POV-Surgery: A Dataset for Egocentric Hand and Tool Pose Estimation During Surgical Activities

Rui Wang, Sophokles Ktistakis, Siwei Zhang, Mirko Meboldt, and Quentin Lohmeyer(✉)

ETH Zurich, Zurich, Switzerland
{ruiwang46,ktistaks,meboldtm,qlohmeyer}@ethz.ch, siwei.zhang@inf.ethz.ch

Abstract. The surgical usage of Mixed Reality (MR) has received growing attention in areas such as surgical navigation systems, skill assessment, and robot-assisted surgeries. For such applications, pose estimation for hand and surgical instruments from an egocentric perspective is a fundamental task and has been studied extensively in the computer vision field in recent years. However, the development of this field has been impeded by a lack of datasets, especially in the surgical field, where bloody gloves and reflective metallic tools make it hard to obtain 3D pose annotations for hands and objects using conventional methods. To address this issue, we propose POV-Surgery, a large-scale, synthetic, egocentric dataset focusing on pose estimation for hands with different surgical gloves and three orthopedic surgical instruments, namely scalpel, friem, and diskplacer. Our dataset consists of 53 sequences and 88,329 frames, featuring high-resolution RGB-D video streams with activity annotations, accurate 3D and 2D annotations for hand-object pose, and 2D hand-object segmentation masks. We fine-tune the current SOTA methods on POV-Surgery and further show the generalizability when applying to real-life cases with surgical gloves and tools by extensive evaluations. The code and the dataset are publicly available at http://batfacewayne.github.io/POV_Surgery_io/.

Keywords: Hand Object Pose Estimation · Deep Learning · Dataset · Mixed Reality

1 Introduction

Understanding the movement of surgical instruments and the surgeon's hands is essential in computer-assisted interventions and has various applications, including surgical navigation systems [27], surgical skill assessment [10,15,23] and robot-assisted surgeries [8]. With the rising interest in using head-mounted Mixed Reality (MR) devices for such applications [1,7,21,28], estimating the 3D

R. Wang and S. Ktistakis—Denotes co-first authorship.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 440–450, 2023.
https://doi.org/10.1007/978-3-031-43996-4_42

pose of hands and objects from the egocentric perspective becomes more important. However, this is more challenging compared to the third-person viewpoint because of the constant self-occlusion of hands and mutual occlusions between hands and objects. While the use of deep neural networks and attention modules has partly addressed this challenge [13,18,19,22], the lack of egocentric datasets to train such models has hindered progress in this field. Most existing datasets that provide 3D hand or hand-object pose annotations focus on the third-person perspective [11,20,29]. FPHA [9] proposed the first egocentric hand-object video dataset by attaching magnetic sensors to hands and objects. However, the attached sensors pollute the RGB frames. More recently, H2O [17] proposed an egocentric video dataset with hand and object pose annotated with a semi-automatic pipeline, based on 2D hand joint detection and object point cloud refinement. However, this pipeline is not applicable to the surgical domain because of the large domain gap between the everyday scenarios in [9,17] and surgical scenarios. For instance, surgeons wear surgical gloves that are often covered with blood during the surgery process, which presents great challenges for vision-based hand keypoint detection methods. Moreover, these datasets focus on large, everyday objects with distinct textures, whereas surgical instruments are often smaller and have featureless, highly reflective metallic surfaces. This results in noisy and incomplete object point clouds when captured with RGB-D cameras. Therefore, in a surgical setting, the annotation approaches proposed in [11,17] are less stable and reliable. Pioneer work in [14] introduces a small synthetic dataset with blue surgical gloves and a surgical drill, following the synthetic data generation approach in [13]. However, being a single-image dataset, it ignores the strong temporal context in surgical tasks, which is crucial for accurate and reliable 3D pose estimation [17,24]. Surgical cases have inherent task-specific information and temporal correlations during surgical instrument usage, such as cutting firmly and steadily with a scalpel. Moreover, it lacks diversity, focusing only on one unbloodied blue surgical glove and one instrument, and only provides low-resolution image patches.

