

EndoX: A GPU-Accelerated Endoscopic Perception Simulation Framework

Kwan Tung Lam^{*1,2}, Ping Shu Ho^{*1,✉}, Patt Phurtivilai^{1,3}, and Ka Chun Cheung¹

¹ NVIDIA AI Technology Center, NVIDIA Corporation, USA

² The Chinese University of Hong Kong, Hong Kong, China

³ The University of Hong Kong, Hong Kong, China

cliffh@nvidia.com

Abstract. Endoscopic examination of gastrointestinal organs increasingly relies on 3D computer vision techniques such as depth estimation, surface reconstruction, and missing region detection. Yet rigorous method development for endoscopy remains challenging, as ground-truth data cannot be readily acquired during clinical procedures. We present *EndoX*, a GPU-accelerated synthetic data generation (SDG) framework and interactive simulator for medical endoscopy that enables efficient generation of fully annotated datasets, addressing the data scarcity gap in algorithm development. EndoX automates anatomically accurate 3D mesh generation from computed tomography (CT) data and employs physics-based rendering to produce RGB imagery together with geometric modalities: depth, surface normals, optical flow, occlusion, and camera coverage maps, with precise ground-truth geometry. For validation, we constructed organ-specific datasets by navigating a virtual capsule endoscope through anatomically accurate CT-derived organ meshes, with sensor and lighting parameters configured to approximate real capsule endoscopes. Quantitative evaluations on depth estimation, pose estimation, and 3D Gaussian Splatting demonstrate that models trained on EndoX data achieve competitive accuracy, and qualitative assessment confirms transferability to clinical endoscopy. The framework is designed for scalable deployment on GPU nodes via containerization. A sample dataset comprising three video sequences (12,000 frames) with paired ground-truth annotations and 3D models is available. The source code of the extension for NVIDIA Omniverse, documentation, and the sample dataset are available at <https://beatricelam111.github.io/EndoX/>.

Keywords: Endoscopy · Synthetic dataset generation · Simulated gastrointestinal tract.

1 Introduction

Three-dimensional (3D) computer vision techniques—including depth estimation, surface reconstruction, and coverage analysis—could transform gastroin-

* These authors contributed equally to this work.

testinal (GI) endoscopy by enabling quantitative assessment of luminal geometry and ensuring complete mucosal inspection. However, the absence of ground-truth geometric data in real procedures poses a fundamental obstacle to rigorous algorithm training, validation, and benchmarking. Unlike scenes where depth sensors or LiDAR can provide supervision, the confined intraluminal environment precludes direct geometric measurement during endoscopy, leaving algorithm development reliant on qualitative evaluation or small-scale phantom studies.

General endoscopic datasets such as Kvasir-Capsule [1] and Galar [2] support disease detection but lack spatial annotations. Specialized datasets have been introduced for depth estimation [3], visual odometry, and Simultaneous Localization and Mapping (SLAM) research [4,5]. Yet inter-patient variability, tissue deformability, inconsistent lighting across endoscopic systems, and limited anatomical coverage complicate ground-truth acquisition and algorithm robustness. Moreover, existing datasets lack the multimodality and diversity required for modern computer vision models and 3D reconstruction algorithms.

Addressing these gaps requires photorealistic imagery, spatial annotations, and 3D ground truth. **Contribution:** We propose *EndoX*, a GPU-accelerated synthetic data generation framework and high-fidelity simulator for medical endoscopy that: (i) automates anatomically accurate 3D mesh generation from CT data across diverse patient anatomies; (ii) employs real-time physically-based rendering (PBR) for optically accurate RGB imagery alongside co-registered geometric ground-truth modalities, with customizable sensor, lighting, material, and trajectory parameters; and (iii) is designed for scalable deployment for high-throughput data production with the containerized headless version to be publicly released upon acceptance.

1.1 Related Work

Endoscopic Datasets for 3D algorithms. Existing GI endoscopic datasets provide inadequate support for the development of 3D computer vision algorithms, constrained by limited diversity. C3VD [6] and C3VDv2 [7] provide annotated imagery with 3D ground truth, but manual acquisition using colon phantoms inherently limits data diversity. The EndoSLAM dataset [4] is based on *ex-vivo* recordings in porcine organs and virtual environments using VR-Caps [8]. While providing ground-truth 3D models and camera poses, the experimental setting requires careful sample handling, precise robotic calibration, and manual tissue preparation—yielding modest data quantities insufficient to cover diverse clinical scenarios. Moreover, they lack geometric ground-truth such as depth and normal maps. The EndoNeRF dataset [9] enables evaluation of neural rendering techniques but comprises only two surgical cases without high-quality 3D geometric ground truth.

