Instance Relation Graph Guided  
Source-Free Domain Adaptive Object Detection

Vibashan VS, Poojan Oza, and Vishal M. Patel  
Johns Hopkins University, Baltimore, MD, USA  
{vvishnu2, poza2, vpatel36}@jhu.edu

Abstract

Unsupervised Domain Adaptation (UDA) is an effective approach to tackle the issue of domain shift. Specifically, UDA methods try to align the source and target representations to improve generalization on the target domain. Further, UDA methods work under the assumption that the source data is accessible during the adaptation process. However, in real-world scenarios, the labelled source data is often restricted due to privacy regulations, data transmission constraints, or proprietary data concerns. The Source-Free Domain Adaptation (SFDA) setting aims to alleviate these concerns by adapting a source-trained model for the target domain without requiring access to the source data. In this paper, we explore the SFDA setting for the task of adaptive object detection. To this end, we propose a novel training strategy for adapting a source-trained object detector to the target domain without source data. More precisely, we design a novel contrastive loss to enhance the target representations by exploiting the objects relations for a given target domain input. These object instance relations are modelled using an Instance Relation Graph (IRG) network, which are then used to guide the contrastive representation learning. In addition, we utilize a student-teacher to effectively distill knowledge from source-trained model to target domain. Extensive experiments on multiple object detection benchmark datasets show that the proposed approach is able to efficiently adapt source-trained object detectors to the target domain, outperforming state-of-the-art domain adaptive detection methods. Code and models are provided in https://viudomain.github.io/irg-sfda-web/.

1. Introduction

In recent years, object detection has seen tremendous advancements due to the rise of deep networks [12, 42, 44, 45, 53, 79]. The major contributor to this success is the availability of large-scale annotated detection datasets [10, 13, 15, 43, 73], as it enables the supervised training of deep object detector models. However, these models often have poor generalization when deployed in visual domains not encountered during training. In such cases, most works in the literature follow the Unsupervised Domain Adaptation (UDA) setting to improve generalization [7, 14, 23, 24, 57, 62]. Specifically, UDA methods aim to minimize the domain discrepancy by aligning the feature distribution of the detector model between source and target domain [9, 19, 28, 56, 59]. To perform feature alignment, UDA methods require simultaneous access to the labeled source and unlabeled target data. However in practical scenarios, the access to source data is often restricted due to concerns related to privacy/safety, data transmission, data proprietary etc. For example, consider a detection model trained on large-scale source data, that performs poorly when deployed in new devices having data with different visual domains. In such cases, it is far more efficient to transmit the source-trained detector model (~500-1000MB) for adaptation rather than transmitting the source data (~10-100GB) to these new devices [27, 37]. Moreover, transmitting only source-trained model alleviates many privacy/safety, data proprietary concerns as well [41, 47, 70]. Hence, adapting the source-trained model to the target domain without having access to source data is essential in the case of practical deployment of detection models. This motivates us to study Source-Free Domain Adaptation (SFDA) setting for adapting object detectors (illustrated in Fig. 1).

The SFDA is a more challenging setting than UDA. Specifically, on top of having no labels for the target data,
the source data is not accessible during adaptation. Therefore, most SFDA methods for detection consider training with pseudo-labels generated by source-trained model [27, 40]. During our initial SFDA training experiments, we identified two key challenges. Firstly, noisy pseudo-labels generated by the source-trained model due to domain shift can result in suboptimal distillation of target domain information into the source-trained model [11, 46]. Secondly, Fig. 2 shows object proposals for an image from FoggyCityscapes, predicted by a detector model trained on Cityscapes. Here, all the proposals have Intersection-over-Union $>0.9$ with respective ground-truth bounding boxes and each proposal is assigned a prediction with a confidence score. Noticeably, the proposals around the bus instance have different predictions, e.g., car with 18%, truck with 93%, and bus with 29% confidence. This indicates that the pooled features are prone to a large shift in the feature space for a small shift in the proposal location. This is because, the source-trained model representations would tend to be biased towards source data, resulting in weak representation for the target data. Therefore, we consider two major challenges in SFDA training: 1) Effectively distill target domain information into source-trained model 2) Enhancing the target domain feature representations.

Motivated by [46], we utilize mean-teacher [61] framework to effectively distill of target domain knowledge into source-trained model. However, the key challenge of enhancing the target domain feature representations remained. To address this, we turned to contrastive representation learning (CRL) methods, has been shown to learn high-quality representations from images in an unsupervised manner [5, 6, 69]. CRL methods achieve this by forcing representations to be similar under multiple views (or augmentations) of an anchor image and dissimilar to all other images. In classification, the CRL methods assume that each image contains only one object. On the contrary, for object detection, each image is highly likely to have multiple object instances. Furthermore, the CRL training also requires large batch sizes and multiple views to learn high-quality representations, which incurs a very high GPU/memory cost, as detection models are computationally expensive. To circumvent these issues, we propose an alternative strategy which exploits the architecture of the detection model like Faster-RCNN [54]. Interestingly, the proposals generated by the Region Proposal Network (RPN) of a Faster-RCNN essentially provide multiple views for any object instance as shown in Fig. 3 (a). In other words, the RPN module provides instance augmentation for free, which could be exploited for CRL, as shown in Fig. 3 (b). However, RPN predictions are class agnostic, it is challenging to form positive (same class)/negative pairs (different class), which is essential for CRL.

