Exploring Object Relation in Mean Teacher for Cross-Domain Detection[*]

Qi Cai [†], Yingwei Pan [‡], Chong-Wah Ngo [§], Xinmei Tian [†], Lingyu Duan [¶], and Ting Yao [‡]

[†] University of Science and Technology of China, Hefei, China   [‡] JD AI Research, Beijing, China

[§] City University of Hong Kong, Kowloon, Hong Kong   [¶] Peking University, Beijing, China

{cqcaiqi, panyw.ustc, tingyao.ustc}@gmail.com, cscwngo@cityu.edu.hk, xinmei@ustc.edu.cn, lingyu@pku.edu.cn

Abstract

Rendering synthetic data (e.g., 3D CAD-rendered images) to generate annotations for learning deep models in vision tasks has attracted increasing attention in recent years. However, simply applying the models learnt on synthetic images may lead to high generalization error on real images due to domain shift. To address this issue, recent progress in cross-domain recognition has featured the Mean Teacher, which directly simulates unsupervised domain adaptation as semi-supervised learning. The domain gap is thus naturally bridged with consistency regularization in a teacher-student scheme. In this work, we advance this Mean Teacher paradigm to be applicable for cross-domain detection. Specifically, we present Mean Teacher with Object Relations (MTOR) that novelly remolds Mean Teacher under the backbone of Faster R-CNN by integrating the object relations into the measure of consistency cost between teacher and student modules. Technically, MTOR firstly learns relational graphs that capture similarities between pairs of regions for teacher and student respectively. The whole architecture is then optimized with three consistency regularizations: 1) region-level consistency to align the region-level predictions between teacher and student, 2) inter-graph consistency for matching the graph structures between teacher and student, and 3) intra-graph consistency to enhance the similarity between regions of same class within the graph of student. Extensive experiments are conducted on the transfers across Cityscapes, Foggy Cityscapes, and SIM10k, and superior results are reported when comparing to state-of-the-art approaches. More remarkably, we obtain a new record of single model: 22.8% of mAP on Syn2Real detection dataset.

1. Introduction

Deep Neural Networks have been proven to be highly effective for learning vision models on large-scale dataset-
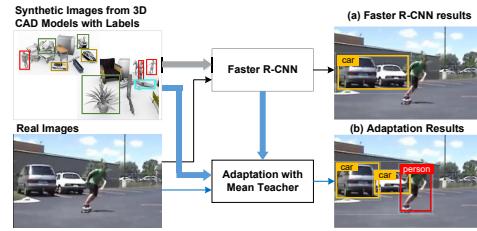

Figure 1. Object detection on one real image by (a) directly applying Faster R-CNN trained on images from 3D CAD models and (b) domain adaptation of Mean Teacher in this work.

s. To date in the literature, there are various datasets (e.g., ImageNet [41] and COCO [25]) that include well-annotated images useful for developing deep models across a variety of vision tasks, e.g., recognition [15, 47], detection [13, 40], and semantic segmentation [2, 27]. Nevertheless, given a new dataset, the typical first step is still to perform intensive manual labeling, which is cost expensive and time consuming. An alternative is to utilize synthetic data which is largely available from 3D CAD models [34], and the ground truth could be freely and automatically generated. However, many previous experiences have also shown that reapplying a model learnt on synthetic data may hurt the performance on real data due to a phenomenon known as "domain shift" [50]. Take the object detection results shown in Figure 1 (a) as an example, the model trained on synthetic data from 3D CAD fails to accurately localize the objects such as person and car. As a result, unsupervised domain adaptation, which aims to utilize labeled examples from the source domain and numerous unlabeled examples in the target domain to reduce the prediction error on the target data, can be a feasible solution for this challenge.

A recent pioneering practice [9] in unsupervised domain adaptation is to directly simulate this task as semi-supervised learning. The basic idea is to develop Mean Teacher [48], the state-of-the-art technique in semi-supervised learning, to work in cross-domain recognition task by pursuing the consistency of two predictions under perturbations of inputs (e.g., different augmentations of image). As such, the domain gap is naturally bridged via the consistency regularization in Mean Teacher, which enforces the predictions of two models (i.e., teacher and student) to
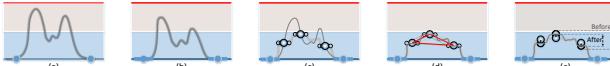

Figure 2. A sketch of cross-domain binary classification task with two labeled examples/regions in source domain (large blue dots) and three unlabeled examples/regions of one image in target domain (blue circle), demonstrating how the choice of the unlabeled target samples affects the unified fitted function across domains (gray curve). (a) A model with no regularization is flexible to fit any function that correctly classifies only labeled source data. (b) A model trained with augmented labeled source data (small blue dots) learns to produce consistent results around labeled data. (c) Mean Teacher [9] locally enforces the predictions to be consistent to the noise around each individual target sample, pursuing additional local smoothing of fitted function (gray curve). (d) Mean Teacher with inter-graph consistency simultaneously adapts target samples to make the holistic graph structure of them resistant to the noise. (e) Mean Teacher with intra-graph consistency enforces additional consistency across target samples of same class, further improving fitted function with long-range smoothing.

be consistent to the perturbations/noise around each unlabeled target sample (Figure 2 (c)). Mean teacher aims for learning a more smooth domain-invariant function than the model trained with no regularization (Figure 2 (a)) or only augmented labeled source data (Figure 2 (b)). In this paper, we novelly consider the use of Mean Teacher for cross-domain detection from the viewpoint of both region-level and graph-structured consistencies. The objective of region-level consistency is to align the region-level classification results of teacher and student models for the identical teacher-generated region proposals, which in turn implicitly enforces the consistency of object localization. The inspiration of graph-structured consistency is from the rationale that the inherent relations between objects within one image should be invariant to different image augmentations. In the context of Mean Teacher, this kind of graph-structured consistency (i.e., inter-graph consistency) is equivalent to matching the graph structures between teacher and student models (Figure 2 (d)). Another kind of graph-structured consistency, i.e., intra-graph consistency, is additionally exploited to reinforce the similarity between image regions of same class within the graph of student model (Figure 2 (e)).

