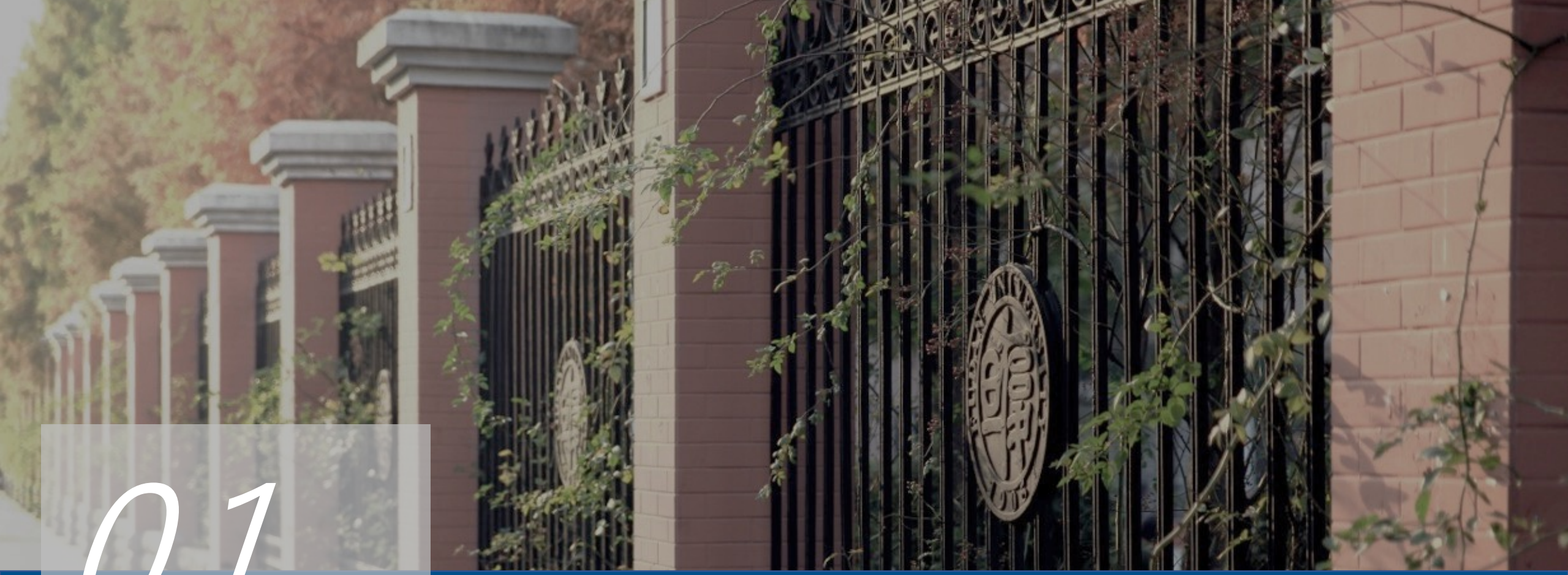
MSMDFusion: Fusing LiDAR and Camera at Multiple Scales with Multi-Depth Seeds for 3D Object Detection

Yang Jiao, Zequn Jie, Shaoxiang Chen, Jingjing Chen,

Lin Ma and Yu-Gang Jiang

Contents



01

- Background

Background

Multi-Sensor Perception

- LiDAR and camera are widely equipped on autonomous driving cars
  - NuScenes Cars – 1 LiDAR & 6 cameras
  - Waymo Cars – 4 LiDAR & 5 cameras

- LiDAR

  - Accurate spatial measurement
  - Long range sparsity

- Camera

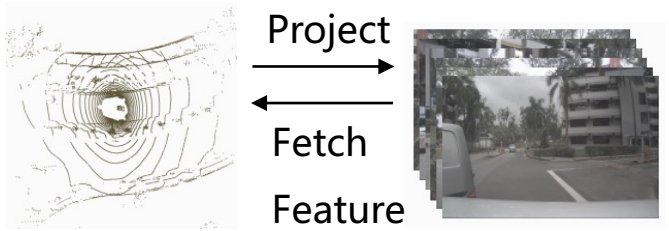
  - Rich semantics
  - Ill-posed depth estimation

Making the best of the two worlds !!!

Background

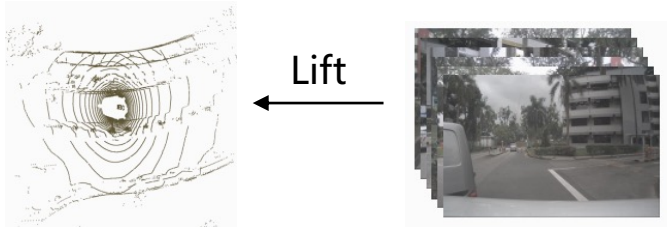
Two LiDAR-camera Fusion Paradigms

- Projecting LiDAR to camera (**L2C**)



  - Accurate position projection
  - Waste informative dense image semantics

- Lifting camera to LiDAR (**C2L**)



