, K., Wu, B., He, Z., Kitani, K., Vajda, P.: Cross-domain adaptive teacher for object detection (2021). https://doi.org/10.48550/ARXIV.2111.13216, https://arxiv.org/abs/2111.13216 3, 4

[7] Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015) 1

[8] Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation (2019) 4

[9] Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: CVPR (2019) 3

[10] Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 4

[11] Tan, J., Lu, X., Zhang, G., Yin, C., Li, Q.: Equalization loss v2: A new gradient balance approach for long-tailed object detection (2021), https://arxiv.org/abs/2012.08548 2

[12] Wrenninge, M., Unger, J.: Synscapes: A photorealistic synthetic dataset for street scene parsing (2018) 1

[13] Wu, J., Chen, J., He, M., Wang, Y., Li, B., Ma, B., Gan, W., Wu, W., Wang, Y., Huang, D.: Target-relevant knowledge preservation for multi-source domain adaptive object detection (2022). https://doi.org/10.48550/ARXIV.2204.07964, https://arxiv.org/abs/2204.07964 1, 4, 5

[14] Xu, C.D., Zhao, X.R., Jin, X., Wei, X.S.: Exploring categorical regularization for domain adaptive object detection (2020) 4

[15] Yao, X., Zhao, S., Xu, P., Yang, J.: Multi-source domain adaptation for object detection (2021). https://doi.org/10.48550/ARXIV.2106.15793, https://arxiv.org/abs/2106.15793 1, 4, 5

[16] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning (2020) 1

[17] Zhao, H., Zhang, S., Wu, G., Moura, J.M.F., Costeira, J.P., Gordon, G.J.: Adversarial multiple source domain adaptation. In: Advances in Neural Information Processing Systems (2018) 4