By consolidating the idea of region-level and graph-structured consistencies into Mean Teacher for facilitating cross-domain detection, we present a novel Mean Teacher with Object Relations (MTOR), as shown in Figure 3. The whole framework consists of teacher and student modules under the same backbone of Faster R-CNN [40]. Specifically, each labeled source sample is only passed through student module to conduct supervised learning of detection, while each unlabeled target sample will be fed into both teacher and student with two random augmentations, enabling the measure of the consistency between them to the induced noise. During training, with the same region proposals generated by teacher, two relational graphs are con-

structed via calculating the feature similarity between each pair of regions for teacher and student. The whole MTOR is then trained by the supervised detection loss in student model plus three consistency regularizations, i.e., region-level consistency to align the region-level predictions, inter-graph consistency to match the graph structures between teacher and student, and intra-graph consistency to enhance the similarity between regions of same class in student. With both region-level and graph-structured consistencies, our MTOR could better build invariance across domains and thus obtain encouraging detection results in Figure 1 (b).

2. Related Work

Object Detection. Recent years have witnessed remarkable progress in object detection with deep learning. R-CNN [14] is one of the early works that exploits a two-stage paradigm for object detection by firstly generating region proposals with selective search and then classifying the proposals into foreground classes/background. Later Fast R-CNN [13] extends such paradigm by sharing convolution features across region proposals to significantly speed up the detection process. Faster R-CNN [40] advances Fast R-CNN by replacing selective search with an accurate and efficient Region Proposal Networks (RPN). Next, a few subsequent works [7, 8, 18, 22, 23, 33, 46] strive to improve the accuracy and speed of two-stage detectors. Another line of works builds detectors in one-stage manner by skipping region proposal stage. YOLO [37] jointly predicts bounding boxes and confidences of multiple categories as regression problem. SSD [26] further improves it by utilizing multiple feature maps at different scales. Numerous extensions to the one-stage scheme have been proposed, e.g. [10, 24, 38, 39]. In this work, we adopt Faster R-CNN as the detection backbone for its robustness and flexibility.

Domain Adaptation. As for the literature on domain adaptation, while it is quite vast, the most relevant category to our work is unsupervised domain adaptation in deep architectures. Recent works have involved discrepancy-based methods that guide the feature learning in DCNNs by minimizing the domain discrepancy with Maximum Mean Discrepancy (MMD) [28, 29, 30]. Another branch is to exploit the domain confusion by learning a domain discriminator [11, 12, 44, 49]. Later, self-ensembling [9] extends Mean Teacher [48] for domain adaptation and establishes new records on several cross-domain recognition benchmarks. All of the aforementioned works focus on the domain adaptation for recognition, and recently much attention has been paid to domain adaptation in other tasks, e.g., object detection [4, 35] and semantic segmentation [5, 16, 53]. For domain adaptation on object detection, [45] uses transfer component analysis to learn the common transfer components across domains and [35] aligns the region features with subspace alignment. More Recently, [4] constructs a
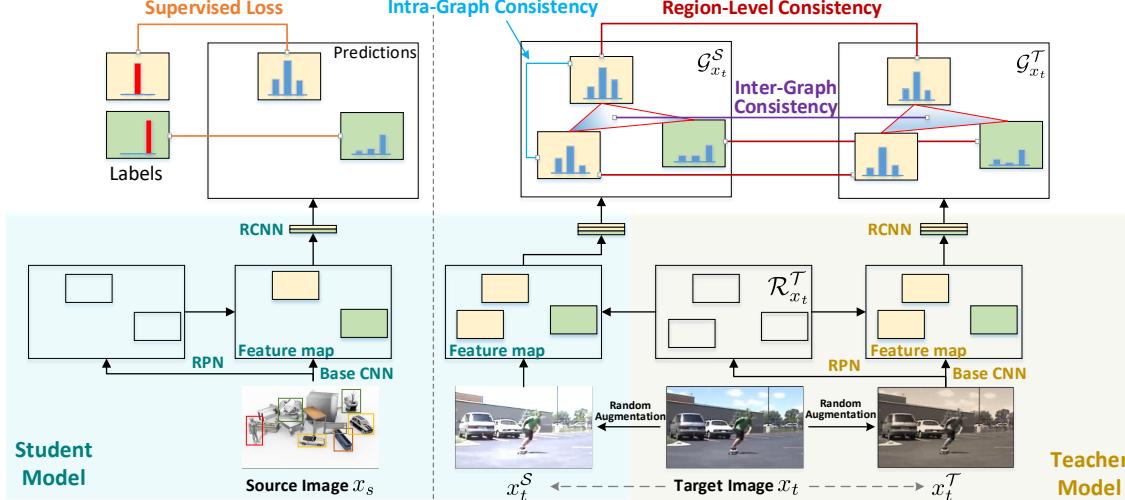
Figure 3. The overview of Mean Teacher with Object Relations (MTOR) for cross-domain detection, with teacher and student models under the same backbone of Faster R-CNN (better viewed in color). Each labeled source image is fed into student model to conduct the supervised learning of detection. Each unlabeled target image x_t is firstly transformed into two perturbed samples, i.e., $x_t^{\mathcal{S}}$ and $x_t^{\mathcal{T}}$, with different augmentations and then we inject the two perturbed samples into student and teacher model separately. During training, with the same set of teacher-generated region proposals $\mathcal{R}_{x_t}^{\mathcal{T}}$ that shares between teacher and student, two relational graphs, i.e., $\mathcal{G}_{x_t}^{\mathcal{T}}$ and $\mathcal{G}_{x_t}^{\mathcal{S}}$, are constructed via calculating the feature similarity between each pair of regions for teacher and student, respectively. Next, three consistency regularization are devised to facilitate cross-domain detection in Mean Teacher paradigm from region-level and graph-structured perspectives: 1) *Region-Level Consistency* to align the region-level predictions between teacher and student; 2) *Inter-Graph consistency* for matching the graph structures between teacher and student, and 3) *Intra-Graph Consistency* to enhance the similarity between regions of same class within the graph of student. The whole MTOR is trained by minimizing the supervised loss on labeled source data plus the three consistency losses on unlabeled target data in an end-to-end manner. Note that the student model is optimized with stochastic gradient descent and the weights of teacher are the exponential moving average of student model weights.

domain adaptive Faster R-CNN by learning domain classifiers on both image and instance levels.

Summary. Similar to previous work [4], our approach aims to leverage additional unlabeled target data for learning domain-invariant detector for cross-domain detection. The novelty is on the exploitation of Mean Teacher to bridge domain gap with consistency regularization in the context of object detection, which has not been previously explored. Moreover, the object relation between image regions is elegantly integrated into Mean Teacher paradigm to boost cross-domain detection.

