CalibBEV: LiDAR-Camera Calibration via BEV Alignment

Filippo D'Addeo[*1], **Lorenzo Cipelli**[*2], Adriano Cardace[3],
Emanuele Ghelfi[4], Andrea Zinelli[4], Massimo Bertozzi[2]

[1]University of Bologna, [2]University of Parma,
[3]Stanford University, [4]VisLab srl, an Ambarella Inc. company

WACV 2026
TUCSON, AZ
3/6 – 3/10

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

UNIVERSITÀ DI PARMA

Ambarella™
VISLAB.IT
Extend Your Vision

Introduction

Autonomous Vehicles (AV) achieve great environmental perception by exploiting heterogeneous sensors. What do we need to enable perception?

- **Accurate Sensors Calibration** → guarantee data consistency

- **Multi-Modal Sensor Fusion** → enables scene understanding and downstream perception tasks

- **Real-Time Extrinsic Calibration** → fixed calibration is risky

$R|t$

Image-to-Point Cloud (I2P) Registration task → estimate the rigid transformation that aligns a LiDAR point cloud with a reference camera

I2P - State of the art

Previous SOTA approaches [1,2,3,4] rely on initial **classification** task followed by a **RANSAC-based correspondence-matching** phase
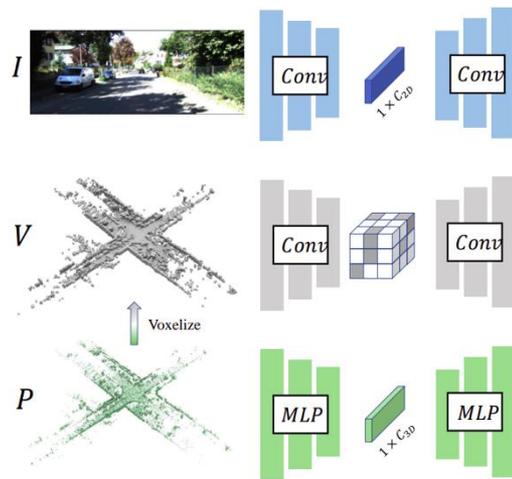

However, these methods often **struggle with matching** RGB and point cloud features, as **different modalities** and architectures do **not** inherently produce a **shared feature space** for seamless **feature matching**.
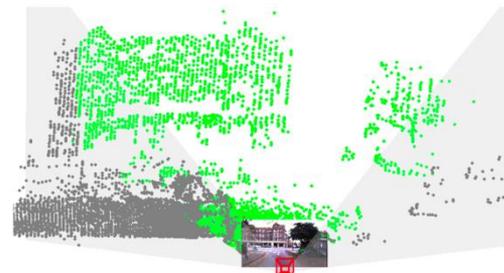
[1] DeepI2P: Image-to-Point Cloud Registration via Deep Classification
[2] CorrI2P: Deep Image-to-Point Cloud Registration via Dense Correspondence
[3] VP2P-Match: Differentiable Registration of Images and LiDAR Point Cloud withVoxelPoint-to-Pixel Matching
[4] CurrI2P: inter and intra modality similarity curriculum learning for image-to-point cloud registration