  - Complement both geometrics and semantics
  - Suffer from depth estimation error

- Representatives of **L2C** and **C2L**
  - L2C:
    - ✓ PointPainting (CVPR 2020), PointAugmenting (CVPR 2021), TransFusion (CVPR 2022), DeepInteraction (NIPS 2022)
  - C2L
    - ✓ MVP (NIPS 2021), BEVFusion-Alibaba (NIPS 2022), BEVFusion-MIT (ICRA 2023)

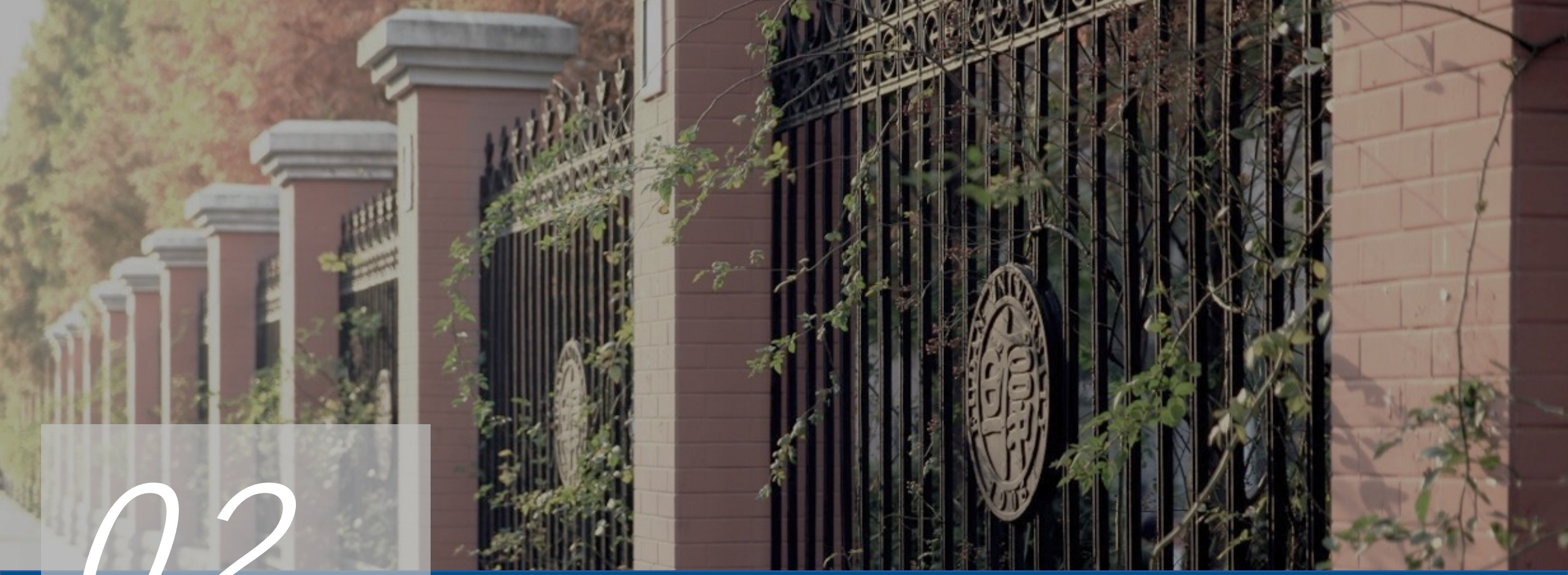
Background

Performance Comparison of L2C and C2L

- Performances on NuScenes detection task

Method	Paradigm	Conference	mAP	NDS
PointAugmenting [1]	L2C	CVPR 2021	66.8	71.0
TransFusion [2]	L2C	CVPR 2022	68.9	71.7
DeepInteraction [3]	L2C	NIPS 2022	70.8	73.4
MVP [4]	C2L	NIPS 2021	66.4	70.5
BEVFusion-ALI [5]	C2L	NIPS 2022	69.2	71.8
BEVFusion-MIT [6]	C2L	ICRA 2023	70.2	72.9

[1] PointAugmenting: Cross-Modal Augmentation for 3D Object Detection (CVPR 2021)
[2] TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers (CVPR 2022)
[3] DeepInteraction: 3D Object Detection via Modality Interaction (NIPS 2022)
[4] Multimodal Virtual Point 3D Detection (NIPS 2021)
[5] BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework (NIPS 2022)
[6] BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation (ICRA 2023)



