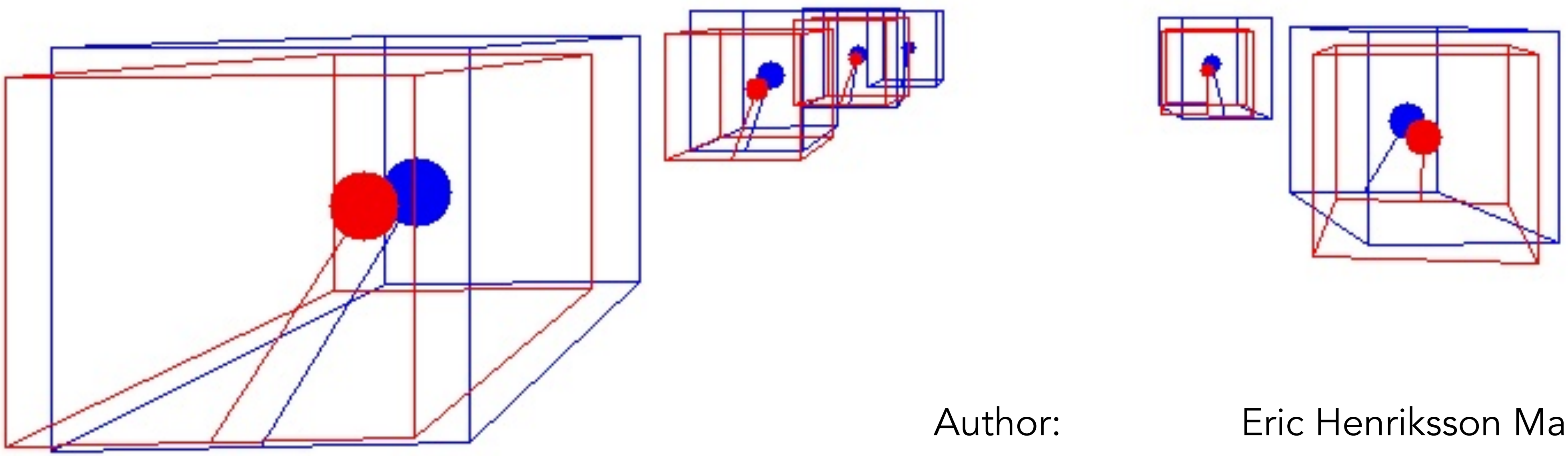
MSc Thesis in Computer Vision

Lightweight Monocular 3D Vehicle Detection
in Calibrated and Uncalibrated Scenarios

19/09/2022



Author: Eric Henriksson Martí

Supervisors: Dimosthenis Karatzas
Marçal Rossinyol

A look into the

Project's background

Images borrowed from AllRead MachineLearning Technologies S.L.

Can we go beyond 2D vehicle detection while still relying on low-resource environments?

What we mean by

Object pose recovery

6D pose estimation

- Main goal: Object pose in space
  ↳ Location and full orientation
- Instance level
- No sensors – camera alignment constraints

3D bounding box detection

- Main goal: Amodal 3D bounding boxes
  ↳ Location, dimensions, and orientation
- Category level
- Sensors – camera alignment constraints

What we mean by

Object pose recovery

6D pose estimation

- Main goal: Object pose in space
  ↳ Location and full orientation
- Instance level
- No sensors – camera alignment constraints

3D bounding box detection

- Main goal: Amodal 3D bounding boxes
  ↳ Location, dimensions, and orientation
- Category level
- Sensors – camera alignment constraints

What we mean by

Object pose recovery

6D pose estimation

- Main goal: Object pose in space
  ↳ Location and full orientation
- Instance level
- No sensors – camera alignment constraints

3D bounding box detection

- Main goal: Amodal 3D bounding boxes
  ↳ Location, dimensions, and orientation
- Category level
- Sensors – camera alignment constraints

A comparison between

2D and 3D vehicle detection

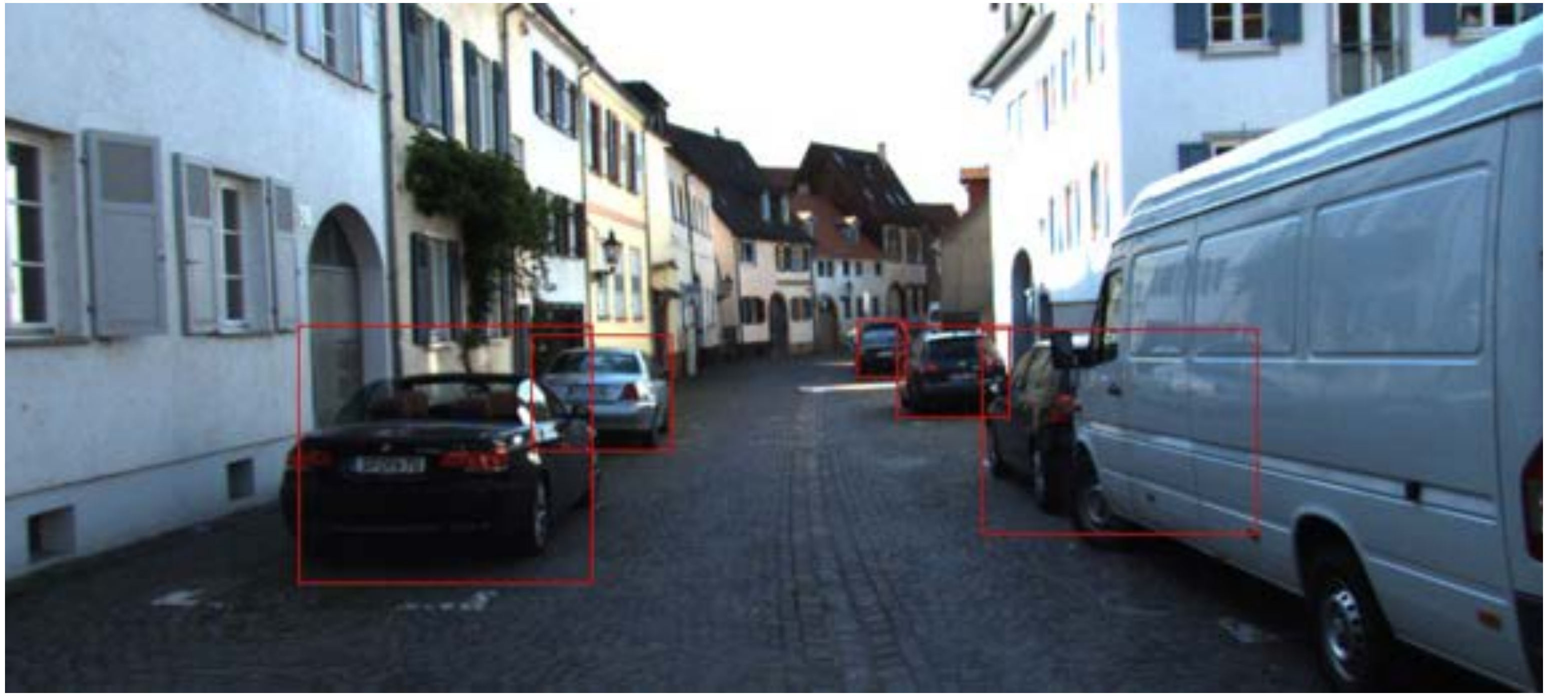


Image borrowed from the KITTI vision benchmark suite

Enables: Classification, counting, basic tracking and speed monitoring

?

No information beyond the image plane

A comparison between

2D and 3D vehicle detection

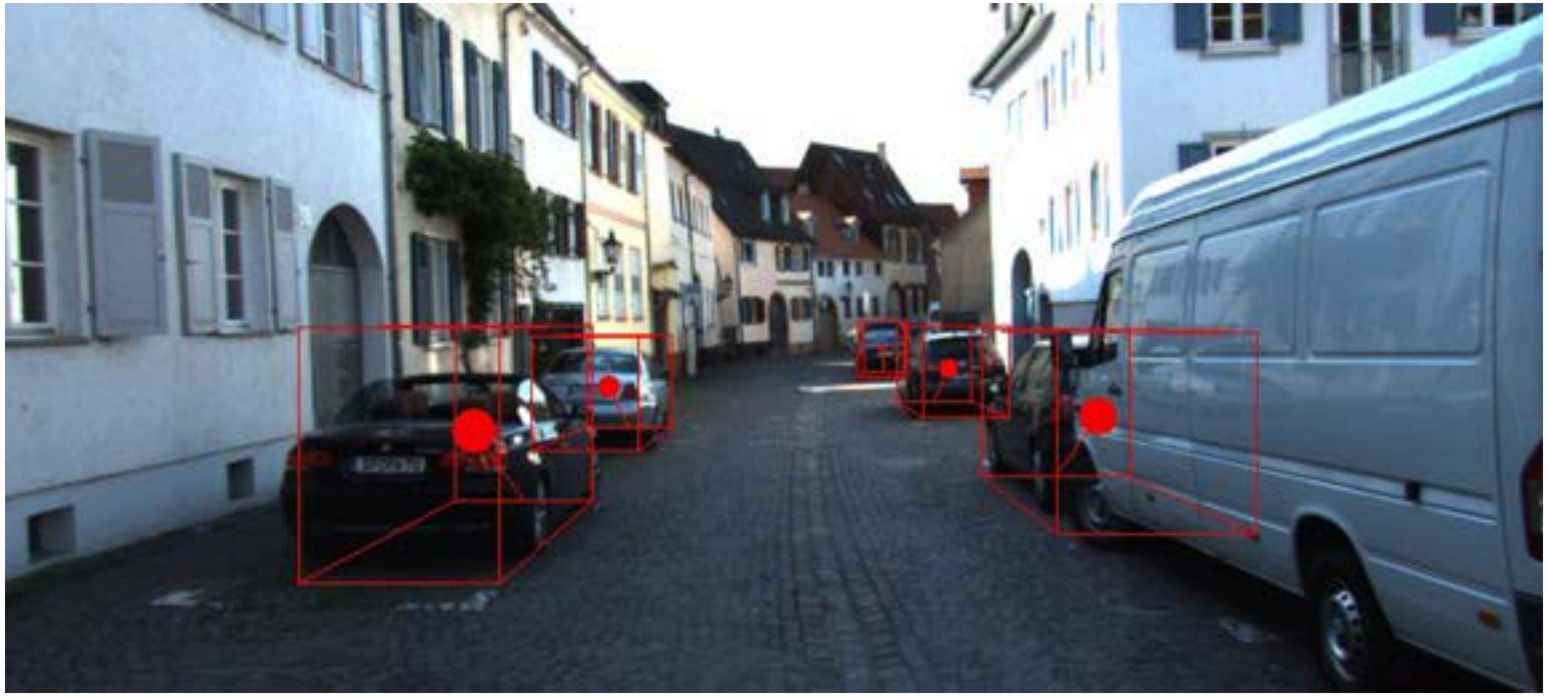
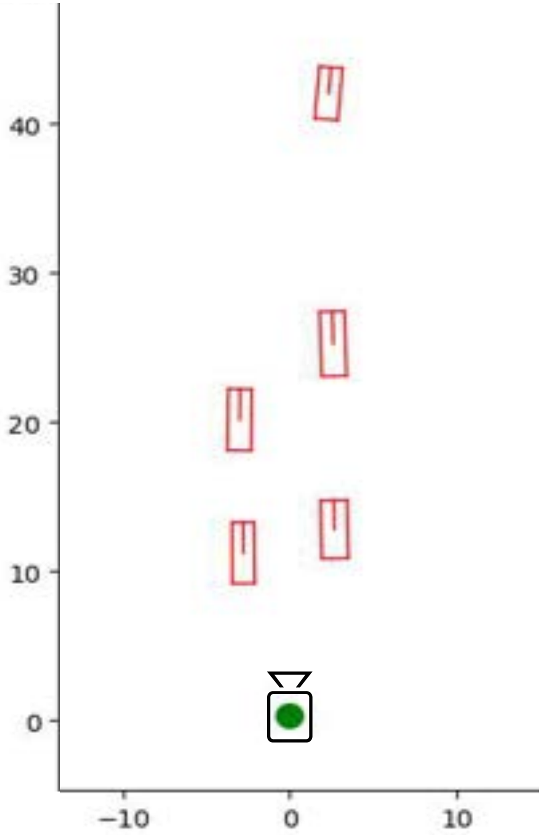


Image borrowed from the KITTI vision benchmark suite



Enables: Classification, counting, advanced tracking and speed monitoring

+

Localization, estimation of occupied volumes, etc.

An overview of

Kinds of inputs

RGB images

- Easily accessible
- No explicit depth information

Stereo images

- Better depth perception
- Two cameras are not always available

3D shape priors

- Better interpretations of object geometries
- Variety in vehicle designs

Depth maps

- Direct depth information
- Requires either RGB-D sensors or pre-computed maps

LiDAR point clouds

- Direct depth information
- Expensive and bulky sensors

An overview of

Kinds of inputs

RGB images

- Easily accessible
- No explicit depth information

Stereo images

- Better depth perception
- Two cameras are not always available

3D shape priors

- Better interpretations of object geometries
- Variety in vehicle designs

Depth maps

- Direct depth information
- Requires either RGB-D sensors or pre-computed maps

LiDAR point clouds

- Direct depth information
- Expensive and bulky sensors

An overview of

Kinds of inputs

RGB images

- Easily accessible
- No explicit depth information

Stereo images

- Better depth perception
- Two cameras are not always available

3D shape priors

- Better interpretations of object geometries
- Variety in vehicle designs

Depth maps

- Direct depth information
- Requires either RGB-D sensors or pre-computed maps

LiDAR point clouds

- Direct depth information
- Expensive and bulky sensors

An overview of

Kinds of inputs

RGB images

- Easily accessible
- No explicit depth information

Stereo images

- Better depth perception
- Two cameras are not always available

3D shape priors

- Better interpretations of object geometries
- Variety in vehicle designs

Depth maps

- Direct depth information
- Requires either RGB-D sensors or pre-computed maps

LiDAR point clouds

- Direct depth information
- Expensive and bulky sensors

An overview of

Kinds of inputs

RGB images

- Easily accessible
- No explicit depth information

Stereo images

- Better depth perception
- Two cameras are not always available

3D shape priors

- Better interpretations of object geometries
- Variety in vehicle designs

Depth maps

- Direct depth information
- Requires either RGB-D sensors or pre-computed maps

LiDAR point clouds

- Direct depth information
- Expensive and bulky sensors

The challenge of

Monocular 3D detection

Image borrowed from the KITTI vision benchmark suite

Where is the black Renault Megane located in the image?

How far is it from the camera?

What space does it occupy in the image?

The challenge of

Monocular 3D detection

Image borrowed from the KITTI vision benchmark suite

Where is the black Renault Megane located in the image?

How far is it from the camera?

What space does it occupy in the image?

The challenge of

Monocular 3D detection

Image borrowed from the KITTI vision benchmark suite

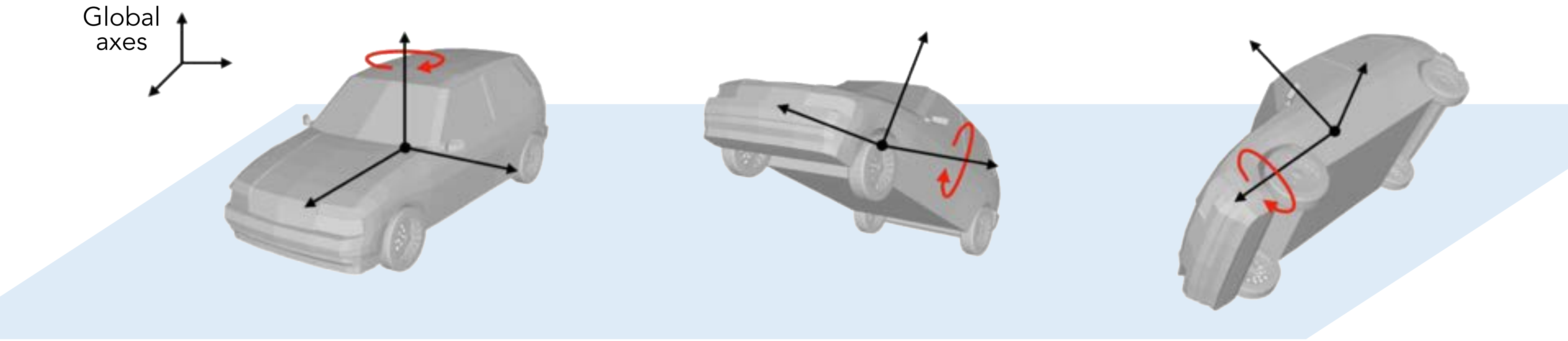
Where is the black Renault Megane located in the image?

How far is it from the camera?

What space does it occupy in the image?

The basics of

3D bounding box representation



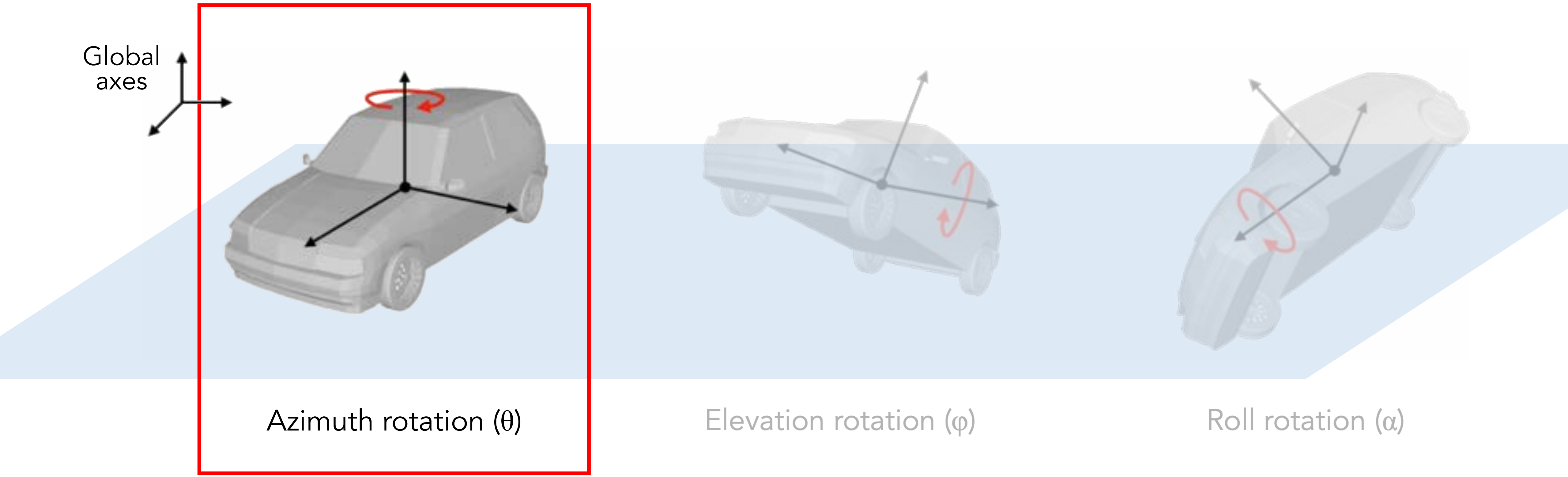
Azimuth rotation (θ)

Elevation rotation (φ)

Roll rotation (α)

