Black-box Unsupervised Domain Adaptation with Bi-directional Atkinson-Shiffrin Memory

Jingyi Zhang   Jiaxing Huang   Xueying Jiang   Shijian Lu*
S-lab, Nanyang Technological University
{Jingyi.Zhang, Jiaxing.Huang, Shijian.Lu}@ntu.edu.sg, xueying003@e.ntu.edu.sg

Abstract

Black-box unsupervised domain adaptation (UDA) learns with source predictions of target data without accessing either source data or source models during training, and it has clear superiority in data privacy and flexibility in target network selection. However, the source predictions of target data are often noisy and training with them is prone to learning collapses. We propose BiMem, a bi-directional memorization mechanism that learns to remember useful and representative information to correct noisy pseudo labels on the fly, leading to robust black-box UDA that can generalize across different visual recognition tasks. BiMem constructs three types of memory, including sensory memory, short-term memory, and long-term memory, which interact in a bi-directional manner for comprehensive and robust memorization of learnt features. It includes a forward memorization flow that identifies and stores useful features and a backward calibration flow that rectifies features' pseudo labels progressively. Extensive experiments show that BiMem achieves superior domain adaptation performance consistently across various visual recognition tasks such as image classification, semantic segmentation and object detection.

1. Introduction

Unsupervised domain adaptation (UDA) has been studied extensively in recent years, aiming to alleviate data collection and annotation constraint in deep network training [8, 67, 25, 57, 75, 88, 12, 58, 52, 56, 93]. However, most existing UDA methods could impair data privacy and confidentiality [43] as they require to access source data or source-trained models during training. In addition, most existing UDA imposes the same network architecture ($i.e.$, the source model architecture) in adaptation which limits the flexibility of selecting different target networks in UDA [43]. Black-box UDA only requires the initial predictions of target data provided by black-box source models [43] ($i.e.$, only the model predictions of target data are available [43]) for domain adaptation, as shown in Table 1. It has attracted increasing attention in recent years [43, 44] due to its advantages in data privacy and flexibility of allowing different target networks regardless of the source-trained black-box models.

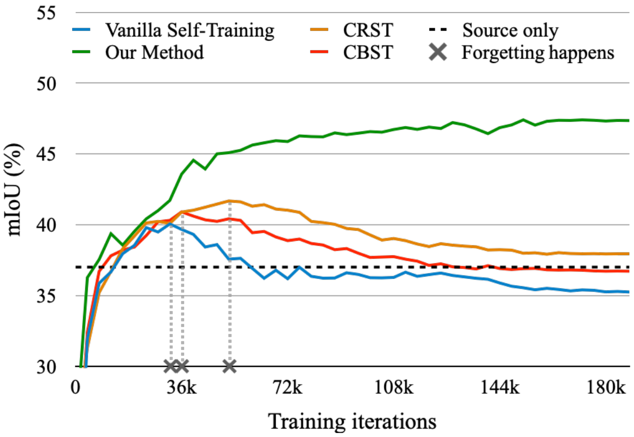

Figure 1: Self-training including Vanilla self-training [34], and advanced CBST [92] and CRST [93] tends to 'forget' learnt useful features due to the accumulation of noisy pseudo labels along the training process in black-box UDA – it learns well at the early adaptation stage but collapses with the adaptation moving on and the finally adapted models in [34, 92] cannot even compare with the *Source only*. We introduce bi-directional memory to remember useful and representative features learnt during adaptation, which helps calibrate noisy pseudo labels on the fly and leads to stabler black-box UDA without collapse. The experiments were conducted over domain adaptive semantic segmentation task GTA5 → Cityscapes and the evaluations were performed over the Cityscapes validation data.

However, the black-box predictions of the target data are prone to errors due to the cross-domain discrepancy, lead-

[*] Corresponding author

Comparisons of Different UDA Setups					
Adaptation Setups	Source Data	Source-Trained Model	Source-Predicted Target Labels	Target Data	Privacy Risk
Conventional UDA	✓	✓	✓	✓	High
Source-free UDA	✗	✓	✓	✓	Medium
Black-box UDA	✗	✗	✓	✓	Low

Table 1: Comparison of different UDA setups: Black-box UDA better preserves data privacy, requiring neither source data nor source-trained models but just source-predicted labels of target data during domain adaptation. It also allows different target networks regardless of source networks.

ing to a fair amount of false pseudo labels. As a result, self-training [34, 92, 93] with the pseudo-labelled target data is often susceptible to collapse as illustrated in Fig. 1. We argue that the learning collapse is largely attributed to certain 'forgetting' in network training. Specifically, the self-training with noisy pseudo labels learns well at the early stage and the trained model outperforms the *Source only* over the test data as shown in Fig. 1. However, the performance deteriorates gradually and drops even lower than that of the *Source only* as the training moves on. This shows that the self-training learns useful target information and adapts well towards the target data at the early training stage, but then gradually forgets the learnt useful target knowledge and collapses with the accumulation of pseudo-label noises along the training process.

Inspired by Atkinson-Shiffrin memory [3], we propose BiMem, a bi-directional memorization mechanism that constructs three types of memories to address the 'forgetting' problem in black-box UDA. The three types of memory interact in a bi-directional manner including a forward memorization flow and a backward calibration flow as illustrated in Fig. 2. In the forward memorization flow, the short-term memory actively identifies and stores hard samples (*i.e.*, samples with high prediction uncertainty) from the sensory memory which buffers fresh features from the current training batch. Meanwhile, the long-term memory accumulates features from the sensory and short-term memory, leading to comprehensive memorization that captures fresh yet representative features. In backward calibration flow, we calibrate the pseudo labels of memorized features progressively where the short-term memory is calibrated by the long-term memory while the sensory memory is corrected by both short-term and long-term memory, leading to robust memorization via a progressive calibration process. Hence, BiMem builds comprehensive and robust memory that allows to learn with more accurate pseudo labels and produce better adapted target model as illustrated in Fig. 1.

The contributions of this work can be summarized in three aspects. *First*, we design BiMem, a general black-box UDA framework that works well on different visual recognition tasks. To the best of our knowledge, this is the first work that explores and benchmarks black-box UDA over different visual recognition tasks. *Second*, we design three types of memory that interact in a bi-directional manner, leading to less 'forgetting' of useful and representative features, more accurate pseudo labeling of target data on the fly, and better adaptation in black-box UDA. *Third*, extensive experiments over multiple benchmarks show that BiMem achieves superior performance consistently across different computer vision tasks including image classification, semantic segmentation, and object detection.

