



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Understanding Model Behavior in Monocular Polyp Sizing

Xinqi Xiong¹, Andrea Dunn Beltran¹, Junmyeong Choi¹, Sarah K. McGill¹,
Marc Niethammer², and Roni Sengupta¹

¹ University of North Carolina at Chapel Hill, Chapel Hill NC 27514, USA
xxiong@cs.unc.edu

² University of California San Diego, La Jolla CA 92093, USA

Abstract. Accurate polyp size stratification guides surveillance decisions, with lesions larger than 5 mm typically requiring closer follow-up. However, monocular colonoscopy lacks a reliable metric reference. We present a diagnostic audit of binary polyp size classification (≤ 5 mm vs. > 5 mm) across multiple public multi-center datasets, model families, and patient-stratified cross-validation. Across architectures and input modalities, including RGB appearance, relative depth, and photometry, model performance is moderately consistent, suggesting reliance on cues correlated with examination behavior rather than true metric scales. By providing ground-truth scale at varying granularities, we quantify the potential improvement from perfect scale information and show that current depth estimation and global calibration offer limited gains. We further demonstrate that segmentation errors under distribution shift eliminate most of this potential, with oracle scale under predicted masks recovering only baseline performance. These results highlight metric scale and mask robustness as two independent bottlenecks and provide reusable evaluation tools—such as oracle scale ladders, shortcut partitions, and mask substitution—for auditing future polyp sizing pipelines. Our code is publicly accessible at <https://github.com/anaxqx/polyp-sizing-audit>.

Keywords: Monocular Polyp Sizing · Shortcut Learning · Diagnostic Audit · Robustness

1 Introduction

Colorectal cancer is the second leading cause of cancer death [27]. The clinical management of colorectal polyps is risk-stratified by size: diminutive polyps (≤ 5 mm) often allow for a 7–10 year surveillance interval if only 1–2 are found, while small polyps (6–9 mm) may require a 3–5 year follow-up depending on the total count and pathology. Large polyps (≥ 10 mm) are considered advanced findings that necessitate immediate resection and a mandatory 3-year surveillance interval [14]. However, in current clinical practice, physicians inspect the polyp visually and qualitatively estimate the size, often leading to a 40% error rate [9]. This inaccuracy introduces systematic bias and directly alters the surveillance intervals for approximately 10% of patients [8].

Automated polyp sizing could reduce inter-observer variability and support consistent surveillance decisions [28]. However, monocular endoscopy presents a geometric ambiguity: without a metric reference, a 3 mm polyp at a 10 mm distance is visually indistinguishable from a 6 mm polyp at 20 mm [15]. This scale ambiguity is further compounded by clinical examination behavior. Endoscopists naturally move the camera closer to diminutive polyps to confirm their status while evaluating larger lesions from a wider field of view. This can induce a spurious correlation between apparent scale and the size label in naturalistic training data. Furthermore, due to variability in endoscopy devices and imaging protocols in different clinical centers, the models trained on one dataset often struggle to generalize.

Prior automated approaches fall into three categories: explicit 3D reconstruction methods [16,12,26], reference-based methods [29], and end-to-end learned models [18,17,28,21,25]. 3D reconstruction methods require recovering full 3D geometry from monocular endoscopy, a problem that remains challenging in feature-poor, deformable colon environments [6]. In effect, they attempt to solve a harder problem as a prerequisite for a simpler one. Reference-based methods that use instruments such as forceps or water jets as scale anchors report low measurement error in controlled settings [29,21], but require the reference object to be present during the procedure, limiting scalability. End-to-end learned models report high accuracy on single-center datasets [18,17], but most implementations are not publicly available, and it remains unclear whether their reported gains reflect genuine recovery of metric scale or reliance on procedural shortcuts. We replicate representative approaches [17,28,25], and find that models across input modalities converge to the same performance ceiling, with diagnostic evidence pointing to shortcut reliance rather than geometric reasoning [13,10,7].

