DARTH: Holistic Test-time Adaptation for Multiple Object Tracking

Mattia Segu[1,2], Bernt Schiele[2], Fisher Yu[1]
[1] ETH Zurich, [2] Max Planck Institute for Informatics, Saarland Informatics Campus
segum@ethz.ch, schiele@mpi-inf.mpg.de, i@yf.io

Abstract

Multiple object tracking (MOT) is a fundamental component of perception systems for autonomous driving, and its robustness to unseen conditions is a requirement to avoid life-critical failures. Despite the urge of safety in driving systems, no solution to the MOT adaptation problem to domain shift in test-time conditions has ever been proposed. However, the nature of a MOT system is manifold - requiring object detection and instance association - and adapting all its components is non-trivial. In this paper, we analyze the effect of domain shift on appearance-based trackers, and introduce DARTH, a holistic test-time adaptation framework for MOT. We propose a detection consistency formulation to adapt object detection in a self-supervised fashion, while adapting the instance appearance representations via our novel patch contrastive loss. We evaluate our method on a variety of domain shifts - including sim-to-real, outdoor-to-indoor, indoor-to-outdoor - and substantially improve the source model performance on all metrics. Project page: https://www.vis.xyz/pub/darth.

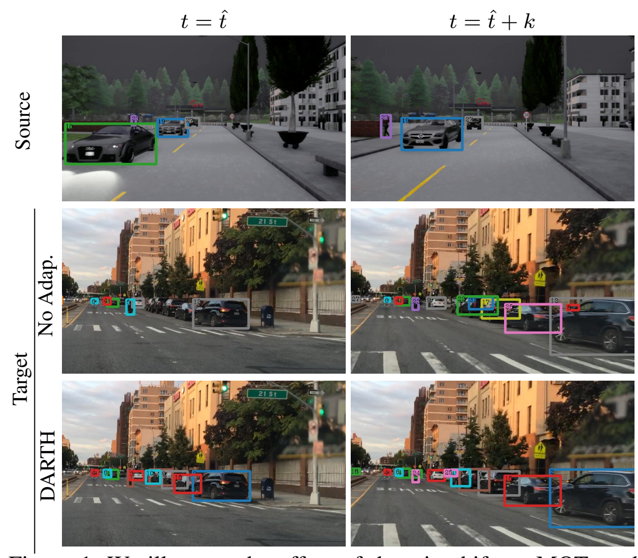

Figure 1. We illustrate the effect of domain shift on MOT, and how our test-time adaptation technique (DARTH) counteracts it. The top row shows the in-domain performance of a model trained on the synthetic dataset SHIFT [60] (Source); the same model (No Adap.) suffers from domain shift when deployed on the real-world BDD100K [70] (Target); the bottom row shows the benefits of DARTH. Each row shows two frames spaced by $k=2$ seconds; boxes of the same color correspond to the same tracking ID.

1. Introduction

Multiple object tracking (MOT) represents a cornerstone of modern perception systems for challenging computer vision applications, such as autonomous driving [17], video surveillance [16], behavior analysis [26], and augmented reality [46]. Laying the ground for safety-critical downstream perception and planning tasks - e.g. obstacle avoidance, motion estimation, prediction of vehicles and pedestrians intentions, and the consequent path planning - the robustness of MOT to diverse conditions is of uttermost importance.

However, domain shift [28] could result in life-threatening failures of MOT pipelines, due to the perception system's inability to understand previously unseen environments and provide meaningful signals for downstream planning. To the best of our knowledge, despite the urge of addressing domain adaptation for MOT to enable safer driving and video analysis, no solution has ever been proposed.

This paper analyzes the effect of domain shift on MOT, and proposes a test-time adaptation solution to counteract it. We focus on appearance-based tracking, which shows state-of-the-art performance across a variety of datasets [19], outperforms motion-based trackers in complex scenarios - i.e. BDD100K [70] - and complements motion cues for superior tracking performance [75]. Since appearance-based trackers [31, 66, 1, 45] associate detections through time based on the similarity of their learnable appearance embeddings, domain shift threatens the performance of both their detection and instance association stages (Table 1).

Test-time adaptation (TTA) offers a practical solution to domain shift by adapting a pre-trained model to any unlabeled target domain in absence of the original source domain. However, current TTA techniques are tailored to classification tasks [63, 8, 64, 40] or require altering the source training procedure [61, 35, 42], and they have been shown to struggle in complex scenarios [35]. Consequently, the development of TTA solutions for MOT is non-trivial. While recent work further investigates TTA for object detection [33, 57], solving TTA for detection is not sufficient to recover MOT systems (see SFOD [33] in Table 5), as instance association plays an equally crucial role in tracking.

To this end, we introduce a holistic test-time adaptation framework that addresses the manifold nature of MOT (Figure 2). We propose a detection consistency formulation to adapt object detection in a self-supervised fashion and enforce its robustness to photometric changes, since tracking benefits from consistency of detection results in adjacent frames. Moreover, we adapt instance association and learn meaningful instance appearance representations on the target domain by introducing a patch contrastive loss, which enforces self-matching of the appearance of detected instances under differently augmented views of the same image. Finally, we update the teacher as an exponential moving average (EMA) of the student model to benefit from the adapted student representations and gradually improve the detection targets for our consistency loss.

We name DARTH our test-time Domain Adaptation method for Recovering multiple object Tracking Holistically. To the best of our knowledge, our proposal is the first solution to the domain adaptation problem for MOT. We evaluate DARTH on a variety of domain shifts across the driving datasets SHIFT [60] and BDD100K [70], and the pedestrian datasets MOT17 [39] and DanceTrack [58], showing substantial improvements over the source model performance on all the evaluated metrics and settings.

We summarize our contributions: (i) we study the domain shift problem for MOT and introduce the first test-time adaptation solution; (ii) we propose a detection consistency formulation to adapt object detection and enforce its consistency to photometric changes; (iii) we introduce a patch contrastive approach to adapt the appearance representations for better data association.

