Cross-domain Detection via Graph-induced Prototype Alignment
Supplementary Material

Minghao Xu[1,2] Hang Wang[1,2] Bingbing Ni[1,2,3*] Qi Tian[4] Wenjun Zhang[1]
[1]Shanghai Jiao Tong University, Shanghai 200240, China
[2]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[3]Huawei Hisilicon [4]Huawei Noah's Ark Lab
{xuminghao118, wang–hang, nibingbing, zhangwenjun}@sjtu.edu.cn
nibingbing@hisilicon.com tian.qi1@huawei.com

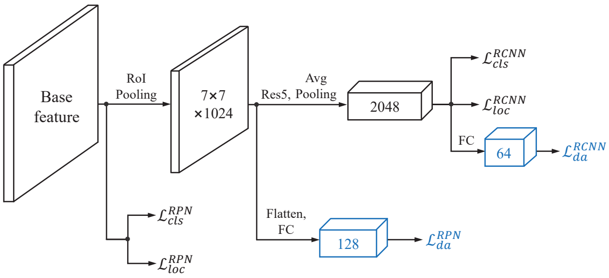

Base feature
RoI Pooling
7×7 ×1024
Res5, Avg Pooling
2048
$\mathcal{L}_{cls}^{RCNN}$
$\mathcal{L}_{loc}^{RCNN}$
FC
64
$\mathcal{L}_{da}^{RCNN}$
$\mathcal{L}_{cls}^{RPN}$
$\mathcal{L}_{loc}^{RPN}$
Flatten, FC
128
$\mathcal{L}_{da}^{RPN}$

Figure 1. The architecture of *head* network. It is built on the basis of Faster R-CNN with ResNet-50 backbone.

1. Network Architecture

In this work, we instantiate the proposed Graph-induced Prototype Alignment (GPA) framework using Faster R-CNN detector with ResNet-50 backbone. For clarity, we split the whole network architecture into two parts: (1) the *backbone* network for feature extraction over entire images, and (2) the *head* network for bounding box recognition (classification and regression) and domain adaptation learning, which is presented in Figure 1.

The whole framework is composed of two stages, Region Proposal Network (RPN) and Region-based CNN (RCNN). For RPN, by utilizing the base feature extracted with ResNet-50 backbone, the classification and localization losses, $\mathcal{L}_{cls}^{RPN}$ and $\mathcal{L}_{loc}^{RPN}$, are defined, and the $7 \times 7 \times 1024$ feature map of each region proposal is generated through RoI pooling. After flattening the feature map, a fully-connected layer outputs the 128-dimensional feature vector which derives foreground and background prototypes, and the domain alignment loss $\mathcal{L}_{da}^{RPN}$ is calculated with these prototypes. For RCNN, a 2048-dimensional feature vector is generated via average pooling, and the classification and localization losses, $\mathcal{L}_{cls}^{RCNN}$ and $\mathcal{L}_{loc}^{RCNN}$, are defined on such basis. By using another fully-connected layer, the 64-dimensional feature vector is produced to derive prototypes of each category, and, based on these prototypes, the domain alignment loss $\mathcal{L}_{da}^{RCNN}$ is calculated.

