UniBEV: Multi-modal 3D Object Detection with Uniform BEV Encoders for Robustness against Missing Sensor Modalities

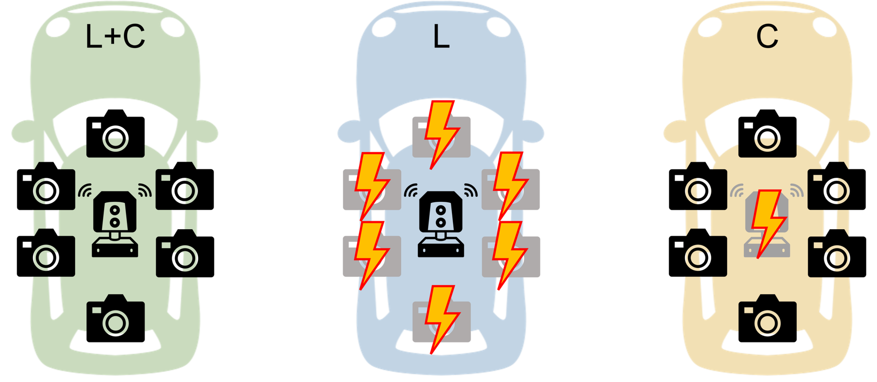
Jeju IV2024 35th IEEE Intelligent Vehicles Symposium

Shiming Wang*,†, Holger Caesar*, Liangliang Nan†, Julian F. P. Kooij*,†
*Intelligent Vehicles Group, †3DUU Lab

TU Delft

Motivation

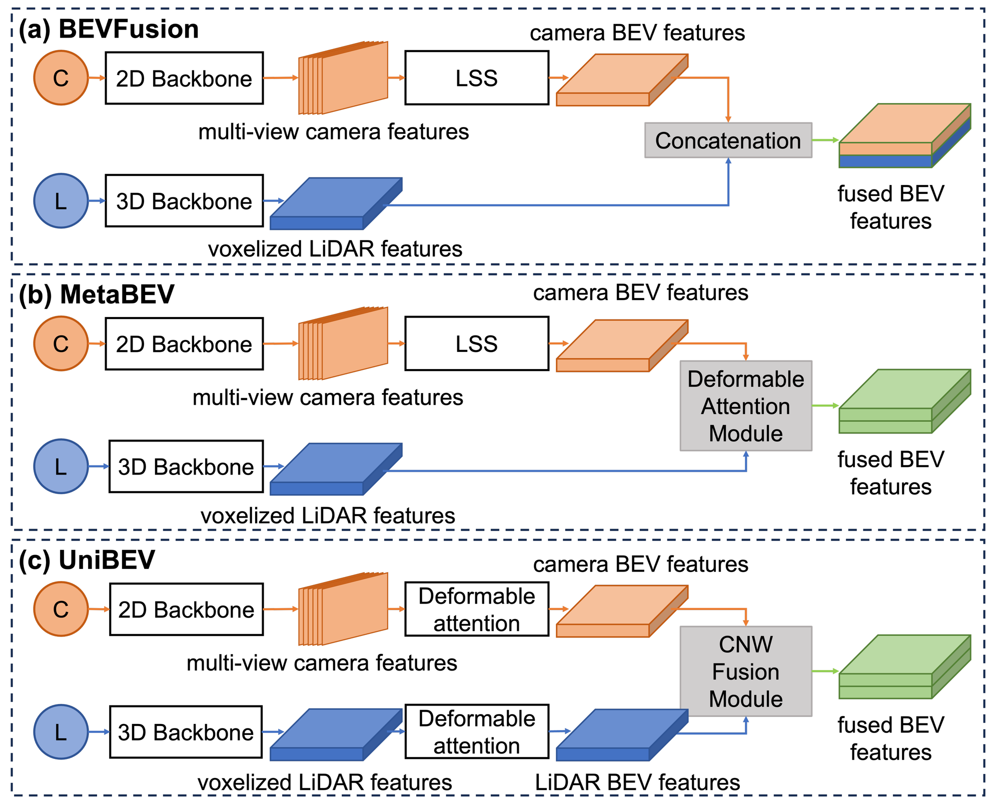
- Task: robust multi-modal fusion for 3D object detection
- Existing approaches fail catastrophically when one sensor modality is missing
- **Our goal:** a robust object detector which fuses LiDAR and camera, but also works when one sensor input is missing without needing to load a different set of model parameters