Figure 2. (a) Object predictions by Cityscapes-trained model on the FoggyCityscapes image. (b) Corresponding ground truth. Here, the proposals around the bus instance have inconsistent predictions, indicating that instance features are prone to large shift in the feature space, for a small shift in the proposal location.

Figure 3. (a) Class agnostic object proposals generated by Region Proposal Network (RPN). (b) Cropping out RPN proposals will provide multiple contrastive views of an object instance. We utilize this to improve target domain feature representations through RPN-view contrastive learning. However as RPN proposals are class agnostic, it is challenging to form positive (same class)/negative pairs (different class), which is essential for CRL.
2. Related works

Unsupervised Domain Adaption. Unsupervised domain adaptation for object detection was first explored by Chen et al. [8]. Chen et al. [8] proposed adversarial-based feature alignment for a Faster-RCNN network at image and instance level to mitigate the domain shift. Later, Saito et al. [56] proposed a method that performs strong local feature alignment and weak global feature alignment based on adversarial training. Instead of utilizing an adversarial-based approach, Khodabandeh et al. [30] proposed to mitigate domain shift by pseudo-label self-training on the target data. Self-training using pseudo-labels ensures that the detection model learns target representation. Recently, Hsu et al. [25] explored domain adaptation for one-stage object detection, where he utilized a one-stage detection framework to perform object center-aware training while performing adversarial feature alignment. Also, there exists multiple UDA work for object detection [2, 19, 50, 55, 59, 66, 67]; however, all these works assume you have access to labeled source and unlabeled target data.

Source-Free Domain Adaption. In a real-world scenario, the source data is not often accessible during the adaptation process due to privacy regulations, data transmission constraints, or proprietary data concerns. Many works have addressed the source-free domain adaptation (SFDA) setting for classification [38, 41], 2D and 3D object detection [21, 22, 27, 64] and video segmentation [48] tasks. First for the classification task, the SFDA setting was explored by Liang et al. [41] proposed source hypothesis transfer, where the source-trained model classifier is kept frozen and target generated features are aligned via pseudo-label training and information maximization. Following the segmentation task Liu et al. [47] proposed a self-supervision and knowledge transfer-based adaptation strategy for target domain adaptation. For object detection task, [40] proposed a pseudo-label self-training strategy and [27] proposed self-supervised feature representation learning via previous models approach.

Contrastive Learning. The huge success in unsupervised feature learning is due to contrastive learning which has attributed to huge improvement in many unsupervised tasks [5, 27, 52]. Contrastive learning generally learns a discriminative feature embedding by maximizing the agreement between positive pairs and minimizing the agreement with negative pairs. In [5, 17, 52], in batch of an image, an anchor image undergoes different augmentation and these augmentations for that anchor forms positive pair and negative pairs are sampled from other images in the given batch. Later, in [31] exploiting the task-specific semantic information, intra-class features embedding is pulled together and repelled away from cross-class feature embedding. In this way, [31] learned a more class discriminative feature representation. All these works are performed for the classification task, and these methods work well for large batch size tasks [5, 31]. Extending this to object detection tasks generally fails as detection models are computationally expensive. To overcome this, we exploit graph convolution networks to guide contrastive learning for object detection.

Graph Convolution Neural Networks (GNNs). Graph Convolution Neural Networks was first introduced by Gori [16] to process the data with a graph structure using neural networks. The key idea is to construct a graph with nodes and edges relating to each other and update node/edge features, i.e., a process called node feature aggregation. In recent years, different GNNs have been proposed (e.g., GraphConv [49], GCN [35], etc.) with a unique feature aggregation rule which is shown to be effective on various tasks. Recent works in image captioning [51, 76], scene graph parsing [72] etc. try to model inter-instance relations by IoU based graph generation. For these applications, IoU based graph is effective as modelling the interaction between objects is essential and can be achieved by simply constructing a graph based on object overlap. However, the problem arises with IoU based graph generation when two objects have no overlap and in these cases, it disregards the object relation. For example, see Fig. 3 (a), where the proposals for the left sidecar and right sidecar has no overlap; as a result, IoU based graph will output no relation between them. In contrast for the CRL case, they need to be treated as a positive pair. To overcome these issues, we propose a learnable graph convolution network to models inter-instance relations present within an image.