2. Related Work

Unsupervised Domain Adaptation (UDA) has been extensively studied in various visual recognition tasks for mitigating the data annotation constraint in deep network training [12, 8, 92, 57, 17, 67, 78, 75, 91, 88, 36, 24, 49, 87, 86, 22, 1, 50, 84, 6, 23, 13, 38, 18, 74, 79, 76, 81, 83, 20, 85]. Conventional UDA requires to access the source data in training, which may not be valid while facing concerns in data privacy and confidentiality. Source-free UDA [42, 21, 48, 11, 37, 73, 77, 80, 31, 63, 10, 68, 77, 32, 35] addresses the data privacy concerns by adopting a source-trained model instead of source data in domain adaptation, but it still requires source models from which source data could be recovered via certain generation techniques [14]. Beyond data privacy, most conventional and source-free UDA imposes the same network architecture in domain adaptation which precludes the flexibility of selecting different target networks in UDA. We study black-box UDA, a new UDA setup that only requires initial predictions of target data during domain adaptation, hence having little concern of data privacy but great flexibility in target network selection.

Black-box UDA learns with source predictions of target data without requiring either source data or source models during domain adaptation [43]. It has recently attracted increasing attention as it has little data privacy concerns and allows flexible selection of target networks. Most existing black-box UDA studies focus on image classification tasks [45, 44, 43]. For example, [44] splits target data into two parts according to the prediction confidence

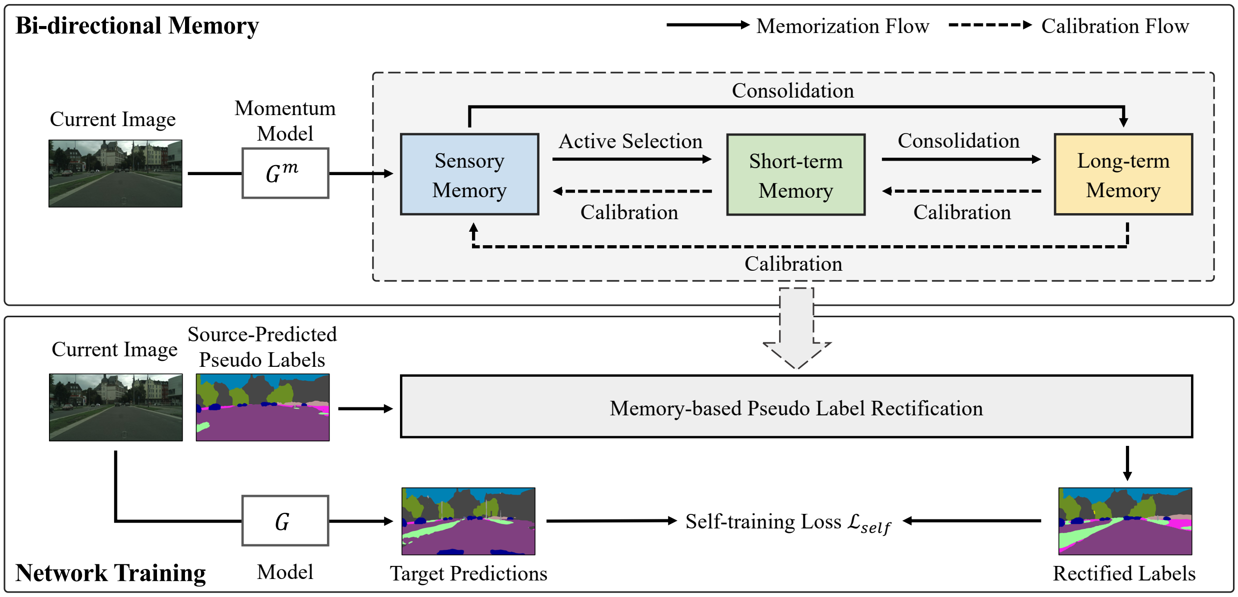

Figure 2: Overview of the proposed BiMem. BiMem constructs three types of memory that interact in a bi-directional manner via a forward memorization flow and a backward calibration flow as shown in the top part. In the forward memorization flow, input images are fed to momentum model G^m and the generated features are exploited to construct three types of memory for comprehensive memorization, capturing fresh yet representative information during the adaptation. In the backward calibration flow, the pseudo labels of memorized features are calibrated progressively for robust memorization. Thus, BiMem builds comprehensive and robust memory that allows to rectify source-predicted pseudo labels conditioned on the features stored in BiMem as shown in the bottom part, leading to stable Black-box UDA and better adapted target models.

and handles unconfident data by adopting a semi-supervised learning strategy. [43] distills knowledge via adaptive label smoothing and fine-tunes the distilled model to fit the target distribution. Beyond image classification, ATP [70] introduces negative pseudo labels into self-training and proposes information propagation for tackling black-box UDA for semantic segmentation. Different from these studies [44, 43], BiMem focuses on addressing the intrinsic 'forgetting' issue in black-box UDA and it is generic and achieves superior performance across various visual recognition tasks.

Memory-based Learning has been widely explored in computer vision [41, 72, 29, 46, 15, 21, 33, 62, 47, 69, 60, 89, 27, 51, 4, 71, 30, 19]. For the task of UDA, several recent studies employ memory to facilitate cross-domain adaptation [21, 66, 26]. For example, [21] tackles the source-free UDA challenge by memorizing historical models which helps the target model to benefit from the source-domain knowledge. [66] stores historical data to generate category-specific attention maps for learning discriminative target representations, which effectively facilitates unsupervised domain adaptation for object detection. Different from these studies [21, 66, 15], BiMem relies on neither source data nor source models but constructs three types of memory using only target data, which interact with each other in a bi-directional manner for black-box UDA. It aims to build comprehensive yet robust memorization to mitigate the intrinsic 'forgetting' issue in black-box UDA.

