Robust Driver Head Pose Estimation in Naturalistic Conditions from Point-Cloud Data

Tiancheng Hu, Sumit Jha and Carlos Busso

TEXAS ANALOG CENTER OF EXCELLENCE

THE UNIVERSITY OF TEXAS AT DALLAS

UT Dallas MSP Multimodal Signal Processing Laboratory

Motivation

- **Head pose estimation is an important task**
  - With applications in areas including
    - advanced driver assistance system [Murphy-Chutorian et al. 2007]
    - visual attention modelling [Ba and Odobez 2009]
    - gaze estimation system [Zhang el al. 2015]
- **In the automotive domain, it is a challenging task due to**

(a) Occlusion (b) Extreme head pose (c) Illumination

Time-of-Flight Depth Camera

- **Time-of-flight camera**
  - Utilize active infrared lighting
  - Calculate distance between camera and object based on round trip time
  - Immune to illumination change!
- **We adopt a pico flexx camera, which provides**

point cloud data

grayscale image

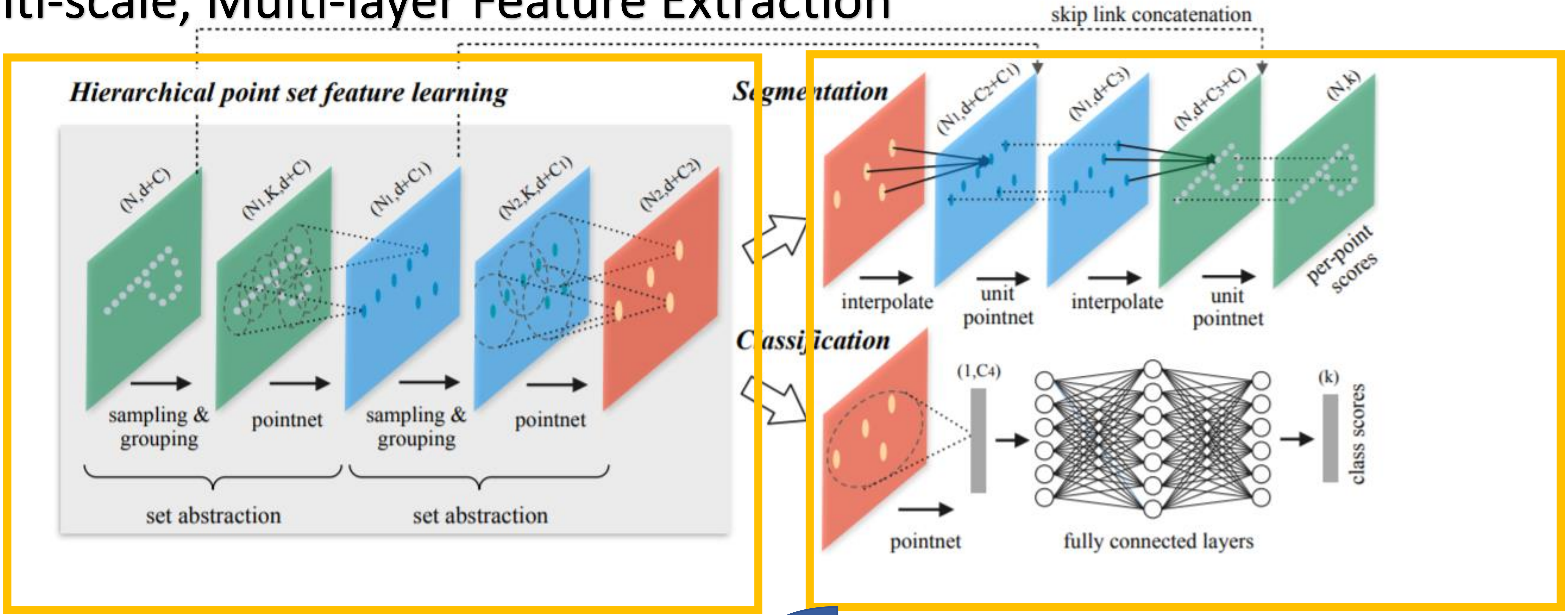
Image credit:
https://pmdtec.com/picofamily/flexx/

Related Work

- **Point Cloud Processing**
  - Wu et al. (2015) represents point clouds as 3D voxel grids and use 3D CNNs to process them
  - Su et al. (2015) renders 2D images from point cloud at different angles and process the set of 2D images using CNNs
  - PointNet [Qi et al. 2017 ICCV] and PointNet++ [Qi et al. 2017 NIPS] directly process 3D point-cloud data without converting to any other intermediate representation
- **Driver Head Pose Estimation**
  - Borghi et al. (2017) utilizes a multi-modal approach, with CNNs trained on RGB, depth map, and optical flow data which are then fused to predict head pose
  - Schwarz et. al (2017) proposes a CNN based model which fuses information from infrared images and depth maps and regresses head pose

PointNet++ [Qi et al. 2017 NIPS]

