CRKD: Enhanced Camera-Radar Object Detection with Cross-modality Knowledge Distillation

Lingjun Zhao , Jingyu Song*†, Katherine A. Skinner University of Michigan, Ann Arbor, MI USA*

- Problem/Objective
  - Object Detection in BEV

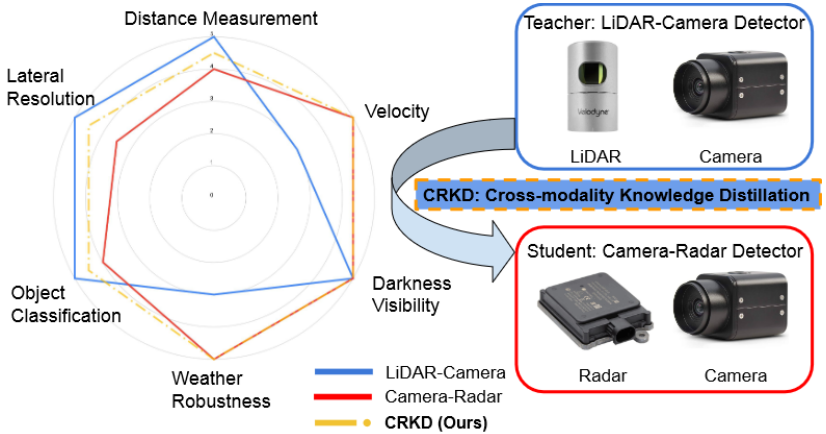
- Contribution/Key Idea
  - First cross-modality KD framework(LC→CR distillation)
  - Novel Loss design
  - Improve the mAP and NDS of student detectors by 3.5% and 3.2% respectively

CRKD: Enhanced Camera-Radar Object Detection with Cross-modality Knowledge Distillation

Lingjun Zhao , Jingyu Song*†, Katherine A. Skinner University of Michigan, Ann Arbor, MI USA*

