# Supplementary Material: Monocular 3D Object Detection for Autonomous Driving

Xiaozhi Chen[1], Kaustav Kundu[2], Ziyu Zhang[2], Huimin Ma[1], Sanja Fidler[2], Raquel Urtasun[2]
[1]Department of Electronic Engineering, Tsinghua University
[2]Department of Computer Science, University of Toronto
{chenxz12@mails., mhmpub@}tsinghua.edu.cn, {kkundu, zzhang, fidler, urtasun}@cs.toronto.edu

## Abstract

*In this supplementary material we include a larger set of additional experiments and visualizations. We start with comparisons to the state-of-the-art in the KITTI test set, followed by an in-depth analysis of 2D and 3D bounding box recall as a function of number of proposals as well as the distance to the obstacle. The latter is a very important metric in the context of autonomous driving. We then show an ablation study of the features employed followed by some additional visualizations.*

## 1. Object Detection and Orientation Estimation Performance

Fig 1 and Fig. 2 show a comparison to all published monocular methods on the KITTI benchmark. Our approach achieves the highest AP and AOS scores across all categories and difficulty levels. Note that we also provide a video visualizing our results in 2D and 3D.

## 2. Proposal Recall

In this section we report recall of our proposals in several regimes.

**2D Bounding Box Recall:** We show recall versus the number of proposals in Fig. 3, and recall versus IoU overlap threshold in Fig. 4, Fig. 5, and Fig. 6, by fixing the number of proposals to 500, 1000 and 2000, respectively. We also report average recall (AR) as a function of the number of proposals in Fig. 7. Our approach outperforms all monocular methods by a large margin, while being competitive with 3DOP [1], which uses stereo imagery, and thus is not a fair comparison.

**Recall vs Distance:** We report recall as a function of the distance from the ego-car in Fig. 8. It can be seen that our approach achieves very high recall even when the distance is quite large (higher than 40m). This shows the advantage of our approach in the setting of autonomous driving.

**3D Bounding Box Recall:** We also compare 3D bounding box recall of our monocular approach with 3DOP [1], which, however, exploits stereo imagery. Fig. 9 shows 3D box recall as a function of the number of proposals. We set the 3D IoU overlap threshold to 0.25 for all categories. Although we do not exploit any depth features, our approach achieves similar 3D recall as 3DOP on *Car*. For small objects, i.e., *Pedestrian* and *Cyclist*, our results are also promising.

## 3. Ablation Study of features

We conduct a detailed analysis of different types of features on *Car* proposals and show recall plots in Fig. 10, Fig. 11, and Fig. 12. Note that each feature helps improve performance. We also study their effect on car detection and orientation estimation. As shown in Table. 1, each type of feature improves AP and AOS similar to their behaviors on proposal recall.

Table 1: **Ablation study of features on Object Detection and Orientation Estimation:** AP and AOS for *Car* on validation set of KITTI.

| Approach | AP | | | AOS | | |
|---|---|---|---|---|---|---|
| | Easy | Moderate | Hard | Easy | Moderate | Hard |
| Loc | 83.59 | 77.50 | 69.41 | 81.56 | 75.09 | 66.39 |
| +ClsSeg | 87.96 | 86.76 | 78.16 | 86.40 | 84.66 | 75.72 |
| +Context | 93.17 | 88.23 | 79.34 | 91.59 | 86.15 | 76.94 |
| +Shape | 93.52 | 88.51 | 79.62 | 91.49 | **86.38** | **77.15** |
| +InstSeg | **93.89** | **88.67** | **79.68** | **91.90** | 86.28 | 77.09 |

## 4. Visualization

We visualize some qualitative results in Fig. 13 and Fig. 14. We show top 50 proposals, 2D detections and 3D detections for each example. We also show some failure examples in Fig. 15. Most failures are propagated from errors in class semantic segmentation. Road segmentation particularly affects the results as our approach infers extent of the 3D space from the road region.

## 5. Video

We refer the reader to the attached video for more visualizations of results. We note that to create the video no temporal information is used, and all results are obtained from using a single monocular image.

## References

[1] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals for accurate object class detection. In *NIPS*, 2015. 1

