Reconstructing 3D Hand-Instrument Interaction from a Single 2D Image in Medical Scenes

Miao Xu[1,2], Xiangyu Zhu[1], Jinlin Wu[2], Ming Feng[3], Zelin Zang[2], Hongbin Liu[1,2], and Zhen Lei[1,2,4⋆]

[1] Institute of Automation, Chinese Academy of Sciences, China
[2] Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science, Hong Kong SAR
[3] Peking Union Medical College Hospital, China
[4] School of Artificial Intelligence, University of Chinese Academy of Sciences, China
xumiao2021@ia.ac.cn

Abstract. Capturing the hand movements of physicians and their interactions with medical instruments plays a critical role in behavior analysis and surgical skill assessment. However, hand-instrument interaction in medical contexts is far more challenging than in general tasks. The weak texture and reflective properties of surgical instruments frequently result in failures in pose estimation. Moreover, the long and thin shape characteristics of the instruments and the sparse points of the reconstructed hand lead to difficulties in accurately grasping the instrument or may result in spatial penetration during interaction. To address failures in pose estimation, we build 3D models of medical instruments as priors to optimize instrument pose estimation. To resolve the issues of inaccurate grasping and minimize spatial penetration, we propose a contact-point-centered interaction module by refining the surface details of the fingers to optimize the hand-instrument relationship computation. Experiments on medical scenario data demonstrate that our method achieves state-of-the-art performance across multiple evaluation metrics. Additionally, the 3D models developed in this work encompass a wide range of surgical instruments, based on real medical devices, and we will release them at https://github.com/xumiao66/MedIns-3D to support and promote further research.

Keywords: Hand-object Interaction · Hand Pose · Deep Learning.

1 Introduction

The integration of medical practices and artificial intelligence has garnered growing attention from researchers. Computer-assisted interventions, which have diverse applications including surgical navigation systems [28], objective evaluation of surgical skills [6,12,24], and advancements in robot-assisted surgery [5], rely heavily on the precise analysis of surgical instrument trajectories and surgeon

[⋆] Corresponding author: zlei@nlpr.ia.ac.cn hand movements. This analysis is crucial for optimizing the effectiveness and accuracy of such interventions. At the same time, surgical instrument trajectories and surgeon hand movements are complementary and inseparable, sharing a semantic relationship within the surgical space. Only by integrating the two and representing them in an interactive state can surgical skills be better demonstrated, thereby enhancing computer-assisted analysis.

In the domain of computational artificial intelligence, the existing methods for estimating the 3D pose of hands and objects using 2D images can be effectively segregated into two primary classifications: optimization-based methods[1,10,25,29] and learning-based techniques [3,17,4,9,8]. Optimization-based approaches refine the pose of both the hand and object by considering their contact surfaces while adhering to physical constraints such as attraction and repulsion. The process of estimating the contact surface between the hand and object typically proves to be time-intensive [7]. To tackle this challenge, Tse *et al.* [25] introduced a graph-based network to expedite the estimation of the contact surface. Contrastingly, learning-based methodologies devise integrated models for simultaneous estimation of the poses of hands and objects. These methods commonly leverage a readily available hand model, such as MANO [23], and rely on the prior availability of a 3D object model. Consequently, they can directly predict the poses of the hand and object based on these foundational principles. Initial efforts [3,8] incorporate dual-stream backbones for independent estimation of hand and object poses, albeit at the expense of heightened model complexity. The objects explored in these fields are mainly bottles, bowls, and similar items, where pose relationships are relatively easy to compute, and the hand-object interactions are clearer.

However, these methods are not robust enough to be applied in medical scenarios. In a single hand image, particularly in real surgical scenarios, hands often encounter various occlusions, including self-occlusions and occlusions caused by the instrument, making hand motion regression exceedingly challenging. For medical instruments, characteristics such as weak textures, reflective surfaces, and slender structures render many general object pose estimation methods unsuitable for direct application. Even after addressing these challenges, establishing meaningful interaction between the hand and medical instruments remains difficult. In medical scenarios, hand-instrument interaction needs to ensure that the hand securely grasps the instrument and avoids spatial misalignment and penetration errors, while the long and thin shape characteristics of instruments lead to significant challenges.

