Towards Robust Adaptive Object Detection under Noisy Annotations

Xinyu Liu[1] Wuyang Li[1] Qiushi Yang[1] Baopu Li[2] Yixuan Yuan[1,*]
[1]City University of Hong Kong [2]Baidu USA LLC
{xliu423-c, wuyangli2-c, qsyang2-c}@my.cityu.edu.hk, baopuli@baidu.com
yxyuan.ee@cityu.edu.hk

Abstract

Domain Adaptive Object Detection (DAOD) models a joint distribution of images and labels from an annotated source domain and learns a domain-invariant transformation to estimate the target labels with the given target domain images. Existing methods assume that the source domain labels are completely clean, yet large-scale datasets often contain error-prone annotations due to instance ambiguity, which may lead to a biased source distribution and severely degrade the performance of the domain adaptive detector de facto. In this paper, we represent the first effort to formulate noisy DAOD and propose a Noise Latent Transferability Exploration (NLTE) framework to address this issue. It is featured with 1) Potential Instance Mining (PIM), which leverages eligible proposals to recapture the miss-annotated instances from the background; 2) Morphable Graph Relation Module (MGRM), which models the adaptation feasibility and transition probability of noisy samples with relation matrices; 3) Entropy-Aware Gradient Reconcilement (EAGR), which incorporates the semantic information into the discrimination process and enforces the gradients provided by noisy and clean samples to be consistent towards learning domain-invariant representations. A thorough evaluation on benchmark DAOD datasets with noisy source annotations validates the effectiveness of NLTE. In particular, NLTE improves the mAP by 8.4% under 60% corrupted annotations and even approaches the ideal upper bound of training on a clean source dataset. [1]

1. Introduction

Recent years have witnessed great progress in domain adaptive object detection (DAOD) [6, 17, 19, 23, 38, 47, 49, 57, 58]. It alleviates the performance drop of the detectors when applied to unseen domains due to the domain shift.

(a) (b) (c)

Figure 1. Examples of noisy annotations in Cityscapes dataset. **Miss-annotated** samples: The *bicycle* in (a); the *rider* and *car* in (c). **Class-corrupted** samples: The *rider* and *bicycle* are labeled as *person* in (a); the *motorcycle* is labeled as *bicycle* in (b).

Most DAOD methods are constructed with domain adversarial training [10], in which a domain classifier is proposed to train the feature extractor to perform a domain-invariant transformation of images from different domains. However, existing methods are all built with an ideal condition that a clean source domain is accessible, which is impractical in many real-world applications [22, 26]. The annotations can be noisy due to various reasons, including ambiguous objects caused by occlusion or obscureness, limited crowdsourcing or labeling time, low quality labeled web-crawled images, etc. [8, 30] Frustratingly, the noisy class annotations occur frequently, even in benchmark DAOD source datasets such as Cityscapes, as shown in Fig. 1. The noisy annotations can be categorized into two groups: *miss-annotated* instances (Fig. 1 (a), (c)) and *class-corrupted* instances (Fig. 1 (a), (b)). More specifically, it has been studied that addressing the classification error is critical to the detector [2, 53], thus the noisy class labels in the source dataset could severely damage the domain adaptive detectors.

The intuitive solution for solving the noisy DAOD problem is to combine approaches in learning with noisy labels for classification and domain adaptive object detection. However, this direct combination may encounter several challenges. Firstly, existing methods in learning with noisy labels for image classification [13, 35, 45] minimize or totally filter out the impact of noisy annotated samples during training the network. While in DAOD, the source images with noisy labels are still useful for aligning with target domain as the domain discriminator is class-agnostic, and the target images could benefit source dataset denoising in reverse. If source samples with rich domain-specific

*Corresponding author. This work was supported by Hong Kong Research Grants Council (RGC) General Research Fund 11211221 (CityU 9043152).

[1]Code is available at https://github.com/CityU-AIM-Group/NLTE.

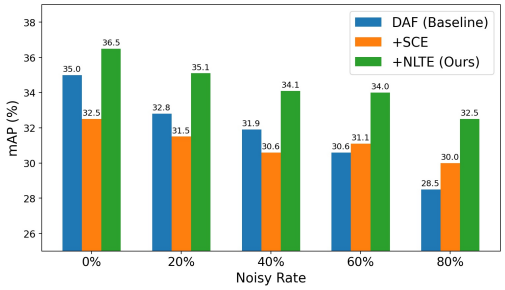


Figure 2. Performance comparison of baseline domain adaptive detector DAF [6], adding noise-robust learning method SCE [45], and adding the proposed NLTE under different noisy rates in Pascal VOC & Noisy Pascal VOC → Clipart1k.

information are filtered out by these methods, the adaptation process will be seriously affected [52]. Secondly, these methods are designed for datasets that are corrupted by class-conditional noise between foreground categories with approximately balanced distributions [29, 45, 56, 59], while noisy DAOD contains diversified noise and imbalanced foreground-background ratio, thus is more complicated and intractable via these methods. Finally, the essentially differed optimization perspective between detection and classification (*i.e.*, the detection task requires multiple losses to work synergistically) [24] makes it non-trival to extend existing noise-robust approaches into DAOD frameworks. Hence they could suffer from underfitting [29] and require elaborately tuning when adopted to the detection framework. We adopt the noise-robust learning approach SCE [45] into the domain adaptive object detector DAF [6] and display the results in Fig. 2. Although mAPs are improved under large noisy rates, SCE severely affects the detector under clean scenario and low noisy rates, which is undesirable for real world applications.

To address the critical yet undeveloped noisy DAOD issue, we propose a novel Noise Latent Transferability Exploration (NLTE) framework to simultaneously address the negative impact caused by *miss-annotated* and *class-corrupted* samples, which facilitates the training of domain adaptive object detectors under noisy source annotations. Specifically, to mine potential *miss-annotated* samples for enriching the source semantic features, we propose Potential Instance Mining (PIM), which looks into the background proposals and dynamically recaptures eligible instances according to their prediction uncertainty. Secondly, considering that the latent domain-related knowledge and semantic information of *class-corrupted* samples are important and attainable for domain alignment, we propose a Morphable Graph Relation Module (MGRM) to leverage their transferable representations for progressively enhancing the discrimination ability of the model. We first conduct intra-domain feature aggregation with morphable graphs, then generate a global relation matrix which is updated by the aggregated node features to model the class-wise transition probability across domains. Afterwards, local relation matrices are built to explore the alignment feasibility of noisy samples, and the transition probabilities of noisy samples are regularized by the global relation matrix. Finally, as both *miss-annotated* and *class-corrupted* noises are contributive to learn domain-invariant representations, we propose an Entropy-Aware Gradient Reconcilement (EAGR) strategy for harmonizing the adaptation procedure of noisy and clean samples. It affiliates class confidence into the discriminator, then enforces the gradients of clean and noisy samples to be consistent towards a domain-invariant direction. Experiments are conducted on both synthetic noisy datasets and real-world scenarios. NLTE outperforms various possible baselines, which validates its effectiveness. With 60% noisy rate, NLTE can significantly improve the mAP of the baseline domain adaptive detector by 8.4%, and only drops by 2% when compared with the clean scenario.

