Multi-sensor data fusion for VRU detection

Ljubomir Jovanov
imec-IPI-Ghent University

Delft, the Netherlands

21 June 2023

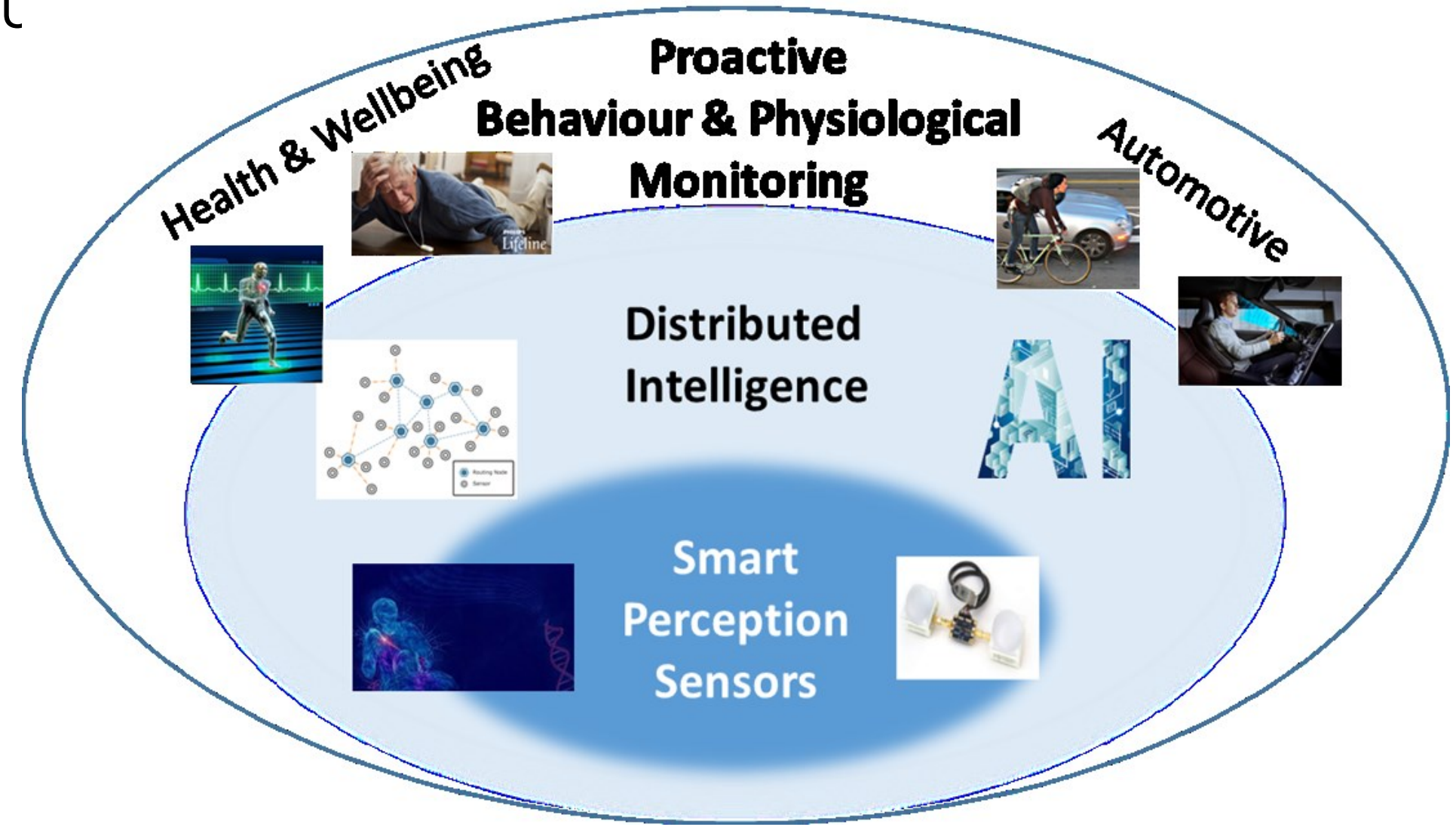
Workshop 2023
ePicture This

Organized by Penta projects:
2020005 Mantis Vision
2021004 Imagination

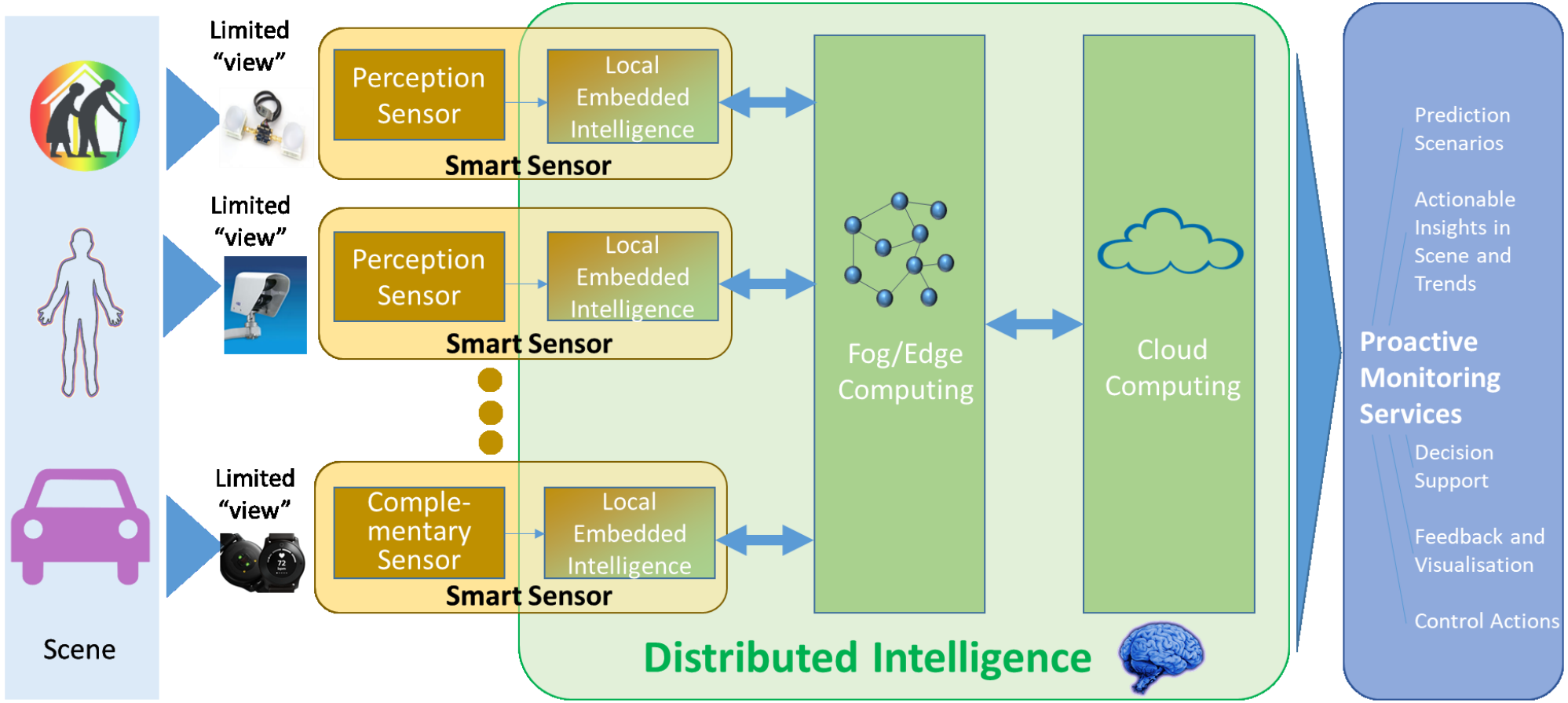
Agenda

- Introduction
- Semi-supervised radar based VRU detection
- Semantic-guided radar-video fusion
- Thermal video fusion for VRU Detection
- Future work-integration
- Conclusions