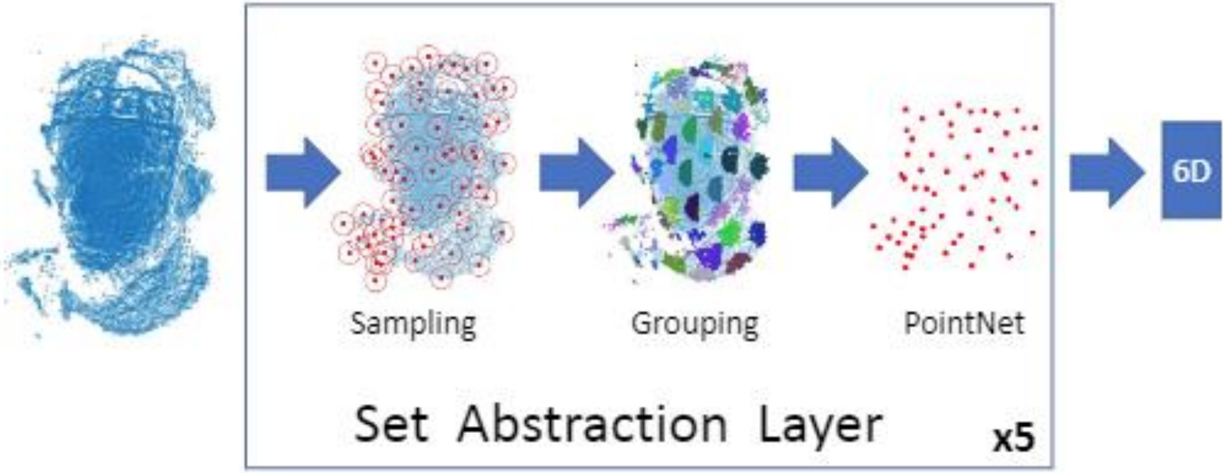
Multi-scale, Multi-layer Feature Extraction



Task specific layers

Model – Directly Process Point Cloud Data

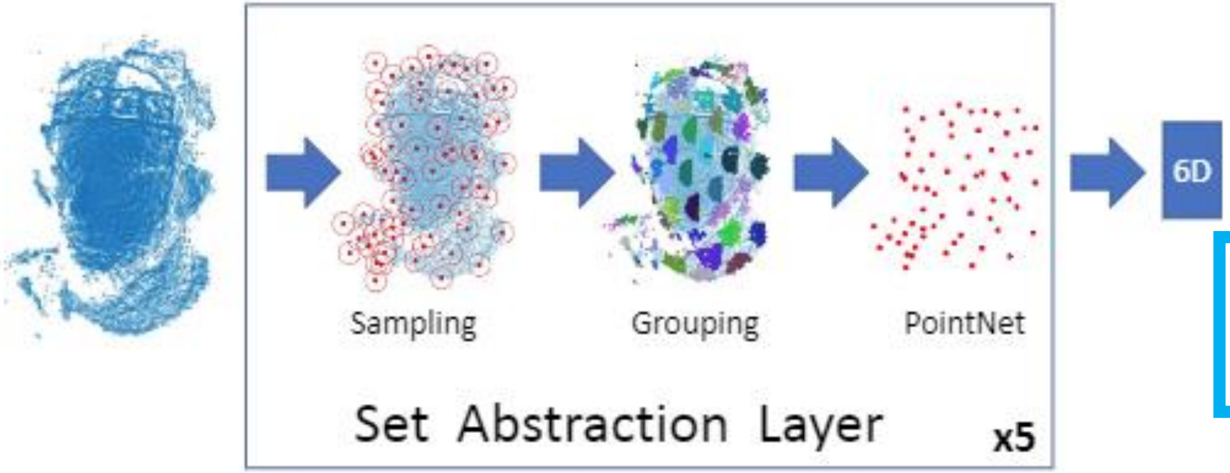
- **Set Abstraction Layer**
  - **Sampling**: iterative farthest point sampling, resulting in ***N*** "anchor points"
    - Motivation: Point clouds are usually large and contain redundant points
    - Goal: Decrease redundancy while maintaining useful information
  - **Grouping**: group points within a radius ***R*** of the "anchor points"
    - Motivation: CNN captures the local features of a neighborhood
    - Goal: Capture the relationship between anchor points and the neighborhood



Represent rotation in 6D, according to Zhou et al. 2019

Model – Directly Process Point Cloud Data

- **Set Abstraction Layer**
  - **PointNet**: multi-layer perceptron to lift the feature up to a higher dimension
    - Motivation: Just as many deep learning networks, we need to extract high-level feature
    - Goal: find a discriminative feature representation
  - Each set abstraction layer has different ***N*** and ***R*** values to capture features of different scale
  - Multiple set abstraction layer stacked together to extract high-level feature



