Task-specific Inconsistency Alignment for Domain Adaptive Object Detection

Liang Zhao   Limin Wang ✉

State Key Laboratory for Novel Software Technology, Nanjing University, China

liangzhao@smail.nju.edu.cn, lmwang@nju.edu.cn

Abstract

*Detectors trained with massive labeled data often exhibit dramatic performance degradation in some particular scenarios with data distribution gap. To alleviate this problem of domain shift, conventional wisdom typically concentrates solely on reducing the discrepancy between the source and target domains via attached domain classifiers, yet ignoring the difficulty of such transferable features in coping with both classification and localization subtasks in object detection. To address this issue, in this paper, we propose **Task-specific Inconsistency Alignment** (TIA), by developing a new alignment mechanism in separate task spaces, improving the performance of the detector on both subtasks. Specifically, we add a set of auxiliary predictors for both classification and localization branches, and exploit their behavioral inconsistencies as finer-grained domain-specific measures. Then, we devise task-specific losses to align such cross-domain disagreement of both subtasks. By optimizing them individually, we are able to well approximate the **category-** and **boundary-wise** discrepancies in each task space, and therefore narrow them in a decoupled manner. TIA demonstrates superior results on various scenarios to the previous state-of-the-art methods. It is also observed that both the classification and localization capabilities of the detector are sufficiently strengthened, further demonstrating the effectiveness of our TIA method. Code and trained models are publicly available at* https://github.com/MCG-NJU/TIA.

1. Introduction

Object detection [13,14,25,28] is manifesting a high demand for massive annotated data, which however, due to either economic or technical reasons, is struggling to be fulfilled in some scenarios. An alternative is to transfer knowledge from a *source* domain depicting general or synthetic scenes to the *target* domain describing particular scenes of interest. Yet, as a consequence of the *domain shift* [33], the
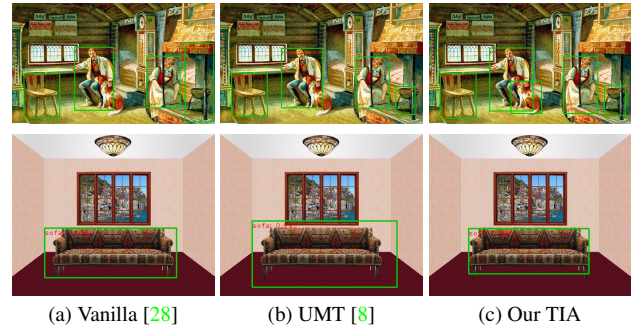

✉: Corresponding author.

(a) Vanilla [28]   (b) UMT [8]   (c) Our TIA

Figure 1. Exampled images from PASCAL VOC [10] → Clipart [18]. Compared with vanilla detector [28], both UMT [8] and TIA identify more foreground objects (Row 1), yet deliver *lower* as well as *higher* quality bounding boxes (Row 2), respectively.

performance of the detector would typically suffer dramatic degradation. A practical strategy to cope with this dilemma is to adopt Unsupervised Domain Adaptation (UDA). Generally, by narrowing the divergence in pixel or feature-level between the source and target domains, a detector trained on source labeled domain can be then well-generalized to unlabeled target domain. This classic strategy of domain alignment, which originated from cross-domain classification [11,20,21,30,37,39], establishes a solid foundation for downstream domain adaptive detection [3,4,8,16,29,32].

Often, as an extension of domain adaptive classifiers, existing domain adaptive detectors focus solely on decreasing the generalization error of their classifiers. Yet, they tend to ignore the **potential improvement** of their localization errors [4,19]. As shown in Fig. 1, compared to vanilla detector, it is observed that the state-of-the-art domain adaptive detector (*i.e.* UMT [8]) is capable of correctly identifying and classifying more foreground objects, but delivering relatively lower quality bounding boxes for them. One possible reason is that, by applying domain alignment via an external binary classifier, the resulted transferable (*i.e.* cross-domain invariant) features grown in the classification space might be harmful for the localization in regression space. Intuitively, the regression space is usually continuous and sparse and has no obvious decision boundaries, hence significantly differs from the classification space.

Motivated by this observation, we argue that *the transferable features induced by previous adaptive detectors fail to cope well with both classification and localization subtasks*. Therefore this paper for the first time, explicitly develops the feature alignment in separate task spaces, in order to seek consistent performance gains on both classification and localization branches. Prevalent two-stage detectors generate a single coupled region of interest (ROI) feature for both subtasks, hindering us from directly applying conventional alignment for each task's feature separately. To overcome this issue, we resort to build multiple auxiliary classifiers and localizers and introduce their behavioral inconsistencies to constitute two task-specific discriminators. In this way, we are able to realize a new decoupled and fine-grained feature alignment by optimizing them separately.

Specifically, we design a general **Task-specific Inconsistency Alignment** (TIA) module to exploit the inconsistency among these new auxiliary predictors and apply it to both subtasks of detectors. Therein, two task-specific losses are devised so that the behavioral disagreement among predictors can be better perceived and easily optimized. In particular, for classification, we establish a stable approximation to the diversity of auxiliary classifiers' decision boundaries with the aid of Shannon Entropy (SE), for effectively shrinking the cross-domain **category-wise** discrepancies. Meanwhile for localization, in consideration of the continuity and sparsity of the regression space, we leverage the Standard Deviation (SD) practically to harvest the ambiguity of various localizers' predictions at each boundary. This allows the **boundary-wise** perception of localizers to be efficiently promoted. Overall, by maximining these two losses, we are able to directly perform inconsistency alignments independently in fully decoupled task spaces, thereby consistently advancing the transferability of features for both classification and localization tasks.

