Supplementary Material for the Paper "Enriching Object Detection with 2D-3D Registration and Continuous Viewpoint Estimation"

Christopher Bongsoo Choy[†], Michael Stark[‡], Sam Corbett-Davies[†], Silvio Savarese[†]
[†]Stanford University, [‡]Max Planck Institute for Informatics
{chrischoy, scorbett, ssilvio}@stanford.edu, [‡]stark@mpi-inf.mpg.de

In the following, we present additional quantitative and qualitative results that accompany the experiments described in Sect. 6 of the main paper, "Enriching Object Detection with 2D-3D Registration and Continuous Viewpoint Estimation".

1. 2D-3D Matching as an Object Detector

It this section, we provide qualitative examples and plots for the experiment "2D-3D Matching as an Object Detector" (Sect. 6.2 in the main paper).

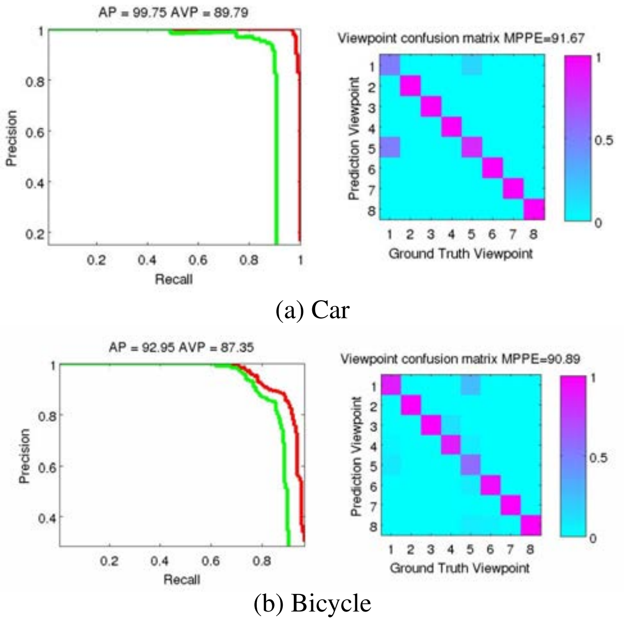
To recapitulate, we run our ensemble of NZ-WHO templates on the 3D Object Classes dataset [2], without the fine-tuning stage. Fig. 1 gives the corresponding detection average precision, average viewpoint precision, viewpoint confusion matrix and mean precision in pose estimation results. Specifically, we followed the detection and viewpoint estimation criteria of [4] where a detection is correct iff intersection over union is at least 0.5 and viewpoint estimation is correct iff detection is correct *and* azimuth of the viewpoint prediction falls into the correct viewpoint bin.

Fig. 2 shows successful detection and viewpoint estimation results for car, Fig. 3 for bicycle. Fig. 4 and Fig. 5 show failure cases, which are mostly due to confused front and back views for cars, and slanted bicycle poses.

2. Enriching Existing Detections

It this section, we provide qualitative examples and plots for the experiment "Enriching Existing Detections" (Sect. 6.3 in the main paper).

To recapitulate, we enrich object detection bounding boxes from a state-of-the-art detector (R-CNN [1]) with 2D-3D registration and continuous viewpoint estimation. We present detection average precision, average viewpoint precision, viewpoint confusion matrix and mean precision in pose estimation results in Fig. 6, following the evaluation criteria of [3] for detection and viewpoint estimation. Fig. 7 to Fig. 10 give qualitative results.



(a) Car

(b) Bicycle

Figure 1: Detection and pose estimation on 3D Object Classes [2] car (a) and bicycle (b). Average Precision (red) and Average Viewpoint Precision (green) are given in the left plot and viewpoint confusion table and MPPE are given in the right plot. The viewpoint index 1 is front, 2 is front-right, . . . , 8 is front-left.

We use 15% context (extending the proposal region by 15% on the top, right, left, and bottom). We run our pipeline on the PASCAL3D+ [3] dataset, including our fine-tuning stage. If our method fails to estimate viewpoint (confidence score is below a threshold), it outputs 0 azimuth, 0 elevation, 0 yaw. Note that for both categories and different numbers of viewpoints, our method has difficulty distinguishing front and back views, but yields little confusion between neighboring views. The failure cases given in Fig. 9 and Fig. 10 are mostly caused by truncation, occlusion, and unusual object shape.

Please visit https://www.youtube.com/watch?v=YKtioOXY8yQ for a biref summary of the paper and MCMC fine-tuning visualizations.

