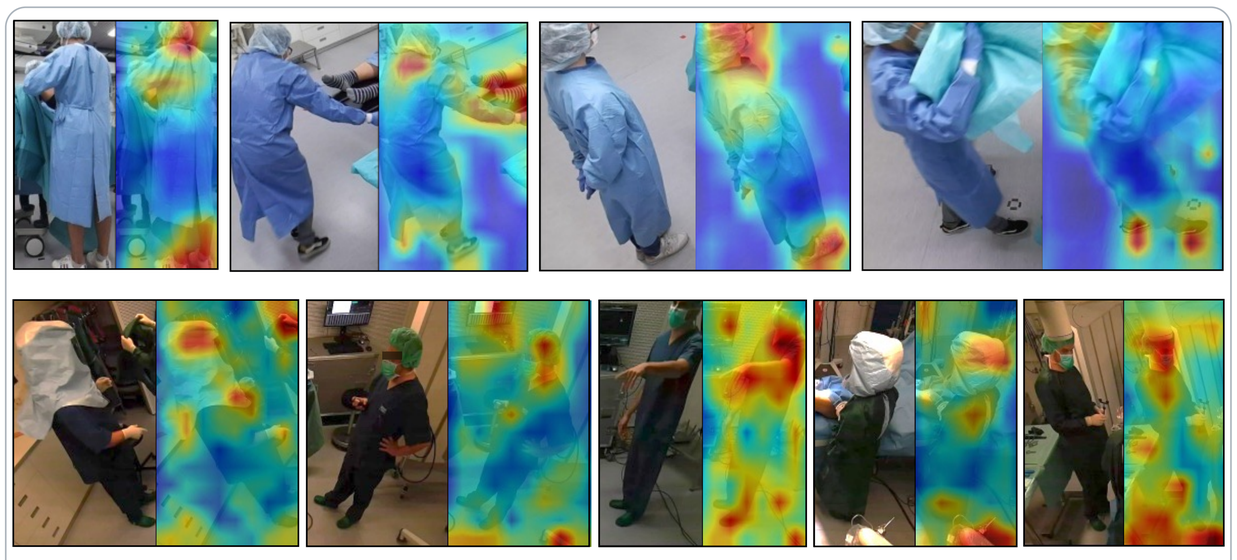
Mitigating Biases in Surgical Operating Rooms with Geometry

Tony D. Wang[1], Tobias Czempiel[2], Christian Heiliger[3], Nassir Navab[1,4], Lennart Bastian[1,4]

[1]Computer Aided Medical Procedures, Technical University of Munich, Germany
[2]UCL Hawkes Institute, Dept. Computer Science, University College London
[3]Minimally Invasive Surgery, University Hospital of Munich (LMU), Germany
[4]Munich Center for Machine Learning, Germany

Introduction

- Deep neural networks are essential for intelligent systems in the OR but are prone to learning **spurious correlations** instead of meaningful features, which can be a **vulnerability in this safety-critical domain**.
- In the OR, this issue **is confounded by homogeneous surgical smocks and gowns** that obscure identifying landmarks. As a result, **models** learn to **"cheat" by focusing on unreliable shortcuts** like a person's shoes or eyewear.
- We propose to mitigate these biases by shifting from **appearance-based data to geometric representations**. By encoding personnel as 3D point cloud sequences, we can learn **robust shape and articulated motion** patterns that are invariant to standardized attire



- On the simplified 4D-OR[1] dataset (left), saliency maps reveal the model learns a shortcut, consistently focusing on non-robust cues like **shoes** that are visible due to a **lack of realism**.
- On the more realistic MM-OR[2] dataset (right) with homogenous attire, these **shortcuts vanish**. The resulting unfocused activation maps show the model **failing to find reliable features**, suggesting that RGB-based recognition is not robust for clinical settings.

