UNIVERSITY OF SURREY

Introduction to Birds-Eye-View Mapping

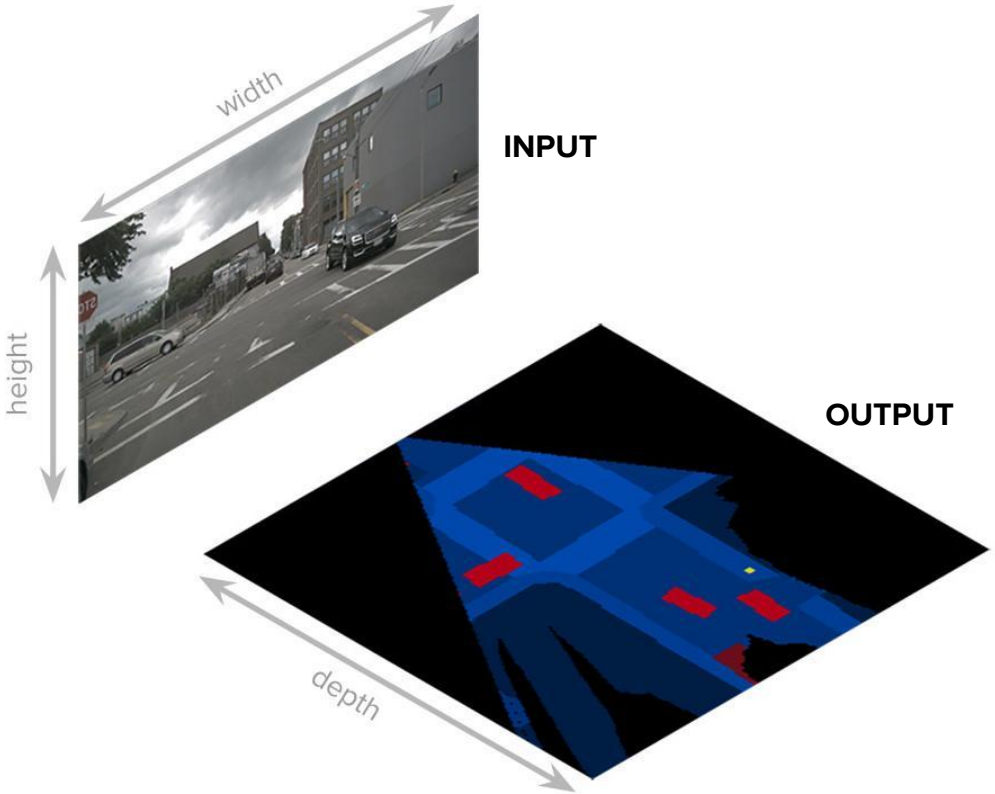
Avi Saha
Jaime Spencer
Chris Russell
Simon Hadfield
Richard Bowden

OVERVIEW

BEV Mapping

1. **Goal:** learn a model that takes a monocular input **image** and generates a semantically segmented **BEV** of the scene.

2. **Relevance:** autonomous navigation and planning on the fly.



Related Work

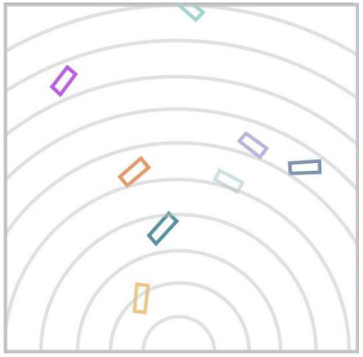
BEV Object Detection

Related work

1. **Task:** given image, predict BEV bounding box in camera coordinates.

INPUT

OUTPUT

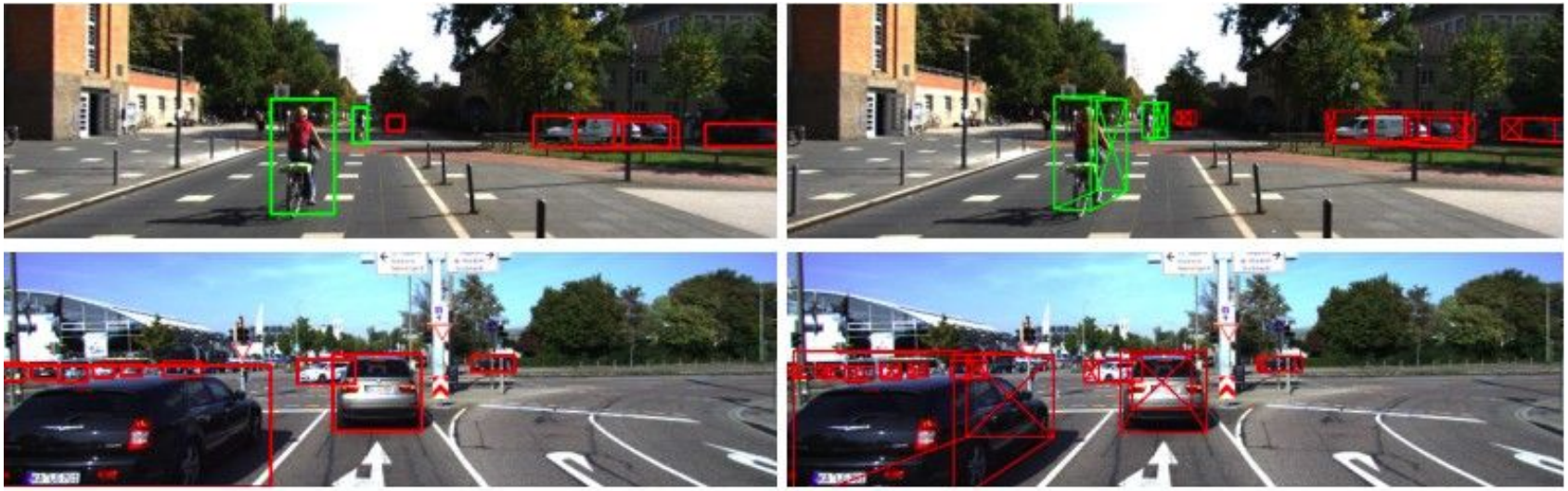


BEV Object Detection

Related work

1. **Approach 1**: detect in image, then regress 3D pose, e.g. Mousavian et al. (2017)
    - **Limitations**: no global scene reasoning in 3D as each 3D object proposal is generated independently.

STAGE 1: object detection | **STAGE 2: pose regression**

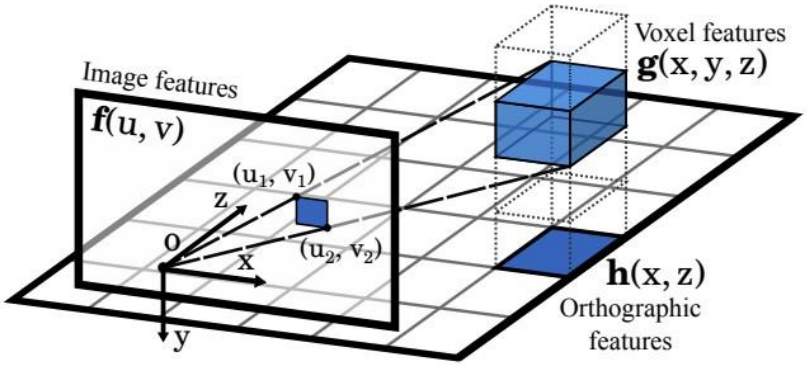


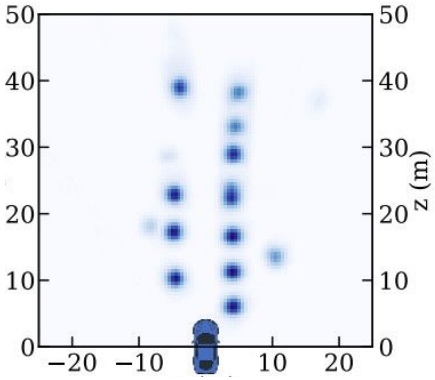
Mousavian, A., Anguelov, D., Flynn, J. and Kosecka, J., 2017. 3d bounding box estimation using deep learning and geometry.

BEV Object Detection

Related work

1. **Approach 2**: project 3D grid to image, reason across all scene objects in BEV, Roddick et al. (2019)
    - **Limitations**: image context available to each BEV voxel is dependent upon distance from camera.



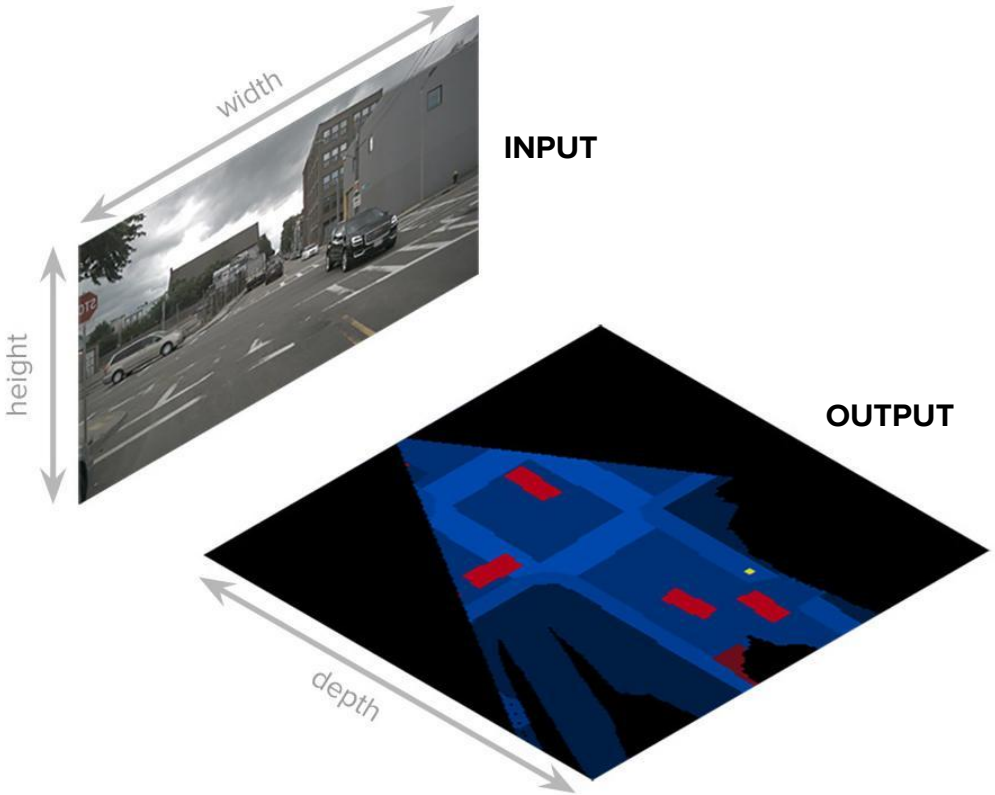


Roddick, T., Kendall, A. and Cipolla, R., 2018. Orthographic feature transform for monocular 3d object detection.