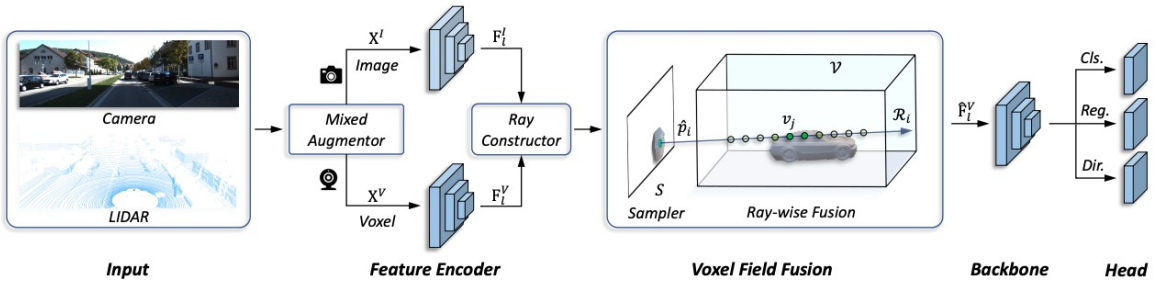
02

- Methodology

Methodology

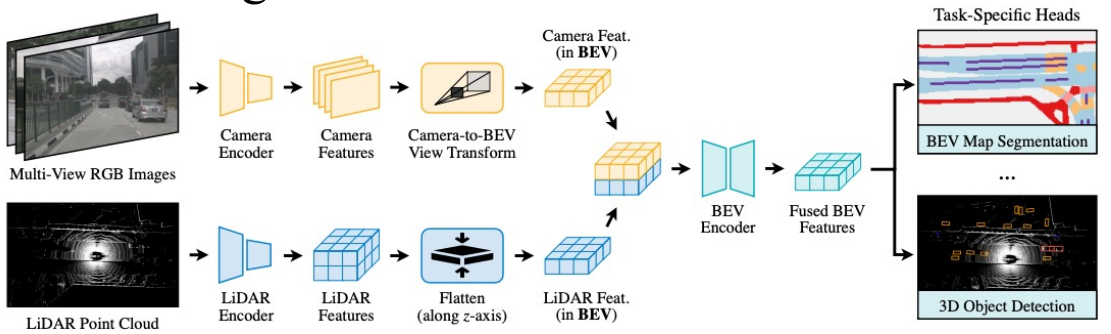
Different "seeds" lift granularities

- Raw pixels as seeds



***Framework of VFF* [7]**

- Feature grids as seeds



***Framework of BEVFusion-MIT* [6]**

- Better geometric description
- Sparse sampling due to computation cost limitation

- Better semantic description
- Information loss due to downsampling

- Depth information in lift "seeds" is under-explored
- Fine-grained cross-modal interaction in voxel space is non-trivial

[7] Voxel Field Fusion for 3D Object Detection (CVPR 2022)

Methodology

Why depth information important

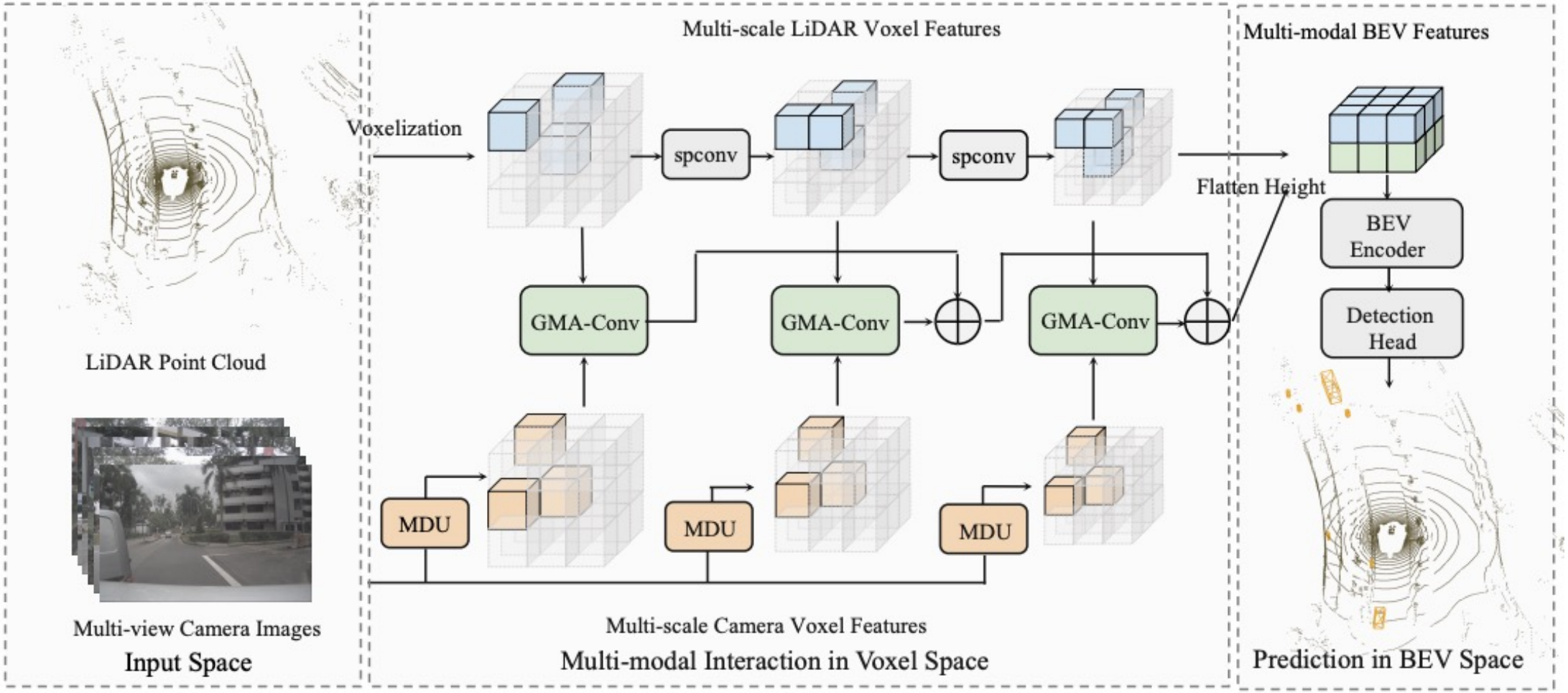
- Determine spatial locations of lifted seeds
- Provide color-insensitive cues

