Rethinking Pseudo-Labels: Multi-Granularity Supervision for Domain Adaptive Object Detection

Weimin Lu[1]*, Haohao Ma[1]*, and Qiqi Ge[1]✉

School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai, China
luweimin@sjtu.edu.cn, haohaoma@sjtu.edu.cn, geqq@sjtu.edu.cn

Abstract. Domain Adaptive Object Detection (DAOD) adapts detectors from labeled source domains to unlabeled target domains. Current teacher-student methods rely solely on proposal-level pseudo-boxes as supervision, which suffer from substantial noise under domain shift. We make an intriguing empirical observation: image-level predictions, obtained by aggregating proposal scores, achieve significantly higher accuracy than proposal-level pseudo-boxes. This superiority stems from aggregation's inherent noise reduction and robustness to localization errors. Despite this finding, image-level predictions have been completely overlooked as a supervision source in DAOD, though extensively used in weakly supervised detection. We propose Multi-Granularity Pseudo-Labeling (MGPL), which for the first time incorporates image-level supervision alongside proposal-level pseudo-boxes in teacher-student DAOD. Our framework supervises the student detector at both granularities: proposal-level boxes provide spatial precision while image-level predictions offer reliable category information. We further introduce coarse-to-fine collaborative filtering, where image-level predictions guide class-adaptive threshold selection for proposals, reducing false negatives while maintaining precision. Extensive experiments demonstrate state-of-the-art performance across multiple benchmarks. For example, on Pascal VOC → Clipart1k, we achieve 51.4% mAP, surpassing previous best and fully-supervised Oracle by 2.3% and 6.4% respectively. Our work reveals the untapped potential of multi-granularity learning in domain adaptation. Code will be made publicly available.

Keywords: Domain Adaptive Object Detection · Image Level Prediction · Multi Granularity

1 Introduction

Object detectors trained on labeled source domains typically suffer severe performance degradation when deployed to target domains with different data distributions [8,33]. Acquiring accurate annotations in new domains is prohibitively

[1] * Equal contribution. ✉ Corresponding author.

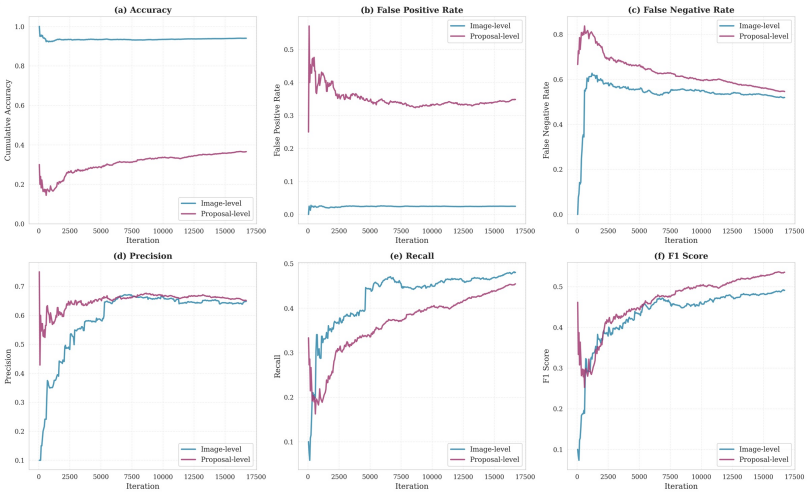

Fig. 1: Comparison of image-level and proposal-level pseudo supervision during the mutual learning stage. Iterations are counted from the beginning of this stage. Image-level predictions consistently yield higher accuracy and precision with comparable recall, indicating more reliable pseudo labels.

expensive and time-consuming. To address the domain gap, Domain Adaptive Object Detection (DAOD) has emerged as a critical research direction. Early approaches explored adversarial learning [8,33,45,49] or image-to-image translation strategies [51, 54] to align feature distributions across domains. Despite notable improvements, these methods did not fully exploit self-training on unlabeled target data. Recently, teacher-student frameworks [4,7,11,12,17,19,20,24,27,38,50] have become the prevailing paradigm in DAOD, where a teacher model generates proposal-level pseudo-boxes on target images, filters them by confidence thresholds, and uses them to supervise a student detector. These methods have achieved remarkable success, with some even surpassing fully-supervised baselines trained directly on target domains [4, 27].

However, teacher-student methods face two fundamental challenges inherent to proposal-level pseudo-labeling. First, pseudo-box quality cannot be guaranteed under domain shift—proposals often suffer from substantial classification and localization errors, introducing noise that degrades student learning [21, 22, 32]. Second, fixed confidence thresholds for filtering pseudo-labels make it difficult for models to learn from challenging instances such as small or occluded objects [27,32], which tend to receive lower confidence scores and are thus filtered out. These limitations raise a critical question: *Are proposal-level pseudo-boxes the optimal choice for supervision in DAOD?*

To answer this question, we draw inspiration from weakly supervised object detection (WSOD), where only image-level labels are available during training [2,36,37]. We investigate whether coarse-grained image-level predictions can serve as an additional source of pseudo-supervision in DAOD, which is easier to

obtain. As shown in Fig. 1, on the challenging Clipart1k benchmark, image-level predictions reach higher accuracy than proposal-level pseudo-boxes. However, relying solely on image-level predictions would discard essential spatial localization information. This motivates us to explore a multi-granularity framework that combines both supervision sources to address the inherent limitations of single-granularity pseudo-labeling.

