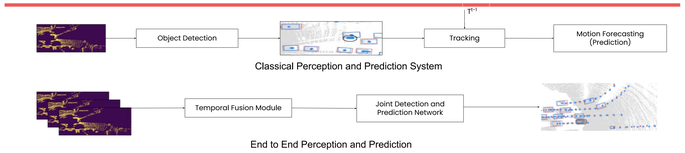
Multi-View Fusion of Sensor Data for Improved Perception and Prediction in Autonomous Driving

Sudeep Fadadu* , Shreyash Pandey* , Darshan Hegde, Yi Shi, Fang-Chieh Chou, Nemanja Djuric, Carlos Vallespi-Gonzalez

WACV WAIKOLOA, HI JAN 4-8

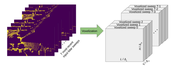
Joint Perception and Prediction



- **Utilizing raw LiDAR sensor** data directly for prediction, yields better performance
- Framework is trained end-to-end for prediction task which **reduces compounding error** between modules and reduces latency

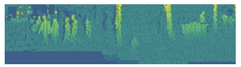
LiDAR Representations for Joint Network

Voxelized Bird's Eye View (BEV)



- Preservation of **metric space and straightforward fusion of historical LiDAR data**.
- **Loss of fine-grained** information

Range View (RV)



- LiDAR's **native, dense representation**, which is very helpful for detecting **small objects**.
- **Fusing historical LiDAR** data in RV is challenging due to distortions

Contributions

- Realizing strengths and weaknesses of these representations, we propose **an efficient data fusion framework**
- Multi-view encoding and processing of LiDAR data separately in BEV and RV frames, before fusing the two views in a common BEV feature space.
- We propose computationally efficient sensor fusion of the **camera RGB data with LiDAR in the RV frame**, before projecting the learned features to the BEV frame.
- This proposal is evaluated against the state-of-the-art **MultiXNet**[1].
  - Fusion method is quite general and can be applied to improve other BEV- and RV-based methods.

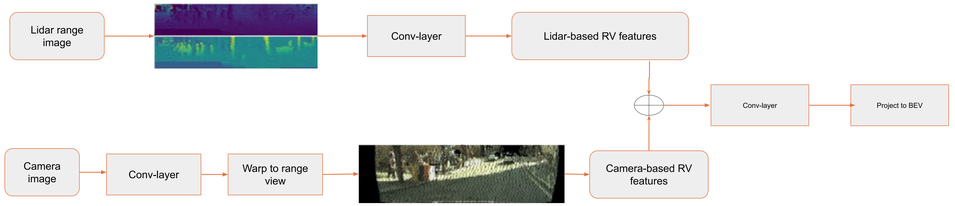
[1] Nemanja Djuric. et al. Multixnet: Multiclass multistage multimodal motion prediction. In IEEE Intelligent Vehicles Symposium (IV), 2021

Multiview Fusion

- BEV is sparsely populated and does not use raw LIDAR data like intensity.
- RV methods operate in LiDAR's native, dense representation, providing full access to the non-quantized sensor information.
- We hypothesize that RV representation is also better than BEV for efficiently fusing information from sensors that natively capture data in RV, like camera.
- To fuse these multiple views together, we propose a point-based feature projection which gathers features from multiple views based on point location in each view.

$$F_t(x) = \frac{\sum_{i=1}^{N} 1_{x=\mathcal{P}_t(L_i)} F_s\big(\mathcal{P}_s(L_i)\big)}{\sum_{i=1}^{N} 1_{x=\mathcal{P}_t(L_i)}}.$$

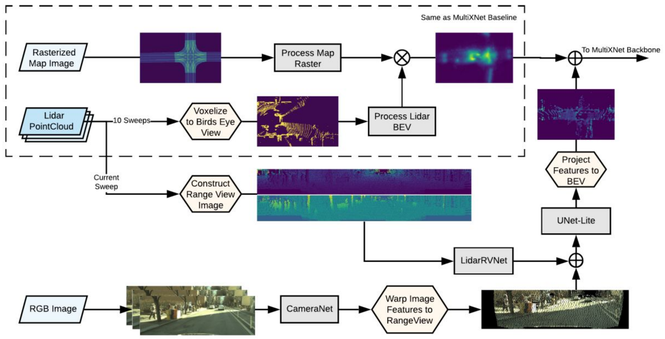
Here F_t and F_s are feature maps in the target and source views, respectively, represented as 2D grids and indexed by a cell index, x denotes a cell index within the target grid, scalar N is the total number of LiDAR points in a sweep, while 1_c is an indicator function that equals 1 if condition c holds true and 0 otherwise. $\mathcal{P}_t$ and $\mathcal{P}_s$ are projection operators that project LiDAR point L_i onto the corresponding view and return the cell index in that view.



