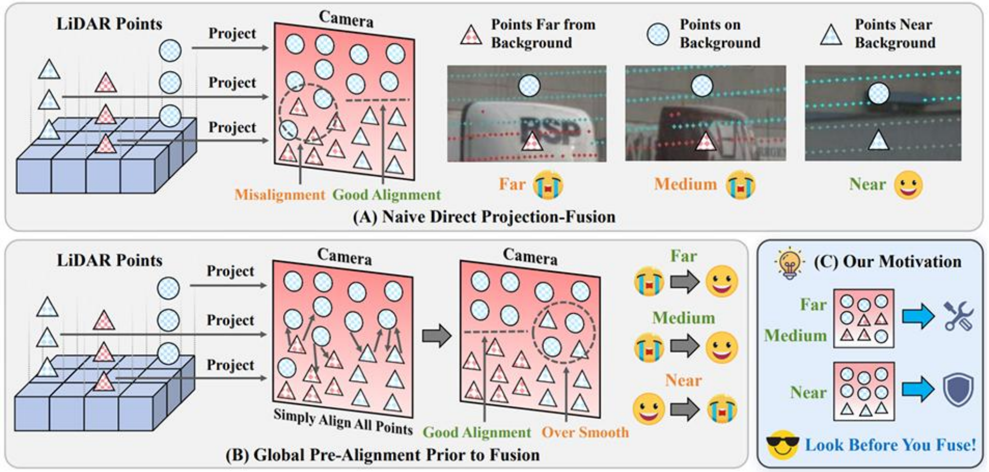
Look Before You Fuse: 2D-Guided Cross-Modal Alignment for Robust 3D Detection

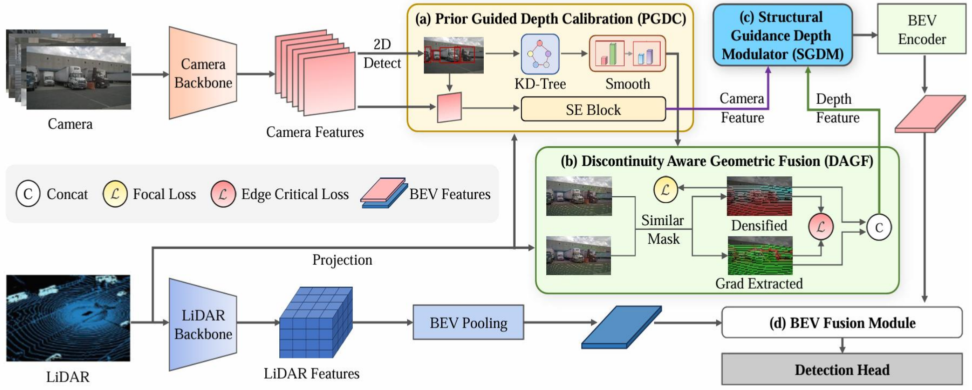
Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences
[2]University of Science and Technology of China
Xiang Li[1,2] Zhangchi Hu[2] Xu Xiao[1,2] Bin Kong[1†]

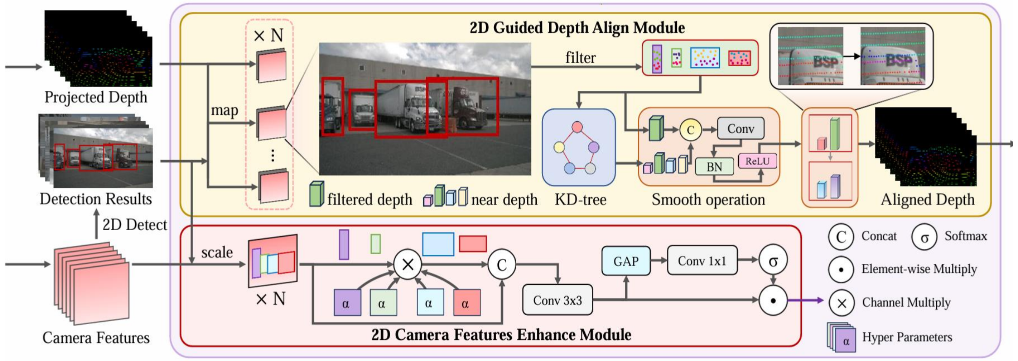
CVPR JUNE 3-7, 2026 DENVER COLORADO

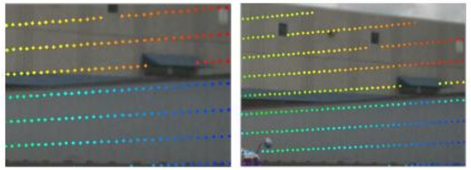
Motivation-Prevailing methods often overlook that misalignment is not randomly distributed, but originates from two pri mary sources: extrinsic calibration errors and motion induced distortions.