We propose Multi-Granularity Pseudo-Labeling (MGPL), a teacher-student framework that synergizes coarse-grained (image-level) and fine-grained (proposal-level) pseudo-labels to effectively bridge the domain gap. In our framework, the teacher generates both image-level predictions and proposal-level pseudo-boxes on target images. The student detector is supervised at both granularities: image-level predictions guide the model to learn global distribution and classification information, while proposal-level pseudo-boxes provide precise spatial localization cues. These two granularities of supervision complement and mutually enhance each other. On one hand, coarse-grained learning establishes a strong baseline and can inform adaptive threshold selection for proposal filtering, thereby generating higher-quality fine-grained pseudo-labels and reducing false positives. On the other hand, fine-grained learning with spatial constraints enables the model to produce more accurate image-level predictions. This collaborative mechanism allows our framework to leverage the reliability of image-level predictions while maintaining the spatial precision of proposal-level supervision, leading to substantial performance gains over existing approaches. We summarize our main contributions as follows:

- We propose a multi-granularity pseudo-labeling framework for teacher-student DAOD, where the teacher generates both image-level (coarse-grained) and proposal-level (fine-grained) pseudo-labels, enabling mutual enhancement across granularities.
- We introduce a dual-competition aggregation mechanism to derive reliable image-level predictions from RPN proposals, providing accurate coarse-grained supervision without additional annotation cost.
- We propose a coarse-to-fine collaborative filtering strategy that adaptively refines proposal-level pseudo boxes using image-level predictions, improving supervision reliability without extra hyperparameters.
- Extensive experiments demonstrate that our method achieves state-of-the-art performance across multiple domain adaptation benchmarks. For instance, on Pascal VOC→Clipart1k, we achieve 51.4% mAP, surpassing the previous best method and the fully-supervised Oracle by 2.3% and 6.4%, respectively.

2 Related Work

2.1 Domain Adaptive Object Detection

Unsupervised domain adaptive object detection (DAOD) aims to transfer knowledge from a labeled source domain to an unlabeled target domain. Early approaches mainly relied on adversarial learning to align feature distributions

across domains. Domain Adaptive Faster R-CNN (DA-Faster) [8] introduced domain classifiers at both image and instance levels with gradient reversal layers to learn domain-invariant representations. Subsequent works explored more sophisticated alignment strategies, such as selective cross-domain alignment [55], strong–weak distribution alignment [33], and memory-guided category-aware adaptation in MeGA-CDA [39]. Another line of research attempted to mitigate domain shift through image-to-image translation. For example, Progressive Domain Adaptation [18] leveraged CycleGAN [54] to translate images between domains, while ForkGAN [51] further extended this idea to adverse weather scenarios. Although these approaches improved cross-domain performance, adversarial alignment and generative models often introduce additional training complexity and instability.

More recently, teacher-student frameworks have become the dominant paradigm for DAOD, inspired by their success in semi-supervised learning [29, 38]. Mean Teacher for Object Relation (MTOR) [3] first extended the mean-teacher framework to cross-domain detection through region-level and graph-based consistency constraints. Unbiased Mean Teacher (UMT) [11] combined the mean-teacher framework with CycleGAN-based augmentation to reduce source bias. Adaptive Teacher (AT) [27] further introduced weak–strong augmentation and adversarial feature alignment, achieving substantial improvements. Subsequent works improved pseudo-label reliability from various perspectives, including uncertainty-guided learning in Probabilistic Teacher (PT) [7], contrastive representation learning in Contrastive Mean Teacher (CMT) [4], classification–localization consistency modeling in Harmonious Teacher (HT) [12], masked retraining in MRT [50], and inter-class relation modeling in CAT [20].

Despite their effectiveness, most teacher-student DAOD methods rely almost exclusively on proposal-level pseudo-boxes as supervision for target-domain training. While proposals provide fine-grained spatial information, their quality can degrade under domain shift and they may suffer from high false negative rates when fixed confidence thresholds are applied. Meanwhile, some studies explore multi-granularity representations for domain adaptation [53], but their notion of "multi-granularity" typically refers to different levels of visual features rather than different levels of supervision.

2.2 Image-Level Supervision in Object Detection

Image-level supervision has been widely studied in weakly supervised object detection (WSOD), where only image-level labels are available during training without bounding box annotations. Early WSOD methods formulated detection as a multiple instance learning (MIL) problem [2]. Later approaches improved localization accuracy through iterative refinement and proposal clustering, such as OICR [37] and PCL [36]. Subsequent methods including C-MIL [40] and WSOD2 [48] further enhanced localization performance through spatial constraints and contextual modeling. H2FA [46] bridged weakly supervised detection and domain adaptation through holistic and hierarchical feature alignment.

Beyond WSOD, image-level supervision has also been explored in other detection-related tasks. Class activation maps (CAMs) [52] enable object localization from image-level labels and have been used to generate pseudo-labels for detection [18]. Other studies investigate image-level contrastive learning to learn more discriminative representations for detection pre-training [1]. Image-level signals have also been leveraged in semi-supervised object detection, where they serve as auxiliary supervision to complement scarce instance-level annotations [29]. Furthermore, multi-label classification objectives trained on image-level labels have been shown to improve proposal quality in two-stage detectors by providing richer semantic context beyond what individual proposals can capture [37]. These findings collectively suggest that image-level supervision encodes complementary semantic information that is systematically underexploited in fully supervised and domain adaptive detection pipelines.

However, the learning scenario of DAOD differs fundamentally from WSOD. In WSOD, image-level labels constitute the only supervision signal and the detector must learn object localization from scratch. In contrast, DAOD already obtains proposal-level pseudo-boxes from the teacher model, while image-level predictions can be naturally derived from the same teacher without requiring additional annotation cost. This observation suggests that image-level predictions can serve as complementary supervision to proposal-level pseudo-boxes, potentially improving robustness under domain shift.

3 Proposed Method

3.1 Overview

We adopt the teacher-student framework for unsupervised domain adaptive object detection. The teacher model, updated via exponential moving average (EMA), generates pseudo-labels on weakly-augmented target images to supervise the student model trained on strongly-augmented images. We follow standard practices to update the teacher with momentum 0.9996 and employ image-level adversarial feature alignment to reduce the domain gap.

Unlike existing methods that rely solely on proposal-level pseudo-boxes, our framework leverages both image-level predictions and proposal-level pseudo-boxes as complementary supervision signals. The key insight is that image-level predictions, obtained by aggregating RPN proposals, provide more reliable categorical information while proposal-level boxes offer precise spatial localization. As illustrated in Fig. 2, our framework consists of three core components:

- Image-Level Prediction (Sec. 3.2): We aggregate RPN proposals to derive image-level predictions without introducing additional parameters or architectural modifications.
- Coarse-to-Fine Collaborative Filtering (Sec. 3.3): Image-level predictions guide class-adaptive threshold selection for proposal filtering, reducing false negatives while maintaining precision.

