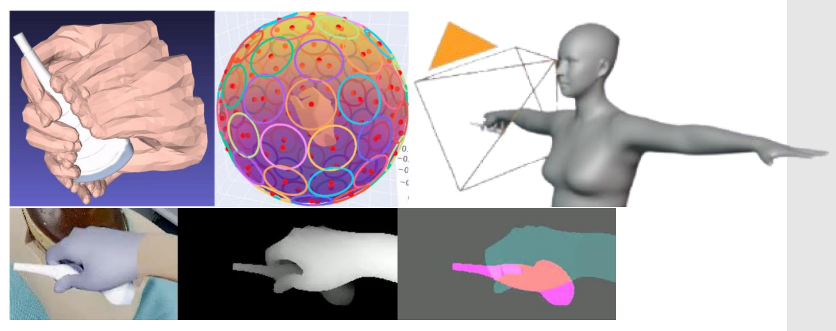
HUP-3D: A 3D multi-view synthetic dataset for assisted-egocentric hand-ultrasound-probe pose estimation

Manuel Birlo, Razvan Caramalau, Philip J. "Eddie" Edwards, Brian Dromey, Matthew J. Clarkson, Danail Stoyanov

Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London, Charles Bell House, 43–45 Foley Street, London W1W 7TY, UK

Introduction

- Egocentric markerless 3D joint pose estimation has potential applications in mixed reality medical education
- The ability to understand hand and probe movements enables tailored guidance and mentoring applications
- Our synthetic dataset HUP-3D, intended for training of state-of-the-art deep learning 3D pose estimators, includes over 31k sets of RGB, depth, and segmentation mask frames, incl. pose-related reference data, emphasizing image diversity and complexity
- HUP-3D is highly suitable for training and optimization of state-of-the-art deep learning-based 3D pose estimation models



Method

1. Grasp generation

- We adopted a strategy focused on generating synthetic grasp images, avoiding the complexities associated with annotating real images
- We adapted a generative model for joint 3D grasp generation to a more clinical scenario.
- Our grasp generation process employs two sequential networks based on the MANO hand model:
  - An encoder-decoder network that generates initial coarse hand poses
  - A subsequent neural network dedicated to fine-tuning the initial coarse poses, specifically enhancing accuracy in hand-tool interaction regions

2. Grasp rendering

- Using Blender, an open-source 3D graphics software for grasp rendering, we tailored our rendering pipeline to accommodate the grasp poses $\psi := [\gamma^{**} \in \mathbb{R}^3, \theta^{**}_{full\_pose} \in \mathbb{R}^3]$ produced by the generative model (see Fig. 1)
- Our rendering approach incorporates a SMPL-H body model, a MANO right hand model $M_{Vert} := [\gamma \in \mathbb{R}^3, \theta_{wrist} \in \mathbb{R}^3]$, and the probe model's vertex data Ω_{Vert}.
- To enhance the diversity of camera perspectives, we transitioned from the purely egocentric viewpoint strategy to the sphere-based methodology which captures both egocentric and non-egocentric images. This method, illustrated in Figs. 1 (lower part) and 2 (b), involves distributing camera positions around a sphere, creating a varied perspective landscape around the right hand
- The rendering model outputs a comprehensive set of images foreach grasp, including RGB-D and segmentation maps, as well as ground truth annotations. Sample frames from the HUP-3D dataset are shown in Fig. 3.