Method-(a) Prior Guided Depth Calibration (PGDC) and (b) Discontinuity Aware Geometric Fusion (DAGF), proactively mitigate multi-sensor feature misalignment before view transformation. And (c) Structural Guidance Depth Modulator (SGDM) intelligently fuses image features and dense geometric representation, predicting an accurate depth distribution.







original points over-smoothed points

Method	Standard		w/ Noise (Drop)	
	mAP (%)	NDS (%)	mAP (%)	NDS (%)
Baseline	68.5	71.4	60.8 (-7.7)	65.7 (-5.7)
TransFusion	69.4	72.3	66.3 (-3.1)	70.1 (-2.2)
RobBEV	69.7	72.5	66.8 (-2.9)	70.5 (-2.0)
GraphBEV	70.1	72.9	68.7 (-1.4)	70.4 (-2.5)
Ours	**71.3**	**73.7**	**70.3** (-1.0)	**72.4** (-1.3)

The originally structured point cloud distribution is distorted at the boundary after global align.

PGDC	DAGF	SGDM	mAP (%)	NDS (%)	LT (ms)
×	×	×	67.9	71.0	+0.0
✓	×	✓	69.8	72.5	+13.0
×	✓	✓	69.0	71.6	+7.0
✓	✓	✓	**71.5**	**73.6**	+15.0

PGDC proactively rectifies projection errors at object boundaries using 2D priors,which DAGF then propagates into a dense,edge-preserving geometric representation.Finally,SGDM utilizes this refined structure to precisely gate feature projection. Optimal performance is only achieved when these three modules work in synergy; the absence of anyone will lead to a decline in performance.

Results:Robust against poor 2D detection quality

Method	Present at	mAP	NDS	C	T	CV	B	Tr	Ba	M	Bi	P	TC
TransFusion-L [1]	CVPR 22	65.5	70.2	86.2	54.8	26.5	70.1	42.3	72.1	69.8	53.9	86.4	70.3
SAFDNet [41]	CVPR 24	66.3	71.0	87.6	60.8	26.6	78.0	43.5	69.7	75.5	58.0	87.8	75.0
BEVFusion-PKU [15]	NeurIPS 22	67.9	71.0	87.3	59.8	28.9	73.5	41.2	73.8	74.6	59.8	85.4	68.2
LION-Mamba [21]	NeurIPS 24	68.0	72.1	87.9	64.9	28.5	77.6	44.4	71.6	75.6	59.4	89.6	80.8
FSHNet [17]	CVPR 25	68.1	71.7	88.7	61.4	26.3	79.3	47.8	72.3	76.7	60.5	89.3	78.6
BEVFusion-MIT [20]	ICRA 23	68.5	71.4	88.2	61.7	30.2	75.1	41.5	72.5	76.3	64.2	87.5	81.0
UniMamba [11]	CVPR 25	68.5	72.6	88.7	64.7	28.7	**79.7**	**47.9**	72.3	74.6	59.1	89.7	79.5
M3Net [4]	AAAI 25	69.0	72.4	89.0	64.5	30.3	77.9	47.5	73.2	76.5	61.4	89.2	80.4
BEVDiffuser [38]	CVPR 25	69.2	71.9	88.5	63.5	31.0	75.3	46.2	73.2	77.5	62.8	87.9	80.5
GraphBEV [31]	ECCV 24	70.1	72.9	**89.8**	64.2	31.2	75.8	43.5	75.6	79.3	66.3	88.6	80.9
Ours	-	**71.5**	**73.6**	**89.8**	**68.5**	**35.1**	77.2	45.5	**78.0**	**80.5**	**68.3**	**90.1**	**82.0**

