Towards Online Domain Adaptive Object Detection

Vibashan VS, Poojan Oza, and Vishal M. Patel
Johns Hopkins University, Baltimore, MD, USA

{vvishnu2,poza2,vpatel36}@jhu.edu

Abstract

Existing object detection models assume both the training and test data are sampled from the same source domain. This assumption does not hold true when these detectors are deployed in real-world applications, where they encounter new visual domains. Unsupervised Domain Adaptation (UDA) methods are generally employed to mitigate the adverse effects caused by domain shift. Existing UDA methods operate in an offline manner where the model is first adapted toward the target domain and then deployed in real-world applications. However, this offline adaptation strategy is not suitable for real-world applications as the model frequently encounters new domain shifts. Hence, it is critical to develop a feasible UDA method that generalizes to the new domain shifts encountered during deployment time in a continuous online manner. To this end, we propose a novel unified adaptation framework that adapts and improves generalization on the target domain in both offline and online settings. Specifically, we introduce MemXformer - a cross-attention transformer-based memory module where items in the memory take advantage of domain shifts and record prototypical patterns of the target distribution. Further, MemXformer produces strong positive and negative pairs to guide a novel contrastive loss, which enhances target-specific representation learning. Experiments on diverse detection benchmarks show that the proposed strategy producs state-of-the-art performance in both offline and online settings. To the best of our knowledge, this is the first work to address online and offline adaptation settings for object detection. Source code: https://github.com/Vibashan/memXformer-online-da

1. Introduction

The ability to train deep network models on large-scale annotated datasets [35, 14, 41, 29] has accelerated the progress for multiple computer vision tasks such as classification [35, 19, 13], segmentation [45, 72, 54], and detection [48, 47, 42]. Despite this success, these models have limited generalization capabilities [56, 23, 17]. Specifically, the model performance drops when the test data (target domain) is sampled from a different distribution than
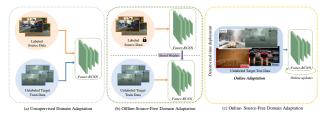

Figure 1. **Left:** Unsupervised Domain Adaptation - labeled source data and unlabeled target data are available during adaptation. **Middle:** Source-Free Domain Adaptation - source-trained model is adapted to the target domain. **Right:** Online Source-Free Domain Adaptation - source-trained model is adapted to target distribution shift during real-world deployment via online updates.

that of the training data (source domain) [1]. For example, when a model is deployed in real-world applications such as autonomous navigation, it could encounter images with weather-based degradations, camera artifacts, etc., unknown during training.

Unsupervised Domain Adaptation (UDA) methods [15, 58, 51, 9, 25, 24, 8, 28, 50] are generally employed to improve model generalization under domain shift condition. Existing UDA methods assume that both labelled source data and unlabeled target data are available during adaptation. This scenario is often not feasible in current real-world applications, as the labelled source data is often restricted due to privacy regulations, data transmission constraints, or proprietary data concerns. To overcome this drawback, recently, some works have explored Source-Free Domain Adaptation (SFDA) [37, 34, 68, 40, 39] setting, where a source-trained model is adapted towards the target domain without requiring access to the source data. However, in both UDA and SFDA settings, adaptation is performed in an offline manner where the model is first adapted towards the target domain and then deployed in real-world applications. In addition, it is often impossible to have prior knowledge about the target domain in most real-world applications. In other words, the deployed model could encounter a diverse set of target domains and offline adaptation to every distribution shift would be infeasible. Therefore, we propose a unified adaptation framework which utilizes a source-trained detector and adapts to the target domain in both offline and online manner.

In recent years, few works have explored various test-time adaptation settings where adaptation is performed during test-time [66, 70, 59]. Wang [66] proposed a fully test-time adaptation strategy which performs entropy minimization during test-time and only updates the model batch-norm parameters for the classification task. However, extending TENT to detection framework [66] has two critical drawbacks: 1) TENT use a very large batch size during test-time adaptation, which is not feasible during real-time deployment as images arrive one by one sequentially. 2) Updating only the batch norm parameter of a network with batch size 1 essentially degrades the model performance [70]. Although existing test-time adaptation settings are closer to online-SFDA settings, they are not suitable for adapting a detection model during real-world deployment. To overcome these issues, we explore an Online Source-Free Domain Adaptation (Online-SFDA)setting, where a model is adapted to any distribution shifts encountered during deployment in an online manner with batch size 1. Fig. 1 illustrate the online source-free domain adaptation setting for detection and its differences against the other adaptation settings.

Source-free domain adaptive object detection is relatively new and a more challenging setting than UDA. Existing SFDA methods [39, 27] for detection adapt to the target domain by training on the pseudo-labels generated by the source-trained model. Due to domain shifts, these generated pseudo-labels are noisy and training a model on top of them would lead to noise overfitting [44, 12]. To alleviate these issues, we employ a mean-teacher framework where the student model is supervised using pseudo-labels generated by the teacher network and the teacher network is slowly updated via the exponential moving average (EMA) of student weights. Therefore, the student network is trained on consistent pseudo-labels leading to less overfitting and the teacher network is a gradual ensemble of target adapted student weights [44]. However, this strategy is inefficient in learning two critical aspects required for optimal online adaptation: 1) They fail to learn robust target feature representation, 2) They fail to fully exploit the online target samples. Hence, we propose a novel memory module and a contrastive loss to fully utilize online target samples and learn robust target feature representation.

