JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

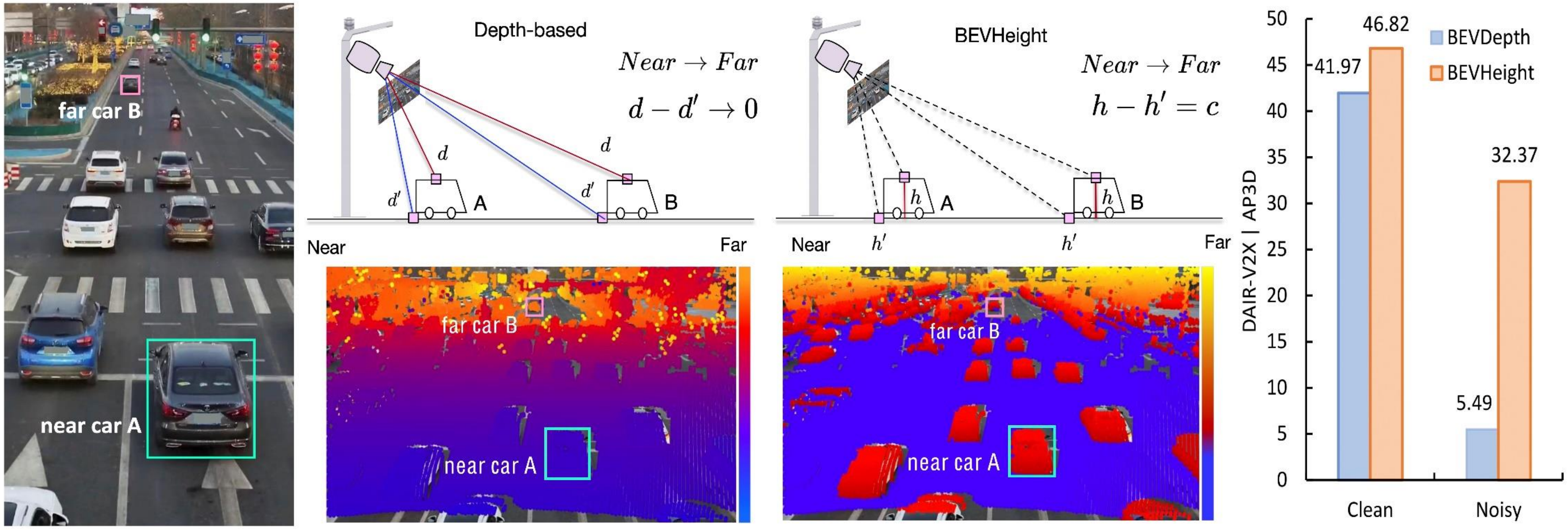
THU-PM-096

BEVHeight: A Robust Framework for Vision-based Roadside3D Object Detection

Lei Yang[1], Kaicheng Yu[2], Tao Tang[3], Jun Li[1], Kun Yuan[4], Li Wang[1], Xinyu Zhang[1*], Peng Chen[2]

[1]Tsinghua University [2] Alibaba Group [3] Sun Yat-sen University [4] Peking University

Preview



Road-side Image

(a) Depth-based Detector

(b) Height-based Detector

(c) Performance

- **Background:** Most AD systems neglect leveraging roadside cameras to enhance perception beyond visual range.
- **Motivation:** The depth difference between the car and the ground decreases as distance increases, while the height difference remains constant. This is superior for the network to detect objects in roadside view.
- **Method:** We propose BEVHeight, by regressing ground height instead of pixel-wise depth, achieving accurate and robust roadside 3D object detection.
- **Experiments:** Our method outperforms the best approach by 4.85% on clean settings and 26.88% on noisy settings.

Background

- Autonomous driving faces great safety challenges due to the inevitably physical occlusion and limited receptive field.
- Roadside perception has a longer perceptual range and greater robustness to occlusion.
- Roadside perception facilitate a safer autonomous driving.

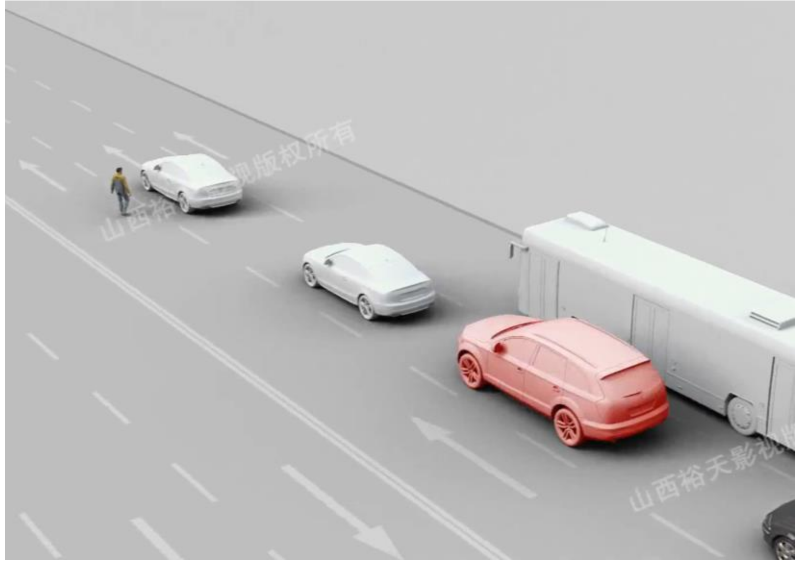


Fig. 1: The inevitably physical occlusion in vehicle-side perception

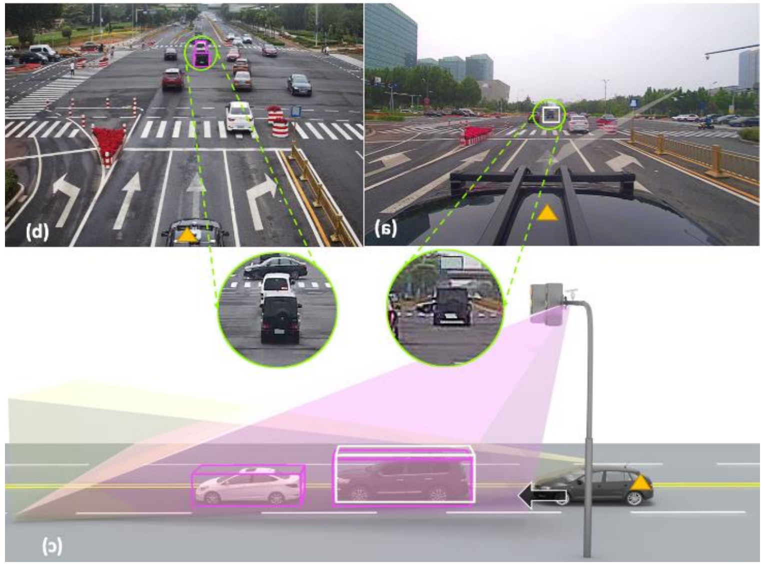


Fig. 2: The comparison of (a) vehicle view and (b) roadside camera view with a pitch angle.

