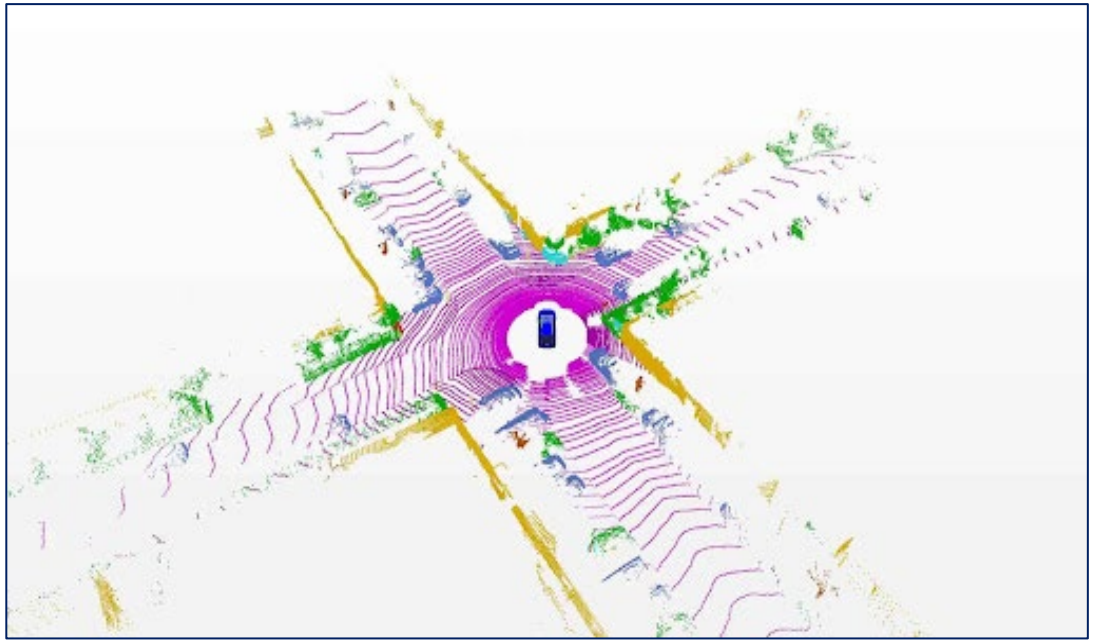
Perception-aware Multi-sensor Fusion for 3D LiDAR Semantic Segmentation

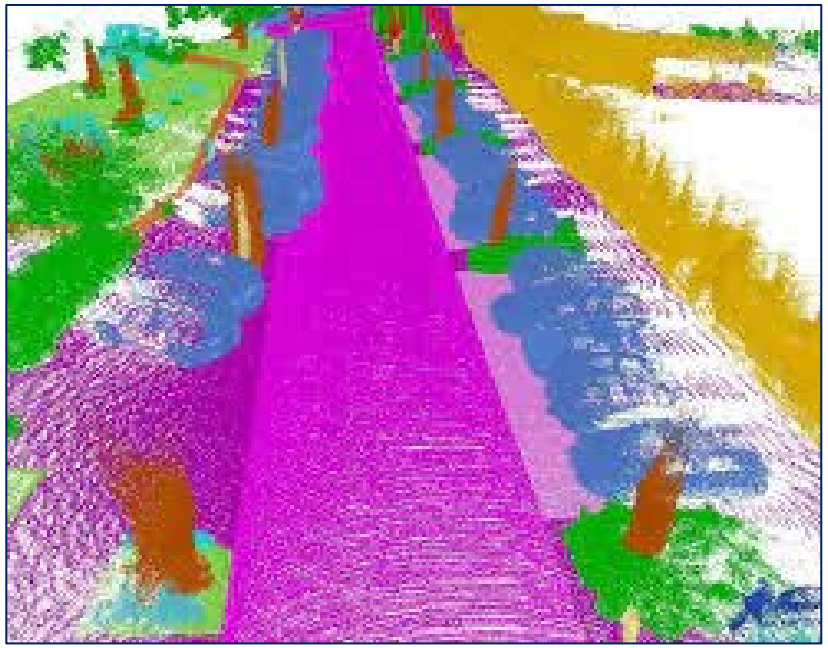
Zhuang, Z., Li, R., et al.

ICCV (2021)

Contents

Background





➢ 3D Semantic Segmentation

- Assign a class label to each data point in the input data
- Help the cars understand the environment, so as to make better decision planning

Behley, Jens, et al. "Semantickitti: A dataset for semantic scene understanding of lidar sequences." ICCV. 2019.