2. Related Works

2.1. Learning with Noisy Annotations

To train a robust model under noisy annotations, different methods have been proposed and can be approximately divided into three categories. The first category is loss correction or adjustment methods [1, 14, 35, 36, 40, 42, 51]. They tried to adjust the loss for each training sample or discard unreliable samples. However, they are prone to false corrections which may further affect the training process, and will suffer from limited eligible data under a high noisy rate. The second is to design symmetric losses that are robust to noise [11, 29, 45, 56, 59]. Generalized Cross Entropy (GCE) [56] combined the merit of MAE [11] and cross entropy loss. Symmetric Cross Entropy (SCE) [45] utilized a weighted summation of the cross entropy loss and the reverse cross entropy loss to make the classification loss symmetric. However, they may only capable to handle certain noisy rates and could collapse under clean scenarios, meanwhile need arduous tuning when applied to the detection task which requires multi-task learning. The last category is to learn a noise transition matrix to rectify the predictions with extra network components [12, 25, 48, 54]. Xia *et al*. [48] approximated the instance-dependent matrix for an instance by a combination of the matrix for the parts of the instance. Li *et al*. [25] consistently estimated the transition matrices without anchor points. However, these methods are based on the assumption that the labels have strong correlations and are designed for ad-hoc situations, such that they are not suitable for the noisy DAOD scenario.

2.2. Domain Adaptive Object Detection

Unsupervised domain adaptive object detection has been widely utilized for narrowing down the domain gaps between labeled source data and unlabelled target data [5, 6,

16, 17, 19, 38, 47, 49, 57, 58]. Chen *et al*. [6] initially used image-level and instance-level adaptations jointly to reduce the domain gap. Xu *et al*. [49] aligned local prototypes and designed a class-reweighted contrastive loss for bridging the domain gap. Zhao *et al*. [57] designed an auxiliary multi-label learning branch for incorporating class to ensure consistent category information between domains. Wu *et al*. [47] disentangled domain-invariant and domain-specific representations based on vector decomposition. However, all previous works assume the source data is clean thus the source category information is reliable for adaptation, while we propose to achieve domain adaptive object detection with a noisy annotated source dataset.

2.3. Learning with Noisy Annotations for Instance Recognition

Chadwick and Newman [4] improved the co-teaching [14] framework for training the object detector with noisy labels. Yang *et al*. [50] described different roles of noisy class labels in the instance segmentation task and used cross entropy or symmetric losses to train them. Li *et al*. [24] utilized the outputs of two diverged classifiers to rectify the bounding boxes and class labels, then trained the model with the corrected labels. However, the previous works attempted to train noise-robust detectors within the same domain, while we propose to tackle the problem of performance deterioration caused by the noisy annotations in the source dataset in the DAOD setting and achieve robust domain adaptive detection.

3. Methodology

3.1. Overview

Problem setting. In noisy DAOD, we are given noisy labeled source dataset $\mathcal{D}^s = \{x_i^s, (y^s, \tilde{y}^s)_i\}_{i=1}^{N_s}$ and unlabeled target dataset $\mathcal{D}^t = \{x_i^t\}_{i=1}^{N_t}$. The labels $y^s, \tilde{y}^s$ both contain box and class annotations, and source and target domains share an identical label space. However, the class annotations in $\tilde{y}^s$ contain noise. The objective is to learn a domain-adaptive object detector that can detects objects in $\mathcal{D}^t$.

How label noise affects domain adaptive object detection. For a domain adaptive detector, it attempts to learn a transformation υ that aligns the conditional distributions of image variable X and label variable Y from different domains, such that $P^t_{\upsilon(X^t)|Y^t} = P^s_{\upsilon(X^s)|Y^s}$. Then with the $P^s_{X^sY^s}$ estimated from $\mathcal{D}^s$ and the distribution of the drawn samples from $P^s_{X^s}$, $P^t_{X^t}$, we are able to estimate $P^t_{Y^t}$, which is the marginal class distribution in the target domain. However, if the source dataset is noisy, then $P^t_{Y^t}$ is computed based on a biased joint distribution $\tilde{P}^s_{X^s\tilde{Y}^s}$ and the learned transformation $\tilde{\upsilon}$ will degrade the performance of domain adaptive object detection.

Concept of the proposed method. Although correcting all noisy labels in the source domain could fully recover the detection performance, it is unattainable practically and may not be optimal for achieving effective domain adaptation. To this end, we build a detection framework that jointly mines miss-annotated samples for recapturing missing semantics and explores the intrinsic positive impact on improving the generalization ability of the detector for class-corrupted samples rather than intuitively correcting the noisy annotations, which is illustrated in Fig. 3.

3.2. Potential Instance Mining

As miss-annotated samples may cause semantic deficiency and limited domain-invariant representations, we propose PIM to recapture potential foreground instances from background in virtue of the Region Proposal Network (RPN). As RPN is class-agnostic, the predicted objectness score of each proposal represents the uncertainty of the existence of an object within the proposal. Therefore, if the proposals have larger objectness scores than thresholds and no intersection with the ground truth boxes, we select them as eligible candidate proposals $\overline{\mathbf{P}}^s$:

$$\overline{\mathbf{P}}^s = \{\overline{p}_i \mid obj(\overline{p}_i) > \tau, \overline{p}_i \notin \mathbf{P}^s, \forall_j IoU(\overline{p}_i, p_j) = 0\}, \tag{1}$$

where τ is the threshold. PIM is also utilized in the target domain to mine confident positive samples $\overline{\mathbf{P}}^t$ for more effective domain alignment. Through the PIM mechanism, only highly-confident proposals are preserved such that missing objects would get recaptured, which simultaneously increases the number of correctly labeled instances for enhancing the discrimination ability and enriches the diversity of source semantic features.

3.3. Morphable Graph Relation Module

To explore the embedded domain knowledge and semantic information within class-corrupted samples, we propose MGRM to model the adaptation feasibility and transition probability of these samples. It regularizes the category-wise relations between noisy local prototypes and global prototypes with morphable graphs. The graphs are built upon features from original proposals generated by RPN $\mathbf{P}^s, \mathbf{P}^t$ and proposals explored by PIM $\overline{\mathbf{P}}^s, \overline{\mathbf{P}}^t$. We omit the domain superscript for clearer explanation if the operations are conducted on both domains.

