Table 1: **Performance of VeXKD on nuScenes for 3D object detection and BEV map segmentation tasks.** "L" and "C" denote the LiDAR and the Camera modality, respectively. "L+C" denotes multi-modal fusion model. "L+C → C", "L → C", and "L+C → L" represent the knowledge distillation from the teacher model to the respective single-modal student. "+" indicates the addition of cross-modal KD methods to the above student models. The FPS results are evaluated on GTX 4090 GPU with batch size of one. "*" denotes our re-implementation results. No test-time augmentation is applied during testing.

Method	Modality	GFLOPs	FPS	nuScenes test		nuScenes val		
				mAP	NDS	mAP	NDS	mIoU
TransFusion	L+C	972	1.8	68.9	71.6	67.5	71.3	–
BEVFusion	L+C	507	2.3	70.2	72.9	68.5	71.4	62.9
TransFusion-L	L	340	5.8	65.5	70.2	65.1	70.1	–
BEVFusion-L	L	322	5.4	–	–	64.7	69.3	48.6
CenterPoint	L	308	11.3	60.3	67.3	57.4	65.6	48.6
+S2M2-SSD	L+C → L	308	11.3	63.6(+3.3)	69.6(+2.3)	–	–	–
+Unidistill	L+C → L	308	11.3	63.9(+3.6)	70.1(+2.9)	59.7	67.5	–
+VeXKD(Ours)	L+C → L	308	11.3	**65.1**(+4.8)	**70.5**(+3.2)	**64.2**	**69.6**	**52.1**
FCOS3d	C	2008	1.7	34.3	41.5	29.5	37.2	–
BEVDet-Tiny	C	370	6.3	–	–	33.3	41.0	56.8*
BEVDet-R50	C	184	14.0	28.9	38.4	28.6	37.2	56.4*
+Unidistill	L+C → C	184	14.0	29.6(+0.7)	39.3(+0.9)	–	–	–
+VeXKD(Ours)	L+C → C	184	14.0	**35.8**(+6.9)	**42.6**(+4.2)	**34.7**	**40.6**	**60.7**
BEVFormer-S	C	1152	2.6	40.9	46.2	37.5	44.8	61.8*
+Unidistill	L+C → C	1152	2.6	–	–	37.7*	45.5*	–
+BEVDistill	L → C	1152	2.6	–	–	38.6	45.7	–
+VeXKD(Ours)	L+C → C	1152	2.6	**42.5**(+1.6)	**48.3**(+2.1)	**41.2**	**47.7**	**64.2**

Table 2: Supplementary Experiment. Performance of More Camera Students on the NuScenes val set.

Method	Modality	Network Component	GFLOPs	FPS	mAP	NDS
BEVFormer	C	ResNet-101	1311	1.9	41.6	51.7
+ VeXKD(Ours)	L+C → C	ResNet-101	1311	1.9	**43.8**	**53.7**
BEVDepth-R50	C	ResNet-50	662	7.3	35.1	47.5
+ VeXKD(Ours)	L+C → C	ResNet-50	662	7.3	**38.2**	**49.7**