3. Method

3.1. Preliminaries of Atkinson-Shiffrin Memory

Human memory is powerful in encoding, storing and retrieving information, which plays a fundamental role in human learning. In the field of Neuroscience, Atkinson-Shiffrin memory [3] models human memory by three types of memory including sensory memory, short-term memory and long-term memory. Among the three, sensory memory stores sensory information (*e.g.*, visual information) which resides for a very brief period of time (*e.g.*, 0.5–1.0 second) and then decays. Short-term memory buffers information selected from sensory memory, where the information in short-term memory has a longer duration (*e.g.*, 18-20 seconds) before being lost. Long-term memory consolidates, compacts and stores information from short-term memory, where the information in long-term memory is relatively permanent. Inspired by the powerful human memory mechanism in human learning, we follow Atkinson-Shiffrin memory theory and build three types of memory for tackling the 'forgetting' issue in Black-box UDA. Specifically, we propose a bi-directional memorization mechanism that enables the three types of memory to interact with each other, leading to comprehensive and robust memorization, less 'forgetting' of useful and representative information, and better adaptation in black-box UDA.

3.2. Task Definition

This work focuses on black-box UDA for visual recognition tasks including image classification, semantic segmentation, and object detection. As shown in Table 1, in training, black-box UDA does not access either source data or source-trained model but just pseudo-label predictions of unlabelled target data. It involves a *black-box predictor* that is pre-trained with certain large-scale training data and stored on cloud as a service provider (for pseudo label prediction). In our study, we employ a source-trained model to simulate the black-box predictor which just provides pseudo-label prediction of target data but is not accessible while training the domain adaptation network as in [43]. We take the single-source scenario for the simplicity of illustrations, which involves a target domain $\mathcal{D}_t = \{X_t, \hat{Y}_t\}$, where the pseudo label $\hat{Y}_t$ of target sample X_t is inferred from one black-box source model G_s. The goal of black-box UDA is to train a target model G that well performs on X_t.

Black-box Source Model Generation. We train a source model G_s that will act as a black-box predictor to provide the initial predictions of target data for domain adaptation as in [43]. Generally, the source model G_s is trained with the labelled data in source domain $\mathcal{D}_s = \{X_s, Y_s\}$ by a supervised loss:

$$\mathcal{L}_{sup} = l(G_s(X_s), Y_s), \tag{1}$$

where $l(\cdot)$ denotes a task-related loss, *e.g.*, the standard cross-entropy loss for image classification.

3.3. BiMem

BiMem constructs three types of memory including sensory memory, short-term memory and long-term memory, which interact in a bi-directional manner for mitigating the 'forgetting' of useful and representative features during black-box adaptation. Specifically, *Sensory Memory* stores fresh information along the adaptation by buffering the features of current batch. *Short-term Memory* actively identifies and stores the hard samples from sensory memory according to samples' uncertainty. *Long-term Memory* consolidates sensory and short-term memories via category-wise compaction and accumulation of all the features dequeued from sensory memory and short-term memory, which memorizes the global and representative information along the whole adaptation process.

Taking a batch of K target samples $x_t = \{x_t^k\}_{k=1}^K$ ($x_t^k \in X_t$) and the corresponding source-predicted labels $\hat{y}_t = \{\hat{y}_t^k\}_{k=1}^K$ ($\hat{y}_t^k \in \hat{Y}_t$) as an example. We elaborate the memory construction and update in *Forward Memorization*, and memory calibration in *Backward Calibration*.

Algorithm 1 The proposed BiMem for black-box UDA.

```
Require: Target data $X_t$ with source-predicted pseudo labels $\hat{Y}_t$
Ensure: Learnt target model $G$
Initialize $G$ and $G^m$
for $iter = 1$ to $Max\_Iter$ do
    Update the momentum model: $G^m \leftarrow G$
    Sample $\{x_t, \hat{y}_t\} \subset \{X_t, \hat{Y}_t\}$ and encode $x_t$: $f_{sm} = G^m(x_t)$
    Forward Memorization Flow:
    Update sensory memory: replace old features $f_{sm}^*$ by $f_{sm}$
    Update short-term memory: dequeue oldest features $f_{st}^*$ and enqueue new features selected by Eq. 2
    Update long-term memory: accumulate $f_{sm}^*$ and $f_{st}^*$ by Eqs. 3-4
    Backward Calibration Flow:
    Calibrate short-term memory by long-term memory as in Eqs. 5-6
    Correct sensory memory by long-term memory and the calibrated short-term memory as in Eq. 7
    Network Training:
    Acquire $\overline{y}_t$ by denoising $\hat{y}_t$ with sensory memory by Eq. 8
    Optimize target model $G$ with $\{x_t, \overline{y}_t\}$ by Eq. 9
end for
return $G$
```

Forward Memorization aims to construct comprehensive memories that capture fresh yet representative information during the adaptation.

We employ a momentum model to encode images for stable memorization as the slow and smooth update mechanism in momentum model allows generating consistent features along the training process. Specifically, we feed x_t into the momentum model G^m (the moving averaged of G, *i.e.*, $\theta_{G^m} \leftarrow \gamma\ \theta_{G^m} + (1-\gamma)\theta_G$, and γ is a momentum coefficient) to acquire the momentum features $f_{sm} = \{f_{sm}^k\}_{k=1}^K$ and the corresponding category predictions $\{\{p_{sm}^{(k,c)}\}_{c=1}^C\}_{k=1}^K$, where C denotes the number of categories.

Sensory memory buffers $\{f_{sm}, p_{sm}\}$ and it is updated in every iteration, *i.e.*, the features of previous batch of samples f_{sm}^* will be replaced by the features of current batch of samples f_{sm}, which allows sensory memory to capture fresh knowledge learnt by the model.

Active Selection. Short-term memory actively selects and stores hard features from the sensory memory. We identify hard samples that are difficult to be classified and generally with high classification uncertainty, *i.e.*, the sample with high uncertainty is considered as a hard sample. Specifically, we measure the uncertainty of K samples by their prediction entropy:

$$\mathcal{H}(f_{sm}^k) = -\sum_{c=1}^{C} p_{sm}^{(k,c)}\ \log\ p_{sm}^{(k,c)}, \tag{2}$$

where the Top-N samples with the highest entropy (*i.e.*, lowest certainty) are selected and stored in short-term memory.

