CaT: Weakly Supervised Object Detection with Category Transfer Supplementary Material

Tianyue Cao[1] Lianyu Du[1] Xiaoyun Zhang[1*] Siheng Chen[1,2] Ya Zhang[1,2] Yan-Feng Wang[1,2]
Cooperative Medianet Innovation Center, Shanghai Jiao Tong University[1] Shanghai AI Laboratory[2]
{vanessa_, dulianyu, xiaoyun.zhang, sihengc, ya_zhang, wangyanfeng}@sjtu.edu.cn

1. Qualitative Results

Figure 1 shows some qualitative results. Figure (c) shows the detection results of our method with 5 overlapping classes between COCO-65 and Pascal VOC. It gets higher quality box boundaries than WSDDN [1]. Our method shows advantages in terms of: (1) *part domination* (column 1-3): WSDDN tends to focus on the discriminative parts and the predicted boxes may only contain parts of the objects. Our method can alleviate the problem by transferring bounding box knowledge from the fully-supervised dataset. (2) *object missing* (column 4-5): WSDDN may miss some objects under complex or dark environments. We use the semantic graph to leverage the correlations between categories to improve the detection ability.

2. Visualization of the Semantic Graph

Intra-dataset graph. Figure 3 (a), (b), and (d) shows the category co-occurrence matrices of Pascal VOC, KITTI, and COCO datasets. We introduce a threshold to build the binary intra-dataset graphs. For example, (c) shows the intra-dataset graph of KITTI dataset using 0.4 as threshold.

Inter-dataset graph. Figure 2 (a) and (b) shows the semantic similarity matrices between the categories of Pascal VOC and COCO datasets, KITTI and COCO datasets, respectively. The elements are the cosine similarities between the word2vec vectors of two categories. For example, in (a), the similarity between *'car'* and *'bus'* is high (0.6), and the similarity between *'car'* and *'cake'* is low (0.0). We normalize the similarity matrices to obtain inter-dataset graphs.

3. Ablation Study

Table 1 shows the performance using different graph convolution layers. The overlapping classes between the fully-supervised and weakly-supervised datasets are set to be 5 and fixed. Our method achieves the best performance using two graph convolution layers. Too much layers may lead to oversmoothing problem.

*Xiaoyun Zhang is the corresponding author.

number of GCN layer	mAP(%)	CorLoc(%)
1	61.6	79.1
2	63.0	80.0
3	62.4	79.2
4	58.8	76.3

Table 1. mAP(%) and CorLoc(%) using different graph convolution layers. Our method achieves the best performance using two graph convolution layers.

τ^w_{rpn}	mAP(%)	CorLoc(%)
0.3	58.6	75.6
0.4	63.0	80.0
0.5	61.4	77.0
0.7	56.9	75.2

Table 2. mAP(%) and CorLoc(%) using different RPN threshold τ^w_{rpn} during training for the weakly-supervised dataset. Our method achieves the best performance when $\tau^w_{\text{rpn}} = 0.4$.

Table 2 shows the performance using different region proposal network (RPN) training threshold τ^w_{rpn} for the weakly-supervised dataset. Our model achieves the best performance with $\tau^w_{\text{rpn}} = 0.4$. With too small τ^w_{rpn}, the RPN extracts too much proposals, which leads to inaccurate locations of regions of interests and high difficulty of training the box offset. With too large τ^w_{rpn}, the RPN misses some positive proposals since it is trained without the non-overlapping categories in the weakly-supervised dataset.