Contrastive Learning (CL) [5, 6, 18, 10, 31] aims to learn high-quality features from unlabeled data by forcing similar object instances to stay close and push dissimilar ones apart in an unsupervised manner. This is especially useful for online-SFDA as source-labelled data are unavailable during adaptation. Existing CL methods are designed for classification tasks where they operate on image-level features and require multiple image views (or augmentations) [5] to learn robust feature representation. Consequently, obtaining these large sets of views through input augmentations

is computationally expensive for adapting detector models. However in detector models, it is possible to obtain different views for an object in an input image without heavy input augmentations. More precisely, the detector provides multiple object proposals generated by Region Proposal Network (RPN), which in turn provides multiple cropped views around the object instance at different locations and various scales. Therefore, applying CL loss on RPN cropped views guides the model to learn object-level feature representation on the target domain. Note this CL loss is used to supervise the student network, where the object-level features are obtained from the student RoI features. However to perform contrastive learning, these student RoI features require positive and negative pairs. To this end, we propose *MemXfromer*, a cross-attention transformer-based memory module where items in the memory record prototypical patterns of the continuous target distribution. The proposed MemXformer solves two important problems for online adaptation: 1) store the target distribution during online adaptation, which are utilized for future adaptation. 2) stored temporal ensemble of target representations provides positive and negative pairs to guide the contrastive learning process. Further, we introduce a cross-attention based read and write technique which models better target distribution and provides strong positive and negative pairs for contrastive learning. Note that the proposed method is not only suitable for online adaptation but also for offline adaptation. In a nutshell, this paper makes the following contributions:

- To the best of our knowledge, this is the first work to consider both online and offline adaptation settings for detector models.
- We propose a novel unified adaptation framework which makes the detector models robust against online target distribution shifts.
- We introduce the MemXformer module, which stores prototypical patterns of the target distribution and provides contrastive pairs to boost contrastive learning on the target domain.
- We consider multiple detection benchmarks for experimental analysis and show that the proposed method outperforms existing UDA, and SFDA methods for both online and offline settings.

2. Related works

Unsupervised domain adaptation. Existing unsupervised domain adaptation methods can be categorized into three groups based on adversarial training [7, 50, 53, 63], self-training [30, 67] and image-to-image translation [33, 49]. The first domain adaptive object detection was studied in [7], where they followed an adversarial-based strategy to perform feature alignment at both image-level and instance-level to mitigate the domain shift. Later, Saito [50] pro-

posed an adversarial-based strategy where strong alignment of the local features and weak alignment of the global features. Kim [33], introduced an image-to-image translation-based where multiple target domain images are created by stylizing the labelled source images. Multiple discriminators are used to performing adversarial alignment to reduce domain discrepancy by utilizing these target-styled source images. In [30], a pseudo-label based training strategy was formulated to counter noise in pseudo-labels to perform robust training of object detectors on the target domain. However, all these works assume to have access to labelled source data and unlabeled target data during adaptation, and they operate in an offline setting.

Source-free domain adaptation. In the source-free domain adaptation setting, we have a source-trained model which adapts to the target domain without having access to source data. Multiple works have addressed the source-free domain adaptation (SFDA) setting for classification [40, 38], segmentation [43, 36, 64] and object detection [39, 27, 61, 21, 22] tasks. In detail, for classification task [39] proposed a self-supervised method to learn target domain representation via information maximization. Further for segmentation [43, 36] and object detection [39, 27], the proposed methods are based on pseudo-label self-training to learn target-specific representation. However, similar to existing UDA works, these SFDA methods operate in an offline setting. Thus, we explore online adaptation, which is a more practical way to tackle domain shifts for real-world applications.

Online adaptation. Sun [55] proposed a Test-time training (TTT) strategy, where a model is trained on source data along with an auxiliary task (eg: rotation prediction) which is utilized during test-time to fine-tune the model on target test distribution. The major drawback of this adaptation strategy is training an auxiliary task along with source training just to perform adaptation during test-time is not a feasible solution and effective solution for real-world application. Later, Wang [66] proposed a fully test-time adaptation setting, where the given source trained model adapts to the target domain by entropy minimization during test-time in an online manner by entropy minimization. In this way, Tent [66] adapts to the target domain with test-time loss. Here, the major limitation of [66] is a requirement of a large batch size during test-time adaptation, which is not feasible during real-time deployment as images arrive one by one sequentially. Although existing test-time adaptation settings provide close resemblance to online-SFDA settings, these test-time settings are not suitable for adapting a detection model during real-world deployment. Therefore in this work, we explore both online and offline adaptation settings for the object detection tasks.

Contrastive representation learning. Contrastive representation learning has shown huge progress towards unsu-pervised feature learning. The standard way of formulating contrastive learning for an anchor is by pulling together the feature embedding of anchor's positive pairs and pushing apart from the anchor's negative pair [46, 5, 18]. These positive and negative pairs are formed by augmenting the anchor image and sampling from the input batch of images. Thus, for a given anchor, the positive pair are augmented anchor images and the negative pairs are other images from the batch [46, 5, 18]. On top of this, by exploiting the task-specific label information, [31] performed contrastive learning in a supervised manner. Nonetheless, all these tasks require a large batch size to perform contrastive learning effectively and it is not feasible to have more than one image during online adaptation. Thus, we propose a memory-based contrastive learning framework suitable for adapting object detectors during deployment in an online manner.