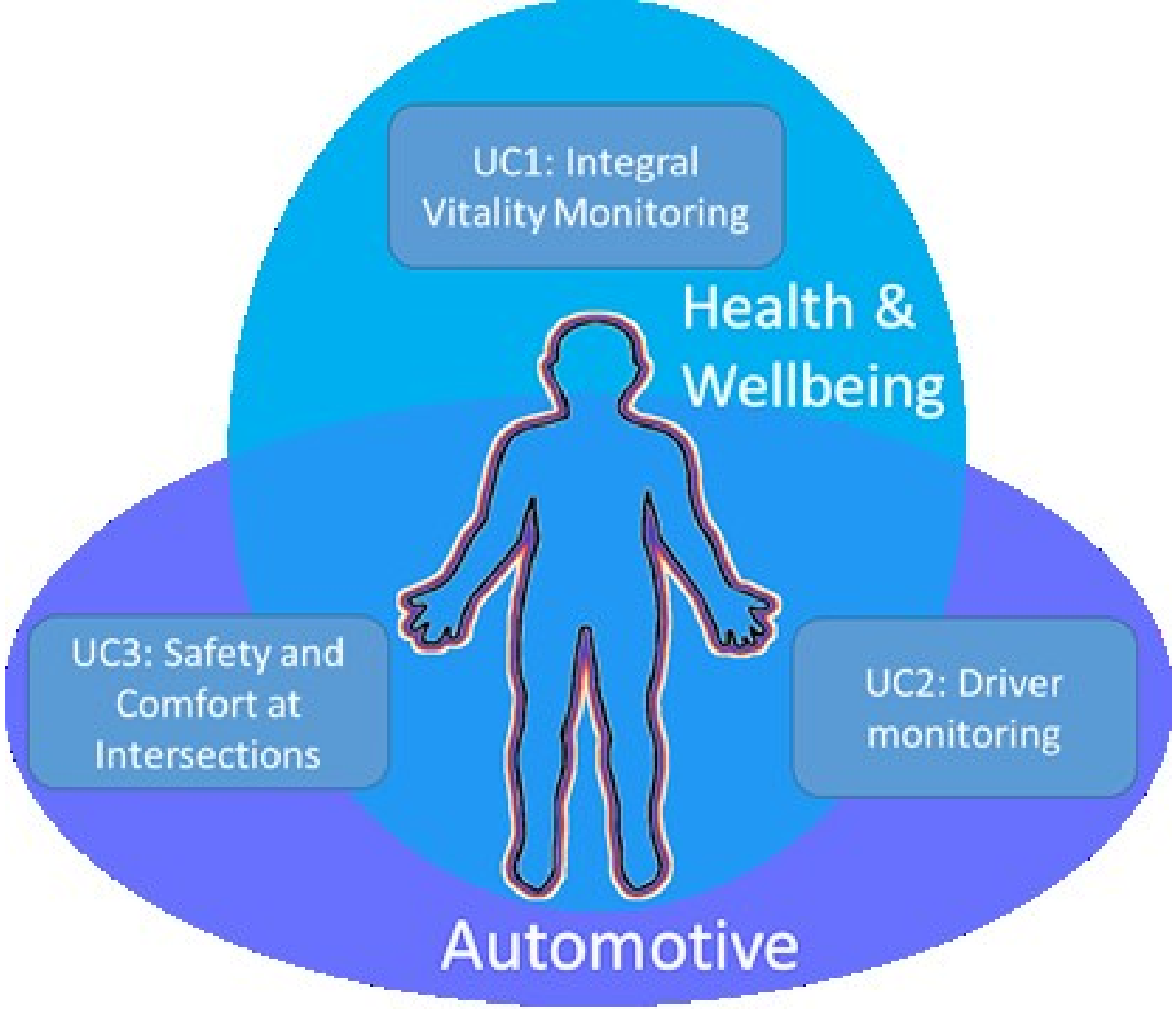
INTRODUCTION



CONCEPTUAL ARCHITECTURE

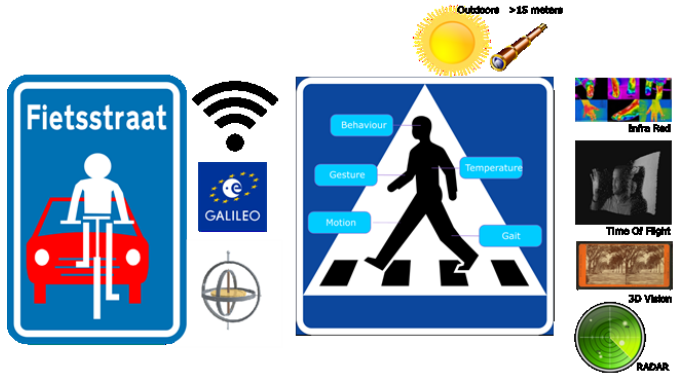


USE CASES

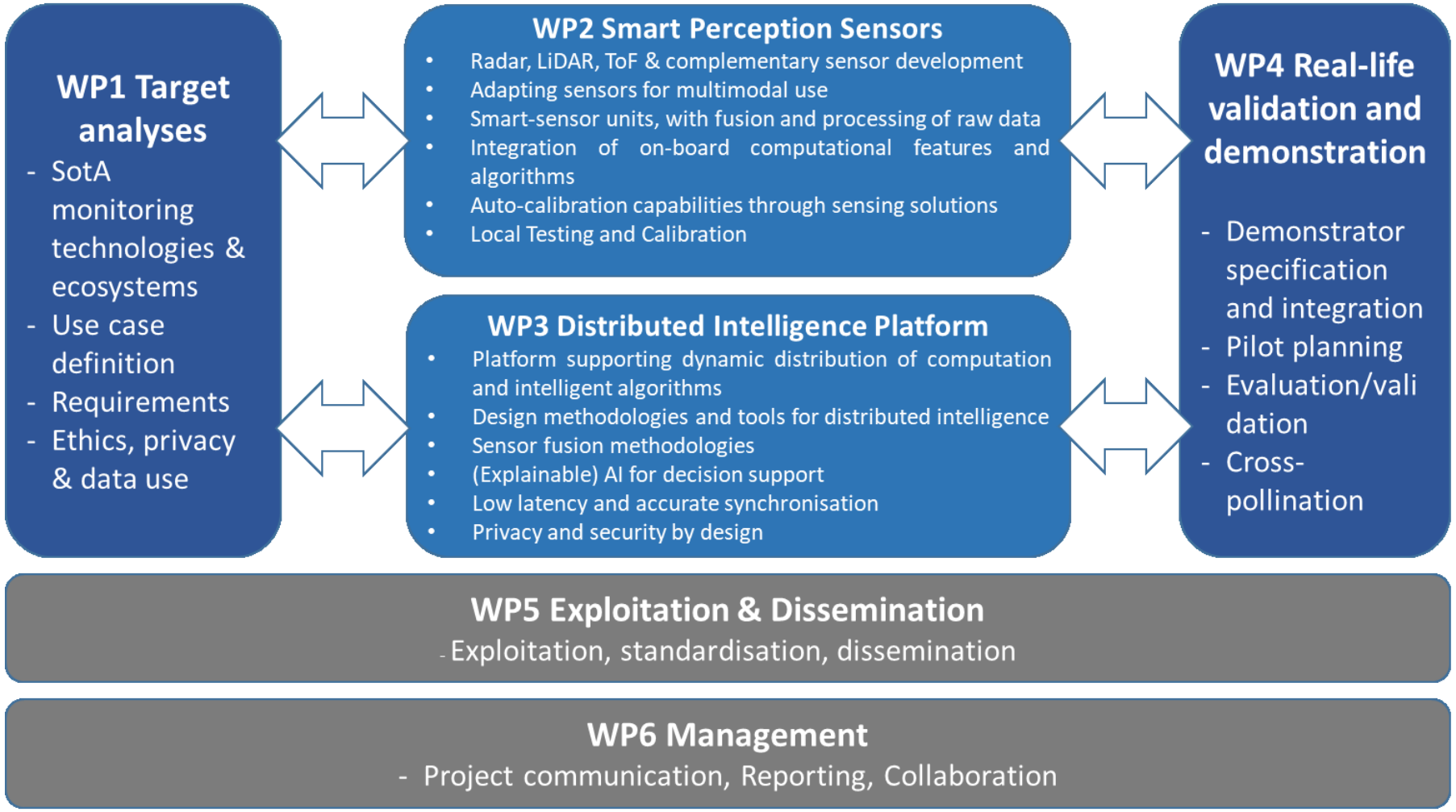

UC1: Integral Vitality Monitoring
Health & Wellbeing
UC3: Safety and Comfort at Intersections
UC2: Driver monitoring
Automotive

UC3: Safety and comfort at intersections

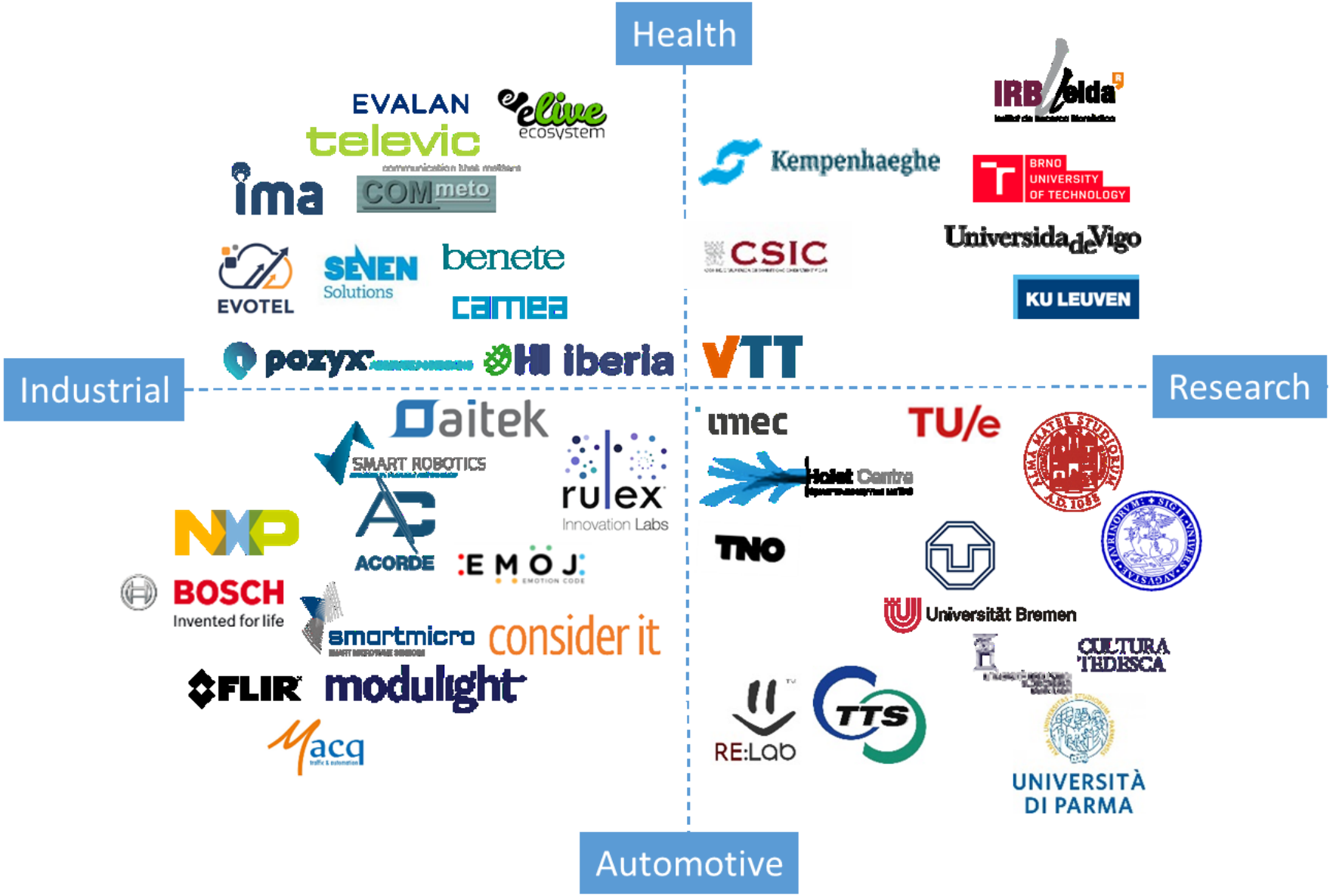
- Provide safety and comfort for all road users - including Vulnerable Road Users (VRUs) like pedestrians and bikers - at a road intersection. The use case will demonstrate the capability to detect the presence of traffic participants, determine their positions and track their motion and intent with high reliability. Specifically, for VRUs this concerns body gestures analysis to derive their intent.