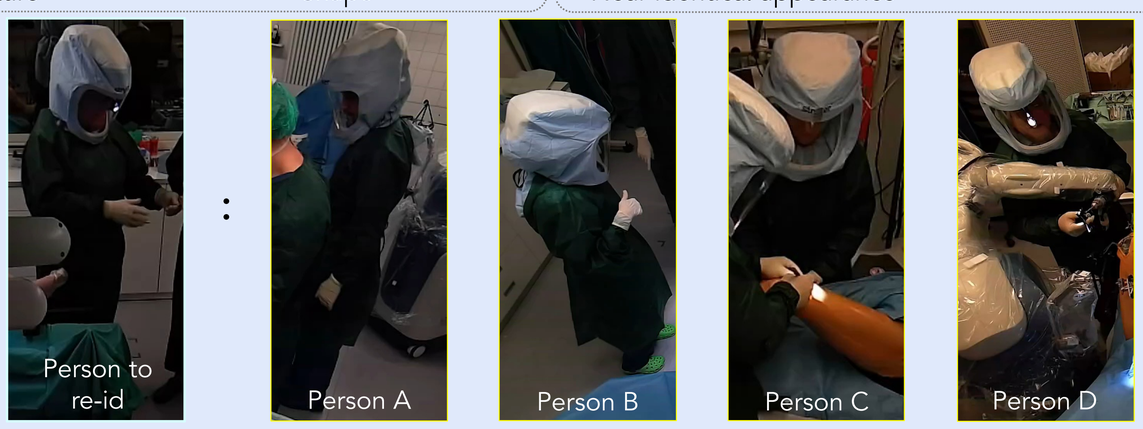
Methodology

We conduct an ablation study comparing **RGB** and **point cloud features** within **identical network architectures** on two surgical datasets.
This setup allows us to **fairly quantify** how each data source **performs on person re-identification**, a task where RGB cues are often ambiguous (as shown below)

RGB

How to re-identify:
- Face
- Attire
- Complexion
- Shape

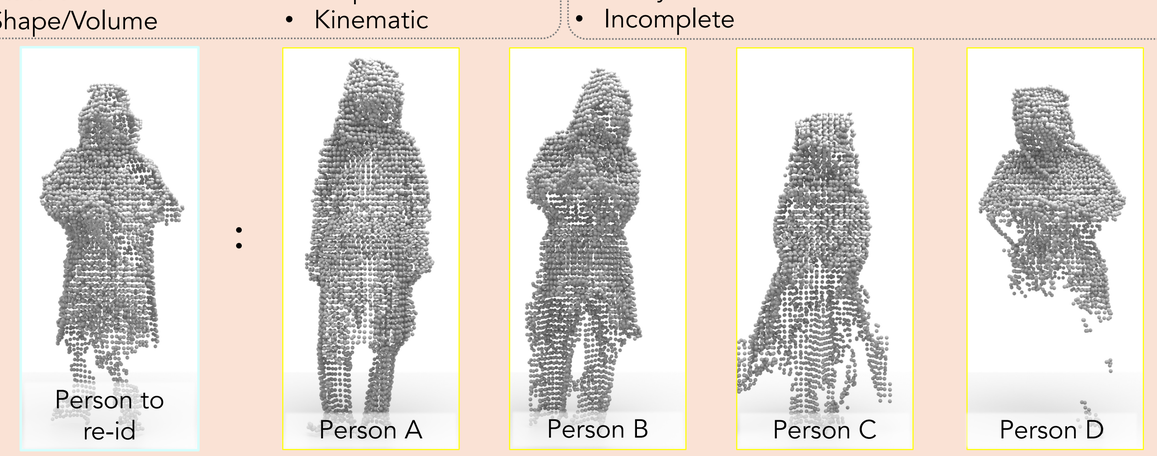
Difficulties:
- Ambiguous scale
- Near identical appearance