3. Proposed method

The online-SFDA setting considers a source-trained model with parameters Θ_{src} and adapts to any target distribution shifts during real-world deployment as illustrated in Fig. 1. Let us consider a stream of online target data denoted as $\mathcal{T} = \{x_1, x_2, .., x_n\}$, where x_n is the nth online sample. Since these samples arrive sequentially, the model gets adapted to each sample and the adapted weights are used for future online samples. Specifically, the model parameters during adaptation on the nth sample x_n, i.e. $\Theta_{src}^{(n)}$, are initialized with the model parameters updated through online adaptation of previous x_{n-1}^{th} sample. To summarize, online-SFDA performs continuous online adaptation, i.e., adaptation will be continued as long as there is a stream of data and necessity.

Student-teacher training. In online-SFDA, the model parameters need to be continuously updated in an online unsupervised manner. Consequently, the model risks forgetting the original hypothesis learned through supervised source training [44, 12]. To overcome this, prior works [57, 44] have employed a student-teacher framework. Specifically, the student parameters (Θ_{std}) are adapted to the target domain by minimizing the detection loss supervised through the teacher-generated pseudo-labels. The adapted student parameters are then transferred to the teacher parameters (Θ_{tch}) via Exponential Moving Average (EMA). This can be formally written as:

$$\mathcal{L}_{pl}(x_n) = \mathcal{L}_{rpn}(x_n, \tilde{y}_n) + \mathcal{L}_{rcnn}(x_n, \tilde{y}_n) \quad (1)$$

$$\Theta_{std}^{(n+1)} \leftarrow \Theta_{std}^{(n)} + \gamma \frac{\partial(\mathcal{L}_{pl}(x_n))}{\partial \Theta_{std}^{(n)}} \quad (2)$$

$$\Theta_{tch}^{(n+1)} \leftarrow \alpha \Theta_{tch}^{(n)} + (1 - \alpha)\Theta_{std}^{(n+1)}, \quad (3)$$

where x_n and $\tilde{y}_n$ are the nth test sample and corresponding pseudo-label generated by teacher network, $\mathcal{L}_{pl}$ is the pseudo-label supervision loss, γ is the student learning rate, and α is teacher EMA rate. However, the student-teacher framework is still not sufficient to learn robust features to
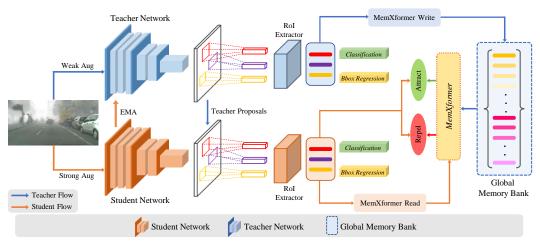
Figure 2. Overview of the proposed online-SFDA training pipeline. The detection network is adapted to online target distribution shifts by improving target representations through contrastive training. Specifically, the proposed MemXformer records prototypical patterns of the target distribution shift and provides strong positive and negative pairs to guide the contrastive learning process. distribution .

mitigate target distribution shifts. Hence, we explore contrastive learning-based strategies further to improve the robustness of feature representations in an online setting.

Contrastive Learning (CL). SimCLR [5] is a commonly used CL framework which learns representations for an image by maximizing agreement between differently augmented views of the same sample. For given an anchor image x_i, the SimCLR loss can be written as:

$$\mathcal{L}_{\text{SimCLR}}(x_i) = -\log \left(\frac{\exp(\text{sim}(z_i, z_j))}{\sum_{l=1, \ni l \neq i}^{2N} \exp(\text{sim}(z_i, z_l))} \right), \tag{4}$$

where N is the batch size, z_i and z_j are the features of two different augmentations of the same sample x_i, whereas z_l represents the feature of the lth batch sample x_l, where $l \neq i$. Also, $\text{sim}(\cdot, \cdot)$ indicates a similarity function, e.g. cosine similarity. Note that in general, the CRL framework assumes that each image contains one category/object [5]. Moreover, it requires large batch sizes that could provide multiple positive/negative pairs for training [6]. In contrast for object detection, each image will have multiple objects and a large batch size or multiple views are computationally not feasible. Hence, existing CRL methods are more suited for classification tasks.

3.1. Memory-based contrastive learning

Though existing contrastive learning methods like Sim-CLR are exceptional at learning high-quality representations, they are more suitable for the classification task. For detection, these CL methods require large batch size and heavy input augmentation, which are computationally expensive to apply for online parameter updates (discussed in Sec. 1). Therefore, we utilize a computationally efficient memory-based approach to make contrastive learning feasible and effective for online model updates. The proposed online-SFDA strategy is illustrated in Fig. 2.

MemXformer. A cross-attention transformer-based memory module which stores target distribution shift and guides the contrastive learning for target domain representation during online adaptation. Specifically, we employ a Global Memory Bank $M = \{m^i \in \mathbb{R}^{1 \times C}\}_{i=1}^{N_l}$, where N_l is number of memory items and C is memory item feature dimension. These memory items are used to store target representation and record prototypical patterns of the target distribution during the adaptation. In addition, these memory items are used to retrieve strong positive and negative pairs for guiding contrastive learning. The MemXformer module has two operations: write and read, which are based on cross-attention. In the MemXformer write operation, the teacher RoI features are used to update the memory elements appropriately. In the MemXformer read operation, the student RoI features are queried to the memory and a weighted sum of similar memory elements is retrieved, which essentially provide strong positive pairs. The read and write operations of MemXformer are illustrated in Fig. 3.