In summary, our contributions fall into threefold: (1) We empirically observe that the resulted features guided by existing feature alignment methods fail to improve the performance of both classification and localization tasks in domain adaptive object detection. To the best of our knowledge, we are the first to address this dilemma by developing domain adaptation into these two branches and directly performing alignment in these two task spaces (not feature space) independently. (2) To effectively perform alignment in task spaces, we propose to build a set of auxiliary predictors and use their behavioral inconsistency for cross-domain alignment. These new inconsistency measures are task-specific and finer-grained, thus expected to better capture the domain difference. (3) Exhaustive experiments have been conducted on various domain shift scenarios, demonstrating superior performance of our framework over state-of-the-art domain adaptive detectors. As shown in Fig. 1 (c), our TIA makes significant progress in both tasks.

2. Related Work

Unsupervised Domain Adaptation (UDA). In light of the basic assumption [1], extensive domain adaption methods have been proposed [11, 20, 21, 30, 37, 39], aiming at learning transferable features to shrink the discrepancy across domains. Recently, several methods [20, 21, 30, 39] have embraced the *consensus regularization* [26] strategy derived from semi-supervised learning. Generally, multiple classifiers with varying initializations are introduced and the inconsistency among their outputs are viewed as an *indicator*, for measuring the divergence between domains. In this way, [20] reduces this disagreement and diversifies the constructed multiple feature embeddings at the same time. [30] then simplifies this procedure, by iteratively maximizing and minimizing the disagreement. On top of them, [21] introduces the wasserstein metric for mining the natural notion of dissimilarity among predictions, while [39, 41] extend the form of [30] and explore in detail the scoring disagreement in the multi-class case. These methods are further generalized to downstream domain adaptation tasks, including semantic segmentation [27, 44] and keypoint detection [19, 45]. In contrast, object detection is a more challenging task in that it is structurally complex and requires the simultaneous optimization of two unparalleled subtasks. Hence, our TIA delves into the *task-specific alignment* and investigates in depth how to accurately bound then reduce both the category-wise disparities and boundary-wise ambiguity within individual task spaces.

UDA for Object Detection. Along the lines of domain adaptive classifiers, the focus of domain adaptive detectors is mostly on bridging the pixel or feature-level divergence between the two domains. Many methods [3, 8, 17, 18, 42] leverage the labeled target-like images generated by CycleGAN [46] to pursue a pixel-level consistency. Yet far more methods [3, 4, 8, 16, 29, 32] are devoted to incrementally reinforcing a feature-level consistency. Nearly all of them explicitly integrate domain-adversarial neural network [11] into the detector, thereby accomplishing feature alignment with simply domain classifiers. [4] initially carries out domain alignment on both backbone features (image-level) and ROI features (instance-level). After that, massive methods [3, 8, 16, 29, 32] continuously strengthen these two alignments, and further improve the performance of the detector with multi-scale [16], contextual [3, 29], spatial attention [22], category attention [38] and cross-domain topological relations [2] information. In addition, [43] and [42] concentrate on enhancing the cross-domain performance of the region proposal network (RPN) to generate high-quality ROIs, whereby the former enforces collaborative training with [30] and self-training on RPN and region proposal classifier, yet the latter construct one set of learnable RPN prototypes for alignment. Problematically, almost all of existing domain adaptive detectors specialize in regulating de-
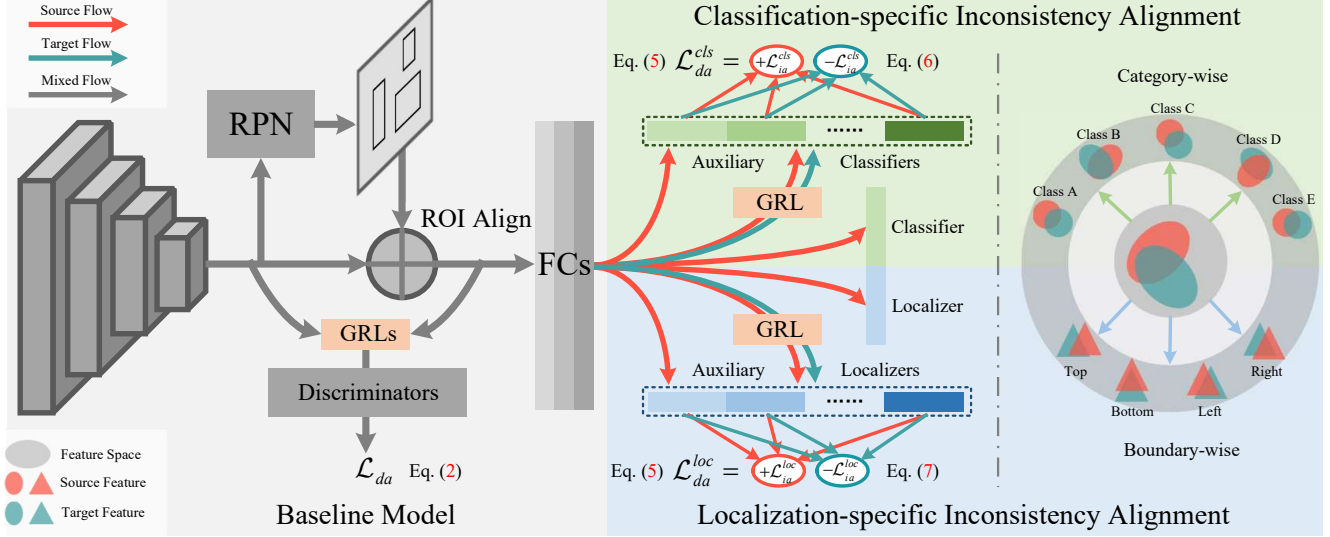
Figure 2. **Framework overview.** Best viewed in color. We develop the high-level feature alignment into separate task spaces, by applying the proposed Task-specific Inconsistency Alignment module to both the classification (*green* part) and localization (*blue* part) branches of the baseline detector (*gray* part). In each branch, the behavioral inconsistency of multiple auxiliary predictors is optimized via the corresponding inconsistency-aware loss, for essentially bridging the category-wise or boundary-wise margins between domains.

cision boundaries of classifiers within detectors, yet ignore the behavioral anomalies of their localizers. In contrast, our TIA first takes this problem into account and develops the general feature alignment into independent task spaces, leading to guaranteed accuracy for each label predictor.