3. Fine Tuning

Lastly, we provide qualitative examples of the fine-tuning stage of our pipeline based on MCMC sampling (Sect. 5 in the main paper) in Fig. 11. Please note that, while the visual difference appear subtle, our method often manages to improve upon the initial pose estimate.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[2] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *ICCV*, 2007.

[3] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014.

[4] Y. Xiang and S. Savarese. Estimating the aspect layout of object categories. In *CVPR*, 2012.

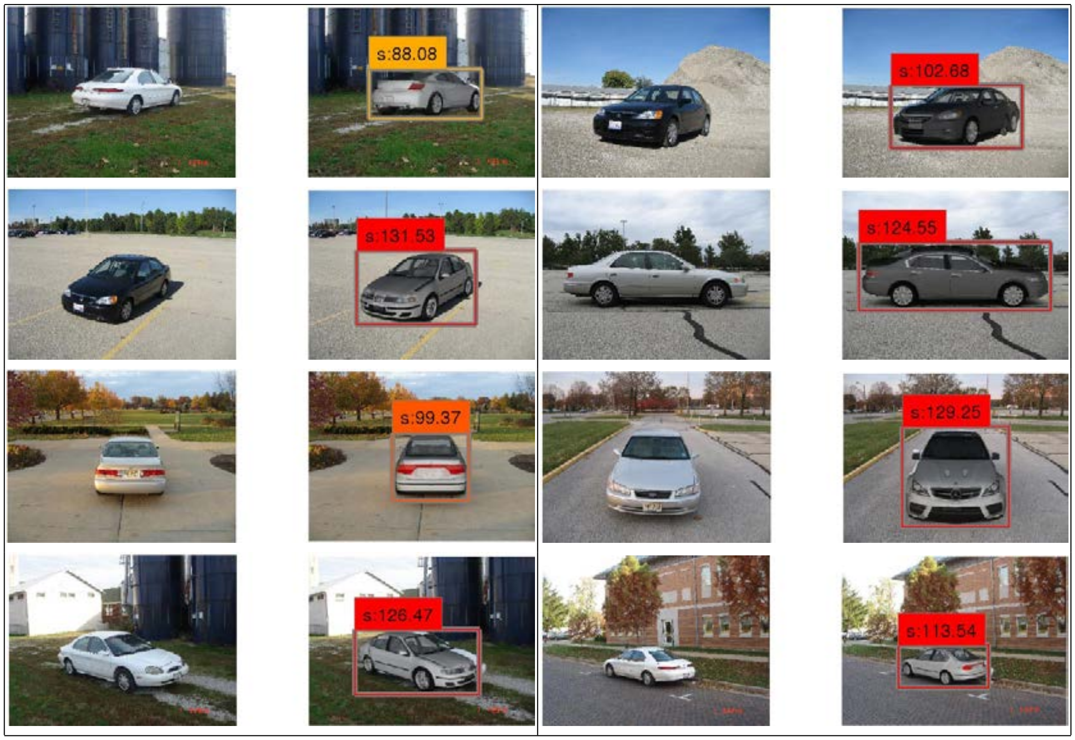


Figure 2: Successful detection results on 3D Object Classes [2] cars. Original image (left) and overlaid detection result (right).

