# KidneyDepth: A Synthetic Kidney Dataset for Metric Depth Estimation in Ureteroscopy

Laura Oliva-Maza[1,3][0000−0001−5382−3025], Florian Steidle[1][0000−0001−6935−9810], Julian Klodmann[1][0000−0002−9428−7211], Klaus Strobl[1][0000−0001−8123−0606], Arkadiusz Miernik[2][0000−0001−5894−9647], and Rudolph Triebel[1,3][0000−0002−7975−036X]

[1] Institute of Robotics and Mechatronics, German Aerospace Center (DLR), Wessling, Germany
[2] Department of Urology, Faculty of Medicine, University of Freiburg - Medical Centre, Freiburg, Germany
[3] Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
laura.olivamaza@dlr.de

**Abstract.** Monocular Metric Depth Estimation (MDE) in endoscopic images is a crucial step to improve navigation during medical procedures, as it enables the estimation of dense, real-scale 3D maps of the organs. For instance, in monocular flexible ureteroscopy (fURS), accurate navigation and real-scale information are essential for locating and removing kidney stones efficiently. Currently, the most promising approach to infer depth from single passive cameras is by supervised training of large neural networks, so-called foundation models for MDE. However, the depth output of these models is biased when the training data domain does not fit the goal domain (both camera and scene). At the same time, one of the greatest challenges in medical imaging is the lack of annotated datasets, as obtaining real ground-truth (e.g., depth data) is difficult. To overcome this, simulation has become a valuable tool in ureteroscopic imaging research. In this study, we introduce KidneyDepth, a synthetic dataset designed to reduce the gap between simulated and real-world 3D imaging. It includes a variety of shapes (e.g. mesh from CT scan, geometric primitive forms) along with different textures and lighting conditions, generated by BlenderProc2 [7]. To assess the effectiveness of KidneyDepth, we fine-tune two state-of-the-art MDE models (Depth Anything V2 and ZoeDepth) and test their performance on both simulated and real ureteroscopic images. Additionally, we evaluate the validity of their output by using the inferred depths in the context of a RGB-D SLAM system. Our results show that training models on a synthetic dataset with diverse structures and lighting conditions improves depth estimation in real endoscopic images and our simulations show that these RGB-D images enhance overall SLAM accuracy. The KidneyDepth dataset can be found in https://zenodo.org/records/14893421.

**Keywords:** Monocular metric depth · Dataset · Navigation · Ureteroscopy.

## 1 Introduction

Flexible ureteroscopy (fURS) is a common treatment for removing kidney stones. In this procedure a flexible ureteroscope is inserted through the urinary tract into the kidney to inspect it and remove stones. Depending on the stone's size, it may need to be fragmented using a laser before extracting the smaller pieces, or, if small enough, it can be removed through the working channel without fragmentation. Knowing the exact size of the stone is important to determine whether laser fragmentation is necessary or not [25]. To prevent the need for a follow-up intervention, it is essential that all stones and fragments are located and removed during the procedure.

The use of fURS is growing due to its benefits for the patient [9], but it presents several challenges for the surgeon. These include a steep learning curve, uncomfortable positioning and difficulty maintaining spatial awareness within the organ using only monocular endoscopic live video [3]. Robotic systems like [13], [22] and [23] address these issues by improving ergonomics and allow solo surgery. Moreover, monocular visual Simultaneous Localization and Mapping (vSLAM) algorithms can assist the surgeon in surgical navigation without additional sensors [18]. However, the primary challenge with monocular vSLAM is the lack of real-scale information.

Metric depth from monocular images can be estimated using either self-supervised models or foundation models. Self-supervised models focus on learning depth from unlabeled data (e.g. stereo image pairs or image sequences) while foundation models are pre-trained on large, diverse datasets and then fine-tuned to a specific task (e.g. metric depth estimation). The main limitation of self-supervised models is their poor generalization across different datasets, often requiring retraining with each new dataset [4]. In the medical domain, these models also face challenges in generalizing accross different patients and anatomies [10]. Also, obtaining ground truth depth maps for fine-tuning remains difficult.

To overcome these challenges, we introduce KidneyDepth, a dataset design for metric depth estimation consisting of various shapes (ranging from real CT scans of the kidney to primitive forms like cylinders and tori), lighting conditions (static lights and lights mimicking the endoscope's illumination, i.e., attached to the moving camera), and diverse materials (with different reflection properties). We fine-tune ZoeDepth [4] and Depth Anything V2 [26] on this dataset, achieving accurate depth map estimations for real kidney images. These models are also evaluated in an RGB-D SLAM system across different trajectories in a simulation model with new materials not seen during training.