The basics of

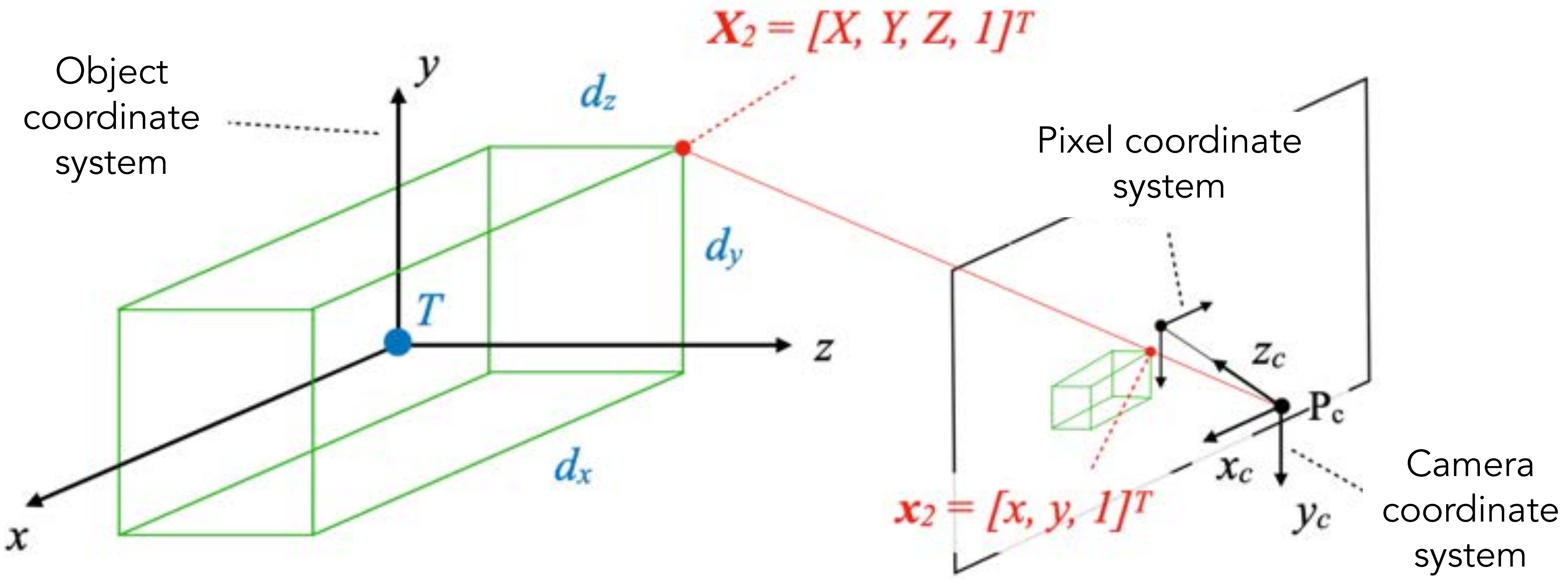
3D bounding box representation



"Flat earth" assumption

The basics of

3D bounding box representation



$$\mathbf{x}_i = K[R\ T]\mathbf{X}_i$$

A matter of

Calibration and 3D labeling

Ideally we have access to:

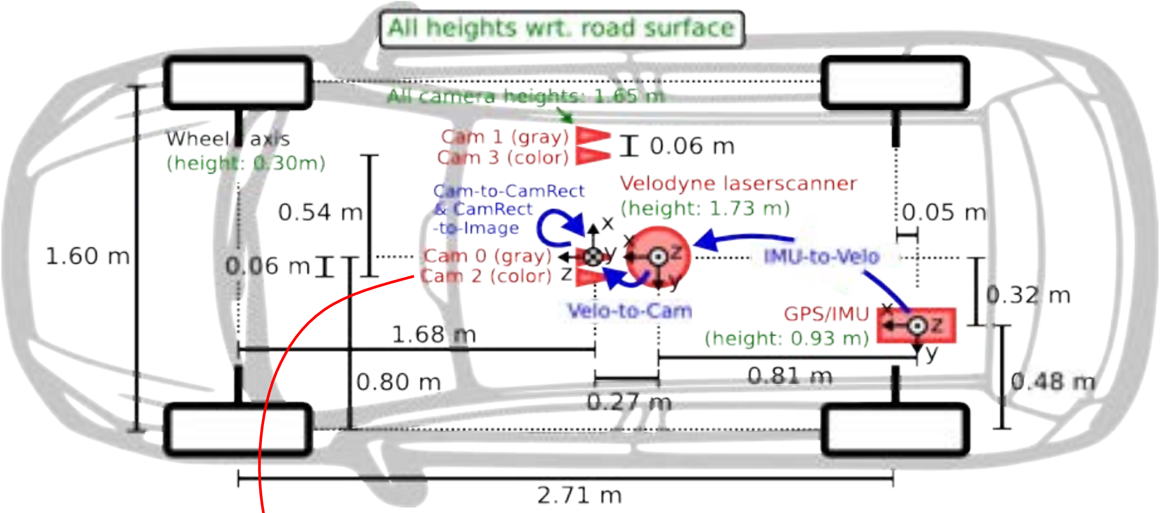


Image borrowed from the KITTI vision benchmark suite

$$P2 = \begin{bmatrix} 707.049 & 0 & 604.081 & 45.758 \\ 0 & 707.049 & 180.507 & -0.345 \\ 0 & 0 & 1 & 0.005 \end{bmatrix}$$

Camera **extrinsics** and **intrinsics**

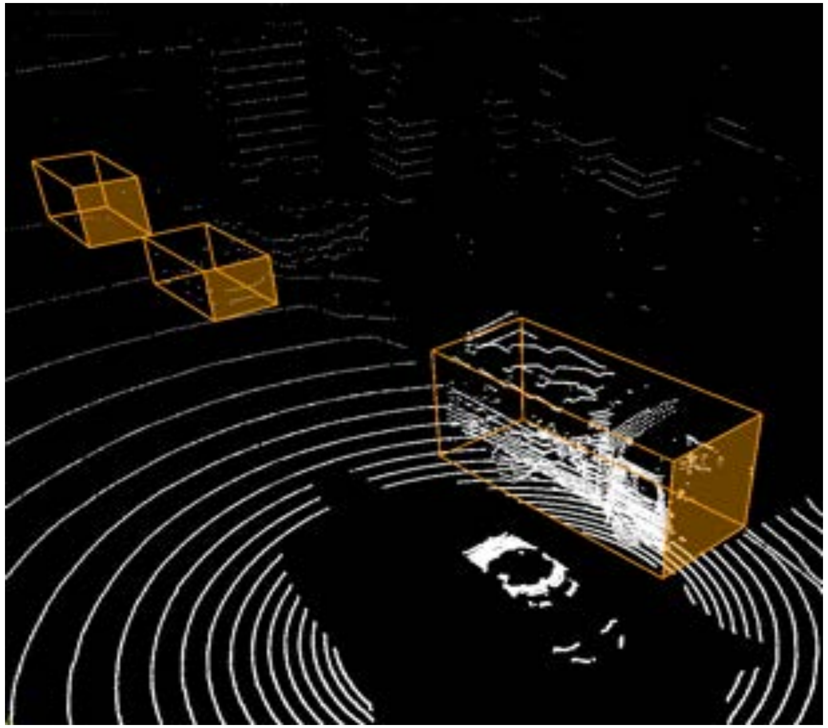


Image borrowed from Motional's NuScenes visualizer

3D bounding box annotations

A matter of

Calibration and 3D labeling

Ideally we have access to:

However, this is usually not the case

↓

Need to simplify the problem

2.71 m

Image borrowed from the KITTI vision benchmark suite

$$P2 = \begin{bmatrix} 707.049 & 0 & 604.081 & 45.758 \\ 0 & 707.049 & 180.507 & -0.345 \\ 0 & 0 & 1 & 0.005 \end{bmatrix}$$

Camera **extrinsics** and **intrinsics**

Image borrowed from Motional's NuScenes visualizer

3D bounding box annotations

The idea behind

CenterNet: Objects as points

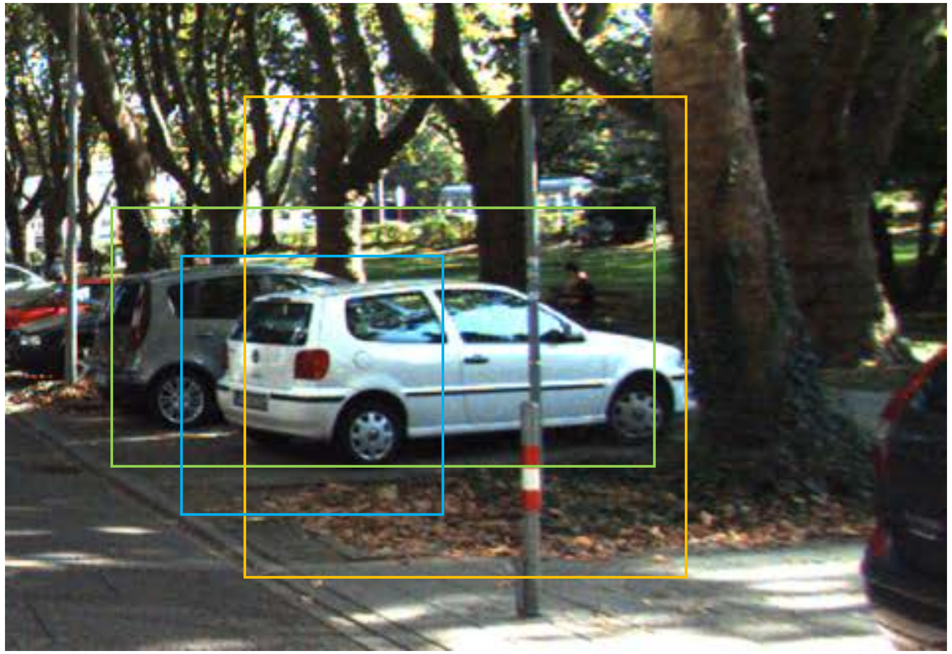


Image borrowed from the KITTI vision benchmark suite

Anchor boxes + refinement + NMS

Object center points + parameters

The idea behind

CenterNet: Objects as points

Image borrowed from the KITTI vision benchmark suite

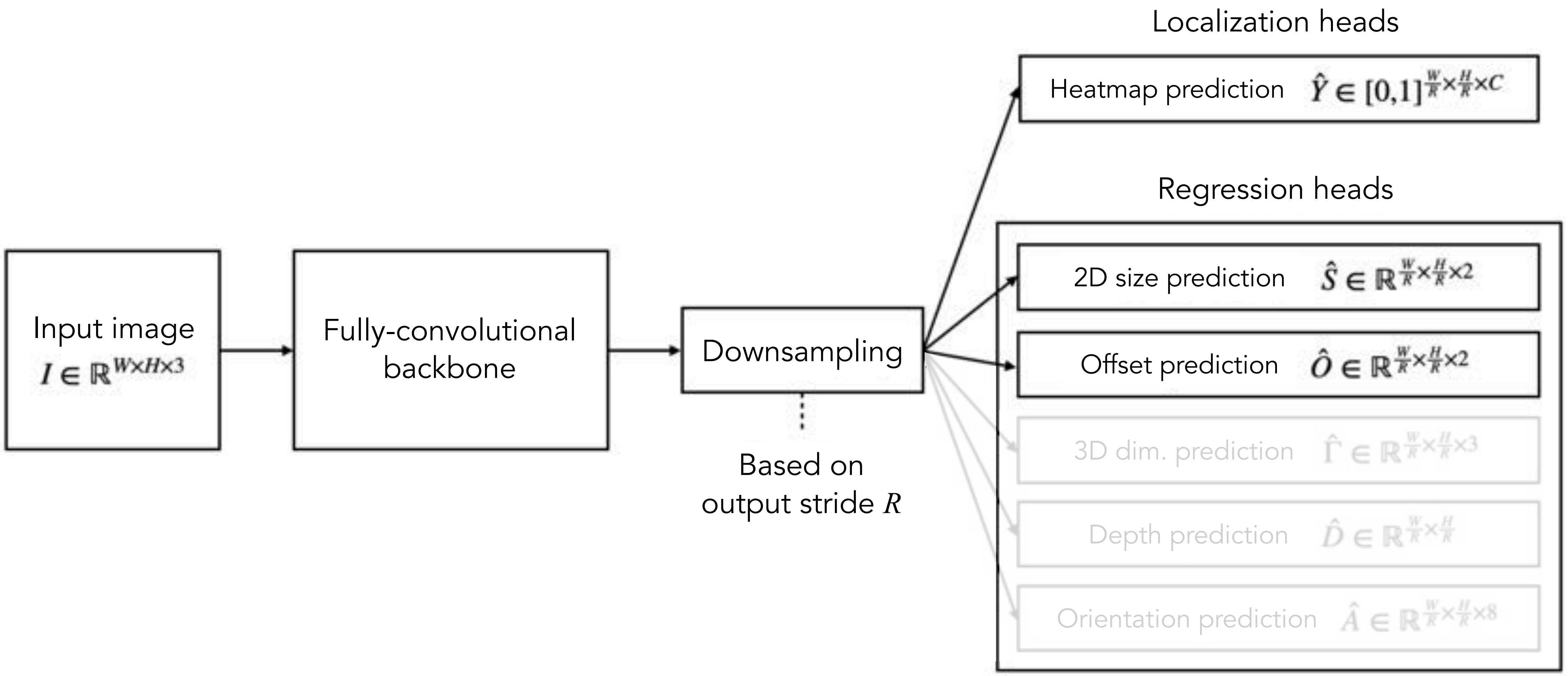
Anchor boxes + refinement + NMS

→

Object center points + parameters

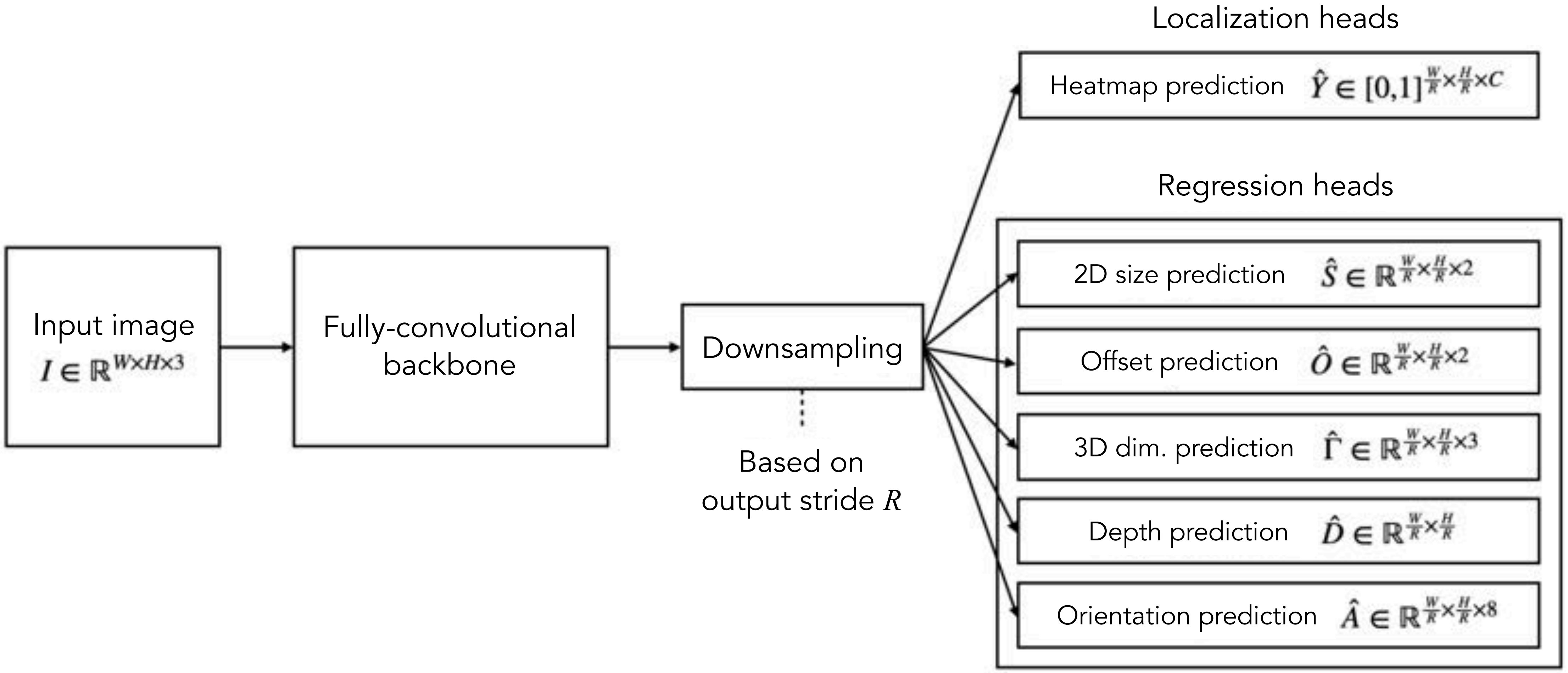
The idea behind

CenterNet: Objects as points



The idea behind

CenterNet: Objects as points



Part 1

Calibrated 3D detection case

Dataset: **KITTI 3D object detection**

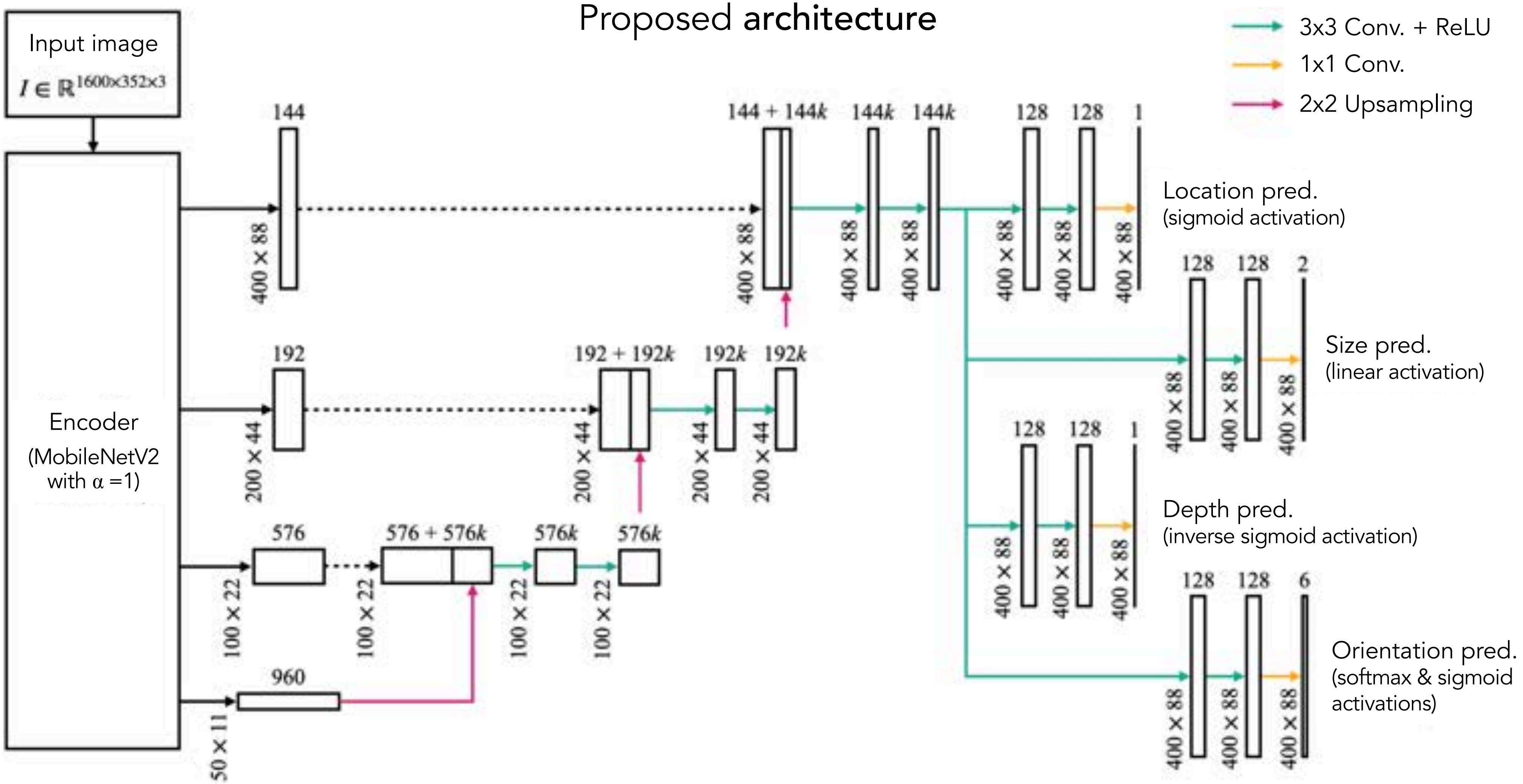
- Car-mounted camera
- RGB images
- Objects divided in difficulty groups
- Only cars considered
- 7,481 training images
  - ↳ 3,682 training
  - ↳ 3,799 validation