BEV Mapping

Related work

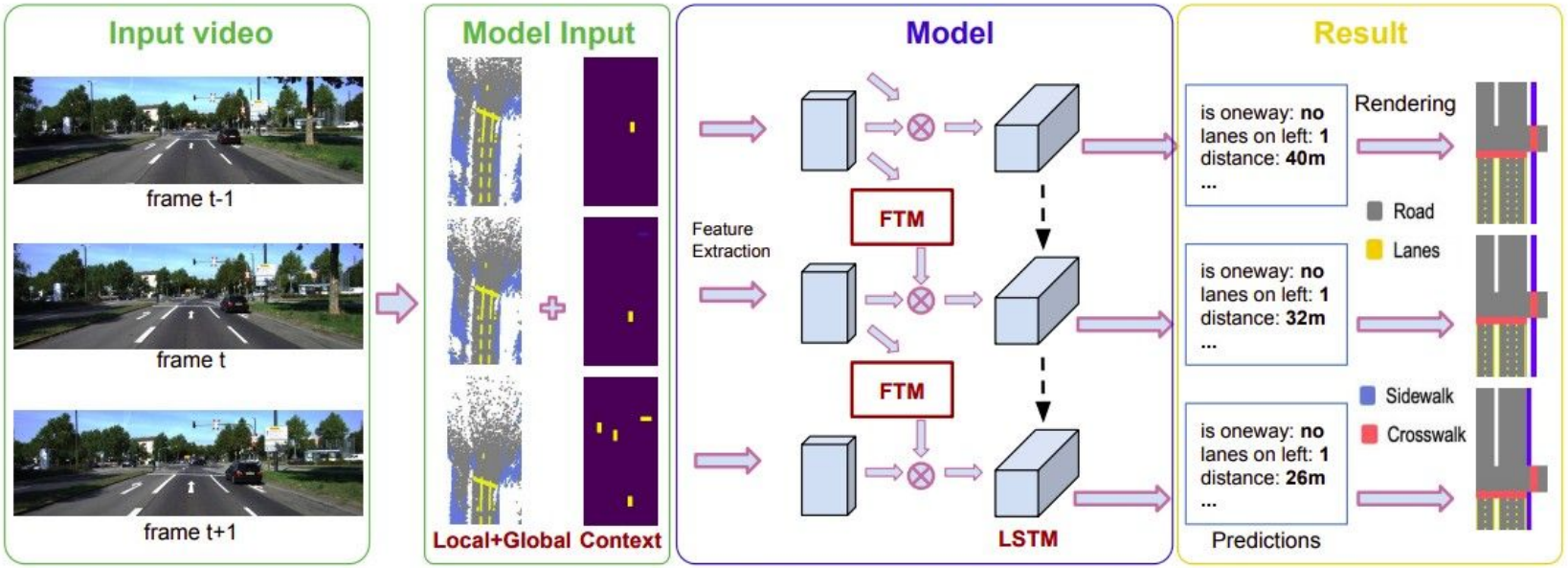
1. **Task:** given image, generate semantic BEV map.



BEV Mapping

Related work

1. **Approach 1:** explicit depth reasoning, Liu et al. (2020)
    - **Limitations**: requires depth and segmentation maps as additional input

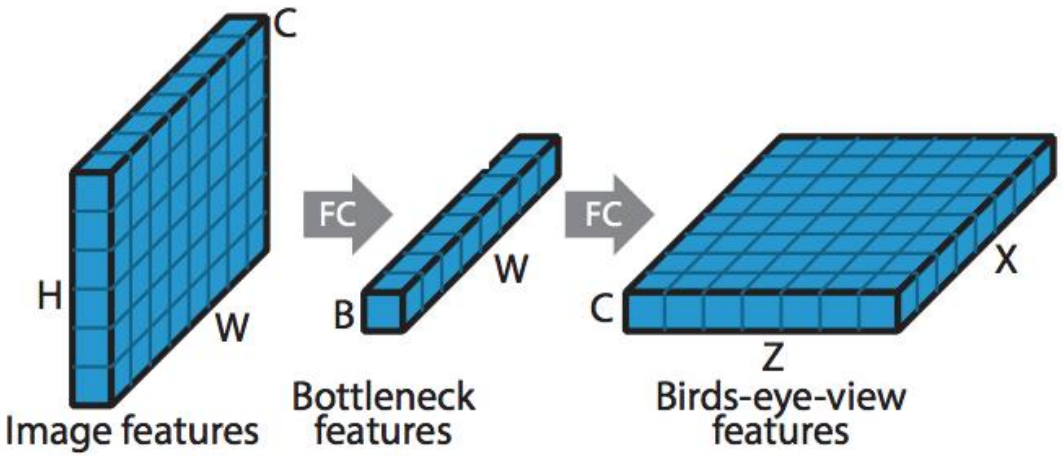


Liu, B., Zhuang, B., Schulter, S., Ji, P. and Chandraker, M., 2020. Understanding road layout from videos as a whole.

BEV Mapping

Related work

1. **Approach 2**: implicit depth reasoning, Roddick and Cipolla (2020)
    - **Limitations**: bottleneck tends to ignore small objects



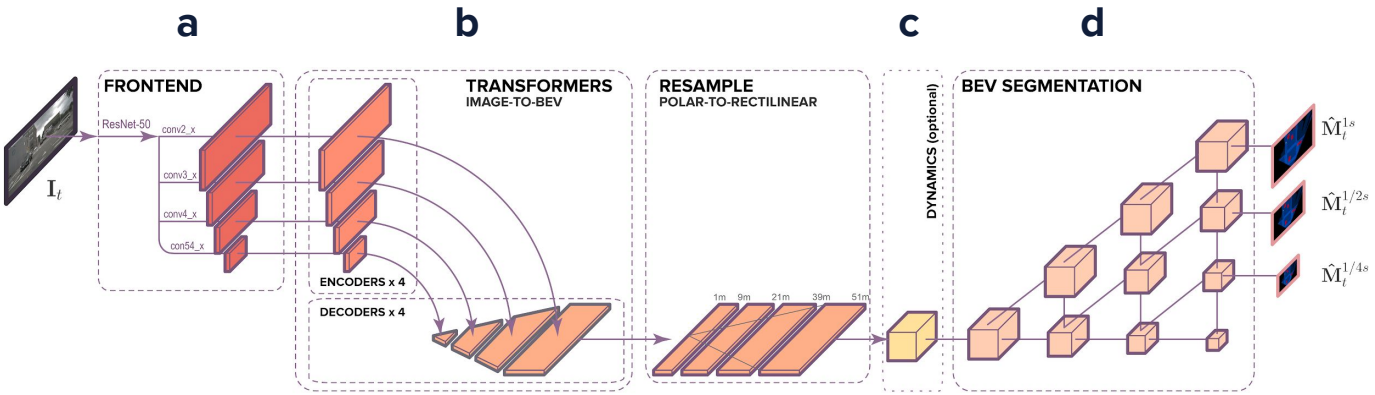
Roddick, T. and Cipolla, R., 2020. Predicting semantic map representations from images using pyramid occupancy networks.

Methodology

Saha, A., Mendez, O., Russell, C. and Bowden, R., 2022, May. Translating images into maps.

OVERVIEW

Approach



MODEL ARCHITECTURE

Our end-to-end approach

a. Construct spatial representations in the image-plane

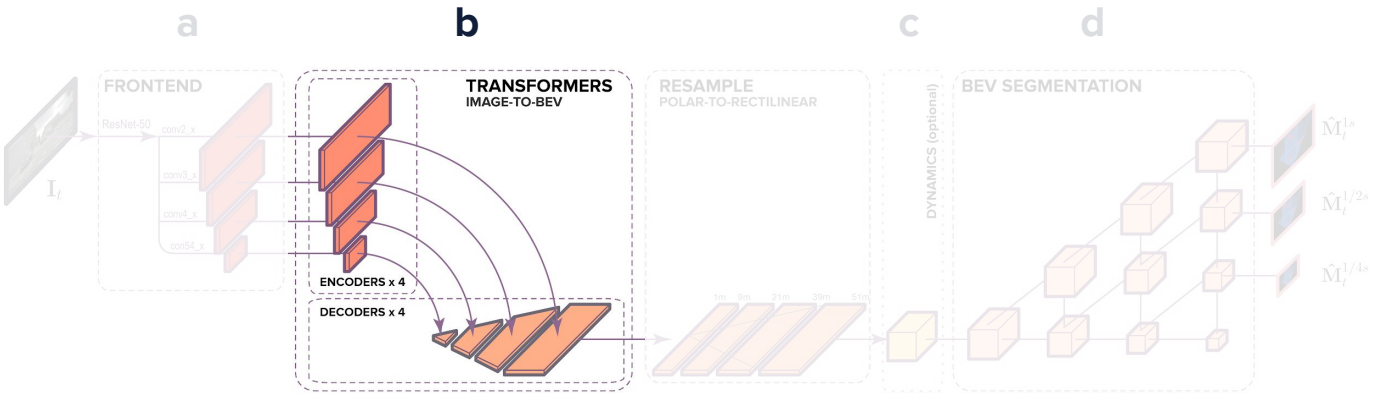
b. Transform image-plane representations to BEV