I2P - State of the art

Newer methods focus on feature consistency by introducing **graph correspondence learning** [5] and **implicit correspondence learning** [6]
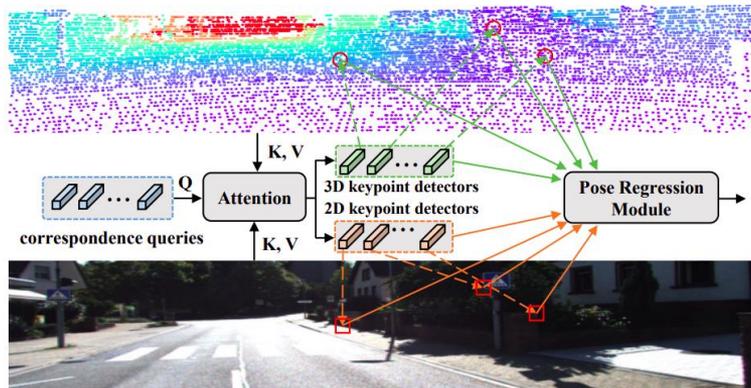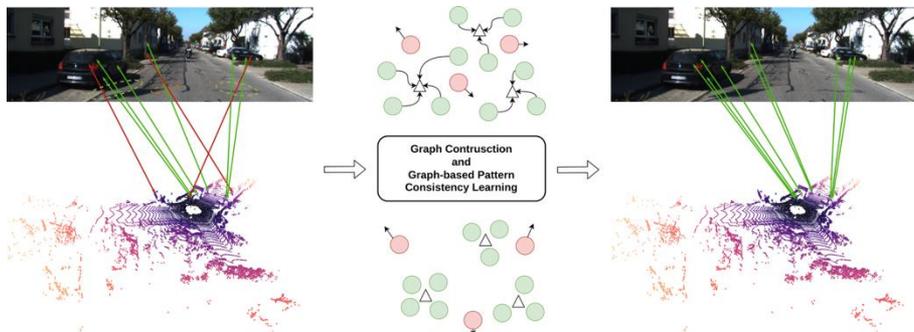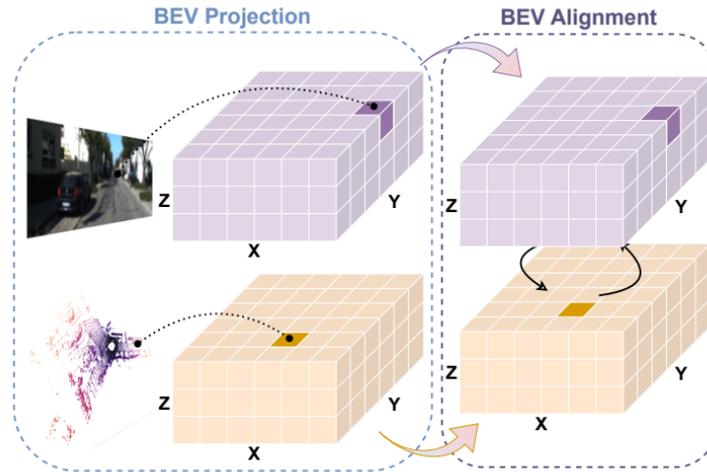