To fill this gap, we propose a novel synthetic data generation pipeline that goes beyond single image cases to synthesize realistic temporal sequences of surgical tool manipulation from an egocentric perspective. It features a body motion capture module to model realistic body movement sequences during artificial surgeries and a hand-object manipulation generation module to model the grasp evolution sequences. With the proposed pipeline, we generate POV-Surgery, a large synthetic egocentric video dataset of surgical activities that features surgical gloves in diverse textures (green, white, and blue) with various bloodstain patterns and three metallic tools that are commonly used in orthopedic surgeries.

In summary, our contributions are:

- A novel, easy-to-use, and generalizable synthetic data generation pipeline to generate temporally realistic hand-object manipulations during surgical activities.
- POV-Surgery: the first large-scale dataset with egocentric sequences for hand and surgical instrument pose estimation, with diverse, realistic surgical glove

textures, and different metallic tools, annotated with accurate 3D/2D hand-object poses and 2D hand-object segmentation masks.
– Extensive evaluations of existing state-of-the-art (SOTA) hand pose estimation methods on POV-Surgery, revealing their shortcomings when dealing with the unique challenges in surgical cases from egocentric view.
– Significantly improved performance for SOTA methods after fine-tuning them on the POV-Surgery training set, on both our synthetic test set and a *real-life test set*.

2 Method

We focus on three tools commonly employed in orthopedic procedures - the scalpel, friem, and diskplacer - each of which requires a unique hand motion. The scalpel requires a side-to-side cutting motion, while the friem uses a quick downward punching motion, similar to using an awl. Finally, the diskplacer requires a screwing motion with the hand. Our pipeline to capture these activities and generate egocentric hand-object manipulation sequences is shown in Fig. 1.

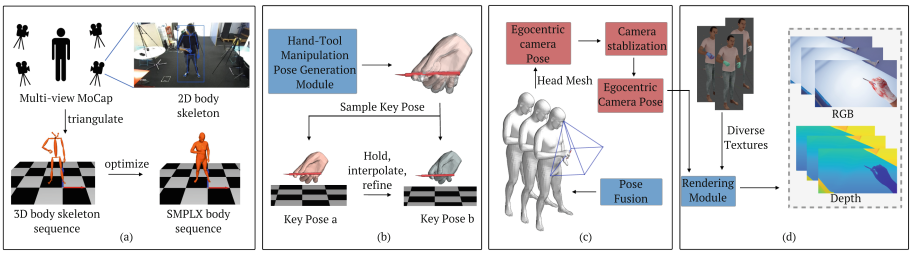

Fig. 1. The proposed pipeline to generate synthetic data sequences. (a) shows the multi-stereo-cameras-based body motion capture module. (b) indicates the optimization-based hand-object manipulation sequence generation pipeline. (c) presents the fused hand-body pose and the egocentric camera pose calculation module. (d) highlights the rendering module with which RGB-D sequences are rendered with diverse textures.

2.1 Multi-view Body Motion Capture

To capture body movements during surgery, we used four temporally synchronized ZED stereo cameras on participants during simulated surgeries. The intrinsic camera parameters were provided by ZED SDK and the extrinsic parameters between the four different cameras were calibrated with a chessboard. We adopt the popular [5] [6] module for SMPLX body reconstruction. OpenPose [2] with hand - and face-detection modules is first used to detect 2D human skeletons with a confidence threshold of 0.3. The 3D keypoints are obtained via triangulation with camera pose, regularized with bone length. The SMPLX body meshed

is optimized by minimizing the 2D re-projection and triangulated 3D skeleton errors. Moreover, we enforce a large smoothness constraint, which regularizes the body and hand pose by constraining the between-frame velocities. It vastly reduces the number of unrealistic body poses.