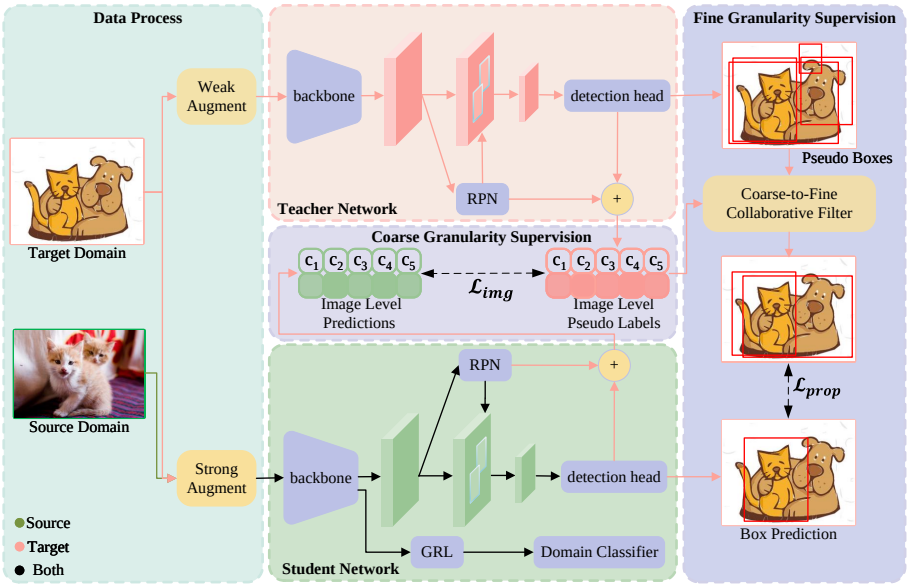

Fig. 2: Overview of the proposed Multi-Granularity Pseudo-Labeling (MGPL) framework. MGPL establishes dual supervision pathways at the image level and instance level, which are tightly coupled by a coarse-to-fine collaborative filtering mechanism. By integrating global semantic guidance with proposal-level localization cues, the framework produces more reliable and consistent pseudo-label supervision for cross-domain object detection.

– Multi-Granularity Supervised Learning (Sec. 3.4): We train the student detector with both image-level and proposal-level pseudo-supervision, enabling the model to benefit from complementary signals at different granularities.

3.2 Image-Level Prediction

To obtain reliable image-level predictions without introducing additional parameters, we propose a dual-competition aggregation mechanism built upon standard Faster R-CNN outputs (Fig. 3). Given N proposals $\{r_i\}_{i=1}^{N}$ with classification logits $x_i \in \mathbb{R}^C$ and objectness score s_i^{obj}, the prediction is derived through two successive competition stages followed by weighted aggregation.

Intra-Proposal Class Competition. We first select the top-k scoring classes for each proposal and normalize over them:

$$P_{i,c} = \frac{\exp(x_i^c) \cdot \mathbf{1}[c \in \mathcal{T}_k(x_i)]}{\sum_{c' \in \mathcal{T}_k(x_i)} \exp(x_i^{c'})}, \tag{1}$$

where $\mathcal{T}_k(x_i)$ denotes the index set of the top-k classes by logit value. This restricts competition to the most plausible categories, suppressing interference from low-scoring classes and yielding sharper semantic probabilities $P_{i,:}$.

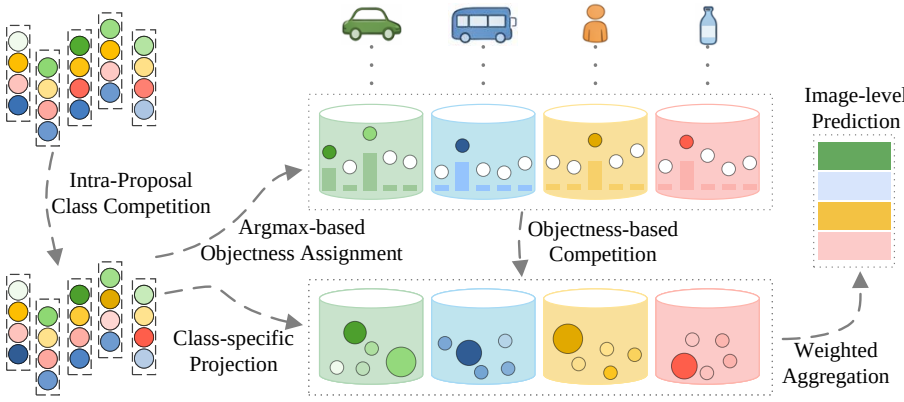

Fig. 3: Illustration of the proposed image-level prediction mechanism. Each proposal first undergoes intra-proposal class competition to obtain semantic probabilities, followed by argmax-based objectness assignment that routes proposal quality to a single class channel. Proposals then compete within each category according to objectness, and the final image-level prediction is obtained via quality-weighted aggregation. The object class logits are from the detection head, and the objectness logits are from the RPN.

Argmax-based Objectness Assignment. For each proposal, we identify its most confident category

$$c_i^* = \arg\max_c x_i^c, \tag{2}$$

and assign its objectness score only to that category:

$$\hat{O}_{i,c} = s_i^{obj} \cdot \mathbf{1}[c = c_i^*], \tag{3}$$

forming a sparse class-wise objectness matrix. This operation routes proposal quality to a single semantic channel without discarding the semantic distribution $P_{i,:}$.

Objectness-based Competition. Within each category, proposals compete according to their objectness scores:

$$w_i^c = \frac{\exp(\hat{O}_{i,c})}{\sum_{j=1}^{N} \exp(\hat{O}_{j,c})}. \tag{4}$$

This class-wise normalization produces reliability-aware weights, highlighting high-quality proposals while suppressing noisy ones.

