Applied Human Factors and Ergonomics (AHFE2025), Vol. 199, 2025, 828–836
https://doi.org/10.54941/ahfe1006895

Interactive Visualization for Human-in-the-Loop 3D-to-2D Pose Annotation

Yike Zhang[1] and Eduardo Davalos[2]

[1]St. Mary's University, San Antonio, TX 78209, USA
[2]Trinity University, San Antonio, TX 78209, USA

ABSTRACT

Aligning 3D objects with their poses in 2D images has traditionally relied on manual trial-and-error rendering, where human annotators repeatedly adjust parameters until the object appears to match the scene. This process is not only slow and labour-intensive, but also cognitively demanding, leading to human fatigue and inconsistent results. The reliance on such tedious workflows makes it difficult to scale annotations across entire video sequences, while the increased likelihood of error limits the reliability of the generated data. To address this gap, we present an interactive 3D-to-2D visualization and annotation tool that aids in accurate human annotation of 3D object poses. To our knowledge, this is the first system that allows users to directly manipulate 3D objects to support alignment to a 2D real-world scene, providing an intuitive 3D graphical user interface for annotating object positions and orientations. The tool integrates visual cues with spatial context to support swift and accurate pose annotation. By offering real-time visualization, depth estimation, and both single- and multi-object linked pose annotation, the proposed tool establishes a practical foundation for generating reliable pose data. By reducing the burden of numerical trial-and-error rendering and making pose annotation more intuitive, this tool advances human involvement in dataset generation, enabling researchers to more efficiently and accurately create the data needed to drive progress in downstream AI and vision-based applications. This interactive tool is available at https://github.com/InteractiveGL/vision6D.

Keywords: Cognitive support, Interactive pose annotation, 3D-to-2D visualization, Pose estimation, 6D pose, Augmented reality interfaces, Annotation tools

INTRODUCTION

Annotating 3D object poses in 2D images remains a time-consuming and cognitively demanding process, particularly when extended across long video sequences or complex object interactions. Traditional pose annotation workflows often rely on repetitive trial-and-render methods, requiring users to manually adjust object positions, render results, and visually verify correctness in a laborious iterative loop (Hodan, 2017; Drost, 2017). This lack of immediate feedback and ergonomic support makes the task tedious, error-prone, and difficult to scale. Although automated pose estimation algorithms can generate initial predictions, their results often require

extensive numerical correction due to occlusions, lighting inconsistencies, and ambiguous object geometries (Hodan, 2018). Existing tools, however, rarely integrate humans into the refinement loop in an intuitive or interactive way (Guan, 2024). Users are left to interpret numerical outputs, matrix parameters, or static overlays without meaningful guidance. As a result, human annotators face high cognitive load and reduced sense of control when engaging with complex numerical data. This gap highlights the need for a human-centered approach that not only leverages computational accuracy but also empowers users through interaction design, feedback mechanisms, and ergonomic visualization of the pose annotation process.

To address this challenge, our work introduces a novel human-in-the-loop pose annotation tool that integrates computational estimation with an interactive, ergonomically designed interface. Rather than treating pose retrieval as a fully automated numerical problem, this tool repositions humans at the center of the process, allowing them to visualize, adjust, and refine 3D poses interactively in real time. This approach not only supports the creation of high quality, domain-specific datasets but also reduces cognitive load by providing immediate visual feedback of alignment between 3D models and the corresponding 2D images. By enabling so, the system addresses principles of human factors and ergonomics, providing users to build trust in the annotation process with real time visualization and pose updates, and achieve greater efficiency when working with complex real-world objects. Beyond the technical scope, the design choices of our system underscores the importance of user experience in data annotation systems. Our proposed interface emphasizes intuitiveness, accessibility, and adaptability across various domains including supporting robot-assisted applications in medicine (Zhang, 2024; Zhang, 2025), agriculture (Wakchaure, 2023), and autonomous systems where custom pose datasets are often essential (Fonteles, 2024). By shifting the narrative from matrix manipulation and calculation to interactive human-computer collaboration, our approach leverages AR-inspired design principles to create an annotation workflow that is not only technically robust but also ergonomically supportive of human cognitive and perceptual capabilities.

