MonoCD: Monocular 3D Object Detection with Complementary Depths

Longfei Yan[1], Pei Yan[1], Shengzhou Xiong[1], Xuanyu Xiang[1], Yihua Tan[1]

[1]School of Artificial Intelligence and Automation,
Huazhong University of Science and Technology

Background

- Monocular 3D object detection(Mono3D) has attracted widespread attention (e.g., in autonomous driving and robotics) due to its potential to accurately obtain object 3D localization from a single image.
- Advantage:
  - Lower cost
  - Simpler configuration
- Challenge:
  - Object depth estimation

Background

- Center-based Detectors:
  - Efficient ✓
  - Local cues
- Transformer-based Detectors
  - Inefficient
  - Global cues ✓

Background

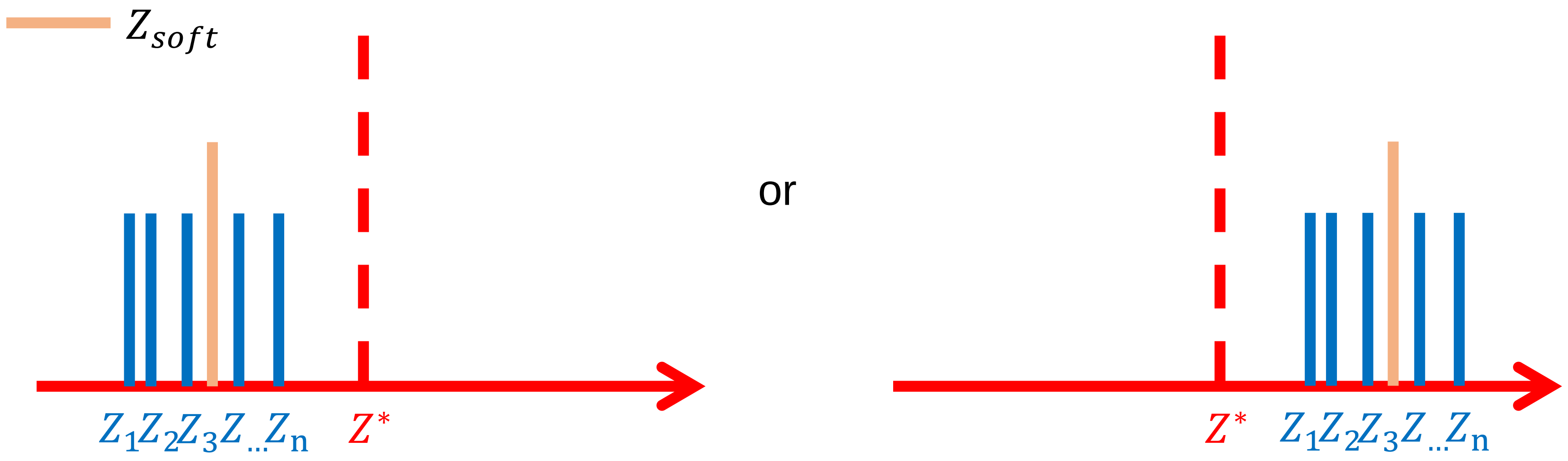
- Center-based Detectors:
  - Tendency: Explore multiple depth cues and formulate them as an ensemble to mitigate the insufficiency of single information

- MonoFlex (2021CVPR): 4 depths per object
- MonoGround (2022CVPR): 7 depths per object
- MonoDDE (2022CVPR): 20 depths per object

Motivation

- Is the greater the number of predicted depths, the better?

  - We observe a coupling phenomenon that existing multiple predicted depths tend to consistently overestimate or underestimate the true depth values



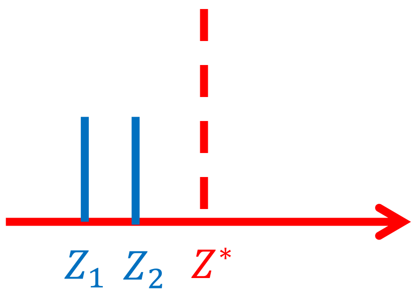
Motivation

- Current multi-depth prediction methods are based on **local cues**

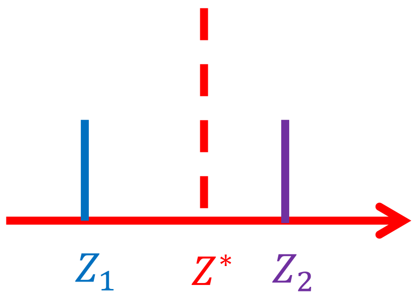
  - We attribute this coupling phenomenon to the fact that the local depth cues they used are all derived from the same **local features** around the object in the CenterNet paradigm.

