Discriminating Known from Unknown Objects via Structure-Enhanced Recurrent Variational AutoEncoder: Supplementary Material

Aming Wu, Cheng Deng*
School of Electronic Engineering, Xidian University, Xi'an, China
amwu@xidian.edu.cn, chdeng@mail.xidian.edu.cn

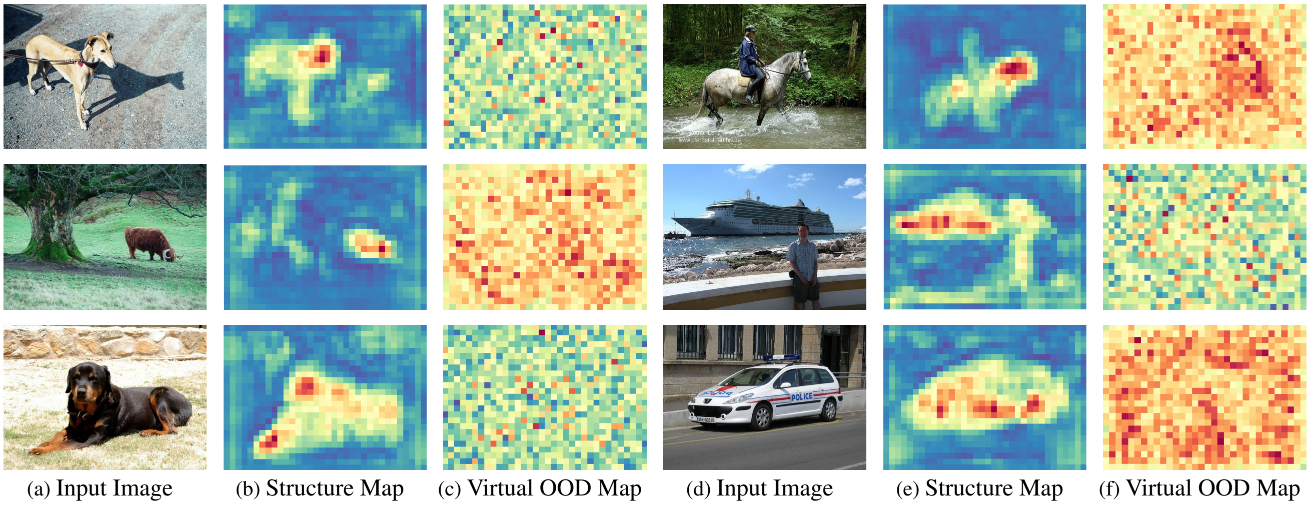

(a) Input Image (b) Structure Map (c) Virtual OOD Map (d) Input Image (e) Structure Map (f) Virtual OOD Map

Figure 1. Visualization of the Structure-Enhanced map and Virtual OOD map based on the ID data (PASCAL VOC). For each feature map, the channels corresponding to the maximum value are selected for visualization.

1. Further Discussion of the OOD Map

To reduce the impact of lacking unknown data for supervision, we propose a method of cycle-consistent conditional VAE to synthesize virtual OOD maps, which is beneficial for improving the ability of distinguishing OOD objects from ID objects. Fig. 6 in the submitted paper and Fig. 1 in the supplementary material separately show some visualization examples from OOD data and ID data.

Compared with the Structure-Enhanced map, we can see that **the synthesized OOD map contains plentiful information that significantly deviates from the object-related features**. Since there is no OOD information available, to obtain sufficient OOD content, it is necessary to enlarge the gap between ID features and OOD features. To this end, we separately define the loss $\mathcal{L}_{\text{dis}}$ and $\mathcal{L}_{\text{cycle}}$. Meanwhile, we also insert the label into the latent space to force a constrained representation, which is instrumental in further enlarging the gap between ID and OOD features.

In the experiments, we separately evaluate our method on OOD-OD, Open-Vocabulary Detection, and Incremental Object Detection. The significant performance gains over baselines indicate the effectiveness of our method.