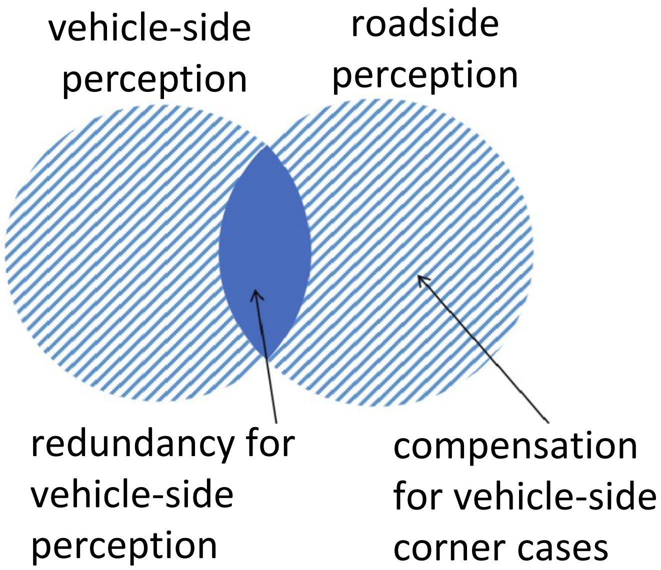


Fig. 3: the redundancy complementarity in vehicle and roadside platforms.

Background

Vision-based roadside 3D object detection have two challenges：

- **Various camera’s specifications**, such as roll, pitch and mounting height.
- An increase in obstacle density.

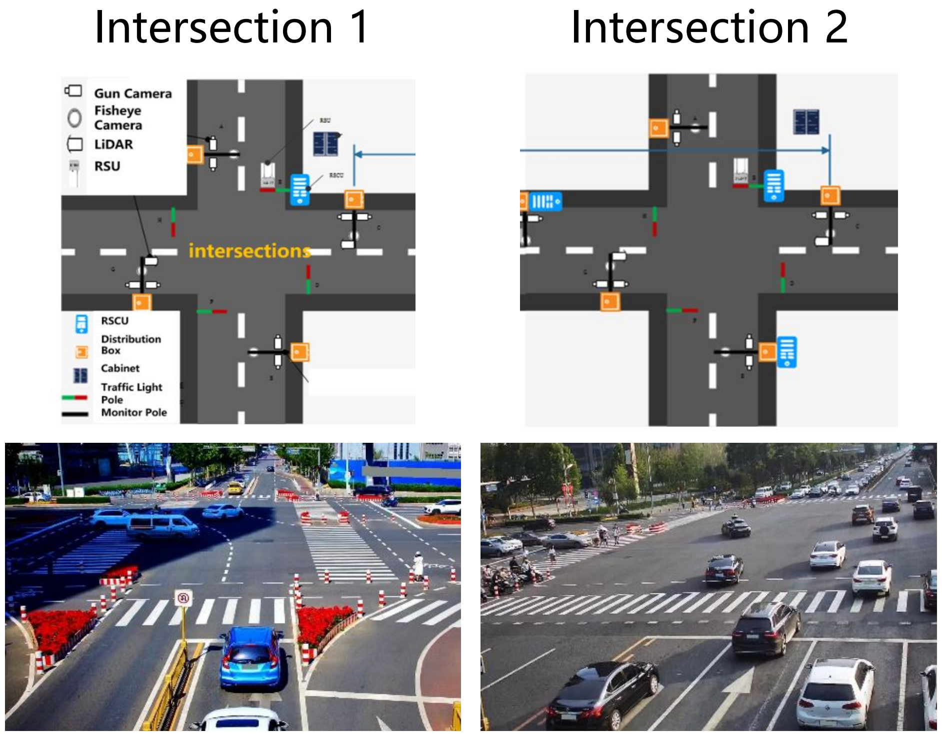


Fig. 4. The images from different roadside cameras.

