
PolypSense3D: A Multi-Source Benchmark Dataset for Depth-Aware Polyp Size Measurement in Endoscopy

Ruyu Liu^{a,b} Lin Wang^a Mingming Zhou^a Jianhua Zhang^{c,d,*} Haoyu Zhang^{a,*}
Xiufeng Liu^b Xu Cheng^d Sixian Chan^e Yanbing Shen^f Sheng Dai^f
Yuping Yang^g Yaochu Jin^g Lingjuan Lv^h

^a Hangzhou Normal University ^b Technical University of Denmark

^c Hohai University ^d Tianjin University of Technology

^e Zhejiang University of Technology ^f Sir Run Run Shaw Hospital Zhejiang

^g Westlake University ^h Sony AI

* Corresponding authors

Correspondence to: Jianhua Zhang <zjh@ieee.org>, Haoyu Zhang <haoyu.zhang@hznu.edu.cn>
{ruyu.liu, xiufeng, xu.cheng}@ieee.org, {2023111011034, 2024112011018}@stu.hznu.edu.cn
{shenyb2006, colon}@zju.edu.cn, {yanyuping, jinyaochu}@westlake.edu.cn, Lingjuan.Lv@sony.com

Abstract

Accurate polyp sizing during endoscopy is crucial for cancer risk assessment but is hindered by subjective methods and inadequate datasets lacking integrated 2D appearance, 3D structure, and real-world size information. We introduce **PolypSense3D**, the first multi-source **benchmark dataset** specifically targeting depth-aware polyp size measurement. It uniquely integrates over 43,000 frames from virtual simulations, physical phantoms, and clinical sequences, providing synchronized RGB, dense/sparse depth, segmentation masks, camera parameters, and millimeter-scale size labels derived via a novel forceps-assisted in-vivo annotation technique. To establish its value, we benchmark state-of-the-art segmentation and depth estimation models. Results quantify significant domain gaps between simulated/phantom and clinical data and reveal substantial error propagation from perception stages to final size estimation, with the best fully automated pipelines achieving an average Mean Absolute Error (MAE) of 0.95 mm on the clinical data subset. Publicly released under CC BY-SA 4.0 with code and evaluation protocols, PolypSense3D offers a standardized platform to accelerate research in robust, clinically relevant quantitative endoscopic vision. The benchmark dataset and code are available at: <https://github.com/HNUicda/PolypSense3D> and <https://doi.org/10.7910/DVN/K13H89>.

1 Introduction

Endoscopy is the cornerstone of early gastrointestinal cancer screening, enabling the detection and removal of high-risk polyps and saving millions of lives annually [52, 36]. However, a critical limitation persists: while endoscopists visualize polyp morphology, accurately measuring polyp size in real-time remains a significant challenge. This information gap directly impacts clinical decision-making, as polyp diameter strongly correlates with malignancy risk – polyps over 20 mm carry up to 15% risk, compared to <2% for those under 10 mm [16]. Precise sizing is therefore essential for risk assessment and tailoring surveillance strategies [16]. Current methods, including subjective visual estimation or error-prone ex-vivo measurements [10], are inadequate, often leading to misjudgments and potentially inappropriate patient management. Consequently, an automated,

objective, and accurate method for in-vivo polyp size measurement is an urgent clinical imperative and presents a significant challenge for quantitative computer vision.

Recent computer vision advances have sped up progress in endoscopic image analysis [37], particularly in 2D segmentation and 3D reconstruction. Datasets like Kvasir-SEG [19] and CVC-ColonDB [51] have hatched sophisticated 2D polyp segmentation algorithms [44, 14, 23]. However, focusing solely on pixel-level masks, these datasets fundamentally lack the 3D spatial context (depth) and real-world scale information indispensable for physical size measurement. A 200-pixel polyp mask yields no true size about whether the lesion is 3 mm or 15 mm, a distinction critical for clinical action. Conversely, while 3D perception datasets like SimCol3D [41] or EndoMapper [1], derived from simulations [40, 6] or SLAM on phantoms [39], provide valuable geometric information for navigation or scene understanding, they typically lack fine-grained annotations of specific polyp instances and, crucially, omit the corresponding ground-truth physical size measurements needed for metrology. This creates a significant bottleneck for machine learning research targeting clinical size estimation: **segmentation datasets lack the necessary geometric and scale information**, while **3D reconstruction datasets lack the specific, instance-level size labels**. Therefore, no existing public resource provides the integrated multi-modal data (appearance, structure, scale) required to develop, train, and rigorously benchmark machine learning models for the specific, clinically vital task of accurate, depth-aware polyp size estimation directly from endoscopic video.