Therefore, this paper proposes a user-centric framework for pose annotation that highlights the role of human factors in enabling reliable, efficient, and intuitive dataset generation. Our proposed work focuses on designing a pose annotation system where technology enhances and assists human performance (Lucchese, 2024). The tool represents a step forward in addressing the gap between automated and non-flexible pose estimation algorithms and human-centered refinement, equipping users to actively shape outcomes with greater efficiency. The highlights of our proposed augmented reality 6D pose annotation interactive tool are summarized below:

1. **Immediate and Intuitive Feedback:** The interactive visualization provides immediate, continuous feedback, reducing cognitive load and supporting users in forming a clear mental model of the 3D-2D alignment.

2. **Cognitive Support for 3D Reasoning**: By making depth cues explicit, the system supports human perceptual limitations in interpreting 3D structure from 2D views, minimizing errors caused by ambiguity.
3. **Precision with Reduced Frustration**: The single-object annotation mode enables focused, high-precision interaction, reducing task complexity and minimizing accidental misalignment.
4. **Linking Poses with Context Preservation**: By linking multi-object poses in the annotation tool, the system maintains spatial consistency, helping users preserve context and avoid repetitive manual corrections. This reduces annotation fatigue and supports efficient workflows in complex scenes.

PROPOSED HUMAN-IN-THE-LOOP FRAMEWORK

The proposed framework has followed the Dual Coding Theory (DCT) (Clark, 1987; Buonocore, 2025) which states that human cognition benefits from the non-verbal processing systems that visual representation complement abstract information. In this context, augmented reality-assisted systems deliver clear and intuitive visual cues that can enhance user understanding, engagement, and performance in computer-assisted tasks (Davalos, 2024). Within the domain of Human-Computer Interaction (HCI), the interactions with digital content has been broadly categorized into modalities such as sensor-based input, including devices like keyboards, mouse, and other peripherals (Nizam, 2018). Furthermore, Nielsen's Usability Heuristics (Krawiec, 2020) emphasize the importance of consistency, feedback, and error prevention, demonstrating that the system design should align with user expectations and natural behaviours. Guided by these principles, we developed an interface that incorporates both visual feedback and mouse-keyboard interaction, as illustrated in Figure 1, to support efficient and user-centered pose annotation within the proposed system.

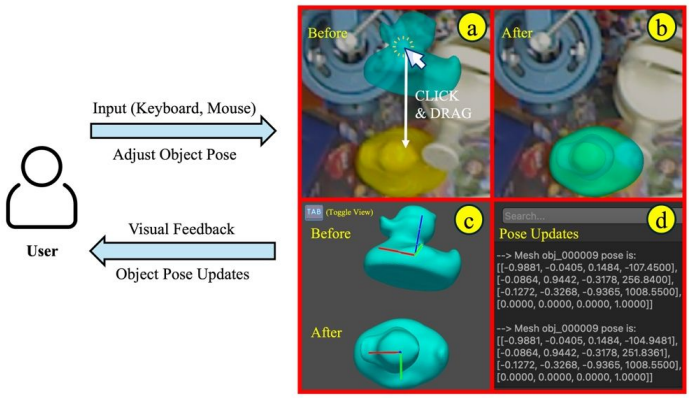

Figure 1: Input/output user interface explained in the proposed pose annotation system. Click-and-Drag gesture, keyboard hotkeys, and multi-view utilities makes data annotation more intuitive and user friendly.