2. Related Work

Multiple Object Tracking. Tracking-by-detection, i.e. detecting objects in individual frames of a video and associating them over time, is the dominant paradigm in MOT. Motion- [3, 4, 18, 6, 75], appearance- [31, 66, 1, 45], and query-based [38, 59, 71] trackers are commonly used to associate the instances detected by an object detector. In this work, we focus on domain adaptation of appearance-based trackers, building on the state-of-the-art QDTrack [45, 19]. QDTrack introduces a quasi-dense paradigm for learning appearance representations, exceeding the association ability of motion- and query-based trackers. Moreover, appearance provides a complementary cue to motion [75]. Nevertheless, Table 1 shows that domain shift threatens both object detection and the learned appearance representations of QDTrack, negatively affecting instance association in MOT. Previous work partially investigated MOT under diverse conditions [20] and limited labels [36]. Our paper provides the first comprehensive analysis of domain shift in MOT, and introduces an holistic framework to counteract its effect on the object detection and data association stages of appearance-based trackers.

Source	Target	DetA	MOTA	HOTA	IDF1	AssA
SHIFT	SHIFT	46.9	48.4	55.2	60.6	65.8
	BDD100K	12.0	-66.4	17.3	18.5	28.9
MOT17	MOT17	57.2	68.2	57.1	68.5	57.4
	DanceTrack	52.4	57.2	21.5	19.5	9.0
	BDD100K	23.2	10.5	27.2	33.3	32.4
MOT17 (+CH)	MOT17	59.8	71.7	59.7	71.6	58.7
	DanceTrack	61.8	74.0	31.1	29.6	15.8
	BDD100K	32.4	28.3	33.7	41.7	35.4
DanceTrack	DanceTrack	68.5	79.2	43.5	42.3	28.0
	MOT17	24.7	23.3	32.6	35.4	43.5
	BDD100K	9.3	-16.0	14.1	12.3	21.8
BDD100K	BDD100K	36.5	14.2	39.6	48.2	43.3
	MOT17	28.6	31.4	36.0	43.5	45.8
	DanceTrack	41.9	41.6	18.0	17.0	7.9

Table 1. **Domain shift in MOT.** We assess the impact of domain shift on the performance of a QDTrack model based on Faster R-CNN with a ResNet-50 backbone. In green the performance on the source domain. The SHIFT → BDD100K metrics are averaged across all object categories; only the pedestrian category is considered for all other experiments. CH: CrowdHuman.

Test-time Adaptation. Differently from unsupervised domain adaptation (UDA) [65], which assumes the availability of target samples when training on the source domain, test-time adaptation aims at adapting a source pre-trained model on any unlabeled target domain in absence of the original source domain. A popular approach to TTA consists in learning, together with the main task, an auxiliary task with easy self-supervision on the target domain, e.g. geometric transformations prediction [15, 21, 61], colorizing images [73, 30], solving jigsaw puzzles [42]. However, such techniques require to alter the training procedure on the source domain to also learn the auxiliary task. Recent approaches allow instead to perform fully test-time adaptation without altering the source training. [53, 41, 40, 54] show the benefits of simply tuning on the target domain the batch normalization statistics of a frozen model. Tent [63] minimizes the output self-entropy on the target domain to learn the shift and scale parameters of the batch normalization (BN) layers while using the batch statistics. Such techniques do not finetune the task-specific head while altering its expected input distribution, deteriorating the model performance under severe distribution shifts [35]. AdaContrast [8] and CoTTA [64] instead enforce prediction consistency under augmented views, learning global representations on the target domain for image classification. In contrast, the combination of our detection consistency formulation and our patch contrastive learning enables DARTH to simultaneously learn global and local representations on the target domain, while adapting respectively the task-specific detection and appearance heads.

Domain Adaptation for Object Detection. Object detection [48, 47] plays a key role in tracking-by-detection. Several works [43] focus on the unsupervised domain adaptation problem for object detection, adopting traditional techniques such as adversarial feature learning [10, 51, 25, 56], image-to-image translation [72, 7, 49], pseudo-label self-training [27, 29, 50], and mean-teacher training [5, 14]. However, such techniques require the availability of the labeled source domain. A more practical test-time adaptation solution [32] shows promising results by self-training with high-confidence pseudo-labels, though it was only applied to arbitrary and mild domain discrepancies such as Cityscapes [12] to Foggy Cityscapes [52] or to BDD100K [70] daytime. Finally, object detection adaptation techniques do not seamlessly extend to MOT adaptation, since the latter requires a further data association stage and detection consistency through time.

3. DARTH

We here introduce DARTH, our test-time adaptation method for MOT. We first introduce the TTA setting (Section 3.1) and give an overview of DARTH (Section 3.2). We further detail our patch contrastive learning and detection consistency formulation in Sections 3.3 and 3.4.

3.1. Test-time Adaptation for MOT

Test-time adaptation addresses the problem of adapting a model previously trained on a source domain $\mathcal{S} = \{(x_s^i, y_s^i,)\}_{i=1}^{N_s}$ to an unlabeled target domain $\mathcal{T} = \{x_t^i\}_{i=1}^{N_t}$, without accessing the source domain.

In this work, we tackle the TTA problem for MOT, building on the state-of-the-art appearance-based tracker QDTrack [45]. Following the tracking-by-detection paradigm, modern MOT methods [45, 6, 75] rely on a detection stage and a data association stage. QDTrack extends a Faster R-CNN [48] detector with an additional embedding head, and learns appearance similarity via a multi-positive contrastive loss that enforces discriminative instance representations. Under domain shift, all the components of the tracking pipeline fail, with significant performance drops on both detection and association metrics (Table 1).

3.2. Overview

MOT systems are composed of an object detection and a data association stage, tightly-coupled with each other. Adapting the one does not necessarily have a positive effect on the other (Table 6). To address this problem, we introduce DARTH, a holistic TTA framework that addresses the manifold nature of MOT by emphasizing the importance of the whole and the interdependence of its parts.