To fill this critical gap in data resources for quantitative endoscopic vision research, we introduce **PolypSense3D**, the first comprehensive, multi-source benchmark dataset specifically engineered for developing and evaluating depth-aware polyp size measurement techniques. PolypSense3D bridges virtual simulation, physical modeling, and real clinical scenarios, providing a unique resource comprising over 43,000 frames. It distinctively features synchronized RGB images, corresponding dense (virtual/physical-estimated) or sparse (clinical) depth maps, precise 2D segmentation masks, calibrated camera intrinsic parameters, and, most importantly, **verifiable millimeter-scale ground-truth polyp size annotations**. Data spans realistic virtual polyps derived from CT data (sizes 1.79-20.52mm), precisely fabricated polyps in 3D-printed phantoms offering single textures (sizes 4.0-14.89mm), and challenging in-vivo clinical cases (sizes 1.39-9.98mm) annotated using our novel, reproducible biopsy-forceps-assisted technique (detailed in Section 3.3 and validated in Appendix D). By providing all essential components – appearance (RGB), structure (depth), camera geometry (intrinsic), and physical scale (size labels) – within a unified framework, PolypSense3D enables the ML community to tackle the full pipeline from perception (segmentation, depth) to measurement (size quantification). It serves as a much-needed standardized platform to benchmark algorithms, quantify domain shifts, analyze error propagation, and ultimately accelerate the development of robust, clinically relevant quantitative endoscopic vision systems.

The primary contributions of this work are:

- **The PolypSense3D Benchmark Dataset:** We construct and release the first large-scale, multi-source (virtual, physical, clinical) benchmark dataset specifically designed for developing and evaluating depth-aware polyp size measurement algorithms. It features over 16k/43k+ (with polys/total) frames with synchronized RGB, depth, segmentation masks, camera parameters, and ground-truth millimeter-scale size annotations. The dataset’s dense virtual and sparse clinical depth enable a pre-training/fine-tuning workflow. This is designed to resolve the scale ambiguity in self-supervised methods and adapt large models for clinically relevant, end-to-end measurement tasks.
- **Novel Verifiable In-Vivo Annotation Protocol:** We develop, detail, and provide initial validation (see Appendix D) for a reproducible biopsy-forceps-assisted annotation strategy, enabling quantitative size and sparse depth estimation directly from standard clinical endoscopic video sequences. This sparse annotation process is both challenging and highly valuable, adding real-world complexity to the benchmark dataset.
- **Comprehensive Benchmarking for ML Models:** We establish strong baseline results on PolypSense3D by evaluating representative state-of-the-art segmentation and depth estimation models. Our analysis quantifies performance across the challenging virtual-physical-clinical domain shifts and investigates error propagation dynamics, demonstrating the dataset’s utility for rigorous ML model evaluation and highlighting key research challenges. We also investigated the adaptability of foundational models in real-world scenarios, emphasizing that large-scale foundational models can be initially trained using dense data and subsequently fine-tuned with sparse clinical annotations to better align with the requirements of real-world applications.

- **Open Resources for Transparency and Reproducibility:** We publicly release the PolypSense3D dataset, associated annotation tools (including code for the forceps-based annotation method), baseline implementation code, and evaluation protocols under permissive open-source licenses (CC BY-SA 4.0, MIT License), ensuring transparency and promoting reproducible research within the NeurIPS community and beyond.

2 Related Work

Our work builds upon advancements in endoscopic computer vision, specifically in 2D polyp segmentation and 3D perception. We review existing work in these areas to precisely situate the contribution of PolypSense3D.

Table 1: Summary of public endoscopic datasets for polyp analysis. Abbreviations: Polyp Segmentation (PS), Depth Estimation (DE), 3D Reconstruction (3DR), Dense Depth Map (DDM), Sparse Point Cloud (SPC).

Dataset	Type	Organ	Task	Image Count	Resolution(s)	Depth Type	Depth Source	Mask
ASU-Mayo [3]	Clinical	Colon	PS	18902	—	—	—	✓
ETIS-Larib [48]	Clinical	Colon	PS	196	1225 × 966	—	—	✓
CVC-ClinicDB [4]	Clinical	Colon	PS	612	576 × 768	—	—	✓
CVC-ColonDB [51]	Clinical	Colon	PS	380	574 × 500	—	—	✓
Kvasir-SEG [19]	Clinical	Colon	PS	1000	Mixed	—	—	✓
UCL Depth [40]	Virtual	Colon	DE	16016	256 × 256	DDM	CT	—
Zhang et al. [68]	Virtual	Colon	DE / 3DR	Seqs.	—	DDM	CT	—
VR-Caps [18]	Virtual	Colon	DE / 3DR	Seqs.	—	DDM	CT	—
Yang et al. [62]	Virtual	Colon	DE	4500	—	DDM	CT	—
EndoSLAM [39]	Physical	Colon+	Disp. Est.	42700+	Mixed	SPC	3D Scanner	—
C3VD [6]	Physical	Colon	DE / 3DR	10015	—	DDM	3D Model	—
EndoMapper [1]	Clinical	Colon	3DR	59+ Seqs.	320 × 240	SPC	SLAM	—
LDPolypVideo [34]	Clinical	Colon	PS	33k / 40k	560 × 480	—	—	✓
SUN-SEG [21]	Clinical	Colon	PS	49k / 158k	1158 × 1008	—	—	✓
PolypSense3D	Mixed	Colon	PS/DE/3DR	16k+ / 43k+	Mixed	Mixed	Mixed	✓