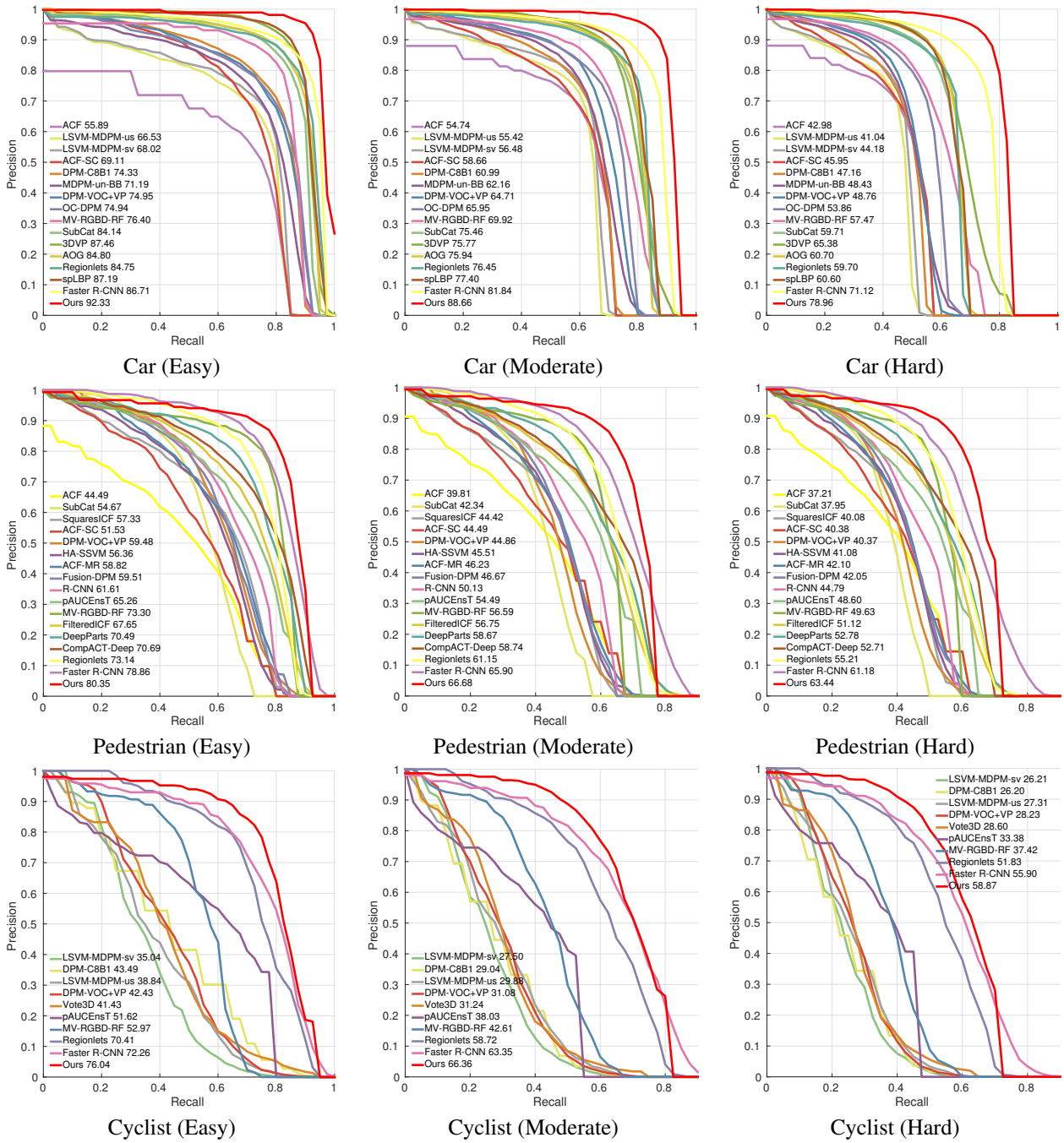



Figure 1: **Precision vs Recall curves on KITTI test set**. The number next to the label indicates the Average Precision (AP).

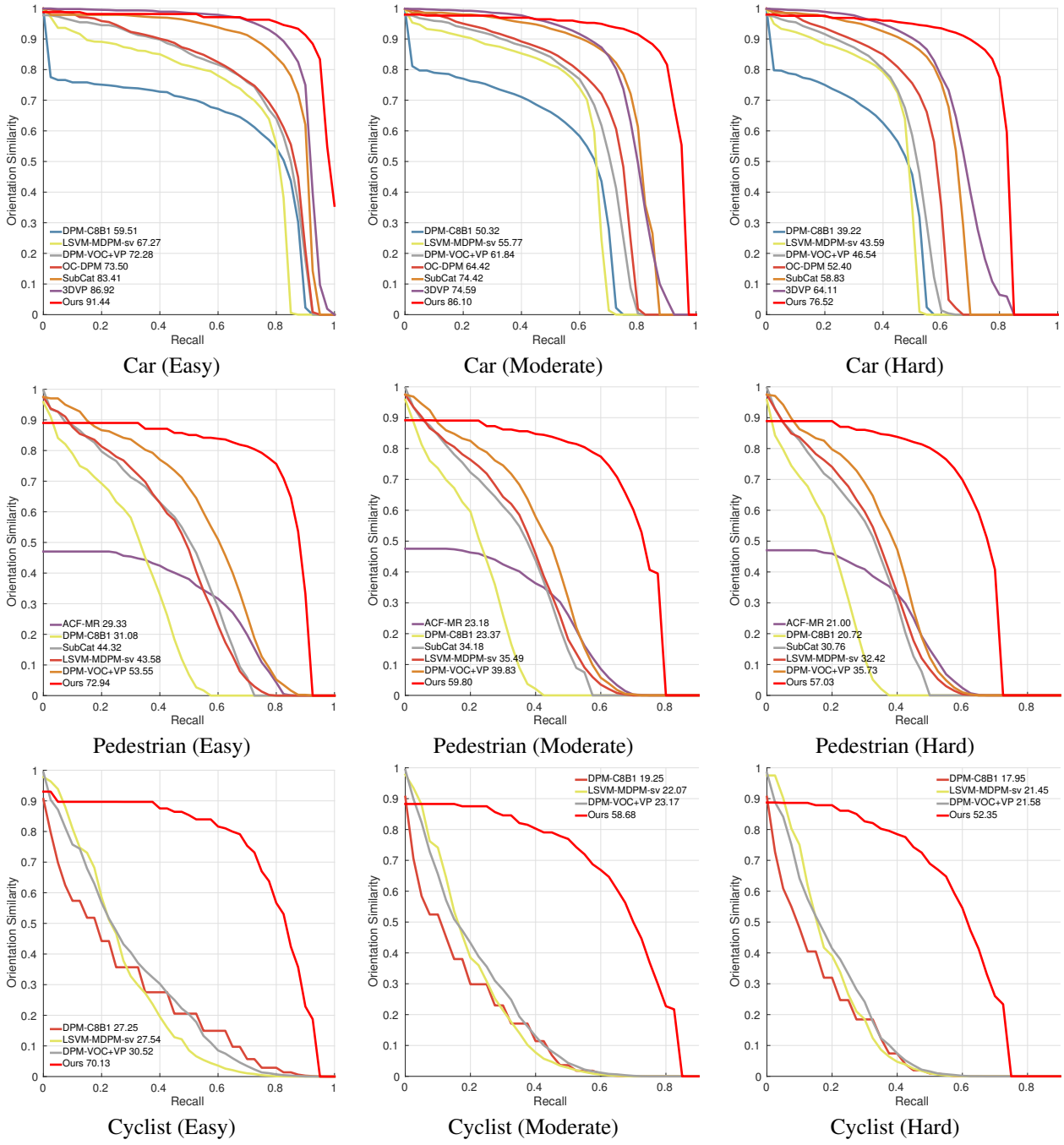



Figure 2: **Orientation Similarity vs Recall curves on KITTI test set**. The number next to the label indicates the Average Orientation Similarity (AOS).