Despite growing interest in automated polyp sizing, the field lacks a controlled analysis of what limits current models and where future effort should be directed. We address this gap with a diagnostic benchmark for monocular polyp size classification (≤ 5 mm vs. >5 mm), evaluated across two public multi-center datasets [4,23], multiple model families, and patient-stratified cross-validation.

In this paper, we identify three primary factors that severely limit progress in monocular polyp size estimation. First, collecting large-scale, size-annotated multi-center data remains costly and logistically challenging, restricting the diversity of training distributions and resulting in poor generalization [4,28,21,23]. Second, recent foundation models for metric depth estimation [3,5] show strong zero-shot performance in natural scenes but fail under specular, textureless, and optically distorted endoscopic conditions. Third, although polyp segmentation models report high in-domain accuracy [11], it remains unclear whether this performance transfers to downstream tasks such as sizing under domain shift.

Our contributions are:

1. We show that models across architectures and input modalities (appearance, relative depth, photometry) reach a similar performance plateau (Macro-F1 ≈ 0.75). We provide diagnostic evidence that this ceiling reflects shortcut reliance on examination behavior rather than genuine geometric reasoning.

2. We introduce oracle scale probes at multiple granularities (per-frame, per-polyp, global) to isolate the performance headroom attributable to scale ambiguity (+16.1 pp), and show that current metric depth estimators do not recover this headroom in clinical data.
3. We demonstrate that segmentation quality acts as an independent bottleneck: replacing ground-truth masks with predictions from a segmentation model not trained on the target datasets erases the gains from oracle scale (-15.3 pp), indicating that in-domain segmentation benchmarks do not predict downstream sizing robustness under distribution shift.

2 Problem Formulation

We study monocular polyp sizing under a geometric scale ambiguity. Under a pinhole camera model, the apparent diameter A (pixels) of a polyp with true diameter S at distance Z is

$$A = \frac{fS}{Z}, \quad (1)$$

where f is the focal length. Since A depends on both S and Z , S is not identifiable from monocular appearance without a metric reference.

Clinical examination behavior further complicates this setting. Endoscopists often move closer to small polyps and view larger lesions from farther away, which can couple S and Z through behavior B ($S \rightarrow B \rightarrow Z$). As a result, the apparent scale A and procedure context B can correlate with the size label S , providing shortcuts that may remain stable across environments.

Motivated by this formulation, we audit monocular sizing via three independently controllable factors: data-model bias, metric scale, and mask quality. Each factor is evaluated by a targeted intervention in Sec. 4.1–4.3.

3 Experimental Setup

We establish an experimental protocol that isolates the contribution of each input modality from architectural choice. This section describes the data, evaluation, and diagnostic probes.

3.1 Datasets and Evaluation

We aggregate two public datasets with endoscopist-reported visual size estimates: Real-Colon [28], with 351,264 frames, 60 patients, and 132 polyps, and SUN-SEG [23], with 49,136 frames, 99 patients, and 100 polyps. The combined dataset contains 400,400 frames from 159 patients but only 232 unique polyps (of which 147 are ≤ 5 mm and 85 are > 5 mm). Given the high temporal correlation between adjacent video frames, the effective sample size is defined by the number of unique polyps rather than the total frame count.

Table 1: Binary polyp size classification (≤ 5 mm vs. > 5 mm). Results are mean $\pm$ std over 5-fold polyp-stratified CV. (Scale regimes follow Sec. 4.2: *None* uses depth as-is; *Global* fits one scale factor on the training set; *Oracle-polyp* & *Oracle-frame* fit per-clip & per-frame factors. Pred masks are from PolypPVT.)