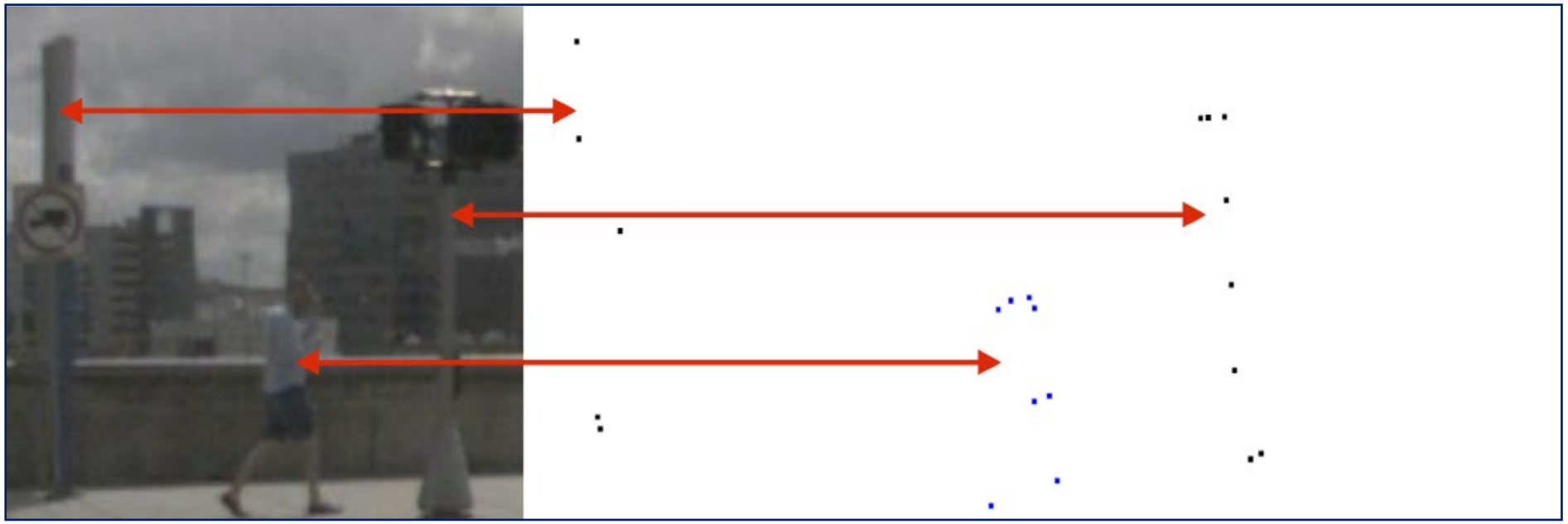
Motivation

Night

Tunnel

- **Camera-Based Methods**
- Dense information (i.e., texture and color)
- Easily disturbed by light (i.e., darkness and overexposure), cause safety issues