Ideally LiDAR and Camera (L+C) are present.
"Missing modality" cases:
only LiDAR (L) or Camera (C) is available

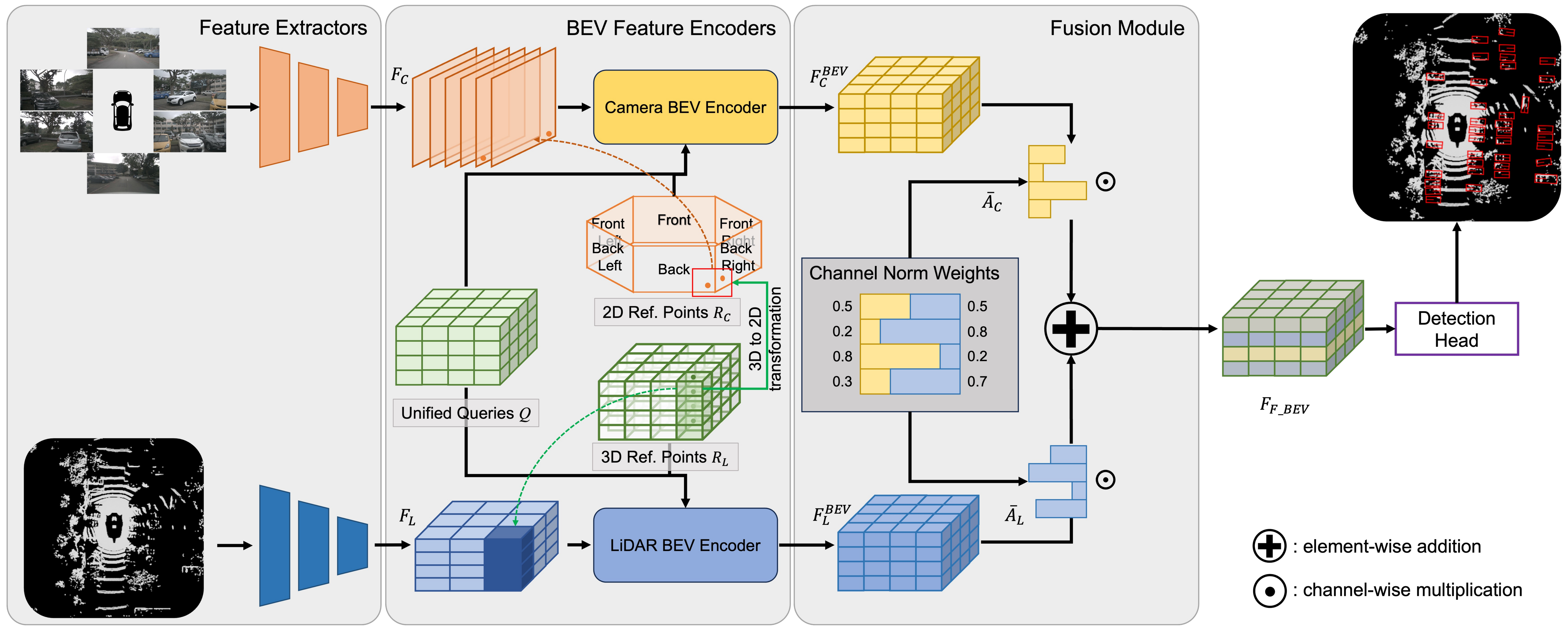
Contributions

- We propose UniBEV, a multi-modal 3D object detector designed for ***robustness against missing modalities***
  - Uniform architecture across sensors, each backbone creates BEV map in a shared feature space
  - BEV maps fused by a new Channel Normalized Weights (CNW) module to align features across modalities, and to learn how much each channel can rely on each modality
- We investigate the impact of various feature fusion strategies: concatenation, averaging and CNW
- We explore the impact of the probability for dropping sensor modalities during the training process



