IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

Predicting Class Distribution Shift for Reliable Domain Adaptive Object Detection

Nicolas Harvey Chapman[1], Feras Dayoub[2], Will Browne[1] and Christopher Lehnert[1]

***Abstract*—Unsupervised Domain Adaptive Object Detection (UDA-OD) uses unlabelled data to improve the reliability of robotic vision systems in open-world environments. Previous approaches to UDA-OD based on self-training have been effective in overcoming changes in the general appearance of images. However, shifts in a robot's deployment environment can also impact the likelihood that different objects will occur, termed class distribution shift. Motivated by this, we propose a framework for explicitly addressing class distribution shift to improve pseudo-label reliability in self-training. Our approach uses the domain invariance and contextual understanding of a pre-trained joint vision and language model to predict the class distribution of unlabelled data. By aligning the class distribution of pseudo-labels with this prediction, we provide weak supervision of pseudo-label accuracy. To further account for low quality pseudo-labels early in self-training, we propose an approach to dynamically adjust the number of pseudo-labels per image based on model confidence. Our method outperforms state-of-the-art approaches on several benchmarks, including a 4.7 mAP improvement when facing challenging class distribution shift. Code available at https://github.com/nhcha6/ClassDistributionPrediction**

***Index Terms*—Object Detection, Deep Learning for Visual Perception; Visual Learning**

I. Introduction

OBJECT detection is a crucial component of many robotic systems, from self-driving cars to service robots. Existing object detectors based on deep learning require the collection of large, annotated datasets for training. However, in open-world deployment a robot will encounter changes in object appearance due to factors such as weather, lighting conditions, or image corruptions [1]. Furthermore, shifts in a robot's deployment environment can impact the likelihood that different objects will occur, termed class distribution shift [2]. Due to the high costs of annotation, it is infeasible to gather labelled data for all potential conditions and environments [3]. Thus, it is inevitable that an object detector deployed on a robot will face the problem of domain shift, where the images being processed do not match those used for training. Unfortunately, the performance of deep learning-based object detectors degrades significantly when facing such domain shift [1]. To address this issue, Unsupervised Domain Adaptive Object Detection (UDA-OD) has been proposed to adapt a model from a known source domain to a shifted target domain using only unlabelled data. This strategy can help the model generalise and improve its performance in the target domain, without the need for expensive labelling.

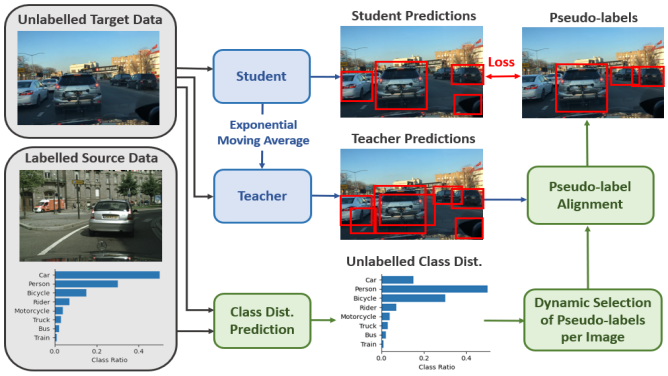

Fig. 1. Our framework for explicitly incorporating class distribution shift into self-training to improve pseudo-label reliability. As per the standard implementation of Mean Teacher, confident detections from a teacher model are used as pseudo-labels to train a student model using unlabelled data. The Exponential Moving Average (EMA) of the weights of the student are then used to update the teacher to make it more stable during training. Traditionally, a static confidence threshold is defined for all classes to generate pseudo-labels. Our method (shown in green) instead predicts the class distribution of the unlabelled data, and selects confidence thresholds to align the class distribution of the pseudo-labels with this prediction. To further address the poor performance of the teacher model in the target domain, we dynamically adjust the number of pseudo-labels per image as teacher confidence increases.

Self-training methods have recently produced State-Of-The-Art (SOTA) results on UDA-OD benchmarks [4], [5]. Such approaches leverage the Mean Teacher framework [6] to enforce consistency between a student and teacher model on unlabelled images (Section III-B). Key to this framework is the use of confident detections from the teacher as pseudo-labels for training the student. These methods were initially used for Semi-Supervised Object Detection (SSOD) [7], [8], where the labelled and unlabelled data come from the same distribution. Thus, when applied to a challenging domain adaptation problem, the reliability of pseudo-labels generated by the teacher may degrade significantly [5]. Focusing on this challenge, recent work has adapted Mean Teacher to UDA-OD by improving pseudo-label reliability in the presence of domain shift [4], [5], [9], [9]–[11].

While promising, existing UDA-OD methods and benchmarks focus largely on changes to the appearance of images [1]. Thus, the benefit of explicitly addressing class distribution shift during self-training is unexplored. Furthermore, in robotics there are opportunities to use contextual cues to generate a prediction for the likelihood of object occurrence. For example, one expects an autonomous vehicle to encounter more pedestrians on city streets than on the highway [2]. Motivated by this, we propose a framework for explicitly

Manuscript received: February 5, 2023; Revised May 8, 2023; Accepted June 14, 2023. This paper was recommended for publication by Editor Markus Vincze upon evaluation of the Associate Editor and Reviewers' comments.

[1]Nicolas Harvey Chapman, Will Browne and Christopher Lehnert are with the School of Electrical Engineering and Robotics, Queensland University of Technology, Brisbane, Australia (will.browne@qut.edu.au; c.lehnert@qut.edu.au; nicolasharvey.chapman@hdr.qut.edu.au).

[2]Feras Dayoub is with the School of Computer Science and the Australian Institute of Machine Learning at the University of Adelaide, Adelaide, Australia (feras.dayoub@adelaide.edu.au).

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

addressing class distribution shift during self-training. We find that our approach can significantly increase the reliability of pseudo-labels produced by Mean Teacher, leading to improved domain adaptation.