2.1 2D Polyp Segmentation Datasets and Methods

Significant progress in 2D endoscopic polyp segmentation has been enabled by high-quality public datasets such as Kvasir-SEG [19], CVC-ClinicDB [4], CVC-ColonDB [51], ETIS-Larib [48], and the ASU-Mayo Clinic dataset [3]. These resources have encouraged the development of sophisticated algorithms, initially dominated by Convolutional Neural Network (CNN) based encoder-decoder architectures like U-Net and its variants [44, 7, 72], which excel at capturing fine-grained spatial details. Innovations include attention mechanisms and specialized modules focusing on boundary refinement or feature enhancement [14, 2, 64, 55, 56, 46]. More recently, Transformer-based architectures [13, 50, 59, 66, 67] and large-scale pre-trained models like the Segment Anything Model (SAM) [23] and its adaptations [11, 26] have demonstrated strong capabilities, particularly in handling global context and challenging cases like small or occluded polyps [22].

Despite impressive gains in segmentation accuracy, existing 2D datasets suffer from a critical limitation for clinical metrology: they **lack associated depth or scale information**. Segmentation masks are purely planar representations (pixel coordinates) without any link to the polyp’s real-world physical dimensions. This fundamental **absence of calibrated spatial context and ground-truth size annotations** prevents the direct use of these datasets for developing or validating algorithms aimed at quantitative polyp measurement, necessitating datasets that integrate these missing components.

2.2 Endoscopic 3D Perception Datasets and Methods

Endoscopic 3D perception techniques, including depth estimation and 3D reconstruction, aim to recover the spatial structure of the colon [41, 63]. Resources supporting this research include: **Virtual synthetic datasets** [40, 68, 18], often generated from CT models, provide paired RGB-depth data but may lack photorealism, exhibit non-physiological scaling, or omit realistic lesion modeling. **Physical phantom datasets** [39, 6] use 3D-printed models to offer greater realism but are often limited to

sparse depth annotations due to sensor capabilities. **Clinical datasets** [1] provide authenticity but typically lack ground-truth depth maps due to intraoperative constraints.

Methodologically, traditional geometric approaches like Structure-from-Motion (SfM) and SLAM [35, 49, 15, 24] struggle with the texture-poor and deformable nature of endoscopic scenes, often yielding sparse, scale-ambiguous reconstructions. Learning-based methods have shown promise, leveraging supervised training on synthetic data [33, 42, 38, 29, 62] or self-supervised signals like photometric consistency from video [31, 69, 12, 47, 43, 30, 17]. However, supervised methods face the synthetic-to-real domain gap, while self-supervised methods can be sensitive to illumination changes and lack metric scale.

Crucially, both existing 3D perception paradigms and datasets primarily focus on global scene reconstruction or dense depth estimation across the entire view, rather than providing the specific tools needed for accurate, lesion-centric metrology. They generally lack datasets containing numerous, well-defined polyp instances that are simultaneously annotated with **both precise 3D location/structure and verified ground-truth physical size**. This disconnect prevents the direct evaluation of algorithms on the clinically vital task of polyp measurement. PolypSense3D is specifically designed to fill this gap by providing multi-source data with jointly annotated segmentation, depth, camera parameters, and critically, ground-truth millimeter-scale polyp sizes.

3 PolypSense3D Dataset Construction

To create a robust benchmark for depth-aware polyp size measurement, PolypSense3D integrates data from three distinct yet complementary sources. Each source utilizes a tailored pipeline, detailed below, designed to maximize fidelity while addressing specific research needs (e.g., controlled ground truth vs. real-world complexity). Exhaustive protocols, parameter details, and specific validation studies (e.g., for the clinical annotation protocol in Appendix D) for each source are provided in Appendix A (Virtual), Appendix B (Physical), and Appendix C (Clinical).

3.1 Virtual Simulation Dataset: Controlled Environment with Dense GT

By operating on the virtual camera of simulator, we built a high-fidelity virtual colon environment with Polyps in Unity based on Vr-Caps [18]. This allows for precise ground truth and systematic variation. Using anonymized patient CT scans from The Cancer Imaging Archive (TCIA)¹, we reconstructed anatomical models via InVesalius², refined them in Blender for anatomical accuracy, and accurately scaled them to realistic human dimensions (validation in Appendix A.2.1). We then designed and procedurally textured 30 distinct polyp models, varying shape, size (1.79-20.52mm), and morphology (details in Appendix A.2.2), embedding them in clinically relevant locations (ascending, transverse, sigmoid colon, near folds, bends) with varied orientations. These virtual polyp models provide a solid and diverse foundation for validation. It is important to note that this dataset is not restrictive, and our modeling pipeline will be open-sourced, allowing researchers to customize polyp variants according to their specific needs.

A custom physics-based controller simulated realistic endoscope dynamics (forward/backward translation, pitch/yaw rotation), mimicking capsule endoscopy movement patterns (controller details in Appendix A.3). Adjustable lighting (intensity, cone angle, position) and camera parameters (FOV set to 78.4°, typical focal length 200mm equivalent) were tuned to approximate clinical conditions while avoiding rendering artifacts (Appendix A.4). Using Unity Recorder, we captured over 32,000 synchronized frames including: (1) High-resolution RGB images with polyps; (2) Dense, per-pixel metric depth maps (16-bit PNG, derived from geometry buffer); (3) Calibrated camera intrinsics per frame; (4) 6DoF camera pose trajectory logs; (5) Automatically generated pixel-perfect polyp segmentation masks (derived from model geometry), subsequently reviewed and confirmed by experienced clinical endoscopists. This yields a rich dataset ideal for supervised training and controlled evaluation of segmentation, depth, and size estimation.