3. Proposed method

3.1. Preliminaries

Background. UDA [9, 24, 62] considers labeled source and unlabeled target domain datasets for adaptation. Let us formally denote the labeled source domain dataset as \( D_s = \{x_s^n, y_s^n\}_{n=1}^{N_s} \), where \( x_s^n \) denotes the \( n^{th} \) source image and \( y_s^n \) denotes the corresponding ground-truth, and the unlabeled target domain dataset as \( D_t = \{x_t^n\}_{n=1}^{N_t} \), where \( x_t^n \) denotes \( n^{th} \) the target image without the ground-truth annotations. In contrast, the SFDA setting [34, 40, 41, 47] considers a more practical scenario where the access to the source dataset is restricted and only a source-trained model \( \Theta \) and the unlabeled target data \( D_t \) are available during adaptation.

Mean-teacher based self-training. Self-training adaptation strategy updates the detection model on unlabeled target data using pseudo-labels generated by the source-trained model [30]. Reliable pseudo-labels are selected based on a confidence threshold and the pseudo-label supervision loss for the detection model is expressed as:

\[
\mathcal{L}_{SL} = L_{cls}^{rpn}(x_t^n, \hat{y}_i^n) + L_{reg}^{rpn}(x_t^n, y_t^n)
+ L_{cls}^{reg}(x_t^n, \hat{y}_i^n) + L_{reg}^{reg}(x_t^n, y_t^n),
\]

\( L_{cls}^{rpn} \) and \( L_{reg}^{rpn} \) denote the classification and regression loss of the proposed graph convolution network, respectively. The graph convolution network is trained with the pseudo-labels generated by the source-trained model. However, the mean-teacher based self-training strategy fails as it does not consider the domain shift. To overcome this, we propose a learnable graph convolution network to models inter-instance relations present within an image.
where pseudo label $\tilde{y}^n_t$ is obtained by filtering low confidence predictions. However, due to domain shift, the pseudo labels generated by the source-trained model are still noisy. Therefore, to effectively distill target domain information to the source-trained model, it is necessary to improve the quality of the pseudo labels [11, 46].

To this end, we utilize mean-teacher [61] which consists of student and teacher networks with parameters $\Theta_s$ and $\Theta_t$, respectively. In the mean-teacher, the student is trained with pseudo labels generated by the teacher and the teacher is progressively updated via Exponential Moving Average (EMA) of student weights. Furthermore, motivated by semi-supervised techniques [11, 46], the student and teacher networks are fed with strong and weak augmentations, respectively and consistency between their predictions improves detection on target data. Hence, the overall student-teacher self-training based object detection framework updates can be formulated as:

$$\Theta_s \leftarrow \Theta_s + \gamma \frac{\partial (L_{st}^{sl})}{\partial \Theta_s}$$  \hspace{1cm} (2)$$

$$\Theta_t \leftarrow \alpha \Theta_t + (1 - \alpha) \Theta_s$$ \hspace{1cm} (3)

where $L_{st}^{sl}$ is the student loss computed using the pseudo-labels generated by the teacher network. The hyperparameters $\gamma$ and $\alpha$ are student learning rate and teacher EMA rate, respectively. Although the student-teacher framework enables target knowledge distillation into the source-trained model, it is still not sufficient to learn high-quality target features. Hence, to enhance the features in the target domain, we utilize contrastive representation learning.

### 3.2. Graph-guided contrastive learning

The Contrastive Representation Learning framework assumes one category per image and relies on large batch sizes for generating positive/negative pairs [5, 6]. However, this approach fails in object detection tasks as detection models are computationally expensive for large batch sizes, and images contain multiple object instances. To overcome the challenges discussed earlier, we exploit the architecture of Faster-RCNN to design a novel contrastive learning strategy as shown in Fig. 4. As we discussed in Sec. 1, RPN by default, provides augmentation for each instance in an image. As shown in Fig. 3, cropping out the RPN proposals will provide multiple different views around each instance in an image. This property can be exploited to learn contrastive representation by maximizing the agreement between proposal features for the same instance and disagreement of the proposal features for different instances. However, RPN predictions are class agnostic and the unavailability of ground truth boxes for target domain makes it difficult to know which proposals belong to which instance. Consequently, for a given proposal as an anchor, sampling positive/ negative pairs become a challenging task. To this end, we introduce an Instance Relation Graph.
Graph (IRG) network that models inter-instance relations between the RPN proposals. IRG then provides pairwise labels by inspecting similarities between two proposals to identify positive/negative proposal pairs.