To address the aforementioned challenges, we construct a medical instrument 3D model dataset (MedIns-3D), including surgical scalpels, scissors, and forceps as priors to constrain the pose estimation pipeline. Subsequently, we proposed a contact-point-centered interaction framework (CPCI) for reconstructing hand-instrument interaction, eliminating the need for post-processing. In this framework, we utilize the MANO [23] hand model as a prior to reconstruct hands of physicians in images of complex medical scenarios, obtaining precise hand motions and poses. For instruments, we initialize the estimation process using

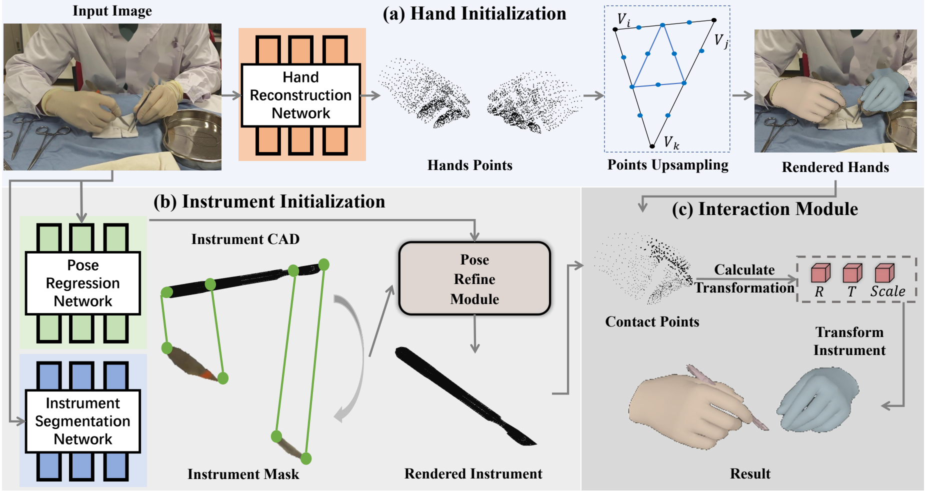

Fig. 1. Brief view of our method. It is a Hand-instrument Interaction network that comprises an Instrument Initialization module for extracting instrument pose, a Hand Initialization module for reconstructing hands, and an Interaction Module for transforming the instrument into the hand space for interaction.

the 3D models from our constructed dataset, enabling accurate pose estimation under weak textures and strong reflectivity. Subsequently, we propose a contact-point-centered interaction module to estimate the contact points between the hand and instruments, where a point up-sampling strategy is employed to constrain the fine-grained hand-instrument relationship computation by refining the surface details of the fingers.

The contributions of this work can be summarized as follows:

- We propose a contact-point-centered interaction (CPCI) framework that jointly reconstructs both hands, estimates instrument poses, and explicitly models hand-instrument interactions in a unified manner.
- We introduce a hand mesh upsampling strategy to refine finger surface geometry, enabling more accurate contact region modeling and improving interaction fidelity.
- We construct and will release a collection of 3D surgical instrument models (MedIns-3D), which serve as geometric priors for pose estimation and can benefit future research in surgical scene understanding.