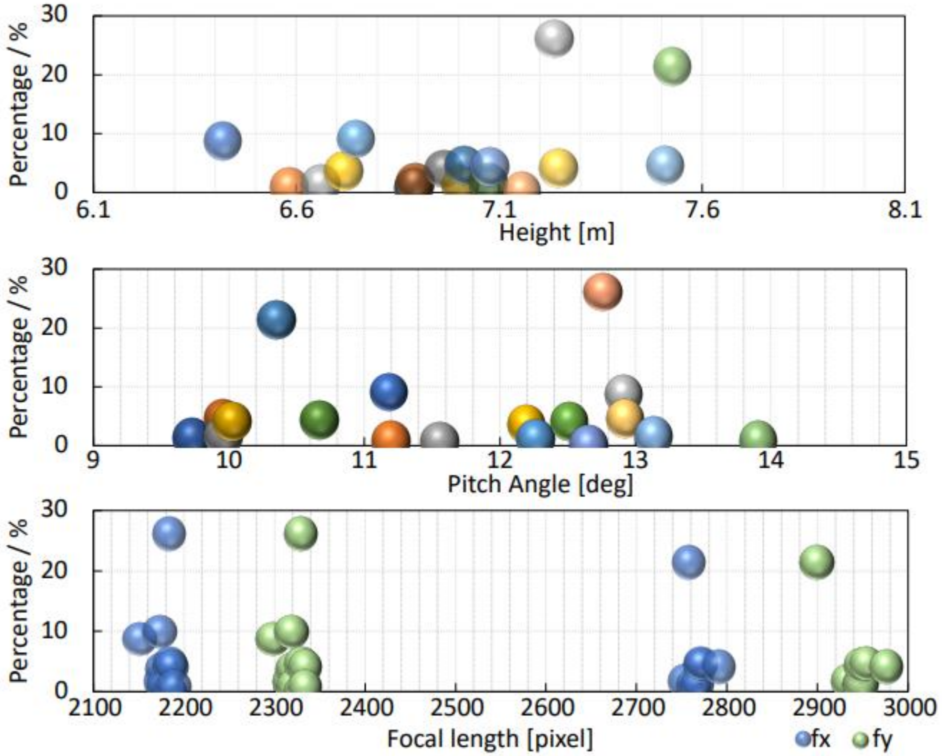
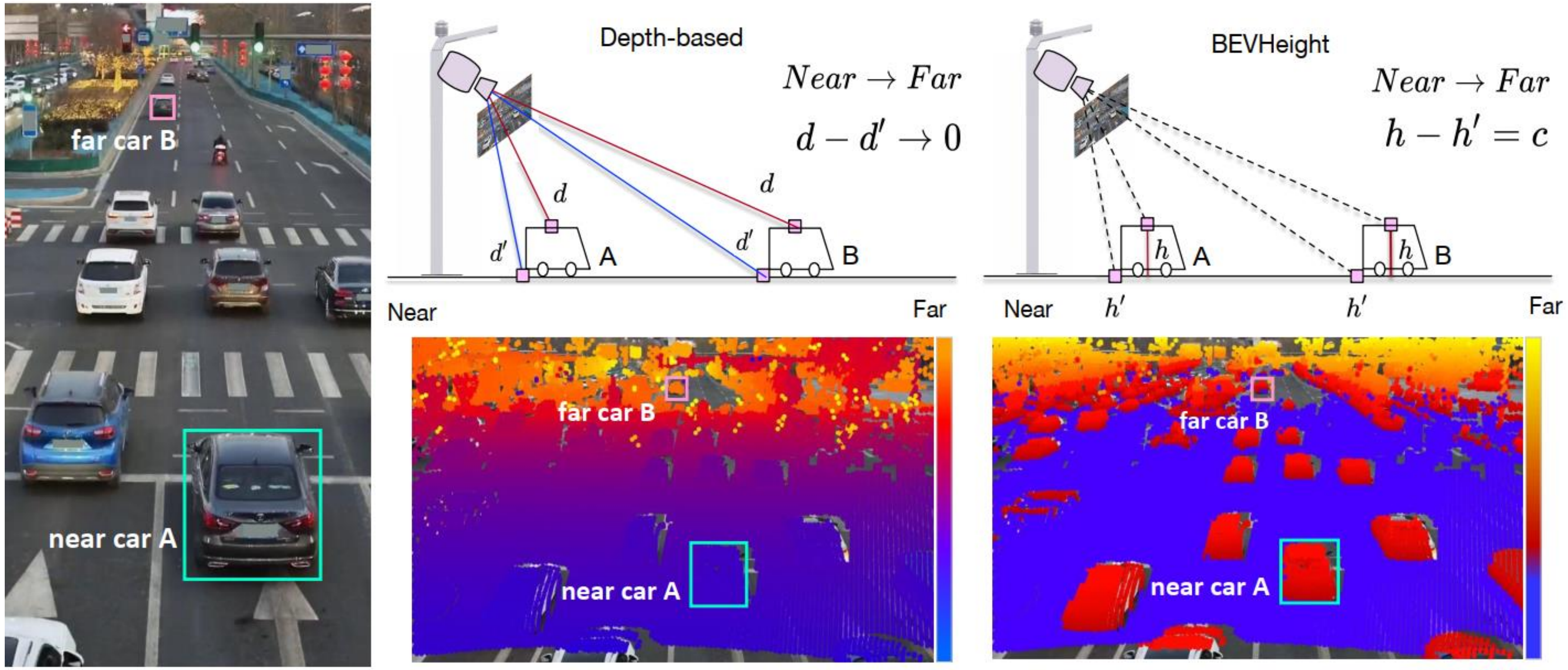


Fig. 5. The diversity of roadside camera’s specifications in Rope3D dataset.

Motivation

Principle from 2D to 3D



Road-side Image　　(a) Depth-based Detector　　(b) Height-based Detector

Depth

↓

Height

a) The depth differences between points on the car roof and surrounding ground quickly shrink when the car moves away from the camera, making it sub-optimal to optimize especially for far objects.

b) The height difference between the same points remains agnostic regardless of the distance, and visually is superior for the network to detect objects.

Motivation

Comparing the depth and height

Distribution:

The range of depth is over 200 meters while the height is within 5 meters, which makes height much easier to learn.

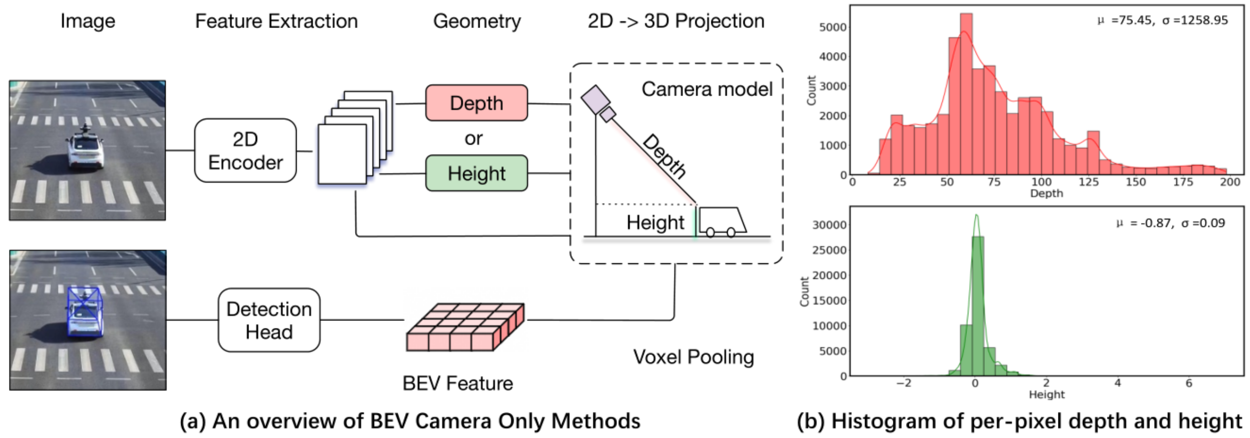


Fig. 7. The comparison of predicting height and depth.

Analysis when extrinsic parameters change:

Compared with depth, the noisy setting of height has larger overlap with its original distribution, which demonstrates height estimation is more robust.