Short-term memory works as a FIFO queue with a fixed size of M ($M>N$), where N oldest features f_{st}^{*} will be dequeued and the fresh N features selected via Eq. 2 will be enqueued to update the short-term memory $f_{st}=\{f_{st}^{m}\}_{m=1}^{M}$. Note that the short-term feature f_{st}^{m} is stored with its category prediction p_{st}^{m}. Such FIFO update strategy can avoid GPU memory explosion as the features are uncompressed before storing in short-term memory.

Consolidation. Long-term memory stores global and representative information along the whole adaptation process by accumulating all the features dequeued from sensory memory and short-term memory.

Inspired by human memory consolidation mechanism [61], we consolidate temporary memories ($i.e.$, sensory and short-term memories) into long-term memory for more stable and sustained long-term memorization. Specifically, we compact the features dequeued from sensory and short-term memories into category-wise feature centroids $\delta_{lt}=\{\delta_{lt}^{c}\}_{c=1}^{C}$, where the feature centroid δ_{lt}^{c} of each category is computed as the following:

$$\delta_{lt}^{c}=\frac{\sum_{f\in f_{sm}^{*}\cup f_{st}^{*}} f\cdot\mathbb{1}(\hat{c}=c)}{\sum_{f\in f_{sm}^{*}\cup f_{st}^{*}}\mathbb{1}(\hat{c}=c)}, \tag{3}$$

where $\mathbb{1}$ is an indicator function that returns '1' if f belongs to c-th category, and '0' otherwise.

To allow the long-term memory to memorize all the information along the adaptation process, we update it with δ_{lt} in a momentum way:

$$\delta_{lt}\leftarrow(1-\gamma')\,\delta_{lt}+\gamma'\,\delta_{lt}^{*}, \tag{4}$$

where δ_{lt}^{*} denotes the old long-term features and γ' is a coefficient for smooth feature update in the long-term memory.

With the backward calibration (described in following paragraphs), long-term memory consolidates the calibrated sensory and short-term memories iteratively, the features stored in which tends to gradually move closer to the true feature centroid of each category while the adaptation moves on. Meanwhile, the consolidation operations specially consider the features of hard samples (with calibrated pseudo labels) that are generally sparse, which allows the long-term features to be more representative.

Backward Calibration rectifies the memories progressively, aiming to suppress the false pseudo labels predicted for the stored features for robust memorization. Specifically, we first employ the representative information accumulated in the long-term memory to correct short-term memory. Then, long-term memory and the corrected short-term memory collaborate to calibrate sensory memory.

We calibrate short-term memory by re-weighting the predicted category probability $p_{st}=\{p_{st}^{c}\}_{c=1}^{C}$ of short-term features f_{st} as the following:

$$\overline{p}_{st}^{c}=w^{c}\otimes p_{st}^{c}, \tag{5}$$

where $\otimes$ denotes the element-wise multiplication and $w^{c}\in\{w^{c}\}_{c=1}^{C}$ is the calibration weight for the corresponding c-th category probability. The calibration weight w for each short-term feature f_{st}^{m} is calculated according to the distance between the short-term feature and the category centroids stored in the long-term memory. Generally, if the short-term feature f_{st}^{m} is far from the c-th long-term feature δ^{c}, this feature should be assigned with a low probability of belonging to the c-th category. Therefore, the calibration weight in Eq. 5 is defined as the following:

$$w^{(m,c)}=\text{Softmax}(-||f_{st}^{m}-\delta_{lt}^{c}||_{1}), \tag{6}$$

where $w^{c}=\{w^{(m,c)}\}_{m=1}^{M}$ and $||\cdot||_{1}$ denotes $L1$ distance and the softmax operation is performed along the category dimension.

Next, we employ long-term memory and the calibrated short-term memory to correct sensory memory by assigning each sensory feature $f_{sm}^{k}\in f_{sm}$ a new category probability vector. Specifically, we first compute the centroids $\delta_{st}=\{\delta^{c}\}_{c=1}^{C}$ of hard features over the calibrated short-term memory $\{f_{st},\overline{p}_{st}\}$ by adopting Eq. 3. With the long-term feature centroids δ_{lt} and the short-term feature centroids δ_{st}, we assign a new category probability to each sensory feature f_{sm}^{k} as the following:

$$\overline{p}_{sm}^{(k,c)}=\text{Softmax}[(-||f_{sm}^{k}-\delta_{lt}^{c}||_{1})+(-||f_{sm}^{k}-\delta_{st}^{c}||_{1})], \tag{7}$$

where the softmax operation works along the category dimension.

3.4. Network Training

With the comprehensive and robust memorization built by BiMem, we calibrate the source-predicted pseudo labels $\hat{y}_{t}$ of each sample in $\{x_{t}^{k}\}_{k=1}^{K}$ as shown in the bottom part of Fig. 2. For sample x_{t}^{k}, we read out its calibrated category-wise probabilities $\overline{p}_{sm}^{(k,c)}$ (acquired by Eq. 7) from sensory memory. Next, the pseudo label $\hat{y}_{t}^{k}$ of x_{t}^{k} is denoised by re-weighting its category-wise probability $\hat{p}^{c}$:

$$\overline{y}_{t}^{k}=\underset{c}{\text{argmax}}(\,\overline{p}_{sm}^{(k,c)}\otimes\hat{p}^{(k,c)}), \tag{8}$$

where the denoised pseudo labels $\overline{y}_{t}$ of a batch of K samples x_{t} can be obtained by rectifying each sample independently.

With the rectified label $\overline{y}_{t}$, the model G is optimized with the unsupervised self-training loss defined as the following:

$$\mathcal{L}_{self}=l(G(x_{t}),\overline{y}_{t}), \tag{9}$$

where $l(\cdot)$ denotes a task-related loss, $e.g.$, the standard cross-entropy loss for image classification. The overall training objective is to minimize the unsupervised self-training loss $\mathcal{L}_{self}$, $i.e.$, $\arg\min_{G}\mathcal{L}_{self}$, as detailed in Algorithm 1.