- Best Performing → LC 조합
  - LiDAR 비싸다
  - 반면, CR은 대부분의 차량에 장착돼있음

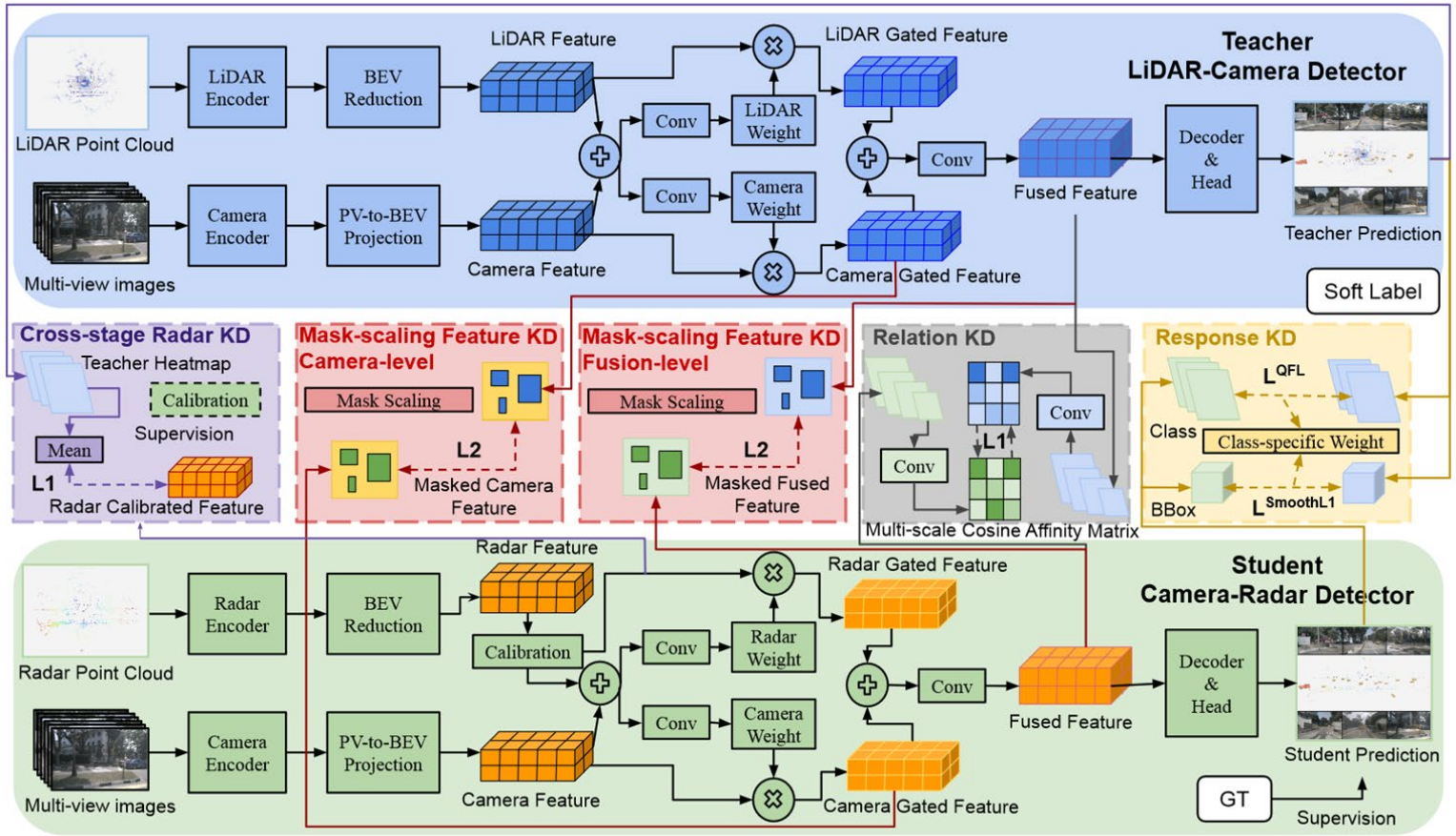
- 기존 모델들
  - L or LC → L/C only
  - Radar 모달리티에 집중한 적이 없음



CRKD: Enhanced Camera-Radar Object Detection with Cross-modality Knowledge Distillation

Lingjun Zhao , Jingyu Song*†, Katherine A. Skinner University of Michigan, Ann Arbor, MI USA*

