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 [13, 43, 45, 46, 55, 81]. The major contributor to this success is the availability of large-scale annotated detection datasets [11, 14, 16, 44, 75], 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, 15, 24, 25, 50, 59, 64]. Specifically, UDA methods aim to minimize the domain discrepancy by aligning the feature distribution of the detector model between source and target domain [9, 20, 29, 58, 61]. 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 [28, 38]. Moreover, transmitting only source-trained model alleviates many privacy/safety, data proprietary concerns as well [42, 48, 72]. 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).

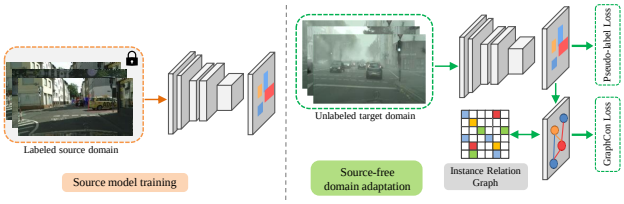
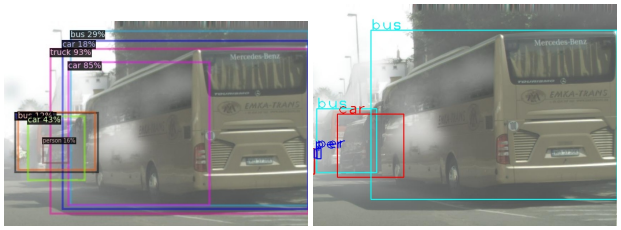


Figure 1. **Left:** Supervised training of detection model on the source domain. **Right:** Source-Free Domain Adaptation (SFDA) setup, i.e., the source-trained model is adapted to the target domain in the absence of source data with pseudo-label self-training and proposed Instance Relation Graph (IRG) network guided contrastive loss.

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 train-



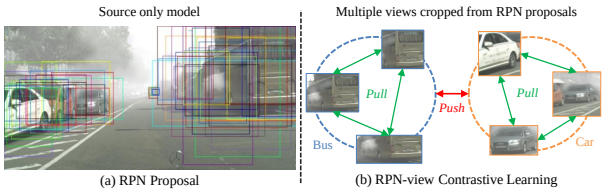
(a) Prediction (b) Ground truth

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.

ing with pseudo-labels generated by source-trained model [28, 41]. However, these pseudo-labels are noisy due to domain shift and the training on noisy pseudo-label is sub-optimal solution [12, 47]. In order to avoid such a scenario, we need to consider two challenges of the SFDA training, *1) Effectively distill target domain information into source-trained model and 2) Enhancing the target domain feature representations*.

A critical challenge is improving the features of the target domain data. Consider Fig. 2, which shows object proposals for an image from FoggyCityscapes [60], predicted by a detector model trained on Cityscapes [11]. 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 biassed towards source data, resulting in weak representation for the target data. To this end, we utilize the Contrastive Representation Learning (CRL) framework [5, 10, 18, 32] and design a novel contrastive loss to enhance the feature representations of the target domain.

CRL has been shown to learn high-quality representations from images in an unsupervised manner [5, 6, 71]. 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



(a) RPN Proposal (b) RPN-view Contrastive Learning

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.

Faster-RCNN [56]. 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 and without the ground-truth annotations for target domain, it is impossible to know which of these proposals would form positive (same class)/negative pairs (different class), which is essential for CRL. To this end, we propose a Graph Convolution Network (GCN) based network that models the inter-instance relations for generated RPN proposals. Specifically, each node corresponds to a proposal and the edges represent the similarity relations between the proposals. This learned similarity relations are utilized to extract information regarding which proposals would form positive/negative pairs and are used to guide CRL. By doing so, we show that such graph-guided contrastive representation learning is able to enhance representations for the target data.

Our contributions are summarized as follows:

- We investigate the problem of source-free domain adaptation for object detection and identify some of the major challenges that need to be addressed.
- We introduced an Instance Relation Graph (IRG) framework to model the relationship between proposals generated by the region proposal network.
- We propose a novel contrastive loss which is guided by the IRG network to improve the representations for the target data.
- The effectiveness of the proposed method is evaluated on multiple object detection benchmarks comprising of visually distinct domains. Our method outperforms existing source-free domain adaptation methods and many unsupervised domain adaptation methods.

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*. [58] 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*. [31] 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. Later, Kim *et al*. [34] proposed an image-to-image generation based adaptation strategy where given source and target domain, the proposed method generates target like source images. The generated target-like images are then used to train the detection model; as a result, the detection network learn target features. Recently, Hsu *et al*. [26] explored domain adaptation for one-stage object detection, where he utilized a one-stage object detection framework to perform object center-aware training while performing adversarial feature alignment. There exists multiple UDA work for object detection [2, 20, 52, 57, 61, 68, 69]; however, all these works assume you have access to labeled source and unlabeled target data.

Source-Free Domain Adaptation. 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 [39, 42], 2D and 3D object detection [22, 23, 28, 66] and video segmentation [49] tasks. First for the classification task, the SFDA setting was explored by Liang *et al*. [42] 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*. [48] proposed a self-supervision and knowledge transfer-based adaptation strategy for target domain adaptation. For object detection task, [41] proposed a pseudo-label self-training strategy and [28] 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, 28, 54]. Contrastive learning generally learns a discriminative feature embedding by maximizing the agreement between positive pairs and minimizing the agreement with negative pairs. In [5, 18, 54]. 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 [32] exploiting the task-specific semantic information, intra-class features embedding is pulled together and repelled away from cross-class feature embedding. In this way, [32] 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, 32]. 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 [17] 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 [51], GCN [36], each with a unique feature aggregation rule which is shown to be effective on various tasks. Recent works in image captioning [53, 78], scene graph parsing [74] 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 araises 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, 25, 64] 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 [35, 41, 42, 48] considers a more practical scenario where the access to the source dataset is restricted and only a source-trained model Θ and the unlabeled target data D_t are available during adaptation.