c. Construct spatiotemporal representation in BEV-plane (optional)

d. Semantically segment BEV representation

OVERVIEW

Approach



MODEL ARCHITECTURE

Our end-to-end approach

a. Construct spatial representations in the image-plane

b. Transform image-plane representations to BEV

c. Construct spatiotemporal representation in BEV-plane (optional)

d. Semantically segment BEV representation

MOTIVATION

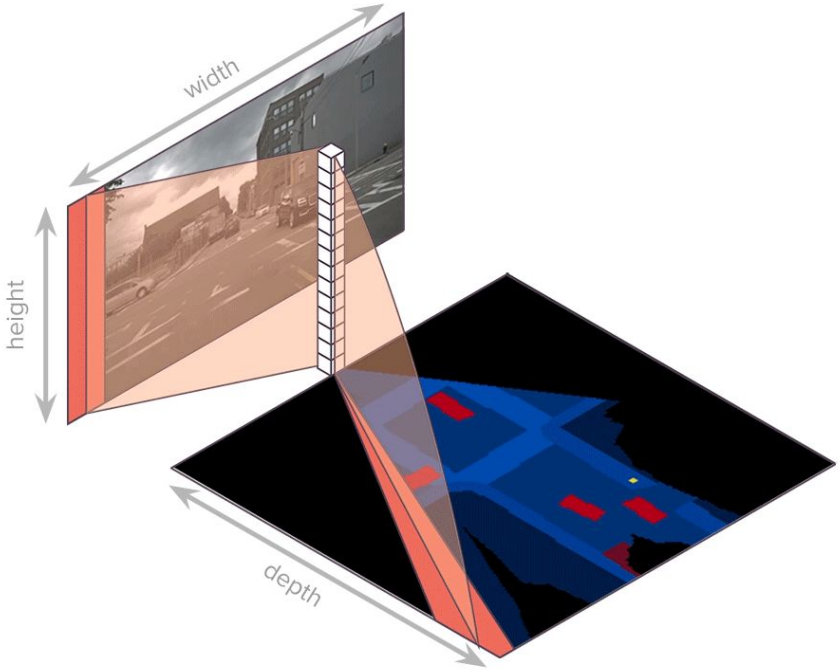
Image-to-BEV Translation

1. Image-to-BEV mapping requires image-pixel correspondence in BEV.

2. Horizontal component that each element maps to is fixed.

3. 1-1 correspondence between vertical scan line and the associated polar ray

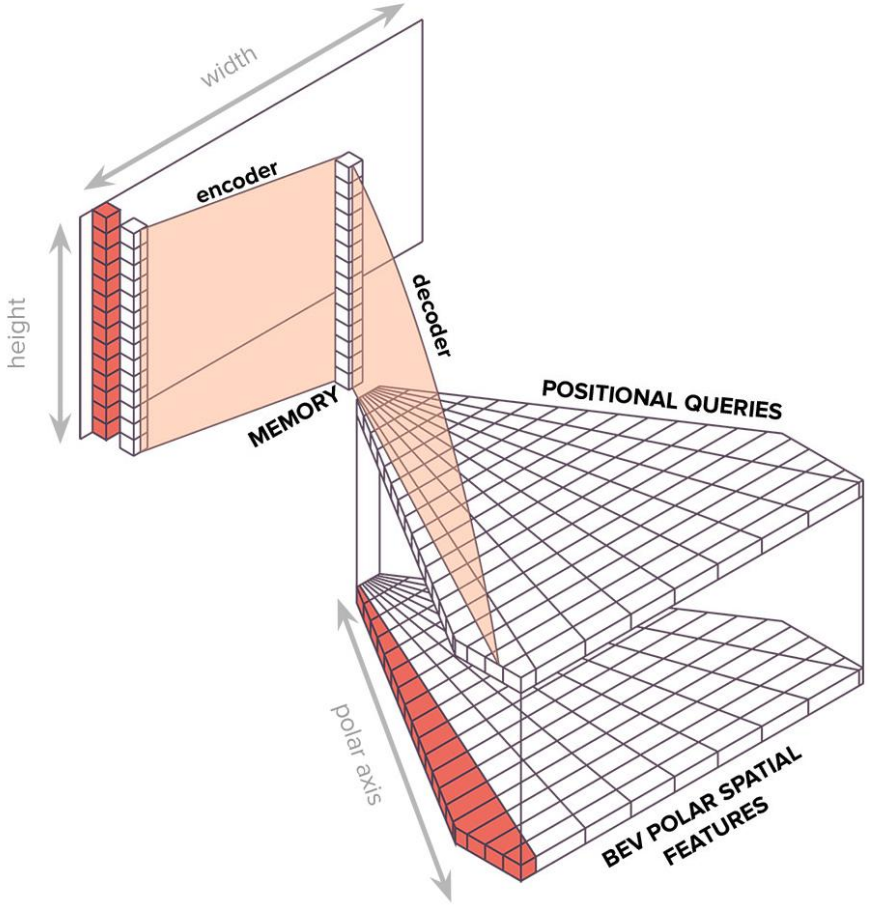
4. We treat the mapping process as a set of **sequence-to-sequence translations between scanlines in the image and rays in BEV**



METHOD

Image-to-BEV w. Transformers

1. **Transformer Encoder**:
   a. Encode vertical dependencies in image-column.
   b. Maintain spatial structure of encoded image-column in “memory”.
2. **Transformer Decoder**:
   a. Learn alignment between memory and positional BEV polar ray queries.
   b. Use alignment to distribute features in memory across polar ray.



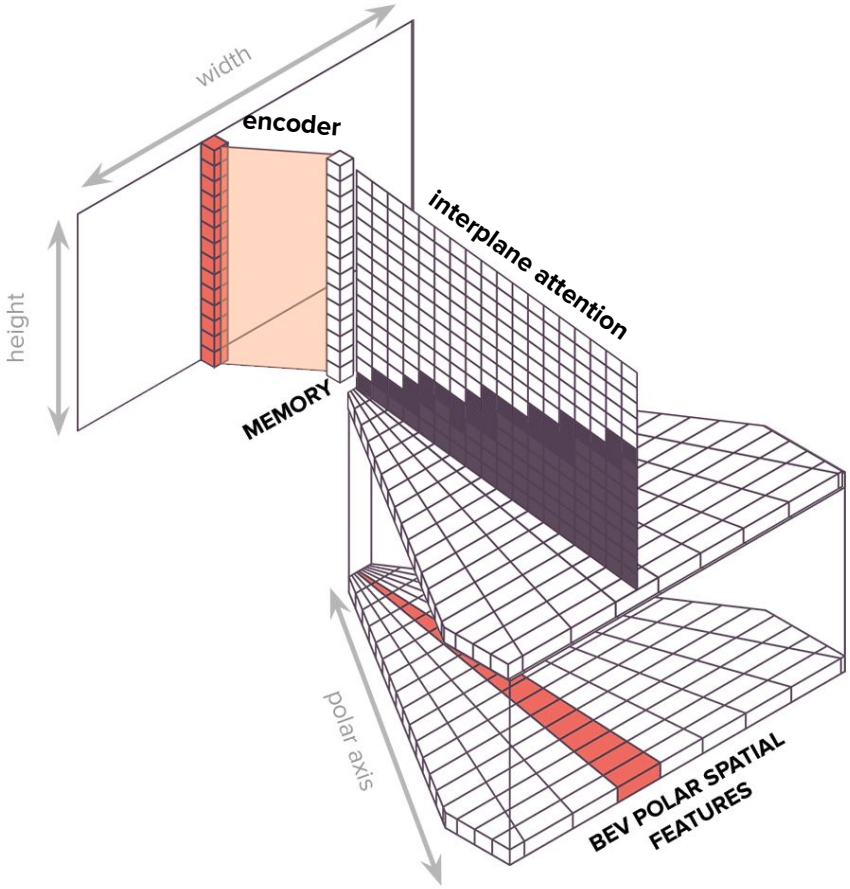
METHOD

Exploiting regularities in data

1. Transformers known to overfit - therefore limit where model looks

2. In urban environments, **depth monotonically increases with height**.

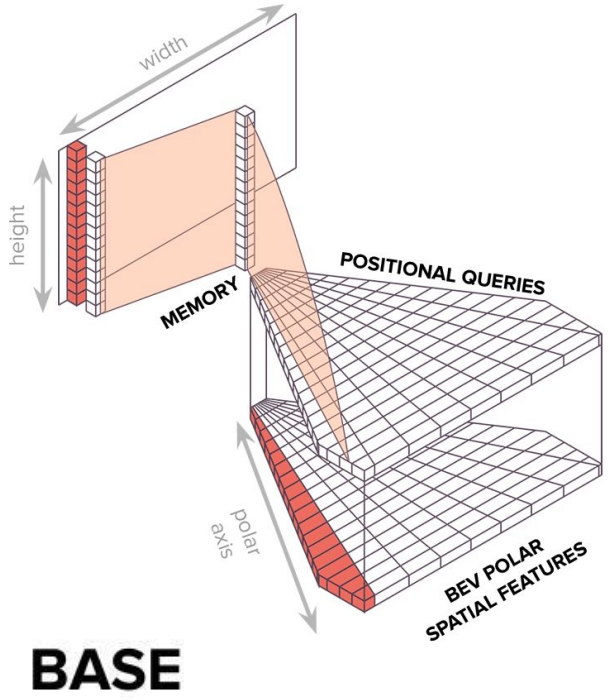
3. Enforce monotonic relationship on attention between encoder-decoder.

