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, and 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
manuel.birlo.18@ucl.ac.uk

Abstract. We present HUP-3D, a 3D multiview multimodal synthetic dataset for hand ultrasound (US) probe pose estimation in the context of obstetric ultrasound. Egocentric markerless 3D joint pose estimation has potential applications in mixed reality medical education. The ability to understand hand and probe movements opens the door to tailored guidance and mentoring applications. Our dataset consists of over 31k sets of RGB, depth, and segmentation mask frames, including pose-related reference data, with an emphasis on image diversity and complexity. Adopting a camera viewpoint-based sphere concept allows us to capture a variety of views and generate multiple hand grasps poses using a pre-trained network. Additionally, our approach includes a software-based image rendering concept, enhancing diversity with various hand and arm textures, lighting conditions, and background images. We validated our proposed dataset with state-of-the-art learning models and we obtained the lowest hand-object keypoint errors. The supplementary material details the parameters for sphere-based camera view angles and the grasp generation and rendering pipeline configuration. The source code for our grasp generation and rendering pipeline, along with the dataset, is publicly available at https://manuelbirlo.github.io/HUP-3D/.

Keywords: Egocentric 3D joint hand and tool pose estimation · Synthetic datasets · Obstetrics ultrasound.

1 Introduction

The ability to infer hand and tool pose information from video data in clinical setups opens the door to several potentially useful applications aimed at assisting clinicians through context-specific evaluation of their movement. Novel methods focusing on marker-free video-based clinical skill assessment have been proposed such as surgical hand and tool pose estimation [24], surgical tool movement analysis [25] and skill assessment in robotic surgery [26]. Accurate estimation of hand and tool pose can enhance the precision and effectiveness of ultrasound procedures, leading to improved diagnostic outcomes and training methodologies.

Solutions for skill assessment that rely on physical motion sensors have been developed in fields such as hand motion analysis for endovascular procedures [24] and guidance for using ultrasound (US) probes in obstetrics [25]. With the emerging trend of mixed reality head-mounted displays, dominated by the Microsoft HoloLens 2 [1], egocentric pose estimation methods arise, such as probe tracking for US-guided procedures [27].

In the context of obstetric US, we explore 3D joint hand-tool pose estimation with applications in mixed reality-based medical education, where analysis of hand and probe movements could facilitate holographic assistive guidance. Such innovations aim to help standardize clinical training protocols. Standardized target scanning planes are used for training but there is a lack of universally accepted competence measures [18]. By distinguishing between novice and expert clinician movements, machine learning-based pose estimation is a powerful tool for developing standardized training approaches. This technique supports established clinical practices to estimate fetal development through biometry. It offers a pathway to more uniform and effective clinical training [17].

Image datasets required for machine learning-based model training and subsequent pose estimation can be categorized into real and synthetic images. Real dataset generating methods often employ marker-free methods to capture hand grasp information directly from image data [6,7,8,9,10]. Although real images offer the advantage of authentic context [10,13], they pose challenges in generating accurate ground truth due to the labor-intensive nature of manual annotations and potential biases from sensor use [7]. Marker-free approaches, by avoiding markers on the tool and/or hand, mitigate the risk of pose prediction bias, making them a popular choice for reducing inaccuracies associated with additional visible sensors or markers [8,7].

Synthetic images, however, offer built-in ground truth from 3D models, simulating realistic grasping scenarios with the benefits of easy scalability and generalizability to real images [6]. Furthermore, synthetic ground truth proves useful for addressing mutual occlusions resulting from hand-tool interactions.

When creating training images for pose estimation, it is crucial to account for the dataset's generalizability, particularly the expected camera location and range of viewpoints. Applications vary, with some utilizing non-egocentric views for capturing hand grasps [6,10]. Mixed and augmented reality setups using head-mounted devices like the Microsoft HoloLens 2, require consideration of egocentric perspectives for pose estimation from device-recorded camera data [7,8].

When capturing clinical instruments such as US probes in synthetic images, the realism of hand grasps is constrained by specific hand-tool contact areas and orientations. This requirement, previously noted in the context of orthopedic surgical tools such as drills [7] and other instruments such as scalpels and diskplacers [8], poses a challenge. Traditional grasp generation techniques, like those from robotic grasping software [14] and used in similar studies [6,7], encounter difficulties due to the specific dimensions and clinically relevant grasp positions of the US probe. Consequently, we employed more flexible solutions such

[1] https://www.microsoft.com/en-us/hololens

as a generative model for machine learning-based grasp generation [9], which has been validated in clinical environments [8].

