Supplementary - Seeking Similarities over Differences: Similarity-based Domain Alignment for Adaptive Object Detection

Farzaneh Rezaeianaran[1] Rakshith Shetty[1] Rahaf Aljundi[2]
Daniel Olmeda Reino[2] Shanshan Zhang[3] Bernt Schiele[1]
[1]Max Planck Institute for Informatics, Saarland Informatics Campus [2]Toyota Motor Europe [3]Nanjing University of Science and Technology

Abstract

In this supplementary material we present qualitative results and additional experiments to provide more details on our approach. In the first section 1, we provide qualitative results on different domain shift scenarios and then in the next section (Sec. 2) we study the performance of our framework when a fixed number of clusters is used for the grouping stage instead of using an adaptive number of clusters. In addition, we study another aspect of our grouping mechanism and show that our model improves the instance alignment by providing a trade-off between foreground and background groups during the training.

1. Qualitative Results

Figures 1, 2 and 3 show qualitative results on three different domain shift scenarios. As shown in Figure 1, both our ViSGA models recover more objects compared to Faster RCNN in the first row. In addition, as we can see in the second column, similarity-based grouping produces fewer false positives compared to the spatial-based model. In Figure 2 we provide more results when more classes are available in the dataset. As all three columns show, the ViSGA (cosine) model successfully detects most of the objects compared to ViSGA (IoU) and Faster RCNN. In addition, from the last column we can observe that ViSGA (cosine) performs better on very far objects compared to the IoU model by producing a lower number of false positives. Finally, in figure 3, our ViSGA model clearly performs better in the first set of four images. In the next set and last column ViSGA (similar to Faster RCNN) misses. However, also in this set our model provides better performance in comparison with Faster R-CNN in terms of both recovering objects and producing a lower number of false positives.

2. Additional Analyses

2.1. Comparing performance of adaptive and fixed number of groups

Figure 4 shows experimental results of ViSGA using fixed vs. an adaptive number of clusters on *Sim2Real* and *Foggy* scenarios. The rightmost point in the figure, with 256 clusters, shows the result of a model trained without a group aggregation mechanism and proposals are directly fed into the discriminator. As we can see, for both of the domain shift scenarios, fixing the clusters to 50 gives the best fixed-cluster performance. However, if we adaptively create the clusters during instance alignment we reach higher performance for both cases compared to the previous fixed-cluster models (+0.3% on Sim2Real and +0.8% on Foggy). Therefore, these experimental results provide additional evidence that allowing an adaptive number of clusters boost the detection performance during training.

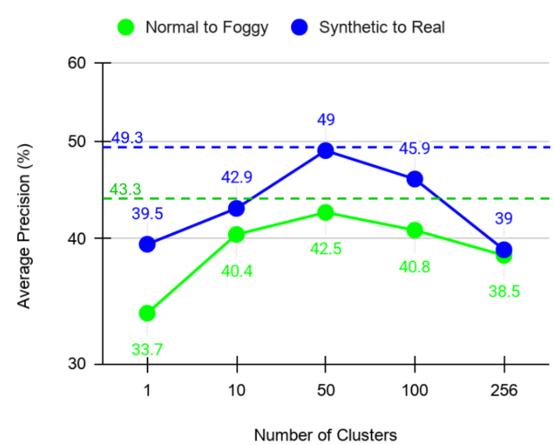

Figure 4. Sensitivity analysis of using fixed number of clusters. Horizontal dashed lines show the results of models trained using adaptive number of clusters. As shown in the figure adaptively creating groups during training brings more improvement compared to fixed-clusters.

Figure 1. Qualitative results. Sim2Real scenario. First row: Faster R-CNN, second row: ViSGA (IoU) and, last row: ViSGA (cosine). True positives, False positives and missed objects are shown as cyan, purple, and red boxes respectively.