2. Qualitative Detection Results

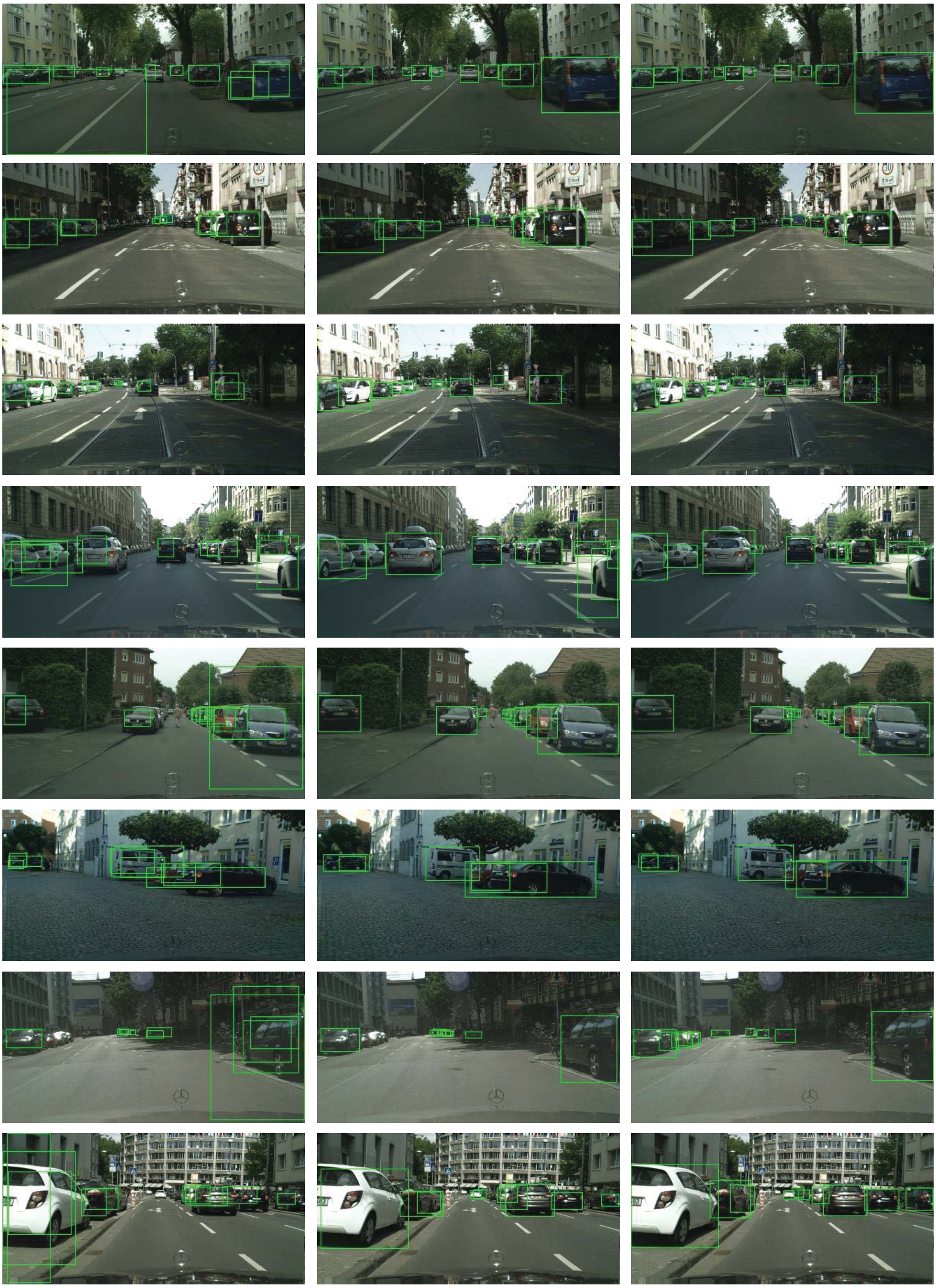
In Figure 2, we present more detection results on the task SIM 10k → Cityscapes, and this task aims for vehicle detection. As shown in the figure, the Source-only model produces many bounding boxes greatly biasing from objects, since the generated features are not discriminative enough. DA [1] localizes objects more precisely, but some false positives are produced by this method, *e.g.* the second figure of the fifth and sixth rows. In the results of our approach, these false positives are effectively alleviated, and our model can accurately localize those objects with small scale and in severe occlusion, *e.g.* the third figure of the first row.

Figure 3 displays several groups of detection results on the task Cityscapes → Foggy Cityscapes. On this task, eight common categories of two datasets are used for evaluation. In the results of Source-only and DA [1], quite a few bounding boxes are assigned with false labels, and several objects are undetected. For example, in the last row, a bus is misclassified as car by the DA model. Our approach correctly detects most of the objects and predicts bounding boxes more accurately. Just as shown in the fifth row, a train is undetected using Source-only and DA model, while it is precisely localized by our method.