GTA5 → Cityscapes Semantic Segmentation

Methods	Road	SW	Buil.	Wall	Fence	Pole	TL	TS	Veg.	Ter.	Sky	PR	Rider	Car	Truck	Bus	Train	Mot.	Bike	mIoU
Source only [16]	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	25.4	49.9	17.2	25.9	**6.5**	25.3	36.0	36.6
CBST [92]	82.7	22.4	70.3	29.1	21.9	21.7	25.9	23.3	76.5	22.3	76.9	55.2	26.4	65.5	36.7	43.3	0.0	26.9	37.1	40.3
CRST [93]	87.1	48.7	74.7	27.4	16.5	**38.5**	35.8	23.9	84.9	36.1	67.3	**60.9**	24.4	82.1	23.1	29.3	0.4	23.7	7.4	41.7
SFDA [48]	92.2	51.8	81.0	3.8	23.6	23.2	18.9	25.8	83.9	35.1	84.1	52.3	24.9	81.4	29.5	38.3	0.1	32.8	40.4	43.3
UR [11]	92.7	53.8	81.5	10.4	23.2	25.6	16.5	30.8	84.3	36.9	83.5	55.6	24.9	82.7	32.5	40.2	0.5	32.7	44.6	44.8
HCL [21]	93.3	58.0	81.9	23.8	24.5	24.9	8.5	31.4	84.2	37.4	84.6	57.4	24.2	84.1	29.1	39.9	0.0	33.1	47.5	45.7
TAP [70]	83.6	25.8	81.9	30.2	25.2	27.9	**36.2**	28.7	84.8	34.4	77.5	62.2	35.7	81.5	32.3	16.8	0.0	41.7	**53.5**	45.3
DINE [43]	88.2	44.2	**83.5**	14.1	**32.4**	23.5	24.6	**36.8**	**85.4**	38.3	**85.3**	59.8	**27.4**	84.7	30.1	42.2	0.0	**42.7**	45.3	46.7
BiMem	**94.2**	**59.5**	81.7	**35.2**	22.9	21.6	10.0	34.3	85.2	**42.4**	85.0	56.8	26.4	**85.6**	**37.2**	**47.4**	0.2	39.9	50.9	**48.2**

Table 2: Experiments on semantic segmentation over black-box UDA task GTA5 → Cityscapes.

SYNTHIA → Cityscapes Semantic Segmentation

Methods	Road	SW	Buil.	Wall*	Fence*	Pole*	TL	TS	Veg.	Sky	PR	Rider	Car	Bus	Mot.	Bike	mIoU	mIoU*
Source only [16]	55.6	23.8	74.6	**9.2**	0.2	24.4	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	33.5	38.6
CBST [92]	76.7	30.5	69.7	8.4	0.3	31.6	0.1	**23.2**	78.4	75.7	50.1	20.1	74.1	18.6	10.0	21.3	36.8	42.1
CRST [93]	65.9	26.4	71.3	6.7	0.1	33.8	10.8	24.1	81.6	79.9	53.2	15.9	74.8	19.7	12.9	21.9	37.5	43.0
SFDA [48]	77.2	**32.7**	74.0	0.7	**0.4**	34.9	13.9	20.9	82.8	79.1	52.5	**21.4**	74.7	14.7	11.2	23.1	38.8	44.5
UR [11]	74.2	29.0	75.3	0.2	0.0	38.6	17.8	15.0	81.4	53.6	65.9	13.5	74.6	30.0	33.9	**26.6**	39.5	45.4
HCL [21]	**83.8**	32.5	**80.7**	0.3	0.2	28.3	12.1	5.6	84.1	**81.4**	60.3	15.0	**82.9**	25.4	16.3	25.5	40.8	46.6
DINE [43]	77.5	29.6	79.5	4.3	0.3	39.0	21.3	13.9	81.8	68.9	66.6	13.9	71.7	33.9	**34.2**	18.6	40.9	47.0
BiMem	78.8	30.5	80.4	5.9	0.1	**39.2**	**21.6**	15.0	**84.7**	74.3	**66.8**	14.1	73.3	**36.0**	32.3	21.8	**42.2**	**48.4**

Table 3: Experiments on semantic segmentation over black-box UDA task SYNTHIA → Cityscapes. Following previous studies [92, 93], mIoU is evaluated on 16 classes while mIoU* is evaluated on 13 classes.

Cityscapes → Foggy cityscapes Object Detection

Methods	pers.	rider	car	truck	bus	train	mot.	bike	mAP
Source only [90]	37.7	39.1	44.2	17.2	26.8	5.8	21.6	35.5	28.5
WSOD [25]	39.9	40.1	50.6	16.5	34.9	8.1	25.2	38.2	31.7
WST [28]	40.7	40.4	53.8	17.1	35.0	5.1	30.2	39.1	32.8
SFOD [40]	39.2	39.3	51.8	21.7	33.6	12.5	31.2	**42.9**	34.0
LODS [39]	41.5	42.0	54.5	20.5	37.2	23.9	27.1	40.8	35.9
DINE [43]	41.4	40.9	55.0	21.8	38.5	25.7	28.2	40.4	36.5
BiMem	**42.2**	**42.5**	**56.9**	**23.4**	**39.7**	**28.5**	**32.4**	41.3	**38.4**

Table 4: Experiments on object detection over black-box UDA task Cityscapes → Foggy Cityscapes.

SYNTHIA → Cityscapes Object Detection

Methods	person	rider	car	bus	mot.	bike	mAP
Source only [90]	32.7	19.4	34.7	17.2	5.1	21.8	21.8
WSOD [25]	36.8	21.5	39.3	19.9	4.5	23.2	24.2
WST [28]	35.8	21.9	40.0	23.1	3.9	23.4	25.6
SFOD [40]	39.0	26.5	42.2	24.5	2.9	25.4	26.7
LODS [39]	37.3	30.5	39.6	27.9	6.3	25.8	27.9
DINE [43]	**41.2**	26.0	45.3	**28.4**	**6.8**	26.7	29.1
BiMem	40.8	**36.7**	**49.6**	27.5	5.7	**28.9**	**31.5**

Table 5: Experiments on object detection over black-box UDA task SYNTHIA → Cityscapes.

4. Experiment

This section presents experiments including datasets, implementation details, benchmarking over the tasks of semantic segmentation, object detection and image classification, as well as discussion of specific parameters and designs. More details are to be described in the ensuing subsections.