3. Mean Teacher in Semi-Supervised Learning

We briefly review semi-supervised learning with Mean Teacher [48]. Mean Teacher consists of two models with the same network architecture: a student model $f_{\mathcal{S}}$ parameterized by $w_{f_{\mathcal{S}}}$ and a teacher model $f_{\mathcal{T}}$ parameterized by $w_{f_{\mathcal{T}}}$. The main idea behind Mean Teacher is to encourage predictions of teacher and student consistent under small perturbations of inputs or network parameters. In other words, with the inputs of two different augmentations for the same unlabeled sample, teacher and student models should produce similar predicted probabilities. Specifically, in the standard setting of semi-supervised learning, we have access to labeled set $\mathcal{X}_L = \{(x_l, y_l)\}$ and unlabeled set $\mathcal{X}_U = \{x_u\}$. Given two perturbed samples $x_u^{\mathcal{S}}$ and $x_u^{\mathcal{T}}$ of the same unla-

beled sample x_u, the consistency loss penalizes the difference between the student's prediction $f_{\mathcal{S}}(x_u^{\mathcal{S}}; w_{f_{\mathcal{S}}})$ and the teacher's $f_{\mathcal{T}}(x_u^{\mathcal{T}}; w_{f_{\mathcal{T}}})$, which is typically computed as the Mean Squared Error:

$$\mathcal{L}_{cons}(x_u) = ||f_{\mathcal{S}}(x_u^{\mathcal{S}}; w_{f_{\mathcal{S}}}) - f_{\mathcal{T}}(x_u^{\mathcal{T}}; w_{f_{\mathcal{T}}})||_2^2. \quad (1)$$

The student is trained using gradient descent, while the weights of the teacher $w_{f_{\mathcal{T}}}$ at t-th iteration are the exponential moving average of the student weights $w_{f_{\mathcal{S}}}$: $w_{f_{\mathcal{T}}}^t = \alpha \cdot w_{f_{\mathcal{T}}}^{t-1} + (1 - \alpha) \cdot w_{f_{\mathcal{S}}}^{t-1}$. α is a smoothing coefficient parameter that controls the updating of teacher weights.

Hence, the total training loss in Mean Teacher is composed of supervised cross entropy loss on labeled samples and consistency loss of unlabeled samples, balanced with the tradeoff parameter λ:

$$\mathcal{L} = \sum_{(x_l, y_l) \in \mathcal{X}_L} L_{CE}(x_l, y_l) + \lambda \cdot \sum_{x_u \in \mathcal{X}_U} L_{cons}(x_u). \quad (2)$$

4. Mean Teacher in Cross-Domain Detection

In this paper we remold Mean Teacher in the detection backbone (e.g., Faster R-CNN) for cross-domain detection by integrating the object relations into the measure of consistency regularization between teacher and student. An overview of our Mean Teacher with Object Relations (M-TOR) framework is depicted in Figure 3. We begin this

section by elaborating the problem formulation. Then, a region-level consistency, which is different from the generic consistency at image-level in primal Mean Teacher, is provided to facilitate domain adaptation at region-level. In addition, two kinds of graph-structured consistencies (inter-graph and intra-graph consistencies) are introduced to explore object relation in Mean Teacher, enabling the interaction between regions, which further enhance domain adaptation. Finally, the overall objective combining various consistencies along with its optimization strategy are provided.

4.1. Problem Formulation

In unsupervised domain adaptation, we are given N_s labeled images $\mathcal{D}_s = \{(x_s, B_s)\}$ in source domain and N_t unlabeled images $\mathcal{D}_t = \{x_t\}$ in target domain, where B_s denotes the bounding box annotation for source image x_s. The ultimate goal of cross-domain detection is to design domain-invariant detectors depending on $\mathcal{D}_s$ and $\mathcal{D}_t$.

Inspired by the recent success of consistency-based methods in semi-supervised learning [1, 20, 48] and Mean Teacher in cross-domain recognition [9], we formulate our cross-domain detection model in a Mean Teacher paradigm by enforcing the predictions of teacher and student models consistent under perturbations of input unlabeled target sample. Accordingly, each labeled source sample x_s is passed through student module to perform supervised learning of detection. Meanwhile, each unlabeled target sample x_t is firstly transformed into two perturbed samples (i.e., $x_t^{\mathcal{T}}$ and $x_t^{\mathcal{S}}$) with different augmentations, and then fed into teacher and student models separately. This enables the measure of consistency between student and teacher. During training, different from Mean Teacher in cross-domain recognition [9] that solely encourages generic image-level consistency, we consider the consistency at a finer granularity (i.e., region-level), which is tailored for object detection. Moreover, two graph-structured consistencies are especially designed to exploit object relations in the context of Mean Teacher, which further boosts adaptation by aligning the results depending on the inherent relations between objects.

Specifically, given the identical set of region proposals $\mathcal{R}_{x_t}^{\mathcal{T}} = \{r_t\}$ generated by teacher model $F^{\mathcal{T}}$, we construct two relational graphs $\mathcal{G}_{x_t}^{\mathcal{T}}$ and $\mathcal{G}_{x_t}^{\mathcal{S}}$ to learn the affinity matrix that captures the relation between any pair of regions in teacher and student, respectively. Note that we use $\mathcal{G}_{x_t} \in \{\mathcal{G}_{x_t}^{\mathcal{T}}, \mathcal{G}_{x_t}^{\mathcal{S}}\}$ for simplicity, i.e., $\mathcal{G}_{x_t}$ denotes the graph in either teacher $\mathcal{G}_{x_t}^{\mathcal{T}}$ or student $\mathcal{G}_{x_t}^{\mathcal{S}}$. More precisely, by treating each region in teacher/student as one vertex, the relational graph is constructed as $\mathcal{G}_{x_t} = \{\mathcal{V}_{x_t}, \mathcal{E}_{x_t}\}$, where $\mathcal{V}_{x_t}$ denotes the set of predictions for all region proposals in teacher/student and $\mathcal{E}_{x_t}$ is a $(|\mathcal{V}_{x_t}| \times |\mathcal{V}_{x_t}|)$ affinity matrix whose entry measures the similarities between every two regions. $\mathcal{E}_{x_t}$ is symmetric, and represents an undirected weighted graph. On the basis of two constructed relation-al graphs, we make the detection backbone—Faster R-CNN transferable across domains in Mean Teacher paradigm with three consistency regularization: 1) region-level consistency (Section 4.2) to align the region-level predictions of the vertices in teacher and student graphs sharing the same spatial location, 2) inter-graph consistency (Section 4.3) for matching the graph structures (i.e., the affinity matrices) of teacher and student graphs, and 3) intra-graph consistency (Section 4.4) to enhance the similarity between regions belonging to the same class within the graph of student.