Intra-domain graph feature aggregation. Given proposals $\mathbf{P} \in \mathbb{R}^{N\times D} \leftarrow \{\mathbf{P}, \overline{\mathbf{P}}\}$ after PIM, we first construct them as intra-domain undirected graphs $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$. Specifically, the vertices correspond to the proposals within each domain, and the edges are defined as the feature cosine similarity between them ($e_{ii'} = \frac{p_i \cdot p_{i'}}{\|p_i\|_2 \cdot \|p_{i'}\|_2}$). Afterwards, we apply intra-domain aggregation to enhance the feature representation within each domain, shown as follows:

$$p_i \leftarrow \sigma\Big(\sum_{i' \in \text{Neighbour}(i)} (w_i p_i e_{ii'} + p_i)\Big), \quad p_i \in \mathbf{P}, \tag{2}$$

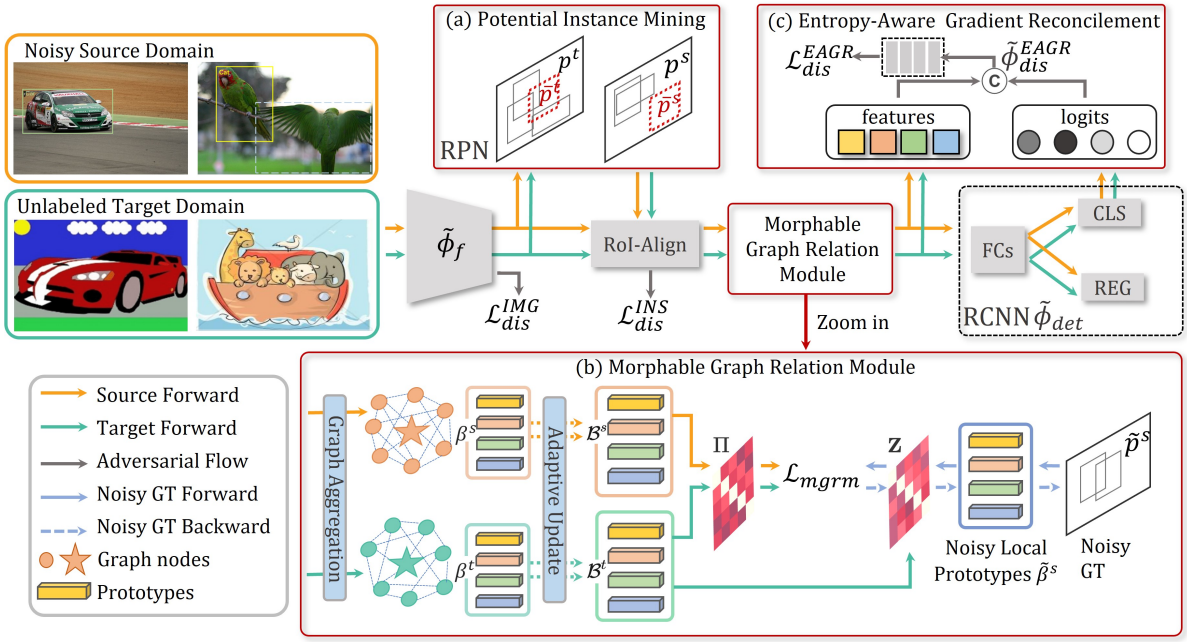


Figure 3. Overview of our NLTE framework, which includes PIM, MGRM and EAGR. ⓒ is the concatenation operation.

where the Neighbour(i) denotes the proposals within the same domain as i, σ is an activation function, and w_i is a learnable weight that maps the original feature dimension D to D'. After the aggregation, proposals that share the common features could improve the feature representation and generate robust proposal features $\mathbf{P}$ with enhanced adaptive contexts within each domain.

Global relation matrix construction. To model the semantic relationship and effectively explore the category-wise transition probability between source and target domains, we introduce a global relation matrix that represents the category-wise affinity between domains. Specifically, considering the source dataset contains noisy annotations and the target dataset is unlabeled, we first utilize confident proposals after aggregation $\mathbf{P}$ to assemble as batch-wise prototypes, which correspond to the class-wise feature centroids:

$$\{\beta_{(u)}\}_{u=1}^{C} = \frac{1}{Card(\mathbf{P}_{(y')})} \sum_{\substack{y'=u \\ p_{(y',i)} \in \mathbf{P}_{(y)}}} p_{(y',i)}, \quad (3)$$

where $Card$ is the cardinality and y' is the most confident category. Then, the correspondence between local and global prototypes is characterized according to their semantic correlation, and an adaptive update operation for generating global prototypes $\{\mathcal{B}^s_{(u)}\}_{u=1}^{C}$ and $\{\mathcal{B}^t_{(v)}\}_{v=1}^{C}$ is conducted:

$$\{\mathcal{B}_{(u)}\}_{u=1}^{C} = \sum_{m=1}^{C} (1 - \tau_{(m,u)})\beta_{(m)} + \tau_{(m,u)}\mathcal{B}_{(u)}, \quad (4)$$

where $\tau_{(m,u)}$ is the cosine similarity between the m-th batch-wise prototype and the u-th global prototype. With this adaptive update process, the representation of global prototypes $\{\mathcal{B}^s_{(u)}\}_{u=1}^{C}$ and $\{\mathcal{B}^t_{(v)}\}_{v=1}^{C}$ can be strengthened via robust and compact batch-wise local features. Finally, the global relation matrix $\Pi \in \mathbb{R}^{C\times C}$ is constructed by the cosine similarity between the prototypes, where each entry $\pi_{u,v}$ represents the affinity between the u-th prototype in the source domain and the v-th prototype in the target domain.

Transition probability regularization. During alignment, the class-wise domain knowledge of class-corrupted samples are inequilibrium with correctly-labeled samples. To mitigate this, the transition probabilities of the class-corrupted samples are expected to be regularized by the intrinsic class-wise correspondence between source and target domain. Therefore, we directly extract noisy source proposals features from $\tilde{\mathbf{P}}^s_{(\tilde{y})}$ regarding to their corresponding noisy labels $\tilde{y}$, and generate noisy source local prototype similar to the batch-wise prototypes in Eq. (3):

$$\{\tilde{\beta}^s_{(u)}\}_{u=1}^{C} = \frac{1}{Card(\tilde{\mathbf{P}}^s_{(\tilde{y})})} \sum_{\substack{\tilde{y}=u \\ \tilde{p}^s_{(\tilde{y},i)} \in \tilde{\mathbf{P}}^s_{(\tilde{y})}}} \tilde{p}^s_{(\tilde{y},i)}. \quad (5)$$

Then, we build local relation matrix $\mathbf{Z} \in \mathbb{R}^{C\times C}$ between $\tilde{\beta}^s_{(u)}$ and $\mathcal{B}^t_{(v)}$ to model the transferability of noisy source samples. Each entry is the class-wise transition probability $z_{u,v} = \frac{\tilde{\beta}^s_{(u)} \cdot \mathcal{B}^t_{(v)}}{\|\tilde{\beta}^s_{(u)}\|_2 \cdot \|\mathcal{B}^t_{(v)}\|_2}$. We use $\ell 1$ loss to regularize such transition probability between local relation matrix and global relation matrix:

$$\mathcal{L}_{mgrm} = \frac{1}{r} \sum_{r\in\mathbb{1}(\mathbf{Z})} |z_r - \pi_r|, \quad (6)$$

where $\mathbb{1}(\mathbf{Z})$ refers to the non-zero columns within $\mathbf{Z}$, which indicates the existence of the r-th category within the

batch. Different from other methods that build category-wise graphs [44, 58] or maintain batch-wise graphs with fixed shapes [49] to model the relationship between source and target domains, our proposed MGRM combines the semantic knowledge between source and target domains, and use it to regularize the transition probability of noisy features implicitly. Therefore, the transferable representations of feasible adapted noisy samples can be extracted for achieving effective semantic alignment.