To increase image diversity for egocentric applications in a scalable way, we broadened the approach to multi-view by allowing camera movement around a sphere's surface, centered on the hand. This method supports both egocentric distances and a mix of egocentric and non-egocentric viewpoints.

Our contributions can be summarized as follows:

- A scalable synthetic multi-modal (RGB-D, segmentation maps) image generation pipeline capable of producing a wide variety of realistic hand-ultrasound-probe grasp frames, without prior external data recording
- A novel sphere-based camera viewpoint generation that enhances frame generalizability by combining egocentric head-hand distances with non-egocentric camera viewpoints.
- HUP-3D: A diverse multimodal synthetic dataset tailored for joint 3D hand and tool pose estimation, featuring the Voluson™ C1-5-D[2] ultrasound probe commonly used in obstetrics, including a variety of hand poses, textures, backgrounds, lighting, and camera angles
- Lowest hand and object 3D pose estimation errors for a synthetic dataset with a trained state-of-the-art model, HOPE-net [30].

2 Method

We focus on potential medical education applications in the context of US obstetrics, but maintain a high degree of flexibility toward other use cases. Our synthetic image generation pipeline is split into two sections: grasp generation and grasp rendering. These are described in the following subsections. A graphical overview of our pipeline is shown in Fig. 1.

2.1 Grasp generation

To achieve automated annotation we adopted a strategy focused on generating synthetic grasp images, avoiding the complexities associated with annotating real images. This approach allowed us to maintain a clear and manageable rendering workflow. The underlying motivation in pursuing a purely synthetic image generation approach is to explore the possibility of creating a sufficiently large variety of training images to allow generalizability to real images for joint 3D hand and tool pose prediction. Our initial feasibility study incorporated the use of a robotic grasping tool [14] which turned out to be error-prone in our application and did not produce a sufficiently large variety of plausible hand grasps due to restrictions in terms of hand dexterity. We then adapted the generative model proposed in [9] for joint 3D grasp generation to a more clinical scenario. Our grasp generation process employs two sequential networks based on the MANO hand model [28]: an encoder-decoder network that generates initial coarse hand

[2] https://services.gehealthcare.com/gehcstorefront/p/5499513

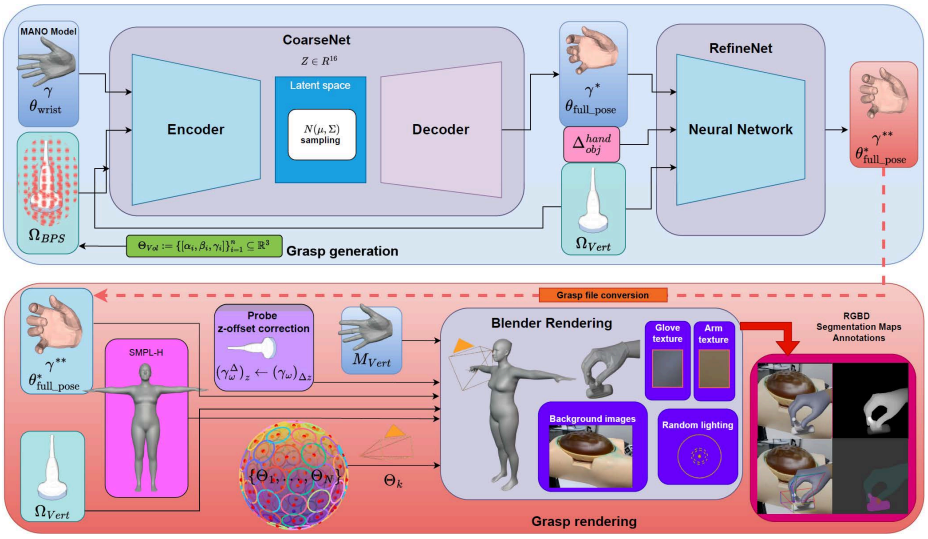

Fig. 1: Grasp Generation (blue) and Rendering Pipeline (red): The process begins with a MANO hand model initialization and a BPS-encoded Voluson model point cloud. CoarseNet generates initial hand poses, further refined by RefineNet for precise hand-probe alignment. In the rendering phase, the optimized 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.