4.2. Region-Level Consistency

Unlike [9] that pursues image-level consistency to perturbations of inputs in recognition, we facilitate Mean Teacher in cross-domain detection by exploiting region-level consistency under the identical region proposals between teacher and student. The design of region-level consistency helps to reduce the local instance variances such as scale, color jitter, random noise, *etc*, which in turn implicitly enforces the consistency of object localization.

Technically, given the two perturbed samples $x_t^{\mathcal{T}}$ and $x_t^{\mathcal{S}}$ of one unlabeled target sample x_t, they are fed into teacher and student detectors under the same backbone (i.e., Faster R-CNN) separately. Faster R-CNN is a two-stage detector consisting of three major components: a Base Convolution Neural Network (Base CNN) for feature extraction, a Region Proposal Network (RPN) to generate candidate region proposals, and a Region-based Convolution Neural Network (RCNN) for classifying each region. Hence, with the input of $x_t^{\mathcal{T}}$, the Base CNN of teacher $F_{Conv}^{\mathcal{T}}$ firstly produces output feature map $\boldsymbol{f}_{x_t}^{\mathcal{T}}$. Next, depending on the output feature map $\boldsymbol{f}_{x_t}^{\mathcal{T}}$, a set of region proposals $\mathcal{R}_{x_t}^{\mathcal{T}} = \{r_t\}$ are generated via RPN in teacher $F_{RPN}^{\mathcal{T}}$:

$$\boldsymbol{f}_{x_t}^{\mathcal{T}} = F_{Conv}^{\mathcal{T}}(x_t^{\mathcal{T}}), \mathcal{R}_{x_t}^{\mathcal{T}} = F_{RPN}^{\mathcal{T}}(\boldsymbol{f}_{x_t}^{\mathcal{T}}). \quad (3)$$

For each region proposal $r_t \in \mathcal{R}_{x_t}^{\mathcal{T}}$, a ROI pooling layer is utilized to extract a fixed-length vector $\boldsymbol{f}_r^{\mathcal{T}}$ from the feature map $\boldsymbol{f}_{x_t}^{\mathcal{T}}$, which represents the region feature of r_t in teacher. The RCNN in teacher $F_{RCNN}^{\mathcal{T}}$ further takes each region feature $\boldsymbol{f}_r^{\mathcal{T}}$ as input and classifies it into one of the C foreground categories and a catch-all background class. Here the prediction of each region is the probability distribution over background plus foreground categories, which is denoted as $\boldsymbol{d}_r^{\mathcal{T}} = F_{RCNN}^{\mathcal{T}}(\boldsymbol{f}_r^{\mathcal{T}})$. As such, by accumulating the predicted results of all region proposals, the entire detection output of $x_t^{\mathcal{T}}$ in teacher is denoted as $\mathcal{V}_{x_t}^{\mathcal{T}} = \{\boldsymbol{d}_r^{\mathcal{T}}\}$. Similarly, for student model $F^{\mathcal{S}}$, another perturbed image $x_t^{\mathcal{S}}$ is fed into its Base CNN $F_{Conv}^{\mathcal{S}}$ to produce the feature map $\boldsymbol{f}_{x_t}^{\mathcal{S}}$. Note that instead of generating another set of region proposals for $x_t^{\mathcal{S}}$ via RPN in student, we directly take the region proposals from teacher $\mathcal{R}_{x_t}^{\mathcal{T}}$ as the ones in student:

$$\boldsymbol{f}_{x_t}^{\mathcal{S}} = F_{Conv}^{\mathcal{S}}(x_t^{\mathcal{S}}), \mathcal{R}_{x_t}^{\mathcal{S}} = \mathcal{R}_{x_t}^{\mathcal{T}}. \quad (4)$$

That is, we endow teacher and student with the same set of region proposals, enabling the interaction between teacher and student for measuring region-level consistency. Given region proposals $\mathcal{R}_{x_t}^{\mathcal{S}}$ and feature map $\boldsymbol{f}_{x_t}^{\mathcal{S}}$, we can acquire the region feature $\boldsymbol{f}_r^{\mathcal{S}}$ for each region proposal and the corresponding probability distribution $\boldsymbol{d}_r^{\mathcal{S}} = F_{RCNN}^{\mathcal{S}}(\boldsymbol{f}_r^{\mathcal{S}})$, leading to the entire detection results in student $\mathcal{V}_{x_t}^{\mathcal{S}} = \{\boldsymbol{d}_r^{\mathcal{S}}\}$.

As such, the region-level consistency is measured as the distance between the prediction of teacher $\mathcal{V}_{x_t}^{\mathcal{T}}$ and that of student $\mathcal{V}_{x_t}^{\mathcal{S}}$. To focus more on foreground samples and stabilize the training in the challenging cross-domain detection scenario, we follow [9] and adopt confidence thresholding to filter out background region proposals and low-confidence foreground region proposals with noise. For each region proposal $r_t \in \mathcal{R}_{x_t}^{\mathcal{T}}$ of teacher model, we compute the confidence as $q_r^{\mathcal{T}} = \max_{j \in \mathcal{C}}(\boldsymbol{d}_{rj}^{\mathcal{T}})$, where $\mathcal{C}$ is the set of C foreground categories and $\boldsymbol{d}_{rj}^{\mathcal{T}}$ is the predicted probability of j-th foreground category. If $q_r^{\mathcal{T}}$ is below the confidence threshold ϵ, we eliminate the region proposal in $\mathcal{R}_{x_t}^{\mathcal{T}}$. With the refined region proposal $(\mathcal{R}_{x_t}^{\mathcal{T}})$, and the corresponding region-level predictions of teacher and student $(\mathcal{V}_{x_t}^{\mathcal{T}} = \{\boldsymbol{d}_r^{\mathcal{T}}\}$ and $\mathcal{V}_{x_t}^{\mathcal{S}} = \{\boldsymbol{d}_r^{\mathcal{S}}\})$, the **R**egion-level **C**onsistency **L**oss (RCL) is calculated as the average of Mean Squared Error between the region-level predictions of teacher and student for all region proposals:

$$L_{x_t}^{RCL} = \frac{1}{|R_{x_t}^{\mathcal{T}}|} \cdot \sum_{r \in R_{x_t}^{\mathcal{T}}} ||\boldsymbol{d}_r^{\mathcal{T}} - \boldsymbol{d}_r^{\mathcal{S}}||_2^2. \tag{5}$$

4.3. Inter-Graph Consistency

The region-level consistency only individually aligns the predictions of each region proposal in teacher and student, while leaving the relations between regions unexploited. Thus, inspired from graph structure exploitation [31, 32, 51, 52] in computer vision tasks, we devise a novel graph-structured regularization, i.e., inter-graph consistency, to measure the consistency of graph structures under perturbations of inputs by matching the affinity matrices of graphs constructed in teacher and student models. The rationale of inter-graph consistency is that the inherent relations between objects within each image should be invariant to different image augmentations.