3.4. Entropy-Aware Gradient Reconcilement

Given noisy annotated source domain data $\mathcal{D}^s$, it implicitly comprises a clean subset $\mathcal{D}^s_{cln} = \{x^s_i, y^s_i\}^{N_{cln}}_{i=1}$ and a subset with both *miss-annotated* and *class-corrupted* samples $\mathcal{D}^s_{cpt} = \{x^s_i, \tilde{y}^s_i\}^{N_{cpt}}_{i=1}$, which are drawn from the clean and noisy joint distributions $P^s_{X^sY^s}$ and $\tilde{P}^s_{X^s\tilde{Y}^s}$, respectively. To magnify the effect of learning domain-invariant representations within $\mathcal{D}^s_{cpt}$, we propose an Entropy-Aware Gradient Reconcilement (EAGR) strategy, which first affiliates the class confidence information into the discrimination process, then enforces the gradients of noisy samples to be consistent with the clean ones.

Entropy-aware alignment. To alleviate the performance deterioration on the target domain, the domain adaptive detector conducts a min-max game to yield a saddle-point solution $\left(\hat{\phi}_f, \hat{\phi}_{det}, \hat{\phi}_{dis}\right)$:

$$\begin{aligned}\left(\hat{\phi}_f, \hat{\phi}_{det}\right) &= \arg\min_{\phi_f, \phi_{det}} \mathcal{L}_{det} - \mathcal{L}_{dis}, \\ \left(\hat{\phi}_{dis}\right) &= \arg\min_{\phi_{dis}} \mathcal{L}_{dis},\end{aligned} \tag{7}$$

where $\hat{\phi}_f$, $\hat{\phi}_{det}$, and $\hat{\phi}_{dis}$ refer to the optimal parameters of the feature extractor, the detector, and the discriminator respectively, which compose the entire domain adaptive detection framework $\hat{\phi}$. However, the noisy labels in the source domain will cause an incompatible optimation between (ϕ_f, ϕ_{det}) and (ϕ_{dis}) as the discriminator is class-agnostic, resulting in an insufficient upper boundary of the source risk [28, 43]. A natural solution is to directly map category onto features for discrimination [28, 57], but it could further magnify the effect of noisy labels under the noisy DAOD setting if the detector is biased. Hence, we build an entropy-aware discriminator to alleviate this effect. Specifically, for each source and target proposal feature $p^s_i \in \mathbf{P}^s, p^t_i \in \mathbf{P}^t$ and their corresponding logits $\eta^s_i \in \boldsymbol{\eta}^s, \eta^t_i \in \boldsymbol{\eta}^t$ generated with ϕ_{det}, we concatenate them and feed into a discriminator ϕ^{EAGR}_{dis}, as shown in Fig. 3(c). The loss function of the discriminator is written as:

$$\begin{aligned}\mathcal{L}^{EAGR}_{dis} = -\sum_{i,j} & z log(\phi^{EAGR}_{dis}(p^s_i \copyright \eta^s_i)) + \\ & (1-z) log(\phi^{EAGR}_{dis}(p^t_j \copyright \eta^t_j)),\end{aligned} \tag{8}$$

where z refers to the domain label, which is 1 for source and 0 for target. Considering the entropy criterion $H(\eta) = -\sum_{u=1}^{C} \eta_u \log(\eta_u)$ that quantifies the uncertainty of classifier predictions, the concatenated logits are softly conditioned on the pooled RoI features [21, 31, 34] to implicitly associate each instance to several most related categories. Hence, category information is preserved for discriminators in aligning class-wise semantic features within each domain, meanwhile providing entropy-aware gradients for the subsequent gradient concilement process.

Gradient reconcilement. Given a domain adaptive detector with parameters ϕ and objective function $\mathcal{L}$, we have gradients for different roles of the proposals:

$$\begin{aligned}G^s_{cln} &= \mathbb{E}_{x^s \in D^s_{cln}} \frac{\partial \mathcal{L}[(x^s, y^s); \phi]}{\partial \phi}, \\ G^s_{cpt} &= \mathbb{E}_{x^s \in D^s_{cpt}} \frac{\partial \mathcal{L}[(x^s, \tilde{y}^s); \phi]}{\partial \phi}, \\ G^t &= \mathbb{E}_{x^t \in D^t} \frac{\partial \mathcal{L}[(x^t); \phi]}{\partial \phi},\end{aligned} \tag{9}$$

where G^s_{cln}, G^s_{cpt}, G^t are the gradients provided by clean proposals, noisy proposals, and target proposals, respectively. After the entropy-aware alignment, the gradients G^s_{cln}, G^s_{cpt}, and G^t are conditioned to the class-wise information provided by both noisy labels and the entropy of classifier predictions. Considering that both G^s_{cln} and G^t optimize the feature extractor $\hat{\phi}_f$ and detector $\hat{\phi}_{det}$ in the direction towards learning domain-invariant representations, the value of their inner product $G^s_{cln} \cdot G^t$ is expected to be sufficiently large. Nevertheless, the direction of G^s_{cpt} could not be determined as they are produced by noisy samples. To reliably characterize the domain-invariant portion within G^s_{cpt}, we simultaneously maximize $G^s_{cpt} \cdot G^s_{cln}$ and $G^s_{cpt} \cdot G^t$ to encourage the direction of gradients provided by noisy and clean samples to be consistent. Thus, we are expected to maximize the summation of the above terms:

$$\arg\max_{\phi_f, \phi_{det}, \phi_{dis}} \left(G^s_{cln} \cdot G^s_{cpt} + G^s_{cln} \cdot G^t + G^s_{cpt} \cdot G^t\right). \tag{10}$$

As we cannot directly optimize Eq. (10) with SGD as clean and noisy gradients cannot be explicitly split, meanwhile computing the Hessians (second order derivatives) is computational prohibitive, inspired by [33, 41], we utilize the first-order meta update of the network as an approximation, which could maximize the above inner product between gradients over iterations and avoid splitting G^s_{cln} and G^s_{cpt} :

$$(\phi_f, \phi_{det}) \leftarrow (\phi_f, \phi_{det}) + \lambda(\Delta\tilde{\phi}_f, \Delta\tilde{\phi}_{det}), \tag{11}$$

where $(\Delta\tilde{\phi}_f, \Delta\tilde{\phi}_{det})$ denotes the residual of the parameters before and after multi-step training of the network and λ is the meta weight. The detailed proof of the approximation is provided in the supplementary. With EAGR, the semantic and discrimination information are harmonized into the backpropagation process, and the gradients of distinct samples are encouraged to achieve coherence. Therefore, both clean and noisy samples would be contributive towards learning a domain-invariant object detector.