Quality-weighted Aggregation. Finally, the image-level prediction is obtained as

$$\hat{y}_c = \sum_{i=1}^{N} w_i^c \cdot P_{i,c}. \tag{5}$$

The resulting vector $\hat{y} \in \mathbb{R}^C$ reflects the presence probability of each category in the image.

Algorithm 1 Coarse-to-Fine Collaborative Filtering

Input: Proposals P with classification logits $X \in \mathbb{R}^{N\times C}$, Image-level prediction $\hat{y} \in \mathbb{R}^C$, Histogram bins K
Output: Filtered pseudo-boxes P_{pseudo}, keep mask M

$S \leftarrow \text{Softmax}(X, \text{dim} = 1)$ ▷ $[N, C]$ class probabilities
Joint $\leftarrow S \odot \hat{y}^\top$ ▷ $[N, C]$ collaborative scores
$r_i \leftarrow \max_c \ \text{Joint}_{i,c}$ **for** $i = 1, \ldots, N$ ▷ per-proposal score $[N]$
$c_i^* \leftarrow \arg\max_c \ \text{Joint}_{i,c}$ ▷ predicted class $[N]$
Compute empirical distribution $\{p_k, t_k\}_{k=1}^K$ of $\{r_i\}$
for $k = 1, \ldots, K$ **do**
    $\omega(t_k) \leftarrow \sum_{j\le k} p_j$ ▷ cumulative weight
    $\mu(t_k) \leftarrow \sum_{j\le k} p_j \cdot t_j$ ▷ cumulative mean
    $\sigma_B^2(t_k) \leftarrow \frac{(\mu_T\cdot\omega(t_k)-\mu(t_k))^2}{\omega(t_k)\cdot(1-\omega(t_k))}$
end for
$k^* \leftarrow \arg\max_k \ \sigma_B^2(t_k)$
$\tau^* \leftarrow t_{k^*}$ ▷ adaptive threshold
$M \leftarrow \{\, i : r_i \ge \tau^* \,\}$ ▷ keep mask
$P_{\text{pseudo}} \leftarrow \{\, p_i : i \in M \,\}$
return P_{pseudo}, M

Overall, the proposed mechanism decouples semantic discrimination from proposal quality assessment: semantic competition operates within proposals, while objectness-based competition selects reliable proposals within each category. This separation yields robust image-level predictions without introducing additional learnable parameters.

3.3 Coarse-to-Fine Collaborative Filtering

Standard pseudo-labeling applies a fixed confidence threshold τ to filter proposals, which simultaneously over-filters difficult objects and under-filters categories absent from the image. We replace the hand-tuned threshold with one derived automatically from the collaborative score distribution.

Collaborative Score. Given N proposals with classification logits $\{x_i \in \mathbb{R}^C\}$ and image-level predictions $\hat{y}$, we first convert logits to class probabilities via softmax:

$$P_{i,c} = \frac{\exp(x_i^c)}{\sum_{c'} \exp(x_i^{c'})}, \quad P \in \mathbb{R}^{N\times C}. \tag{6}$$

We then compute a per-proposal collaborative score by element-wise coupling with the image-level prediction:

$$r_i = \max_c \left(P_{i,c} \cdot \hat{y}_c\right), \tag{7}$$

with predicted class $c_i^* = \arg\max_c(P_{i,c} \cdot \hat{y}_c)$. This naturally suppresses proposals whose predicted category is absent from the image (where $\hat{y}_c \approx 0$).

Adaptive Threshold via Inter-Class Variance Maximization. For a candidate threshold t, we define the cumulative weight and mean of scores below t:

$$\omega(t) = \sum_{r_i \le t} p(r_i), \qquad \mu(t) = \sum_{r_i \le t} r_i \cdot p(r_i), \tag{8}$$

where $\mu_T = \sum_i r_i \cdot p(r_i)$ is the total mean. The optimal threshold maximally discriminates the two groups by between-class variance:

$$\sigma_B^2(t) = \frac{\big(\mu_T \cdot \omega(t) - \mu(t)\big)^2}{\omega(t)\big(1-\omega(t)\big)}, \qquad \tau^* = \arg\max_t \ \sigma_B^2(t). \tag{9}$$

A proposal is retained if $r_i \ge \tau^*$. This requires no manually tuned hyperparameters and adapts to each image's score distribution. The complete procedure is outlined in Algorithm 1.

3.4 Multi-Granularity Supervised Learning

To effectively leverage the complementary strengths of coarse-grained semantic cues and fine-grained localization signals, we train the detector with multi-granularity supervision on the target domain. Specifically, the student model is optimized using a combination of supervised learning on labeled source data and pseudo-supervised learning on unlabeled target data.

The overall training objective is defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{src}} + \lambda_{\text{img}}\mathcal{L}_{\text{img}} + \lambda_{\text{prop}}\mathcal{L}_{\text{prop}}, \tag{10}$$

where $\mathcal{L}_{\text{src}}$ denotes the standard supervised detection loss on the source domain, while $\mathcal{L}_{\text{img}}$ and $\mathcal{L}_{\text{prop}}$ represent the image-level and proposal-level pseudo-supervision on target data, respectively.

Source supervision. The source loss $\mathcal{L}_{\text{src}}$ follows the standard Faster R-CNN training objective and consists of the RPN loss, classification loss, and bounding box regression loss. This term ensures that the detector maintains reliable detection capability through fully supervised learning on the labeled source domain.

Image-level supervision. To introduce global semantic guidance during adaptation, we supervise the student's image-level predictions using the teacher's probabilistic predictions as soft pseudo-labels. The image-level loss is formulated as a binary cross-entropy loss:

$$\mathcal{L}_{\text{img}} = -\frac{1}{C}\sum_{c=1}^{C}\left[\hat{y}_c^t \log(\hat{y}_c^s) + (1-\hat{y}_c^t)\log(1-\hat{y}_c^s)\right], \tag{11}$$

where $\hat{y}_c^t$ and $\hat{y}_c^s$ denote the teacher and student image-level predictions for category c, respectively, and C is the number of object categories. This supervision encourages the student to learn global category distributions that are more robust to domain shift.