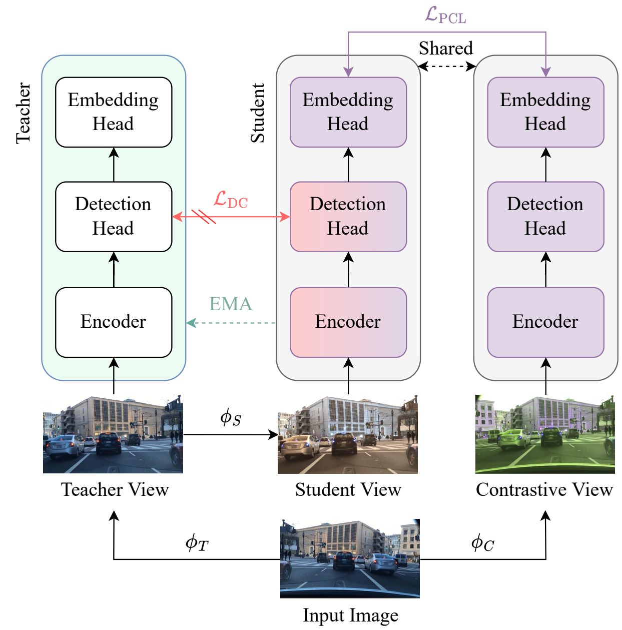

Figure 2. Schematic representation on the target domain of DARTH, our test-time adaptation method for MOT. Our patch contrastive loss $\mathcal{L}_{\text{PCL}}$ between the siamese student's instance embeddings adapts instance association. Our detection consistency loss $\mathcal{L}_{\text{DC}}$ enforces consistency to photometric changes. The EMA updates to the teacher gradually improve the detection targets for our consistency loss. ϕ_T, ϕ_S, and ϕ_C are the image transformations described in Section 3.2. '\\' = stop gradient.

Architecture. DARTH relies on a teacher model and a siamese student (Figure 2). Given a set of QDTrack weights $\hat{\theta}$ trained on the source domain following [45], the student network is defined by a set of weights $\theta := \hat{\theta}$. The teacher shares the same architecture with the student and its weights ξ are initialized from the student weights $\hat{\theta}$ and updated as an EMA of the student parameters θ during adaptation: $\xi \leftarrow \tau\xi + (1-\tau)\theta$. τ is the momentum of the update. The momentum teacher provides the targets to our detection consistency loss (Section 3.4) between teacher and student detection outputs under two differently augmented versions (views) of the same image. The siamese student enables learning discriminative appearance representations via our patch contrastive loss (Section 3.3) between the detections of two views of the same image. At inference time, we use our DARTH-adapted model to detect objects and extract instance embeddings, and apply the standard QDTrack inference strategy described in [45] to track objects in a video.

Views Definition. Figure 2 illustrates a schematic view of our framework and of the generation process of the different input views. Given an input image x, we apply a geometric augmentation ϕ_T to generate the teacher view x_T, and apply a subsequent photometric augmentation ϕ_S to produce the student view x_S. We generate x_S from x_T to satisfy the assumption of geometric alignment of teacher and student views in our detection consistency loss. The contrastive view x_C used in the siamese pair is independently generated by applying a sequence ϕ_C of geometric and photometric augmentations on the original input image. We ablate on the impact of different augmentation strategies in Section 4.3. Details on the choice and parameters of geometric and photometric augmentations are in the Appendix.

3.3. Patch Contrastive Learning

To adapt the data association stage and learn discriminative appearance representations on the target domain, we introduce a novel patch contrastive learning (PCL) formulation, whose functioning is illustrated in Figure 3.

Localizing Objects. The goal of this step is identifying on the two views object regions over which learning instance-discriminative appearance representations, and filter out false positive detections. Given an image x from the target domain $\mathcal{T}$ and the set of K detections $D=\{d_i\}_{i=1}^{K}$ extracted by the teacher detector, we filter the detections by retaining only those with confidence higher than a threshold γ, i.e. $\hat{D}=\{d \in D | \text{conf}(d) \geq \gamma\}$. We then generate the student and contrastive views x_S and x_C by respectively applying on x the image transformations $\hat{\phi}_S = \phi_T \circ \phi_S$ and ϕ_C, and coherently warping the detections to $\hat{D}_S$ and $\hat{D}_T$.

Quasi-dense Formulation. We then phrase the patch contrastive learning problem as quasi-dense self-matching of the contrastive-view regions of interest (RoIs) R_C - i.e. Faster R-CNN region proposals - to the student-view proposals R_S. Since the student- and contrastive-view detections $\hat{D}_S$ and $\hat{D}_C$ are generated by augmenting the same teacher detections $\hat{D}$, instance correspondences between x_S and x_C are known in advance. In particular, we output image-level features through the student encoder, use the region proposal network (RPN) to generate RoIs from the two images and RoI Align [23] to pool their feature maps at different levels in the Feature Pyramid Network (FPN) [34] according to their scales. For each RoI we extract deeper appearance features via the additional embedding head. A RoI in a view x_i is considered a positive match to a detection $\hat{D}_i$ on the same view if they have Intersection over Union (IoU) higher than $\alpha_1 = 0.7$, negative if lower than $\alpha_2 = 0.3$. The matching of RoIs under the two views x_S and x_C is positive if both regions are associated to the same teacher detection $\hat{D}$; negative otherwise.

Patch Contrastive Learning. Assuming that V positive RoIs are proposed on the student view x_S as training samples and K RoIs on the contrastive view x_C as contrastive targets, we use the non-parametric softmax [68, 62] with cross-entropy to optimize the appearance embeddings of each training sample. We here only show the loss for one training sample, but average it over all of them:

$$\mathcal{L}_{\text{embed}} = -\sum_{\mathbf{k}^+} \log \frac{\exp(\mathbf{v} \cdot \mathbf{k}^+)}{\exp(\mathbf{v} \cdot \mathbf{k}^+) + \sum_{\mathbf{k}^-} \exp(\mathbf{v} \cdot \mathbf{k}^-)}, \quad (1)$$

where $\mathbf{v}$ are RoI embeddings on x_S, and $\mathbf{k}^+$, $\mathbf{k}^-$ are their positive and negative targets on x_C.