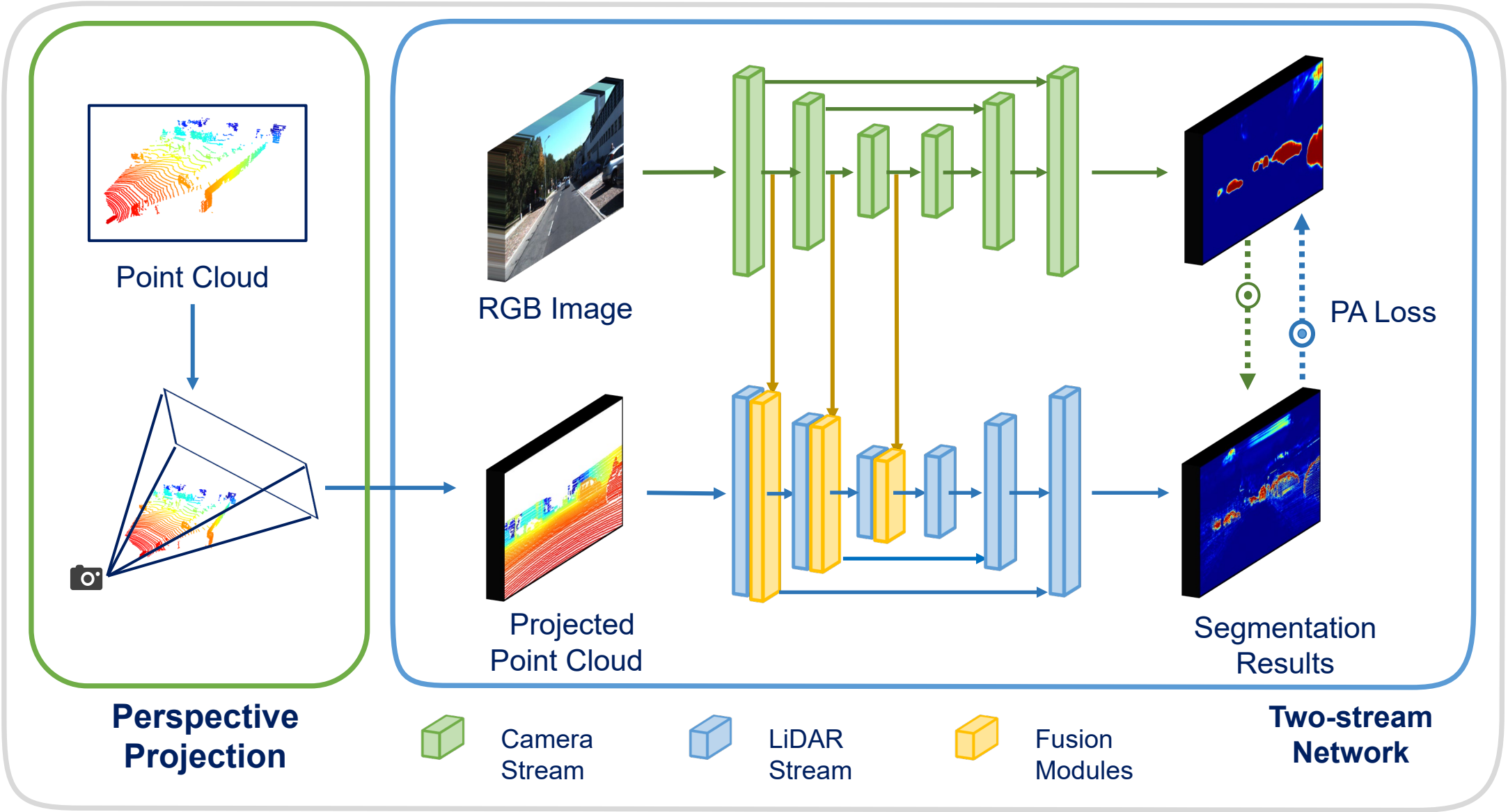
Motivation



➢ **LiDAR-Based Methods**

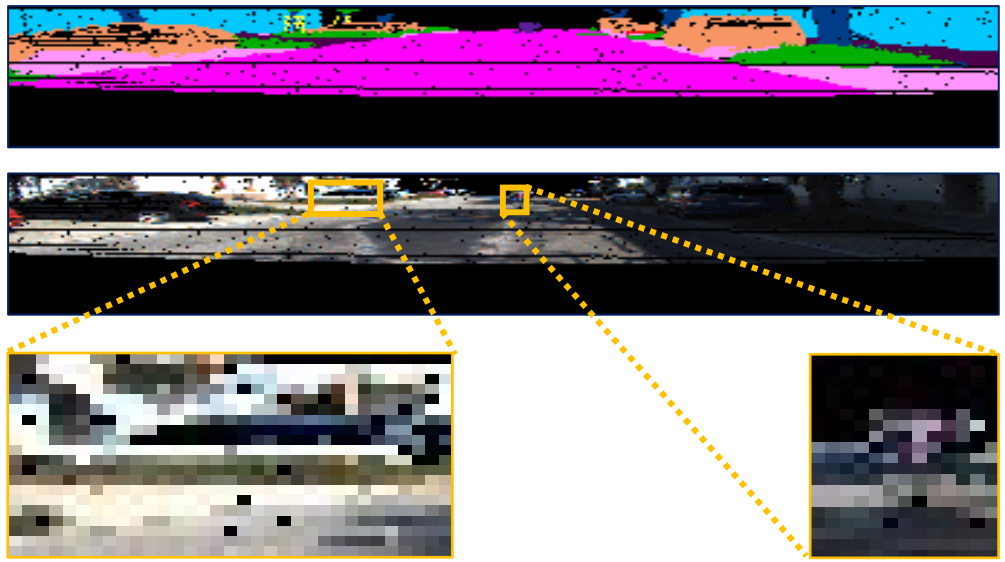
- Highly robust to light variations
- Sparse information, cannot distinguish objects with similar shape

Method

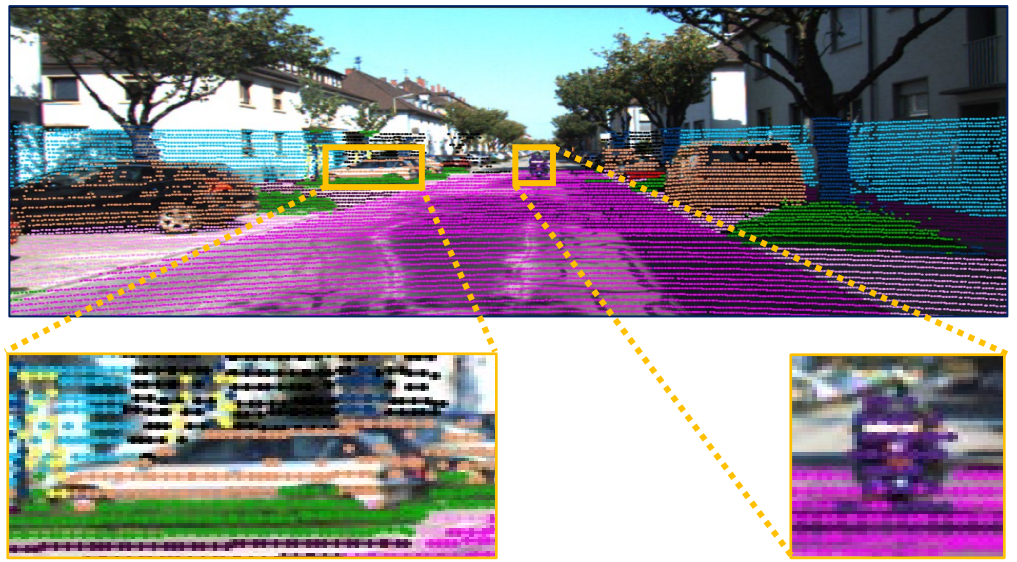

Point Cloud
RGB Image
PA Loss
Projected
Point Cloud
Segmentation
Results
Perspective
Projection
Camera
Stream
LiDAR
Stream
Fusion
Modules
Two-stream
Network

Method

➢ Perception-Aware Loss



Spherical Projection

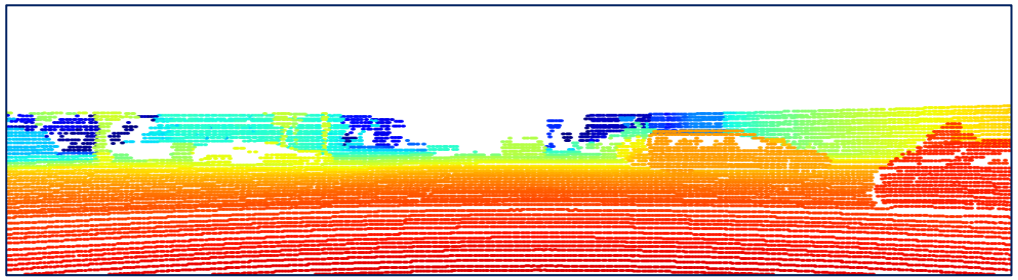


Perceptual Projection (Ours)