Proposal-level supervision. For fine-grained classification and localization learning, the proposal-level loss $\mathcal{L}_{\text{prop}}$ supervises the student's detection head using the filtered pseudo-boxes described in Sec. 3.3. Specifically, the pseudo-boxes are treated as supervision signals for the standard detection losses in Faster R-CNN, enabling the model to learn object localization and classification on the target domain.

The weighting coefficients λ_{img} and λ_{prop} control the relative contributions of image-level and proposal-level supervision during training. The detailed hyperparameter settings are described in Sec. 4.2.

4 Experiments

4.1 Datasets and Evaluation

We evaluate our proposed method MGPL on three widely adopted domain adaptive object detection benchmarks, covering diverse types of domain shift.

Pascal VOC → Clipart1k. Pascal VOC [13] is a large-scale real-image dataset containing 20 object categories. Following prior work, we use the combined training split of Pascal VOC 2007 and 2012, comprising 16,551 images in total. Clipart1k [18] is a clipart-style image dataset that shares the same 20 categories as Pascal VOC but exhibits a significant visual domain gap. The training and test sets each contain 500 images. This benchmark is designed to evaluate adaptation from photographic to artistic image styles, representing one of the most challenging domain shifts due to the drastic difference in image appearance.

Cityscapes → Foggy Cityscapes. Cityscapes [9] is an urban street scene dataset collected from 50 cities, containing 2,975 training images and 500 validation images. For detection, 8 object categories are used, with bounding box annotations converted from segmentation masks. Foggy Cityscapes [34] is a synthetic dataset derived from Cityscapes by artificially adding fog at three levels $(0.02, 0.01, 0.005)$ to simulate adverse weather conditions. Following the standard protocol, we train on all fog-density splits and evaluate on the corresponding validation set. This benchmark targets adaptation from normal weather to adverse weather conditions.

KITTI → Cityscapes. KITTI [15] is a driving dataset captured with different camera sensors and in different geographic locations from Cityscapes. We use the training split of 7,481 images from KITTI as the source domain and Cityscapes as the target domain. Since only the "car" category is shared between the two datasets, evaluation is restricted to this single category. This benchmark reflects the challenge of cross-camera and cross-scene domain adaptation.

Evaluation Metric. Following standard practice in domain adaptive object detection, we report the mean Average Precision at an IoU threshold of 0.5 (mAP_{50}) for all benchmarks. For Pascal VOC → Clipart1k and KITTI → Cityscapes, per-class AP is also reported where applicable.

4.2 Implementation Details

Detection framework and backbone. Our approach is built upon the AT framework [27], using Faster R-CNN [31] as the base detector. Following standard protocols in domain adaptive object detection benchmarks, we employ VGG-16 [35] as the backbone for Cityscapes-related tasks (Cityscapes → Foggy Cityscapes and KITTI → Cityscapes), and ResNet-101 [16] for Clipart1k-related benchmarks (Pascal VOC → Clipart1k).

Training setup. All experiments are conducted on two Tesla V100 GPUs, with 4 source images and 4 target images per iteration. We first pre-train the student detector on labeled source-domain data for 40k iterations. The teacher network is then initialized with the student weights, followed by a mutual-learning stage of 100k iterations using both labeled source data and unlabeled target data.

The model is optimized using Stochastic Gradient Descent (SGD) with a learning rate of 0.005 and momentum 0.9. The teacher parameters are updated via an exponential moving average (EMA) of the student parameters with a decay coefficient $\alpha = 0.9996$, following [4,20,27]. Weak–strong augmentation [29] is applied to both source and target images. We set $k = 3$ for the top-k class competition in image-level prediction, which we find robust across benchmarks.

For loss balancing, the weights for proposal-level and image-level supervision are set to $\lambda_{\text{prop}} = 0.5$ and $\lambda_{\text{img}} = 0.5$, respectively. The adversarial alignment loss weight follows the AT baseline [27] with $\lambda_{\text{dis}} = 0.1$. Our implementation is developed based on Detectron2 [43].

4.3 Main Results

Pascal VOC → Clipart1k. Table 1 presents the per-class AP and overall mAP_{50} on the Clipart1k test set. This is the most challenging benchmark due to the large appearance gap between photographic and clipart images. The source-only Faster R-CNN baseline achieves only 28.8% mAP, reflecting the severity of the domain shift. Compared to the strong baseline AT (45.7% mAP), our MGPL consistently improves performance by leveraging richer multi-granularity pseudo-label supervision. The image-level supervision in MGPL provides global semantic context that is complementary to instance-level pseudo-labels, and the adaptive filtering mechanism reduces noise from low-quality target predictions—both crucial for bridging the real-to-artistic gap. MGPL achieves a new state-of-the-art result on this benchmark, outperforming recent methods such as TIA [49], CMT [4], and CAT [20].

Cityscapes → Foggy Cityscapes. Table 2 reports per-class AP on all 8 categories for all fog-density split. This is the most commonly used benchmark in domain adaptive object detection. The source-only model achieves 18.4% mAP, while the Oracle (fully supervised target) reaches 46.3% mAP. Among previous methods, mean-teacher-based approaches such as AT (49.8%) and CMT [4] (51.9%) have steadily pushed performance above the Oracle. Our MGPL further improves upon AT by introducing complementary image-level supervision alongside instance-level pseudo-labels. The image-level branch captures global

