Imperial College London

APSS Seminar 5

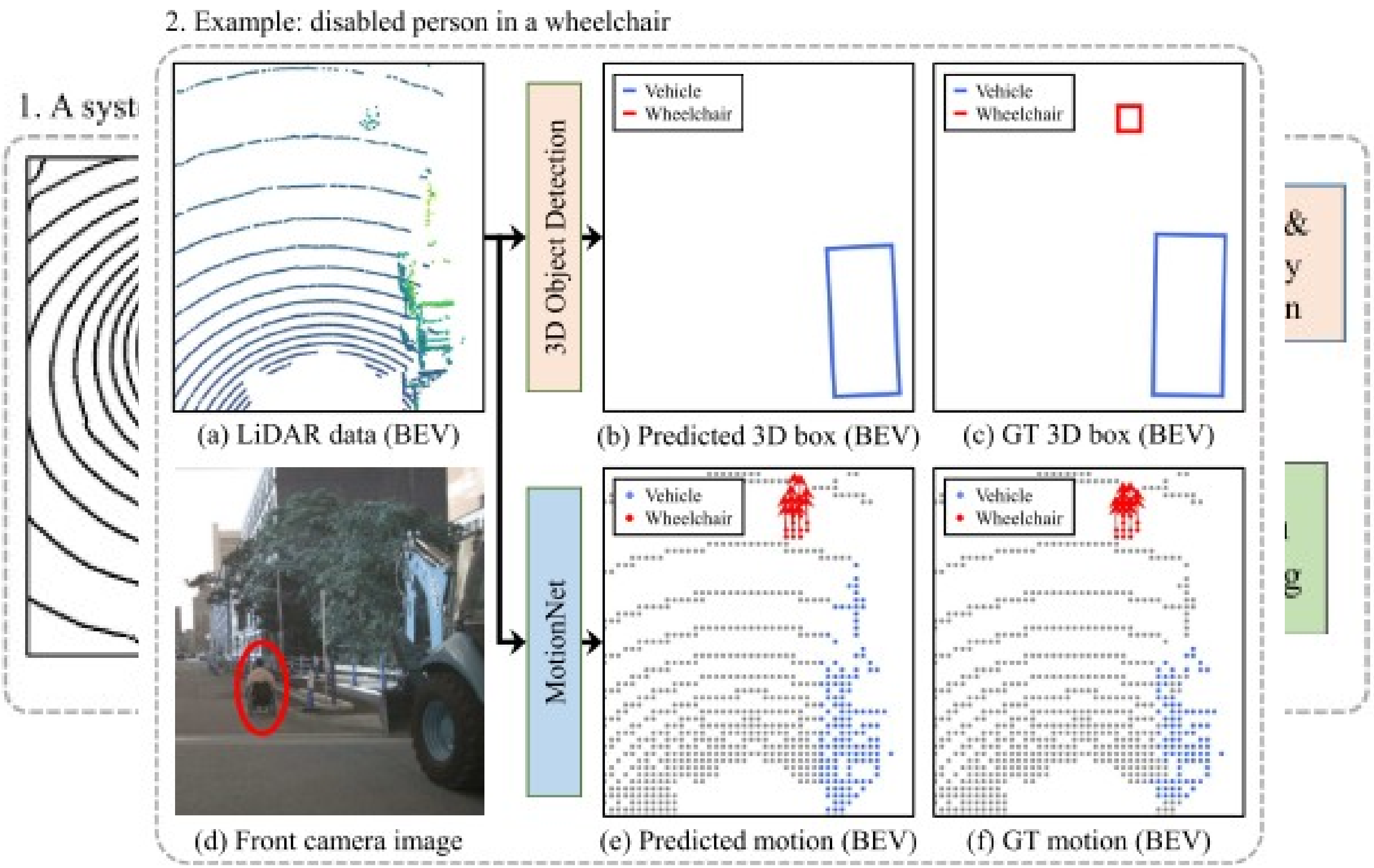
MotionNet: Joint Perception and Motion Prediction for Autonomous Driving

Based on Bird's Eye View Maps

(Charles) Chengzeng You
Connected and autonomous vehicles
Department of computing
Email: chengzeng.you19@imperial.ac.uk