2.2 Hand-Object Manipulation Sequence Generation

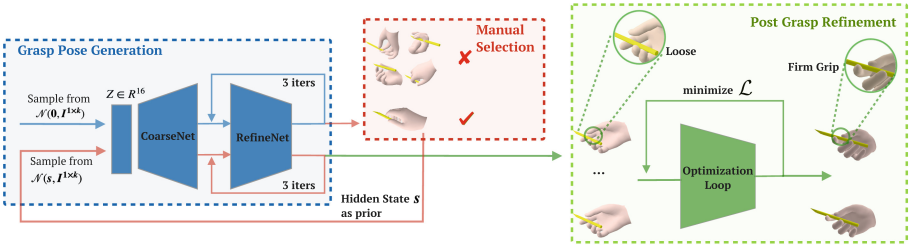

Fig. 2. The Hand manipulation sequence generation pipeline consists of three components: grasp pose generation, pose selection, and pose refinement, highlighted in blue, red, and green, respectively. (Color figure online)

There are two critical differences between the surgical tool and everyday object grasping: surgical tools require to be held in specific poses. Moreover, a surgeon would hold firmly and steadily with a particular pose for some time span during surgeries.

To address this issue, we generate each instrument manipulation sequence by firstly modeling the key poses that are surgically plausible, and then interpolating in between to model pose evolution. The key pose generation pipeline is shown in Fig. 2. The part highlighted in blue is the pose generation component based on GrabNet [25]. We provide the 3D instrument models to GrabNet with arbitrary initial rotation and sample from a Gaussian distribution in latent space to obtain diverse poses. 500 samples are generated for scalpel, diskplacer, and friem, respectively, followed by manual selection to get the best grasping poses as templates. With a pose template as prior, we perform the re-sampling near it to obtain diverse and similar hand-grasping poses as key poses for each sequence. To improve the plausibility of the grasping pose and hand-object interactions, inspired by [16], an optimization module is adopted for post-processing, with the overall loss function defined as:

$$\mathcal{L} = \alpha \cdot L_{penetr} + \beta \cdot L_{contact} + \gamma \cdot L_{keypoint}, \tag{1}$$

L_{penetr}, $L_{contact}$, and $L_{keypoint}$ denote the penetration loss, contact loss, and keypoint loss, respectively. And α, β, γ are object-specific scaling factors to balance the loss components. For example, the weight for penetration is smaller for the scalpel than the friem and diskplacer to account for the smaller object

size. The penetration loss is defined as the overall penetration distance of the object into the hand mesh:

$$L_{penetr} = \frac{1}{|\mathcal{P}_{in}^{o}|} \sum_{p \in \mathcal{P}_{in}^{o}} \min_{i} \|p - \mathcal{V}_i\|_2^2 , \tag{2}$$

where $\mathcal{P}_{in}$ denotes the vertices from the object mesh which are inside the hand mesh, and $\mathcal{V}_i$ denotes the hand vertex. The $\mathcal{P}_{in}^{o}$ is defined as the dot product of the vector from the hand mesh vertices to their nearest neighbors on the object mesh. To encourage hand-object contact, a contact loss is defined to minimize the distance from the hand mesh to the object mesh.

$$L_{contact} = \sum_{j} \min_{i} \|\mathcal{V}_i - \mathcal{P}_j\|_2^2, \tag{3}$$

where $\mathcal{V}$ and $\mathcal{P}$ denote vertices from the hand and object mesh, respectively. In addition, we regularize the optimized hand pose by the keypoint displacement, which penalizes hand keypoints that are far away from the initial hand keypoints:

$$L_{keypoint} = \sum_{i} \|K_i - k_i\|^2, \tag{4}$$

where K is the refined hand keypoint position and k is the source keypoint position. After the grasping pose refinement, a small portion of the generated hand poses are still unrealistic due to the poor initialization. To this end, a post-selection technique similar to [13,26] is further applied to discard the unrealistic samples with hand-centric interpenetration volume, contact region, and displacement simulation.