3. Methodology

Following the regular settings of unsupervised domain adaption, we define a labeled source domain $\mathcal{D}_s$ and an unlabeled target domain $\mathcal{D}_t$. Our goal is to establish a knowledge transfer from $\mathcal{D}_s$ to $\mathcal{D}_t$ for object detection, with a favorable generalization over the target domain being guaranteed. In this section, we present technical details of the proposed framework, its overall architecture is illustrated in Fig. 2. We first briefly review the baseline model (left *gray* part) and, on top of it, thoroughly describe the proposed task-specific inconsistency alignment (right *blue* and *green* parts). In the end, some theoretical insights will be raised to explain how our method functions to improve the transferability of both subtasks within the detector.

3.1. Baseline Model

Our framework is implemented on the basis of the popular two-stage detector Faster R-CNN [28], and the *gray* areas in Fig. 2 represent the detector's core structure. Images from both domains are firstly fed into the backbone to yield image-level features, followed by RPN to deliver plentiful proposals, which are then aggregated with the backbone features through ROI Align [14] to generate a certain number of ROIs. With the two ROI predictors on the right

of FCs, the total detection loss can be formally defined as

$$\mathcal{L}_{det} = \mathcal{L}_{rpn} + \mathcal{L}_{roi}. \qquad (1)$$

To pursue the semantic consistency for subsequent modules, we adhere to the mainstream practice of aligning features on the source and target domains, at both mid-to-upper layers of the backbone (*i.e.* image-level) and ROI layer (*i.e.* instance-level). Similar to [3, 4, 32], all these feature alignments are realized by adversarial training, in terms of the domain-adversarial neural network (DANN) [11]. Specifically, features are conveyed via a Gradient Reversal Layer (GRL) to the discriminator D_k for distinguishing their domain labels. The objective is as follows:

$$\mathcal{L}_{da} = \sum_{k=1}^{K} \left(\frac{1}{n_s} \sum_{i=1}^{n_s} L\Big(D_k\big(f_{k,i}\big), d_{k,i}^s\Big) + \frac{1}{n_t} \sum_{i=1}^{n_t} L\Big(D_k\big(f_{k,i}\big), d_{k,i}^t\Big) \right),$$
$$(2)$$

where L is normally a binary cross-entropy loss, $f_{k,i}$ denotes the i-th feature output from the k-th layer and $d_{k,i}$ indicates its corresponding domain label, n_s and n_t refer to the total number of features within a mini-batch in source and target domains, respectively, and K represents the total number of feature alignments. After the above domain adaptation loss being minimized, the sign of gradient back-propagated from discriminator to generator (*e.g.* backbones) is inverted by GRL, guiding generator to deliver cross-domain invariant features so as to confuse discriminator and maximize the loss. The overall objective of the baseline model can be formulated as

$$\mathcal{L} = \mathcal{L}_{det} + \lambda_1 \mathcal{L}_{da}, \qquad (3)$$

where λ_1 is the trade-off parameter.

Following [3, 8], we further interpolate the input to encourage the pixel-level consistency. Specifically, we augment the source domain, by mixing original source images with the target-like source images generated using Cycle-GAN [46]. In summary, we build a very competitive baseline model with feature-level and pixel-level consistency.

3.2. Task-specific Inconsistency Alignment

Conventional object detectors yield a *single* ROI feature after FCs for both tasks of classification and localization, making it difficult to apply previous feature alignment in this *coupled* space. An intuitive way to perform task-specific alignment is to simply duplicate FCs and then align their outputs to each predictor in a DANN [11] manner. However, as discussed in Sec. 5.1, such an alternative poorly decouples the task spaces and leads to insufficient alignments. More importantly, it still suffers from the lack of task-specific treatment, especially for localization task.

Following [30, 39, 41], we propose the Task-specific Inconsistency Alignment to directly shrink the task-specific divergence between source and target domains. This module can be applied to both the classification and localization heads independently, as illustrated in the *blue* and *green* regions. Rather than externally attaching additional discriminators, we use a set of auxiliary predictors to estimate the inconsistency of each domain. By aligning them, our method can not only yield an easier approximation to domain distance, but also come up with a more natural and direct solution to perform alignment in each task space independently for detectors with multiple prediction heads.

Auxiliary predictors. The core of our idea is employing multiple auxiliary predictors to construct an alignment mechanism between domains. Therefore, apart from the primitive classifier C^p and localizer L^p, two additional sets of auxiliary classifiers C^a and localizers L^a composed of N classifiers $C_i^a (1 \leq i \leq N)$ and M localizers $L_j^a (1 \leq j \leq M)$ respectively, are constructed on top of FCs. To ensure a high prediction accuracy, they are all trained with labeled source data as in the primary predictors by the objective:

$$\mathcal{L}_{roi} = \frac{1}{n_s} \sum_{i=1}^{n_s} \left(\sum_{j=1}^{N} \mathcal{L}^{cls}\left(C_j^a(\hat{r}_i), y_i^s\right) + \sum_{j=1}^{M} \mathcal{L}^{loc}\left(L_j^a(\hat{r}_i), b_i^s\right) \right),$$
(4)

where $\hat{r}_i$ represents the higher-level feature of ROI patches r_i processed by FCs, y_i and b_i indicate the corresponding category label and bounding box, respectively. For $\mathcal{L}^{cls}$ and $\mathcal{L}^{loc}$, the traditional cross-entropy and smooth-L1 losses are used. Notably, the gradients of these auxiliary predictors are *detached* when back-propagating to avoid affecting the training of primitive predictors. In addition, to use these auxiliary predictors to perform inconsistency alignment between source and target domains, some GRLs are inserted between FCs and them to adversarially train the proposed

task-specific inconsistency-aware losses.