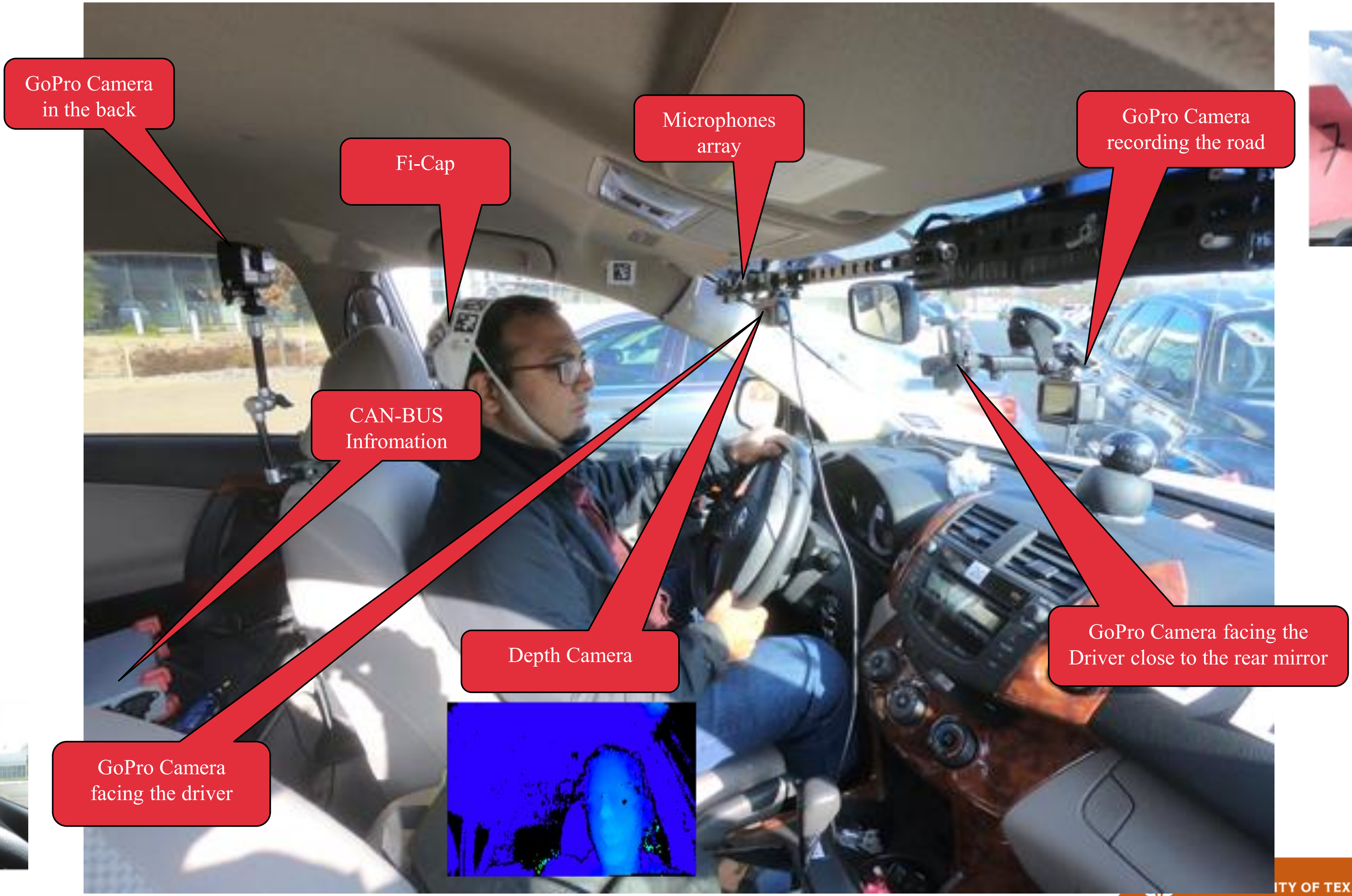
Represent rotation in 6D, according to Zhou et al. 2019

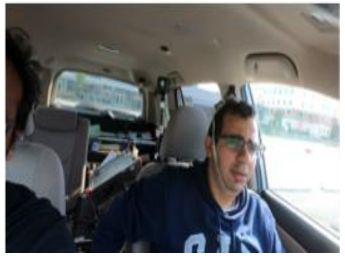
-0.4879	-0.3358	-0.8057
-0.8123	-0.1632	0.5600
-0.3195	0.9277	-0.1931

Dataset

- **Multimodal Driver Monitoring (MDM) Dataset**
  - 4 GoPro RGB cameras, 1 pico flexx depth camera
  - Naturalistic driving with a diverse range of head poses
  - Head pose labels provided by Fi-Cap [Jha and Busso 2018]
  - 59 subjects (27 male 32 female, mostly college students)
    - Used 22 in this study, total duration 17 hours 39 minutes

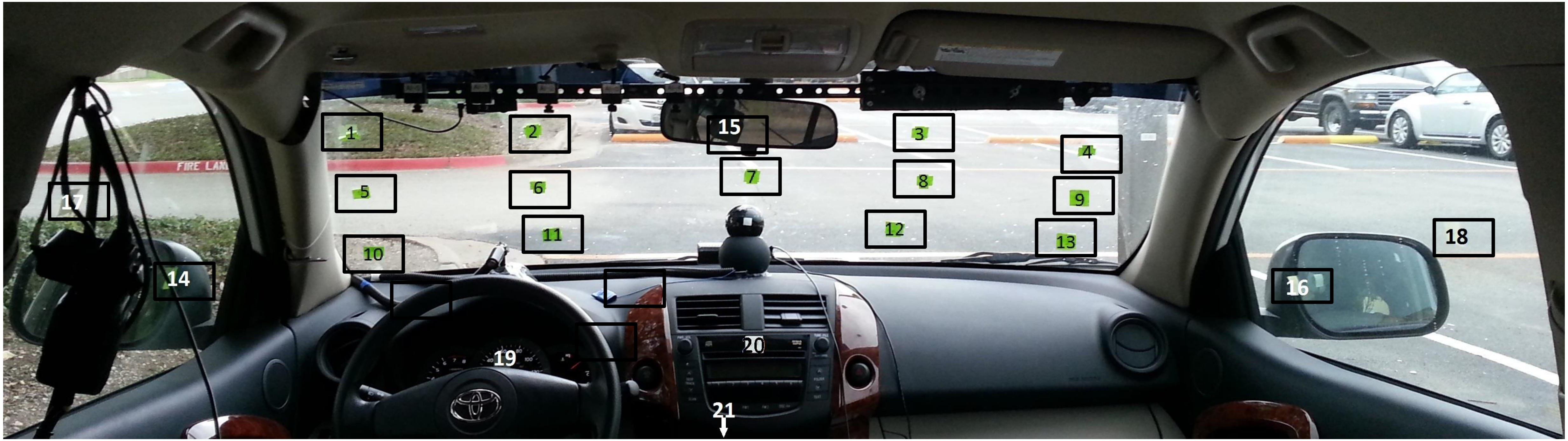
Setup - Sensors





Example data from different sensors

Setup - Markers



Protocol – Phase 1

- **Phase 1: Natural Gaze – (Parked Vehicle)**
  - Subject asked to look at target markers on the windshield in random order
  - Subject asked to look at a trackable marker that the researcher moves around in front of the car