poses and a subsequent neural network dedicated to fine-tuning these poses, specifically enhancing accuracy in hand-tool interaction regions. The encoder, which samples from a normal distributed 16-dimensional latent space, requires encoded point cloud representations [19] of the probe model (Ω_{BPS}), together with the MANO right-hand model's initial translation $\gamma \in \mathbb{R}^3$ and hand wrist orientation $\theta_{wrist} \in \mathbb{R}^3$. Defined Euler angles Θ_{Vol} for probe meshes Ω_{BPS} were used for precise grasp pose control. Originally, the model described in [9] was trained with ordinary objects (like mugs, cameras etc.). However, we extended its capability to include the Voluson US probe. The decoder outputs an initial hand pose $[\gamma, \theta_{full\_pose}]$, which is subsequently refined through a neural network utilizing the vertices of the probe model Ω_{Vert} and the vertex distances Δ_{obj}^{hand} between hand and probe. This refined pose, expressed as $\Psi := [\gamma^{**}, \theta_{full\_pose}]$, forms the foundation for our grasp rendering approach detailed in Sec. 2.2.

[2] https://www.meshlab.net/

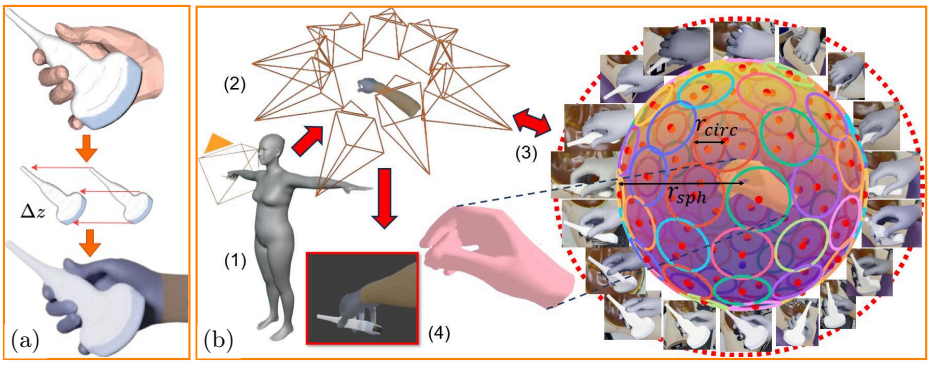

Fig. 2: (a) Schematic grasp conversion from generative model to rendering software, including probe offset (Δz) correction. (b) Grasp rendering overview: (1) SMPL-H body model grasping the probe, showing egocentric and non-egocentric views. (2) Right arm and sphere-based camera orientations with remaining 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.

2.2 Grasp rendering

Using Blender, an open source 3D graphics software [15] for grasp rendering, as demonstrated in [6,7], we tailored our rendering pipeline to accommodate the grasp poses Ψ produced by the generative model outlined in Sec. 2.1. Additionally, this rendering approach incorporates a SMPL-H body model [16], a MANO right hand model M_{Vert}, and the probe model's vertex data Ω_{Vert}. The grasp rendering pipeline can be seen in the lower part of Fig. 1. 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. We position the probe at the origin of the rendering software's world coordinate system, but the modified hand grasp pose Ψ necessitates an offset Δz along the positive z-axis, which was calculated through polygon offset analysis. We adjust the translation offset of the probe by $\gamma_{\omega}^{\Delta} = \gamma_{\omega} + (0, 0, -\Delta z)$. Fig. 2(a) illustrates this correction.

To enhance the diversity of camera perspectives, we transitioned from the purely egocentric viewpoint strategy to the sphere-based methodology outlined in Sec. 2.2, illustrated in Figs. 1 and 2(b). This provides challenging examples such as mutual occlusions between hands and tools, improving the generalizability of the resulting pose estimation model.