Mean-teacher based self-training. Self-training adaptation strategy updates the model on unlabeled target data using pseudo labels generated by the source-trained model. The pseudo labels are filtered through confidence threshold and the reliable ones are used to supervise the detector training [31]. More formally, the pseudo label supervision loss

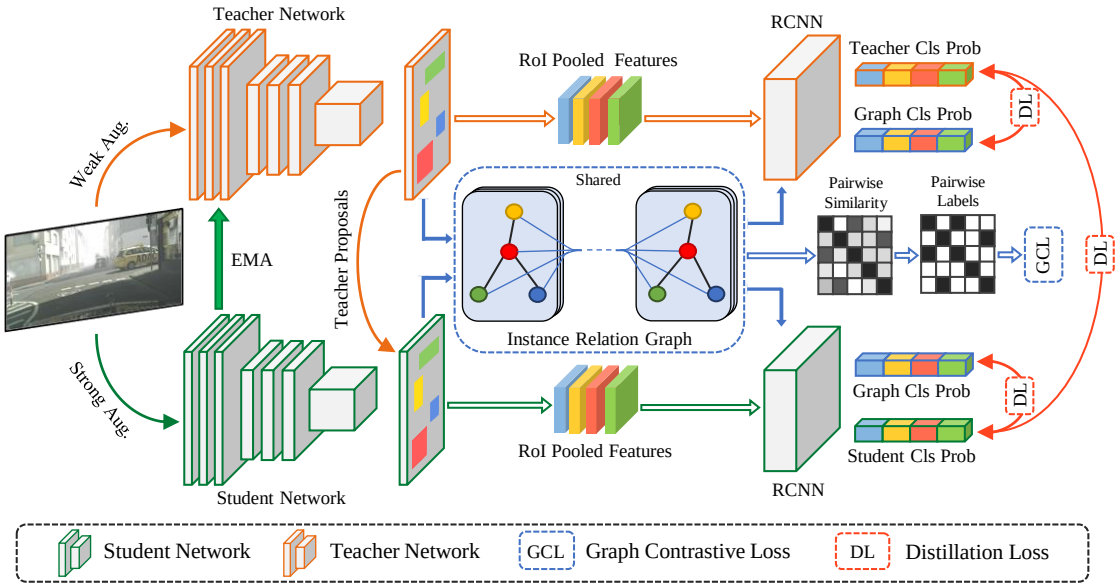


Figure 4. Overall architecture of our method. We follow a student-teacher framework for the detector model training. The proposed Instance Relation Graph (IRG) network models the relation between the object proposals generated by the detector. Using the inter-proposal relations learned by IRG, we generate pairwise labels to identify positive/negative pairs for contrastive learning. The IRG network is regularized with distillation loss between student and teacher model.

for the object detection model can be given as:

$$\mathcal{L}_{SL} = \mathcal{L}^{rpn}_{cls}(x^n_t, \tilde{y}^n_t) + \mathcal{L}^{rpn}_{reg}(x^n_t, \tilde{y}^n_t) + \mathcal{L}^{roi}_{cls}(x^n_t, \tilde{y}^n_t) + \mathcal{L}^{roi}_{reg}(x^n_t, \tilde{y}^n_t), \quad (1)$$

where $\tilde{y}^n_t$ is the pseudo label obtained after filtering low confident predictions. Even after filtering out low confidence predictions, the pseudo labels generated by the source-trained model are still noisy due to the domain shift. Therefore, to effectively distill knowledge from a source-trained model, the pseudo-labels quality need to be improved [12, 47].

To this end, we utilize mean-teacher [63] which consists of student and teacher networks with parameters Θ_s and Θ_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 [12, 47], 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:

$$\Theta_s \leftarrow \Theta_s + \gamma \frac{\partial(\mathcal{L}^{st}_{SL})}{\partial \Theta_s}, \quad (2)$$

$$\Theta_t \leftarrow \alpha\Theta_t + (1-\alpha)\Theta_s, \quad (3)$$

where $\mathcal{L}^{st}_{SL}$ is the student loss computed using the pseudo-labels generated by the teacher network. The hyperparameters γ and α are student learning rate and teacher EMA rate, respectively. Although the student-teacher framework enables knowledge distillation with noisy pseudo-labels, it is not sufficient to learn high-quality target features, as discussed earlier. Hence, to enhance the features in the target domain, we utilize contrastive representation learning.

Contrastive Representation Learning (CRL). SimCLR [5] is a commonly used CRL framework, which learns representations for an image by maximizing agreement between differently augmented views of the same sample via a contrastive loss. More formally, given an anchor image x_i, the SimCLR loss can be written as:

$$\mathcal{L}_{\text{SimCLR}} = -\log\left(\frac{\exp(\text{sim}(r_i, r_j))}{\sum_{k=1,\exists k\neq i}^{2N}\exp(\text{sim}(r_i, r_k))}\right), \quad (4)$$

where N is the batch size, r_i and r_j are the features of two different augmentations of the same sample x_i, whereas r_k represents the feature of k^{th} batch sample x_k, where $k \neq i$. Also, $\text{sim}(\cdot,\cdot)$ indicates a similarity function, e.g. cosine similarity. Note that, in general the CRL framework assumes that each image contains one category [5]. Moreover, it requires large batch sizes that could provide multiple positive/negative pairs for the training [6].

3.2. Graph-guided contrastive learning

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 (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, 70, 76]. The proposed IRG network utilizes GCN to learn the relationship between the RPN proposals. Let us denote IRG as $\mathcal{G} : \mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, 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 $\mathcal{E} = [e_{ij}]_{m \times m}$, 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:

$$e_{ij} = \frac{\exp(S_{ij})}{\sum \exp(S_{ij})}, \text{ where } S_{ij} = f(v_i) \cdot g(v_j)^T, \quad (5)$$

where, f and g are learnable function helps to model relation between nodes.

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:

$$\tilde{F} = \text{ReLU}(\mathcal{E}FW), \quad (6)$$

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:

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

where KL denotes the Kullback–Leibler divergence, σ denotes softmax operator. Therefore, minimizing $\mathcal{L}_{GDL}$ supervises the IRG network which inturn learns the instance relation matrix ($\mathcal{E}$).

3.2.2 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 it's corresponding learned relation between them, $e_{ij} \in \mathcal{E}$. Now, one can obtain positive/negative pairs by simply setting a threshold ϵ 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:

$$M_{ij} = \begin{cases} 0, & e_{ij} < \epsilon \\ 1, & e_{ij} \geq \epsilon, \end{cases} \quad (7)$$

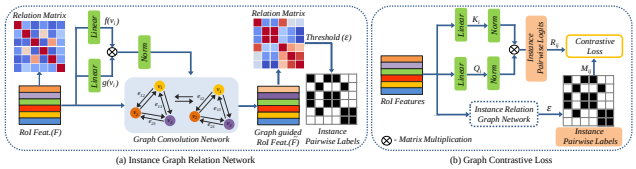


Figure 5. **(a) Instance Graph Relation Network:** Given proposal RoI features, the IRG models and improves the similarity relations between proposals. Thresholding the learned relation matrix generates instance pairwise labels used to obtain positive (white)/negative (black) pairs for computing the contrastive loss. **(b) Graph Contrastive Loss:** Projecting RoI features as keys and queries and performing transpose multiplication provides instance pair wise logits. The generated instance pairwise logits and instance pairwise labels are used to compute the contrastive loss.