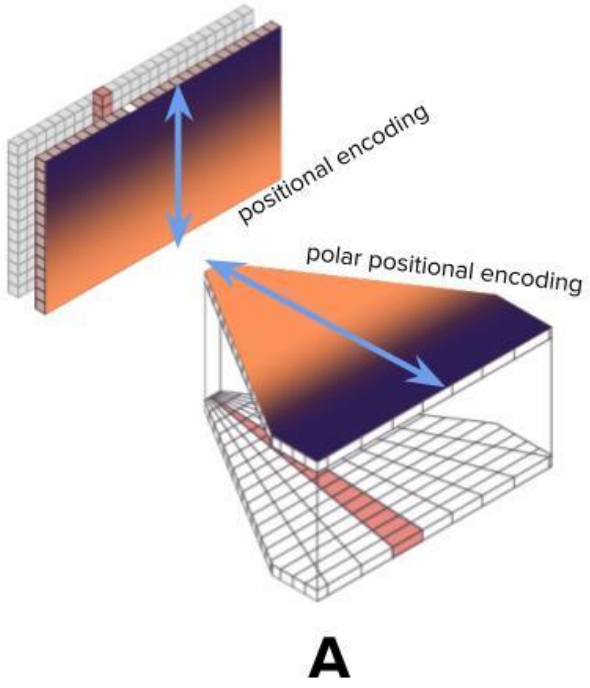


METHOD

Exploiting regularities in data

1. Transformer should be **data efficient** + capture regularities in data.





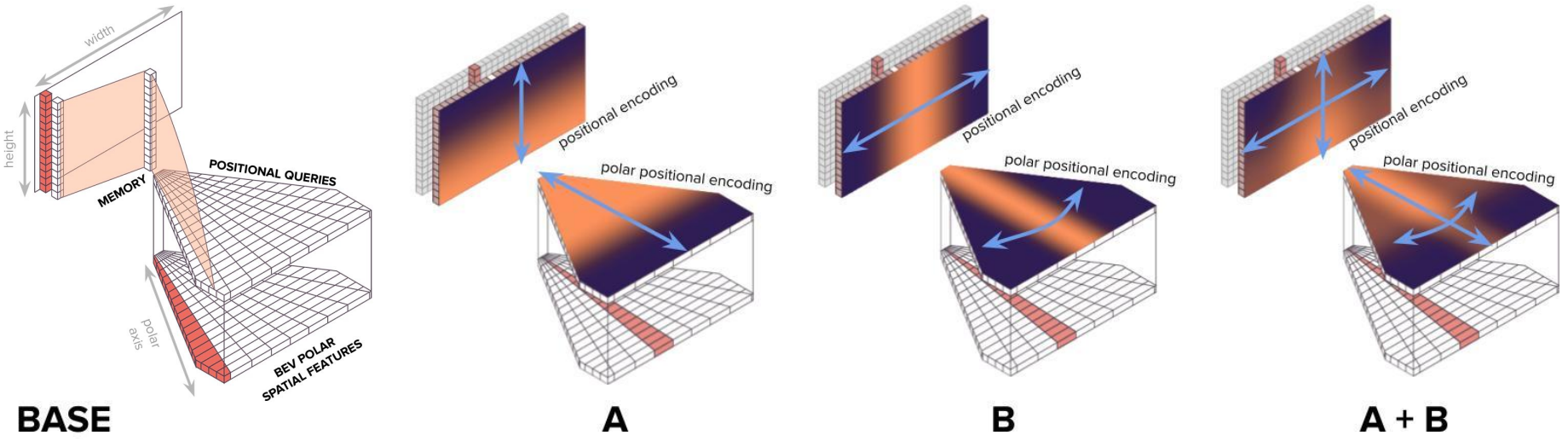
Convolutional

BASE + A = Polar Agnostic

METHOD

Exploiting regularities in data

1. **Object distribution** varies across angular domain —> use polar positional information

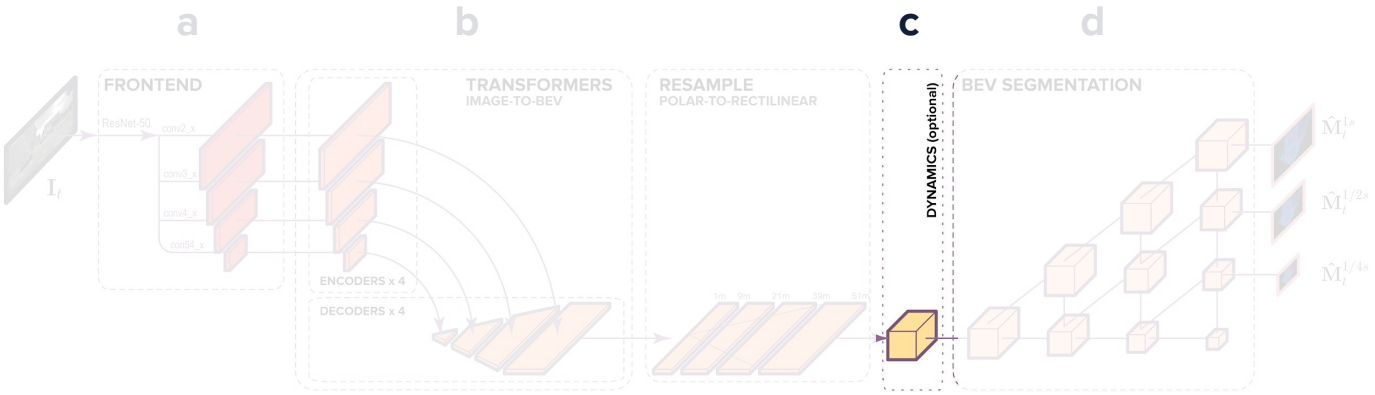


BASE **A** **B** **A + B**

Convolutional **BASE + A = Polar Agnostic** **BASE + A + B = Polar Adaptive**

METHOD

Overview



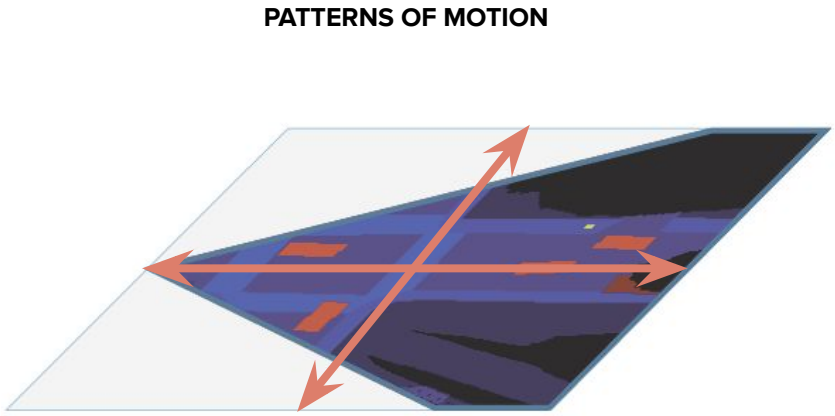
MODEL ARCHITECTURE

2. **Our end-to-end approach**
   a. Construct spatial representations in the image-plane
   b. Transform image-plane representations to BEV
   c. Construct spatiotemporal representation in BEV-plane (optional)
   d. Semantically segment BEV representation

METHOD

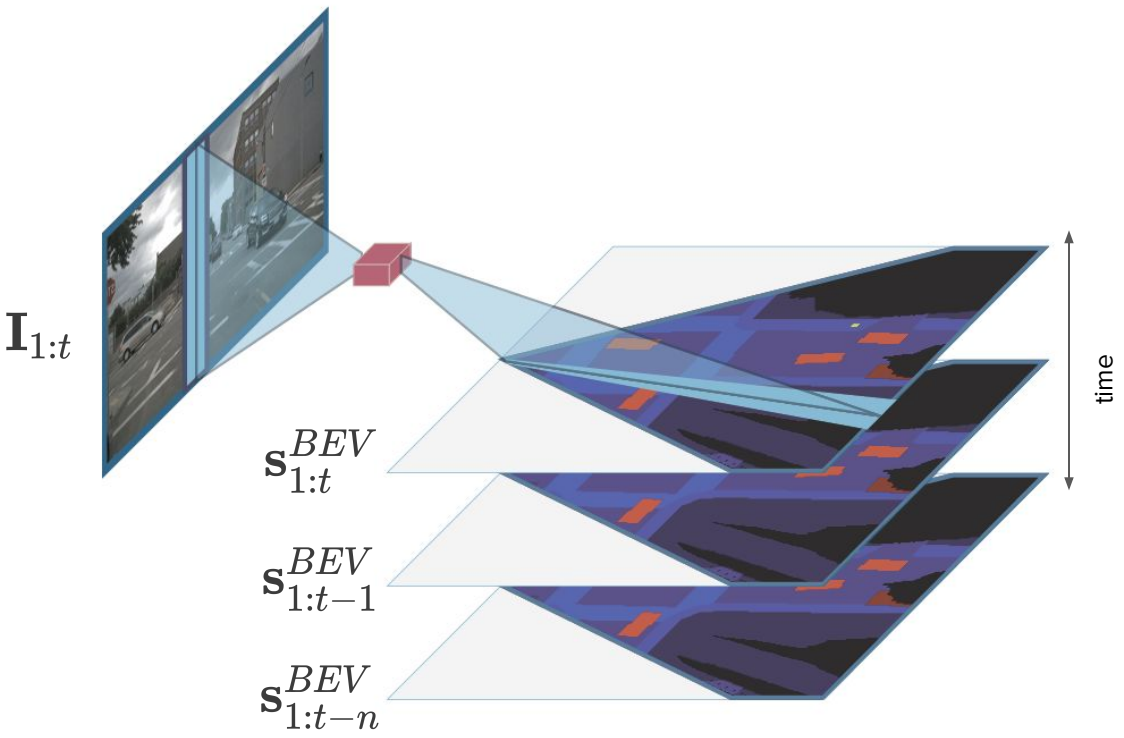
Learning grid-aligned motion

1. **Principal patterns of motion**:
   a. Parallel to ego-vehicle
   b. Perpendicular to ego-vehicle
2. Grid like motion can be learnt with **factorised 3D convolutions**.