WORK BREAKDOWN STRUCTURE



Cons

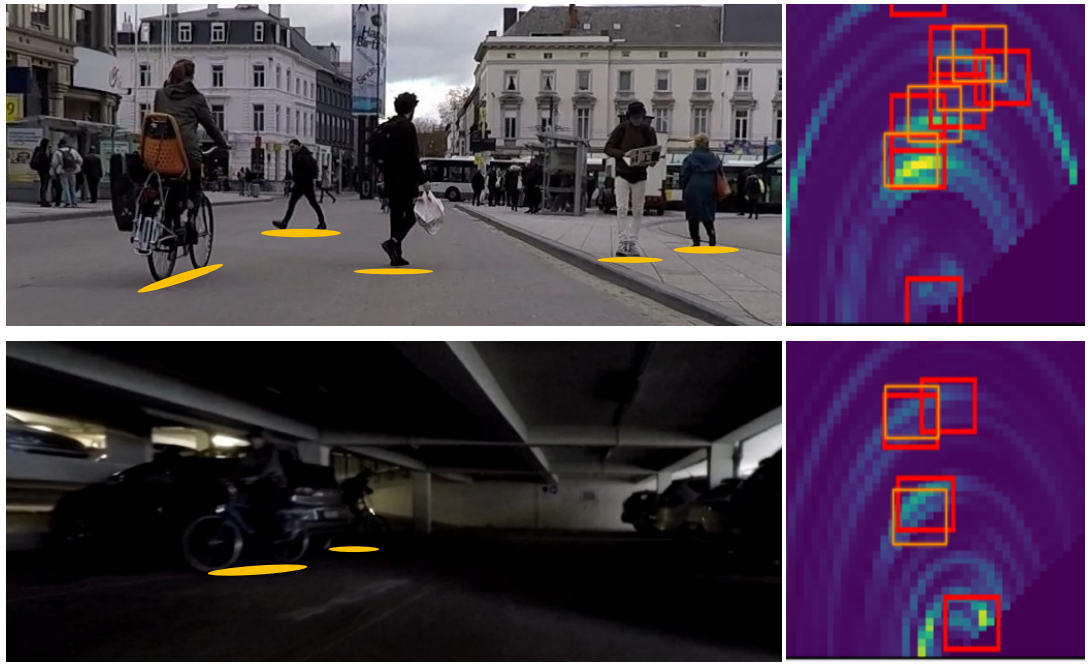


43 Partners
7 Countries

Semi-supervised radar based VRU detection

SEMI-SUPERVISED RADAR BASED VRU DETECTION

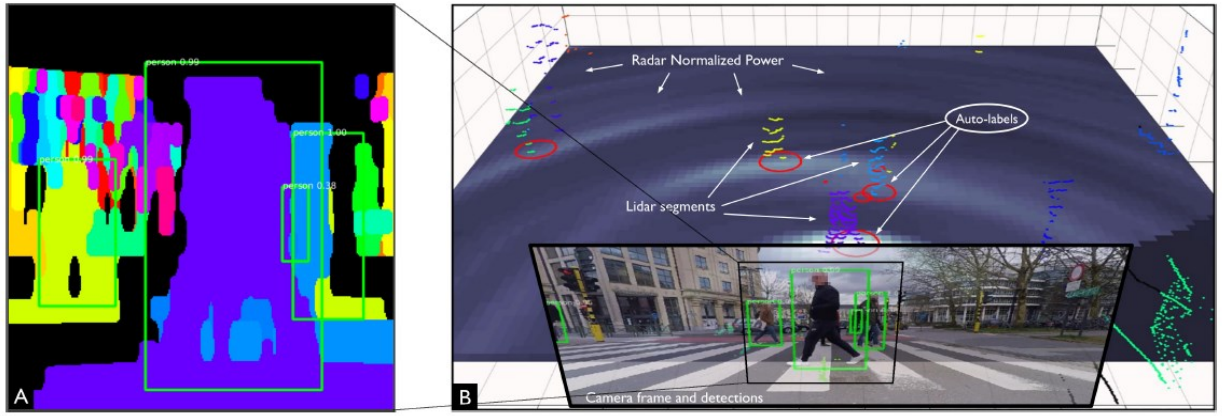
- Key component of perception modules
  - Robustness
  - Lower price
- Problem
  - Requiring expert knowledge
  - More time-consuming than general image
  - Lack of annotated raw radar datasets



An example of radar scan and the synchronized camera image in different light condition.

RELATED WORK

- Previous solutions
  - Weakly supervised learning
    - Introducing external sensor (LiDAR) [1]
    - Automatically generating labels [3]
  - Several concerns
    - Imperfect sensors with known failure modes:
      - Extreme weather/lighting condition recordings are unavailable
      - Occlusion when matching camera and lidar detections
    - Imperfect generated labels

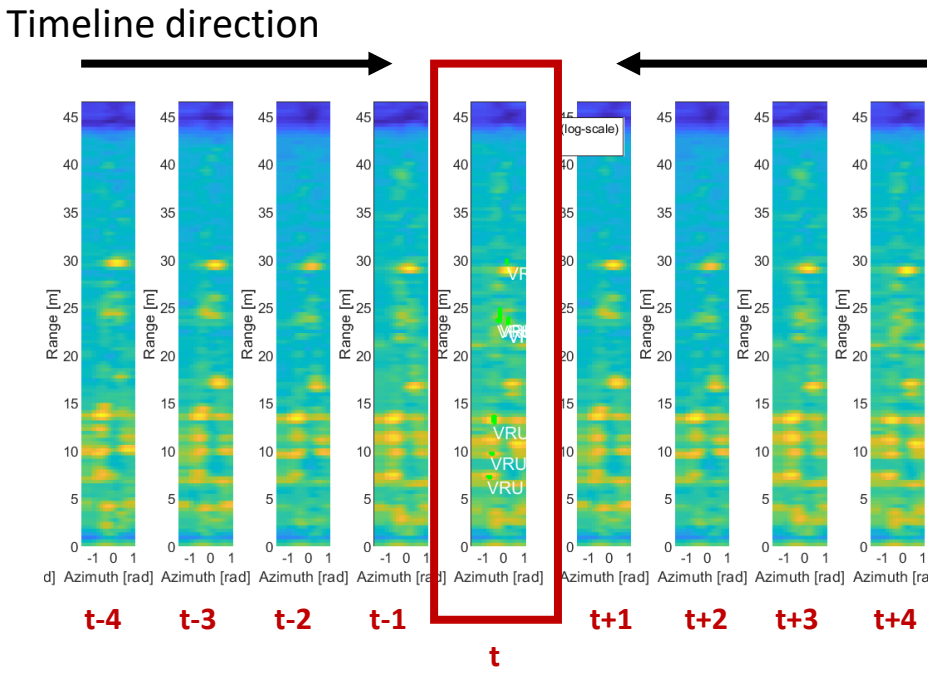


Outline

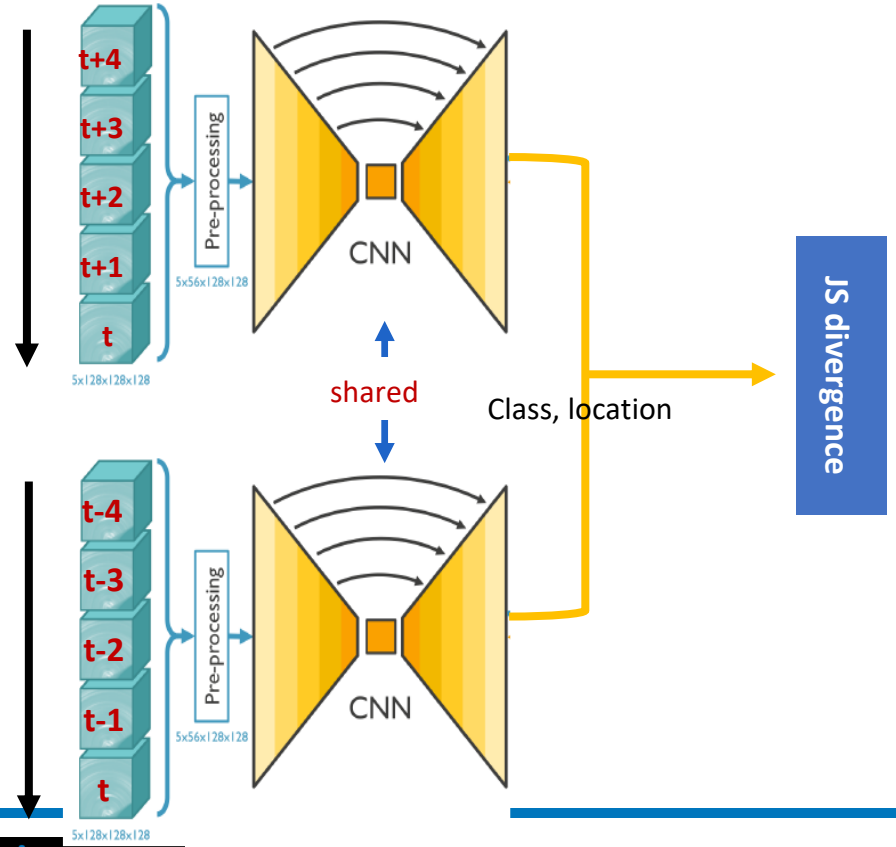
- Proposed solution
  - Semi-supervised learning
    - Spatio-temporal consistency
    - Overall objective function
    - Self-weighted mechanism
    - Network architecture
- Experiments
  - Datasets
  - Implementation details
  - Experimental results

SPATIO-TEMPORAL CONSISTENCY (**TRAINING STAGE**)

- Exploring radar features consistency by using **unlabeled data**
  - Temporal consistency