Analogously to [45], we reformulate Eq. (1) to avoid considering each negative target $\mathbf{k}^-$ multiple times per training sample $\mathbf{v}$, while only once the positive one $\mathbf{k}^+$:

$$\mathcal{L}_{\text{embed}} = \log[1 + \sum_{\mathbf{k}^+} \sum_{\mathbf{k}^-} \exp(\mathbf{v} \cdot \mathbf{k}^- - \mathbf{v} \cdot \mathbf{k}^+)]. \quad (2)$$

We further adopt an L2 auxiliary loss to constrain the logit magnitude and cosine similarity:

$$\mathcal{L}_{\text{aux}} = \left(\frac{\mathbf{v} \cdot \mathbf{k}}{||\mathbf{v}|| \cdot ||\mathbf{k}||} - \mathbb{1}_{\{\mathbf{k} \in \{\mathbf{k}^+\}\}} \right)^2, \quad (3)$$

where $\mathbb{1}$ is the indicator function and $\mathbf{k}$ an RoI embedding such that $\mathbf{k} \in \{\mathbf{k}^-\} \cup \{\mathbf{k}^+\}$. We calculate the auxiliary loss over all positive pairs and three times more negative pairs.

3.4. Detection Consistency

While our PCL adapts the local appearance representations to the target domain and improves instance association, not imposing any additional constraint might let the global image features deviate from the distribution expected by the detection head and damage the overall performance (Table 6). Inspired by self-supervised representation learning for image classification [22], we introduce a detection consistency (DC) loss between predictions of the teacher and student detection heads under different image augmentations to adapt object detection to the target domain, while EMA updates to the teacher model gradually incorporate the improved student representations and enable better targets for the consistency loss.

A by-product of our self-consistency to different augmentations is fostering better global representations on the target domain, complementary to the local representations learned via our PCL. Moreover, tracking-by-detection is negatively affected by flickering of detections through time, and domain shift exacerbates this issue. We find that enforcing detection consistency under different photometric augmentations stabilizes detection outputs in adjacent frames, significantly improving MOTA [2] (Table 7).

In particular, our detection consistency loss is composed of an RPN- and an RoI-consistency component applied on the RPN and RoI heads in Faster R-CNN. Notice that our method applies to other two-stage detectors, and extends to one-stage detectors by ignoring the RPN consistency loss. We now present the details of our method.

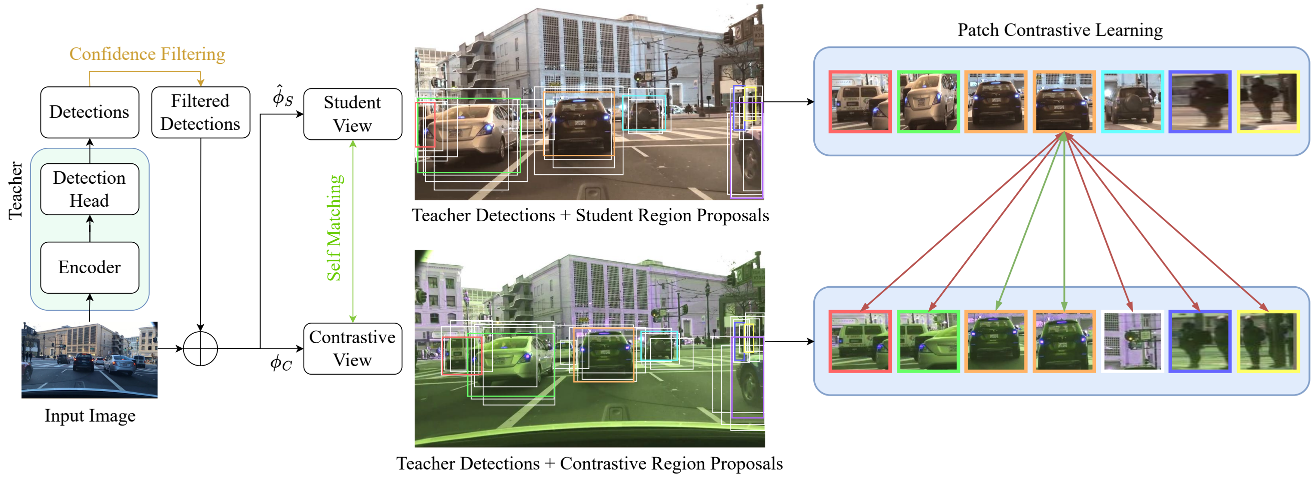

Figure 3. We here illustrate our novel patch contrastive formulation (Section 3.3). First, we identify object regions by applying the teacher detector on the input image and filtering the detections based on their confidence. We then apply on the input image and detected bounding boxes the transformations $\hat{\phi}_S$ and ϕ_C to generate the student and contrastive view, deriving association pseudo-labels by considering a match as positive when proposed regions (white) on different views match to a same teacher detection (identified by the same color across both views). Finally, we apply a multi-positive patch contrastive loss on the projections of the proposed regions obtained via the student embedding head. We here show an example of positive (green) and negative (red) matches for one of the student-view proposed regions.

Views Definition. We contextualize our choices on the views generation protocol (Section 3.2). We generate the teacher image x_T by applying a geometric augmentation ϕ_T on the input image x. Since the teacher predictions should provide high-quality targets for the student, we do not further corrupt x_T with photometric augmentations. Moreover, our DC loss requires geometric alignment of x_T and x_S. We thus generate x_S by applying a photometric augmentation ϕ_S on the teacher view x_T to satisfy geometric alignment and allow consistency under photometric changes.

RPN Consistency. We implement RPN consistency as an $\mathcal{L}_2$ loss between the teacher ξ and student θ RPN regression - i.e. displacement w.r.t. anchors - and classification outputs on x_T and x_S. Inspired by model compression [9], we control the regression consistency by a threshold $\epsilon = 0.1$ on the difference between the teacher and student classification outputs. We define the RPN consistency loss as:

$$\mathcal{L}_{\text{DC}}^{\text{RPN}} = \frac{1}{N}\sum\left(\left\|s_\xi - s_\theta\right\|_2^2 + \mathbb{1}_{\{s_\xi > s_\theta + \epsilon\}}\left\|r_\xi - r_\theta\right\|_2^2\right), \quad (4)$$

where $\mathbb{1}$ is the indicator function, N the number of anchors, s the RPN classification logits, r the RPN regression output.