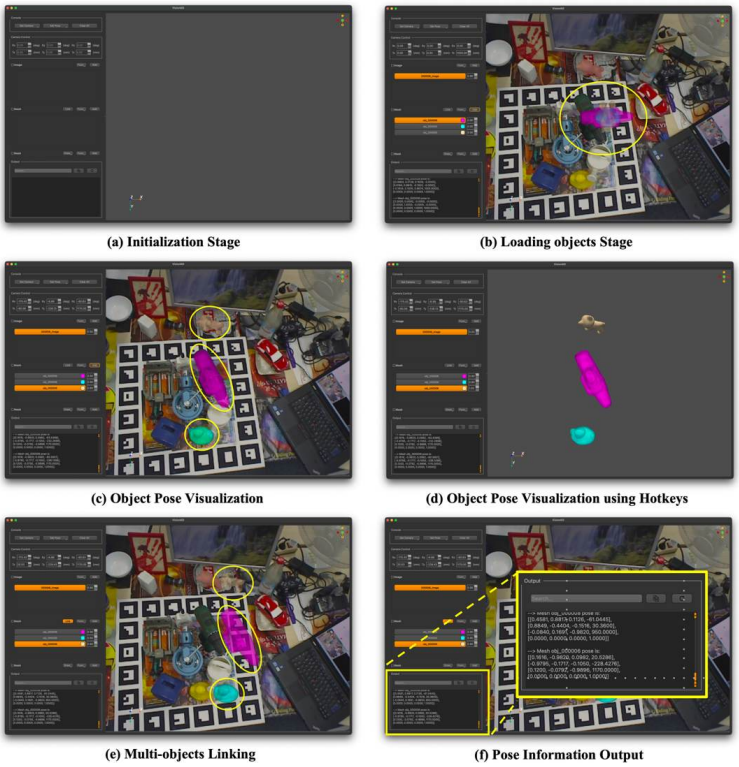

Figure 2: Walkthrough of the proposed system. Images (a-b-c-d-e) indicate example steps in obtaining object poses through interactive visualization and registration. Image (f) shows immediate updates of the current pose information and prior pose history.

Users interact with the proposed pose annotation system through sensor-based inputs, such as keyboard and mouse, to register and adjust 3D object poses. Figure 1 (a) illustrates the click-and-drag interaction, where the user aligns a 3D model with the corresponding 2D scene. Figure 1 (b) shows the successful registration outcome, while Figure 1 (c) compares the targeting object's pose before and after adjustment, enabling users to visually verify improvements in alignment. This view is accessed efficiently via a keyboard shortcut (Tab key), reducing the number of steps required and supporting ergonomic interaction. Finally, Figure 1 (d) demonstrates the system's output of the pose matrix, which is updated in real-time to provide immediate visual feedback. This feedback loop allows users to continuously monitor the results of their actions, evaluate pose accuracy, and decide whether further refinements are needed. Figure 2 illustrated below provides a step-by-step walkthrough of the primary functions implemented in the proposed system using a widely-used pose estimation dataset (Brachmann, 2020), including loading and displaying 2D images and 3D objects on the canvas, adjusting and refining object positions, and the visual feedback of the pose annotations and associated information. Specifically, Figure 2 (a) shows the initialization stage, where the user can import the target 2D image and the corresponding 3D models from there. Figure 2 (b) demonstrates how both the image and

the objects are displayed in the shared canvas environment, allowing the user to visually align the 3D objects directly to the 2D scene. In Figure 2 (c), object position and orientation can be manipulated through straightforward input controls (keyboard and mouse), supported by real-time feedback to ensure the customizable pose adjustments. Figure (d) illustrates the use of hotkeys that allow users to quickly inspect the object poses independently of the 2D image scene. Figure 2 (e) shows the multi object linking function in the system, where the user can easily control multiple objects simultaneously for efficient pose registration. Finally, Figure 2 (f) outputs the real-time pose information alongside a record of historical pose matrices, supporting both transparency and traceability of the annotated process.

The overall system design emphasizes consistency and immediate visual feedback, reducing cognitive load while enabling users to progressively refine 3D object poses with confidence and control. By combining visual representations with sensor-based interaction, the system supports efficient, ergonomic, and user-centered workflows for pose annotation related tasks.

RESULTS AND FINDINGS

To evaluate the effectiveness and usability of the proposed pose annotation system, Table 1 presents six samples from human annotations on object poses generated from our interactive interface. The first column lists the case number, followed by the corresponding human-annotated object poses via visual cues provided from the system. For comparison, the third column provides the ground-truth pose matrices obtained from the pose estimation dataset named LINEMOD-O (Brachmann, 2020). To quantify annotation performance, we report two widely-used pose error metrics: E_R, the angular distance error (in degrees) (Teyssandier, 2006) which measures the rotational difference between the annotated and ground-truth poses; and E_T, the Euclidean distance (Liberti, 2014) in millimeters, which reflects the transitional deviation between the two matrices.

Table 1: Samples of human-in-the-loop pose annotations using the proposed system.