Endoscopy Simulators for Synthetic Data Generation. SDG for medical AI/ML research has emerged in response to the demand for large-scale, high-quality, privacy-preserving datasets for training and validating deep learning models. Researchers have developed endoscopy simulators to generate synthetic datasets

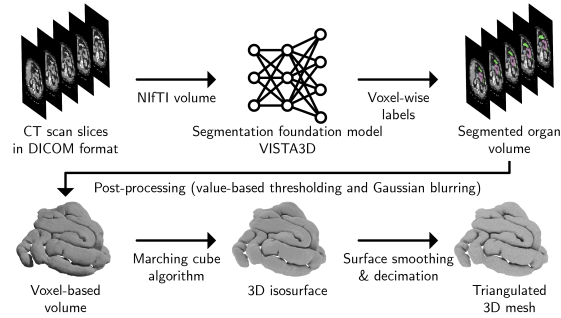


Fig. 1. CT-to-mesh reconstruction pipeline converting 2D CT slices into a 3D mesh model.

that facilitate algorithm development for advanced medical imaging applications. Recent works such as VR-Caps [8] and SimIntestine [10] build virtual GI environments using Unity and Blender, respectively, enabling endoscopic camera navigation and data capture. However, these simulators face an inherent trade-off: VR-Caps relies on rasterization in Unity, achieving real-time performance at the cost of physically accurate rendering, while SimIntestine employs a path tracing renderer in Blender, offering photorealism but at prohibitive offline computation costs. SimuScope [11] attempts to bridge this gap by pairing Unity with Stable Diffusion v1.5 [12], fine-tuned via low-rank adaptation, for image-to-image translation with higher fidelity, yet the probabilistic nature of diffusion models limits temporal consistency across generated frames. None of these approaches efficiently delivers physically based, photorealistic imagery at scale, nor provides an automated pipeline for scalable SDG in a headless, containerized application.

2 Methods

2.1 Generation of 3D Models

EndoX automates generation of 3D GI organ meshes from clinical CT data in DICOM format (Fig. 1), supporting scalable creation of anatomical models for SDG. Clinical CT DICOM sequences are converted to volumetric NIfTI format and segmented using VISTA3D [13], producing voxel-wise binary masks that delineate organ boundaries. The masks are applied to each organ volume, followed by value-based thresholding to remove irrelevant voxels (e.g., air and imaging artifacts) and Gaussian blurring to reduce surface noise. A 3D isosurface is then extracted from each binary volume using the marching cubes algorithm [14] with the VTK package [15], retaining only the largest continuous structure. Iterative windowed-sinc smoothing and quadratic edge collapse decimation are applied to reduce mesh imperfections and decrease face count while preserving anatomical topology. The refined mesh is exported in Wavefront OBJ format.

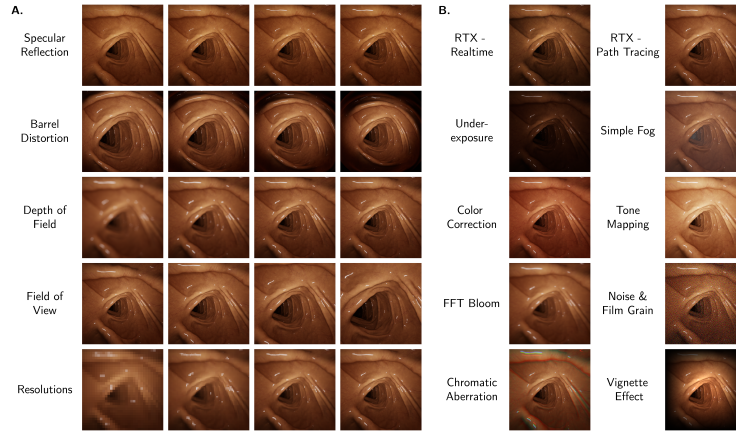


Fig. 2. A: Rendered views under varying camera intrinsics configurations using RTX Path Tracing renderer. B: Post-processing effects applied directly to the rendered views (RTX Path Tracing); RTX Real-Time mode shown for reference only.