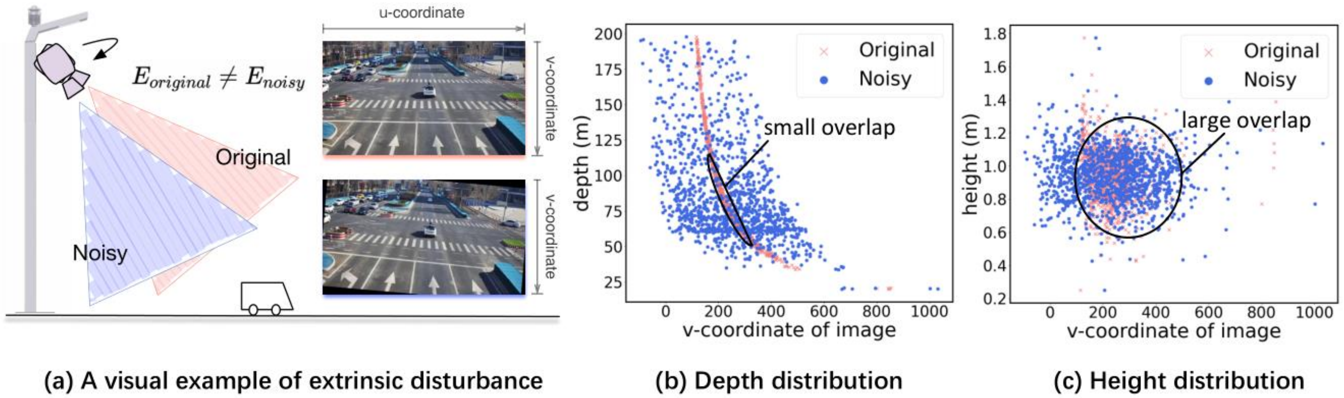


Fig. 8. The correlation between the object's row coordinates on the image with its depth and height.

Proposed Method

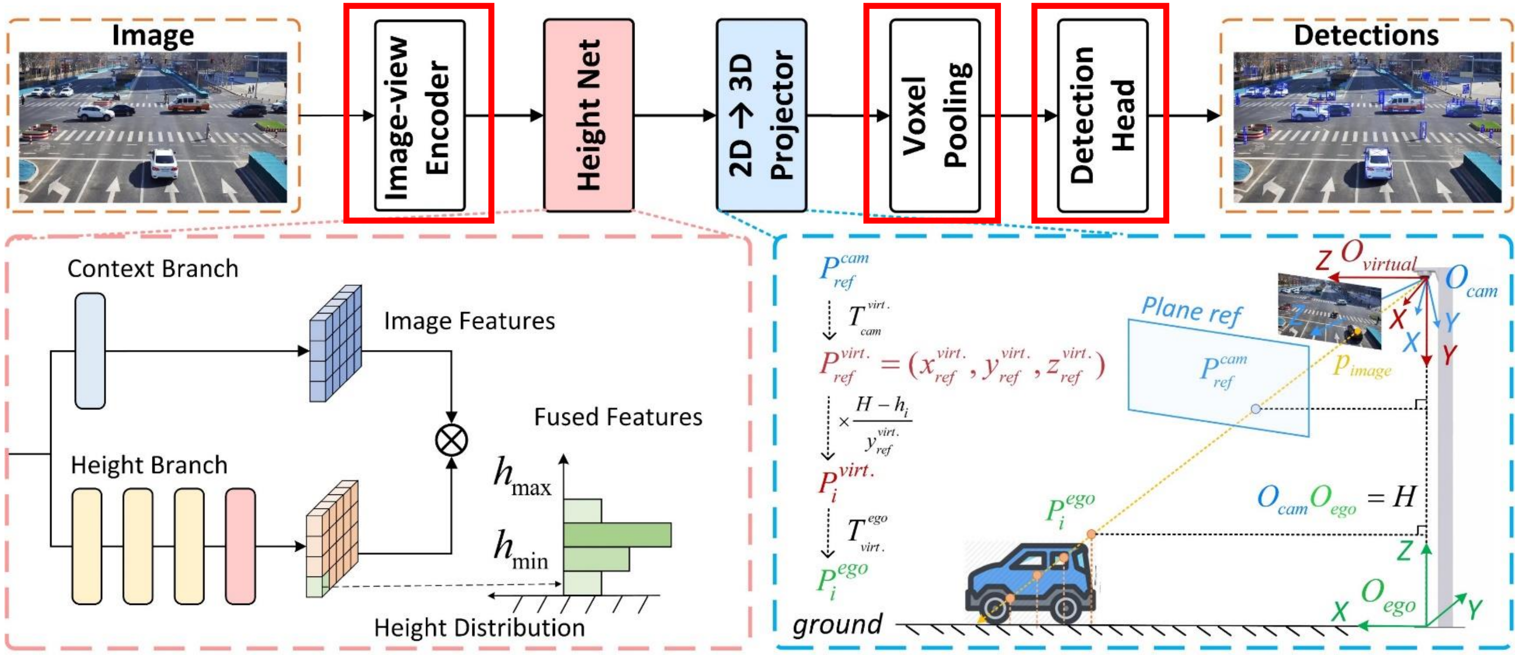
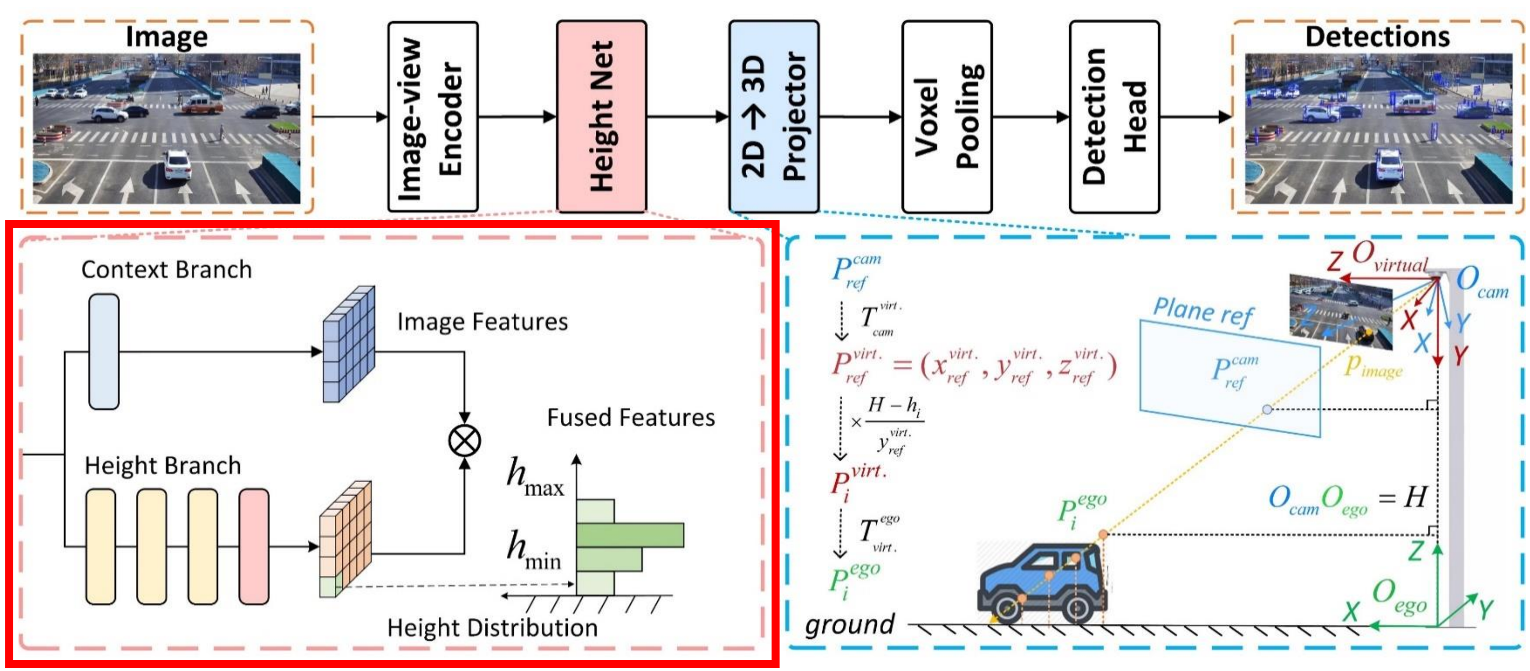