For each hand-object manipulation sequence, we select 30 key grasping poses, hold on, and interpolate in between to model pose evolution within the sequence. The number of frames for the transition phase between every two key poses is randomly sampled from 5 to 30. The interpolated hand poses are also optimized via the pose refinement module with the source keypoint in $L_{keypoint}$ defined as the interpolated keypoints between two key poses.

2.3 Body and Hand Pose Fusion and Camera Pose Calculation

In previous sections, we individually obtained the body motion and hand-object manipulation sequences. To merge the hand pose into the body pose to create a whole-body grasping sequence, we established an optimization-based approach based on the SMPLX model. The vertices to vertices loss is defined as:

$$L_{V2V} = \sum_{\hat{v}_i \in P_{hand}} \left\| v_{M(i)} - (R\hat{v}_i + T) \right\|_2^2 , \tag{5}$$

where $\hat{v}$ is the vertices in the target grasping hand, v is the vertices in the SMPLX body model, with M being the vertices map from MANO's right hand to SMPLX

body. R and T are the rotation matrix and translation vector applied to the right hand. The right-hand pose of SMPLX, R, and T are optimized with the Trust Region Newton Conjugate Gradient method (Trust-NCG) for 300 iterations to obtain an accurate and stable whole-body grasping pose. R and T are then applied to the grasped object. The egocentric camera pose for each frame is calculated with head vertices position and head orientation. Afterwards, outlier removal and moving average filter are applied to the camera pose sequence to remove temporal jitterings between frames.

2.4 Rendering and POV-Surgery Dataset Statistics

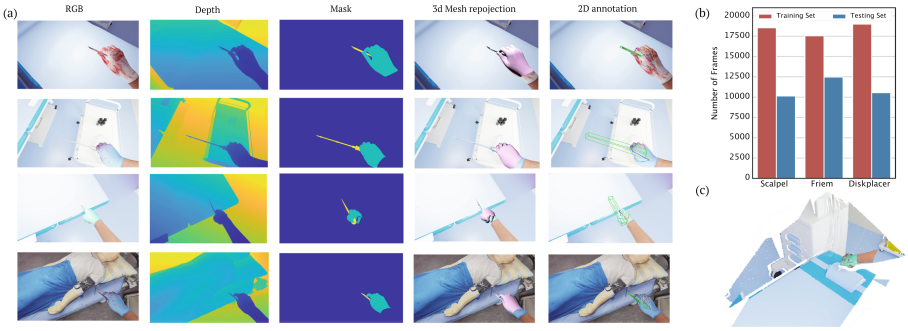

Fig. 3. (a) Dataset samples for RGB-D sequences and annotation. An example of the scalpel, friem, and diskplacer, is shown in the first three rows. The fourth row shows an example of the new scene and blood glove patterns that only appear in the test set. (b) shows the statistics on the number of frames for each surgical instrument in the training and testing sets. (c) shows a point cloud created from an RGB-D frame with simulated Kinect noise.

We use blender [3] and bpycv packages to render the RGB-D sequences and instance segmentation masks. Diverse textures and scenes of high quality are provided in the dataset: it includes 24 SMPLX textures featuring blue, green, and white surgical gloves textures with various blood patterns and a synthetic surgical room scene created by artists. The Cycle rendering engine and de-noising post-processing are adopted to produce high-quality frames. POV-Surgery provides clean depth maps for depth-based methods or point-cloud-based methods, as the artifact of real depth cameras can be efficiently simulated via previous works as [12]. A point cloud example generated from an RGB-D frame with added simulated Kinect depth camera noise is provided in Fig. 3. The POV-Surgery dataset consists of 36 sequences with 55,078 frames in the training set and 17 sequences with 33,161 frames in the testing set, respectively. Three bloodied glove textures and one scene created from a room scanning of a surgery room are only used in the testing set to measure generalizability. Fig. 3 shows the ground truth data samples and the dataset statistics.