2.2 Simulator with physics-based rendering

EndoX is built on the NVIDIA Omniverse platform, using GPU-accelerated real-time ray tracing with PBR for scalable, high-fidelity image synthesis.

Environment setup and navigation path. Generated 3D meshes are imported into an interactive simulation environment. Centerlines of GI tracts are automatically extracted using a Voronoi diagram-based method [16] with the VMTK package [17] to define a virtual sensor navigation path. Alternatively, user-supplied coordinate arrays in *.csv* or *.npy* format can be loaded and interpolated as custom paths.

RTX renderer and post-processing stack. Leveraging the NVIDIA RTX renderer with Omniverse Replicator, EndoX employs zero-copy data extraction, directly accessing GPU memory regions without CPU-mediated transfer to enable simultaneous extraction of multiple arbitrary output variables (AOV) in a single rendering pass, with multi-GPU load balancing for near-linear scaling. Custom image modalities requiring additional array operations are accelerated using NVIDIA Warp [18], which JIT-compiles Python functions to CUDA kernels, preserving the zero-copy pipeline until the final transfer to host memory to save rendered output. The post-processing stack can be configured to simulate optical imperfections common in clinical endoscopes—tone mapping, color correction, FFT bloom, chromatic aberration, vignetting, noise and film grain, and exposure control (Fig. 2B)—and can be combined for domain randomization across synthetic frames.

Sensor modeling and data acquisition. The virtual sensor is fully configurable: camera intrinsics and light source properties are exposed as adjustable param-

eters (Fig. 2A). As a representative use case, we model a capsule endoscope to address the scarcity of capsule endoscopy datasets. Commercial capsule endoscopes have an angle of view ranging from 140 to 174 degrees; here we adopt a 140-degree angle of view with 512×512 image resolution. Virtual light sources are configured and attached as child objects of the capsule. The configured capsule traverses the defined navigation path, capturing RGB and geometric modalities per frame, with its world transformation matrix recorded as ground-truth pose.

Rendered image modalities. *Surface normals* are extracted per frame and encoded in two coordinate systems. In the world coordinate system, the X , Y , Z normal components are mapped to the R, G, B channels, each linearly scaled from $[-1, 1]$ to $[0, 65535]$ as a 16-bit color image. For the camera coordinate system, world normals $\mathbf{n}_{\text{world}}$ are transformed via the camera rotation matrix $\mathbf{R}$ as $\mathbf{n}_{\text{camera}} = \mathbf{n}_{\text{world}} \mathbf{R}^\top$ and encoded identically. *Depth* is computed per-pixel by normalizing the distance between the camera and the surface intersection point as $\text{Depth} = (d - d_{\text{near}}) / (d_{\text{far}} - d_{\text{near}})$, where d is the distance from the camera to the ray-mesh intersection, and d_{near} , d_{far} are the near and far clipping planes. *Optical flow* quantifies per-pixel displacement between consecutive frames. For pixel (x, y) , the motion vector is:

$$\mathbf{M}_{xy} = -[(\mathbf{T}_{xy, \text{curr}} + \mathbf{J}_{\text{curr}}) - (\mathbf{T}_{xy, \text{prev}} + \mathbf{J}_{\text{prev}})] \quad (1)$$

where $\mathbf{T}_{xy, \text{curr}}$ and $\mathbf{T}_{xy, \text{prev}}$ are screen-space pixel coordinates in the current and previous frames, and $\mathbf{J}_{\text{curr}}$, $\mathbf{J}_{\text{prev}}$ are projection jitter offsets in temporal anti-aliasing. This formulation corrects for the jitter, ensuring motion vectors reflect only actual scene and camera movement.

Computation of occlusion frames. Occlusion frames are computed as a custom post-processing modality to quantify scene visibility at each pixel. For each camera view, the near clipping plane is incrementally advanced from d_{near} to d_{far} in fixed steps. At each step, an 8-bit image encoding world-space positions is captured. The initial frame establishes a per-pixel reference (foreground). In subsequent frames, a pixel is considered occluded if its value differs from the reference beyond a tunable threshold ϵ and is not $(0, 0, 0)$ (background), indicating that a further mesh surface has become visible as d_{near} advances. The resulting binary occlusion map $C_{i,j}$, formalized as:

$$C_{i,j} = \begin{cases} 1 & \text{if } \exists k \in \{1, \dots, N\} : \|\mathbf{p}_{k,i,j} - \mathbf{p}_{0,i,j}\| > \epsilon \text{ and } \mathbf{p}_{k,i,j} \neq \mathbf{0} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $\mathbf{p}_{k,i,j}$ is the world position at step k and pixel (i, j) , $\mathbf{p}_{0,i,j}$ is the reference value, and N is the total number of near-plane advancements.