3.2.1 Inconsistency Alignment Mechanism

Previous DANN-based methods [11] rely merely on attached binary discriminators to optimize task-agnostic losses. In contrast, our method optimizes fine-grained category- and boundary-wise multi-class losses [9, 40] for inconsistency alignment between domains, by means of discriminators composed of various auxiliary predictors. By nature, our objective, the *alignment* to the inconsistency of auxiliary predictors' behavior (*e.g.* decision boundaries of classifiers), essentially characterizes a more precise estimation to the *margins* across domains [39]. To better perceive this disagreement and perform alignment, we construct an integral and adversarial, single-stage training mechanism with GRL, to cope with detectors that are too sophisticated to perform multi-stage iterative optimization like [30].

Specifically, we initially detect the behavioral **inconsistency** of auxiliary predictors trained on the source domain over the *target* domain, and maximize the proposed **task**-specific **i**nconsistency-**a**ware loss $\mathcal{L}_{ia}^{task}$. With GRL, the gradients back-propagated to generator (*i.e.* FCs) are reversed hence the loss is actually minimized for generator. In this adversarial training, the framework reaches a dynamic *equilibrium* in which the predictors are diversified to better discriminate the discrepancy between domains, yet the generator yields sufficiently transferable features to discourage the judgments of these predictors. In addition, the behavioral **consistency** over the *source* domain of auxiliary predictors is also leveraged in a similar way. We **maximize** the consistency-aware loss (the negative of $\mathcal{L}_{ia}^{task}$), so as to simultaneously diversifying the source domain distribution and strengthening the predictors' capabilities. The entire domain adaptation objective can be described as follows:

$$\mathcal{L}_{da}^{task} = -\frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}_{ia}^{task}\left(P_1^a(\hat{r}_i), P_2^a(\hat{r}_i), ..., P_N^a(\hat{r}_i)\right)$$
$$- \frac{1}{n_s} \sum_{i=1}^{n_s} (-\mathcal{L}_{ia}^{task})\left(P_1^a(\hat{r}_i), P_2^a(\hat{r}_i), ..., P_N^a(\hat{r}_i)\right).$$
(5)

where $task \in \{cls, loc\}$, and $P \in \{C, L\}$, the specific inconsistency measure will be explained in next subsections.

3.2.2 Classification-specific Loss

The first question is how to **capture** this behavioral disagreement among decision boundaries of auxiliary classifiers. Different distances including L1 [30], Kullback-Leibler (KL) [39], and Sliced Wasserstein Discrepancy (SWD) [21] have been utilized to measure the discrepancy between outputs of a *pair* of classifiers, but they are hard to generalize to handle multi-classifier situations. For the score distribution constituted by auxiliary classifications on
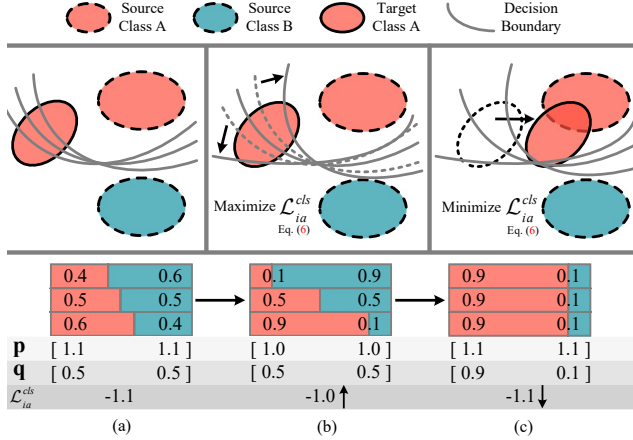
Figure 3. Illustration of the effect when maximin the $\mathcal{L}_{ia}^{cls}$ over target domain in a toy example with two classes and three auxiliary classifiers. Best viewed in color. (a) Initially, the behavior of classifiers is basically consistent with similar decision boundaries; after executing a maximin optimization, we find that: (b) the decision boundaries of the classifiers are mutually exclusive, making the probability distribution on each category sharper and the entropy lower, hence maximizing the loss; (c) the disparity between generated features on class A is shrunk, flattening the probability distribution and increasing the entropy, hence minimizing the loss.

each category, a simple assessment of its *sharpness* or *flatness* is expected. Considering the stability of the optimization and also inspired by [6, 34], we bound it with Shannon Entropy (SE). Concretely, for a probability matrix $\mathbf{M} \in \mathbf{R}^{N \times C}$ of auxiliary predictions, each of its column vectors $\mathbf{m}_i \in \mathbf{R}^N (1 \leq i \leq C)$ represents the predicted probabilities of all classifiers on a particular class i. We can calculate an entropy vector $\mathbf{p} \in \mathbf{R}^C$ — each of whose elements is the entropy calculated from the corresponding softmaxed $\mathbf{m}_i$ — to describe the category-wise variations among various decision boundaries of multiple auxiliary classifiers. Formally, the SE-driven classification-specific inconsistency-aware loss $\mathcal{L}_{ia}^{cls}$ is defined as follows:

$$\mathcal{L}_{ia}^{cls} = -\mathbf{p} \cdot \mathbf{q} = -\sum_{i=1}^{C} \Big(\sum_{j=1}^{N} -\hat{m}_{ij} \log \hat{m}_{ij} \Big) \cdot \Big(\frac{1}{N} \sum_{j=1}^{N} m_{ij} \Big),$$
(6)

where $\hat{\mathbf{m}}_i = \text{softmax}(\mathbf{m}_i)$, and $\mathbf{q}$ indicates the average probability vector. It is notable that the inner product operation between entropy vector and average probability vector is crucial, as weighting the entropy by the confidence of distinct classes keeps the attention on the proper category.