Write. To update the memory elements, we consider only the teacher network RoI features $\mathcal{F}_t = \{f_t^i \in \mathbb{R}^{1 \times C}\}_{i=1}^{N_f}$, where N_f is number of RoI features and C is RoI feature dimension. The teacher RoI features are considered because in the student-teacher framework, the teacher pipeline has input with weak augmentations resulting in accurate RPN proposals compared to the student pipeline. As shown in Fig. 3 (a), first the teacher RoI features are projected as *key* $K_t = \{k_t^i\}_{i=1}^{N_f}$ and *value* $V_t = \{v_t^i\}_{i=1}^{N_f}$ using two FC layer with weight W_k and W_v, respectively. Now each memory items are considered as *query* $Q_m = \{m^j\}_{j=1}^{N_l}$ and we compute a cross-attention map S_t between the teacher RoI features and memory items as follows:

$$\begin{aligned} k_t^i &= W_k \cdot f_t^i, \\ v_t^i &= W_v \cdot f_t^i, \end{aligned} \tag{5}$$

$$s_t^{(i,j)} = \frac{\exp\left(m^j \left(k_t^i\right)^T\right)}{\sum_{l \in M} \exp\left(m^l \left(k_t^i\right)^T\right)}, \qquad (6)$$

where the cross-attention map S_t is a 2D matrix of size $N_m \times N_f$ and $s_t^{i,j}$ represents how j^{th} memory items is related to i^{th} teacher RoI features. We utilize this cross-attention map S_t and V_t to update j^{th} memory item using following equation:

$$m^j \leftarrow F\left(m^j + \sum_{i \in V} s_t^{(i,j)} v_t^i\right). \qquad (7)$$

where $F(.)$ is L_2 norm. Therefore, using attention-based weighted average and global memory bank update for each online sample makes the MemXformer effectively store and model the target distribution.

Read. To read the memory elements, we consider only the student network RoI features $\mathcal{F}_s = \{f_s^i \in \mathbb{R}^{1 \times C}\}_{i=1}^{N_f}$, where N_f is number of RoI features and C is RoI feature dimension. In addition, the MemXformer Read operation is performed to obtain strong positive pairs given student RoI features as a query. As shown in Fig. 3 (b), first the student RoI features are projected as *query* $Q_s = \{q_s^i\}_{i=1}^{N_f}$ by one FC layer with weight W_q. Now each memory items are considered as *key* $K_m = \{m^j\}_{j=1}^{N_l}$ and we compute a cross-attention map S_s between the student RoI features and memory items as follows:

$$q_s^i = W_k \cdot f_s^i, \qquad (8)$$

$$s_s^{(i,j)} = \frac{\exp\left(q_s^i \left(m^j\right)^T\right)}{\sum_{l \in M} \exp\left(q_s^i \left(m^l\right)^T\right)}, \qquad (9)$$

where the cross-attention map S_s is a 2D matrix of size $N_m \times N_f$ and given i^{th} student RoI features as query, the s_t^i th row presents N_l memory items attention score. Therefore, given i^{th} student RoI features as query, we generate its corresponding positive pair by attention guided weighted sum of most similar memory items. Thus, utilizing the cross-attention map S_s and considering memory items as *value* $V_m = \{m^j\}_{j=1}^{N_l}$, we compute the strong positive pair for i^{th} student RoI features using following equation:

$$p_s^i = \sum_{j \in M} s_s^{(i,j)} m^j. \qquad (10)$$

where $\mathcal{P}_s = \{p_s^i\}_{i=1}^{N_f}$ corresponds to set of strong positive pair for student RoI features $\mathcal{F}_s$. In detail, the retrieved positive pairs are temporal ensembles of the prototypical target distribution, which gives more information regarding the online target distribution shifts. This essentially guides contrastive learning to model the target distribution.

Negative Pair Mining. As explained earlier from MemXformer read operation, we obtain a set of strong positive pairs for a given student RoI feature. These strong positive pairs are essentially an ensemble of most similar memory items. However, these ensembled similar memory items
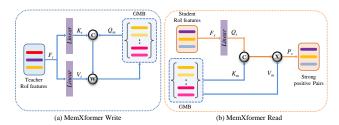

Figure 3. MemXformer Write and Read operations.

also contain dissimilar memory items but are scaled with less attention weights. This restricts the contrastive learning capability to effectively model the target domain representation. To mitigate the dissimilar item's effect on CL, we propose negative pair mining. Specifically in negative pair mining, given a student RoI feature as query and cross-attention map S_s, we mine the least similar 10% of the memory items and label them as negative pairs $\mathcal{N}_s = \{m_i^n\}_{i=1}^{N_s}$. As a result, by performing negative pair mining, we obtain N_s negative samples for one positive sample, where N_s is top 10% of least similar memory items.

Memory contrastive loss. Given student RoI feature f_s^i as anchor, utilizing MemXformer Read operation and negative pair mining we obtain strong positive $\mathcal{P}_s$ and negative pairs $\mathcal{N}_s$ from MemXformer. Therefore, given an image x_n with student RoI feature $\mathcal{F}_s$, the MemCLR loss is calculated as:
$$\mathcal{L}_{\text{MemCLR}}(x_n) =$$
$$- \log \left\{ \frac{1}{|\mathcal{F}_s|} \sum_{i \in \mathcal{F}_s} \frac{\exp(f_s^i \cdot p_s^i)}{\exp(f_s^i \cdot p_s^i) + \sum_{n \in \mathcal{N}_s} \exp(f_s^i \cdot m^n)} \right\},$$
Therefore, minimizing the MemCLR loss guided by strong positive and negative pairs enhance the student model to learn better target representation in an online-SFDA setting.