Model	Input	Scale	Mask	Macro-F1	> 5 mm Rec.
<i>Data-model bias (cross-family comparison)</i>					
ResNet18	RGB	—	GT bbox	0.751 ± 0.058	0.673 ± 0.141
ViT-B	RGB	—	GT bbox	0.778 ± 0.097	0.664 ± 0.275
CNN3	Rel. depth [24]	None	GT bbox	0.725 ± 0.048	0.733 ± 0.099
MLP (w. geom)	Photometry	—	—	0.752 ± 0.055	0.704 ± 0.126
MLP (w.o. geom)	Photometry	—	—	0.694 ± 0.057	0.616 ± 0.133
ResNet18+V-REx	RGB	—	GT bbox	0.753 ± 0.058	0.663 ± 0.140
<i>Metric scale ladder (CNN3; baseline Macro-F1=0.725)</i>					
CNN3	Metric depth [22]	None	GT bbox	0.727 ± 0.055	0.751 ± 0.094
CNN3	Rel. depth [24]	Global	GT bbox	0.716 ± 0.057	0.742 ± 0.105
CNN3	Rel. depth [24]	Oracle-polyp	GT bbox	0.792 ± 0.033	0.778 ± 0.104
CNN3	Rel. depth [24]	Oracle-frame	GT bbox	0.886 ± 0.028	0.876 ± 0.047
<i>Mask quality robustness (CNN3 baseline=0.725; ResNet18 baseline=0.751)</i>					
CNN3	Rel. depth [24]	Oracle-frame	GT bbox $\times 0.8$	0.852 ± 0.035	0.831 ± 0.055
CNN3	Rel. depth [24]	Oracle-frame	GT bbox $\times 1.2$	0.859 ± 0.027	0.847 ± 0.050
CNN3	Rel. depth [24]	None	Pred mask	0.572 ± 0.025	0.559 ± 0.076
CNN3	Rel. depth [24]	Oracle-frame	Pred mask	0.736 ± 0.030	0.735 ± 0.053
ResNet18	RGB	—	GT bbox $\times 0.8$	0.758 ± 0.072	0.659 ± 0.161
ResNet18	RGB	—	GT bbox $\times 1.2$	0.753 ± 0.058	0.644 ± 0.163
ResNet18	RGB	—	Pred mask	0.643 ± 0.070	0.505 ± 0.175

All experiments use 5-fold cross-validation stratified by patient: all polyps from the same patient appear exclusively in the same fold to prevent data leakage. Metrics (Macro-F1 and > 5 mm Recall) are reported as mean $\pm$ standard deviation across folds. We focus on the 5 mm threshold, which corresponds to the clinically actionable boundary between diminutive and small polyps [14]. The 10 mm threshold is excluded because over 85% of polyps fall below it, creating a class imbalance that precludes reliable evaluation.

3.2 Diagnostic Probes and Interventions

We design three lightweight probes, each restricted to a single input modality: (A) a ResNet18 on masked RGB, (B) the depth-based sizing three-block CNN of Itoh et al. [17] on relative depth maps from PPSNet [24], and (C) a two-layer MLP on 51 handcrafted geometric and photometric, and texture features. A ViT-B on masked RGB is included to assess the effect of higher capacity. All models share a unified training pipeline: AdamW($\eta = 10^{-4}$, $\lambda = 10^{-4}$), 30 epochs with cosine annealing, and class-weighted cross-entropy.

To disentangle metric scale from segmentation quality, we apply controlled interventions to the CNN3 depth probe. Scale correction multiplies relative depth

by a metric factor derived from Eq. 1, an operation that is only meaningful for depth-based input. For the RGB probe, there is no natural mechanism to inject scale information, and the MLP features do not include raw depth values. For scale, we derive a deterministic diagnostic factor from the ground-truth size label via Eq. 1. From $A = fS/Z$, the per-frame ratio Z/f is recovered as S_p/A_i , with S_p the labeled polyp diameter and A_i the apparent diameter in frame i . Since the estimated depth $\bar{d}_i$ shares an arbitrary scale with Z/f , a frame-specific correction factor $c_i = (S_p/A_i)/\bar{d}_i$ aligns the two. *Oracle-frame* applies c_i per frame; *Oracle-polyp* averages c_i over a polyp clip; *Global* averages c_i across the training fold. We also evaluate zero-shot metric depth from MetricCol [22]. For masks, we replace ground-truth bounding boxes with PolypPVT [11] predictions and evaluate $\pm 20\%$ bounding-box perturbations.