Imperial College London

1. Introduction



Imperial College London

1. Introduction

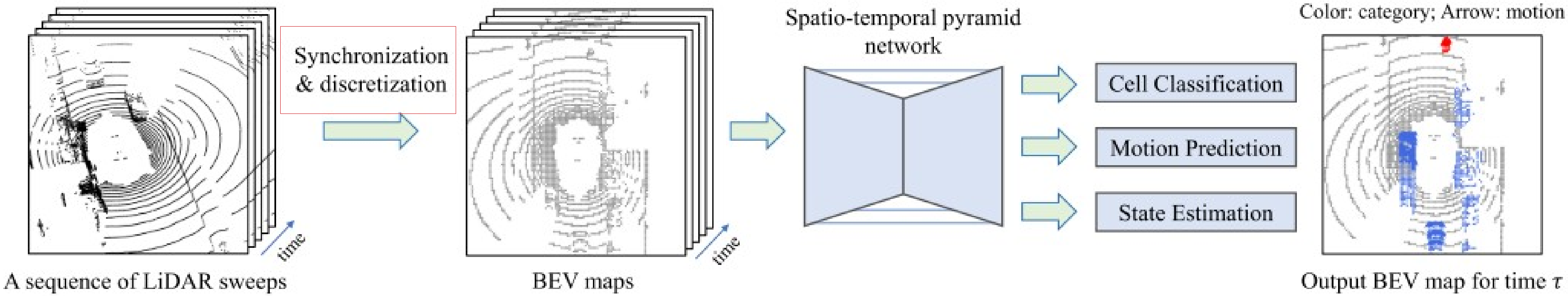
1.1 Contributions

- Propose a novel model, called MotionNet, for joint perception and motion prediction based on BEV maps.
- Propose a novel spatio-temporal pyramid network to extract spatio-temporal features in a hierarchical fashion.
- Develop spatial and temporal consistency losses to constrain the network training, enforcing the smoothness of predictions both spatially and temporally.
- Extensive experiments and in-depth analysis

Imperial College London

2. Methodology

2.1 Ego-motion compensation

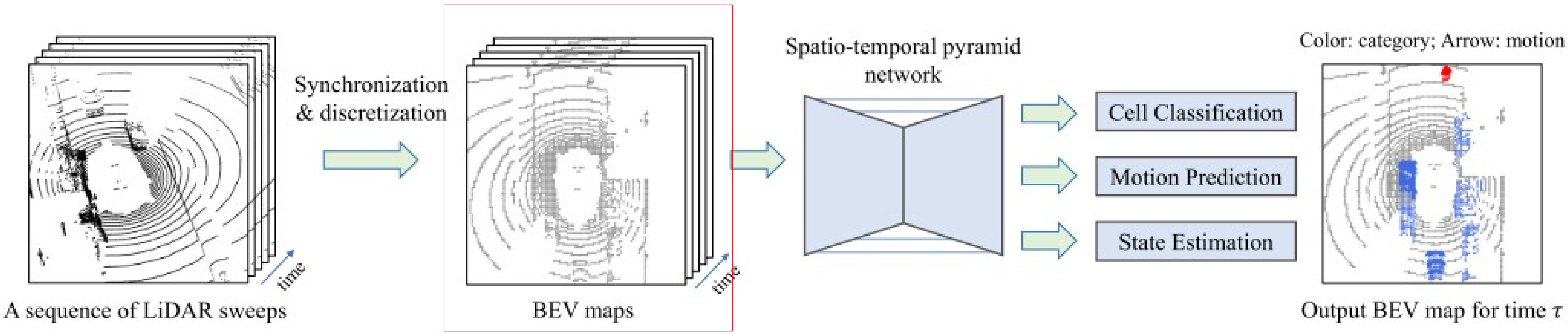


synchronize all the past frames to the current one.

Imperial College London

2. Methodology

2.2 BEV-map-based representation

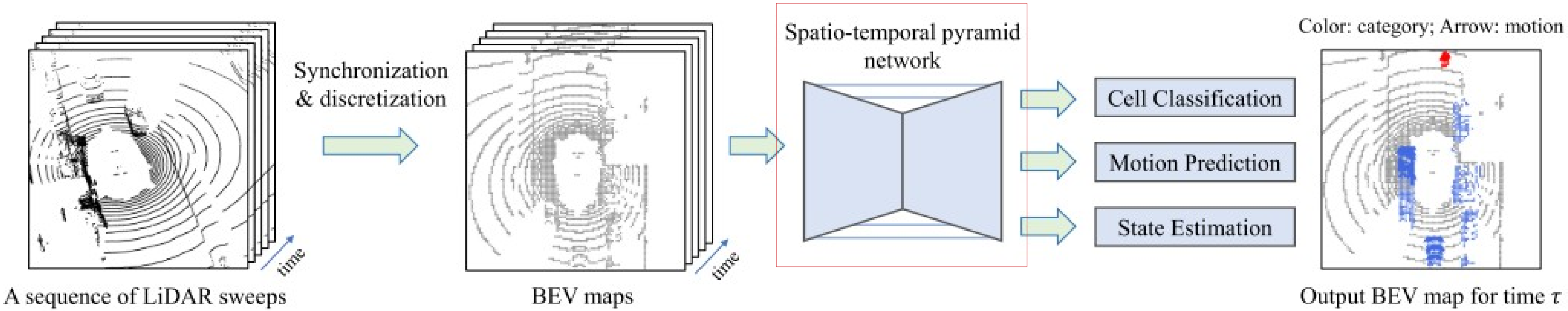


1.quantize the 3D points into regular voxels.
2.represent the 3D voxel lattice as a 2D pseudo-image, with the height dimension corresponding to image channels.