N	Human Annotated Pose Matrix	Ground-Truth Pose Matrix	E_R	E_T
1	[[0.9114, 0.4089, 0.0468, –24.7800], [0.4027, –0.8626, –0.3063, –17.2300], [–0.0849, 0.2980, –0.9508, 1106.5400], [0.0000, 0.0000, 0.0000, 1.0000]]	[[0.8956, 0.4354, 0.0932, –24.8778], [0.4377, –0.8226, –0.3634, –15.6402], [–0.0816, 0.3662, –0.9271, 1097.1863], [0.0000, 0.0000, 0.0000, 1.0000]]	4.9	9.5
2	[[0.9671, 0.2441, 0.0722, –37.4400], [0.2544, –0.9157, –0.3112, 166.7612], [–0.0098, 0.3193, –0.9476, 986.5400], [0.0000, 0.0000, 0.0000, 1.0000]]	[[0.9573, 0.2739, 0.0930, –35.0693], [0.2892, –0.9135, –0.2862, 168.3097], [0.0066, 0.3009, –0.9537, 978.1587], [0.0000, 0.0000, 0.0000, 1.0000]]	2.5	8.9

Continued

Table 1: Continued

N	Human Annotated Pose Matrix	Ground-Truth Pose Matrix	E_R	E_T
3	[[0.4806, 0.8723, 0.0897, –69.4300], [0.8433, –0.4317, –0.3201, 23.8133], [–0.2406, 0.2295, –0.9431, 1026.5400], [0.0000, 0.0000, 0.0000, 1.0000]]	[[0.4624, 0.8798, 0.119, –71.9268], [0.8614, –0.4121, –0.300, 22.5920], [–0.2147, 0.2409, –0.9475, 1015.1082], [0.0000, 0.0000, 0.0000, 1.0000]]	2.3	11.8
4	[[–0.9922, –0.0902, 0.0862, –105.6925], [–0.1143, 0.9342, –0.3378, 252.1267], [–0.0501, –0.3450, –0.9373, 1006.5400], [0.0000, 0.0000, 0.0000, 1.0000]]	[[–0.9758, –0.1923, 0.1043, –106.5101], [–0.2141, 0.9374, –0.2747, 250.2484], [–0.0449, –0.2904, –0.9559, 1002.0606], [0.0000, 0.0000, 0.0000, 1.0000]]	6.8	5.0
5	[[0.0553, 0.9935, 0.0994, 187.1882], [0.9614, –0.0261, –0.2740, –62.5707], [–0.2697, 0.1107, –0.9566, 1146.5400], [0.0000, 0.0000, 0.0000, 1.0000]]	[[0.0194, 0.9854, 0.1690, 189.8814], [0.9582, 0.0300, –0.2846, –62.4703], [–0.2856, 0.1674, –0.9436, 1160.3044], [0.0000, 0.0000, 0.0000, 1.0000]]	4.6	14.0
6	[[0.2716, –0.9614, –0.0433, –81.7160], [–0.9546, –0.2634, –0.1393, –233.5716], [0.1225, 0.0792, –0.9893, 1156.5388], [0.0000, 0.0000, 0.0000, 1.0000]]	[[0.3063, –0.9519, –0.0153, –78.8694], [–0.9506, –0.3049, –0.0603, –232.6228], [0.0527, 0.033, –0.9981, 1156.5381], [0.0000, 0.0000, 0.0000, 1.0000]]	5.3	3.0

In summary, these error measures E_R and E_T demonstrate not only the accuracy of human annotations achieved through the proposed system but also the usability of applying simple and intuitive visual cues for pose alignment. Unlike traditional numerical manipulation of pose matrices (Horn, 1987; Triggs, 2000), the system enables users to engage with the task through natural perceptual and hand-motor interactions, lowering the cognitive demands associated with abstract pose transformations. By providing immediate visual feedback and intuitive sensor-based controls, the human-in-the-loop interface supports cognitive ergonomics, allowing users to verify, refine, and confirm pose annotations immediately with reduced mental effort. This design strengthens user confidence and trust in the annotation process by providing transparent and clear visual feedback, while ensuring final results that are reliable and comparable in accuracy to ground-truth poses. From a human factors perspective, these findings highlight the importance of embedding ergonomic principles into the pose annotation tools. The results suggest that the proposed system not only preserves technical rigor in 3D object pose estimation but also enhances workflow efficiency and supports user engagement via prompt feedback.

LIMITATIONS AND FUTURE WORK

One noticeable limitation of the proposed pose annotation tool lies in handling round and symmetric texture-less objects, such as spheres or cylinders. From a human factors perspective, these objects impose significant cognitive challenges because they lack distinctive features or markers for annotators to confidently determine orientation. The resulting rotational ambiguities can lead to multiple possible pose solutions that appear equally valid to the human eye. Those ambiguities not only slow down the annotation process but may also contribute to user frustration. One possible solution is to texture the object and thus, they will have distinguishable features when users are annotating the poses.