Fig. 9. The overall framework of BEVHeight

Image-view Encoder	extracts the 2D high-dimensional image features from a RGB image.
Voxel Pooling	transforms the 3D volume features into the BEV features along the height direction.
Detection Head	predicts the 3D bounding box consisting of location, dimension, and orientation.

Proposed Method



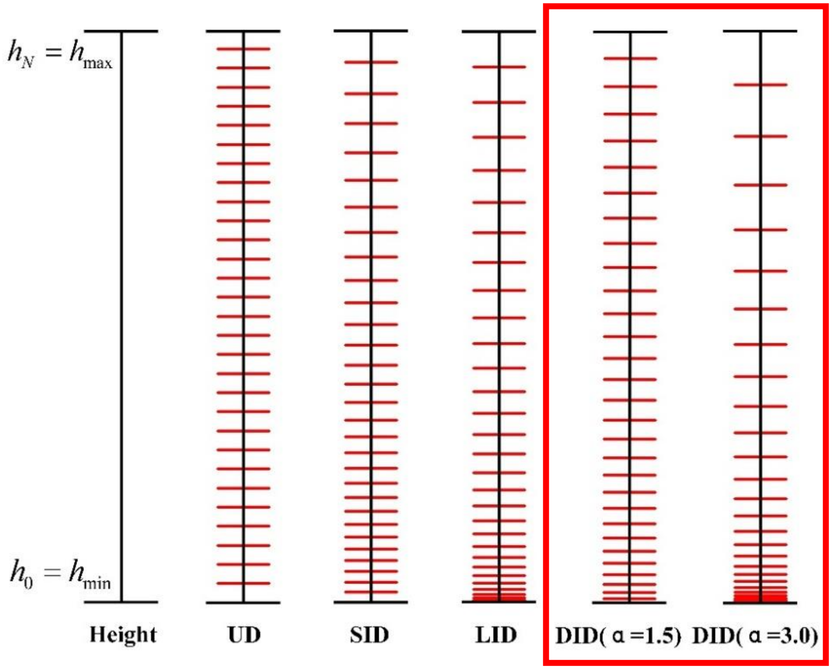


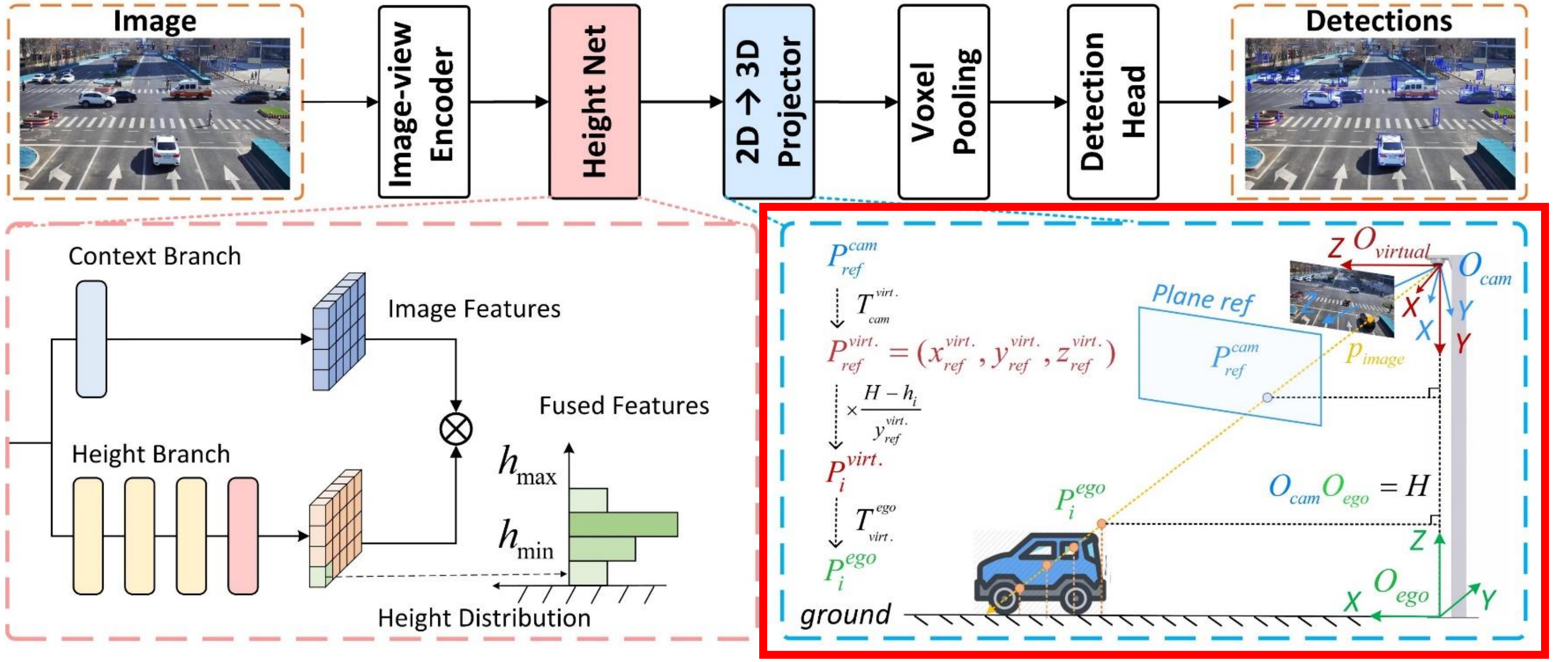
Fig. 10. Height Discretization Methods.

HeightNet	generate bins-like height distribution and context features.