[5] GraphI2P: Image-to-Point Cloud Registration with Exploring Pattern of Correspondence via Graph Learning
[6] ICLM: Implicit Correspondence Learning for Image-to-Point Cloud Registration
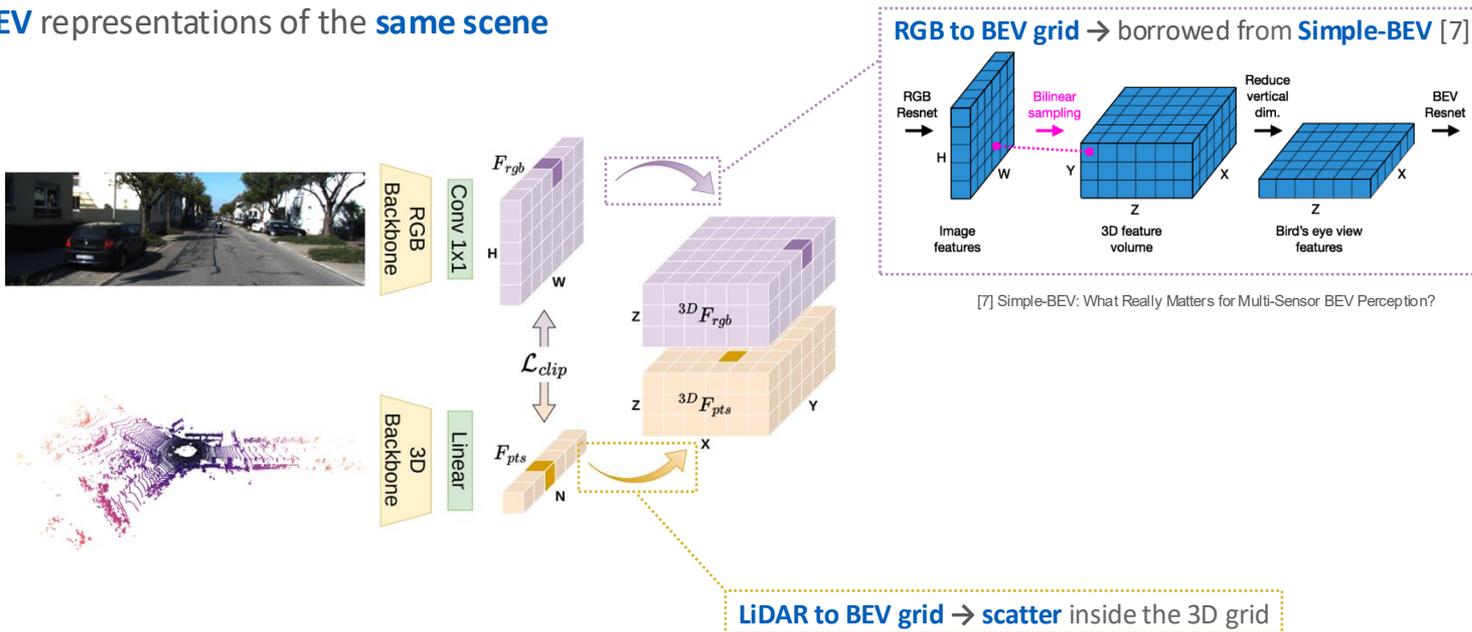
Key Idea 🔑

Taking inspiration from the detection literature:
formulate the I2P registration task as **Bird's Eye View**
(BEV) task

Method - 1
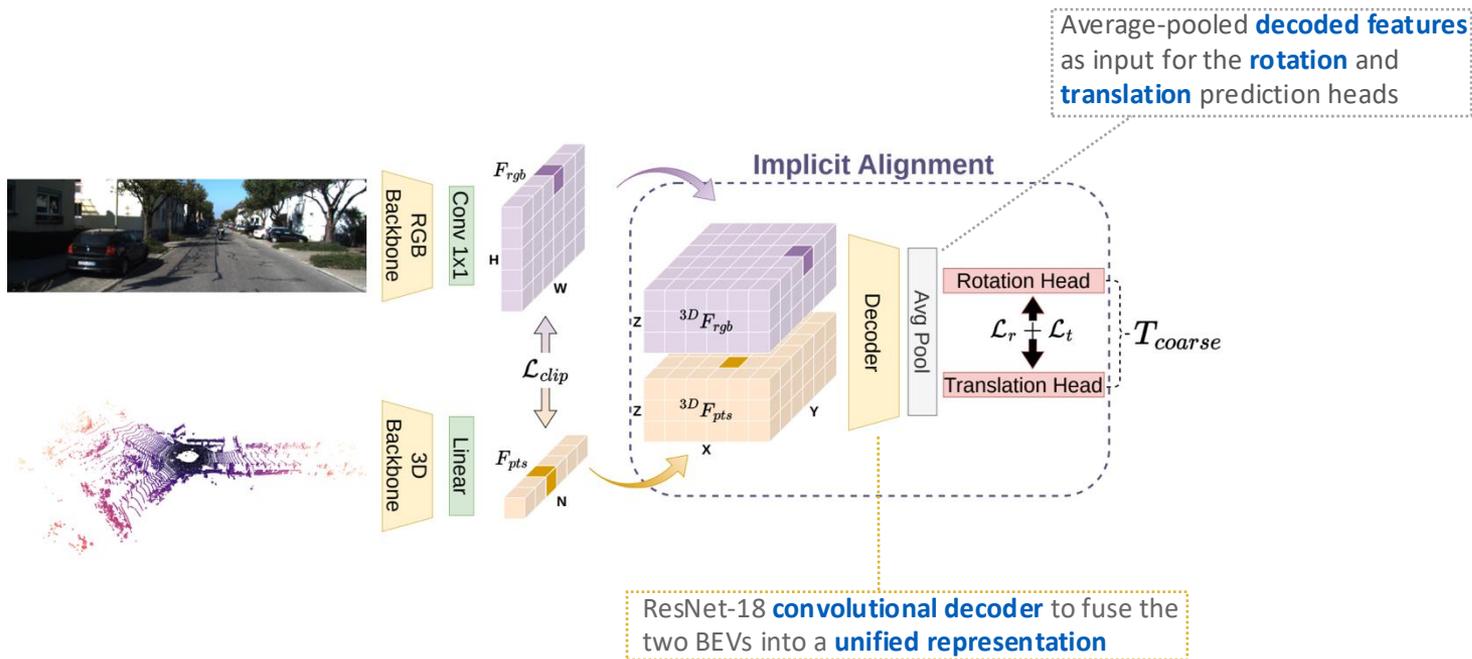
Start with two **BEV** representations of the **same scene**

