QCRAFT   UNIVERSITY OF CALIFORNIA SANTA CRUZ

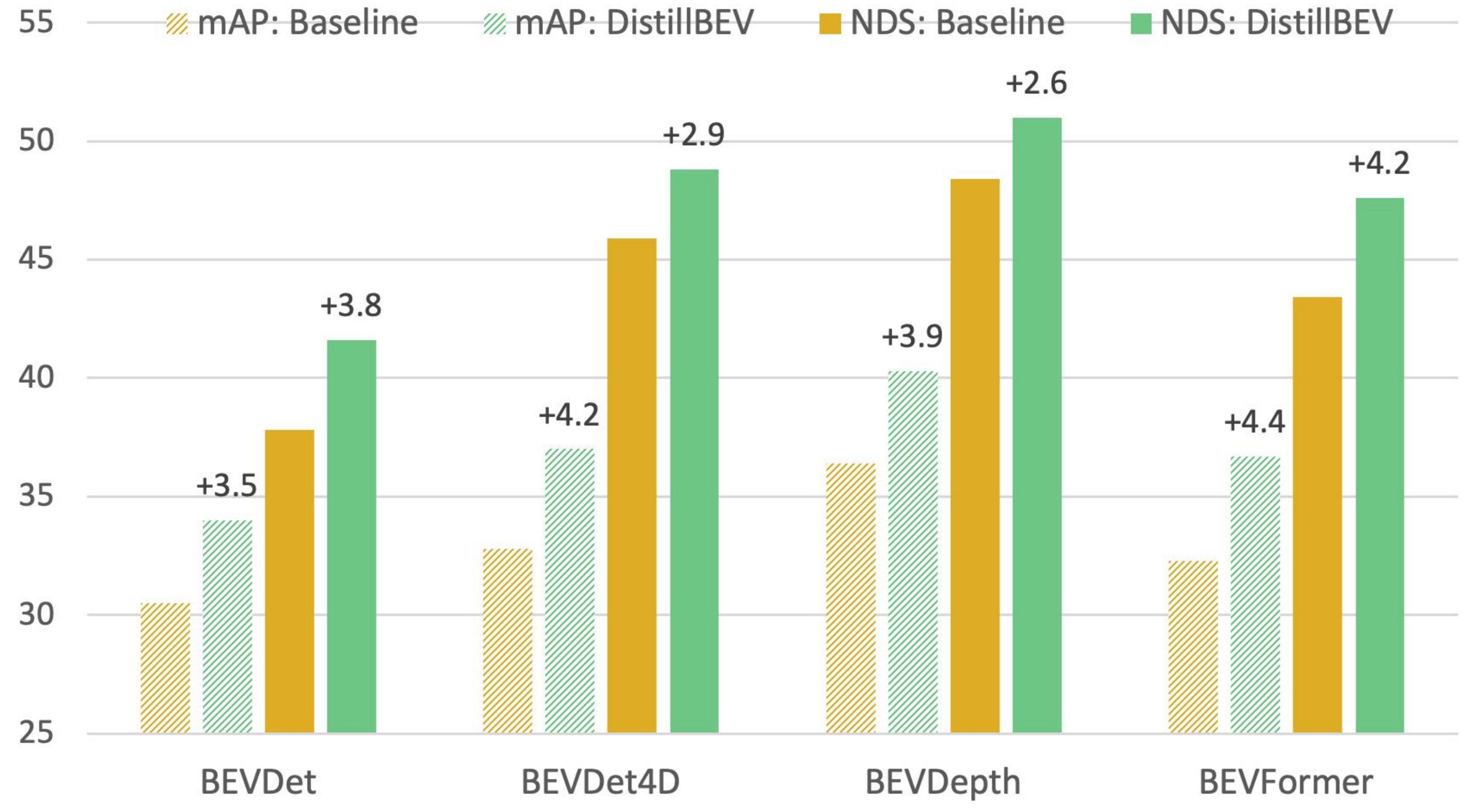
DistillBEV: Boosting Multi-Camera 3D Object Detection with Cross-Modal Knowledge Distillation