Comparison of our UniBEV and other relevant works

UniBEV Architecture



Key Designs

- Deformable attention-based BEV feature encoders are uniformly applied to all sensor modalities
- Shared queries for shared attention across modalities, helps aligning the feature maps
- CNW computes weighted average of all available (non-missing) sensor BEV maps
  - CNW learns a per-channel weight for each modality, as one modality could be more reliable for fusion
  - When a modality is missing (i.e., sensor failure), CNW does not weigh the BEV map of remaining sensor
- Modality Dropout (MD) training strategy to expose network for 50% of the time to only LiDAR or camera

Quantitative Results

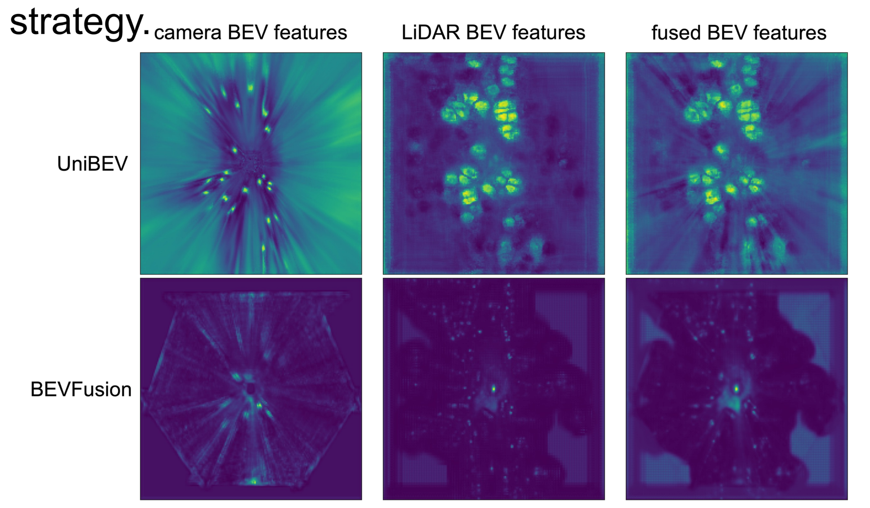
Three test cases: *L*iDAR+*C*amera, *L*iDAR-only, *C*amera-only. Robustness *Summary* averages L+C, L and C.