2D Prior Source	mAP (%)	NDS (%)
Random Priors	68.5	71.2
No 2D Priors	69.0	71.6
Full-Image Prior	69.4	71.8
YOLO-X Priors	70.3	72.5
Realistic 2D Priors (YOLOv9)	71.5	73.6
Ground Truth 2D Priors	73.5	74.2

(d) Edge weight map G (BEVFusion) (e) Edge weight map G (GraphBEV) (f) Edge weight map G (Ours)

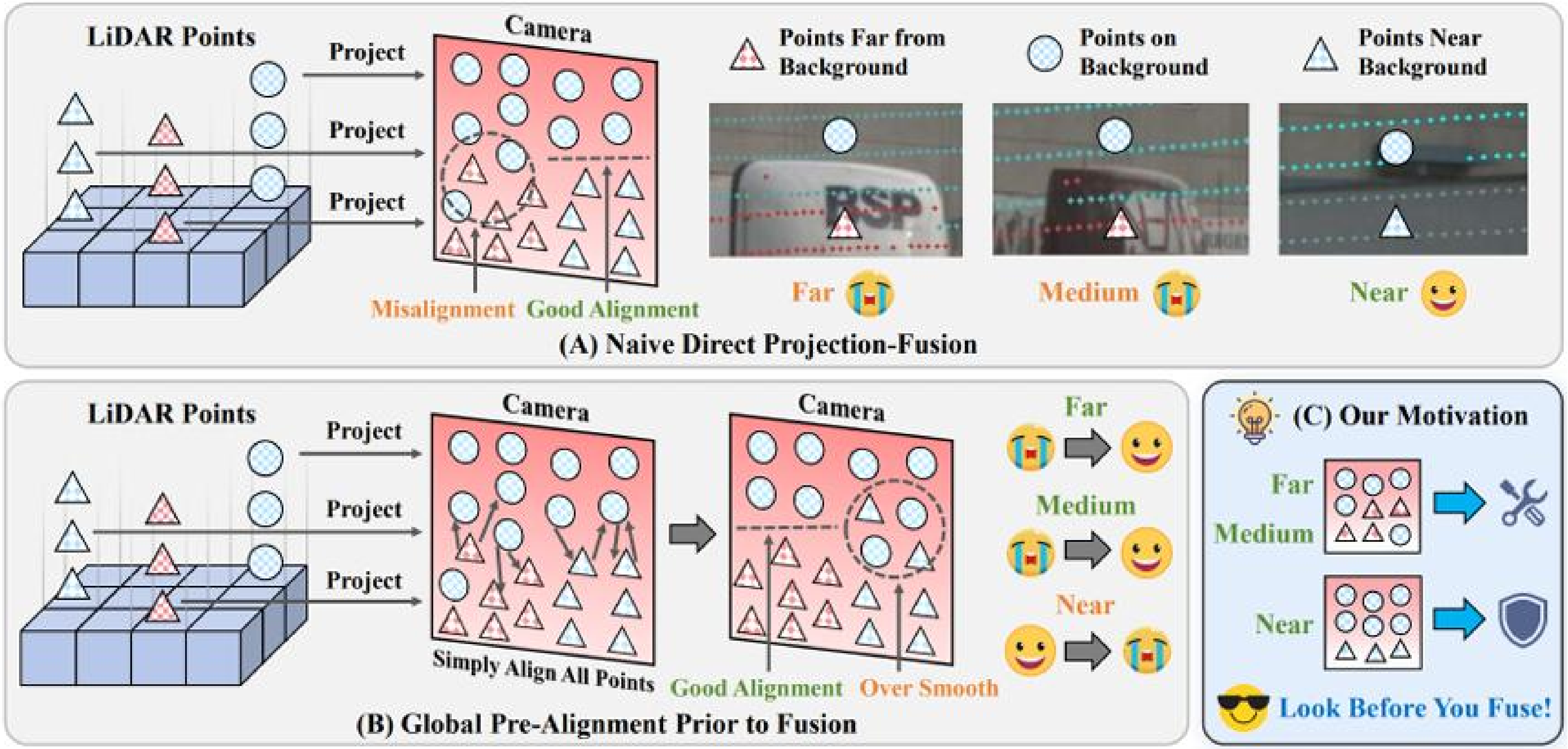
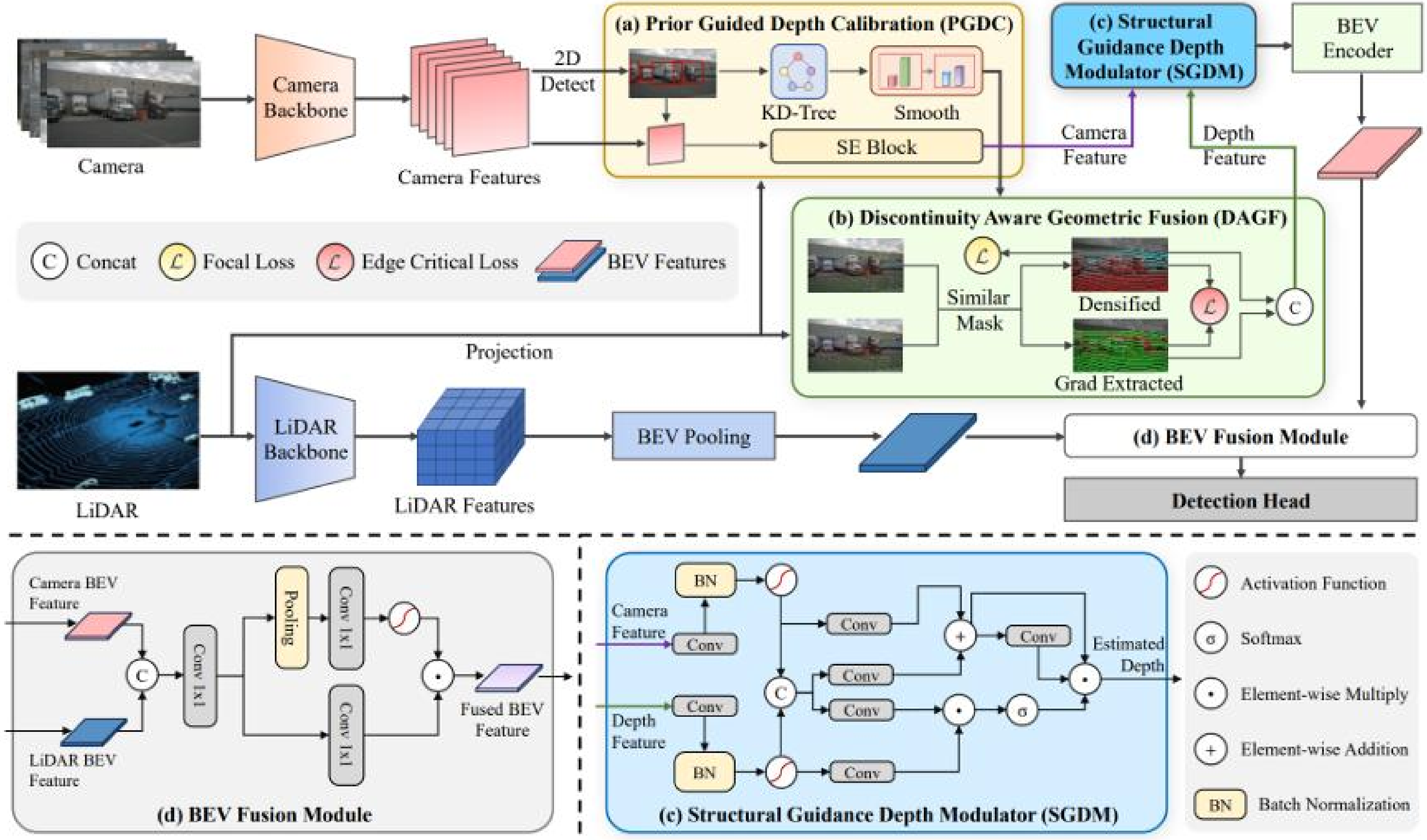


