OpenDriveLab 浦驾

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Leveraging Vision-Centric objects for 3D Object Detection

Linyan Huang, Zhiqi Li, Chonghao Sima, Wenhai Wang, Jingdong Wang, Yu Qiao, Hongyang Li

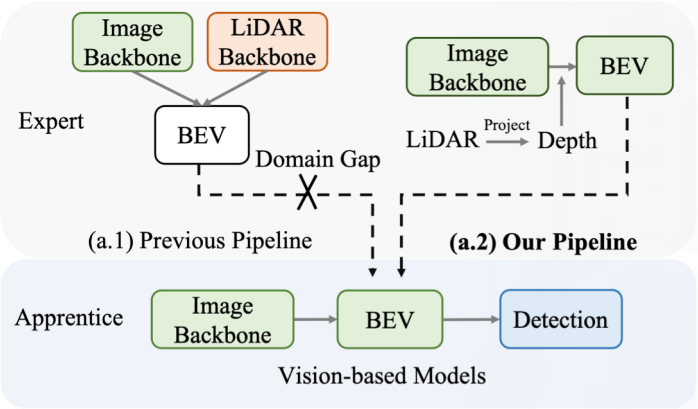
Shanghai AI Lab, Nanjing University, CUHK, Baidu

NeurIPS 2023

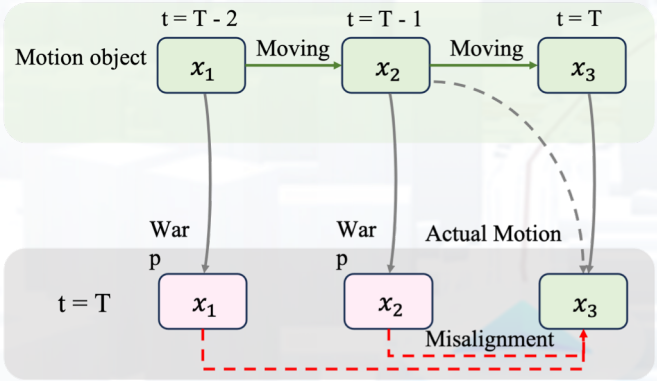
VCD

https://github.com/OpenDriveLab/Birds-eye-view-Perception

Comparison



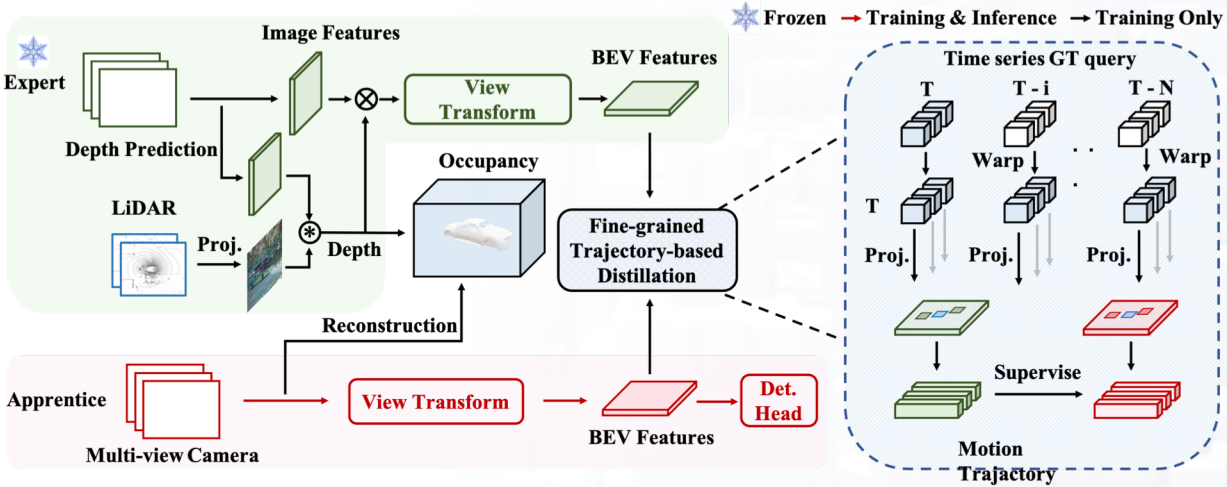
Motion Misalignment



VCD

https://github.com/OpenDriveLab/Birds-eye-view-Perception

Overview



- Better Expert
- Occupancy Reconstruction
- Motion Trajectory Module

VCD

https://github.com/OpenDriveLab/Birds-eye-view-Perception

Experiments

Methods	Backbone	Image Size	Frames	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVDet [23]	ResNet-50	256 × 704	1	0.298	0.379	0.725	0.279	0.589	0.860	0.245
PETR [39]	ResNet-50	384 × 1056	1	0.313	0.381	0.768	0.278	0.564	0.923	0.225
BEVDet4D [22]	ResNet-50	256 × 704	2	0.322	0.457	0.703	0.278	0.495	0.354	0.206
BEVDepth [35]	ResNet-50	256 × 704	2	0.351	0.475	0.639	0.267	0.479	0.428	0.198
BEVStereo [34]	ResNet-50	256 × 704	2	0.372	0.500	0.598	0.270	0.438	0.367	0.190
STS [54]	ResNet-50	256 × 704	2	0.377	0.489	0.601	0.275	0.450	0.446	0.212
VideoBEV [19]	ResNet-50	256 × 704	8	0.422	0.535	0.564	0.276	0.440	0.286	0.198
SOLOFusion [43]	ResNet-50	256 × 704	16+1	0.427	0.534	0.567	0.274	0.411	**0.252**	**0.188**
StreamPETR [51]	ResNet-50	256 × 704	8	0.432	0.540	0.581	0.272	0.413	0.295	0.195
Baseline*	ResNet-50	256 × 704	8+1	0.401	0.515	0.595	0.279	0.489	0.291	0.198
VCD-A	ResNet-50	256 × 704	8+1	0.426	0.540	0.547	0.271	0.433	0.268	0.207
Baseline*†	ResNet-50	256 × 704	8+1	0.418	0.542	0.522	0.267	0.428	0.262	**0.188**
VCD-A†	ResNet-50	256 × 704	8+1	**0.446**	**0.566**	**0.497**	**0.260**	**0.350**	0.257	0.203