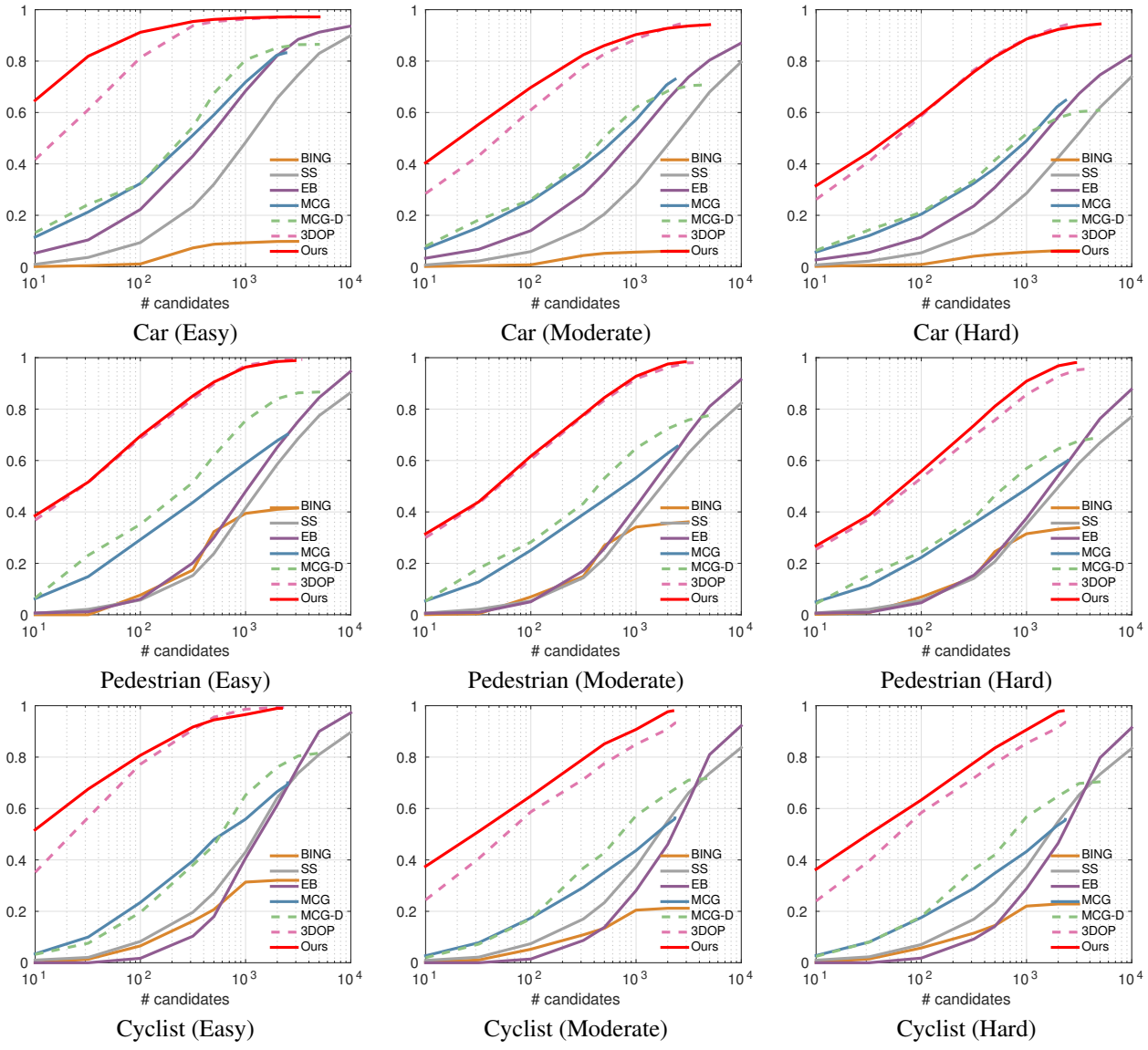



Figure 3: **2D bounding box Recall vs #Candidates**. We use an overlap threshold of 0.7 for Car, and 0.5 for Pedestrian and Cyclist. From left to right are for easy, moderate, and hard objects, respectively.