Our framework involves firstly predicting the class distribution of unlabelled data in novel deployment environments. Due to their strong understanding of image context and resistance to domain shift, we find that pre-trained joint vision and language models are useful for this task [12]. We then use Adapative label distribution aware Confidence Thresholding (ACT) [13] (Section III-C) to incorporate our prediction into Mean Teacher. Proposed for SSOD, ACT improves performance on low probability classes and provides weak supervision of the teacher model by aligning the distribution of pseudo-labels with that of the labelled data. By using our prediction instead of the labelled prior, the reliability of pseudo-labels used in Mean Teacher improves substantially. However, even with a perfectly accurate class distribution, we find that the pseudo-labels generated with ACT remain unreliable early in self-training. To address this, we propose an approach for dynamically adjusting the number of pseudo-labels per image based on the confidence of the teacher model. Extensive experiments show that our proposed method returns SOTA performance on several benchmarks, including a 4.3 mAP improvement when adapting from a small to large scale dataset. On a novel scenario containing more challenging class distribution shift, we return a 4.7 mAP improvement.

To summarise, this letter makes the following contributions:

- A framework for explicitly addressing class distribution shift during self-training, leading to an improvement in pseudo-label reliability.
- A method to predict the class distribution of unlabelled data using a pre-trained joint vision and language model.
- An approach for dynamically adjusting the number of pseudo-labels per image to account for the confidence of the teacher in the target domain.
- Experimental results showing that the proposed method returns SOTA performance on several benchmarks, including scenarios with realistic class distribution shift.

II. Related Work

A. Self-training in Unsupervised Domain Adaptive Object Detection

Initially proposed for semi-supervised learning [6], self-training methods have recently produced SOTA results on UDA-OD benchmarks [4], [5]. Several augmentations have been made to the standard Mean Teacher framework [6] to improve pseudo-label reliability under domain shift. Methods have been proposed for merging patches from labelled and unlabelled images [9], selecting optimal pseudo-labels to balance true positive and false negative detections [11] and performing style transfer between the source and target domains [5]. Recognising that pseudo-labels are inevitably unreliable, Chen *et al.* [4] propose a probabilistic self-training framework that does not require confidence thresholds [14]. Instead, they implement an entropy focal loss that encourages the student to pay more attention to high certainty detections. While promising, these methods fail to leverage the rich contextual information available during robotic deployment and focus on shifts in the general appearance of images. Consequently, the potential for contextual and class distribution shift is largely ignored. Motivated by this, Xu *et al.* [15] enforce consistency between the object detector and an image-level multilabel classifier, which is more robust to arbitrary changes in image background. However, this method was proposed for use in domain alignment instead of self-training. Cai *et al.* [10] model the relationships between objects in a scene using a graph structure, and enforces consistency between student and teacher graphs. While incorporating contextual information improves robustness, the pseudo-labels and relational graphs generated using arbitrary thresholds remain unreliable. Furthermore, these methods do not consider the impact of class distribution shift on self-training. In response, we model image context to predict the class distribution of unlabelled data in novel deployment environments. This prior can be incorporated into self-training via ACT, which generates pseudo-labels to match the predicted class distribution [13]. This method has shown promising results in SSOD, and we propose a series of changes to optimise it for domain adaptation.

B. Class Imbalance in Self-training

It is well established that class imbalance leads to severe confirmation bias during self-training, as dominant classes are predicted with high confidence and subsequently reinforced [8]. Solutions to this problem have been proposed for SSOD, such as the weighting of low probability classes [8], [14] or ACT to align the distribution of pseudo-labels with a known class distribution [13]. However, existing methods fail to address the class imbalance problem in the presence of domain shift.

C. Class Distribution Shift in UDA-OD

Existing UDA-OD benchmarks often focus on appearance changes due to factors such as weather, lighting conditions, or image corruptions [1], but do not adequately assess realistic class distribution shifts. Ignoring class distribution shift is a critical limitation, as they are standard in open-world robotic deployment and can significantly impact the performance of object detection algorithms [2], [16], [17]. In particular, the Cityscapes [18] to Foggy Cityscapes [19] adaptation scenario contains no class distribution shift as the target domain is an augmented version of the source images [1]. Similarly, the cross camera adaptation and simulation to real scenarios only contain a single class [1]. The final benchmark studies adaptation from a small to large-scale dataset [20], reflecting the challenge of deploying a robot in a new environment with a broad array of conditions. We find that the shift in location produces significant class distribution shift, making it useful for this work. However, the impact of class distribution shift on UDA-OD is poorly evaluated using existing benchmarks. In turn, the benefits of explicitly addressing this problem during self-training are largely unexplored.

D. Pre-trained Joint Vision and Language Models

The abundance of captioned images has allowed the learning of transferable image representations using natural language

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

supervision. One notable example is Contrastive Language-Image Pretraining (CLIP), which trains a model to predict the images and text that occur together using 400 million image-text pairs [12]. Because natural language descriptions are highly generalisable, CLIP can be easily transferred to new tasks and exhibits superior resistance to domain shift compared to supervised pre-training methods [12], [21]–[23]. Language descriptions also contain a wealth of contextual information [24]–[26], making them useful for modelling the deployment environment of a robot. Recent work proposes using the similarity between an image and a series of text prompts produced by CLIP as a representation of image context [23]. This representation is a strong predictor of object occurrence in an image, helping to improve semantic segmentation performance in open-world settings. In this work, we leverage the domain invariance and contextual detail of these image-text similarity scores to predict the class distribution of unlabelled data.

III. Preliminaries

A. Problem Formulation

In UDA-OD, a set of n_l labelled images $D_l = \{X_l, Y_l\}$ containing n_c classes from the the source domain and n_u unlabelled images $D_u = \{X_u\}$ from the target domain are provided. The goal of UDA-OD is to use both the labelled source and unlabelled target data to optimise performance of the object detector in the target domain.