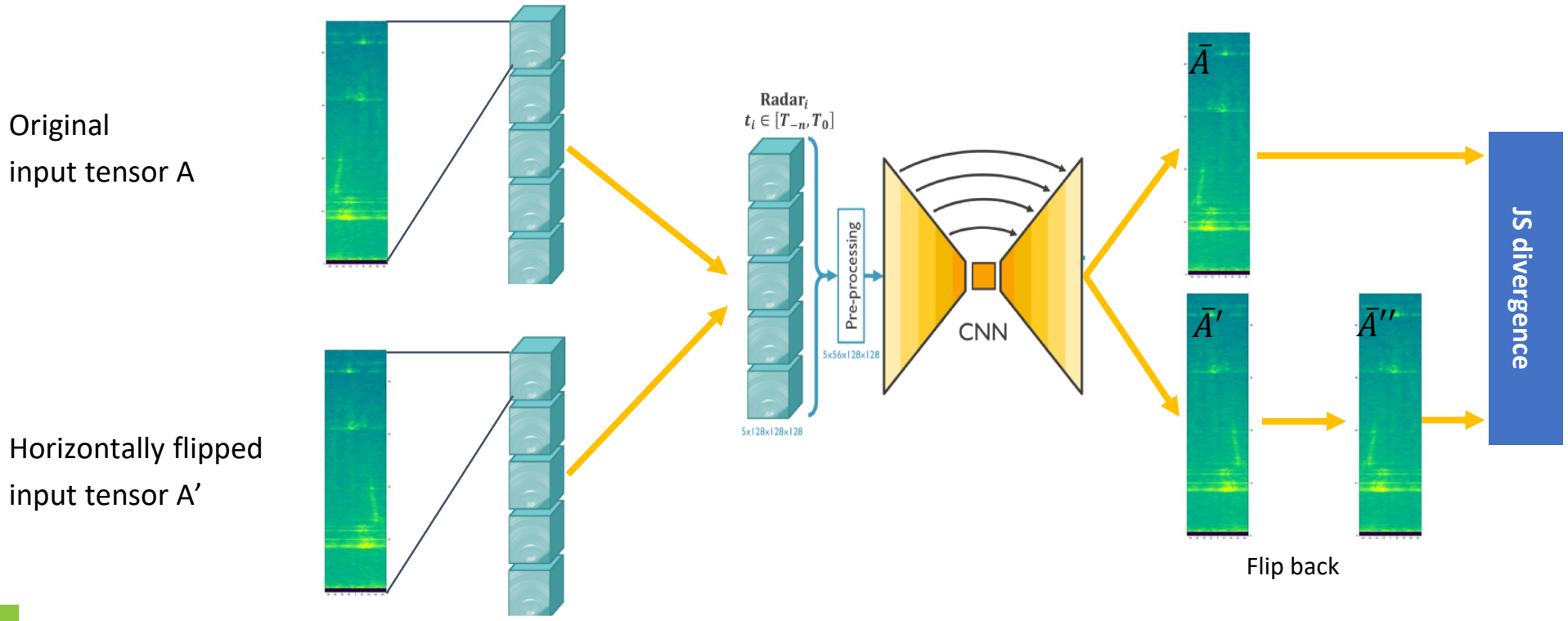


Detecting the objects at time t from different timeline direction



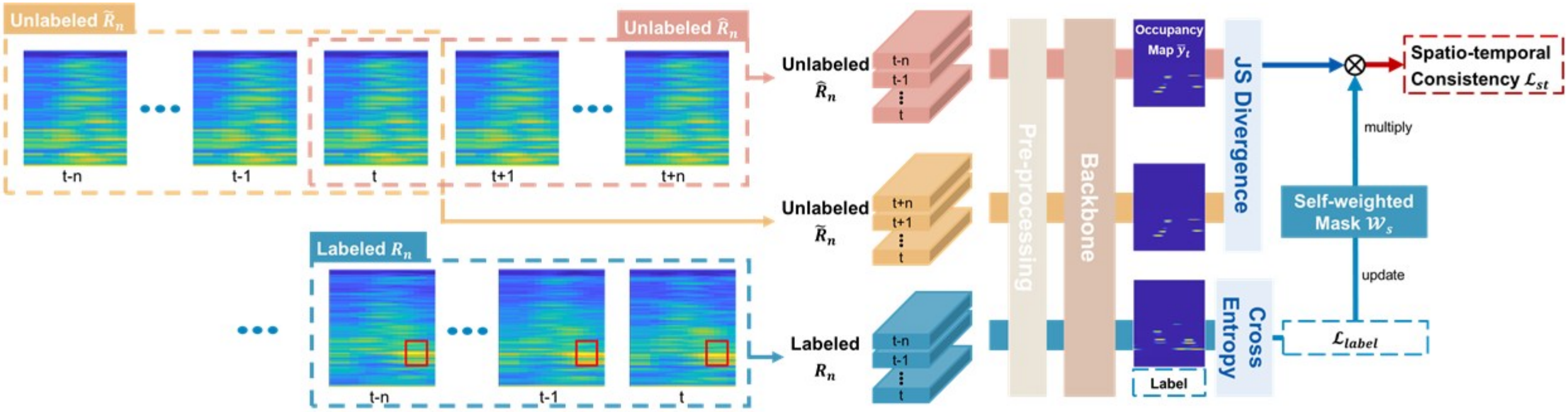
SPATIO-TEMPORAL CONSISTENCY (**TRAINING STAGE**)

- Exploring the radar feature consistency by using **unlabeled data**
  - Spatial consistency



Different from data augmentation, we learn the symmetric relationship without labels

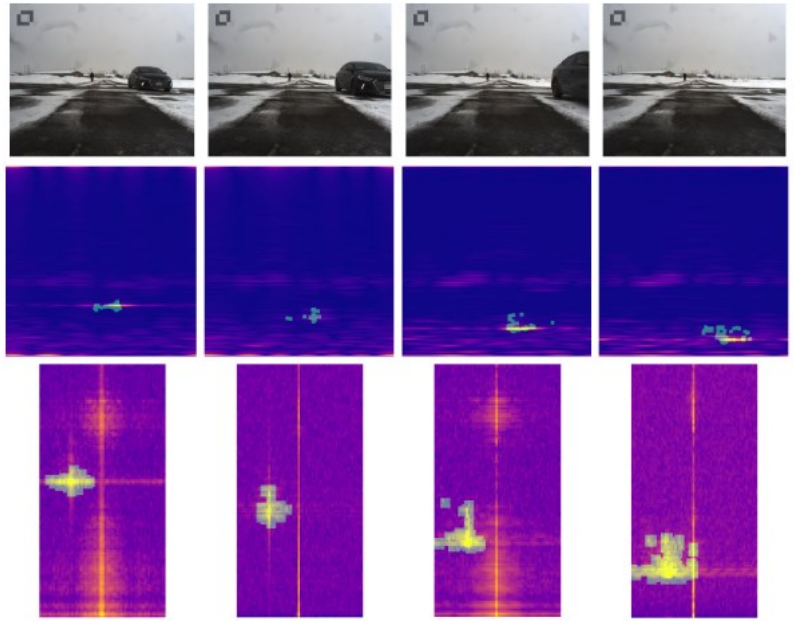
NETWORK ARCHITECTURE



Data flow

DATASETS

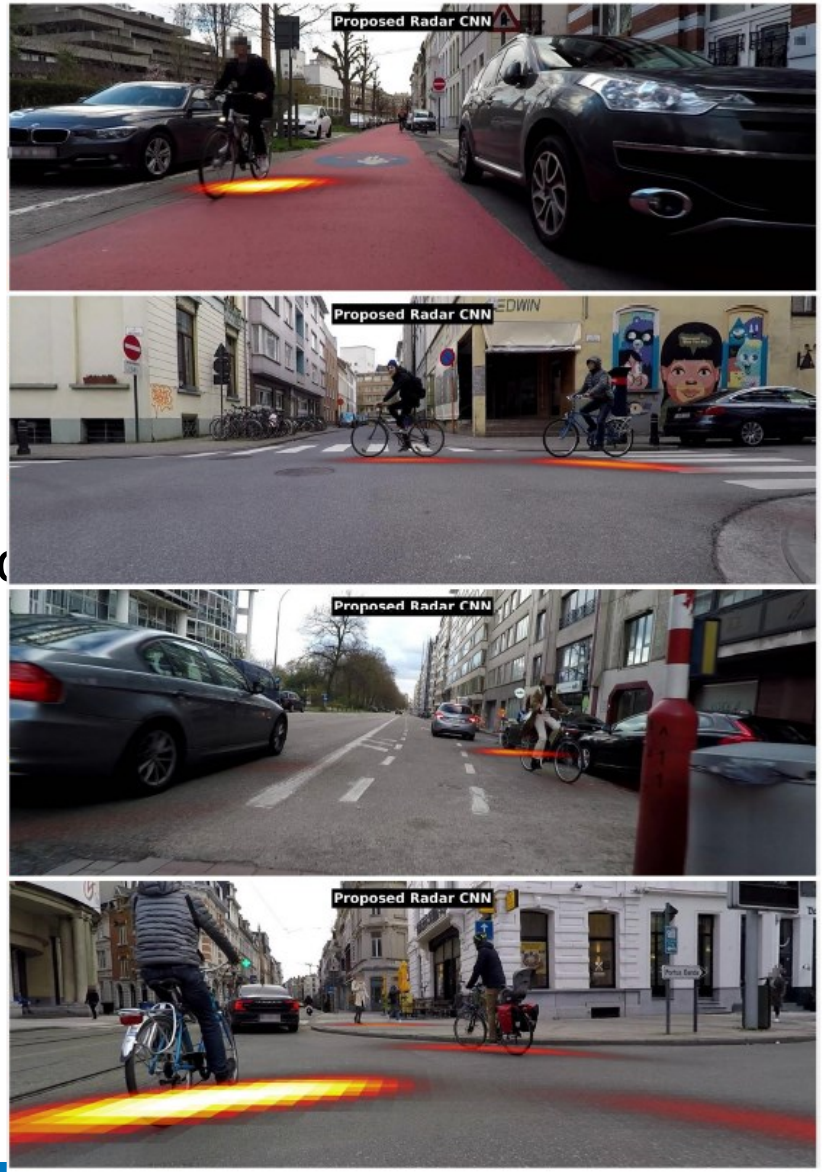
- Problem:
  - Semantic segmentation
- CARRADA dataset [2]:
  - 12726 synchronized FMCW radar & camera frames
  - Mounted on a stationary car
  - Controlled environment
  - Targets including cars, pedestrians, and cyclists
  - Semantic segmentation labels:
    - Range-Doppler (RD)
    - Range-Angle (RA)



Ouaknine, Arthur, et al. "CARRADA Dataset: Camera and Automotive Radar with Range-Angle-Doppler Annotations." *arXiv preprint arXiv:2005.01456* (2020).

DATASETS

- Problem:
  - Vulnerable Road Users (VRUs) detection
- Ghent Zuid radar dataset [3]:
  - Captured by synchronized FMCW radar, camera, and
  - Mounted on a moving vehicle (more challenging)
  - Public road scenes (more challenging)
  - 1840 labeled frames, 4988 annotated VRUs
  - 6955 unlabeled frames



EXPERIMENTAL RESULTS

- Semantic segmentation
  - Our model was trained only on 30% labeled data, and 70% unlabeled data
  - Clearly improved the performance, especially the pixel recall rate

TABLE I

SEMANTIC SEGMENTATION PERFORMANCE COMPARISON ON CARRADA DATASET. THE LEFT/RIGHT DIGITS IN EACH ENTRY DENOTE THE PIXEL ACCURACY/RECALL. L: LABELED DATA, U: UNLABELED DATA. NOTE THAT FCN-8S RESULTS ARE BORROWED FROM [13], AND THE MODEL WAS TRAINED BY SINGLE RADAR FRAME AS INPUT.

Data	Method	Training data	Pixel Accuracy / Recall				
			Background	Pedestrian	Cyclist	Car	Mean
RD	FCN-8s [2]	All	99.9 / 99.7 %	**53.6** / 67.9 %	28.6 / 36.8 %	60.6 / 79.4 %	60.7 / 71.0 %
	FCN-8s *	L: 30%	99.9 / 99.6 %	49.6 / 41.3 %	32.1 / 68.1 %	59.2 / 69.2 %	60.2 / 69.6 %
	Ours	L: 30%, U: 70%	99.9 / 99.2 %	52.8 / **89.1** %	**35.5** / **78.8** %	**63.7** / **93.0** %	**63.0** / **90.0** %
RA	FCN-8s [2]	All	99.9 / 99.9 %	28.2 / 11.4 %	2.8 / 5.2 %	50.3 / 39.9 %	45.3 / 39.1 %
	FCN-8s *	L: 30%	99.9 / 99.9 %	26.4 / 13.4 %	2.1 / 4.7 %	48.7 / 35.1 %	44.3 / 38.3 %
	Ours	L: 30%, U: 70%	99.9 / 99.9 %	**33.8** / **20.1** %	**5.7** / **10.4** %	**53.2** / **46.3** %	**48.2** / **44.2** %