Computation of coverage maps. The coverage map records whether each mesh vertex is visible from the camera during traversal, computed via view-projection transforms and Z-buffer visibility testing. Each world-space vertex $\mathbf{v}_w = [x \ y \ z \ 1]$

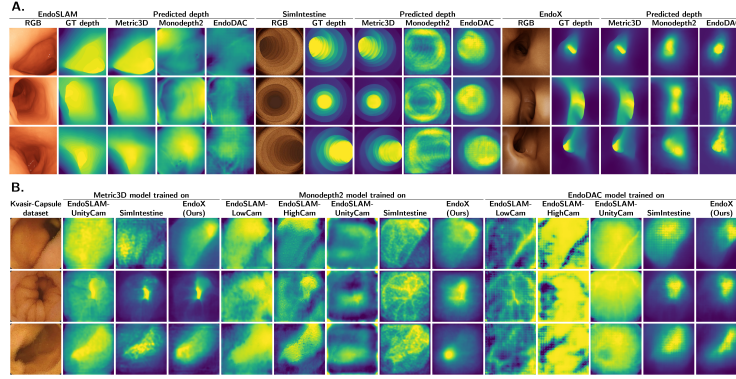


Fig. 3. A. Ground-truth and predicted depth maps of synthetically trained models evaluated on synthetic frames. B. Predicted depth maps of the models tested on clinical Kvasir-Capsule images.

is projected to normalized device coordinates (NDC) by $\mathbf{v}_{\text{NDC}} = [x' \ y' \ z'] / w'$, where $[x' \ y' \ z' \ w'] = \mathbf{v}_w V P$ with view matrix V and projection matrix P (row-vector convention), then mapped to screen space as $\mathbf{v}_{\text{screen}} = [\frac{W}{2}(x_{\text{ndc}} + 1), \frac{H}{2}(1 - y_{\text{ndc}})]$, where W and H are image dimensions in pixels. Mesh faces are rasterized via barycentric interpolation, and a Z-buffer $\mathcal{Z}$ (initialized to $+\infty$) is updated at each pixel (p_x, p_y) with the smallest encountered z . For each vertex v_i with projected depth z_i and screen position $(p_{x,i}, p_{y,i})$, the visibility boolean is defined with floating-point tolerance $\tau = 1 \times 10^{-8}$ as:

$$\gamma_i = \begin{cases} 1 & \text{if } |z_i - \mathcal{Z}(p_{x,i}, p_{y,i})| < \tau \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

3 Evaluated Tasks and Results

Evaluation with depth estimation models. Metric3D v2 [19], Monodepth2 [20], and EndoDAC [21] were each independently trained and evaluated on the small intestine frames from three synthetic datasets: EndoSLAM-UnityCam (ES-U),

Table 1. Performance of depth estimation models trained with synthetic datasets.

	Metric3D v2			Monodepth2			EndoDAC		
	ES-U	SI	EndoX	ES-U	SI	EndoX	ES-U	SI	EndoX
RMSE ↓	0.0082	0.0798	0.0108	0.0506	0.1947	0.0539	0.0477	0.1247	0.0897
RMSLE ↓	0.0953	0.1992	0.0680	0.5405	0.7182	0.3018	0.5122	0.4129	0.3238
Abs. Rel. ↓	0.0682	0.1616	0.0530	0.5334	0.5375	0.2611	0.5066	0.3509	0.2996
Sq. Rel. ↓	0.0007	0.0122	0.0007	0.0324	0.0805	0.0180	0.0285	0.0347	0.0420
a1 ↑	0.9742	0.7347	0.9875	0.3262	0.2635	0.5191	0.3375	0.4926	0.6051
a2 ↑	0.9940	0.9772	0.9999	0.5891	0.5343	0.8572	0.6183	0.7820	0.8436
a3 ↑	0.9977	0.9941	0.9999	0.7765	0.7083	0.9796	0.8022	0.8916	0.9352