For each grasp produced by our module, we generate a synthetic image for each camera view angle $\Theta_k \in \{\Theta_1, \ldots, \Theta_N\}$, covering N positions around the sphere. Our rendering scene setup uses two shades of clinical gloves, varied scene lighting, and eight backgrounds (a lab with a SPACE-FAN ultrasound fetus

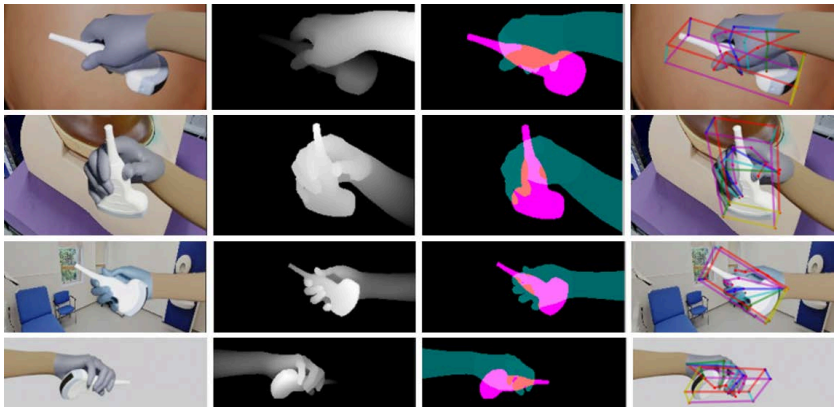

Fig. 3: Sample frames from the HUP-3D dataset, grouped columnwise, from left to right: RGB, depth, segmentation map, and ground truth annotations.

model[3], consultation rooms, a white background, and real abdomens of pregnant women). The rendering model outputs a comprehensive set of images for each 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.

Camera view angle sphere concept Our methodology diverges from traditional egocentric viewpoints by implementing a sphere-based camera view setup to capture both egocentric and non-egocentric images, enhancing dataset diversity. This method, inspired by [3], involves distributing camera positions around a sphere, creating a varied perspective landscape around the right hand. The sphere is divided into horizontal segments, determined by latitude angles, to evenly distribute viewpoints. Specifically, the number of latitude segments N^{ϕ} and circles per segment $N_{circ}^{(i)}$ are calculated to ensure comprehensive coverage:

$$N^{\phi} = \left\lfloor \frac{\pi}{2\arcsin\left(\frac{r_{circ}}{r_{sph}}\right)} \right\rfloor, \quad N_{circ}^{(i)} = \left\lfloor \frac{\pi r_{sph}\sin(\theta_i)}{r_{circ}} \right\rfloor, \; i \in \{1, 2, \ldots, N^{\phi}\} \quad (1)$$

The division of the sphere into latitude floors was chosen to control camera placement and minimize frame redundancy, rather than to achieve perfect circle uniformity. The number of circles is determined in relation to the sphere's radius r_{sph} and the circle's radius r_{circ}. The circles are placed sinusoidally from the top to the bottom of the sphere. For each segment i and each circle j within, camera locations are defined by their spherical coordinates $(\theta_i, \phi_j^{(i)})$, ensuring a near-uniform spread of angles:

$$(\theta_i, \phi_j^{(i)}) \quad \text{with} \quad \phi_j^{(i)} = j \cdot \frac{2\pi}{N_{circ}^{(i)}}, \quad j \in \{0, 1, \ldots, N_{circ}^{(i)} - 1\} \quad (2)$$

[3] https://www.kyotokagaku.com/en/products_data/us-7_en/

This structured approach facilitates the generation of camera angles Θ_k, utilized in our subsequent rendering process. Figs. 1 and 2b visually demonstrate this concept, showcasing the strategic camera placement and the diverse grasp views it enables.

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 dataset for clinical applications, presenting 3 possible modalities, RGB-DS (color, depth and segmentation maps). Only POV-Surgery [8] contains a higher number, but with less samples per tool (29k) and just firs-person view.

Dataset	# frames	Source (Real/ Synth)	Viewpoints (Single/Multi/Ego)	Annotations	Modalities	Clinical (no. of tools)
HO-3D [10]	77.5k	Real	Single	automatic	RGB	-
ObMan [6]	153k	Synth	Multi	automatic	RGB-DS	-
ContactPose [11]	2.9M	Real	Multi	semi-automatic	RGB-D	-
Hein et al. [7]	10.5k	Synth	Ego	automatic	RGB-DS	1
POV-Surgery [8]	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.

3 Experiment

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 [10]. As mentioned before, our dataset consists of 31,680 image sets from 11 realistic hand-object grasps. 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). This will ensure the generalizability of the tested DL model.

3.1 3D hand-probe pose estimation

There have been extensive DL methods proposed for 3D hand-object pose estimation in the computer vision community. One of these competitive baselines is HOPE-net [30], originally tested on real data. HOPE-net extends the capabilities of residual convolutional neural networks [29] with an adaptive Graph U-Net module [31]. This module manages to reduce the highly non-linear regression of the 3D hand and object coordinates.