Table 1. Quantitative results (mAP) for Cityscapes → FoggyCityscapes. S: Source only, O: Oracle, UDA: Unsupervised Domain Adaptation, SFDA: Source-Free Domain Adaptation.

Type	Method	prsn	rider	car	truck	bus	train	mcycle	bicycle	mAP
S	Source Only	29.3	34.1	35.8	15.4	26.0	9.09	22.4	29.7	25.2
UDA	DA Faster [8]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
	D&Match [34]	30.8	40.5	44.3	27.2	38.4	34.5	28.4	32.2	34.6
	MAF [20] (ICCV 2019)	28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
	Robust DA [31] (ICCV 2019)	35.1	42.1	49.1	30.0	45.2	26.9	26.8	36.0	36.4
	MTOR [2]	30.6	41.4	44.0	21.9	38.6	40.6	28.3	35.6	35.1
	SWDA [58]	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
	CDN [62]	35.8	45.7	50.9	30.1	42.5	29.8	30.8	36.5	36.6
	Collaborative DA [77]	32.7	44.4	50.1	21.7	45.6	25.4	30.1	36.8	35.9
	iFAN DA [80]	32.6	48.5	22.8	40.0	33.0	45.5	31.7	27.9	35.3
	Instance DA [80]	33.1	43.4	49.6	21.9	45.7	32.0	29.5	37.0	36.5
	Progressive DA [27]	36.0	45.5	54.4	24.3	44.1	25.8	29.1	35.9	36.9
	Categorical DA [73]	32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6	37.4
	MeGA CDA [27]	37.7	49.0	52.4	25.4	49.2	46.9	34.5	39.0	41.8
	Unbiased DA [12]	33.8	47.3	49.8	30.0	48.2	42.1	33.0	37.3	40.4
SFDA	SFOD [41]	21.7	44.0	40.4	32.2	11.8	25.3	34.5	34.3	30.6
	SFOD-Mosaic [41]	25.5	44.5	40.7	**33.2**	22.2	**28.4**	34.1	39.0	33.5
	HCL [28]	26.9	**46.0**	41.3	33.0	25.0	28.1	**35.9**	40.7	34.6
	LODS [40]	34.0	45.7	48.8	27.3	39.7	19.6	33.2	37.8	35.8
	Mean-Teacher [63]	33.9	43.0	45.0	29.2	37.2	25.1	25.6	38.2	34.3
	IRG (Ours)	**37.4**	45.2	**51.9**	24.4	**39.6**	25.2	31.5	**41.6**	**37.1**
O	Oracle	38.7	46.9	56.7	35.5	49.4	44.7	35.9	38.8	43.1

where ϵ 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 [65]. For given i^{th} RoI features, we obtain key, query and pairwise logits as follows:

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

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 calculated as:

$$\mathcal{L}_{GCL} = \sum_{i \in I} -\log \left\{ \frac{1}{|P(i)|} \sum_{p \in P(i)} \frac{\exp\left(q_i (k_p)^T\right)}{\sum_{a \in A(i)} \exp\left(q_i (k_a)^T\right)} \right\}, \tag{8}$$

By training with the proposed loss $\mathcal{L}_{GCL}$, 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 objective

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:

$$\mathcal{L}_{SFDA} = \mathcal{L}_{SL}^{st} + \mathcal{L}_{GDL} + \mathcal{L}_{GCL}, \tag{9}$$

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.

4.1. Implementation details

Following the SFDA setting [35, 41], we adopt Faster-RCNN [56] with ImageNet [37] pre-trained ResNet50 [19] 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 α 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.

4.2. Quantitative comparison

4.2.1 Adaptation to adverse weather

Description. 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 [11] consist of 2,975 training images and 500 validation images with 8 object categories: *person, rider, car, truck, bus, train, motorcycle and bicycle*. The FoggyCityscapes [60] 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 dataset.

Results. Table 1 provides the quantitative comparison with the existing UDA and SFDA methods for Cityscape→FoggyCityscapes adaptation scenario. From Table 1, we can infer that the proposed method outperforms most of the existing UDA methods such as SWDA [58], InstanceDA [69], and CategoricalDA [73]. However, compared MeGA-CDA [67] and Unbiased DA [12] methods, our proposed method produces a competitive performance with a drop of 2.5-3.5 mAP. But it is worth noting that,

Table 2. Quantitative results for Sim10K → Cityscapes and KITTI → Cityscapes. S: Source only, UDA: Unsupervised Domain Adaptation, SFDA: Source-Free domain adaptation.

Type	Method	Sim10k → City	Kitti → City
		AP of Car	AP of Car
S	Source Only	32.0	33.9
UDA	DA Faster [8]	38.9	38.5
	Selective DA [79]	43.0	42.5
	MAF [20]	41.1	41.0
	Robust DA [31]	42.5	42.9
	Strong-Weak [58]	40.1	37.9
	ATF [21]	42.8	42.1
	Harmonizing [3]	42.5	-
	Cycle DA [77]	41.5	41.7
	MeGA CDA [67]	44.8	43.0
	Unbiased DA [12]	43.1	-
SFDA	SFOD [41]	42.3	43.6
	SFOD-Mosaic [41]	42.9	44.6
	Mean-teacher [63]	39.7	41.2
	IRG (Ours)	**45.2**	**46.9**

Table 3. Quantitative results for PASCAL-VOC → Watercolor. S: Source only, UDA: Unsupervised Domain Adaptation, SFDA: Source-Free domain adaptation.

Type	Method	bike	bird	car	cat	dog	prsn	mAP
S	Source only	68.8	46.8	37.2	32.7	21.3	60.7	44.6
UDA	DA Faster [8]	75.2	40.6	48.0	31.5	20.6	60.0	46.0
	BDC Faster [58]	68.6	48.3	47.2	26.5	21.7	60.5	45.5
	BSR [33]	82.8	43.2	49.8	29.6	27.6	58.4	48.6
	WST [33]	77.8	48.0	45.2	30.4	29.5	64.2	49.2
	SWDA [58]	71.3	52.0	46.6	36.2	29.2	67.3	50.4
	HTCN [3]	78.6	47.5	45.6	35.4	31.0	62.2	50.1
	I^3Net [4]	81.1	49.3	46.2	35.0	31.9	65.7	51.5
	Unbiased DA [12]	88.2	55.3	51.7	39.8	43.6	69.9	55.6
SFDA	PL [31]	74.6	46.5	45.1	27.3	25.9	54.4	46.1
	SFOD [41]	**76.2**	44.9	49.3	**31.6**	30.6	55.2	47.9
	Mean-teacher [63]	73.6	47.6	46.6	28.5	29.4	56.6	47.1
	IRG (Ours)	75.9	**52.5**	**50.8**	30.8	**38.7**	**69.2**	**53.0**

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 [41] and HCL [28], the proposed method provides improvement of 4.3 mAP and 3.2 mAP, respectively. We also compared with mean-teacher 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.