3 Experiment

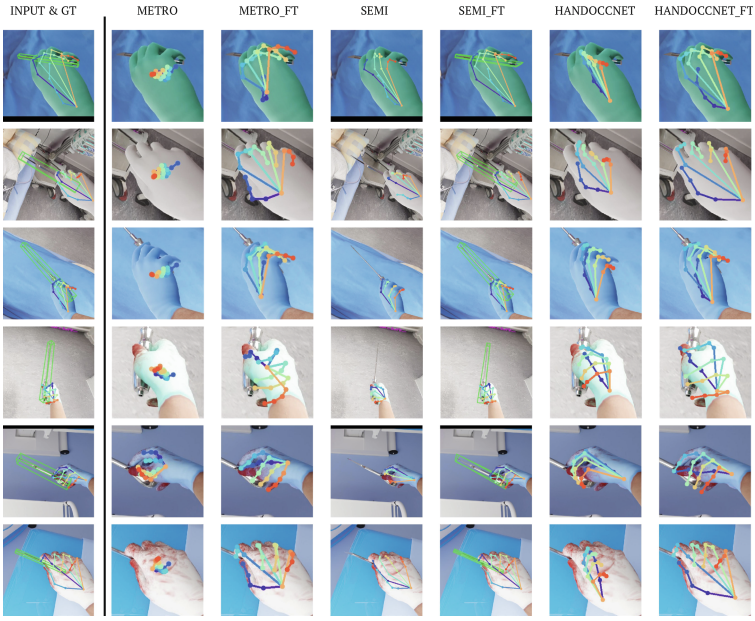

Fig. 4. Qualitative results of METRO [18], SEMI [19], and HANDOCCNET [22] on the test of of POV-Surgery. The FT denotes fine-tuning. We show the 2D re-projection of the predicted 3D hand joints and object control bounding box overlayed on the input image.

We evaluate and fine-tune two state-of-the-art hand pose estimation methods: [18,22], and one hand-object pose estimation [19] method on our dataset with provided checkpoints in their official repositories. 6 out of 36 sequences from the training set are selected as the validation set for model selection. We continue to train their checkpoints on our synthetic training set, with a reduced learning rate (10^{-5}) and various data augmentation methods such as color jittering, scale and center jittering, hue-saturation-contrast value jittering, and motion blur for better generalizability. Afterwards, we evaluate the performance of those methods on our testing set. We set a baseline for object control point error in pixels: 41.56 from fine-tuning [19]. The hand quantitative metrics are shown in Table 1 and qualitative visualizations are shown in Fig. 4, where we highlight the significant performance improvement for existing methods after fine-tuning them on the POV-Surgery dataset.

To further evaluate the generalizability of the methods fine-tuned on our dataset, we collect 6,557 real-life images with multiple surgical gloves, tools, and backgrounds as the *real-life test set*. The data capture setup with four stereo

Table 1. The evaluation result of different methods on the test set of POV-Surgery, where the $_{ft}$ denotes fine-tuned on the training set. P_{2d} denotes the 2D hand joint re-projection error (in pixels). MPJPE and PVE denote the 3D Mean Per Joint Position Error and Per Vertex Error, respectively. PA denotes procrustes alignment.

