Synthetic Skeletal Pose Pre-training to Mitigate Data Scarcity in In-Cabin 2D-to-3D Pose Lifting

Thummanoon Kunanuntakij, Dominik Schörkhuber, Margrit Gelautz
Computer Vision Lab, TU Wien

Motivation

- Driver-related factors cause ~87.7% of traffic accidents [1]. DMS are now mandatory in new EU vehicles since 2024.
- 3D pose estimation is valuable but bottlenecked by annotation cost. 3D labels require calibrated multi-view triangulation, unlike 2D poses which can be labeled from single images.
- Synthetic data is the proposed solution.

Datasets

- **Real: Drive&Act** [2]
  - Near Infrared images from the *center_mirror* view was used.
  - 3D labels via triangulation.
  - 15 subjects: #1-#8 for training, #10 for validation, #11-14 for testing.
- **Synthetic: SimulatedCabin IR 1M** [3]
  - 1M poses, generated via inverse kinematics, three front-facing views used (*OMS_01, Dashboard, Front*).
  - Both use COCO-style 13 upper-body keypoints, neck-centered.

Methods

- Three-stage pipeline:
  Driver Localization: **Faster R-CNN**[4]
  ↳ 2D Pose Estimation: **HRNet** [5]
  ↳ 3D Pose Estimation: **2D-to-3D Lifter**
- Faster R-CNN and HRNet are COCO-pretrained and fixed.
- The training is done only on the lifting stage.
- Five architectures tested:
  SimpleBL [6], **SemGCN** [7], **GraFormer** [8], **GraphMLP** [9], **JointFormer** [10].
- Two training modes compared: **from-scratch** vs. **synthetic pre-training** then fine-tuning.
- Progressive real-world data regime: **5%, 10%, 25%, 50%, and 100%** of one subject, then **2**, **4**, and **8 subjects.**

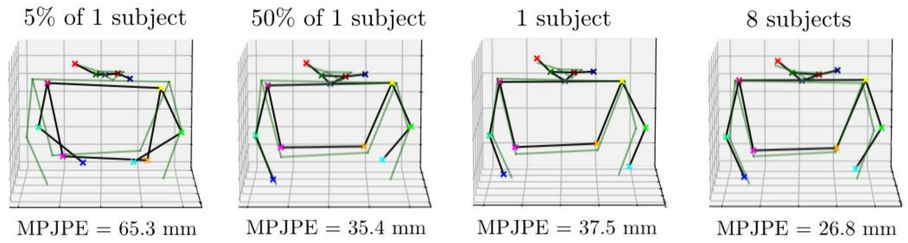
Key Results

- **Fine-tuning with 5% real data**
  - GraFormer MPJPE drops from 90.0 → 70.9 mm with pre-training (19.1 mm improvement).
  - SimpleBL shows the most dramatic gain: 404.3 → 88.0 mm (316.3 mm reduction).
- The gap narrows as real data increases.
- **At full data:** GraFormer sees no improvement (both are 48.1 mm).
- **Best overall results:** MPJPE 46.0 mm from Pre-trained SimpleBL and JointFormer at full data.
- **Hybrid models:** GraphMLP and GraFormer benefit less from pre-training but are more data-efficient overall.

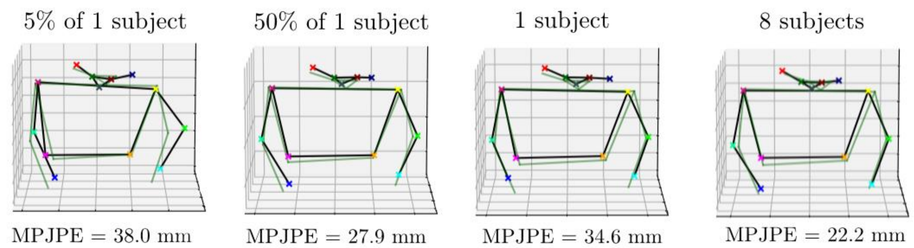
Example Estimation

(A sample frame from Drive&Act Subject#12)