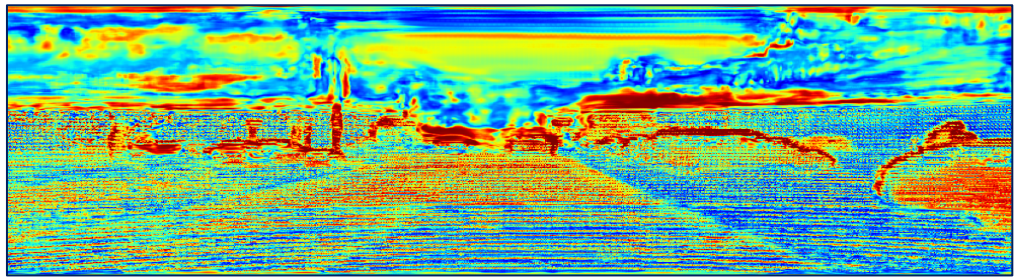
- Existing methods project image into point cloud space for processing, and lose a lot of image perception information (texture, shape)
- We project the point cloud into the image space and performing feature fusion

Method

➢ Perception-Aware Loss

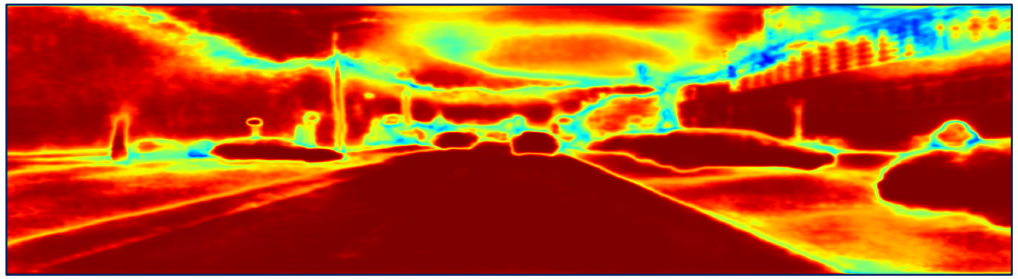


Projected Point Cloud



Confidence Map

RGB Image



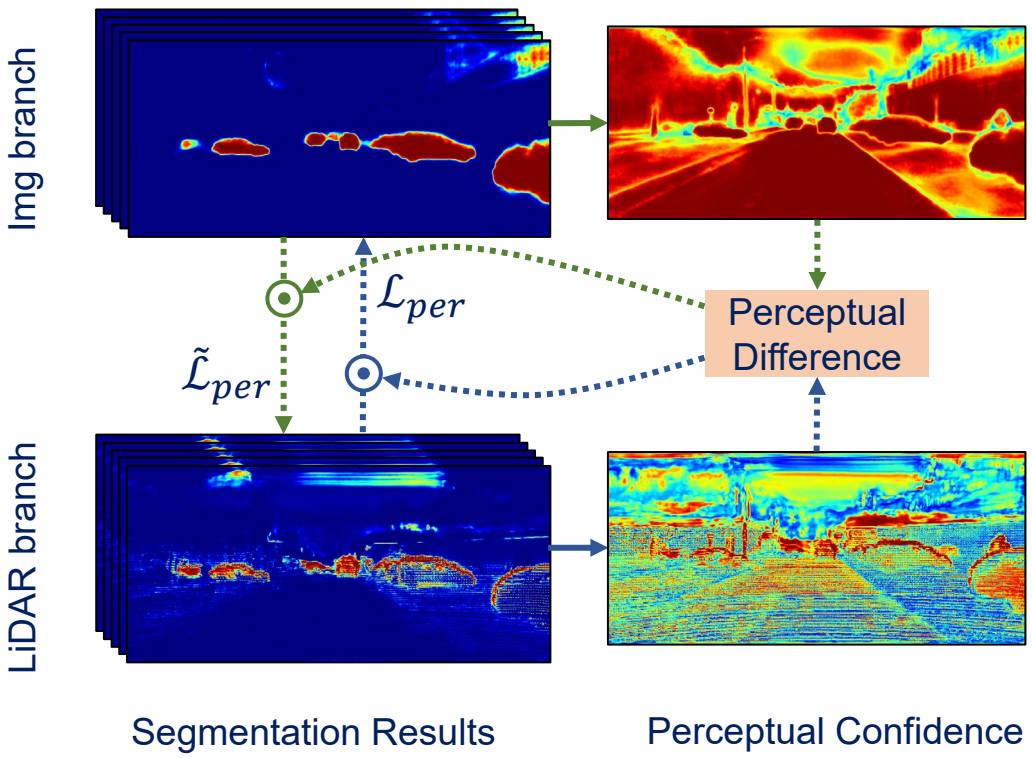
Confidence Map

Advantages: Enable the use of dense information of image

Disadvantages: Only capture local feature, because point cloud is very sparse

Method

➢ Perception-Aware Loss



Definition of **Entropy**

$$\widetilde{\mathbf{E}}_{h,w} = -\frac{1}{\log S}\sum_{s=1}^{S}\widetilde{\mathbf{O}}_{s,h,w}\log(\widetilde{\mathbf{O}}_{s,h,w})$$