B. Mean Teacher

The Mean Teacher framework [6] encourages consistent predictions by a student and teacher model on augmented versions of the unlabelled data. This process leads the student and teacher to learn a solution from the source domain that is transferable to the target domain. An overview of this method is provided in Figure 1. The student and teacher models share the same network architecture, but different weights θ^s and θ^t respectively. For a given unlabelled sample x_u, a weakly augmented version x_u^t is created for input into the teacher model by applying horizontal flipping and multi-scaling [13]. After non-maximum suppression (NMS), a confidence threshold is applied to the teacher's predictions $f(x_u^t, \theta^t)$ to generate pseudo-labels y_u. Strong augmentations including color jittering, grayscale, gaussian blurring and cutout patches [13] are applied to create x_u^s, which is used to generate the student's predictions $f(x_u^s, \theta^s)$. In turn, an unsupervised loss can be calculated as:

$$\mathcal{L}_u = \mathcal{L}_{cls}(f(x_u^s, \theta^s), y_u) + \mathcal{L}_{reg}(f(x_u^s, \theta^s), y_u) \tag{1}$$

where $\mathcal{L}_{cls}$ is the classification loss of the object detector, and $\mathcal{L}_{reg}$ is the box regression loss. A standard supervised loss of the same form is calculated using a labelled sample (x_l, y_l) from the source domain:

$$\mathcal{L}_l = \mathcal{L}_{cls}(f(x_l^s, \theta^s), y_l) + \mathcal{L}_{reg}(f(x_l^s, \theta^s), y_l) \tag{2}$$

The overall loss for a batch containing both labelled and unlabelled samples is then calculated as:

$$\mathcal{L} = \mathcal{L}_l + \lambda \mathcal{L}_u \tag{3}$$

where λ is a hyperparameter to weight the unsupervised loss. This loss is used to update the student via gradient descent, and the exponential moving average (EMA) of the student's weights are used to update the teacher model. That is, after each training iteration t the weights of the teacher model are calculated as:

$$\theta_t^t = \alpha \theta_{t-1}^t + (1 - \alpha)\theta_s \tag{4}$$

where α controls the stability of the teacher model.

C. Adaptive Label Distribution-aware Confidence Thresholding (ACT)

ACT aims to select confidence thresholds for Mean Teacher to overcome the class imbalance problem. It does so by aligning the class distribution of the pseudo-labels with that of the labelled data [13]. The method firstly calculates the class ratio of the labelled data $\boldsymbol{r^l} = [r_1^l, ..., r_{n_c}^l]$ and the number of objects per image n_o^l. Using this class distribution prior, the expected number of objects in the unlabelled data for each class c is calculated as:

$$n_c^u = r_c^l \cdot n_o^l \cdot n_u \tag{5}$$

The confidence threshold for class c is then selected such that n_c^u pseudo-labels are generated. That is:

$$t_c = \boldsymbol{p_c^{sort}}[n_c^u] \tag{6}$$

where $\boldsymbol{p_c^{sort}}$ is a list of the confidence scores predicted for class c by the teacher, sorted in descending order. ACT can be used with any self-training approach based on Mean Teacher, but specific changes to the standard implementation are recommended by the authors [13]. A proportion of the pseudo-labels are defined as reliable and used to calculate the hard classification and box regression losses defined in (1). To deal with the noise present in the unreliable pseudo-labels, a soft classification loss is calculated using the cross-entropy between student and teacher classification predictions [13]. This results in a reformulation of (1) as:

$$\begin{aligned}\mathcal{L}_u = \mathcal{L}_{cls}(f(x_u^s, \theta^s), y_{ur}) + \mathcal{L}_{reg}(f(x_u^s, \theta^s), y_{ur}) \\ + \widehat{\mathcal{L}}_{cls}(f(x_u^s, \theta^s), y_{uu})\end{aligned} \tag{7}$$

where y_{ur} and y_{uu} refer to the reliable and unreliable pseudo-labels, and $\widehat{\mathcal{L}}_{cls}$ denotes the soft classification loss.

D. Contrasting Language-Image Pretraining (CLIP)

Given a batch of n images X_i and associated texts X_t, CLIP is trained to predict which possible image-text pairs $X_i \times X_t$ actually occurred. Images and texts are input to an image encoder I and text encoder T to extract their respective embeddings. These embeddings are then projected into a shared, multi-modal representation space with learnt projection matrices W_i and W_t. L2-normalisation is then applied to extract the final multi-modal embeddings $Z_i \in \mathbb{R}^{n \times d}$ and $Z_t \in \mathbb{R}^{n \times d}$:

$$Z_i = \|I(X_i) \cdot W_i\|_2 \tag{8}$$

$$Z_t = \|T(X_t) \cdot W_t\|_2 \tag{9}$$

The cosine similarity of the shared embeddings is then calculated, and multiplied by a learnt temperature parameter t to calculate similarity scores $S \in \mathbb{R}^{n \times n}$ for each text and image pair in the batch:

$$S = (Z_i \cdot Z_t^T) * exp(t) \tag{10}$$

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

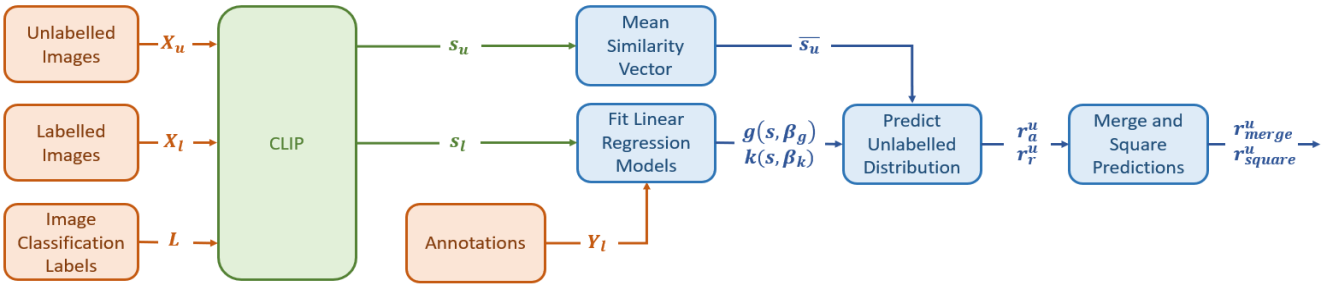

Fig. 2. Our proposed method for predicting the class ratio of the unlabelled data. CLIP is used to calculate the similarity between the labelled images X_l and a series of natural language labels L of the form "a photo of class c". Using the labelled similarity vector $\boldsymbol{s_l}$ as a domain invariant representation of semantic context, two linear regression models are fit to predict the number of instances $g(\boldsymbol{s}, \beta_g)$ and the class ratio $k(\boldsymbol{s}, \beta_k)$ in each labelled image. To make a prediction for the class ratio of the unlabelled images X_u, CLIP is used to extract the similarity vectors $\boldsymbol{s_u}$. The mean similarity vector $\overline{\boldsymbol{s_u}}$ is then calculated and input to the linear regression models to generate two distinct predictions for the class ratio of the entire unlabelled dataset. These predictions are merged by calculating the geometric mean, and the relative change in class ratio squared to account for persistent underestimation.

During training, a cross-entropy loss is applied to the similarity scores to encourage similarity between true image-text pairs and discourage similarity between false pairs. Our work uses the pretrained model to generate similarity scores between images and image classification labels.

IV. Method

A. Class Distribution Prediction

We aim to predict the class ratio of the unlabelled data $\boldsymbol{r^u} \in \mathbb{R}^{n_c}$, such that it can be used as prior for performing ACT. An overview of this approach can be found in Figure 2. We start by using CLIP to generate a representation of the semantic context of each image. To do so, we calculate the similarity between the labelled images X_l and a series of natural language prompts L. The performance of vision-language models on downstream tasks is sensitive to the format of natural language labels, leading to the development of prompt optimisation techniques [25], [27]. We implement a simple approach to generate L where each class label c is converted to a prompt of the form "a photo of c" [12]. For each labelled image, this results in a vector of similarity scores $\boldsymbol{s_l} \in \mathbb{R}^{n_c}$ that characterise how the image relates to the text prompts [23]. As the natural language descriptions used to generate the similarity vectors are resistant to domain shift [21], [22], this results in a contextual representation that is consistent across the source and target data.

We investigate two linear regression models, g and k, for predicting the class ratio given the similarity vector $\boldsymbol{s}$ as a domain invariant input. The absolute model $g(\boldsymbol{s}, \beta_g)$ is optimised using the labelled data to predict the absolute number of instances of each class in an image. By normalising the output of $g(\boldsymbol{s}, \beta_g)$ to sum to 1, this model can be used to predict the class ratio. The relative model $k(\boldsymbol{s}, \beta_k)$ is optimised using the labelled data to directly predict the ratio of each class in an image. The relative model does not consider how many objects occur in an image, instead learning how likely a class is to occur relative to other classes.

To make a prediction for the class distribution of the unlabelled dataset, we use CLIP to extract the similarity vector s_u for each unlabelled image. We then calculate the mean similarity vector across the unlabelled dataset $\overline{\boldsymbol{s_u}} \in \mathbb{R}^{n_c}$ and input it to the linear regression models:

$$\boldsymbol{r_a^u} = \frac{g(\overline{\boldsymbol{s_u}}, \beta_g)}{|g(\overline{\boldsymbol{s_u}}, \beta_g)|} \tag{11}$$

$$\boldsymbol{r_r^u} = k(\overline{\boldsymbol{s_u}}, \beta_k) \tag{12}$$

The relative and absolute models produce distinct predictions $\boldsymbol{r_a^u} \in \mathbb{R}^{n_c}$ and $\boldsymbol{r_r^u} \in \mathbb{R}^{n_c}$ for the class ratio, either of which can be used in our framework. However, neither prediction is consistently more accurate than the other, and the optimal prediction to use for a specific scenario cannot be determined without labelled data. To ensure consistent performance across all scenarios, we therefore average the output of the two models. To do so, we propose using the geometric mean, as it is more appropriate than the arithmetic mean for summarising the central tendency of ratios. We term the element-wise geometric mean of the class ratio predictions $\boldsymbol{r_{merge}^u} \in \mathbb{R}^{n_c}$ as the merged prediction:

$$\boldsymbol{r_{merge}^u} = \sqrt{\boldsymbol{r_a^u} \cdot \boldsymbol{r_r^u}} \tag{13}$$

Secondly, we found that the merged prediction persistently underestimates the shift from the labelled class ratio $\boldsymbol{r^l}$ to the unlabelled class ratio $\boldsymbol{r^u}$. A heuristic solution to this underestimation problem is to square the relative change in the class ratio predicted by $\boldsymbol{r_{merge}^u}$. Formally, we express $\boldsymbol{r_{merge}^u}$ relative to $\boldsymbol{r^l}$ and calculate the squared prediction $\boldsymbol{r_{square}^u}$ for the class ratio as:

$$\boldsymbol{r_{square}^u} = \boldsymbol{r^l} \left(\frac{\boldsymbol{r_{merge}^u}}{\boldsymbol{r^l}} \right)^2 = \frac{\boldsymbol{r_a^u} \cdot \boldsymbol{r_r^u}}{\boldsymbol{r^l}} \tag{14}$$

Finally, we apply L1 normalisation to $\boldsymbol{r_{square}^u}$ so the distribution sums to 1.