RGB to BEV grid → borrowed from **Simple-BEV** [7]

[7] Simple-BEV: What Really Matters for Multi-Sensor BEV Perception?

LiDAR to BEV grid → **scatter** inside the 3D grid

Method - 2

Average-pooled **decoded features** as input for the **rotation** and **translation** prediction heads

Implicit Alignment

F_{rgb}

$\mathcal{L}_{clip}$

F_{pts}

$^{3D}F_{rgb}$

$^{3D}F_{pts}$

Rotation Head

$\mathcal{L}_r$ $\mathcal{L}_t$

Translation Head

T_{coarse}

ResNet-18 **convolutional decoder** to fuse the two BEVs into a **unified representation**

Method - 3.1

The whole model is trained in a two-step fashion:

1) **Train Implicit Alignment** 🔥

2) Keep Implicit Alignment frozen ❄️ while training Explicit Alignment 🔥
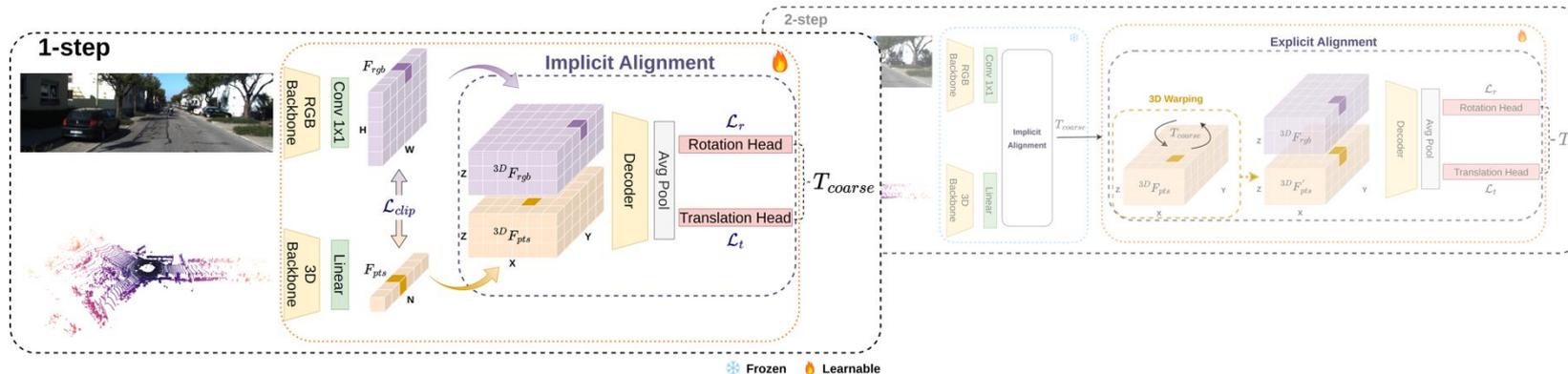
Method - 3.2

