Supplementary Materials of Object as Query: Lifting any 2D Object Detector to 3D Detection

1. Experiments with More Architectures

We provide more experiments with different 2D detectors and feature extractors in this section. The 2D detector part of MV2D is pretrained on nuImages [1], then the 2D detector part and 3D detector part is jointly trained on nuScenes *train* set. 3D object detection performance and model latency are evaluated on nuScenes *val* set [1]. All the models are trained for 24 epochs without CBGS. For model latency, we only consider the latency of network forward pass and ignore the pre-processing and post-processing time (e.g., image loading and format converting). The latency is evaluated on a single NVIDIA RTX 3090 GPU with batch size 1. We provide the detailed model architectures below.

1.1. Model Architecture

2D detector Without loss of generality, we choose 3 kinds of 2D detectors, including Faster R-CNN [4], a single-stage anchor-based 2D detector RetinaNet [3] and a single-stage anchor-free 2D detector YOLOX [2].

Feature pyramid For models with ResNet-50 backbone and Faster R-CNN detector, the feature pyramid is built to produce feature maps with downsample stride $\{4, 8, 16, 32, 64\}$. For models with ResNet-50 backbone and RetinaNet/YOLOX detector, the feature pyramid is built to produce feature maps with downsample stride $\{8, 16, 32, 64, 128\}$.

Decoder layers The decoder in MV2D contains 6 decoder layers by default. We also experiment with different numbers of decoder layers.

1.2. Performance Comparison

We equip MV2D with different 2D detectors and feature extractors, then evaluate their latency and performance on nuScenes *val* set. The results are listed in Table 1. As demonstrated by the results, under 1408×512 input resolution, MV2D with Faster R-CNN as 2D detector achieves the highest performance of 41.4% mAP and 51.1% NDS with the inference latency of 380ms. When using RetinaNet as 2D detector, the performance drops slightly, obtaining 41.1% mAP and 50.9% NDS. With a faster single-stage detector YOLOX, the inference latency is reduced to 246ms with a decent performance of 41.3% mAP and 50.0% NDS. These results suggest that MV2D can adapt to different 2D detectors and a very lightweight 2D detector can still work. Under a smaller input resolution of 800×320, MV2D with YOLOX achieves 37.7% mAP and 47.0% NDS and reduces the latency to 115ms. These experiments show that MV2D can generalize well to other architectures.

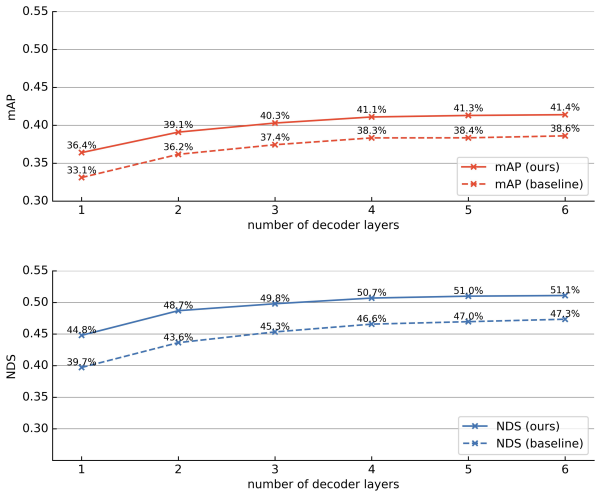

Figure 1. Comparison on different decoder layers.

In Figure 1, we evaluate the performance of different numbers of decoder layers. MV2D is based on Faster R-CNN as 2D detector with 1408×512 input resolution, and the baseline method is based on fixed object queries with the same input resolution. With 1 decoder layer, MV2D achieves 36.4% mAP and 44.8% NDS. With 2 decoder layers, mAP and NDS improve by 2.7% and 3.9% respectively. As the number of decoder layers increases, the mAP and NDS also increase. It can be seen that MV2D with 2 decoder layers outperforms the baseline method with 6 decoder layers on both mAP and NDS.

Resolution	Backbone	2D Detector	Latency	mAP	NDS	mATE	mASE	mAOE	mAVE	mAAE
800×320	ResNet-50	YOLOX	115ms	0.377	0.470	0.737	0.280	0.532	0.427	0.213
1408×512	ResNet-50	YOLOX	246ms	0.413	0.500	0.697	0.271	0.496	0.402	0.203
1408×512	ResNet-50	RetinaNet	368ms	0.411	0.509	0.696	0.272	0.418	0.387	0.185
1408×512	ResNet-50	Faster R-CNN	380ms	0.414	0.511	0.694	0.272	0.427	0.396	0.172

Table 1. Performance comparison on nuScenes *val* set.