Imperial College London

2. Methodology

2.3 Spatio-temporal pyramid network(STPN)



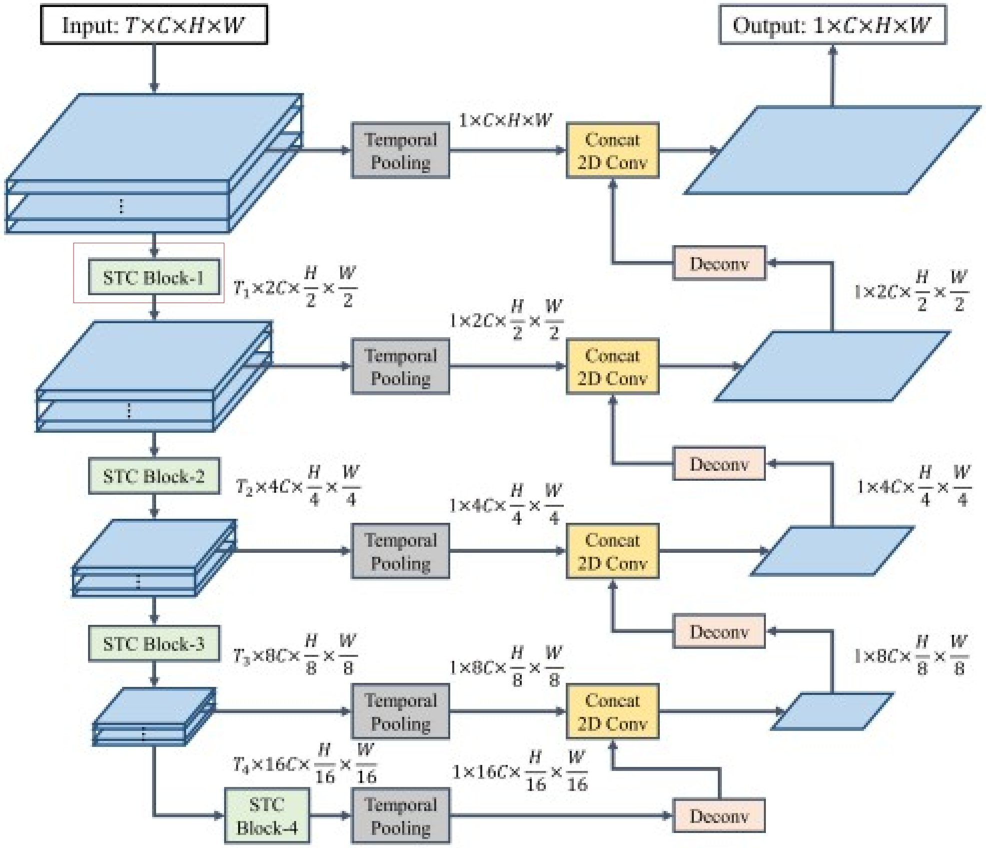
extract features along both the spatial and temporal dimensions in a hierarchical fashion.