¹<https://www.cancerimagingarchive.net>

²<https://invesalius.github.io/>

3.2 Physical Phantom Dataset: Bridging the Reality Gap

To introduce real-world textures, lighting, and sensor characteristics, we created a physical benchmark. A 50cm section of the transverse colon was 3D printed using White Resin material on a UnionTech SLA lite600 3D printer (further design details in Appendix B.1), based on CT data [39]. This phantom features an interlocking modular design for assembly and a semi-open structure for external visibility during experiments. 13 solid polyps (diameters 4.00-14.89 mm) were physically embedded at known locations. We acquired video data using a commercially available small-diameter veterinary endoscope connected via USB (eCap software), manually navigating the lumen.

Crucially, we performed meticulous camera calibration *prior* to data acquisition using Zhang's method [70] implemented in MATLAB. A $4 \times 5 \times 3$ mm checkerboard was determined optimal after evaluating multiple sizes and distances (details in Appendix B.2). Calibration was performed at a working distance of 22mm, capturing views across all image quadrants. The resulting intrinsic parameters (see Appendix B.2 for final parameters used) were used for all subsequent image undistortion and metric calculations. This physical dataset (13 videos) provides realistic appearance data paired with known polyp sizes and locations, ideal for evaluating sim-to-real transfer and validating scale estimation from camera intrinsics.

3.3 Real Clinical Dataset: In-Vivo Complexity and Sparse Cues

To provide data from the target environment, we collected videos during routine colonoscopies using standard clinical equipment, specifically an Olympus EVIS EXERA III endoscopic system with an Olympus CF-H290I colonoscope, adhering to approved institutional protocols (IRB approval details in Section 8). The clinical endoscope was calibrated using the same rigorous procedure as the phantom endoscope (details in Appendix B.2). We extracted stable video segments ("freeze-frames") showing detected polyps during inspection.

Obtaining dense, metric ground truth in-vivo is infeasible; therefore, we developed and rigorously applied a novel **biopsy-forceps-assisted annotation protocol**, whose validation regarding reproducibility and inter-operator variability is detailed in Appendix D:

1. **Reference Frame Selection & Segmentation:** To minimize the effects of camera motion and intestinal peristalsis, stable and clear frames are first selected by trained annotators under the supervision of experienced endoscopists. Segmentation masks are then manually refined with the assistance of prompt-based models such as SAM [23], ensuring accurate and consistent delineation of polyp boundaries.
2. **Size Annotation (Proximal Comparison):** In designated frames captured by the physician, fully opened biopsy forceps (calibrated 5mm physical tip-to-tip distance) are positioned adjacent to the polyp, aiming for the same apparent depth plane. Annotators precisely mark the forceps tips (illustrated in Appendix C.3). The pixel distance d_{pixel} yields a local scale factor: $scale = 5\text{mm}/d_{pixels}$ (mm/pixel). This scale factor is applied to the refined segmentation mask's dimensions (e.g., major axis L_{pixels}^{major}) to compute the polyp's physical size (e.g., $L_{mm}^{major} = L_{pixels}^{major} \times scale$). This method grounds the measurement in a known physical reference visible within the frame.
3. **Sparse Depth Annotation (Extension Contact):** In separate stable frames, forceps with clear 5mm gradation markings are used. The physician extends the forceps so the tip gently contacts the polyp apex or adjacent flat mucosa. The annotator measures the visible pixel length of the extended marked portion of the forceps. Using the known physical length per gradation (5mm), calibrated camera intrinsics, and a carefully validated perspective rectification and regression model (detailed in Appendix C.3), the metric depth Z of that contact point is calculated. This provides sparse but metrically accurate depth points on or near the lesion.

Quality control involved cross-checking annotations between multiple trained annotators and physician review of selected cases (details in Appendix C.3 and Appendix D). While this protocol introduces potential operator variability and provides only sparse depth, it offers the first feasible method, to our knowledge, for obtaining quantitative size and partial depth labels directly from standard in-vivo colonoscopy videos, creating a uniquely challenging and realistic dataset reflecting true clinical conditions, particularly for smaller polyps prevalent in screening.

Table 2: Summary Statistics of the PolypSense3D Dataset. See Appendix E for detailed breakdown.