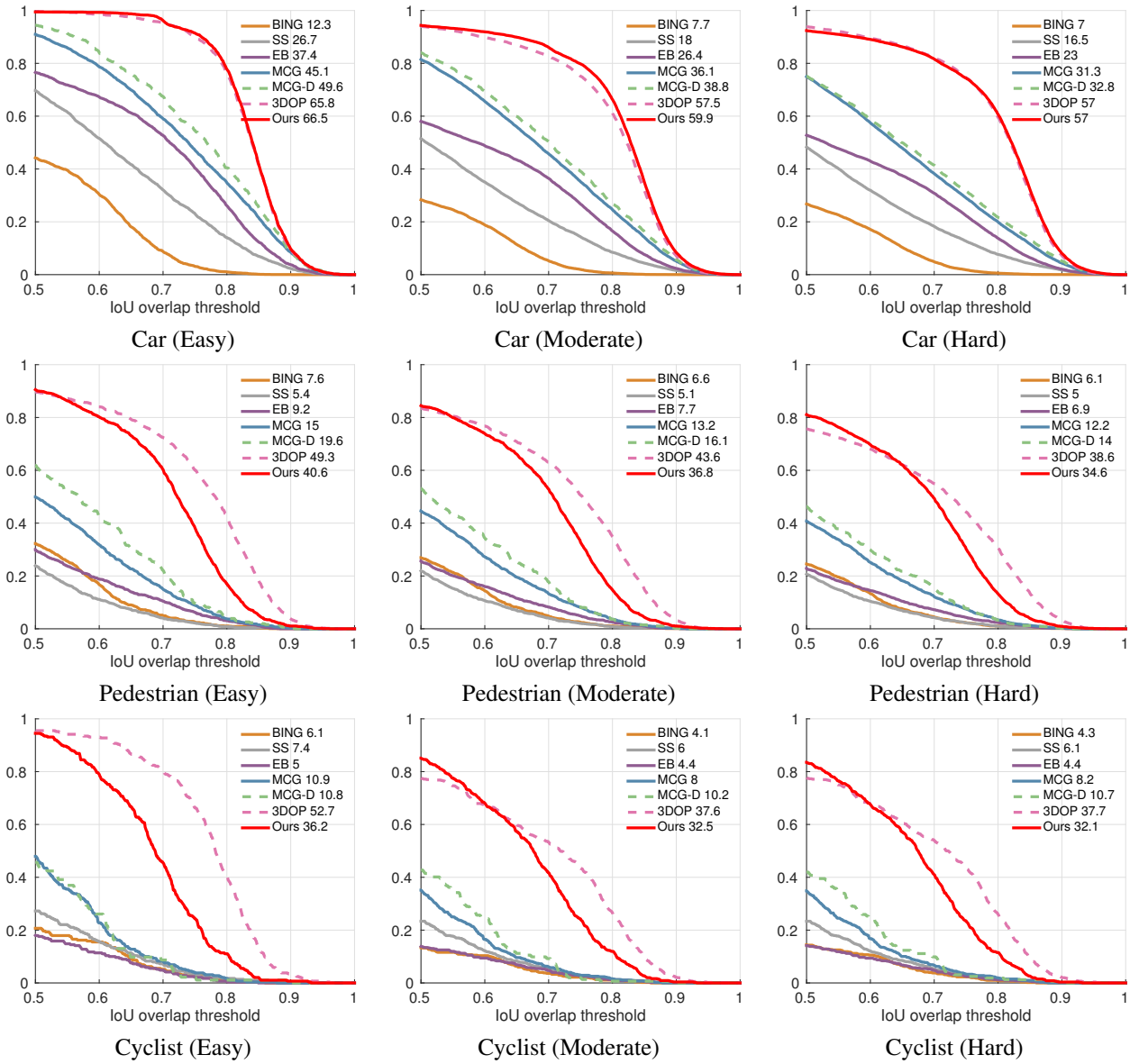



Figure 4: **2D bounding box Recall vs IoU for 500 proposals**. The number next to the label indicates the average recall (AR).

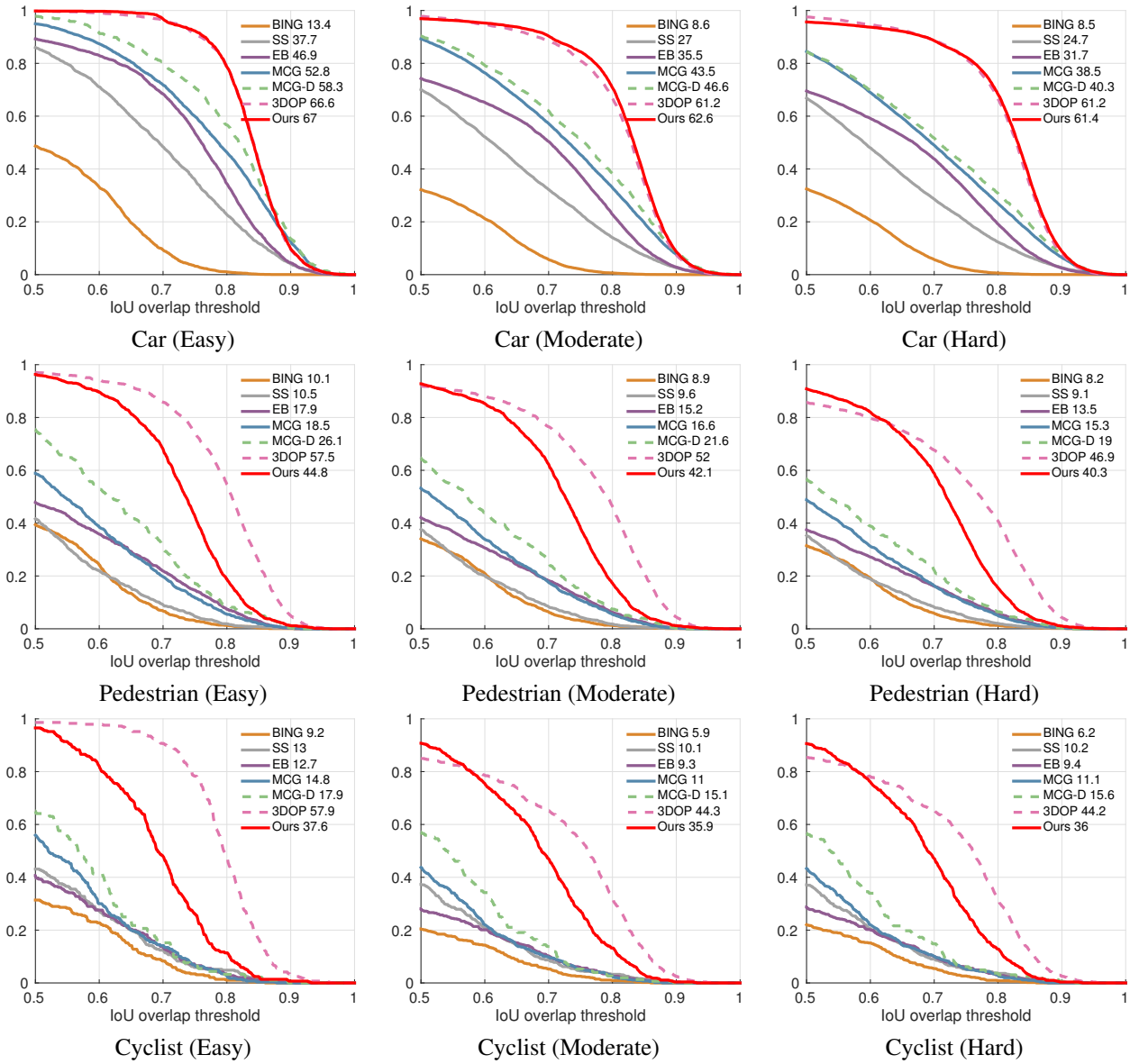



Figure 5: **2D bounding box Recall vs IoU for 1000 proposals**. The number next to the label indicates the average recall (AR).