Table 1: Detection results (%) on Pascal VOC→Clipart1k.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mtr	prsn	plant	shp	sofa	train	tv	mAP
Source† [27]	23.0	39.6	20.1	23.6	25.7	42.6	25.2	0.9	41.2	25.6	23.7	11.2	28.2	49.5	45.2	46.9	9.1	22.3	38.9	31.5	28.8
Oracle† [27]	33.3	47.6	**43.1**	38.0	24.5	**82.0**	**57.4**	**22.9**	48.4	49.2	37.9	**46.4**	41.1	54.0	73.7	39.5	**36.7**	19.1	53.2	52.9	45.0
I3Net [6]	30.0	67.0	32.5	21.8	29.2	62.5	41.3	11.6	37.1	39.4	27.4	19.3	25.0	67.4	55.2	42.9	19.5	36.2	50.7	39.3	37.8
ICR-CCR [44]	28.7	55.3	31.8	26.0	40.1	63.6	36.6	9.4	38.7	49.3	17.6	14.1	33.3	74.3	61.3	46.3	22.3	24.3	49.1	44.3	38.3
HTCN [5]	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
DM [23]	25.8	63.2	24.5	42.4	47.9	43.1	37.5	9.1	47.0	46.7	26.8	24.9	48.1	78.7	63.0	45.0	21.3	36.1	52.3	53.4	41.8
UMT [11]	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
TIA [49]	**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
AT‡ [27]	33.1	66.1	35.3	44.9	57.5	44.9	51.0	5.8	59.5	54.9	34.6	23.5	64.3	84.0	75.4	51.5	17.1	30.3	43.3	37.2	45.7
CMT [4]	39.8	56.3	38.7	39.7	60.4	35.0	56.0	7.1	**60.1**	60.4	35.8	28.1	**67.8**	84.5	**80.1**	55.5	20.3	32.8	42.3	38.2	47.0
CAT [20]	40.5	64.1	38.8	41.0	60.7	55.5	55.6	14.3	54.7	59.6	46.2	20.3	58.7	**92.9**	62.6	57.5	22.4	40.9	49.5	46.0	49.1
MGPL (Ours)	37.4	**71.5**	41.2	**45.5**	**63.1**	71.9	50.6	17.3	56.3	50.5	**47.1**	16.2	52.3	92.1	75.2	**62.9**	30.7	**46.8**	37.4	**62.4**	**51.4**

† Results from AT [27]. ‡ AT performance is reproduced following [4].

domain characteristics and helps the model better distinguish foreground from background in foggy scenes where per-object visibility is degraded. This leads to consistent improvements across multiple categories, particularly those with heavy fog occlusion such as bus, truck, and train.

KITTI → Cityscapes. Table 2 also reports the AP on the "car" category for the KITTI → Cityscapes benchmark. This benchmark isolates cross-camera and cross-city domain shift without the confound of multi-class imbalance. The source-only model achieves 40.3% AP, far below the Oracle at 64.4% AP. Mean-teacher methods significantly outperform adversarial feature alignment approaches, and MGPL further improves upon the AT baseline. The multi-granularity supervision in MGPL proves beneficial even in this single-category setting: image-level pseudo-labels capture scene-level context that helps the detector attend to car instances appearing in diverse lighting and road conditions across the two datasets. Our method achieves competitive performance compared to recent state-of-the-art methods.

4.4 Analysis and Ablation Study

In this section, we conduct three analyses on the Pascal VOC→Clipart1k benchmark to examine component contributions, supervision balance, and robustness to pseudo-label noise.

Component analysis of multi-granularity supervision. Table 3 reports the main ablation results. Compared with the instance-level baseline (AT, 45.7%), adding *image-level supervision alone* yields 46.1%, suggesting that global semantic cues provide complementary guidance beyond proposal-level pseudo boxes. When combining image-level and instance-level supervision (MGPL w/o Filter), performance increases substantially to 50.6%, indicating that the two granularities are highly synergistic: instance-level pseudo labels contribute spatially precise localization, while image-level signals regularize category learning at the scene level. Finally, enabling the filtering strategy further improves performance to 51.4%, demonstrating that explicitly controlling the reliability of coarse-grained supervision is beneficial for stable cross-domain training.

Sensitivity to supervision balance. We further study the influence of the image-level loss weight λ_{img} in Table 4. Performance improves as λ_{img} increases

Table 2: Comparison (%) with existing methods on Cross-Weather adaptation Cityscapes→Foggy Cityscapes (C→F) and Cross-Fov adaptation KITTI→Cityscapes (K→C).

Methods	Venue	C→F									K→C
		Person	Rider	Car	Truck	Bus	Train	Motor	Bicycle	mAP	mAP
Source† (F-RCNN)	–	27.9	33.4	40.4	12.1	23.2	10.1	20.7	30.9	24.8	40.3
DA-Faster [8]	CVPR 2018	29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0	41.9
VDD [42]	ICCV 2021	33.4	44.0	51.7	33.9	52.0	34.7	34.2	36.8	40.0	–
SCAN [25]	AAAI 2022	41.7	43.9	57.3	28.7	48.6	48.7	31.0	37.3	42.1	45.8
TIA [49]	CVPR 2022	**52.1**	38.1	49.7	37.7	34.8	46.3	**48.6**	31.1	42.3	44.0
LRA [30]	TNNLS 2024	45.6	47.1	59.7	31.2	52.4	44.6	28.1	39.5	43.5	49.4
SIGMA++ [26]	TPAMI 2023	46.4	45.1	61.0	32.1	52.2	44.6	34.8	39.9	44.5	49.5
CIGAR [28]	CVPR 2023	46.1	47.3	62.1	27.8	56.6	44.3	33.7	41.3	44.9	48.5
OADA [47]	ECCV 2022	47.8	46.5	62.9	32.1	48.5	50.9	34.3	39.8	45.4	47.8
PT [7]	ICML 2022	43.2	52.4	63.4	33.4	56.6	37.8	41.3	48.7	47.1	47.8
MTM [41]	AAAI 2024	51.0	53.4	67.2	37.2	54.4	41.6	38.4	47.7	48.9	–
AT‡ [27]	CVPR 2022	46.3	55.9	64.3	38.5	61.1	39.3	40.8	52.3	49.8	–
SOCCER [10]	ACMMM 2024	51.7	57.7	**68.6**	38.2	51.6	47.5	41.6	51.7	51.1	–
CMT [4]	CVPR 2023	47.0	55.7	64.5	39.4	**63.2**	51.9	40.3	53.1	51.9	64.3
DSD-DA [14]	ICML 2024	49.0	**59.6**	65.3	35.7	61.0	46.5	43.9	57.3	52.3	49.3
MGPL (Ours)	–	51.7	56.2	64.7	**40.6**	62.2	**52.6**	44.2	**57.4**	**52.7**	64.3
Oracle† (F-RCNN)	–	41.2	49.1	61.6	32.6	56.6	49.0	37.9	42.4	46.3	**64.4**

† Results from PT [7]. ‡ AT performance is reproduced following [4].