References

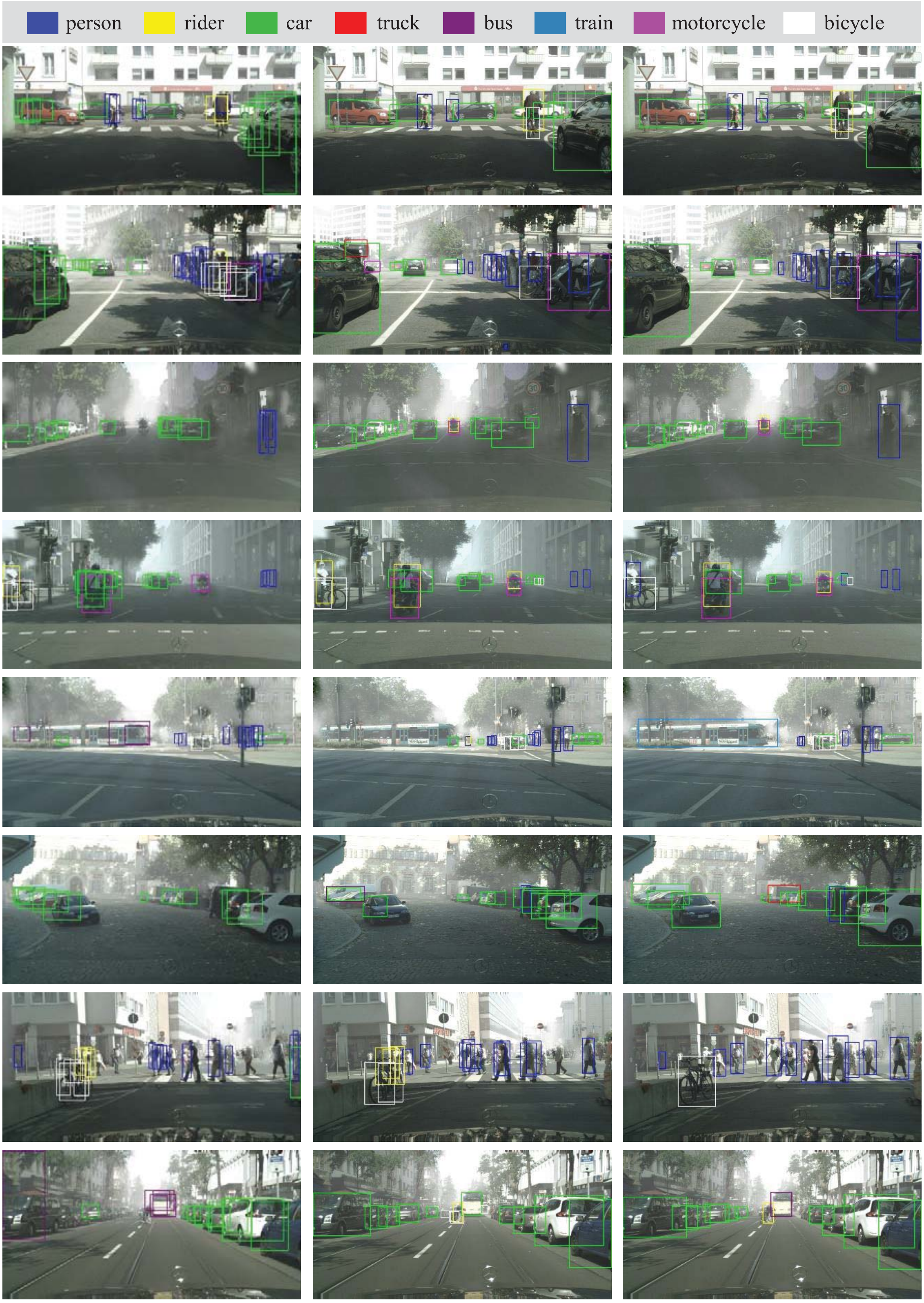
[1] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster R-CNN for object detection in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

*The corresponding author is Bingbing Ni.



(a) Source-only (b) DA (c) GPA (Two-stage Alignment)

Figure 2. The detection results on the task SIM 10k → Cityscapes, in which Source-only, DA [1] and our method are evaluated.



(a) Source-only (b) DA (c) GPA (Two-stage Alignment)

Figure 3. The detection results on the task Cityscapes → Foggy Cityscapes, in which Source-only, DA [1] and our method are evaluated.