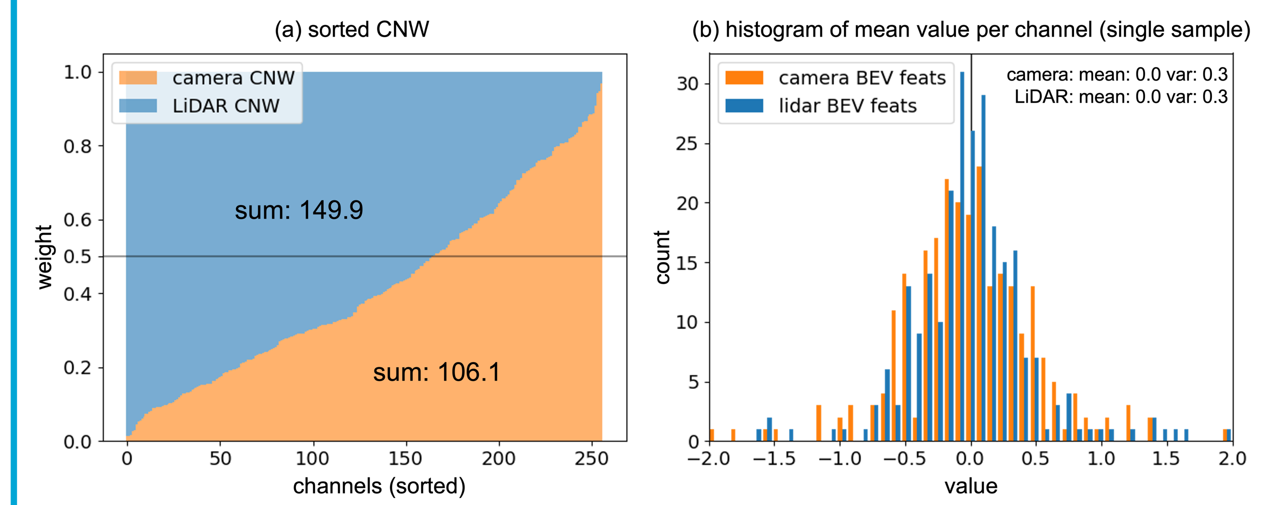
Methods	Training Modality	L+C		L		C		Summary		Inference speed
		NDS	mAP	NDS	mAP	NDS	mAP	NDS	mAP	
BEVFusion [1]	L+C (MD)	65.3	58.7	60.6	49.1	29.6	22.6	51.8	43.5	0.7 FPS
MetaBEV [2]	L+C (MD)	67.5	62.5	65.2	57.8	33.6	25.9	55.4	48.7	1.4 FPS
UniBEV (ours)	L+C (MD)	**68.5**	**64.2**	**65.3**	**58.2**	**42.4**	**35.0**	**58.7**	**52.5**	**1.6 FPS**

*: all methods are training with same protocols, e.g. no CBGS data augmentation and all with MD training strategy.

Visualization of BEV Feature Maps

- UniBEV's camera and LiDAR BEV features more clearly discern similar object locations than BEVFusion [1]
- Lift-Splat-Shoot(LSS)-based BEVFusion [1] and MetaBEV [2] enforce an inductive bias on its camera BEV features not present in its LiDAR features, as exhibited by the hexagon-shaped outline





Visualization of CNW's Learned Weights

(a) CNW's LiDAR weights (149.9) > camera wights (106.1) → More reliance on LiDAR than on camera
(b) Distribution of the average channel activations is the same for both modalities → CNW does not just scale channels to compensate for different magnitudes