Image borrowed from the KITTI vision benchmark suite

Part 1

Calibrated 3D detection case

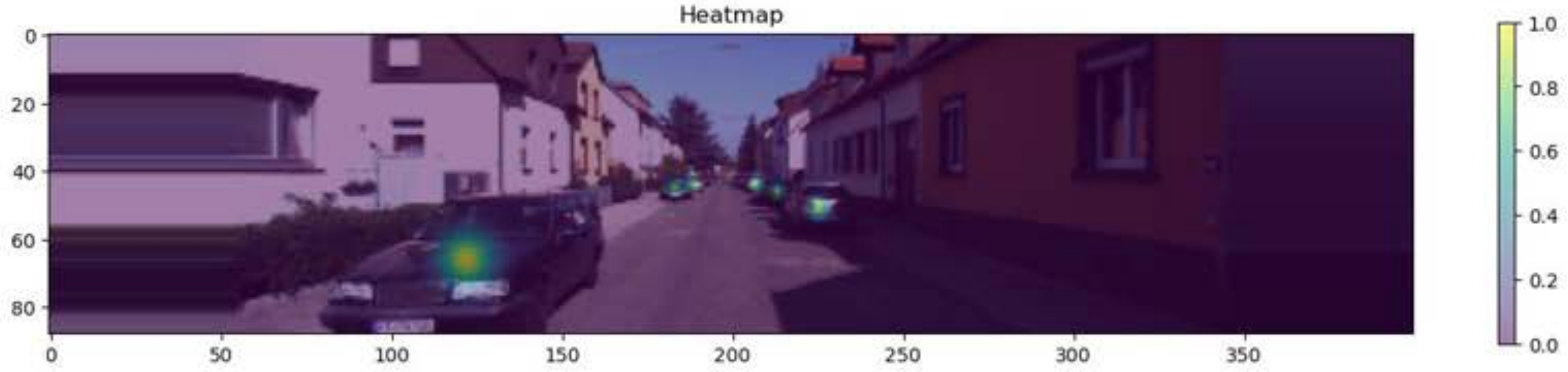
Proposed **architecture**



Part 1

Calibrated 3D detection case

Object center localization prediction (Trying both 2D box centers and projected 3D box centers)

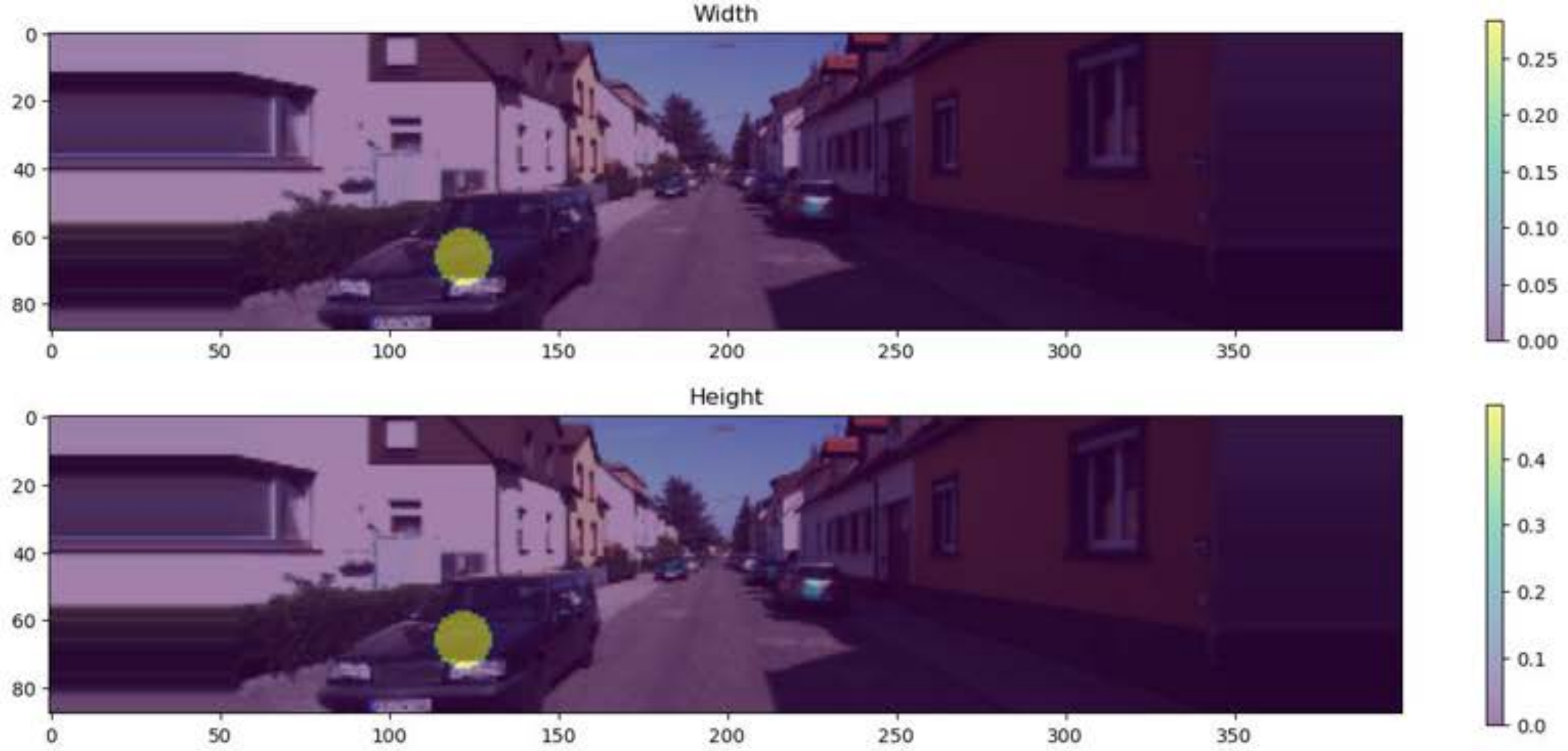


$$L_{loc} = \frac{-1}{N} \sum_{xy} \begin{cases} (1 - \hat{Y}_{xy})^{\alpha} \log(\hat{Y}_{xy}) & \text{if } Y_{xy} = 1 \\ (1 - Y_{xy})^{\beta} (\hat{Y}_{xy})^{\alpha} \log(1 - \hat{Y}_{xy}) & \text{otherwise} \end{cases}$$

Part 1

Calibrated 3D detection case

Size prediction

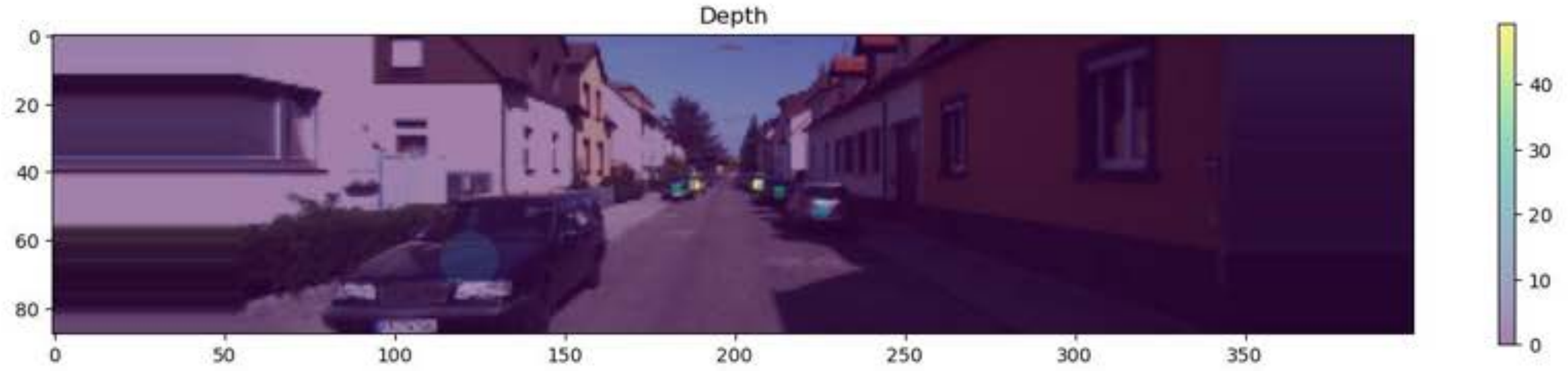


$$L_{size} = \frac{1}{N}\sum_{k=1}^{N}\left|\hat{S}_k - S_k\right|$$

Part 1

Calibrated 3D detection case

Depth prediction



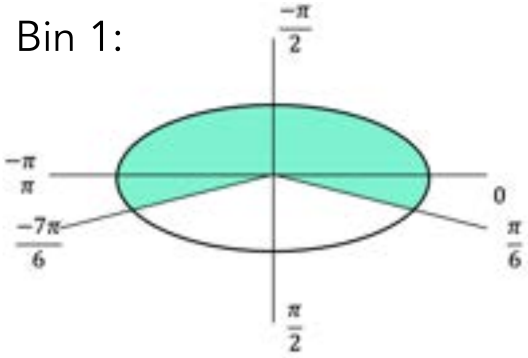
$$L_{dep} = \frac{1}{N}\sum_{k=1}^{N}\left|\frac{1}{\sigma(\hat{d}_k)} - 1 - d_k\right|$$

Part 1

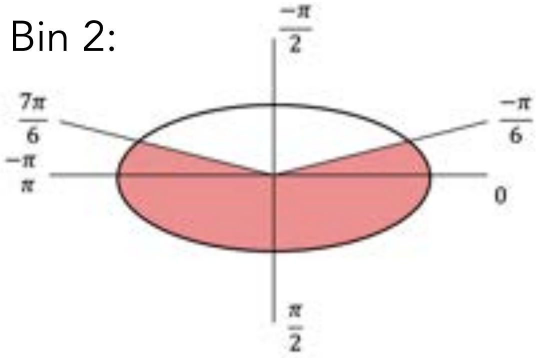
Calibrated 3D detection case

Orientation prediction

Bin 1:



Bin 2:

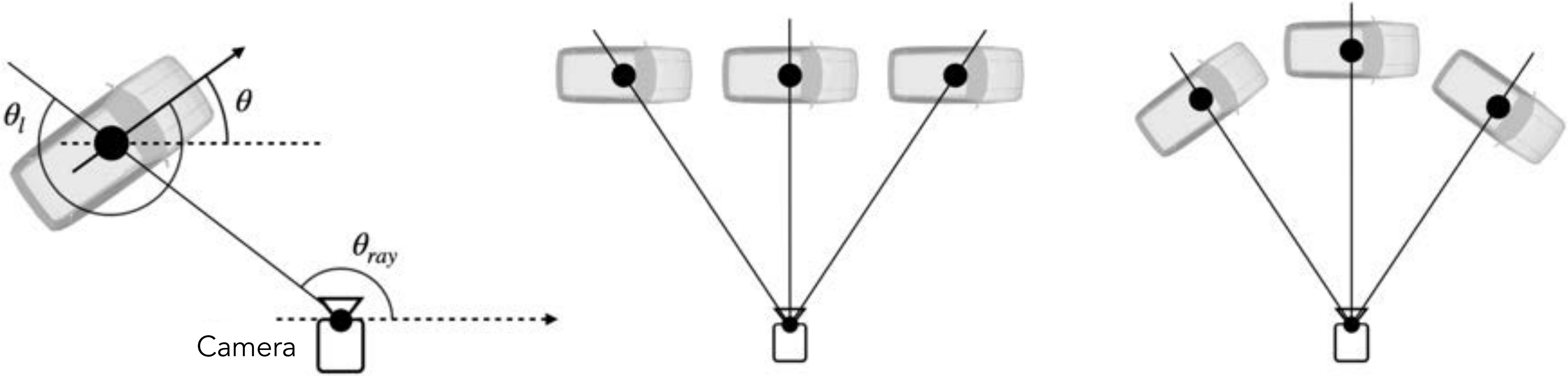


$$L_{ori} = \frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{2}(softmax(\hat{b}_{k_i}, c_{k_i}) + w \times c_{k_i}\ |(sin(\Delta\theta_{lk_i}), cos(\Delta\theta_{lk_i})) - (sin(\theta_{l_k} - m_{k_i}), cos(\theta_{l_k} - m_{k_i}))|)$$

Part 1

Calibrated 3D detection case

Global orientation vs **local orientation**

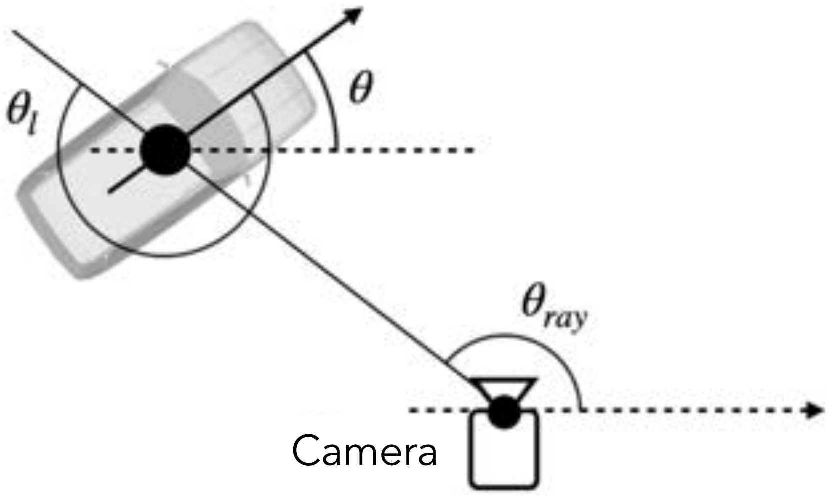


$$\theta = \theta_{ray} + \theta_l$$

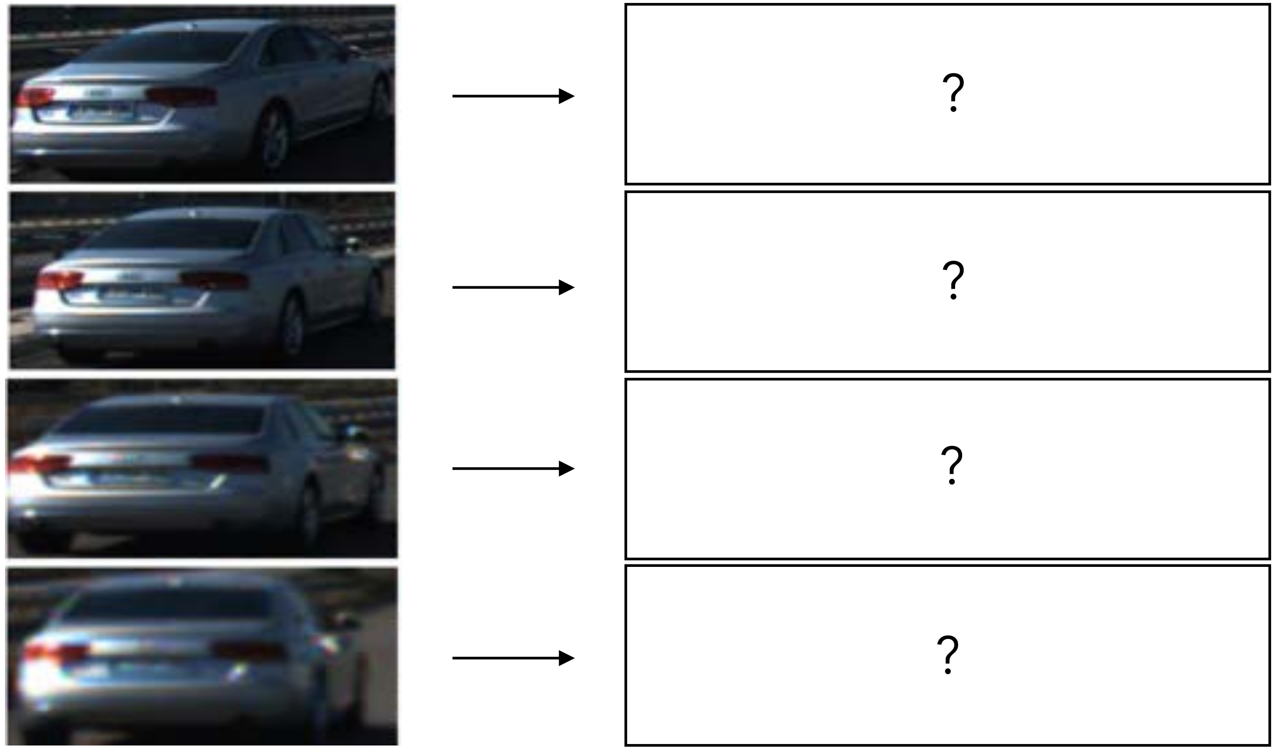
Part 1

Calibrated 3D detection case

Global orientation vs local orientation



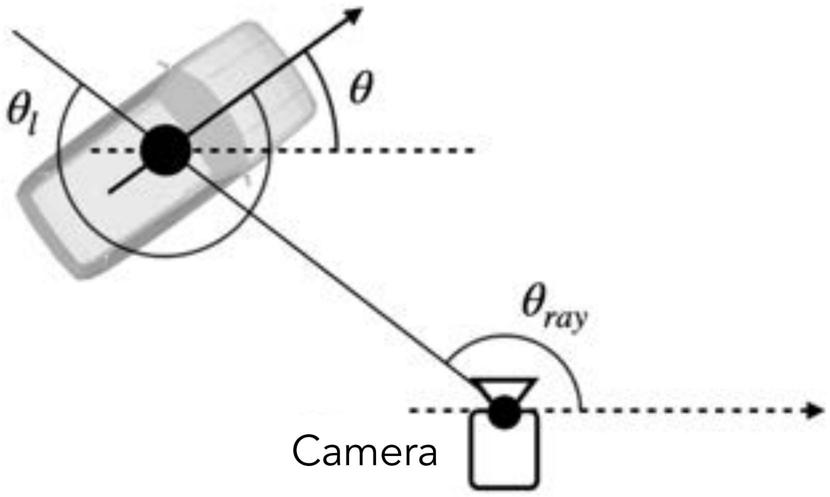
$$\theta = \theta_{ray} + \theta_l$$



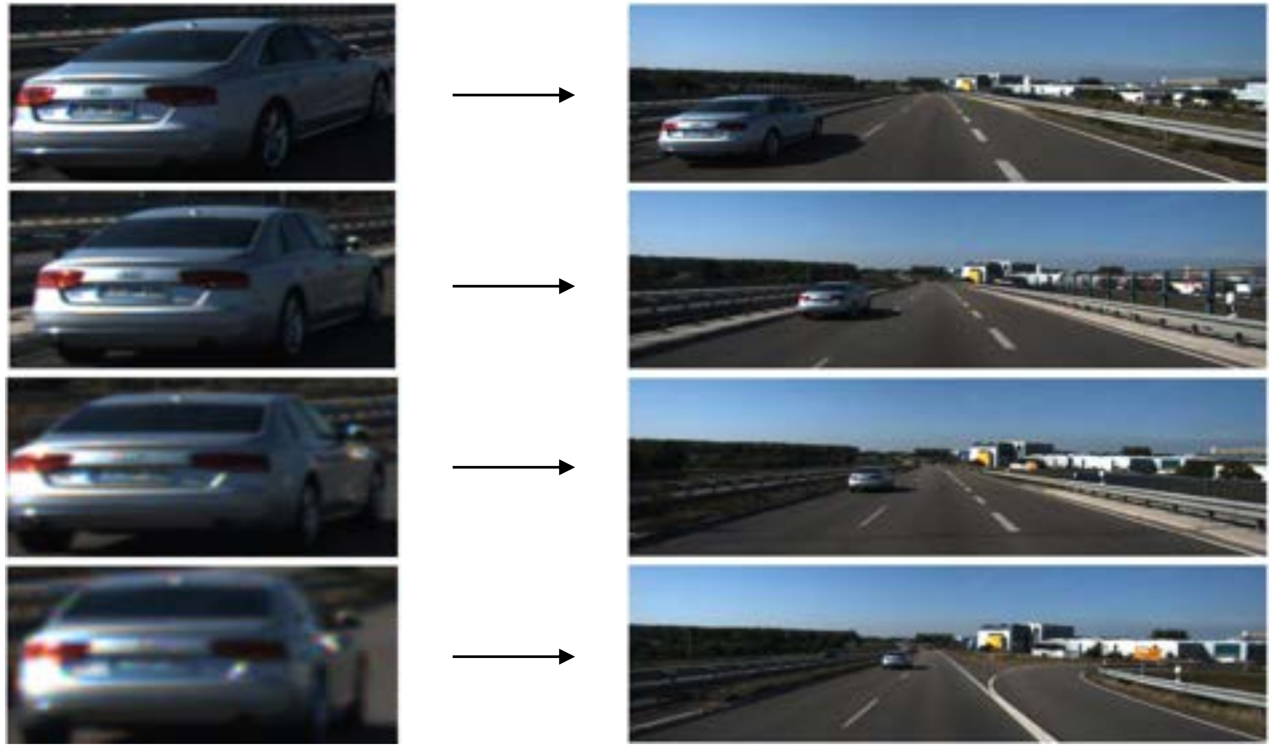
Part 1

Calibrated 3D detection case

Global orientation vs local orientation



$$\theta = \theta_{ray} + \theta_l$$



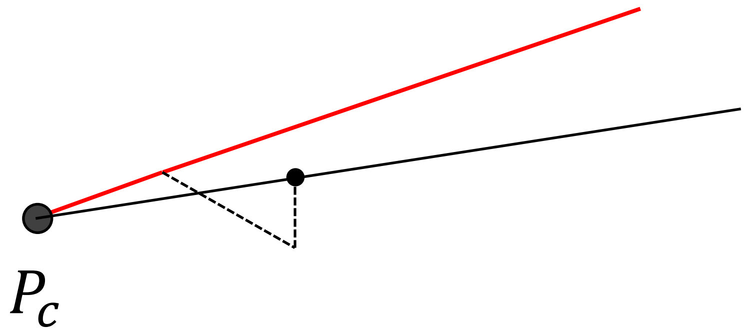
Images borrowed from Mousavian et al. (*)

