Black-Box Domain Adaptation for Object Detection with Retention-Driven Knowledge Compression

Supplementary Material

Supplement Contents

Due to space limitations, this supplementary material provides additional details not included in the main paper and presents more experimental results. The contents are organized as follows:

A introduces more recent works;

B provides the relationship between lifelong learning and RDKC;

C provides more dataset details;

D provides more implementation details;

E provides quantitative experiments on more datasets, resource consumption comparisons, and qualitative experiments on architectural versatility.

A. Supplement of Recent Works

Unsupervised domain adaptation (UDA) assumes that the model has simultaneous access to labeled data from the source domain and unlabeled data from the target domain. In object detection, UDA is often formulated as unsupervised domain adaptive object detection (UDAOD). Existing UDAOD works address this task through diverse strategies: DA-Faster [2] and HTCN [1] leverage instance-level and image-level representations to align feature distributions across source and target domains; ODSC [16] adopts image translation to generalize knowledge and bridge source-target gaps; ASTOD [19] employs self-training to generate instance-level pseudo labels for adaptation.

Source-free domain adaptation (SFDA) assumes that only unlabeled target data and a pre-trained source model are accessible during the adaptation stage. SFDA is often formulated as source-free domain adaptive object detection (SFOD). Existing SFOD methods primarily adhere to three technical paradigms: pseudo-label refinement, knowledge distillation, and domain alignment. For pseudo-label refinement, SED [11] serves as a representative example, which identifies a confidence threshold to filter pseudo labels via self-entropy descent. Within the knowledge distillation paradigm, LODS [13] achieves effective bidirectional knowledge transfer by first enhancing target domain styles and then disregarding them. For domain alignment, methods such as IRG [20] and LPLD [24] leverage contrastive loss to refine target representations, exploiting object relational cues to enhance alignment. Although these methods perform well in source-free settings, they face considerable challenges under black-box conditions. Since the downloaded predictions remain static throughout the adaptation process, relying solely on self-contrastive signals between the teacher and student models is insufficient for effective adaptation. While SFDA methods such as LPLD [24] attempt to filter and refine low-confidence predictions using the teacher model, the reliability of the teacher is undermined by pervasive noise in black-box environments. Notably, such noise is not limited to low-confidence predictions, it also frequently appears in high-confidence regions [23]. Without effectively suppressing this noisy knowledge, the performance of the student model can degrade significantly.

B. Supplement of the Relationship between Lifelong Learning and RDKC

Human memory in lifelong learning [10, 15] operates as a dynamic and adaptive system that sustains lifelong knowledge acquisition by continuously balancing three interdependent processes: (1) active forgetting, which identifies and eliminates outdated or low-value information through molecular and neural mechanisms to reallocate cognitive resources for new learning; (2) selective stabilization, which strengthens relevant neural pathways to preserve high-value knowledge and frequently used skills; and (3) collaborative integration among specialized modules, where distinct memory subsystems act as parallel learners whose outputs from systems of varying robustness are synergistically combined to enhance adaptive reasoning and generalization across contexts.

Recent studies [14, 21] have demonstrated that integrating lifelong learning into machine learning effectively addresses the limitations of conventional models in continual training scenarios. Traditional approaches often require storing large volumes of past data to mitigate forgetting in sequential tasks or rely on complex incremental training pipelines, both of which incur high storage costs and low training efficiency. To enhance practical applicability, [21] introduces a memory-optimized learning paradigm without dependence on whole previous predictions, where prior knowledge is retained through posterior parameter adjustment (active forgetting) and parameter importance quantification (stability preservation), enabling scene-specific learning via compact knowledge compression. Furthermore, [14] provides theoretical evidence of its effectiveness across multiple domains, including visual classification and reinforcement learning, thereby offering strong theoretical support for integrating lifelong learning into machine learning.