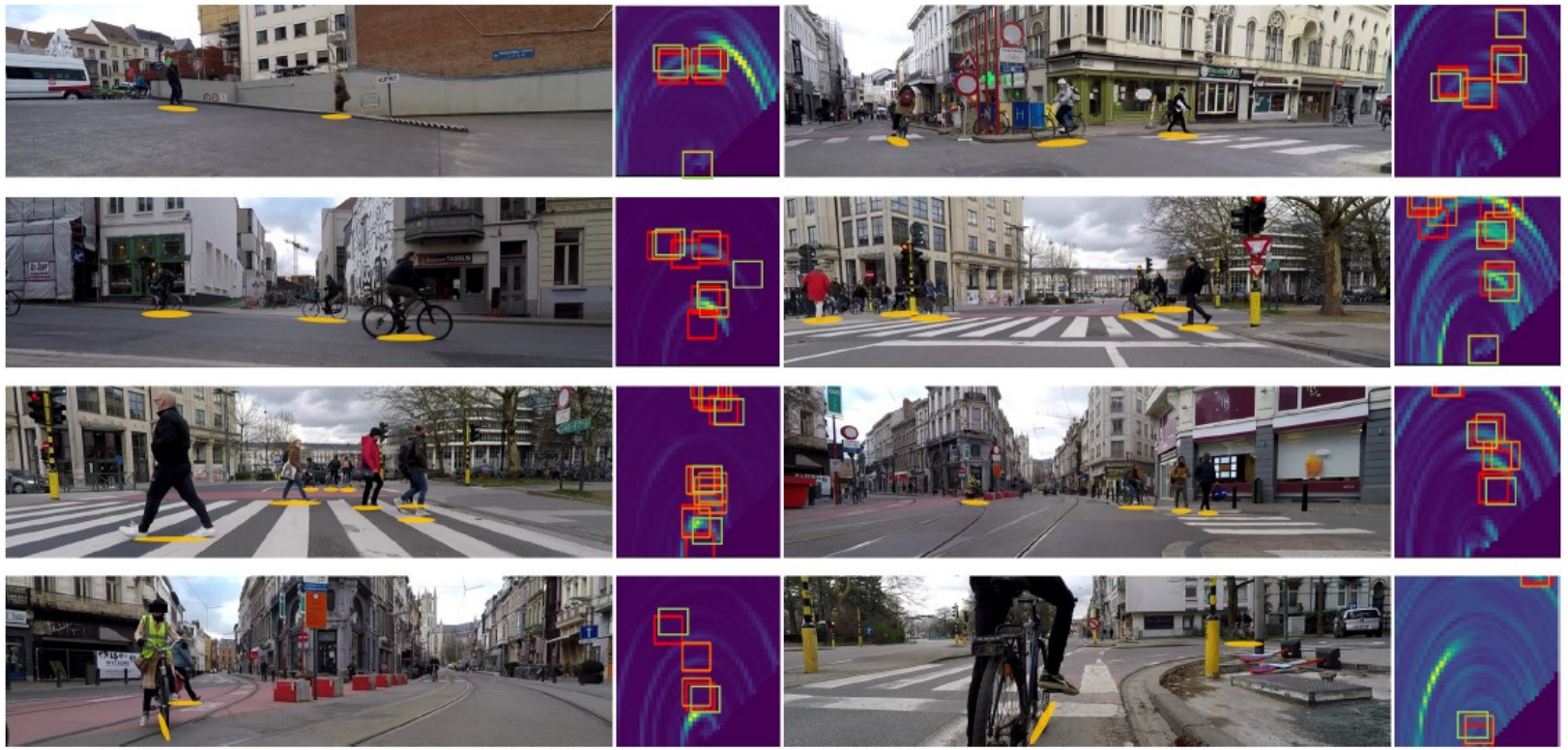
EXPERIMENTAL RESULTS

- Vulnerable Road Users (VRUs) detection
  - Our model was trained on only 1351 manual labeled frames and 6955 unlabeled frames
  - Without using external sensors, we still can achieve comparable

TABLE II

VRU DETECTION PERFORMANCE COMPARISON ON THE DATASET RELEASED BY [1]. NOTE THAT ONLY CNN-AUTO * AND CNN-AUTO ** USE AUTO LABELS AS TRAINING DATA, WHICH ARE GENERATED BY USING THE SYNCHRONIZED EXTERNAL CAMERA AND LIDAR DATA [1]. ALL THE METHODS ARE USING U-NET [15] AS THE NETWORK BACKBONE.

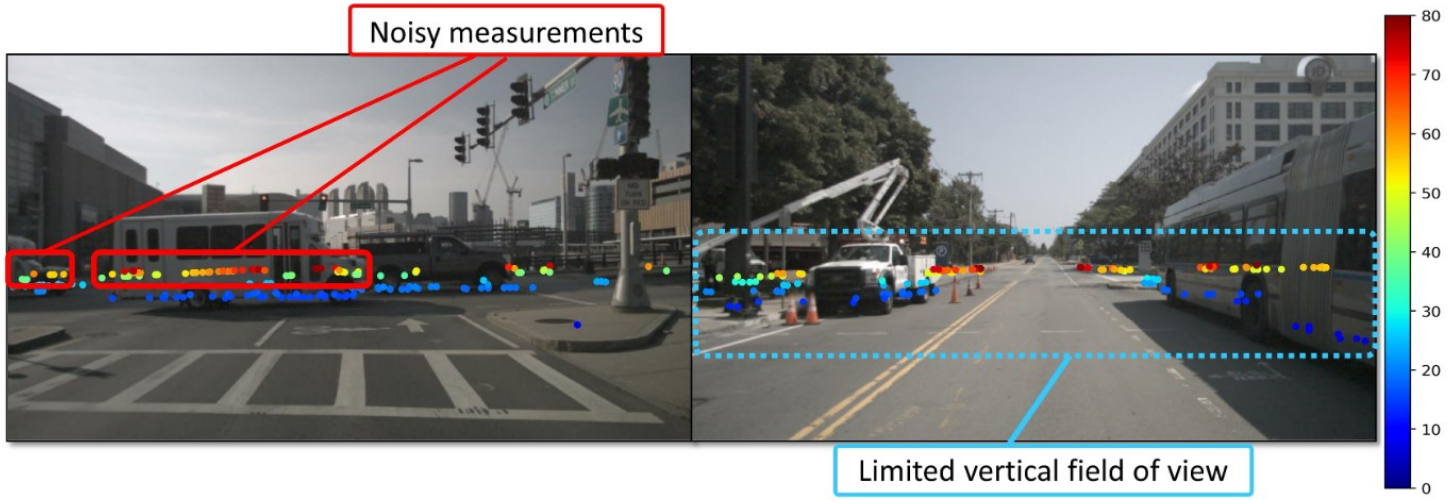
Method	Training data			AP
	Manual labeled	Auto labeled	Unlabeled	
CNN-auto * [1]	1351	6955	-	0.556
CNN-auto ** [1]	1351	6955	-	0.600
CNN-manual [1]	1351	-	-	0.513
Ours	1351	-	6955	**0.667**



Semi-supervised radar video fusion for VRU detection

PROBLEM STATEMENT

- 2D object detection using Radar + Vision
- nuScenes dataset:
  - 6 cameras, 5 radars, and 1 LiDAR, covered 360 degree
  - 1000 scenes (20s video per each)
  - Only 850 scenes is annotated (850 train, 150 test)
    - 765 for training, 85 for evalu
  - Radar point clouds:
    - Sparseness
    - Limited vertical field of view
    - Noise
    - Inconsistency with LiDAR

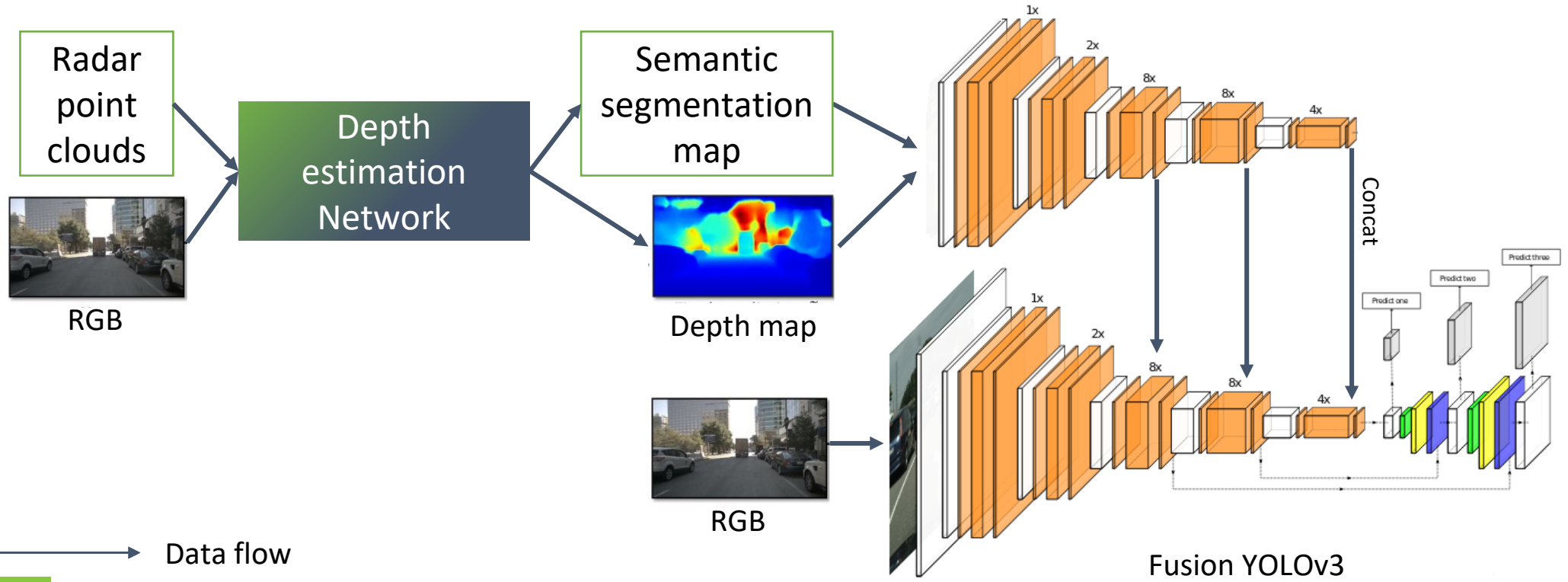


Challenge

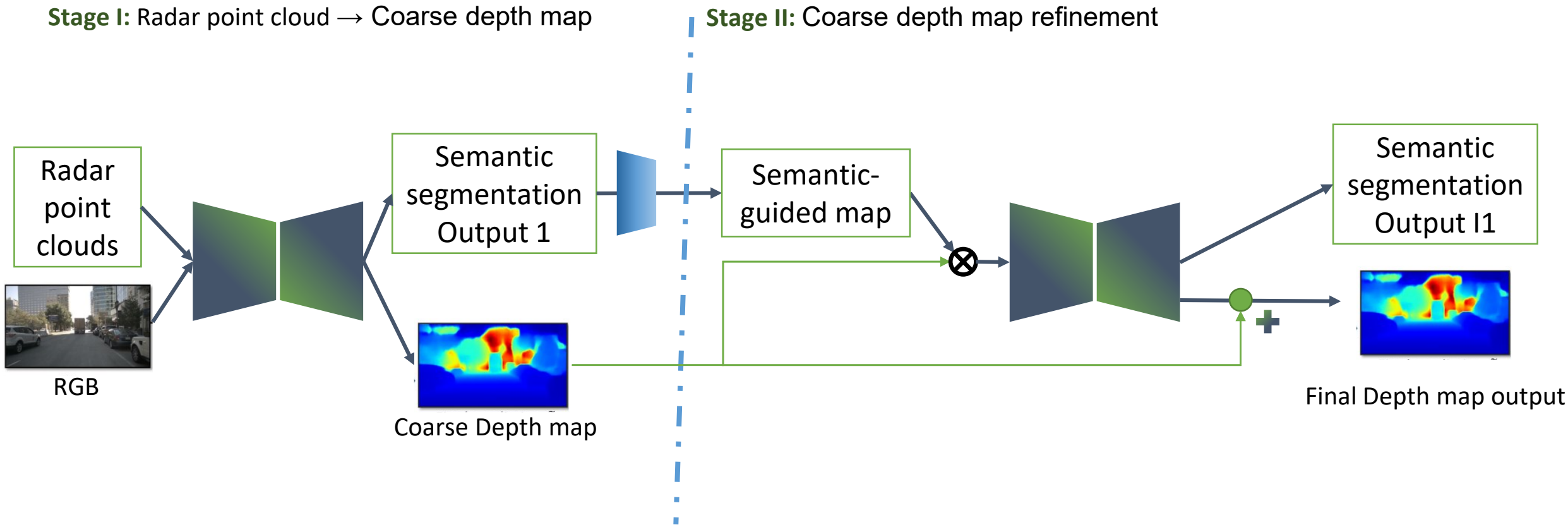
- What we have already known is…
  - Daytime and Nighttime performance
    - Daytime: only small or no space for improvement
    - Nighttime: radar can enhance the performance significantly
  - Detection range
    - LiDAR has much more points (1000) than radar (100)
    - But LiDAR's detection range (~80m) is shorter than radar (200-300m)

