Densely Semantic Enhancement for Domain Adaptive Region-free Detectors

Bo Zhang, Tao Chen, Bin Wang, Xiaofeng Wu, Liming Zhang, and Jiayuan Fan

Research Center of Smart Networks and Systems, School of Information Science and Technology, Fudan University

Abstract

Unsupervised domain adaptive object detection aims to adapt a well-trained detector from its original source domain with rich labeled data to a new target domain with unlabeled data. Previous works focus on improving the domain adaptability of region-based detectors, e.g., Faster-RCNN, through matching cross-domain instance-level features that are explicitly extracted from a region proposal network (RPN). However, this is unsuitable for region-free detectors such as single shot detector (SSD), which perform a dense prediction from all possible locations in an image and do not have the RPN to encode such instance-level features. As a result, they fail to align important image regions and crucial instance-level features between the domains of region-free detectors.

Problem Formulation

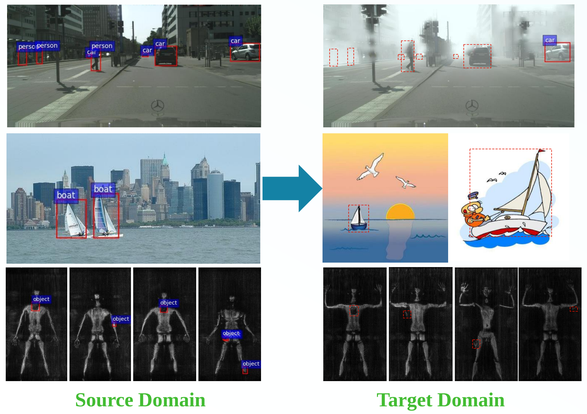
Unsupervised domain adaptation for object detection:

1) All instances with rich annotations from the source domain: $\{X^s, Y^s, B^s\}$
2) Some unlabeled instances from the target domain: $\{X^t\}$
3) Adapt the detector that is well-trained on the source domain to the unlabeled target domain.

Challenges

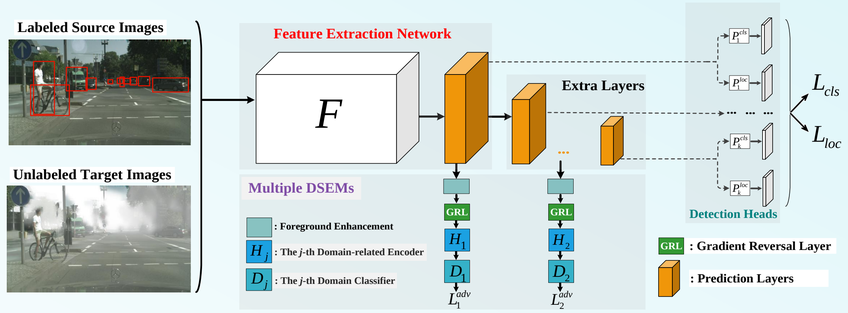
When the previous domain adaptive researches, originally explored to meet the requirements of region-based detector architecture, are applied to region-free detectors such as SSD, there are still two major **challenges** as follows.

1) Unlike region-based detectors employing the RPN to produce proposals from an input image, region-free detectors first produce spatially-dense feature representations without the encoding process of instance-specific features. However, these dense representations often contain extensive background information. As a result, aligning the dense representations between domains becomes difficult, due to large variations in background appearance and scene layout.
2) Most region-free detectors recognize objects with different scales using multi-layer features. Besides, the input image contains complex multi-instance information. For these reasons, the cross-domain adaptation should fully consider the matching of features from different semantic levels and multiple spatial instances.

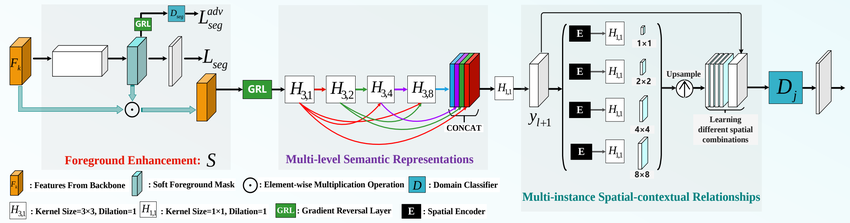


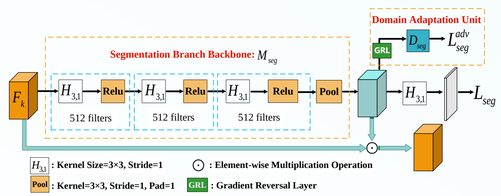
Methodology

Overview: the developed domain adaptive region-free detector, which is mainly composed of a feature extraction network F, multiple detection heads P and multiple DSEMs. Here, SSD is selected as the baseline detector. Our implementation aligns the Conv4_3 and the Conv6_2 layers of SSD. Red bounding boxes in the source images represent the object annotations.