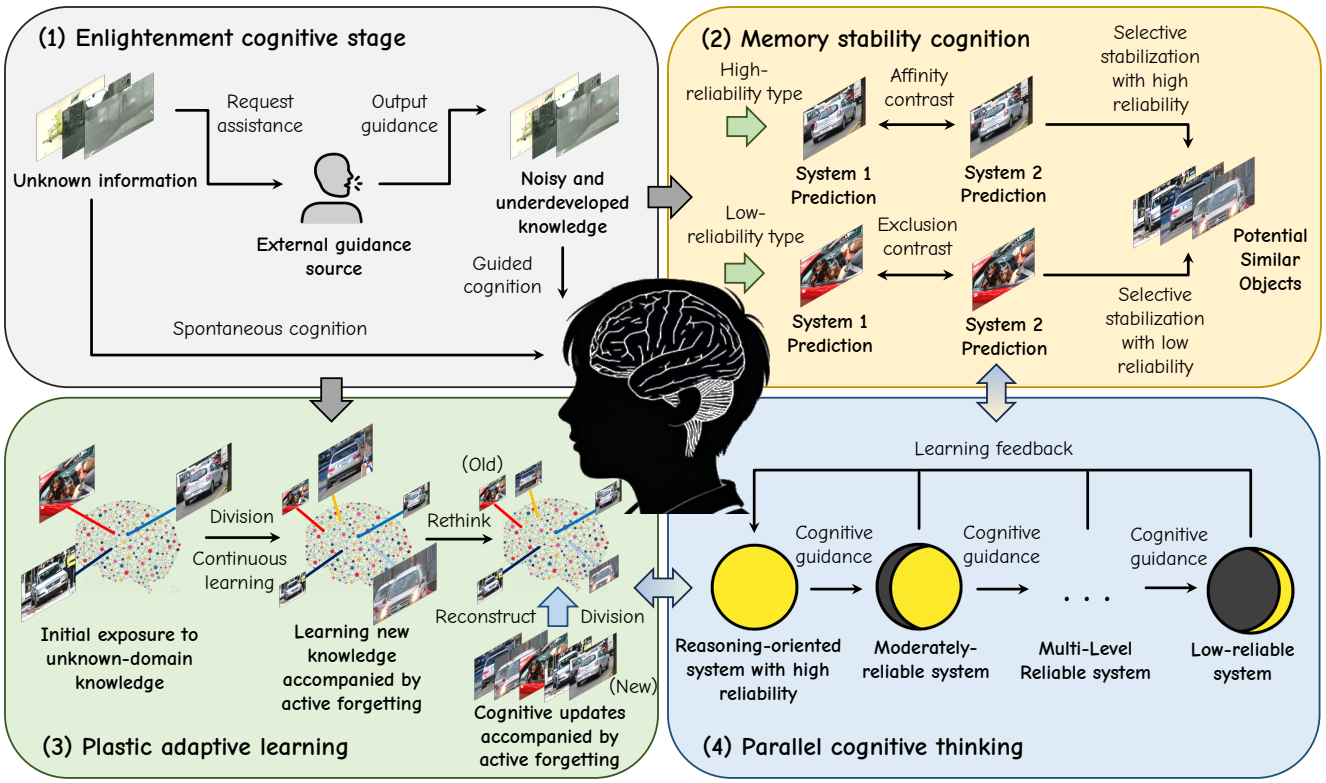

Figure 1. Conceptual and learning principle diagram. Biological learners initially rely on external guidance and self-reflection to acquire knowledge in unfamiliar domains, gradually transitioning to independent reasoning as familiarity increases. The brain cognitive system comprises multiple parallel subsystems of varying reliability, where rational cognition often guides emotional reasoning. Through adaptive learning, the brain continuously reshapes its understanding by selectively retaining reliable knowledge and actively forgetting outdated information, allowing new cognition to dominate while leveraging stable memories to guide future learning. Our proposed RDKC framework simulates the full cognitive process: Black-box initialization reflects the guided cognition, knowledge distillation reflects the spontaneous cognition, Memory Retention aligns with the plastic adaptive learning, and Scene Compression aligns with the memory stability cognition. Moreover, the teacher-student architecture embodies the parallel cognitive thinking.

In lifelong learning, during the enlightenment cognitive stage, biological learners exhibit immature capabilities when facing unfamiliar domains, they require external guidance combined with self-reflection to acquire new knowledge, *e.g.*, a child learns under adult supervision, while a fruit fly imitates the behavior of its peers for biological information exchange. Once the domain becomes familiar, independent reasoning and contextual adaptability dominate. Humans form distinct understandings of events through dynamic interactions between rational and emotional cognition across different cognitive stages. Therefore, as modeled in [14, 21], the brain's cognitive system is organized into multiple parallel cognitive thinking subsystems, where more reliable cognitive processes influence reasoning-oriented subsystems, *e.g.*, rational thinking is typically more dependable than purely emotional reasoning). Meanwhile, the cognitive system continuously acquires new knowledge but does not retain all of it. Through plastic adaptive learning, the brain performs cognitive comparisons, active forgetting outdated perceptions while selectively reshaping cognition with new insights. In other words, the latest understanding always dominates the learning process. Moreover, memory stability cognition guides new learning by leveraging prior knowledge, enabling selective stabilization through reliability judgments across subsystems.