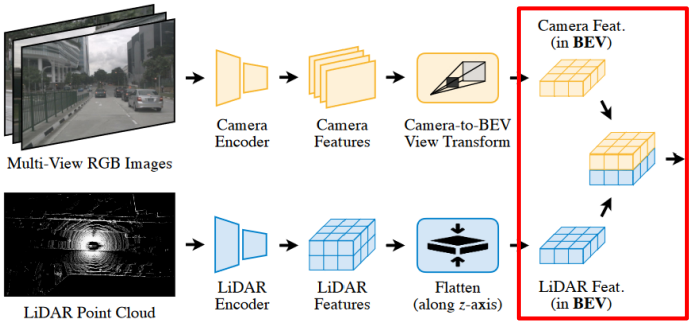
CVPR 2024

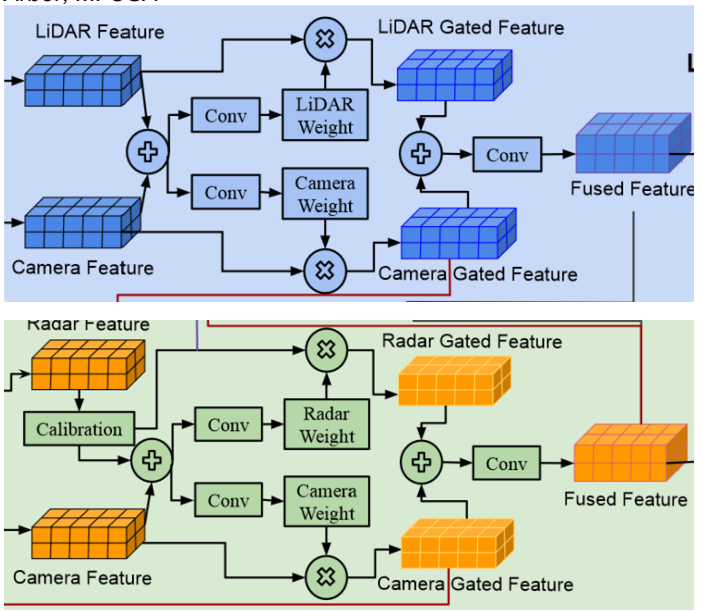


CRKD: Enhanced Camera-Radar Object Detection with Cross-modality Knowledge Distillation

Lingjun Zhao , Jingyu Song*†, Katherine A. Skinner University of Michigan, Ann Arbor, MI USA*

1. Gated Fusion





- fusion 전에 더 각각의 특성에 attention 된 modality feature map을 만들 수 있음
- 각각의 gated feature을 구한 후 fuser

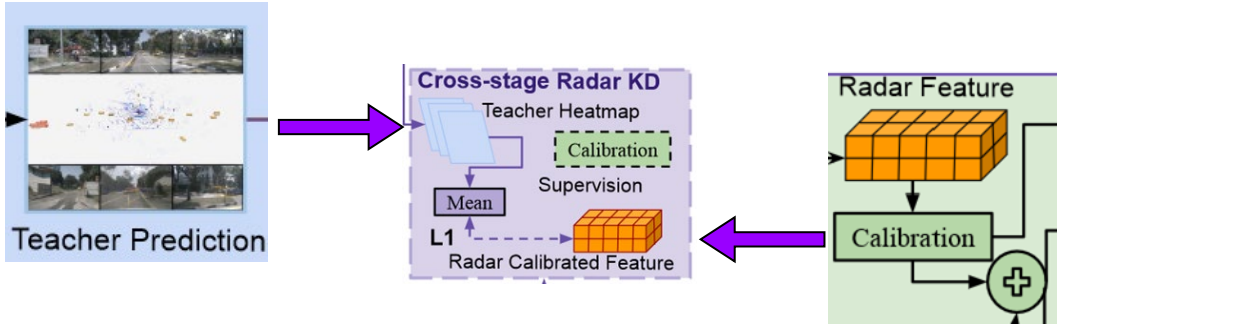
$$\tilde{F}_{M_1} = F_{M_1} \times \sigma(\mathrm{Conv}_{M_1}(\mathrm{Concat}(F_{M_1}, F_{M_2}))),$$

$$\tilde{F}_{M_2} = F_{M_2} \times \sigma(\mathrm{Conv}_{M_2}(\mathrm{Concat}(F_{M_1}, F_{M_2}))),$$

CRKD: Enhanced Camera-Radar Object Detection with Cross-modality Knowledge Distillation

Lingjun Zhao , Jingyu Song*†, Katherine A. Skinner University of Michigan, Ann Arbor, MI USA*

2. Cross-Stage Radar Distillation (CSRD)



compensate noise by calibration

- LiDAR / Radar는 둘다 point cloud
  - Radar는 sparse / scene-level info (object-level info with V measurement)
  - LiDAR는 denser / geometry-level info

- Radar feature map ←→ scene-level objectness heatmap

$$\mathcal{L}_{csrd} = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} \| \hat{Y}_{i,j}^{T} - \hat{F}_{r}^{S}{}_{i,j} \|_1,$$