Point Cloud Sequence

How to re-identify:
- Stature
- Shape/Volume
- Proportions
- Kinematic

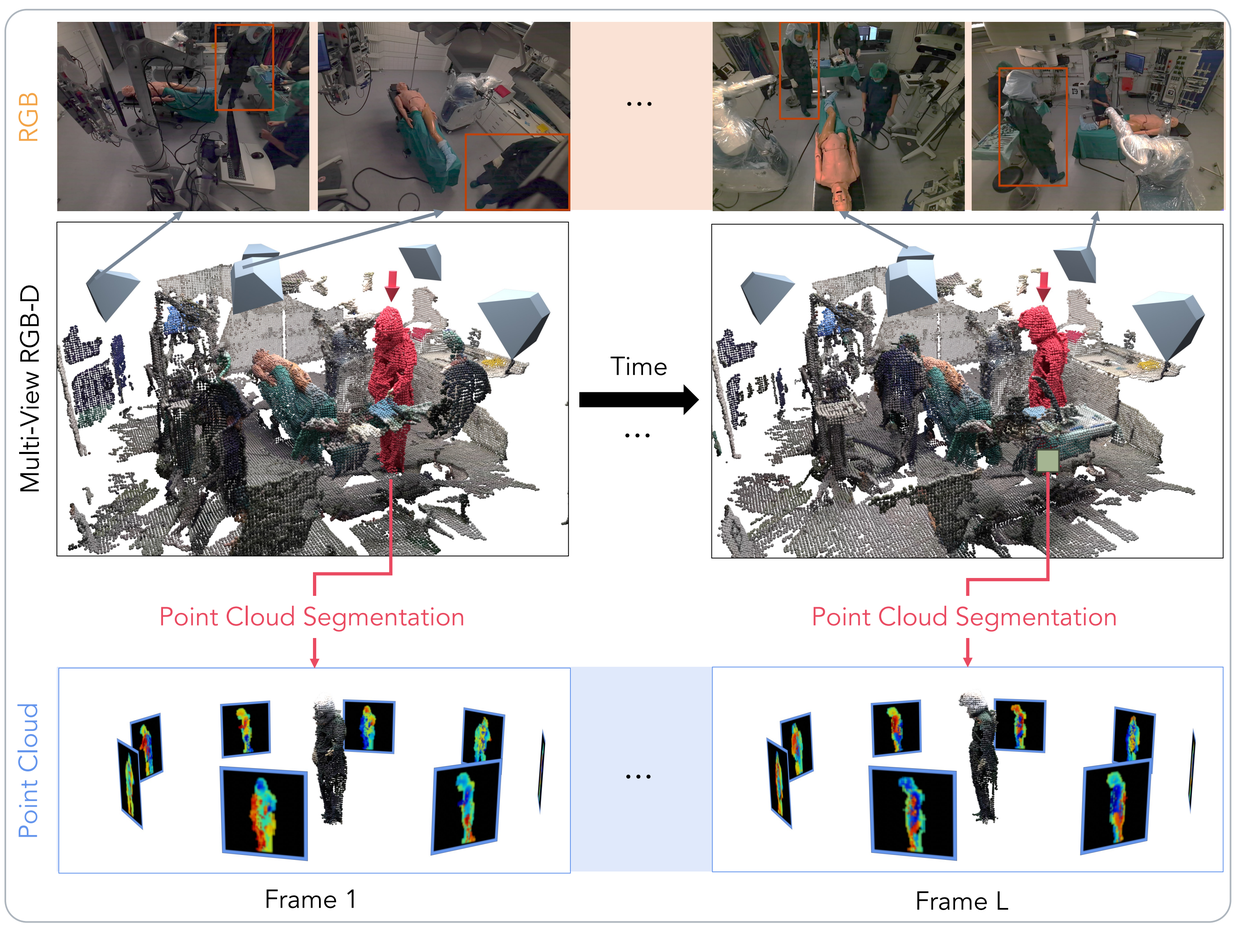
Difficulties:
- Noisy
- Incomplete



Can you re-identify the person?

Dataset

- We use two surgical datasets: the simplified 4D-OR[1] with 5 individuals across 10 procedures, and the more authentic MM-OR[2] dataset, which features 13 individuals across 11 surgeries.
- To create our inputs, individuals are first segmented from the 3D scene using a weakly-supervised method[3]. We then use these segmentations to generate paired sequences of both cropped RGB images and 3D point clouds for each person.