4 Results and Discussion

4.1 Data–Model Bias

Table 1 (top) summarizes performance across probes. Despite spanning three input modalities, all models fall within a narrow band near Macro-F1 ≈ 0.75 (ResNet18: 0.751; CNN3: 0.725; MLP: 0.752). Increasing capacity yields no reliable gain. ViT-B on masked RGB reaches 0.778 but shows high cross-fold variance in > 5 mm recall (0.664 ± 0.275), consistent with overfitting at this sample size. Similarly, invariant training (V-REx [19]) does not change the plateau ($+0.002$), reflecting directionally similar confounds across both datasets [1]

We next audit the predictive signals behind this saturation. Mutual information (MI) on the MLP features indicates that the apparent size of the polyp in the image (`apparent_area_frac`, the ratio of bounding-box area to image area, MI= 0.195) is more informative than photometric features of background illumination (MI= 0.043). Removing geometric features (bounding-box dimensions, bounding-box area, and compactness) reduces Macro-F1 by 5.6 pp ($0.752 \rightarrow 0.694$), confirming that models rely on apparent size, which conflates true scale S with distance Z . Together, these diagnostics suggest that the observed plateau is driven by missing metric information in the input, rather than insufficient capacity.

Implications. Increasing training data diversity is the most direct way to reduce procedure-linked cues. However, collecting large-scale, size-annotated multi-center data is costly, and our effective sample size of 232 polyps remains modest for deep learning models. The cross-fold instability of higher-capacity controls is consistent with sensitivity to procedure-specific signals and a sign of overfitting.

While synthetic data generation can provide the required variability in polyp size, camera pose, intrinsics, and lighting, models trained solely on synthetic data often suffer from sim-to-real domain gaps that limit their transfer to clinical environments [6]. Nevertheless, even with expanded data, the geometric ambiguity in Eq. 1 remains: without a metric reference, more data can likely shift the plateau

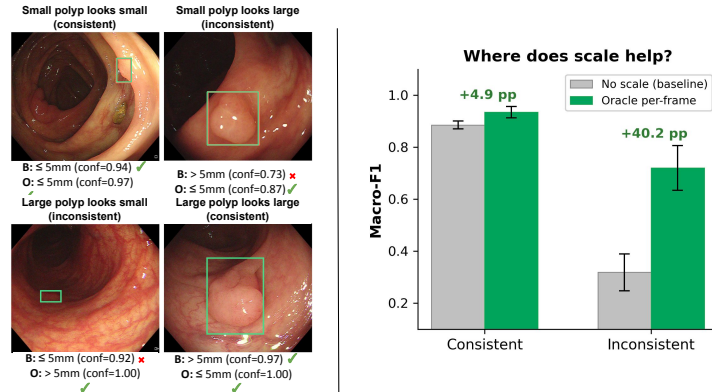


Fig. 1: Effect of oracle per-frame scale on actual and apparent size consistent and inconsistent frames. (Left) Examples where apparent size matches (consistent) or contradicts (inconsistent) the true size, with predictions from the baseline (B) and oracle (O) models. (Right) Macro-F1 gains from oracle scale, concentrated in shortcut-inconsistent frames ($+40.2 \pm 15.4$ pp vs. $+4.9 \pm 2.2$ pp across 5 folds; paired t -test $p < 0.01$)

but cannot remove the identifiability limit. Next, we test whether providing a ground-truth metric scale improves performance.

4.2 Metric Scale

We conduct scale interventions on the CNN3 depth probe because scale correction operates on the depth map directly: the oracle multiplies relative depth by a metric factor, which is only meaningful for depth-based input. For the RGB probe, scale information has no natural injection point, and for the MLP, the handcrafted features do not include raw depth values. Table 1 (middle) reports the effect of progressively resolving scale ambiguity. Oracle per-frame scale yields a $+16.1$ pp gain ($0.725 \rightarrow 0.886$). A per-polyp oracle scale provides a smaller $+6.7$ pp gain ($0.725 \rightarrow 0.792$), indicating that Z varies substantially even within a single polyp clip, likely due to endoscope motion.

