Predicting Semantic Map Representations from Images using Pyramid Occupancy Networks

Supplementary Material

A. Modifications to competing networks

In Section 5 of the main paper, we compare our approach to two recent works from the literature: the Variational Encoder-Decoder of Lu *et al*. [1] and the View Parsing Network of Pan *et al*. [2]. Both works tackle a closely related task to ours, but use other datasets and presume slightly different input dimensions and output map resolutions. In order to compare our work directly, we must therefore make minor architectural changes, which we consider to be the minimum possible for compatibility with our datasets. In the interest of transparency, we detail these changes below:

VED: We modify the bottleneck to use dimensions of 3×6×128 for NuScenes and 4×7×128 for Argoverse to account for the different input aspect ratios. We add an additional decoder layer (identical to previous layers) to increase the resolution from 64×64 to 128×128 and then bilinearly upsample to our output size of 196×200.

VPN: We increase the transformer module bottleneck dimension to 29×50 for NuScenes and 38×50 for Argoverse, and then upsample the output to 196×200 using the authors' existing code.

Since we consider a multilabel prediction setting unlike the single label prediction task addressed in the original works, we train both methods using the balanced cross entropy loss described in Section 3.1.

B. Precision-Recall curves for Argoverse and NuScenes experiments

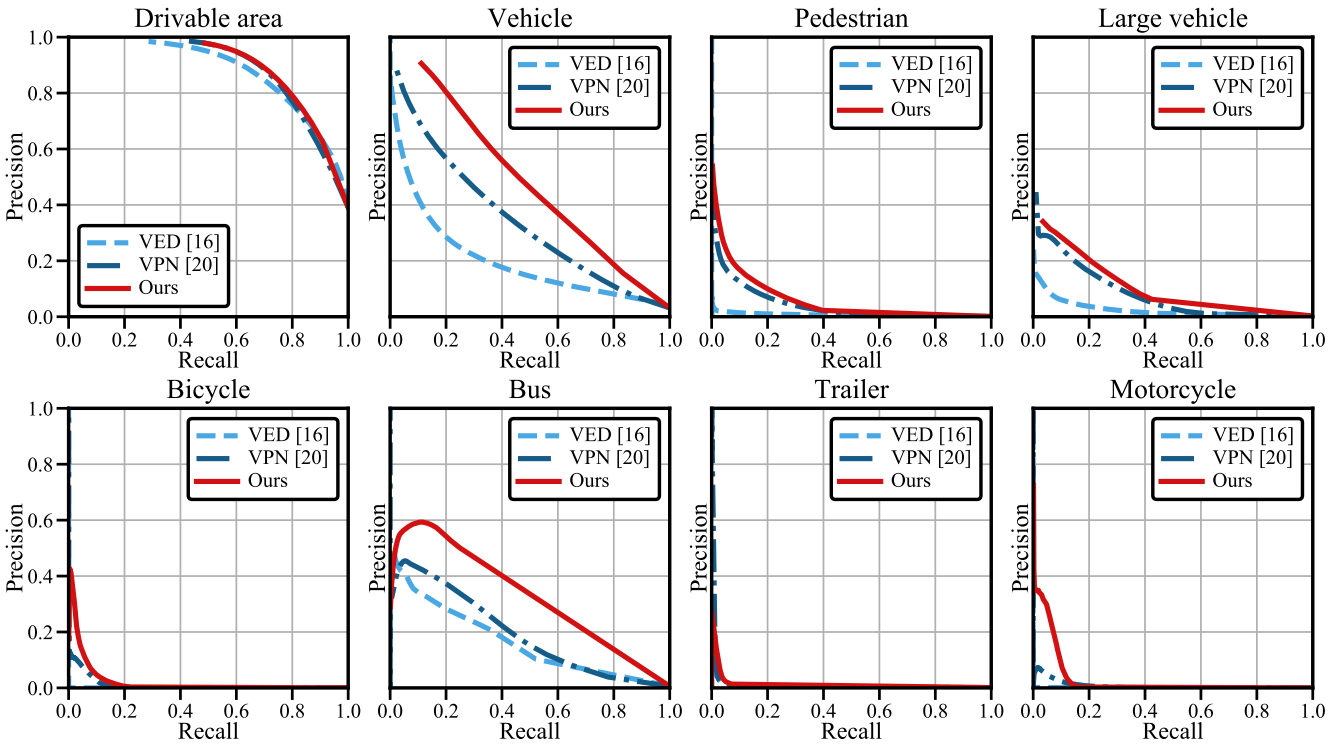

Figure 1. Precision-recall curves for the Argoverse dataset. Best results are curves which occupy the top-right corner of the graph. Our method is more discriminative across all semantic categories, in particular for the *vehicle* and *bus* classes.