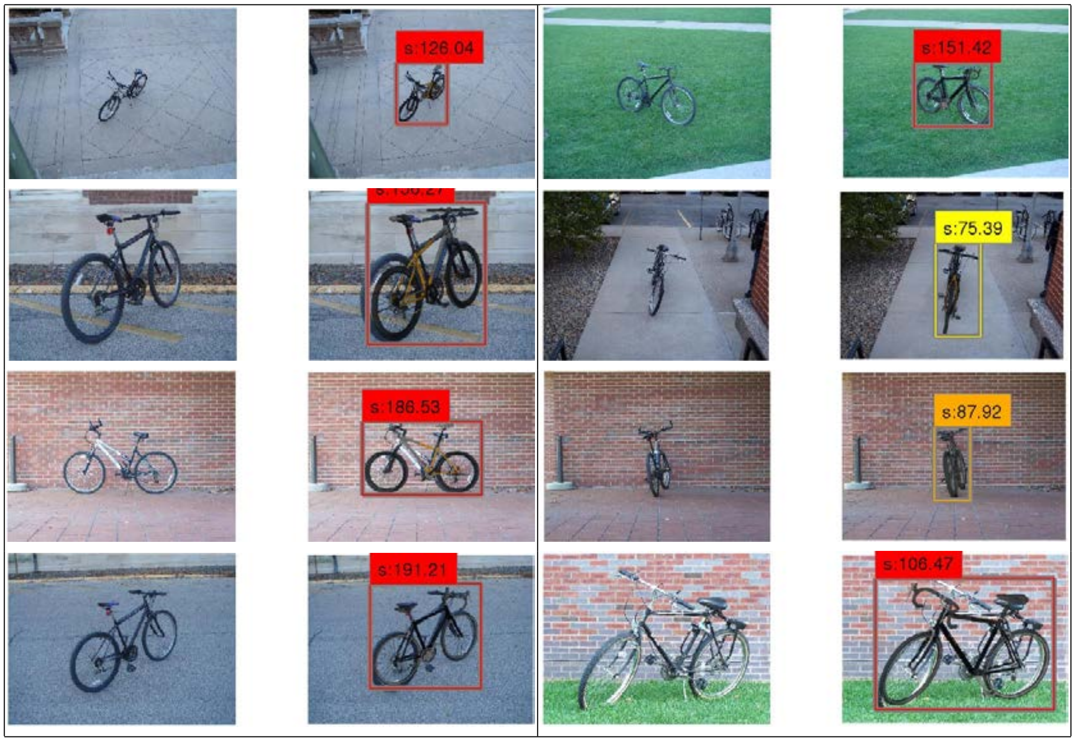


Figure 3: Successful detection results on 3D Object Classes [2] bicycles. Original image (left) and overlaid detection result (right).

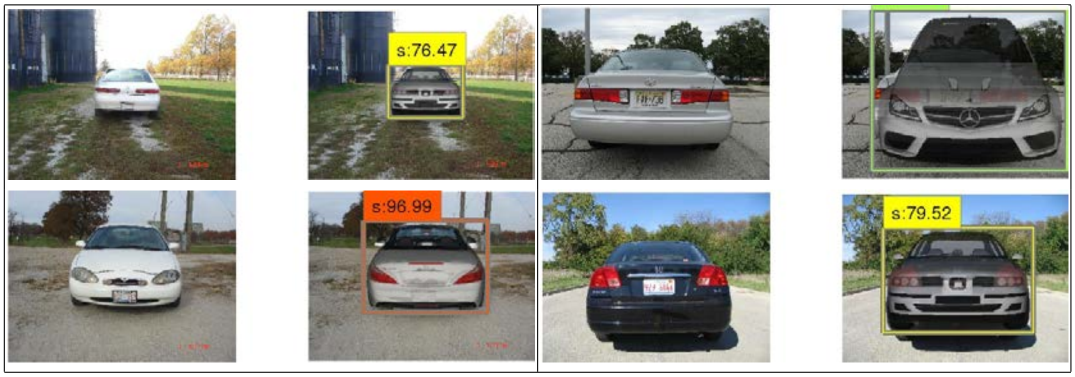


Figure 4: Failed detection or pose estimation on 3D Object Classes [2] cars. Original image (left) and overlaid detection result (right).