Motivation

- Can we design a network that has both large and small biased depths?
- Can we utilize global features while ensuring real-time performance to avoid falling into coupling?



coupling



complementary

Core Idea

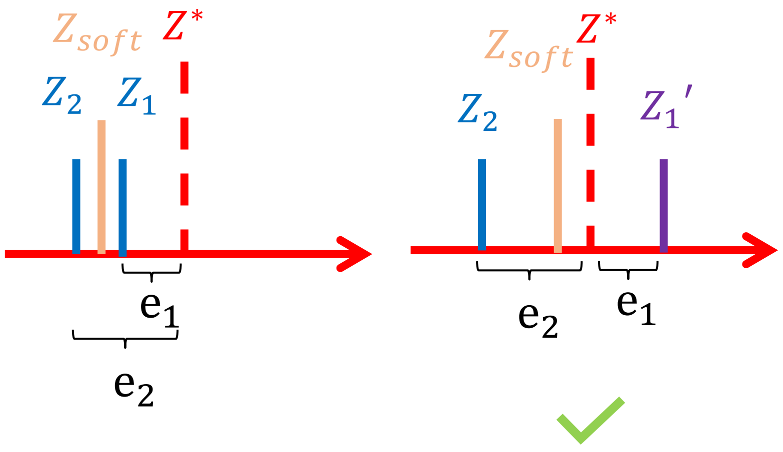
- Why we need multiple depths with different error signs?

  - It's better to listen and synthesize different opinions (different error signs)

proof:

Define two different depth prediction branches $\hat{z}_1$ and $\hat{z}_2$ as follows:

$$\begin{cases} \hat{z}_1 = Z^* + e_1 \\ \hat{z}_2 = Z^* + e_2 \end{cases}, e_1 e_2 > 0$$



*The **coupling depths error E_1** of $\hat{z}_1$ and $\hat{z}_2$ can be formulated as:*

$$E_1 = |\omega_1 e_1 + \omega_2 e_2|$$

By changing only the sign of the error in $\hat{z}_1$ without changing the magnitude of the error we get:

$$\hat{z}_1{}' = Z^* - e_1$$

*The **complementary depths error E_2** of $\hat{z}_1{}'$ and $\hat{z}_2$ can be formulated as:*

$$E_2 = |\omega_1 e_1 - \omega_2 e_2|$$

By mathematical transformations we further express E_1 and E_2 as:

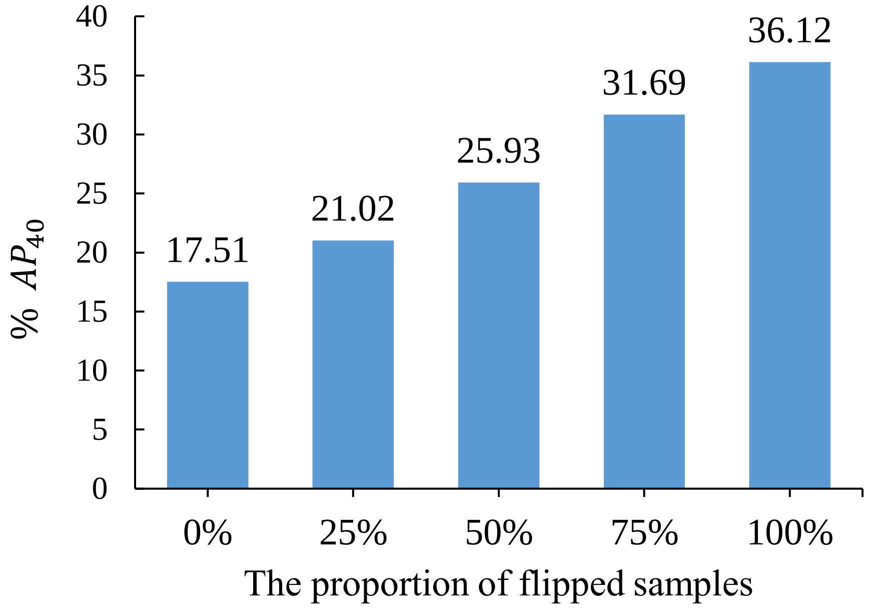
$$\begin{cases} E_1 = \sqrt{(\omega_1 e_1)^2 + 2\omega_1\omega_2 e_1 e_2 + (\omega_2 e_2)^2} \\ E_2 = \sqrt{(\omega_1 e_1)^2 - 2\omega_1\omega_2 e_1 e_2 + (\omega_2 e_2)^2} \end{cases}, e_1 e_2 > 0$$