METHOD

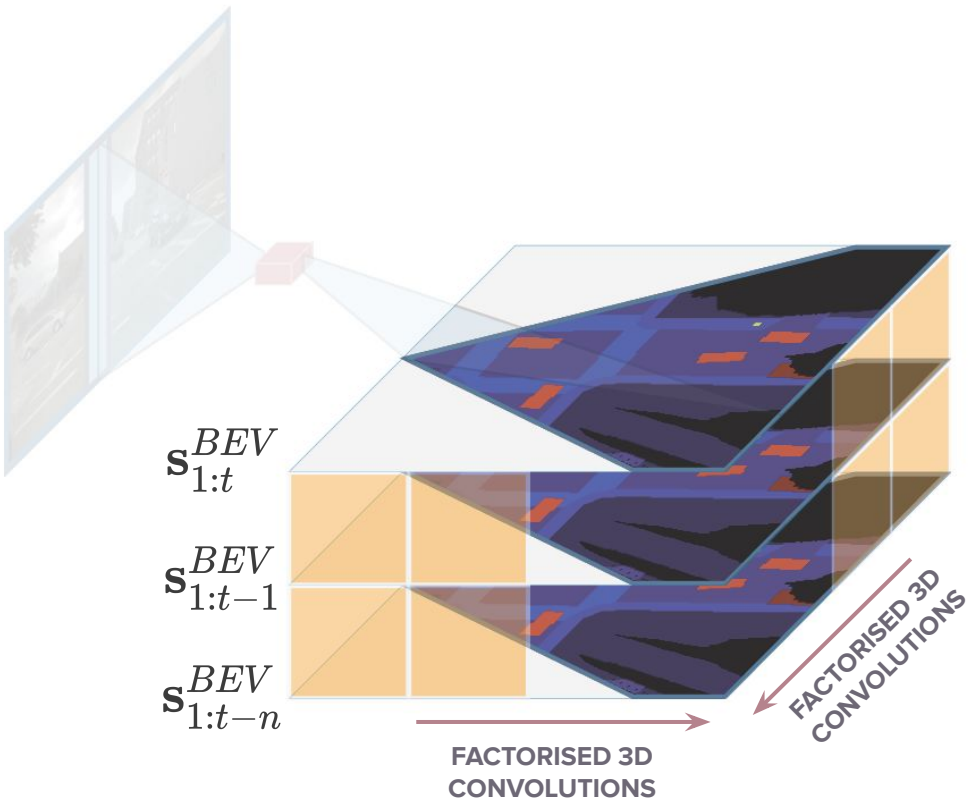
Constructing spatiotemporal representations



1. Generate BEV features for multiple timesteps

METHOD

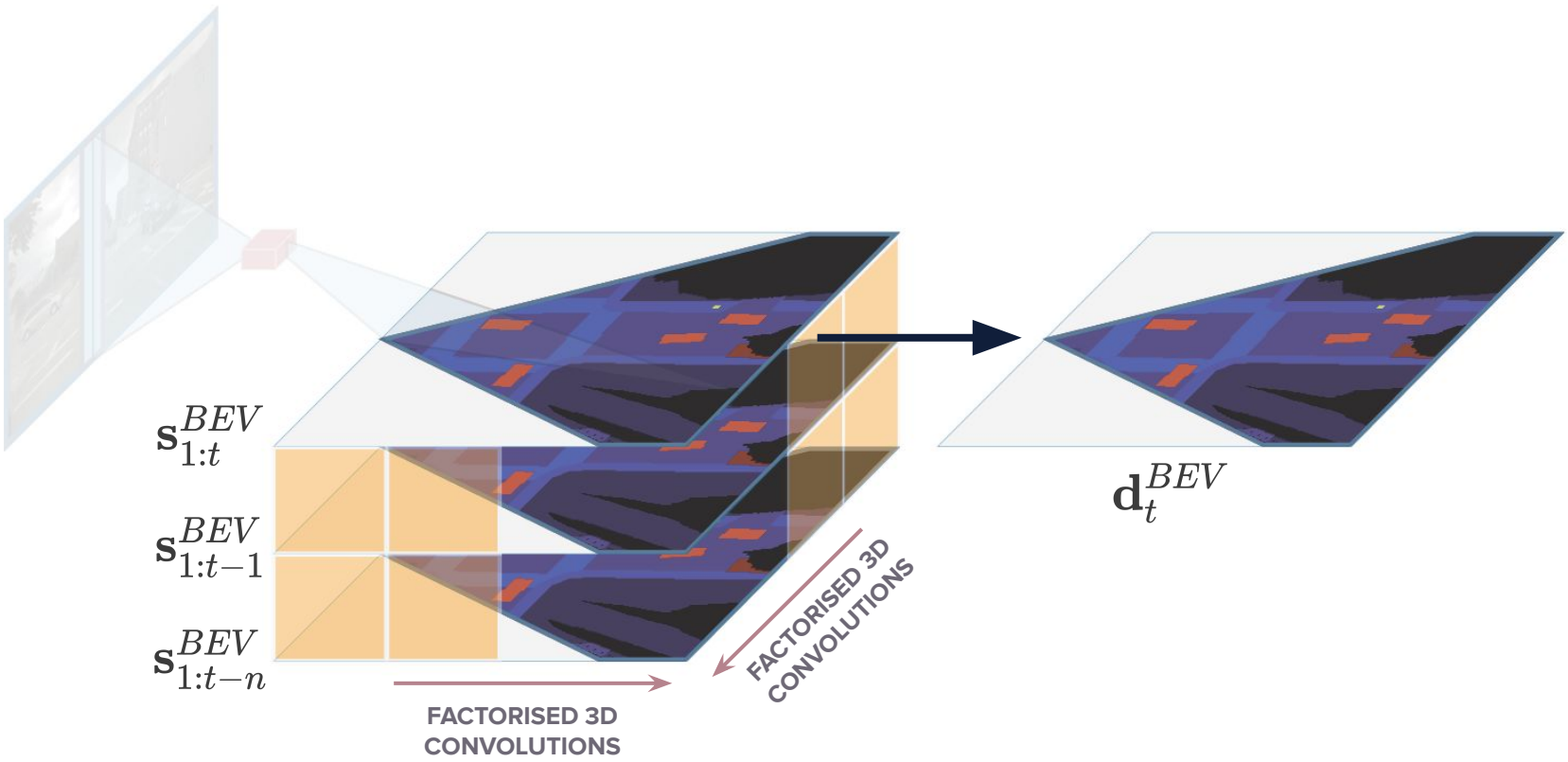
Constructing spatiotemporal representations



1. Generate BEV features for multiple timesteps
2. Learn dynamics using factorised 3D convolutions aligned to grid-like motion

METHOD

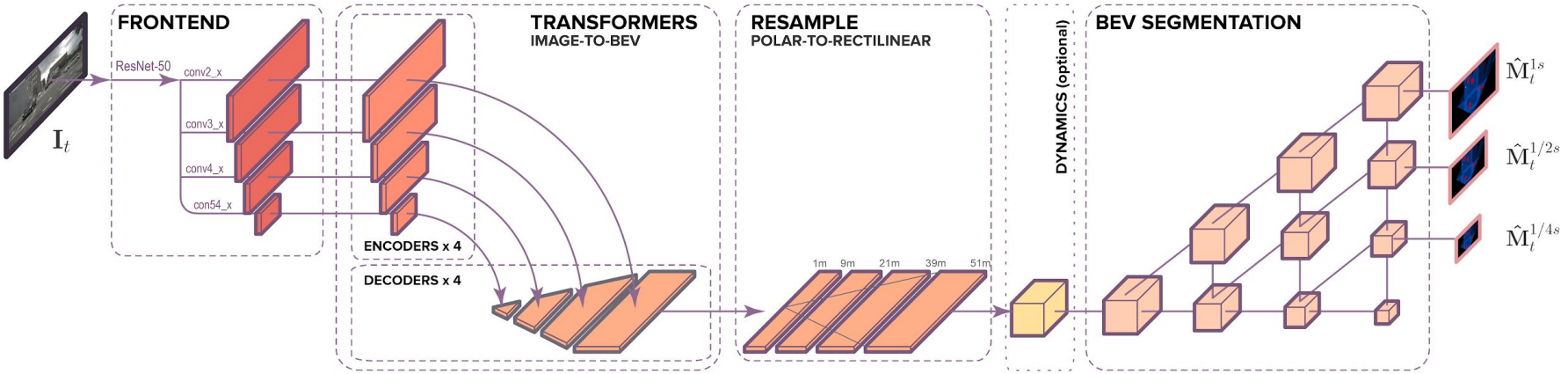
Constructing spatiotemporal representations



1. Generate BEV features for multiple timesteps
2. Learn dynamics using factorised 3D convolutions aligned to grid-like motion
3. Aggregate into spatiotemporal representation for single timestep

END-TO-END FORMULATION

Model architecture



1. **Frontend** extracts spatial features at multiple scales.
2. **Transformers** translate each scale of features to polar spatial representations at different depth ranges.
3. **Resampling** operation converts polar spatial features to rectilinear coordinate frame.
4. **3D Convolutions** learn dynamics to build a spatiotemporal BEV representation
5. **BEV Segmentation** network decodes BEV features into semantic occupancy grids.
6. **Dice Loss** applied to semantic maps at multiple scales.