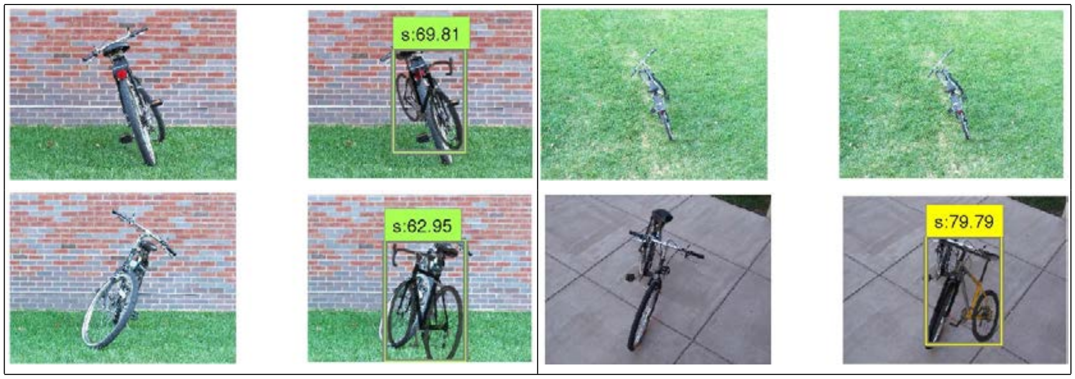
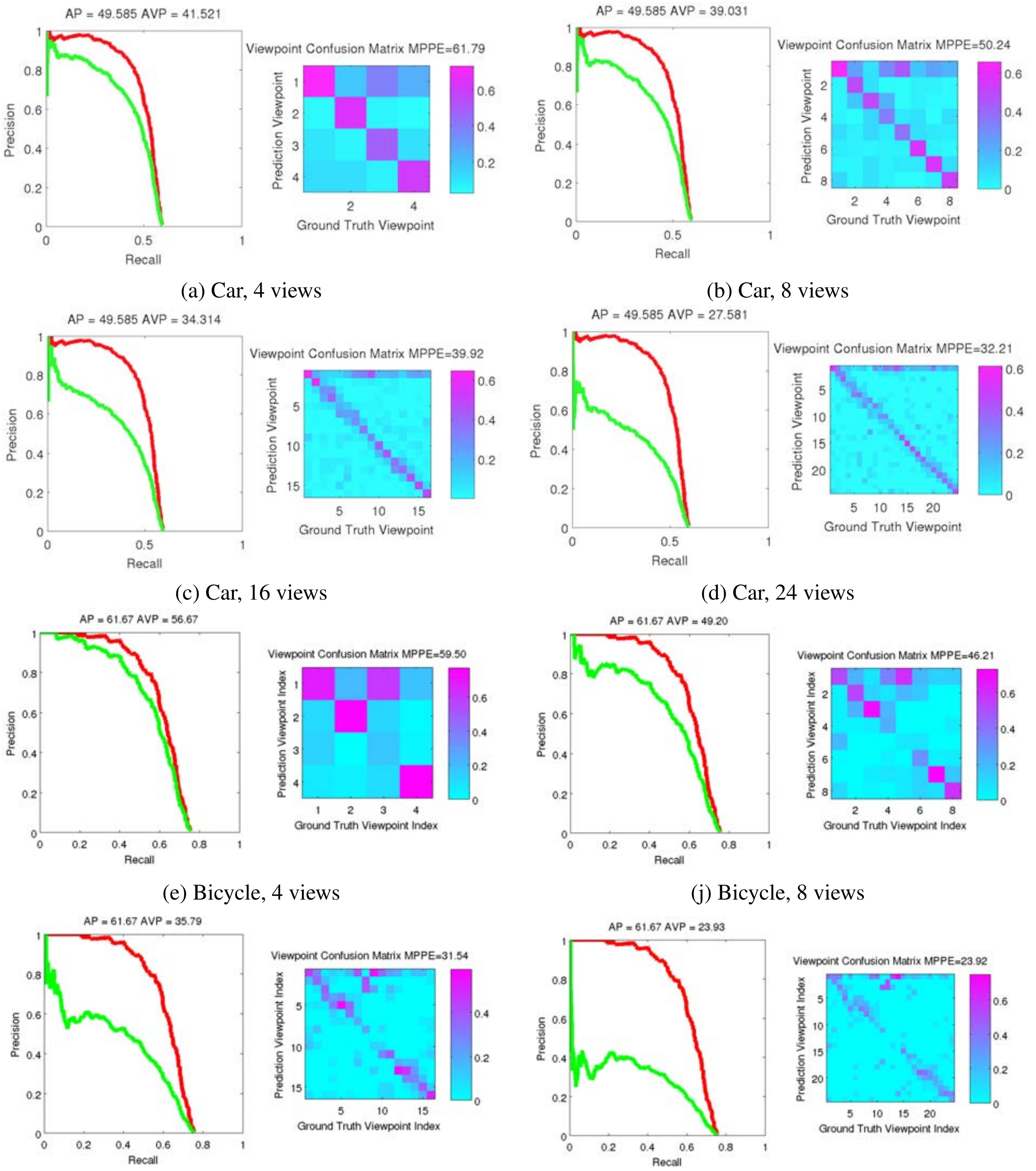


Figure 5: Failed detection or pose estimation on 3D Object Classes [2] bicycles. Original image (left) and overlaid detection result (right).



(a) Car, 4 views

(b) Car, 8 views

(c) Car, 16 views

(d) Car, 24 views

(e) Bicycle, 4 views

(j) Bicycle, 8 views

(g) Bicycle, 16 views

(h) Bicycle, 24 views

Figure 6: Detection and pose estimation on PASCAL3D+ [3] cars and bicycles. Average Precision (red) and Average Viewpoint Precision (green) are given in the left plot and viewpoint confusion table and MPPE are given in the right plot. The viewpoint index 1 is front, 2 is front-right, . . . , $n - 1$ is front-left for number of viewpoints n.