Zeyu Wang*   Dingwen Li*   Chenxu Luo   Cihang Xie   Xiaodong Yang

ICCV23 PARIS

Introduction

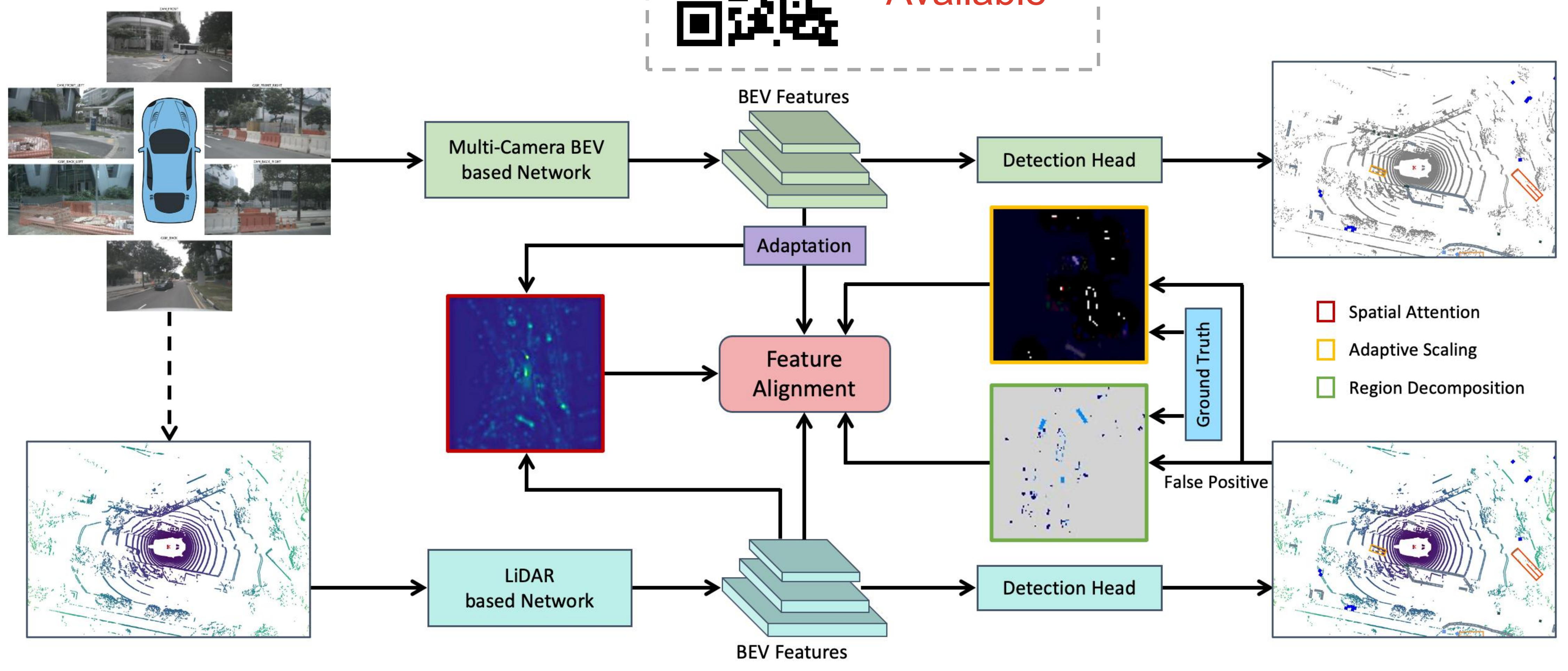
- A performance gap remains between camera BEV detectors and Lidar detectors.
- DistillBEV makes the distillation focused and balanced.
- Our approach is generalizable to various combinations of teacher and student models.



Enabled by DistillBEV, a variety of multi-camera BEV based detectors achieve significant performance boost.