Motivated by lifelong learning, RDKC simulates the underlying process and pioneers its use in black-box unsupervised visual tasks. Although external knowledge can guide self-learning, the internal mechanisms remain invisible. Therefore, we model Guided Cognition as the black-box initialization and Spontaneous Cognition as a knowledge distillation process, together forming the Enlightenment Cognitive Stage. The human cognitive system consists of multiple parallel thinking subsystems. Considering the computational cost, we simplify it into two subsystems: a high-robustness and a low-robustness subsystem, modeled respectively by a teacher–student framework. Our Memory

Algorithm 1 RDKC for black-box domain adaptive object detection.

Input: Unlabeled target domain D_t; teacher model $\mathcal{M}^{tea}$ with its parameters Θ_{tea}; student model $\mathcal{M}^{stu}$ with its parameters Θ_{stu}, training epoch e; distillation epoch e_d; hyperparameters λ and η; cloud API interface API; and strong augmentation $\mathcal{A}_s$ and weak augmentation $\mathcal{A}_w$.

Batch upload the data x_i from D_t and download the noisy hard predictions P_s from API. ▷ *Black-box initialization*

========== Simple Knowledge Distillation for Teacher Model Initialization ==========

for $epoch \leftarrow 1$ **to** e_d **do**

  $x_i^w \leftarrow \mathcal{A}_w(x_i)$, $\hat{y}_i \leftarrow$ Non-Maximum Suppression(P_s).

  Optimize and update Θ_{tea} by minimizing the task-specific loss $\mathcal{L}_{task}$ using x_i^w and $\hat{y}_i$. ▷ *Eqs. (1) and (4)*

end for

Freeze Θ_{tea}.

========== Black-box Training ==========

Initialize $\Theta_{stu} \leftarrow \Theta_{tea}$. ▷ *Student model initialization*

for $epoch \leftarrow 1$ **to** e **do**

  $x_i^w \leftarrow \mathcal{A}_w(x_i)$, $x_i^s \leftarrow \mathcal{A}_s(x_i)$, $Sco_{i,j}^{tea} \leftarrow \mathcal{M}^{tea}(x_i^w)$, $Sco_{i,j}^{stu} \leftarrow \mathcal{M}^{stu}(x_i^s)$, $\hat{y}_i \leftarrow$ Non-Maximum Suppression($Sco_{i,j}^{tea}$).

  With Memory Retention:

    Conduct a coarse-grained assessment of teacher model proposals by using IoU local comparison and λ. ▷ *Eq. (5)*

    Conduct a Score optimization for different types of $Sco_{i,j}^{tea}$. ▷ *Eqs. (6) and (7)*

    Minimize the memory optimization loss $\mathcal{L}_{MR}$ with $Sco_{i,j}^{stu}$ and the optimized $Sco_{i,j}^{tea}$. ▷ *Eq. (8)*

  With Scene Compression :

    Assign the scene compression weight $\mathcal{W}_{i,j}$ with $Sco_{i,j}^{stu}$ and the optimized $Sco_{i,j}^{tea}$. ▷ *Eq. (9)*

    Construct a combination of $\mathcal{W}_{i,j}$ and $\mathcal{L}_{MR}$. ▷ *Eq. (10)*

  Optimize and update Θ_{stu} by minimizing $\mathcal{W}_i\mathcal{L}_{MR}$ and $\mathcal{L}_{task}$ using x_i^w, η, and $\hat{y}_i$. ▷ *Eqs. (2) and (11)*

  Update Θ_{tea} by using Θ_{stu} with Exponential Moving Average. ▷ *Eq. (3)*

end for

Output: Student model $\mathcal{M}^{stu}$ with its parameters Θ_{stu}.

Table 1. Results of computational cost comparison on the *Cityscapes* → *Foggy-Cityscapes* using the ResNet-50 backbone, built upon the Mean-Teacher framework [17]. The batch size is 1.

Method	Space (MiB)	Time (s/epoch)	mAP
SEAL	19473MiB	514.6s	36.3
LPLD†	13841MiB	288.1s	32.6
RDKC	14695MiB	304.1s	40.7

Retention module categorizes memory types and processes new and old memories accordingly. Through contrastive learning, it selectively performs active forgetting–based reconstruction. Moreover, by combining memory-type partitioning with contrastive learning across subsystems, Scene Compression achieves selective stabilization of subsequent learning, primarily guided by high-value knowledge. As illustrated in Figure 1, we draw an analogy between the process of lifelong learning and RDKC, highlighting the complete learning principle.