We also evaluate algorithmic alternatives. Replacing relative depth with metric depth, MetricCol [22], produces no measurable change ($0.725 \rightarrow 0.727$, within fold variance). Global calibration, which fits a single scale factor on the training set, slightly degrades performance ($0.725 \rightarrow 0.716$), indicating that scale variation is not captured by a population-level constant.

To localize where scale helps, we partition test frames into two groups based on whether the apparent size agrees with the true label (Fig. 1). In the *consistent* group, polyps that appear large in the image are genuinely large, and polyps that appear small are genuinely small. In the *inconsistent* group, apparent size contradicts the true physical size (e.g., a large polyp viewed from far away appears

small). We split groups at the per-fold median of bounding-box area relative to image area. The oracle gain concentrates in inconsistent frames ($+40.2 \pm 15.4$ pp, $0.319 \rightarrow 0.720$) and is small in consistent frames ($+4.9 \pm 2.2$ pp, $0.886 \rightarrow 0.935$); a paired t -test over the 5 fold-level gain differences yields $p < 0.01$, confirming that oracle scale selectively helps when apparent size contradicts the true label. This pattern is consistent with metric scale resolving the S/Z ambiguity when apparent size contradicts the true label.

Implications. In routine colonoscopy, camera-to-polyp distance varies continuously with scope advancement, angulation, and insufflation, yet few datasets report camera intrinsics or explicit metric references [28,21,23,2]. Recording intrinsics would benefit not only reconstruction methods but also physics-informed sizing models. However, knowing the focal length alone does not resolve ambiguity if the distance remains unknown. Reliable progress, therefore, likely requires either reference-based anchors (e.g., instrument tips) or calibrated sensing that provides a stable metric scale at the frame level.

Beyond deployment-time sensing, reference-based measurements during data collection could address a more fundamental issue: size labels in current datasets are endoscopist visual estimates and can be subject to rounding near clinical thresholds. Training and evaluation on tool-verified labels would reduce label noise at the decision boundary where sizing errors are most consequential.

Off-the-shelf metric depth models provide limited gains. Existing endoscopic metric depth methods [22,20] acquire scale from synthetic supervision and adapt to clinical data through self-supervised fine-tuning or data augmentation. In our evaluation, MetricCol [22], which follows this paradigm, yields no measurable improvement over relative depth ($0.725 \rightarrow 0.727$), suggesting that synthetically derived scale does not transfer reliably across the heterogeneous intrinsics and acquisition conditions present in multi-center data. Test-time adaptation may improve robustness to low-level shifts such as lighting and tissue appearance, but examination behavior introduces a semantic-level confound that is unlikely to be resolved without explicit metric scale supervision. Accurate metric depth estimation for endoscopy could substantially reduce this bottleneck, but reliable methods that generalize across clinical settings remain limited.

The oracle results assume accurate foreground isolation using a ground-truth mask. We evaluate this dependency next.

4.3 Mask Quality

Table 1 (bottom) evaluates the interaction between scale and segmentation quality. PolypPVT [11] reports mDice > 0.85 on curated benchmarks but achieves IoU ≈ 0.3 on Real-Colon, where full-procedure artifacts are prevalent. Replacing ground-truth bounding boxes with these predicted masks reduces CNN3 performance by 15.3 pp ($0.725 \rightarrow 0.572$) (Fig. 2). With oracle per-frame scale under-predicted masks, performance recovers only to 0.736, which is close to the unscaled baseline (0.725). This indicates that segmentation errors can absorb most of the oracle-scale gain.