Protocol – Phase 2

- **Phase 2: Natural task – Driving**
  - Subject asked to follow navigation on a smart phone
    - Multiple destinations in sequence
  - Subject asked to change radio channels when driving

Protocol – Phase 3

- **Phase 3: Natural Gaze – Driving**
  - Subject asked to look at landmarks on the road and answer questions
  - Subject asked to look at points on the windshield

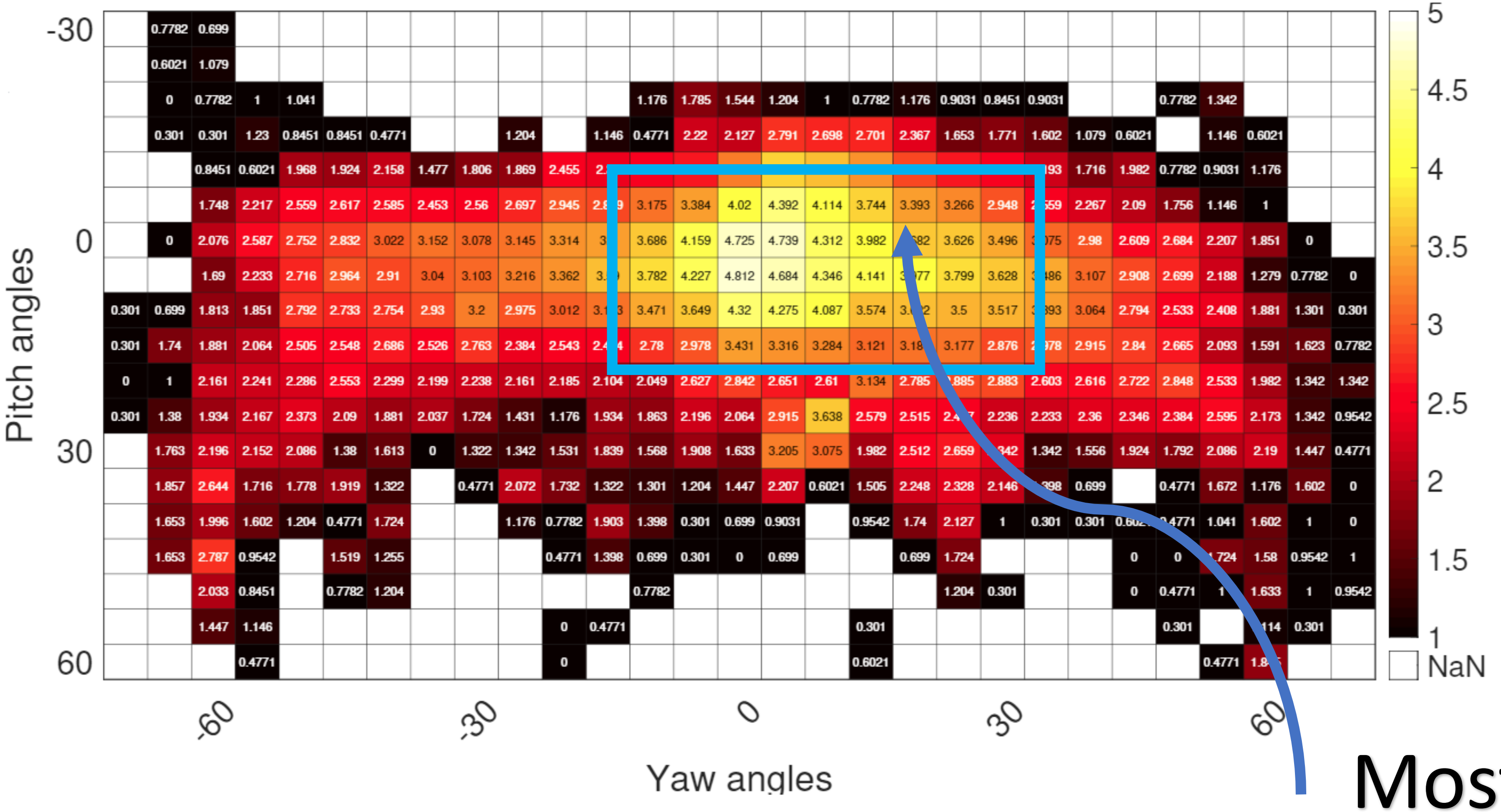
Training Details

- **Point Cloud Preprocessing**
  - Distance-based filtering -> grid-based sampling -> 5000 points -> normalized (centroid at (0,0,0), all points in unit sphere)
- **Rotation Representation**
  - Represent rotation in 6D, according to Zhou et al. 2019
  - Easily convertible to full rotation matrix
- **Training Detail**
  - Train:14 subject; development: 4 subject; test: 4 subject
  - Adam optimizer, learning rate = 0.001 with learning rate decay of 0.7 per 2 million steps
  - L2 loss

Baseline Approaches

- **OpenFace 2.0** [Baltrušaitis et al. 2018]
  - State-of-the-art (SOTA) toolkit for face analysis, including head pose estimation
- **Face Alignment Network (FAN)** [Bulat and Tzimiropoulos, 2017]
  - One of the SOTAs for facial landmark estimation
  - Use singular value decomposition to get rotation from landmarks
- **For Both**
  - Full resolution (1920x1080) RGB image captured from GoPro is used as input
  - To avoid difference in angle definition:
    Subject-wise transformation applied between baseline prediction and ground truth