Method	P_{2d} ↓	MPJPE ↓	PVE ↓	PA-MPJPE ↓	PA-PVE ↓
METRO [18]	95.11	77.46	75.06	23.43	22.34
SEMI [19]	77.91	115.67	112.10	12.68	12.76
HandOCCNet [22]	64.70	95.19	90.83	11.71	11.13
METRO$_{ft}$	30.49	14.90	13.80	6.36	4.34
SEMI$_{ft}$	**13.42**	15.14	14.69	**4.29**	**4.23**
HandOCCNet$_{ft}$	13.80	**14.35**	**13.73**	4.49	4.35

cameras is shown in Fig. 5. We adopt a top-down-based method from [4] with manually selected hand bounding boxes for 2D hand joint detection. [5] is used to reconstruct 3D hand poses from different camera observations. We project the hand pose to the egocentric camera view and manually select the frames with accurate hand predictions to obtain reliable 2D hand pose ground truth. We show quantitative examples of the indicated methods and the PCP curve in Fig. 5. After fine-tuning on our synthetic dataset significant performance improvements are achieved for SOTA methods on the *real-life test set*. Particularly, we observe a similar performance improvement for unseen purple-texture gloves, showing the effectiveness of our POV-Surgery dataset towards the challenging egocentric surgical hand-object interaction scenarios in general.

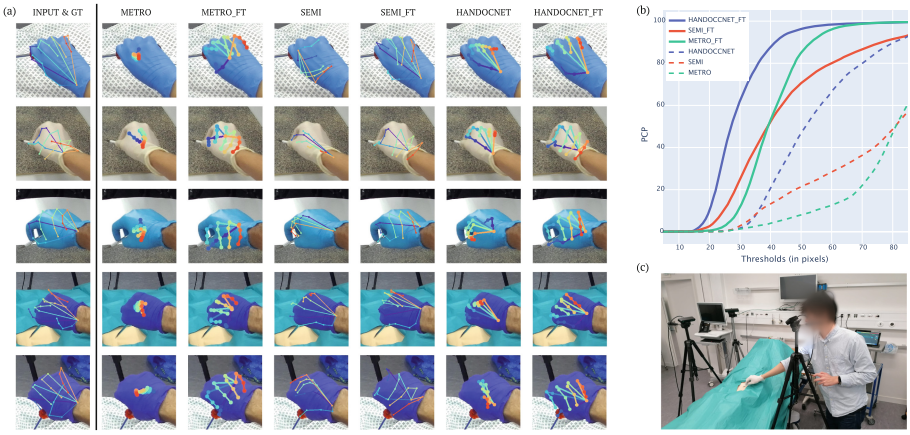

Fig. 5. (a) Ground truth and qualitative results of different methods on the *real-life test set*. (b) Accuracy with different 2D pixel error thresholds, showing large performance improvement after fine-tuning on POV-Surgery (c) Our multi-camera real-life data capturing set-up.

4 Conclusion

This paper proposes a novel synthetic data generation pipeline that generates hand-tool manipulation temporal sequences. Using the data generation pipeline and focusing on three tools used in orthopedic surgeries: scalpel, diskplacer, and friem, we propose a large, synthetic, and temporal dataset on egocentric surgical hand-object pose estimation, with 88,329 RGB-D frames and diverse bloody surgical gloves patterns. We evaluate and fine-tune three current state-of-the-art methods on the POV-Surgery dataset. We prove the effectiveness of the synthetic dataset by showing the significant performance improvement of the SOTA methods in real-life cases with surgical gloves and tools.

Acknowledgement. This work is part of a research project that has been financially supported by Accenture LLP. Siwei Zhang is funded by Microsoft Mixed Reality & AI Zurich Lab PhD scholarship. The authors would like to thank PD Dr. Michaela Kolbe for providing the simulation facilities and the students participating in motion capture.