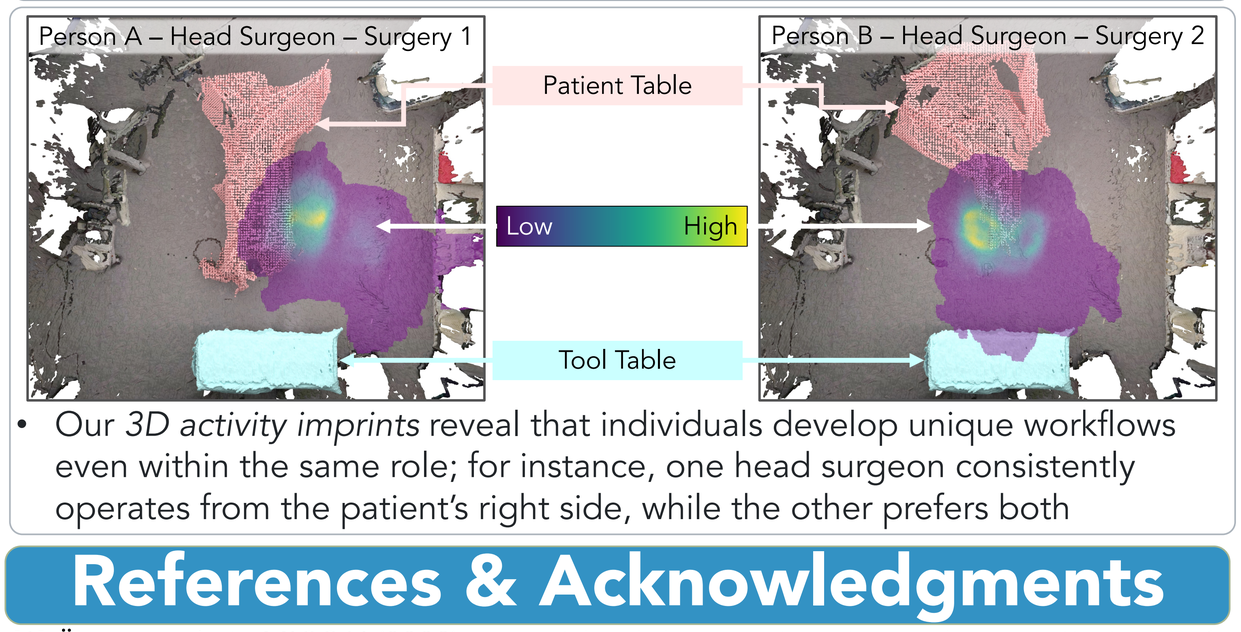


Experiments & Results

Our experiments confirm that **geometric cues are more robust** than RGB for re-identifying surgical personnel. While both modalities perform well on the **4D-OR** dataset, the RGB model's performance degrades on the more authentic **MM-OR** dataset, where standardized attire limits visual diversity. This weakness is further highlighted in cross-domain generalization.

TEST	TRAIN: 4D-OR	TRAIN: MM-OR	TRAIN: SUSTech1K
4D-OR	**96.91 ± 0.7**	54.69 ± 3.1	65.14 ± 1.1
	95.95 ± 0.5	**89.44 ± 0.4**	**90.06 ± 0.3**
MM-OR	46.98 ± 2.5	73.23 ± 3.6	62.82 ± 1.8
	65.92 ± 0.7	**85.75 ± 2.0**	**78.10 ± 0.7**

- Intra-domain: **On 4D-OR**, both methods achieve **>95% macro accuracy**; however, on **MM-OR**, the RGB model's accuracy drops to **73.23%** while our point cloud method maintains **85.74%**.
- Cross-domain: When training on the general-purpose **SUSTech1K** dataset[4], our point cloud model retains **90.06%** accuracy on 4D-OR, whereas the RGB model's performance falls to **65.14%**.



- Our *3D activity imprints* reveal that individuals develop unique workflows even within the same role; for instance, one head surgeon consistently operates from the patient's right side, while the other prefers both

References & Acknowledgments

[1] Özsoy et al., in MICCAI (2022)
[2] Özsoy et al., in CVPR (2025)
[3] Bastian et al., in MICCAI (2023)
[4] Shen et al., in CVPR (2023)

Acknowledgements: This work was partly supported by the state of Bavaria through Bayerische Forschungsstiftung (BFS) under Grant AZ-1592-23-ForNeRo and the German Federal Ministry for Economic Affairs and Climate Action (BMWK) through the Central Innovation Programme for SMEs (ZIM) under Grant KK 5389102BA3.