What does our MSMDFusion do

- Provide a multi-scale fusion framework for comprehensive LiDAR geometrics and camera semantics fusion
- Enhance depth reliability (*Multi-Depth Unprojection, MDU*)
- Design a fine-grained LiDAR-camera interaction operation in voxel space for filtering noises (*Gated Modality-Aware Convolution, GMA-Conv*)

Methodology

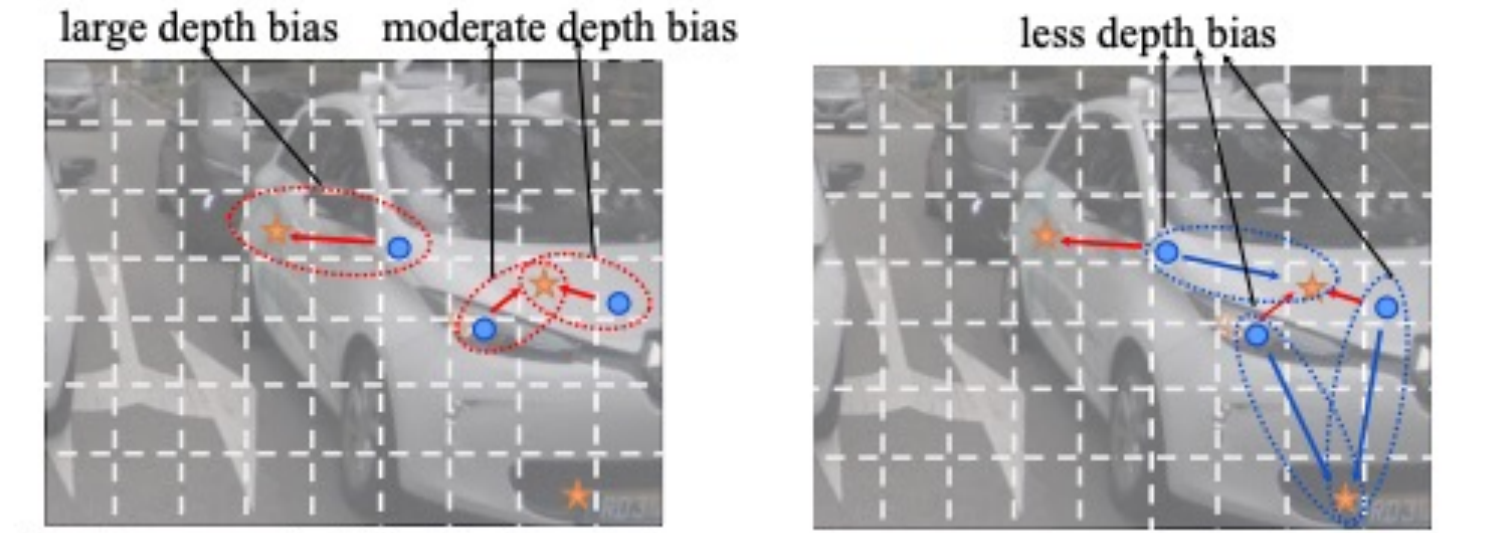
Framework of MSMDFusion



Methodology

Multi-Depth Unprojection (MDU)

- Preliminary: MVP nearest neighbor depth assignment strategy
- Motivation: Spatial proximity in 2D images is not always consistency with that in 3D space
- Solution: K nearest neighbor depth assignment to explore more reliable depth prediction



(a) Nearest neighbor retrieval (b) K Nearest neighbor retrieval (Ours)

Methodology

Gated Modality-Aware Convolution (GMA-Conv)

- Motivation: Lifted seeds are inherently noisy due to inaccurate depth estimation
- Challenge: Designing a fine-grained cross-modal interactor is restricted by huge computation and memory cost

◆ **A simple computation and space complexity estimation**

LiDAR voxels = N

Camera voxels = M

($N, M \approx 10^5$ for a LiDAR frame)