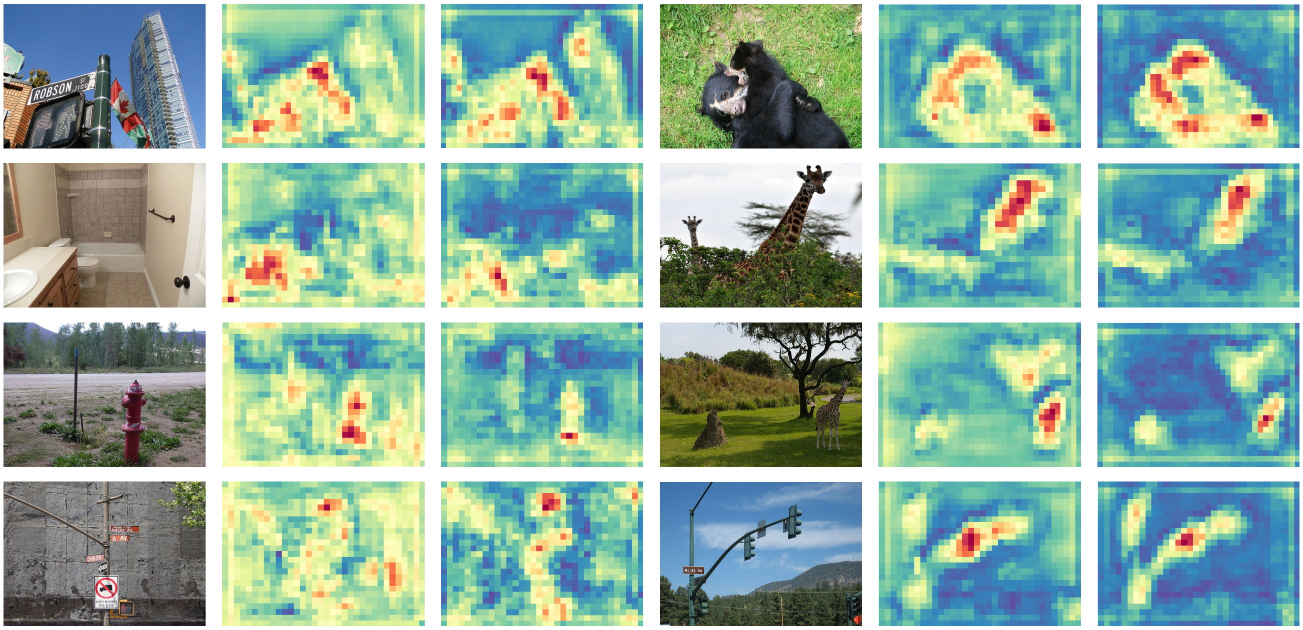
2. Experimental Details of OVD and IOD

In this paper, to further demonstrate the effectiveness of our method, we verify our method on other different tasks, i.e., OVD and IOD. Here, we directly plug our method into two baseline methods and do not calculate the uncertainty loss. The training settings are the same as the baselines.

It is worth noting that to sufficiently exploit the synthesized features, we train a binarized classifier, i.e., the output of the known category is 1, and the output of the synthesized features is 0. By minimizing the cross-entropy loss, the discrimination ability of the object classifier could be enhanced effectively.

3. Analysis of the LoG Operator

To improve the localization performance, we explore utilizing the LoG algorithm to perform a convolution operation on the feature maps extracted by the backbone network,



(a) Input Image (b) Source Map (c) Structure Map (d) Input Image (e) Source Map (f) Structure Map

Figure 2. Comparisons between the Structure-Enhanced map and the Source map extracted by the backbone network. For each feature map, the channels corresponding to the maximum value are selected for visualization.

which is beneficial for enhancing object-related information in the extracted low-level features.

To further indicate the effectiveness of this operation, we make a visualization analysis. Fig. 2 shows some visualization examples. We can see that compared with the Source map, by the LoG operation, the object-related information in the new feature maps is much stronger. This indicates that using the LoG operation is indeed helpful for purifying object-related information and strengthening the ability of object localization.

4. Ablation Analysis of Hyper-Parameters

For our method, we utilize the hyper-parameter α for the loss $\mathcal{L}_{\text{in}}$ (Eq. (4)), hyper-parameter λ and τ for the loss $\mathcal{L}$ (Eq. (8)). Since the uncertainty loss $\mathcal{L}_{\text{uncertainty}}$ is directly related to the current task, the value of τ should be set larger than α and λ. Meanwhile, if α and λ are set to a small value, the role of the two corresponding losses will be weakened in optimization. Thus, it is meaningful to set proper values for these hyper-parameters. Here, we take BDD-100k as the ID data and MS-COCO as the OOD data to perform an ablation analysis of hyper-parameters. And we only change these hyper-parameters and keep other modules unchanged.

Analysis of α. The hyper-parameter α in Eq. (4) is to balance the detection loss and the loss that aims to minimize the KL-divergence between the prediction probabilities from H_t and P_{in}. In the experiments, we observe that when α is set to 0.01, 0.001, and 0.0001, the performance of FPR95 is 33.12%, 32.23%, and 32.56%.