B. Dynamic Selection of Pseudo-labels per Image

Even with a perfectly accurate class distribution, we find that under significant domain shift the pseudo-labels generated by the teacher remain unreliable early in self-training. Thus, instead of estimating the average number of objects per image in the unlabelled data, we set this value to target a mean confidence score τ across the generated pseudo-labels. This accounts for the initial poor performance of the teacher in the target domain, and allows the number of pseudo-labels per image to increase as the model becomes more confident. We use Algorithm 1 to calculate the number of objects per image required to deliver a mean pseudo-label confidence score of τ. In this algorithm, the average number of objects per image

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

$n_o \in \mathbb{R}$ is iteratively increased by Δn_o. At each step, pseudo-labels y_u are generated for the entire unlabelled dataset by aligning them with the predicted class ratio $\boldsymbol{r^u}$ and the current number of objects per image n_o. The mean confidence score p_{mean} is then calculated across the pseudo-labels, and the loop terminates when this value falls below the target value τ.

Algorithm 1 Dynamic Selection of Objects per Image

$\boldsymbol{r^u}$ = PREDICTCLASSRATIO(D_l, D_u)
$n_o = 0$
while True **do**
  $n_o = n_o + \Delta n_o$
  y_u = ACT($\boldsymbol{r^u}$, n_o)
  p_{mean} = MEANCONFIDENCESCORE(y_u)
  if $p_{mean} < \tau$ **then**
    return $n_o - \Delta n_o$

V. EXPERIMENTS

A. Experimental Settings

Scenarios. We validate our proposed method using two scenarios previously used for UDA-OD, and one novel scenario that captures a more extreme class distribution shift.

- Adaptation from small to large dataset (C2B): as in previous work [4], [11], [15], we utilise Cityscapes [18] as a small source dataset and BDD100k daytime [20] images as the large and diverse target domain. The shift in location in this scenario produces significant class distribution shift.
- Adaptation from normal to foggy weather (C2F): this commonly studied scenario uses Cityscapes [18] as the source domain and Foggy Cityscapes [19] as the target domain. To compare with the optimal performance of the current state of the art [4], we use all three levels of fog.
- Adaptation from daytime to night (C2N): we propose a novel benchmark to explore the impact of more extreme class distribution shift on UDA-OD. We utilise Cityscapes [18] as the source domain and BDD100k night [20] as the target domain.

Network Architecture. A standard object detection architecture is used across all UDA-OD works to enable fair comparison between methods [1]. Following this trend, we use Faster-RCNN [28] as the detector architecture with VGG-16 [29] pre-trained on ImageNet [30] as the backbone. The inference rate of this network is quoted at 5Hz on a Nvidia K40 GPU [28], while our experiments return 20Hz on a Nvidia GeForce RTX 3090. However, our method could be easily extended to novel architectures, making it useful for diverse robotic applications [13].

Implementation Details. We follow the default implementation provided in [13] to test our method, which leverages MMDetection [31]. Thus, we use a batch size of 16 for the labelled and unlabelled data and a learning rate of 0.016. The optimiser used is SGD, with a momentum rate of 0.9 and a weight decay of 0.0001. For all experiments, the model is pre-trained using source data for 4,000 iterations before beginning self-training. For the C2B and C2N scenarios, we implement 42,000 iterations of self-training. Due to slower convergence on C2F, we self-train for 72,000 iterations. Lastly, we use a target mean confidence score τ of 0.6 and a step size Δn_o of 0.1 for selecting the number of objects per image.

TABLE I
RESULTS OF ADAPTATION FROM SMALL TO LARGE-SCALE DATASET (C2B) AND CITYSCAPES TO BDD100K NIGHT (C2N) SCENARIO. "SOURCE ONLY" AND "TARGET ONLY" REFER TO THE MODELS TRAINED BY ONLY USING LABELLED SOURCE DATA AND LABELLED TARGET DATA.

Methods	Reference	C2B (mAP)	C2N (mAP)
Source only [4]	-	20.6	3.8
ICR-CCR [15]	CVPR'20	26.9	-
SFOD [11]	AAAI'21	29.0	-
PT [4]	ICML'22	34.9	19.5
LabelMatch [13]	-	30.3	11.1
Ours	-	**39.2**	**24.2**
Target only [4]	-	51.7	32.7

Evaluation. Standard mean average precision (mAP) at an IOU threshold of 0.5 is used to compare performance with that reported by existing works. In additional to implementing our approach, we conduct the following evaluations:

- For comparison on the novel C2N scenario, we implement Probabilistic Teacher [4] as per the released code and report the peak performance recorded across 24,000 iterations. This method is the current state of the art on the C2B and C2F scenarios, and thus provides a strong baseline for evaluating our method under more extreme class distribution shift.
- We evaluate the original ACT [13] on all scenarios using the same implementation as our method.
- Lastly, for the C2N scenario, we evaluate the model trained for 48,000 iterations on the source data, and target data only.

B. Performance Comparison

Our method matches or outperforms existing approaches across all three scenarios (Table I, II). Strong performance occurs on scenarios with more extreme class distribution shift, a case that has been largely overlooked in existing benchmarks. We outperform the strongest baseline Probabilistic Teacher [4] by 4.3 mAP on the previously studied C2B scenario (Table I). We further return a 4.7 mAP improvement on the novel C2N scenario (Table I), which is more challenging due to extreme class distribution shift (Figure 5) and the discrete appearance change from daytime to night. A clear improvement is also noted on both scenarios relative to using the original ACT [13]. These results highlight that by modelling shifts in a robot's deployment context, pseudo-label reliability can be improved in Mean Teacher. Furthermore, it emphasises that class distribution shift must be considered when designing UDA-OD methods for robotics applications.