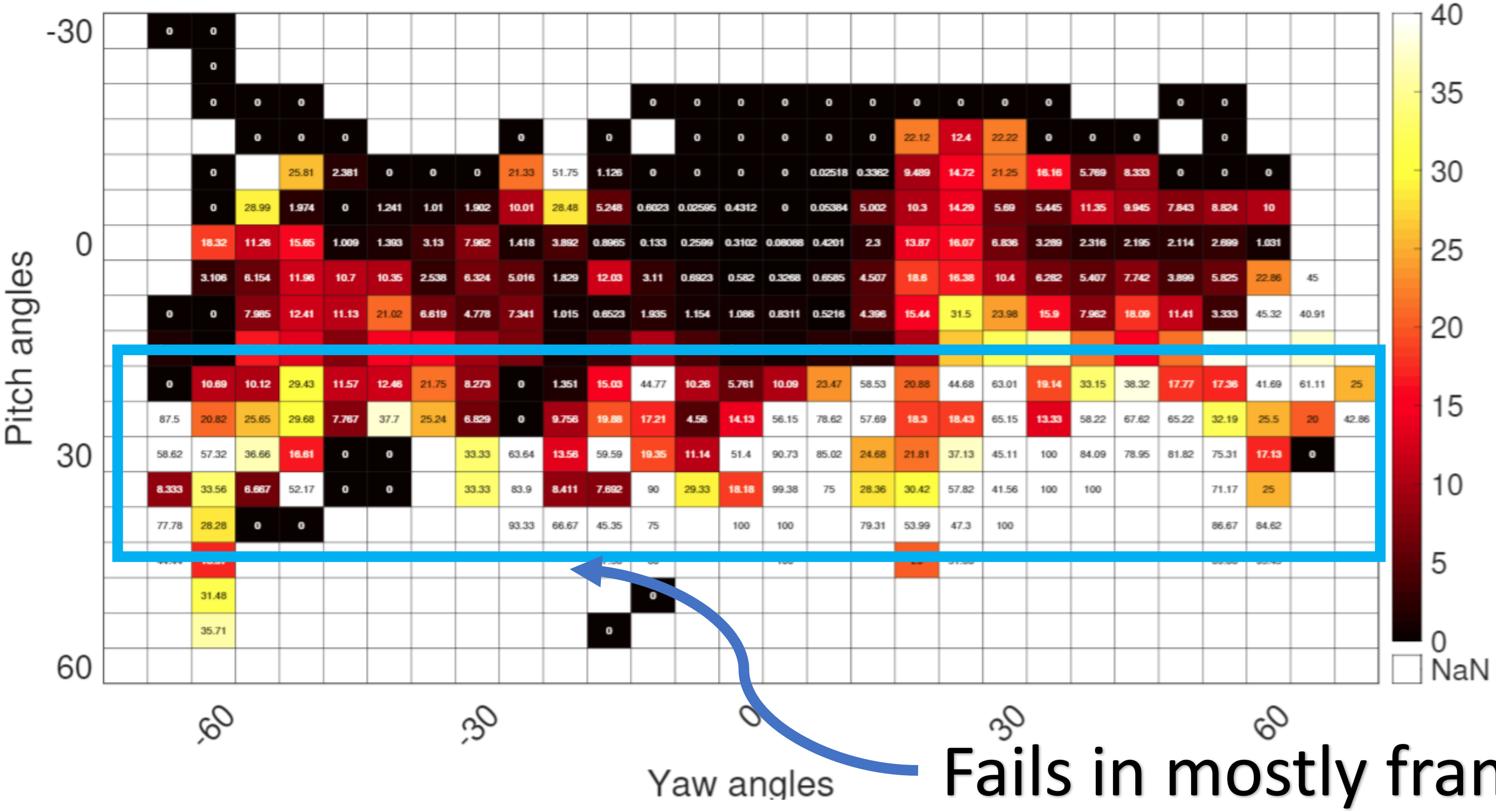
Result – Test Set Ground Truth Distribution