In particular, for the graph constructed in teacher $\mathcal{G}_{x_t}^{\mathcal{T}} = \{\mathcal{V}_{x_t}^{\mathcal{T}}, \mathcal{E}_{x_t}^{\mathcal{T}}\}$, the affinity matrix of teacher $\mathcal{E}_{x_t}^{\mathcal{T}}$ is obtained by defining each entry as the similarity between two regions. For instance, given two region proposals $r_m, r_n \in \mathcal{R}_{x_t}^{\mathcal{T}}$, the entry $(\mathcal{E}_{x_t}^{\mathcal{T}})_{m,n}$ in $\mathcal{E}_{x_t}^{\mathcal{T}}$ is calculated as the cosine similarity between the region representations $(\boldsymbol{f}_{r_m}^{\mathcal{T}}$ and $\boldsymbol{f}_{r_n}^{\mathcal{T}})$:

$$(\mathcal{E}_{x_t}^{\mathcal{T}})_{m,n} = \frac{\boldsymbol{f}_{r_m}^{\mathcal{T}} \cdot \boldsymbol{f}_{r_n}^{\mathcal{T}}}{||\boldsymbol{f}_{r_m}^{\mathcal{T}}||_2 \cdot ||\boldsymbol{f}_{r_n}^{\mathcal{T}}||_2}. \tag{6}$$

Similarly, we achieve the affinity matrix of student $\mathcal{E}_{x_t^{\mathcal{S}}}$ by measuring the cosine similarities between every two regions

in student. Accordingly, the Int**Er**-**G**raph Consistency **L**oss (EGL) is defined as the Mean Squared Error between the affinity matrices of graphs in teacher and student models:

$$L_{x_t}^{EGL} = \frac{1}{|R_{x_t}^{\mathcal{T}}|^2} \cdot ||\mathcal{E}_{x_t}^{\mathcal{S}} - \mathcal{E}_{x_t}^{\mathcal{T}}||_2^2. \tag{7}$$

4.4. Intra-Graph Consistency in Student

Inspired from self-labeling [21, 42] for domain adaptation, the inter-graph consistency is devised to further reinforce the similarity between regions of same class within the graph of student with the supervision from teacher. Specifically, since no label is provided for target samples in unsupervised domain adaptation settings, we directly utilize the teacher to assign each region proposal $r_t \in \mathcal{R}_{x_t}^{\mathcal{T}}$ a "pseudo" label: $\hat{l}_r = arg \max_{j \in C}(\boldsymbol{d}_{rj}^{\mathcal{T}})$. Next, a $(|\mathcal{R}_{x_t}^{\mathcal{T}}| \times |\mathcal{R}_{x_t}^{\mathcal{T}}|)$ supervision matrix $M_{x_t}^{\mathcal{T}}$ is naturally generated to indicate whether two regions belong to the same category:

$$(M_{x_t}^{\mathcal{T}})_{(m,n)} = \begin{cases} 1 & \text{if } \hat{l}_{r_m} = \hat{l}_{r_n}, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

where $\hat{l}_{r_m}$ and $\hat{l}_{r_n}$ denote the pseudo labels of two regions $r_m, r_n \in \mathcal{R}_{x_t}^{\mathcal{T}}$, respectively. Thus, given the the affinity matrix of student $\mathcal{E}_{x_t}^{\mathcal{S}}$ and the supervision matrix $M_{x_t}^{\mathcal{T}}$, the intr**A**-**G**raph consistency **L**oss (AGL) is defined as:

$$L_{x_t}^{AGL} = \frac{\sum\limits_{1 \leq m,n \leq |\mathcal{R}_{x_t}^{\mathcal{T}}|} (M_{x_t}^{\mathcal{T}})_{(m,n)} \cdot (1 - (\mathcal{E}_{x_t}^{\mathcal{S}})_{(m,n)})}{max(1, \sum\limits_{1 \leq m,n \leq |\mathcal{R}_{x_t}^{\mathcal{T}}|} (M_{x_t}^{\mathcal{T}})_{(m,n)})}. \tag{9}$$

Note that $L_{x_t}^{AGL}$ is triggered when at least two regions share the same pseudo label in $\mathcal{R}_{x_t}^{\mathcal{T}}$. By minimizing the inter-graph consistency loss, the similarity between regions with the same pseudo label in student is enhanced, pursuing lower intra-class variation within the graph of student.

4.5. Optimization

Training Objective. The overall training objective of our MTOR integrates the supervised loss $\mathcal{L}_{sup}$ on labeled source data $\mathcal{D}_s$ and three consistency losses, i.e., region-level consistency $L_{x_t}^{RCL}$ in Eq.(5), inter-graph consistency $L_{x_t}^{EGL}$ in Eq.(7) and intra-graph consistency $L_{x_t}^{AGL}$ in Eq.(9) on unlabeled target data $\mathcal{D}_t$:

$$\mathcal{L} = \sum_{(x_s, B_s) \in \mathcal{D}_s} \mathcal{L}_{sup}(x_s, B_s) + \lambda \cdot \sum_{x_t \in \mathcal{D}_t} (L_{x_t}^{RCL} + L_{x_t}^{EGL} + L_{x_t}^{AGL}), \tag{10}$$

where λ is the tradeoff parameter.

Weights Update. The student network $F^{\mathcal{S}}$ is optimized with standard SGD algorithm by minimizing $\mathcal{L}$. The weights of teacher network $F^{\mathcal{T}}$ at iteration t are updated as the exponential moving average of student weights:

$$w_{F^{\mathcal{T}}}^t = \alpha \cdot w_{F^{\mathcal{T}}}^{t-1} + (1 - \alpha) \cdot w_{F^{\mathcal{S}}}^{t-1}, \tag{11}$$

where α denotes smoothing coefficient parameter.

5. Experiments

We conduct extensive evaluations of our MTOR for cross-domain detection in two different domain shift scenarios, including one normal-to-foggy weather transfer in urban scene (*Cityscapes* [6] → *Foggy Cityscapes* [43]) and two synthetic-to-real transfers (i.e., *SIM10k* [19] → *Cityscapes* and 3D CAD-rendered images → real images in *Syn2Real* detection dataset [34]).