The **perceptual confidence**

$$\tilde{\mathbf{C}} = \mathbf{1} - \tilde{\mathbf{E}}$$

Mutual difference

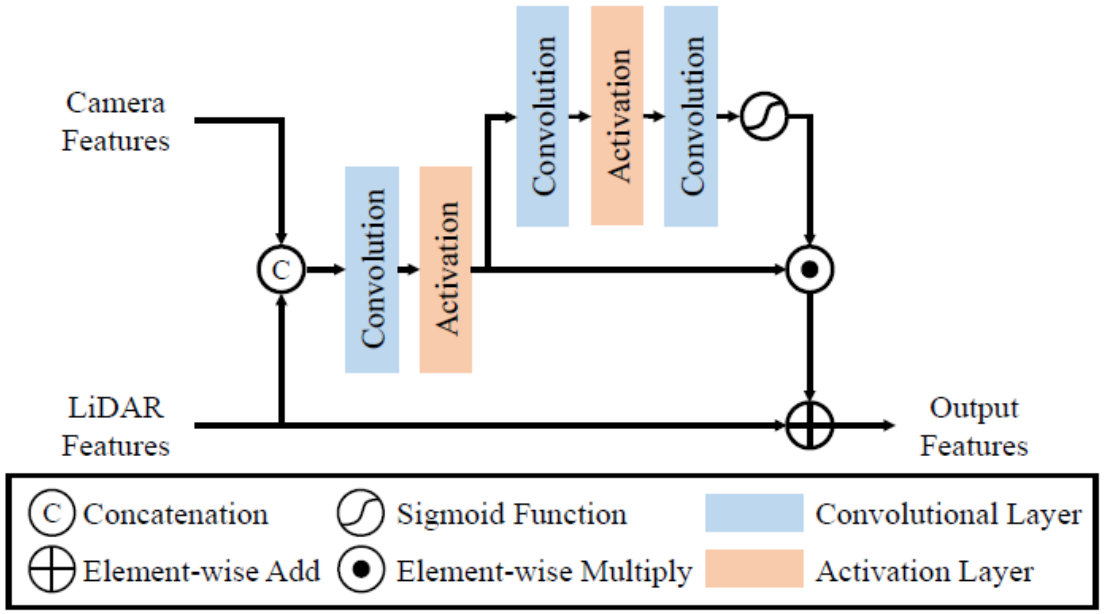
$$\widetilde{\Omega}_{h,w} = \begin{cases} \max(\widetilde{\mathbf{C}}_{h,w} - \mathbf{C}_{h,w}, 0), & \text{if } \widetilde{\mathbf{C}}_{h,w} > \tau, \\ 0, & \text{otherwise.} \end{cases}$$

Perception-aware loss

$$\widetilde{\mathcal{L}}_{per} = \frac{1}{Q}\sum_{h=1}^{H}\sum_{w=1}^{W}\widetilde{\Omega}_{h,w} D_{KL}(\widetilde{\mathbf{O}}_{:,h,w}||\mathbf{O}_{:,h,w}),$$

Method

➢ Residual-Based Network



Residual fusion module

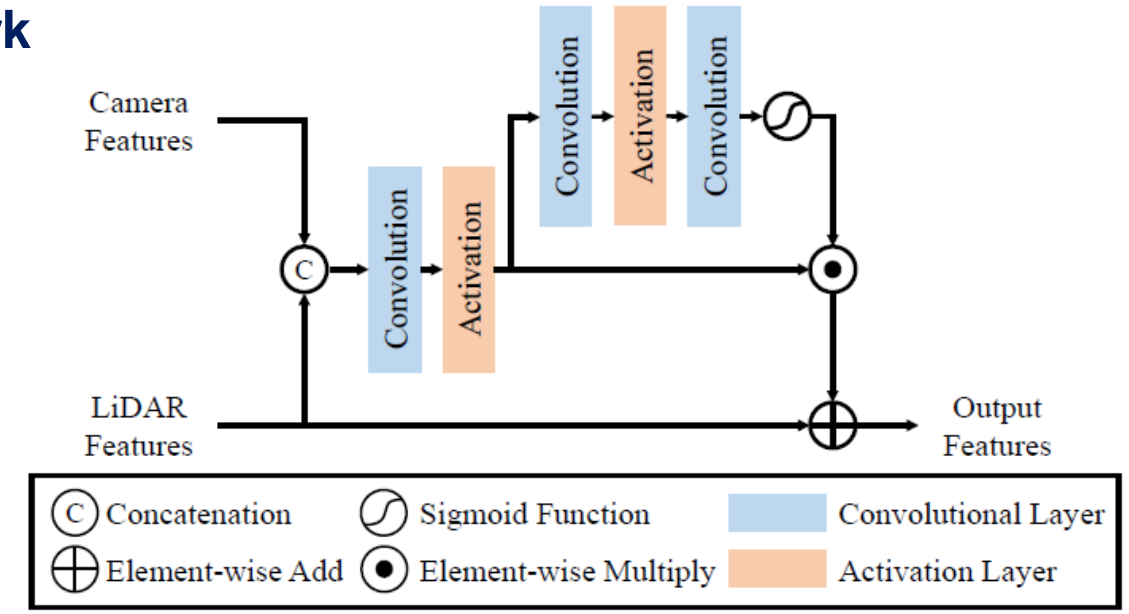
- Fusion feature is the supplement of the original LiDAR feature