## 2 Related Work

### 2.1 Monocular Depth Estimation

Monocular Metric Depth Estimation (MDE) from endoscopic images can significantly enhance 3D reconstruction and navigation during minimally-invasive surgical procedures; in effect, this contribution applied to a monocular endoscope

can be visualized as an upgraded sensor, yielding RGB-D data for improved mapping and navigation within body organs.

Self-supervised models for visual depth estimation learn to infer depth by using unlabeled data and applying e.g. structure-from-motion methods [14], photometric consistency[21], and temporal consistency between video frames [6]. Yet as previously mentioned, self-supervised models suffer from poor generalization.

In contrast, supervised training of foundation models has the potential to address these limitations. They are pre-trained on large, diverse datasets and can be subsequently fine-tuned for specific tasks or datasets. For instance, Depth Anything V2 [26] uses a transformer-based architecture trained on both synthetic and real-world data, while Zoedepth [4] combines multi-scale learning with depth and semantic information. Both models were evaluated on medical datasets in [10], without fine-tuning. However, estimating accurate metric depth requires fine-tuning these models on domain-specific datasets, which is hindered by the lack of ground-truth depth maps in the medical domain.

### 2.2 Datasets

Most existing datasets for real endoscopic procedures focus on laparoscopy [20] or colonoscopy [2], [15]. A key challenge in acquiring useful real-world data is the difficulty of capturing ground truth information required for supervised training, leading some studies to focus on ex-vivo scenarios [19] or simulations.

Simulations offer significant advantages, such as the ability to accurately generate segmentation masks, endoscope poses, point clouds, and depth maps that are perfectly aligned with the corresponding synthetic images. However, the primary challenge lies in bridging the simulation-reality gap. Approaches such as SimuScope [17] generate synthetic endoscopic images with realistic lighting and textures. Other methods, like [12], use sim-to-real transfer techniques to train depth estimation models, while [16] leverages cinematic rendering for photorealistic medical images. Additionally, [11] applies neural stylization to anatomical meshes for improved endoscopic image generation. To the best of the authors' knowledge, no existing datasets currently focus on ureteroscopic images.

### 2.3 Visual Navigation in the Urinary Tract

Accurate visual navigation within the kidney is crucial to ensure no stones or fragments are left behind. To facilitate this, [1] uses structure from motion to estimate a point cloud of the kidney's interior, registered with pre-operative CT scans. Meanwhile, [27] and [8] use electromagnetic sensors for metric pose tracking of the ureteroscope with high accuracy. In contrast, [18] extends ORB-SLAM3 [5] to accurately estimate the ureteroscope's pose and a up-to-scale map of the kidney using only the endoscope images. Our goal is to generate metrically-accurate kidney models and precise metric motion estimates in real time, using only a thin, flexible monocular endoscopic camera.

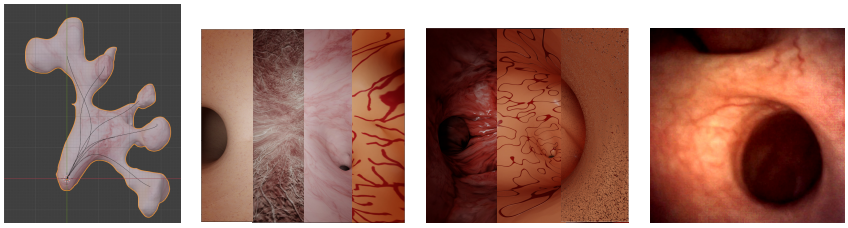
## 3 Methodology and Results

In this section we will focus on explaining the dataset content and the different evaluations performed on it. The experiments have been performed with a 13th Gen Intel® Core™ i9-13900K CPU (24 cores), 64GB RAM, and a Nvidia GeForce RTX 4090 24GB GPU.

### 3.1 KidneyDepth

Given the absence of ureteroscopic image datasets with depth maps, we introduce KidneyDepth, a synthetic dataset developed for metric depth estimation. The dataset serves three main goals: (1) generation of ground truth maps to train depth estimation models, (2) evaluation of the models on real endoscopic kidney images, and (3) SLAM trajectory evaluation.