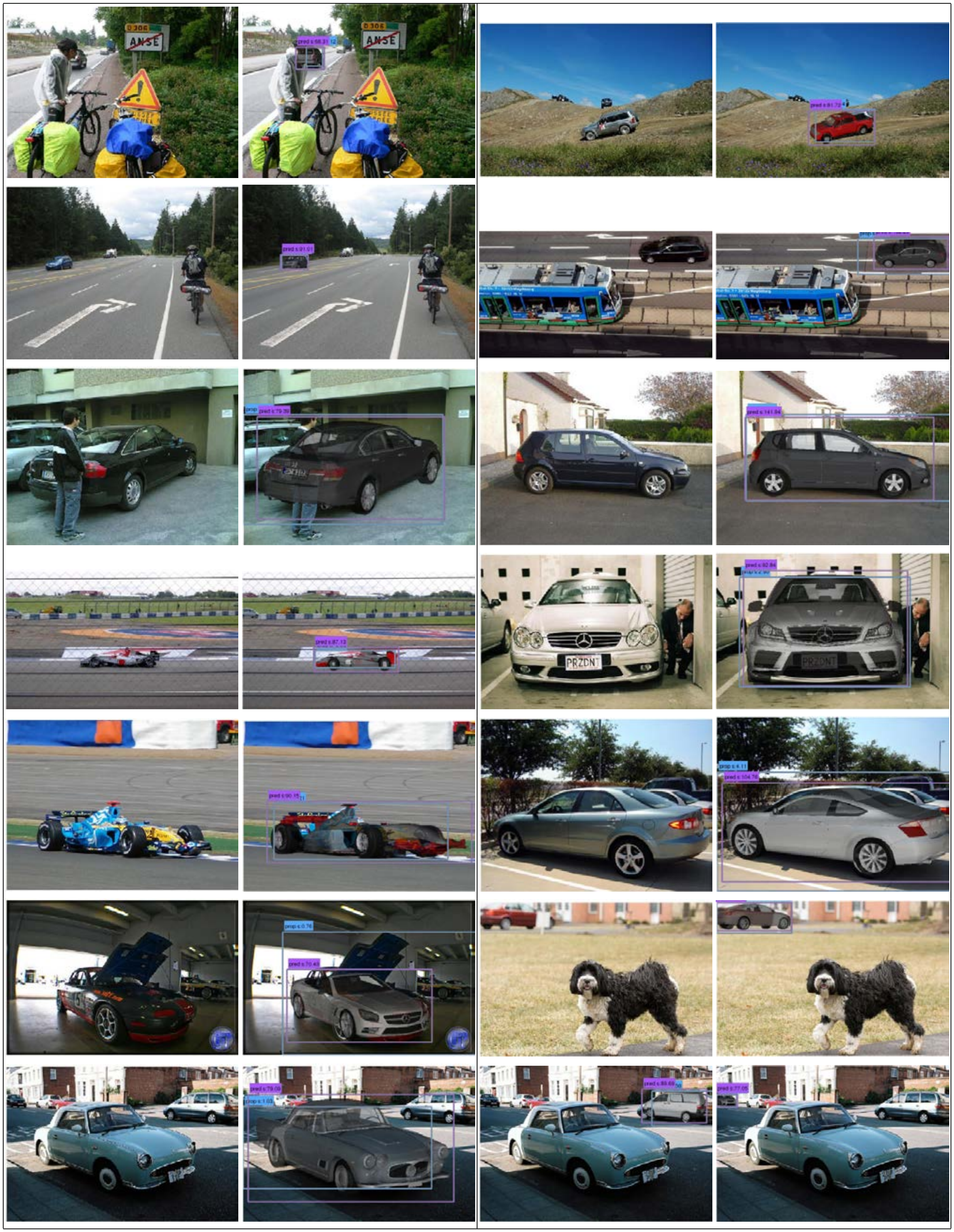


Figure 7: Successful detection and pose estimation results on PASCAL3D+ [3] cars. Original image (left) and overlaid detection result (right). Proposal bounding box (blue) and predicted bounding box (purple).