EXPERIMENTS

Qualitative results

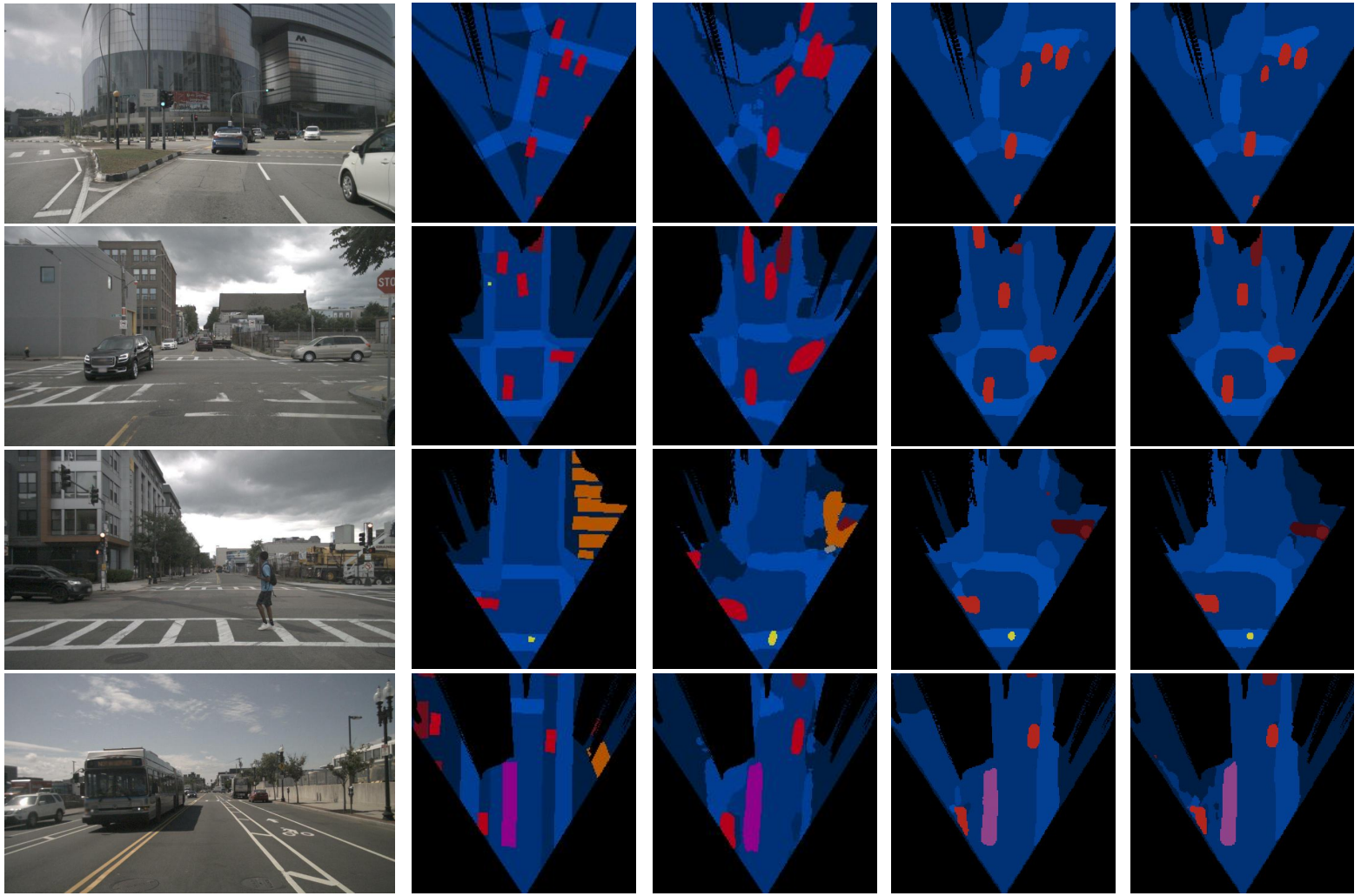


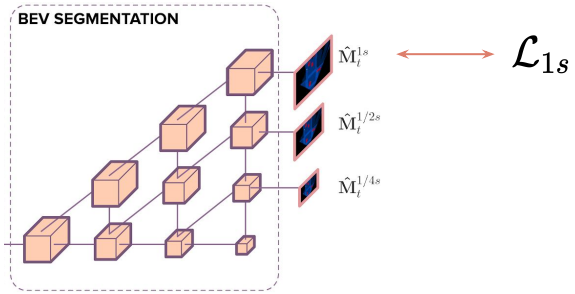
Image | Ground truth | PON | Our spatial model | Our spatiotemporal

Roddick, T. and Cipolla, R., 2020. PON: Predicting semantic map representations from images using pyramid occupancy networks.

ABLATIONS

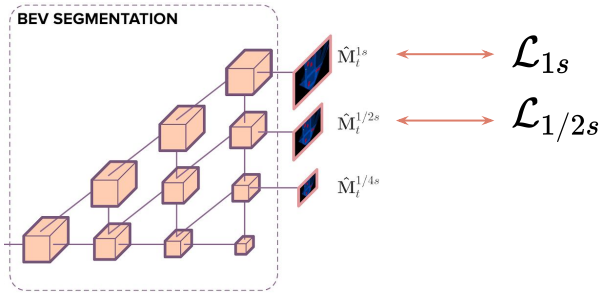
1. **Multi-scale supervision:** increases IoU by emulating an Earth Mover's Distance.

supervision scales: 1



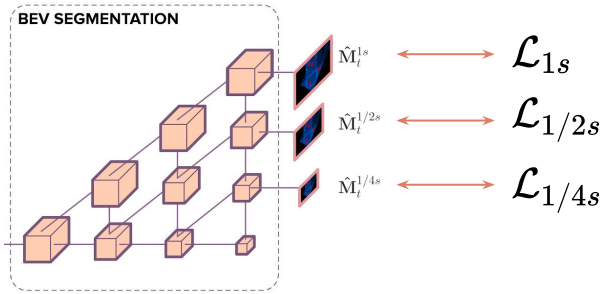
IoU = 17.3%

supervision scales: 1, 1/2



IoU = 21.6%

supervision scales: 1, ½, ¼

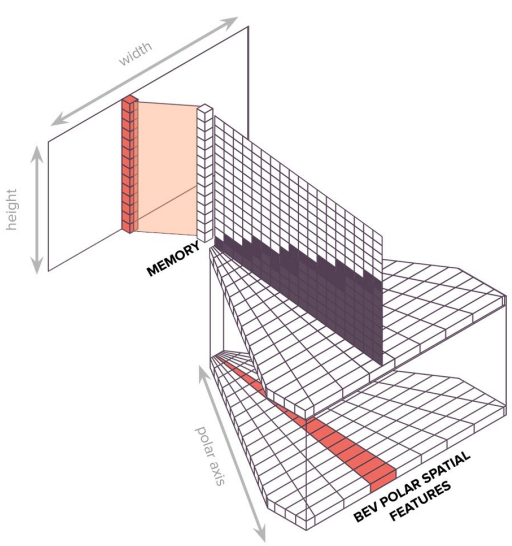


IoU = **24.7%**

ABLATIONS

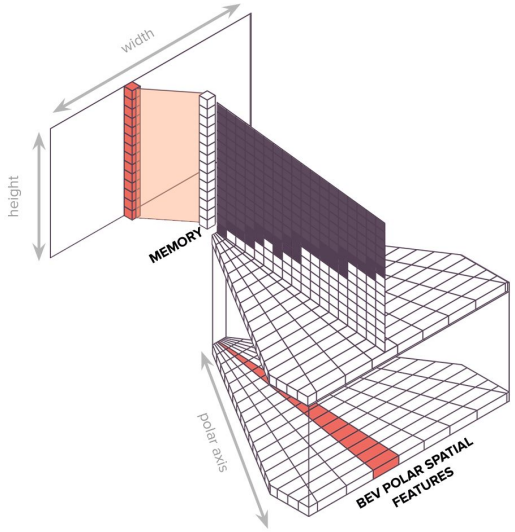
1. **Where to look?** looking downwards better than looking up, but looking in both directions is best.

Looking Up



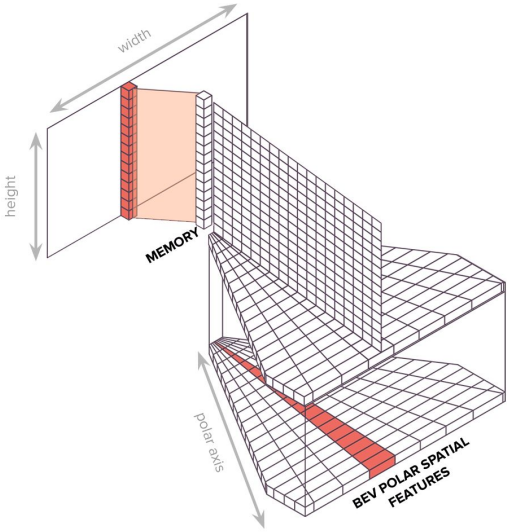
IoU = 22.1%

Looking Down



IoU = 23.0%

Looking Both Ways

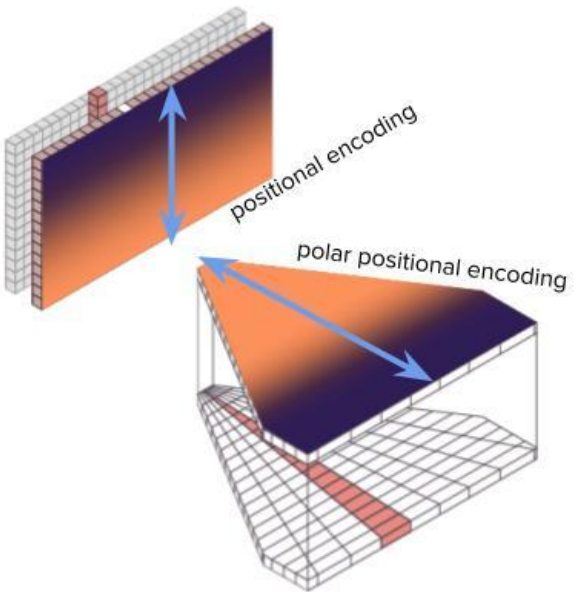


IoU = **24.7%**

ABLATIONS

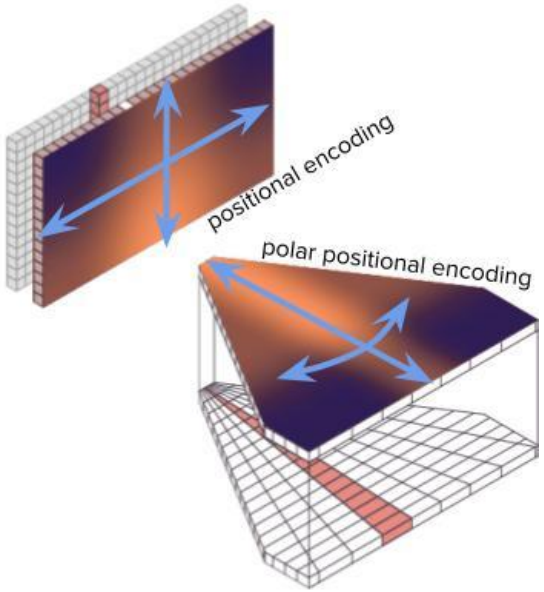
1. **Polar-agnostic vs. Polar-adaptive Translations:** polar-positional information increases object discriminativeness in the image-plane.