- Context branch consists of a squeeze-and-excitation(SE) layer.
- Height branch contains three residual blocks and a DCN layer.
- A dynamic-increasing discretization strategy (DID) with adjustable size.

$$h_i = \left\lfloor N \times \sqrt[\alpha]{\frac{h - h_{min}}{h_{max} - h_{min}}} \right\rfloor$$

Proposed Method



2D->3D Projector	Push the 2D features into 3D volume features.

- We design a virtual coordinate system leveraging the height predictions.
- We adopt a reference plane to simplify the computation.

$$P_i^{ego} = T_{virt.}^{ego} \frac{H - h_i}{y_{ref}^{virt.}} T_{cam}^{virt.} K^{-1}[u, v, 1]^T$$

Algorithm 1 Height-based 2D to 3D projector

Parameters Definition:

O, X, Y, Z: coordinate system, where $O_{virt.}$ has the same origin as O_{cam} with Y-axis prependicular to the ground.
T_A^B: transformation matrix from coordinate A to B.
K: the camera's intrinsic matrix.
H: the distance from the origin of the virtual coordinate system to the ground.
h_i: the height from the ground of i-th height bin.
P_{ref}^B: the pixel (u, v) projected from reference plane A in coordinate B
P_i^A: the pixel (u, v) projection point on i-th height bin in the coordinate system A.

Input:
$F^{fused} = \left\{f_1^{fused}, \ldots, f_{\frac{H}{16}\times\frac{W}{16}}^{fused}\right\}$, $f_m^{fused} \in R^{C_H \times C_c}$
H; K; $T_{cam}^{virt.}$; T_{cam}^{ego}
Output:
F_{wedge} is the 3D wedge-shaped volume features.
Begin:

```
F_wedge = {}
for f_m^fused in F^fused do
    u, v ← m
    P_ref^cam = K^{-1}[u, v, 1]^T
    P_ref^virt. = {x_ref^virt., y_ref^virt., z_ref^virt.} = T_cam^virt. P_ref^cam
    for i ← 0 to C_H do
        P_i^virt. = (H − h_i)/y_ref^virt. P_ref^virt.
        P_i^ego = T_virt.^ego P_i^virt.
        F_wedge ← F_wedge ∪ associate(P_i^ego, f_m^fused[i])
    end for
end for
return F_wedge
```

End

Experiments

Comparisons with state-of-the-arts

Tab. 1: Comparisons with SOTA methods on the DAIR-V2X-I val set.

Method	Modality	Vehicle$_{(IoU=0.5)}$			Pedestrian$_{(IoU=0.25)}$			Cyclist$_{(IoU=0.25)}$		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
PointPillars [1]	L	63.07	54.00	54.01	38.53	37.20	37.28	38.46	22.60	22.49
SECOND [6]	L	71.47	53.99	54.00	55.16	52.49	52.52	54.68	31.05	31.19
MVXNet [5]	LC	71.04	53.71	53.76	55.83	54.45	54.40	54.05	30.79	31.06
ImvoxelNet [4]	C	44.78	37.58	37.55	6.81	6.746	6.73	21.06	13.57	13.17
BEVFormer [3]	C	61.37	50.73	50.73	16.89	15.82	15.95	22.16	22.13	22.06
BEVDepth [2]	C	75.50	63.58	63.67	34.95	33.42	33.27	55.67	55.47	55.34
BEVHeight	C	**77.78**	**65.77**	**65.85**	**41.22**	**39.29**	**39.46**	**60.23**	**60.08**	**60.54**

L, C denotes LiDAR, camera respectively.

DAIR-V2X-I:
Our method significantly outperforms the SOTA by a large margin;
Vehicle +2.19% Pedestrian +5.87%
Cyclist +4.61%

Tab. 2: Comparisons with SOTA methods on the Rope3D val set.

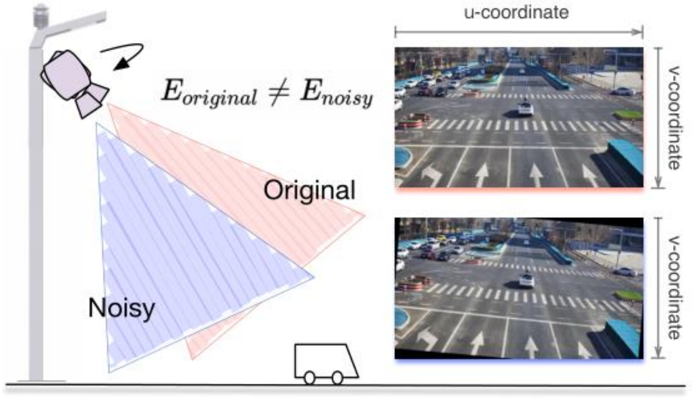
Method	IoU = 0.5				IoU = 0.7			
	Car		Big Vehicle		Car		Big Vehicle	
	AP	Rope	AP	Rope	AP	Rope	AP	Rope
M3D-RPN [1]	54.19	62.65	33.05	44.94	16.75	32.90	6.86	24.19
Kinematic3D [2]	50.57	58.86	37.60	48.08	17.74	32.9	6.10	22.88
MonoDLE [6]	51.70	60.36	40.34	50.07	13.58	29.46	9.63	25.80
MonoFlex [11]	60.33	66.86	37.33	47.96	33.78	46.12	10.08	26.16
BEVFormer [5]	50.62	58.78	34.58	45.16	24.64	38.71	10.05	25.56
BEVDepth [4]	69.63	74.70	45.02	54.64	42.56	53.05	21.47	35.82
BEVHeight	**74.60**	**78.72**	**48.93**	**57.70**	**45.73**	**55.62**	**23.07**	**37.04**

AP and Rope denote $AP_{3D|R40}$ and $Rope_{score}$ respectively.

Rope3D:
Ours method is also better than the SOTA under large-scale dataset.
Car +4.97%
Big Vehicle +3.91%

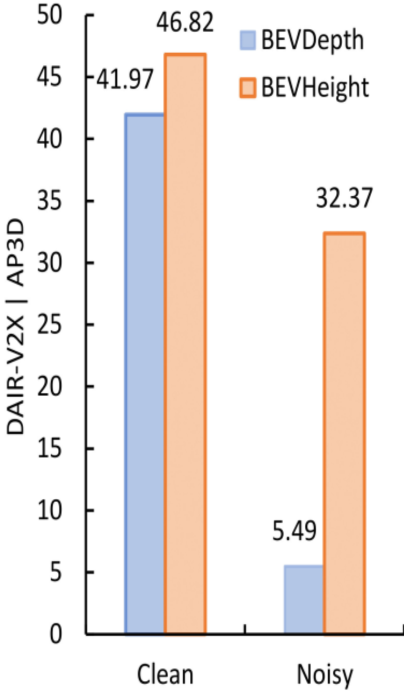
Experiments

Comparisons on robustness settings



26.88% ↑

Our BEVHeight maintains the best performance under the disturbed roll and pitch angles.