Imperial College London

2. Methodology

2.3 Spatio-temporal pyramid network(STPN)

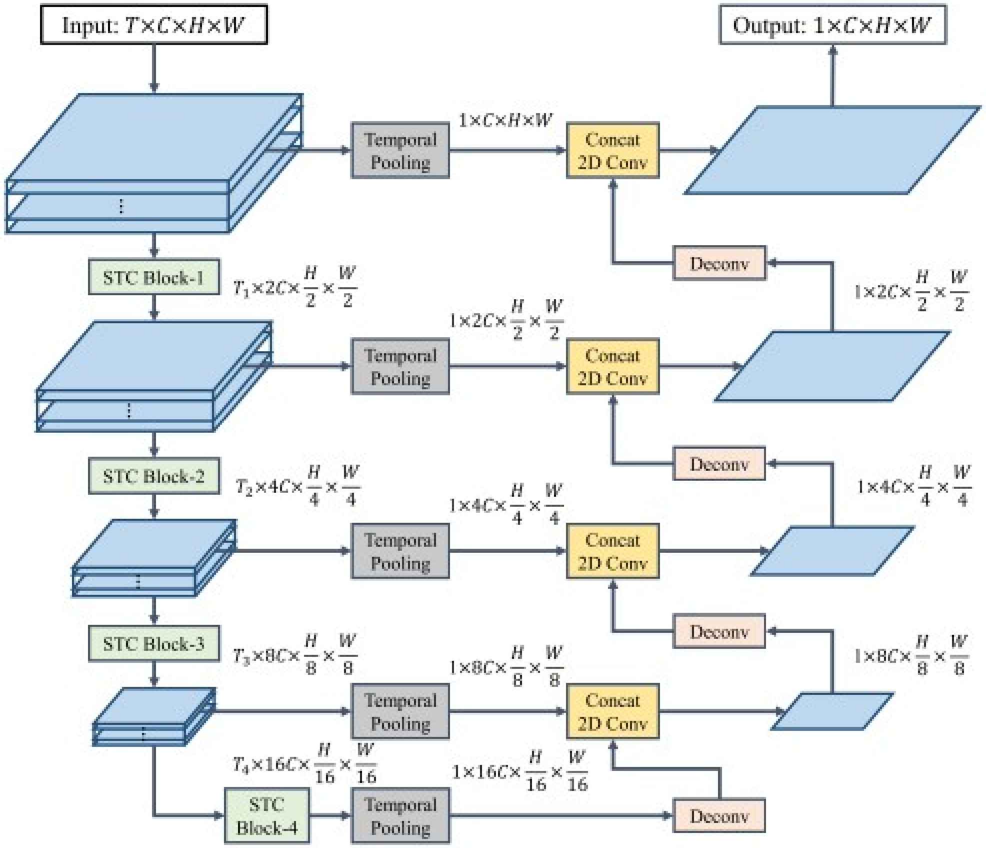
spatio-temporal convolution (STC) block = 2D convolutions + degenerate 3D convolution



Imperial College London

2. Methodology

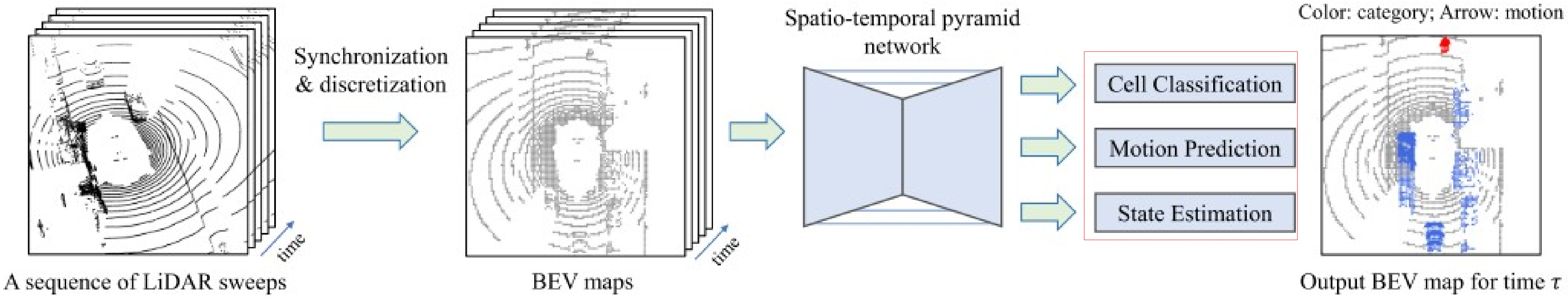
2.3 Spatio-temporal pyramid network(STPN)