Methods	Backbone	Image Size	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
FCOS3D† [52]	R101-DCN	900 × 1600	0.358	0.428	0.690	0.249	0.452	1.434	0.124
DETR3D† [53]	V2-99	900 × 1600	0.412	0.479	0.641	0.255	0.394	0.845	0.133
UVTR [33]	V2-99	900 × 1600	0.472	0.551	0.577	0.253	0.391	0.508	0.123
BEVDet4D† [22]	Swin-B [41]	900 × 1600	0.451	0.569	0.511	**0.241**	0.386	0.301	0.121
BEVFormer [36]	V2-99	900 × 1600	0.481	0.569	0.582	0.256	0.375	0.378	0.126
PolarFormer [28]	V2-99	900 × 1600	0.493	0.572	0.556	0.256	0.364	0.439	0.127
BEVDistill [11]	ConvNeXt-B	900 × 1600	0.496	0.594	0.475	0.249	0.378	0.313	0.125
PETRv2 [40]	RevCol [4]	640 × 1600	0.512	0.592	0.547	0.242	0.360	0.367	0.126
BEVDepth [35]	ConvNeXt-B	640 × 1600	0.520	0.609	0.445	0.243	0.352	0.347	0.127
AeDet† [15]	ConvNeXt-B	640 × 1600	0.531	0.620	0.439	0.247	0.344	0.292	0.130
SOLOFusion [43]	ConvNeXt-B	640 × 1600	0.540	0.619	0.453	0.257	0.376	0.276	0.148
StreamPETR [51]	ConvNeXt-B	640 × 1600	**0.550**	**0.631**	0.493	**0.241**	**0.343**	**0.243**	0.123
Baseline*	ConvNeXt-B	640 × 1600	0.522	0.610	0.457	0.253	0.391	0.271	0.142
VCD-A	ConvNeXt-B	640 × 1600	0.548	**0.631**	**0.436**	0.244	**0.343**	0.290	**0.120**

- VCD-A Results on nuScenes val
- VCD-A Results on nuScenes test
- Consistency Improvement

VCD

https://github.com/OpenDriveLab/Birds-eye-view-Perception

Experiments

Methods	Venue	Backbone	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVFusion [38]	NeurIPS 2022	LiDAR & Image	0.642	0.680	-	-	-	-	-
FUTR3D [10]	Arxiv 2022	LiDAR & Image	0.645	0.683	-	-	-	-	-
UVTR [33]	NeurIPS 2022	LiDAR & Image	0.654	0.702	0.332	0.258	**0.268**	0.212	**0.177**
CMT [57]	Arxiv 2023	LiDAR & Image	**0.679**	0.708	-	-	-	-	-
VCD-E	-	Image	0.677	**0.711**	**0.308**	**0.254**	0.317	**0.189**	0.201