Table 3: Component ablation on Pascal VOC→Clipart1k with ResNet-101 backbone.

Method	Image-level	Filtering	Instance-level	mAP (%)
Instance-level Only (AT)	✗	✗	✓	45.7
Image-level Only	✓	✗	✗	46.1
MGPL (w/o Filter)	✓	✗	✓	50.6
MGPL (Full)	✓	✓	✓	**51.4**

from 0.1 to 0.5, and peaks at $\lambda_{\text{img}} = 0.5$. When λ_{img} becomes too large, performance slightly drops, indicating that overly dominant coarse-grained supervision may weaken the benefit of spatially precise instance-level learning. This unimodal trend supports the need for balanced multi-granularity optimization.

Robustness to pseudo-label noise. To evaluate robustness to corrupted supervision, we inject noise into pseudo-box labels on the Pascal VOC→Clipart1k benchmark. Specifically, at each noise ratio $r \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, we randomly flip a fraction r of pseudo-box class labels to incorrect categories before training.

As shown in Fig. 4, all methods degrade as noise increases, while MGPL degrades more slowly. AT drops from 45.7% to 38.6% ($\Delta = -7.1\%$) under full corruption, whereas MGPL w/o Filter is more stable ($\Delta = -4.5\%$). Moreover, the performance gap between MGPL (Full) and AT widens from +5.7% to +9.3%

Table 4: Sensitivity analysis of the image-level supervision weight λ_{img} on Pascal VOC→Clipart1k.

λ_{img}	0.1	0.3	0.5	1.0	2.0
mAP (%)	48.5	50.5	**51.4**	51.0	50.8

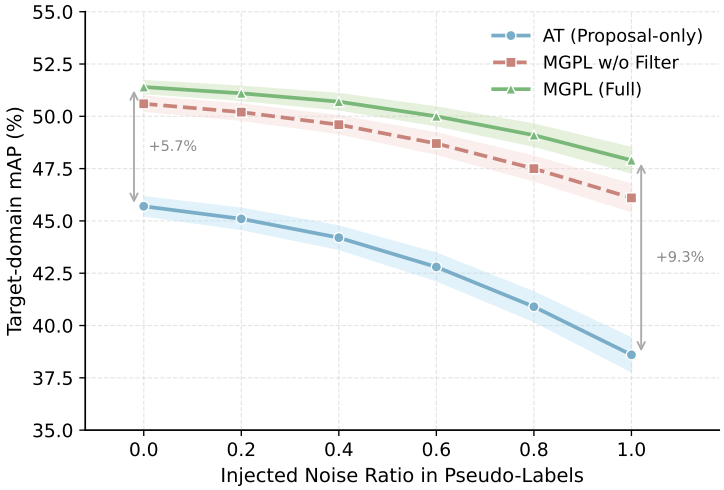

Fig. 4: Robustness to pseudo-label noise on Pascal VOC→Clipart1k. Shadows depict the standard deviation across three runs. We randomly flip a fraction r of pseudo-box class labels before training. MGPL shows slower performance degradation than the proposal-only baseline AT, and its margin over AT increases from +5.7% ($r = 0$) to +9.3% ($r = 1.0$), indicating stronger robustness under corrupted supervision.

as noise intensifies, suggesting that MGPL becomes relatively more advantageous when pseudo labels are less reliable. We attribute this robustness to (i) noise dilution through image-level aggregation across proposals and (ii) the stable semantic guidance provided by $\mathcal{L}_{\text{img}}$, which mitigates error amplification during EMA-based updates.

Summary. Overall, MGPL benefits from (1) synergistic image-level and instance-level supervision, (2) reliability control via filtering, and (3) a balanced supervision weight that stabilizes optimization, while also exhibiting strong robustness under severe pseudo-label corruption.

5 Conclusion

In this paper, we presented MGPL, a teacher-student framework that, for the first time, incorporates image-level predictions as a complementary supervi-

sion signal alongside proposal-level pseudo-boxes for unsupervised domain adaptive object detection. By supervising the student detector at dual granularities and introducing a coarse-to-fine collaborative filtering mechanism, MGPL consistently outperforms existing state-of-the-art methods across three standard benchmarks, achieving 51.4% mAP on Pascal VOC→Clipart1k and surpassing the fully-supervised Oracle by 6.4%. We believe the core idea of leveraging coarse-grained predictions to guide fine-grained learning holds broader potential for other structured prediction tasks under domain shift.

Limitations. Despite the promising results, our image-level prediction mechanism relies on the quality of RPN proposals from a Faster R-CNN detector, making it inherently tied to two-stage detection architectures; extending this dual-granularity supervision to one-stage or transformer-based detectors (e.g., DETR variants) is non-trivial and remains future work.

References

1. Bai, Y., Chen, X., Kirillov, A., Yuille, A., Berg, A.C.: Point-level region contrast for object detection pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16061–16070 (2022)
2. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2846–2854 (2016)
3. Cai, Q., Pan, Y., Ngo, C.W., Tian, X., Duan, L., Yao, T.: Exploring object relation in mean teacher for cross-domain detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11457–11466 (2019)
4. Cao, S., Joshi, D., Gui, L.Y., Wang, Y.X.: Contrastive mean teacher for domain adaptive object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23839–23848 (2023)
5. Chen, C., Zheng, Z., Ding, X., Huang, Y., Dou, Q.: Harmonizing transferability and discriminability for adapting object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8869–8878 (2020)
6. Chen, C., Zheng, Z., Huang, Y., Ding, X., Yu, Y.: I3net: Implicit instance-invariant network for adapting one-stage object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12576–12585 (2021)
7. Chen, M., Chen, W., Yang, S., Song, J., Wang, X., Zhang, L., Yan, Y., Qi, D., Zhuang, Y., Xie, D., et al.: Learning domain adaptive object detection with probabilistic teacher. arXiv preprint arXiv:2206.06293 (2022)
8. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3339–3348 (2018)
9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
10. Cui, Y., Li, L., Zhang, J., Yan, C., Wang, H., Wang, S., Jin, H., Wu, L.: Stochastic context consistency reasoning for domain adaptive object detection. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 1331–1340 (2024)