The whole model is trained in a two-step fashion:

1) Train Implicit Alignment 🔥

2) **Keep Implicit Alignment frozen** ❄️ **while training Explicit Alignment** 🔥
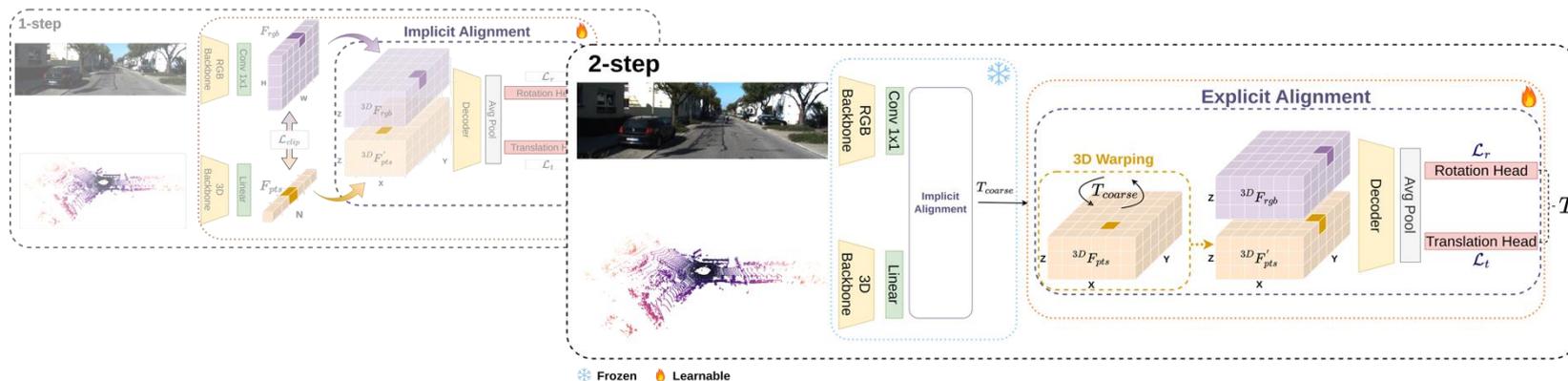
The **Explicit Alignment** is introduced to **warp** the 3D grid and **refine** the final estimate by leveraging the predicted coarse calibration matrix
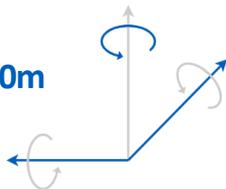
Results – Quantitative Analysis - 1

Comparison against the **state-of-the-art**

Method	KITTI			nuScenes		
	RTE(m)↓	RRE(°)↓	Acc.↑	RTE(m)↓	RRE(°)↓	Acc.↑
Grid Cls. + PnP [18]	3.64 ± 3.46	19.19 ± 28.96	11.22	3.02 ± 2.40	12.66 ± 21.01	2.45
DeepI2P (3D) [18]	4.06 ± 3.54	24.73 ± 31.69	3.77	2.88 ± 2.12	20.65 ± 12.24	2.26
DeepI2P(2D) [18]	3.59 ± 3.21	11.66 ± 18.16	25.95	2.78 ± 1.99	4.80 ± 6.21	38.10
CorrI2P [31]	3.78 ± 65.16	5.89 ± 20.34	72.42	3.04 ± 60.76	3.73 ± 9.03	49.00
CurrI2P (CorrI2P) [23]	1.55 ± 7.99	3.61 ± 10.88	-	2.16 ± 4.20	2.98 ± 5.35	-
VP2P-Match [42]	0.75 ± 1.13	3.29 ± 7.99	83.04	0.89 ± 1.44	2.15 ± 7.03	88.33
CurrI2P (VP2P-Match) [23]	0.53 ± 1.00	2.11 ± 9.48	-	1.04 ± 1.64	2.67 ± 8.61	-
RelaI2P [9]	0.72 ± 1.45	2.92 ± 6.67	85.60	-	-	-
ICLM [20]	$\underline{0.20 \pm 0.21}$	$\underline{1.24 \pm 2.34}$	97.49	0.63 ± 0.44	2.13 ± 3.75	90.94
GraphI2P [2]	0.32 ± 0.81	1.65 ± 1.32	$\underline{99.61}$	$\underline{0.49 \pm 1.22}$	$\underline{1.73 \pm 1.63}$	$\underline{99.48}$
CalibBEV (ours)	$\mathbf{0.04 \pm 0.10}$	$\mathbf{0.61 \pm 0.52}$	**99.96**	$\mathbf{0.04 \pm 0.08}$	$\mathbf{0.54 \pm 0.45}$	**99.98**