Attention module

- Remove the noise of fusion features, because the image feature is easily be affected by light

Method

➢ Residual-Based Network



For l-th fusion module, we have

$$\begin{cases} \mathbf{F}_l^{fuse} = f_l([\tilde{\mathbf{F}}_l; \mathbf{F}_l]) \\ \mathbf{F}_l^{out} = \tilde{\mathbf{F}}_l + \sigma\left(g_l\left(\mathbf{F}_l^{fuse}\right)\right) \odot \mathbf{F}_l^{fuse} \end{cases}$$

where

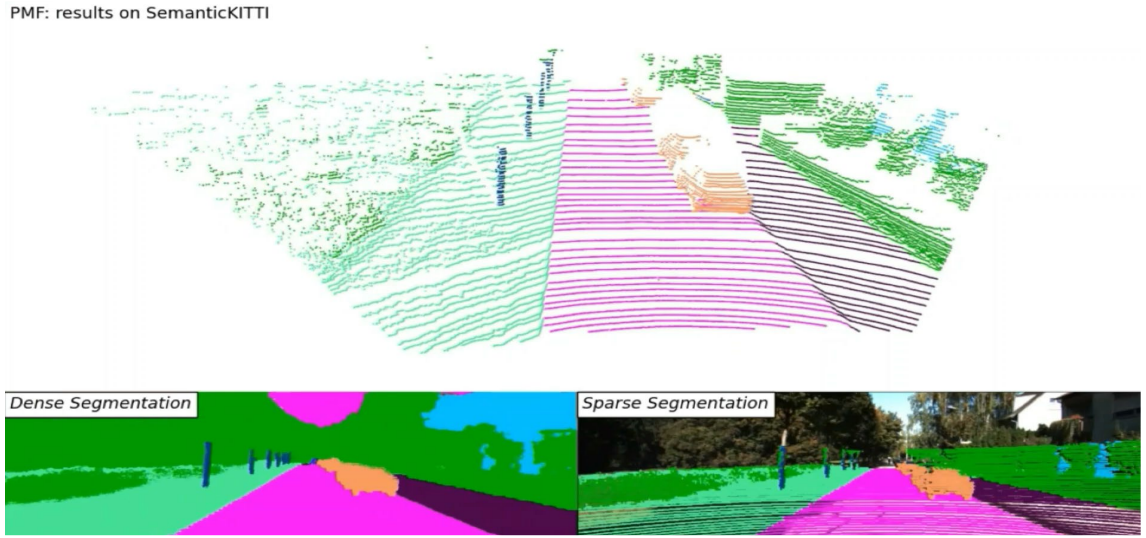
$\tilde{\mathbf{F}}_l$ denotes the features from the LiDAR-stream, $\mathbf{F}_l$ denotes the features from the camera stream, $\mathbf{F}_l^{fuse}$ denotes the fused features, $\mathbf{F}_l^{out}$ denotes output features, $[\cdot;\cdot]$ denotes the concatenation operation, $f_l(\cdot)$ and $g_l(\cdot)$ denote the convolutional layers, $\sigma(\cdot)$ denotes the sigmoid function

Experiment

➢ **Results on SemanticKITTI :**

- Day Scenes
- Only 90° Front View
- 64 Lines LiDAR

(~35k pts/frame)



Method	Input	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign	mIoU (%)
#Points (k)	-	6384	44	52	101	471	127	129	5	21434	974	8149	67	6304	1691	20391	882	8125	317	64	-
RandLANet [22]	L	92.0	8.0	12.8	**74.8**	46.7	52.3	46.0	0.0	93.4	32.7	73.4	0.1	84.0	43.5	83.7	57.3	**73.1**	48.0	27.3	50.0
RangeNet++ [36]	L	89.4	26.5	48.4	33.9	26.7	54.8	69.4	0.0	92.9	37.0	69.9	0.0	83.4	51.0	83.3	54.0	68.1	49.8	34.0	51.2
SequeezeSegV2 [50]	L	82.7	15.1	22.7	25.6	26.9	22.9	44.5	0.0	92.7	39.7	70.7	0.1	71.6	37.0	74.6	35.8	68.1	21.8	22.2	40.8
SequeezeSegV3 [51]	L	87.1	34.3	48.6	47.5	47.1	58.1	53.8	0.0	**95.3**	43.1	**78.2**	0.3	78.9	53.2	82.3	55.5	70.4	46.3	33.2	53.3
SalsaNext [12]	L	90.5	44.6	49.6	**86.3**	54.6	74.0	81.4	0.0	93.4	40.6	69.1	0.0	84.6	53.0	83.6	64.3	64.2	54.4	39.8	59.4
MinkowskiNet [10]	L	95.0	23.9	50.4	55.3	45.9	65.6	82.2	0.0	94.3	**43.7**	76.4	0.0	87.9	57.6	87.4	67.7	71.5	**63.5**	**43.6**	58.5
SPVNAS [46]	L	**96.5**	44.8	**63.1**	59.9	64.3	72.0	**86.0**	0.0	93.9	42.4	75.9	0.0	**88.8**	**59.1**	**88.0**	67.5	73.0	**63.5**	**44.3**	62.3
Cylinder3D [59]	L	**96.4**	**61.5**	**78.2**	66.3	**69.8**	**80.8**	**93.3**	0.0	94.9	41.5	78.0	**1.4**	87.5	50.0	86.7	**72.2**	68.8	63.0	42.1	**64.9**
PointPainting* [47]	L+C	94.7	17.7	35.0	28.8	55.0	59.4	63.6	0.0	**95.3**	39.9	77.6	0.4	87.5	55.1	87.7	67.0	72.9	61.8	36.5	54.5
RGBAL* [33]	L+C	87.3	36.1	26.4	64.6	54.6	58.1	72.7	0.0	95.1	45.6	77.5	**0.8**	78.9	53.4	84.3	61.7	72.9	56.1	41.5	56.2
PMF (Ours)	L+C	95.4	**47.8**	62.9	68.4	**75.2**	**78.9**	71.6	0.0	**96.4**	**43.5**	**80.5**	0.1	**88.7**	**60.1**	**88.6**	**72.7**	**75.3**	**65.5**	43.0	**63.9**