3. Mask-Scaling Feature Distillation (MSFD)

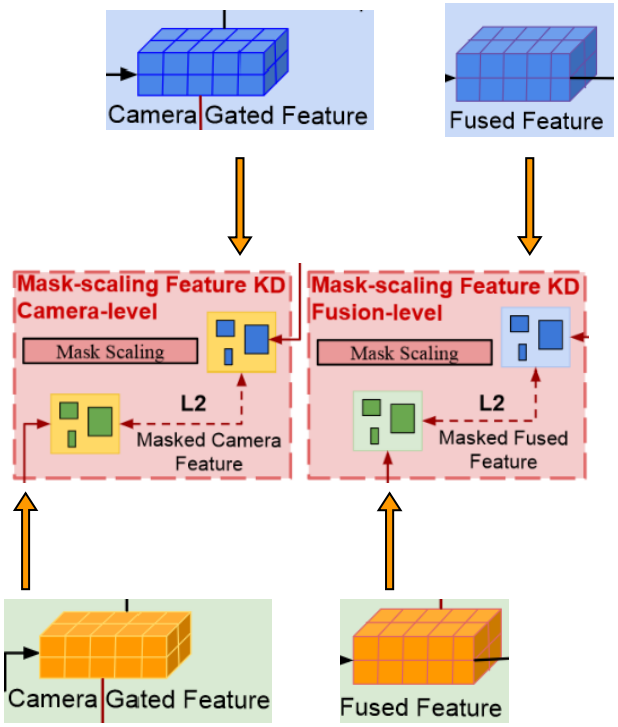
- Feature distillation
  - Feature distill 바로 적용은 foreground와 background의 imbalance

  ⟶ foreground의 정보만 추출하기 위해 Mask 사용

- BEV 상에서는 예측이 어려움(오차가 큼)
  - 거리/속도에 따라 다른 Mask 크기를 키워서 적용

count for the potential misalignment. We increase the width and length of the mask by α and β if the objects are in the range groups $[r_1, r_2]$ and $[r_2, \infty]$. We also increase the width and length by α and β if the velocities along that axis are within $[v_1, v_2]$ and $[v_2, \infty]$. In practice, We clip the in-

$$\mathcal{L}_{msfd} = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} M_{i,j} \| F_{i,j}^{T} - F_{i,j}^{S} \|_2,$$

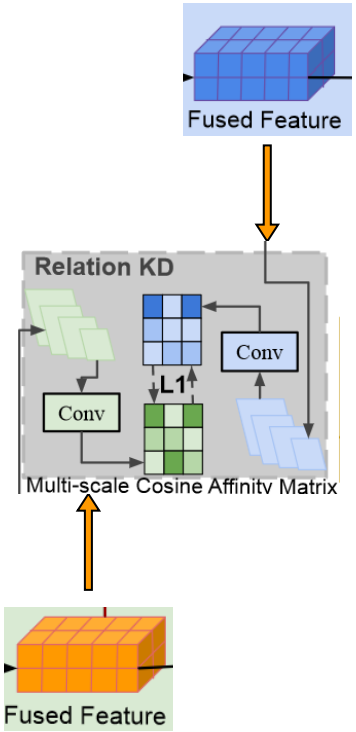


4. Relation Distillation (ReID)

- Affinity Matrix를 적용
  - cosine similarity 비교
  - Monodistill 아이디어 차용

$$C_{i,j} = \frac{F_i^\top F_j}{\|F_i\|_2 \cdot \|F_j\|_2},$$

$$\mathcal{L}_{reld} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \|C_{i,j}^T - C_{i,j}^S\|_1,$$



CRKD: Enhanced Camera-Radar Object Detection with Cross-modality Knowledge Distillation

Lingjun Zhao , Jingyu Song*†, Katherine A. Skinner University of Michigan, Ann Arbor, MI USA*