Interestingly, our approach remains competitive with existing methods on the normal to foggy scenario (C2F) (Table II), where there is a large change in image appearance but no class distribution shift. The average performance across five trials of our method is identical to the optimal performance returned by Probabilistic Teacher. The benefit of our approach is prominent in low probability classes such as train, truck

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

TABLE II
RESULTS OF ADAPTATION FROM NORMAL TO FOGGY WEATHER (C2F) [4]. DUE TO VARIABILITY IN TRAINING AND SIMILAR PERFORMANCE TO PROBABILISTIC TEACHER, WE RUN OUR METHOD FIVE TIMES AND REPORT THE MEAN MAP AND STANDARD DEVIATION. "0.01", "0.02" AND "ALL" IN THE COLUMN OF "SPLIT" REPRESENT THE LEVEL OF FOG INCLUDED IN THE TARGET DOMAIN. SEE TABLE I FOR DETAILS OF NAMING CONVENTIONS.

Methods	Split	Reference	truck	car	rider	person	train	motor	bicycle	bus	mAP
Source only [4]	ALL	-	12.1	40.4	33.4	27.9	10.1	20.7	30.9	23.2	24.8
MTOR [10]	0.02	CVPR'19	21.9	44.0	41.4	30.6	40.2	31.7	33.2	43.4	35.1
SW [32]	0.02	CVPR'19	24.5	43.5	42.3	29.9	32.6	30.0	35.3	32.6	34.3
DM [33]	UN	CVPR'19	27.2	40.5	40.5	30.8	34.5	28.4	32.3	38.4	34.6
PDA [34]	ALL	WACV'20	24.3	54.4	45.5	36.0	25.8	29.1	35.9	44.1	36.9
GPA [35]	0.01	CVPR'20	24.7	54.1	46.7	32.9	41.1	32.4	38.7	45.7	39.5
ATF [36]	UN	ECCV'20	23.7	50.0	47.0	34.6	38.7	33.4	38.8	43.3	38.7
HTCN [37]	0.02	CVPR'20	31.6	47.9	47.5	33.2	40.9	32.3	37.1	47.4	39.8
ICR-CCR [15]	ALL	CVPR'20	27.2	49.2	43.8	32.9	36.4	30.3	34.6	36.4	37.4
CF [38]	UN	CVPR'20	30.8	52.1	46.9	34.0	29.9	34.7	37.4	43.2	38.6
iFAN [39]	UN	AAAI'20	27.9	48.5	40.0	32.6	31.7	22.8	33.0	45.5	35.3
SFOD [11]	ALL	AAAI'21	25.5	44.5	40.7	33.2	22.2	28.4	34.1	39.0	33.5
MeGA [40]	UN	CVPR'21	25.4	52.4	49.0	37.7	**46.9**	34.5	39.0	49.2	41.8
UMT [5]	0.02	CVPR'21	34.1	48.6	46.7	33.0	46.8	30.4	37.3	56.5	41.7
PT [4]	ALL	ICML'22	33.4	**63.4**	**52.4**	**43.2**	37.8	**41.3**	**48.7**	56.6	**47.1**
LabelMatch [13]	ALL	-	33.4	61.3	49.8	38.7	27.4	32.7	44.5	55.8	43.0
Ours	ALL	-	**36.3±3.8**	61.2±0.7	49.2±2.3	41.9±0.9	44.2±5.9	37.3±2.1	47.7±0.5	**59.2±0.2**	**47.1±1.3**
Target only [4]	ALL	-	32.6	61.6	49.1	41.2	49.0	37.9	42.4	56.6	46.3

TABLE III
FINAL VALIDATION PERFORMANCE ON THE C2B SCENARIO WHEN USING ALTERNATIVE CLASS RATIOS AS THE PRIOR AND DIFFERENT APPROACHES TO SETTING THE NUMBER OF PSEUDO-LABELS PER IMAGE.

Obj. per Img.	Class Ratio	mAP
Labelled Data	Labelled Data	30.3
Dynamic (Ours)	Labelled Data	35.4
Dynamic (Ours)	Predicted (Ours)	39.2
Dynamic (Ours)	Unlabelled Data	40.2

TABLE IV
SENSITIVITY OF OUR APPROACH ON THE C2B SCENARIO TO DIFFERENT SETTINGS OF THE TARGET MEAN CONFIDENCE THRESHOLD τ.

τ	0.4	0.5	0.6	0.7	0.8
mAP	29.9	36.1	39.2	40.5	23.3

and bus, indicating that we are effectively mitigating the class imbalance problem. Furthermore, we return a 4.1 mAP improvement on this scenario relative to the original ACT, emphasising the benefit of accounting for pseudo-label reliability by dynamically updating the number of objects per image.

C. Impact on Pseudo-label Reliability

To further explore the proposed approach, we assess the pseudo-label reliability generated using alternative threshold selection techniques on the C2B scenario. To do so, we take the teacher model after 1000 iterations of self-training and calculate the F1 score of the generated pseudo-labels when different class ratios and number of objects per image are used (Figure 3a). This plot firstly highlights the benefit of using an accurate class distribution prediction to inform pseudo-label selection. When the predicted class ratio is used, there is a clear shift to higher F1 scores relative to using the naive prior provided by the labelled data. We further see that targeting a mean confidence score of 0.6 results in an F1 score that is close to the peak. Consequently, the proposed approach results in an improvement in the F1 score of 0.06 relative to the original ACT. We also plot how the selected number of objects per image evolves through training for the C2B and C2N scenarios, highlighting how the proposed method responds to the variable reliability of the teacher model (Figure 3b-c). As the pseudo-label accuracy improves with training, the number of objects per image is dynamically increased. We also see that for the more challenging C2N scenario, the number of pseudo-labels per image is lower to account for the poorer performance of the teacher in the target domain.