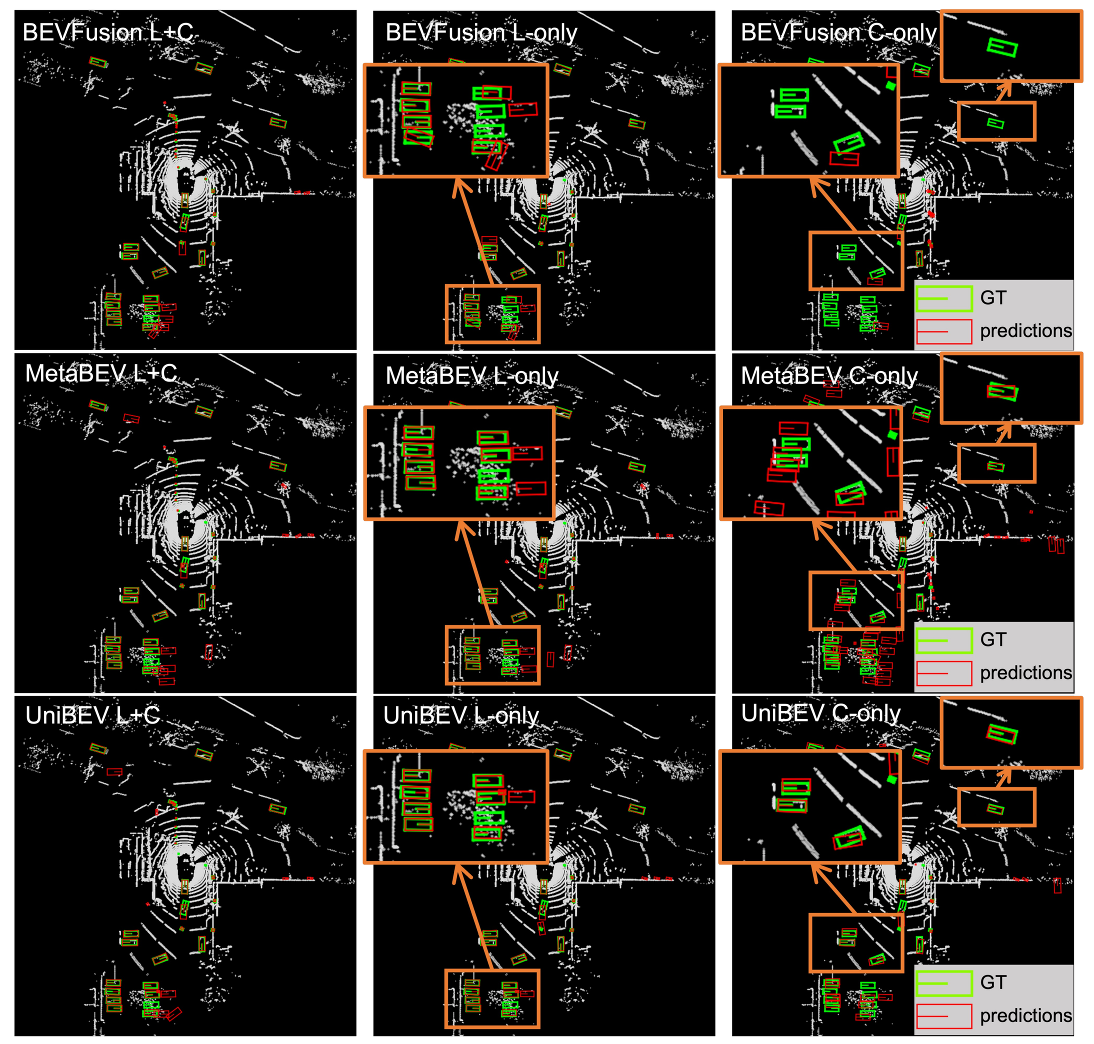
Ablation: different fusion modules in UniBEV

Methods	Encoder Dimensions	mAP			
		L+C	L	C	Summary
Concatenation	128	63.8	57.6	34.4	51.9
Average	256	64.1	57.6	**35.1**	52.3
CNW (ours)	256	**64.2**	**58.2**	35.0	**52.5**

Ablation: probability of dropping L or C during training

p_L	p_C	mAP			
		L+C	L	C	Summary
0	1	63.2	45.5	**36.0**	48.2
0.25	0.75	64.0	57.8	35.8	**52.5**
0.50	***0.50***	**64.2**	58.2	35.0	**52.5**
0.75	0.25	63.8	**58.3**	33.2	51.8
1	0	60.8	55.9	3.0	39.9

Qualitative Results



Conclusions and Insights

- UniBEV is more robust than baselines, achieving 52.5 % mAP on nuScenes (averaged for LiDAR+Camera, Lidar-only and Camera-only test cases) without loss of inference speed
- No trade-off: UniBEV also outperforms SoTA in just the "regular" LiDAR+Camera test setting
- LiDAR is a more informative sensor compared to camera on nuScenes, our CNW module captures this property in its learned fusion weights

Arxiv
GitHub
IV Group
3DUU Lab

Acknowledgement: This work was supported by the 3D Urban Understanding (3DUU) Lab funded by the TU Delft AI Initiative.

[1] Liang, Tingting, et al. BevFusion: A simple and robust lidar-camera fusion framework. NeurIPS 2022.
[2] Ge, Chongjian, et al. MetaBEV: Solving sensor failures for 3d detection and map segmentation. CVPR 2023.