4.1. Datasets

We evaluate BiMem over multiple datasets across three widely studied computer vision tasks as listed:

Black-box UDA for Semantic Segmentation: We study two domain adaptive semantic segmentation tasks GTA5 [53] → Cityscapes [9] and SYNTHIA [54] → Cityscapes.

Black-box UDA for Object Detection: We study two domain adaptive detection tasks Cityscapes [9] → Foggy Cityscapes [59] and SYNTHIA [54] → Cityscapes [9].

Black-box UDA for Image Classification: We study two domain adaptive image classification tasks Office-Home [65] and Office-31 [55]. Office-home consists of 12 adaptation tasks across 4 domains: Art, Clipart, Product and Real-world. Office-31 includes 6 adaptation tasks across 3 domains: Amazon, DSLR and Webcam.

We provide more details of the involved datasets in the appendix.

4.2. Implementation Details

Semantic Segmentation: We adopt DeepLab-V2 [7] with ResNet-101 [16] as the segmentation network as in [64, 92].

Object Detection: We adopt deformable-DETR [90] with ResNet-50 [16] as detection network as in [5, 90].

Image Classification: Following [43], we adopt ResNet-

Office-Home Classification													
Methods	Ar → Cl	Ar→ Pr	Ar → Re	Cl→ Ar	Cl→ Pr	Cl→ Re	Pr→ Ar	Pr→ Cl	Pr→ Re	Re → Ar	Re→ Cl	Re→ Pr	Avg.
Source only [16]	44.1	66.9	74.2	54.5	63.3	66.1	52.8	41.2	73.2	66.1	46.7	77.5	60.6
NLL-OT [2]	49.1	71.7	77.3	60.2	68.7	73.1	57.0	46.5	76.8	67.1	52.3	79.5	64.9
NLL-KL [82]	49.0	71.5	77.1	59.0	68.7	72.9	56.4	46.9	76.6	66.2	52.3	79.1	64.6
HD-SHOT [42]	48.6	72.8	77.0	60.7	70.0	73.2	56.6	47.0	76.7	67.5	52.6	80.2	65.3
SD-SHOT [42]	50.1	75.0	78.8	63.2	72.9	76.4	60.0	48.0	79.4	69.2	54.2	81.6	67.4
DINE [43]	52.2	78.4	81.3	65.3	76.6	78.7	62.7	49.6	82.2	69.8	55.8	84.2	69.7
BiMem	**54.5**	**78.8**	**81.4**	**66.7**	**78.7**	**79.6**	**65.9**	**53.6**	**82.3**	**73.6**	**57.8**	**84.9**	**71.5**

Table 6: Experiments on image classification over black-box UDA task Office-Home.

Office-31 Classification							
Methods	A → D	A→ W	D → A	D→ W	W→ A	W→ D	Avg.
Source only [16]	79.9	76.6	56.4	92.8	60.9	98.5	77.5
NLL-OT [2]	88.8	85.5	64.6	95.1	66.7	98.7	83.2
NLL-KL [82]	89.4	86.8	65.1	94.8	67.1	98.7	83.6
HD-SHOT [42]	86.5	83.1	66.1	95.1	68.9	98.1	83.0
SD-SHOT [42]	89.2	83.7	67.9	95.3	71.1	97.1	84.1
DINE [43]	91.6	86.8	72.2	96.2	73.3	98.6	86.4
BiMem	**92.8**	**88.2**	**73.9**	**96.8**	**75.3**	**99.4**	**87.7**

Table 7: Experiments on image classification over black-box UDA task Office-31.

50 [16] for the tasks Office-Home and Office-31.

We provide more implementation details in the appendix.

4.3. Black-box UDA for Semantic Segmentation

We evaluate BiMem over two black-box domain adaptive semantic segmentation tasks GTA5 → Cityscapes and SYNTHIA → Cityscapes. As there is little prior black-box UDA research on semantic segmentation, we benchmark BiMem by reproducing conventional UDA methods [92, 93] and source-free UDA methods [48, 11, 21]. In addition, we rerun the SOTA black-box UDA approach [43] (designed for image classification task) on semantic segmentation task for benchmarking. Tables 2 and 3 present the experimental results. It can be observed that BiMem achieves superior segmentation performance clearly, largely because BiMem builds comprehensive and robust memorization that helps calibrate the noisy pseudo label and mitigate the 'forgetting' issue in Black-box UDA.

4.4. Black-box UDA for Object Detection

We evaluate BiMem over two black-box domain adaptive object detection tasks Cityscapes → Foggy Cityscapes and SYNTHIA → Cityscapes. Similar to semantic segmentation benchmarking, we reproduce UDA-based object detection methods [25, 28, 40, 39] and rerun the SOTA black-box UDA approach [43] (designed for image classification task) for comparisons. As shown in Tables 4-5, BiMem achieves superior detection performance as well, indicating the superior generalization of the proposed BiMem.

Row	Forward Memorization Flow			Backward Calibration Flow			mIoU
No.	SM → ST	SM → LT	ST → LT	SM ← ST	SM ← LT	ST ← LT	
1							35.2
2	✓			✓			44.3
3		✓			✓		44.5
4	✓	✓		✓	✓		46.4
5	✓	✓	✓	✓	✓		47.5
6	✓	✓		✓	✓	✓	47.4
7	✓	✓	✓	✓	✓	✓	**48.2**

Table 8: Ablation study of BiMem over GTA5 → Cityscapes semantic segmentation task, where 'SM', 'ST' and 'LT' stand for sensory memory, short-term memory and long-term memory, respectively.

4.5. Black-box UDA for Image Classification

Following [43], we evaluate BiMem over two popular black-box UDA classification tasks Office-Home and Office-31. Tables 6-7 show experimental results. It can be observed that BiMem outperforms state-of-the-art methods clearly. The superior performance is largely attributed to the memory mechanism in BiMem which helps produce more accurate pseudo labels while learning domain adaptive models.

4.6. Ablation Studies

We conduct extensive ablation studies to examine how different BiMem designs contribute to the overall performance. Table 8 shows experimental results over domain adaptive semantic segmentation task GTA5 → Cityscapes. We can see that the conventional self-training [34] in the 1st Row does not perform well due to the 'forgetting' issue in black-box UDA.