5.1. Dataset and Experimental Settings

Dataset. The *Cityscapes* dataset (**C**) is a popular semantic understanding benchmark in urban street scenes with pixel-level annotation, containing 2,975 images for training and 500 images for validation. Since it is not dedicated for detection, we follow [4] and generate the bounding box annotations by the tightest rectangles of each instance segmentation mask for 8 categories (*person* plus 7 kinds of *transports*). *Foggy Cityscapes* (**F**) is a recently proposed synthetic foggy dataset which simulates fog on real scenes. Each foggy image is rendered with clear image and depth map from *Cityscapes*. Thus the annotations and data split in *Foggy Cityscapes* are inherited from *Cityscapes*. *SIM10k* (**M**) dataset contains $10k$ images rendered from computer game—Grand Theft Auto V (GTA 5) with bounding box annotations for *cars*. The *Syn2Real* detection dataset is the largest synthetic-to-real object detection dataset to date with over $70k$ images in the training, validation and testing domains. The training domain consists of $8k$ synthetic images (**S**) which are generated from 3D CAD models. Each object is rendered independently and placed on a white background. The validation domain includes 3,289 real images from *COCO* [25] (**O**) and the testing domain contains 60,863 images from video frames in *YTBB* [36] (**Y**).

Normal-to-Foggy Weather Transfer. We follow [4] and evaluate **C** → **F** for transfer across different weather conditions. The training set in *Cityscapes* is taken as source domain. We use the training set in *Foggy Cityscapes* as target domain and results are reported on its validation set.

Synthetic-to-Real Image Transfer. We consider two directions for synthetic-to-real transfers: **M** → **C** and **S** → **O/Y**. For **M** → **C**, we utilize the entire *SIM10k* as source domain and leverage *Cityscapes* training set as target domain. The results are reported on *Cityscapes* validation split. For **S** → **O/Y** on *Syn2Real* detection dataset, we take the training set (synthetic images) as source domain and the validation set (*COCO*)/testing set (*YTBB*) as target domain. Since the annotations of testing set are not publicly available, we submit results to online testing server for evaluation.

Implementation Details. For **C** → **F** and **M** → **C**, we adopt the 50-layer ResNet [15] pre-trained on ImageNet [41] as the basic architecture of Faster R-CNN backbone. For the more challenging **S** → **O/Y**, the Faster R-CNN backbone is mainly constructed on 152-layer ResNet. For all transfers, we utilize "image-centric" sampling strategy [13]. Each input image is resized such that its scale (shorter edge) is 600 pixels. Each mini-batch contains 2 images per GPU, one from the source domain and the other from the target domain. We train on 4 GPUs (so effective mini-batch size is 8) and each image has 128 sampled anchors, with a ratio of 1:3 of positive to negatives [13]. We implement MTOR based on MXNet [3]. Specifically, the network weights are trained by SGD optimizer with 0.0005 weight decay and 0.9 momentum. The learning rate and maximum training epoch are set as 0.001 and 10 for all experiments. The confidence threshold ϵ is empirically set to 0.98 for **C** → **F** and **M** → **C**, and 0.99 for **S** → **O/Y**. The tradeoff parameter λ in Eq.(10) and the smooth coefficient parameter α in Eq.(11) is set as 1.0 and 0.99, respectively. Moreover, our MTOR is firstly pre-trained on labeled source data. For data augmentations on target images, we firstly augment each target image with the same spatial perturbation including random cropping, padding, or flipping. Next, we additionally perform two different kinds of image augmentations with random color jittering (i.e., brightness, contrast, hue and saturation augmentations) or PCA noise, resulting in two perturbed target samples, one for student and the other for teacher. Following [4], we report mAP with a IoU threshold of 0.5 for evaluation.

Compared Approaches. To empirically verify the merit of our MTOR, we compare the following methods: (1) **Source-Only** directly exploits the Faster R-CNN model trained on source domain to detect objects in target samples. (2) **DA**[4] designs two domain classifiers to alleviate both image-level and region-level domain discrepancy, which are further enforced with a consistency regularizer. (3) **MTOR** is the proposal in this paper. Moreover, we design three degraded variants trained with region-level consistency (**MTOR$_R$**), region-level plus inter-graph consistency (**MTOR$_{RE}$**), and region-level plus intra-Graph consistency (**MTOR$_{RA}$**). (4) **Train-on-target** is an oracle run that trains Faster R-CNN on all the labeled target samples.

5.2. Performance Comparison and Analysis

Normal-to-Foggy Weather Transfer. Table 1 shows the performance comparisons on *Foggy Cityscapes* validation set for **C** → **F** transfer. Overall, the results with regard to mAP score indicate that our proposed MTOR achieves superior performance against state-of-the-art technique (DA). In particular, the mAP of MTOR can achieve 35.1%, making 3.1% absolute improvement over the best competitor DA. The performances of Source-only which trains Faster R-CNN only on the labeled source data can be regarded as a lower bound without adaptation. By additionally incorporating the domain classifier in both image and region level, DA leads to a large performance boost over Source-only, which basically indicates the advantage of alleviat-

Table 1. The mean Average Precision (mAP) of different models on *Foggy Cityscapes* validation set for **C** → **F** transfer.

	RCL	EGL	AGL	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
Source-only				25.7	35.9	36.0	19.4	30.8	9.7	**29.0**	28.9	26.9
DA [4]				29.2	40.4	43.4	19.7	38.3	28.5	23.7	32.7	32.0
MTOR$_R$	✓			**30.8**	**41.5**	44.1	21.6	37.8	35.1	26.7	**35.8**	34.2
MTOR$_{RE}$	✓	✓		28.7	40.1	**45.9**	**22.9**	38.0	38.6	26.9	34.9	34.5
MTOR$_{RA}$	✓		✓	29.6	41.2	43.7	22.2	38.4	**40.9**	27.8	35.3	34.9
MTOR	✓	✓	✓	30.6	41.4	44.0	21.9	**38.6**	40.6	28.3	35.6	**35.1**
Train-on-target				31.4	42.6	51.7	28.8	43.4	40.2	31.7	33.2	37.9

Table 2. The Average Precision (AP) of *car* on *Cityscapes* validation set for **M** → **C** transfer.