C. Supplement of Dataset Details

Cityscapes [3] is an urban street-scene dataset with 5,000 finely annotated images, among which 2,925 are used for training and 500 for validation. *Foggy-Cityscapes* [7] is a synthetic variant of *Cityscapes* generated by simulating fog, available in three different visibility levels. *Sim10k* [9] is a synthetic street-view dataset comprising 10,000 images rendered from a video game, annotated with vehicle bounding boxes as a substitute for manually labeled real-world data. *KITTI* [6] contains 7,481 real-world driving images. While it shares similar urban scenes with *Cityscapes*, it differs in terms of camera configurations and environmental settings. *Pascal-VOC* [5] features real-world objects, such as birds, cats, and chairs-captured in diverse scenes. *Watercolor* [8] is artistic dataset composed of 2,000 watercolor paintings.

D. Supplement of Implementation Details

Our pseudocode for the training process is shown in Algorithm 1. Following [4, 20, 24], weak augmentation includes only image resizing, whereas strong augmentation incorporates additional operations such as color jittering, grayscale transformation, Gaussian blurring, and random erasing. All input images are resized such that the shorter side is set to 600 pixels, and the longer side does not exceed 1333 pixels, while preserving the original aspect ratio. The batch size is fixed at 1.

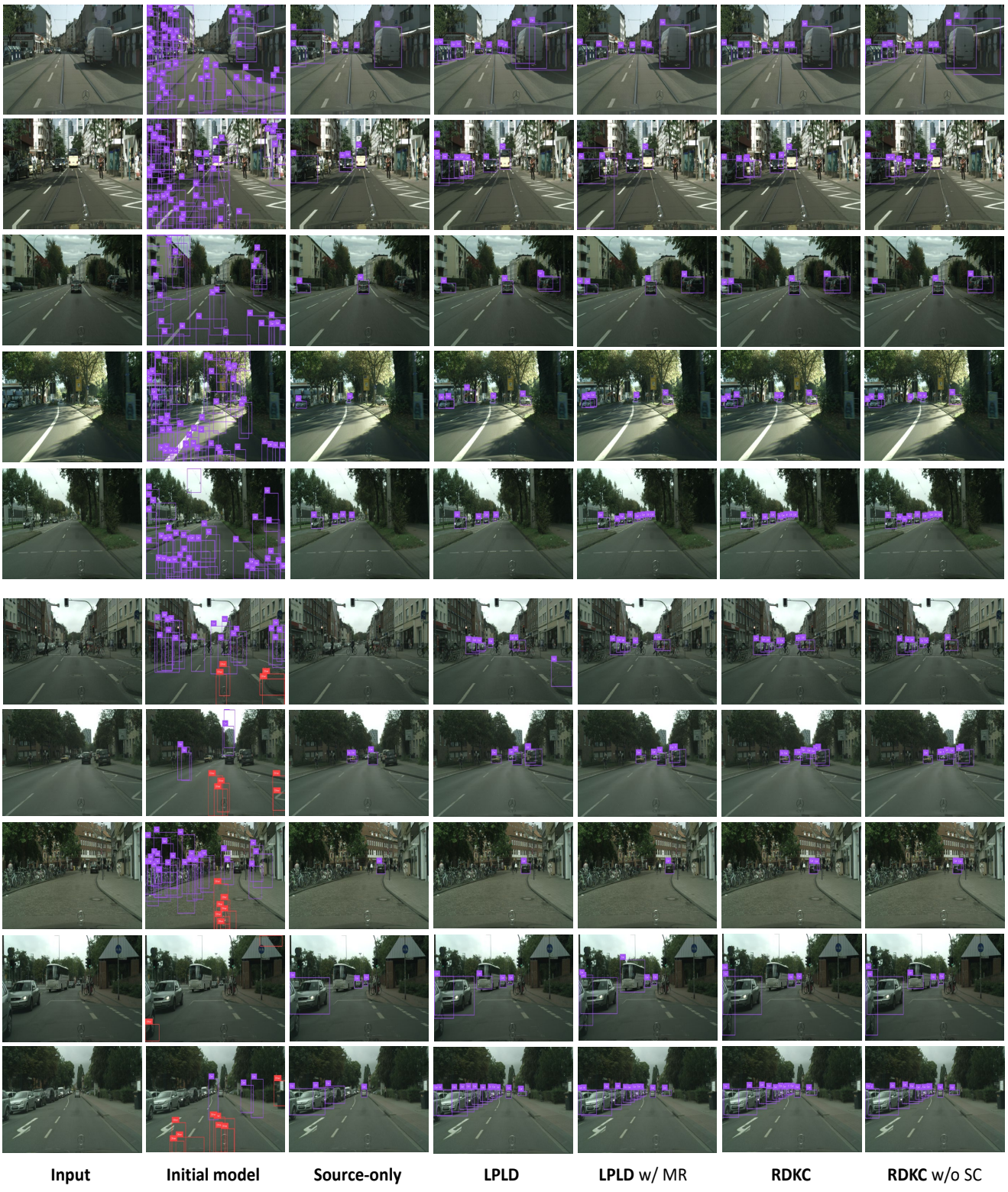

Figure 2. Qualitative results for the synthetic-to-real adaptation *Sim10K* → *Cityscapes* (Rows 1–5) and the cross-camera adaptation *KITTI* → *Cityscapes* (Rows 6–10). (Zooming in for a clearer view)

Table 2. Quantitative results (mAP) of different architecture models in weather adaptation scenario *Cityscapes* → *Foggy-Cityscapes* under the BBDA setting. † denotes methods originally designed for other settings but adapted to the BBDA scenario.

Baseline Detector	**Faster-RCNN**			**Deformable DETR**
Source Model → **Target Model**	**ResNet-50** → **ResNet-50**	**ResNet-50** → **VGG-16**	**ResNet-50** → **ResNet-101**	**ResNet-50** → **ResNet-50**
Source-only	25.2	25.2	25.2	28.5
DINE	35.2	34.5	35.8	36.5
BiMem	34.6	34.5	35.3	38.4
IRG†	33.1	32.4	33.7	-
LPLD†	32.6	33.2	34.1	-
DRU†	-	-	-	38.2
SEAL	36.3	36.4	37.6	37.4
RDKC	**40.7**	**40.4**	**41.8**	**42.0**

E. More Qualitative and Quantitative Results

Resource Consumption Comparisons. For a fair comparison, we set up to run 10 epochs in the compared works with a batch-size of 1, the number of num-workers is 8, and the same scenario *Cityscapes* → *Foggy-Cityscapes*. In addition, we run these methods on the same machine with a GeForce RTX4090 24G GPU.