We first study how the interaction between sensory memory and short-/long-term memory affects the adaptation performance. Specifically, on top of the conventional self-training, further including the memorization and calibration flows between sensory memory and either short-term memory or long-term memory improves the performance clearly, as shown in Rows 2-3. This shows that either the hard features captured in short-term memory or the representative features accumulated in long-term memory provides useful information for denoising the initial source-predicted pseudo labels. Besides, we can see that the network performs clearly better when sensory memory interacts with both short-term memory and long-term memory in Row 4, demonstrating that short-term memory and long-term memory complement each other as they capture different features with complementary information.

Next, we investigate how the memorization and calibration flows between short-term and long-term memories benefit Black-box adaptation. Specifically, the memorization flow from short-term memory to long-term memory that specially compacts and accumulates the hard features into long-term memory brings clear improvements as shown in Row 5, indicating that the long-term features become more representative as hard features are generally rare. Besides, employing long-term memory to denoise pseudo labels in short-term memory also improves the performance by a large margin as shown in Row 6, showing that hard features in short-term memory are noisy and can be calibrated by long-term features effectively. At last, BiMem that includes all the memorization and calibration flows performs clearly the best, demonstrating that the proposed bi-directional flow enables comprehensive yet robust memorization which leads to stable and effective Black-box UDA.

4.7. Discussion

Generalization across Different Tasks: Our BiMem is general and works well over various computer vision tasks consistently without any task-specific modifications and fine-tuning as shown in Sections 4.3-4.5. The superior generalization capability of BiMem is largely attributed to its task-agnostic underlying mechanism and designs that enable BiMem to work consistently across different tasks.

Analysis of the 'Forgetting' Issue in Black-box UDA: We examine the source of the 'forgetting' illustrated in Fig. 1 with a controlled experiment. Specifically, we split the target data into two subsets according to their initial pseudo labels $\hat{Y}_t$ (predicted by the black-box predictor). This produces target data with correct initial pseudo label (*i.e.*, $X_t^{\text{correct}} = X_t[\hat{Y}_t = Y_t]$) and target data with incorrect initial pseudo labels (*i.e.*, $X_t^{\text{incorrect}} = X_t[\hat{Y}_t \neq Y_t]$), where Y_t denotes the ground-truth of X_t and $X_t = X_t^{\text{correct}} \cup X_t^{\text{incorrect}}$. The splitting allows training models with full data but evaluating them over decomposed data X_t^{correct} and $X_t^{\text{incorrect}}$ separately. As Fig. 3 shows, for vanilla self-training, the mIoU of X_t^{correct} increases stably in the left graph while the mIoU of $X_t^{\text{incorrect}}$ increases at the early stage but decreases gradually as shown in the right graph. This shows that the overall performance degradation at the later training stage mainly comes from $X_t^{\text{incorrect}}$, indicating that vanilla self-training learns useful information to generate correct predictions for $X_t^{\text{incorrect}}$ at the early training stage but tends to forget these information at a later training stage. Differently, BiMem builds comprehensive and robust memorization that memorizes and calibrates useful and representative information on the fly, leading to stabler black-box UDA without performance degradation and training collapse.

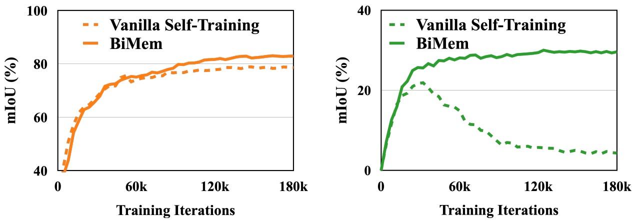

Figure 3: Analysis of the 'forgetting' issue in black-box UDA. We split the target data into two portions according to their initial pseudo-label predictions Y_t by black-box predictor. On the target data whose initial pseudo-label predictions are correct in the left graph, the vanilla self-training (trained using all target data) performs consistently well along the training process. But on the target data whose initial pseudo-label predictions are incorrect in the right graph, the vanilla self-training performs well at the starting stage and then collapses with training moving on, indicating that the vanilla self-training gradually forgets the earlier learnt useful knowledge. As a comparison, the proposed BiMem can remember the learnt useful knowledge and performs consistently well for both portions of target data.

We provide more discussions, theoretical insights and qualitative analysis in the appendix.

5. Conclusion

This paper presents BiMem, a general black-box UDA framework that works well for various visual recognition tasks. BiMem constructs three types of memory that interact in a bi-directional manner with a forward memorization flow and a backward calibration flow, resulting in comprehensive yet robust memorization that captures useful and representative information during black-box adaptation. In this way, BiMem enables to effectively calibrate the noisy pseudo labels conditioned on the memorized information, mitigating the 'forgetting' issue in black-box UDA and leading to stable and effective adaptation. Extensive experiments over multiple benchmarks show that BiMem achieves superior black-box UDA performance consistently across various vision tasks including classification, segmentation, and detection. Moving forwards, we plan to further extend our BiMem to other vision tasks such as pose estimation and person re-identification, and investigate other usages of BiMem in addition to label calibration.

Acknowledgement. This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15384–15394, 2021. 2

[2] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019. 7

[3] Richard C Atkinson and Richard M Shiffrin. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, volume 2, pages 89–195. Elsevier, 1968. 2, 3

[4] Eden Belouadah and Adrian Popescu. Il2m: Class incremental learning with dual memory. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 583–592, 2019. 3

[5] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019. 6

[6] Lin Chen, Huaian Chen, Zhixiang Wei, Xin Jin, Xiao Tan, Yi Jin, and Enhong Chen. Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7181–7190, 2022. 2

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 6

[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. 1, 2

[9] 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. 6

[10] Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7212–7222, 2022. 2

[11] Francois Fleuret et al. Uncertainty reduction for model adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9613–9623, 2021. 2, 6, 7

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 1, 2

[13] Huan Gao, Jichang Guo, Guoli Wang, and Qian Zhang. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9913–9923, 2022. 2

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[15] 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. 3

[16] 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. 6, 7

[17] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6668–6677, 2019. 2

[18] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022. 2

[19] Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, and Rong Jin. Learning position and target consistency for memory-based video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4144–4154, 2021. 3

[20] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021. 2

[21] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 3, 6, 7

[22] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Rda: Robust domain adaptation via fourier adversarial attacking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8988–8999, 2021. 2