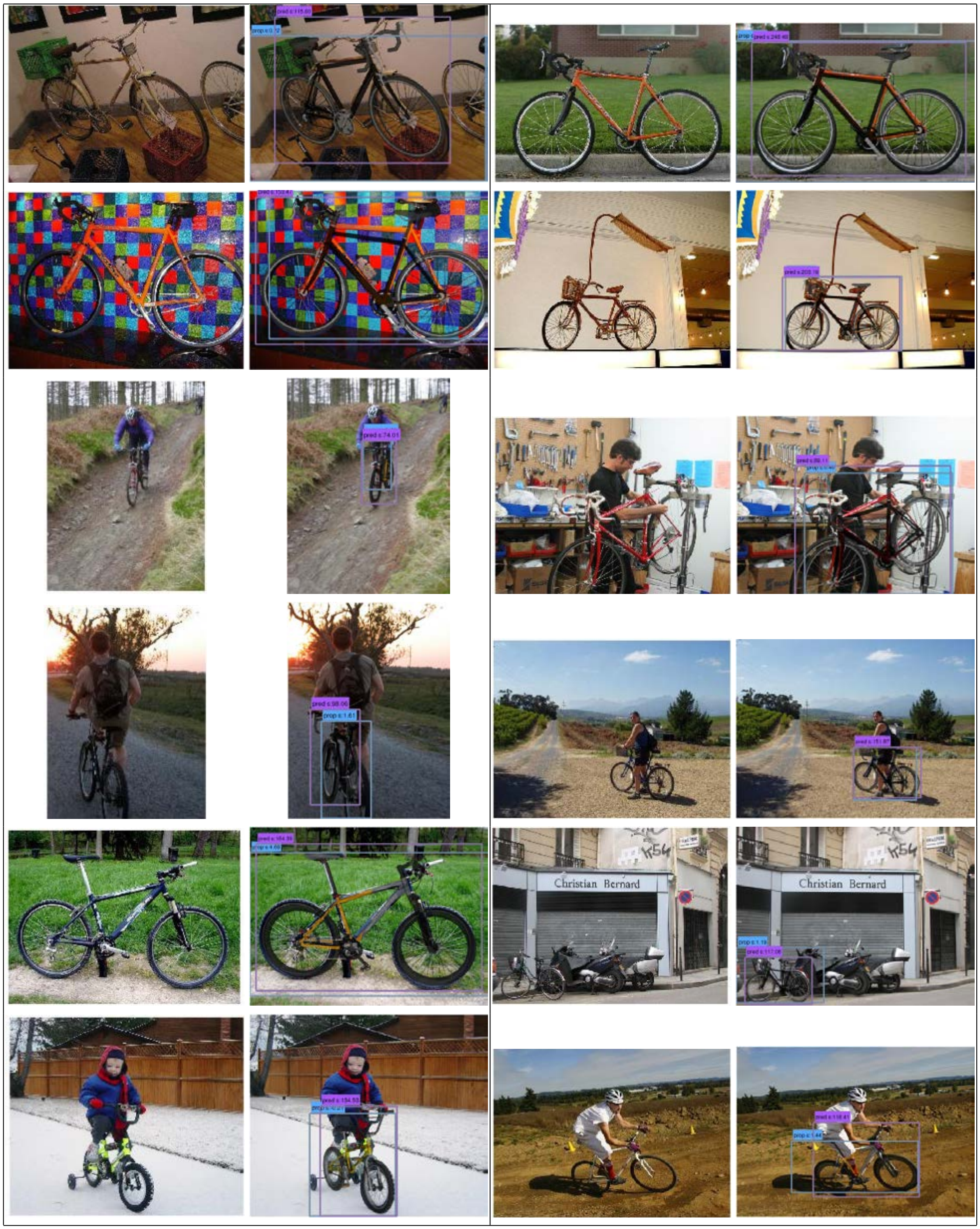


Figure 8: Successful detection and pose estimation results on PASCAL3D+ [3] bicycles. Original image (left) and overlaid detection result (right). Proposal bounding box (blue) and predicted bounding box (purple).