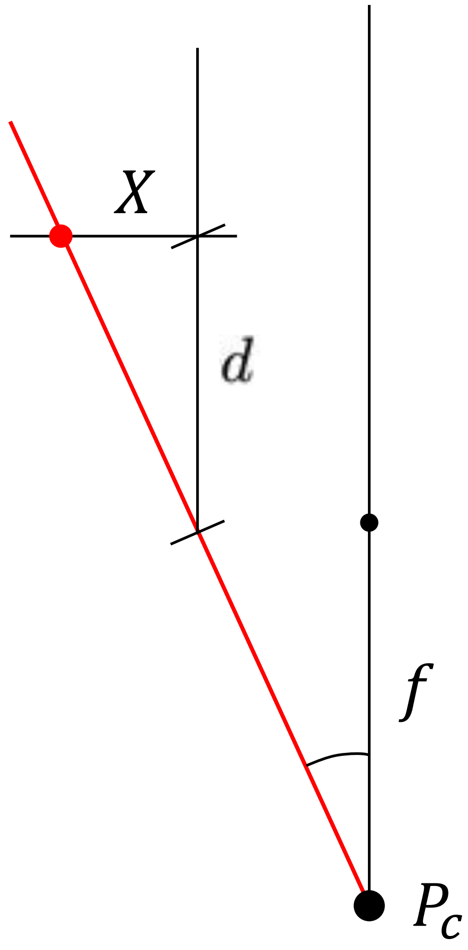
(*) Mousavian, A., Anguelov, D., Flynn, J., & Kosecka, J. (2017). 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 7074-7082).

Part 1

Calibrated 3D detection case

Decoding 3D center location





[...]

Part 1

Calibrated 3D detection case

Experiments

- Different kinds of **encoders** and **layer width multipliers**:

Encoder model		Parameters	Depth	Number of filters			
				Stride 4	Stride 8	Stride 16	Stride 32
MobileNetV2	$\alpha = 0.35$	224,224	100	48	96	192	336
	$\alpha = 0.5$	417,888	100	96	96	288	480
	$\alpha = 1$	1,528,656	100	144	192	576	960
EfficientNetB0		3,144,899	119	144	240	672	1152

- Different kinds of **up-sampling operations**:

  Transp. convolutions vs nearest neighbor interpolation vs bilinear interpolation

- Different kinds of **activations** for hidden layers:

  ReLU vs leaky ReLU

Part 1

Calibrated 3D detection case

Experiments

- Different kinds of **encoders** and **layer width multipliers**:

Encoder model		Parameters	Depth	Number of filters			
				Stride 4	Stride 8	Stride 16	Stride 32
MobileNetV2	$\alpha = 0.35$	224,224	100	48	96	192	336
	$\alpha = 0.5$	417,888	100	96	96	288	480
	$\alpha = 1$	1,528,656	100	144	192	576	960
EfficientNetB0		3,144,899	119	144	240	672	1152

- Different kinds of **up-sampling operations**:

  Transp. convolutions vs **nearest neighbor interpolation** vs **bilinear interpolation**

- Different kinds of **activations** for hidden layers:

  ReLU vs **leaky ReLU**

Part 1

Calibrated 3D detection case

Experiments

- Different kinds of **encoders** and **layer width multipliers**:

Encoder model		Parameters	Depth	Number of filters			
				Stride 4	Stride 8	Stride 16	Stride 32
MobileNetV2	$\alpha = 0.35$	224,224	100	48	96	192	336
	$\alpha = 0.5$	417,888	100	96	96	288	480
	$\alpha = 1$	1,528,656	100	144	192	576	960
EfficientNetB0		3,144,899	119	144	240	672	1152

- Different kinds of **up-sampling operations**:

  Transp. convolutions vs nearest neighbor interpolation vs bilinear interpolation

- Different kinds of **activations** for hidden layers:

  ReLU vs leaky ReLU

Part 1

Calibrated 3D detection case

Experiments

Data augmentation
- Random **horizontal flipping** (50% chance)
- Random **brightness** alterations (80 to 120% of original value)

Training conditions

- Maximum **100 epochs** w/ early stopping set to a patience of 7 epochs
- 1st stage: **Frozen encoder weights**
  Batch size: 8, initialization: Xavier, **learning rate: 1e-3** (w/ *0.25 reductions), **optimizer: Adam**
- 2nd stage: **Unfrozen encoder weights**
  Batch size: 8, initialization: Xavier, **learning rate: 1e-4** (w/ *0.25 reductions), **optimizer: Adam**
- Random searches carried out in special cases

Part 1

Calibrated 3D detection case

Experiments

Data augmentation
- Random **horizontal flipping** (50% chance)
- Random **brightness** alterations (80 to 120% of original value)

Training conditions

- Maximum **100 epochs** w/ early stopping set to a patience of 7 epochs
- 1st stage: **Frozen encoder weights**

  Batch size: 8, initialization: Xavier, **learning rate: 1e-3** (w/ *0.25 reductions), **optimizer: Adam**
- 2nd stage: **Unfrozen encoder weights**

  Batch size: 8, initialization: Xavier, **learning rate: 1e-4** (w/ *0.25 reductions), **optimizer: Adam**
- Random searches carried out in selected case

Part 1

Calibrated 3D detection case

Results

Encoder model			AP			AOS			BEV AP			Parameters	Inference
	α	k	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard		time (ms)
MobileNetV2	0.35	0.5	72.0 / 72.8	52.3 / 52.9	46.3 / 46.3	70.2 / 71.6	51.1 / 51.3	43.5 / 44.5	28.4 / 31.3	19.3 / 23.8	15.7 / 19.8	1,657,450	28.5
		0.75	72.2 / 72.7	51.4 / 51.7	43.8 / 44.2	70.2 / 70.3	49.5 / 49.5	41.9 / 42.1	27.8 / 31.6	18.7 / 22.4	17.1 / 20.6	2,119,630	30.1
		1	68.9 / 72.9	46.2 / 50.4	40.0 / 43.4	66.2 / 69.0	44.1 / 47.3	37.9 / 40.5	23.5 / 28.2	15.9 / 21.2	12.3 / 17.5	2,697,010	31.7
	0.5	0.5	67.9 / 68.9	48.2 / 49.3	42.2 / 42.5	64.3 / 66.5	44.9 / 47.0	40.0 / 40.2	18.5 / 19.5	14.3 / 14.6	11.1 / 11.8	2,804,530	28.6
		0.75	70.6 / 72.4	49.1 / 49.0	41.8 / 41.8	68.3 / 70.1	47.2 / 48.9	39.9 / 39.7	31.5 / 31.4	21.9 / 23.4	20.5 / 19.5	3,736,426	29.5
		1	72.5 / 72.5	50.8 / 51.1	43.6 / 43.5	69.9 / 69.7	48.8 / 48.9	41.6 / 41.3	30.5 / 28.8	21.4 / 21.2	17.5 / 20.5	4,894,114	32.6
	1	0.5	75.6 / 76.3	57.8 / 59.0	44.2 / 46.1	70.1 / 74.0	53.1 / 56.4	42.8 / 44.0	27.4 / 31.4	19.7 / 22.9	19.0 / 21.1	7,816,346	35.1
		0.75	75.1 / 76.3	53.5 / 54.0	45.8 / 46.3	72.3 / 73.4	51.2 / 51.3	43.5 / 43.7	32.4 / 33.3	23.2 / 25.8	21.2 / 21.8	11,493,950	35.8
		1	74.9 / 76.0	54.8 / 54.6	46.4 / 46.7	72.8 / 73.4	53.0 / 52.4	44.7 / 44.6	27.4 / 34.4	19.3 / 23.6	17.6 / 21.6	16,070,114	36.6
EfficientNetB0	-	0.5	83.8 / 87.0	64.5 / 65.8	55.8 / 57.3	80.2 / 85.3	63.1 / 64.0	52.8 / 55.3	30.0 / 33.2	26.5 / 29.7	17.8 / 20.6	11,197,861	42.7
		0.75	81.7 / 82.2	63.0 / 63.2	54.5 / 54.5	79.3 / 79.3	60.5 / 60.2	51.8 / 51.6	23.9 / 28.7	18.1 / 22.7	16.4 / 20.8	16,275,829	43.2
		1	82.3 / 75.1	60.8 / 59.7	53.0 / 52.2	79.9 / 72.7	58.4 / 57.2	50.6 / 49.6	28.9 / 34.0	20.6 / 23.7	18.9 / 21.7	22,597,381	43.5

Part 1

Calibrated 3D detection case

Results

Differences between encoder models

Encoder model	α	k	AP Easy	AP Medium	AP Hard	AOS Easy	AOS Medium	AOS Hard	BEV AP Easy	BEV AP Medium	BEV AP Hard	Parameters	Inference time (ms)
MobileNetV2	0.35	0.5	72.0 / 72.8	52.3 / 52.9	46.3 / 46.3	70.2 / 71.6	51.1 / 51.3	43.5 / 44.5	28.4 / 31.3	19.3 / 23.8	15.7 / 19.8	1,657,450	28.5
		0.75	72.2 / 72.7	51.4 / 51.7	43.8 / 44.2	70.2 / 70.3	49.5 / 49.5	41.9 / 42.1	27.8 / 31.6	18.7 / 22.4	17.1 / 20.6	2,119,630	30.1
		1	68.9 / 72.9	46.2 / 50.4	40.0 / 43.4	66.2 / 69.0	44.1 / 47.3	37.9 / 40.5	23.5 / 28.2	15.9 / 21.2	12.3 / 17.5	2,697,010	31.7
	0.5	0.5	67.9 / 68.9	48.2 / 49.3	42.2 / 42.5	64.3 / 66.5	44.9 / 47.0	40.0 / 40.2	18.5 / 19.5	14.3 / 14.6	11.1 / 11.8	2,804,530	28.6
		0.75	70.6 / 72.4	49.1 / 49.0	41.8 / 41.8	68.3 / 70.1	47.2 / 48.9	39.9 / 39.7	31.5 / 31.4	21.9 / 23.4	20.5 / 19.5	3,736,426	29.5
		1	72.5 / 72.5	50.8 / 51.1	43.6 / 43.5	69.9 / 69.7	48.8 / 48.9	41.6 / 41.3	30.5 / 28.8	21.4 / 21.2	17.5 / 20.5	4,894,114	32.6
	1	0.5	75.6 / 76.3	57.8 / 59.0	44.2 / 46.1	70.1 / 74.0	53.1 / 56.4	42.8 / 44.0	27.4 / 31.4	19.7 / 22.9	19.0 / 21.1	7,816,346	35.1
		0.75	75.1 / 76.3	53.5 / 54.0	45.8 / 46.3	72.3 / 73.4	51.2 / 51.3	43.5 / 43.7	32.4 / 33.3	23.2 / 25.8	21.2 / 21.8	11,493,950	35.8
		1	74.9 / 76.0	54.8 / 54.6	46.4 / 46.7	72.8 / 73.4	53.0 / 52.4	44.7 / 44.6	27.4 / 34.4	19.3 / 23.6	17.6 / 21.6	16,070,114	36.6
EfficientNetB0	-	0.5	83.8 / 87.0	64.5 / 65.8	55.8 / 57.3	80.2 / 85.3	63.1 / 64.0	52.8 / 55.3	30.0 / 33.2	26.5 / 29.7	17.8 / 20.6	11,197,861	42.7
		0.75	81.7 / 82.2	63.0 / 63.2	54.5 / 54.5	79.3 / 79.3	60.5 / 60.2	51.8 / 51.6	23.9 / 28.7	18.1 / 22.7	16.4 / 20.8	16,275,829	43.2
		1	82.3 / 75.1	60.8 / 59.7	53.0 / 52.2	79.9 / 72.7	58.4 / 57.2	50.6 / 49.6	28.9 / 34.0	20.6 / 23.7	18.9 / 21.7	22,597,381	43.5

Part 1

Calibrated 3D detection case

Results

Effects of different filter multipliers k

Encoder model			AP			AOS			BEV AP			Parameters	Inference
	α	k	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard		time (ms)
MobileNetV2	0.35	0.5	72.0 / 72.8	52.3 / 52.9	46.3 / 46.3	70.2 / 71.6	51.1 / 51.3	43.5 / 44.5	28.4 / 31.3	19.3 / 23.8	15.7 / 19.8	1,657,450	28.5
		0.75	72.2 / 72.7	51.4 / 51.7	43.8 / 44.2	70.2 / 70.3	49.5 / 49.5	41.9 / 42.1	27.8 / 31.6	18.7 / 22.4	17.1 / 20.6	2,119,630	30.1
		1	68.9 / 72.9	46.2 / 50.4	40.0 / 43.4	66.2 / 69.0	44.1 / 47.3	37.9 / 40.5	23.5 / 28.2	15.9 / 21.2	12.3 / 17.5	2,697,010	31.7
	0.5	0.5	67.9 / 68.9	48.2 / 49.3	42.2 / 42.5	64.3 / 66.5	44.9 / 47.0	40.0 / 40.2	18.5 / 19.5	14.3 / 14.6	11.1 / 11.8	2,804,530	28.6
		0.75	70.6 / 72.4	49.1 / 49.0	41.8 / 41.8	68.3 / 70.1	47.2 / 48.9	39.9 / 39.7	31.5 / 31.4	21.9 / 23.4	20.5 / 19.5	3,736,426	29.5
		1	72.5 / 72.5	50.8 / 51.1	43.6 / 43.5	69.9 / 69.7	48.8 / 48.9	41.6 / 41.3	30.5 / 28.8	21.4 / 21.2	17.5 / 20.5	4,894,114	32.6
	1	0.5	75.6 / 76.3	57.8 / 59.0	44.2 / 46.1	70.1 / 74.0	53.1 / 56.4	42.8 / 44.0	27.4 / 31.4	19.7 / 22.9	19.0 / 21.1	7,816,346	35.1
		0.75	75.1 / 76.3	53.5 / 54.0	45.8 / 46.3	72.3 / 73.4	51.2 / 51.3	43.5 / 43.7	32.4 / 33.3	23.2 / 25.8	21.2 / 21.8	11,493,950	35.8
		1	74.9 / 76.0	54.8 / 54.6	46.4 / 46.7	72.8 / 73.4	53.0 / 52.4	44.7 / 44.6	27.4 / 34.4	19.3 / 23.6	17.6 / 21.6	16,070,114	36.6
EfficientNetB0	-	0.5	83.8 / 87.0	64.5 / 65.8	55.8 / 57.3	80.2 / 85.3	63.1 / 64.0	52.8 / 55.3	30.0 / 33.2	26.5 / 29.7	17.8 / 20.6	11,197,861	42.7
		0.75	81.7 / 82.2	63.0 / 63.2	54.5 / 54.5	79.3 / 79.3	60.5 / 60.2	51.8 / 51.6	23.9 / 28.7	18.1 / 22.7	16.4 / 20.8	16,275,829	43.2
		1	82.3 / 75.1	60.8 / 59.7	53.0 / 52.2	79.9 / 72.7	58.4 / 57.2	50.6 / 49.6	28.9 / 34.0	20.6 / 23.7	18.9 / 21.7	22,597,381	43.5

Part 1

Calibrated 3D detection case

Results

Effects of different filter multipliers k

Encoder model			AP			AOS			BEV AP			Parameters	Inference
	α	k	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard		time (ms)
MobileNetV2	0.35	0.5	72.0 / 72.8	52.3 / 52.9	46.3 / 46.3	70.2 / 71.6	51.1 / 51.3	43.5 / 44.5	28.4 / 31.3	19.3 / 23.8	15.7 / 19.8	1,657,450	28.5
		0.75	72.2 / 72.7	51.4 / 51.7	43.8 / 44.2	70.2 / 70.3	49.5 / 49.5	41.9 / 42.1	27.8 / 31.6	18.7 / 22.4	17.1 / 20.6	2,119,630	30.1
		1	68.9 / 72.9	46.2 / 50.4	40.0 / 43.4	66.2 / 69.0	44.1 / 47.3	37.9 / 40.5	23.5 / 28.2	15.9 / 21.2	12.3 / 17.5	2,697,010	31.7
	0.5	0.5	67.9 / 68.9	48.2 / 49.3	42.2 / 42.5	64.3 / 66.5	44.9 / 47.0	40.0 / 40.2	18.5 / 19.5	14.3 / 14.6	11.1 / 11.8	2,804,530	28.6
		0.75	70.6 / 72.4	49.1 / 49.0	41.8 / 41.8	68.3 / 70.1	47.2 / 48.9	39.9 / 39.7	31.5 / 31.4	21.9 / 23.4	20.5 / 19.5	3,736,426	29.5
		1	72.5 / 72.5	50.8 / 51.1	43.6 / 43.5	69.9 / 69.7	48.8 / 48.9	41.6 / 41.3	30.5 / 28.8	21.4 / 21.2	17.5 / 20.5	4,894,114	32.6
	1	0.5	75.6 / 76.3	57.8 / 59.0	44.2 / 46.1	70.1 / 74.0	53.1 / 56.4	42.8 / 44.0	27.4 / 31.4	19.7 / 22.9	19.0 / 21.1	7,816,346	35.1
		0.75	75.1 / 76.3	53.5 / 54.0	45.8 / 46.3	72.3 / 73.4	51.2 / 51.3	43.5 / 43.7	32.4 / 33.3	23.2 / 25.8	21.2 / 21.8	11,493,950	35.8
		1	74.9 / 76.0	54.8 / 54.6	46.4 / 46.7	72.8 / 73.4	53.0 / 52.4	44.7 / 44.6	27.4 / 34.4	19.3 / 23.6	17.6 / 21.6	16,070,114	36.6
EfficientNetB0	-	0.5	83.8 / 87.0	64.5 / 65.8	55.8 / 57.3	80.2 / 85.3	63.1 / 64.0	52.8 / 55.3	30.0 / 33.2	26.5 / 29.7	17.8 / 20.6	11,197,861	42.7
		0.75	81.7 / 82.2	63.0 / 63.2	54.5 / 54.5	79.3 / 79.3	60.5 / 60.2	51.8 / 51.6	23.9 / 28.7	18.1 / 22.7	16.4 / 20.8	16,275,829	43.2
		1	82.3 / 75.1	60.8 / 59.7	53.0 / 52.2	79.9 / 72.7	58.4 / 57.2	50.6 / 49.6	28.9 / 34.0	20.6 / 23.7	18.9 / 21.7	22,597,381	43.5