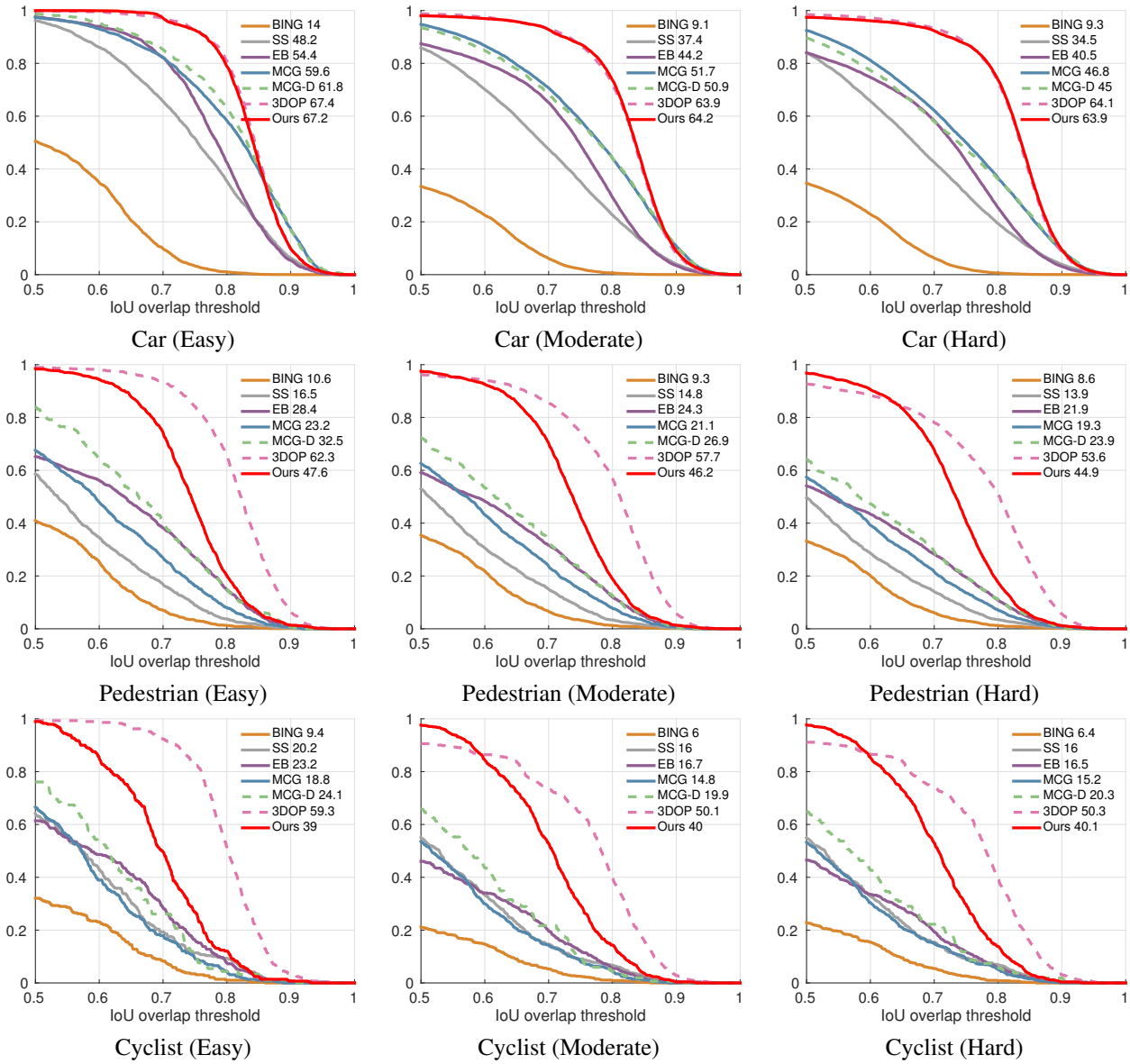



Figure 6: **2D bounding box Recall vs IoU for 2000 proposals**. The number next to the label indicates the average recall (AR).