To achieve these goals, we created three sub-datasets: KidneyDepthMetric for fine-tuning depth estimation models, featuring various shapes (*ct*: mesh extracted from a CT scan of the kidney using ImFusion Suite[4]- Figure 1a, *ct_stones*: *ct* with stones, *cylinder* and *tori*), materials (*skin*, *flesh*, *marble* and *painted* - Figure 1b), and light sources (*dynamic*: flashlight moving with the camera and *static*: distributed lights inside the corresponding mesh); KidneyDepthSLAM for evaluating SLAM trajectories, consisting of longer sequences within the *ct* and *ct_stones* mesh with different materials (*bloody organ*, *vessels* and *thin vessels* - Figure 1c); and KidneyDepthfURS, which includes real ureteroscopic images recorded at the University Hospital Freiburg (Figure 1d).



(a) (b) (c) (d)

**Fig. 1.** (a) Kidney mesh from CT scan with camera trajectories, (b) Materials of KidneyDepthMetric: *skin*, *flesh*, *marble* and *painted* with the *dynamic* light inside the *ct* mesh. (c) Materials of KidneyDepthSLAM: *bloody organ*, *vessels* and *thin vessels* with the *dynamic* light inside the *ct* mesh (d) Example of KidneyDepthfURS.

The simulated images, camera poses and depth maps are generated using BlenderProc2 [7] an open-source framework for generating synthetic data by rendering realistic 3D scenes with configurable camera paths, lighting, and materials. The dataset can be found in https://zenodo.org/records/14893421.

[4] ImFusion GmbH, www.imfusion.com

### 3.2 Metric Depth from Monocular

We fine tuned Depth Anything V2 [26] and Zoedepth [4], using the KidneyDepthMetric dataset split into training (70%), validation (20%), and test sets (10%). The models were fine-tuned for 40 epochs with augmentations like noise, brightness adjustments, and rotations/scale shifts. Depth Anything V2 was trained on two Nvidia GeForce RTX 3090 GPUs, while Zoedepth used one.

Depth Anything V2 [26] is a monocular depth estimation model that uses an encoder-decoder architecture with multi-scale fusion and attention mechanisms to generate depth maps. Trained on both real-world and synthetic datasets with L1 and scale-aware losses, it provides accurate depth predictions. Similarly, Zoedepth [4] estimates metric depth by integrating appearance and geometry features through a deep network with a metrics bin module. Also trained on real-world and synthetic datasets, it utilizes photometric, gradient, scale-invariant, and edge-aware loss functions.

We have fine-tuned both models on different subsets of KidneyDepthMetric to evaluate how shape, material, and light source affect the depth estimation. For the evaluation, we used KidneyDepthSLAM images, where the material is unknown to the models. We computed $\delta_1 = n\left(\max\left(\frac{\hat{D}}{D}, \frac{D}{\hat{D}}\right)\right) < 1.25$ and $AbsRel = \frac{1}{N}\sum_i \frac{|\hat{d}_i - d_i|}{\hat{d}_i}$, where $d_i$ is a pixel from the predicted depth map $D$, $\hat{d}_i$ is its corresponding pixel on the ground truth map $\hat{D}$, $N$ is the total number of pixels and $n$(condition) is the percentage of pixels that satisfy the condition. $\delta_1$ measures the percentage of predicted pixels with a deviation of no more than 25%, while $AbsRel$ calculates the average percentage difference between predicted and true distances. These metrics have been computed for different configurations (Table 1). "Material: *mat*" represents a fine-tune where only the material *mat* has been seen in different shapes and light conditions, "Light: *light*" different materials and shapes that have been seen under the light source *light*, and "Shape: *shape*" where the shape *shape* has been seen with different materials and light sources. For "Complete" all the images have been used.

In Table 1 we observe that, when the models are fine tuned on a single material ("Material: *mat*"), they do not generalize well to new materials. When trained with multiple materials ("Complete"), the models generalize better, as each material interacts with light differently, increasing the likelihood that new materials share properties with those already learned. Concerning light sources, the results highlight that the light in the simulation must closely resemble real-world conditions, observe how the configuration "Light: *dynamic*" and "Complete" results in similar accuracy, indicating that "Light: *static*" does not contribute significantly during training. A similar behavior is observed with shapes.