Part 1

Calibrated 3D detection case

Results

Difference in accuracy per difficulty level

Encoder model	α	k	AP Easy	AP Medium	AP Hard	AOS Easy	AOS Medium	AOS Hard	BEV AP Easy	BEV AP Medium	BEV AP Hard	Parameters	Inference time (ms)
MobileNetV2	0.35	0.5	72.0 / 72.8	52.3 / 52.9	46.3 / 46.3	70.2 / 71.6	51.1 / 51.3	43.5 / 44.5	28.4 / 31.3	19.3 / 23.8	15.7 / 19.8	1,657,450	28.5
		0.75	72.2 / 72.7	51.4 / 51.7	43.8 / 44.2	70.2 / 70.3	49.5 / 49.5	41.9 / 42.1	27.8 / 31.6	18.7 / 22.4	17.1 / 20.6	2,119,630	30.1
		1	68.9 / 72.9	46.2 / 50.4	40.0 / 43.4	66.2 / 69.0	44.1 / 47.3	37.9 / 40.5	23.5 / 28.2	15.9 / 21.2	12.3 / 17.5	2,697,010	31.7
	0.5	0.5	67.9 / 68.9	48.2 / 49.3	42.2 / 42.5	64.3 / 66.5	44.9 / 47.0	40.0 / 40.2	18.5 / 19.5	14.3 / 14.6	11.1 / 11.8	2,804,530	28.6
		0.75	70.6 / 72.4	49.1 / 49.0	41.8 / 41.8	68.3 / 70.1	47.2 / 48.9	39.9 / 39.7	31.5 / 31.4	21.9 / 23.4	20.5 / 19.5	3,736,426	29.5
		1	72.5 / 72.5	50.8 / 51.1	43.6 / 43.5	69.9 / 69.7	48.8 / 48.9	41.6 / 41.3	30.5 / 28.8	21.4 / 21.2	17.5 / 20.5	4,894,114	32.6
	1	0.5	75.6 / 76.3	57.8 / 59.0	44.2 / 46.1	70.1 / 74.0	53.1 / 56.4	42.8 / 44.0	27.4 / 31.4	19.7 / 22.9	19.0 / 21.1	7,816,346	35.1
		0.75	75.1 / 76.3	53.5 / 54.0	45.8 / 46.3	72.3 / 73.4	51.2 / 51.3	43.5 / 43.7	32.4 / 33.3	23.2 / 25.8	21.2 / 21.8	11,493,950	35.8
		1	74.9 / 76.0	54.8 / 54.6	46.4 / 46.7	72.8 / 73.4	53.0 / 52.4	44.7 / 44.6	27.4 / 34.4	19.3 / 23.6	17.6 / 21.6	16,070,114	36.6
EfficientNetB0	-	0.5	83.8 / 87.0	64.5 / 65.8	55.8 / 57.3	80.2 / 85.3	63.1 / 64.0	52.8 / 55.3	30.0 / 33.2	26.5 / 29.7	17.8 / 20.6	11,197,861	42.7
		0.75	81.7 / 82.2	63.0 / 63.2	54.5 / 54.5	79.3 / 79.3	60.5 / 60.2	51.8 / 51.6	23.9 / 28.7	18.1 / 22.7	16.4 / 20.8	16,275,829	43.2
		1	82.3 / 75.1	60.8 / 59.7	53.0 / 52.2	79.9 / 72.7	58.4 / 57.2	50.6 / 49.6	28.9 / 34.0	20.6 / 23.7	18.9 / 21.7	22,597,381	43.5

Part 1

Calibrated 3D detection case

Results

Difference in accuracy per difficulty level

Encoder model			AP			AOS			BEV AP			Parameters	Inference
	α	k	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard		time (ms)
MobileNetV2	0.35	0.5	72.0 / 72.8	52.3 / 52.9	46.3 / 46.3	70.2 / 71.6	51.1 / 51.3	43.5 / 44.5	28.4 / 31.3	19.3 / 23.8	15.7 / 19.8	1,657,450	28.5
		0.75	72.2 / 72.7	51.4 / 51.7	43.8 / 44.2	70.2 / 70.3	49.5 / 49.5	41.9 / 42.1	27.8 / 31.6	18.7 / 22.4	17.1 / 20.6	2,119,630	30.1
		1	68.9 / 72.9	46.2 / 50.4	40.0 / 43.4	66.2 / 69.0	44.1 / 47.3	37.9 / 40.5	23.5 / 28.2	15.9 / 21.2	12.3 / 17.5	2,697,010	31.7
	0.5	0.5	67.9 / 68.9	48.2 / 49.3	42.2 / 42.5	64.3 / 66.5	44.9 / 47.0	40.0 / 40.2	18.5 / 19.5	14.3 / 14.6	11.1 / 11.8	2,804,530	28.6
		0.75	70.6 / 72.4	49.1 / 49.0	41.8 / 41.8	68.3 / 70.1	47.2 / 48.9	39.9 / 39.7	31.5 / 31.4	21.9 / 23.4	20.5 / 19.5	3,736,426	29.5
		1	72.5 / 72.5	50.8 / 51.1	43.6 / 43.5	69.9 / 69.7	48.8 / 48.9	41.6 / 41.3	30.5 / 28.8	21.4 / 21.2	17.5 / 20.5	4,894,114	32.6
	1	0.5	75.6 / 76.3	57.8 / 59.0	44.2 / 46.1	70.1 / 74.0	53.1 / 56.4	42.8 / 44.0	27.4 / 31.4	19.7 / 22.9	19.0 / 21.1	7,816,346	35.1
		0.75	75.1 / 76.3	53.5 / 54.0	45.8 / 46.3	72.3 / 73.4	51.2 / 51.3	43.5 / 43.7	32.4 / 33.3	23.2 / 25.8	21.2 / 21.8	11,493,950	35.8
		1	74.9 / 76.0	54.8 / 54.6	46.4 / 46.7	72.8 / 73.4	53.0 / 52.4	44.7 / 44.6	27.4 / 34.4	19.3 / 23.6	17.6 / 21.6	16,070,114	36.6
EfficientNetB0	-	0.5	83.8 / 87.0	64.5 / 65.8	55.8 / 57.3	80.2 / 85.3	63.1 / 64.0	52.8 / 55.3	30.0 / 33.2	26.5 / 29.7	17.8 / 20.6	11,197,861	42.7
		0.75	81.7 / 82.2	63.0 / 63.2	54.5 / 54.5	79.3 / 79.3	60.5 / 60.2	51.8 / 51.6	23.9 / 28.7	18.1 / 22.7	16.4 / 20.8	16,275,829	43.2
		1	82.3 / 75.1	60.8 / 59.7	53.0 / 52.2	79.9 / 72.7	58.4 / 57.2	50.6 / 49.6	28.9 / 34.0	20.6 / 23.7	18.9 / 21.7	22,597,381	43.5

Part 1

Calibrated 3D detection case

Results

Difference in accuracy per difficulty level

Encoder model	α	k	AP Easy	AP Medium	AP Hard	AOS Easy	AOS Medium	AOS Hard	BEV AP Easy	BEV AP Medium	BEV AP Hard	Parameters	Inference time (ms)
MobileNetV2	0.35	0.5	72.0 / 72.8	52.3 / 52.9	46.3 / 46.3	70.2 / 71.6	51.1 / 51.3	43.5 / 44.5	28.4 / 31.3	19.3 / 23.8	15.7 / 19.8	1,657,450	28.5
		0.75	72.2 / 72.7	51.4 / 51.7	43.8 / 44.2	70.2 / 70.3	49.5 / 49.5	41.9 / 42.1	27.8 / 31.6	18.7 / 22.4	17.1 / 20.6	2,119,630	30.1
		1	68.9 / 72.9	46.2 / 50.4	40.0 / 43.4	66.2 / 69.0	44.1 / 47.3	37.9 / 40.5	23.5 / 28.2	15.9 / 21.2	12.3 / 17.5	2,697,010	31.7
	0.5	0.5	67.9 / 68.9	48.2 / 49.3	42.2 / 42.5	64.3 / 66.5	44.9 / 47.0	40.0 / 40.2	18.5 / 19.5	14.3 / 14.6	11.1 / 11.8	2,804,530	28.6
		0.75	70.6 / 72.4	49.1 / 49.0	41.8 / 41.8	68.3 / 70.1	47.2 / 48.9	39.9 / 39.7	31.5 / 31.4	21.9 / 23.4	20.5 / 19.5	3,736,426	29.5
		1	72.5 / 72.5	50.8 / 51.1	43.6 / 43.5	69.9 / 69.7	48.8 / 48.9	41.6 / 41.3	30.5 / 28.8	21.4 / 21.2	17.5 / 20.5	4,894,114	32.6
	1	0.5	75.6 / 76.3	57.8 / 59.0	44.2 / 46.1	70.1 / 74.0	53.1 / 56.4	42.8 / 44.0	27.4 / 31.4	19.7 / 22.9	19.0 / 21.1	7,816,346	35.1
		0.75	75.1 / 76.3	53.5 / 54.0	45.8 / 46.3	72.3 / 73.4	51.2 / 51.3	43.5 / 43.7	32.4 / 33.3	23.2 / 25.8	21.2 / 21.8	11,493,950	35.8
		1	74.9 / 76.0	54.8 / 54.6	46.4 / 46.7	72.8 / 73.4	53.0 / 52.4	44.7 / 44.6	27.4 / 34.4	19.3 / 23.6	17.6 / 21.6	16,070,114	36.6
EfficientNetB0	-	0.5	83.8 / 87.0	64.5 / 65.8	55.8 / 57.3	80.2 / 85.3	63.1 / 64.0	52.8 / 55.3	30.0 / 33.2	26.5 / 29.7	17.8 / 20.6	11,197,861	42.7
		0.75	81.7 / 82.2	63.0 / 63.2	54.5 / 54.5	79.3 / 79.3	60.5 / 60.2	51.8 / 51.6	23.9 / 28.7	18.1 / 22.7	16.4 / 20.8	16,275,829	43.2
		1	82.3 / 75.1	60.8 / 59.7	53.0 / 52.2	79.9 / 72.7	58.4 / 57.2	50.6 / 49.6	28.9 / 34.0	20.6 / 23.7	18.9 / 21.7	22,597,381	43.5

Part 1

Calibrated 3D detection case

Results

The effects of data augmentation

Encoder model			AP			AOS			BEV AP			Parameters	Inference
	α	k	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard		time (ms)
MobileNetV2	0.35	0.5	72.0 / 72.8	52.3 / 52.9	46.3 / 46.3	70.2 / 71.6	51.1 / 51.3	43.5 / 44.5	28.4 / 31.3	19.3 / 23.8	15.7 / 19.8	1,657,450	28.5
		0.75	72.2 / 72.7	51.4 / 51.7	43.8 / 44.2	70.2 / 70.3	49.5 / 49.5	41.9 / 42.1	27.8 / 31.6	18.7 / 22.4	17.1 / 20.6	2,119,630	30.1
		1	68.9 / 72.9	46.2 / 50.4	40.0 / 43.4	66.2 / 69.0	44.1 / 47.3	37.9 / 40.5	23.5 / 28.2	15.9 / 21.2	12.3 / 17.5	2,697,010	31.7
	0.5	0.5	67.9 / 68.9	48.2 / 49.3	42.2 / 42.5	64.3 / 66.5	44.9 / 47.0	40.0 / 40.2	18.5 / 19.5	14.3 / 14.6	11.1 / 11.8	2,804,530	28.6
		0.75	70.6 / 72.4	49.1 / 49.0	41.8 / 41.8	68.3 / 70.1	47.2 / 48.9	39.9 / 39.7	31.5 / 31.4	21.9 / 23.4	20.5 / 19.5	3,736,426	29.5
		1	72.5 / 72.5	50.8 / 51.1	43.6 / 43.5	69.9 / 69.7	48.8 / 48.9	41.6 / 41.3	30.5 / 28.8	21.4 / 21.2	17.5 / 20.5	4,894,114	32.6
	1	0.5	75.6 / 76.3	57.8 / 59.0	44.2 / 46.1	70.1 / 74.0	53.1 / 56.4	42.8 / 44.0	27.4 / 31.4	19.7 / 22.9	19.0 / 21.1	7,816,346	35.1
		0.75	75.1 / 76.3	53.5 / 54.0	45.8 / 46.3	72.3 / 73.4	51.2 / 51.3	43.5 / 43.7	32.4 / 33.3	23.2 / 25.8	21.2 / 21.8	11,493,950	35.8
		1	74.9 / 76.0	54.8 / 54.6	46.4 / 46.7	72.8 / 73.4	53.0 / 52.4	44.7 / 44.6	27.4 / 34.4	19.3 / 23.6	17.6 / 21.6	16,070,114	36.6
EfficientNetB0	-	0.5	83.8 / 87.0	64.5 / 65.8	55.8 / 57.3	80.2 / 85.3	63.1 / 64.0	52.8 / 55.3	30.0 / 33.2	26.5 / 29.7	17.8 / 20.6	11,197,861	42.7
		0.75	81.7 / 82.2	63.0 / 63.2	54.5 / 54.5	79.3 / 79.3	60.5 / 60.2	51.8 / 51.6	23.9 / 28.7	18.1 / 22.7	16.4 / 20.8	16,275,829	43.2
		1	82.3 / 75.1	60.8 / 59.7	53.0 / 52.2	79.9 / 72.7	58.4 / 57.2	50.6 / 49.6	28.9 / 34.0	20.6 / 23.7	18.9 / 21.7	22,597,381	43.5

Part 1

Calibrated 3D detection case

Results

Settling for the middle ground

Encoder model	α	k	AP Easy	AP Medium	AP Hard	AOS Easy	AOS Medium	AOS Hard	BEV AP Easy	BEV AP Medium	BEV AP Hard	Parameters	Inference time (ms)
MobileNetV2	0.35	0.5	72.0 / 72.8	52.3 / 52.9	46.3 / 46.3	70.2 / 71.6	51.1 / 51.3	43.5 / 44.5	28.4 / 31.3	19.3 / 23.8	15.7 / 19.8	1,657,450	28.5
		0.75	72.2 / 72.7	51.4 / 51.7	43.8 / 44.2	70.2 / 70.3	49.5 / 49.5	41.9 / 42.1	27.8 / 31.6	18.7 / 22.4	17.1 / 20.6	2,119,630	30.1
		1	68.9 / 72.9	46.2 / 50.4	40.0 / 43.4	66.2 / 69.0	44.1 / 47.3	37.9 / 40.5	23.5 / 28.2	15.9 / 21.2	12.3 / 17.5	2,697,010	31.7
	0.5	0.5	67.9 / 68.9	48.2 / 49.3	42.2 / 42.5	64.3 / 66.5	44.9 / 47.0	40.0 / 40.2	18.5 / 19.5	14.3 / 14.6	11.1 / 11.8	2,804,530	28.6
		0.75	70.6 / 72.4	49.1 / 49.0	41.8 / 41.8	68.3 / 70.1	47.2 / 48.9	39.9 / 39.7	31.5 / 31.4	21.9 / 23.4	20.5 / 19.5	3,736,426	29.5
		1	72.5 / 72.5	50.8 / 51.1	43.6 / 43.5	69.9 / 69.7	48.8 / 48.9	41.6 / 41.3	30.5 / 28.8	21.4 / 21.2	17.5 / 20.5	4,894,114	32.6
	1	0.5	75.6 / 76.3	57.8 / 59.0	44.2 / 46.1	70.1 / 74.0	53.1 / 56.4	42.8 / 44.0	27.4 / 31.4	19.7 / 22.9	19.0 / 21.1	7,816,346	35.1
		0.75	75.1 / 76.3	53.5 / 54.0	45.8 / 46.3	72.3 / 73.4	51.2 / 51.3	43.5 / 43.7	32.4 / 33.3	23.2 / 25.8	21.2 / 21.8	11,493,950	35.8
		1	74.9 / 76.0	54.8 / 54.6	46.4 / 46.7	72.8 / 73.4	53.0 / 52.4	44.7 / 44.6	27.4 / 34.4	19.3 / 23.6	17.6 / 21.6	16,070,114	36.6
EfficientNetB0	-	0.5	83.8 / 87.0	64.5 / 65.8	55.8 / 57.3	80.2 / 85.3	63.1 / 64.0	52.8 / 55.3	30.0 / 33.2	26.5 / 29.7	17.8 / 20.6	11,197,861	42.7
		0.75	81.7 / 82.2	63.0 / 63.2	54.5 / 54.5	79.3 / 79.3	60.5 / 60.2	51.8 / 51.6	23.9 / 28.7	18.1 / 22.7	16.4 / 20.8	16,275,829	43.2
		1	82.3 / 75.1	60.8 / 59.7	53.0 / 52.2	79.9 / 72.7	58.4 / 57.2	50.6 / 49.6	28.9 / 34.0	20.6 / 23.7	18.9 / 21.7	22,597,381	43.5

Part 1

Calibrated 3D detection case

Results

Differences between up-sampling methods

Up-sampling method	AP			AOS			BEV AP		
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
Transposed convolutions	72.4	49.0	41.8	70.1	48.9	39.7	31.4	23.4	19.5
Nearest neighbor interp.	72.6	52.5	44.8	68.6	49.0	41.4	32.1	21.8	17.8
Bilinear interpolation	75.8	54.8	46.6	73.7	52.4	44.3	25.8	18.2	14.3

Differences between activation functions

Hidden layer activations	AP			AOS			BEV AP		
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
Standard ReLU	72.4	49.0	41.8	70.1	48.9	39.7	31.4	23.4	19.5
Leaky ReLU	72.8	50.7	43.6	69.4	47.6	40.7	30.4	20.9	17.2

Part 1

Calibrated 3D detection case

Results

Differences between up-sampling methods

Up-sampling method	AP			AOS			BEV AP		
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
Transposed convolutions	72.4	49.0	41.8	70.1	48.9	39.7	31.4	23.4	19.5
Nearest neighbor interp.	72.6	52.5	44.8	68.6	49.0	41.4	32.1	21.8	17.8
Bilinear interpolation	75.8	54.8	46.6	73.7	52.4	44.3	25.8	18.2	14.3

Differences between activation functions in hidden layers

Hidden layer activations	**AP**			**AOS**			**BEV AP**		
	Easy	**Medium**	**Hard**	**Easy**	**Medium**	**Hard**	**Easy**	**Medium**	**Hard**
Standard ReLU	72.4	49.0	41.8	70.1	48.9	39.7	31.4	23.4	19.5
Leaky ReLU	72.8	50.7	43.6	69.4	47.6	40.7	30.4	20.9	17.2

Part 1

Calibrated 3D detection case

Results

Comparative between systems

Detection system	AP			AOS			BEV AP			Parameters
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard	
Original CenterNet	90.2±1.2	80.4±1.4	71.1±1.6	85.3±1.7	75.0±1.6	66.2±1.8	31.4±3.7	26.5±1.6	23.8±2.9	20,615,323
Ours (*)	72.8±0.8	50.6±1.1	42.3±1.8	71.4±1.4	50.1±1.8	41.4±1.8	31.4±2.2	23.6±2.0	19.8±2.3	3,736,426
Ours (**)	32.1±1.6	20.2±1.8	18.9±1.8	31.17±1.6	20.2±1.8	18.9±1.9	31.6±2.5	23.5±2.2	19.5±2.6	3,476,832