Imperial College London

2. Methodology

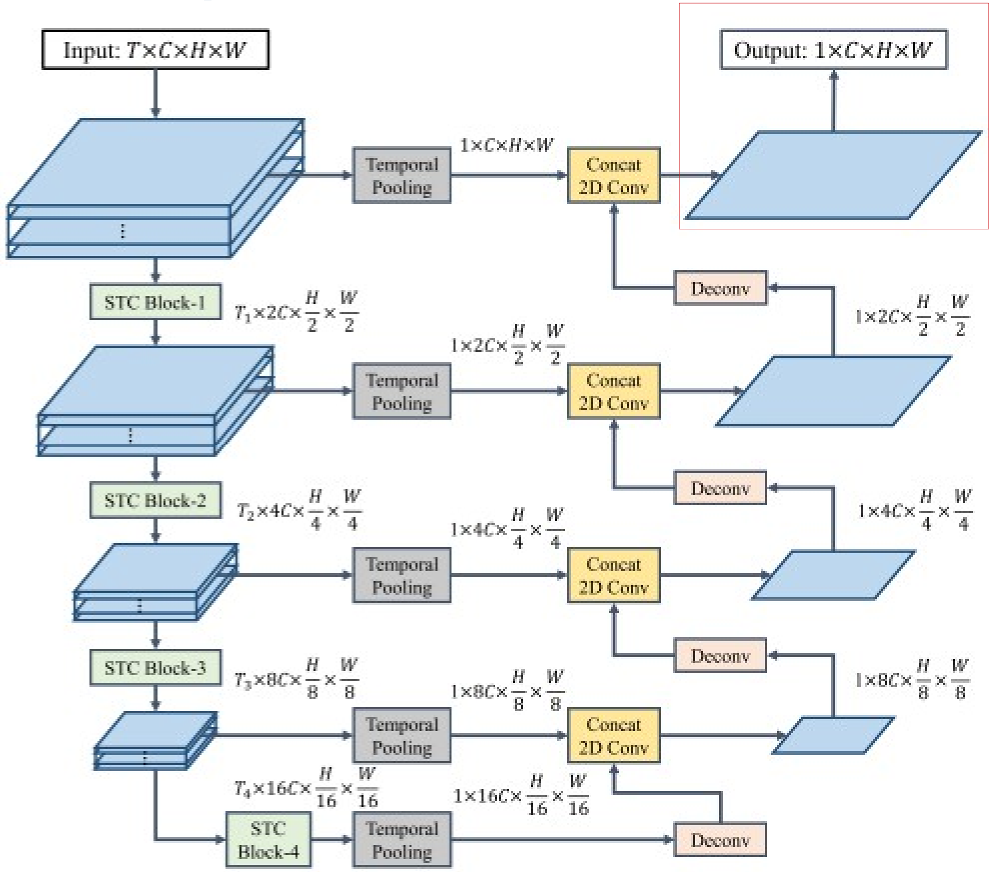
2.4 Output heads



Imperial College London

2. Methodology

2.4 Output heads

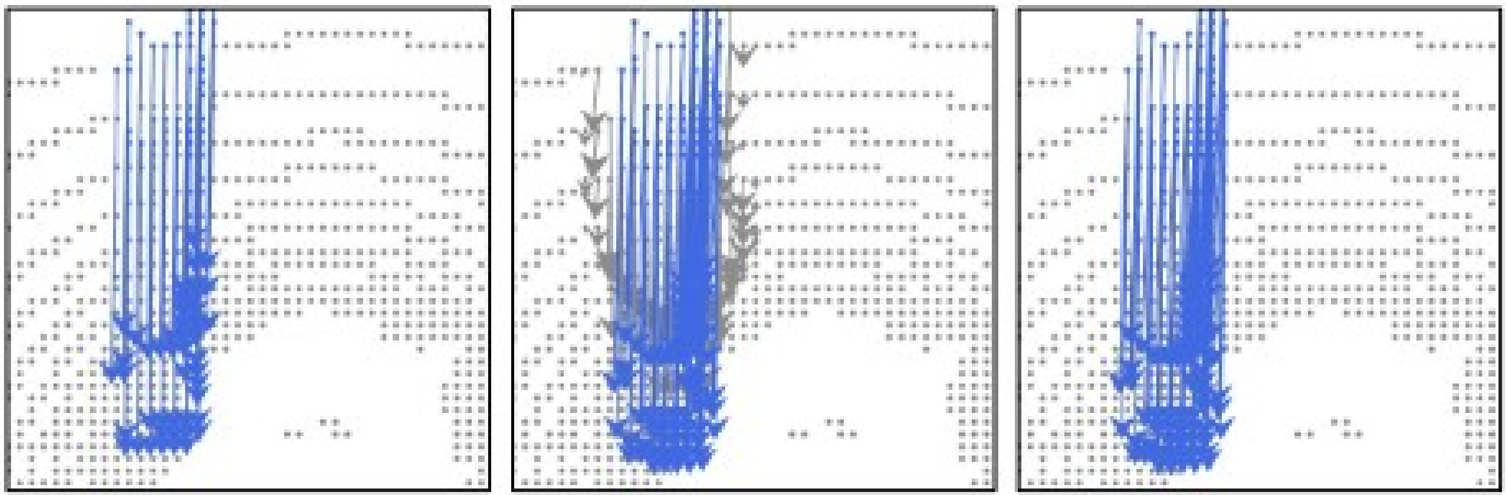


For cell-classification head:
H × W × C
The predicted cell positions:
$\{X^{(\tau)}\}_{\tau=t}^{t+N}$
For motion-prediction head:
N × H × W × 2
For the state-estimation head:
H×W