D. Class Distribution Prediction

We further investigate the accuracy of the proposed class ratio prediction method to assess if it is robust to different forms of domain shift. To do so, we utilise the four driving datasets used to assess UDA-OD performance; Cityscapes [18], foggy Cityscapes [19], BDD100k daytime and BDD100k night [20]. The permutation of these datasets into labelled and unlabelled pairs leads to 12 scenarios for predicting class distribution shift (Figure 5). Additionally, we use the Wanderlust dataset [17] to assess realistic class distribution shift encountered across time by an egocentric agent. This dataset contains object annotations of an egocentric videostream collected by a graduate student over nine months of their life. The seven most common classes of bicycle, bus, car, chair, dining table, person and potted plant are considered. Temporally coherent splits of the training data are created by splitting it at the 2.5%, 5%, 10%, 25%. 50%, 75% and 90% mark of the datastream. We use the labels for all data before the split, resulting in 7 scenarios for assessing class distribution prediction (Figure 4).

For each scenario, we report the Kullback-Leibler (KL) divergence between the labelled class ratio, our merged and squared predictions, and the true class ratio. Generally, the squared prediction is much more accurate than using the labelled data alone, reducing KL divergence by 3-4 fold. On the driving scenarios, our method is robust to large and small class distribution shifts and significant appearance changes due to foggy weather and night driving (Figure 5). This emphasises that the representation of semantic context used to predict the class ratio is sufficiently resistant to domain shift. On the Wanderlust scenarios, our prediction remains accurate across a variable labelling budget (Figure 4).

We also investigate the regression models to ascertain why the squared prediction is consistently more accurate than the

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

Fig. 3. a) The F1 score of pseudo-labels generated by the teacher model after 1000 iterations of self-training using different class distribution priors. The solid lines highlight how F1 score varies with respect to mean pseudo-label confidence when using different class ratios as prior. The dotted lines show the mean confidence score selected using each method. The pseudo-labels generated by our method (purple) have an F1 score of 0.3, while those generated using ACT (green) return 0.24. b) Evolution of the number of pseudo-labels per image while training the proposed method on the C2N and C2B scenarios. The dotted lines show the true number of objects per image in the target dataset for each scenario. c) Evolution of the number of pseudo-labels per image while training on the C2B scenario with different settings of the mean pseudo-label confidence τ. The black dotted line shows the true number of objects per image.

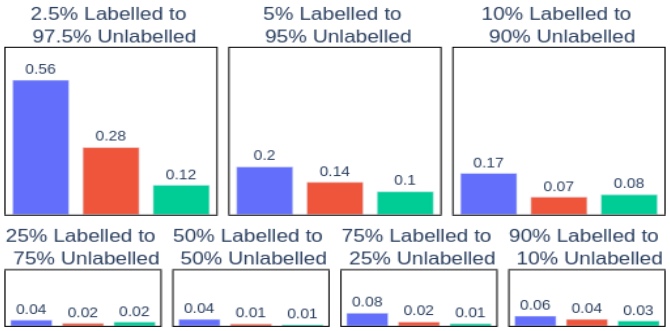

Fig. 4. KL divergence of alternative class ratio prediction methods for different temporal splits of the Wanderlust dataset. Each subplot shows an adaptation scenario, defined by the proportion of the data considered labelled. See Figure 4 for a description of the methods assessed.

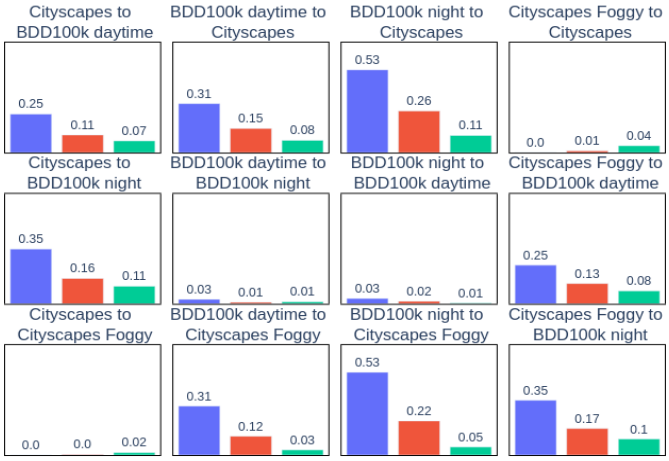

Fig. 5. KL divergence of alternative class ratio prediction methods from the true distribution on different combinations of the driving datasets. The title of each subplot shows the adaptation scenario. The blue bar represents the KL divergence between the labelled and unlabelled class ratio, with a larger value corresponding to scenarios with greater class distribution shift. The red and green bars show the KL-divergence between the unlabelled class ratio and the proposed merged and squared predictions respectively. Smaller values correspond to a more accurate prediction of the class ratio.

merged prediction (Figure 5, 4). We find that the intercept coefficients of these models vary significantly across datasets. This implies that the class ratio is influenced by exogenous factors not captured by the CLIP similarity scores. As a result, a model fit to the labelled data will consistently be in error when making predictions on the shifted, unlabelled data. However, this intercept error is strongly correlated with the class distribution shift between the datasets ($r = 0.98$ for the driving scenarios in Figure 5). Thus, by predicting the relative change in class ratio between datasets, we are indirectly estimating the intercept error. Scaling up the predicted change in class ratio therefore improves accuracy by incorporating this prediction of the intercept error into our method. In future, prompt learning techniques [25], [27] could be investigated to make the CLIP similarity scores more expressive of the class ratio, avoiding the need to heuristically account for the intercept error.

E. Ablations

Using the C2B scenario, we study the independent effect of using a more accurate class ratio and the dynamic adjustment of the number of objects per image. We first introduce the dynamic adjustment of the number of pseudo-labels per image, but continue using the labelled class distribution prior. The resulting 5.1 mAP improvement relative to the original ACT [13] can therefore be accredited to using the mean confidence of the teacher to select the number of pseudo-labels per image (Table III). When the predicted prior is introduced instead of the labelled class ratio, a further improvement of 3.8 mAP results. Lastly, we test the performance of our method if the true class ratio of the unlabelled data was used as the prior, finding that it leads to only a minor increase in performance, from 39.2 to 40.2 mAP (Table III).