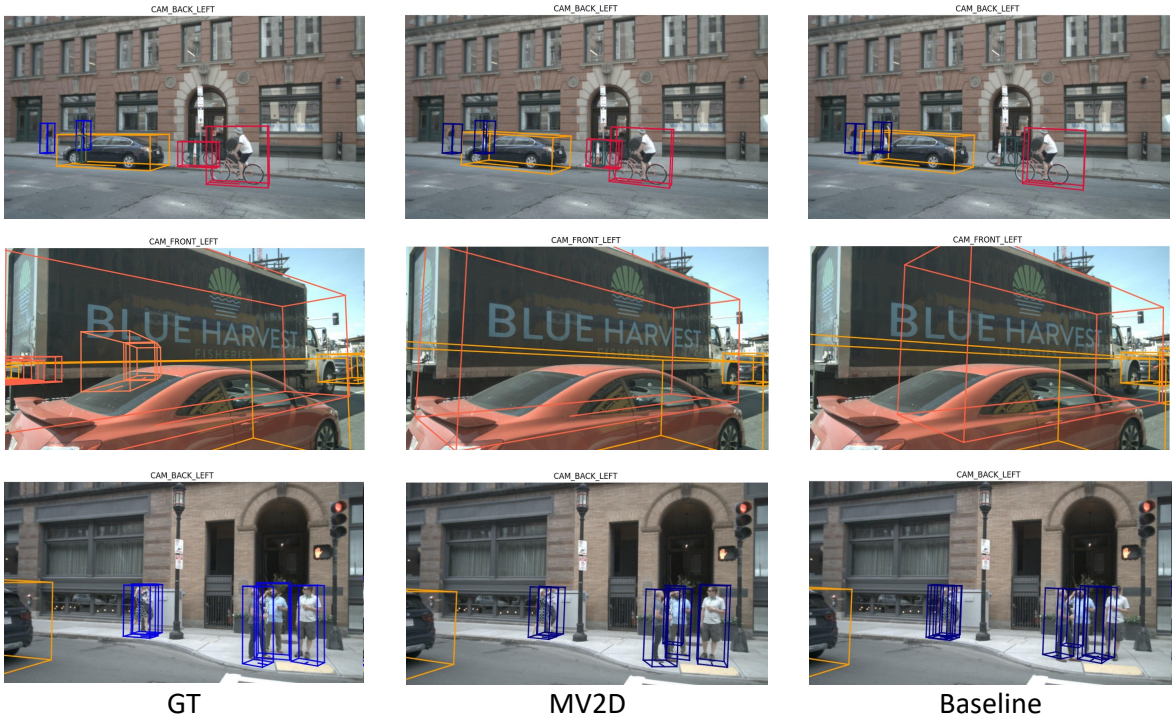

Figure 2. Some qualitative results.

2. More Visualizations

2.1. Qualitative Results

We compare MV2D with the baseline method (using fixed object queries and gathering information from all image regions) and show the qualitative results. The 3D object detection results are illustrated in Figure 2. Row 1 to row 3 is drawn from different data samples. Line 1 to line 3 represent ground truth, the MV2D detection results, and the baseline detection results, respectively. The baseline method might fail to detect (the bicycle in row 1) or mislocate (the truck in row 2 and the persons in row 3) some objects in 3D space. However, most of these objects can be detected by a 2D detector in the image space. Thus 2D detection can provide rich evidence about object existence and location. By exploiting this information, MV2D can generate more accurate 3D detection results.

2.2. Failure Case Analysis

We also analyze the failure case of MV2D. Some examples are shown in Figure 3. From row 1 to row 2, MV2D sometimes splits a "truck" object into a "truck" object and a "trailer" object. From row 3 to row 4, if objects are heavily occluded, the 2D detector might fail to detect them successfully, causing false negatives in MV2D. From row 5 to row 6, if there are extreme lighting conditions or large motion blurs, the 2D detector can also fail to detect some objects and impair the performance of MV2D.

References

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1

[2] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding YOLO series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 1

[3] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 1

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1

Figure 3. Illustration of some failure cases.