2.2. How does ViSGA help the adaptation process?

As we observed in the previous part 2.1, for the Sim2Real scenario, the model with 50 (fixed number) clusters performs best. Now, in this section we study how our grouping mechanism clusters foreground (fg) and background (bg) proposals into groups when performing alignment. We show this by comparing the ratio of fg groups to fg proposals and bg groups to bg proposals. If a group contains a higher number of fg (bg) proposals, we label it as a fg (bg) group. As we can see in figure 5 for 50 clusters, our ViSGA mechanism keeps more fg groups (0.76) than bg groups (0.15) which provides a desirable trade-off between fg and bg groups and leads to the best performance at this point according to figure 4. In addition, we can see that, in the left of this figure (5), fg proposals are grouped heavily which causes a performance drop in figure 4. On the other hand, for higher cluster numbers in this figure (5), bg proposals are not grouped enough and as a result we observe sup-optimal performance of models in the right part of figure 4.

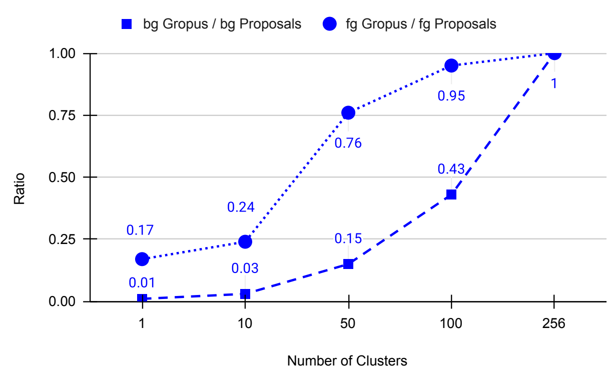

Figure 5. Groups to Proposals ratios analysis on Sim2Real scenario.

t-SNE Visualization: In tSNE visualizations figure 6, we notice an effective domain alignment – especially for background samples. The strong alignment of the source and target background proposals allows the classifier to ignore the specifics of source/target background proposals and fo-