The illustration of the DSEM. Note that the domain discrepancy in the foreground enhancement can also be reduced via an adversarial loss L_{seg}^{adv}. Our implementation sets the layer number to 3. Gradient Reversal Layer (GRL) is employed to simply implement adversarial training process.





The illustration of the semantic segmentation branch in DSEM. The purpose of this branch is to produce a domain invariant mask to enhance the foreground region of the features extracted by the region-free detector such as SSD.

Results and Discussions

Table 1. Adaptation from PASCAL VOC

Method		G	L	DN-8	DN-32	P-A	mAP
Region-free	SSD [13]						27.6
	L-SSD		√				29.5
	G-SSD	√					29.5
	G-L-SSD	√	√				31.5
	LW-SSD	√	√				34.3
	Proposed SSD			√			38.1
				√	√		40.1
				√	√	√	**42.2**
	RDet [15]						26.0
	Proposed RDet			√			33.1
				√	√		39.8
				√	√	√	**43.5**
	SSD* [13]						34.3
	Proposed SSD*			√	√		**44.3**
	RDet* [15]						35.6
	Proposed RDet*			√	√		48.9
	WST+BSR [37]						**35.7**

Two important findings can be observed:

1) Applying **some Faster RCNN based domain adaptation modules** to the SSD detector only achieves very limited performance improvement. However, by employing the DSEM to align the conv4_3 layer **(DN-8)**, the SSD+DSEM compares favorably against G-L-SSD and LW-SSD. Further, aligning both the conv4_3 and conv6_2 (DN-32) layers increases the results from **38.1% to 40.1%.**
2) Large-scale data are indispensable for improving DNNs based detectors. We studied the impact of large-scale data on domain adaptive detectors. SSD* and RDet* refer to results that COCO model replaces ImageNet-pre-trained model to initialize the network weights. It can be seen that employing COCO model has significantly increased the mAP by 6.7~9.6%. In other words, the pre-trained model on large-scale data can significantly relieve the domain discrepancy.

Table 2. Adaptation from PASCAL VOC to Comic2k

Method	DN-8	DN-32	P-A	Target Domain							Source Domain
				bicycle	bird	car	cat	dog	prsn	mAP	mAP
SSD [14]				21.7	12.8	**34.4**	11.0	14.6	44.4	23.1	81.4
SSD+DSEMs (ours)	√			39.7	15.2	22.6	14.9	25.9	50.3	28.1	81.1
	√	√		49.6	18.2	26.6	**28.8**	30.8	46.3	33.4	80.1
	√	√	√	**57.8**	**22.2**	32.2	28.5	**32.9**	**56.8**	**38.4**	79.5
ADDA [32]				39.5	9.8	17.2	12.7	20.4	43.3	23.8	\
DD+MRL [45]				\	\	\	\	\	\	34.5	\
WST+BSR [46]				50.6	13.6	31.0	7.5	16.4	41.4	26.8	\
DT [47]				43.6	13.6	30.2	16.0	26.9	48.3	29.8	\

Table 3. Faster RCNN+DSEM

Method	Backbone	mAP
Faster RCNN [11]	VGG16	23.5
G-L-Faster [51]	VGG16	31.5
Faster RCNN+DSEM (ours)	VGG16	**33.1**

To comprehensively investigate the effectiveness of DSEM on the region-based detectors, we conduct the experiments of applying the DSEM to Faster RCNN in Table 3.

We give a visual explanation via Grad-CAM, illustrating which features regions are fed into the domain classifier to perform the alignment process. The domain-related evidence represented by our DSEM is visualized in the right Fig. The visual results show that, compared with typical domain adaptation modules, DSEM can capture more important regions and the encoded features are semantically rich. Thus, the backbone detection network can focus on multiple instances to deceive the domain classifier.