Table 1. Results (%) of Pascal VOC and Noisy Pascal VOC with different noisy rates (NR) → Clipart1k.

Pascal VOC & Noisy Pascal VOC → Clipart1k																							
NR	Methods	aero	bcycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hrs	bike	prsn	plnt	sheep	sofa	train	tv	mAP	Imprv.
0%	DAF	29.0	45.1	33.3	25.8	28.6	48.0	39.8	12.3	35.3	50.3	22.9	17.4	33.4	33.8	59.2	44.8	20.7	26.0	45.3	49.6	35.0	0.0
	+SCE	26.5	46.4	35.9	24.3	30.9	38.3	34.9	3.1	31.7	49.8	18.2	17.8	25.2	45.4	53.9	43.0	15.7	26.4	43.3	39.3	32.5	-2.5
	+CP	30.3	49.2	29.8	33.2	34.1	45.8	41.1	9.7	35.8	50.7	23.6	14.4	31.7	36.9	54.6	45.8	18.6	29.9	44.8	43.5	35.2	+0.2
	+GCE	31.9	53.2	27.9	25.8	31.0	41.9	39.3	4.3	34.5	46.7	18.1	18.4	30.2	39.1	55.0	44.1	18.1	21.1	43.2	40.7	33.2	-1.8
	+NLTE	39.1	50.3	33.6	34.7	35.0	40.5	44.2	5.9	36.8	45.8	23.1	17.3	31.8	39.5	60.7	45.4	17.9	28.4	49.0	51.3	**36.5**	**+1.5**
20%	DAF	34.0	39.1	32.0	27.3	32.2	39.3	38.9	2.9	34.9	44.9	20.6	14.2	30.8	36.6	53.8	43.8	17.6	23.6	42.8	46.1	32.8	0.0
	+SCE	23.3	42.3	33.1	27.3	28.8	42.4	35.1	4.0	33.0	44.2	14.6	19.4	27.0	40.9	51.1	45.2	14.9	25.8	41.6	34.8	31.5	-1.3
	+CP	29.3	39.6	29.1	28.0	29.4	34.2	42.4	3.9	35.0	39.4	21.2	12.5	32.2	38.9	57.2	43.0	18.6	27.9	40.2	45.0	32.3	-0.5
	+GCE	24.0	42.2	32.4	29.4	31.5	45.5	39.9	6.7	36.5	38.0	16.7	15.3	30.4	37.9	53.6	44.1	13.5	24.6	46.9	43.7	32.6	-0.2
	+NLTE	33.1	47.5	35.5	28.2	33.7	53.8	43.8	4.2	34.2	48.4	19.3	14.6	29.7	47.2	57.1	42.5	17.7	27.7	40.0	44.5	**35.1**	**+2.3**
40%	DAF	24.5	39.4	29.1	26.9	32.8	46.5	40.0	4.7	36.1	42.0	21.3	10.6	27.8	37.3	52.8	39.7	17.5	26.9	36.0	46.2	31.9	0.0
	+SCE	17.9	42.9	29.7	21.8	26.9	41.5	34.2	8.2	29.1	38.8	19.3	19.2	28.9	48.6	50.7	42.5	10.6	20.4	41.6	40.3	30.6	-1.3
	+CP	24.0	40.9	31.1	22.0	31.2	33.0	40.8	4.4	34.6	36.8	18.5	13.8	29.6	41.3	51.7	38.6	14.2	27.8	26.0	37.8	29.9	-2.0
	+GCE	27.0	36.2	31.1	26.0	33.8	42.9	41.2	2.6	37.0	45.3	19.0	17.2	33.0	42.9	54.6	45.7	17.6	21.4	41.9	49.1	33.3	+1.4
	+NLTE	32.8	45.5	30.8	29.8	35.7	43.2	43.0	6.4	32.7	45.9	19.8	10.8	31.1	43.4	56.4	43.3	19.6	24.8	42.5	43.9	**34.1**	**+2.2**
60%	DAF	29.4	33.5	29.7	29.0	27.7	39.5	38.0	2.7	31.9	41.5	19.8	12.9	30.2	37.0	49.7	37.2	12.8	25.5	40.8	44.2	30.6	0.0
	+SCE	22.2	44.0	31.3	28.0	29.8	48.7	31.6	11.0	29.1	30.7	19.7	9.2	25.8	55.9	51.9	41.3	5.7	21.8	49.0	34.9	31.1	+0.5
	+CP	32.2	42.1	31.5	26.3	31.9	42.4	40.5	2.7	31.8	45.2	20.0	12.2	26.5	38.1	51.1	42.3	11.0	25.6	38.4	41.4	31.7	+1.1
	+GCE	26.7	43.3	33.0	28.6	33.8	51.8	38.0	6.3	33.8	41.7	22.2	13.9	33.4	44.9	53.1	43.9	14.5	22.8	38.6	43.7	33.3	+2.7
	+NLTE	33.0	51.9	32.2	31.7	29.9	39.7	43.6	11.0	36.4	40.7	27.0	11.8	30.3	35.3	55.9	42.2	20.8	30.1	34.5	41.2	**34.0**	**+3.4**
80%	DAF	28.2	34.0	29.6	20.8	27.7	45.0	34.4	1.4	31.5	34.1	19.9	9.3	26.2	33.3	46.0	37.4	17.5	20.4	30.6	41.9	28.5	0.0
	+SCE	19.5	32.9	28.9	23.1	34.3	50.6	31.5	4.3	29.5	35.9	19.5	12.6	23.9	56.2	52.6	38.0	8.2	21.7	41.8	35.5	30.0	+1.5
	+CP	25.2	36.1	27.5	29.8	32.5	29.1	34.3	3.2	31.4	37.7	22.3	7.6	30.4	36.5	46.8	35.4	19.9	27.0	29.6	39.1	29.1	+0.6
	+GCE	25.8	32.8	29.2	21.1	28.8	50.0	33.5	1.3	28.9	34.1	21.7	7.7	27.2	46.8	50.1	37.9	7.3	20.2	42.5	36.0	29.1	+0.6
	+NLTE	36.0	45.4	33.5	30.3	27.3	40.5	40.6	2.6	28.3	51.7	20.4	9.5	30.8	43.1	56.6	42.1	17.7	23.3	31.2	38.4	**32.5**	**+4.0**

3.5. Framework Optimization

The framework is trained with the following objective function:

$$\mathcal{L} = \mathcal{L}_{det} + \lambda_{mgrm}\mathcal{L}_{mgrm} + \mathcal{L}_{dis}^{DAF} + \mathcal{L}_{dis}^{EAGR}, \quad (12)$$

where $\mathcal{L}_{det}$ denotes the loss of Faster R-CNN [37] which consists of RPN loss and RCNN loss. $\mathcal{L}_{dis}^{DAF}$ contains the discrimination components in DAF [6]. Hereafter, we adopt meta update in Eq. (11) for achieving gradient reconcilement. During inference, the input images are consecutively fed into (ϕ_f, ϕ_{det}) to obtain the detection results.