For future work, integrating 2D-to-3D Point-and-Perspective (PnP) registration technique (Yu, 2024) could provide initial automated predictions that serve as starting points for the object pose annotation. Using PnP would allow annotators to have an alternative route to pose annotation that leverages from indicating, via clicking, the matching 3D points on the object and their corresponding 2D point on the scene image. By offering this additional functionality in the manual registration process, we can further reduce cognitive workload while maintaining user agency and control. Moreover, instead of annotating poses in individual images, the system can evolve toward a hybrid workflow to assist video pose annotation where initial human input is propagated automatically across a video using geometric tracking and camera localization, such as SLAM-based approaches (Alsadik, 2021; Khairuddin, 2015). In such scenarios, users would only refine and adjust propagated annotations, thereby improving efficiency while minimizing overall manual effort. Ultimately, these suggestions and directions opens many avenues for broader exploration of AR-driven visual pose annotation support and potentially design human-in-the-loop collaborative systems that balance machine efficiency and human insight.

CONCLUSION

Beyond the immediate benefits to object pose estimation research, the proposed cross-platform system demonstrates how ergonomically informed design can enhance the usability and accessibility of general pose annotation tasks. Our work aim to shift the perspective from the traditional numerical object pose matrices calculation and manipulation to weigh more on human factors that emphasizing immediate visual feedback and prompt adjustment of poses using the proposed human-in-the-loop system. Finally, this system demonstrates that it is essential to design tools that support and enhance human cognition and decision-making. Instead of requiring users to interpret or direct manipulate transformation matrices, the system provides intuitive pose visualization and interactive controls that align with natural human perceptual and motor capabilities.

SUPPLEMENTARY MATERIAL

An example video demonstrating robust and efficient registration of a 3D object to a 2D scene is available: https://github.com/InteractiveGL/vision6D?tab=readme-ov-file#3d-to-2d-visualization-and-annotation-desktop-app-for-6d-pose-estimation-related-tasks

REFERENCES

Alsadik, Bashar, and Samer Karam. (2021). The Simultaneous Localization and Mapping (SLAM)—An Overview. Journal of Applied Science and Technology Trends, 2, 120–131. https://doi.org/10.38094/jastt204117

Brachmann, Eric, 2020, "6D Object Pose Estimation using 3D Object Coordinates [Data]", https://doi.org/10.11588/DATA/V4MUMX, heiDATA, V1.

Buonocore, S., Granata, E., Tulino, A., Gironimo, G. (2025). Augmented Reality Aids Logistics: Augmenting workers' abilities during customs inspections. In: Tareq Z. Ahram, Waldemar Karwowski and Pei-Luen Rau (eds) Human-Computer Interaction & Emerging Technologies. AHFE (2025) International Conference. AHFE Open Access, vol. 195. AHFE International, USA. http://doi.org/10.54941/ahfe1006259

Clark, J. M., & Paivio, A. (1987). A dual coding perspective on encoding processes. In Springer eBooks (pp. 5–33). https://doi.org/10.1007/978-1-4612-4676-3_1

Davalos, Eduardo, Yike Zhang, Ashwin T. S., Joyce Horn Fonteles, Umesh Timalsina, and Gautam Biswas. (2024). 3D Gaze Tracking for Studying Collaborative Interactions in Mixed-Reality Environments. In Companion Proceedings of the 26th International Conference on Multimodal Interaction (ICMI '24 Companion), San Jose, Costa Rica, 175–183. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3686215.3688380

Drost, Bertram, Markus Ulrich, Paul Bergmann, Philipp Härtinger, and Carsten Steger. (2017). Introducing MVTec ITODD — A Dataset for 3D Object Recognition in Industry. In 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2200–2208. https://doi.org/10.1109/ICCVW.2017.257

Fonteles, J. et al. (2024). A First Step in Using Machine Learning Methods to Enhance Interaction Analysis for Embodied Learning Environments. In: Olney, A. M., Chounta, IA., Liu, Z., Santos, O. C., Bittencourt, I. I. (eds) Artificial Intelligence in Education. AIED 2024. Lecture Notes in Computer Science (), vol. 14830. Springer, Cham. https://doi.org/10.1007/978-3-031-64299-9_1