Figure 1. LiDAR points from the distant wall (cyan) are incorrectly projected onto the foreground vehicle (which should appear red) due to a sharp depth change. In contrast, the boundary between background objects like the wall (light cyan) and garage (dark cyan) is correctly projected thanks to a gradual depth transition. This shows that misalignment is most severe at abrupt foreground-background boundaries. And our motivation is to rectify misaligned points while maintaining already aligned points.

Prevailing methods often overlook that misalignment is not randomly distributed, but originates from two primary sources: extrinsic calibration errors and motion induced distortions. The resulting geometric projection er rors are depth-dependent, being negligible for nearby objects but significantly exacerbated at greater depths. This depth-dependent distortion creates the most severe feature misalignment at boundaries between foreground objects and their backgrounds. Our approach is built upon the core insight that critical misalignments are concentrated at foreground-background boundaries where sharp disconti nuities in depth occur, as illustrated in Fig. 1(A), and we can leverage robust 2D object priors to identify these specific regions, locating real misalignment and accurately solving it while maintaining already aligned regions, as illustrated in Fig. 1(C).



This paper proposes a multi-modal 3D detection framework based on the "Look Before You Fuse" principle, designed to address the spatial misalignment between LiDAR and camera features through proactive pre-alignment.

The overall architecture consists of three synergistic modules:
First, **Prior Guided Depth Calibration (PGDC)** leverages 2D object detection priors to actively locate and correct erroneous point cloud depth projections at foreground-background boundaries, while simultaneously enhancing image features in the target regions. Second, **Discontinuity Aware Geometric Fusion (DAGF)** denoises and densifies the calibrated sparse depth map, extracting a structurally aware dense geometric representation that accurately captures sharp depth transitions (e.g., object edges). Finally, the **Structural Guidance Depth Modulator (SGDM)** employs a gated attention mechanism to intelligently fuse the enhanced visual features with the dense geometric cues, predicting a highly accurate pixel-level depth distribution.
Building on this foundation, the system accurately projects image features into a unified Bird's Eye-View (BEV) space and fuses them with LiDAR BEV features, ultimately achieving 3D object detection that is both highly accurate and robust.