### 3.2.1 Instance Relation Graph (IRG):

Graph Convolution Network (GCN) is an effective way to understand the relationship and propagate information between the nodes [1, 68, 74]. The proposed IRG network utilizes GCN to learn the relationship between the RPN proposals by considering each RPN proposal features as a node. Let us denote IRG as $\mathcal{G} : \mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is nodes and $\mathcal{E}$ is edges of the graph network. The nodes in $\mathcal{V}$ corresponds to RoI features extracted from RPN proposals and edges $e_{i,j} \in \mathcal{E}$ encodes relationship between the $i^{th}$ and the $j^{th}$ proposals. We then aim to learn relation matrix $\mathcal{E}$, to find the relationship between the RPN proposals. Both the student and teacher networks share the IRG network for modeling relationships between object proposals.

**Nodes.** The nodes in IRG represent features of the RPN proposals obtained from RoI feature extractor. The nodes in $\mathcal{G}$ are denoted as $\mathcal{V} = \{v_1, v_2, ..., v_m\}$, where $v_m$ is the feature of the $m^{th}$ instance. Here, $m$ is the total number of RPN proposals. We set $m$ to 300 for both teacher and student. The teacher pipeline has input with weak augmentations; thus, the teacher RPN proposals are better and more consistent than strongly augmented student RPN. Hence, we use teacher RPN proposals to extract RoI features and construct IRG for both student and teacher networks.

**Edges.** The edges in the graph $\mathcal{G}$ are denoted as $\mathcal{E} = \{e_{i,j}\}_{m \times m}$, where $e_{ij}$ is the edge of the $v_i^{th}$ and $v_j^{th}$ nodes, denoting the relation of corresponding instances in the feature space and can be formally represented as:

$$ e_{ij} = \frac{\exp(S_{ij})}{\sum \exp(S_{ij})}, \text{ where } S_{ij} = f(v_i) \cdot g(v_j)^T, $$

where, $f$ and $g$ are learnable function.

### 3.2.2 Graph Distillation Loss (GDL).

Let us denote the input features to IRG as $F \in \mathbb{R}^{m \times d}$ where $m$ denoting the number of proposal instances and $d$ denoting the feature dimension of the RoI features. The features $F$ are then passed through graph convolution layers of IRG to model the inter-instance relations. The output features $\tilde{F}$ are calculated as:

$$ \tilde{F} = \text{ReLU}(\mathcal{E} F W), $$

where $W$ is a learnable weight matrix. Subsequently, both features $F$ and $\tilde{F}$ are fed into the RCNN classification layer to obtain class logits for each proposal. Let us denote the student and the teacher class logits corresponding to features $F$ as $Z_{st}$ and $Z_{te}$, respectively. Similarly, let us denote student and teacher class logits corresponding to IRG output features $\tilde{F}$ as $\tilde{Z}_{st}$ and $\tilde{Z}_{te}$, respectively. To supervise the IRG network parameters, we minimize the discrepancy between class logits $Z$ and $\tilde{Z}$ for both student and teacher pipeline in an end-to-end manner. In addition, we also minimize the discrepancy between student and teacher class logits $Z_{st}$ and $Z_{te}$ to maintain consistency between both pipelines. We denote this discrepancy as GDL which can be formally written as:

$$ L_{GDL} = \text{KL}(\sigma(Z_{st}), \sigma(\tilde{Z}_{st})) + \text{KL}(\sigma(Z_{te}), \sigma(\tilde{Z}_{te})) + \text{KL}(\sigma(Z_{st}), \sigma(Z_{te})), $$

where KL denotes the Kullback–Leibler divergence, $\sigma$ denotes softmax operator. Therefore, minimizing $L_{GDL}$ supervises the IRG network which in turn learns the instance relation matrix ($\mathcal{E}$).

### 3.2.3 Graph Contrastive Loss (GCL)

**Instance pairwise labels.** In order to utilize the contrastive loss, we need to understand the relation of the given anchor proposal with other RPN proposals to form positive/negative pairs. As mentioned earlier, this relation matrix ($\mathcal{E}$) is obtained from the IRG network, which learns how proposals are related to each other. For instance pairwise label generation, let us consider proposal instances $i$ and $j$ and its corresponding learned relation between them, $e_{ij} \in \mathcal{E}$. Now, one can obtain positive/negative pairs by simply setting a threshold $\epsilon$ on normalized $\mathcal{E}$ where the $e_{ij} > \epsilon$ would indicate that they are highly related, forming a positive pair and vice versa for the negative pairs. The pairwise labels between $i^{th}$ and $j^{th}$ proposal instances, denoted as $M_{ij}$, can be given as:

$$ M_{ij} = \begin{cases} 
0, & e_{ij} < \epsilon \\
1, & e_{ij} \geq \epsilon,
\end{cases} $$

where $\epsilon$ is a hyper parameter. Thus, for a given anchor proposal we obtain its corresponding positive and negative proposal pairs from $M_{ij}$.