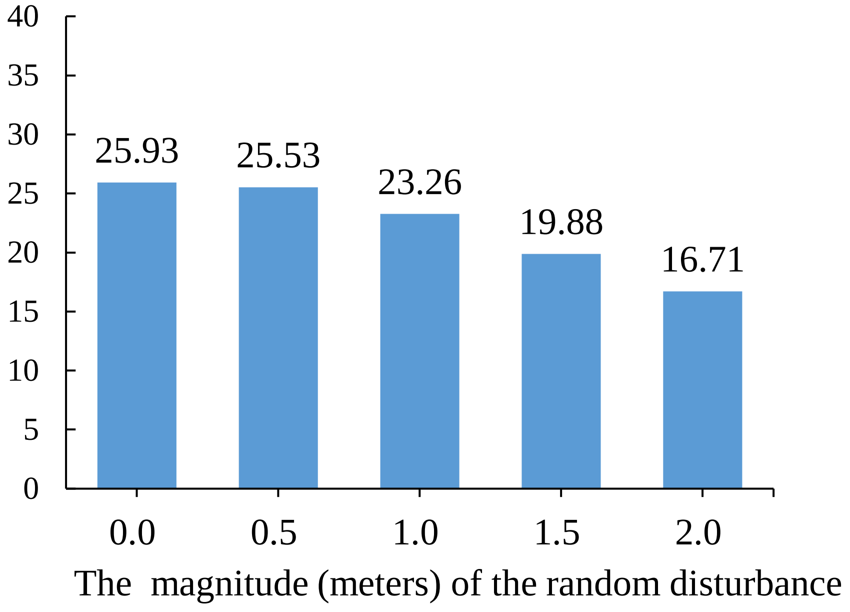
$$E_2 < E_1$$

Core Idea

- Why we need multiple depths with different error signs?

  - Potential to improve existing methods (Take MonoFlex as an example)