5. Response Distillation (RespD)

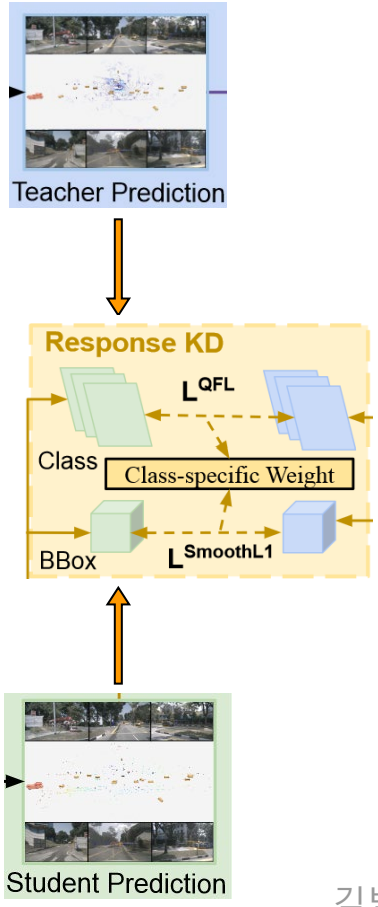
- Response distillation은 이미 3D OD에서 많이 쓰임
  - Radar는 도플러 효과로 속도 측정에 유리
    → dynamic object에 더 큰 가중치

$$\mathcal{L}_{cls} = \sum_{i=1}^{K} QFL(P_{C_i}^T, P_{C_i}^S) \times w_i,$$

$$+$$

$$\mathcal{L}_{reg} = \sum_{i=1}^{K} Smooth\mathcal{L}1(P_{B_i}^T, P_{B_i}^S) \times w_i,$$

K: # of tasks in Centerhead
Wi: class-specific weight



CRKD: Enhanced Camera-Radar Object Detection with Cross-modality Knowledge Distillation

Lingjun Zhao , Jingyu Song*†, Katherine A. Skinner University of Michigan, Ann Arbor, MI USA*

- **Experiments**

Set	Method	Modality	Backbone	Resolution	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
val	BEVFormer-S [35]	C	R101	900 × 1600	37.5	44.8	0.725	0.272	**0.391**	0.802	0.200
	BEVDet [17]	C	R50	256 × 704	29.8	37.9	0.725	0.279	0.589	0.860	0.245
	RCM-Fusion [22]	C+R	R101	900 × 1600	44.3	52.9	-	-	-	-	-
	CenterFusion [44]	C+R	DLA34	450 × 800	33.2	45.3	0.649	**0.263**	0.535	0.540	**0.142**
	CRAFT [24]	C+R	DLA34	448 × 800	41.1	51.7	0.494	0.276	0.454	0.486	0.176
	RCBEV [73]	C+R	SwinT	256 × 704	37.7	48.2	0.534	0.271	0.558	0.493	0.209
	BEVFusion [39]	C+R	SwinT	256 × 704	43.2	54.1	0.489	0.269	0.512	**0.313**	0.171
	UVTR (L2C) [31]	C◇	R101	900 × 1600	37.2	45.0	0.735	0.269	0.397	0.761	0.193
	BEVDistill (BEVFormer-S) [5]	C◇	R50	640 × 1600	38.6	45.7	0.693	0.264	0.399	0.802	0.199
	UniDistill (LC2C) [72]	C◇	R50	256 × 704	26.5	37.8	-	-	-	-	-
	BEVSimDet [71]	C◇	SwinT	256 × 704	40.4	45.3	0.526	0.275	0.607	0.805	0.273
	X3KD (LC2C) [26]	C◇	R50	256 × 704	39.0	50.5	0.615	0.269	0.471	0.345	0.203
	DistillBEV (BEVDet) [55]	C◇	R50	256 × 704	34.0	41.6	0.704	0.266	0.556	0.815	0.201
	X3KD (L2CR) [26]	C+R◇	R50	256 × 704	42.3	53.8	-	-	-	-	-
	CRKD	C+R◇	R50	256 × 704	43.2	54.9	0.450	0.267	0.442	0.339	0.176
	CRKD	C+R◇	SwinT	256 × 704	**46.7**	**57.3**	**0.446**	**0.263**	0.408	0.331	0.162
test	BEVFormer-S [35]	C	R101	900 × 1600	40.9	46.2	0.650	0.261	0.439	0.925	0.147
	BEVDet† [17]	C	SwinT	640 × 1600	42.4	48.2	0.528	**0.236**	0.395	0.979	0.152
	RCM-Fusion [22]	C+R	R101	900 × 1600	**49.3**	58.0	0.485	0.255	0.386	0.421	0.115
	CenterFusion† [44]	C+R	DLA34	450 × 800	32.6	44.9	0.631	0.261	0.516	0.614	0.115
	CRAFT† [24]	C+R	DLA34	448 × 800	41.1	52.3	0.467	0.268	0.456	0.519	0.114
	RCBEV [73]	C+R	SwinT	256 × 704	40.6	48.6	0.484	0.257	0.587	0.702	0.140
	UVTR (L2C) [31]	C◇	V2-99	900 × 1600	45.2	52.2	0.612	0.256	**0.385**	0.664	0.125
	X3KD (LC2C) [26]	C◇	R101	640 × 1600	45.6	56.1	0.506	0.253	0.414	**0.366**	0.131
	UniDistill (LC2C) [72]	C◇	R50	256 × 704	29.6	39.3	0.637	0.257	0.492	1.084	0.167
	X3KD (L2CR) [26]	C+R◇	R50	256 × 704	44.1	55.3	-	-	-	-	-
	CRKD	C+R◇	SwinT	256 × 704	48.7	**58.7**	**0.404**	0.253	0.425	0.376	**0.111**