**Instance pairwise logits.** As shown in Fig. 5, the RoI features $v_i$ are projected as key $k_i$ and query $q_i$ inorder to model better correlation among the RoI features [63]. For
given \(i^{th}\) RoI features, we obtain key, query and pairwise logits as follows:

\[
k_i = W_k \cdot v_i, \quad q_i = W_q \cdot v_i, \quad R_{ij} = q_i(k_j)^T,
\]

where \(W_k\) and \(W_q\) are linear layer weights and \(k_i, q_i\) and \(R_{ij}\) are key, query and instance pairwise logits. To this end, the contrastive loss can be computed from the instance pairwise logits \((R_{ij})\) and instance pairwise labels \((M_{ij})\).

**Contrastive loss.** Considering any \(i^{th}\) proposal as an anchor, where \(i \in I \equiv \{1, 2, ..., m\}\), let us define a set consisting of all the samples excluding the anchor as \(A(i) \equiv I \setminus \{i\}\). Further, using pairwise labels from \(M\), we can create a positive pair set defined as \(P(i) \equiv \{p \in I : M_{ij} = 1\} \setminus \{i\}\). For given \(i^{th}\) proposal, the Graph Contrastive Loss (GCL) can be formulated as:

\[
\mathcal{L}_{\text{GCL}} = \sum_{i \in I} - \log \left( \frac{1}{|P(i)|} \sum_{p \in P(i)} \exp \left( q_i(k_p)^T \right) \right) \sum_{a \in A(i)} \exp \left( q_i(k_a)^T \right).
\]

By training with the proposed loss \(\mathcal{L}_{\text{GCL}}\), the student network is encouraged to learn high-quality feature representations on the target domain. We show that it improves the detector’s performance by conducting experimental analysis in Sec. 4. Note that GCL is used only to update the student network parameters, whereas the teacher network parameters are updated via EMA.

### 3.3. Overall loss function

So far, we have introduced an Instance Relation Graph (IRG), Graph Distillation Loss (GDL), and Graph Contrastive Loss (GCL) to effectively tackle the source-free domain adaptation problem for detection. Then overall objective of our proposed SFDA method is formulated as:

\[
\mathcal{L}_{\text{SFDA}} = \mathcal{L}_{\text{SL}} + \mathcal{L}_{\text{GDL}} + \mathcal{L}_{\text{GCL}},
\]

### 4. Experiments and Results

To validate the effectiveness of our method, we compare our model performance with existing state-of-the-art UDA and SFDA methods on four different domain shift scenarios: 1) Adaptation to adverse weather, 2) Real to artistic, 3) Synthetic to real, and 4) Cross-camera. Note that in UDA we have access to both source and target domain data. However, in SFDA, we have access only to source-trained model and not the source domain data for adaptation.

**Implementation details:** Following the SFDA setting [34, 40], we adopt Faster-RCNN [54] with ImageNet [36] pre-trained ResNet50 [18] as the backbone. In all of our experiments, the input images are resized with a shorter side to be 600 while maintaining the aspect ratio and the batch size to 1. The source model is trained using SGD optimizer with a learning rate of 0.001 and momentum of 0.9 for 10 epochs. For the proposed framework, the teacher network EMA momentum rate \(\alpha\) is set equal to 0.9. In addition, the pseudo-labels generated by the teacher network with confidence greater than the threshold \(T = 0.9\) are selected for student training. We utilize the SGD optimizer to train the student network with a learning rate of 0.001 and momentum of 0.9 for 10 epochs. We report the mean Average Precision (mAP) with an IoU threshold of 0.5 for the teacher network on the target domain during the evaluation.

### 4.1. Quantitative comparison

#### 4.1.1 Adaptation to adverse weather

Given a model trained on clear weather condition, we aim to perform adaptation to images in adverse weather conditions like fog/haze etc. The Cityscapes [10] consist of 2,975 training and 500 validation images with 8 object categories: \textit{person, rider, car, truck, bus, train, motorcycle and bicycle}. The FoggyCityscapes [58] consist of images that are rendered from the Cityscapes dataset by integrating fog and depth information. To this end, a model trained on Cityscapes is adapted to FoggyCityscapes without having access to the Cityscapes.

**Results.** Table 1 provides the quantitative comparison with the existing UDA and SFDA methods for Cityscapes \(\rightarrow\) FoggyCityscapes adaptation scenario. From Table 1, we can infer that the proposed method outperforms most of the existing UDA methods such as SWDA [56], InstanceDA [67], and CategoricalDA [71]. However, compared with MeGA-CDA [65] and Unbiased DA [11] methods, our proposed method produces a competitive performance with a drop of 2.5-3.5 mAP. But it is worth noting that, these method make use of labelled source data during adaptation whereas our proposed method only has access to source-trained model. Furthermore, compared with existing SFDA methods, SFOD [40] and HCL [27], the proposed method provides improvement of 3.5 mAP and 2.4 mAP, respectively. We also compared with mean-teacher