Imperial College London

2. Methodology

2.4 Output heads



(a) Ground-truth (b) Before suppression (c) After suppression

2. Methodology

2.5 Loss function

For the classification and state-estimation heads: **cross-entropy loss**

$$-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

For the motion-prediction head: **weighted smooth L1 loss**

$$L_{1;smooth} = \begin{cases} |x| & \text{if } |x| > \alpha; \\ \frac{1}{|\alpha|}x^2 & \text{if } |x| \leq \alpha \end{cases}$$

Imperial College London

2. Methodology

2.5 Loss function - Spatial consistency loss

for the cells belonging to the same rigid object, their predicted motions should be very close without much divergence.

$$L_s = \sum_{k} \sum_{(i,j),(i',j') \in o_k} \left\| X_{i,j}^{(\tau)} - X_{i',j'}^{(\tau)} \right\|$$

$\|\cdot\|$ is the smooth L1 loss
o_k denotes the object instance
$X_{i,j}^{(\tau)}$ is the predicted motion at position (i, j) and time τ

Imperial College London

2. Methodology

2.5 Loss function - Spatial consistency loss

For each object, there will be no sharp change of motions between two consecutive frames.

$$L_{\mathrm{ft}} = \sum_{k} \left\| X_{o_k}^{(\tau)} - X_{o_k}^{(\tau+\Delta t)} \right\|$$

$X_{o_k}^{(\tau)}$ denotes the overall motion of object k

Imperial College London

2. Methodology

2.5 Loss function - Background temporal consistency loss

$$L_{\mathrm{bt}} = \sum_{(i,j)\in X^{(\tau)}\cap T(\widetilde{X}^{(\tau-\Delta t)})} \left\| X_{i,j}^{(\tau)} - T_{i,j}\left(\widetilde{X}^{(\tau-\Delta t)}\right) \right\|$$

$X^{(\tau)}$ $\widetilde{X}^{(\tau)}$ are the predictions with current time being t and t + Δt

T is a rigid transformation

Imperial College London

2. Methodology

2.5 Loss function - overall loss function

$$L = L_{\mathrm{cls}} + L_{\mathrm{motion}} + L_{\mathrm{state}} + \alpha L_{\mathrm{s}} + \beta L_{\mathrm{ft}} + \gamma L_{\mathrm{bt}}$$

α, β and γ are the balancing factors

3. Experiments

3.1 Setup - Dataset

NuScenes LiDAR: 850 scenes in all, 500 scenes for training, 100 for validation and 250 for testing.

Adaption: for each cell inside a bounding box, its motion is computed as:

$$\mathbf{Rx + c\Delta - x}$$

x: cell position
R: yaw rotation with respect to the box center
cΔ: displacement of box

3. Experiments

3.1 Setup - Implementation details

Point clouds region: [−32, 32] × [−32, 32] × [−3, 2] meters.
Voxel resolution: $(\Delta x, \Delta y, \Delta z) = (0.25, 0.25, 0.4)$ m.
Temporal information: 5 frames of synchronized point clouds, where 4 are from the past timestamps and 1 corresponds to the current time.

5 cell categories: background, vehicle (comprising car and bus), pedestrian, bicycle and others.

Imperial College London

3. Experiments

3.1 Setup - Evaluation criteria

For motion prediction, dividing the cells into 3 groups: **static, slow (≤ 5m/s), and fast (> 5m/s)**.

In each group, **average L2 distances** between the estimated displacements and the ground-truth displacements.

For the classification, measure the performance with two metrics: **(1) overall cell classification accuracy (OA)** ;**(2) mean category accuracy (MCA)**

Imperial College London

3. Experiments

3.2 Comparison with state-of-the-art methods

- **Baselines**

(1) **Static Model**, which assumes the environment is static.
(2) **FlowNet3D [28] and HPLFlowNet [12]**, which estimate the scene flow between two point clouds.
(3) **PointRCNN [46] + Kalman filter [17]** to track the objects and predict their future trajectories.
(4) **LSTM-Encoder-Decoder** [44], which estimates the multi-step OGMs(occupancy grid maps).