4.2.2 Realistic to artistic data adaptation

Description. Here, we consider adaptation to dissimilar domains [58], 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 [14] and adapt to two target domains, namely, Clipart [29] and Watercolor [29]. 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 [29] and DANN [15] 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 [58] and I^3Net [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.

4.2.3 Synthetic to real-world adaptation

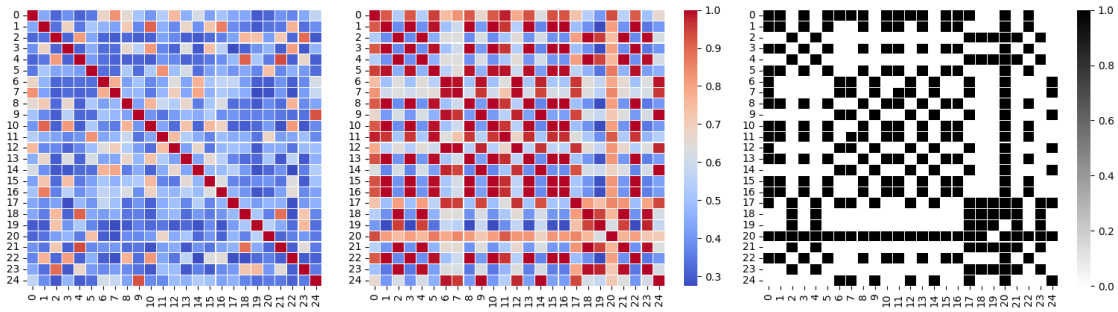
Description. The cost of generating and labeling synthetic data is low compared to real-world data. Hence, it makes sense to train a detector on synthetic images and transfer the knowledge to real-world data. However, the style shift between synthetic to real domain makes it challenging. Here, we consider such scenario where we adapt a model trained on the synthetic data, Sim10K [30], to a real-world data, Cityscapes [11] 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 unlabeled training images 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 [77], Unbiased DA [12] etc. by considerable margin. Under SFDA setting, the proposed method produces state-of-the-art performance by improving ∼1 mAP compared to SFOD [41].

4.2.4 Cross-camera adaptation

Description. 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 [16], is adapted to target, Cityscapes [11]. 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



(a) Proposal relation before IRG (b) Proposal relation after IRG (c) Masked relation after IRG

Figure 6. Relation matrix analysis for 25 proposal RoI features before and after passing through IRG network and corresponding masked 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.

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

Type	Method	aero	bcycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	prsn	plnt	sheep	sofa	train	tv	mAP
S	Source only	35.6	52.5	24.3	23.0	20.0	43.9	32.8	10.7	30.6	11.7	13.8	6.0	36.8	45.9	48.7	41.9	16.5	7.3	22.9	32.0	27.8
	DANN [15]	24.1	52.6	27.5	18.5	20.3	59.3	37.4	3.8	35.1	32.6	23.9	13.8	22.5	50.9	49.9	36.3	11.6	31.3	48.0	35.8	31.8
	DAF [8]	15.0	34.6	12.4	11.9	19.8	21.1	23.3	3.10	22.1	26.3	10.6	10.0	19.6	39.4	34.6	29.3	1.00	17.1	19.7	24.8	19.8
	ADDA [29]	20.1	50.2	20.5	23.6	11.4	40.5	34.9	2.3	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
UDA	BDC Faster [58]	20.2	46.4	20.4	19.3	18.7	41.3	26.5	6.40	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
	CRDA [73]	28.7	55.3	31.8	26.0	40.1	63.6	36.6	9.4	38.7	49.3	17.6	14.1	33.3	74.3	61.3	46.3	22.3	24.3	49.1	44.3	38.3
	Unbiased DA [12]	39.5	60.0	30.5	39.7	37.5	56.0	42.7	11.1	49.6	59.5	21.0	29.2	49.5	71.9	66.4	48.0	21.2	13.5	38.8	50.4	41.8
	PL [63]	18.3	48.4	19.2	22.4	12.8	38.9	36.1	5.2	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
	SFOD [41]	20.1	51.5	26.8	**23.0**	24.8	64.1	**37.6**	10.3	36.3	20.0	18.7	**13.5**	26.5	49.1	37.1	32.1	**10.1**	17.6	42.6	30.0	29.5
SFDA	Mean-teacher [63]	**22.3**	42.3	23.8	21.7	23.5	60.7	33.2	9.1	24.7	**16.7**	12.2	13.1	26.8	**73.6**	43.9	34.5	9.09	24.3	37.9	42.2	29.1
	IRG (Ours)	20.3	**47.3**	**27.3**	19.7	**30.5**	**54.2**	36.2	**10.3**	**35.1**	20.6	20.2	12.3	**28.7**	53.1	**47.5**	**42.4**	9.09	21.1	**42.3**	**50.3**	**31.5**

methods like Cycle DA [77], MeGA CDA [67] and Unbiased DA [12] by considerable margin. Further in SFDA setting, the proposed method produce state-of-the-art performance by improving around 1.1 mAP compared to SFOD.

4.3. 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 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.

Table 5. Ablation study on FoggyCityscapes.

Method	PL	GDL	GCL	prsn	rider	car	truc	bus	train	mcycle	bcycle	mAP
Source Only	✗	✗	✗	25.8	33.7	35.2	13.0	28.2	9.1	18.7	31.4	24.4
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	45.0	**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**

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 conduct

extensive experiments on multiple detection benchmarks to show that the proposed method efficiently adapts a source-trained object detector to the target domain, outperforming the state-of-the-art source-free domain adaptation and many unsupervised domain adaptation methods.