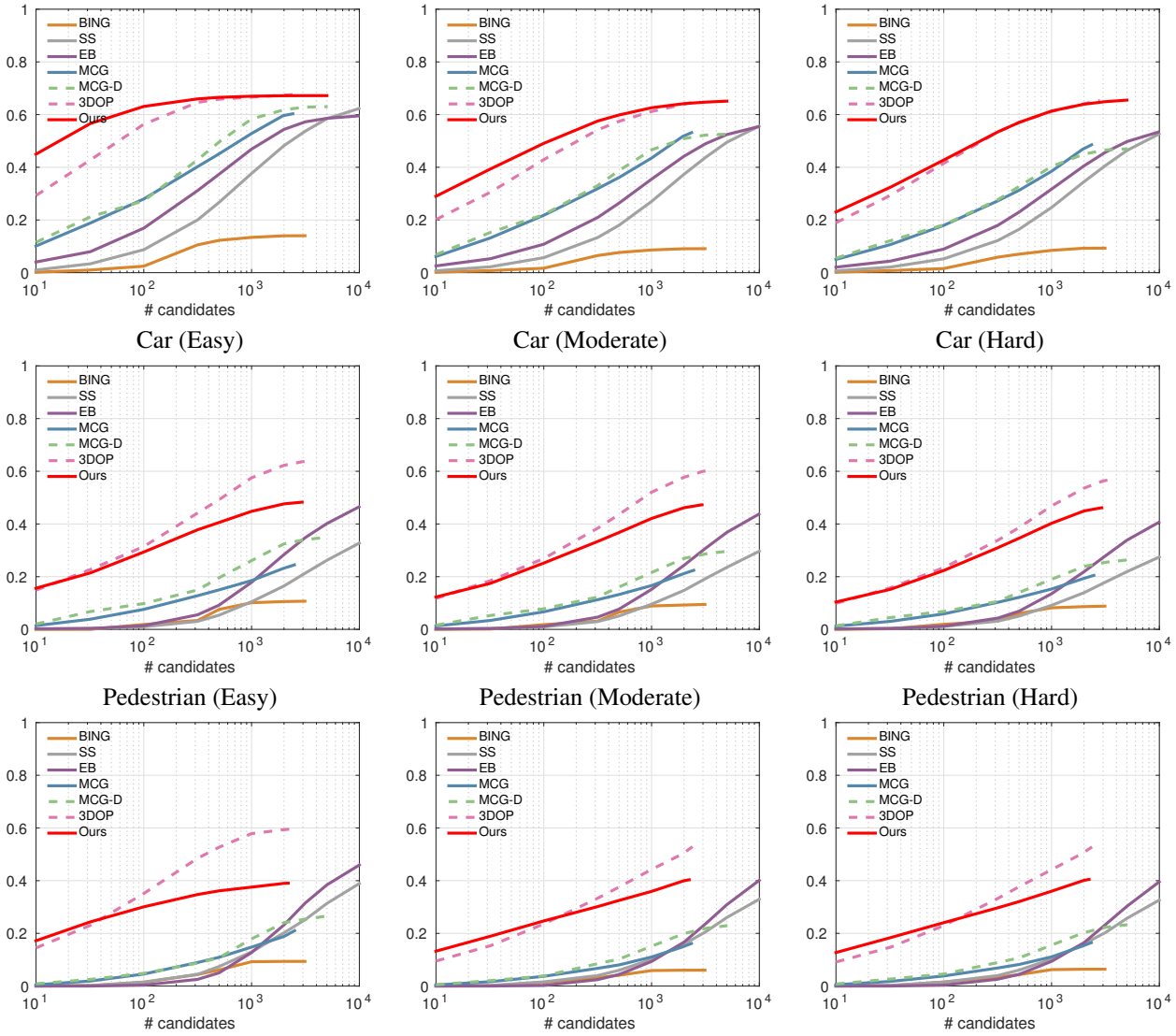



Figure 7: **Average Recall (AR) vs #Candidates for 2D bounding boxes**. Note that the comparison to 3DOP and MCG-D is unfair as we use a monocular image while they exploit depth information.

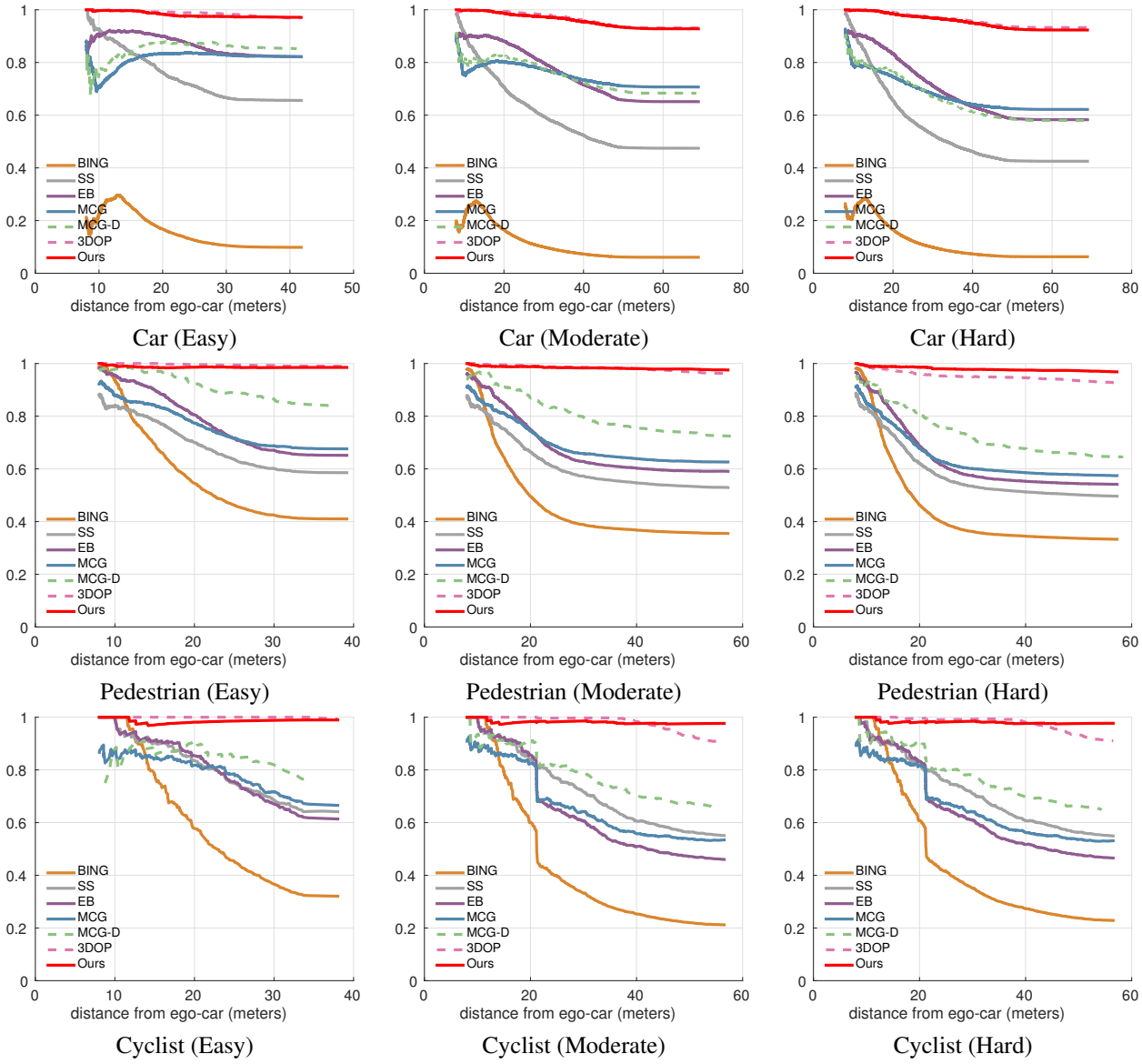



Figure 8: **2D bounding box Recall vs Distance using 2000 proposals**. We use an overlap threshold of 0.7 for Car, and 0.5 for Pedestrian and Cyclist.