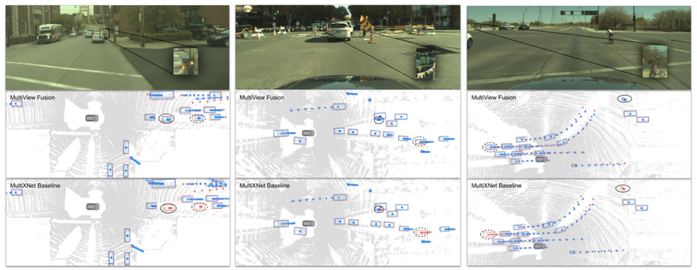
- We use similar method to LaserNet++[2] to associate range view pixels to camera features by projecting lidar points into the camera image

[2] Gregory P. Meyer, et al. Sensor fusion for joint 3d object detection and semantic segmentation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019

Proposed Fusion Architecture applied to MultiXNet[1]



Qualitative Results



Qualitative comparison of ours (middle raw) and MultiXNet (bottom)

Quantitative Results and Analysis

- Proposed method shows **improved detection and prediction** metrics over MultiXNet for all classes.
- The improvement is more **pronounced on smaller objects** which benefit from the fusion of high resolution data.
- Camera fusion is compared against continuous fusion (ContFuse[3]) which is the current state-of-the-art for LiDAR-Camera fusion. Proposed method **outperforms our implementation of ContFuse** by a considerable margin.
- The inference latency increases only slightly and the proposed fusion is **much faster than ContFuse**.
- We see significant **increase for actors at a longer ranges**.

Table 1: Evaluation on ATG4D using detection AP (%), prediction DE (cm), and latency (ms)

	Vehicles					Pedestrians					Bicyclists					
	$AP_{0.7}$ ↑				DE↓	$AP_{0.1}$ ↑				DE↓	$AP_{0.3}$ ↑				DE↓	Lat.↓
	Full	In Camera FOV				Full	In Camera FOV				Full	In Camera FOV				
Method		0-75m	0-25	25-50	50-75		0-75m	0-25	25-50	50-75		0-75m	0-25	25-50	50-75		
MultiXNet	**84.2**	83.9	**92.6**	85.5	68.6	80	88.6	88.9	87.3	89.5	88.0	57	84.6	82.2	87.4	78.7	**76.8**	**49**	35.2
LC-MV-Lite	**84.7**	**84.5**	**92.6**	**86.1**	**70.1**	**79**	**89.4**	**89.8**	**88.5**	**90.1**	**88.7**	**57**	**87.3**	**85.2**	**90.8**	**81.5**	**76.8**	**49**	43.4

Table 2: Evaluation on nuScenes using detection AP (%), prediction DE (cm), and latency (ms)

	Vehicles					Pedestrians					Bicyclists					
	$AP_{0.7}$ ↑				DE↓	$AP_{0.1}$ ↑				DE↓	$AP_{0.3}$ ↑				DE↓	Lat.↓
	Full	In Camera FOV				Full	In Camera FOV				Full	In Camera FOV				
Method		0-50m	0-25	25-50			0-50m	0-25	25-50			0-50m	0-25	25-50		
MultiXNet	60.6	60.9	80.0	43.3	108	65.5	63.8	72.9	54.1	87	31.7	32.1	41.4	25.2	191	32.3
ContFuse	60.9	61.9	79.4	46.0	109	67.2	71.6	78.0	64.3	82	33.5	44.1	53.5	36.1	188	63.6
LC-MV	61.1	61.5	79.6	45.3	**107**	71.0	70.4	79.1	60.6	82	38.2	38.1	53.3	25.8	187	37.4
LC-MV-ResNet	**62.4**	**63.7**	**80.8**	**48.6**	**107**	**72.6**	**76.7**	**83.5**	**69.5**	**73**	**40.9**	**52.3**	**64.5**	**43.2**	204	46.2
LC-MV-Lite	**62.9**	**63.3**	**81.2**	47.0	**107**	71.4	73.1	80.4	64.6	80	39.8	40.4	53.6	29.5	**179**	38.3

[3] Ming Liang,et al. Deep continuous fusion for multi-sensor 3d object detection. In Proceedings of the European Conference on Computer Vision (ECCV)

Authors are with Aurora Innovation, work done while at Uber ATG.
* Equal Contribution