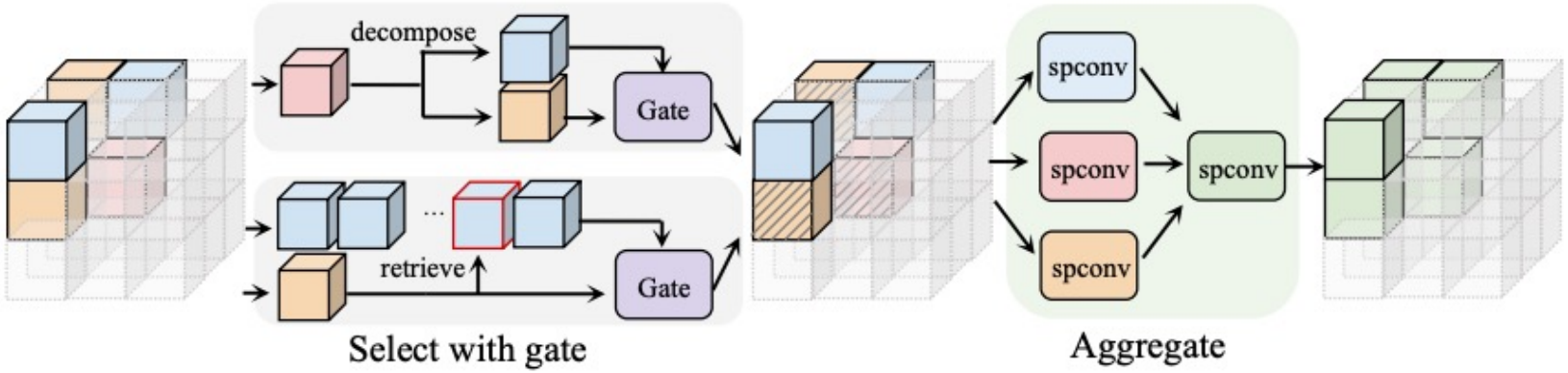
$$\underbrace{\mathcal{O}(N^2), \mathcal{O}(M^2), \mathcal{O}(MN)}_{\textit{Intractable !!!}}$$

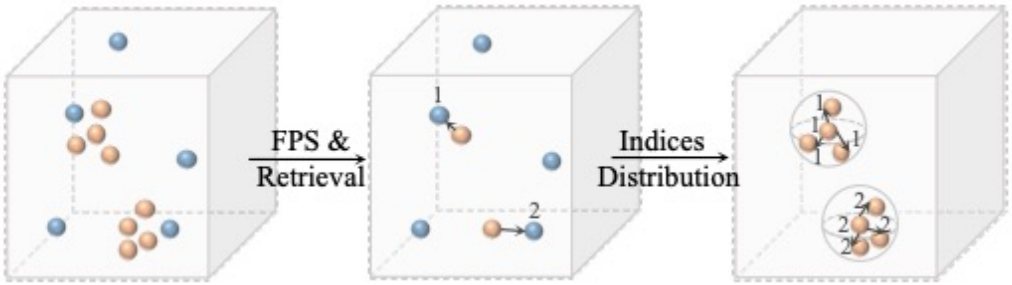
- Standard Cross-Attention : $\mathcal{O}(MN)$
- Local Attention like Swin-Transformer : $\mathcal{O}(MN)$ in retrieving neighbors

Methodology

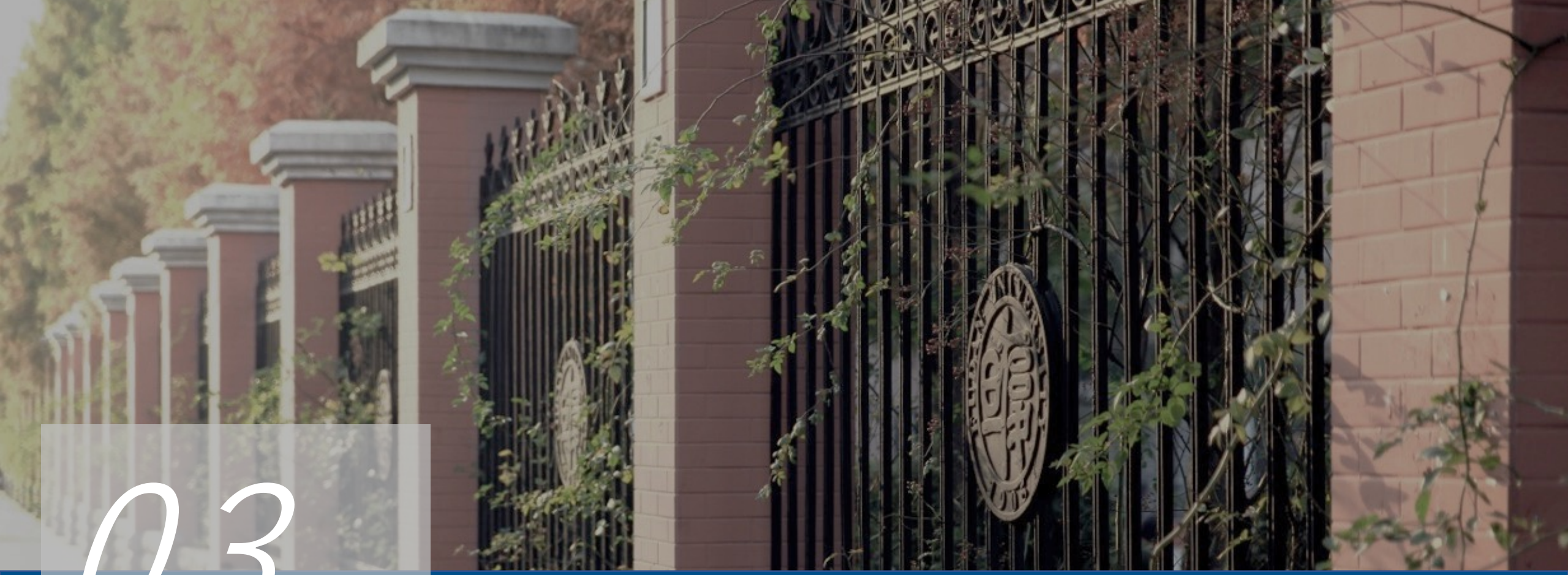
Gated Modality-Aware Convolution (GMA-Conv)