Tab. 3: Comparisons on robustness settings.

Model	Disturbed		Vehicle$_{(IoU=0.5)}$			Pedestrian$_{(IoU=0.25)}$			Cyclist$_{(IoU=0.25)}$		
	roll	pitch	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
BEVFormer [3]			61.37	50.73	50.73	16.89	15.82	15.95	22.16	22.13	22.0
	✓		50.65	42.90	42.95	10.16	9.41	9.47	13.62	13.71	13.08
		✓	46.40	38.26	38.37	9.12	8.44	8.55	8.99	8.43	8.42
	✓	✓	19.24	16.35	16.47	3.93	3.43	3.52	4.93	4.98	4.98
BEVDepth [2]			71.56	60.75	60.85	21.55	20.51	20.75	40.83	40.66	40.26
	✓		34.82	28.32	28.35	4.49	4.36	4.39	10.48	9.51	9.73
		✓	14.04	11.41	11.49	3.01	2.67	2.75	6.43	6.23	6.83
	✓	✓	11.84	9.48	9.54	2.16	1.84	1.89	4.31	4.14	4.26
BEVHeight			75.58	63.49	63.59	26.93	25.47	25.78	47.97	47.45	48.12
	✓		66.06	54.99	55.14	18.66	17.63	17.78	34.45	26.93	27.68
		✓	68.49	56.98	57.11	17.94	16.87	17.09	34.48	27.82	28.67
	✓	✓	62.64	51.77	51.9	14.38	14.01	14.09	31.28	25.24	26.02

Experiments

Ablation Studies

Dynamic Discretization strategy (DID):

Our dynamic discretization is effective.

The hype-parameter α is necessary to achieve the most appropriate discretization.

Tab. 4: Ablation on dynamic discretization.

Spacing		Veh.$_{(IoU=0.5)}$			Ped.$_{(IoU=0.25)}$			Cyc.$_{(IoU=0.25)}$		
DID (α)	UD	Easy	Mid	Hard	Easy	Mid	Hard	Easy	Mid	Hard
	✓	75.63	63.75	63.85	25.82	25.47	25.35	47.52	47.47	47.19
✓(1.5)		76.24	64.54	64.13	26.47	25.79	25.72	48.55	48.21	47.96
✓(2.0)		**76.61**	**64.71**	**64.76**	**27.34**	**26.09**	**25.33**	**49.68**	**48.84**	**48.58**

Latency:

The BEVHeight is more efficient because of much less height bins in the smaller height range.

Tab. 5: Latency of BEVHeight and BEVDepth.

Methods	Backbone	Range	Number of bins	Latency (ms)	FPS
BEVDepth [16]	R50	1 - 104m	206	82	12.2
BEVHeight	R50	-1 - 1m	90	77	13.0
BEVDepth [16]	R101	1 - 104m	206	68	14.7
BEVHeight	R101	-1 - 1m	90	62	16.1

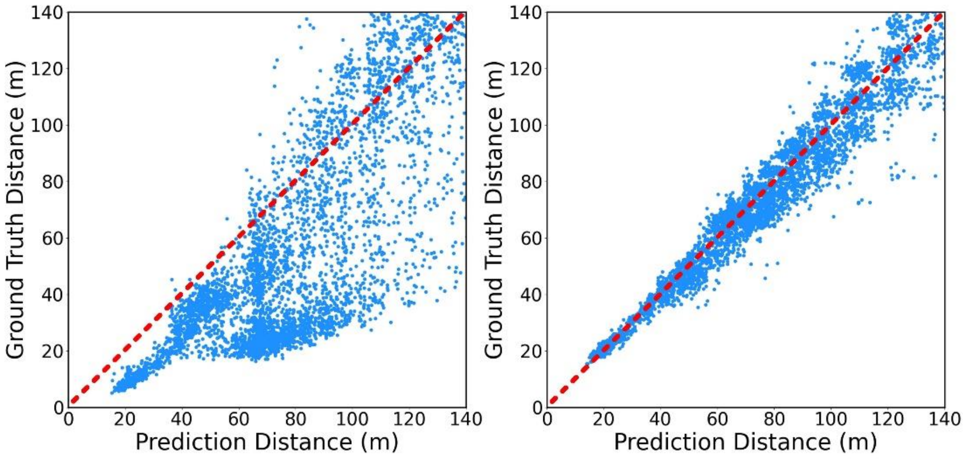
Measured on a V100 GPU. Image shape 864×1536.

Experiments

Ablation Studies

Analysis on Distance Error:

Height estimation in BEVHeight exhibits superior accuracy compared to depth estimation in roadside scenarios, minimizing errors.



(a) BEVDepth Distance Correlation (b) BEVHeight Distance Correlation

Fig. 11. Empirical analysis of the distance correlation

Effectiveness on multi depth-based Detectors:

Replacing the depth-based projection in BEVDepth, our method achieves a performance increase of 2.19%, 5.87%, 4.61% on vehicle, pedestrian and cyclist. Similarly, our approach surpasses BEVDet by 8.56%, 5.35%, 8.60% respectively.