References

[1] Cai, H., Zheng, V.W., Chang, K.C.C.: A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Transactions on Knowledge and Data Engineering **30**(9), 1616–1637 (2018) 5

[2] Cai, Q., Pan, Y., Ngo, C.W., Tian, X., Duan, L., Yao, T.: Exploring object relation in mean teacher for cross-domain detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11457–11466 (2019) 3, 6

[3] Chen, C., Zheng, Z., Ding, X., Huang, Y., Dou, Q.: Harmonizing transferability and discriminability for adapting object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8869–8878 (2020) 7

[4] Chen, C., Zheng, Z., Huang, Y., Ding, X., Yu, Y.: I3net: Implicit instance-invariant network for adapting one-stage object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12576–12585 (2021) 7

[5] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 2, 3, 4

[6] Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029 (2020) 2, 4

[7] Chen, Y.H., Chen, W.Y., Chen, Y.T., Tsai, B.C., Frank Wang, Y.C., Sun, M.: No more discrimination: Cross city adaptation of road scene segmenters. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1992–2001 (2017) 1

[8] Chen, Y., Li, W., Sakaridis, C., Dai, D., Gool, L.V.: Domain adaptive faster r-cnn for object detection in the wild. 2018 IEEE Conference on Computer Vision and Pattern Recognition pp. 3339–3348 (2018) 2, 6, 7, 8

[9] Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3339–3348 (2018) 1, 3

[10] Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 539–546. IEEE (2005) 2

[11] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016) 1, 2, 6, 7

[12] Deng, J., Li, W., Chen, Y., Duan, L.: Unbiased mean teacher for cross-domain object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4091–4101 (2021) 2, 4, 6, 7, 8

[13] Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6569–6578 (2019) 1

[14] Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010) 1, 7

[15] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The Journal of Machine Learning Research **17**(1), 2096–2030 (2016) 1, 7, 8

[16] Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013) 1, 7

[17] Gori, M., Monfardini, G., Scarselli, F.: A new model for learning in graph domains. In: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. vol. 2, pp. 729–734. IEEE (2005) 3

[18] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020) 2, 3

[19] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 6

[20] He, Z., Zhang, L.: Multi-adversarial faster-rcnn for unrestricted object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6668–6677 (2019) 1, 3, 6, 7

[21] He, Z., Zhang, L.: Domain adaptive object detection via asymmetric tri-way faster-rcnn. In: Proceedings of the European Conference on Computer Vision (2020) 7

[22] Hegde, D., Patel, V.: Attentive prototypes for source-free unsupervised domain adaptive 3d object detection. arXiv preprint arXiv:2111.15656 (2021) 3

[23] Hegde, D., Sindagi, V., Kilic, V., Cooper, A.B., Foster, M., Patel, V.: Uncertainty-aware mean teacher for source-free unsupervised domain adaptive 3d object detection. arXiv preprint arXiv:2109.14651 (2021) 3

[24] Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: International Conference on Machine Learning. pp. 1989–1998 (2018) 1

[25] Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016) 1, 3

[26] Hsu, C.C., Tsai, Y.H., Lin, Y.Y., Yang, M.H.: Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In: European Conference on Computer Vision. pp. 733–748. Springer (2020) 3

[27] Hsu, H.K., Yao, C.H., Tsai, Y.H., Hung, W.C., Tseng, H.Y., Singh, M., Yang, M.H.: Progressive domain adaptation for object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 749–757 (2020) 6

[28] Huang, J., Guan, D., Xiao, A., Lu, S.: Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. arXiv preprint arXiv:2110.03374 (2021) 1, 2, 3, 6, 7

[29] Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-domain weakly-supervised object detection through progressive domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5001–5009 (2018) 1, 7, 8

[30] Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S.N., Rosaen, K., Vasudevan, R.: Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? arXiv preprint arXiv:1610.01983 (2016) 7

[31] Khodabandeh, M., Vahdat, A., Ranjbar, M., Macready, W.G.: A robust learning approach to domain adaptive object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 480–490 (2019) 3, 6, 7

[32] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. arXiv preprint arXiv:2004.11362 (2020) 2, 3, 13

[33] Kim, S., Choi, J., Kim, T., Kim, C.: Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6092–6101 (2019) 7

[34] Kim, T., Jeong, M., Kim, S., Choi, S., Kim, C.: Diversify and match: A domain adaptive representation learning paradigm for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12456–12465 (2019) 3, 6

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

[36] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016) 3

[37] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012) 6

[38] Kundu, J.N., Venkat, N., Babu, R.V., et al.: Universal source-free domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4544–4553 (2020) 1

[39] Li, R., Jiao, Q., Cao, W., Wong, H.S., Wu, S.: Model adaptation: Unsupervised domain adaptation without source data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9641–9650 (2020) 3

[40] Li, S., Ye, M., Zhu, X., Zhou, L., Xiong, L.: Source-free object detection by learning to overlook domain style. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8014–8023 (2022) 6

[41] 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

[42] Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International Conference on Machine Learning. pp. 6028–6039. PMLR (2020) 1, 3

[43] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 1

[44] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 1

[45] Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: A survey. International journal of computer vision **128**(2), 261–318 (2020) 1

[46] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016) 1

[47] Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. arXiv preprint arXiv:2102.09480 (2021) 2, 4

[48] Liu, Y., Zhang, W., Wang, J.: Source-free domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1215–1224 (2021) 1, 3

[49] Lo, S.Y., Oza, P., Chennupati, S., Galindo, A., Patel, V.M.: Spatio-temporal pixel-level contrastive learning-based source-free domain adaptation for video semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 3

[50] Lo, S.Y., Wang, W., Thomas, J., Zheng, J., Patel, V.M., Kuo, C.H.: Learning feature decomposition for domain adaptive monocular depth estimation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2022) 1

[51] Morris, C., Ritzert, M., Fey, M., Hamilton, W.L., Lenssen, J.E., Rattan, G., Grohe, M.: Weisfeiler and leman go neural: Higher-order graph neural networks. In: Proceedings of

the AAAI Conference on Artificial Intelligence. vol. 33, pp. 4602–4609 (2019) 3

[52] Nair, N.G., Patel, V.M.: Confidence guided network for atmospheric turbulence mitigation. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 1359–1363. IEEE (2021) 3

[53] Nguyen, K., Tripathi, S., Du, B., Guha, T., Nguyen, T.Q.: In defense of scene graphs for image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1407–1416 (2021) 3

[54] Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 3

[55] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016) 1

[56] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015) 2, 6

[57] RoyChowdhury, A., Chakrabarty, P., Singh, A., Jin, S., Jiang, H., Cao, L., Learned-Miller, E.: Automatic adaptation of object detectors to new domains using self-training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 780–790 (2019) 3

[58] Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6956–6965 (2019) 1, 3, 6, 7, 8