Metric	Virtual Simulation	Physical Phantom	Clinical Data	Total / Overall
Total Frames (approx.)	~32k	~13 videos	~11k	~43k+
Frames with Annotated Polyps (approx.)	~16k (50%)	~74 (cut from 13 videos)	~438	~16k
Unique Polyp Instances	30 (Models)	13 (Embedded)	53 (from 127 Patients)	114
Size Range (mm, approx. diameter)	1.79–20.52	4.00–14.89	1.39–9.98	1.39–32.8
Size Distribution (<10 / 10–20 / >20 mm)	50.0% / 43.3% / 6.7%	69.2% / 30.8% / 0.0%	100% / 0.0% / 0.0%	80.2% / 17.71% / 2.1%
Depth Annotation Type	Dense (GT)	-	Sparse (Forceps GT)	Mixed
Resolution(s)	320×320(Scalable)	640×480	1157×1006	Multiple
Camera Intrinsics Provided	Yes (Full)	Yes (Calibrated)	Yes (Calibrated)	Yes

4 Dataset Analysis and Statistics

PolypSense3D offers a substantial and diverse resource, comprising over 43,000 frames across three progressively challenging subsets, synthetic, physical, and clinical subsets. These subsets are designed to reflect increasing levels of real-world complexity. The synthetic subset provides dense, precise annotations under idealized conditions but lacks realistic textures and lighting variability. The physical dataset, built from 3D-printed models, introduces real and accurate geometry while remaining limited in dynamic content and texture richness. The clinical subset poses the greatest challenge, featuring uncontrolled endoscope motion, variable illumination, specular highlights, complex mucosal textures, tool occlusions, and motion blur from real procedures.

A distinctive feature of the clinical data is its multimodal richness: a single polyp may correspond to multiple frames annotated with RGB images, depth maps, and segmentation masks. In total, nearly 16,000 polyp-containing frames are annotated, enabling fine-grained evaluation of detection, measurement, and spatial reasoning algorithms. Table 2 provides a high-level overview, and detailed statistics on polyp characteristics, imaging resolutions, and camera parameters are included in Appendix E.

5 Benchmark Experiments and Evaluation

We conducted experiments using PolypSense3D to establish baselines for depth-aware polyp sizing. Our evaluation aims to: (1) assess the performance of state-of-the-art models on each dataset component, (2) evaluate model robustness across the three progressively challenging subsets, synthetic (idealized with dense annotations), physical (real geometry with static but limited texture), and clinical (highly realistic with motion, lighting, and occlusion challenges), and (3) analyze how upstream errors in segmentation and depth estimation propagate to final size measurements.

5.1 Experimental Setup

Tasks & Metrics: (1) *2D Polyp Segmentation* (mDice, mIoU, Recall, Precision, F1 Score); (2) *Monocular Depth Estimation* (RMSE($10 \times e - 3$), AbsRel, Log10, $\delta < 1.25^x$); (3) *End-to-End Polyp Size Estimation* (MAE mm), using outputs from (1) and (2). For key mean metrics, standard deviations are reported in Appendix G to indicate result variability. **Baseline Models:** Segmentation: MSNet[28], PraNet [14], SAM [23], MobileSAM [65], VM-UNet [45]. Depth: DAM-V1 [60], DAM-V2 [61], ZoeDepth [5], DAM V2-mini ³ [61]. Rationale included SOTA status and diverse architectures. **Implementation:** Models trained/fine-tuned on our training split were evaluated across all test splits (7:3 ratio). These models were all trained on our Virtual Simulation Dataset. Details regarding training procedures, hyperparameters, and computational resources are provided in Appendix F. The size estimation pipeline is detailed in Appendix G.1.

5.2 Benchmark Results

We measure polyp size in metric scale by combining contour segmentation and absolute depth estimation (see Appendix G.1 for details). Comprehensive evaluations are conducted on all three proposed datasets for: (1) polyp segmentation, (2) depth estimation, and (3) integrated size measurement. Full per-polyp results and extended error analyses are available in Appendix G.

³<http://github.com/RubyQianru/Depth-Anything-V2-Mini>

Virtual Simulation Evaluation: On the virtual test set, segmentation models exhibited strong performance (Table 3). Among them, Sam2-unet achieved the highest mDice of 0.9399, indicating excellent adaptability to synthetic endoscopic imagery. MSNet and PraNet followed closely. MobileSAM and VM-UNet yielded lower scores, likely due to MobileSAM’s lightweight design optimized for speed, and VM-UNet’s shallow architecture with limited representational capacity. Depth estimation models, trained on densely labelled data, have achieved remarkably strong performance (Table 10), with DAM V2-mini showing the lowest RMSE (0.185e-03). This level of accuracy in depth estimation holds significant promise for substantially enhancing polyp size measurement. Combining predicted segmentation (MSNet) and ground truth depth yielded an average size MAE of 0.43mm (Table 6), while using ground truth segmentation and predicted depth resulted in a higher MAE of 4.95mm, quantifying error propagation even in this controlled setting. Detailed per-polyp results are in Table 16 (Appendix G).

Physical Phantom Evaluation: Applying models to the physical phantom revealed a clear sim-to-real gap. Compared to the synthetic dataset, the lack of realistic texture in the 3D-printed environment led to a general drop in performance across tasks. As shown in Table 3, segmentation accuracy decreased for all models, though polyp-specific methods such as MSNet and Sam2-unet still maintained satisfactory results. These findings align with the qualitative visualizations provided in Appendix G.2. Depth estimation was particularly affected by the inability to acquire dense, full-frame ground-truth depth on this platform, limiting both training and evaluation to qualitative visualizations (also in Appendix G.2). Compared to the synthetic domain, estimated depth maps on physical data exhibited degraded detail and smoothness. These limitations in segmentation and depth estimation propagated to polyp size measurement. As shown in Table 7, size estimation errors on the phantom were notably higher than in the virtual setting, with all methods yielding an Abs.Error above 1mm. This underscores the challenge of domain transfer from ideal simulation to physical systems with realistic geometry but limited texture and sensing variation. A per-polyp breakdown is provided in Table 17 (Appendix G).