RoI Consistency. We feed the teacher proposals into the student RoI head and enforce an $\mathcal{L}_2$ consistency loss with the final teacher regression - i.e. displacement w.r.t. region proposals - and classification outputs. For each RoI classification output - the logits p - we subtract the mean over the class dimension to get the zero-mean classification result, $\tilde{p}$. Given K sampled RoIs, C classes including background, and the bounding box regression result t, we derive the RoI consistency loss as:

$$\mathcal{L}_{\text{DC}}^{\text{RoI}} = \frac{1}{K \cdot C}\sum\left(\left\|\tilde{p}_\xi - \tilde{p}_\theta\right\|_2^2 + \left\|t_\xi - t_\theta\right\|_2^2\right) \quad (5)$$

3.5. Total Loss

The entire framework is jointly optimized under a weighted sum of the individual losses:

$$\begin{aligned}\mathcal{L} &= \gamma_1\mathcal{L}_{\text{embed}} + \gamma_2\mathcal{L}_{\text{aux}} + \gamma_3\mathcal{L}_{\text{DC}}^{\text{RPN}} + \gamma_4\mathcal{L}_{\text{DC}}^{\text{RoI}} \quad (6)\\ &= \mathcal{L}_{\text{PCL}} + \gamma_3\mathcal{L}_{\text{DC}}^{\text{RPN}} + \gamma_4\mathcal{L}_{\text{DC}}^{\text{RoI}}, \quad (7)\end{aligned}$$

where γ_1, γ_2, γ_3 and γ_4 are set to 0.25, 1.0, 1.0 and 1.0. $\mathcal{L}_{\text{PCL}} = \gamma_1\mathcal{L}_{\text{embed}} + \gamma_2\mathcal{L}_{\text{aux}}$ is the total PCL loss.

In Section 4.3 we ablate on the need for each individual component, showing the importance of a holistic adaptation solution for MOT that emphasizes the importance of the whole and the interdependence of its parts.

4. Experiments

We provide a thorough experimental analysis of the benefits of our proposal. We detail the experimental setting in Section 4.1, evaluate DARTH on a variety MOT adaptation benchmarks (Section 4.2), and ablate on different method components and data augmentation strategies (Section 4.3). Further experimental results are in the Appendix.

4.1. Experimental Setting

We tackle the offline TTA problem for MOT. Each model is initially supervised on the Source dataset, and adapted/tested on the combined validation set of the Target dataset. Only the categories shared across both datasets are considered. To evaluate the impact of domain shift on the individual components of MOT systems and how each TTA method can address them, we choose a set of 5 metrics here ordered by the extent to which they measure the detection (left) or association (right) performance: DetA [37], MOTA [2], HOTA [37], IDF1 [2], AssA [37].

Benchmark. We validate DARTH on a variety of domain shifts across the driving datasets SHIFT [60] and BDD100K [70], and the pedestrian datasets MOT17 [39] and DanceTrack [58]. The *sim-to-real* gap provided by SHIFT → BDD100K offers a comprehensive scenario to analyze the impact of domain shift on multi-category multiple object tracking. By training and adapting on both datasets only on the set of shared categories - i.e. pedestrian, car, truck, bus, motorcycle, bicycle - we can assess how different adaptation methods deal with class imbalance. Moreover, we analyze the *outdoor-to-indoor* shift on MOT17 → DanceTrack and BDD100K → DanceTrack, and *indoor-to-outdoor* shift in the opposite direction. Finally, we investigate how trackers trained on small datasets can be improved via large amounts of unlabeled and diverse data (*small-to-large*) in MOT17 → BDD100K and DanceTrack → BDD100K, while the opposite direction tells us more about the generality of trackers trained on large-scale driving datasets. Experiments on additional domain shift settings are reported in the Appendix.

Baselines. Although no method for TTA of MOT was previously proposed, we compare against extensions to QDTrack [45] of popular TTA techniques for image classification and object detection: the No Adaptation (No Adap.) baseline, which applies the source pre-trained model directly on the target domain without further finetuning; Tent [63], originally proposed for image classification, we extend it to adapt the encoder's batch normalization parameters by minimizing the entropy of the RoI classification head; SFOD [33], a TTA method for object detection which adapts a student model on the confidence-filtered detections of a source model on the target domain; Oracle, the optimal baseline provided by an oracle model trained directly on the target domain with full supervision and access to the privileged information provided by the target labels.

Implementation Details. We build on the state-of-the-art appearance-based tracker, QDTrack [45]. QDTrack equips an object detector with a further embedding head to learn instance similarities. As object detector, we use the Faster R-CNN [48] architecture with a ResNet-50 [24] backbone and FPN [34]. Our embedding head is a *4conv1fc* head with group normalization [67] to extract 256-dimensional features. For additional source model implementation details and tracking algorithm, refer to the original paper [45].

During the adaptation phase, the teacher model is updated as an EMA of the student weights with a momentum $\tau = 0.998$. For our patch contrastive loss we sample 128 RoIs via IoU-balanced sampling [44] from the student view and 256 from the contrastive view, with a positive-negative ratio of 1.0 for the contrastive targets. We use the SGD optimizer, with an initial learning rate of 0.001 decayed following a dataset-dependent step schedule. The gradients' norm is clipped to 35. Further dataset- and method-specific hyperparameters are reported in the Appendix.

Method	Source	Target	DetA	MOTA	HOTA	IDF1	AssA
No Adap.	SHIFT	BDD100K	12.0	-66.4	17.3	18.5	28.9
Tent [63]			0.1	0.0	0.7	0.2	4.5
SFOD [33]			12.4	-57.3	17.7	19.0	29.1
DARTH			**15.2**	**8.3**	**20.6**	**23.7**	**33.1**
Oracle	BDD100K	BDD100K	29.6	35.8	35.1	56.0	42.6

Table 2. **State of the art on SHIFT → BDD100K.** We compare DARTH (ours) against baseline TTA methods for adapting QDTrack from the synthetic driving dataset SHIFT to the real-world BDD100K. Metrics are averaged across all object categories.

4.2. DARTH

Domain Shift in MOT. We analyze the effect of different types of domain shift on a QDTrack model pre-trained on a given source domain (Table 1). Sim-to-real drastically affects all the components of the tracking pipeline, with the detection accuracy (DetA) dropping by -74.4%, the association accuracy (AssA) more than halving, and the MOTA suffering a catastrophic -118.8. Interestingly, MOT17 → DanceTrack provides a contextual shift fatal to the AssA (-84.3%), while the DetA remains stable. This can be explained by the identical clothing of dancers in DanceTrack, causing problems to embedding heads learned on datasets where diverse clothing is a discriminative feature. Inversely, indoor trackers trained on DanceTrack fail to generalize their DetA, but retain high AssA on outdoor datasets. These findings call for a solution that addresses adaptation of the tracking pipeline as a whole.