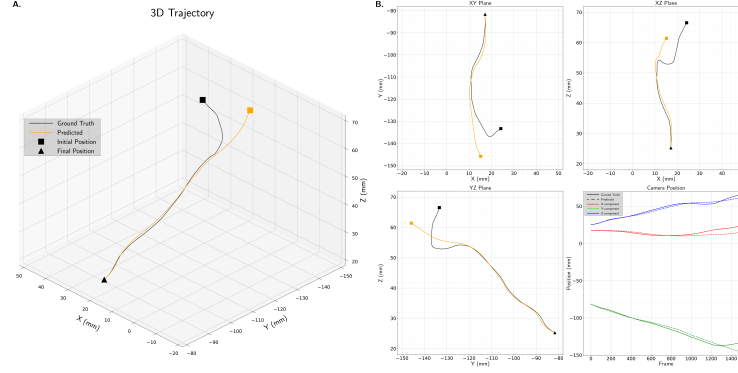


Fig. 4. Pose estimation results using Monodepth2 trained on EndoX small intestine data. A. 3D trajectories. B. 2D plane projections of trajectories and per-axis component comparison between ground-truth and predicted trajectories.

SimIntestine (SI), and EndoX. For each dataset, data were split 7:2 for training and testing, with a further 9:1 training/validation split; each model was trained on the training split of a given dataset and evaluated against the ground-truth depth maps of the corresponding test split. Performance is reported in Table 1 using standard error metrics and threshold accuracy a_i , defined as the percentage of pixels satisfying $\max(d_{\text{pred}}/d_{\text{gt}}, d_{\text{gt}}/d_{\text{pred}}) < 1.25^i$ for $i \in \{1, 2, 3\}$.

Results confirm that EndoX synthetic data effectively supports depth estimation model training across both supervised and self-supervised paradigms. For Metric3D v2, EndoX achieves the lowest Abs. Rel. (0.053) and near-perfect a_1 (0.987). For Monodepth2 and EndoDAC, EndoX-trained models consistently yield the best threshold accuracies and relative errors, while ES-U retains marginal RMSE advantages. Across all architectures, EndoX performs strongly on relative and ratio-based metrics, suggesting that EndoX produces well-calibrated RGB–depth pairs suitable for depth learning. Predictions on synthetic frames compared against the ground truth are shown in Fig. 3A. Since ground-truth depth is unavailable for clinical capsule endoscopy, the trained models were qualitatively evaluated on the Kvasir-Capsule dataset [1] (Fig. 3B). The models trained on EndoX data produce smoother, higher-contrast depth maps with better structural detail preservation compared to those trained on EndoSLAM or SimIntestine data, indicating that EndoX closely approximates the visual characteristics of real endoscopic imagery.

Evaluation with pose estimation model. Pose estimation was evaluated using Monodepth2 [20] trained on a 7,000-frame sequence from the EndoX small intestine dataset and tested on a separate 1,000-frame sequence. The predicted trajectory is compared against the ground-truth path in Fig. 4, showing both 3D and per-axis component decomposition. The predicted trajectory follows the ground truth, with an absolute trajectory error (ATE) of 3.92 mm. The per-component plots confirm that the model captures the dominant motion trends

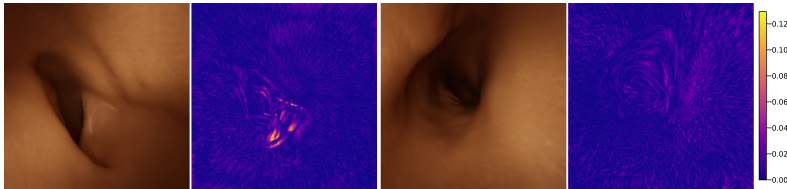


Fig. 5. Representative novel views by 3DGS models and corresponding loss maps.

along each axis, though minor drift accumulates over the sequence. These results validate that EndoX synthetic data, with accurate camera pose annotations, provides a suitable training signal for self-supervised visual odometry in endoscopy.

Evaluation with 3D reconstruction technique. We evaluated 3D reconstruction on the EndoX small intestine dataset by projecting each pixel into 3D point clouds via Open3D using ground-truth depth and training a vanilla 3D Gaussian Splatting (3DGS) model. Novel views were rendered from held-out test poses and quantitatively compared against ground-truth frames. The model achieves a PSNR of 40.27 dB, SSIM of 0.978, and LPIPS of 0.231, confirming high photometric fidelity of the reconstructed scene. Representative novel-view rendering and per-pixel MAE loss map are shown in Fig. 5, with values in $[0, 0.24]$ and an average MAE of 0.007 across test views. A video showing the rendered 3DGS is provided in the supplementary materials.