- Solution: Utilizing the redundancy of lifted seeds





- Blue – LiDAR
- Yellow – Camera
- Red – LiDAR and Camera



03

- Experiments

Experiments

Comparison with SOTA methods

- On nuScenes detection track

Method	Modality	mAP	NDS	Car	Truck	C.V.	Bus	Trailer	Barrier	Motor.	Bike	Ped.	T.C.
PointPillar [8]	L	40.1	55.0	76.0	31.0	11.3	32.1	36.6	56.4	34.2	14.0	64.0	45.6
CenterPoint [27]	L	60.3	67.3	85.2	53.5	20.0	63.6	56.0	71.1	59.5	30.7	84.6	78.4
TransFusion-L [1]	L	65.5	70.2	86.2	56.7	28.2	66.3	58.8	78.2	68.3	44.2	86.1	82.0
PointPainting [22]	LC	46.4	58.1	77.9	35.8	15.8	36.2	37.3	60.2	41.5	24.1	73.3	62.4
3D-CVF [29]	LC	52.7	62.3	83.0	45.0	15.9	48.8	49.6	65.9	51.2	30.4	74.2	62.9
PointAugmenting [23]	LC	66.8	71.0	87.5	57.3	28.0	65.2	60.7	72.6	74.3	50.9	87.9	83.6
MVP [28]	LC	66.4	70.5	86.8	58.5	26.1	67.4	57.3	74.8	70.0	49.3	89.1	85.0
TransFusion [1]	LC	68.9	71.7	87.1	60.0	33.1	68.3	60.8	78.1	73.6	52.9	88.4	86.7
VFF [10]	LC	68.4	72.4	86.8	58.1	32.1	70.2	61.0	73.9	**78.5**	52.9	87.1	83.8
BEVFusion [13]	LC	69.2	71.8	88.1	60.9	34.4	69.3	62.1	78.2	72.2	52.2	89.2	85.2
BEVFusion [16]	LC	70.2	72.9	**88.6**	60.1	**39.3**	69.8	63.8	80.0	74.1	51.0	89.2	86.5
UVTR [9]	LC	67.1	71.1	87.5	56.0	33.8	67.5	59.5	73.0	73.4	54.8	86.3	79.6
DeepInteraction [26]	LC	70.8	73.4	87.9	60.2	37.5	70.8	63.8	80.4	75.4	54.5	90.3	87.0
MSMDFusion (Ours)	LC	**71.5**	**74.0**	88.4	**61.0**	35.2	**71.4**	**64.2**	**80.7**	76.9	**58.3**	**90.6**	**88.1**

- On nuScenes tracking track

Method	Modality	AMOTA ↑	AMOTP ↓	TP ↑	FP ↓	FN ↓	Recall ↑
CenterPoint [27]	L	63.8	0.606	95877	18612	22928	67.5
TransFusion [1]	LC	71.8	0.551	96775	16232	21846	75.8
UVTR [9]	LC	70.1	0.686	98434	15615	20190	75.0
BEVFusion [16]*	LC	74.1	0.403	99664	19997	19395	77.9
MSMDFusion (Ours)	LC	**74.0**	**0.549**	**98624**	**14789**	**19853**	**76.3**

Experiments

Efficacy of our lifted seeds (virtual points)