	RCL	EGL	AGL	car AP
Source-only				39.4
DA [4]				41.9
MTOR$_R$	✓			45.9
MTOR$_{RE}$	✓	✓		46.1
MTOR$_{RA}$	✓		✓	46.3
MTOR	✓	✓	✓	**46.6**
Train-on-target				58.6

Figure 4. Examples of detection results on *COCO* for **S** → **O**.

ing the domain discrepancy over the source and target data. Note that for fair comparison, we re-implemented DA based on the same 50-layer ResNet architecture. However, the performances of DA are still lower than our MTOR$_R$, which utilizes region-level consistency regularization in Mean Teacher paradigm. This confirms the effectiveness of enforcing region-level consistency under perturbations of unlabeled target samples for cross-domain detection. In addition, by further integrating object relations into Mean Teacher paradigm through graph-structured consistency from inter-graph or intra-graph perspective, our MTOR$_{RE}$ and MTOR$_{RA}$ improve MTOR$_R$. The results demonstrate the advantage of inter-graph consistency to match the graph structures between teacher and student, and intra-graph consistency to enhance the similarity between regions of same class in student. By simultaneously utilizing region-level and two graph-structured consistencies, MTOR further boosts up the performances, which indicates the merit of jointly exploiting inter-graph and intra-graph consistencies in Mean Teacher paradigm.

Synthetic-to-Real Image Transfer. The performance comparisons for synthetic-to-real transfer task on **M** → **C** are summarized in Table 2. Our MTOR exhibits better performance than other runs. In particular, the AP of *car* for MTOR can reach 46.6%, making the absolute improvement over DA by 4.7%. Similar to the observations in normal-to-foggy weather transfer, MTOR$_R$ performs better than DA by aligning region-level predictions in Mean Teacher and the performance is further improved by incorporating inter-graph and intra-graph consistency in MTOR$_{RE}$ and MTOR$_{RA}$. Combining all the three consistency regularizations, our MTOR achieves the best performance.

We further evaluate our approach for **S** → **O/Y** transfer on the more challenging *Syn2Real* detection dataset. Table 3 shows the performance comparisons on **S** → **O** trans-

fer. A clear performance improvement is achieved by our proposed MTOR over other baselines. Similar to the observations on the transfers across *SIM10k*, *Cityscapes*, and *Foggy Cityscapes*, MTOR$_R$ performs better than DA by taking region-level consistency on target samples into account for cross-domain detection. Moreover, MTOR$_{RE}$ and MTOR$_{RA}$ exhibit better performance than MTOR$_R$ by additionally pursuing inter-graph and intra-graph consistency respectively, and further performance improvement is attained when exploiting region-level consistency plus two graph-structured consistencies by MTOR. We also submitted our MTOR, Source-only, and DA to online evaluation server and evaluated the performances on official testing set. Table 3 summaries the performances on official testing set *YTBB* for **S** → **Y** transfer. The results clearly show that our MTOR outperforms two other baselines.
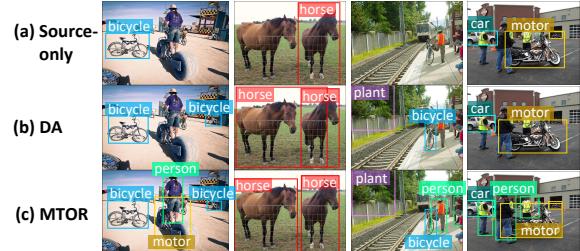
Qualitative Analysis. Figure 4 showcases four examples of detection results on *COCO* for **S** → **O** transfer by three approaches, i.e., Source-only, DA and our MTOR. The exemplar results clearly show that our MTOR can generate more accurate detection results by exploring region-level and graph-structured consistency in Mean Teacher paradigm to boost cross-domain detection. For instance, MTOR correctly detects *person* in the fourth image which is missed in Source-only and DA.

Effect of the Parameters λ and α. To clarify the effect of tradeoff parameter λ in Eq.(10) and smoothing coefficient parameter α in Eq.(11), we show the performance curves with different tradeoff/smoothing coefficient parameters in Figure 5. As shown in the figure, we can see that both mAP curves of λ and α are generally like the "∧" shapes when λ varies in a range from 0.1 to 5.0 and α varies in a range from 0.92 to 0.9999. The best performance is achieved when λ is 1.0 and α is about 0.98.

Error Analysis of Highest Confident Detections. To

Table 3. The mean Average Precision (mAP) of different models on *Syn2Real* detection dataset for **S → O/Y** transfers.

	RCL	EGL	AGL	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbd	train	truck	mAP
mAP on validation set (COCO) for S → O transfer:																
Source-only				30.0	**25.3**	31.3	14.0	17.3	1.9	25.6	18.5	14.7	14.7	21.1	2.2	18.1
DA [4]				30.3	24.1	31.3	14.0	17.4	1.3	27.4	18.9	**17.5**	14.5	21.8	**3.1**	18.5
MTOR$_R$	✓			32.0	22.8	29.1	15.3	**20.8**	0.6	**32.4**	22.2	0.5	18.2	**36.9**	0.6	19.3
MTOR$_{RE}$	✓	✓		33.3	21.2	32.9	13.1	18.1	**3.1**	32.2	24.0	1.4	**20.5**	34.4	0.6	19.6
MTOR$_{RA}$	✓		✓	35.4	24.0	32.1	14.9	19.1	1.8	31.6	**24.2**	3.7	18.9	31.7	2.0	20.0
MTOR	✓	✓	✓	**35.5**	24.9	**32.9**	**15.4**	19.1	1.8	31.4	21.8	14.4	18.9	30.4	1.7	**20.7**
Train-on-target				84.5	52.2	77.5	58.7	76.1	28.9	65.4	71.9	49.2	70.5	83.8	52.5	64.3
mAP on official testing set (YTBB) for S → Y transfer:																
Source-only				28.4	18.4	23.8	28.4	35.8	3.6	35.7	**8.6**	8.4	14.8	6.4	5.2	18.1
DA [4]				38.0	16.1	23.3	30.7	33.0	4.7	34.8	6.1	**15.7**	14.0	**9.8**	**9.5**	19.6
MTOR (Ours)	✓	✓	✓	**42.8**	**21.0**	**31.3**	**33.3**	**42.9**	**10.2**	**38.5**	7.2	12.9	**18.0**	7.2	8.2	**22.8**
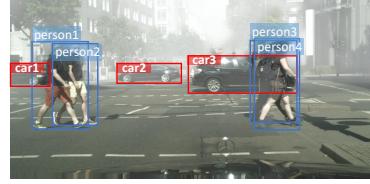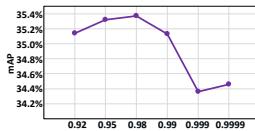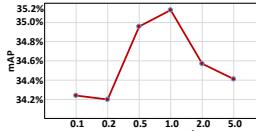
Figure 5. Effect of parameters λ and α on **C → F** transfer.
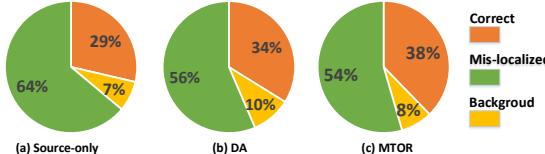
Figure 6. Error analysis of highest confident detections on **C → F**.