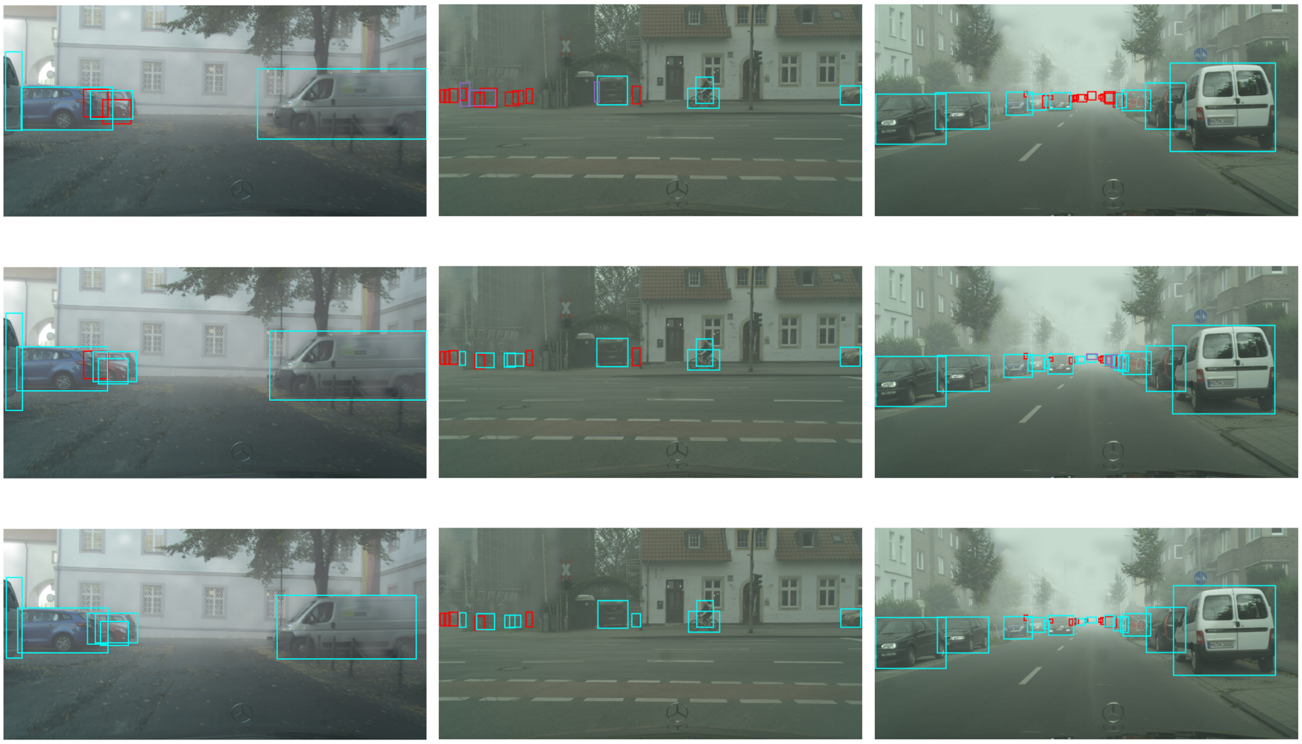

Figure 2. Qualitative results. Foggy scenario. First row: Faster R-CNN, second row: ViSGA (IoU) and, last row: ViSGA (cosine). True positives, False positives and missed objects are shown as cyan, purple, and red boxes respectively.

cus on the foreground class. This is in fact what we showed in figure 5 where we compare the number of grouped proposals for foreground vs. background. The figure shows that a large number of background proposals collapsed in a single group compared to a relaxed grouping of foreground proposals giving them flexibility to maintain their distinctive features. This translates in better results by our method (ViSGA).

Figure 3. Qualitative results. Cross Camera scenario. Odd rows: Faster R-CNN and even rows: ViSGA (cosine). True positives, False positives and missed objects are shown as cyan, purple, and red boxes respectively.

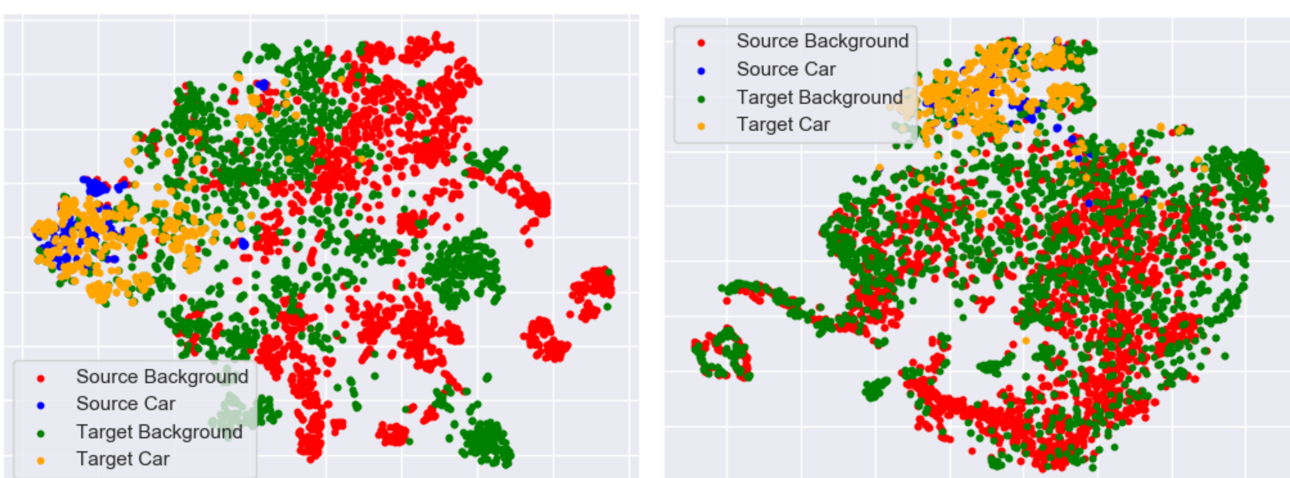

Figure 6. tSNE visualization of feature embeddings. Left: Source only model. Right: ViSGA.