Overall loss. We illustrate our overall architecture for online source-free domain adaptation in Fig. 2. The proposed method utilizes a global memory bank to perform memory-based contrastive learning to robustify the representations under varying target distribution shifts. Therefore, the overall online-SFDA loss for any online sample x_n can be calculated as:

$$\mathcal{L}_{FTTA}(x_n) = \mathcal{L}_{pl}^{st}(x_n) + \mathcal{L}_{\text{MemCLR}}(x_n).$$

4. Experiments and Results

To validate the proposed method, we consider four domain shift scenarios where the source train model is adapted to the unlabelled target domain, typically used for comparison in UDA and SFDA literature. Specifically, we evaluate the proposed method with the existing UDA, SFDA and Test-time works under four domain shifts, 1) clear-weather to foggy-weather, 2) real to artistic, 3) synthetic to real, and 4) cross-camera adaptation. Note that, to show the effectiveness of our proposed approach, we evaluate both online and offline settings. Specifically, the offline setting follows the standard SFDA setting. The source-trained model is adapted towards the target domain using an unlabelled target train-set for multiple iterations and evaluated on the target test-set. Whereas in the online setting, the model

Table 1. Quantitative results (mAP) for Cityscapes → FoggyCityscapes. S: Source-only, O: Oracle, UDA: Unsupervised Domain Adaptation, SFDA: Source-Free Domain Adaptation, O-SFDA: Online Source-Free Domain Adaptation.

Type	Method	Offline	Online	prsn	rider	car	truck	bus	train	mcycle	bicycle	mAP
S	Source-Only	✓	✗	29.3	34.1	35.8	15.4	26.0	9.09	22.4	29.7	25.2
	DA Faster [7] (CVPR 2018)	✓	✗	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
	Selective DA [73] (CVPR 2019)	✓	✗	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
	D&Match [33] (CVPR 2019)	✓	✗	30.8	40.5	44.3	27.2	38.4	34.5	28.4	32.2	34.6
UDA	MAF [20] (ICCV 2019)	✓	✗	28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
	Robust DA [30] (ICCV 2019)	✓	✗	35.1	42.1	49.1	30.0	45.2	26.9	26.8	36.0	36.4
	MTOR [2] (CVPR 2019)	✓	✗	30.6	41.4	44.0	21.9	38.6	40.6	28.3	35.6	35.1
	Strong-Weak [50] (CVPR 2019)	✓	✗	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
	Categorical DA [69] (CVPR 2020)	✓	✗	32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6	37.4
	MeGA CDA [26] (CVPR 2021)	✓	✗	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	41.8
	Unbiased DA [12] (CVPR 2021)	✓	✗	33.8	47.3	49.8	30.0	48.2	42.1	33.0	37.3	40.4
	SFOD [39] (AAAI 2021)	✓	✗	25.5	44.5	40.7	**33.2**	22.2	28.4	34.1	39.0	33.5
SFDA	HCL [27] (NeurIPS 2021)	✓	✗	26.9	**46.0**	41.3	33.0	25.0	28.1	**35.9**	40.7	34.6
	Mean-Teacher [57]	✓	✗	33.9	43.0	45.0	29.2	37.2	25.1	25.6	38.2	34.3
	MemCLR (Ours)	✓	✗	**37.7**	42.8	**52.4**	24.5	**40.6**	**31.7**	29.4	**42.2**	**37.7**
O-SFDA	Tent [65] (ICLR 2021)	✗	✓	31.2	38.6	37.1	20.2	23.4	10.1	21.7	33.4	26.8
	MemCLR (Ours)	✗	✓	**32.1**	**41.4**	**43.5**	**21.4**	**33.1**	**11.5**	**25.5**	**32.9**	**29.8**
O	Oracle	✓	✗	38.7	46.9	56.7	35.5	49.4	44.7	35.9	38.8	43.1

is adapted towards the target domain in an online manner where the target test samples are seen only once and finally evaluated on the target test-set. This essentially simulates the real-world scenario where you see the target samples only once and adaptation needs to be continuous.

4.1. Implementation details

For the Online adaptation setting, we adopt Faster-RCNN [48] with ResNet50 [19] as the backbone pre-trained on ImageNet [35]. In all of our experiments, the input images are resized with a shorter side to be 600 pixels while maintaining the aspect ratio. We set the batch size to 1 for all experiments. For the student-teacher framework, the weight momentum update parameter α of the EMA for the teacher model is set equal to 0.99. The pseudo-labels generated by the teacher network with confidence greater than the threshold T=0.9 are selected for student training. We utilize an SGD optimizer to train the student network with a learning rate of 0.001 and momentum of 0.9 for both online and offline training. The Global Memory Bank contains N_m memory items, which are set to 1024. Further, the source model is trained using an SGD optimizer with a learning rate of 0.001 and momentum of 0.9 for 10 epochs. We report the mean Average Precision (mAP) with an IoU threshold of 0.5 for the teacher network on the distribution-shifted target domain test data during the evaluation.

4.1.1 Clear-weather to foggy-weather adaptation

When the source-trained models are deployed in real-world applications such as autonomous navigation, they are likely to encounter data from multiple weather conditions such as fog, haze, etc. In most cases, the deployed detector models would be trained for clear weather conditions. We propose to formulate this as an online adaptation problem, as it is difficult to pre-determine what kind of weather conditions will occur. Subsequently, we update the detector model in an online manner to adapt to any weather shifts the model might observe after deployment. To evaluate the

Table 2. Quantitative results for Sim10K → Cityscapes and KITTI → Cityscapes.