Methods	Backbone	mAP	NDS
BEVFusion	ResNet-50	0.598	0.662
BEVFusion	ConvNext-B	0.597	0.665
VCD-E	ResNet-50	0.611	0.656
VCD-E	ConvNext-B	**0.664**	**0.693**

- VCD-E Results on nuScenes val
- Gains of different image backbone on multi-modal models

VCD

https://github.com/OpenDriveLab/Birds-eye-view-Perception

Experiments

Expert	Paradigm	mAP	NDS
-	-	0.297	0.409
CenterPoint [60]	CM	0.281	0.420
Transfusion [2]	CM	0.292	0.435
BEVDepth [35]	UM	0.341	0.442
VCD-E	UM	**0.354**	**0.459**

Methods	mAP	NDS
Baseline [35]	0.297	0.409
FitNet [44]	0.318	0.421
CWD [46]	0.311	0.412
BEVDistill [11]	0.316	0.439
VCD-A	**0.354**	**0.459**

- The performance gains of the apprentice
- Effect of different distillation methods

VCD

https://github.com/OpenDriveLab/Birds-eye-view-Perception

Experiments

Expert	Paradigm	mAP	NDS
-	-	0.297	0.409
CenterPoint [60]	CM	0.281	0.420
Transfusion [2]	CM	0.292	0.435
BEVDepth [35]	UM	0.341	0.442
VCD-E	UM	**0.354**	**0.459**

Methods	mAP	NDS
Baseline [35]	0.297	0.409
FitNet [44]	0.318	0.421
CWD [46]	0.311	0.412
BEVDistill [11]	0.316	0.439
VCD-A	**0.354**	**0.459**

- The performance gains of the apprentice
- Effect of different distillation methods

VCD

https://github.com/OpenDriveLab/Birds-eye-view-Perception

Experiments

Temporal Length	Distill	mAP (%)↑	NDS (%) ↑	mATE↓	mAOE↓	mAVE↓
1	✗	26.6	37.9	0.815	0.645	0.556
	✓	30.1 (+3.5)	41.5(+3.6)	0.732	0.629	0.476
2	✗	26.9	38.4	0.804	0.706	0.461
	✓	31.3 (+4.4)	43.2 (+4.8)	0.717	0.615	0.403
4	✗	28.4	39.8	0.748	0.739	0.432
	✓	33.0 (+4.6)	44.1 (+4.3)	0.707	0.632	0.389
8	✗	29.7	40.9	0.762	0.714	0.415
	✓	35.4 (+5.7)	45.9 (+5.0)	0.690	0.625	0.370

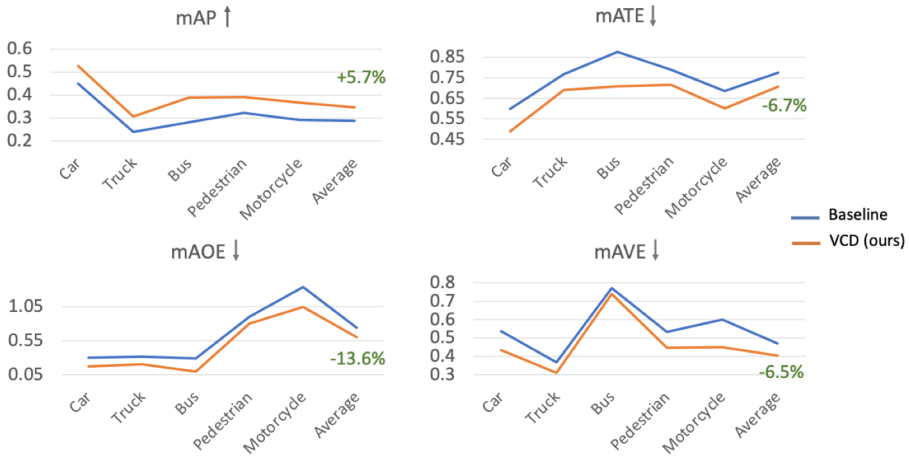
Trajectory Length	Distill	mAP (%)	NDS (%)
-	✗	29.7	40.9
0	✓	31.8	42.1
1	✓	33.1	44.5
3	✓	34.6	45.6
5	✓	**35.4**	**45.9**
9	✓	33.9	44.7

- The performance gains on different temporal lengths
- The performance gains of different trajectory length

VCD

https://github.com/OpenDriveLab/Birds-eye-view-Perception

Experiments



- **The Effects of VCD on Movable Objects**

VCD

https://github.com/OpenDriveLab/Birds-eye-view-Perception

Visualization

- Visualization of the Predictions

OpenDriveLab 浦驾

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Thanks

opendrivelab.com | X @OpenDriveLab