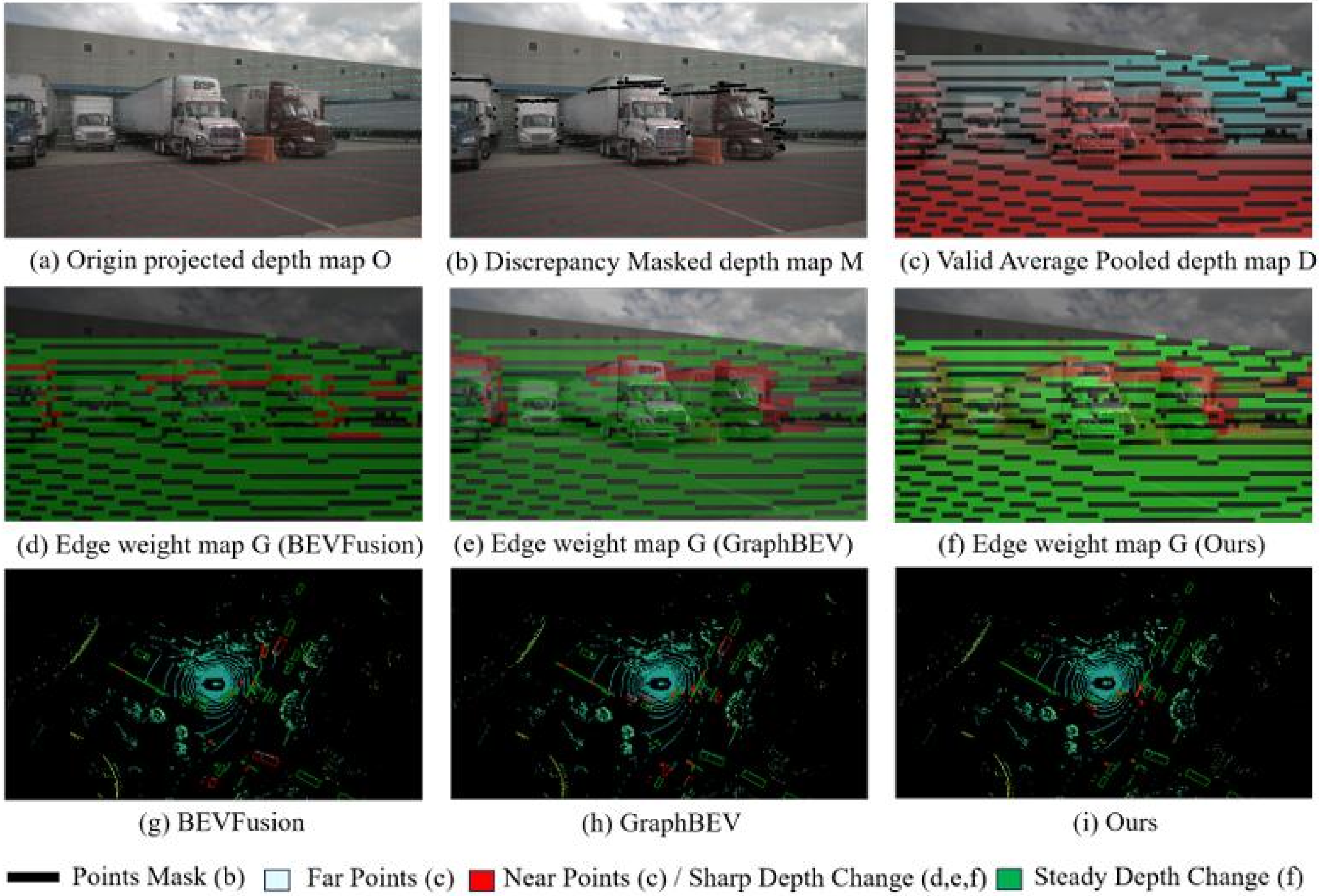


Figure 4. (a) Represents the original projected depth. (b) Represents the projected depth after applying Discrepancy Masking. (c) Shows the block-wise depth map after Block-based Densification; (d), (e), and (f) are the final depth change magnitude maps of different methods after Block-based Gradient Extraction. (g), (h) and (i) are visualization of detection results, in which green boxes are True Positives (TP), solid red boxes are False Positives (FP), and dashed red boxes are False Negatives (FN). It is obvious that our method (f) outperforms GraphBEV (e) and BEVFusion (d) because we accurately delineated regions with drastic depth variations and avoided over-smoothing.

Table 1. Performance comparison of 3D object detection methods on the nuScenes validation set across 10 classes: Car (C), Truck (T), Construction Vehicle (CV), Bus (B), Trailer (Tr), Barrier (Ba), Motorcycle (M), Bicycle (Bi), Pedestrian (P), and Traffic Cone (TC).

Method	Present at	mAP	NDS	C	T	CV	B	Tr	Ba	M	Bi	P	TC
TransFusion-L [1]	CVPR 22	65.5	70.2	86.2	54.8	26.5	70.1	42.3	72.1	69.8	53.9	86.4	70.3
SAFDNet [41]	CVPR 24	66.3	71.0	87.6	60.8	26.6	78.0	43.5	69.7	75.5	58.0	87.8	75.0
BEVFusion-PKU [15]	NeurIPS 22	67.9	71.0	87.3	59.8	28.9	73.5	41.2	73.8	74.6	59.8	85.4	68.2
LION-Mamba [21]	NeurIPS 24	68.0	72.1	87.9	64.9	28.5	77.6	44.4	71.6	75.6	59.4	89.6	80.8
FSHNet [17]	CVPR 25	68.1	71.7	88.7	61.4	26.3	79.3	47.8	72.3	76.7	60.5	89.3	78.6
BEVFusion-MIT [20]	ICRA 23	68.5	71.4	88.2	61.7	30.2	75.1	41.5	72.5	76.3	64.2	87.5	81.0
UniMamba [11]	CVPR 25	68.5	72.6	88.7	64.7	28.7	**79.7**	**47.9**	72.3	74.6	59.1	89.7	79.5
M3Net [4]	AAAI 25	69.0	72.4	89.0	64.5	30.3	77.9	47.5	73.2	76.5	61.4	89.2	80.4
BEVDiffuser [38]	CVPR 25	69.2	71.9	88.5	63.5	31.0	75.3	46.2	73.2	77.5	62.8	87.9	80.5
GraphBEV [31]	ECCV 24	70.1	72.9	**89.8**	64.2	31.2	75.8	43.5	75.6	79.3	66.3	88.6	80.9
Ours	-	**71.5**	**73.6**	**89.8**	**68.5**	**35.1**	77.2	45.5	**78.0**	**80.5**	**68.3**	**90.1**	**82.0**

Table 3. Ablation study of our three proposed modules on nuscenes dataset. LT is an abbreviation for latency.