Clinical Data Evaluation: The clinical test set represents the most challenging scenario due to in vivo factors such as variable lighting, fluid interference, tissue motion, and camera shake. These conditions significantly degrade performance, as evidenced by further drops in segmentation accuracy (e.g., MSNet mDice 0.6989, Sam2-unet mDice 0.7346; Table 3). Dense ground-truth depth is unavailable; instead, we employed biopsy forceps contact points for sparse but reliable supervision (Appendix C.3.2). These annotations confirmed a notable decline in depth accuracy compared to the synthetic dataset, which features dense depth supervision and near-ideal imaging conditions (Table 5). Despite increased segmentation and depth errors, automated size estimation remained competitive. The average MAE reached 0.95mm, outperforming physician visual estimates (1.84mm Table 8). This highlights the potential of algorithmic measurements to assist clinical polyp assessment, even under realistic constraints. Per-case results are provided in Table 18 (Appendix G).

Performance Across Multi-source Datasets

Performance consistently declined from the Unity to the Physical and Clinical datasets, reflecting increasing real-world complexity. The Unity subset, built with 3D-modeled polyps in a clean, artifact-free environment, enables precise ground-truth generation. The Physical set

Table 3: Segmentation Performance on Unity, Physical, and Clinical Datasets

Dataset	Method	mDice	mIoU	Recall	Precision	FI Score
Unity	MSNet[28]	0.9301	0.9000	0.9816	0.9169	0.9412
	PraNet [14]	0.9062	0.8525	0.9733	0.8676	0.9062
	SAM [23]	0.8803	0.8180	0.8552	0.9418	0.8803
	MobileSAM [65]	0.8341	0.7054	0.8776	0.8334	0.8341
	VM-Unet [45]	0.8432	0.7288	0.9172	0.8044	0.8432
	ASPS [25]	0.8951	0.8321	0.9409	0.8717	0.8951
Physical	Sam2-unet [57]	0.9399	0.9010	0.9723	0.9243	0.9428
	MSNet[28]	0.8812	0.8432	0.9425	0.8457	0.8847
	PraNet [14]	0.8523	0.8028	0.9106	0.8138	0.8523
	SAM [23]	0.8693	0.8218	0.9603	0.8351	0.8693
	MobileSAM [65]	0.6874	0.5091	0.6304	0.9887	0.6874
	VM-Unet [45]	0.2546	0.1459	0.3631	0.1960	0.2546
Clinical	ASPS [25]	0.7526	0.6834	0.7599	0.7649	0.7540
	Sam2-unet [57]	0.8983	0.8501	0.9981	0.8513	0.9064
	MSNet[28]	0.6989	0.6250	0.7441	0.7201	0.7068
	PraNet [14]	0.7183	0.6316	0.7082	0.7999	0.7183
	SAM [23]	0.7992	0.7032	0.8569	0.7978	0.7992
	MobileSAM [65]	0.2728	0.1437	0.2131	0.9586	0.2728
Physical	VM-Unet [45]	0.0634	0.0328	0.0780	0.0526	0.0634
	ASPS [25]	0.5179	0.4239	0.5574	0.5586	0.5183
	Sam2-unet [57]	0.7346	0.6665	0.8720	0.7208	0.7524

Table 4: Benchmark on Unity Dataset: Depth Estimation Performance

Method	Error Metrics ↓			Accuracy Metrics ↑	
	RMSE	abs.REL	Log10	δ_1	δ_2
DAM V1 [60]	0.242	0.020	0.009	0.998	0.999
DAM V2 [61]	0.216	0.015	0.006	0.996	0.999
ZoeDepth [5]	0.213	0.015	0.006	0.998	0.999
DAM V2-mini [61]	0.185	0.011	0.005	0.997	0.999

Table 5: Benchmark on Clinical Dataset: Depth Estimation Performance

Method	Error Metrics ↓				Accuracy Metrics ↑	
	RMSE	abs.REL	Log10	MAE	δ_1	δ_2
DAM V1 [60]	10.983	0.428	0.249	8.348	0.098	0.390
DAM V2 [61]	8.089	0.383	0.152	5.757	0.463	0.732
ZoeDepth [5]	12.481	0.491	0.319	9.784	0.122	0.268
DAM V2-mini [61]	8.801	0.359	0.162	6.069	0.415	0.659

adds real imaging and geometry via 3D-printed models but lacks texture diversity. The Clinical set poses the greatest challenge, with motion blur, lighting variation, tool occlusions, and complex tissue textures. These results highlight the need for segmentation models with stronger cross-domain robustness.

Table 6: Benchmark on Unity Dataset: Per-Polyp Size Estimation (mm). Full table in Appendix G (Table 16).

