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, 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.

However, the black-box predictions of the target data are prone to errors due to the cross-domain discrepancy, lead-
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.
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 for Black-box UDA.
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 illustration, which involves a target domain \( D_t = \{ X_t, Y_t \} \), where the pseudo label \( Y_t \) 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 \( G_t \) 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 \( D_s = \{ X_s, Y_s \} \) by a supervised loss:

\[
L_{sup} = l(G_s(X_s), Y_s),
\]

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 \( x_t = \{x^k_t\}_{k=1}^K \) \((x^k_t \in X_t)\) and the corresponding source-predicted labels \( \hat{y}_t = \{ \hat{y}^k_t \}_{k=1}^K \) \((\hat{y}^k_t \in Y_t)\) as an example. We elaborate the memory construction and update in Forward Memorization, and memory calibration in Backward Calibration.

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 \( \gamma \) is a momentum coefficient) to acquire the momentum features \( \hat{f}_{sm} = \{ \hat{f}^k_{sm} \}_{k=1}^K \) and the corresponding category predictions \( \{ \hat{p}^c_{sm} \}_{c=1}^C \) \( \{ \hat{p}^c_{sm} \}_{c=1}^C \) \( \kappa_{c=1}^C \), 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:

\[
\mathcal{H}(f^k_{sm}) = - \sum_{c=1}^C p^{(k,c)}_{sm} \log p^{(k,c)}_{sm},
\]

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

---

**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 \( G \)

1: Initialize \( G \) and \( G^m \)
2: for \( iter = 1 \) to \( \text{Max}_\text{Iter} \) do
3: Update the momentum model: \( G^m \leftarrow G \)
4: Sample \( \{ x_t, \hat{y}_t \} \subset \{ X_t, Y_t \} \) and encode \( x_t: f_{sm} = G^m(x_t) \)
5: **Forward Memorization Flow:**
6: Update sensory memory: replace old features \( f^*_{sm} \) by \( f_{sm} \)
7: Update short-term memory: dequeue oldest features \( f^{k-1}_{sm} \) and enqueue new features selected by Eq. 2
8: Update long-term memory: accumulate \( f^*_{sm} \) and \( f^{s*}_{sm} \) by Eqs. 3-4
9: **Backward Calibration Flow:**
10: Calibrate short-term memory by long-term memory as in Eqs. 5-6
11: Correct sensory memory by long-term memory and the calibrated short-term memory as in Eq. 7
12: **Network Training:**
13: Acquire \( \theta_G \) by denoising \( \hat{y}_t \) with sensory memory by Eq. 8
14: Optimize target model \( G \) with \( \{ x_t, \hat{y}_t \} \) by Eq. 9
15: end for
16: return \( G \)
Short-term memory works as a FIFO queue with a fixed size of $M$ ($M > N$), where $N$ oldest features $f_{st}^m$ 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 long-term memory for efficient feature update in the long-term memory. Generally, if the short-term feature $f_{st}^m$ is far from the $c$-th long-term feature $\delta^c$, 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:

$$w^{(m,c)} = \text{Softmax}(-||f_{st}^m - \delta^c||_1),$$  

where $w^{(m,c)} = \{w^{(m,c)}\}_{m=1}^M$ 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 $\delta_{lt}$ and the short-term feature centroids $\delta_{st}$, we assign a new category probability to each sensory feature $f_{sm}^k$ as the following:

$$\overline{p}_{sm}^{(k,c)} = \text{Softmax}([-||f_{sm}^k - \delta_{lt}^c||_1] + (-||f_{sm}^k - \delta_{st}^c||_1)],$$  

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 $\overline{y}_t$ of each sample in $\{x_{st}^k\}_{k=1}^K$ as shown in the bottom part of Fig. 2. For sample $x_{st}^k$, we read out its calibrated category-wise probabilities $p_{sm}^{(k,c)}$ (acquired by Eq. 7) from sensory memory. Next, the pseudo label $\overline{y}_t^k$ of $x_{st}^k$ is denoised by re-weighting its category-wise probability $\overline{p}_{st}$:

$$\overline{y}_t^k = \argmax_c (p_{sm}^{(k,c)} \odot \overline{p}_{st}^c),$$  

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:

$$L_{\text{self}} = l(G(x_t), \overline{y}_t),$$  

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 $L_{\text{self}}$, i.e., $\arg \min_G L_{\text{self}}$, as detailed in Algorithm 1.
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 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-50 [16] as the classification network as in [5, 16].
Table 6: Experiments on image classification over black-box UDA task Office-Home.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A → D</td>
<td>79.9</td>
<td>88.8</td>
<td>89.4</td>
<td>86.5</td>
<td>92.2</td>
<td>91.6</td>
<td>92.8</td>
</tr>
<tr>
<td>A → W</td>
<td>76.6</td>
<td>85.5</td>
<td>86.8</td>
<td>83.1</td>
<td>87.2</td>
<td>86.8</td>
<td>88.2</td>
</tr>
<tr>
<td>D → A</td>
<td>56.4</td>
<td>64.6</td>
<td>65.1</td>
<td>66.1</td>
<td>72.2</td>
<td>73.9</td>
<td>73.9</td>
</tr>
<tr>
<td>D → W</td>
<td>92.8</td>
<td>95.1</td>
<td>94.8</td>
<td>95.1</td>
<td>96.2</td>
<td>96.8</td>
<td>96.8</td>
</tr>
<tr>
<td>W → A</td>
<td>60.9</td>
<td>66.7</td>
<td>67.1</td>
<td>68.9</td>
<td>73.3</td>
<td>98.6</td>
<td>99.4</td>
</tr>
<tr>
<td>W → D</td>
<td>98.5</td>
<td>97.8</td>
<td>98.7</td>
<td>98.1</td>
<td>98.6</td>
<td>86.4</td>
<td>87.7</td>
</tr>
<tr>
<td>Avg.</td>
<td>77.5</td>
<td>83.2</td>
<td>83.6</td>
<td>83.0</td>
<td>84.1</td>
<td>86.4</td>
<td></td>
</tr>
</tbody>
</table>

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

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
pseudolabels. 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 compactifies 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 pseudolabels 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 pseudolabels $Y_t$ (predicted by the black-box predictor). This produces target data with correct initial pseudolabel ($i.e., X_t^{\text{correct}} = X_t[\bar{Y}_t = Y_t]$) and target data with incorrect initial pseudolabels ($i.e., X_t^{\text{incorrect}} = X_t[\bar{Y}_t \neq Y_t]$), 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 $X_t^{\text{correct}}$ and $X_t^{\text{incorrect}}$ separately. As Fig. 3 shows, for vanilla self-training, the mIoU of $X_t^{\text{correct}}$ 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.

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 pseudolabels 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


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


[75] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen. Tvt: Transferable vision transformer for unsupervised do-