2 Methodology

2.1 Hand Initialization

In this paper, the parametric hand model MANO [23] is utilized to set the initial hand geometry. This model proficiently translates the pose parameter $\theta \in R^{J\times3}$, where J represents per-bone parts, and the shape parameter $\beta \in R^{10}$ onto a template mesh $\hat{M}$ comprising vertices V. This mapping, denoted as ω, relies on linear blending skinning with associated weights $W \in R^{|V| \times J}$. The resultant posed hand mesh M can be derived using the following expression:

$$M = \omega(\hat{M}, W, \theta, \beta). \tag{1}$$

Physicians often operate with one hand holding a medical instrument, yet images frequently depict both hands. If both hands overlap with the instrument in the image, ambiguity arises, complicating the determination of contact points. To address this problem, we perform the reconstruction of both hands simultaneously inspired by the method [30]. As illustrated in Figure 1, given an input image I, we first use a CNN backbone $\mathcal{E}$ to extract image features F. These features are then disentangled into two feature maps, representing the left and right hands separately:

$$\mathbf{F}_l, \mathbf{F}_r = \mathcal{E}(I). \tag{2}$$

From these two feature maps, we independently regress the parameters for both hands as well as the weak perspective camera parameters:

$$P_l, P_r, P_c = \mathcal{R}(\mathbf{F}_l, \mathbf{F}_r), \tag{3}$$

where P_l and P_r represent the left and right hands parameters. P_c denotes the camera parameters. $\mathcal{R}$ is the regression module. Finally, the parameters are fed into the MANO layer to generate 3D hand models. The original MANO mesh [23], which comprises 778 vertices and 1538 faces, possesses a constrained capacity to accurately represent nuanced details [2]. To refine the surface details of the fingers and obtain more accurate interaction points for subsequent interaction operations, we then split the vertices of fingers by taking the midpoints of the three edges of each face three times, thereby upsampling the vertices.

2.2 Instrument Initialization

For medical instruments, which are characterized by their slender shapes, weak textures, and reflective surfaces, reconstructing their 3D shape from monocular images is highly challenging. Even if a coarse 3D shape is reconstructed, it often negatively impacts subsequent pose estimation results.

To optimize the pose estimation of instruments and achieve more precise hand-instrument interaction results, detailed 3D models of medical instruments are required as priors. While datasets such as MedShapeNet [14] have released a wide range of 3D medical shapes, they primarily focus on anatomical and diagnostic structures, and do not provide sufficiently detailed or interaction-ready models of handheld surgical instruments. Therefore, we collect a medical instrument 3D model dataset (MedIns-3D), consisting of 36 different types of medical instruments, including scalpels, surgical scissors, scalpel handle, forceps, and surgical blades. Subsequently, we modeled these instruments to obtain a series of 3D models with textures. Figure 2 shows some examples from the dataset.

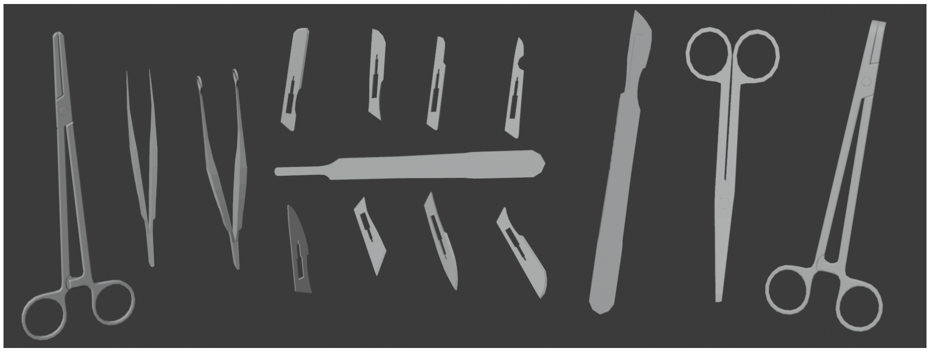

Fig. 2. Examples of the 3D instrument models we have constructed, which are based on real medical instruments and include detailed texture.