Polar-Agnostic



IoU = 24.7%

Polar-Adaptive

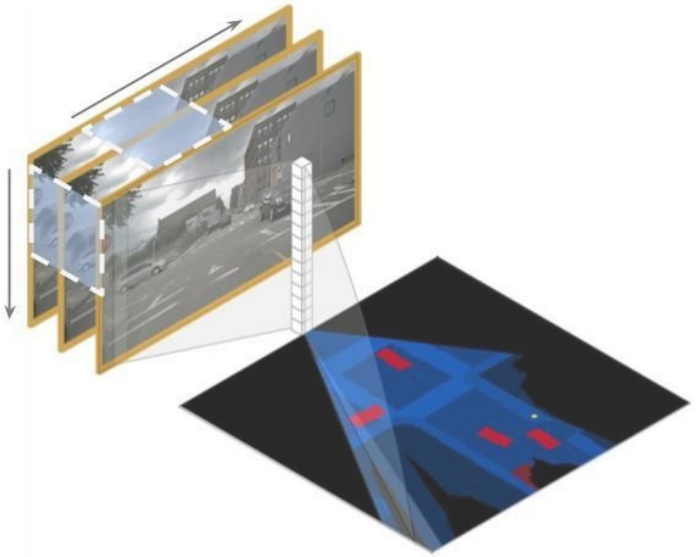


IoU = **25.7%**

ABLATIONS

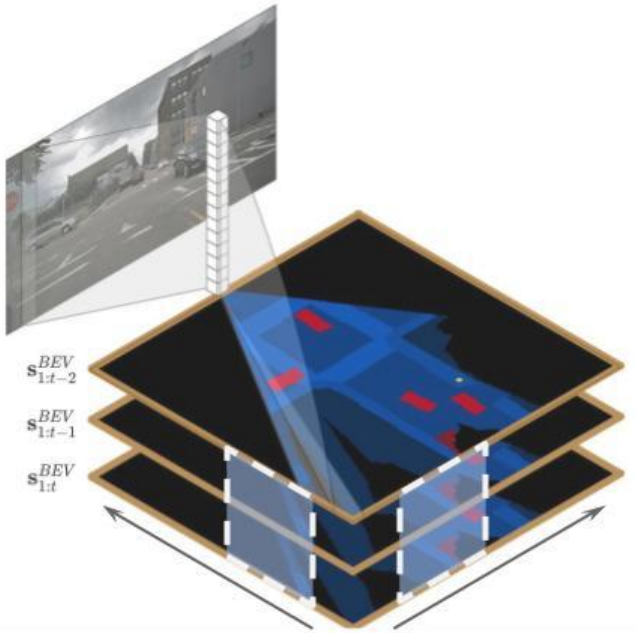
1. **Learning dynamics in BEV vs. image-plane:** motion-specific kernels are better suited to the grid-like motion seen in BEV.

Dynamics in image



IoU = 18.3%

Dynamics in BEV

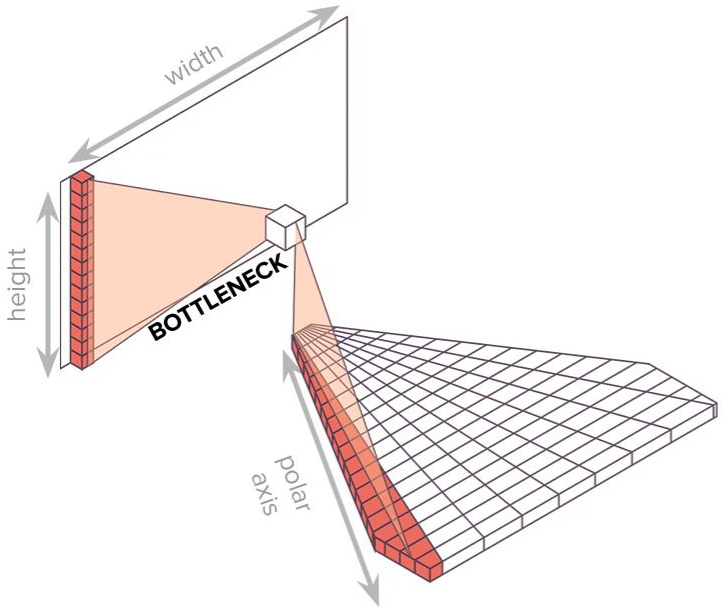


IoU = **25.4%**

ABLATIONS

Attention vs Compression

COMPRESSION

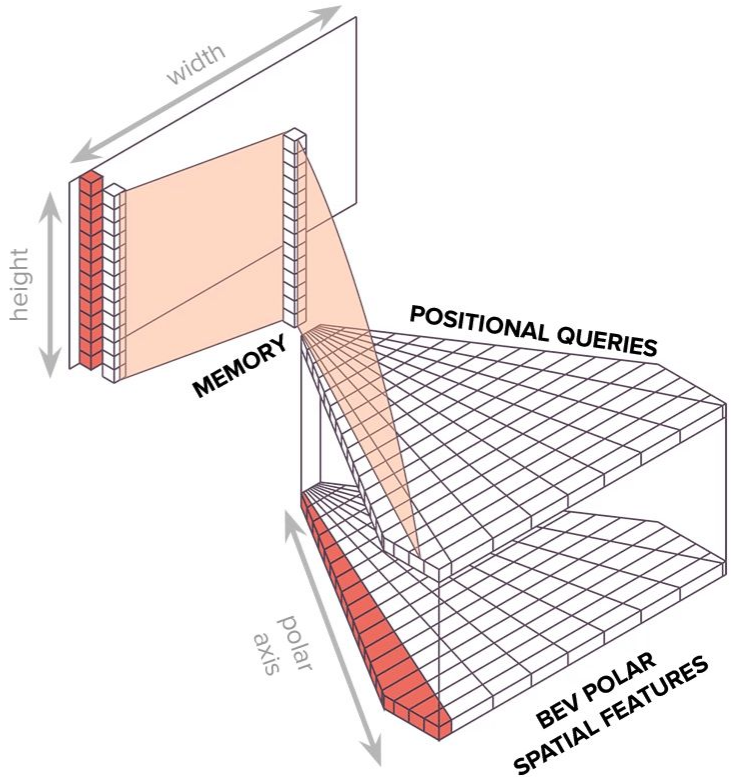


1. Compresses image-features into a **bottleneck** using a fully-connected layer.

2. Expands bottleneck along polar axis using another fully-connected layer.

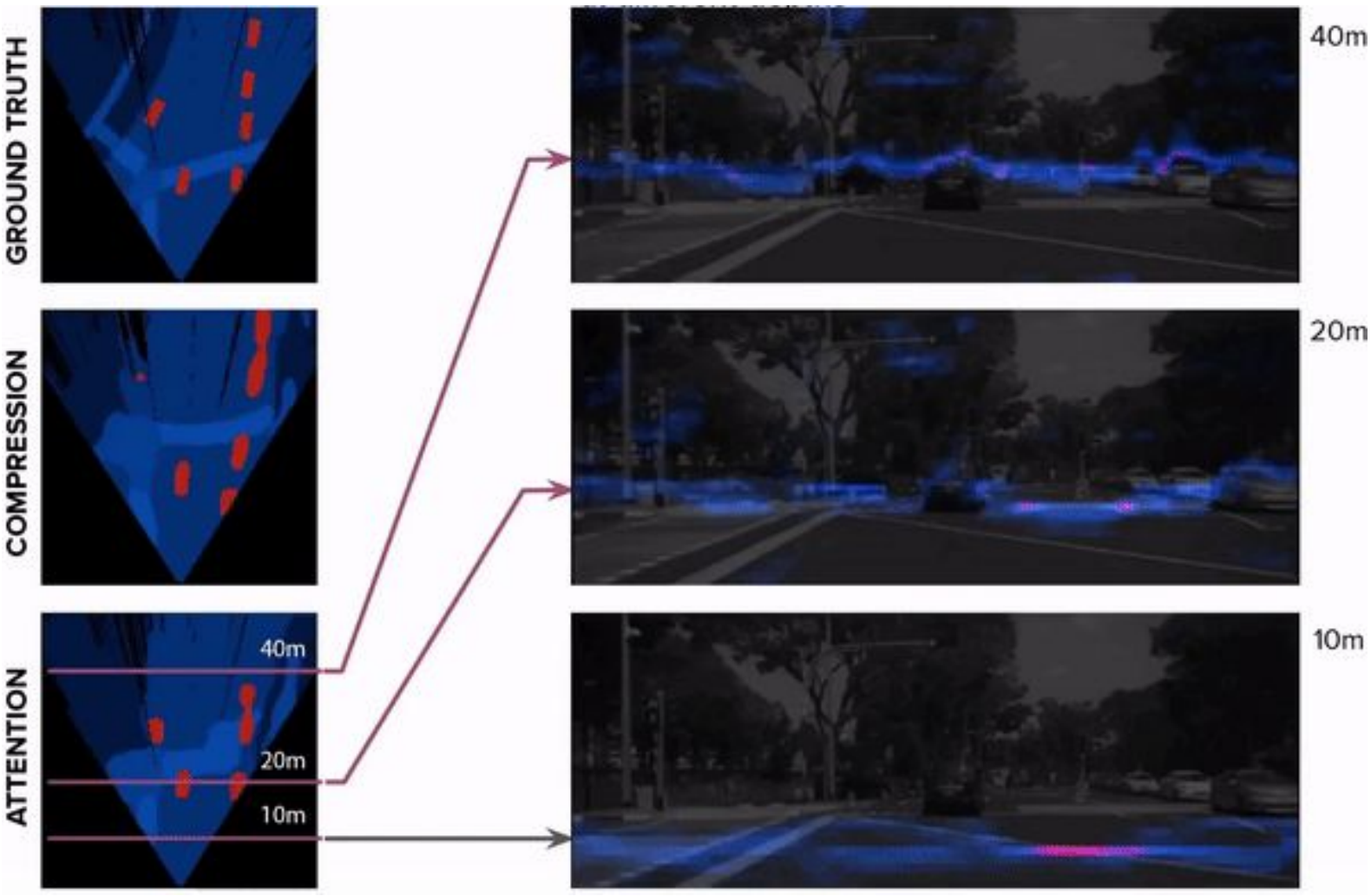
Roddick, T. and Cipolla, R., 2020. Predicting semantic map representations from images using pyramid occupancy networks.