4. Experiments

In this section, we first introduce the experimental setup of synthetic noise and real-world noise, then compare NLTE with the baseline DAF [6] and equipping different noise-robust training approaches [35, 45, 56]. Also, NLTE is compared with existing DAOD methods [3, 6, 19, 27, 32, 46, 47, 49, 55, 57, 60] on the real-world noise setting.

4.1. Experimental Setup

4.1.1 Datasets

Pascal VOC & Noisy Pascal VOC. Pascal VOC [9] contains 16,551 images with 20 distinct object categories. As it contains few instances per image and has been extensively verified by human annotators, we consider it as a clean dataset with no noise. Based on the clean Pascal VOC, we randomly add synthetic label noise with different rates to mimic the annotation mistakes. Specifically, we randomly select a portion of samples and substitute them to another random label. Note that if a label is substituted to background, the corresponding instance is directly removed.

Clipart1k & Watercolor2k. Clipart1k [20] contains 1k graphical images and shares the same 20 categories as Pascal VOC. All images are used for both adversarial training and testing. Watercolor2k [20] shares 6 common categories as the Pascal VOC dataset. We use the 1k training set for adversarial training and the remaining 1k for testing.

Cityscapes & Foggy Cityscapes. Cityscapes [7] contains 2,975 images for training and 500 images for validation. As shown in Fig. 1, Cityscapes dataset contains noisy annotations itself, so we treat it as noisy annotated and directly use the training set as the source domain. Foggy Cityscapes [39] is a fog-rendered Cityscapes dataset and we follow [6, 19, 23, 49] to use the validation set as the target domain. As the validation set only contains 500 images, we manually check all images and consider it as a clean dataset.

4.1.2 Training Details

For all experiments, DA Faster R-CNN (DAF) [6] with backbone ResNet-50 [15] is utilized as our baseline UDA object detector. SGD optimizer is used for training the model for 7 epochs, with an initial learning rate 1×10^{-3} and decays by 0.1 after 5 and 6 epochs. Following the common practice [23, 38], we resize the shorter side of the image to 600 during both training and testing unless specified. λ_{mgrm} is set to 0.1. Experiments are conducted on NVIDIA V100 GPUs and PyTorch is used for the implementation.

Table 2. Results (%) of Pascal VOC and Noisy Pascal VOC with different noisy rates (NR) → Watercolor2k.

Pascal VOC & Noisy Pascal VOC → Watercolor2k									
NR	Methods	bcycle	bird	car	cat	dog	prsn	mAP	Imprv.
0%	DAF	65.8	40.4	35.3	30.0	21.5	44.1	39.6	0.0
	+SCE	65.3	36.9	38.3	25.8	18.9	43.2	37.9	-1.7
	+CP	67.1	39.1	34.5	27.2	22.9	45.3	39.4	-0.2
	+GCE	67.3	37.0	39.7	21.9	21.3	46.4	38.9	-0.7
	+NLTE	73.7	36.9	39.9	26.8	22.6	45.3	**40.9**	**+1.3**
20%	DAF	69.1	36.5	25.8	31.0	16.1	44.9	37.2	0.0
	+SCE	62.4	42.6	33.2	32.2	18.5	46.5	39.2	+2.0
	+CP	72	36.5	21.3	18.3	21.1	41.5	35.1	-2.1
	+GCE	62.7	42.5	40.1	26.2	18.8	44.9	39.2	+2.0
	+NLTE	73.7	37.1	35.3	28.1	21.2	44.5	**40.0**	**+2.8**
40%	DAF	68.0	32.9	20.5	19.8	13.6	39.4	32.4	0.0
	+SCE	64.5	36.6	37.8	14.1	14.0	42.8	35.0	+2.6
	+CP	66.0	36.6	17.8	24.0	18.2	39.8	33.7	+1.3
	+GCE	64.3	40.0	34.7	21.3	19.0	43.8	37.2	+4.8
	+NLTE	75.7	37.2	32.5	22.6	24.3	43.1	**39.2**	**+6.8**
60%	DAF	58.6	35.6	16.7	18.8	11.5	40.1	30.2	0.0
	+SCE	68.1	36.3	31.8	21.9	19.7	41.3	36.5	+6.3
	+CP	68.4	30.3	24.0	22.8	9.6	38.7	32.3	+2.1
	+GCE	73.7	33.0	28.7	24.3	20.4	41.2	36.9	+6.7
	+NLTE	69.5	35.4	27.4	28.4	19.8	51.5	**38.6**	**+8.4**
80%	DAF	56.8	36.7	15.6	19.0	14.8	37.8	30.1	0.0
	+SCE	69.4	37.4	22.6	24.3	16.6	34.6	34.2	+4.1
	+CP	49.1	36.1	16.6	13.7	10.1	36.9	27.1	-3.0
	+GCE	62.8	34.3	14.5	13.4	10.7	40.6	29.4	-0.7
	+NLTE	72.7	41.4	6.6	30.5	14.1	47.9	**35.6**	**+5.5**

4.2. Synthetic Noise

Pascal VOC & Noisy Pascal VOC → Clipart1k. We list the results of using Pascal VOC and Noisy Pascal VOC with different noisy rates as the source domain, and Clipart1k as the target domain in Table 1. It is shown that the performance of the baseline domain adaptive detector [6] drops consistently as the noisy rate increases, and drops from 35.0% to 28.5% under 80% noisy annotations. With loss adjustment method CP [35], the performance shows limited improvement and even drops by 0.5% and 2.0% at 20% and 40% noisy rates. With symmetric loss methods SCE [45] and GCE [56], the detector performs better than CP under noisy settings, but they significantly deteriorate the performance of the detector under clean scenario, causing the mAP drops by 2.5% and 1.8%, respectively. However, adding our proposed NLTE not only achieves robust adaptive detection under different noisy rates (2.3% mAP improvement at 20% and 4.0% mAP improvement at 80%), but also guarantees the performance of the domain adaptive detector when the source annotations are clean.