SHIFT → BDD100K. We analyze the impact of different TTA adaptation strategies on this sim-to-real setting in Table 2, and report each metric averaged across all object categories. Compared to the SFOD baseline, which produces only marginal improvements, DARTH effectively boosts all the components of the MOT system, with a noteworthy +74.7 MOTA over the non-adapted source model (No Adap.). This result highlights the effectiveness of DARTH under severe domain shift and in class-imbalanced conditions. Notably, using Tent [63] out-of-the-box fails in this scenario. While in other settings (Tables 3 to 5) Tent's failure is less striking, we argue that it is expected since: (i) the entropy minimization objective harms localization; (ii) object detectors commonly keep the encoder's ImageNet [13] normalization statistics frozen, while Tent updates the batch statistics during adaptation and the model cannot cope with such a large internal distribution shift. Recent work also

Method	Source	Target	DetA	MOTA	HOTA	IDF1	AssA
No Adap.	MOT17	DT	52.4	57.2	21.5	19.5	9.0
Tent [63]			32.6	27.7	11.9	10.9	4.6
SFOD [33]			53.5	59.0	22.0	20.3	9.3
Ours			**57.2**	**70.1**	**31.6**	**32.8**	**17.7**
Oracle	DT	DT	68.5	79.2	43.5	42.3	28.0
No Adap.	MOT17 (+ CH)	DT	61.8	74.0	31.1	29.6	15.8
Tent [63]			25.5	26.7	12.2	11.3	6.0
SFOD [33]			62.5	74.1	30.1	27.5	14.7
Ours			**64.7**	**78.9**	**35.4**	**35.3**	**19.6**
Oracle	DT	DT	68.5	79.2	43.5	42.3	28.0
No Adap.	DT	MOT17	24.7	23.3	32.6	35.4	43.5
Tent [63]			18.9	-4.8	26.0	25.1	37.4
SFOD [33]			25.1	23.7	33.1	35.7	44.3
Ours			**26.4**	**25.5**	**34.3**	**37.9**	**45.2**
Oracle	MOT17	MOT17	57.2	68.2	57.1	68.5	57.4

Table 3. **State of the art on MOT17 → DanceTrack and DanceTrack → MOT17.** We compare DARTH (ours) against baseline TTA methods for multiple object tracking across pedestrian tracking datasets. DT: DanceTrack; CH: CrowdHuman.

Method	Source	Target	DetA	MOTA	HOTA	IDF1	AssA
No Adap.	BDD100K	MOT17	28.6	31.4	36.0	43.5	45.8
Tent [63]			17.3	-86.8	24.6	23.9	35.9
SFOD [33]			**29.6**	31.7	35.4	42.4	42.8
Ours			29.4	**32.6**	**36.6**	**44.4**	**45.9**
Oracle	MOT17	MOT17	57.2	68.2	57.1	68.5	57.4
No Adap.	BDD100K	DT	41.9	41.6	18.0	17.0	7.9
Tent [63]			9.9	-45.9	6.1	4.7	3.8
SFOD [33]			43.8	42.3	18.1	17.0	7.6
Ours			**45.1**	**50.2**	**21.5**	**21.4**	**10.4**
Oracle	DT	DT	68.5	79.2	43.5	42.3	28.0

Table 4. **State of the art on BDD100K → MOT17/DanceTrack.** We compare DARTH (ours) against baseline TTA methods for adapting a pedestrian MOT model trained on the large-scale driving dataset BDD100K to the pedestrian datasets MOT17 and DanceTrack (DT).

shows that Tent deteriorates the source model under strong distribution shift in both image classification [69] and semantic segmentation [74]. Finally, Figure 1 shows qualitative results before and after adaptation with DARTH. While No Adap. fails at consistently detecting across frames the cars on the right side of the road, DARTH successfully recovers missing detections and correctly tracks them.

MOT17 ↔ DanceTrack. We compare different TTA adaptation methods on indoor-outdoor and contextual shifts on the MOT17 and DanceTrack datasets in Table 3. As reported in Table 1, DanceTrack poses a great challenge to the data association of a tracker trained on MOT17. We show that DARTH almost doubles the initial AssA of the non-adapted source model, and increases the MOTA and HOTA by a remarkable +12.9 and +10.1, considerably bridging the gap with an Oracle model directly trained on the target domain DanceTrack. More limited is the performance boost

Method	Source	Target	DetA	MOTA	HOTA	IDF1	AssA
No Adap.	DT	BDD100K	9.3	-16.0	14.1	12.3	21.8
Tent [63]			3.6	-29.5	8.1	5.7	18.6
SFOD [33]			9.5	-23.2	14.8	12.9	23.4
Ours			**12.8**	**-1.5**	**17.8**	**17.4**	**25.1**
No Adap.	MOT17	BDD100K	23.2	10.5	27.2	33.3	32.4
Tent [63]			13.4	-29.5	18.9	19.7	27.2
SFOD [33]			24.5	-7.4	27.8	32.9	32.2
Ours			**31.6**	**21.4**	**32.4**	**40.4**	**33.6**
No Adap.	MOT17 (+ CH)	BDD100K	32.4	**28.3**	33.7	41.7	35.4
Tent [63]			3.6	-29.5	8.1	5.7	18.6
SFOD [33]			34.9	17.0	35.1	41.9	35.8
Ours			**36.3**	23.4	**36.3**	**44.4**	**36.8**
Oracle	BDD100K	BDD100K	36.5	14.2	39.6	48.2	43.3

Table 5. **State of the art on MOT17/DanceTrack → BDD100K.** We compare DARTH (ours) against baseline TTA methods for adapting pedestrian MOT models to the large-scale driving dataset BDD100K. DT: DanceTrack; CH: CrowdHuman.