log10 scale

Most Data in this region

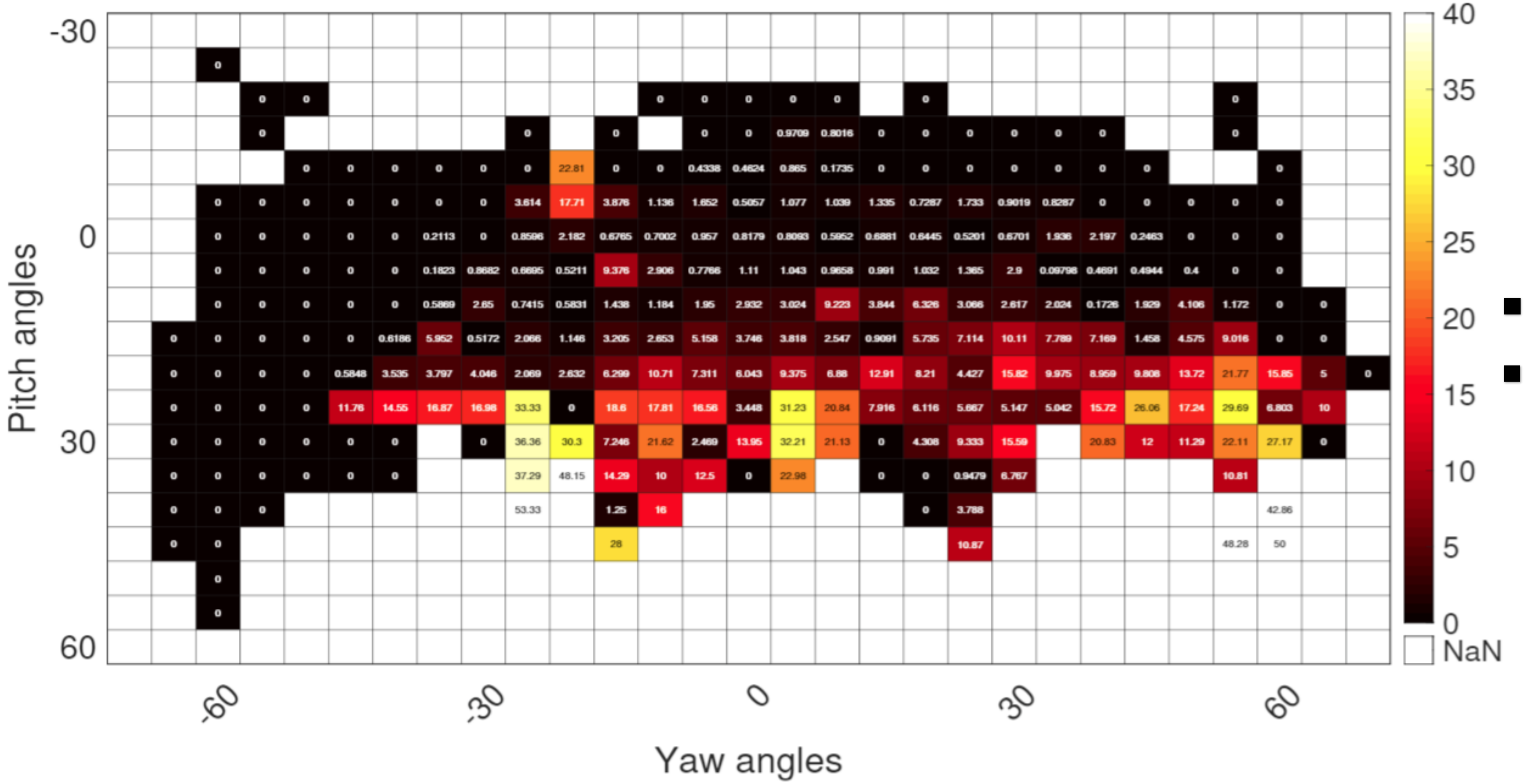
Result – OpenFace 2.0 Detection Failure



4.8%

Fails in mostly frames with larger rotation

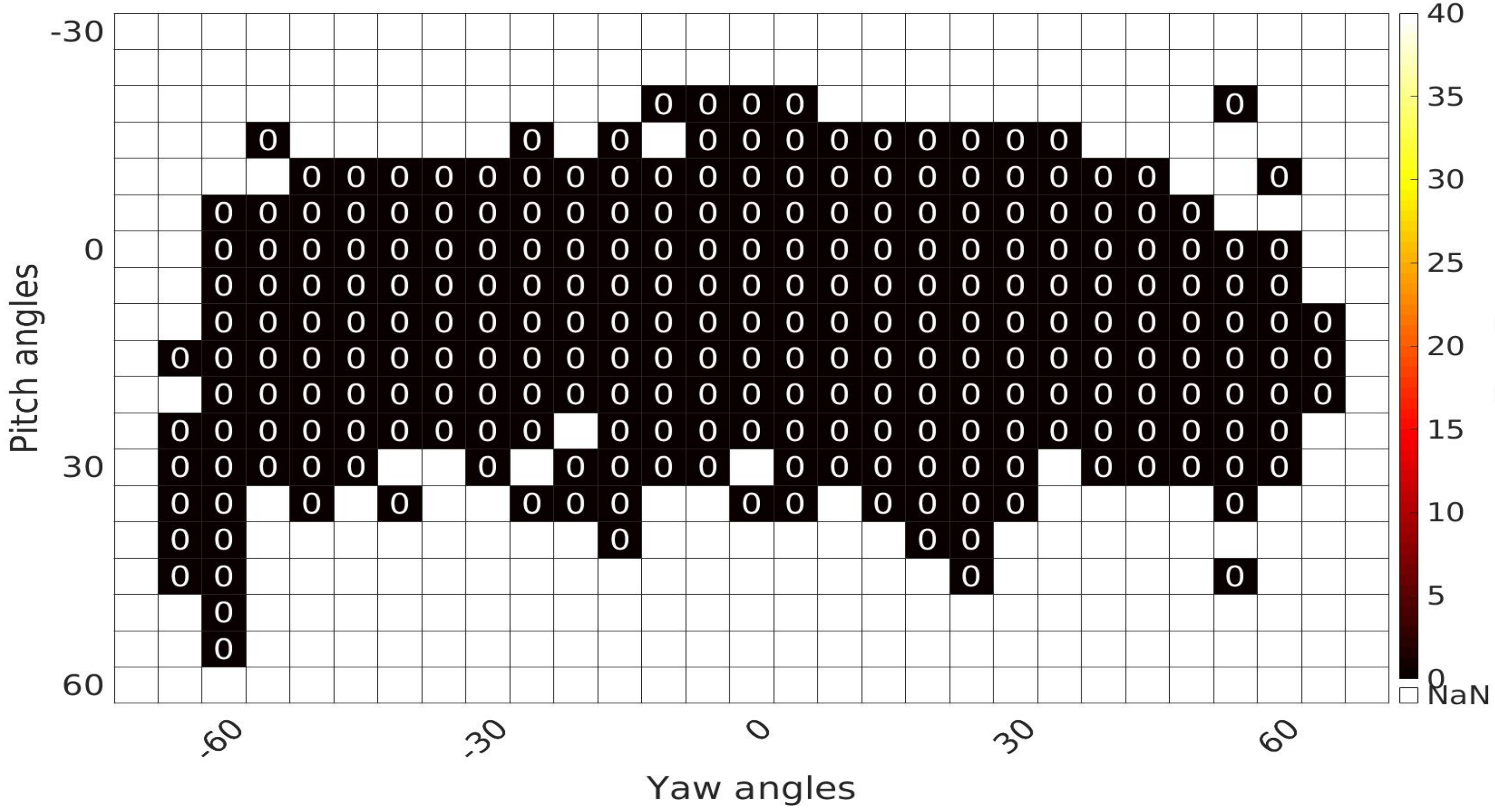
Result – FAN Detection Failure



2.1%

- Better than OpenFace
- Failure also concentrated in larger ground truth rotation