CRKD: Enhanced Camera-Radar Object Detection with Cross-modality Knowledge Distillation

Lingjun Zhao , Jingyu Song*†, Katherine A. Skinner University of Michigan, Ann Arbor, MI USA*

- **Experiments**

Model	Modality	Car	Truck	Bus	Trailer	CV	Ped	Motor	Bicycle	TC	Barrier	mAP↑
Teacher	L+C	88.4	62.4	73.8	40.6	29.2	78.7	75.3	65.8	74.9	72.3	66.1
Baseline	C+R	72.1	37.8	48.9	18.3	12.6	48.4	42.0	33.8	58.8	59.6	43.2
Student	C+R	72.2	41.3	51.0	19.2	15.2	49.0	46.2	35.5	59.1	**60.1**	44.9
CRKD	C+R	**74.8(+2.7)**	**44.1(+6.3)**	**53.6(+4.7)**	**20.6(+2.3)**	**16.9(+4.3)**	**50.6(+2.2)**	**46.8(+4.8)**	**38.2(+4.4)**	**61.5(+2.7)**	**60.1(+0.5)**	**46.7(+3.5)**

Model	Gated	RespD	CSRD	MSFD	RelD	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
Baseline						43.2	54.1	0.489	0.269	0.512	0.313	0.171
CRKD	✓					44.9	55.9	0.464	0.267	0.458	0.304	0.165
	✓	✓				45.7	56.7	0.448	0.262	0.409	0.330	0.166
	✓	✓	✓			46.0	57.0	0.445	0.261	0.407	0.326	0.163
	✓	✓	✓	✓		46.2	57.2	0.439	0.260	0.394	0.332	0.166
	✓	✓	✓	✓	✓	46.7	57.3	0.446	0.263	0.408	0.331	0.162

CRKD: Enhanced Camera-Radar Object Detection with Cross-modality Knowledge Distillation

Lingjun Zhao , Jingyu Song*†, Katherine A. Skinner University of Michigan, Ann Arbor, MI USA*

- **Experiments**

Module	LiDAR	Heatmap	mAP↑	NDS↑
CSRD	✓		44.9	56.3
		✓	46.0	57.0

(a) Ablation study of Cross-stage Radar Distillation (CSRD).

Module	Mask	Mask-scaling	mAP↑	NDS↑
MSFD	✓		46.0	56.7
		✓	46.2	56.9

(b) Ablation study of Mask-scaling Feature Distillation (MSFD).