Imperial College London

3. Experiments

3.2 Comparison with state-of-the-art methods

- **Results**

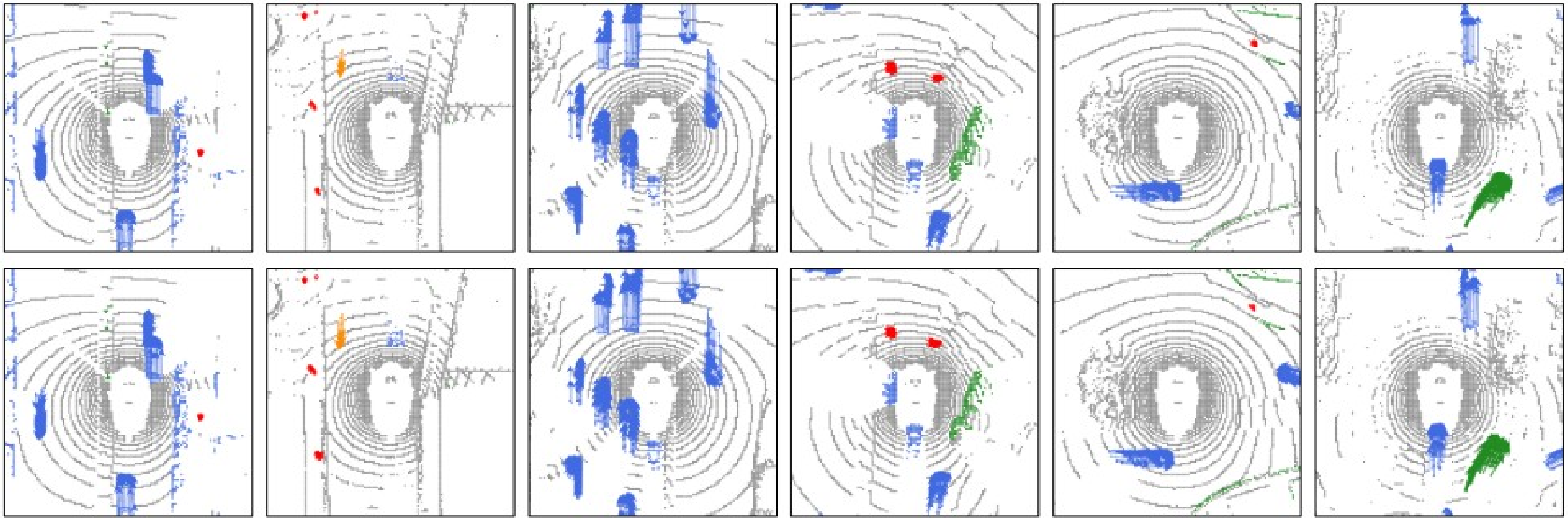
Method	Static		Speed $\leq$ 5m/s		Speed > 5m/s		Classification Accuracy (%)							Infer.
	Mean	Median	Mean	Median	Mean	Median	Bg	Vehicle	Ped.	Bike	Others	MCA	OA	Speed
Static Model	**0**	**0**	0.6111	0.0971	8.6517	8.1412	-	-	-	-	-	-	-	-
FlowNet3D (pretrain) [28]	2.0514	0	2.2058	0.3172	9.1923	8.4923	-	-	-	-	-	-	-	0.434s
FlowNet3D [28]	0.0410	0	0.8183	0.1782	8.5261	8.0230	-	-	-	-	-	-	-	0.434s
HPLFlowNet (pretrain) [12]	2.2165	1.4925	1.5477	1.1269	5.9841	4.8553	-	-	-	-	-	-	-	0.352s
HPLFlowNet [12]	0.0041	0.0002	0.4458	0.0960	4.3206	2.4881	-	-	-	-	-	-	-	0.352s
PointRCNN [46]	0.0204	0	0.5514	0.1627	3.9888	1.6252	98.4	78.7	44.1	11.9	44.0	55.4	96.0	0.201s
LSTM-Encoder-Decoder [44]	0.0358	0	0.3551	0.1044	1.5885	1.0003	93.8	91.0	73.4	17.9	71.7	69.6	92.8	0.042s
MotionNet	0.0256	0	0.2565	0.0962	1.0744	0.7332	97.3	91.1	76.2	20.6	66.1	70.3	96.1	0.019s
MotionNet + L_s	0.0256	0	0.2488	0.0958	1.0110	0.7001	97.5	91.3	76.2	23.7	67.6	71.2	96.3	0.019s
MotionNet + L_{ft}	0.0252	0	0.2515	0.0962	1.0360	0.7136	97.6	90.6	75.3	21.9	65.2	70.1	96.3	0.019s
MotionNet + L_{bt}	0.0240	0	0.2530	0.0960	1.0399	0.7131	97.5	91.1	74.6	25.2	68.0	71.3	96.3	0.019s
MotionNet + L_s + L_{ft} + L_{bt}	0.0239	0	0.2467	0.0961	1.0109	0.6994	97.6	90.7	77.2	25.8	65.1	**71.3**	**96.3**	0.019s
MotionNet + MGDA	0.0222	0	0.2366	0.0953	0.9675	0.6639	97.1	90.5	78.4	22.1	67.4	71.1	95.7	0.019s
MotionNet + $\{L\}$ + MGDA	0.0201	0	**0.2292**	**0.0952**	**0.9454**	**0.6180**	97.0	90.7	77.7	19.7	66.3	70.3	95.8	0.019s

Imperial College London

3. Experiments

3.2 Comparison with state-of-the-art methods

- **Results**



Gray: background; blue: vehicle; red: pedestrian; orange: bicycle; green: others.

