Vector-Decomposed Disentanglement for Domain-Invariant Object Detection: Supplementary Material

Aming Wu[1,2] Rui Liu[1,2] Yahong Han[1,2,3] Linchao Zhu[4] Yi Yang[4]
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Tianjin Key Lab of Machine Learning, Tianjin University, Tianjin, China
[3]Peng Cheng Laboratory, Shenzhen, China [4]ReLER Lab, AAII, University of Technology Sydney
{tjwam, ruiliu, yahong}@tju.edu.cn, {Linchao.Zhu, yi.yang}@uts.edu.au

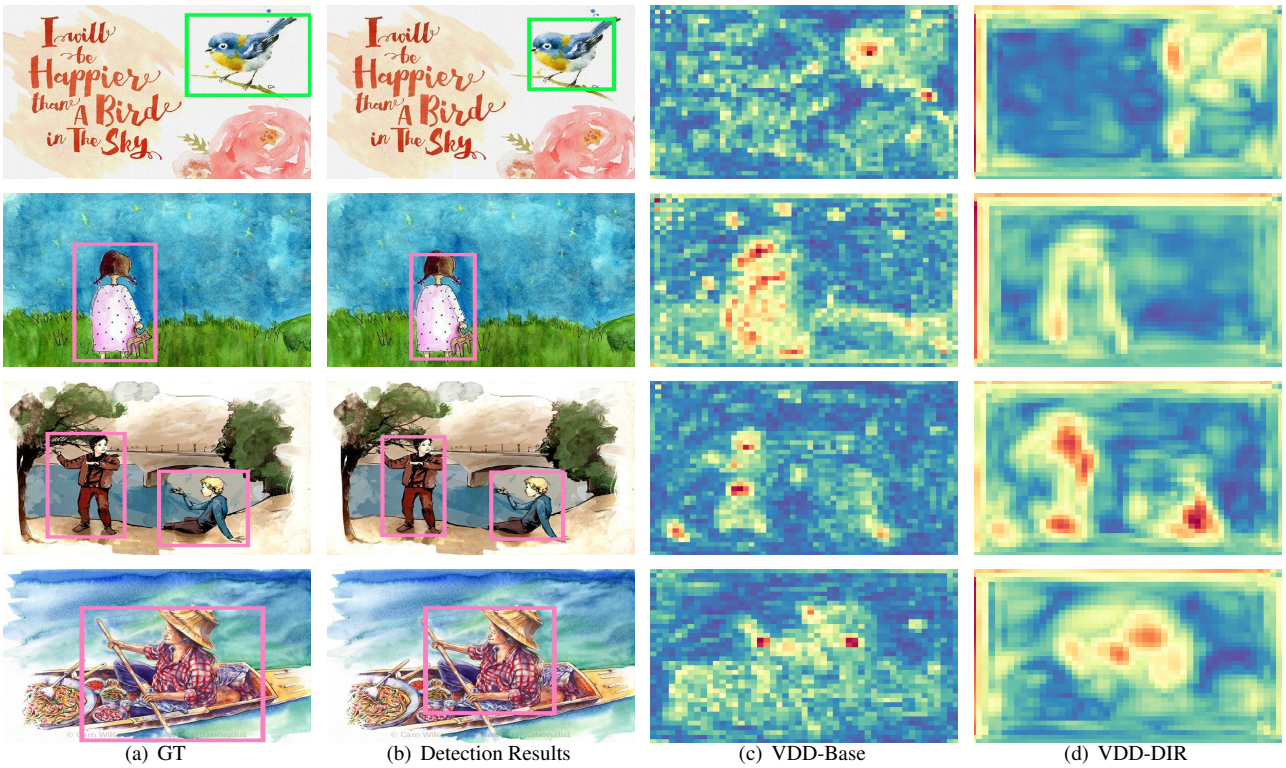
Figure 1: Detection results on the "Cityscapes → FoggyCityscapes" scene. We can see that our method could accurately detect the objects existing in the images, e.g., the bus, car, bicycle, rider, person, and truck.