THE PROPOSED METHOD: RADAR + VISION

- Semantic-guided depth map estimation + Single stage object detection



THE PROPOSED METHOD: DEPTH ESTIMATION NETWORK



[A] Tao, Andrew, Karan Sapra, and Bryan Catanzaro. "Hierarchical multi-scale attention for semantic segmentation." *arXiv preprint arXiv:2005.10821* (2020).

DEPTH ESTIMATION NETWORK

Stage I: Radar point cloud → Coarse depth map

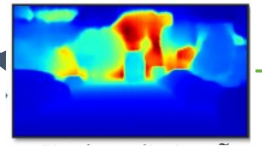
Stage II: Coarse depth map refinement

< 80m

Radar point clouds

RGB

Semantic segmentation Output 1



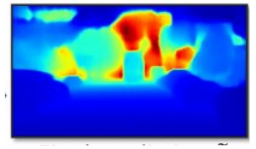
Coarse Depth map

RGB

> 80m

Radar point clouds

Semantic segmentation Output I1



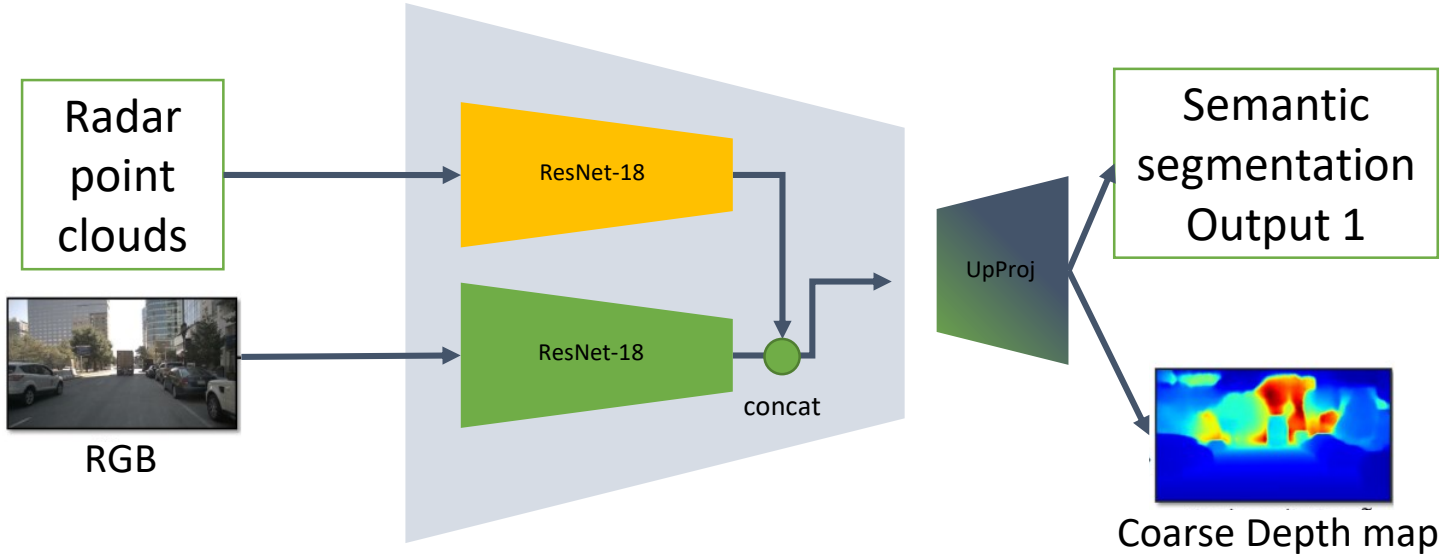
Final Depth map output

→ Data flow

[A] Tao, Andrew, Karan Sapra, and Bryan Catanzaro. "Hierarchical multi-scale attention for semantic segmentation." *arXiv preprint arXiv:2005.10821* (2020).

DEPTH ESTIMATION NETWORK

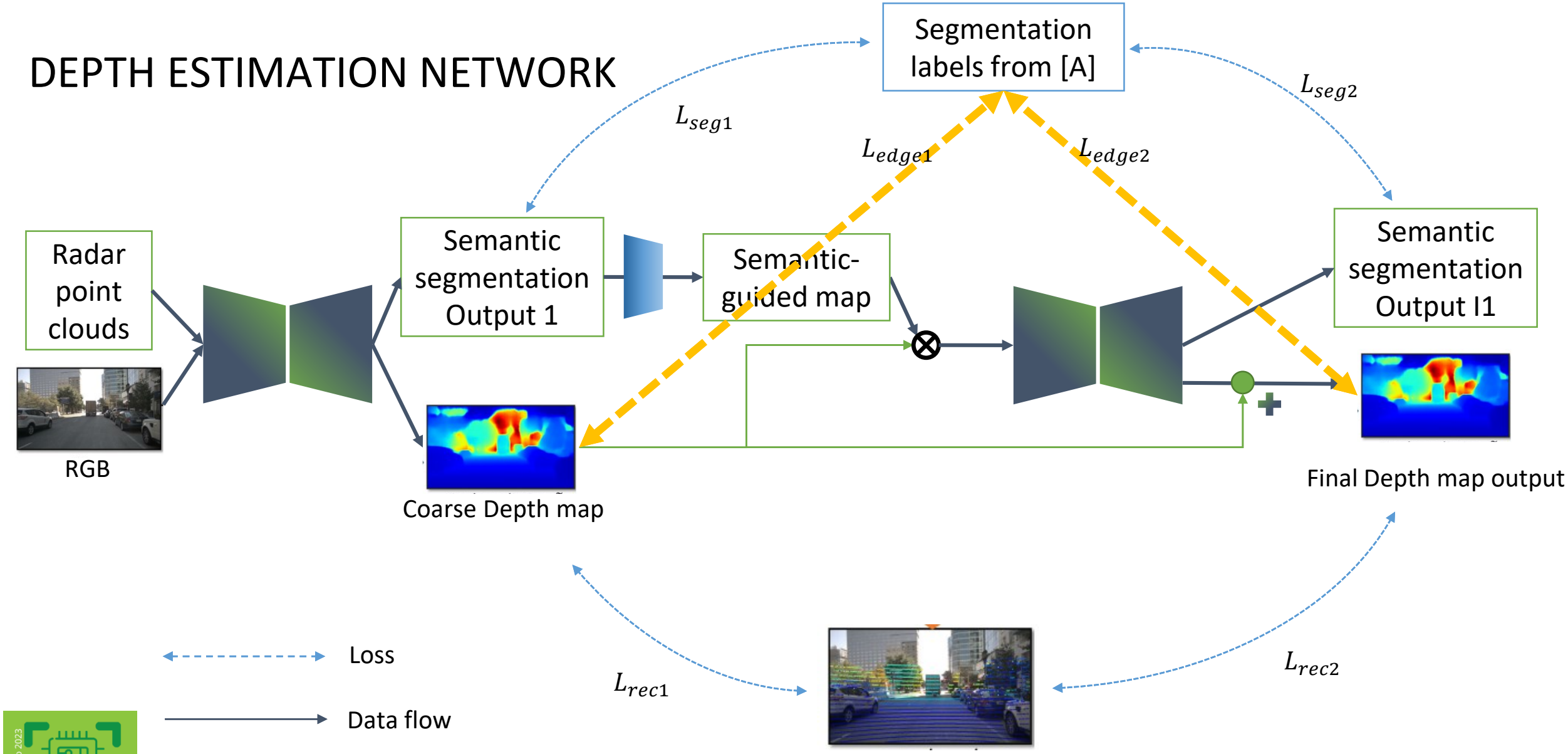
Stage I: Radar point cloud → Coarse depth map



→ Data flow

[A] Tao, Andrew, Karan Sapra, and Bryan Catanzaro. "Hierarchical multi-scale attention for semantic segmentation." *arXiv preprint arXiv:2005.10821* (2020).