Since the optimization of the inconsistency over target domain is our main objective, we take this process as an example, as depicted in Fig. 3. After solving a maximin game on $\mathcal{L}_{ia}^{cls}$, the behavior of auxiliary classifiers changes first, and driving the probability distribution over each category to flow in a sharper and more deterministic direction.

In this case, the decision boundaries of classifiers are diversified, as shown in Fig. 3 (b). Meanwhile, the generated target domain features are shifted towards the source domain features, flattening the probability distribution. In this context, features are aligned by category in the classification space, so that greater transferability and discriminability are achieved at the same time, as illustrated in Fig. 3 (c).

3.2.3 Localization-specific Loss

The second question lies in, how to catch the behavioral disagreement across various localizers in the **regression** space. Unlike classification, the regression space usually exhibits a continuity and *sparsity*, and the predicted locations are normally heterogeneously clustered in certain regions, rendering it challenging to properly assess the dispersion of the predictions. Some domain adaptation methods [19, 45] dealing with keypoint detection consider that, shrinking the regression space by transformations contributes to alleviating the negative impact of sparsity on the adversarial learning of localizers. Besides, the recent proposed [23, 24], in which the *ambiguity* of multiple localizers' predictions on object boundaries are exploited for detecting anomalous bounding boxes, regard top-k values along with their mean value as a robust representation to the ambiguity.

Practically, in this work, we recommend to choose the most straightforward statistic, the standard deviation (SD) to measure the behavioral inconsistency among auxiliary localization results. This choice is attributed to two reasons. First, two-stage detectors since R-CNN [13] have already well constrained the regression space by *linear transformations*. Second, the L2-norm within SD is more sensitive to outliers, which are crucial for representing the behavioral inconsistency of localizers. The SD-driven localization-specific inconsistency-aware loss $\mathcal{L}_{ia}^{loc}$ can be formulated as

$$\mathcal{L}_{ia}^{loc} = \frac{1}{4 \cdot \sqrt{M}} \sum_{i=1}^{4} \left\| m_i - \frac{1}{M} \sum_{j=1}^{M} m_{ij} \right\|_2,$$
(7)

where $\mathbf{m}_i \in \mathbf{R}^M$ denotes the i-th column vector of the prediction matrix $\mathbf{M} \in \mathbf{R}^{M \times 4}$ constructed by M auxiliary localizers, $\|\cdot\|_2$ indicates the L2-norm.

3.2.4 Overall Objective

Combined with the baseline model, the final objective of the proposed framework becomes

$$\mathcal{L} = \mathcal{L}_{det} + \lambda_1 \mathcal{L}_{da} + \lambda_2 \mathcal{L}_{da}^{cls} + \lambda_3 \mathcal{L}_{da}^{loc},$$
(8)

where λ_1, λ_2 and λ_3 are trade-off parameters for balancing various loss components.

3.3. Theoretical Insights

Tracing the roots, extensive unsupervised domain adaptation methods are motivated by the theoretical analysis in [1], which states the following:

Method	aero	bcycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hrs	bike	prsn	plnt	sheep	sofa	train	tv	mAP
DAF [4]	38.0	47.5	27.7	24.8	41.3	41.2	38.2	11.4	36.8	39.7	12.7	12.7	31.9	47.8	55.6	46.3	12.1	25.6	51.1	45.5	34.7
SWDA [29]	26.2	48.5	32.6	33.7	38.5	54.3	37.1	18.6	34.8	58.3	12.5	12.5	33.8	65.5	54.5	52.0	9.3	24.9	54.1	49.1	38.1
SCL [32]	**44.7**	50.0	33.6	27.4	42.2	55.6	38.3	19.2	37.9	69.0	30.1	26.3	34.4	67.3	61.0	47.9	21.4	26.3	50.1	47.3	41.5
HTCN [3]	33.6	58.9	34.0	23.4	**45.6**	57.0	39.8	12.0	39.7	51.3	20.1	20.1	39.1	72.8	61.3	43.1	19.3	**30.1**	50.2	**51.8**	40.3
SAP [22]	27.4	**70.8**	32.0	27.9	42.4	63.5	47.5	14.3	**48.2**	46.1	**31.8**	17.9	43.8	68.0	**68.1**	49.0	18.7	20.4	55.8	51.3	42.2
UMT [8]	39.6	59.1	32.4	35.0	45.1	61.9	48.4	7.5	46.0	**67.6**	21.4	**29.5**	**48.2**	75.9	70.5	**56.7**	25.9	28.9	39.4	43.6	44.1
DBGL [2]	28.5	52.3	34.3	32.8	38.6	66.4	38.2	**25.3**	39.9	47.4	23.9	17.9	38.9	78.3	61.2	51.7	26.2	28.9	56.8	44.5	41.6
Source Only	35.6	52.5	24.3	23.0	20.0	43.9	32.8	10.7	30.6	11.7	13.8	6.0	36.8	45.9	48.7	41.9	16.5	7.3	22.9	32.0	27.8
Baseline	31.9	56.3	33.4	26.3	40.2	53.3	42.7	17.9	42.3	59.1	15.5	23.6	35.1	85.2	63.2	46.3	22.0	28.4	51.0	48.2	41.1
TIA$_{CLS}$	38.3	51.0	**38.3**	33.2	43.0	65.7	43.8	22.2	43.3	57.1	20.9	23.7	38.9	**89.4**	64.2	53.8	**38.2**	25.0	52.4	50.5	44.7
TIA$_{LOC}$	37.5	55.8	35.3	32.2	**45.6**	63.1	44.1	15.6	44.4	62.1	15.1	26.3	38.5	74.3	65.3	46.9	30.7	27.2	55.5	48.9	43.2
TIA	42.2	66.0	36.9	**37.3**	43.7	**71.8**	**49.7**	18.2	44.9	58.9	18.2	29.1	40.7	87.8	67.4	49.7	27.4	27.8	**57.1**	50.6	**46.3**

Table 1. Experimental results (%) of *Real-to-Artistic* scenario, PASCAL VOC $\rightarrow$ Clipart.