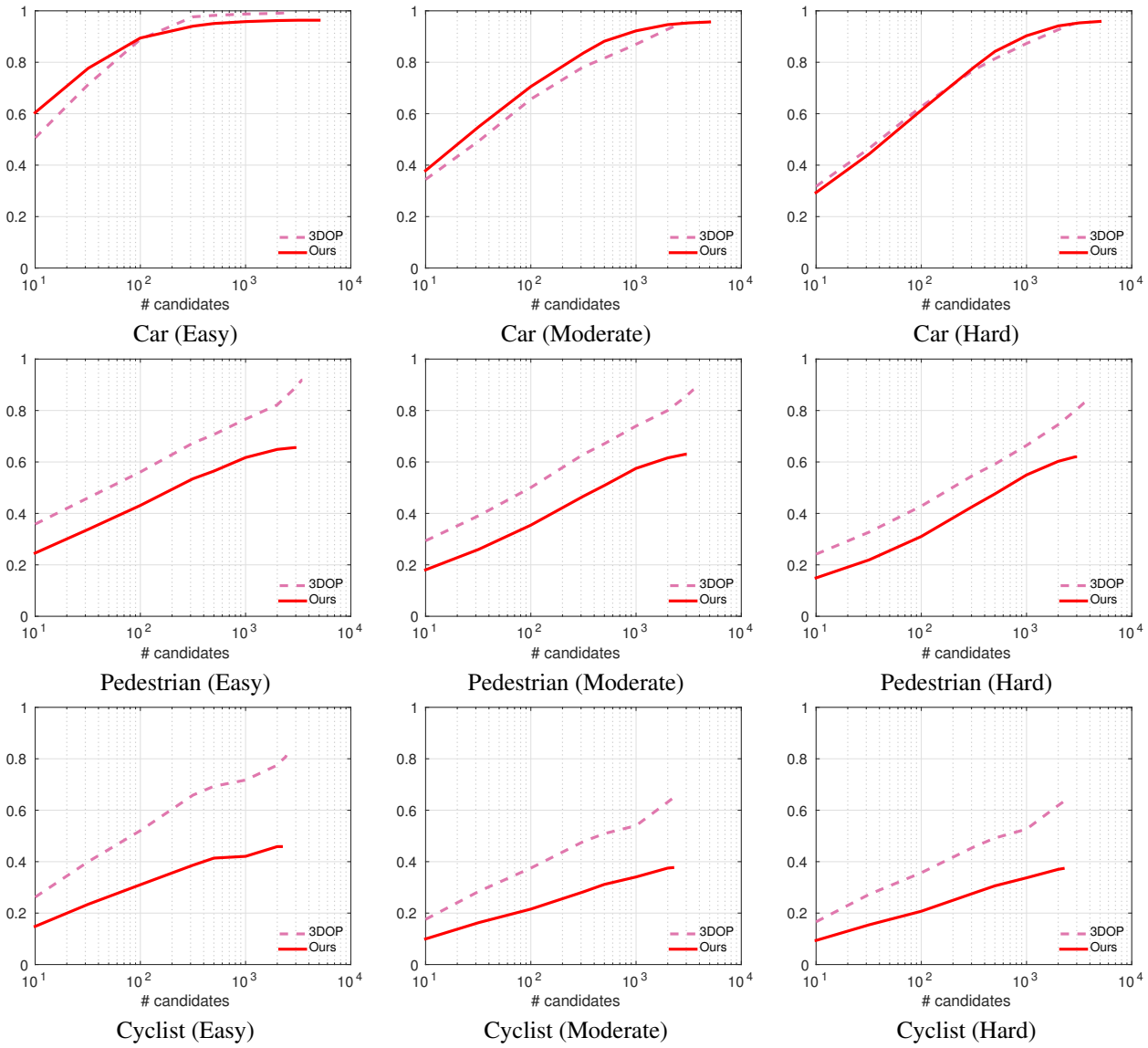
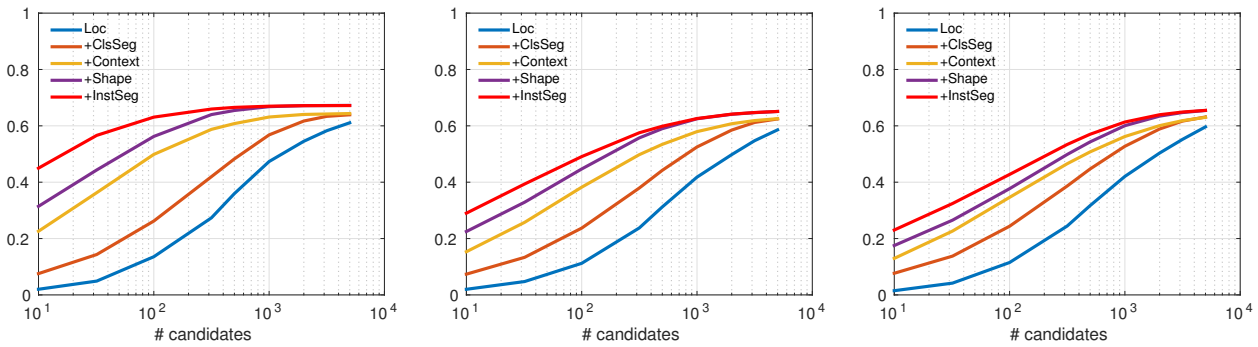
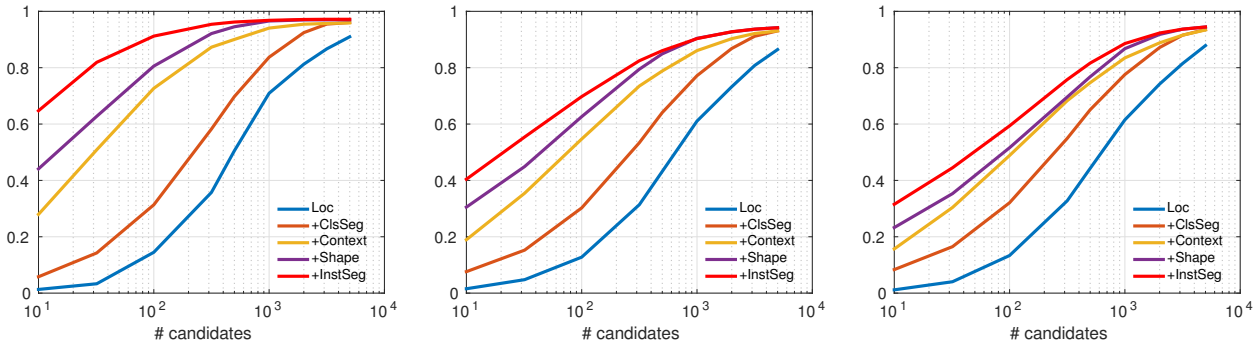



Figure 9: **3D bounding box Recall vs #Candidates at IoU threshold of 0.25**. Note that our monocular approach achieves similar 3D box recall on *Car* with 3DOP, which exploits stereo imagery.