Finally, we visually evaluate the inferred depth of the fine tuned models on KidneyDepthfURS. Showing a big improvement of these models after fine-tuning them on synthetic images (Figure 2). While in simulation, Depth Anything V2 outperforms Zoedepth (Table 1), in real ureteroscopic images (Figure 2), Depth Anything V2 provides finer depth estimates when it is accurate, yet it does fail

**Table 1.** Comparison of Depth Anything V2 and Zoedepth when fine-tuned in different subsets of KidneyDepthMetric and tested on images from KidneyDepthSLAM. The bold values represent the overall best performance within the model.

| | Depth Anything V2 | | Zoedepth | |
|---|---|---|---|---|
| | $\delta_1 \uparrow$ | *AbsRel* $\downarrow$ | $\delta_1 \uparrow$ | *AbsRel* $\downarrow$ |
| Material: *marble* | 0.655 | 0.199 | 0.501 | 0.264 |
| Material: *flesh* | 0.184 | 0.389 | 0.040 | 0.622 |
| Material: *painted* | 0.099 | 0.631 | 0.047 | 0.813 |
| Material: *skin* | 0.282 | 0.418 | 0.264 | 0.423 |
| Light: *dynamic* | 0.766 | 0.156 | 0.788 | 0.150 |
| Light: *static* | 0.542 | 0.27 | 0.378 | 0.339 |
| Shape: *ct & stones* | 0.863 | 0.14 | **0.794** | **0.167** |
| Shape: *cylinder & torus* | 0.203 | 0.406 | 0.224 | 0.44 |
| Complete | **0.869** | **0.139** | 0.773 | 0.179 |

more frequently than Zoedepth (e.g., Figure 2 row 5). Additionally, while Depth Anything V2 operates at 12 fps, Zoedepth is limited to 5 fps.

### 3.3 RGB-D SLAM in Ureteroscopy

To evaluate the effect of the inferred depths, we have extended [18] to work with RGB-D image. The SLAM system was evaluated on the KidneyDepthSLAM sequences by computing the Absolute Trajectory Error (*ATE*), the Scale Error ($\epsilon_{\text{scale}}$), and the percentage of frames where the SLAM successfully tracks ($n_{\text{traj}}$).

The Absolute Trajectory Error is defined as $ATE = \sqrt{\frac{1}{N}\sum_{t=1}^{N} \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2}$ where $N$ is the total number of frames, and $\mathbf{x}_t$ and $\widetilde{\mathbf{x}}_t$ are the estimated and ground truth positions, respectively. This metric, as proposed in [24], evaluates the root mean squared error (RMSE) of the trajectory after alignment.

To measure the scale consistency of the SLAM system, we compute the Scale Error, $\epsilon_{\text{scale}} = \frac{\sum_{t=1}^{N} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|}{\sum_{t=1}^{N} \|\widetilde{\mathbf{x}}_{t+1} - \widetilde{\mathbf{x}}_t\|}$. This ratio quantifies the deviation in scale between the estimated and ground truth trajectories – a value closer to 1 indicates a more accurate scale estimation. These metrics were computed for the trajectories estimated by the SLAM system when using inferred depth from Depth Anything V2 and Zoedepth (both after fine-tuning with "Complete"). Additionally, *ATE* and $n_{\text{traj}}$ were calculated for the trajectories estimated by the monocular SLAM based on [18] after scaling them to correct for scale error, as monocular estimates trajectories up to scale. KidneyDepthSLAM is composed by: *Seq_ 01* and *Seq_ 03* where the camera explores all the calyxes of the *ct* mesh with *dynamic* light and *bloody organ* and *vessels* material respectively; *Seq_ 02*, *Seq_ 04* and *Seq_ 05* where we add camera rotations and explore the *ct_ stones* mesh with *bloody organ*, *vessels* and *thin vessels* material.

Table 2 presents the computed metrics. Zoedepth generally performs better than Depth Anything V2 in estimating scale across most trajectories. Due to