Theorem 1 *Let $\mathcal{H}$ be the hypothesis space and let $\langle \mathcal{D}_s, f_s \rangle$ and $\langle \mathcal{D}_t, f_t \rangle$ be the two domains consisting of a pair of distribution $\mathcal{D}$ and labeling function f. Hence for any $h \in \mathcal{H}$:*

$$\varepsilon_t(h, f_t) \leq \varepsilon_s(h, f_s) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) + \lambda^*, \qquad (9)$$

where ϵ_s (resp. ϵ_t) denotes the disagreement (i.e. error) between the labeling function f_s (resp. f_t) and hypothesis h over the source (resp. target) domain, $d_{\mathcal{H}\Delta\mathcal{H}}$ denotes the $\mathcal{H}\Delta\mathcal{H}$ divergence between domains, λ^* indicates the error of an ideal hypothesis h^*.

Most of existing cross-domain detectors continue the practice in DANN [11] and are dedicated to approximating the optimal $\mathcal{H}$-divergence (including $\mathcal{H}\Delta\mathcal{H}$-divergence) by minimizing the *Jensen-Shannon* divergence [35]. Then, for the two labeling functions (a classifier f^c and a localizer f^l) possessed by all detectors, we have

$$\varepsilon_t(h, f_t^c) \leq \varepsilon_s(h, f_s^c) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) + \lambda^*,$$
$$\varepsilon_t(h, f_t^l) \leq \varepsilon_s(h, f_s^l) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) + \lambda^*. \qquad (10)$$

In this context, by narrowing a single divergence, the target errors of both two labeling functions are restricted, which is however, hard to be done. Since the large differences in classification and regression spaces make it difficult for a single hypothesis to be consistent with both functions simultaneously, and we also empirically found that the target domain error of the localizer is often poorly bounded. In response to this problem, our framework actually **decouples** the optimization of the above **divergence**, and by specifying the hypothesis on each labeling function, consistently reducing the two target errors. Specifically, we have

$$\varepsilon_t(h_1, f_t^c) \leq \varepsilon_s(h_1, f_s^c) + \frac{1}{2} d_{MCSD}^{cls}(\mathcal{D}_s, \mathcal{D}_t) + \lambda^*,$$
$$\varepsilon_t(h_2, f_t^l) \leq \varepsilon_s(h_2, f_s^l) + \frac{1}{2} d_{MCSD}^{loc}(\mathcal{D}_s, \mathcal{D}_t) + \lambda^*, \qquad (11)$$

where d_{MCSD}^{cls} (resp. d_{MCSD}^{loc}) indicates the classification (resp. localization)-specific Multi-Class Scoring Disagreement [39] divergence, which is narrowed when maximining our proposed $\mathcal{L}_{da}^{cls}$ (resp. $\mathcal{L}_{da}^{loc}$).

4. Experiments

4.1. Experimental Setup

Following the default settings in [4, 29], in all experiments, the input image is first resized to have a shorter side length of 600, and then fed into the Faster R-CNN [28] with ROI Align [14]. We train the model using the SGD optimizer with an initial learning rate of 0.001 and divide by 10 every 50k iterations. The batch size is set to 2, one for source domain and one for target domain. For experiments on *Normal-to-Foggy* and *Cross-Camera*, the VGG16 [36] pretrained on ImageNet [7] is employed as the detection backbone, and 70k iterations are trained totally. While for *Real-to-Artistic*, we use the pretrained ResNet101 [15] instead and train a total of 120k iterations. The numbers of auxiliary classifiers (N) and localizers (M) are set to 8 and 4, and the trade-off parameters λ_1, λ_2, and λ_3 are given as 1.0, 1.0 and 0.01, respectively. We report mean average precision (mAP) with a threshold of 0.5 for evaluation.

Various state-of-the-art domain adaptive detectors are introduced for comparison, including DAF [4], SWDA [29], MAF [16], SCL [32], HTCN [3], CST [43], SAP [22], RP-NPA [42], UMT [8], DBGL [2], MeGA [38]. For all these methods, we cite the results in their original paper. To verify the effectiveness of our method, we report the performance of the *Baseline* model and our TIA sequentially. We also train the Faster R-CNN using only the source images, as well as only the annotated target images, and their performance on different scenarios is uniformly referred as *Source Only*, *Target Only*, respectively.

Method	bus	bcycle	car	cycle	person	rider	train	truck	mAP
DAF [4]	35.3	27.1	40.5	20.0	25.0	31.0	20.2	22.1	27.6
SWDA [29]	36.2	35.3	43.5	30.0	29.9	42.3	32.6	24.5	34.3
MAF [16]	39.9	33.9	43.9	29.2	28.2	39.5	33.3	23.8	34.0
SCL [32]	41.8	36.2	44.8	33.6	31.6	44.0	40.7	30.4	37.9
HTCN [3]	47.4	37.1	47.9	32.3	33.2	47.5	40.9	31.6	39.8
CST [43]	45.6	36.8	50.1	30.1	32.7	44.4	25.4	21.7	35.9
SAP [22]	46.8	40.7	**59.8**	30.4	40.8	46.7	37.5	24.3	40.9
RPNPA [42]	43.6	36.8	50.5	29.7	33.3	45.6	42.0	30.4	39.0
UMT [8]	**56.5**	37.3	48.6	30.4	33.0	46.7	46.8	34.1	41.7
MeGA [38]	49.2	39.0	52.4	34.5	37.7	**49.0**	46.9	25.4	41.8
Source Only	22.3	26.5	34.3	15.3	24.1	33.1	3.0	4.1	20.3
Baseline	33.0	**45.7**	47.9	33.3	**45.5**	36.0	35.0	**37.0**	39.2
TIA	52.1	38.1	49.7	**37.7**	34.8	46.3	**48.6**	31.1	**42.3**
Target Only	53.1	36.4	52.8	36.0	36.2	46.5	40.2	34.0	41.9

Table 2. Experimental results (%) of *Normal-to-Foggy* scenario, Cityscapes → Foggy Cityscapes.