The task of the estimator is to map the RGB images to 3D world coordinates of the hand skeleton and the object's boundary corners. For this, we minimize a mean square error loss during training on both 2D and 3D coordinates. In terms of training, we follow the same settings as in the original HOPE-net paper [30].

Quantitatively, once trained, we measure the error in millimeters between the predicted joints and the ground truth. In the test set, we obtain a total error of **8.65 mm** as the mean per joint position error (MPJPE) with 5.33 mm from the hand and 17.05 mm from the object. The testing error is the lowest compared to other clinical data sets such as POV-Surgery [8] (14.35 mm) and Hein et al. [7] (17.02 mm) where even more advanced DL models were used. This proves that a multi-viewpoint dataset helps to estimate the object and the hand location with higher precision. In our experiments, we have also tested a more simple baseline composed just of ResNet-50 [29], where the error was higher at 9.69 mm. In Fig. 4, we visually confirm the accurate predicted keypoints of our HUP-3D test set.

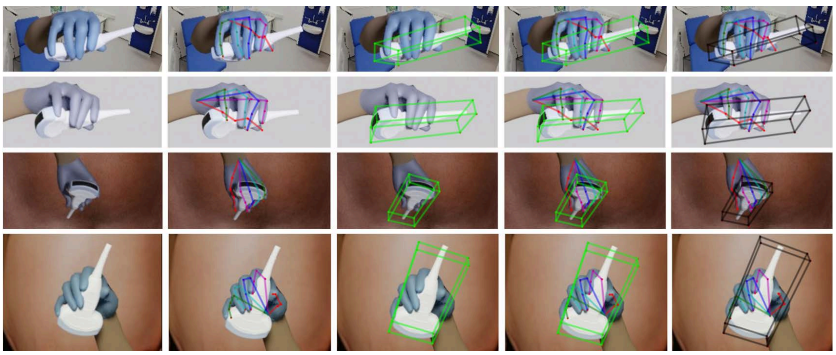

Fig. 4: Qualitative results, shown with 4 test images from HUP-3D: image columns from left to right: RGB, predicted hand joints, predicted probe corners, predicted joints and corners, ground truth of joints and corners

4 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. We hope that the current dataset will facilitate rapid advances in hand and tool pose estimation in obstetric ultrasound.

Ethical Considerations In collecting the data, the right ethics have been considered. The presented data is synthetic and does not involve any data privacy concerns. For the backgrounds and meshes, we have used anonymized public images. The dataset does not pose any known harm.

Acknowledgments. This work was supported in whole, or in part, by the Wellcome/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. For the purpose of open access, the author has applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission.

Disclosure of Interests. Author Danail Stoyanov is employed at Odin Vision Ltd. and Digital Surgery. Neither of these companies were involved in this publication. The other authors declare that they have no conflict of interest.

References

1. Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging Photometric Consistency Over Time for Sparsely Supervised Hand-Object Reconstruction. In *Proceedings of the IEEE/CVF CVPR*, June 2020.
2. H. Jiang, S. Liu, J. Wang, and X. Wang, "Hand-Object Contact Consistency Reasoning for Human Grasps Generation," in *Proceedings of the ICCV*, 2021.
3. Abdulkadir Akin, E. Erdede, Hossein Afshari, Alexandre Schmid, and Yusuf Leblebici. Enhanced Omnidirectional Image Reconstruction Algorithm and Its Real-Time Hardware. In *Proceedings - 15th Euromicro Conference on Digital System Design, DSD 2012*, Sep 2012. https://doi.org/10.1109/DSD.2012.52. ISBN 978-1-4673-2498-4.
4. Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-Supervised 3D Hand-Object Poses Estimation With Interactions in Time. In *Proceedings of the IEEE/CVF CVPR*, pages 14687–14697, June 2021.
5. Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Generalized Feedback Loop for Joint Hand-Object Pose Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):1898–1912, 2020. https://doi.org/10.1109/TPAMI.2019.2907951.
6. Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid, "Learning joint reconstruction of hands and manipulated objects," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
7. Hein, J., Seibold, M., Bogo, F., Farshad, M., Pollefeys, M., Fürnstahl, P. and Navab, N. (2021). Towards markerless surgical tool and hand pose estimation. International Journal of Computer Assisted Radiology and Surgery, 16(5), 799–808. https://doi.org/10.1007/s11548-021-02369-2
8. R. Wang, S. Ktistakis, S. Zhang, M. Meboldt, and Q. Lohmeyer, "POV-Surgery: A Dataset for Egocentric Hand and Tool Pose Estimation During Surgical Activities," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023, pp. 440–450.
9. O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "GRAB: A Dataset of Whole-Body Human Grasping of Objects," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. [Online]. Available: https://grab.is.tue.mpg.de
10. Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A Method for 3D Annotation of Hand and Object Poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
11. Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A Dataset of Grasps with Object Contact and Hand Pose. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 361–378. Springer International Publishing, Cham, 2020. ISBN 978-3-030-58601-0.
12. Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J. Crandall. HOPE-Net: A Graph-Based Model for Hand-Object Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
13. T. Kwon, B. Tekin, J. Stühmer, F. Bogo, and M. Pollefeys, "H2O: Two hands manipulating objects for first person interaction recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10138–10148, 2021.