[23] Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu, and Ling Shao. Category contrast for unsupervised domain adaptation in visual tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1203–1214, 2022. 2

[24] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In *European Conference on Computer Vision*, pages 705–722. Springer, 2020. 2

[25] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object de-

tection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 1, 6, 7

[26] Tarun Kalluri, Astuti Sharma, and Manmohan Chandraker. Memsac: Memory augmented sample consistency for large scale domain adaptation. *arXiv preprint arXiv:2207.12389*, 2022. 3

[27] Jung Uk Kim, Sungjune Park, and Yong Man Ro. Robust small-scale pedestrian detection with cued recall via memory learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3050–3059, 2021. 3

[28] 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 International Conference on Computer Vision*, pages 6092–6101, 2019. 6, 7

[29] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 3

[30] Byungsoo Ko, Geonmo Gu, and Han-Gyu Kim. Learning with memory-based virtual classes for deep metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11792–11801, 2021. 3

[31] 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. 2

[32] Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 615–625, 2021. 2

[33] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 3

[34] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 1, 2, 7

[35] Jonghyun Lee, Dahuin Jung, Junho Yim, and Sungroh Yoon. Confidence score for source-free unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 12365–12377. PMLR, 2022. 2

[36] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 481–497. Springer, 2020. 2

[37] 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. 2

[38] Ruihuang Li, Shuai Li, Chenhang He, Yabin Zhang, Xu Jia, and Lei Zhang. Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11593–11603, 2022. 2

[39] Shuaifeng Li, Mao Ye, Xiatian Zhu, Lihua Zhou, and Lin Xiong. Source-free object detection by learning to overlook domain style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8014–8023, 2022. 6, 7

[40] 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*, volume 35, pages 8474–8481, 2021. 6, 7

[41] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 3

[42] 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. 2, 7

[43] 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*, pages 8003–8013, 2022. 1, 2, 3, 4, 6, 7

[44] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2, 3

[45] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018. 2

[46] Yaoyao Liu, Bernt Schiele, and Qianru Sun. Rmm: Reinforced memory management for class-incremental learning. *Advances in Neural Information Processing Systems*, 34:3478–3490, 2021. 3

[47] Yong Liu, Ran Yu, Fei Yin, Xinyuan Zhao, Wei Zhao, Weihao Xia, and Yujiu Yang. Learning quality-aware dynamic memory for video object segmentation. *arXiv preprint arXiv:2207.07922*, 2022. 3

[48] 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. 2, 6, 7

[49] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12435–12445, 2021. 2

[50] M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14765–14775, 2022. 2

[51] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14372–14381, 2020. 3

[52] Pedro O Pinheiro. Unsupervised domain adaptation with similarity learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8004–8013, 2018. 1

[53] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 6

[54] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 6

[55] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 6

[56] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8050–8058, 2019. 1

[57] 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. 1, 2

[58] 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. 1

[59] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. 6

[60] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016. 3

[61] Larry R Squire, Lisa Genzel, John T Wixted, and Richard G Morris. Memory consolidation. *Cold Spring Harbor perspectives in biology*, 7(8):a021766, 2015. 5

[62] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 3

[63] Jiayi Tian, Jing Zhang, Wen Li, and Dong Xu. Vdm-da: Virtual domain modeling for source data-free domain adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 2

[64] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 6

[65] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 6

[66] Vibashan VS, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4516–4526, 2021. 3

[67] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 1, 2

[68] Fan Wang, Zhongyi Han, Yongshun Gong, and Yilong Yin. Exploring domain-invariant parameters for source free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2022. 2

[69] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6388–6397, 2020. 3

[70] Yuxi Wang, Jian Liang, and Zhaoxiang Zhang. Source data-free cross-domain semantic segmentation: Align, teach and propagate. *arXiv preprint arXiv:2106.11653*, 2021. 3, 6

[71] Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, and Philip S Yu. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9154–9162, 2019. 3

[72] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014. 3

[73] 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. 2

[74] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, and Xinjing Cheng. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8068–8078, 2022. 2

[75] 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. 1, 2

[76] Jinyu Yang, Jingjing Liu, Ning Xu, and Junzhou Huang. Tvt: Transferable vision transformer for unsupervised do-

main adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 520–530, 2023. 2

[77] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8978–8987, 2021. 2

[78] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2

[79] Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, and Guang Chen. Unsupervised domain adaptation for nighttime aerial tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8896–8905, 2022. 2

[80] Hao-Wei Yeh, Baoyao Yang, Pong C Yuen, and Tatsuya Harada. Sofa: Source-data-free feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 474–483, 2021. 2

[81] Fuxun Yu, Di Wang, Yinpeng Chen, Nikolaos Karianakis, Tong Shen, Pei Yu, Dimitrios Lymberopoulos, Sidi Lu, Weisong Shi, and Xiang Chen. Sc-uda: Style and content gaps aware unsupervised domain adaptation for object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 382–391, 2022. 2

[82] Haojian Zhang, Yabin Zhang, Kui Jia, and Lei Zhang. Unsupervised domain adaptation of black-box source models. *arXiv preprint arXiv:2101.02839*, 2021. 7

[83] Jingyi Zhang, Jiaxing Huang, Zhipeng Luo, Gongjie Zhang, Xiaoqin Zhang, and Shijian Lu. Da-detr: Domain adaptive detection transformer with information fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23787–23798, 2023. 2

[84] Jingyi Zhang, Jiaxing Huang, Zichen Tian, and Shijian Lu. Spectral unsupervised domain adaptation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9829–9840, 2022. 2

[85] Jingyi Zhang, Jiaxing Huang, Xiaoqin Zhang, and Shijian Lu. Unidaformer: Unified domain adaptive panoptic segmentation transformer via hierarchical mask calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11227–11237, 2023. 2

[86] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12414–12424, 2021. 2

[87] Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12425–12434, 2021. 2

[88] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13766–13775, 2020. 1, 2

[89] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 751–766, 2018. 3

[90] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 6

[91] Chenfan Zhuang, Xintong Han, Weilin Huang, and Matthew Scott. ifan: Image-instance full alignment networks for adaptive object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13122–13129, 2020. 2

[92] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 1, 2, 6, 7

[93] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5982–5991, 2019. 1, 2, 6, 7