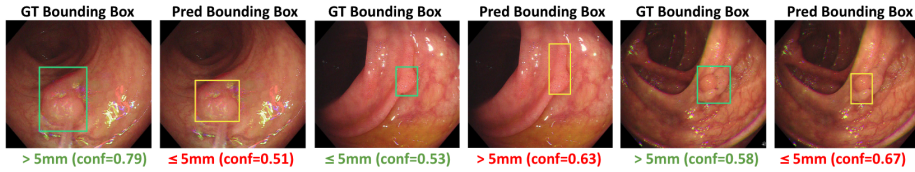


Fig. 2: Effect of mask quality on sizing predictions. Each image pair shows the same polyp under ground-truth (GT) and predicted bounding boxes [11], paired with model predictions from CNN3. Predicted masks introduce localization errors that change the apparent size seen by the model, directly altering the classification outcome (e.g., misclassifying a correctly classified polyp under GT bbox).

Bounding-box perturbations ($\pm 20\%$) further show higher sensitivity for depth-based sizing than for the RGB probe (CNN3 + Oracle: -3.4 pp; ResNet18: within variance). Overall, these results indicate that metric scale and segmentation quality constitute two independent bottlenecks for monocular polyp sizing.

Implications. PolypPVT reports strong overlap scores on curated benchmarks, but Real-Colon contains full-procedure artifacts that can shift mask quality. Small polyps may be poorly localized under specular highlights, larger lesions may be partially segmented due to complex texture, and instrument interaction can introduce partial occlusion. Such conditions are underrepresented in curated training data and affect the foreground region used by sizing pipelines. As a result, in-domain overlap metrics alone may not predict robustness for downstream size estimation under realistic clinical variation. Including sizing accuracy under predicted or perturbed masks can serve as a task-conditioned robustness check.

4.4 Photometry Control

Finally, we test whether near-field illumination falloff can serve as a proxy for scale [22,24]. Under ideal conditions, background luminance P would follow an inverse-square law ($P \approx k/Z^2$), allowing distance to be estimated from brightness. In clinical endoscopy, Auto-Exposure Control continuously adjusts sensor gain and illumination intensity to maintain stable image brightness regardless of distance, which breaks this relationship and limits the usefulness of photometric features as a distance surrogate. Consistent with this, photometry-only correction (A/P) recovers only 17% of the oracle-scale headroom (Macro-F1 0.760).

5 Conclusion

We present a diagnostic audit for monocular polyp size classification (≤ 5 mm vs. > 5 mm) across two multi-center datasets with polyp-stratified cross-validation. Across architectures and input modalities, performance concentrates near Macro-F1 ≈ 0.75 , and diagnostics indicate reliance on stable cues correlated with examination behavior. When apparent size in the image agrees with true physical

size, current models perform well; however, performance degrades substantially when examination behavior causes apparent size to contradict the true label, and oracle scale specifically resolves these cases (+40.1 pp). To our knowledge, this is the first controlled study showing why better monocular depth and segmentation do not necessarily improve polyp sizing. We identify three actionable directions for future work. (i) Metric Scale: relative depth, global scale, and colonoscopy-trained metric depth remain below the oracle frame-level ceiling, motivating endoscopy-specific scale estimation or calibration anchors. (ii) Segmentation Robustness: despite high Dice/IoU, segmentation degrades under distribution shift and impacts sizing, motivating evaluation of downstream robustness and uncertainty-aware refinement. (iii) Data Limitations: current datasets rely on visual estimates and lack ≥ 10 mm polyps. Larger tool-verified datasets or realistic simulations are needed to support scale recovery, depth, and segmentation under tool occlusions and glare. Our reusable evaluation checks (oracle scale ladder, shortcut partitions, and mask substitution) provide a basis for measuring this progress.

Acknowledgments. This work is supported by a National Institute of Health (NIH) project #1R21EB035832 “Next-gen 3D Modeling of Endoscopy Videos”.