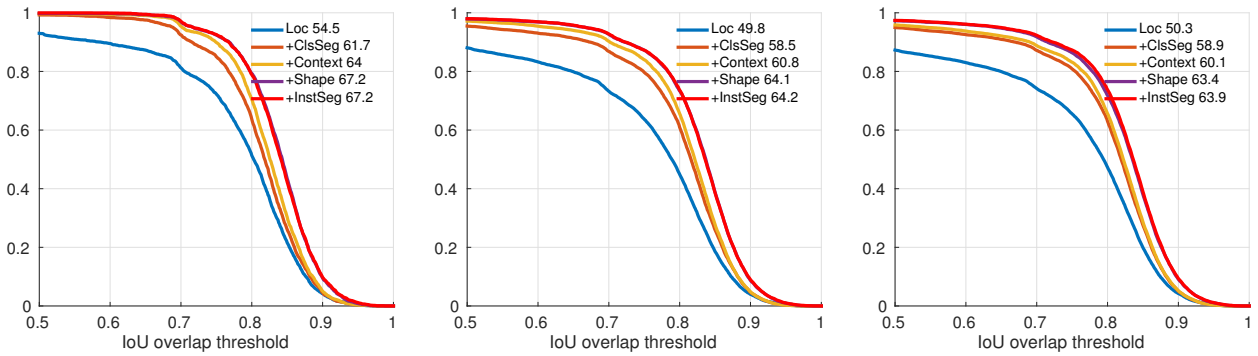

Car (Easy) Car (Moderate) Car (Hard)

Figure 10: **Ablation study of features on car proposals:** Average Recall (AR) vs #Candidates for 2D bounding boxes. The basic model (*Loc*) only uses location prior feature. We then gradually add other types of features: class segmentation, context, shape, and instance segmentation.




Car (Easy) Car (Moderate) Car (Hard)

Figure 11: **Ablation study of features on car proposals:** 2D bounding box Recall vs #Candidates at IoU threshold of 0.7. The basic model (*Loc*) only uses location prior feature. We then gradually add other types of features: class segmentation, context, shape, and instance segmentation.




Car (Easy) Car (Moderate) Car (Hard)

Figure 12: **Ablation study of features on car proposals:** 2D bounding box Recall vs IoU for 2000 proposals. The basic model (*Loc*) only uses location prior feature. We then gradually add other types of features: class segmentation, context, shape, and instance segmentation.

Figure 13: **Qualitative examples of detections results for** *Cars***:** (**left**) top 50 scoring proposals (color from blue to red indicates increasing score), (**middle**) 2D detections, (**right**) 3D detections.

Figure 14: **Qualitative examples of detections results for** *Pedestrians* **and** *Cyclists***:** (**left**) top 50 scoring proposals (color from blue to red indicates increasing score), (**middle**) 2D detections, (**right**) 3D detections.

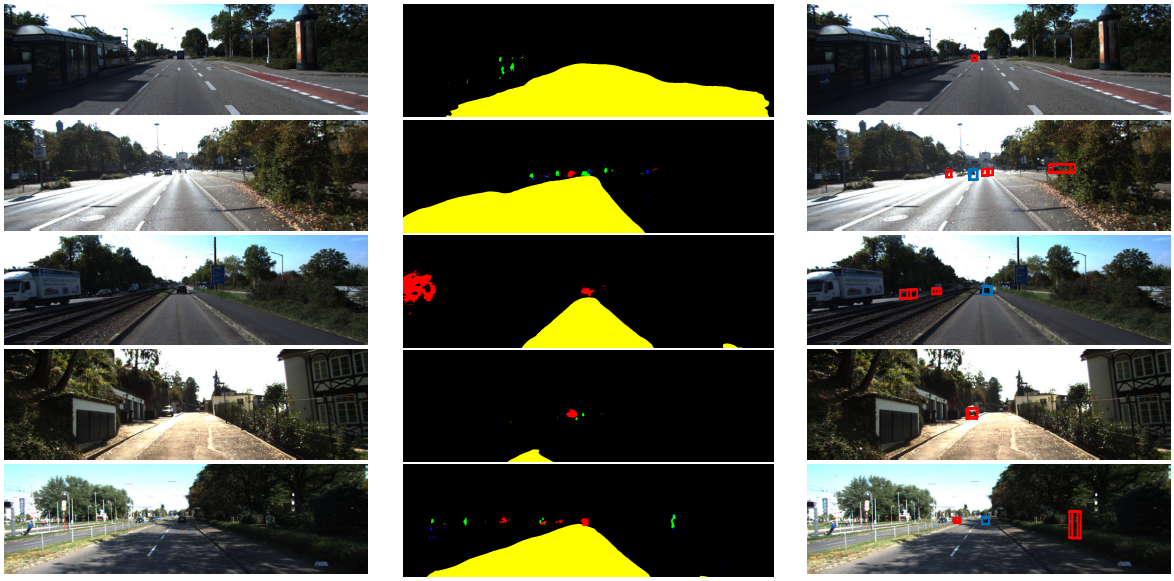


Figure 15: **Qualitative examples of failure cases:** (**left**) input images, (**middle**) semantic segmentation for *Car* (red), *Pedestrian* (green), *Cyclist* (blue), and *Road* (yellow), and (**right**) best 3D proposals among 2K candidates. The correct detections are indicated in blue and the missed detections are in red. Most failure cases are due to class or road segmentation error.