Then we directly utilize the constructed 3D model of the instrument V_o as initialization. By leveraging the accurate 3D model as a prior for the BundleSDF [27], we optimize the pose estimation process to obtain the precise instruments pose of each frame P_{v_o} with high speed:

$$P_{v_o} = \mathcal{B}(I, V_o), \tag{4}$$

where $\mathcal{B}$ is the pose estimation module. I denote the input image. The other branch segments the mask of the instrument. Subsequently, the instrument model is projected back to the image based on the pose obtained earlier. The pose is refined iteratively by minimizing the error between the projected model and the mask.

2.3 Contact-Point-Centered Interaction Module

After the hand and object initializations, we obtain the point clouds of the hand and the instrument, along with their respective transformation matrices relative to the image coordinate system. However, the hand and the instrument are not defined within the same 3D coordinate system. While their projections might appear overlapped in the 2D image plane, placing them in the same 3D space reveals inconsistencies in scale and discrepancies in pose alignment. These issues hinder accurate modeling of hand-instrument interactions and must be resolved to achieve realistic and semantically meaningful results. A contact-point-centered interaction module (CPCI) is designed to combine hands and instruments.

We render the previously obtained 3D hand into the image coordinate system and get 2D hands vertices V_l^{2d} and V_r^{2d}:

$$V_l^{2d} = Render(V_l, P_c), V_r^{2d} = Render(V_r, P_c). \tag{5}$$

It is important to note that the left and right hands are encoded separately, with left-hand points assigned a value of 0 and right-hand points assigned a value of

1. The instrument is also projected into the image coordinate system and gets 2D instrument vertices V_o^{2d}:

$$V_o^{2d} = Render(V_o, P_{v_o}). \tag{6}$$

Then we calculate the 2D overlap between the left and right hands and the instrument. Since physicians typically manipulate instruments using their fingers, the overlap here is computed based solely on the fingertips and the instrument. The overlap area is used to determine whether the left or right hand is involved in the interaction. We then match the points in the overlapping regions based on the closest distance, obtaining a set of corresponding hand-instrument points $\mathbf{X}^h$ and $\mathbf{X}^i$. $\mathbf{X}^h = \{\mathbf{x}_1^h, \mathbf{x}_2^h, ..., \mathbf{x}_n^h\}$ is the set of n 3D points on the reference hand 3D model, $\mathbf{X}^i = \{\mathbf{x}_1^i, \mathbf{x}_2^i, ..., \mathbf{x}_n^i\}$ is the set of n 3D points on the instrument model. Utilizing these corresponding points, we compute the rotation angle R, translation offset T, and a scaling factor S by Least Squares Method. Through S, the scale of the instrument is adjusted. Meanwhile, after corresponding points are established, the module ensures that the instrument is appropriately aligned with the hand and that no physical penetration occurs in 3D space. It continuously monitors the interaction dynamics and adjusts the hand or instrument model if necessary, ensuring that the grasp is physically plausible and semantically meaningful. Finally, the 3D model of the instrument is transformed through R and T, aligning it to the 3D space of hands.

3 Experiments

3.1 Implementation Details

We implement our network using PyTorch. For the backbone architecture of the hand reconstruction network, we trained with ResNet-50 [11]. For a monocular raw RGB input, no cropping or detection is needed; instead, all input images and segmentation maps are resized to 512×512, preserving the original aspect ratio through zero padding. During training, we supervised the model using the L2 loss on both MANO parameters and vertex distances. For the pose regression network and instrument segmentation network, we employ BundleSDF [27] and Segment Anything [13], respectively. Notably, we modify the BundleSDF framework by utilizing the established instrument models to replace the reconstruction step, and output the pose based solely on the current frame.

3.2 Datasets and Metrics

Datasets. We primarily conduct experiments on the POV-Surgery [26] dataset to validate the effectiveness of our method. To evaluate the model's generalization ability, three bloodied glove textures and a scene generated from a 3D scan of a surgical room are exclusively included in the test set.