4. Compared with Faster R-CNN trained on COCO

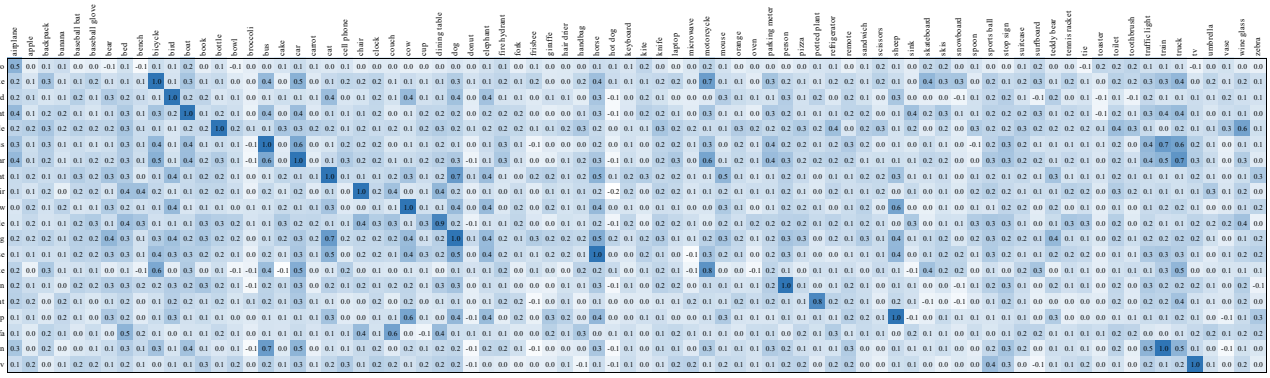
To validate the domain adaptation ability of our method, we compare our method with a Faster R-CNN [2] trained only on COCO dataset and testing on Pascal VOC dataset. The Faster R-CNN trained on COCO achieves 57.7% mAP, while our method achieves 66.0% mAP using the same fully-supervised dataset as source. Our method can alleviate the domain gap problem by using the double-supervision mean teacher network.

(a) ground truth boxes

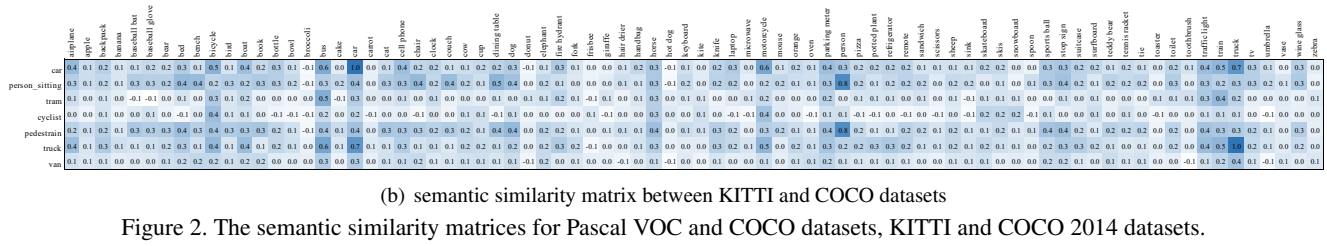
(b) detection results of WSDDN

(c) detection results of our method

Figure 1. Some qualitative results on Pascal VOC test set. (a) - (c) shows the ground truth boxes, the detection results of WSDDN [1], and the detection results of our method with 5 overlapping classes between COCO-65 and Pascal VOC. Our method alleviate the part domination problem (column 1-3) and the object missing problem (column 4,5).



(a) semantic similarity matrix between Pascal VOC and COCO datasets



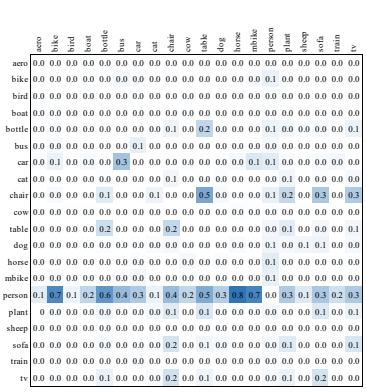
(b) semantic similarity matrix between KITTI and COCO datasets

Figure 2. The semantic similarity matrices for Pascal VOC and COCO datasets, KITTI and COCO 2014 datasets.

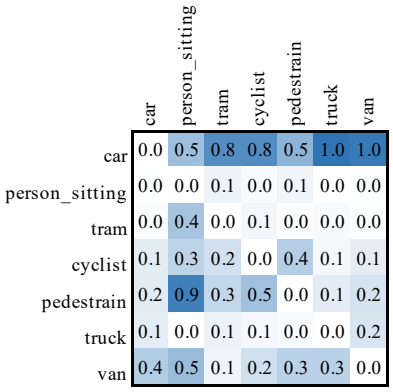
References

[1] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, pages 2846–2854. IEEE Computer Society, 2016.

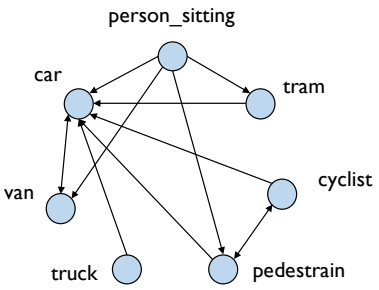
[2] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.