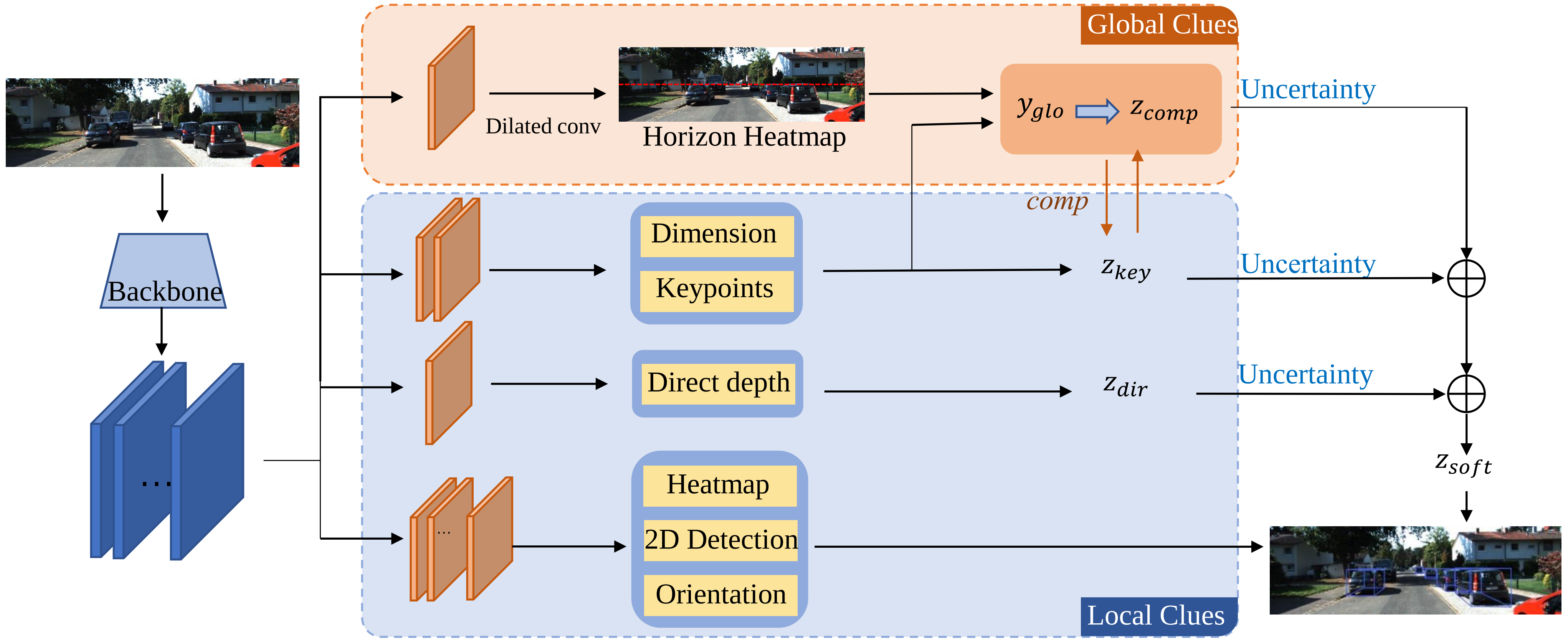




MonoFlex: Zhang Y, Lu J, Zhou J. Objects are different: Flexible monocular 3d object detection. In CVPR, pages 3289-3298, 2021.

Method Overview

- **Global Branch**: predicting the global horizon heatmap of the image, serving as a global cue to generate the prediction of complementary depths (z_{comp}).
- **Local Branch**: predicting local information for each point of interest.



Method

- Introducing a new depth branch with global cue to avoid falling into coupling
  - Global cue comes from that all objects in one image almost lie on the same plane

(u_b, v_b)

$$Ax + By + Cz + D = 0$$

$$y_{glo} = -\frac{D}{A\frac{f_y(u_b - c_u)}{f_x(v_b - c_v)} + C\frac{f_y}{v_b - c_v} + B}$$

$$z = \frac{f_y y_{glo}}{v_b - c_v}$$

(A new geometric cue is introduced)

Method

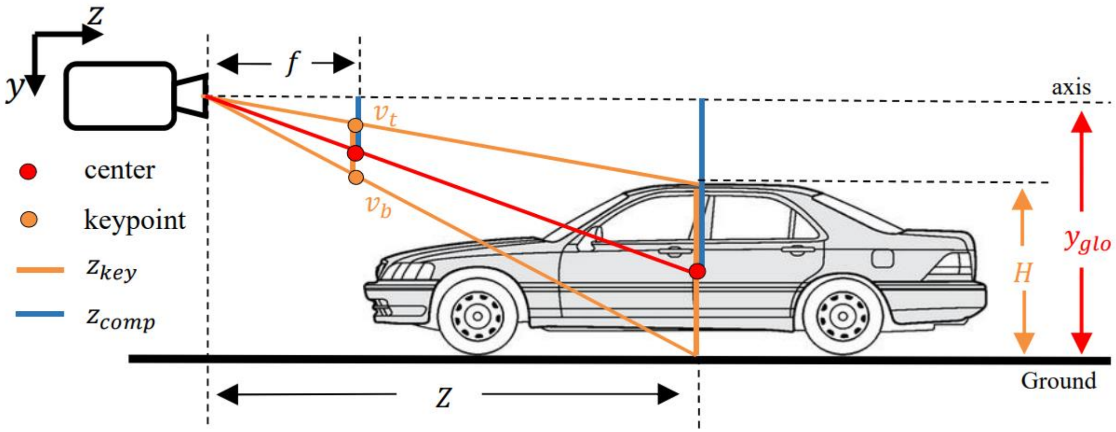
- Achieve complementary form in solving

- Existing

$$z_{key} = \frac{f_y H}{v_b - v_t} \quad (1)$$

- Add

$$z = \frac{f_y y_{glo}}{v_b - c_v} \Rightarrow z_{comp} = \frac{f_y\left(y_{glo} - \frac{1}{2}H\right)}{\frac{1}{2}(v_b + v_t) - c_v} \quad (2)$$



Results

- On the official KITTI 3D benchmark, MonoCD reaches SOTA in most metrics without using additional data while ensuring real-time performance.