Pascal VOC & Noisy Pascal VOC → Watercolor2k. Table 2 shows the experimental results of Pascal VOC and Noisy Pascal VOC → Watercolor2k. In the clean setting, adding all compared methods [35, 45, 56] perform worse than the baseline DAF [6], indicating that they could suffer from underfitting and hurt the detection performance. While the proposed NLTE improves the mAP by 1.3%, which is attributed to its capacity of promoting the generalization ability through fully utilizing the noisy samples instead of correcting them straightforwardly. As the noise rate increases from 20% to 80%, adopting NLTE consistently improves the mAP by 2.8%, 6.8%, 8.4%, and 5.5%, respectively, while the compared methods show unstable improvements. The results evidently suggest that NLTE efficiently boosts the robustness of domain adaptive object detectors.

Table 3. Results (%) of Cityscapes → Foggy Cityscapes. † denotes larger training and testing scales.

Cityscapes → Foggy Cityscapes									
Methods	prsn	rider	car	truck	bus	train	mtor	bike	mAP
Source-only [37]	26.9	38.2	35.6	18.3	32.4	9.6	25.8	28.6	26.9
DAF [6]$_{CVPR'18}$	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
SC-DA [60]$_{CVPR'19}$	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
MTOR [3]$_{CVPR'19}$	30.6	41.4	44.0	21.9	38.6	40.6	28.3	35.6	35.1
GPA [49]$_{CVPR'20}$	32.9	46.7	54.1	24.7	45.7	41.1	32.4	38.7	39.5
MCAR [57]$_{ECCV'20}$	32.0	42.1	43.9	31.3	44.1	43.4	37.4	36.6	38.8
EPM† [19]$_{ECCV'20}$	41.9	38.7	56.7	22.6	41.5	26.8	24.6	35.5	36.0
RPNPA [55]$_{CVPR'21}$	33.3	45.6	50.5	30.4	43.6	42.0	29.7	36.8	39.0
DSS-UDA [46]$_{CVPR'21}$	42.9	51.2	53.6	33.6	49.2	18.9	36.2	41.8	40.9
DIDN [27]$_{ICCV'21}$	38.3	44.4	51.8	28.7	53.3	34.7	32.4	40.4	40.5
VDD [47]$_{ICCV'21}$	33.4	44.0	51.7	33.9	52.0	40.9	32.3	36.8	39.8
SSAL† [32]$_{NeurIPS'21}$	45.1	47.4	59.4	24.5	50.0	25.7	26.0	38.7	39.6
DAF+NLTE (Ours)	37.0	46.9	54.8	32.1	49.9	43.5	29.9	39.6	41.8
DAF+NLTE (Ours)†	43.1	50.7	58.7	33.6	56.7	42.7	33.7	43.3	**45.4**

4.3. Real-world Noise

Cityscapes → Foggy Cityscapes. We consider Cityscapes as a real-world noisy annotated source dataset for DAOD and directly implement DAF+NLTE in the Cityscapes → Foggy Cityscapes benchmark to validate its effectiveness. As listed in Table 3, DAF+NLTE shows promising improvements over other state-of-the-arts that were tailored for the DAOD task with the same (or less) training epochs and under the same settings. Specially, with larger training and testing scales as EPM [19] and SSAL [32], *i.e.*, setting the short side of the images to 800 pixels, DAF+NLTE achieves an mAP of 45.4%, outperforming SSAL by 5.8%. The results indicate that existing DAOD methods may suffer from biased source data and addressing the noisy annotations is arguably important for achieving effective adaptation.

5. Further Analysis

5.1. Ablation Studies

Table 4 presents the ablation study of the proposed modules in NLTE under small (20%) and large noisy rates (80%) in the Noisy Pascal VOC → Clipart1k setting. We observe that with PIM, the mAP is improved by 1.1% under 20% noisy rate and 0.3% under 80% noisy rate, indicating that addressing miss-annotation samples in source data would benefit DAOD. With MGRM and EAGR, the mAP is further improved, which demonstrates the effectiveness of utlizing class-corrupted samples in domain alignment. Finally, we demonstrate that strengthening the domain adaptive detector with all components in NLTE boosts the mAP by 2.3% and 4.0% in 20% and 80% noisy rates, respectively, which verifies the effectiveness of NLTE.

(a) DAF [6] (b) DAF+SCE [45] (c) DAF+CP [35] (d) DAF+GCE [56] (e) DAF+NLTE (Ours)

Figure 4. Qualitative results with noisy rate 20% on Clipart1k (top row) and Watercolor2k (bottom row).

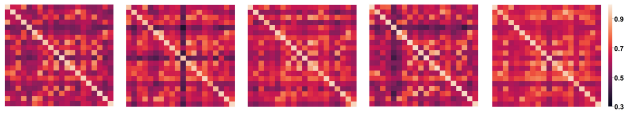
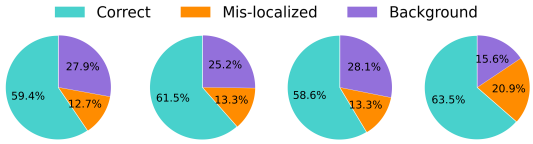


Figure 5. Global relation matrices of Pascal VOC & Noisy Pascal VOC → Clipart1k. From left to right refers to noisy rate 0%, 20%, 40%, 60%, 80%, respectively.



(a) DAF (40%) (b) +NLTE (40%) (c) DAF (60%) (d) +NLTE (60%)

Figure 6. Error analysis of highly confident detections on Noisy Pascal VOC → Watercolor2k. The numbers in brackets refer to noisy rates.

5.2. Qualitative Results

Fig. 4 illustrates the example of detection results with noisy rate 20%. From the figure, NLTE can address the semantic confusion problem via PIM and MGRM and correctly classify obscured objects to avoid false positive detections such as *bottle* and *diningtable* even with large domain shift (top row). Meanwhile, NLTE can also generate accurate bounding boxes for occluded objects such as *person* and leverage noisy annotated samples to learn domain-invariant representations via EAGR (bottom row).

5.3. Visualization of the Relation Matrices

Fig. 5 shows the global relation matrices of Pascal VOC & Noisy Pascal VOC → Clipart1k. The diagonal entries denote the similarity of the prototypes between the same category and others refer to different categories. It is shown that despite the percentage of noisy annotations increases, global relation matrices can still reflect the class-wise transition probability. With MGRM, the transition probabilities of noisy samples are regularized by global relation matrices, and the latent domain-related knowledge and semantic information are conductive to the domain alignment.

Table 4. Ablation studies of the proposed modules in NLTE.

Method	PIM	MGRM	EAGR	20%	80%
Baseline [6]				32.8	28.5
+PIM	✓			33.9	28.8
+PIM+MGRM	✓	✓		34.6	29.3
+PIM+EAGR	✓		✓	34.5	29.5
+PIM+MGRM+EAGR	✓	✓	✓	**35.1**	**32.5**

5.4. Error Analysis of Highly Confident Detections

To further explore the effect of NLTE, we follow [3, 6, 18, 58] to categorize most confident detections into three types: 1) **Correct** (IoU with GT ≥ 0.5), 2) **Mis-localized** ($0.3 \leq$ IoU with GT < 0.5), and 3) **Background** (IoU with GT < 0.3). For each category, we select top-k predictions for analysis, where k is the number of ground truths within the category. Results of mean percentages are shown in Fig. 6. On both 40% and 60% noisy rates, NLTE improves the percentage of correct detections and reduces the percentage of false positives. The analysis demonstrates that adopting NLTE could enhance the ability of the detector in distinguishing different classes under noisy scenarios.