Module	Vanilla	Adapt	mAP↑	NDS↑
RelD	✓		45.9	56.9
		✓	46.2	57.0

(c) Ablation study of Relation Distillation (RelD).

Module	Vanilla	Dynamic	mAP↑	NDS↑
RespD	✓		45.3	56.7
		✓	45.7	56.7

(d) Ablation study of Response Distillation (RespD).

- **Experiments**

Fuser	*In Channels*	*Out Channels*	*mAP↑*	*NDS↑*
Conv	80 + 256	256	43.2	54.1
Gated	64 + 64	64	44.2	54.3
	128 + 128	256	44.4	54.7
	80 + 256	256	44.9	55.9

CRKD: Enhanced Camera-Radar Object Detection with Cross-modality Knowledge Distillation

Lingjun Zhao , Jingyu Song*†, Katherine A. Skinner University of Michigan, Ann Arbor, MI USA*

- **Experiments**

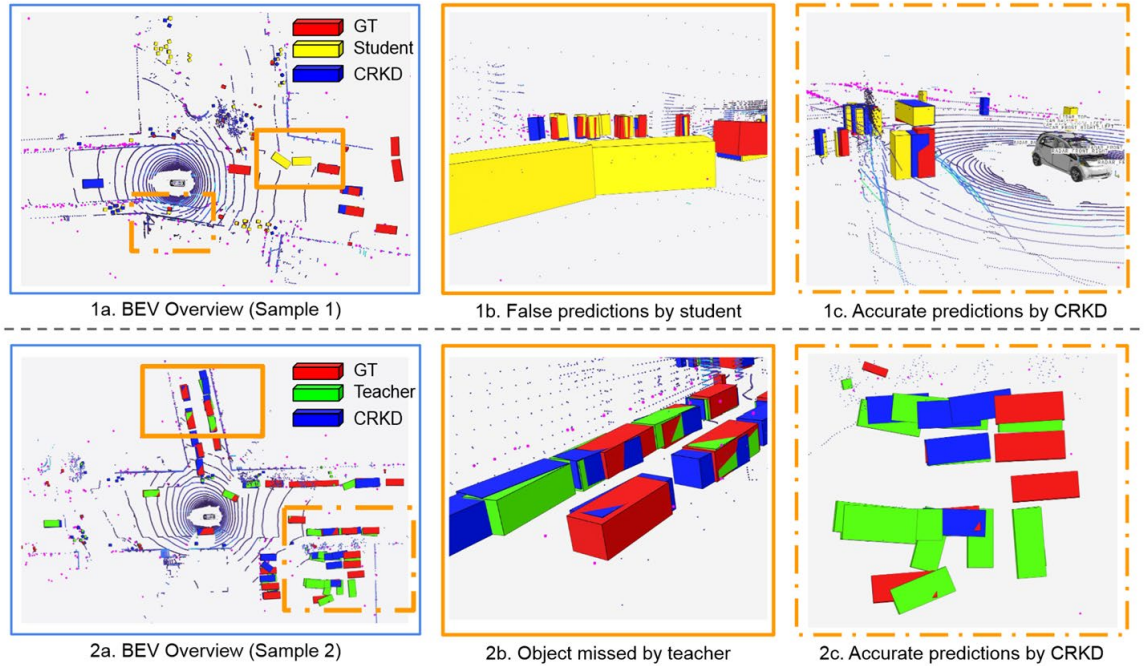


Figure 3. Qualitative results on nuScenes. We show zoomed-in views in panel b and c for the highlighted regions in panel a, with the border dash as the correspondence. We show the ground truth annotation in red, teacher prediction in green, student prediction in yellow, CRKD prediction in blue, and radar points in magenta. In (1a) to (1c), we show an example frame where CRKD has more accurate predictions and fewer false predictions than the student model. In (2a) to (2c) we show another example frame where CRKD even outperforms the LC teacher by detecting a missed car and rejecting several false predictions. Best viewed on screen and in color.