Conclusion. We introduced EndoX, a framework integrating automated anatomical mesh reconstruction with a GPU-accelerated data processing pipeline to generate annotated endoscopic imagery. Evaluations on depth estimation, pose estimation, and 3DGS show that models trained on EndoX data achieve competitive accuracy and generalize to clinical capsule endoscopy.

4 Datasets and Distribution

The example scene in the sample dataset was reconstructed from CT colonography data (CC BY 3.0) [22], with the mesh surface texture adopted from VR-Caps [8]. For each video sequence, we provide: (i) a triangulated *3D mesh* reconstructed from CT data (Wavefront OBJ), and (ii) a vertex-wise *coverage map* in vertex indices array in .npz format; a video visualizing the coverage map is included in the supplementary materials. For every frame, we provide the annotations: *Depth*—distance from the camera plane to the surface, clamped to the near/far clipping planes (0.1–100 mm), linearly normalized, and encoded as a 16-bit grayscale image; *Surface normals*—reported in world and camera coordinates, with the X , Y , Z components stored in the R, G, B channels, each linearly mapped from $[-1, 1]$ to $[0, 65535]$ as a 16-bit color image; *Optical flow*—per-pixel displacement between consecutive frames, where the R and G channels encode X - and Y -direction motion (± 512 px), each linearly mapped to $[0, 65535]$

as a 16-bit color image (the first frame carries no flow value); *Occlusion*—an 8-bit binary mask within the 1–100 mm clipping range, with pixels occluding other mesh faces set to 255 and all others to 0; and *Camera pose*—per-frame world-space translation and rotation stored as `tx`, `ty`, `tz`, `qx`, `qy`, `qz`, `qw` (unit quaternion) in a text file.

Limitations. Quantitative evaluation of sim-to-real transfer on clinical capsule endoscopy was not performed, owing to the absence of suitable real-world datasets with ground-truth geometry. The sample datasets cover only the colon and small bowel as demo cases, though the framework supports any GI organ that can be reconstructed from CT. Pathology-bearing CT inputs are supported for generating imagery with clinical abnormalities; however, physical artifacts such as debris, blood, and surgical tools currently require manual scene authoring.

Disclosure of Interests. Kwan Tung Lam is an intern at NVIDIA Corporation. Ping Shu Ho and Ka Chun Cheung are employees of NVIDIA Corporation. The remaining author has no competing interests to declare that are relevant to the content of this article.

References

1. Smedsrud, P.H., Thambawita, V., Hicks, S.A., Gjestang, H., Nedrejord, O.O., Næss, E., Borgli, H., Jha, D., Berstad, T.J.D., Eskeland, S.L., Lux, M., Espeland, H., Petlund, A., Nguyen, D.T.D., Garcia-Ceja, E., Johansen, D., Schmidt, P.T., Toth, E., Hammer, H.L., de Lange, T., Riegler, M.A., Halvorsen, P.: Kvasir-Capsule, a video capsule endoscopy dataset. *Scientific Data* **8**, 142 (2021).
2. Le Floch, M., Wolf, F., McIntyre, L., Weinert, C., Palm, A., Volk, K., Herzog, P., Kirk, S.H., Steinhäuser, J.L., Stopp, C., Geissler, M.E., Herzog, M., Sulk, S., Kather, J.N., Meining, A., Hann, A., Palm, S., Hampe, J., Herzog, N., Brinkmann, F.: Galar - a large multi-label video capsule endoscopy dataset. *Scientific Data* **12**, 828 (2025).
3. Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zeffiro, T., Xia, W., Zhu, Z., Luo, H., Jia, F., Zhang, X., Li, X., Sharan, L., Kurmann, T., Schmid, S., Sznitman, R., Psychogios, D., Azizian, M., Stoyanov, D., Maier-Hein, L., Speidel, S.: Stereo correspondence and reconstruction of endoscopic data challenge. *arXiv preprint arXiv:2101.01133* (2021).
4. Ozyoruk, K.B., Gokceler, G.I., Bobrow, T.L., Coskun, G., Incetan, K., Almalioglu, Y., Mahmood, F., Curto, E., Perdigoto, L., Oliveira, M., Sahin, H., Araujo, H., Alexandrino, H., Durr, N.J., Gilbert, H.B., Turan, M.: EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Medical Image Analysis* **71**, 102058 (2021).
5. Azagra, P., Sostres, C., Ferrández, Á., Riazuelo, L., Tomasini, C., Barbed, O.L., Morlana, J., Recasens, D., Batlle, V.M., Gómez-Rodríguez, J.J., Elvira, R., López, J., Oriol, C., Civera, J., Tardós, J.D., Murillo, A.C., Lanas, A., Montiel, J.M.M.: Endomapper dataset of complete calibrated endoscopy procedures. *Scientific Data* **10**, 671 (2023).