<table>
<thead>
<tr>
<th>Method</th>
<th>cam</th>
<th>fog</th>
<th>haze</th>
<th>synthetic</th>
<th>train</th>
<th>motorcycle</th>
<th>bicycle</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>UDA Faster [4]</td>
<td>35.5</td>
<td>31.8</td>
<td>40.5</td>
<td>32.1</td>
<td>35.3</td>
<td>20.2</td>
<td>20.0</td>
<td>27.1</td>
</tr>
<tr>
<td>DMD [5]</td>
<td>38.8</td>
<td>40.5</td>
<td>44.3</td>
<td>37.2</td>
<td>38.4</td>
<td>34.5</td>
<td>28.4</td>
<td>32.2</td>
</tr>
<tr>
<td>MTrack [2]</td>
<td>30.6</td>
<td>41.4</td>
<td>44.0</td>
<td>21.9</td>
<td>38.6</td>
<td>40.6</td>
<td>35.3</td>
<td>35.1</td>
</tr>
<tr>
<td>SWDA [56]</td>
<td>29.9</td>
<td>42.3</td>
<td>43.5</td>
<td>24.5</td>
<td>36.2</td>
<td>32.6</td>
<td>30.0</td>
<td>33.3</td>
</tr>
<tr>
<td>CDNN [66]</td>
<td>35.4</td>
<td>45.7</td>
<td>50.9</td>
<td>30.1</td>
<td>42.5</td>
<td>29.8</td>
<td>30.8</td>
<td>36.5</td>
</tr>
<tr>
<td>Collaborative DA [75]</td>
<td>32.7</td>
<td>44.4</td>
<td>50.1</td>
<td>21.7</td>
<td>45.6</td>
<td>25.4</td>
<td>30.1</td>
<td>36.8</td>
</tr>
<tr>
<td>iFACDA [13]</td>
<td>32.6</td>
<td>46.5</td>
<td>22.8</td>
<td>40.0</td>
<td>33.0</td>
<td>45.5</td>
<td>31.7</td>
<td>27.9</td>
</tr>
<tr>
<td>InstantDA [78]</td>
<td>33.1</td>
<td>43.4</td>
<td>49.6</td>
<td>21.9</td>
<td>45.7</td>
<td>32.0</td>
<td>29.5</td>
<td>37.0</td>
</tr>
<tr>
<td>Progressive DA [26]</td>
<td>36.0</td>
<td>45.5</td>
<td>34.4</td>
<td>24.3</td>
<td>44.1</td>
<td>25.8</td>
<td>29.1</td>
<td>35.9</td>
</tr>
<tr>
<td>CategoricalDA [71]</td>
<td>32.9</td>
<td>38.9</td>
<td>49.2</td>
<td>27.2</td>
<td>45.1</td>
<td>36.4</td>
<td>30.3</td>
<td>34.6</td>
</tr>
<tr>
<td>MeGA-CDA [65]</td>
<td>37.7</td>
<td>49.0</td>
<td>52.4</td>
<td>25.4</td>
<td>49.2</td>
<td>46.9</td>
<td>34.5</td>
<td>39.0</td>
</tr>
<tr>
<td>UnbiasedDA [11]</td>
<td>33.8</td>
<td>47.3</td>
<td>49.8</td>
<td>30.0</td>
<td>48.2</td>
<td>42.1</td>
<td>33.9</td>
<td>40.4</td>
</tr>
<tr>
<td>Metrics [9]</td>
<td>21.7</td>
<td>41.6</td>
<td>43.4</td>
<td>32.2</td>
<td>17.5</td>
<td>25.7</td>
<td>18.5</td>
<td>21.1</td>
</tr>
<tr>
<td>SFOD-Mosaic [40]</td>
<td>25.5</td>
<td>44.5</td>
<td>40.7</td>
<td>33.2</td>
<td>22.2</td>
<td>28.4</td>
<td>34.1</td>
<td>39.0</td>
</tr>
<tr>
<td>SFDA [66]</td>
<td>26.9</td>
<td>46.8</td>
<td>44.3</td>
<td>33.0</td>
<td>25.0</td>
<td>28.1</td>
<td>26.9</td>
<td>40.7</td>
</tr>
<tr>
<td>LIDS [39]</td>
<td>34.0</td>
<td>45.7</td>
<td>48.8</td>
<td>27.3</td>
<td>39.7</td>
<td>19.6</td>
<td>33.2</td>
<td>37.8</td>
</tr>
<tr>
<td>Mean-Teacher [61]</td>
<td>33.9</td>
<td>43.0</td>
<td>45.0</td>
<td>29.2</td>
<td>37.2</td>
<td>25.1</td>
<td>25.6</td>
<td>38.2</td>
</tr>
<tr>
<td>IRG (Ours)</td>
<td>37.4</td>
<td>45.2</td>
<td>51.9</td>
<td>24.4</td>
<td>39.6</td>
<td>25.2</td>
<td>31.5</td>
<td>41.6</td>
</tr>
</tbody>
</table>

Table 1. Quantitative results (mAP) for Cityscapes \(\rightarrow\) FoggyCityscapes. S: Source only; O: Oracle, UDA: Unsupervised Domain Adaptation, SFDA: Source-Free Domain Adaptation.
self-training baseline to show that adding the proposed GCL loss is able to enhance the features representation on the target domain, providing an improvement of 3.5 mAP.