Imperial College London

3. Experiments

3.3 Ablation studies

- Number of frames

Frame #	Static	Speed $\leq 5m/s$	Speed $> 5m/s$	MCA	OA	Infer. Speed
2	0.0270	0.2921	1.2445	69.7	95.6	**0.013s**
3	0.0264	0.2738	1.0953	69.6	95.9	0.014s
4	0.0258	0.2597	1.0804	70.2	96.0	0.017s
5	0.0256	**0.2565**	**1.0744**	**70.3**	96.1	0.019s
6	**0.0254**	0.2657	1.1220	69.7	**96.2**	0.021s
7	0.0255	0.2582	1.0779	70.0	**96.2**	0.022s

Imperial College London

3. Experiments

3.3 Ablation studies

- **Ego-motion compensation**

Synch. Strategy	Static	Speed $\leq 5m/s$	Speed $> 5m/s$	MCA	OA
No Synch.	0.0281	0.4245	1.7317	67.1	95.2
ICP [2]	0.0279	0.4073	1.6614	67.4	95.3
GT Synch.	**0.0256**	**0.2565**	**1.0744**	**70.3**	**96.1**

Imperial College London

3. Experiments

3.3 Ablation studies

- **Input data representations**

Data Rep.	Static	Speed ≤ 5m/s	Speed > 5m/s	MCA	OA	Infer. Speed
Voxel	0.0257	0.2546	**1.0712**	69.6	**96.2**	0.107s
Pillar	0.0258	0.2612	1.0747	70.0	96.1	0.096s
BEV	0.0256	0.2565	1.0744	70.3	96.1	0.019s
$(1.0, 1.0, 0.5)\Delta$	**0.0253**	**0.2540**	1.0752	70.1	96.0	0.024s
$(1.0, 1.0, 1.5)\Delta$	**0.0253**	0.2562	1.0726	70.1	95.9	**0.014s**
$(0.5, 0.5, 0.5)\Delta$	0.0261	0.2561	1.0806	70.5	96.1	0.106s
$(0.5, 0.5, 1.0)\Delta$	0.0269	0.2545	1.0761	**71.0**	95.9	0.064s
$(0.5, 0.5, 1.5)\Delta$	0.0257	0.2547	1.0733	70.9	96.0	0.050s

Imperial College London

3. Experiments

3.3 Ablation studies

- Spatio-temporal feature extraction

Block \ Fusion	Early	Mid	Late	Static	Speed $\leq 5m/s$	Speed $> 5m/s$	MCA	OA	Infer. Speed
STC	✓			0.0271	0.2596	1.1002	70.5	96.0	**0.015s**
STC		✓		**0.0256**	**0.2565**	**1.0744**	70.3	**96.1**	0.019s
STC			✓	0.0256	0.2748	1.0838	70.4	96.0	0.019s
C3D [49]		✓		0.0257	0.2624	1.0831	70.5	96.1	0.021s
S3D [54]		✓		0.0267	0.2644	1.1236	70.9	95.9	0.019s
TSM [25]		✓		0.0262	0.2651	1.1241	70.9	96.0	0.018s
CS3D [50]		✓		0.0261	0.2631	1.1787	**71.0**	96.0	0.021s

Imperial College London

3. Experiments

3.3 Ablation studies

- Prediction strategies

	State Head	Relative Offset	J.S. w/ Cls	J.S. w/ State	Static	Speed $\leq$ 5m/s	Speed $>$ 5m/s	MCA	OA
1		✓	✓		0.0284	0.2610	1.0957	69.8	95.0
2	✓		✓	✓	0.0264	0.2621	1.1121	70.2	95.8
3	✓	✓			0.0331	**0.2547**	**1.0601**	70.3	96.1
4	✓	✓	✓		0.0259	0.2564	1.0722	70.3	96.1
5	✓	✓		✓	0.0264	0.2554	1.0657	70.3	96.1
6	✓	✓	✓	✓	**0.0256**	0.2565	1.0744	**70.3**	**96.1**

Imperial College London

4. Limitations

1. small object detection
The classification accuracy for the “bicycle” and ‘pedestrian’ categories are low.
2. cold start problem
For each single scene, cannot make predictions on the first 20 frames.

Thank you