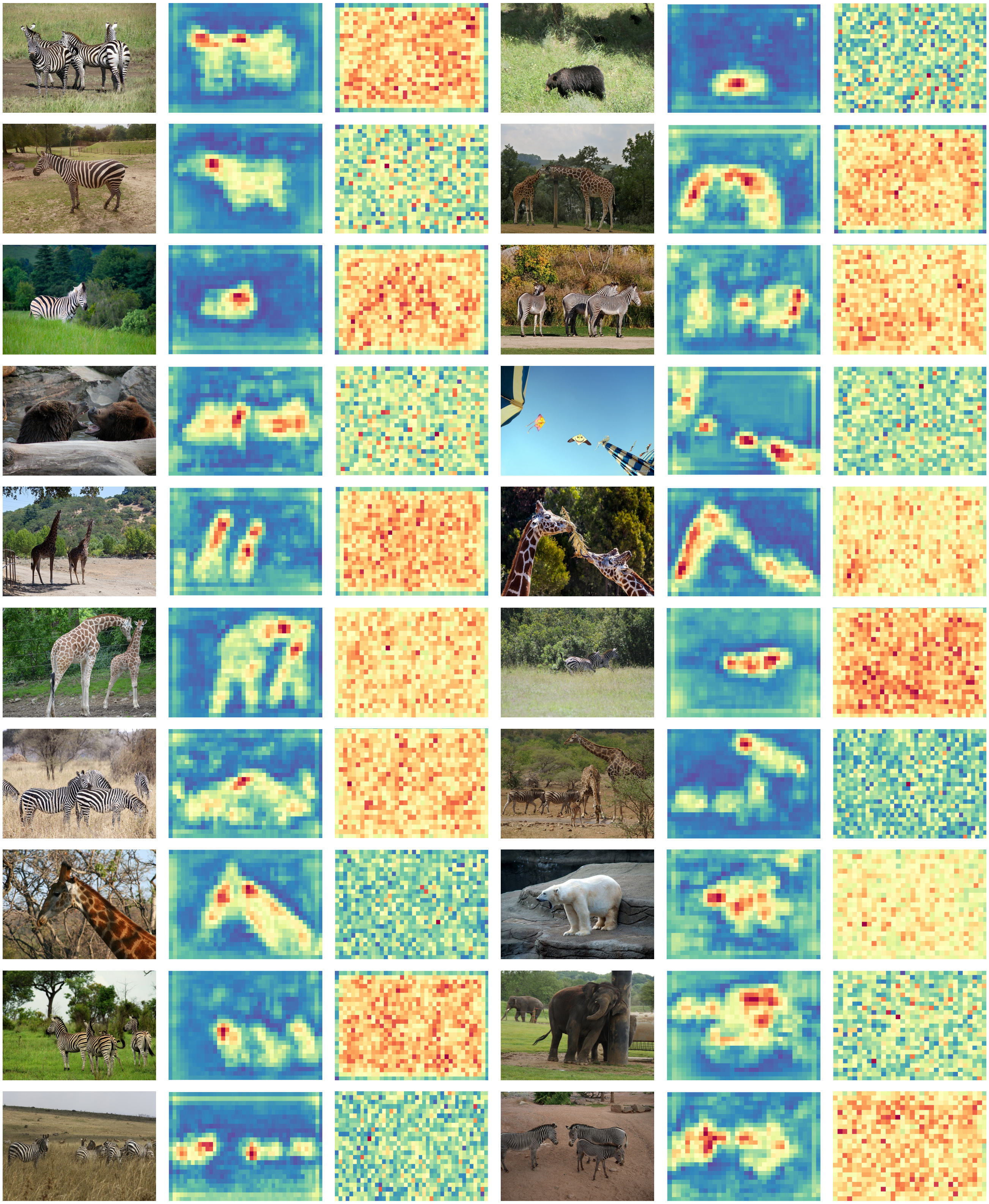
Analysis of λ. The goal of the hyper-parameter λ in Eq. (8) is to weigh the importance of the module of cycle-consistent conditional VAE. In the experiments, we find that when λ is set to 0.01, 0.001, and 0.0001, the corresponding FPR95 performance is 33.41%, 32.23%, and 33.04%.

Analysis of τ. The hyper-parameter τ in Eq. (8) is to constrain the uncertainty loss $\mathcal{L}_{\text{uncertainty}}$. In the experiments, we observe that when τ is set to 0.5, 0.1, and 0.01, the corresponding performance of FPR95 is 34.17%, 32.23%, and 32.76%. This shows that when the role of the uncertainty loss is intensified, the performance of OOD object detection will be decreased. The reason may be that enlarging the uncertainty loss weakens the importance of other losses, e.g., the detection loss.