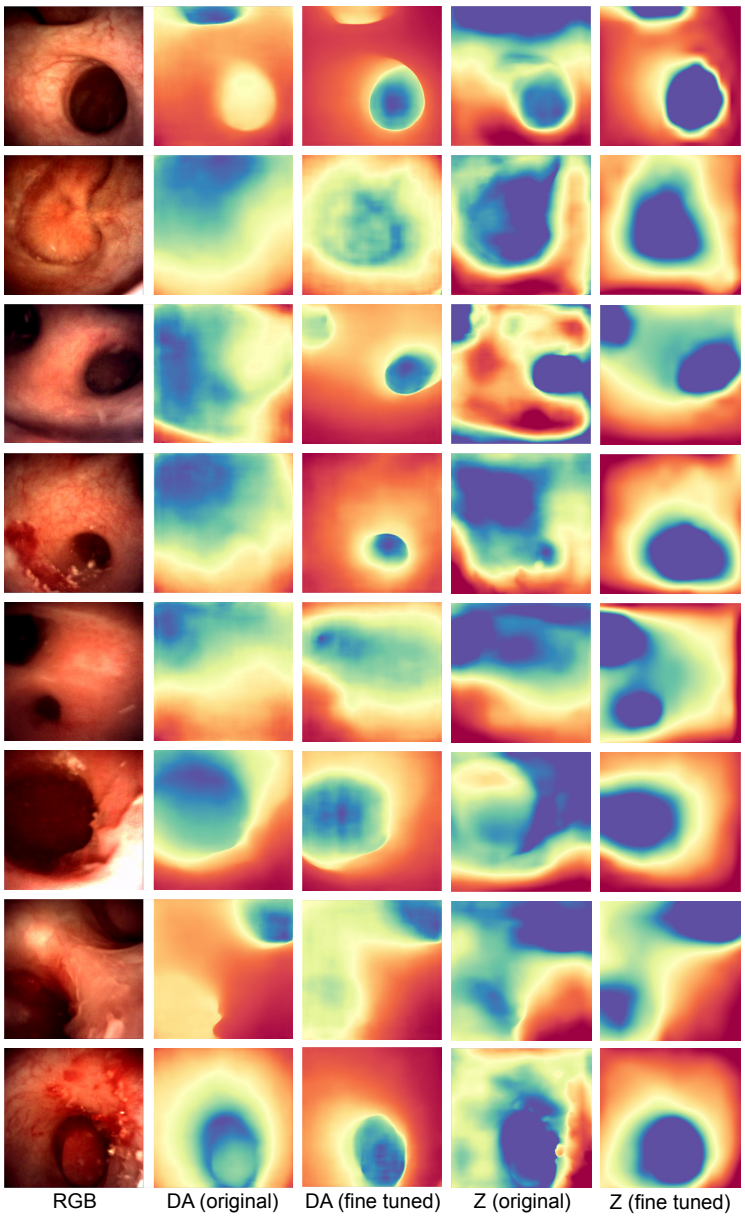



**Fig. 2.** Inferred depths of images from KidneyDepthfURS ("RGB") by using Depth Anything V2 ("DA (original)"), the fine-tuned configuration on "Complete" ("DA (fine tuned)"), Zoedepth ("Z (original)"), and its fine-tuned configuration on "Complete" ("Z (fine tuned)").

inaccuracies in the estimated depth maps, and scale errors, the *ATE* is larger for RGB-D compared to the monocular estimated trajectories after scaling. However, a key advantage of RGB-D is its one-frame initialization, which allows the SLAM system to track the camera trajectories over a greater number of frames.

**Table 2.** Performance of RGB-D SLAM on synthetic sequences. Monocular has been scaled with respect to the ground truth before computing the *ATE*.

| | Depth Anything V2 | | | Zoedepth | | | Monocular | |
|---|---|---|---|---|---|---|---|---|
| Sequence | *ATE* [mm] | $\epsilon_{\text{scale}}$ | $n_{\text{traj}}$ | *ATE* [mm] | $\epsilon_{\text{scale}}$ | $n_{\text{traj}}$ | *ATE* [mm] | $n_{\text{traj}}$ |
| *Seq_ 01* | 2.46 | 0.89 | 99.96% | 1.84 | 0.93 | 100.0% | 0.15 | 95.74% |
| *Seq_ 02* | 2.81 | 0.91 | 99.96% | 2.07 | 0.89 | 99.91% | 0.21 | 95.30% |
| *Seq_ 03* | 2.18 | 0.88 | 99.13% | 0.93 | 1.09 | 100.0% | 0.16 | 98.91% |
| *Seq_ 04* | 1.88 | 0.97 | 99.96% | 1.13 | 1.08 | 100.0% | 0.18 | 99.26% |
| *Seq_ 05* | 1.90 | 1.20 | 93.61% | 2.77 | 1.33 | 99.96% | 0.34 | 61.98% |

In monocular SLAM, features from different frames must be matched to triangulate and create new map points, resulting in sparse maps. In contrast, RGB-D SLAM generates denser maps since the depth of each pixel is known, eliminating the need for matching and triangulation. Map points are used to estimate the camera pose by matching the projected map points to the extracted features in the new image. In *Seq_ 05*, thinner vessels lead to fewer features, fewer matches and a sparser map in monocular SLAM, making it harder to track the camera pose (low $n_{\text{traj}}$). However, RGB-D SLAM generates denser maps, facilitating better pose tracking and maintaining a high $n_{\text{traj}}$.