Method	KITTI		
	RTE(m)↓	RRE(°)↓	Acc.↑
CorrI2P [31]	0.96 ± 3.10	2.87 ± 4.58	85.76
Calibnet [11]	5.88 ± 2.80	10.92 ± 6.09	3.03
LCCNet [24]	0.40 ± 0.29	4.27 ± 3.70	75.75
CalibBEV (ours)	$\mathbf{0.10 \pm 0.06}$	$\mathbf{1.13 \pm 0.62}$	**99.96**

Projection-based **6 DoF**
benchmark: **±20°** and **±1.5m**

Point-based **3 DoF**
benchmark: **±360°** and **±10m**
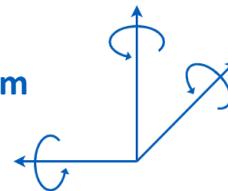
Results – Quantitative Analysis - 2

Zero-shot generalization: train on left camera images and evaluated on the right camera ones on the KITTI dataset

Train	Test	KITTI		
		RTE(m)↓	RRE(°)↓	Acc.↑
L+R	R	0.04 ± 0.10	0.61 ± 0.52	99.96
L	R	0.03 ± 0.02	0.72 ± 0.61	99.89

Differently from any previous work, our model is easily extendable to **multi-camera** configurations

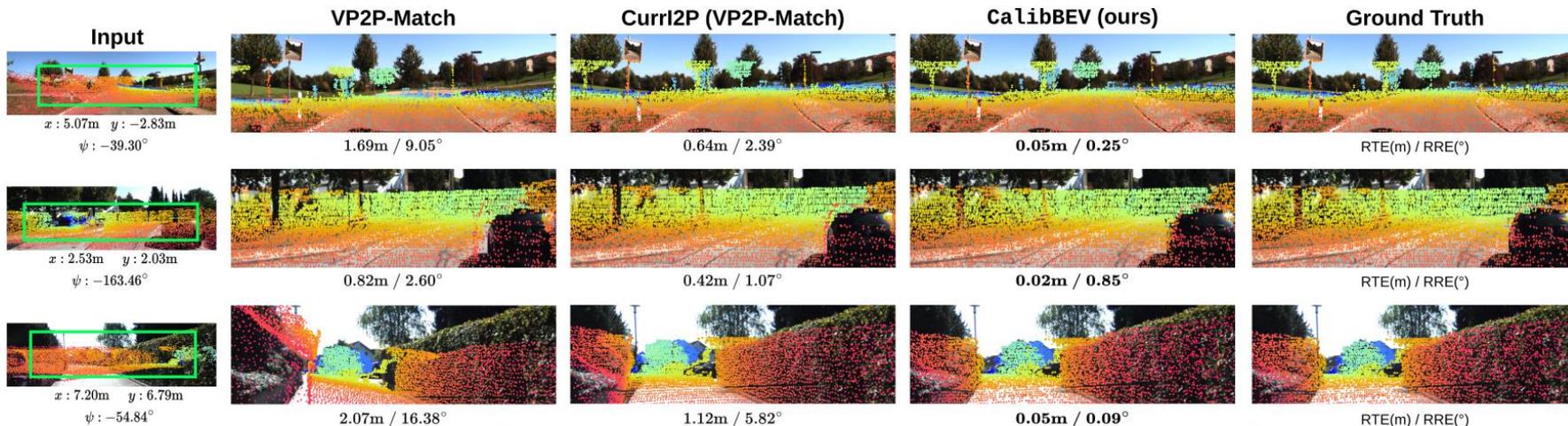| Method	Num. Cams.	nuScenes		
		RTE(m)↓	RRE(^{o})↓	Acc.↑
CalibBEV (ours)	1	0.04 ± 0.08	0.54 ± 0.45	99.98
CalibBEV (ours)	6	$\mathbf{0.04 \pm 0.03}$	$\mathbf{0.28 \pm 0.22}$	**100.0**