As illustrated in Table 1, we recorded the maximum space usage during training and the average runtime per epoch throughout the entire training process. The previous SOTA BBDA method, SEAL [22], is computationally expensive, as it performs fine-grained discrimination across bounding boxes based on sample features and applies targeted similarity learning, leading to significant time overhead in box-level contrastive computation. Compared with previous methods, our RDKC only needs to continuously retains the learning data from the previous epoch in memory space, focusing on in-place adaptive adjustments in the current learning stage without simultaneously comparing all similar sample features. By flexibly utilizing the stored memory data, this learning strategy significantly reduces both memory consumption and training time.

Qualitative Results on Architectural Versatility. BBDA is inherently flexible, allowing for different target model architectures regardless of the source model, as discussed in the Introduction. To validate the architectural flexibility of BBDA, we conduct qualitative experiments to compare different model architectures in the weather adaptation scenario of *Cityscapes* → *Foggy-Cityscapes*, where Faster R-CNN [2] is adopted as the default detector. As shown in Table 2, compared to SFDA-based methods [20, 24] adapted to the BBDA setting, the native BBDA approach [12, 22, 25] demonstrates stronger architectural adaptability, achieving superior performance across diverse model architectures. Compared with previous methods, our RDKC achieves competitive performance across various architectural models, and notably attains the best mAP of 41.8% in the ResNet-50 → ResNet-101. Additionally, to validate the universality of our framework, we conducted supplementary experiments using Deformable DETR [26] as the detector and included comparative evaluations with the SFDA method DRU [18], which adopts Deformable DETR as its default detector. The results demonstrate that our method consistently achieves SOTA performance, attaining the best mAP of 42.0% under the same experimental conditions.

Quantitative Results on Different Datasets. As shown in Figure 2, we supplement additional quantitative analyses under the BBDA setting for scenarios *Sim10K* → *Cityscapes* and *KITTI* → *Cityscapes*. We observe that existing SFDA method LPLD [24], when directly applied to BBDA without integrating MR, suffers from a large number of hallucinated and missed detections. After incorporating MR, hallucinated detections are significantly reduced; however, excessive filtering of low-confidence predictions still leads to frequent misses. In contrast, our proposed RDKC effectively alleviates both hallucination and omission issues, enabling accurate perception of both near-view objects and fine-grained distant targets.