(*) Standard version: Prediction of 2D centers and size, (**) Alternative version: Direct prediction of projected 3D centers

Part 1

Calibrated 3D detection case

Results

Comparative between systems

Detection system	AP			AOS			BEV AP			Parameters
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard	
Original CenterNet	90.2±1.2	80.4±1.4	71.1±1.6	85.3±1.7	75.0±1.6	66.2±1.8	31.4±3.7	26.5±1.6	23.8±2.9	20,615,323
Ours (*)	72.8±0.8	50.6±1.1	42.3±1.8	71.4±1.4	50.1±1.8	41.4±1.8	31.4±2.2	23.6±2.0	19.8±2.3	3,736,426
Ours (**)	32.1±1.6	20.2±1.8	18.9±1.8	31.17±1.6	20.2±1.8	18.9±1.9	31.6±2.5	23.5±2.2	19.5±2.6	3,476,832

(*) Standard version: Prediction of 2D centers and size, (**) Alternative version: Direct prediction of projected 3D centers

OS		
Easy	Medium	Hard
~0.94	~0.93	~0.93
~0.98	~0.99	~0.97
~0.97	~1.00	~1.00

Part 1

Calibrated 3D detection case

Results

Comparative between systems

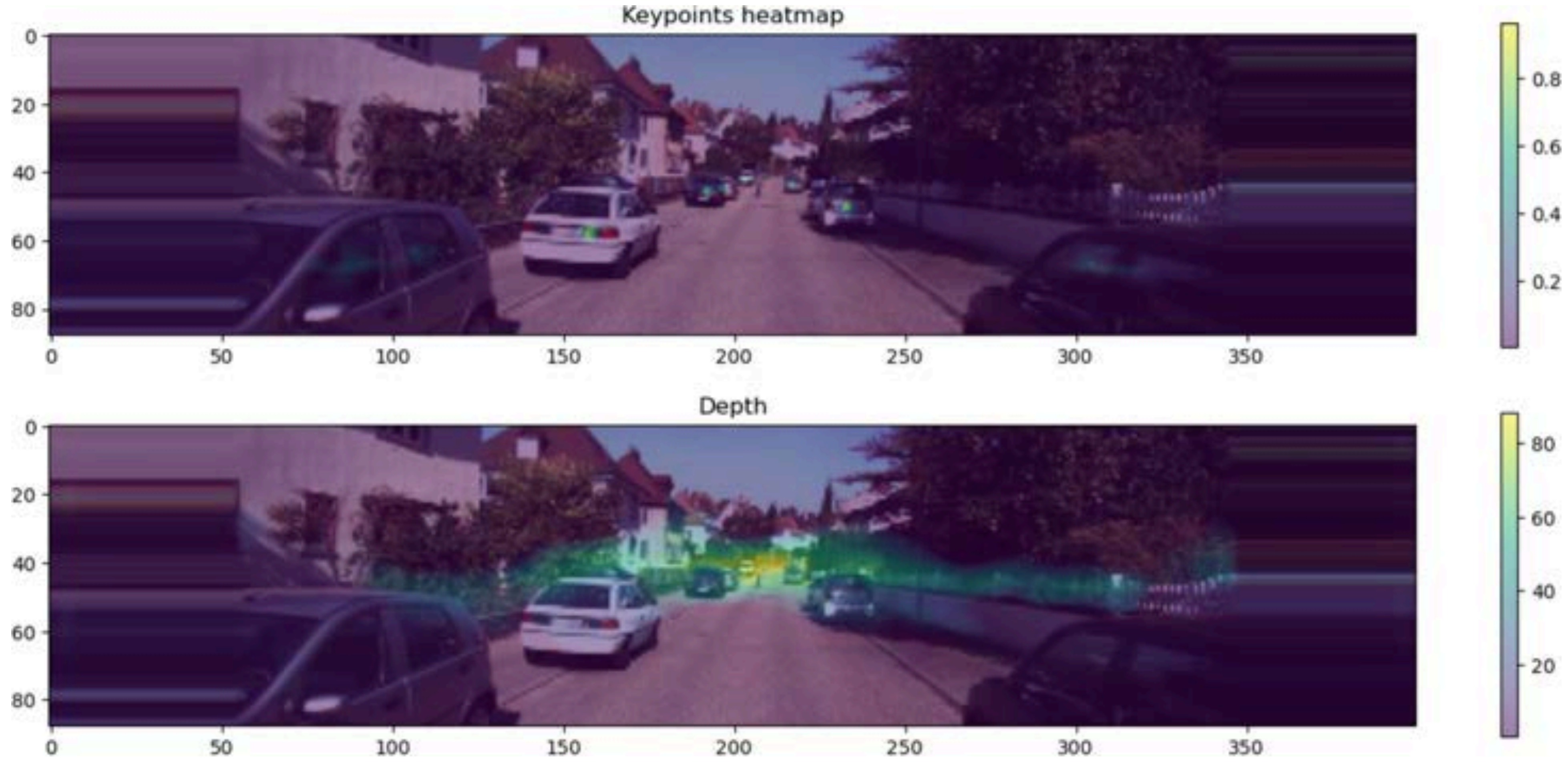
Detection system	AP			AOS			BEV AP			Parameters
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard	
Original CenterNet	90.2±1.2	80.4±1.4	71.1±1.6	85.3±1.7	75.0±1.6	66.2±1.8	31.4±3.7	26.5±1.6	23.8±2.9	20,615,323
Ours (*)	72.8±0.8	50.6±1.1	42.3±1.8	71.4±1.4	50.1±1.8	41.4±1.8	31.4±2.2	23.6±2.0	19.8±2.3	3,736,426
Ours (**)	32.1±1.6	20.2±1.8	18.9±1.8	31.17±1.6	20.2±1.8	18.9±1.9	31.6±2.5	23.5±2.2	19.5±2.6	3,476,832

(*) Standard version: Prediction of 2D centers and size, (**) Alternative version: Direct prediction of projected 3D centers

Part 1

Calibrated 3D detection case

Visual examples



Part 1

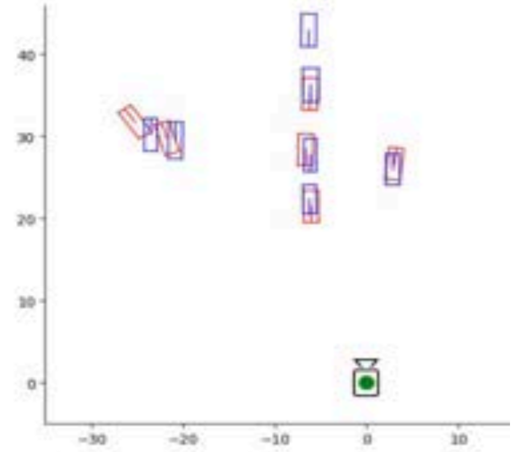
Calibrated 3D detection case

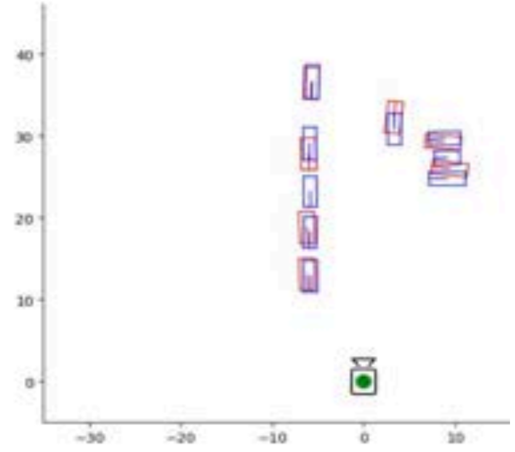
Visual examples

Part 1

Calibrated 3D detection case

Visual examples

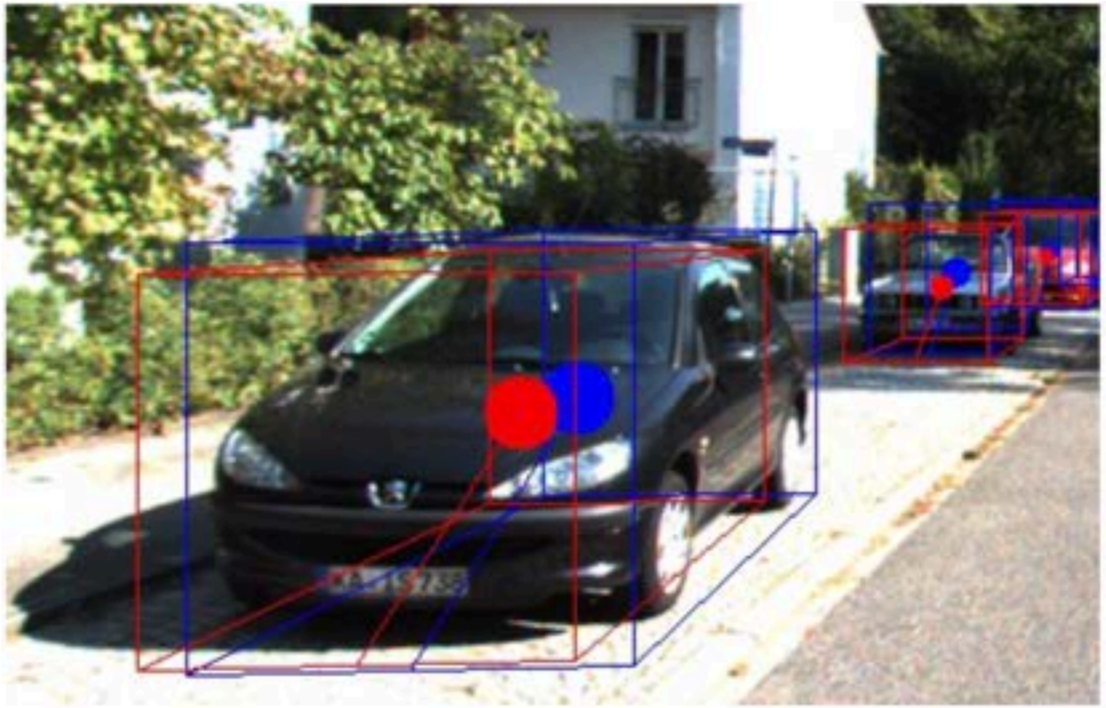




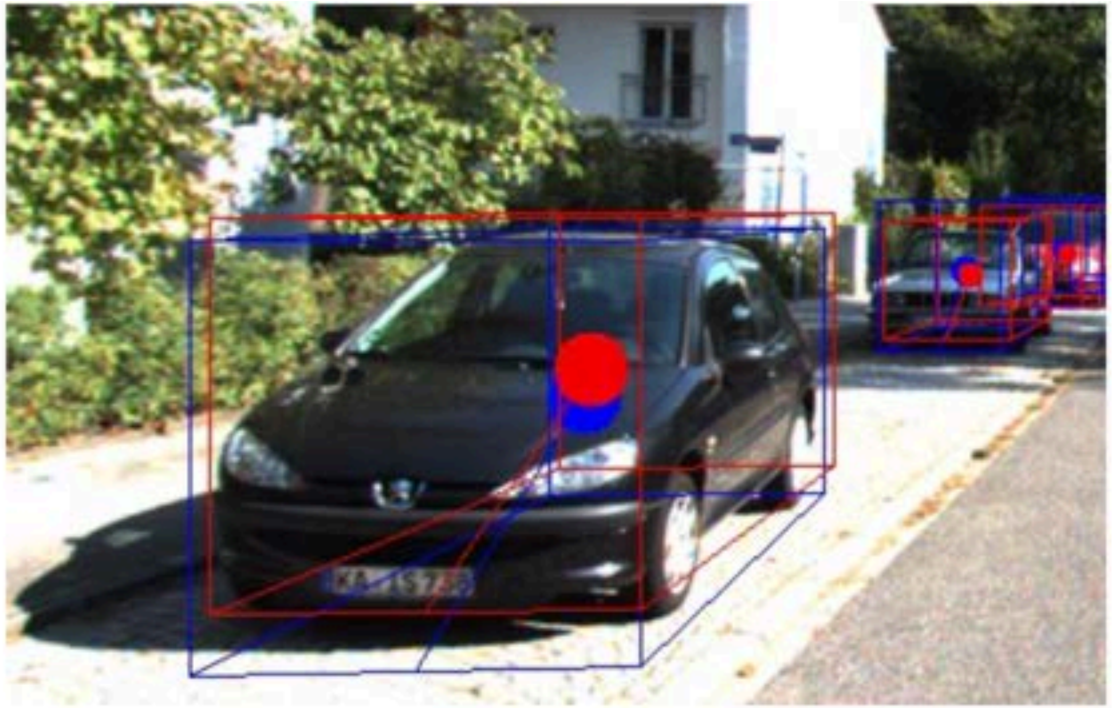
Part 1

Calibrated 3D detection case

Visual examples



Standard version: Prediction of 2D centers



Alternative version: Direct prediction of projected 3D centers

Part 2

Uncalibrated 3D detection case

Dataset: Ko-PER intersection laserscanner and video dataset

- 2 traffic monitoring cams
- Greyscale images
- Only cars considered
- 4-part sequence, totalling 4,831 images per camera
  - ↳ 3,626 training
  - ↳ 1,205 validation

Images borrowed from the Ko-PER intersection laserscanner and video dataset

A lack of calibration is assumed

Part 2

Uncalibrated 3D detection case

Simplifying the problem

4+ point correspondences
between camera images and map

↓

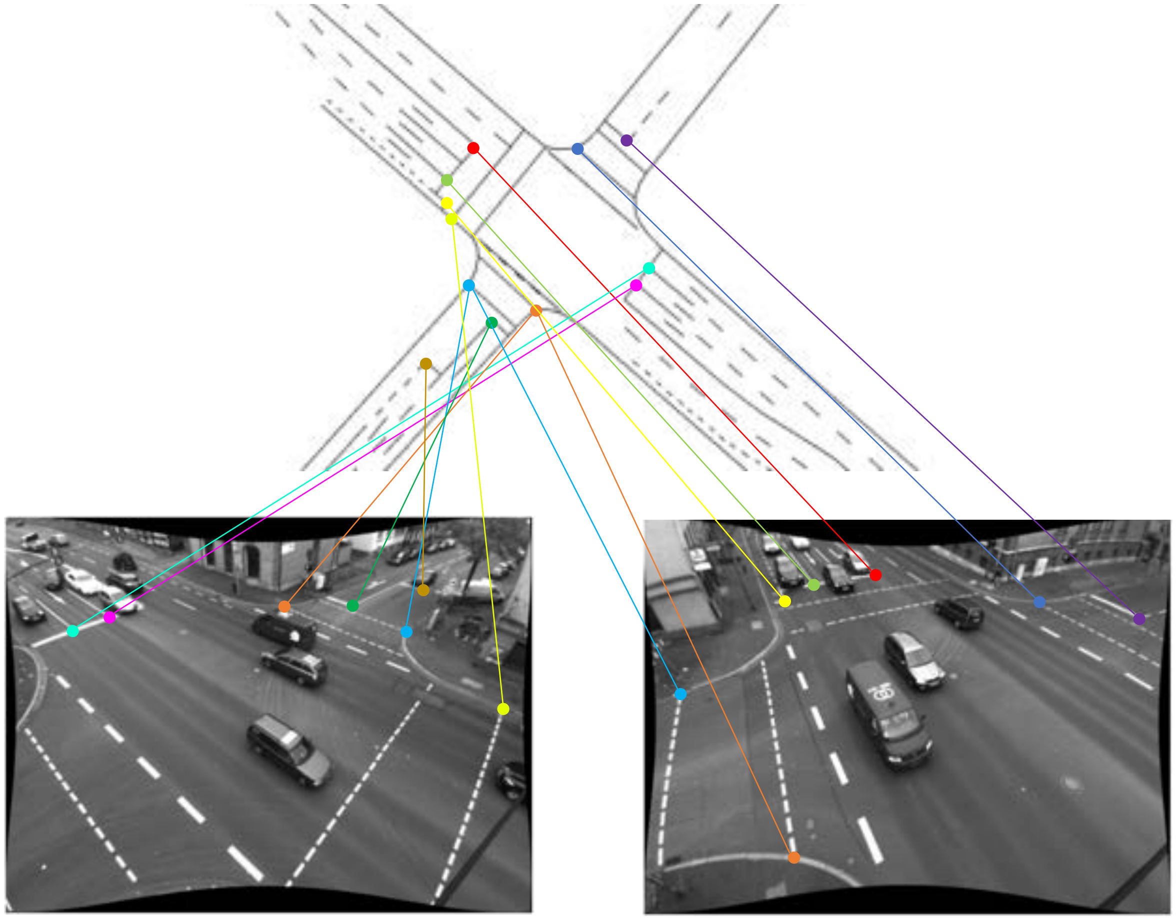
Calculation of homography, incl.
the use of RANSAC

↓

Planar transformation of camera
images and stitching into BEV

↓

Treatment of 3D problem as 2D
detection of r-bounding boxes



Inspired by: Zhu, M., Zhang, S., Zhong, Y., Lu, P., Peng, H., & Lenneman, J. (2021, September). Monocular 3d vehicle detection using uncalibrated traffic cameras through homography. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 3814-3821). IEEE.

Part 2

Uncalibrated 3D detection case

Simplifying the problem

4+ point correspondences between camera images and map

↓

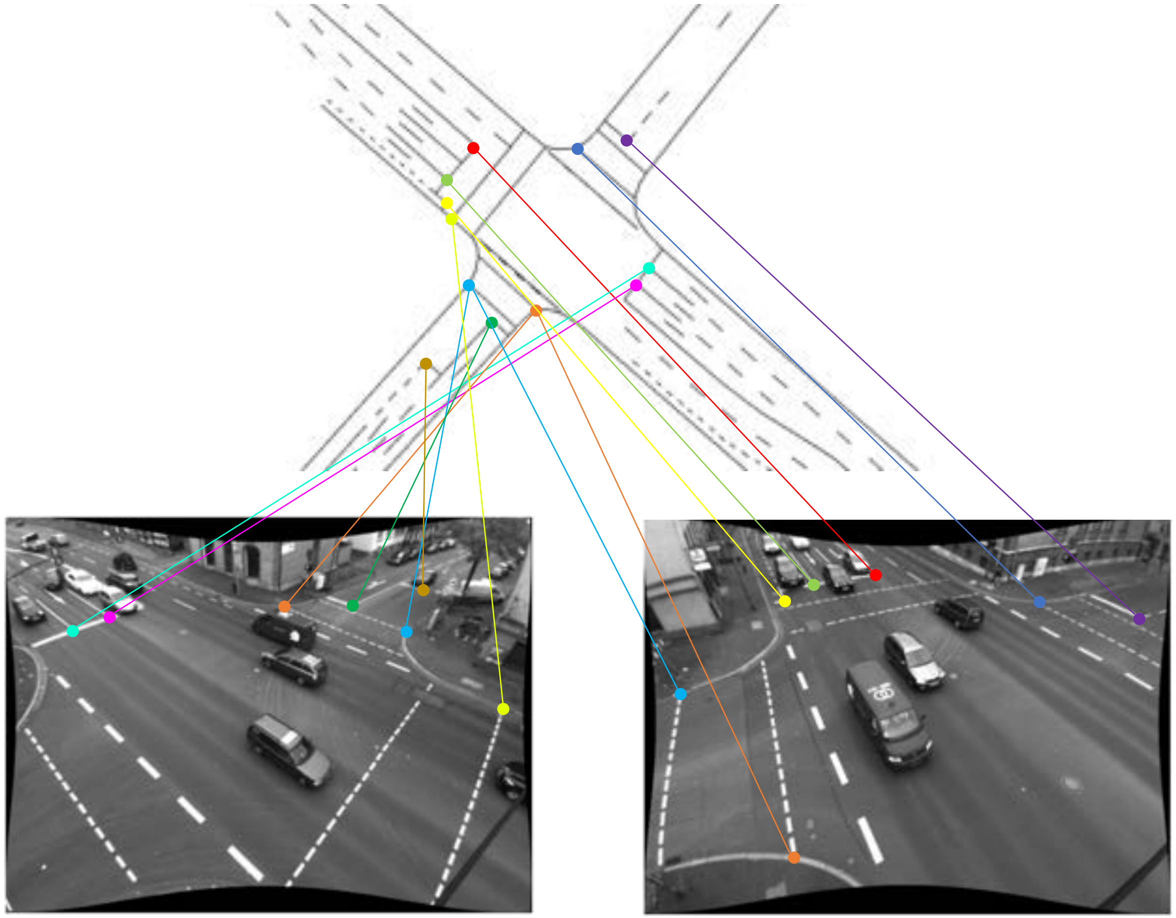
Calculation of homography, incl. the use of RANSAC

↓

Planar transformation of camera images and stitching into BEV

↓

Treatment of 3D problem as 2D detection of r-bounding boxes



Inspired by: Zhu, M., Zhang, S., Zhong, Y., Lu, P., Peng, H., & Lenneman, J. (2021, September). Monocular 3d vehicle detection using uncalibrated traffic cameras through homography. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 3814-3821). IEEE.