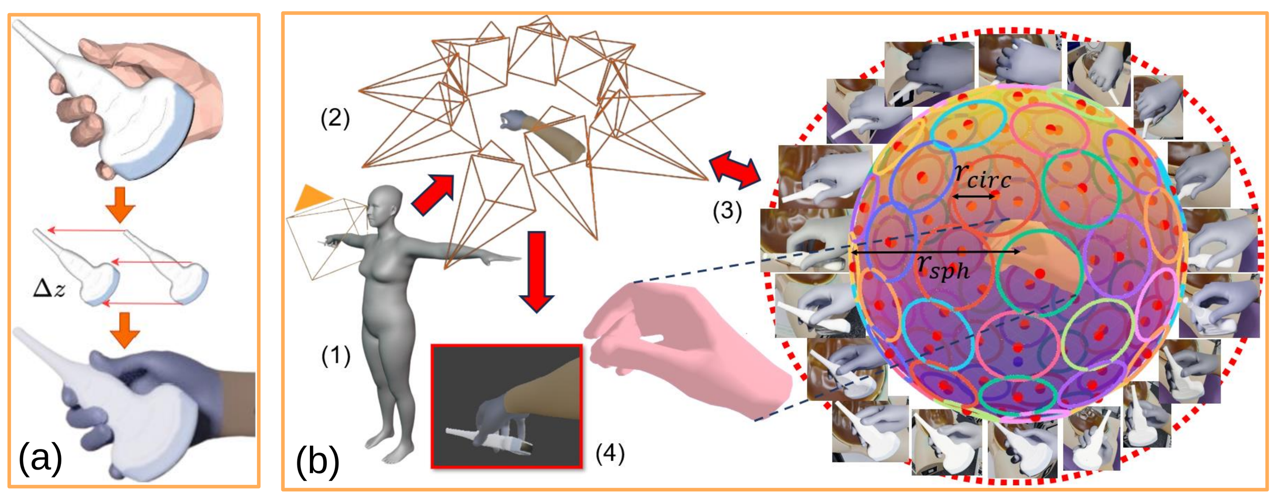


Fig 2: (a) Schematic grasp conversion from generative model to rendering software, including probe offset (Δz) correction: A calibration step is needed, either pre-rendering or pre-grasp generation, to correct small differences between the probe model's world coordinate representation from grasp generation and rendering. (b) Grasp rendering overview: (1) SMPL-H body model grasping the probe, showing egocentric and non-egocentric views. (2) Right arm and sphere-based camera orientations with remaining SMPL-H body parts hidden. (3) Camera angle sphere concept with views at various latitudes, centered on hand mesh; defines sphere (r_{sphr}) and circle (r_{circ}) radii. (4) Rendered hand-probe scene example from a sphere camera position.

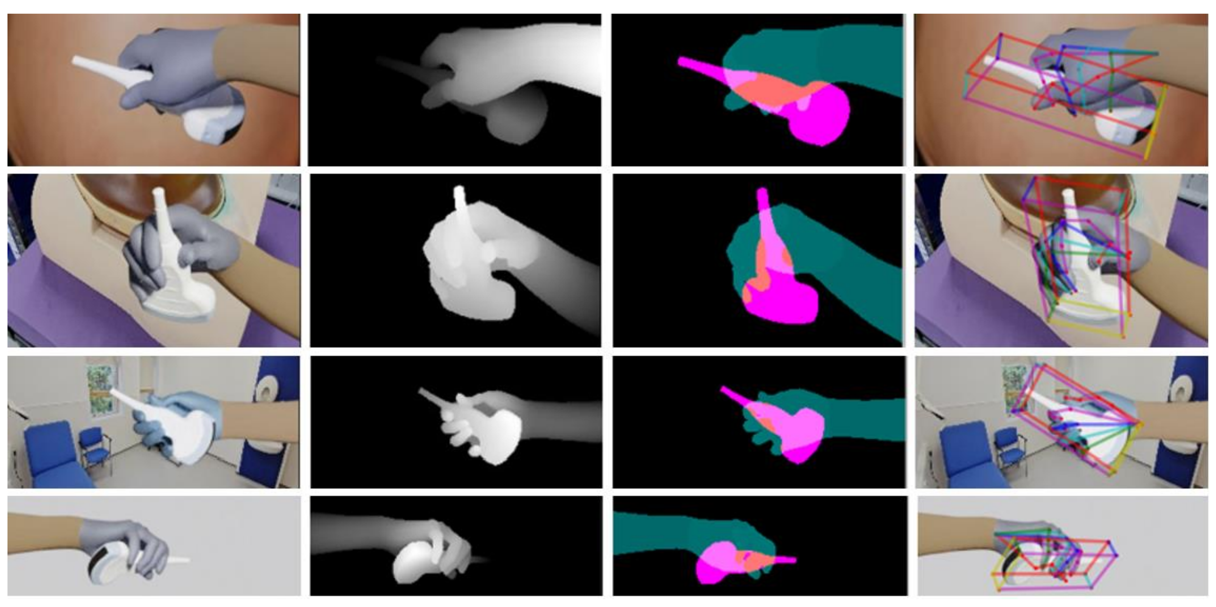


Fig. 4: Sample frames from the HUP-3D dataset, grouped columnwise, from left to right: RGB, depth, segmentation map, and ground truth annotations.

Dataset	# frames	Source (Real/ Synth)	Viewpoints (Single/Multi/Ego)	Annotations	Modalities	Clinical (no. of tools)
HO-3D	77.5k	Real	Single	automatic	RGB	-
ObMan	153k	Synth	Multi	automatic	RGB-DS	-
ContactPose	2.9M	Real	Multi	semi-automatic	RGB-D	-
Hein et al.	10.5k	Synth	Ego	automatic	RGB-DS	1
POV-Surgery	88k	Synth	Ego	automatic	RGB-DS	3
HUP-3D (ours)	31680	Synth	Multi	automatic	RGB-DS	1

Table 1: Dataset comparison: HUP-3D outstands as the first multi-view 3D hand-(clinical) object dataset.