DEPTH ESTIMATION NETWORK



OBJECTIVE FUNCTION

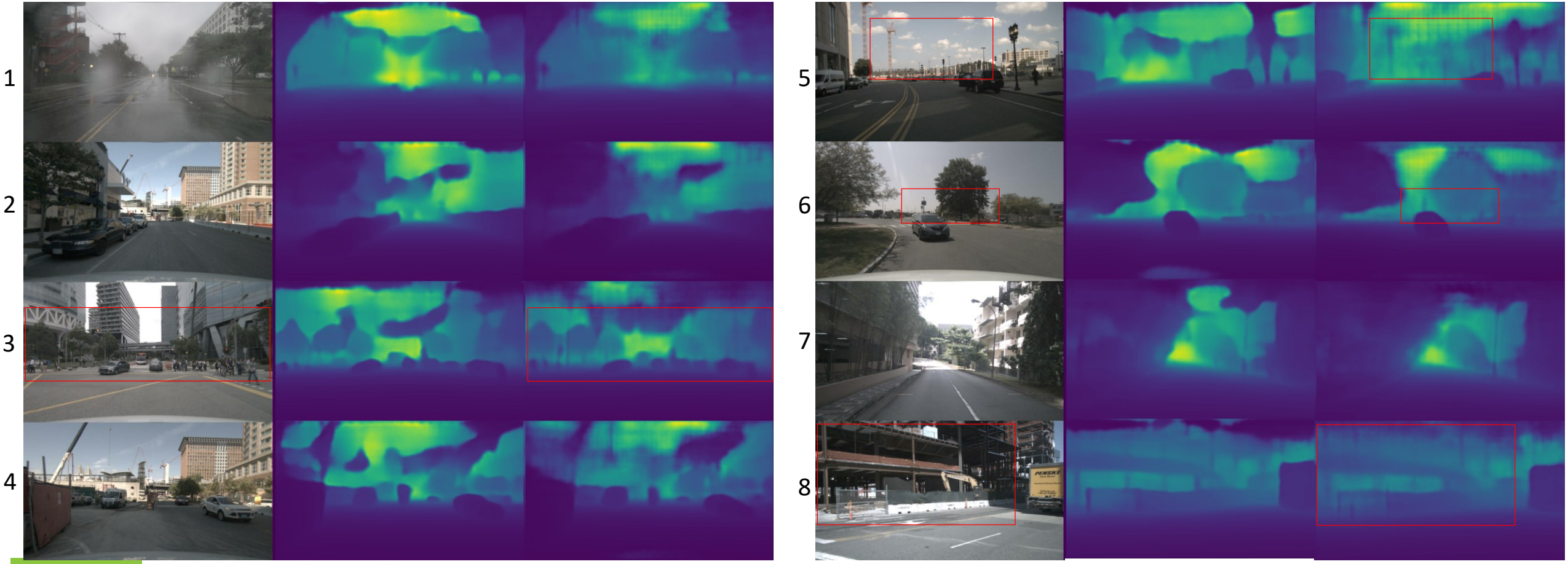
- Depth estimation Network
  - Given output depth map $\tilde{y}_d$, output semantic segmentation $\tilde{y}_s$
  - LiDAR ground truth y_d, segmentation ground truth y_s
  - L_{seg}: cross-entropy loss
  - L_{rec}: L1 loss
  - $L_{edge} = \left|\nabla_u(\tilde{y}_d)e^{-|\nabla_u(y_s)|}\right| + \left|\nabla_v(\tilde{y}_d)e^{-|\nabla_v(y_s)|}\right|$
  - $L_{stage1} = L_{seg1} + \alpha_1 L_{rec1} + \beta_1 L_{edge1}$.....(1)
  - $L_{stage2} = L_{seg2} + \alpha_2 L_{rec2} + \beta_2 L_{edge2}$.....(2)
  - $L_{depth} = e^{-w_1} \cdot L_{stage1} + e^{-w_2} \cdot L_{stage2} + \sum w_i$.....(3)
- Total loss
  - $L_{total} = L_{depth} + L_{yolo}$

EXPERIMENTAL RESULTS

Method	Input data	RMSE	MAE
RGB only	RGB	5.613	0.050
Sparse-to-dense [4]	RGB + radar	5.628	0.047
PnP [5]	RGB + radar	5.406	0.045
CSPN [6]	RGB + radar	5.566	0.048
Lin, J. T. et al [3] (IROS2020)	RGB + radar	**5.18**	**0.040**
Ours	RGB + radar	5.24	0.041

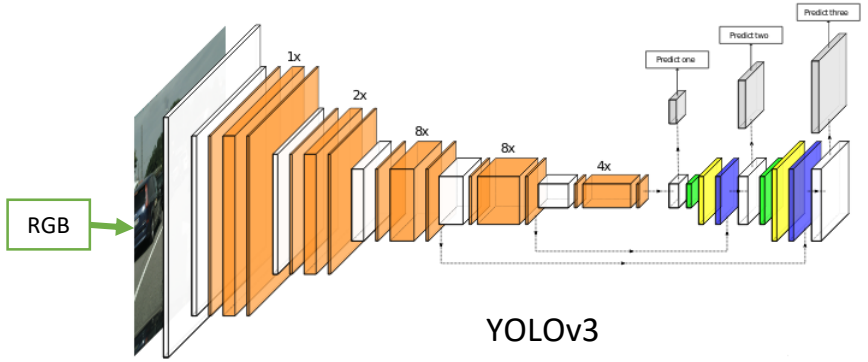
[3] Lin, J. T., Dai, D., & Van Gool, L., "Depth estimation from monocular images and sparse radar data," *IROS2020*.
[4] F. Mal and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," *ICRA2018.*
[5] T.-H. Wang, F.-E. Wang, J.-T. Lin, Y.-H. Tsai, W.-C. Chiu, and M. Sun, "Plug-and-play: Improve depth estimation via sparse data propagation," *ICRA2018.*
[6] X. Cheng, P. Wang, and R. Yang, "Depth estimation via affinity learned with convolutional spatial propagation network," *ECCV2018.*

EXPERIMENTAL RESULTS

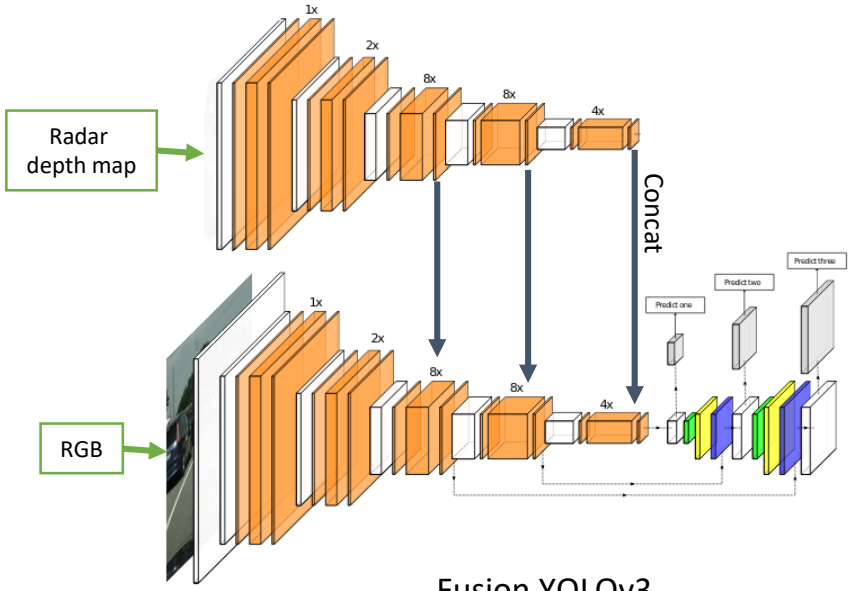


EXPERIMENTAL RESULTS

Method	Network	mAP	mAP@0.50
RRPN [1]	Faster R-CNN	35.45	59.00
Nabati et al. [2]	Faster R-CNN	35.60	60.53
RGB only	YOLOv3	34.69	64.62
RGB + Radar depth map [3] (concat at 4th channel)	YOLOv3	32.10	60.70
RGB + Radar depth map [3]	Fusion YOLOv3	34.95	64.71
RGB + Radar depth map (Ours) (stage I coarse depth map)	Fusion YOLOv3	**35.28**	**65.20**
RGB + Radar depth map (Ours) (stage II final depth map)	Fusion YOLOv3	**35.34**	**65.44**
RGB + Radar depth map + Semantic segmentation map (Ours)	Fusion YOLOv3	-	-



YOLOv3



Fusion YOLOv3

[1] R. Nabati and H. Qi, "Rrpn: Radar region proposal network for object detection in autonomous vehicles," *ICIP2019*.
[2] R. Nabati and H. Qi, "Radar-Camera Sensor Fusion for Joint Object Detection and Distance Estimation in Autonomous Vehicles," *arXiv 2020*.
[3] Lin, J. T., Dai, D., & Van Gool, L., "Depth estimation from monocular images and sparse radar data," *IROS2020*.

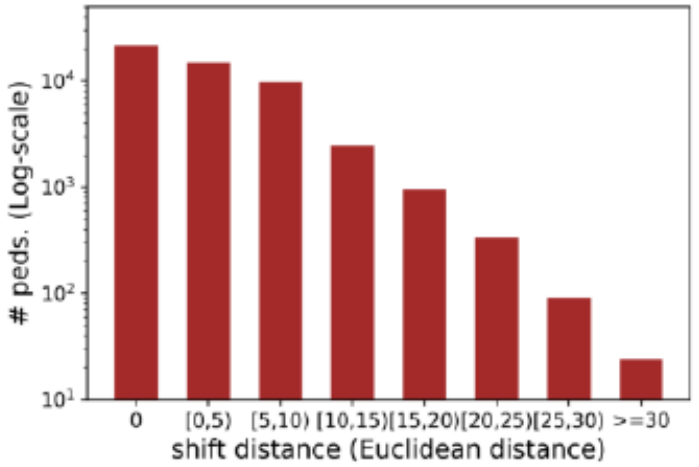
CONCLUSIONS

- Radar point cloud depth map completion
  - First work that introducing radar point cloud and semantic information to estimate dense depth map
  - Depth map containing labeled (< 80m) and unlabeled targets (> 80m)

- Radar + Vision fusion (Fusion YOLOv3)
  - A novel framework that fusing semantic information & radar depth map in multi-scale
  - Suitable for single stage and two stage object detection framework*

Thermal video fusion for VRU Detection

PROBLEM STATEMENT: THERMAL + VISION



- 2D pedestrian detection using Thermal + Vision
- KAIST dataset:
  - 95,328 color & thermal image pairs (512x640)
  - 103,128 annotations
  - 1,182 unique pedestrians
  - Covered day & night scenes
- Subsets:
  - Occlusion: none, partial, heavy
  - Distance: near, medium, far
- Known issue:
  - Unpaired & Poor alignment

ARCHITECTURE

Inference stage

RGB images

Thermal images

Feature maps

VGG16

concat

Cross-Modality Self-Attention Transformer

concat

VGG16

Feature maps

Modality-specific features

Modality-specific features

Region Proposal Network (RGB)

Region Proposal Network (Thermal)

ROI Pooling

ROI Pooling

ROI pairs

Multimodal Fusion Transformer

Classification Results

Bounding Box Results

Two-stream feature extractor

Modality-specific region proposals

Multimodal fusion

Experiment results
ABLATION STUDY

(a) Detection results w/o CAT (b) Detection results w/ CAT (c) Features w/o CAT (d) Features w/ CAT

