A. Implementation details

VEDet model. We use three different backbones to report performance on NuScenes: ResNet-50 and ResNet-101 [3] are initialized from the ImageNet-pretrained weights hosted on OpenMMLab [2]; VoVNetV2-99 [5] is initialized from the depth-pretrained weights released by [9]. The image features and geometric positional encodings have dimension $C = 256$, and are added element-wise as the keys to the transformer decoder, which has $L = 6$ transformer layers. In the transformer layers, we use multi-head attention with 8 heads, dropout rate 0.1 on the residual connection, and 2048 hidden dimensions in the feed-forward network. To predict the classification scores, we use a single linear projection from 256-dim queries to 10-dim class scores; for predicting the 3D box attributes, we use a 2-layer MLP with $[512, 512]$ hidden dimensions interleaved with ReLU activations. The classification and regression heads are both shared across the 6 transformer layers.

Learnable geometry mapping. For the MLP in the learnable geometry mapping, used to make both geometric positional encoding and object queries, we use 1 hidden layer with 1920 dimensions, followed by a ReLU activation and a final projection to $C = 256$ dimensions. Therefore, given Fourier bands $k = 64$, the dimensions go through the following changes: $d_0 \rightarrow_{\text{Fourier}} 1280 \rightarrow_{\text{hidden}} 1920 \rightarrow_{\text{proj}} 256$, where $d_0 = 10$ for both perspective geometry of an image feature $[\mathbf{r}_{(u_i,v_i)}, \bar{\mathbf{q}}, \mathbf{t}]$ and query geometry $[\mathbf{c}_j^v, \bar{\mathbf{q}}^v, \mathbf{t}^v]$.

Query points. We use 900 learnable 3D query points in all experiments. We follow [10] to use object ranges $[-51.2m, -51.2m, -5.0m, 51.2m, 51.2m, 3.0m]$ in XYZ axes of the global BEV space around the vehicle. The query points are normalized to $[0, 1]$ by a sigmoid operation and scaled by their range. The predictions of box center offsets are added to the points before the sigmoid operation.

Virtual view sampling. During training, the range we use to uniformly sample the translation for the virtual query views is $[-0.6m, -1.0m, -0.3m, 0.6m, 1.0m, 0m]$ in XYZ axes. We uniformly sample the yaw angle to be between $[0, 2\pi]$.

Temporal modeling. In the full-version VEDet we concatenate 2 temporal frames at the token dimension. Following [4, 7], we randomly sample one frame from the past $[3, 27]$ frames during training, and use the past 15-th frame during inference. The time interval between consecutive frames is roughly 0.083s.

Optimization. During training, the loss weights we use are $\lambda_{cls} = 2.0$ and $\lambda_{reg} = 0.25$ following [6, 10]. We use the AdamW optimizer [8] with weight decay 0.01. The learning rate is linearly warmed up in the first 500 iterations from $6.77e^{-5}$ ($\frac{1}{3}$ of initial learning rate) to $2e^{-4}$. The learning rate of the pretrained backbone is multiplied by 0.1 compared to all other components, that are trained from scratch. Checkpointing [1] is adopted during training to save GPU memory, bringing the training time of the full-version VEDet (2 frames, 640×1600 images, $V = 4$) to 36 hours on 8 A100 GPUs, for 24 epochs on NuScenes.

Data augmentation. We use data augmentations following [6], in the order shown below:

- Resize. The original images are resized keeping the aspect ratio. The resize factor is sampled uniformly from $[0.564, 0.8]$ for 384×1056 images, $[0.79, 1.1]$ for 512×1408 images, and $[0.94, 1.25]$ for 640×1600 images.

- Crop. Given a crop size $H \times W$ and an intermediate image size $H' \times W'$ after the resizing, the top area $[0, H' - H]$ is cropped to meet the final height H. The left limit of the cropping box is uniformly sampled from $[0, W' - W]$.

- Horizontal flip. With a 50% probability, we flip all N images at the same time, alongside the 3D box annotations. The camera poses and intrinsics are transformed accordingly to reflect the flipping. Concretely, the X coordinate of the camera translation and yaw angle are flipped, while the principal point in the intrinsic matrix has the X-coordinate flipped.

- Global rotation. Without changing the images, the camera poses and 3D box annotations are rotated around the Z axis of the global BEV space. The angle is uniformly sampled from $[-22.5°, 22.5°]$.

- Global scaling. Without changing the images, the camera poses and 3D box annotations are scaled relative to the origin of the global BEV space. The scaling factor is uniformly sampled from $[0.95, 1.05]$.

During testing, no random augmentations are used. The images are resized to the final width while keeping the aspect ratio, and cropped at the bottom-center.

References

[1] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016. 1

[2] MMCV Contributors. MMCV: OpenMMLab computer vision foundation. https://github.com/open-mmlab/mmcv, 2018. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[4] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 1

[5] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13906–13915, 2020. 1

[6] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. *arXiv preprint arXiv:2203.05625*, 2022. 1

[7] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022. 1

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[9] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3142–3152, 2021. 1

[10] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 1