6. Bobrow, T.L., Golhar, M., Vijayan, R., Akshintala, V.S., Garcia, J.R., Durr, N.J.: Colonoscopy 3D video dataset with paired depth from 2D-3D registration. *Medical Image Analysis* **90**, 102956 (2023).
7. Golhar, M.V., Fretes, L.S.G., Ayers, L., Akshintala, V.S., Bobrow, T.L., Durr, N.J.: C3VDv2 - Colonoscopy 3D video dataset with enhanced realism. *arXiv preprint arXiv:2506.24074* (2025).
8. İncetan, K., Celik, I.O., Obeid, A., Gokceler, G.I., Ozyoruk, K.B., Almalioglu, Y., Chen, R.J., Mahmood, F., Gilbert, H., Durr, N.J., Turan, M.: VR-Caps: A virtual environment for capsule endoscopy. *Medical Image Analysis* **70**, 101990 (2021).
9. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2022*. LNCS, vol. 13437, pp. 431–441. Springer, Cham (2022).
10. Singh, S., Bhaumik, B., Chatterjee, S.: SimIntestine: A synthetic dataset from virtual capsule endoscope. *Medical Image Analysis* **105**, 103706 (2025).
11. Martyniak, S., Kaleta, J., Dall’Alba, D., Naskręt, M., Plotka, S., Korzeniowski, P.: SimuScope: Realistic endoscopic synthetic dataset generation through surgical simulation and diffusion models. In: *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2025*. pp. 4268–4278. IEEE Press, New York (2025).
12. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022*. pp. 10684–10695. IEEE Press, New York (2022).
13. He, Y., Guo, P., Tang, Y., Myronenko, A., Nath, V., Xu, Z., Yang, D., Zhao, C., Simon, B., Belue, M., Harmon, S., Turkbey, B., Xu, D., Li, W.: VISTA3D: A unified segmentation foundation model for 3D medical imaging. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2025*. pp. 20863–20873. IEEE Press, New York (2025).
14. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3D surface construction algorithm. In: *14th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH ’87)*. pp. 163–169. ACM, New York (1987).
15. Schroeder, W., Martin, K., Lorensen, B.: *The visualization toolkit*. 4th edn. Kitware, New York (2006).
16. Antiga, L., Ene-Iordache, B., Remuzzi, A.: Centerline computation and geometric analysis of branching tubular surfaces with application to blood vessel modeling. In: *11th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG) 2003*, University of West Bohemia, Plzen (2003).
17. Izzo, R., Steinman, D., Manini, S., Antiga, L.: The vascular modeling toolkit: A Python library for the analysis of tubular structures in medical images. *Journal of Open Source Software* **3**(25), 745 (2018).
18. Macklin, M.: Warp: A high-performance Python framework for GPU simulation and graphics. In: *NVIDIA GPU Technology Conference (GTC) (2022)*.
19. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3Dv2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(12), 10579–10596 (2024).
20. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.: Digging into self-supervised monocular depth estimation. In: *IEEE/CVF International Conference on Computer Vision (ICCV) 2019*. pp. 3827–3837. IEEE Press, New York (2019).

21. Cui, B., Islam, M., Bai, L., Wang, A., Ren, H.: EndoDAC: Efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2024. LNCS, vol. 15006, pp. 208–218. Springer, Cham (2024).
22. Smith, K., Clark, K., Bennett, W., Nolan, T., Kirby, J., Wolfsberger, M., Moulton, J., Vendt, B., Freymann, J.: Data from CT colonography. The Cancer Imaging Archive (2015).