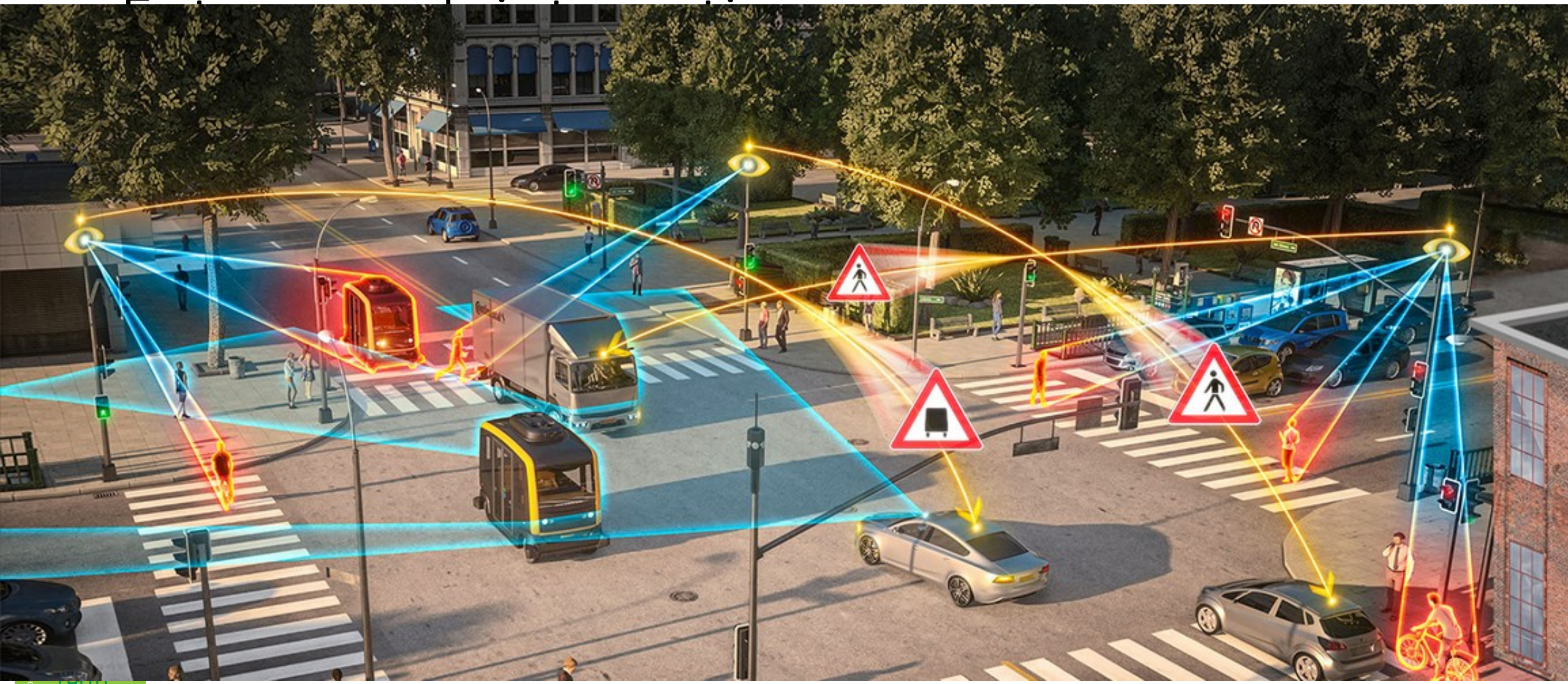
* CAT: Cross-Modality Attention Transformer

EXPERIMENTAL RESULTS

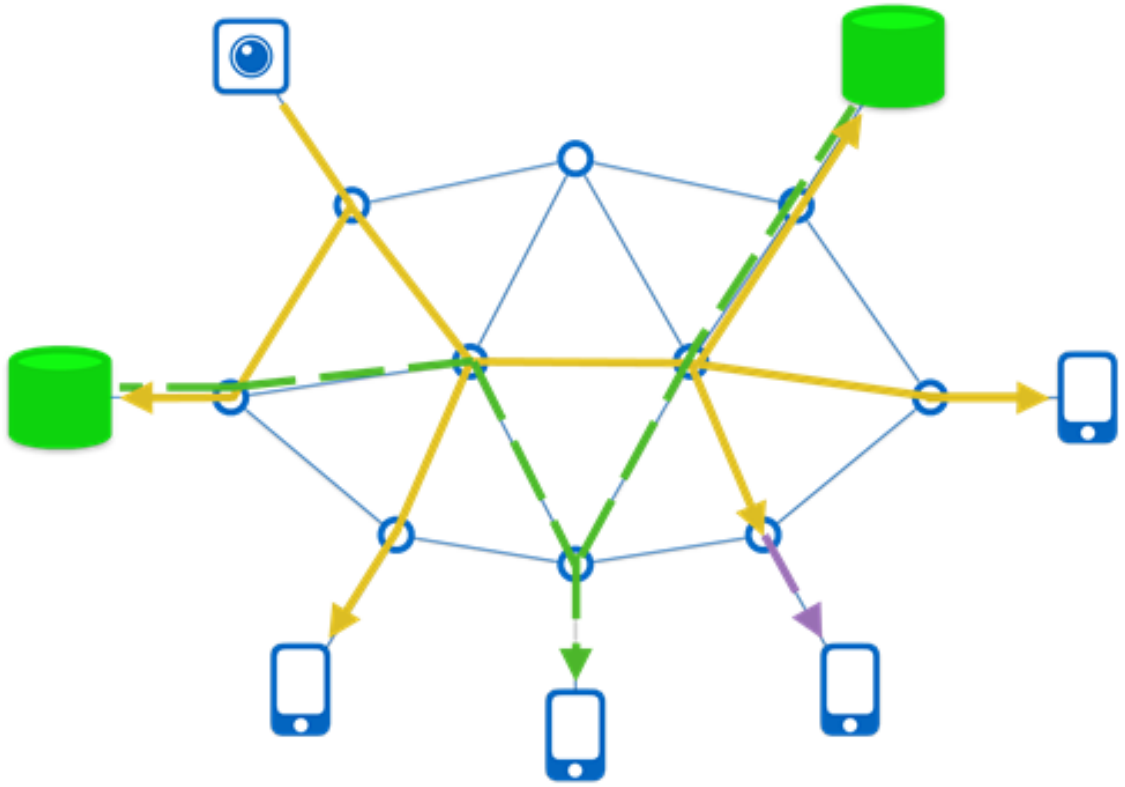
Method	Input	MR (partial occlusion)	MR (heavy occlusion)
Faster-RCNN	RGB	48.59	
Faster-RCNN	Thermal	47.59	
Faster-RCNN	RGB + Thermal	25.78	
Halfway fusion [1]	RGB + Thermal	25.75	
RPN + BDT [2]	RGB + Thermal	29.83	
IAF-RCNN [5]	RGB + Thermal	15.73	
IATDNN + IAMSS [9]	RGB + Thermal	14.95	
MSDS-RCNN [3]	RGB + Thermal	11.63	
GAFF [7]	RGB + Thermal	10.62	
AR-CNN [6]	RGB + Thermal	**9.34**	**30.76**
Ours	RGB + Thermal	**8.02**	**28.74**

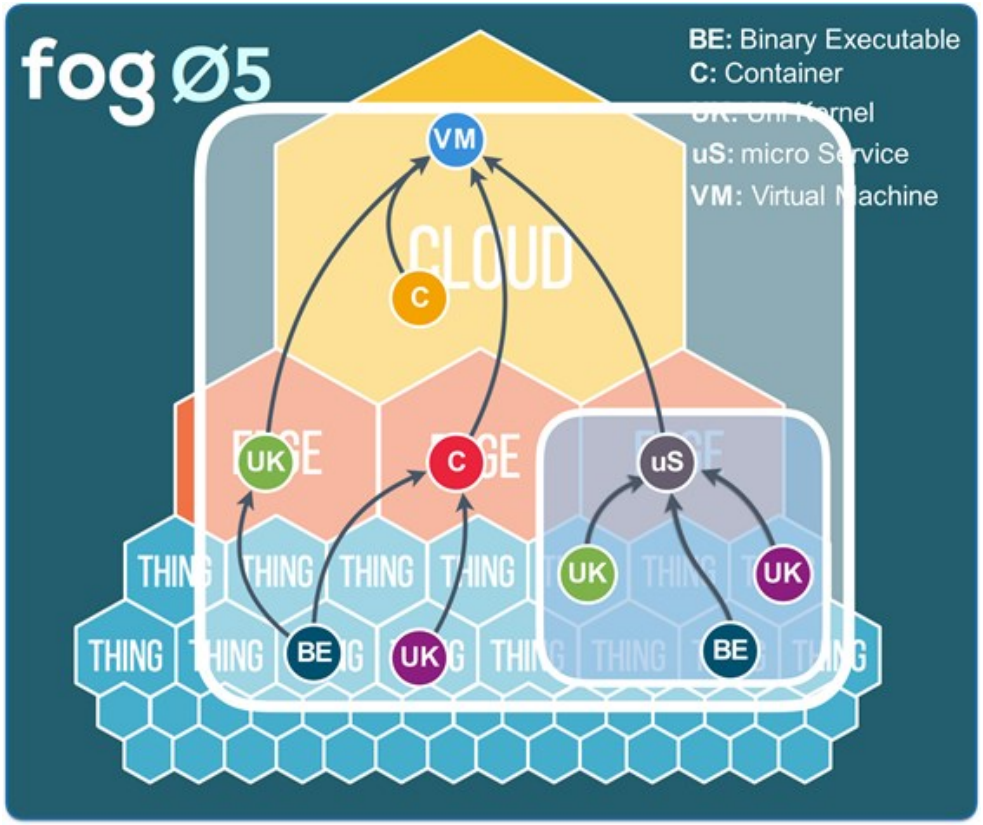
[1] Liu, Jingjing, et al. "Multispectral deep neural networks for pedestrian detection." *BMVC,* 2016.
[2] Konig, Daniel, et al. "Fully convolutional region proposal networks for multispectral person detection." CVPRW, 2017.
[3] Li, Chengyang, et al. "Multispectral pedestrian detection via simultaneous detection and segmentation." BMVC, 2018.
[5] Li, Chengyang, et al. "Illumination-aware faster R-CNN for robust multispectral pedestrian detection." Pattern Recognition, 2019.
[6] Zhang, Lu, et al. "Weakly aligned cross-modal learning for multispectral pedestrian detection." ICCV, 2019.
[7] Zhang, Heng, et al. "Guided attentive feature fusion for multispectral pedestrian detection." WACV, 2021.
[9] Guan, Dayan, et al. "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection." Information Fusion, 2019.

Future work-integration

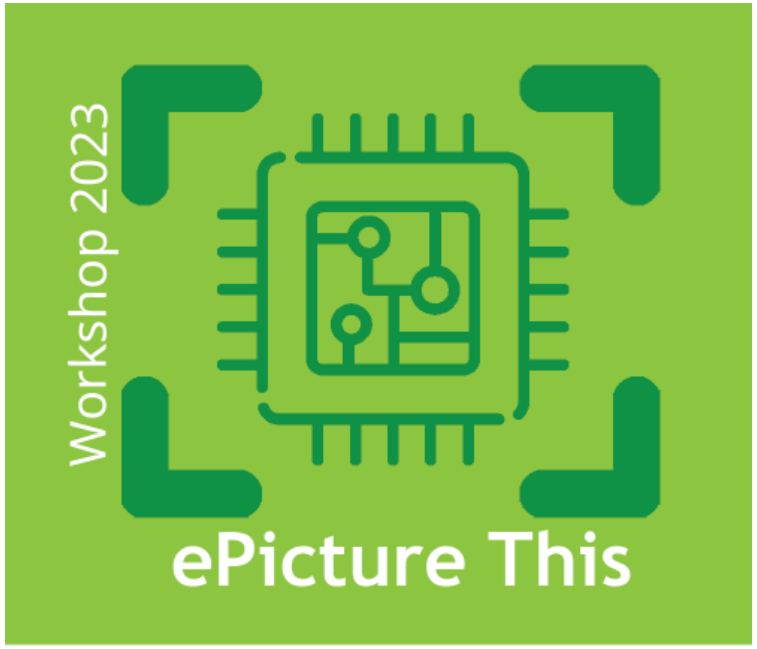


FINAL STEP: INTEGRATION





QUESTIONS



an initiative by PENTA label projects
MANTIS and IMAGINATION with AENEAS support

THANK YOU

Penta Xecs EURIPIDES² European Smart Electronic Systems Aeneas