WACV 2026
TUCSON, AZ   3/6 – 3/10

Inference Time

Inference time: CalibBEV achieves state-of-the-art performance while being faster than the other competitors

- **Implicit Alignment** → 40% faster than CurrI2P (VP2P-Match)

- **Implicit + Explicit Alignment** → just 12% slower than Implicit Alignment, 32% faster than CurrI2P (VP2P-Match)

Method	Explicit	Time(s)↓
VP2P-Match [16]	-	0.2862
CurrI2P (VP2P-Match) [9]	-	0.2503
CalibBEV (ours)		0.1486
CalibBEV (ours)	✓	0.1695

Results – Qualitative Analysis - 1

Input	VP2P-Match	CurrI2P (VP2P-Match)	CalibBEV (ours)	Ground Truth
$x: 5.07\text{m} \quad y: -2.83\text{m}$ $\psi: -39.30°$	1.69m / 9.05°	0.64m / 2.39°	**0.05m / 0.25°**	RTE(m) / RRE(°)
$x: 2.53\text{m} \quad y: 2.03\text{m}$ $\psi: -163.46°$	0.82m / 2.60°	0.42m / 1.07°	**0.02m / 0.85°**	RTE(m) / RRE(°)
$x: 7.20\text{m} \quad y: 6.79\text{m}$ $\psi: -54.84°$	2.07m / 16.38°	1.12m / 5.82°	**0.05m / 0.09°**	RTE(m) / RRE(°)

Results – Qualitative Analysis - 2

Input

CalibBEV (ours)

Ground Truth

$x: -9.91\text{m} \quad y: 6.79\text{m}$
$\psi: -132.12°$
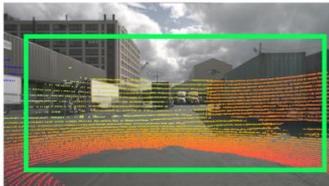
$0.03\text{m} / 0.13°$

RTE(m) / RRE(°)

$x: 7.50\text{m} \quad y: -0.33\text{m}$
$\psi: 45.23°$
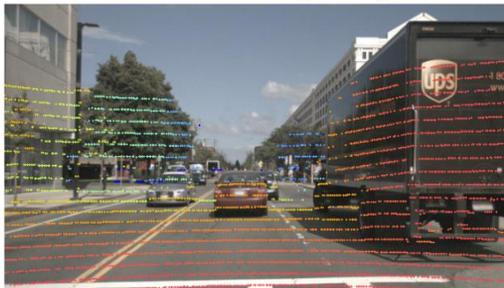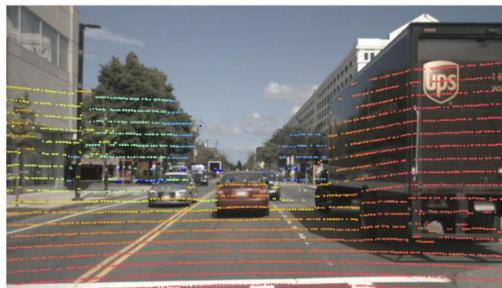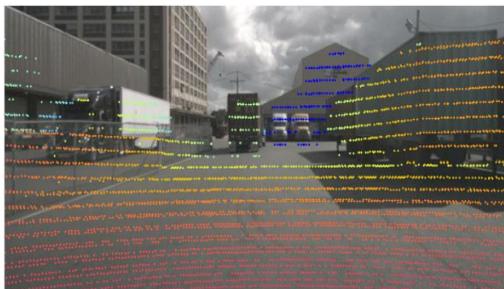
$0.02\text{m} / 0.07°$
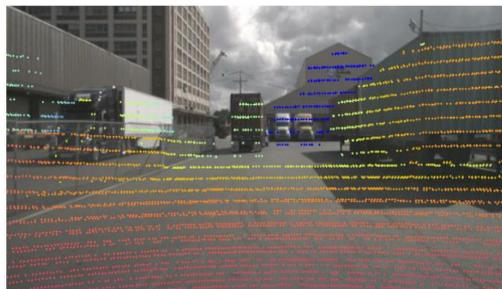
RTE(m) / RRE(°)