### 4.1.2 Realistic to artistic data adaptation:

Here, we consider adaptation to dissimilar domains [56], where a model trained on the real-world images is aimed to perform adaptation towards artistic domain. We consider the model trained on the Pascal-VOC dataset [13] and adapt to two target domains, namely, Clipart [28] and Watercolor [28]. The Clipart dataset contains 1K unlabeled images and has the same 20 categories as Pascal-VOC. The Watercolor consists of 1K training and 1K testing images with six categories.

**Results.** The PASCAL-VOC→Clipart adaptation results are reported in Table 4. Our method outperforms the existing UDA methods such as ADDA [28] and DANN [14] by a margin of 4.7 mAP and 0.3 mAP, respectively. Moreover, the PASCAL-VOC→Watercolor adaptation results are reported in Table 3. Even in this case, our method outperforms the state-of-the-art UDA methods such as SWDA [56] and I²Net [4] by 2.6 mAP and 1.5 mAP, respectively. Furthermore, for both Clipart and Watercolor adaptation scenarios, our method consistently outperforms in every category compared with pseudo-label self-training (PL) and mean-teacher baseline.

### 4.1.3 Synthetic to real-world adaptation

The cost of generating and labeling synthetic data is low compared to real-world data. Hence, it makes sense to train a detector on synthetic images and transfer the knowledge to real-world data. However, the style shift between synthetic to real domain makes it challenging. Here, we consider such scenario where we adapt a model trained on the synthetic data, Sim10K [29], to a real-world data, Cityscapes [10] under SFDA condition, i.e., synthetic data are not available while adapting the model to the real-world images. The model is trained on 10,000 training images of Sim10k rendered by the Grand Theft Auto gaming engine. The target Cityscapes dataset consists of 2,975 training and 500 validation images.

**Results.** We report the results of Sim10K→Cityscapes in Table 2. Note that even though we adapt for only the car category, the proposed GCL training strategy is able to get discriminative positive pairs for different cars and improve the feature representations through contrastive training. Our proposed method outperforms existing UDA method like Cycle DA [75], Unbiased DA [11] etc. by considerable margin. Under SFDA setting, the proposed method produces state-of-the-art performance by improving ~2 mAP compared to SFOD [40].

**Cross-camera adaptation** In real-world scenarios, the target domain data is captured by a camera with configurations different from the source data. To emulate this cross-camera conditions, we consider a model trained on source, KITTI dataset [15], is adapted to target, Cityscapes [10]. The KITTI dataset consists of 7,481 training images, which is used to get the source-trained detector model. The model is then adapted to the target domain dataset, i.e., Cityscapes.

**Results.** KITTI→Cityscapes results are reported in Table 2. Our method outperform existing state-of-the-art UDA methods like Cycle DA [75], MeGA CDA [65] and Unbiased DA [11] by considerable margin. Further in SFDA setting, the proposed method produce state-of-the-art performance by improving around 3.3 mAP compared to SFOD.
instance pairwise labels. We can observe the IRG network models the relationship between the proposal, which maximizes the similarity between similar proposals and vice versa for dissimilar proposals.

Figure 6. Relation matrix analysis for 25 proposal RoI features before and after passing through IRG network and corresponding masked feature representations. Subsequently, addition of distillation loss enhances the model performance from 34.3 mAP to 35.9 mAP. Finally, utilizing graph-guided contrastive learning on the proposal features further helps the model learn high-quality representations, resulting in an increase in performance by 1.9 mAP on the target domain.

**Qualitative analysis.** In Fig. 6, we show the relation matrix for the RoI features before and after it is processed by IRG. For better visualizations, we consider 25 out of 300 RoI features. It can be observed that relation between the proposals are poorly defined and IRG network is able to improve these relations through graph-based feature aggregation.

### 5. Conclusion

In this work, we presented a novel approach for source-free domain adaptive detection using graph-guided contrastive learning. Specifically, we introduced a contrastive graph loss to enhance the target domain representations by exploiting instance relations. We propose an instance relation graph network built on top of a graph convolution network to model the relation between proposal instances. Subsequently, the learned instance relations are used to get positive/negative proposal pairs to guide contrastive learning. We conducted experiments on multiple detection benchmarks to show the effect of various combinations of weak and strong augmentation for a mean-teacher framework. We conducted experiments on multiple detection benchmarks to show the effect of various combinations of weak and strong augmentation for a mean-teacher framework.