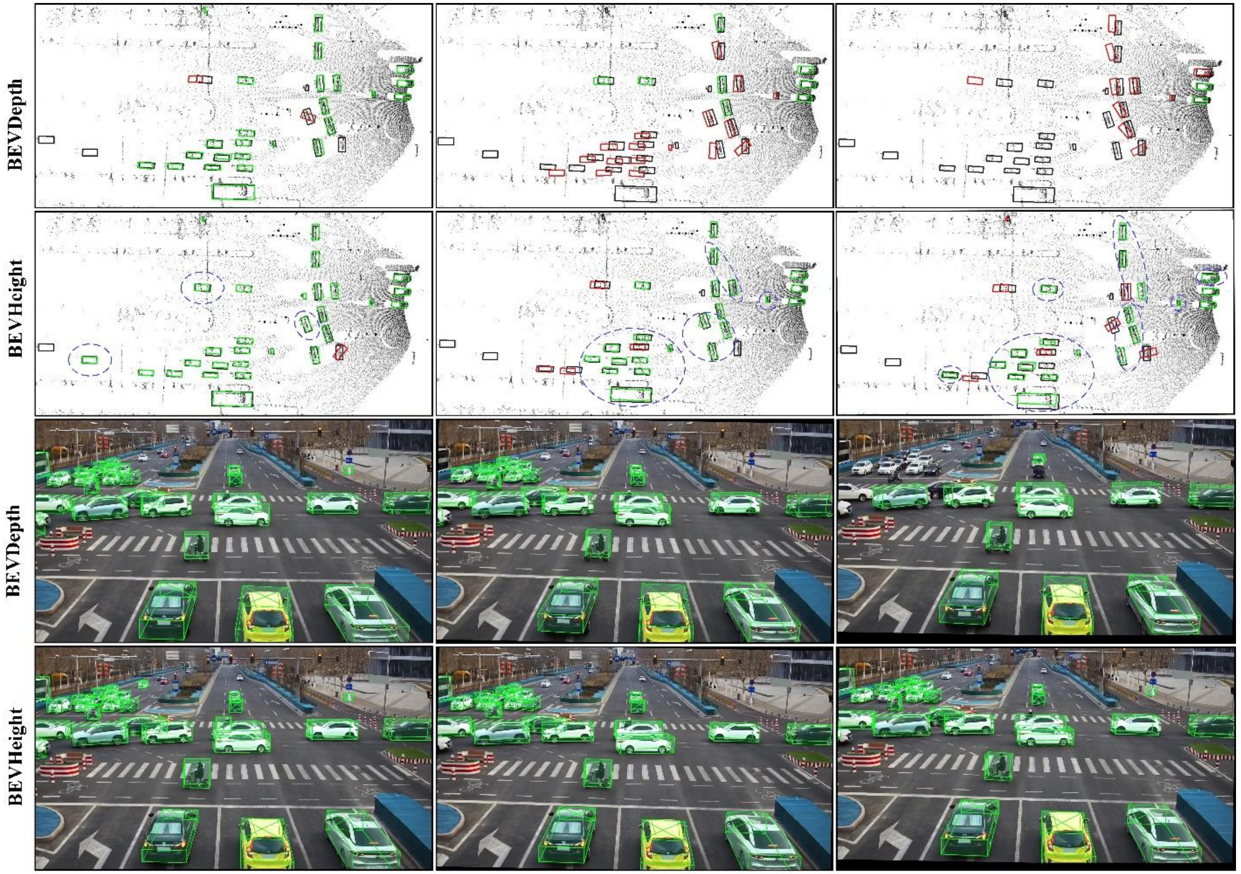
Tab. 6: Ablation studies on different depth-based methods.

Method	VT	Veh.$_{(IoU=0.5)}$ Easy	Mod.	Hard	Ped.$_{(IoU=0.25)}$ Easy	Mod.	Hard	Cyc.$_{(IoU=0.25)}$ Easy	Mod.	Hard
BEVDepth [16]	D	75.50	63.58	63.67	34.95	33.42	33.27	55.67	55.47	55.34
	H	77.78	65.77	65.85	41.22	39.29	39.46	60.23	60.08	60.54
BEVDet [10]	D	59.59	51.92	51.81	12.61	12.43	12.37	34.91	34.32	34.21
	H	69.42	60.48	59.68	18.11	17.81	17.74	44.69	42.92	42.34

VT denotes view transformation, D,H represents depth-based and height-based ones.

Experiments

Qualitative results



(a) Clean (b) Disturbed Roll (c) Disturbed Roll and Pitch

black - ground truth
red - false positive
green - truth positive

- On the clean setting, our BEVHeight fit more closely to the ground truth than that of BEVDepth.
- Under the disturbance of pose angles, our method consistently maintains accurate positioning, while there is a noticeable deviation in the BEVDepth detections when compared to the ground truth.

Discussion

Limitations and Analysis

Limitation:
Our methods are effective on cameras with high installation and bird's-eye-view as in the roadside scenario, and is not ideal on cameras mounted on ego-vehicles.

Analysis:
Fig. 12: (a) shows when the height prediction is equal to the ground-truth, detection is perfect for all cameras; (b) if not, for the same height prediction error, the distance between predicted point and ground-truth is inversely proportional to the camera ground height.

Verification:
BEVHeight surpasses BEVDepth when the camera's height only increases less than 1 meter (on truck platform).

Tab. 7: Comparisons on nuScenes dataset.

Method	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVDepth	0.315	0.367	0.702	0.271	0.621	1.042	0.315
BEVDepth*	0.313	0.354	0.713	0.280	0.655	1.230	0.377
BEVHeight	0.291	0.342	0.722	0.278	0.674	1.230	0.361

* denotes the results we reproduce.

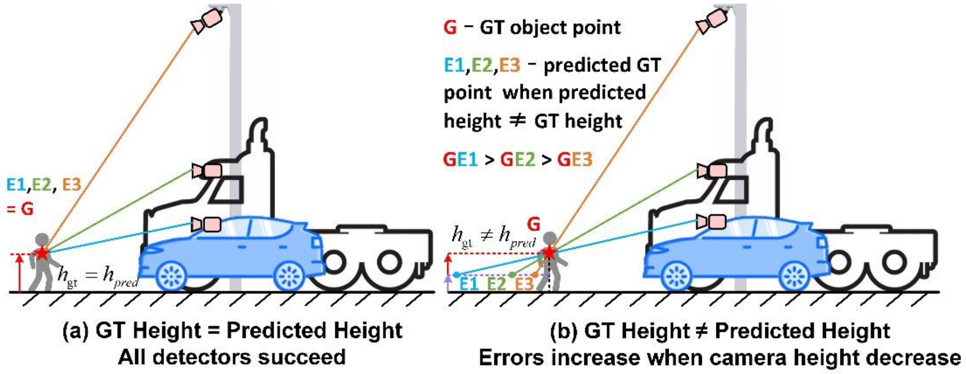


Fig. 12. Distance error analysis caused by same height estimation error on different platform cameras.

Tab. 8: Comparisons on the dataset collected by higher truck.

Method	Car$_{(IoU=0.5)}$			Big Vehicle$_{(IoU=0.5)}$		
	Easy	Mod.	Hard	Easy	Mod.	Hard
BEVDepth [16]	50.05	36.82	36.82	30.15	24.74	24.74
BEVHeight	**51.77**	**40.96**	**40.96**	**34.65**	**29.01**	**29.01**

Conclusion

- we take the advances and challenges of roadside cameras into account, and design an efficient and robust roadside perception framework, **BEVHeight.**

- we implement a lightweight HeightNet and design a novel height-based projection module to achieve the projection from 2D to 3D effectively.

- The proposed detector achieves state-of-the-art results on DAIR-V2X-I and Rope3D dataset, and up to 26.88% improvements on robust settings where external camera parameters change.

Thanks!

Code

Scan ME