1. More Implementation Details

In this paper, we employ three convolutional layers as the domain-invariant feature extractor E_{DIR}. Table 1 shows the parameter settings. Here, the size of the convolutional kernel is set to 5, which is helpful for capturing more domain-invariant information. Besides, we design a network with three fully connected layers as the domain classifier. The output channels of the domain classifier network are separately set to 512, 128, and 2. Our method is trained in an end-to-end way. We can see that our method does not introduce many parameters and computational costs, which further shows the effectiveness of our method.

2. Visualization Analysis

Fig. 1 shows more detection examples. Particularly, we can see that the real traffic scene under foggy weather is full of challenges. Our method could localize and recognize these persons effectively, which shows the effectiveness of



(a) GT (b) Detection Results (c) VDD-Base (d) VDD-DIR

Figure 2: Visualization of feature maps of our vector-decomposed disentanglement. 'VDD-Base' indicates the feature map used for disentanglement. 'VDD-DIR' indicates the extracted DIR. These examples are from the 'Pascal VOC → Watercolor' scene. We can see that our method could extract domain-invariant features effectively.

DIR extractor E_{DIR}
Conv $5 \times 5 \times 512 \times 256$, stride 1, padding 2
Batch Normalization, ReLU, Dropout
Conv $5 \times 5 \times 256 \times 256$, stride 1, padding 2
Batch Normalization, ReLU, Dropout
Conv $5 \times 5 \times 256 \times 512$, stride 1, padding 2
Batch Normalization

Table 1: The parameter details of domain-invariant feature extractor E_{DIR}. Here, we use VGG16 as the backbone of Faster R-CNN.

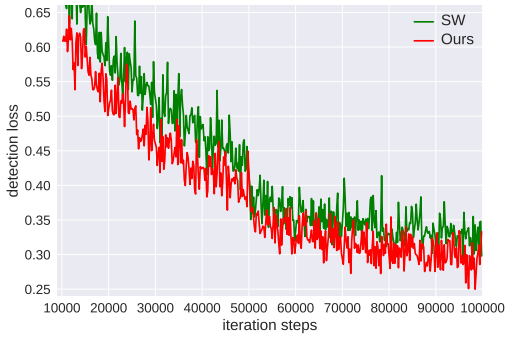


Figure 3: The training loss of our method and SW. Here, we plot the loss from 10,000 steps to 100,000 steps.

our disentangled method.

In Fig. 2, we show more visualization examples. We can see that our method could localize and recognize objects existing in images accurately. Particularly, compared with ground truth, the bounding boxes extracted by our method are much more accurate and contain much less background information. Meanwhile, the visualization results also show that our method could extract the domain-invariant object information effectively. These further demonstrate that our vector-decomposed method could extract domain-invariant features effectively, which leads to the performance improvement.

In Fig. 3, we plot the training loss of our model (i.e., SW-VDD) and SW method on the adaptation from Cityscapes to FoggyCityscapes. Here, we take VGG16 as the backbone network. And we plot the loss from 10,000 steps to 100,000 steps. We can see that our method could converge stably. Meanwhile, the converging position of our method is lower than that of the SW method, which leads to much better detection performance. This further demonstrates the effectiveness of our method.

Figure 4: Detection results on the "Daytime-sunny → Dusk-rainy" and "Daytime-sunny → Night-rainy" scene. The first three rows show the results of the Dusk-rainy scene. The last four rows show the results of the Night-rainy scene. We can see that our method could accurately detect the objects existing in the images, e.g., the bus, car, person, and truck.

3. Results on Dusk-rainy and Night-rainy

Fig. 4 shows more results on dusk-rainy and night-rainy scenes. The images under these two scenes are very challenging. We can see our method detects objects existing in these images accurately. This further demonstrates disentangling DIR is helpful for alleviating the domain-shift impact. And our method is effective for extracting DIR.