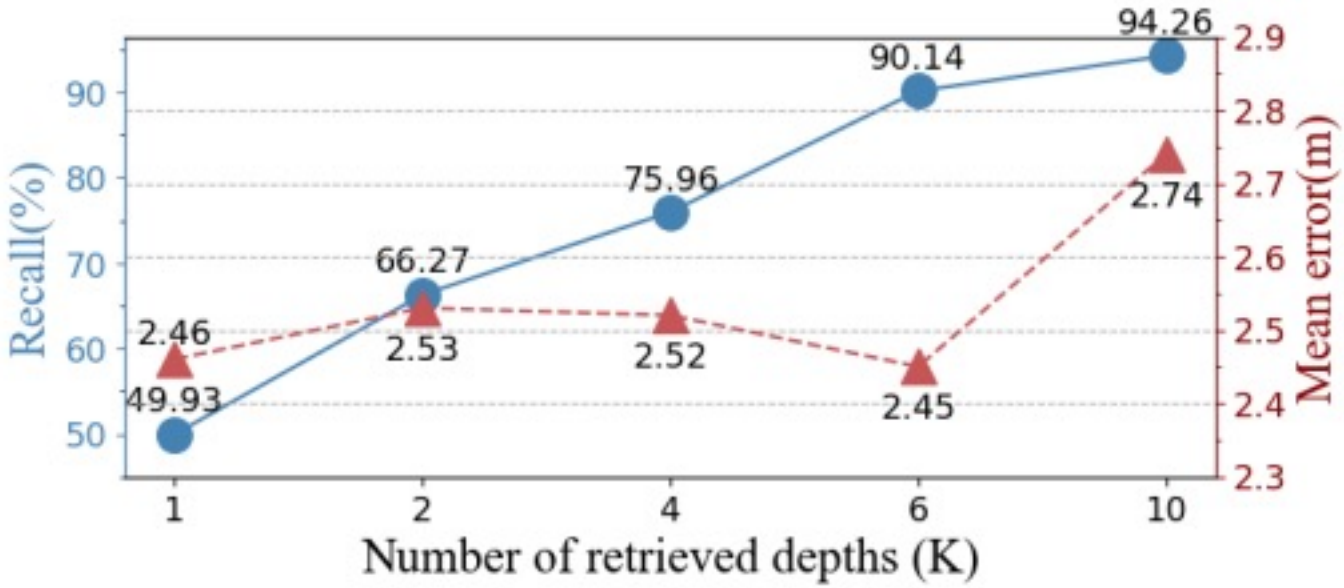
- Number of Virtual point Per Frame (NVPF and performance comparison)

Method	NVPF[2]	mAP	NDS	FPS
BEVFusion [13]	5M	69.2	71.8	0.7
BEVFusion [16]	2M	70.2	72.9	**8.4**
MSMDFusion(Ours)	16k	**71.5**	**74.0**	2.1

Experiments

Effects of KNN retrieval in MDU

- Effects of the number of depths per seed (K) in MDU



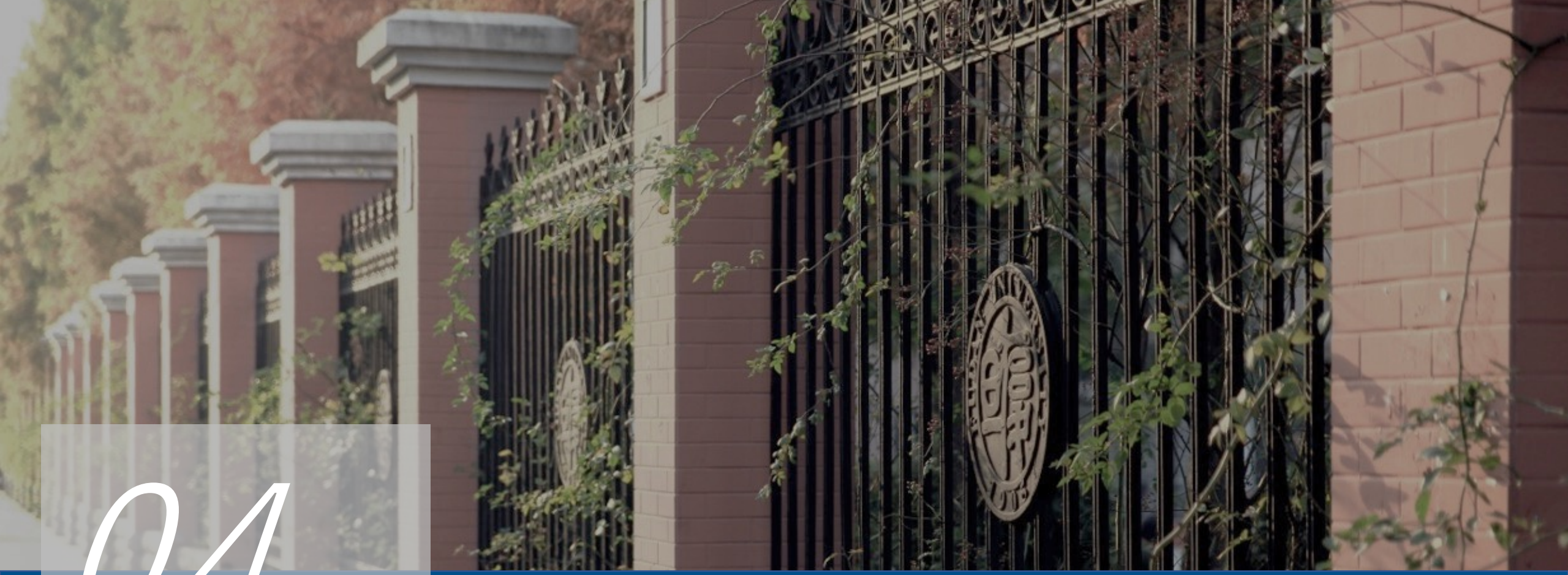
Experiments

Visualization for MDU

(a) LiDAR points

(b) LiDAR points + virtual points (1NN)