References

1. Azimi, E., et al.: An interactive mixed reality platform for bedside surgical procedures. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 65–75. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_7
2. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: OpenPose: real-time multi-person 2D pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
3. Community, B.O.: Blender - a 3D modelling and rendering package. Blender Foundation, Stichting Blender Foundation, Amsterdam (2018). http://www.blender.org
4. Contributors, M.: OpenMMLab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose (2020)
5. Dong, J., Fang, Q., Jiang, W., Yang, Y., Bao, H., Zhou, X.: EasyMocap - make human motion capture easier. Github (2021). https://github.com/zju3dv/EasyMocap
6. Dong, J., Fang, Q., Jiang, W., Yang, Y., Bao, H., Zhou, X.: Fast and robust multi-person 3D pose estimation and tracking from multiple views. In: T-PAMI (2021)
7. Doughty, M., Singh, K., Ghugre, N.R.: SurgeonAssist-Net: towards context-aware head-mounted display-based augmented reality for surgical guidance. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 667–677. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_64
8. Fattahi Sani, M., Ascione, R., Dogramadzi, S.: Mapping surgeons hand/finger movements to surgical tool motion during conventional microsurgery using machine learning. J. Med. Robot. Res. **6**(03n04), 2150004 (2021)
9. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.K.: First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 409–419 (2018)
10. Goodman, E.D., et al.: A real-time spatiotemporal AI model analyzes skill in open surgical videos. arXiv preprint arXiv:2112.07219 (2021)

11. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: HOnnotate: a method for 3D annotation of hand and object poses. In: CVPR (2020)
12. Handa, A., Whelan, T., McDonald, J., Davison, A.J.: A benchmark for RGB-D visual odometry, 3D reconstruction and slam. ICRA (2014)
13. Hasson, Y., et al.: Learning joint reconstruction of hands and manipulated objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11807–11816 (2019)
14. Hein, J., et al.: Towards markerless surgical tool and hand pose estimation. Int. J. Comput. Assist. Radiol. Surgery **16**(5), 799–808 (2021). https://doi.org/10.1007/s11548-021-02369-2
15. Jian, Z., Yue, W., Wu, Q., Li, W., Wang, Z., Lam, V.: Multitask learning for video-based surgical skill assessment. In: 2020 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–8. IEEE (2020)
16. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: Proceedings of the International Conference on Computer Vision (2021)
17. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: 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)
18. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: CVPR (2021)
19. Liu, S., Jiang, H., Xu, J., Liu, S., Wang, X.: Semi-supervised 3D hand-object poses estimation with interactions in time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14687–14697 (2021)
20. Moon, G., Yu, S.I., Wen, H., Shiratori, T., Lee, K.M.: InterHand2. 6M: a dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In: Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 548–564. Springer (2020). https://doi.org/10.1007/978-3-030-58565-5_33
21. Palumbo, M.C., et al.: Mixed reality and deep learning for external ventricular drainage placement: a fast and automatic workflow for emergency treatments. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII, pp. 147–156. Springer (2022). https://doi.org/10.1007/978-3-031-16449-1_15
22. Park, J., Oh, Y., Moon, G., Choi, H., Lee, K.M.: HandOccNet: occlusion-robust 3D hand mesh estimation network. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
23. Saggio, G., et al.: Objective surgical skill assessment: an initial experience by means of a sensory glove paving the way to open surgery simulation? J. Surg. Educ. **72**(5), 910–917 (2015)
24. Sener, F., et al.: Assembly101: a large-scale multi-view video dataset for understanding procedural activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21096–21106 (2022)
25. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: a dataset of whole-body human grasping of objects. In: European Conference on Computer Vision (ECCV) (2020). https://grab.is.tue.mpg.de
26. Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., Gall, J.: Capturing hands in action using discriminative salient points and physics simulation. Int. J. Comput. Vis. **118**(2), 172–193 (2016)

27. Wesierski, D., Jezierska, A.: Instrument detection and pose estimation with rigid part mixtures model in video-assisted surgeries. Med. Image Anal. **46**, 244–265 (2018)
28. Wolf, J., Luchmann, D., Lohmeyer, Q., Farshad, M., Fürnstahl, P., Meboldt, M.: How different augmented reality visualizations for drilling affect trajectory deviation, visual attention, and user experience. Int. J. Comput. Assist. Radiol. Surgery, 1–9 (2023)
29. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: FreiHAND: a dataset for markerless capture of hand pose and shape from single RGB images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 813–822 (2019)