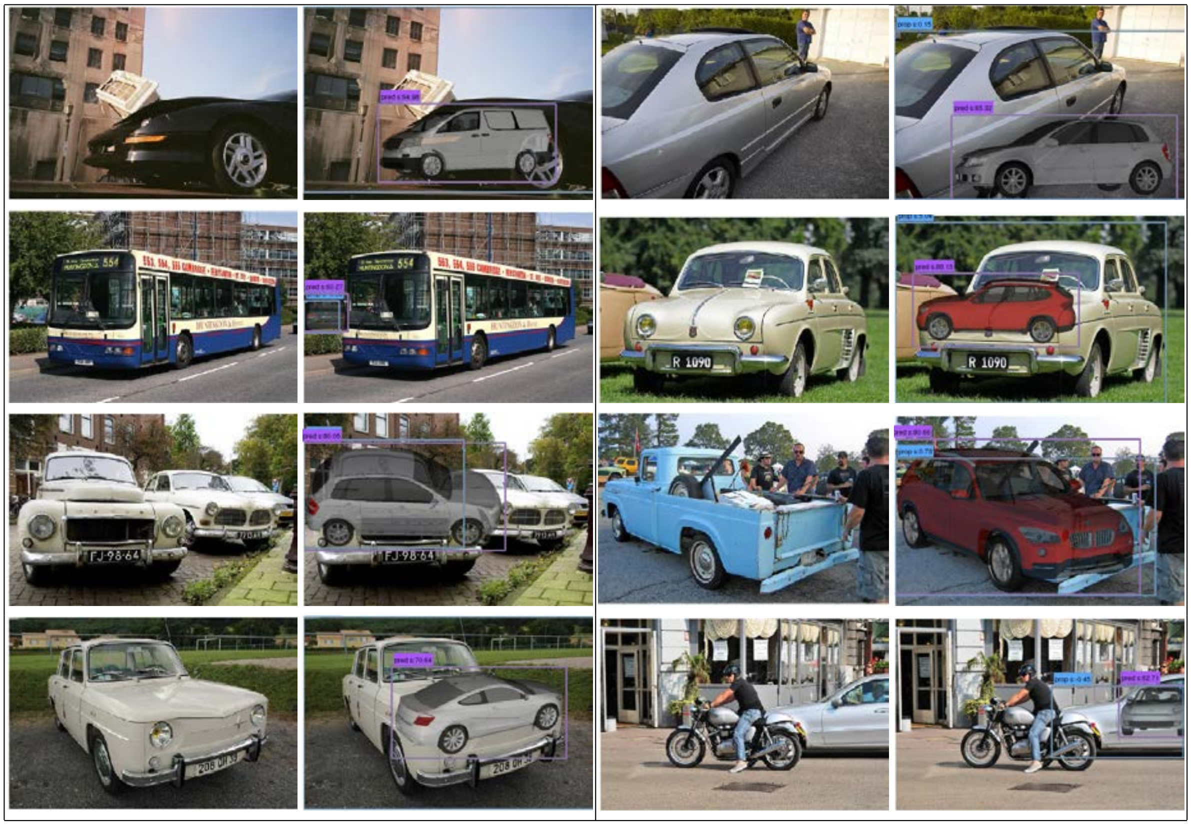


Figure 9: Failed detection or pose estimation result on PASCAL3D+ [3] cars. Original image (left) and overlaid detection result (right). Proposal bounding box (blue) and predicted bounding box (purple).

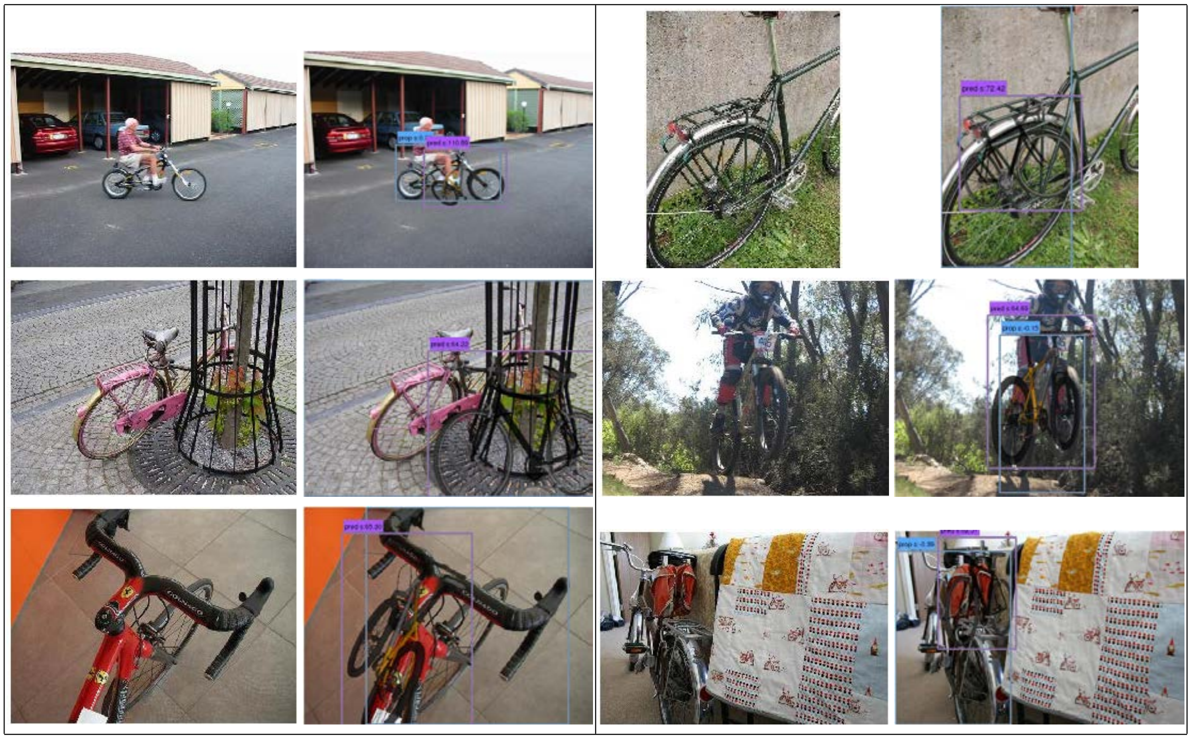


Figure 10: Failed detection or pose estimation result on PASCAL3D+ [3] bicycles. Original image (left) and overlaid detection result (right). Proposal bounding box (blue) and predicted bounding box (purple).