[59] Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3723–3732 (2018) 1

[60] Sakaridis, C., Dai, D., Gool, L.V.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision **126**, 973–992 (2018) 2, 6

[61] Sindagi, V.A., nad R. Yasarla, P.O., Patel, V.M.: Prior-based domain adaptive object detection for hazy and rainy conditions. In: European Conference on Computer Vision (ECCV) (2020) 1, 3

[62] Su, P., Wang, K., Zeng, X., Tang, S., Chen, D., Qiu, D., Wang, X.: Adapting object detectors with conditional domain normalization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 403–419. Springer (2020) 6

[63] Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780 (2017) 4, 6, 7, 8, 13

[64] Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7167–7176 (2017) 1, 3

[65] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 6

[66] VS, V., Oza, P., Patel, V.M.: Towards online domain adaptive object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 478–488 (2023) 3

[67] VS, V., Oza, P., Sindagi, V.A., Gupta, V., Patel, V.M.: Megacda: Memory guided attention for category-aware unsupervised domain adaptive object detection (2021) 6, 7, 8

[68] Vs, V., Poster, D., You, S., Hu, S., Patel, V.M.: Meta-uda: Unsupervised domain adaptive thermal object detection using meta-learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1412–1423 (2022) 3

[69] Wu, A., Han, Y., Zhu, L., Yang, Y.: Instance-invariant domain adaptive object detection via progressive disentanglement. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 3, 6

[70] Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., Weinberger, K.: Simplifying graph convolutional networks. In: International conference on machine learning. pp. 6861–6871. PMLR (2019) 5

[71] Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018) 2

[72] Xia, H., Zhao, H., Ding, Z.: Adaptive adversarial network for source-free domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9010–9019 (2021) 1

[73] Xu, C.D., Zhao, X.R., Jin, X., Wei, X.S.: Exploring categorical regularization for domain adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11724–11733 (2020) 6, 8

[74] Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph r-cnn for scene graph generation. In: Proceedings of the European conference on computer vision (ECCV). pp. 670–685 (2018) 3

[75] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020) 1

[76] Zhang, Z., Cui, P., Zhu, W.: Deep learning on graphs: A survey. IEEE Transactions on Knowledge and Data Engineering (2020) 5

[77] Zhao, G., Li, G., Xu, R., Lin, L.: Collaborative training between region proposal localization and classification for domain adaptive object detection. In: European Conference on Computer Vision. pp. 86–102. Springer (2020) 6, 7, 8

[78] Zhong, Y., Wang, L., Chen, J., Yu, D., Li, Y.: Comprehensive image captioning via scene graph decomposition. In: European Conference on Computer Vision. pp. 211–229. Springer (2020) 3

[79] Zhu, X., Pang, J., Yang, C., Shi, J., Lin, D.: Adapting object detectors via selective cross-domain alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 687–696 (2019) 7

[80] Zhuang, C., Han, X., Huang, W., Scott, M.: ifan: Image-instance full alignment networks for adaptive object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13122–13129 (2020) 6

[81] Zou, Z., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: A survey. arXiv preprint arXiv:1905.05055 (2019) 1

Supplementary material: Instance Relation Graph Guided Source-Free Domain Adaptive Object Detection

Ablation analysis for different loss functions: In Table 6, we present more ablation analysis for Cityscapes→FoggyCityscapes experiment. In Table 6, the first experiment is standard mean-teacher [63] framework for SFDA setting. The second experiment is performed to understand the effect between supervised and our proposed contrastive loss. The supervised contrastive loss [32] is computed between the proposals generated from the student and teacher network. The class information for each proposal is obtained after passing through the RoI feature extractor and classification head. Using the proposal features and class information between student and teacher networks, we compute the SupCon loss. Therefore compared to mean-teacher training, supervised contrastive learning on mean-teacher increase the performance by 1.9 mAP. Without supervised contrastive learning, distillation loss between student and teacher networks degrades the performance from 36.2 mAP to 35.9 mAP. In final, utilizing graph-guided contrastive learning on the proposal features on top of GDL further helps the model learn high-quality representations, increasing performance by 1.9 mAP on the target domain. Note all augmentations are applied on the image level. Also, strong and weak augmentation essentially simulates the domain gap between student and teacher. Exploiting this property and mean-teacher framework, the consistency loss between strongly augmented student prediction and weakly augmented teacher predictions enforces the student network to learn a more robust and domain-invariant feature representation. Strong augmentation: color jitter, grayscale, Gaussian blur, erasing. Weak augmentation: horizontal flip.

Table 6. Ablation study for different loss functions.

Method	PL	SimCLR	GDL	GCL	prsn	rider	car	truc	bus	train	mcycle	bcycle	mAP
Source Only	✗	✗	✗	✗	25.8	33.7	35.2	13.0	28.2	9.1	18.7	31.4	24.4
MT + SW	✓	✗	✗	✗	33.9	43.0	45.0	**29.1**	37.2	25.1	25.5	38.2	34.3
MT + SW	✓	✓	✗	✗	36.1	**45.8**	47.2	28.5	36.6	29.5	27.2	38.8	36.2
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**

Ablations analysis on IRG network: In Table 7, in the second row, we use Kmeans clustering algorithm to find positive proposal pairs instead of IRG network. Specifically, we utilize kmeans algorithm to find positive proposals and then apply CRL loss. From Table 7, we can infer that using learnable IRG to model positive relations is more effective than kmeans and improves the performance by considerable margin. In a next row, we freeze the IRG network and use the original edge weights to compute CRL loss. In frozen IRG experiment, when we use original edge weights to compute GCL loss the performance drops to 36.2 mAP compared to learnable IRG network.

Table 7. Ablation study on IRG network

Method	CRL	prsn	rider	car	truc	bus	train	mcycle	bcycle	mAP
Source Only	✗	25.8	33.7	35.2	13.0	28.2	9.1	18.7	31.4	24.4
Kmeans	✓	36.3	44.5	49.7	26.2	37.9	26.0	32.8	39.3	36.5
IRG(Frozen)	✓	33.9	43.7	47.3	26.8	38.5	27.1	30.2	38.9	36.2
Ours	✓	37.4	45.2	51.9	24.4	39.6	25.2	31.5	41.6	37.1

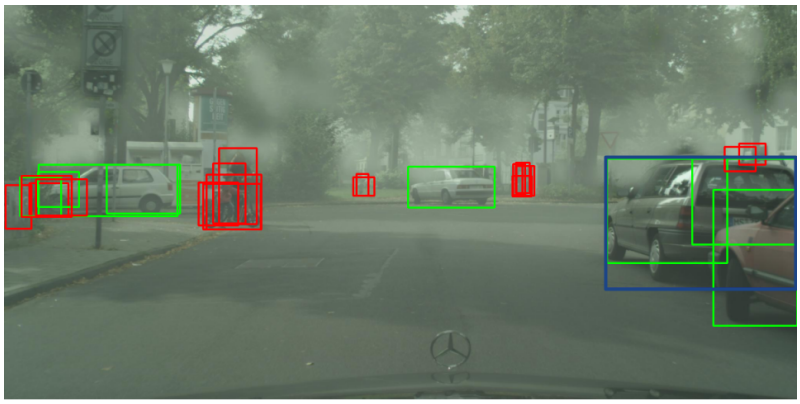


Figure 7. Blue: Query proposal, Green: positive proposals, Red: Negative proposals

Positive proposals visualization: From Fig 7, given a anchor (blue), we visualized it's positive (green) and negative (red) proposals generated by the IRG network.