Experiment

- **Results on nuScenes**
- Day & Night Scene
- 360° View
- 16 Lines LiDAR

  (~35k pts/frame)

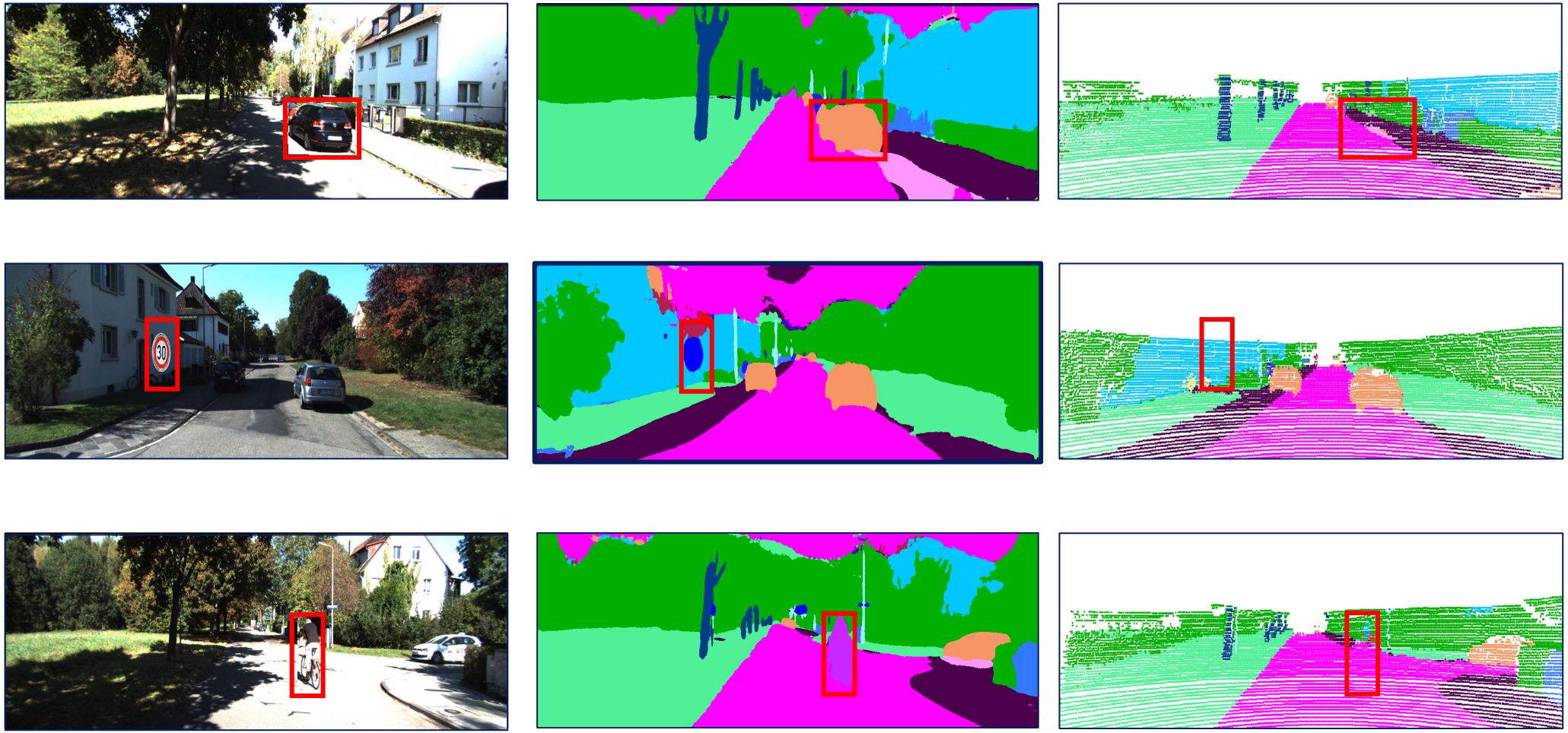
Method	barrier	bicycle	bus	car	construction	motorcycle	pedestrian	traffic-cone	trailer	truck	driveable	other-flat	sidewalk	terrain	manmade	vegetation	mIoU (%)
#Points (k)	1629	21	851	6130	194	81	417	112	370	2560	56048	1972	12631	13620	31667	21948	-
RangeNet++ [36]	66.0	21.3	77.2	80.9	30.2	66.8	69.6	52.1	54.2	72.3	94.1	66.6	63.5	70.1	83.1	79.8	65.5
PolarNet [56]	74.7	28.2	85.3	90.9	35.1	77.5	71.3	58.8	57.4	76.1	96.5	71.1	74.7	74.0	87.3	85.7	71.0
Salsanext [12]	74.8	34.1	85.9	88.4	42.2	72.4	72.2	63.1	61.3	76.5	96.0	70.8	71.2	71.5	86.7	84.4	72.2
Cylinder3D [59]	**76.4**	40.3	**91.3**	**93.8**	51.3	**78.0**	78.9	64.9	62.1	**84.4**	**96.8**	71.6	**76.4**	**75.4**	**90.5**	87.4	76.1
PMF (Ours)	74.1	**46.6**	89.8	92.1	**57.0**	77.7	**80.9**	**70.9**	**64.6**	82.9	95.5	**73.3**	73.6	74.8	89.4	**87.7**	**76.9**