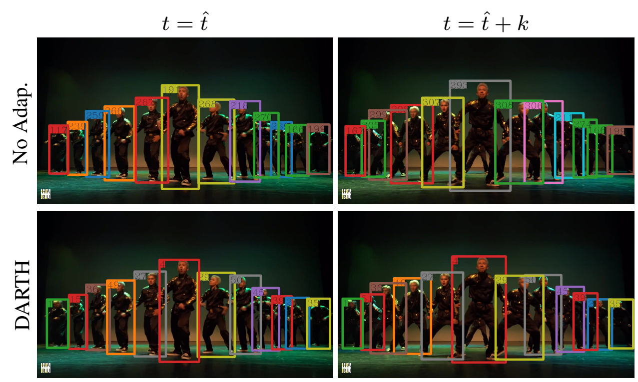

Figure 4. Tracking results on the sequence *0034* of the DanceTrack validation set in the adaptation setting MOT17 → DanceTrack. We analyze 2 frames spaced by $k = 0.5$ seconds and visualize the No Adap. baseline (top row) and DARTH (bottom row). On each row, boxes of the same color correspond to the same tracking ID.

in the opposite direction, where our proposal improves the source model over all metrics, but the DetA gap with the Oracle model remains large. Qualitative results before and after adaptation with DARTH on MOT17 → DanceTrack are shown in Figure 4. The unadapted source model correctly detects the dancers but fails at associating them, while DARTH effectively recovers instance association.

Pedestrians ↔ BDD100K. Data annotation is an expensive procedure, especially in video tasks such as MOT. Being able to train on limited labeled data and generalize to large and diverse unlabeled datasets would save enormous annotation costs and time. Table 5 shows how, after adapting an MOT17 model to BDD100K with DARTH, the gap with the Oracle trained on BDD100K is drastically reduced, with our DARTH model far exceeding the Oracle's MOTA. When pre-training on CrowdHuman, DARTH even ties the Oracle's DetA, although there is still room for improving

EMA	DC	PCL	DetA	MOTA	HOTA	IDF1	AssA
-	-	-	12.0	-66.4	17.3	18.5	28.9
-	-	✓	9.4	-40.5	14.3	14.5	27.6
-	✓	-	12.6	-37.6	18.0	19.5	29.5
✓	✓	-	14.5	6.1	19.7	22.0	31.0
✓	✓	✓	**15.2**	**8.3**	**20.6**	**23.7**	**33.1**

Table 6. **Ablation study on the impact of different method components on DARTH (Average).** We analyze the effect of different method components on DARTH (ours) on SHIFT → BDD100K. We report with a ✓whether exponential moving average (EMA), detection consistency (DC) and Patch Contrastive Learning (PCL) are applied. For each metric we report its average across all object categories. No Adap. is in gray.

data association. In contrast, SFOD only marginally satisfies its objective of improving the DetA, while worsening all tracking-related metrics by not adapting data association.

Our method reports improvements also in the opposite direction (Table 4), where the tracker is first trained on the large scale dataset BDD100K and then asked to generalize to the smaller scale datasets MOT17 and DanceTrack. Nevertheless, the SFOD baseline also shows improvements on BDD100K → MOT17, slightly exceeding DARTH's DetA.

4.3. Ablation Studies

We here ablate on different design choices and components of DARTH, highlighting the importance of a holistic solution to the MOT adaptation problem. Additional ablations and visual results are provided in the Appendix.

Method Components. We ablate on the impact of different method components - i.e. exponential moving average (EMA), detection consistency (DC), and patch contrastive learning (PCL) - on SHIFT → BDD100K in Table 6. We find that applying PCL alone is detrimental, since the newly learned features become incompatible with the unadapted detection head. Applying DC alone produces instead improvements over all metrics, and in particular over the MOTA, hinting at the enhanced consistency of detection results in adjacent frames. Enabling the momentum updates to the teacher (EMA + DC) causes a remarkable boost, meaning that the adapted global representations fostered by DC and gradually injected into the teacher generate better targets for our DC formulation. Finally, the PCL further boosts the performance of EMA + DC, proving how all tracking components are interconnected and a holistic solution is required to achieve the best adaptation performance.

Data Augmentation. We ablate on the effect of different data augmentation strategies to generate the teacher, student, and contrastive views. The results, reported in Table 7, show how applying independent geometric augmentations to the teacher/student and contrastive views already boosts the overall performance. However, a significant additional improvement is caused by adding a subsequent photometric augmentation when generating the student view from the teacher view, making the detection consistency a consistency to photometric augmentations problem. This results in a further +15.1 in MOTA, proving that a by-product of our photometric detection consistency formulation is stabilization of detections through time.

Teacher	Student	Contrastive	DetA	MOTA	HOTA	IDF1	AssA
-	-	-	12.0	-66.4	17.3	18.5	28.9
-	-	-	12.0	-39.9	14.2	13.1	21.7
g	-	g	13.7	-7.4	19.3	21.4	32.3
g	-	g + p	13.5	-5.8	18.9	20.8	31.3
g + p	-	g + p	13.2	-6.8	18.5	20.4	30.5
g	p	g	15.1	7.4	20.2	23.0	32.2
g	p	g + p	**15.2**	**8.3**	**20.6**	**23.7**	**33.1**

Table 7. **Ablation study on different data augmentation settings for DARTH (Average).** We analyze the effect of different data augmentation settings on DARTH on SHIFT → BDD100K. We report the augmentations applied on the Teacher, Student and Contrastive view, chosen from geometric (g) and photometric (p) augmentations as detailed in Section 3.2. For each metric we report its average across all object categories. No Adap. is in gray.

Stronger Source Model. We investigate the impact of a stronger source model by pre-training Faster R-CNN on CrowdHuman (CH) [55] before training QDTrack on MOT17. Although this results in a marginal improvement on the source domain MOT17 (Table 1), it significantly boosts the robustness of the source model by up to +9.4 DetA and +6.8 AssA when tested on DanceTrack or BDD100K compared to the model only trained on MOT17. The experiments on MOT17 (+CH) → DanceTrack (Table 3) and on MOT17 (+CH) → BDD100K (Table 5) demonstrate that, even when starting from a more robust initialization, DARTH still significantly improves No Adap. by up to +4.0 DetA, +4.3 HOTA, and +3.8 AssA.