14. Andrew T. Miller and Peter K. Allen. Graspit! A Versatile Simulator for Robotic Grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004. IEEE.
15. Blender Online Community. *Blender - a 3D modelling and rendering package*. Stichting Blender Foundation, Amsterdam, 2018.
16. Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245, November 2017. Association for Computing Machinery, New York, NY, USA. ISSN 0730-0301.
17. B. P. Dromey, S. Ahmed, F. Vasconcelos, E. Mazomenos, Y. Kunpalin, S. Ourselin, J. Deprest, A. L. David, D. Stoyanov, and D. M. Peebles, "Dimensionless squared jerk: An objective differential to assess experienced and novice probe movement in obstetric ultrasound," *Prenatal Diagnosis*, vol. 11, 2020.
18. Y. Cai, R. Droste, H. Sharma, P. Chatelain, L. Drukker, A. T. Papageorghiou, and J. A. Noble, "Spatio-temporal visual attention modelling of standard biometry plane-finding navigation," *Medical Image Analysis*, vol. 65, 2020.
19. Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF ICCV*, 2019.
20. G. Varol, J. Romero, X. Martin, N. Mahmood, M.J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In: *Proceedings of the IEEE CVPR*, 2017.
21. D. P. Azari, Y. H. Hu, B. L. Miller, B. V. Le, and R. G. Radwin, "Using surgeon hand motions to predict surgical maneuvers," *Human Factors*, vol. 61, 2019, SAGE Publications Sage CA: Los Angeles, CA.
22. X.-H. Zhou, G.-B. Bian, X.-L. Xie, Z.-G. Hou, X. Qu, and S. Guan, "Analysis of Interventionalists' Natural Behaviors for Recognizing Motion Patterns of Endovascular Tools During Percutaneous Coronary Interventions," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, 2019.
23. R. Droste, L. Drukker, A. T. Papageorghiou, and J. A. Noble, "Automatic probe movement guidance for freehand obstetric ultrasound," in *MICCAI 2020: 23rd International Conference*, Springer, 2020.
24. E. D. Goodman, K. K. Patel, Y. Zhang, W. Locke, C. J. Kennedy, R. Mehrotra, S. Ren, M. Y. Guan, M. Downing, H. W. Chen, et al., "A real-time spatiotemporal AI model analyzes skill in open surgical videos," *arXiv preprint arXiv:2112.07219*, 2021.
25. A. Jin et al., "Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks," in *2018 IEEE WACV*, 2018.
26. G. Lajkó, R. Nagyné Elek, and T. Haidegger, "Endoscopic image-based skill assessment in robot-assisted minimally invasive surgery," *Sensors*, vol. 21, MDPI.
27. T. Nguyen, W. Plishker, A. Matisoff, K. Sharma, and R. Shekhar, "HoloUS: Augmented reality visualization of live ultrasound images using HoloLens for ultrasound-guided procedures," *International Journal of Computer Assisted Radiology and Surgery*, Springer vol. 17, 2022.
28. J. Romero, D. Tzionas, and M. J. Black, "Embodied Hands: Modeling and Capturing Hands and Bodies Together," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, Nov. 2017.
29. Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun, "Deep Residual Learning for Image Recognition", 2015, Tech Report, eprint=1512.03385.
30. B. Doosti, S. Naha, M. Mirbagheri and D. Crandall, "HOPE-Net: A Graph-based Model for Hand-Object Pose Estimation", (CVPR), June, 2020.
31. H. Gao and S. Ji, "Graph U-Nets", ICML, 2019