Experiment

- **Adversarial attack experiment**

Experiment

- **Adversarial attack experiment**



Images | Image-Based Results | Ours

Experiment

➢ Ablation Study

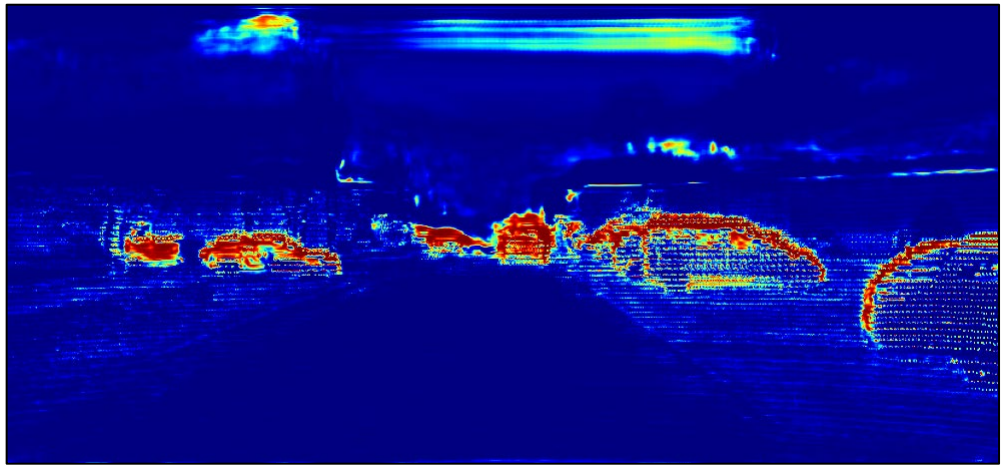
Table. Ablation study for network components on SemanticKITTI validation set. PP denotes perspective projection. RF denotes residual-based fusion module. PL denotes perception-aware loss. The bold number indicates the best result.

Baseline	PP	ASPP	RF	PL	mIoU (%)
✓					57.2
✓	✓				57.6
✓	✓	✓			59.7
✓		✓	✓		55.8
✓	✓	✓	✓		61.7
✓	✓	✓	✓	✓	**63.9**

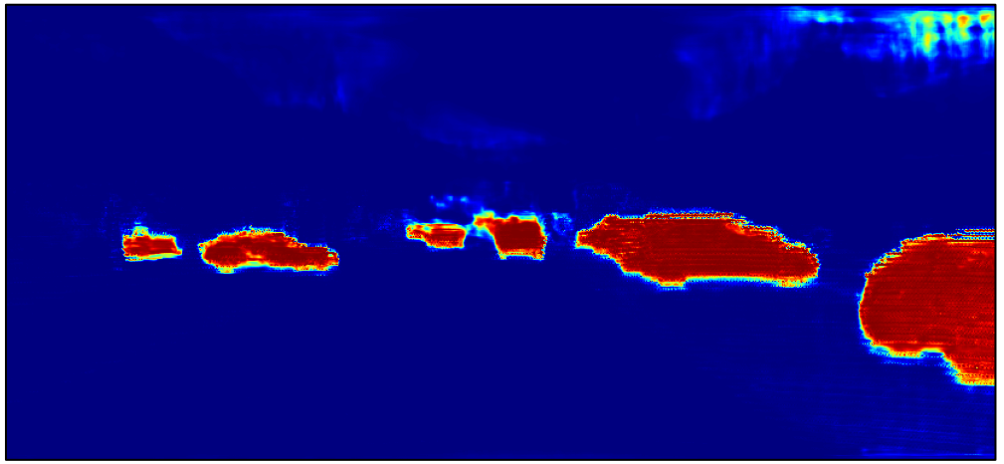
- Perspective projection brings 5.9% improvement over spherical projection with multi-modalities inputs, but only 0.4% improvement with LiDAR only input
- Residual-based fusion modules bring 2.0% improvement to the network
- Perception-aware losses improve the performance of the network by 2.2% in mIoU

Experiment

➢ Ablation Study of Perception-aware Loss



(a) Predictions of Car without PL



(b) Predictions of Car with PL

- Perception-aware losses help the LiDAR-stream to capture the perceptual information from the images
- With perception-aware losses, PMF generates dense predictions that combine both benefits of the images and point clouds

Thanks for Listening