5. Conclusion

Playing a pivotal role in perception systems for safety-critical applications such as autonomous driving, MOT algorithms must cope with unseen conditions to avoid life-critical failures. In this paper, we introduce DARTH, the first domain adaptation method for multiple object tracking. DARTH provides a holistic framework for TTA of appearance-based MOT by jointly adapting all the tracking components and their intrinsic relationship to the target domain. Our detection consistency formulation adapts the object detection stage by learning global representations on the target domain while enforcing detection consistency to view changes. Our patch contrastive loss adapts the appearance representations to the target domain, fostering discriminative local instance representations suitable for downstream association. Experimental results validate the remarkable effectiveness of DARTH, fostering an all-round improvement to MOT in both the object detection and instance association stages on a variety of domain shifts.

References

[1] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019. 1, 2, 13

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 4, 6

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 2, 13

[4] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2017. 2, 13

[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. 3

[6] Jinkun Cao, Xinshuo Weng, Rawal Khirodkar, Jiangmiao Pang, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. *arXiv preprint arXiv:2203.14360*, 2022. 2, 3, 13

[7] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8869–8878, 2020. 3

[8] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022. 1, 2

[9] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017. 5

[10] 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. 3

[11] MMTracking Contributors. MMTracking: OpenMMLab video perception toolbox and benchmark. https://github.com/open-mmlab/mmtracking, 2020. 12

[12] 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. 3

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[14] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 3

[15] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014. 2

[16] Mohamed Elhoseny. Multi-object detection and tracking (modt) machine learning model for real-time video surveillance systems. *Circuits, Systems, and Signal Processing*, 39(2):611–630, 2020. 1

[17] Andreas Ess, Konrad Schindler, Bastian Leibe, and Luc Van Gool. Object detection and tracking for autonomous navigation in dynamic environments. *The International Journal of Robotics Research*, 29(14):1707–1725, 2010. 1

[18] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *Proceedings of the IEEE international conference on computer vision*, pages 3038–3046, 2017. 2, 13

[19] Tobias Fischer, Jiangmiao Pang, Thomas E Huang, Linlu Qiu, Haofeng Chen, Trevor Darrell, and Fisher Yu. Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking. *arXiv preprint arXiv:2210.06984*, 2022. 1, 2, 13

[20] Adrien Gaidon and Eleonora Vig. Online domain adaptation for multi-object tracking. *arXiv preprint arXiv:1508.00776*, 2015. 2

[21] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[22] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 4

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[24] 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, 13, 14

[25] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *European Conference on Computer Vision*, pages 733–748. Springer, 2020. 3

[26] Weiming Hu, Tieniu Tan, Liang Wang, and Steve Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(3):334–352, 2004. 1

[27] 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. 3

[28] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012. 1

[29] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6092–6101, 2019. 3

[30] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pages 577–593. Springer, 2016. 2

[31] Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese cnn for robust target association. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 33–40, 2016. 1, 2, 13

[32] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data. *arXiv preprint arXiv:2012.05400*, 2020. 3

[33] 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. 2, 6, 7, 15

[34] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4, 6, 13, 14

[35] Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2

[36] Zhizheng Liu, Mattia Segu, and Fisher Yu. Cooler: Class-incremental learning for appearance-based multiple object tracking. In *DAGM German Conference on Pattern Recognition*. Springer, 2023. 2

[37] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2):548–578, 2021. 6

[38] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022. 2, 13

[39] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 2, 6

[40] 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. 1, 2

[41] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020. 2

[42] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 1, 2

[43] Poojan Oza, Vishwanath A Sindagi, Vibashan VS, and Vishal M Patel. Unsupervised domain adaptation of object detectors: A survey. *arXiv preprint arXiv:2105.13502*, 2021. 3

[44] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 821–830, 2019. 6

[45] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 164–173, 2021. 1, 2, 3, 4, 6, 13, 14

[46] Youngmin Park, Vincent Lepetit, and Woontack Woo. Multiple 3d object tracking for augmented reality. In *2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 117–120. IEEE, 2008. 1

[47] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3, 6, 13, 14

[49] Adrian Lopez Rodriguez and Krystian Mikolajczyk. Domain adaptation for object detection via style consistency. *arXiv preprint arXiv:1911.10033*, 2019. 3

[50] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 780–790, 2019. 3

[51] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019. 3

[52] 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. 3

[53] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33:11539–11551, 2020. 2

[54] Mattia Segu, Alessio Tonioni, and Federico Tombari. Batch normalization embeddings for deep domain generalization. *Pattern Recognition*, 135:109115, 2023. 2

[55] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 8

[56] Vishwanath A Sindagi, Poojan Oza, Rajeev Yasarla, and Vishal M Patel. Prior-based domain adaptive object detection for hazy and rainy conditions. In *European Conference on Computer Vision*, pages 763–780. Springer, 2020. 3

[57] Samarth Sinha, Peter Gehler, Francesco Locatello, and Bernt Schiele. Test: Test-time self-training under distribution shift. *arXiv preprint arXiv:2209.11459*, 2022. 2

[58] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20993–21002, 2022. 2, 6

[59] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 2, 13

[60] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: A synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21371–21382, 2022. 1, 2, 6

[61] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, pages 9229–9248. PMLR, 2020. 1, 2

[62] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018. 4

[63] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 1, 2, 6, 7, 15

[64] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 1, 2

[65] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020. 2

[66] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 1, 2, 13

[67] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 6

[68] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 4

[69] Fuming You, Jingjing Li, and Zhou Zhao. Test-time batch statistics calibration for covariate shift. *arXiv preprint arXiv:2110.04065*, 2021. 7

[70] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018. 1, 2, 3, 6, 13

[71] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 659–675. Springer, 2022. 2, 13

[72] Dan Zhang, Jingjing Li, Lin Xiong, Lan Lin, Mao Ye, and Shangming Yang. Cycle-consistent domain adaptive faster rcnn. *IEEE Access*, 7:123903–123911, 2019. 3

[73] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

[74] Yizhe Zhang, Shubhankar Borse, Hong Cai, and Fatih Porikli. Auxadapt: Stable and efficient test-time adaptation for temporally consistent video semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2339–2348, 2022. 7

[75] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022. 1, 2, 3, 13, 14