Guan, Jian, Yingming Hao, Qingxiao Wu, Sicong Li, and Yingjian Fang. (2024). A Survey of 6DoF Object Pose Estimation Methods for Different Application Scenarios. Sensors, 24(4), 1076. https://doi.org/10.3390/s24041076

Hodan, Tomas, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiri Matas, and Carsten Rother. (2018). BOP: Benchmark for 6D Object Pose Estimation. arXiv:1808.08319 [cs. CV]. https://arxiv.org/abs/1808.08319

Hodan, Tomas, Pavel Haluza, Stepan Obdrzalek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. (2017). T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects. In IEEE Winter Conference on Applications of Computer Vision (WACV).

Horn, Berthold K. P. (1987). Closed-form solution of absolute orientation using unit quaternions. J. Opt. Soc. Am. A, 4(4), 629–642. https://doi.org/10.1364/JOSAA.4.000629

Khairuddin, A. R., M. S. Talib, and H. Haron. (2015). Review on simultaneous localization and mapping (SLAM). In 2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Penang, Malaysia, 85–90. https://doi.org/10.1109/ICCSCE.2015.7482163

Krawiec, Lukasz, and Helena Dudycz. (2020). A comparison of heuristics applied for studying the usability of websites. Procedia Computer Science, 176, 3571–3580. https://doi.org/10.1016/j.procs.2020.09.029

Liberti, Leo & Lavor, Carlile & Maculan, Nelson & Mucherino, Antonio. (2014). Euclidean Distance Geometry and Applications. SIAM Review. 56. 3–69. 10.1137/120875909

Lucchese, A., Panagou, S., & Sgarbossa, F. (2024). Investigating the impact of cognitive assistive technologies on human performance and well-being: An experimental study in assembly and picking tasks. International Journal of Production Research, 63(6), 2038–2057. https://doi.org/10.1080/00207543.2024.2394090

Nizam, S. S. M., Abidin, R. Z., Hashim, N. C., Lam, M. C., Arshad, H., & Majid, N. a. A. (2018). A review of multimodal interaction technique in augmented reality environment. International Journal on Advanced Science Engineering and Information Technology, 8(4–2), 1460–1469. https://doi.org/10.18517/ijaseit.8.4-2.6824

Teyssandier, Pierre & Le Poncin-Lafitte, Christophe. (2006). Angular distances in metric theories.

Triggs, B., McLauchlan, P. F., Hartley, R. I., Fitzgibbon, A. W. (2000). Bundle Adjustment—A Modern Synthesis. In: Triggs, B., Zisserman, A., Szeliski, R. (eds) Vision Algorithms: Theory and Practice. IWVA 1999. Lecture Notes in Computer Science, vol. 1883. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3–540-44480–7_21

Wakchaure, Manas, B. K. Patle, and A. K. Mahindrakar. (2023). Application of AI techniques and robotics in agriculture: A review. Artificial Intelligence in the Life Sciences, 3, 100057. https://doi.org/10.1016/j.ailsci.2023.100057

Yu, Yingjian, Zi Wang, Zhang Li, and Qifeng Yu. (2024). A comprehensive study on PnP-based pipeline for pose estimation of noncooperative satellite. Acta Astronautica, 224, 486–496. https://doi.org/10.1016/j.actaastro.2024.08.027.

Zhang, Yike, and Jack H. Noble. (2025). Post-mastoidectomy surface multiview synthesis from a single microscopy image. In Medical Imaging 2025: Image-Guided Procedures, Robotic Interventions, and Modeling, eds. Maryam E. Rettmann and Jeffrey H. Siewerdsen, vol. 13408, 1340806. Bellingham, WA: SPIE (International Society for Optics and Photonics). https://doi.org/10.1117/12.3047130

Zhang, Yike, Eduardo Davalos, Dingjie Su, Ange Lou, and Jack H. Noble. (2024). Monocular microscope to CT registration using pose estimation of the incus for augmented reality cochlear implant surgery. In Medical Imaging 2024: Image-Guided Procedures, Robotic Interventions, and Modeling, eds. Jeffrey H. Siewerdsen and Maryam E. Rettmann, vol. 12928, 129282I. Bellingham, WA: SPIE (International Society for Optics and Photonics). https://doi.org/10.1117/12.3008830