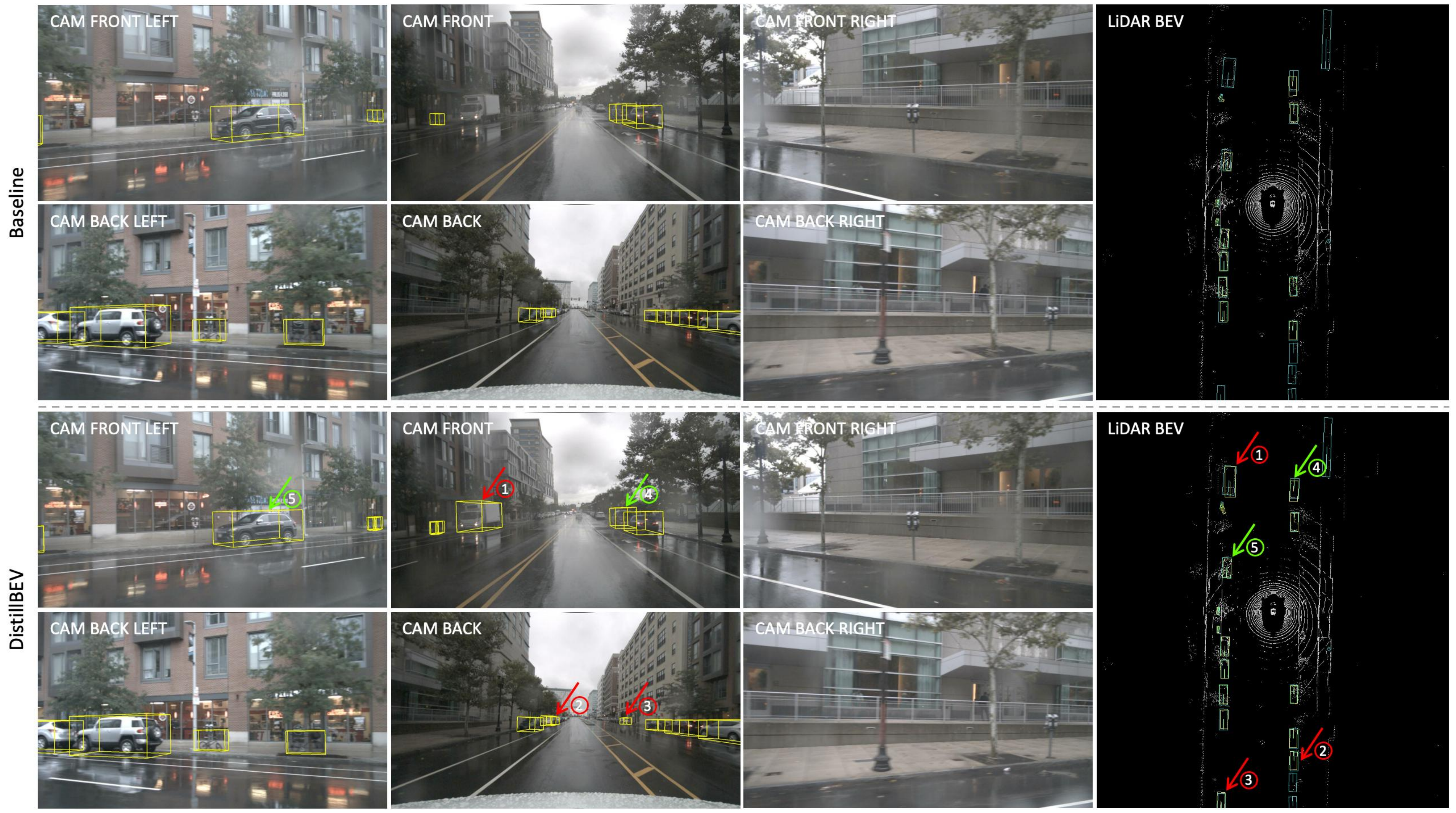
Method

Code Available

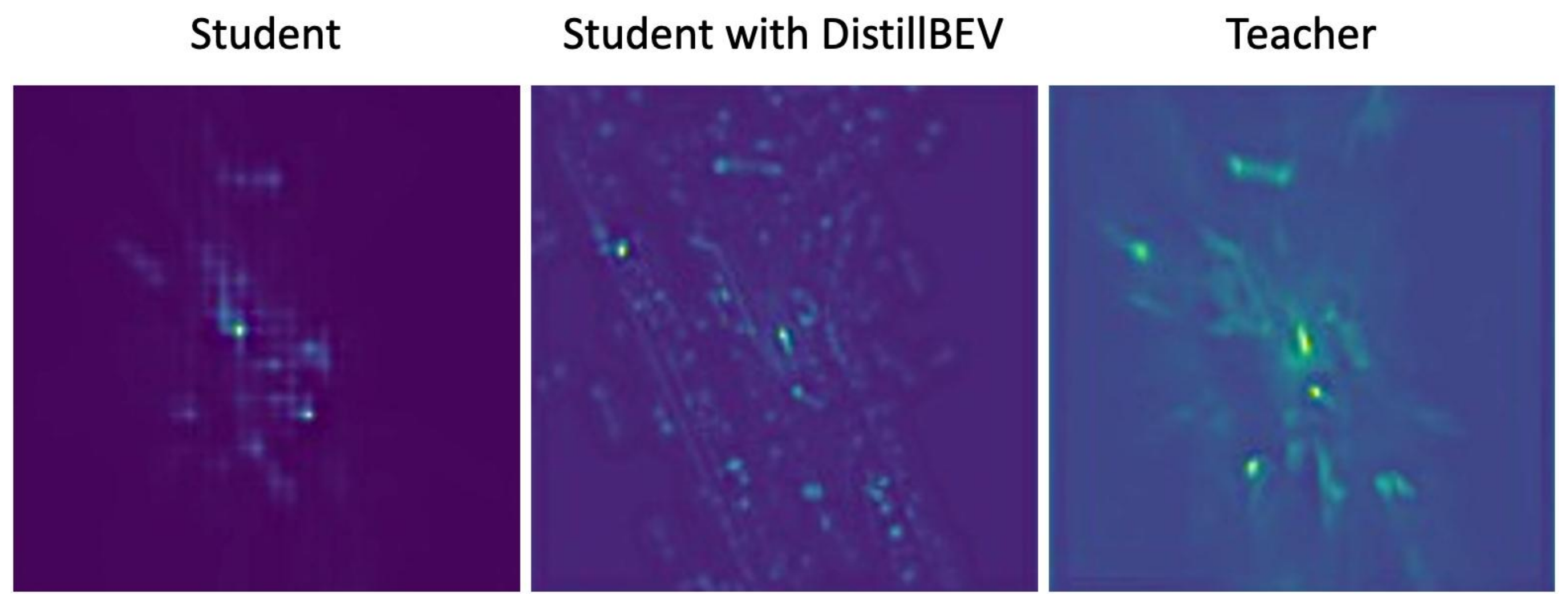


Teacher (bottom branch) guides student (top branch) by BEV feature imitating.

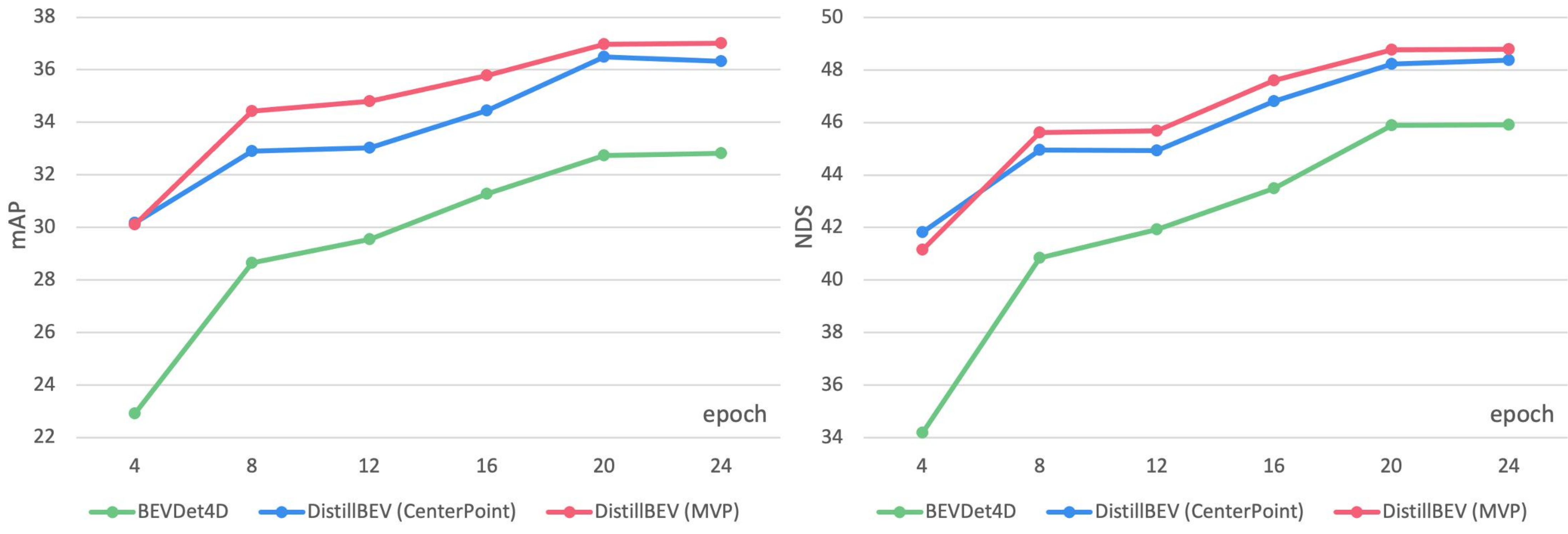
Qualitative Results



Comparison of the baseline and its distilled version. The cyan and yellow boxes denote the ground truth and detection results. Red arrow: missed by baseline; green arrow: more accurate depth by DistillBEV



Guided by DistillBEV, the attention map of student becomes shaper and more similar to the teacher's.