### 4.2. Ablation analysis

We study the impact of the proposed GCL and IRG network by performing an in-depth ablation analysis on Cityscapes → FoggyCityscapes adaptation scenario.

**Quantitative analysis.** The results for Cityscapes → FoggyCityscapes ablation experiments are reported in Table 5. In Table 5, the first three experiments are performed to analyze the effect of various combinations of weak and strong augmentation for a mean-teacher framework in an SFDA setting. More precisely, we input the student and teacher network with Weak-Weak (WW), Strong-Strong (SS) and Strong-Weak (SW) augmented images, respectively. These three experiments show that strong-weak (SW) produces consistent and improved results compared to other variations. This is due to mutual learning between student and teacher networks, where student trains on strong augmentation leading to robust prediction and the teacher supervise the student by good pseudo-labels predicted from the weak augmented images. Furthermore, minimizing the discrepancy between instance relation graph network of student and teacher framework ensures consistency between student and teacher graph proposal feature representations. Subsequently, addition of graph

---

Table 4. Quantitative results (mAP) for PASCAL-VOC → Clipart. S: Source only, UDA: Unsupervised Domain Adaptation, SFDA: Source-Free domain adaptation.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td></td>
<td>35.6</td>
<td>32.8 10.7</td>
<td>30.6 11.7</td>
<td>36.8 55.9 48.7</td>
<td>41.9 16.5</td>
</tr>
<tr>
<td>DANN [14]</td>
<td>24.1 52.5 24.3 32.8 35.4 3.8</td>
<td>35.1 23.6 23.9 13.8 22.5 50.9 49.9 56.3 13.6 31.3 38.0 31.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UDA DAF [9]</td>
<td>15.0 34.6 12.4 11.9 19.8 21.1 23.3 3.10</td>
<td>22.1 26.3 10.6 10.0 19.6 39.4 34.6 29.3 10.0 171.19 24.8 19.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADDA [28]</td>
<td>20.1 50.2 20.5 23.6 11.4 40.5 34.9 2.3</td>
<td>39.7 22.3 27.1 10.4 31.7 53.6 46.6 32.1 18.0 21.1 23.6 18.3 27.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BDC Faster [50]</td>
<td>20.2 46.4 20.4 19.3 18.7 41.3 26.5 6.40</td>
<td>33.2 11.7 26.0 1.7 36.6 41.5 37.7 44.5 10.6 20.4 33.3 15.5 25.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL [61]</td>
<td>18.3 48.4 19.2 22.4 12.8 38.9 36.1 5.2</td>
<td>36.9 24.8 29.3 9.09 34.6 58.6 43.1 34.3 9.09 14.4 26.9 19.8 28.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MT + WW</td>
<td>20.2 46.4 20.4 19.3 18.7 41.3 26.5 6.40</td>
<td>33.2 11.7 26.0 1.7 36.6 41.5 37.7 44.5 10.6 20.4 33.3 15.5 25.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SFDA Mean-teacher [61]</td>
<td>22.3 42.3 23.8 21.7 23.5 60.7 33.2 9.1</td>
<td>24.7 16.7 12.2 13.1 26.5 73.6 43.9 34.5 9.09 24.3 37.9 42.2 29.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRG (Ours)</td>
<td>20.3 47.3 27.3 19.7 30.5 54.2 36.2 10.3 351.0</td>
<td>20.6 20.2 12.3 28.7 53.1 47.5 42.4 9.09 21.1 42.3 50.3 31.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Ablation study on FoggyCityscapes.

| Method | PL DGF GCL gpl rider car truck bus train bicycle mAP |
|--------|-----------------|-------------|-------------|-------------|
| MT + WW | ✓ ✓ ✓ ✓ | 35.8 42.6 43.9 23.1 32.7 11.0 29.9 38.7 32.2 |
| MT + SS | ✓ ✓ ✓ ✓ | 32.8 41.4 43.8 18.2 28.6 11.2 24.6 38.3 29.9 |
| MT + SW | ✓ ✓ ✓ ✓ | 33.9 43.0 45.6 29.1 37.2 25.1 25.5 38.2 34.3 |
| Ours | ✓ ✓ ✓ ✓ | 37.2 43.1 51.0 28.6 40.1 21.2 28.2 37.1 35.9 |
| Ours | ✓ ✓ ✓ ✓ | 37.4 45.2 51.9 24.4 39.6 25.2 31.5 41.6 37.1 |
References


[34] Kim, Y., Cho, D., Han, K., Panda, P., Hong, S.: Domain adaptation without source data. IEEE Transactions on Artifical Intelligence (2021) 3, 6


[40] Li, X., Chen, W., Xie, D., Yang, S., Yuan, P., Pu, S., Zhuang, Y.: A free lunch for unsupervised domain adaptive object detection without source data. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 8474–8481 (2021) 2, 3, 6, 7, 8