4.2. Real to Artistic

In this scenario, we specialize in the migration from trivial real to stylized artistic domains. Typically, to simulate this adaptation, we use both VOC2007-trainval and VOC2012-trainval in PASCAL VOC [10] to construct natural source domain, and Clipart [18] to represent artistic target domain, according to [3, 18, 29]. The Clipart shares 20 categories with PASCAL VOC, totaling 1k images, is employed for both training (without labels) and evaluation.

Tab. 1 shows the results of adaptation from PASCAL VOC to Clipart. It can be observed that our approach outperforms the previous state-of-the-arts with a notable margin (+2.2%), achieving a mAP of 46.3%. Notably, the increase in localization accuracy delivers a consistent improvement over all classes, enabling the highest mean AP with limited categories reaching the highest AP. The overall results showcase that, a finer-grained feature alignment towards high-level abstract semantic inconsistency is essential, especially in such completely dissimilar scene. Also, considering the cross-domain label shifts in the class distribution and the spatial distribution of bounding boxes, our way of shrinking both the category-wise and boundary-wise discrepancies explains the superiority of TIA.

4.3. Normal to Foggy

The capability to accommodate various weather conditions becomes a new expectation for detectors. In this experiment, we use Cityscapes [5] and Foggy Cityscapes [31] as the source and target domains, respectively, to perform a transfer from regular scenes to foggy scenes. Cityscapes comprises 3,475 images, of which 2,975 are training set and the remaining 500 are validation set. Foggy Cityscapes is built on Cityscapes and rendered with the physical model of

Method	KITTI → City	KITTI ← City
DAF [4]	38.5	64.1
SWDA [29]	37.9	71.0
MAF [16]	41.0	72.1
SCL [32]	41.9	72.7
HTCN [3]	42.1	73.2
CST [43]	43.6	-
SAP [22]	43.4	75.2
RPNPA [42]	-	75.1
MeGA [38]	43.0	75.5
Source Only	30.2	53.5
Baseline	42.4	73.0
TIA	**44.0**	**75.9**

Table 3. Experimental results (%) of *Cross-Camera* scenario, KITTI ↔ Cityscapes.

haze, thus both are identical in scenes and annotations. Results are reported in the validation set of Foggy Cityscapes.

According to Tab. 2, our proposed framework TIA obtains the highest mAP (42.3%) over all compared methods, and in particular, our method outperforms the *Target Only* (+0.4%) for the first time. These results demonstrate the importance of aligning task-specific inconsistency. Additionally, taking into account that the benchmark is close to saturation, the performance improvement we achieve relative to the state-of-the-art method (+0.5%) is quite considerable.

4.4. Cross Camera

The domain gap derived from camera differences constitutes a shackle that limits applications of many deep learning algorithms. In this part, we adopt both KITTI [12] which contains 7,481 images and Cityscapes as the source and target domains and transfer them in both adaptation directions. In line with the protocol of [4], we only evaluate detection performance on their common category, *car*.

The AP on detecting cars of various adaptive detectors is reported in Tab. 3. Our method achieves the new state-of-the-art results of 44.0% and 75.9% in both adaptations, also improves +1.6% and +2.9% respectively relative to *Baseline*, manifesting once again the effectiveness and generalization of our approach.

5. Analysis

5.1. Ablation Study

Effect on subtasks. Tab. 1 also demonstrates the effectiveness of TIA on both subtasks of classification and localization. As represented by TIA_{CLS} and TIA_{LOC}, our proposed classification- and localization-specific inconsistency alignment bring consistent improvements (+3.6% and +2.1%). These results indicate that aligning the inconsistency in each task space is effective in enhancing both

Classification	Localization	mAP
-	-	41.1
DANN [11]	DANN [11]	42.6
L1	L1	43.2
KL	L1	43.7
SWD [21]	SWD [21]	44.4
$\mathcal{L}_{ia}^{cls}$ (6)	$\mathcal{L}_{ia}^{loc}$ (7)	46.3

Table 4. Ablation study on the effect on subtasks.

Figure 4. Ablation study on the effect of number of auxiliary predictors N, M.

Figure 5. Error analysis of highest confident detections. (SO refers to *Source Only*)

the category-wise and boundary-wise transferability. Moreover, because of the orthogonality of classification and localization, and boundary-wise alignment always leads to congruent improvement for the detector, revealing the significance of learning a cross-domain localizer.

Effect of inconsistency losses. The superiority of our proposed $\mathcal{L}_{ia}^{cls}$ (6) and $\mathcal{L}_{ia}^{loc}$ (7) is validated by Tab. 4, on the PASCAL VOC → Clipart benchmark. For compared models from third to fifth rows, the number of auxiliary classifiers (N) and localizers (M) is fixed to 2. In comparison to the practical L1 loss used in *MCD* [30], KL divergence used in [39], the recently proposed *SWD* [21], and the intuitive alternative as mentioned in Sec. 3.2, our losses better capture the disagreement and ambiguity of the behavior in auxiliary classifiers and localizers, thus suggesting more precise measures to cross-domain discrepancies. In particular, the use of L1 loss on our framework (shown in the third row) can be regarded as a suitable alternative to MCD [30], since experimentally, it fails to work on detection. Therefore, the improvement of our TIA *w.r.t.* it reveals both the stability of our optimization and the superiority of the proposed losses. Additionally, the way of simply repeating FCs and then aligning their output features to each predictor based on *DANN* [11] (shown in the second row) brings a limited performance gain. This is reasonable since it actually shares substantial layers before FCs, resulting in poor decoupling of the feature space and inadequate task-specific alignment. Meanwhile, it also lacks task-specific treatment and hence still suffers from the rising of localization error in the classification space of the domain discriminator.