Part 2

Uncalibrated 3D detection case

Simplifying the problem

4+ point correspondences between camera images and map

↓

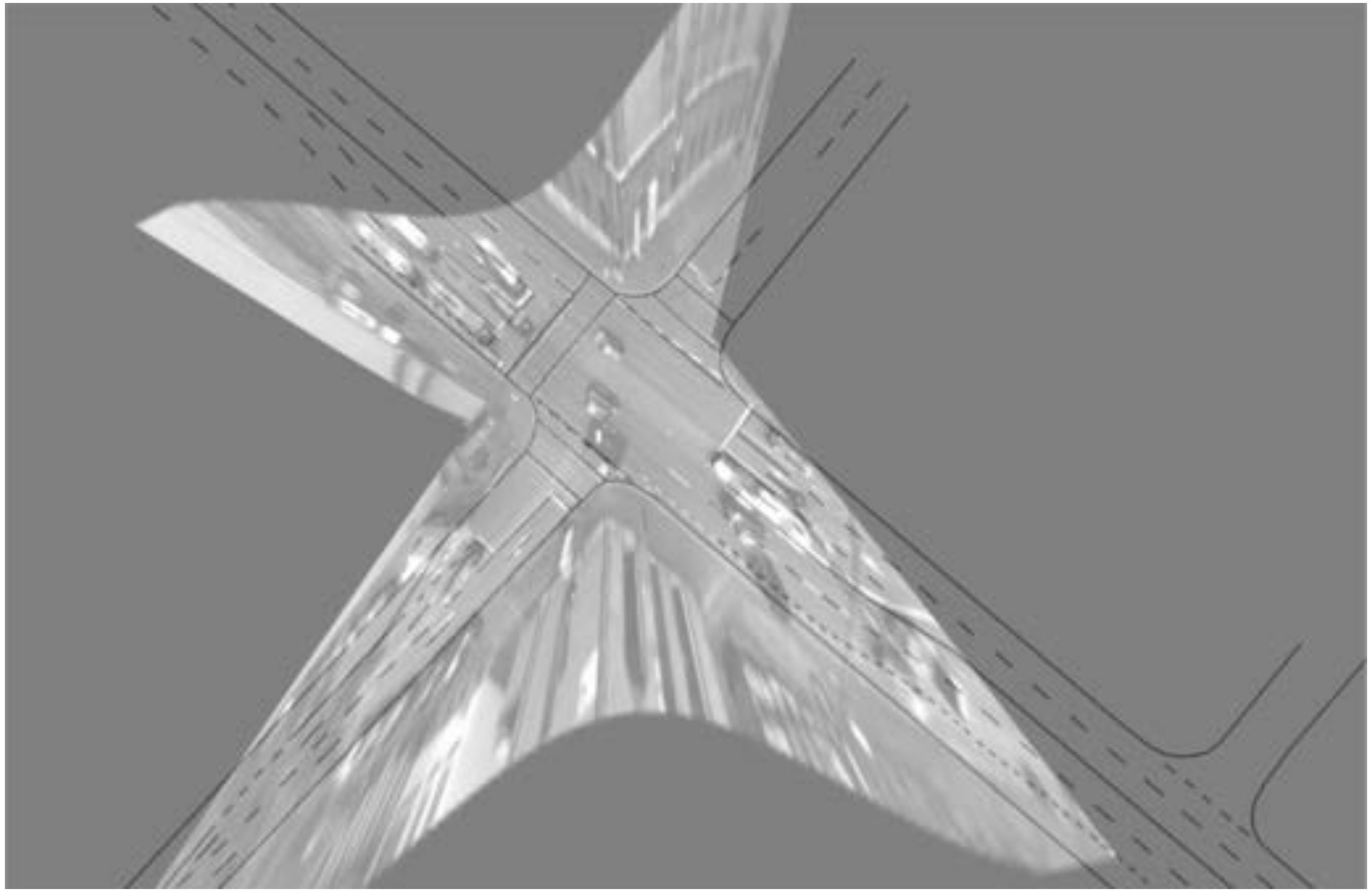
Calculation of homography, incl. the use of RANSAC

↓

Planar transformation of camera images and stitching into BEV

↓

Treatment of 3D problem as 2D detection of r-bounding boxes



Inspired by: Zhu, M., Zhang, S., Zhong, Y., Lu, P., Peng, H., & Lenneman, J. (2021, September). Monocular 3d vehicle detection using uncalibrated traffic cameras through homography. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 3814-3821). IEEE.

Part 2

Uncalibrated 3D detection case

Simplifying the problem

4+ point correspondences between camera images and map

↓

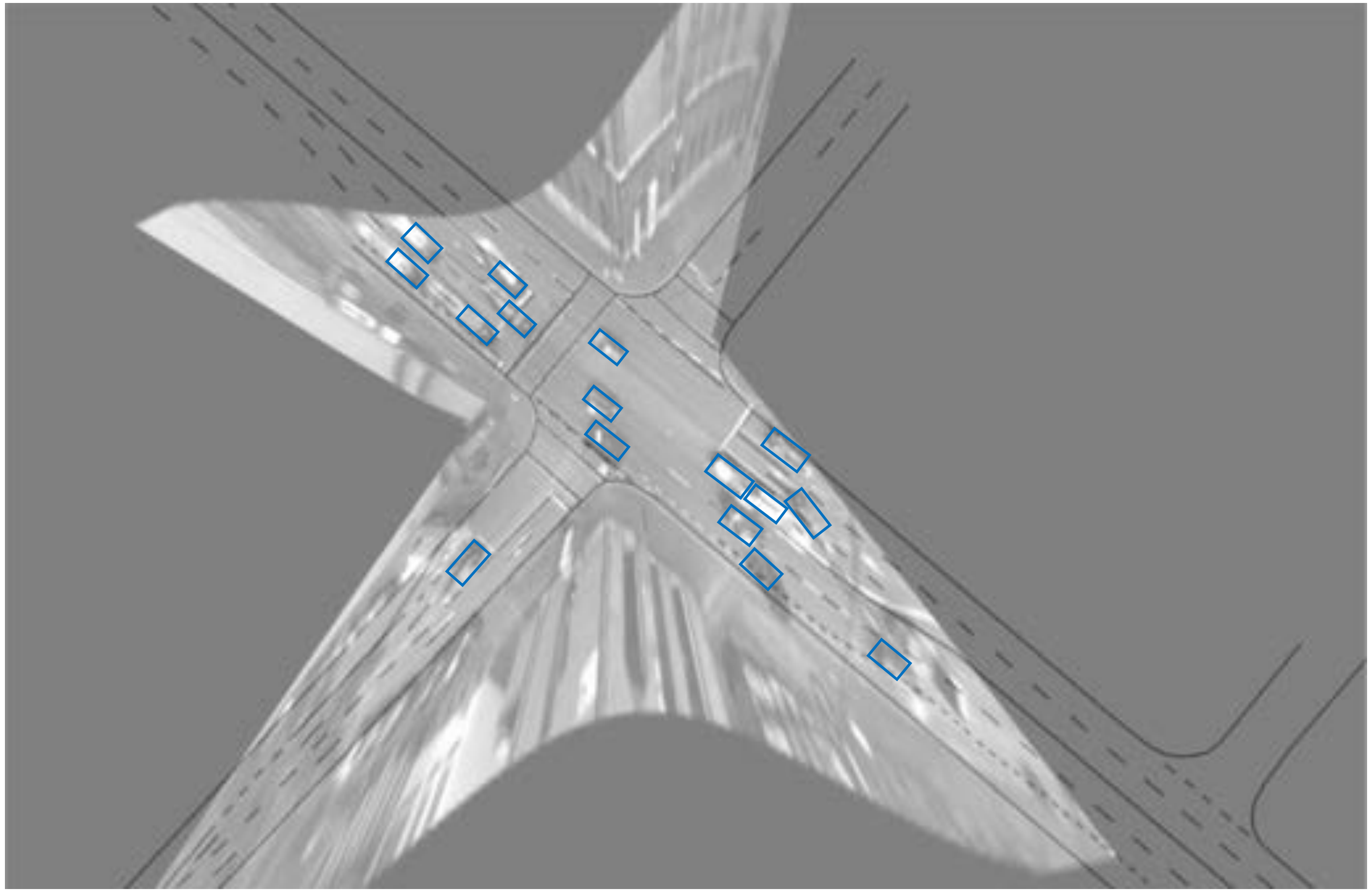
Calculation of homography, incl. the use of RANSAC

↓

Planar transformation of camera images and stitching into BEV

↓

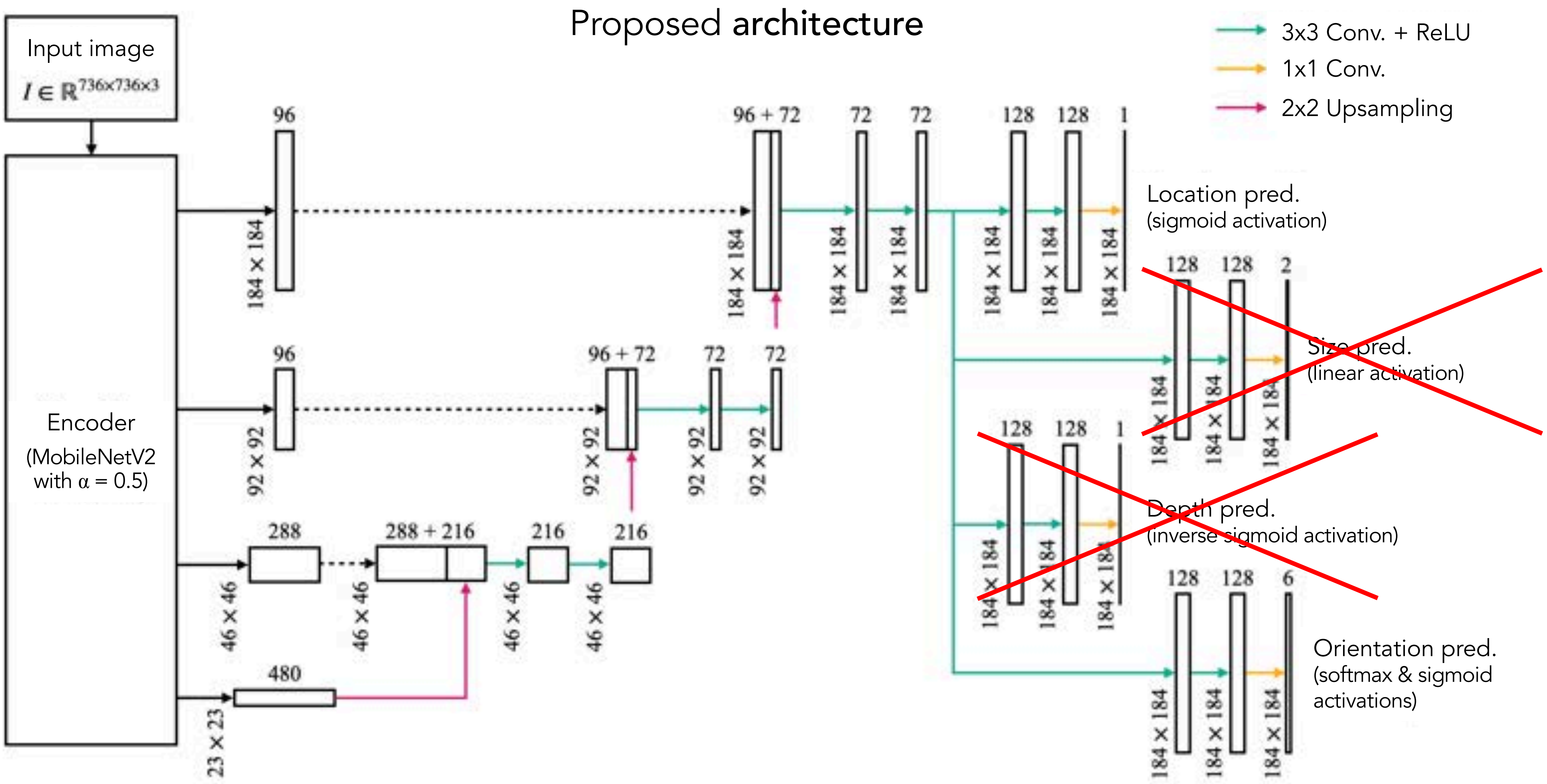
Treatment of 3D problem as 2D detection of r-bounding boxes



Inspired by: Zhu, M., Zhang, S., Zhong, Y., Lu, P., Peng, H., & Lenneman, J. (2021, September). Monocular 3d vehicle detection using uncalibrated traffic cameras through homography. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 3814-3821). IEEE.

Part 2

Uncalibrated 3D detection case



Part 2

Uncalibrated 3D detection case

Experiments

Data augmentation
- Random **horizontal flipping** (50% chance)
- Random **vertical flipping** (50% chance)
- Random **horizontal or vertical flipping** (50% chance)

Training conditions → Same 2 stage approach. Hyperparameters based on random search

Results

BEV AP			
No data augmentation	Random H flipping	Random V flippling	Random H & V flipping
62.37	71.14	59.93	58.09

Part 2

Uncalibrated 3D detection case

Experiments

Data augmentation
- Random **horizontal flipping** (50% chance)
- Random **vertical flipping** (50% chance)
- Random **horizontal or vertical flipping** (50% chance)

Training conditions → Same 2 stage approach. Hyperparameters based on random search

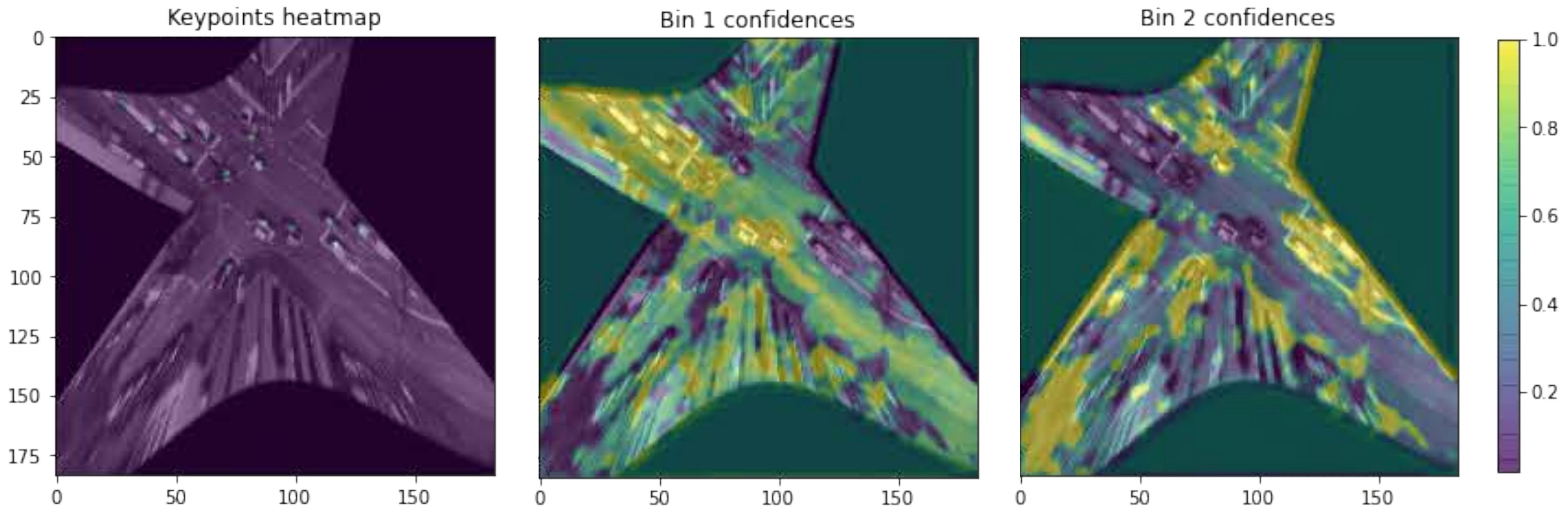
Results

BEV AP			
No data augmentation	**Random H flipping**	**Random V flippling**	**Random H & V flipping**
62.37	71.14	59.93	58.09