References

[1] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8869–8878, 2020. 1

[2] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3339–3348, 2018. 1, 5

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 3

[4] Jinhong Deng, Wen Li, and Lixin Duan. Balanced teacher for source-free object detection. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 34(8):7231–7243, 2024. 3

[5] Mark Everingham, S. M. Eslami, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136, 2015. 3

[6] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 3

[7] Martin Hahner, Dengxin Dai, Christos Sakaridis, Jan-Nico Zaech, and Luc Van Gool. Semantic understanding of foggy scenes with purely synthetic data. In *Proceedings of the IEEE Conference on Intelligent Transportation Systems Conference (ITSC)*, pages 3675–3681, 2019. 3

[8] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5001–5009, 2018. 3

[9] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *Proceedings of the IEEE Conference on International Conference on Robotics and Automation (ICRA)*, pages 746–753, 2017. 3

[10] Dhireesha Kudithipudi, Mario Aguilar-Simon, Jonathan Babb, Maxim Bazhenov, Douglas Blackiston, Josh Bongard, Andrew P Brna, Suraj Chakravarthi Raja, Nick Cheney, Jeff Clune, et al. Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence (Nat. Mach. Intell)*, 4(3):196–210, 2022. 1

[11] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 8474–8481, 2021. 1

[12] Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. Dine: Domain adaptation from single and multiple black-box predictors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7993–8003, 2022. 5

[13] Mingyang Liu, Xinyang Chen, Yang Shu, Xiucheng Li, Weili Guan, and Liqiang Nie. Boosting transferability and discriminability for time series domain adaptation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 1004029–100427, 2024. 1

[14] Wang Liyuan, Zhang Xingxing, Li Qian, Zhang Mingtian, Su Hang, Zhu Jun, and Zhong Yi. Incorporating neuro-inspired adaptability for continual learning in artificial intelligence. *Nature Machine Intelligence (Nat. Mach. Intell)*, 5: 1356–1368, 2023. 1, 2

[15] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks (NN)*, 113:54–71, 2019. 1

[16] Adrian Lopez Rodriguez and Krystian Mikolajczyk. Deep domain adaptive object detection: a survey. In *British Machine Vision Conference (BMVC)*, pages 1808–1813, 2020. 1

[17] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, page 1195–1204, 2017. 3

[18] Long Hoang Pham Duong Nguyen-Ngoc Tran Trinh Le Ba Khanh, Huy-Hung Nguyen and Jae Wook Jeon. Dynamic retraining-updating mean teacher for source-free object detection. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*. 5

[19] Renaud Vandeghen, Gilles Louppe, and Marc Van Droogenbroeck. Adaptive self-training for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 914–923, 2023. 1

[20] Vibashan VS, Poojan Oza, and Vishal M. Patel. Instance relation graph guided source-free domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3520–3530, 2023. 1, 3, 5

[21] Liyuan Wang, Xingxing Zhang, Kuo Yang, Longhui Yu, Chongxuan Li, HONG Lanqing, Shifeng Zhang, Zhenguo Li, Yi Zhong, and Jun Zhu. Memory replay with data compression for continual learning. In *International Conference on Learning Representations (ICLR)*, 2022. 1, 2

[22] Mingxuan Xia, Junbo Zhao, Gengyu Lyu, Zenan Huang, Tianlei Hu, Gang Chen, and Haobo Wang. A separation and alignment framework for black-box domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 16005–16013, 2024. 5

[23] Jianfei Yang, Xiangyu Peng, Kai Wang, Zheng Zhu, Jiashi Feng, Lihua Xie, and Yang You. Divide to adapt: Mitigating confirmation bias for domain adaptation of black-box predictors. In *International Conference on Learning Representations (ICLR)*, 2023. 1

[24] Ilhoon Yoon, Hyeongjun Kwon, Jin Kim, Junyoung Park, Hyunsung Jang, and Kwanghoon Sohn. Enhancing source-free domain adaptive object detection with low-confidence pseudo label distillation. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pages 337–353, 2024. 1, 3, 5

[25] Jingyi Zhang, Jiaxing Huang, Xueying Jiang, and Shijian Lu. Black-box unsupervised domain adaptation with bi-directional atkinson-shiffrin memory. In *Proceedings of*

the IEEE/CVF International Conference on Computer Vision (ICCV), pages 11771–11782, 2023. 5

[26] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021. 5