Result – Proposed Method

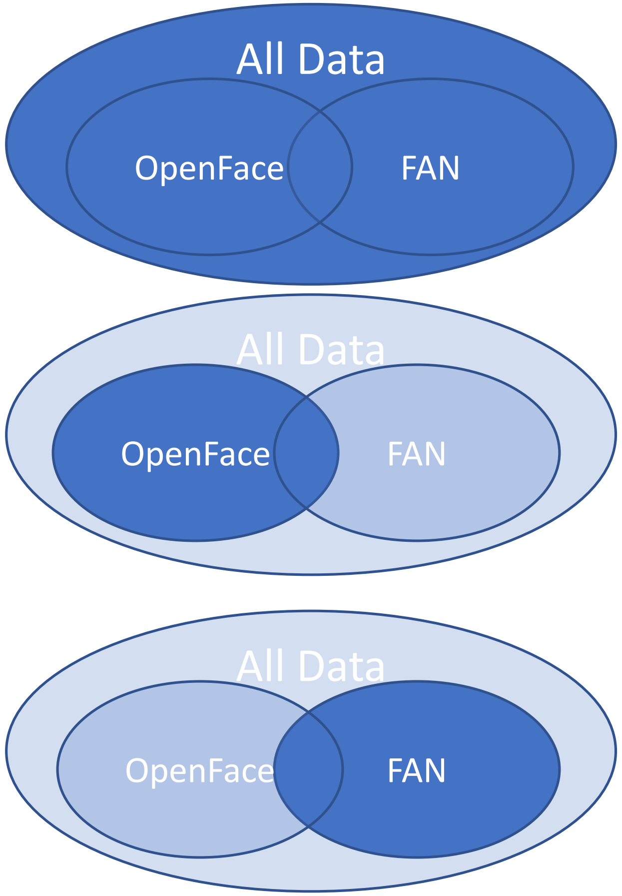


0%

- No head detection needed
- Thus 0% detection failure

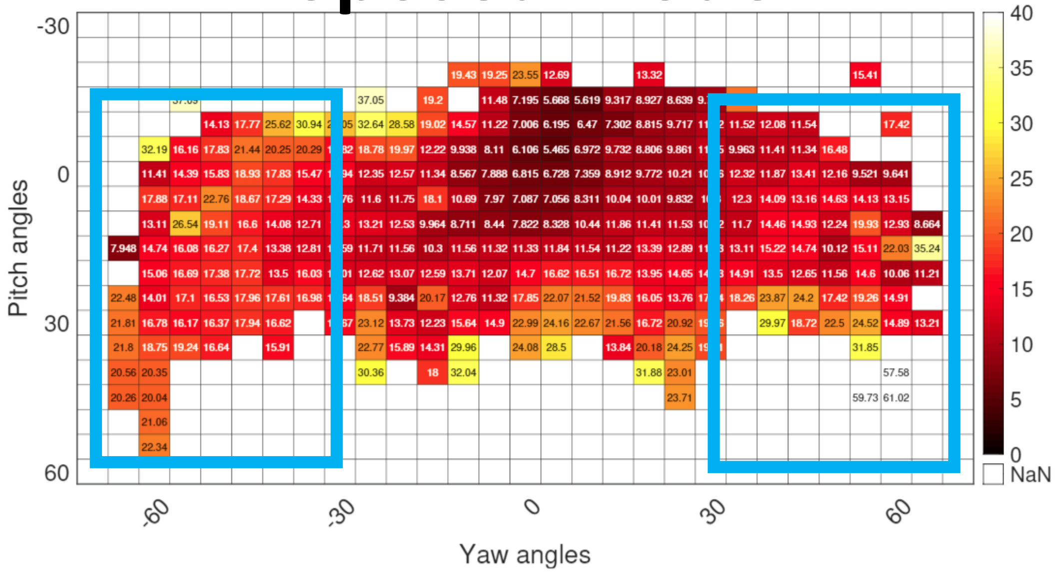
Results – Mean Squared Error

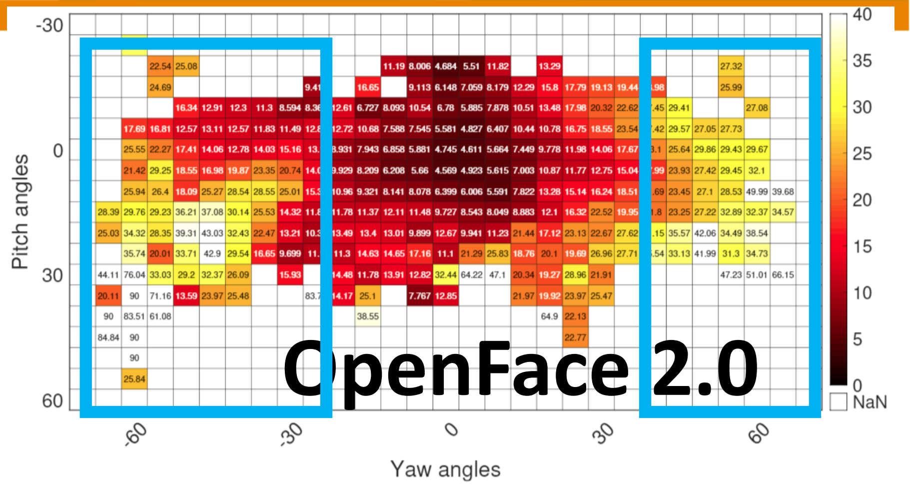
	Errors (Model Set)		
	Roll(°)	Yaw(°)	Pitch(°)
Proposed Model	5.91	7.32	6.68
	Errors (OpenFace Set)		
	Roll(°)	Yaw(°)	Pitch(°)
Proposed Model	**5.48**	7.15	**6.39**
OpenFace 2.0	9.32	**6.21**	8.42
	Errors (FAN Set)		
	Roll(°)	Yaw(°)	Pitch(°)
Proposed Model	**5.66**	**7.24**	**6.53**
FAN	11.30	19.28	8.47