Metrics. For the hand mesh recovery, our primary metric is the mean pervertex position error (PVE). Additionally, we employ Procrustes Analysis (PA)

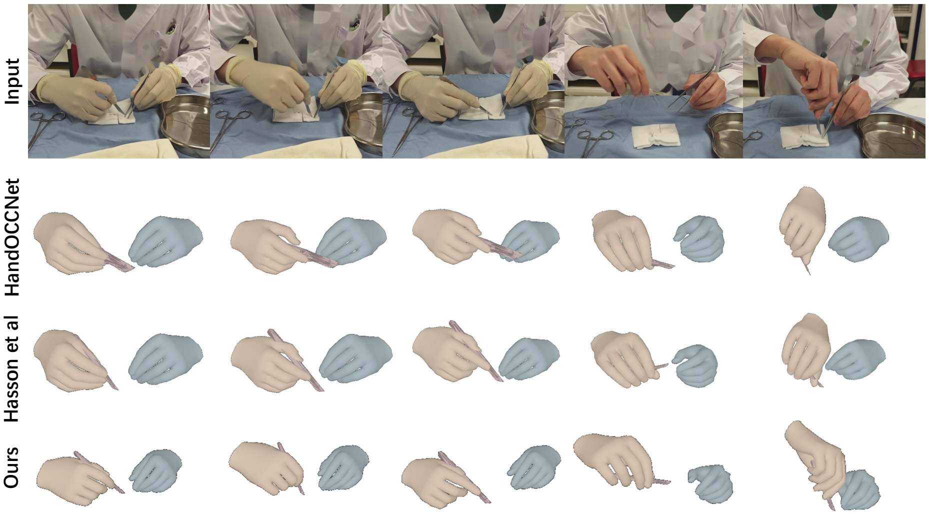

Fig. 3. Qualitative results on in-the-wild videos. Our method outperforms in handling scenarios involving glove-wearing and various gripping techniques.

Table 1. The evaluation result of different methods on the test set of POV-Surgery, including Solid Intersection Volume (IV) and Penetration Depth, measuring penetrations, Proximity Error, evaluating the difference of the hand-object proximity, and HO Motion Consistency, assessing the hand-object motion consistency.

Method	IV↓	Penetration Depth↓	Proximity Error↓	HO Motion Consistency↓
HandOCCNet [19]	2.31	2.60	3.08	21.35
HandOCCNet+TOCH [31]	3.04	2.19	3.14	4.42
Hasson *et al.* [9]	1.96	2.01	3.19	7.26
Hasson *et al.* [10]	1.78	1.85	2.98	4.13
HOISDF [22]	1.52	1.65	2.87	1.85
CPCI w/o UP (ours)	1.14	1.63	2.23	0.50
CPCI (ours)	**1.12**	**1.62**	**2.21**	**0.48**

on the reconstructed mesh and report the PA-PVE after rigid alignment. We also report the mean per joint position error (MPJPE) along with PA-MPJPE and hand error is computed as the mean error between the left and right hands [26]. For the interaction between the hand and instrument, we employ several quantifiers, including Solid Intersection Volume (IV) and Penetration Depth to measure penetrations, Proximity Error to assess the discrepancy in hand-object proximity, and Hand-Object (HO) Motion Consistency to evaluate the consistency of the hand-object motion.

3.3 Hand-instrument Interaction Comparison

To validate the effectiveness of our method in hand-instrument interaction, we compare it with previous state-of-the-art hand-object interaction methods. It is

Table 2. The evaluation result of different methods on the POV-Surgery. P_{2d} denotes the 2D hand joint reprojection error. MPJPE and PVE denote the 3D Mean Per Joint Position Error and Per Vertex Error, respectively. PA denotes Procrustes alignment.