Comparison of the training process between the student model (BEVDet4D) and the distilled versions using CenterPoint and MVP as the teacher models.

Quantitative Results

Teacher	Mode	Student	Mode	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
-	-	BEVDet	C	30.5	37.8	72.1	27.9	57.9	91.4	25.0
CenterPoint	L	BEVDet	C	32.7	40.7	70.9	26.5	56.5	81.2	21.0
MVP	L&C	BEVDet	C	**34.0**	**41.6**	70.4	26.6	55.6	81.5	20.1
-	-	BEVDet4D	C	32.8	45.9	69.5	27.9	50.8	36.5	20.6
CenterPoint	L	BEVDet4D	C	36.3	48.4	66.6	26.8	49.8	34.9	19.9
MVP	L&C	BEVDet4D	C	**37.0**	**48.8**	67.6	26.8	46.1	36.8	20.0
-	-	BEVDepth	C	36.4	48.4	64.9	27.3	49.8	34.9	20.7
CenterPoint	L	BEVDepth	C	38.9	49.8	63.0	26.7	50.4	36.0	20.2
MVP	L&C	BEVDepth	C	**40.3**	**51.0**	62.3	26.6	46.4	35.7	20.7
-	-	BEVFormer	C	32.3	43.4	79.6	28.3	53.1	46.0	21.4
CenterPoint	L	BEVFormer	C	35.5	46.8	71.9	27.7	50.8	39.3	20.0
MVP	L&C	BEVFormer	C	**36.7**	**47.6**	72.1	27.5	50.6	37.6	20.0

Comparison of our approach using various combinations of teacher and student models on the validation set of nuScenes.

Method	Backbone	Mode	mAP	NDS
CenterPoint [42]	-	L	56.4	64.8
MVP [43]	-	L&C	67.1	78.0
FCOS3D [36]	R101	C	34.3	41.5
PETR [23]	R101	C	35.7	42.1
DETR3D [38]	R101	C	34.6	42.5
BEVFormer [21]	R50	C	32.3	43.4
BEVFormer [21]	R101	C	41.6	51.7
BEVDepth [20]	R50	C	35.1	47.5
BEVDepth [20]	R101	C	41.2	53.5
Set2Set [39]	R50	C	37.5	47.9
FitNet [34]	R50	C	37.3	48.0
MonoDistill [6]	R50	C	39.0	49.5
UVTR [19]	R50	C	39.4	50.1
BEVDistill [5]	R50	C	40.7	51.5
BEVDistill [5]	R101	C	41.7	52.4
Ours (BEVFormer)	R50	C	36.7	47.6
Ours (BEVFormer)	R101	C	44.6	54.5
Ours (BEVDepth)	R50	C	40.3	51.0
Ours (BEVDepth)	R101	C	**45.0**	**54.7**

Comparison on the validation set of nuScenes.

Method	Backbone	Mode	mAP	NDS
CenterPoint [42]	-	L	58.0	65.5
MVP [43]	-	L&C	66.4	70.5
FCOS3D [36]	R101	C	35.8	42.8
BEVDet [12]	Swin-B	C	39.8	46.3
DD3D [30]	VoV-99	C	41.8	47.7
DETR3D [38]	VoV-99	C	41.2	47.9
PETR [23]	VoV-99	C	44.1	50.4
BEVDet4D [12]	Swin-B	C	45.1	56.9
BEVFormer [21]	VoV-99	C	48.1	56.9
BEVDistill [5]	ConvNeXt-B	C	49.8	59.4
BEVDepth [20]	VoV-99	C	50.3	60.0
Ours (BEVDepth)	Swin-B	C	**52.5**	**61.2**

Comparison on the test set of nuScenes.

B0	B1	B2	H	mAP	NDS
				32.8	45.9
			✓	34.9	46.2
		✓	✓	36.1	47.3
	✓	✓	✓	36.8	48.4
✓	✓	✓	✓	34.4	46.6

Evaluation of distillation effects at different layers. B2-B0: three preceding intermediate layers of BEV encoders, H: pre-head layer.

Method	FG&BG	Attention	Scaling	FP	mAP	NDS
Baseline					32.8	45.9
Distill					32.9	45.6
	✓				33.4	46.3
	✓	✓			34.6	46.5
	✓	✓	✓		36.8	48.4
	✓	✓	✓	✓	37.8	49.0

Comparison of using different combinations of region decomposition (FG&BG and FP), spatial attention and adaptive scaling.