## 4 Conclusions

We introduce KidneyDepth, a synthetic dataset of ureteroscopic images designed for MDE in monocular flexible ureteroscopy (fURS). The dataset consists of three parts: KidneyDepthMetric, KidneyDepthSLAM, and KidneyDepthfURS. We use the metric depth data in KidneyDepthMetric to fine-tune two state-of-the-art MDE foundation models (Depth Anything V2 and ZoeDepth), and we use KidneyDepthSLAM and KidneyDepthfURS to evaluate these models. With KidneyDepthMetric, we explore which factors impact the learning process, demonstrating that training with diverse materials improves the model's ability to generalize to unseen images. Additionally, we assess the inferred depths of KidneyDepthSLAM in a RGB-D SLAM system based on [18], showing that both models accurately estimate absolute scale and the endoscope trajectories, while also enabling faster initialization. Finally, we evaluate the fine-tuned models on KidneyDepthfURS, demonstrating an improvement in monocular depth estimation for real ureteroscopic images.

By estimating the metric depth of monocular ureteroscopic images, surgeons can accurately evaluate the real-world scale of objects (e.g., kidney stones) in the

images. Integrating the inferred depths into a SLAM system enhances navigation by enabling pose tracking in real scale over a larger number of frames. Our results show that fine-tuning MDE models with synthetic images yields promising performance on real endoscopic images, providing a solid foundation for further advancements in surgical navigation.

**Acknowledgments.** The authors would like to thank ImFusion GmbH.

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

## References

1. Acar, A., Lu, D., Wu, Y., Oguz, I., Kavoussi, N., Wu, J.Y.: Towards navigation in endoscopic kidney surgery based on preoperative imaging. Healthcare Technology Letters **11**(2-3), 67–75 (2024)
2. Azagra, P., Sostres, C., Ferrández, Á., Riazuelo, L., Tomasini, C., Barbed, O.L., Morlana, J., Recasens, D., Batlle, V.M., Gómez-Rodríguez, J.J., et al.: Endomapper dataset of complete calibrated endoscopy procedures. Scientific Data **10**(1), 671 (2023)
3. Bergen, T., Wittenberg, T.: Stitching and surface reconstruction from endoscopic image sequences: a review of applications and methods. IEEE journal of biomedical and health informatics **20**(1), 304–321 (2014)
4. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)
5. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE Transactions on Robotics **37**(6), 1874–1890 (2021)
6. Cheng, K., Ma, Y., Sun, B., Li, Y., Chen, X.: Depth estimation for colonoscopy images with self-supervised learning from videos. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24. pp. 119–128. Springer (2021)
7. Denninger, M., Winkelbauer, D., Sundermeyer, M., Boerdijk, W., Knauer, M., Strobl, K.H., Humt, M., Triebel, R.: 2: A procedural pipeline for photorealistic rendering. Journal of Open Source Software **8**(82), 4901 (2023). https://doi.org/10.21105/joss.04901
8. Fu, Z., Jin, Z., Zhang, C., Dai, Y., Gao, X., Wang, Z., Li, L., Ding, G., Hu, H., Wang, P., et al.: Visual-electromagnetic system: A novel fusion-based monocular localization, reconstruction, and measurement for flexible ureteroscopy. The International Journal of Medical Robotics and Computer Assisted Surgery **17**(4), e2274 (2021)
9. Geraghty, R.M., Jones, P., Somani, B.K.: Worldwide trends of urinary stone disease treatment over the last two decades: a systematic review. Journal of endourology **31**(6), 547–556 (2017)
10. Han, J.J., Acar, A., Henry, C., Wu, J.Y.: Depth anything in medical images: A comparative study. arXiv preprint arXiv:2401.16600 (2024)