Type	Method	Online	Offline	Sim10k → City	Kitti → City
				AP of Car	AP of Car
S	Source-Only	✓	✗	32.0	33.9
	DA Faster [7] (CVPR 2018)	✓	✗	38.9	38.5
	MAF [20] (ICCV 2019)	✓	✗	41.1	41.0
UDA	Robust DA [30] (ICCV 2019)	✓	✗	42.5	42.9
	Strong-Weak [50] (CVPR 2019)	✓	✗	40.1	37.9
	Harmonizing [3] (CVPR 2020)	✓	✗	42.5	-
	Cycle DA [71] (ECCV 2020)	✓	✗	41.5	41.7
	MeGA CDA [62] (CVPR 2021)	✓	✗	44.8	43.0
	Unbiased DA [12] (CVPR 2021)	✓	✗	43.1	-
	SFOD [39] (AAAI 2021)	✓	✗	42.3	43.6
SFDA	Mean-teacher [57]	✓	✗	42.3	43.6
	MemCLR (Ours)	✓	✗	**44.2**	**46.8**
O-SFDA	Tent [65] (ICLR 2021)	✗	✓	32.8	34.5
	MemCLR (Ours)	✗	✓	**37.2**	**38.5**

proposed method under such conditions, we experiment on Cityscapes [11] → FoggyCityscapes [52] dataset. Here, we have a detection model trained on the Cityscapes dataset consisting of 2,975 normal weather images and 500 test images with 8 object categories: *person, rider, car, truck, bus, train, motorcycle and bicycle*. During inference, images from FoggyCityscapes are sequentially sent and the object detection model is adapted in an online manner to improve generalization on foggy/hazy weather. Table 1 provides the comparison of the proposed FTTA method with the state-of-the-art UDA, SFDA, and O-SFDA methods for Cityscape→FoggyCityscapes adaptation scenario. From Table 1, we can infer that UDA and SFDA methods operate in an offline manner, where as O-SFDA operates in an online manner. Firstly, in the online setting our proposed method outperforms existing UDA methods such as SWDA [50], MTOR [2] and InstanceDA [67] by a considerable margin. However, compared to MeGA-CDA [62] and Unbiased DA [12] our proposed method produces competitive performance with a drop of 3-4 mAP. Note that these UDA methods have access to labelled source data, whereas under the SFDA setting, the proposed model only has access to source-trained model. Furthermore, the proposed method outperforms SFDA methods like SFOD [39] and HCL [27] by 1.7 and 0.6 mAP, respectively. Sec-

Table 3. Quantitative results for PASCAL-VOC → Watercolor.

Type	Method	Online	Offline	bike	bird	car	cat	dog	prsn	mAP
S	Source only	✓	✗	68.8	46.8	37.2	32.7	21.3	60.7	44.6
UDA	DA Faster [7] (CVPR 2018)	✓	✗	75.2	40.6	48.0	31.5	20.6	60.0	46.0
	BDC Faster [50] (CVPR 2019)	✓	✗	68.6	48.3	47.2	26.5	21.7	60.5	45.5
	BSR [32] (ICCV 2019)	✓	✗	82.8	43.2	49.8	29.6	27.6	58.4	48.6
	WST [32] (ICCV 2019)	✓	✗	77.8	48.0	45.2	30.4	29.5	64.2	49.2
	SWDA [50] (CVPR 2019)	✓	✗	71.3	52.0	46.6	36.2	29.2	67.3	50.4
	HTCN [3] (CVPR 2020)	✓	✗	78.6	47.5	45.6	35.4	31.0	62.2	50.1
	I³Net [4] (CVPR 2021)	✓	✗	81.1	49.3	46.2	35.0	31.9	65.7	51.5
	Unbiased DA [12] (CVPR 2021)	✓	✗	88.2	55.3	51.7	39.8	43.6	69.9	55.6
SFDA	SFOD [39] (AAAI 2021)	✓	✗	76.2	44.9	49.3	31.6	30.6	55.2	47.9
	Mean-teacher [57]	✓	✗	73.6	47.6	46.6	28.5	29.4	56.6	47.1
	MemCLR (Ours)	✓	✗	70.7	48.5	51.3	31.6	34.0	61.3	49.6
O-SFDA	Tent [65] (ICLR 2021)	✗	✓	62.3	53.4	43.7	29.5	36.4	48.3	45.4
	MemCLR (Ours)	✗	✓	66.1	46.2	47.8	30.8	30.0	55.3	46.1

Table 4. Ablation analysis on Cityscapes→FoggyCityscapes.

Method	Mem items	prsn	rider	car	truck	bus	train	mcycle	bcycle	mAP
Source-only	✗	29.3	34.1	35.8	15.4	26.0	9.09	22.4	29.7	25.2
Student-Teacher	✗	33.1	42.2	44.7	24.0	33.6	17.8	26.8	38.1	32.5
SupCon	✗	33.0	43.1	49.8	26.5	31.1	23.3	27.7	37.2	33.8
MemCLR (Ours)	256	37.2	41.7	51.3	27.5	38.5	28.5	29.6	39.3	36.7
MemCLR (Ours)	512	37.4	45.2	51.9	24.4	39.6	25.2	31.5	41.6	37.1
MemCLR (Ours)	1024	37.7	42.8	52.4	24.5	40.6	31.7	29.4	42.2	37.7

ondly, when compared to the Test-time adaptation based methods such as Tent [66], our best-performing model surpasses it by a huge margin of by 3.0 mAP. Therefore, for Cityscape→FoggyCityscapes adaptation scenario, our proposed method produces state-of-the-art results in both online and offline SFDA settings.