Dataset comparison In Table 1, we enlist the top clinical and non-clinical datasets, together with their properties. HUP-3D is the largest multi-view data set for clinical applications, presenting 3 possible modalities, RGB-DS (color, depth and segmentation maps). Only POV-Surgery contains a higher number, but with less samples per tool (29k) and just firs-person view.

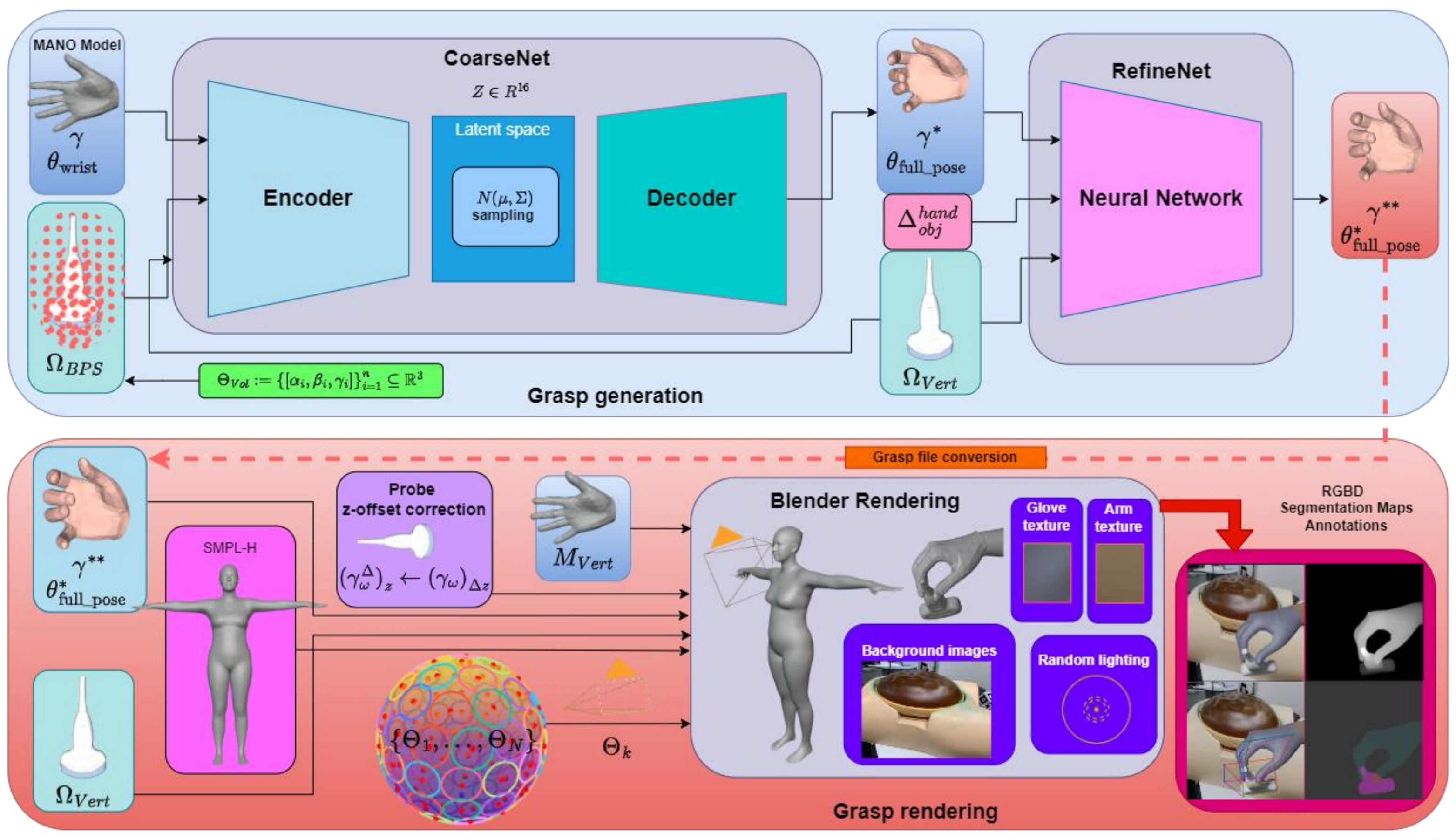


Fig. 1: Grasp Generation (blue) and Rendering Pipeline (red): The process begins with a MANO hand model with initial hand pose $\gamma \in \mathbb{R}^3$ and wrist orientation $\theta_{wrist} \in \mathbb{R}^3$, and BPS-encoded point cloud representations of the probe model Ω_{BPS}. Defined Euler angles Θ_{Vol} for probe meshes Ω_{BPS} were used for precise grasp control. CoarseNet generates initial hand poses, further refined by RefineNet for precise hand-probe alignment. In the rendering phase, the optimized hand pose model vertices, and a SMPL-H model are processed in Blender. Using a multi-viewpoint camera via a spherical layout and centered on the hand and arm, several textures and backgrounds are applied for diverse RGB-D, segmentation maps, and annotations.