Effect of number of auxiliary predictors N, M. To reveal more clearly the impact of number of auxiliary predictors, *i.e.* N or M, on their accordingly tasks, we fix one to 0 and exponentially vary the other from 0 to 32, and Fig. 4 depicts their impact. Obviously, N contributes more to the overall performance than M, suggesting a rational representation of category-wise differences is of greater significance. In addition, in sparse regression spaces, the growth in the number of localizers is of limited benefit in capturing the inconsistency. And we speculate that this is because the localization results are typically heterogeneously clustered
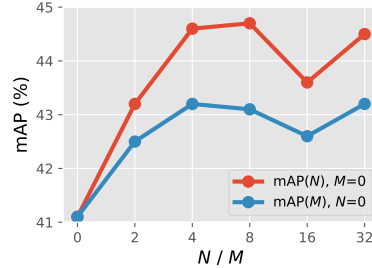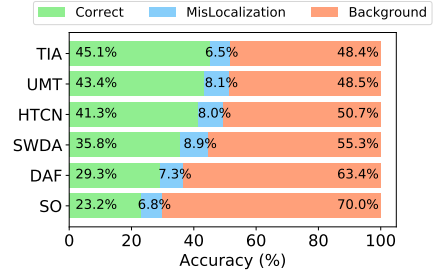
in certain regions.

5.2. Error Analysis

To demonstrate that our framework is capable of promoting the discriminability of features towards both classification and localization tasks, we analyze the accuracies of *Source Only* [28], *DAF* [4], *SWDA* [29], *HTCN* [3], *UMT* [8] and our TIA caused by most confident detections on the Foggy Cityscapes → Cityscapes task. In line with [4], we categorize the detections into 3 types: **Correct** (IoU with GT ≥ 0.5), **MisLocalization** ($0.5 >$IoU with GT ≥ 0.3) and **Background** (IoU with GT <0.3). For each class, we select top-K predictions where K is the number of ground-truth bounding boxes in this class, and report the mean percentage of each type across all categories. As shown in Fig. 5, in comparison with previous mainstream cross-domain detectors, not only does our TIA clearly improve the number of correct detections (*green* color) and reduce the number of false positives, but more importantly, it lowers mislocalization error relative to *Source Only* for the first time (*blue* color). This proves that our TIA boosts the transferability of features while also enhancing their awareness towards both tasks, especially the localization task.

6. Conclusions and Limitations

In this paper, we propose a new method TIA, by developing fine-grained feature alignment in separate task spaces in terms of inconsistency, sufficiently strengthening both the classification and localization capabilities of the detector. Extensive experiments demonstrate the effectiveness of TIA. Nevertheless, the issues of label shift and training stability inherent in domain adaptation still limit TIA, and research in these regards will become future work.

Acknowledgments

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010. 2, 5

[2] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2703–2712, 2021. 2, 6

[3] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8869–8878, 2020. 1, 2, 3, 4, 6, 7, 8

[4] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 1, 2, 3, 6, 7, 8

[5] 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. 7

[6] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3941–3950, 2020. 5

[7] 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

[8] 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. 1, 2, 4, 6, 7, 8

[9] Ürün Dogan, Tobias Glasmachers, and Christian Igel. A unified view on multi-class support vector classification. *J. Mach. Learn. Res.*, 17(45):1–32, 2016. 4

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1, 7

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 1, 2, 3, 4, 6, 8

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 7

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1, 5

[14] 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. 1, 3, 6

[15] 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

[16] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6668–6677, 2019. 1, 2, 6, 7

[17] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 749–757, 2020. 2

[18] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 1, 2, 7

[19] Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. *arXiv preprint arXiv:2103.06175*, 2021. 1, 2, 5

[20] Abhishek Kumar, Prasanna Sattigeri, Kahini Wadhawan, Leonid Karlinsky, Rogerio Feris, William T Freeman, and Gregory Wornell. Co-regularized alignment for unsupervised domain adaptation. *arXiv preprint arXiv:1811.05443*, 2018. 1, 2

[21] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10285–10295, 2019. 1, 2, 4, 8

[22] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 481–497. Springer, 2020. 2, 6, 7

[23] Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11632–11641, 2021. 5

[24] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *arXiv preprint arXiv:2006.04388*, 2020. 5

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Pro-*

ceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017. 1

[26] Ping Luo, Fuzhen Zhuang, Hui Xiong, Yuhong Xiong, and Qing He. Transfer learning from multiple source domains via consensus regularization. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 103–112, 2008. 2

[27] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019. 2

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 1, 3, 6, 8

[29] 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. 1, 2, 6, 7, 8

[30] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018. 1, 2, 4, 8

[31] 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. 7

[32] Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv preprint arXiv:1911.02559*, 2019. 1, 2, 3, 6, 7

[33] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000. 1

[34] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018. 5

[35] Changjian Shui, Qi Chen, Jun Wen, Fan Zhou, Christian Gagné, and Boyu Wang. Beyond $\mathcal{H}$-divergence: Domain adaptation theory with jensen-shannon divergence. *arXiv preprint arXiv:2007.15567*, 2020. 6

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[37] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 1, 2

[38] Vibashan VS, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4516–4526, 2021. 2, 6, 7

[39] Yabin Zhang, Bin Deng, Hui Tang, Lei Zhang, and Kui Jia. Unsupervised multi-class domain adaptation: Theory, algorithms, and practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 4, 6, 8

[40] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413. PMLR, 2019. 4

[41] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5031–5040, 2019. 2, 4

[42] Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12425–12434, 2021. 2, 6, 7

[43] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *European Conference on Computer Vision*, pages 86–102. Springer, 2020. 2, 6, 7

[44] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4):1106–1120, 2021. 2

[45] Xingyi Zhou, Arjun Karpur, Chuang Gan, Linjie Luo, and Qixing Huang. Unsupervised domain adaptation for 3d keypoint estimation via view consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 137–153, 2018. 2, 5

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2, 4