ATTENTION



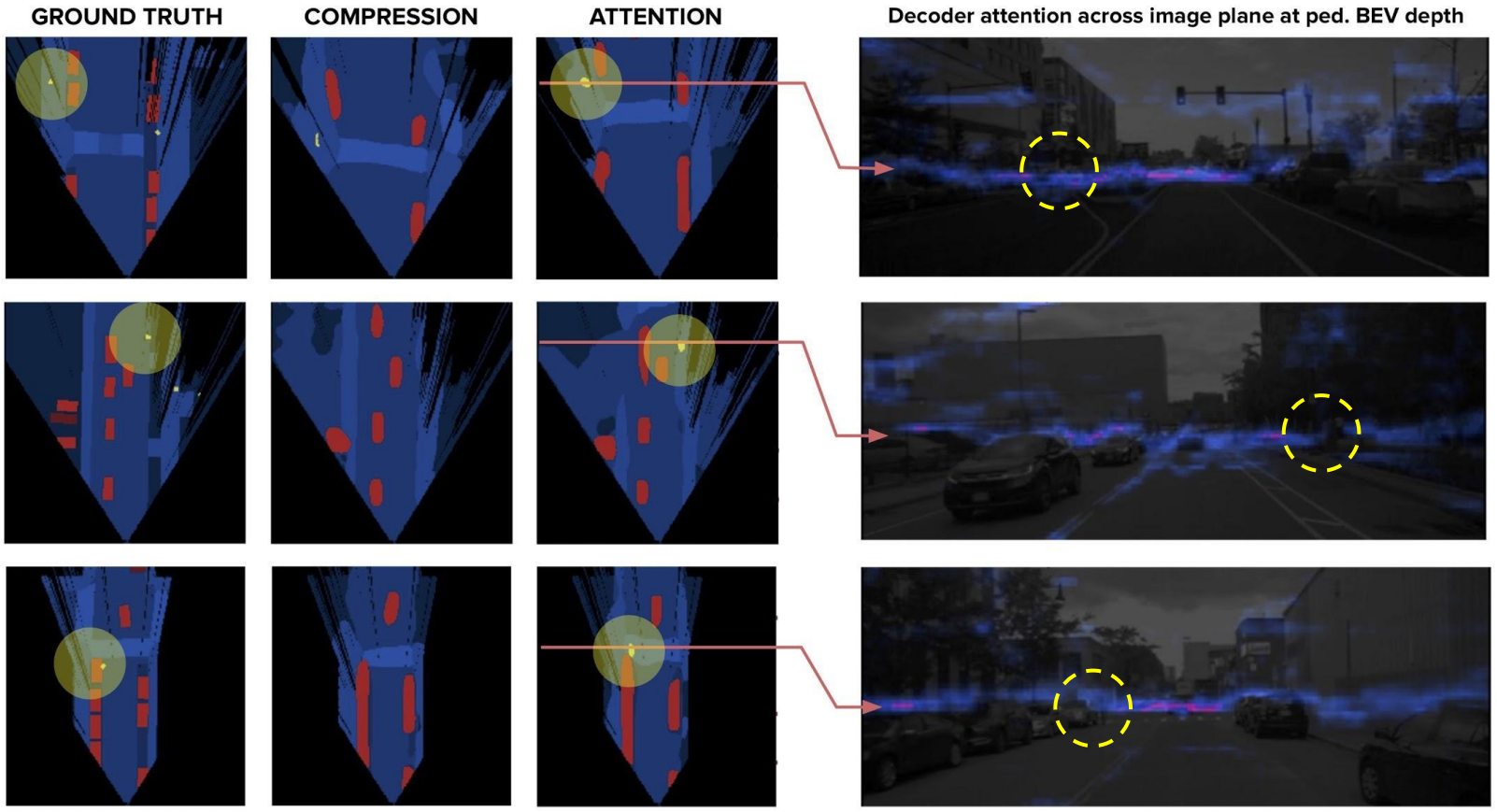
ABLATIONS

Attention vs Compression



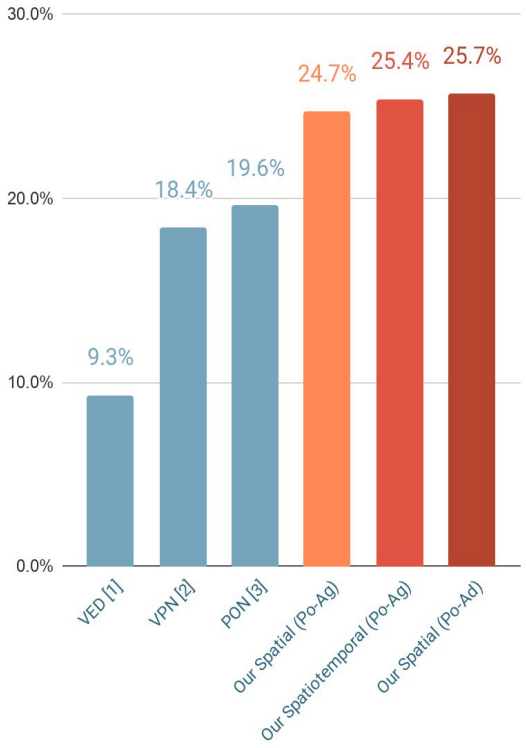
ABLATIONS

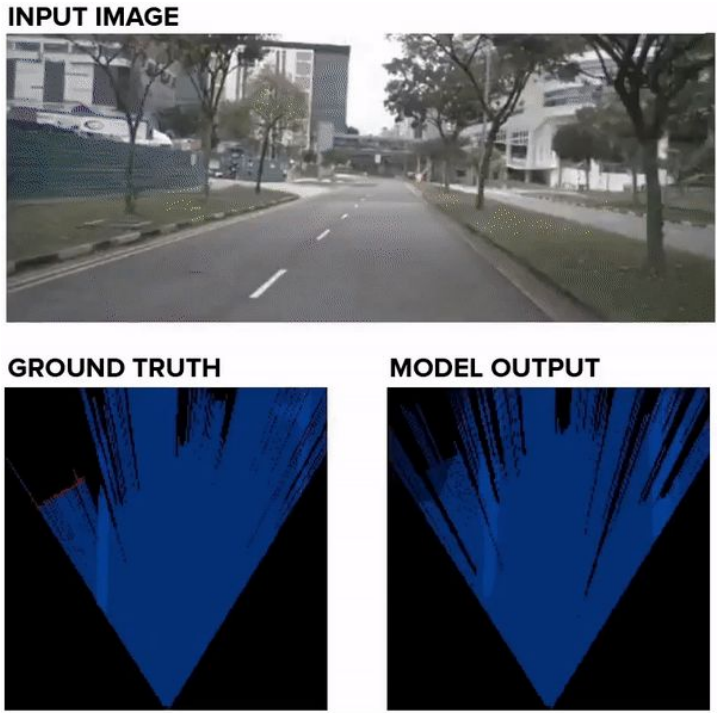
Overcoming tendency to ignore small objects in image



EXPERIMENTS

Comparison to SOTA





Method	Drivable	Ped. Cross	Walkway	Carpark	Bus	Bicycle	Car	Motorcycle	Trailer	Truck	Pedestrian	Traf. Cone	Barrier	Mean
PON [39]	60.4	28.0	**31.0**	18.4	20.8	9.4	24.7	7.0	**16.6**	16.3	**8.2**	5.7	8.1	19.6
VED [26]	54.7	12.0	20.7	13.5	0.0	0.0	8.8	0.0	7.4	0.2	0.0	0	4.0	9.3
VPN [34]	58.0	27.3	29.4	12.3	20.0	4.4	25.5	5.6	**16.6**	17.3	7.1	4.6	10.8	18.4
Our Spatial (polar-agnostic)	70.7	33.9	30.9	30.7	30.6	**15.0**	35.9	**7.9**	**15.1**	21.6	7.7	**6.9**	**14.6**	24.7
Our Spatial (polar-adaptive)	**71.7**	**34.5**	**32.7**	**31.3**	**33.5**	**16.0**	**36.2**	6.8	14.4	**26.5**	8.1	6.7	**15.9**	**25.7**
Our Spatiotemp. (polar-agnostic)	**72.6**	**34.1**	**34.2**	**31.5**	**30.9**	14.7	**38.1**	**7.4**	13.5	**23.1**	**8.3**	**7.0**	14.2	**25.4**

UNIVERSITY OF SURREY

Thanks for watching!
Questions?