(c) LiDAR points + virtual points (6NN)



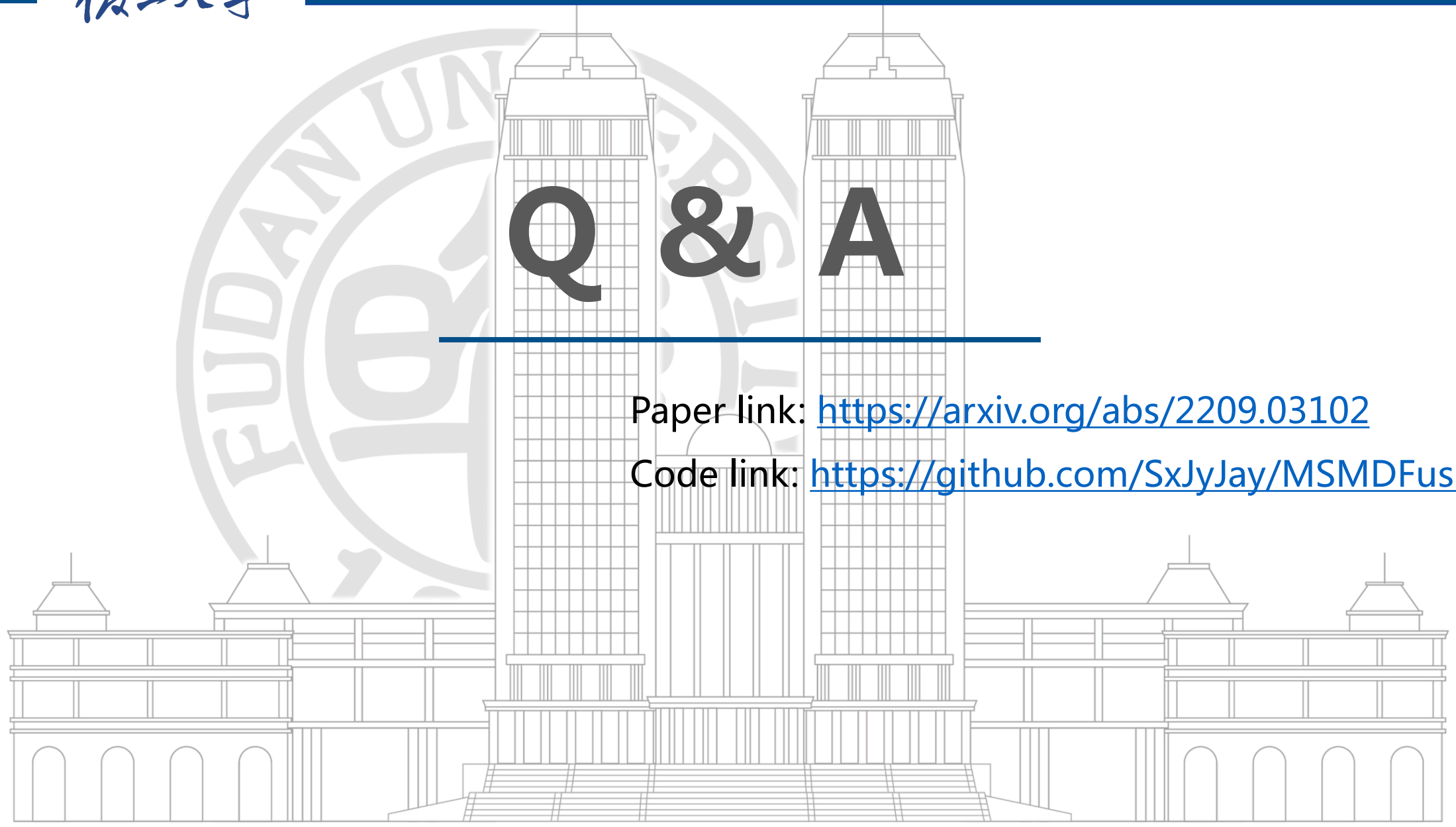
04

- Discussions

Discussions

Problems under-explored

- Application
  - Lightweight multi-sensor fusion framework
  - Hardware-friendly operators
- Performance
  - Intelligent sampler and filter for raw point lift-based multi-modal detector
  - Long-range temporal information utilization to mitigate point-cloud sparsity



Q & A

Paper link: https://arxiv.org/abs/2209.03102

Code link: https://github.com/SxJyJay/MSMDFusion

Thanks for listening

Paper link: https://arxiv.org/abs/2209.03102

Code link: https://github.com/SxJyJay/MSMDFusion