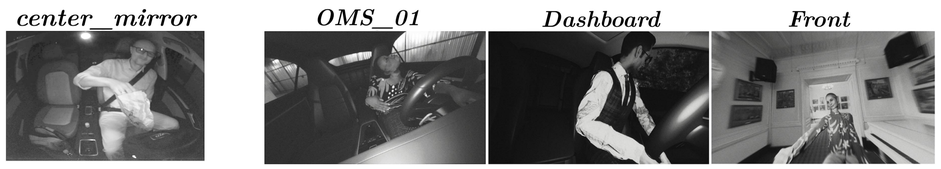
GraFormer trained from scratch



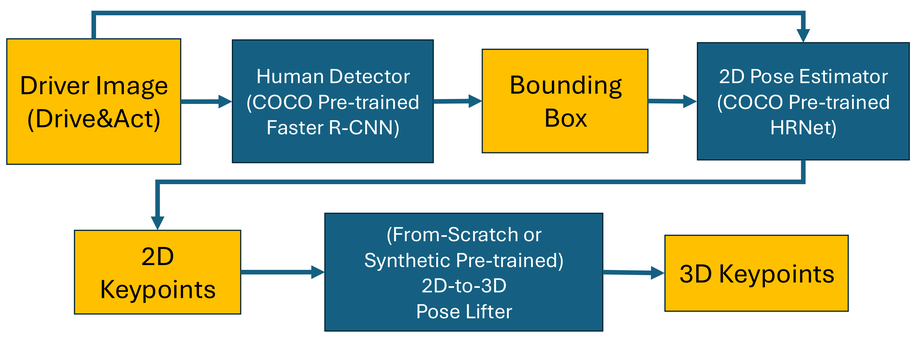
GraFormer with pre-training



Datasets

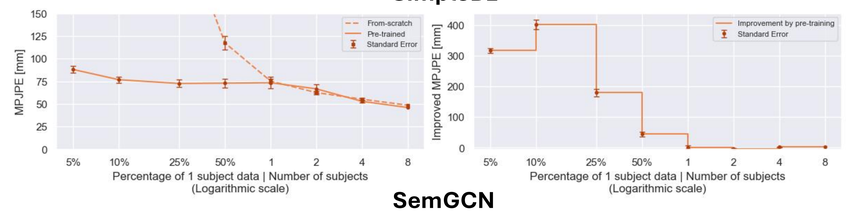


3D Pose Estimation Pipeline

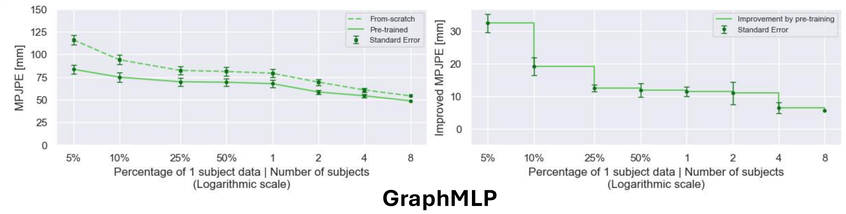


Results

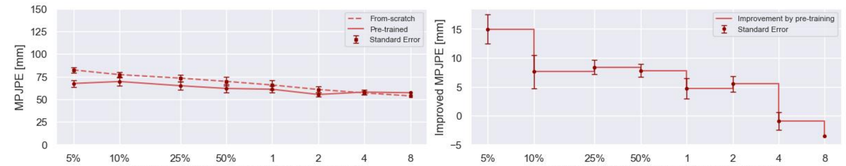
SimpleBL



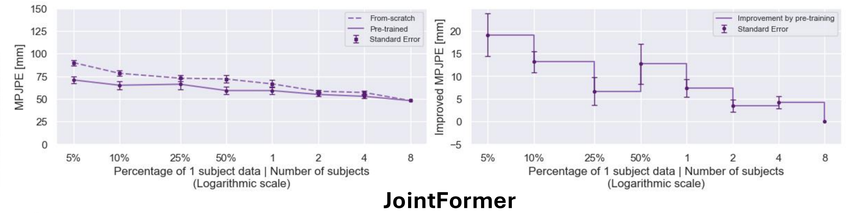
SemGCN



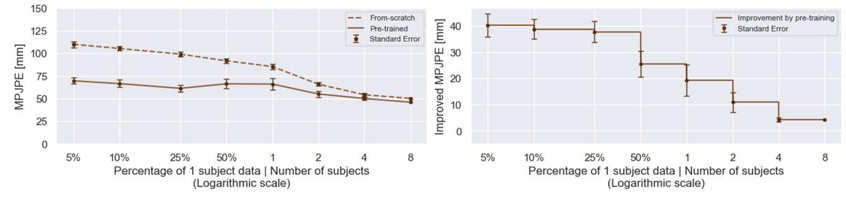
GraphMLP



GraFormer



JointFormer



Conclusion

- Synthetic pre-training consistently helps when real data is scarce, and the benefit diminishes gradually as real data grows.
- It also slightly improves the best achievable result (48.1 mm with GraFormer → 46.0 mm with JointFormer).
- Hybrid architectures benefit less from synthetic pre-training, while showing better data efficiency.

Acknowledgements & Funding
This work was funded by FFG (BMK) in parts under the SyntheticCabin (No. 884336) project and in parts under the UNISCOPE-3D (No. 911019/923852/937153) project. Synthetic datasets were provided by emotion3D.

Reference
[1] Dingus et al., PNAS 2016
[2] Martin et al., ICCV 2019
[3] Sagmeister et al., IV 2023
[4] Ren et al., NeurIPS 2015
[5] Sun et al., CVPR 2019
[6] Martinez et al., ICCV 2017
[7] Zhao et al., CVPR 2019
[8] Zhao et al., CVPR 2022
[9] Li et al., Pattern Recognition 2025
[10] Lutz et al., ICPR 2022