5. More Visualization Analysis

In Fig. 3, we show more visualization examples of our method. We can see that the structure-enhanced maps indeed contain plentiful object-related information, which is beneficial for improving the performance of object localization. Meanwhile, the synthesized virtual OOD maps involve rich information that deviates from the object-related contents, which is instrumental in improving the ability of distinguishing OOD objects.

Finally, in Fig. 4 and 5, we show more detection results about OOD objects and ID objects. We can see that for these images, our method could accurately distinguish OOD objects from ID objects, which further demonstrates that our method could indeed enhance the discrimination ability.



(a) Input Image (b) Structure Map (c) Virtual OOD Map (d) Input Image (e) Structure Map (f) Virtual OOD Map

Figure 3. Visualization of the Structure-Enhanced map and Virtual OOD map based on the OOD data (MS-COCO). For each feature map, the channels corresponding to the maximum value are selected for visualization.

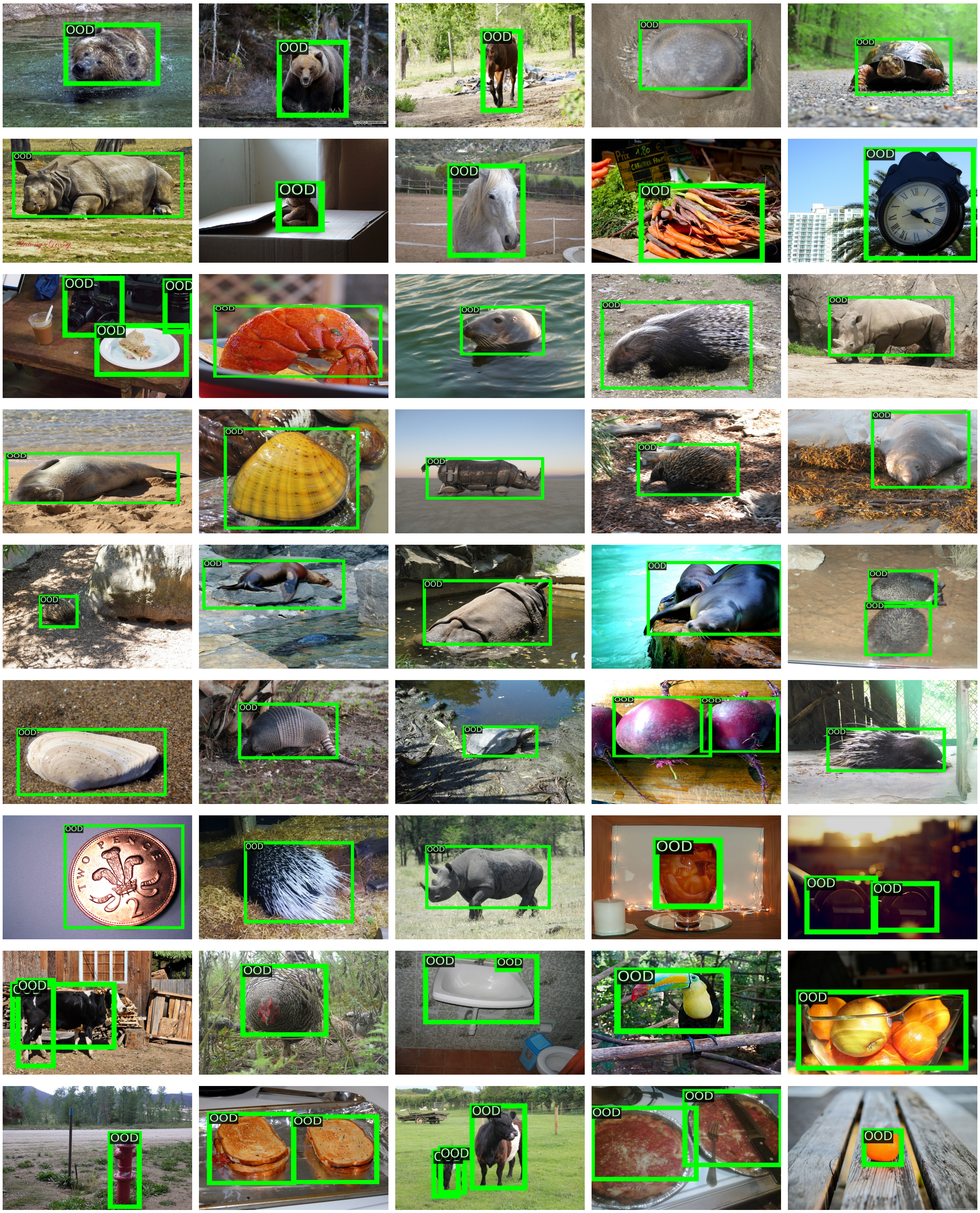


Figure 4. Detection results on the OOD images from MS-COCO. We can see that our method detects OOD objects accurately, which further demonstrates the effectiveness of our method.