Methods, Venues	Extra data	Test, AP_{3D}			Test, AP_{BEV}			Time(ms)
		Eazy	Mod.	Hard	Eazy	Mod.	Hard	
DDMP-3D [33], CVPR2021	Depth	19.71	12.78	9.80	28.08	17.89	13.44	180
Kinematic3D [1], ECCV2020	Video	19.07	12.72	9.17	26.69	17.52	13.10	120
AutoShape [19], ICCV2021	CAD	22.47	14.17	11.36	30.66	20.08	15.59	50
DCD [14], ECCV2022		23.81	15.90	13.21	32.55	21.50	18.25	-
MonoRUn [3], CVPR2021	LiDAR	19.65	12.30	10.58	27.94	17.34	15.24	70
CaDDN [26], CVPR2021		19.17	13.41	11.46	27.94	18.91	17.19	630
MonoDTR [8], CVPR2022		21.99	15.39	12.73	28.59	20.38	17.14	37
SMOKE [18], CVPRW2020	None	14.03	9.76	7.84	20.83	14.49	12.75	30
MonoDLE [21], CVPR21		17.23	12.26	10.29	24.79	18.89	16.00	40
MonoRCNN [29], ICCV2021		18.36	12.65	10.03	25.48	18.11	14.10	70
MonoFlex [40], CVPR2021		19.94	13.89	12.07	28.23	19.75	16.89	35
MonoGround [25], CVPR2022		21.37	14.36	12.62	30.07	20.47	17.74	30
GPENet [35], -		22.41	15.44	12.84	30.31	20.79	18.21	-
MonoJSG [16], CVPR2022		24.69	16.14	13.64	32.59	21.26	18.18	42
MonoCon [17], AAAI2022		22.50	16.46	<u>13.95</u>	31.12	22.10	<u>19.00</u>	25.8
MonoDETR [39], ICCV2023		<u>25.00</u>	<u>16.47</u>	13.58	**33.60**	<u>22.11</u>	18.60	43
MonoCD(Ours)	None	**25.53**	**16.59**	**14.53**	<u>33.41</u>	**22.81**	**19.57**	36
Improvement	*v.s. second-best*	+0.53	+0.12	+0.58	-0.19	+0.70	+0.57	-