6. Conclusion

In this paper, we address the challenging yet undeveloped issue of domain adaptive object detection under noisy annotations. We propose NLTE, which is a robust adaptive detection framework that simultaneously recaptures miss-annotated samples and explores the transferability of class-corrupted samples. It also harmonizes the gradients between samples for learning domain-invariant representations. Compared with intuitively combining the domain adaptive detector and denoising methods, NLTE shows significant superiority under different noisy rates. Besides, our method outperforms other DAOD methods remarkably in the real-world noise scenario, which implies that addressing the noisy annotations is a suitable and effective alternative to promote the performance of domain adaptive detectors.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *ICML*, pages 312–321, 2019. 2

[2] Ali Borji and Seyed Mehdi Iranmanesh. Empirical upper bound in object detection and more. *arXiv preprint arXiv:1911.12451*, 2019. 1

[3] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *CVPR*, pages 11457–11466, 2019. 6, 7, 8

[4] Simon Chadwick and Paul Newman. Training object detectors with noisy data. In *IV*, pages 1319–1325, 2019. 3

[5] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *CVPR*, pages 8869–8878, 2020. 2

[6] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, pages 3339–3348, 2018. 1, 2, 3, 6, 7, 8

[7] 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 *CVPR*, pages 3213–3223, 2016. 6

[8] Viv Cothey. Web-crawling reliability. *J. Assoc. Inf. Sci. Technol.*, 55(14):1228–1238, 2004. 1

[9] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.*, 111(1):98–136, 2015. 6

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015. 1

[11] Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, volume 31, 2017. 2

[12] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *ICLR*, 2017. 2

[13] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *ECCV*, pages 135–150, 2018. 1

[14] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *arXiv preprint arXiv:1804.06872*, 2018. 2, 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[16] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *ICCV*, pages 6668–6677, 2019. 2

[17] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. In *ECCV*, pages 309–324, 2020. 1, 2

[18] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *ECCV*, pages 340–353. Springer, 2012. 8

[19] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *ECCV*, pages 733–748, 2020. 1, 2, 6, 7

[20] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, pages 5001–5009, 2018. 6

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 5

[22] Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Med. Image Anal.*, 65:101759, 2020. 1

[23] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *ECCV*, pages 481–497, 2020. 1, 6

[24] Junnan Li, Caiming Xiong, Richard Socher, and Steven Hoi. Towards noise-resistant object detection with noisy annotations. *arXiv preprint arXiv:2003.01285*, 2020. 2, 3

[25] Xuefeng Li, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. Provably end-to-end label-noise learning without anchor points. *arXiv preprint arXiv:2102.02400*, 2021. 2

[26] Ying Li, Lingfei Ma, Zilong Zhong, Fei Liu, Michael A Chapman, Dongpu Cao, and Jonathan Li. Deep learning for lidar point clouds in autonomous driving: a review. *IEEE Trans. Neural Netw. Learn.*, 2020. 1

[27] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *ICCV*, pages 8771–8780, 2021. 6, 7

[28] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, pages 1647–1657, 2018. 5

[29] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *ICML*, pages 6543–6553, 2020. 2

[30] Winter Mason and Siddharth Suri. Conducting behavioral research on amazon's mechanical turk. *Behav. Res. Methods*, 44(1):1–23, 2012. 1

[31] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 5

[32] Muhammad Akhtar Munir, Muhammad Haris Khan, M Saquib Sarfraz, and Mohsen Ali. Synergizing between self-training and adversarial learning for domain adaptive object detection. *arXiv preprint arXiv:2110.00249*, 2021. 6, 7

[33] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 5

[34] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, pages 2642–2651, 2017. 5

[35] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017. 1, 2, 6, 7, 8

[36] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014. 2

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 6, 7

[38] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 1, 2, 6

[39] 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. 6

[40] Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In *ICML*, pages 5739–5748, 2019. 2

[41] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021. 5

[42] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *ICML*, pages 5907–5915, 2019. 2

[43] Hui Tang and Kui Jia. Discriminative adversarial domain adaptation. In *AAAI*, volume 34, pages 5940–5947, 2020. 5

[44] Kun Tian, Chenghao Zhang, Ying Wang, Shiming Xiang, and Chunhong Pan. Knowledge mining and transferring for domain adaptive object detection. In *ICCV*, pages 9133–9142, 2021. 5

[45] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, pages 322–330, 2019. 1, 2, 6, 7, 8

[46] Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, YangYang Xia, XiShan Zhang, and ShaoLi Liu. Domain-specific suppression for adaptive object detection. In *CVPR*, pages 9603–9612, 2021. 6, 7

[47] Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, and Yi Yang. Vector-decomposed disentanglement for domain-invariant object detection. *arXiv preprint arXiv:2108.06685*, 2021. 1, 2, 3, 6, 7

[48] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. *NeurIPS*, 33, 2020. 2

[49] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, pages 12355–12364, 2020. 1, 2, 3, 5, 6, 7

[50] Longrong Yang, Fanman Meng, Hongliang Li, Qingbo Wu, and Qishang Cheng. Learning with noisy class labels for instance segmentation. In *ECCV*, pages 38–53. Springer, 2020. 3

[51] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *ICML*, pages 7164–7173. PMLR, 2019. 2

[52] Xiyu Yu, Tongliang Liu, Mingming Gong, Kun Zhang, Kayhan Batmanghelich, and Dacheng Tao. Label-noise robust domain adaptation. In *ICML*, pages 10913–10924. PMLR, 2020. 2

[53] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. Varifocalnet: An iou-aware dense object detector. In *CVPR*, pages 8514–8523, 2021. 1

[54] Yivan Zhang, Gang Niu, and Masashi Sugiyama. Learning noise transition matrix from only noisy labels via total variation regularization. *arXiv preprint arXiv:2102.02414*, 2021. 2

[55] Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *CVPR*, pages 12425–12434, 2021. 6, 7

[56] Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018. 2, 6, 7, 8

[57] Zhen Zhao, Yuhong Guo, Haifeng Shen, and Jieping Ye. Adaptive object detection with dual multi-label prediction. In *ECCV*, pages 54–69. Springer, 2020. 1, 2, 3, 5, 6, 7

[58] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *CVPR*, pages 13766–13775, 2020. 1, 2, 5, 8

[59] Xiong Zhou, Xianming Liu, Chenyang Wang, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Learning with noisy labels via sparse regularization. In *ICCV*, pages 72–81, 2021. 2

[60] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *CVPR*, pages 687–696, 2019. 6, 7