We also study the sensitivity of our approach under different settings of the target mean pseudo-label confidence τ. For values between 0.5 and 0.7, our approach maintains performance superior to the strongest baseline (Table IV). These values perform optimally as they lead the number of pseudo-labels per image to converge to the true number of objects per image in the target dataset (Figure 3b-c). We therefore recommend setting $\tau = 0.6$ as a starting point on novel scenarios, before confirming that the number of pseudo-labels per image is converging to an appropriate value. In future work, merging class distribution information with probabilistic self-training approaches [4], [14] may reduce sensitivity to key parameters.

VI. Conclusion

In this work, we propose a framework for explicitly considering class distribution shift to improve the reliability of pseudo-labels in self-training. The resulting method improves upon existing baselines on a range of UDA-OD scenarios, showing strong performance when facing realistic contextual

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

and class distribution shifts. The results motivate further research into how a robot's deployment environment can be modelled to improve pseudo-label reliability in Mean Teacher. Furthermore, we show that class distribution shift must be considered when implementing UDA-OD methods for robotics applications. Lastly, the weak supervision provided by the class distribution prior may benefit other tasks crucial to reliable robotic deployment, such as run-time monitoring or out-of-distribution detection.

REFERENCES

[1] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. 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*.

[2] G. Graffieti, G. Borghi, and D. Maltoni, "Continual learning in real-life applications," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6195–6202, 2022.

[3] B. Liu, "Learning on the job: Online lifelong and continual learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 09, 2020, pp. 13 544–13 549.

[4] M. Chen, W. Chen, S. Yang, J. Song, X. Wang, L. Zhang, Y. Yan, D. Qi, Y. Zhuang, D. Xie *et al.*, "Learning domain adaptive object detection with probabilistic teacher," in *International Conference on Machine Learning*. PMLR, 2022, pp. 3040–3055.

[5] J. Deng, W. Li, Y. Chen, and L. Duan, "Unbiased mean teacher for cross-domain object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4091–4101.

[6] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[7] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister, "A simple semi-supervised learning framework for object detection," *arXiv preprint arXiv:2005.04757*, 2020.

[8] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda, "Unbiased teacher for semi-supervised object detection," *arXiv preprint arXiv:2102.09480*, 2021.

[9] R. Ramamonjison, A. Banitalebi-Dehkordi, X. Kang, X. Bai, and Y. Zhang, "Simrod: A simple adaptation method for robust object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3570–3579.

[10] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao, "Exploring object relation in mean teacher for cross-domain detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 457–11 466.

[11] X. Li, W. Chen, D. Xie, S. Yang, P. Yuan, S. Pu, and Y. Zhuang, "A free lunch for unsupervised domain adaptive object detection without source data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 8474–8481.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[13] B. Chen, W. Chen, S. Yang, Y. Xuan, J. Song, D. Xie, S. Pu, M. Song, and Y. Zhuang, "Label matching semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 381–14 390.

[14] H. Li, Z. Wu, A. Shrivastava, and L. S. Davis, "Rethinking pseudo labels for semi-supervised object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2.

[15] C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei, "Exploring categorical regularization for domain adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 724–11 733.

[16] H. Elsahar and M. Gallé, "To annotate or not? predicting performance drop under domain shift," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 2163–2173.

[17] J. Wang, X. Wang, Y. Shang-Guan, and A. Gupta, "Wanderlust: Online continual object detection in the real world," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[18] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[19] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, 2018.

[20] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving video database with scalable annotation tooling," *arXiv preprint arXiv:1805.04687*, vol. 2, no. 5, p. 6, 2018.

[21] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 437–10 446.

[22] S. Min, N. Park, S. Kim, S. Park, and J. Kim, "Grounding visual representations with texts for domain generalization," in *European Conference on Computer Vision*. Springer, 2022, pp. 37–53.

[23] Z. Zhou, B. Zhang, Y. Lei, L. Liu, and Y. Liu, "ZegCLIP: Towards Adapting CLIP for Zero-shot Semantic Segmentation," *arXiv*, 2022.

[24] B. Krojer, V. Adlakha, V. Vineet, Y. Goyal, E. Ponti, and S. Reddy, "Image retrieval from contextual descriptions," *arXiv preprint arXiv:2203.15867*, 2022.

[25] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[26] Y. Ruan, Y. Dubois, and C. J. Maddison, "Optimal representations for covariate shift," *arXiv preprint arXiv:2201.00057*, 2021.

[27] T. Huang, J. Chu, and F. Wei, "Unsupervised prompt learning for vision-language models," *arXiv preprint arXiv:2204.03649*, 2022.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[31] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[32] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6956–6965.

[33] T. Kim, M. Jeong, S. Kim, S. Choi, and C. Kim, "Diversify and match: A domain adaptive representation learning paradigm for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[34] H.-K. Hsu, C.-H. Yao, Y.-H. Tsai, W.-C. Hung, H.-Y. Tseng, M. Singh, and M.-H. Yang, "Progressive domain adaptation for object detection," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 749–757.

[35] M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang, "Cross-domain detection via graph-induced prototype alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 355–12 364.

[36] Z. He and L. Zhang, "Domain adaptive object detection via asymmetric tri-way faster-rcnn," in *European conference on computer vision*. Springer, 2020, pp. 309–324.

[37] C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, "Harmonizing transferability and discriminability for adapting object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8869–8878.

[38] Y. Zheng, D. Huang, S. Liu, and Y. Wang, "Cross-domain object detection through coarse-to-fine feature adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 766–13 775.

[39] C. Zhuang, X. Han, W. Huang, and M. Scott, "ifan: Image-instance full alignment networks for adaptive object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020.

[40] V. Vs, V. Gupta, P. Oza, V. A. Sindagi, and V. 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*, 2021, pp. 4516–4526.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.