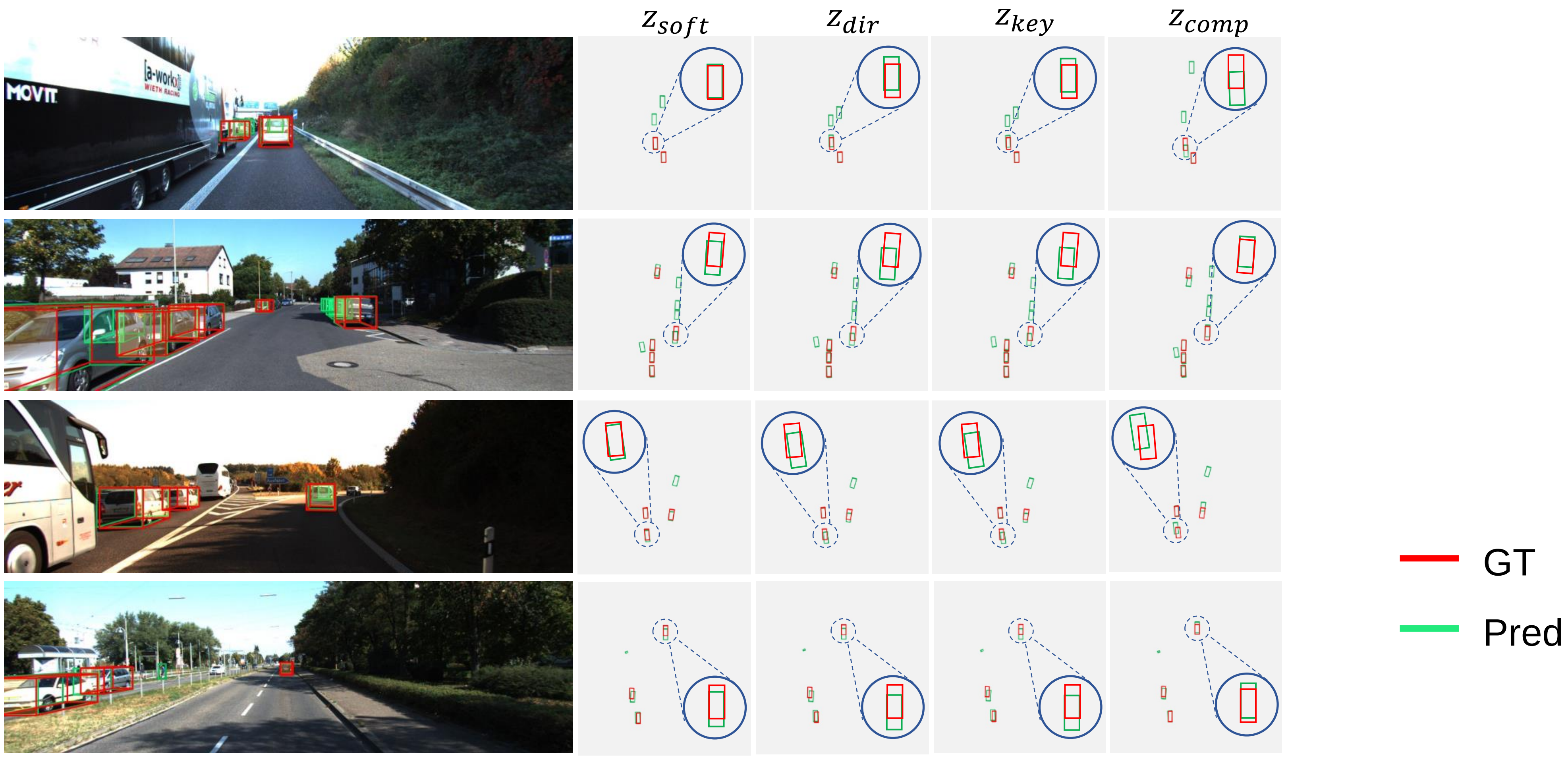
Results

- Proposed complementary depth can also be a plug-and-play module to boost multiple existing monocular 3d object detectors

Method	Val, AP_{BEV}			Val, AP_{3D}		
	Eazy	Mod.	Hard	Eazy	Mod.	Hard
MonoDLE	24.97	19.33	17.01	17.45	13.66	11.68
+ **Ours**	**26.84**	**20.86**	**17.89**	**18.60**	**15.09**	**12.86**
Improvement	+1.87	+1.53	+0.88	+1.15	+1.43	+1.18
MonoFlex	30.51	23.16	19.87	23.64	17.51	15.14
+ **Ours**	**31.49**	**23.56**	**20.12**	**24.22**	**18.27**	**15.42**
Improvement	+0.98	+0.40	+0.25	+0.58	+0.76	+0.28
MonoCon	33.36	24.39	21.03	26.33	19.01	15.98
+ **Ours**	**34.60**	**24.96**	**21.51**	**26.45**	**19.37**	**16.38**
Improvement	+1.24	+0.57	+0.48	+0.12	+0.36	+0.40

Results

- z_{comp} from the global cue branch is significantly different from z_{dir} and z_{key} from the local cue branch and has the opposite error sign, which achieves error neutralization and makes the final predicted box closer to the ground truth



CVPR
SEATTLE, WA JUNE 17-21, 2024

THANK YOU !

Longfei Yan[1], Pei Yan[1], Shengzhou Xiong[1], Xuanyu Xiang[1], Yihua Tan[1]

[1]School of Artificial Intelligence and Automation,
Huazhong University of Science and Technology

華中科技大學
HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

人工智能与自动化学院
School of Artificial Intelligence and Automation,HUST