Supplementary Material:
"The Pedestrian next to the Lamppost"
Adaptive Object Graphs for Better Instantaneous Mapping

Avishkar Saha[1], Oscar Mendez[1], Chris Russell[2], Richard Bowden[1]
[1]Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, UK
[2]*Amazon*, Tubingen, Germany
{a.saha, o.mendez, r.bowden}@surrey.ac.uk, cmruss@amazon.com

A. Losses

Object Network: Our object network consists of losses applied at different layers within it, as shown in Fig. 1 of the main paper. The initial node embeddings $\{\boldsymbol{v}_i^0 | i \in \mathcal{V}\}$ of input graph $\mathcal{G}$ are supervised for object yaw θ, dimensions $\delta = (l, w)$ and label c. After message-passing across this graph, the updated node embeddings $\{\boldsymbol{v}_i' | i \in \mathcal{V}\}$ are used for the object's centroid $p^v = (x, z)$, while the updated edge embeddings $\{\boldsymbol{e}_{ij}' | (i, i) \in \mathcal{E}\}$ are trained to predict the midpoint of the graph's edges, that is, the midpoint between node i and j: $p^e = (x, z)$.

The object's yaw θ is a single scalar. However, as mentioned in the main paper, it is difficult to regress. Instead, we follow Mousavian *et al.* [2, 3, 5] and predict the object's observation angle β as a vector trained with a discrete-continuous loss. First, the observation angle β is defined as follows:

$$\beta = \alpha + \theta, \tag{1}$$

where α is the viewing angle (the polar angle of the ray). To construct the multi-scalar β encoding, the orientation range $[-\pi, \pi]$ is discretised into n overlapping bins. Within each bin, the network estimates the confidence probability c_i of the observation angle falling within the bin and the residual rotation to the bin center m_i. The residual rotation is represented by the sine and cosine of the offset to the bin center. This results in 3 parameters for each bin i: $(c_i, \sin(\beta_i - m_i), \cos(\beta_i - m_i))$. The confidence probabilities are trained with a cross-entropy loss and the residuals with a Smooth L1 Loss:

$$L_\beta = \frac{1}{|\mathcal{V}|} \sum_{k=1}^{|\mathcal{V}|} \sum_{i=1}^{n} CE(\hat{c}_i, c_i) + c_i * SmoothL1(\hat{a}_i, a_i), \tag{2}$$

where CE is the cross-entropy loss, c_i is the ground truth binary variable of the angle β_i falling within the bin i, and $a_i = (\sin(\beta_i - m_i), \cos(\beta_i - m_i))$. In practice, we find $n = 2$ bins sufficient.

The object's BEV dimensions $\delta = (l, w)$ are regressed directly with a Smooth L1 Loss:

$$L_{dim} = \frac{1}{|\mathcal{V}|} \sum_{k=1}^{|\mathcal{V}|} SmoothL1(\hat{\delta}_k, \delta_k), \tag{3}$$

where δ_k is the object's length and width in meters. The object's label c is supervised with a focal loss [1] for classification:

$$L_c = \frac{1}{|\mathcal{V}|} \sum_{k=1}^{|\mathcal{V}|} -\alpha(1 - p_k)^\gamma \log(p_k), \tag{4}$$

where p_k is the class probability of a predicted object. We follow the hyperparameter settings of the paper, with $\alpha = 0.25$ and $\gamma = 2$.

To predict each object's centroid $p^v = (x, z)$, we regress its viewing angle α and depth z directly using Eq. 3. The x-value of the centroid is recovered using both α and z. We follow the same procedure for the midpoint of each edge $p^e = (x, z)$.

Scene Network: To supervise our scene network, we follow Saha *et al.* [4] and apply a Dice loss to each BEV map M_u^{BEV} generated at scale u. In total, the multi-scale Dice loss across all scales U is defined as:

$$L_{scene} = 1 - \frac{1}{C} \sum_{u=1}^{U} \sum_{c=1}^{C} \frac{2 \sum_i^N \hat{m}_i^c m_i^c}{\sum_i^N \hat{m}_i^c + m_i^c + \epsilon}, \tag{5}$$

where $\hat{m}_i^c$ is the predicted sigmoid output of the network at scale u, m_i^c is the ground truth binary variable at scale u and ϵ is a constant which prevents division by zero.

References

[1] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1

[2] Lijie Liu, Jiwen Lu, Chunjing Xu, Qi Tian, and Jie Zhou. Deep fitting degree scoring network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1057–1066, 2019. 1

[3] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017. 1

[4] Avishkar Saha, Oscar Mendez, Chris Russell, and Richard Bowden. Enabling spatio-temporal aggregation in birds-eye-view vehicle estimation. In *Proceedings of the International Conference on Robotics and Automation*, 2021. 1

[5] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 1