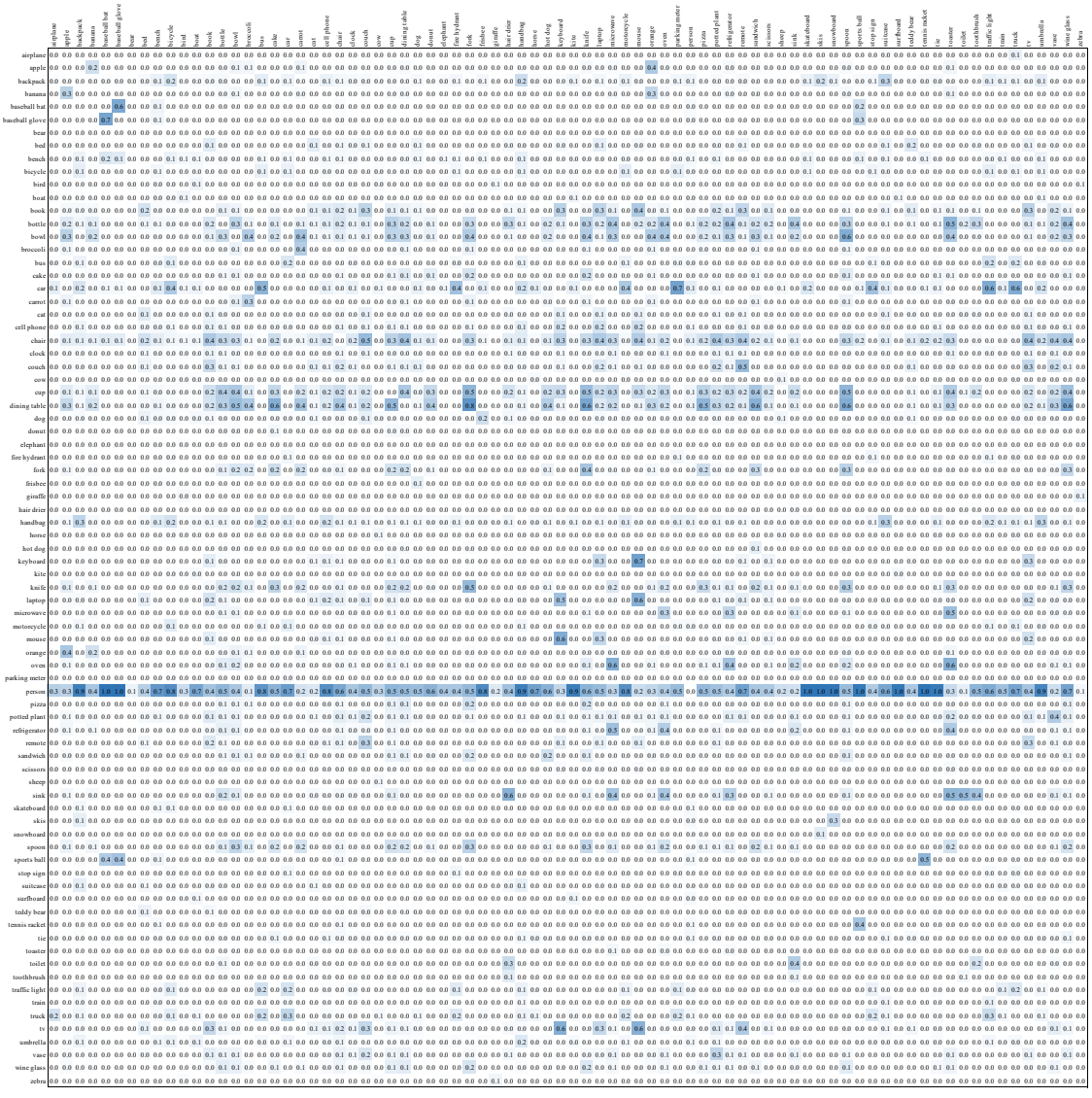
(a) Pascal VOC category co-occurrence matrix



(b) KITTI category co-occurrence matrix



(c) KITTI intra-dataset graph



(d) COCO category co-occurrence matrix

Figure 3. The category co-occurrence matrices for Pascal VOC, KITTI, and COCO datasets.