Figure 7. Visualization of relational graph on *Foggy Cityscapes*.

further clarify the effect of the proposed region-level and graph-structured consistencies in Mean Teacher paradigm, we analyze the accuracies of Source-only, DA and MTOR caused by the highest confident detections on *Foggy Cityscapes* for **C → F** transfer. We follow [4, 17] and categorize the detections into 3 types: **Correct** (IoU with ground-truth ≥ 0.5), **Mis-Localized** ($0.5 >$ IoU with ground-truth ≥ 0.3) and **Background** (IoU with ground-truth < 0.3). For each class, we select top-K predictions where K is the number of ground-truth bounding boxes in this class. We report the mean percentage of each type across all categories in Figure 6. Compared to Source-only, DA and our MTOR clearly improve the number of correct detections (orange color) and reduce the number of false positives (other colors). Moreover, by leveraging region-level and graph-structured consistencies in Mean Teacher, MTOR leads to both smaller mis-localized and background errors than DA.

Visualization of Relational Graph. Figure 7 further shows the visualization of an exemplar relational graph (i.e., the affinity matrix) learned by Source-only, DA and our M-TOR on *Foggy Cityscapes* for **C → F** transfer. For each approach, we extract the region representation of each ground-truth region and construct the relational graph by computing cosine similarity between every two regions. Note that the first three regions belong to *car* class and the rest four regions fall into *person* class. Thus we can clearly see that most intra-class similarities of MTOR are higher than those of Source-only and DA. The results demonstrate the advan-

tage of enforcing intra-graph consistency in MTOR, leading to more discriminative region feature for object detection.

6. Conclusions

We have presented Mean Teacher with Object Relations (MTOR), which explores domain adaptation for object detection in an unsupervised manner. Particularly, we study the problem from the viewpoint of both region-level and graph-structured consistencies in Mean Teacher paradigm. To verify our claim, we have built two relational graphs that capture similarities between pairs of regions for teacher and student respectively. The region-level consistency is to align the region-level predictions between teacher and student, which facilitates domain adaptation at region-level. The inter-graph consistency further matches the graph structures between teacher and student, pursuing a noise-resistant holistic graph structure on target domain. In addition, intra-graph consistency is utilized to enhance the similarity between regions of same class in student, which ideally leads to graph with lower intra-class variation. Experiments conducted on the transfers across *Cityscapes*, *Foggy Cityscapes*, and *SIM10k* validate our proposal and analysis. More remarkably, we achieve state-of-the-art performance of single model on synthetic-to-real image transfer in *Syn2Real* detection dataset.

References

[1] Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. Improving consistency-based semi-supervised learning with weight averaging. *arXiv preprint arXiv:1806.05594*, 2018.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. on PAMI*, 2018.

[3] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. In *Workshop on Machine Learning Systems, NIPS*, 2016.

[4] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, 2018.

[5] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *CVPR*, 2018.

[6] 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*, 2016.

[7] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, 2016.

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017.

[9] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for domain adaptation. In *ICLR*, 2018.

[10] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.

[11] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *ICML*, 2015.

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016.

[13] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

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

[16] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[17] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *ECCV*, 2012.

[18] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *CVPR*, 2018.

[19] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *ICRA*, 2017.

[20] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017.

[21] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, 2013.

[22] Zeming Li, Chao Peng, Gang Yu, Xiangyu Zhang, Yangdong Deng, and Jian Sun. Light-head r-cnn: In defense of two-stage object detector. *arXiv preprint arXiv:1711.07264*, 2017.

[23] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *ICCV*, 2017.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[28] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

[29] Mingsheng Long, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, 2017.

[30] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *NIPS*, 2016.

[31] Yingwei Pan, Yehao Li, Ting Yao, Tao Mei, Houqiang Li, and Yong Rui. Learning deep intrinsic video representation by exploring temporal coherence and graph structure. In *IJCAI*, 2016.

[32] Yingwei Pan, Ting Yao, Tao Mei, Houqiang Li, Chong-Wah Ngo, and Yong Rui. Click-through-based cross-view learning for image search. In *SIGIR*, 2014.

[33] Chao Peng, Tete Xiao, Zeming Li, Yuning Jiang, Xiangyu Zhang, Kai Jia, Gang Yu, and Jian Sun. Megdet: A large mini-batch object detector. In *CVPR*, 2018.

[34] Xingchao Peng, Ben Usman, Kuniaki Saito, Neela Kaushik, Judy Hoffman, and Kate Saenko. Syn2real: A new benchmark forsynthetic-to-real visual domain adaptation. *arXiv preprint arXiv:1806.09755*, 2018.

[35] Anant Raj, Vinay P Namboodiri, and Tinne Tuytelaars. Subspace alignment based domain adaptation for rcnn detector. *BMVC*, 2015.

[36] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *CVPR*, 2017.

[37] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[38] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *CVPR*, 2017.

[39] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[42] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *ICML*, 2017.

[43] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 2018.

[44] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. *CVPR*, 2018.

[45] Behjat Siddiquie, Vlad I Morariu, Fatemeh Mirrashed, Rogerio S Feris, and Larry S Davis. Domain adaptive object detection. In *WACV*, 2013.

[46] Bharat Singh, Hengduo Li, Abhishek Sharma, and Larry S Davis. R-fcn-3000 at 30fps: Decoupling detection and classification. In *CVPR*, 2018.

[47] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[48] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, 2017.

[49] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

[50] Ting Yao, Chong-Wah Ngo, and Shiai Zhu. Predicting domain adaptivity: redo or recycle? In *ACM MM*, 2012.

[51] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.

[52] Ting Yao, Yingwei Pan, Chong-Wah Ngo, Houqiang Li, and Tao Mei. Semi-supervised domain adaptation with subspace learning for visual recognition. In *CVPR*, 2015.

[53] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *CVPR*, 2018.