Method	$P_{2d}\downarrow$	MPJPE↓	PVE↓	PA-MPJPE↓	PA-PVE↓
METRO [15]	30.49	14.90	13.80	6.36	4.34
HandTailor [18]	25.42	13.20	12.48	5.89	4.19
Mesh Graphormer [16]	20.36	12.75	12.68	5.46	4.32
WiLoR [21]	18.48	13.72	12.91	4.33	4.20
SimpleHand [32]	16.52	13.45	12.61	4.32	4.19
SEMI [17]	13.42	15.14	14.69	4.29	4.23
HandOCCNet [19]	13.80	14.35	13.73	4.49	4.35
HaMeR [20]	13.05	13.15	12.55	4.41	**4.18**
CPCI (ours)	**12.08**	**12.21**	**12.25**	**4.21**	4.20

important to note that, in order to ensure fairness, we have replaced the object models used in previous methods with scanned models. HandOCCNet [19] demonstrates excellent performance in hand reconstruction while holding objects. Based on this method, we obtain the pose of the instrument through BundleSDF after reconstruction to enable interaction. Additionally, we use TOCH [31] for post-processing to further conduct comparative analysis. Table 1 reports the performance of our method in Solid Intersection Volume (IV), Penetration Depth, measuring penetrations, and Proximity Error. Our method outperforms other approaches. In Figure 3, we present the qualitative results of our method, and our method avoids spatial penetration and failed grasping.

3.4 Hand Mesh Recovery Comparison

To validate the effectiveness of our method in hand mesh recovery, we compare it with previous methods based on MANO on the POV-Surgery dataset. To ensure fairness, every method is finetuned on the POV-Surgery dataset as the POV-Surgery dataset consists of first-person perspective images with bloody gloves. Table 2 reports the performance of our method in MPJPE, PVE, PA-MPJPE, and PA-PVE. Our method performs well across all metrics.

3.5 Ablation Study

To achieve more stable hand-instrument interaction results, we adopt an upsampling strategy to densify the sparse hand point cloud, thereby obtaining more accurate contact points. To validate the effectiveness of this approach, we conducted another baseline CPCI w/o UP by matching only 778 hand points after projection, with all other aspects remaining the same. The results are presented in Table 1, where it is evident that sparse hand points are insufficient for reliably obtaining stable hand-instrument contact points, leading to larger errors.

4 Conclusion

In this paper, we first build a medical instrument 3D model dataset (MedIns-3D) as priors to constrain the pose estimation pipeline. Second, we also develop an innovative framework that simultaneously reconstructs physicians' hands and estimates the pose of medical instruments. To ensure the hand can grasp the instrument accurately while minimizing spatial penetration as much as possible, we propose a contact-point-centered interaction module to estimate the contact points between the hand and instruments, where a point up-sampling strategy is employed. Finally, extensive experiments validate the effectiveness of our approach, demonstrating superior performance compared to existing methods.

Acknowledgements. This work was supported in part by Chinese National Natural Science Foundation Projects 62276254, U23B2054, the Beijing Science and Technology Plan Project Z231100005923033, Beijing Natural Science Foundation L221013, and InnoHK program.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12417–12426 (2021)
2. Chen, X., Wang, B., Shum, H.Y.: Hand avatar: Free-pose hand animation and rendering from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8683–8693 (2023)
3. Chen, Y., Tu, Z., Kang, D., Chen, R., Bao, L., Zhang, Z., Yuan, J.: Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion. IEEE Transactions on Image Processing **30**, 4008–4021 (2021)
4. Doosti, B., Naha, S., Mirbagheri, M., Crandall, D.J.: Hope-net: A graph-based model for hand-object pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6608–6617 (2020)
5. Fattahi Sani, M., Ascione, R., Dogramadzi, S.: Mapping surgeons hand/finger movements to surgical tool motion during conventional microsurgery using machine learning. Journal of Medical Robotics Research **6**(03n04), 2150004 (2021)
6. Goodman, E.D., Patel, K.K., Zhang, Y., Locke, W., Kennedy, C.J., Mehrotra, R., Ren, S., Guan, M.Y., Downing, M., Chen, H.W., et al.: A real-time spatiotemporal ai model analyzes skill in open surgical videos. arXiv preprint arXiv:2112.07219 (2021)
7. Grady, P., Tang, C., Twigg, C.D., Vo, M., Brahmbhatt, S., Kemp, C.C.: Contactopt: Optimizing contact to improve grasps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1471–1481 (2021)
8. Hampali, S., Sarkar, S.D., Rad, M., Lepetit, V.: Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11090–11100 (2022)

9. Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M., Schmid, C.: Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 571–580 (2020)
10. Hasson, Y., Varol, G., Schmid, C., Laptev, I.: Towards unconstrained joint hand-object reconstruction from rgb videos. In: 2021 International Conference on 3D Vision (3DV). pp. 659–668. IEEE (2021)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
12. 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)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
14. Li, J., Zhou, Z., Yang, J., Pepe, A., Gsaxner, C., Luijten, G., Qu, C., Zhang, T., Chen, X., Li, W., et al.: Medshapenet–a large-scale dataset of 3d medical shapes for computer vision. Biomedical Engineering/Biomedizinische Technik **70**(1), 71–90 (2025)
15. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1954–1963 (2021)
16. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12919–12928 (2021). https://doi.org/10.1109/ICCV48922.2021.01270
17. 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)
18. Lv, J., Xu, W., Yang, L., Qian, S., Mao, C., Lu, C.: Handtailor: Towards high-precision monocular 3d hand recovery. arXiv preprint arXiv:2102.09244 (2021)
19. Park, J., Oh, Y., Moon, G., Choi, H., Lee, K.M.: Handoccnet: Occlusion-robust 3d hand mesh estimation network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1496–1505 (2022)
20. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3d with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9826–9836 (2024)
21. Potamias, R.A., Zhang, J., Deng, J., Zafeiriou, S.: Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. arXiv preprint arXiv:2409.12259 (2024)
22. Qi, H., Zhao, C., Salzmann, M., Mathis, A.: Hoisdf: Constraining 3d hand-object pose estimation with global signed distance fields. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10392–10402. IEEE (2024)
23. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610 (2022)
24. Saggio, G., Lazzaro, A., Sbernini, L., Carrano, F.M., Passi, D., Corona, A., Panetta, V., Gaspari, A.L., Di Lorenzo, N.: Objective surgical skill assessment: An initial experience by means of a sensory glove paving the way to open surgery simulation? Journal of surgical education **72**(5), 910–917 (2015)

25. Tse, T.H.E., Zhang, Z., Kim, K.I., Leonardis, A., Zheng, F., Chang, H.J.: S 2 contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning. In: European Conference on Computer Vision. pp. 568–584. Springer (2022)
26. Wang, R., Ktistakis, S., Zhang, S., Meboldt, M., Lohmeyer, Q.: Pov-surgery: A dataset for egocentric hand and tool pose estimation during surgical activities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 440–450. Springer (2023)
27. Wen, B., Tremblay, J., Blukis, V., Tyree, S., Müller, T., Evans, A., Fox, D., Kautz, J., Birchfield, S.: Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 606–617 (2023)
28. Wesierski, D., Jezierska, A.: Instrument detection and pose estimation with rigid part mixtures model in video-assisted surgeries. Medical image analysis **46**, 244–265 (2018)
29. Yang, L., Zhan, X., Li, K., Xu, W., Li, J., Lu, C.: Cpf: Learning a contact potential field to model the hand-object interaction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11097–11106 (2021)
30. Yu, Z., Huang, S., Fang, C., Breckon, T.P., Wang, J.: Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12955–12964 (2023)
31. Zhou, K., Lal Bhatnagar, B., Lenssen, J.E., Pons-Moll, G.: Toch: Spatio-temporal object correspondence to hand for motion refinement. arXiv preprint arXiv:2205.07982 (2022)
32. Zhou, Z., Zhou, S., Lv, Z., Zou, M., Tang, Y., Liang, J.: A simple baseline for efficient hand mesh reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1367–1376 (2024)