11. Deng, J., Li, W., Chen, Y., Duan, L.: Unbiased mean teacher for cross-domain object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4091–4101 (2021)
12. Deng, J., Xu, D., Li, W., Duan, L.: Harmonious teacher for cross-domain object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23829–23838 (2023)
13. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
14. Feng, Y., Li, S., Gao, Y., Huang, Z., Zhang, Y., Liu, Q., Wang, Y.: Dsd-da: Distillation-based source debiasing for domain adaptive object detection. arXiv preprint arXiv:2311.10437 (2023)
15. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. He, M., Wang, Y., Wu, J., Wang, Y., Li, H., Li, B., Gan, W., Wu, W., Qiao, Y.: Cross domain object detection by target-perceived dual branch distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9570–9580 (2022)
18. Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-domain weakly-supervised object detection through progressive domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5001–5009 (2018)
19. Kennerley, M., Wang, J.G., Veeravalli, B., Tan, R.T.: 2pcnet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11484–11493 (2023)
20. Kennerley, M., Wang, J.G., Veeravalli, B., Tan, R.T.: Cat: Exploiting inter-class dynamics for domain adaptive object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16541–16550 (2024)
21. Khodabandeh, M., Vahdat, A., Ranjbar, M., Macready, W.G.: A robust learning approach to domain adaptive object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 480–490 (2019)
22. Kim, S., Choi, J., Kim, T., Kim, C.: 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. pp. 6092–6101 (2019)
23. Kim, T., Jeong, M., Kim, S., Choi, S., Kim, C.: Diversify and match: A domain adaptive representation learning paradigm for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12456–12465 (2019)
24. Krishna, O., Ohashi, H., Sinha, S.: Mila: memory-based instance-level adaptation for cross-domain object detection. arXiv preprint arXiv:2309.01086 (2023)
25. Li, W., Liu, X., Yao, X., Yuan, Y.: Scan: Cross domain object detection with semantic conditioned adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1421–1428 (2022)
26. Li, W., Liu, X., Yuan, Y.: Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 9022–9040 (2023)

27. Li, Y.J., Dai, X., Ma, C.Y., Liu, Y.C., Chen, K., Wu, B., He, Z., Kitani, K., Vajda, P.: Cross-domain adaptive teacher for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7581–7590 (2022)
28. Liu, Y., Wang, J., Huang, C., Wang, Y., Xu, Y.: Cigar: Cross-modality graph reasoning for domain adaptive object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23776–23786 (2023)
29. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. arXiv preprint arXiv:2102.09480 (2021)
30. Piao, Z., Tang, L., Zhao, B.: Unsupervised domain-adaptive object detection via localization regression alignment. IEEE Transactions on Neural Networks and Learning Systems **35**(11), 15170–15181 (2023)
31. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
32. RoyChowdhury, A., Chakrabarty, P., Singh, A., Jin, S., Jiang, H., Cao, L., Learned-Miller, E.: Automatic adaptation of object detectors to new domains using self-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 780–790 (2019)
33. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6956–6965 (2019)
34. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision **126**(9), 973–992 (2018)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
36. Tang, P., Wang, X., Bai, S., Shen, W., Bai, X., Liu, W., Yuille, A.: Pcl: Proposal cluster learning for weakly supervised object detection. IEEE transactions on pattern analysis and machine intelligence **42**(1), 176–191 (2018)
37. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2843–2851 (2017)
38. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
39. Vs, V., Gupta, V., Oza, P., Sindagi, V.A., Patel, V.M.: 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. pp. 4516–4526 (2021)
40. Wan, F., Liu, C., Ke, W., Ji, X., Jiao, J., Ye, Q.: C-mil: Continuation multiple instance learning for weakly supervised object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2199–2208 (2019)
41. Weng, W., Yuan, C.: Mean teacher detr with masked feature alignment: A robust domain adaptive detection transformer framework. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5912–5920 (2024)
42. Wu, A., Liu, R., Han, Y., Zhu, L., Yang, Y.: Vector-decomposed disentanglement for domain-invariant object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9342–9351 (2021)

43. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. `https://github.com/facebookresearch/detectron2` (2019)
44. Xu, C.D., Zhao, X.R., Jin, X., Wei, X.S.: Exploring categorical regularization for domain adaptive object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11724–11733 (2020)
45. Xu, M., Wang, H., Ni, B., Tian, Q., Zhang, W.: Cross-domain detection via graph-induced prototype alignment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12355–12364 (2020)
46. Xu, Y., Sun, Y., Yang, Z., Miao, J., Yang, Y.: H2fa r-cnn: Holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14329–14339 (2022)
47. Yoo, J., Chung, I., Kwak, N.: Unsupervised domain adaptation for one-stage object detector using offsets to bounding box. In: European conference on computer vision. pp. 691–708. Springer (2022)
48. Zeng, Z., Liu, B., Fu, J., Chao, H., Zhang, L.: Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8292–8300 (2019)
49. Zhao, L., Wang, L.: Task-specific inconsistency alignment for domain adaptive object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14217–14226 (2022)
50. Zhao, Z., Wei, S., Chen, Q., Li, D., Yang, Y., Peng, Y., Liu, Y.: Masked retraining teacher-student framework for domain adaptive object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19039–19049 (2023)
51. Zheng, Z., Wu, Y., Han, X., Shi, J.: Forkgan: Seeing into the rainy night. In: European conference on computer vision. pp. 155–170. Springer (2020)
52. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
53. Zhou, W., Du, D., Zhang, L., Luo, T., Wu, Y.: Multi-granularity alignment domain adaptation for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9581–9590 (2022)
54. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
55. Zhu, X., Pang, J., Yang, C., Shi, J., Lin, D.: Adapting object detectors via selective cross-domain alignment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 687–696 (2019)