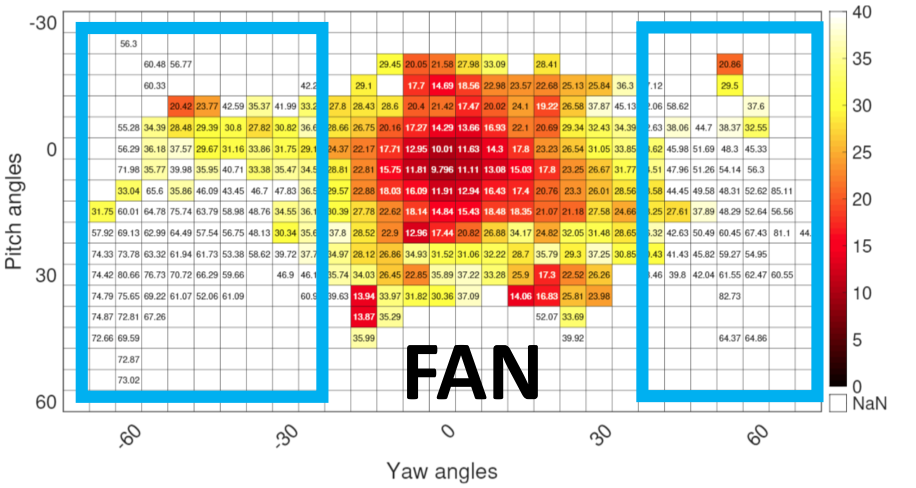


Result – Result Comparison – Geodesic Distance

Proposed Model





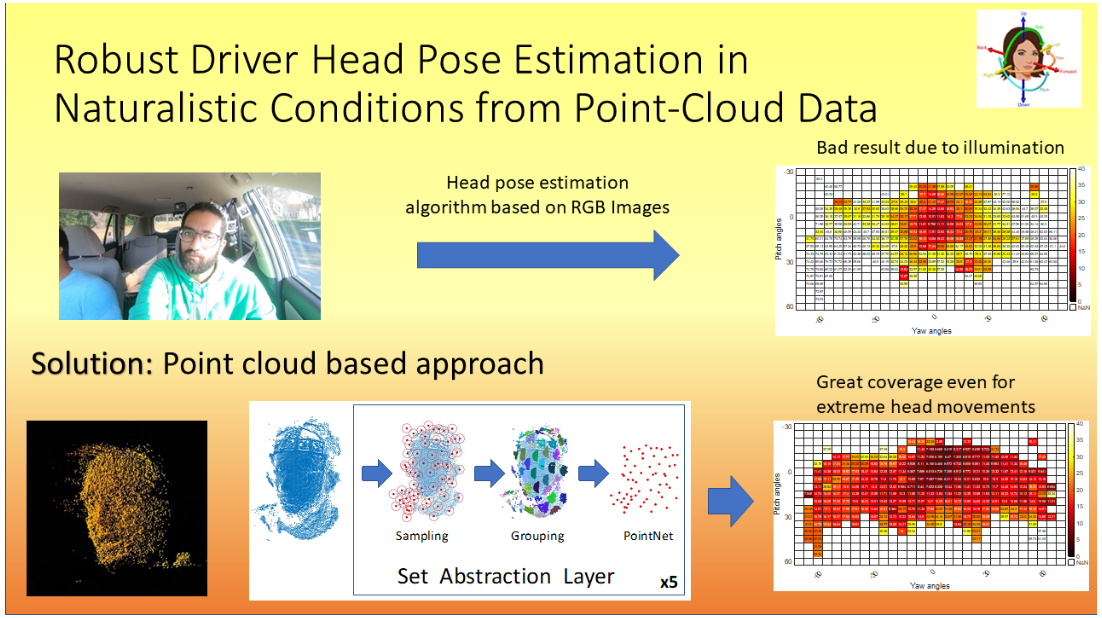


- **Proposed model has lower error overall**
- **Error especially lower in large yaw rotations**

$$\Delta(R_1, R_2) = \| \log(R_1 R_2^T) \|$$

Conclusion & Future Work

- **Ours: 1st deep learning based head pose estimation algorithm directly from point cloud**
- **Evaluated on Multimodal Driver Monitoring dataset**
- **Achieve better performance than the baselines**
- **Model reliable overall, especially in large rotations**
- **Future Work**
  - Multimodal approach (depth, RGB and more) to jointly model head pose
  - Temporal modelling for driver head pose
  - Build a more parameter-efficient model for real-time applications



Thank you

- This work was supported by Semiconductor Research Corporation (SRC) / Texas Analog Center of Excellence (TxACE), under task 2810.014.

Our Research: msp.utdallas.edu