11. Han, J.J., Acar, A., Kavoussi, N., Wu, J.Y.: Meshbrush: Painting the anatomical mesh with neural stylization for endoscopy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 380–390. Springer (2024)
12. Jeong, B.H., Kim, H.K., Son, Y.D.: Depth estimation of endoscopy using sim-to-real transfer. arXiv preprint arXiv:2112.13595 (2021)
13. Klodmann, J., Schlenk, C., Hellings-Kuß, A., Bahls, T., Unterhinninghofen, R., Albu-Schäffer, A., Hirzinger, G.: An introduction to robotically assisted surgical systems: current developments and focus areas of research. Current Robotics Reports **2**(3), 321–332 (2021)
14. Liu, X., Sinha, A., Unberath, M., Ishii, M., Hager, G.D., Taylor, R.H., Reiter, A.: Self-supervised learning for dense depth estimation in monocular endoscopy. In: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis: First International Workshop, OR 2.0 2018, 5th International Workshop, CARE 2018, 7th International Workshop, CLIP 2018, Third International Workshop, ISIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings 5. pp. 128–138. Springer (2018)
15. Ma, R., McGill, S.K., Wang, R., Rosenman, J., Frahm, J.M., Zhang, Y., Pizer, S.: Colon10k: a benchmark for place recognition in colonoscopy. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1279–1283. IEEE (2021)
16. Mahmood, F., Chen, R., Sudarsky, S., Yu, D., Durr, N.J.: Deep learning with cinematic rendering: fine-tuning deep neural networks using photorealistic medical images. Physics in Medicine & Biology **63**(18), 185012 (2018)
17. Martyniak, S., Kaleta, J., Dall'Alba, D., Naskręt, M., Płotka, S., Korzeniowski, P.: Simuscope: Realistic endoscopic synthetic dataset generation through surgical simulation and diffusion models. arXiv preprint arXiv:2412.02332 (2024)
18. Oliva Maza, L., Steidle, F., Klodmann, J., Strobl, K., Triebel, R.: An orb-slam3-based approach for surgical navigation in ureteroscopy. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **11**(4), 1005–1011 (2023)
19. Ozyoruk, K.B., Gokceler, G.I., Bobrow, T.L., Coskun, G., Incetan, K., Almalioglu, Y., Mahmood, F., Curto, E., Perdigoto, L., Oliveira, M., et al.: Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Medical image analysis **71**, 102058 (2021)
20. Recasens, D., Lamarca, J., Fácil, J.M., Montiel, J., Civera, J.: Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. IEEE Robotics and Automation Letters **6**(4), 7225–7232 (2021)
21. Reyes-Amezcua, I., Espinosa, R., Daul, C., Ochoa-Ruiz, G., Mendez-Vazquez, A.: Endodepth: A benchmark for assessing robustness in endoscopic depth prediction. In: MICCAI Workshop on Data Engineering in Medical Imaging. pp. 84–94. Springer (2024)
22. Schlenk, C., Hagmann, K., Steidle, F., Oliva Maza, L., Kolb, A., Hellings-Kuß, A., Schöb, D.S., Klodmann, J., Miernik, A., Albu-Schäffer, A.: A robotic system for solo surgery in flexible ureteroscopy: development and evaluation with clinical users. International Journal of Computer Assisted Radiology and Surgery **18**(9), 1559–1569 (2023)

23. Schlenk, C., Hellings-Kuß, A., Hagmann, K., Budjakoski, N., Maza, L.O., Klodmann, J., Müller-Spahn, S., Steidle, F., Miernik, A., Albu-Schäffer, A.: Coflex top: A teleoperation system for flexible ureteroscopy. IEEE Robotics and Automation Letters (2024)
24. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: 2012 IEEE/RSJ international conference on intelligent robots and systems. pp. 573–580. IEEE (2012)
25. Türk, C., Petřík, A., Sarica, K., Seitz, C., Skolarikos, A., Straub, M., Knoll, T.: Eau guidelines on interventional treatment for urolithiasis. European urology **69**(3), 475–482 (2016)
26. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. arXiv preprint arXiv:2406.09414 (2024)
27. Yoshida, K., Kawa, G., Taniguchi, H., Inoue, T., Mishima, T., Yanishi, M., Sugi, M., Kinoshita, H., Matsuda, T.: Novel ureteroscopic navigation system with a magnetic tracking device: a preliminary ex vivo evaluation. Journal of Endourology **28**(9), 1053–1057 (2014)