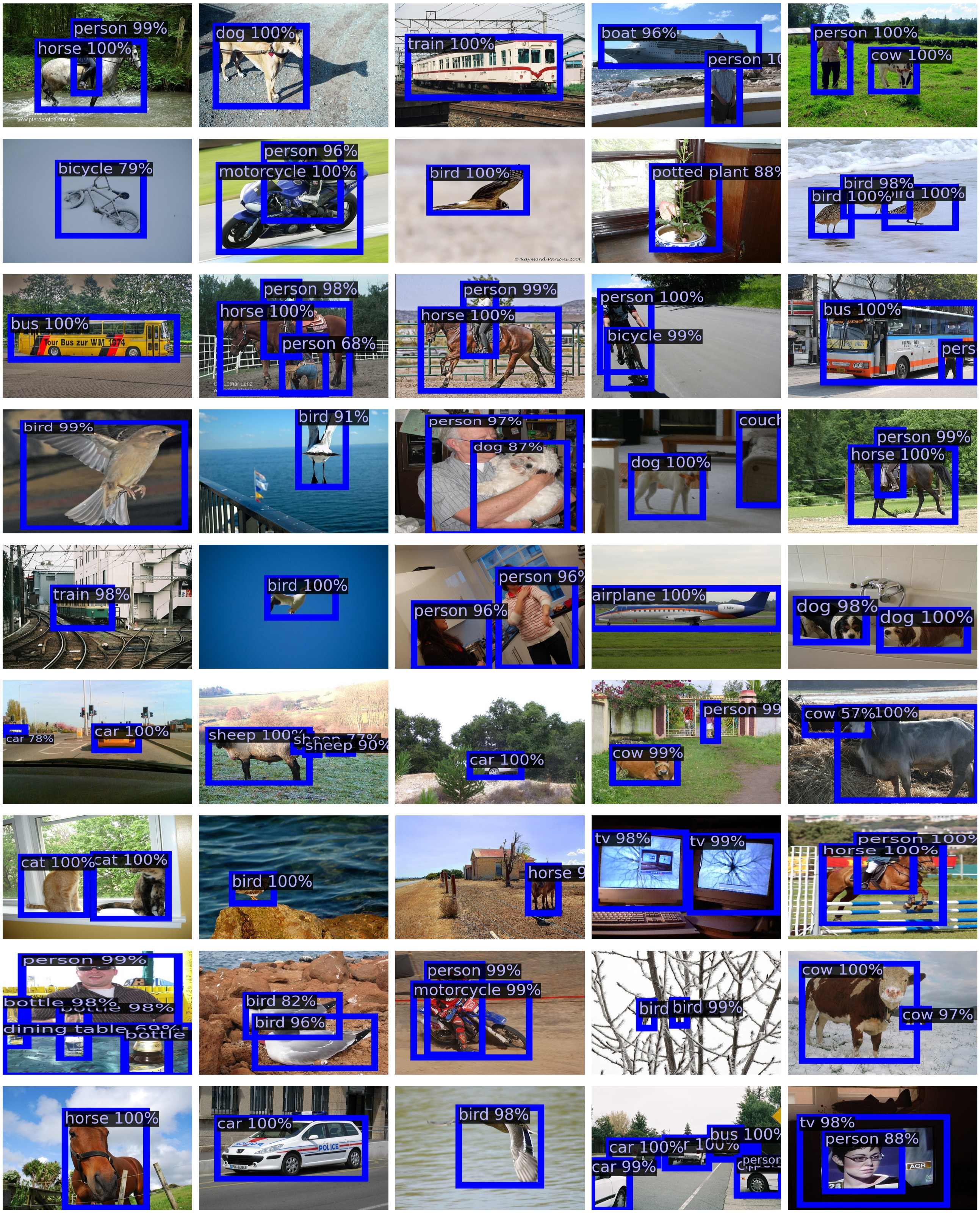


Figure 5. Detection results on In-Distribution dataset, i.e., PASCAL VOC. We can see that our method effectively detects objects in these images, which shows the advantages of our method.