Results – Qualitative Analysis - 3

Input

CalibBEV (ours)

Ground Truth

$x : -5.40\text{m}$ $y : 7.95\text{m}$
$\psi : -172.82°$

$0.04\text{m} / 0.31°$

RTE(m) / RRE(°)

$x : -1.50\text{m}$ $y : 8.07\text{m}$
$\psi : 45.23°$
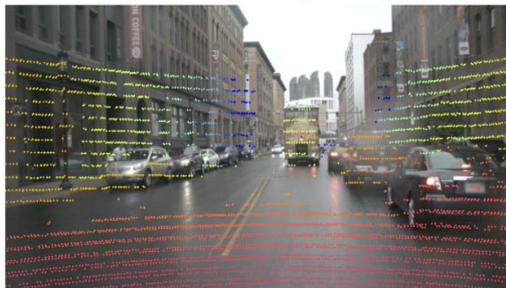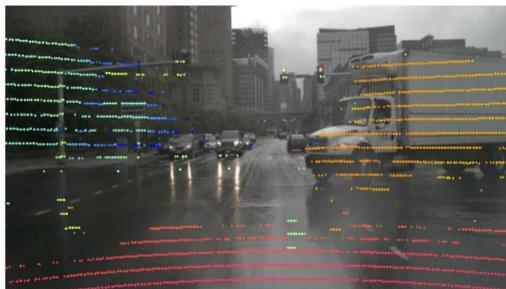
$0.04\text{m} / 0.02°$

RTE(m) / RRE(°)

Conclusions

To summarize:

- CalibBEV is the **first framework** that for the first time frames the I2P task as a **BEV alignment problem**

- Leveraging our **unique** BEV formulation, we propose a **two-step alignment** algorithm

- CalibBEV is **easily extendable** to **multi-camera** configurations, leading to better registration performance

- CalibBEV achieves **state-of-the-art performance** both on KITTI and nuScenes datasets, surpassing previous methods by **significant margin** while being **faster** and **more robust**
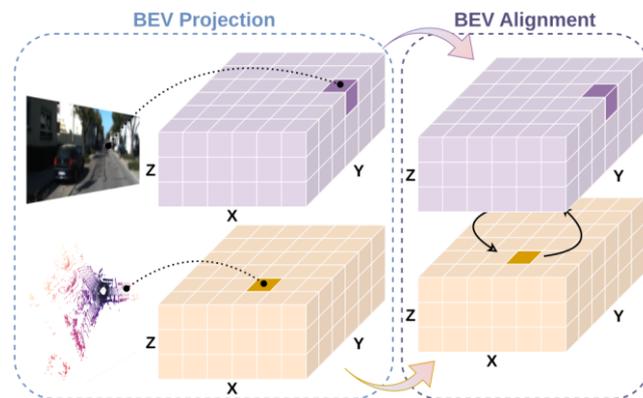
Come to visit us:
Poster 6 @ Session 4

Thank for Your Attention