Polyp ID	Label (Forceps)	Seg(GT)+Depth(Pred)		Seg(Pred)+Depth(Pred)		Doctor Visual Est.(Avg)	
		Predicted	Abs. Error	Predicted	Abs. Error	Predicted	Abs. Error
P1	11.55	13.46	1.92	19.06	7.51	9.26	2.29
...	7.55	2.49
Average	9.57	9.27	4.34	14.42	7.52	7.26	4.88

Table 7: Benchmark on 3D-Printed Dataset: Per-Polyp Size Estimation (mm). Full table in Appendix G (Table 17).

Polyp ID	Label	Seg(GT)+Depth(Pred)		Seg(Pred)+Depth(Pred)	
		Predicted	Abs. Error	Predicted	Abs. Error
P1	6.20	6.22	0.02	8.46	2.26
P13	4.94	6.74	1.08	8.89	3.85
Average	8.03	8.38	1.48	9.48	1.99

Table 8: Benchmark on Clinical Dataset: Per-Case Size Estimation (mm). Full table in Appendix G (Table 18).

Polyp ID	Label (Forceps)	Seg(GT)+Depth(Pred)		Seg(Pred)+Depth(Pred)		Doctor Visual Est.(Avg)	
		Predicted	Abs. Error	Predicted	Abs. Error	Predicted	Abs. Error
P1	3.28	2.17	1.11	1.73	1.55	4.20	0.92
...
P53	5.95	5.99	0.04	4.22	1.73	8.67	2.72
Average	3.28	2.80	1.19	2.66	0.95	4.73	1.84

6 Discussion

PolypSense3D provides a comprehensive benchmark for depth-aware polyp size estimation from two key perspectives. First, it enables systematic evaluation of segmentation and depth estimation performance across three increasingly challenging datasets: synthetic (ideal and richly annotated), physical (real imaging with limited texture), and clinical (dynamic, noisy, and highly realistic). This reveals clear performance degradation, especially when models trained on virtual data are applied to clinical scenarios. Second, the dataset allows analysis of how upstream perception errors propagate to downstream size estimation. Despite these errors, automated measurements remain competitive with physician estimates, underscoring the clinical value of robust, end-to-end methods under real-world conditions.

Limitations and Biases: While PolypSense3D offers unique advantages for benchmarking, certain limitations must be acknowledged. Virtual simulations, despite efforts towards realism, cannot fully replicate complex tissue biomechanics or intricate light-tissue interactions. Physical phantoms utilize simplified materials and lack the dynamic nature of living tissue. Our novel clinical annotation method, while providing invaluable in-vivo quantitative labels currently unavailable elsewhere, relies on physician interaction during the procedure and yields only sparse depth information. This introduces potential operator variability and inherent precision limits tied to operator skill and strict adherence to the protocol; a detailed validation of this protocol, including inter-operator agreement and accuracy assessments, is provided in Appendix D. We have implemented measures like standardized training and cross-checks to mitigate this variability, but it remains a factor reflecting real-world measurement challenges. In annotating and predicting polyp size, our estimation method is limited by the biopsy forceps’ angle of approach, position, and the perspective of the polyp images, which may lead to certain inaccuracies. Additionally, the clinical data also originates from a specific, single-center institutional context, reflecting its patient demographics and endoscopic equipment, which may introduce sampling bias and limit immediate generalizability without adaptation (ethical considerations in Section 8). Users must carefully consider these factors when designing experiments and interpreting results obtained using PolypSense3D.

Future Research Directions Enabled by PolypSense3D: This benchmark dataset is designed to catalyze progress in several key areas of machine learning relevant to the NeurIPS community:

- **Domain Adaptation and Generalization:** The pronounced domain gap between PolypSense3D’s virtual, physical, and clinical subsets provides a challenging, realistic testbed for evaluating and developing novel unsupervised and semi-supervised domain adaptation techniques, with ongoing efforts to expand multi-center, cross-device data benchmark collection. This is crucial for transferring models trained on readily available synthetic or controlled data to real-world clinical settings where labeled data is scarce.
- **Learning with Sparse and Imperfect Labels:** The clinical subset, with its sparse depth annotations derived via the forceps protocol (validation in Appendix D) and associated measurement variability, directly facilitates research into methods robust to sparse, potentially noisy, or weakly supervised labels. This includes areas like few-shot depth completion, uncertainty-aware learning, incorporating geometric priors, or learning robust representations that are less sensitive to label noise. At the same time, these sparse depth measurements can effectively address the inherent scale ambiguity in self-supervised methods, serving as constraints to assist the self-supervised sparse-to-dense depth completion framework with ground truth (GT) size and depth annotations[32]. Additionally, sparse measurement points can act as strong cues to guide or fine-tune large foundational models[27], enabling them to adapt to the clinical domain and recover absolute scale.
- **End-to-End Quantitative Perception:** By providing synchronized multi-modal data including ground-truth size, PolypSense3D enables the direct training and evaluation of end-to-end networks that predict polyp real-scale size from endoscopic images or video snippets[58]. Such approaches could potentially bypass cascaded segmentation and depth estimation steps, mitigating cumulative error propagation.
- **Leveraging Temporal Information:** Although presented frame-wise, the underlying video sequences (especially for clinical and phantom data) open avenues for exploring temporal consistency models. Utilizing information across frames could lead to more robust and temporally stable depth and size estimations, drawing on techniques from video understanding and time-series analysis[8].
- **Uncertainty Quantification for Clinical Trust:** The multi-source nature, inclusion of challenging clinical data, and availability of ground-truth allow for benchmarking methods that quantify prediction uncertainty[71]. Reliable uncertainty estimates for automated size measurements are essential for assessing model trustworthiness and facilitating safer clinical integration.
- **Multi-Task Learning and Synergies:** The dataset naturally supports investigating multi-task learning frameworks that jointly optimize for segmentation, depth estimation, and size prediction, potentially uncovering synergistic benefits between these related tasks.