4.1.2 Synthetic to real world adaptation

Collecting and annotating detection data is computationally intensive, where on top of assigning a category, one needs to add bounding boxes to every object location in the image. On the other hand, creating a synthetic dataset through simulation is much less computation-intensive and generates annotations for free. Hence, training a detector model on a synthetically generated dataset makes sense and then deploying it in real-world conditions. However, stylistic/appearance differences between real and synthetic data limit such deployment due to performance issues. Here, we formulate it as an online adaptation problem to update a synthetic data trained model on the real-world test data. Particularly, we consider a source model trained on Sim10k [29] on 10,000 training images with 58,701 bounding boxes of *car* category, rendered by the gaming engine *Grand Theft Auto*. For real-world test data we use the Cityscapes [11] validation set for online model adaptation. In Table 2, we report Sim10K→Cityscapes adaptation results on the existing UDA, SFDA, and O-SFDA methods. In an offline setting, compared to the existing UDA works such as DAFaster [8], SWDA [50] and RobustDA [30], the proposed method outperforms all of them by a considerable margin. Furthermore, when compared to SFOD [39] the proposed method is better by 0.7 mAP. In an online setting, compared to Tent [66], our proposed method outperforms it by 4.0 mAP. Therefore, our proposed is able to perform well under synthetic to real-world domain shifts.

4.1.3 Cross-camera adaptation

In most real-world applications, it is assumed that both training and test data would be collected using a camera with the same parameters. However, the camera parameters are often different, which causes the collected images to have different appearances, such as radial distor-

tions, tangential distortions, etc. This can cause the model to perform poorly due to changes in the camera parameters. Hence, to tackle any such camera distortions, we formulate the problem as an online adaptation problem and show that the proposed approach succeeds in generalizing to such cases. Here, we have access to only the source model, trained on the KITTI [16] dataset with 7,481 training images with bounding boxes for the *car* category. To emulate cross-camera scenario, we consider online adaptation on the Citsycapes [11] validation set containing 500 images. We report the results of the cross-camera adaptation experiment in Table 2. Similar to Sim→Cityscapes adaptation even for Kitti→Cityscapes adaptation, we show similar performance improvements compared to UDA, SFDA and O-SFDA methods. Specifically, in the O-SFDA setting, the proposed method outperforms Tent [66] by 5.6 mAP. Thus, our proposed method is able to model the cross-camera domain shifts effectively.

4.1.4 Real to artistic adaptation

Here, we evaluate the proposed method for the case where there is a *concept shift* in during inference. By *concept shift*, we refer to the case where there is a complete change in the object, e.g., going from real-world to artistic images. Unlike previous scenarios where the objects go through stylistic/appearance changes, the entire *concept* of an object is different, e.g., a real-world car vs a cartoon car [28]. We show that even in this challenging scenario, the proposed approach is able to improve model generalization through online updates. We consider a model trained on the Pascal-VOC data [14] which adapts to test set of Watercolor [28]. Specifically, the Watercolor consists of 1K training and 1K testing images with six categories. We compare PASCAL-VOC→Watercolor results with the existing methods in Table 3. From Table 3, we can infer that the proposed method outperforms most of the existing UDA methods and SFDA methods in offline settings. Further, in the online setting, when compared to TENT[65] the proposed method is able to outperform by a significant margin. This demonstrates the capability of the proposed method to generalize even for both online and offline settings.

4.2. Ablation analysis

Quantitative analysis. The Cityscapes→FoggyCityscapes ablation experiment results are reported in Table 4 for the offline-SFDA setting. We first consider a student-teacher offline update baseline which, compared to the source-only baseline, provides significant improvements. To have a

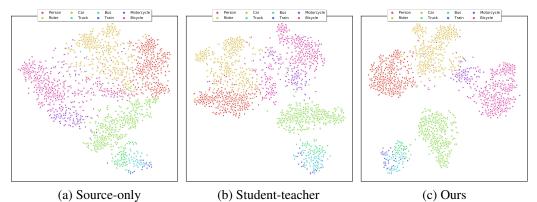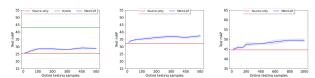(a) Source-only   (b) Student-teacher   (c) Ours

Figure 4. t-SNE [60] visualization of RoI features for source-only, student-teacher and our methods for Cityscapes to FoggyCityScapes online setting. Different colors represent different classes. Compared to the source-only and student-teacher method, our proposed method has learned better classification boundaries and compact feature representation for each category.

(a) Foggy-Cityscapes   (b) Cityscapes   (c) Watercolor

Figure 5. Quantitative comparison is performed to analyze the effect of the order of input sequence during online adaptation. We can observe from the variance that the input sequence order does not affect the model performance much. Note that, in online adaptation, the test samples are seen only once and adaptation happens in an unsupervised manner.

fair comparison, we also consider utilizing supervised contrastive loss [31] for offline updates. In particular, we utilize predictions provided by student-teacher training as label information needed for applying the supervised contrastive loss over object proposals. Denoted as SupCon in Table 4, the addition of supervised contrastive learning further improves the performance by 1.3 mAP. However, the proposed memory-based contrastive learning outperforms the supervised contrastive learning by 1.4 mAP, indicating the utility of the proposed method to learn better target representations. Finally, we analyze the performance of the proposed method by varying global memory bank capacity from 256 to 1024 memory items. As shown in Table 4, memory-based contrastive loss with 1024 memory items performs the best when compared to 256 and 512 memory items. Further, note that our model takes around 1 second to perform online adaptation for one sample.