Experiment: 3D-hand-probe pose estimation

- To support the utility of our proposed dataset HUP-3D, we deploy a deep learning (DL) state-of-the-art model designed for other datasets like HO-3D. In a supervised learning setting, we further split the data as 7 grasps for training (20,160), 2 grasps for validation, and 2 more for testing (5,760)
- We use a state-of-the-art baseline model HOPE-net, which manages to reduce the highly non-linear regression of the 3D hand and object coordinates, and is trained with our HUP-3D dataset, following the same settings as in the original HOPE-net paper

Model/ MPJPE error [mm]	3D Hand Joints	3D Probe box	3D Hand + Probe
DeepPrior++	7.18	22.21	9.69
HOPE-Net	**5.3**	**17.05**	**8.65**

- **Qualitative results:** Visual confirmation of predicted key points of our HUP-3D dataset:

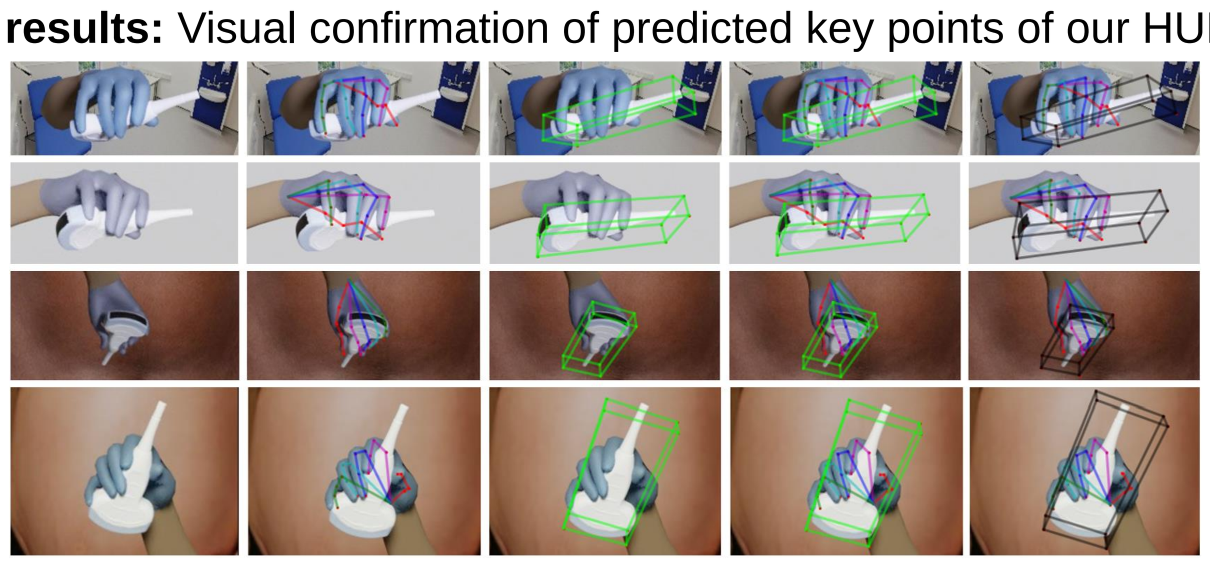


Fig. 5: Qualitative results, shown with 4 test images from HUP-3D: image columns from left to right: RGB, predicted hand joints, predicted probe corners, predicted joints and corners, ground truth of joints and corners

Conclusion and future work

- We introduce HUP-3D, a pioneering 3D hand-object multi-view dataset tailored for obstetric hand US probe grasps
- HUP-3D aims to enhance research in clinical movement analysis via egocentric camera and mixed reality applications
- Our data generation process leverages a versatile model for grasp generation and an efficient automated rendering pipeline, illustrating the benefits of our multi-view camera sphere approach
- A baseline model evaluation confirmed our method's effectiveness, even with significant hand-probe occlusions
- Future efforts will focus on improving real-world applicability by incorporating automatically annotated real images and developing more sophisticated grasp generation techniques that incorporate temporal sequences for better manual interaction and predictions

Acknowledgments

This work was supported in whole, or in part, by the Well-come/EPSRC Centre for Interventional and Surgical Sciences (WEISS)[203145/Z/16/Z], the Department of Science, Innovation and Technology (DSIT) and the Royal Academy of Engineering under the Chair in Emerging Technologies programme.

Contact: Manuel Birlo

manuel.birlo.18@ucl.ac.uk

HUP-3D Project website