PGDC	DAGF	SGDM	mAP (%)	NDS (%)	LT (ms)
×	×	×	67.9	71.0	+0.0
✓	×	✓	69.8	72.5	+13.0
×	✓	✓	69.0	71.6	+7.0
✓	✓	✓	**71.5**	**73.6**	+15.0

Table 5. Ablation study on the impact of 2D detectors.

2D Prior Source	mAP (%)	NDS (%)
Random Priors	68.5	71.2
No 2D Priors	69.0	71.6
Full-Image Prior	69.4	71.8
YOLO-X Priors	70.3	72.5
Realistic 2D Priors (YOLOv9)	71.5	73.6
Ground Truth 2D Priors	73.5	74.2

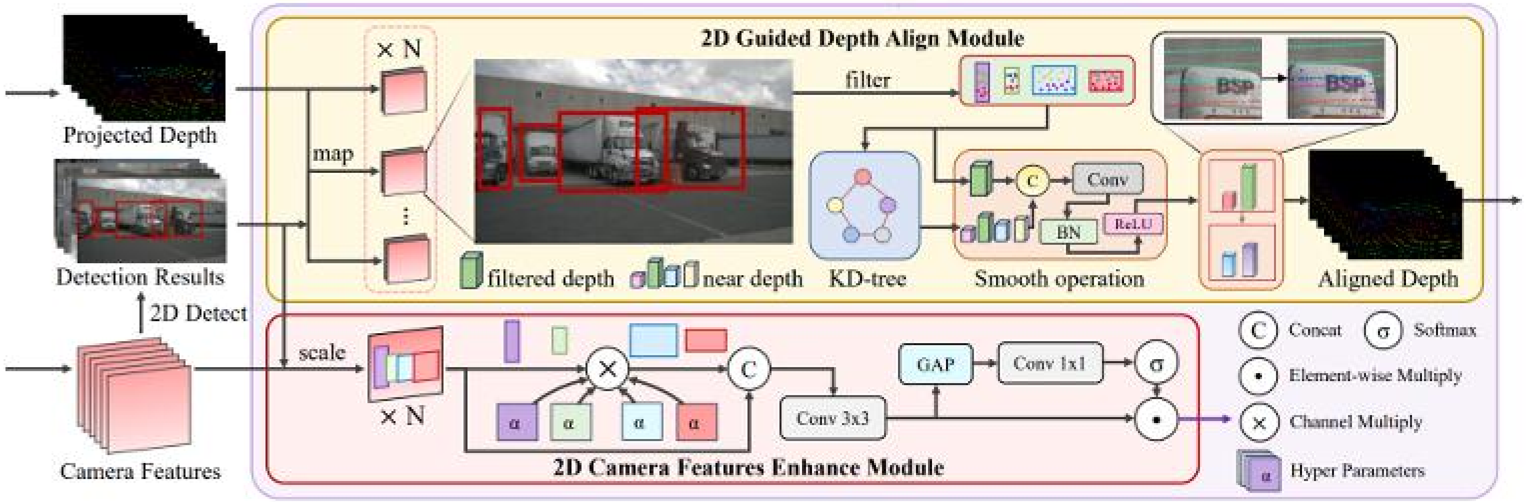
Limitation

Table 5. Ablation study on the impact of 2D detectors.

2D Prior Source	mAP (%)	NDS (%)
Random Priors	68.5	71.2
No 2D Priors	69.0	71.6
Full-Image Prior	69.4	71.8
YOLO-X Priors	70.3	72.5
Realistic 2D Priors (YOLOv9)	71.5	73.6
Ground Truth 2D Priors	73.5	74.2

Table 6. Impact of different 2D prior qualities on PGDC module configurations.

Module	2D Prior Source	mAP (%)	NDS (%)
DAM	No Priors	69.0	71.6
	Random Priors	69.1	71.8
	Full-Image Prior	69.9	72.0
Full Model	No Priors	69.0	71.6
	Random Priors	68.5	71.2
	Full-Image Prior	69.4	71.8
FEM	No Priors	69.0	71.6
	Random Priors	67.4	70.1
	Full-Image Prior	68.6	71.3



FEM is a little tricky and should be considered carefully before put into use. Especially when faced with real world.
But DAM is quite good, and can be used in the worst poor 2D detection condition.