Qualitative analysis. Fig. 4 shows t-SNE visualization for source-only, student-teacher training and the proposed method for the Cityscapes→FoggyCityscapes online-SFDA setting. The t-SNE [60] visualizations are created from the RoI features extracted from the predictions for 500 test images. Due to the distribution shift, the features are dispersed for the source-only baseline and classification boundaries are weak. With the help of student-teacher training, the model learns better classification boundaries, resulting in

better quantitative performance. However, the features in the student-teacher training have a large variance and do not have compact features. Whereas the proposed method has even better classification boundaries and learns compact features for each category, resulting in a more robust model. Further qualitative comparison is performed to analyze the effect of the order of input sequence during online adaptation is shown in Fig. 5. Multiple experiments with changing the order of input sequence are conducted and corresponding performance mean and variance is plotted in Fig. 5. We can observe from the variance that the order of input sequence does not much affect the model's performance. Further, we can observe the model performance increase as it encounters more test samples during online adaptation, showing the MemXformer effectiveness in exploiting online target distribution. Note that, in online adaptation, the test samples are seen only once and adaptation happens in an unsupervised manner.

5. Conclusion

In this work, we introduced a practical domain adaptation setting for the object detection task, which is feasible for real-world settings. Particularly, we proposed a novel unified adaptation framework which makes the detector models robust against online target distribution shifts. Further, We introduce the MemXformer module, which stores prototypical patterns of the target distribution and provides contrastive pairs to boost the contrastive learning on the target domain. We conducted extensive experiments on multiple detection benchmark datasets and compared existing unsupervised domain adaptation, source-free domain adaptation and test-time adaptation methods to show the effectiveness of the proposed approach for both online and offline adaptation of object detection models. We also analyzed multiple aspects of the proposed method in ablation experiments and identified increasing the online adaptation speed further is a potential directions for future research.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[2] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019.

[3] 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*, pages 8869–8878, 2020.

[4] Chaoqi Chen, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. I3net: Implicit instance-invariant network for adapting one-stage object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12576–12585, 2021.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

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

[8] 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*, pages 3339–3348, 2018.

[9] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1992–2001, 2017.

[10] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[11] 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 conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[12] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[16] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[17] Robert Geirhos, Carlos R Medina Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *arXiv preprint arXiv:1808.08750*, 2018.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6668–6677, 2019.

[21] Deepti Hegde and Vishal Patel. Attentive prototypes for source-free unsupervised domain adaptive 3d object detection. *arXiv preprint arXiv:2111.15656*, 2021.

[22] Deepti Hegde, Vishwanath Sindagi, Velat Kilic, A Brinton Cooper, Mark Foster, and Vishal Patel. Uncertainty-aware mean teacher for source-free unsupervised domain adaptive 3d object detection. *arXiv preprint arXiv:2109.14651*, 2021.

[23] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[24] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1989–1998, 2018.

[25] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[26] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection.

In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 749–757, 2020.

[27] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *arXiv preprint arXiv:2110.03374*, 2021.

[28] 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*, pages 5001–5009, 2018.

[29] 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? *arXiv preprint arXiv:1610.01983*, 2016.

[30] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 480–490, 2019.

[31] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[32] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6092–6101, 2019.

[33] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019.

[34] Youngeun Kim, Donghyeon Cho, Kyeongtak Han, Priyadarshini Panda, and Sungeun Hong. Domain adaptation without source data. *IEEE Transactions on Artificial Intelligence*, 2021.

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[36] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7046–7056, 2021.

[37] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4544–4553, 2020.

[38] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9650, 2020.

[39] 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. *arXiv preprint arXiv:2012.05400*, 2020.

[40] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020.

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[42] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[43] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021.

[44] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021.

[45] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[46] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[47] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[49] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 780–790, 2019.

[50] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019.

[51] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018.

[52] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018.

[53] V. A. Sindagi, P. Oza nad R. Yasarla, and V. M. Patel. Prior-based domain adaptive object detection for hazy and rainy conditions. In *European Conference on Computer Vision (ECCV)*, 2020.

[54] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019.

[55] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, pages 9229–9248. PMLR, 2020.

[56] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[57] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017.

[58] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[59] Jeya Maria Jose Valanarasu, Pengfei Guo, Vibashan VS, and Vishal M Patel. On-the-fly test-time adaptation for medical image segmentation. *arXiv preprint arXiv:2203.05574*, 2022.

[60] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[61] Vibashan VS, Poojan Oza, and Vishal M Patel. Instance relation graph guided source-free domain adaptive object detection. *arXiv preprint arXiv:2203.15793*, 2022.

[62] Vibashan VS, Poojan Oza, Vishwanath A Sindagi, Vikram Gupta, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. 2021.

[63] Vibashan Vs, Domenick Poster, Suya You, Shuowen Hu, and Vishal M Patel. Meta-uda: Unsupervised domain adaptive thermal object detection using meta-learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1412–1423, 2022.

[64] Vibashan VS, Jeya Maria Jose Valanarasu, and Vishal M Patel. Target and task specific source-free domain adaptive image segmentation. *arXiv preprint arXiv:2203.15792*, 2022.

[65] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[66] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

[67] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Instance-invariant domain adaptive object detection via progressive disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[68] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9010–9019, 2021.

[69] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11724–11733, 2020.

[70] Zhuliang Yao, Yue Cao, Shuxin Zheng, Gao Huang, and Stephen Lin. Cross-iteration batch normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12331–12340, 2021.

[71] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *European Conference on Computer Vision*, pages 86–102. Springer, 2020.

[72] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[73] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 687–696, 2019.