Part 2

Uncalibrated 3D detection case

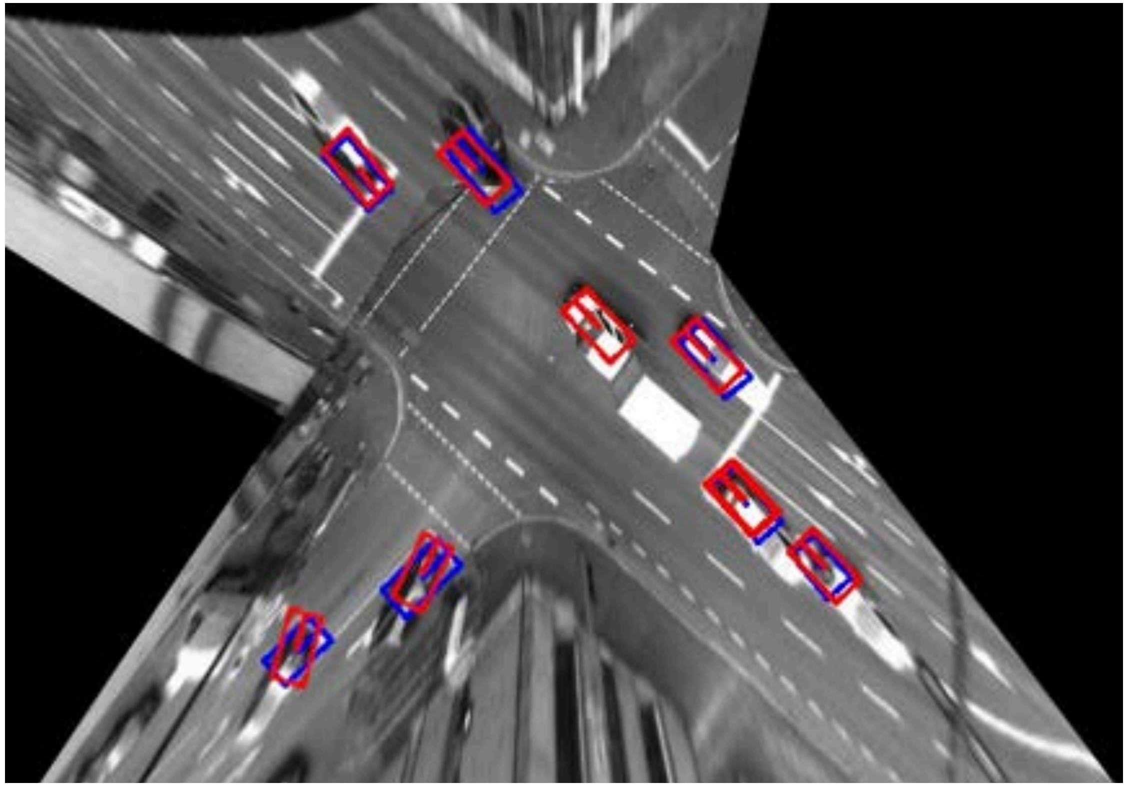
Visual examples

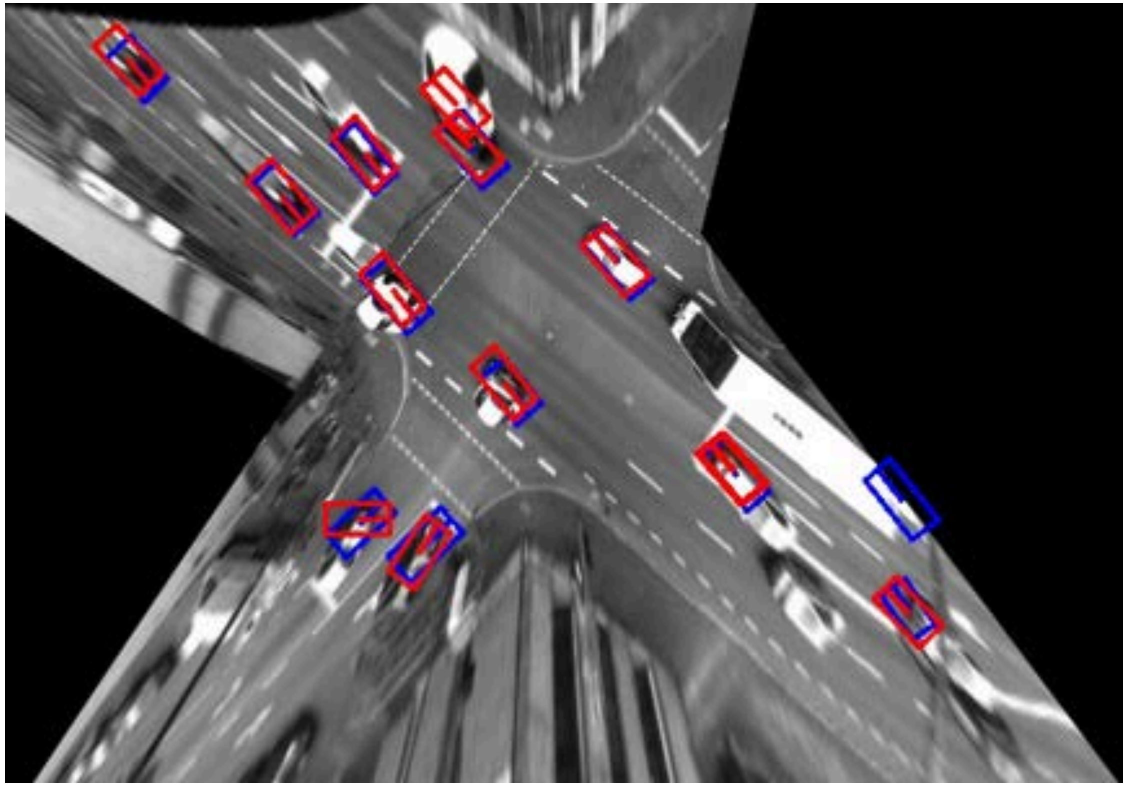


Part 2

Uncalibrated 3D detection case

Visual examples

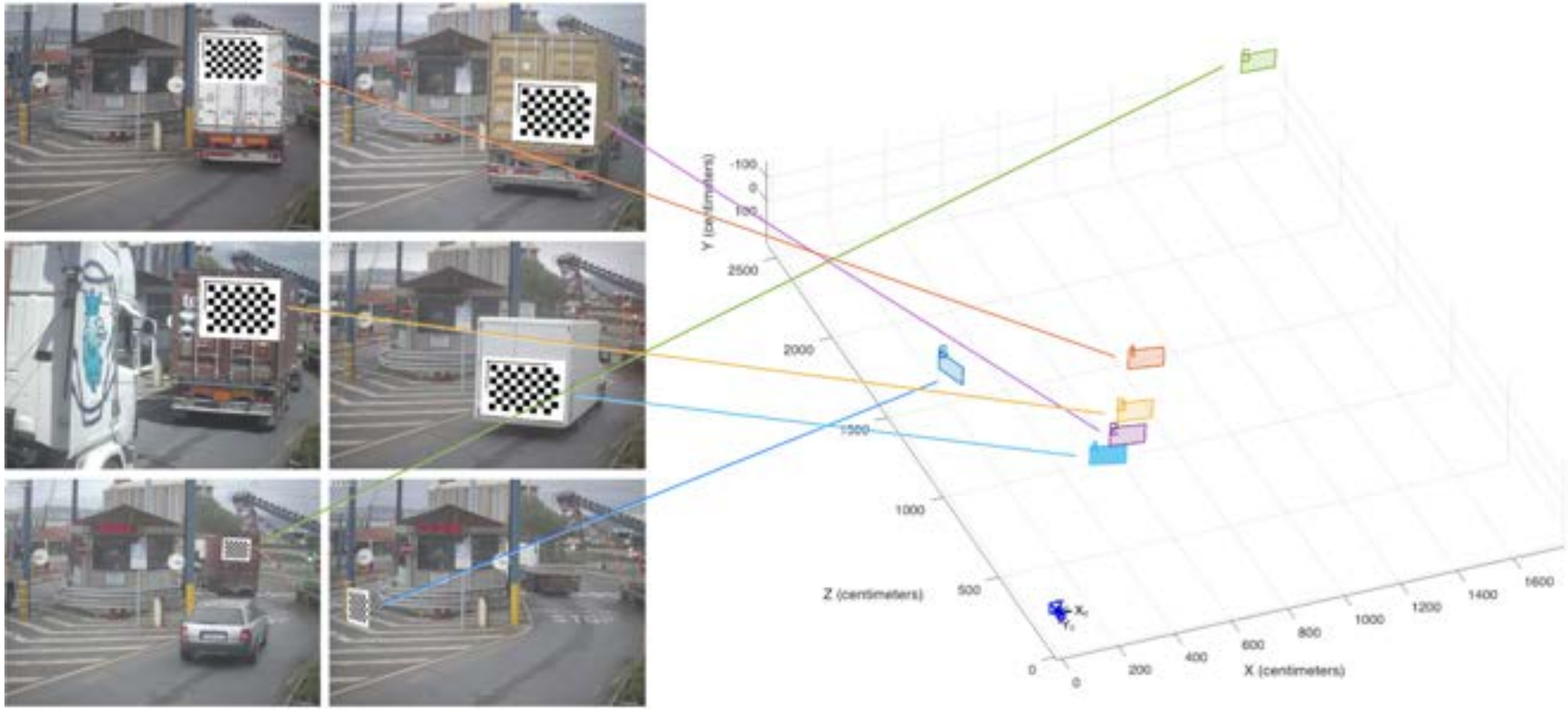




A few comments on the

Usability of the studied systems

Calibration is a must for accurate **full 3D** bounding box detection



It can be estimated for most uncalibrated cameras

A few comments on the

Usability of the studied systems

Still, the **difficult access to 3D annotated data** remains a problem

Possible **sources**

- **Existing datasets** Problem of camera angle and calibration compatibilities
- Annotating **own data** obtained with calibrated cameras and LiDARs
- Using **synthetic data**

All inviable in many scenarios

The proposed **uncalibrated detection approach** can still be considered

- A **satellite image** of most scenes can be obtained
- **Annotating rotated 2D bounding boxes** is (usually) feasible

Generalization?

A few comments on the

Usability of the studied systems

Still, the **difficult access to 3D annotated data** remains a problem

Possible **sources**

- **Existing datasets** — Problem of camera angle and calibration compatibilities
- Annotating **own data** obtained with calibrated cameras and LiDARs
- Using **synthetic data**

All inviable in many scenarios

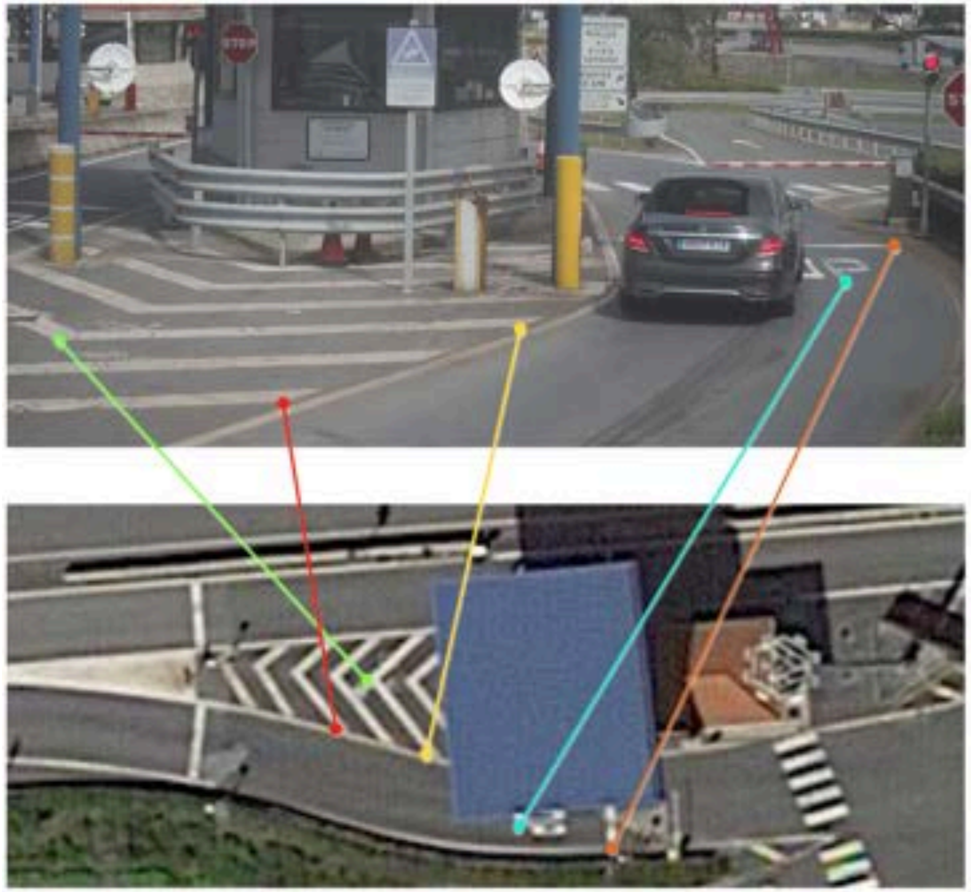
The proposed **uncalibrated detection approach** can still be considered

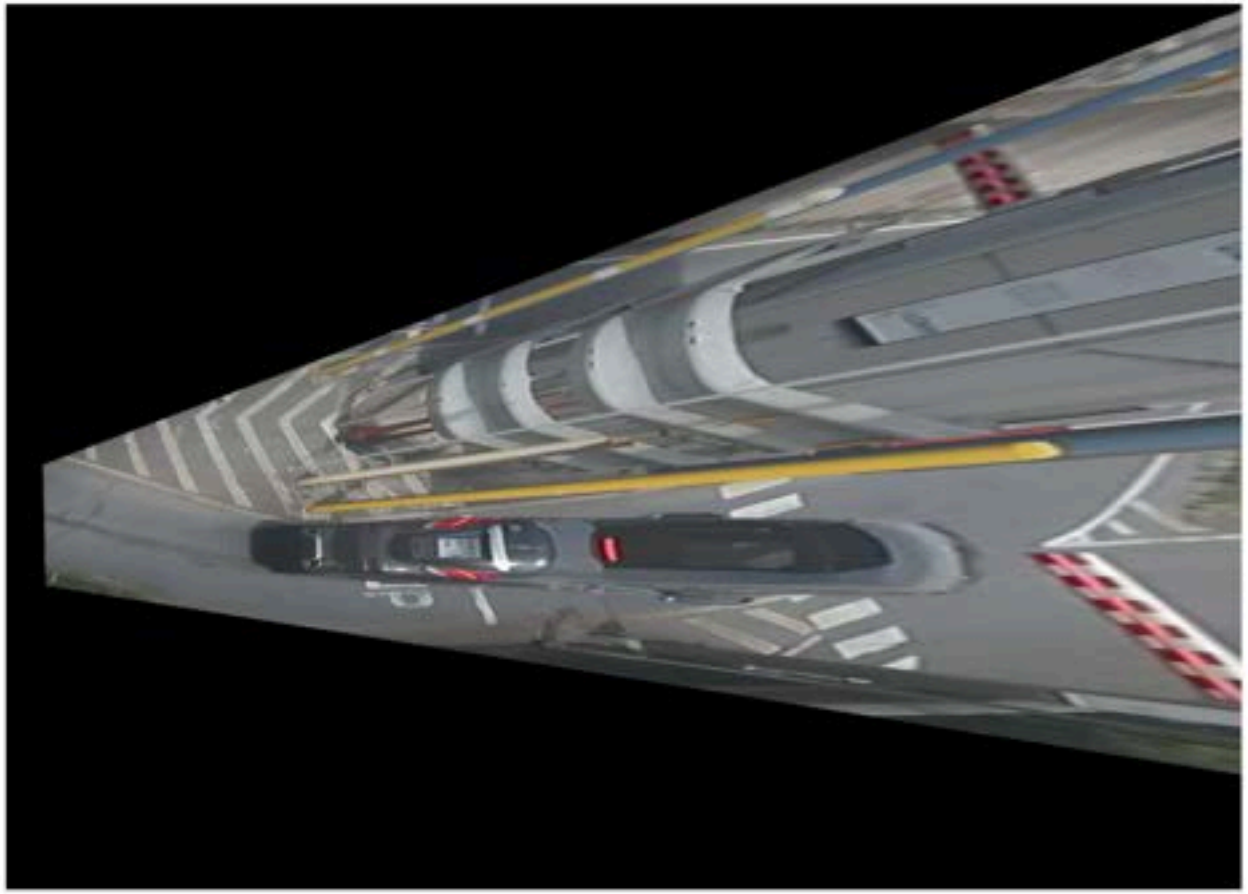
- A **satellite image** of most scenes can be obtained
- **Annotating rotated 2D bounding boxes** is (usually) feasible

Generalization?

A few comments on the

Usability of the studied systems





Finally, some

Key takeaways

- Relying exclusively on **monocular 3D detection** is challenging, but **doable to an extent**
- **Camera calibration** and **3D annotations** are necessary for accurate full 3D detection, but can be difficult to access
- When not available, part of the 3D detection task can still be carried out through a **simplification to a 2D detection problem** of rotated bounding boxes on BEVs
- The **end-to-end differentiability**, **speed** and **adaptability** of the **CenterNet** make it a great fit for the task of 3D detection
- **Lightweight variants** of the CenterNet with simpler feature extraction architectures can still provide **functional** detection capabilities
- There is plenty of progress to be made in terms of **automatization** and **generalization**

Finally, some

Key takeaways

- Relying exclusively on **monocular 3D detection** is challenging, but **doable to an extent**
- **Camera calibration** and **3D annotations** are necessary for accurate full 3D detection, but can be difficult to access
- When not available, part of the 3D detection task can still be carried out through a **simplification to a 2D detection problem** of rotated bounding boxes on BEVs
- The **end-to-end differentiability**, **speed** and **adaptability** of the **CenterNet** make it a great fit for the task of 3D detection
- **Lightweight variants** of the CenterNet with simpler feature extraction architectures can still provide **functional** detection capabilities
- There is plenty of progress to be made in terms of **automatization** and **generalization**

Finally, some

Key takeaways

- Relying exclusively on **monocular 3D detection** is challenging, but **doable to an extent**
- **Camera calibration** and **3D annotations** are necessary for accurate full 3D detection, but can be difficult to access
- When not available, part of the 3D detection task can still be carried out through a **simplification to a 2D detection problem** of rotated bounding boxes on BEVs
- The **end-to-end differentiability**, **speed** and **adaptability** of the **CenterNet** make it a great fit for the task of 3D detection
- **Lightweight variants** of the CenterNet with simpler feature extraction architectures can still provide **functional** detection capabilities
- There is plenty of progress to be made in terms of **automatization** and **generalization**

Finally, some

Key takeaways

- Relying exclusively on **monocular 3D detection** is challenging, but **doable to an extent**
- **Camera calibration** and **3D annotations** are necessary for accurate full 3D detection, but can be difficult to access
- When not available, part of the 3D detection task can still be carried out through a **simplification to a 2D detection problem** of rotated bounding boxes on BEVs
- The **end-to-end differentiability**, **speed** and **adaptability** of the **CenterNet** make it a great fit for the task of 3D detection
- **Lightweight variants** of the CenterNet with simpler feature extraction architectures can still provide **functional** detection capabilities
- There is plenty of progress to be made in terms of **automatization** and **generalization**

Finally, some

Key takeaways

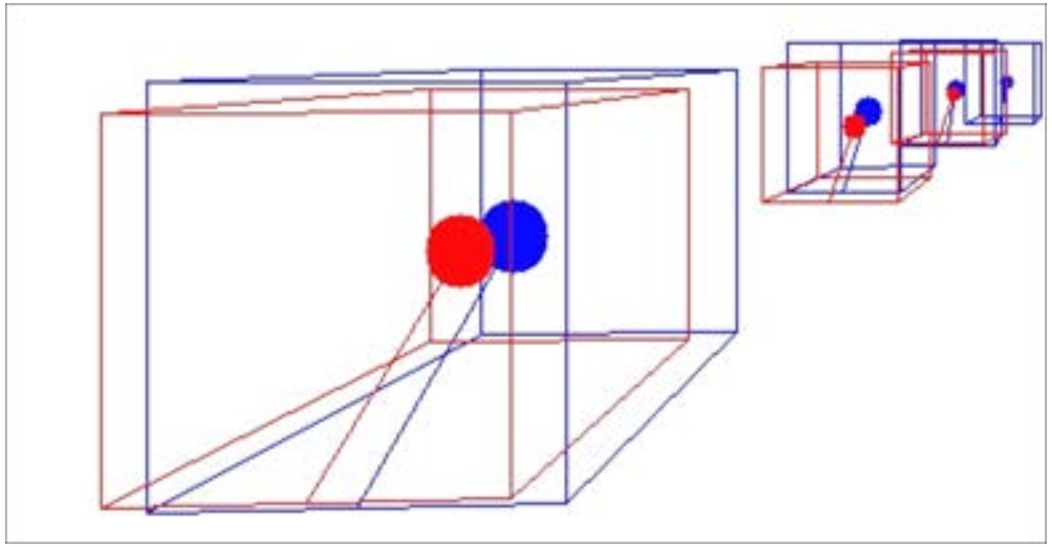
- Relying exclusively on **monocular 3D detection** is challenging, but **doable to an extent**
- **Camera calibration** and **3D annotations** are necessary for accurate full 3D detection, but can be difficult to access
- When not available, part of the 3D detection task can still be carried out through a **simplification to a 2D detection problem** of rotated bounding boxes on BEVs
- The **end-to-end differentiability**, **speed** and **adaptability** of the **CenterNet** make it a great fit for the task of 3D detection
- **Lightweight variants** of the CenterNet with simpler feature extraction architectures can still provide **functional** detection capabilities
- There is plenty of progress to be made in terms of **automatization** and **generalization**

Finally, some

Key takeaways

- Relying exclusively on **monocular 3D detection** is challenging, but **doable to an extent**
- **Camera calibration** and **3D annotations** are necessary for accurate full 3D detection, but can be difficult to access
- When not available, part of the 3D detection task can still be carried out through a **simplification to a 2D detection problem** of rotated bounding boxes on BEVs
- The **end-to-end differentiability**, **speed** and **adaptability** of the **CenterNet** make it a great fit for the task of 3D detection
- **Lightweight variants** of the CenterNet with simpler feature extraction architectures can still provide **functional** detection capabilities
- There is plenty of progress to be made in terms of **automatization** and **generalization**

Thank you!



Eric Henriksson Martí

E-mail: *eckehenrikssonmarti@gmail.com*

LinkedIn: *linkedin.com/in/erichenriksson/*

Mobile: *+34 618 777 779*