PolypSense3D provides the necessary data diversity and annotations to rigorously pursue these and other related ML research questions in quantitative medical vision.

7 Conclusion

Accurate polyp size measurement is a critical unmet need in clinical endoscopy. We introduced PolypSense3D, the first multi-source benchmark dataset specifically designed to facilitate research and evaluation in depth-aware polyp size estimation. By combining virtual, physical, and clinically annotated data with synchronized multi-modal information (RGB, depth, segmentation, size, camera parameters), PolypSense3D provides an essential resource currently missing in the field. Our comprehensive benchmark experiments establish strong baselines, quantify the performance challenges faced by current SOTA models, particularly regarding domain generalization and error propagation, and confirm the dataset’s value for evaluating the complete perception-to-measurement pipeline. We release PolypSense3D publicly with code and protocols, aiming to standardize evaluation and accelerate the development of reliable automated tools for quantitative endoscopic analysis, ultimately benefiting patient care.

8 Ethical Considerations

The intestinal models utilized for both our virtual synthetic dataset and physical phantom in this study were derived from publicly available medical data. The collection of clinical data was conducted by experienced medical professionals at Sir Run Run Shaw Hospital, Zhejiang University School of

Medicine, and received approval from the hospital’s ethics committee. Throughout the clinical data acquisition process, we strictly adhered to relevant regulations, ensuring the anonymization of all data to comprehensively protect patient privacy and personal information. It is crucial to emphasize that this dataset is intended to advance research in the field of in-vivo spatial intelligence perception based on endoscopic images. Its purpose is not for direct clinical diagnosis or medical decision-making, and any clinical application based on this dataset should undergo thorough evaluation and supervision by qualified medical professionals.

9 Acknowledgements

We would like to express our gratitude to Chenyu Yan, Yihao Ying, Tianyu Zhao, Yuanyuan Zhang, Gaoqi Huang, Yibo Wang, and Peng Lu from Tianjin University of Technology, Hangzhou Normal University for their contributions and efforts in the experimental annotation work. The work is supported by the Marie Skłodowska-Curie Postdoctoral Individual Fellowship under Grant No. 101154277, National Natural Science Foundation of China under Grant 62202137, 62306097, Zhejiang Provincial Natural Science Foundation of China under Grant LMS25F020009.

References

- [1] Pablo Azagra, Carlos Sostres, Ángel Ferrández, Luis Riazuelo, Clara Tomasini, O León Barbed, Javier Morlana, David Recasens, Víctor M Batlle, Juan J Gómez-Rodríguez, et al. Endomapper dataset of complete calibrated endoscopy procedures. *Scientific Data*, 10(1):671, 2023.
- [2] Debapriya Banik, Kaushiki Roy, Debotosh Bhattacharjee, Mita Nasipuri, and Ondrej Krejcar. Polyp-net: A multimodel fusion network for polyp segmentation. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2020.
- [3] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.
- [4] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.
- [5] Shariq Farooq Bhat, Reiner Birk, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.
- [6] Taylor L Bobrow, Mayank Golhar, Rohan Vijayan, Venkata S Akshintala, Juan R Garcia, and Nicholas J Durr. Colonoscopy 3d video dataset with paired depth from 2d-3d registration. *Medical image analysis*, 90:102956, 2023.
- [7] Patrick Brandao, Evangelos Mazomenos, Gastone Ciuti, Renato Caliò, Federico Bianchi, Arianna Menciassi, Paolo Dario, Anastasios Koulaouzidis, Alberto Arezzo, and Danail Stoyanov. Fully convolutional neural networks for polyp segmentation in colonoscopy. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, pages 101–107. Spie, 2017.
- [8] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Unsupervised monocular depth and ego-motion learning with structure and semantics, 2019. URL <https://arxiv.org/abs/1906.05717>.
- [9] Marlin Wayne Causey, David E Rivadeneira, and Scott R Steele. Historical and current trends in colon trauma. *Clinics in colon and rectal surgery*, 25(04):189–199, 2012.
- [10] Louis Chaptini, Adib Chaaya, Fedele Depalma, Krystal Hunter, Steven Peikin, and Loren Laine. Variation in polyp size estimation among endoscopists and impact on surveillance intervals. *Gastrointestinal endoscopy*, 80(4):652–659, 2014.