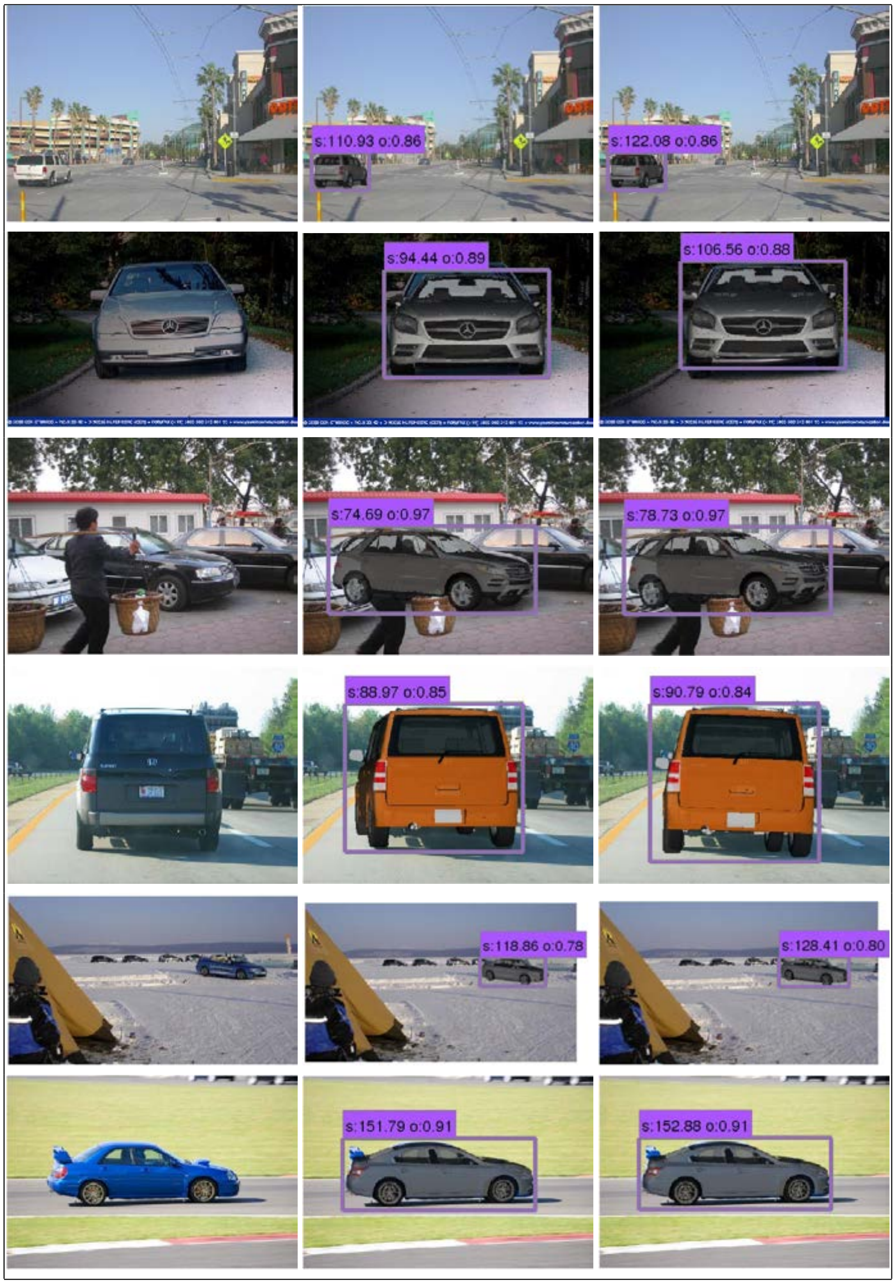


Figure 11: Effect of fine-tuning based on MCMC sampling. From left to right: original image, initial pose estimate, and fine-tuned result. Numbers indicate detection confidence (s) and intersection over union (o), respectively.