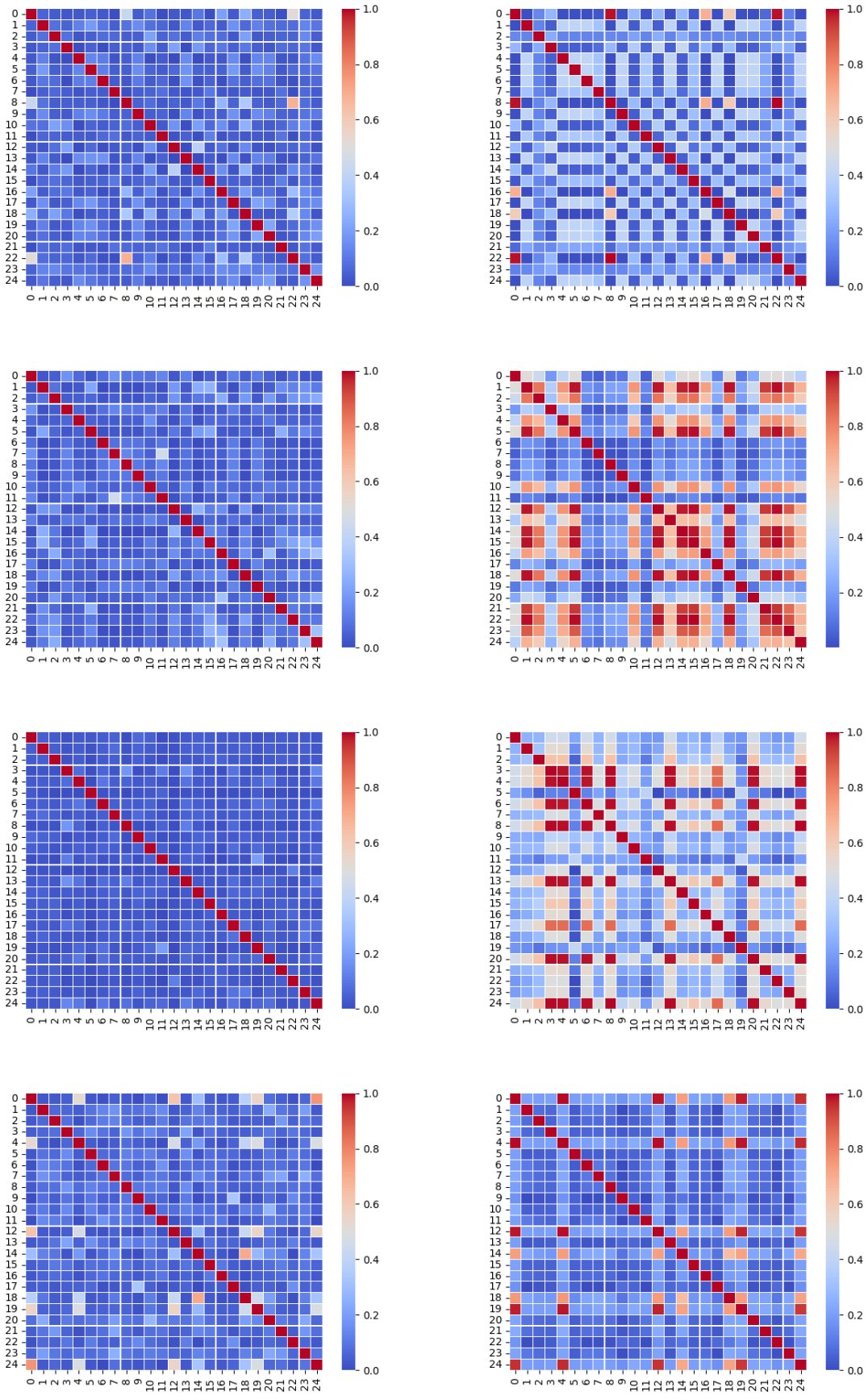
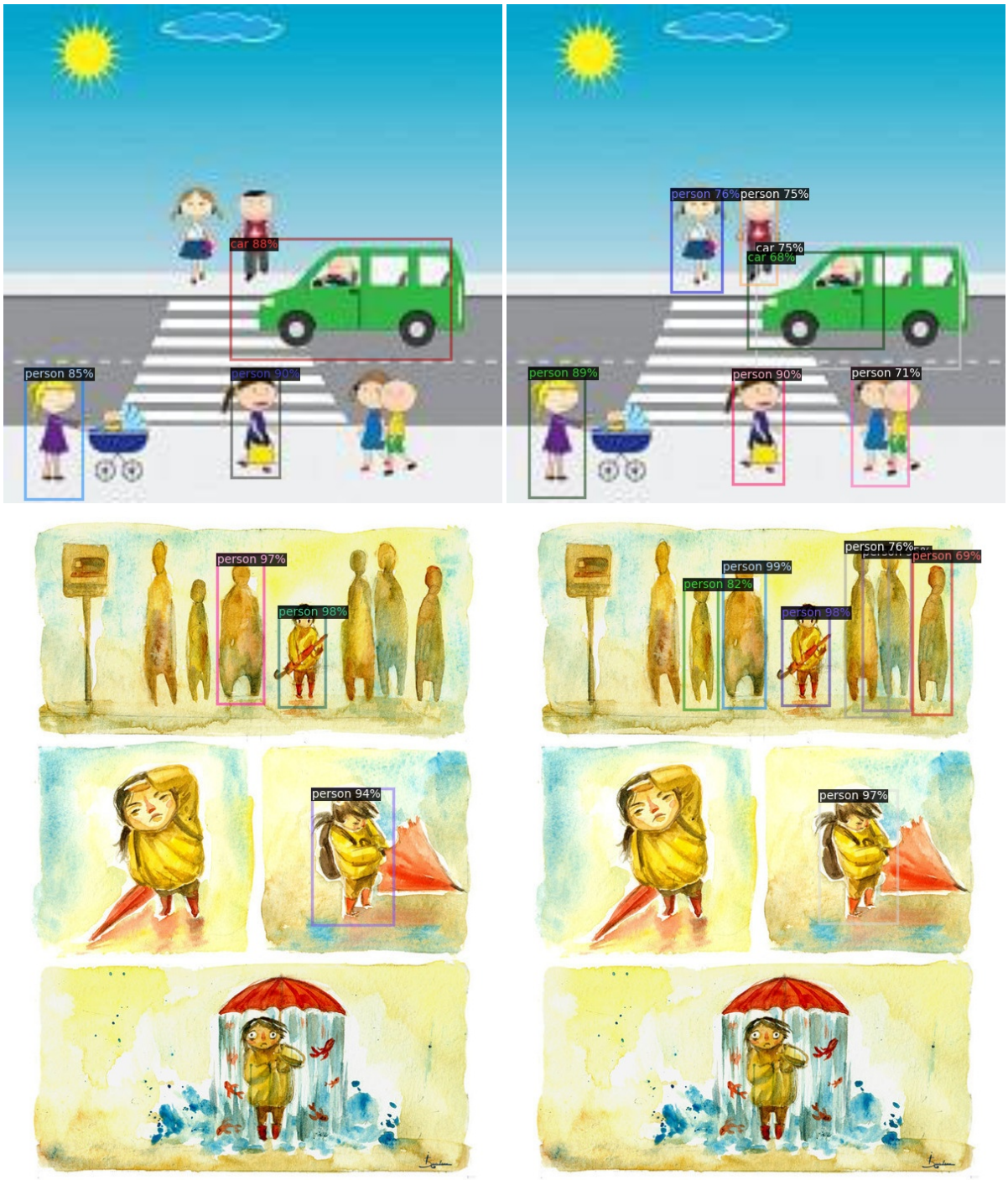
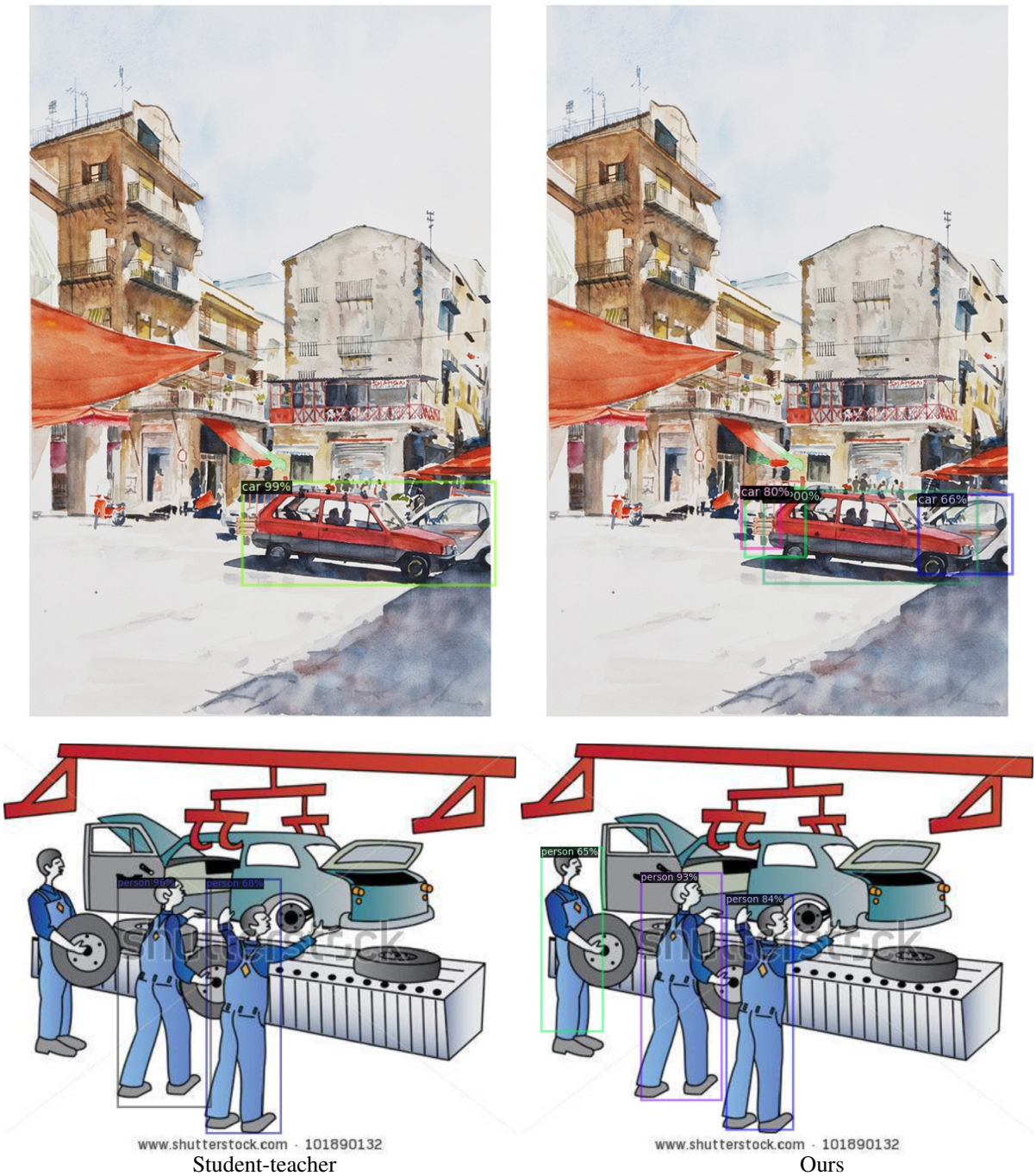


Figure 8. More visualizations of relation matrix of RoI features before and after passing through IRG network.



Student-teacher Ours

Figure 9. More detection visualization for realistic to artistic adaptation. From the above visualization, we can infer that our model efficiently tackles classes' negative transfer and constructs high confidence prediction boxes.



Student-teacher Ours

Figure 10. More detection visualization for realistic to artistic adaptation. From the above visualization, we can infer that RPN fails to generate proposals due to tiny/occluded objects or heavy domain shifts, those instances cannot take full benefits of the proposed contrastive learning strategy.