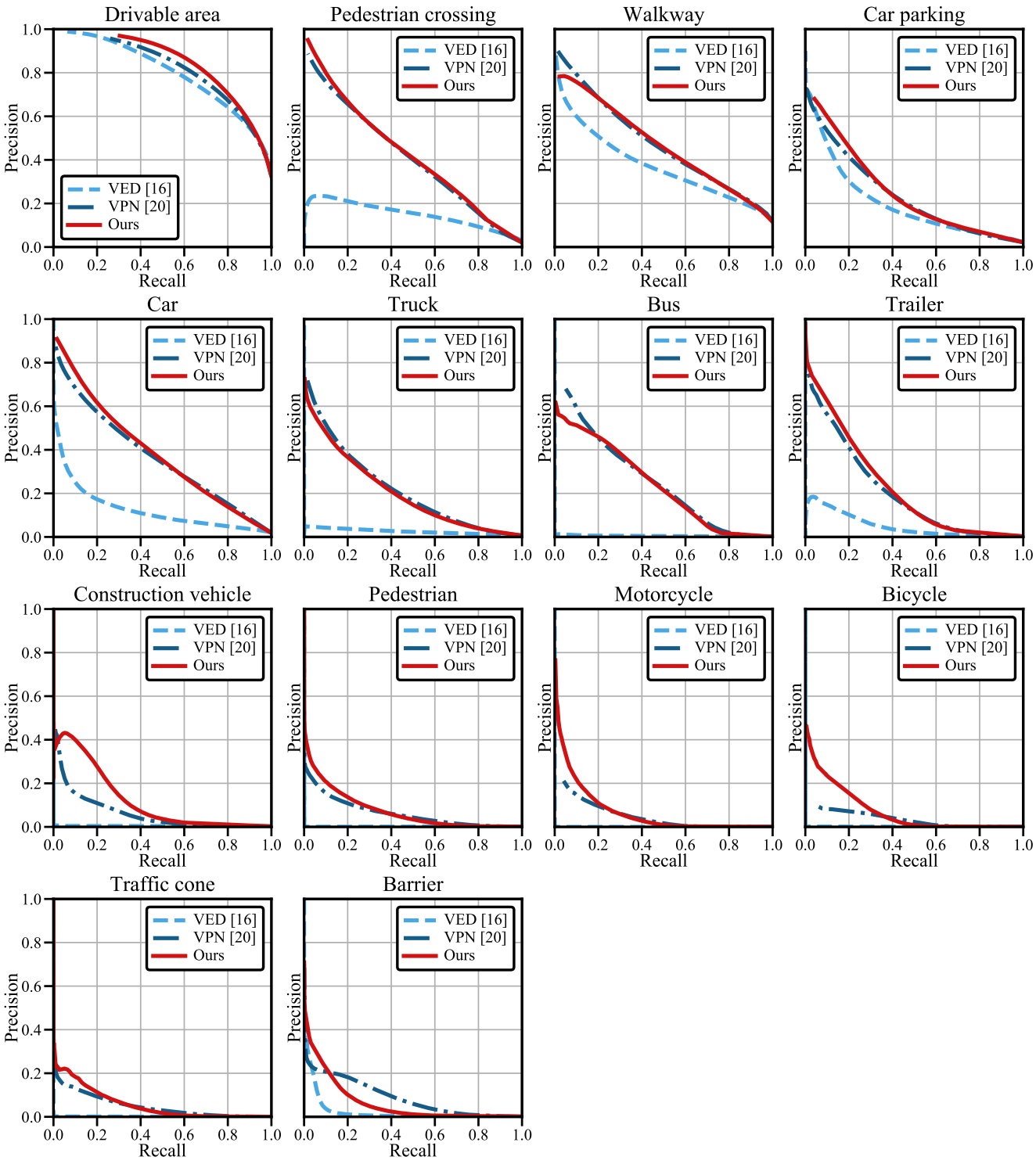

Figure 2. Precision recall curves for the NuScenes dataset.

References

[1] Chenyang Lu, Marinus Jacobus Gerardus van de Molengraft, and Gijs Dubbelman. Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks. *IEEE Robotics and Automation Letters*, 4(2):445–452, 2019. 1

[2] Bowen Pan, Jiankai Sun, Alex Andonian, Aude Oliva, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. *arXiv preprint arXiv:1906.03560*, 2019. 1