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

References

1. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)
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.: Endomap-per dataset of complete calibrated endoscopy procedures. *Scientific Data* **10**(1), 671 (2023)
3. Bhat, S.F., Birkel, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)
4. Biffi, C., Antonelli, G., Bernhofer, S., Hassan, C., Hirata, D., Iwatate, M., Maieron, A., Salvagnini, P., Cherubini, A.: Real-colon: A dataset for developing real-world ai applications in colonoscopy. *Scientific Data* **11**(1), 539 (2024)
5. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. arXiv preprint arXiv:2410.02073 (2024)
6. Bonilla, S., Zhang, S., Psychogyios, D., Stoyanov, D., Vasconcelos, F., Bano, S.: Gaussian pancakes: geometrically-regularized 3d gaussian splatting for realistic endoscopic reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 274–283. Springer (2024)
7. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. *Nature Communications* **11**(1), 3673 (2020)
8. Chaptini, L., Chaaya, A., Depalma, F., Hunter, K., Peikin, S., Laine, L.: Variation in polyp size estimation among endoscopists and impact on surveillance intervals. *Gastrointestinal endoscopy* **80**(4), 652–659 (2014)

9. Cheloff, A.Z., Kim, L., Pochapin, M.B., Shaukat, A., Popov, V.: Accuracy of visual estimation for measuring colonic polyp size: a systematic review and meta-analysis. *Official journal of the American College of Gastroenterology| ACG* pp. 10–14309 (2022)
10. DeGrave, A.J., Janizek, J.D., Lee, S.I.: Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence* **3**(7), 610–619 (2021)
11. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932* (2021)
12. Du, S., Zhang, Q., Zhang, Z., Cai, C., Li, X., Qian, D.: Polyp size estimation by generalizing metric depth estimation and monocular 3d reconstruction. In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. pp. 1–5. IEEE (2024)
13. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**(11), 665–673 (2020)
14. Gupta, S., Lieberman, D., Anderson, J.C., Burke, C.A., Dominitz, J.A., Kaltenbach, T., Robertson, D.J., Shaukat, A., Syngal, S., Rex, D.K.: Recommendations for follow-up after colonoscopy and polypectomy: a consensus update by the us multi-society task force on colorectal cancer. *Official journal of the American College of Gastroenterology| ACG* **115**(3), 415–434 (2020)
15. Hewett, D.G.: Measurement of polyp size at colonoscopy: addressing human and technology bias. *Dig Endosc* **34**(7), 1478–80 (2022)
16. Iranzo, R., Batlle, V.M., Tardós, J.D., Montiel, J.M.: Endometric: Near-light monocular metric scale estimation in endoscopy. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 180–190. Springer (2025)
17. Itoh, H., Oda, M., Jiang, K., Mori, Y., Misawa, M., Kudo, S.E., Imai, K., Ito, S., Hotta, K., Mori, K.: Binary polyp-size classification based on deep-learned spatial information. *International Journal of Computer Assisted Radiology and Surgery* **16**(10), 1817–1828 (2021)
18. Itoh, H., Roth, H.R., Lu, L., Oda, M., Misawa, M., Mori, Y., Kudo, S.e., Mori, K.: Towards automated colonoscopy diagnosis: binary polyp size estimation via unsupervised depth learning. In: *International conference on medical image computing and computer-assisted intervention*. pp. 611–619. Springer (2018)
19. Krueger, D., Caballero, E., Jacobsen, J.H., Zhang, A., Binas, J., Zhang, D., Priol, R.L., Courville, A.: Out-of-distribution generalization via risk extrapolation (rex). In: Meila, M., Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 139, pp. 5815–5826. PMLR (18–24 Jul 2021), <https://proceedings.mlr.press/v139/krueger21a.html>
20. Li, H., Lu, D., Wang, J., Webster III, R.J., Oguz, I.: Endostreamdepth: Temporally consistent monocular depth estimation for endoscopic video streams. *arXiv preprint arXiv:2512.18159* (2025)
21. Liu, R., Wang, L., Mingming, Z., Zhang, J., HAOYU, Z., Liu, X., Cheng, X., Chan, S., Sheng, D., Yan, Y., et al.: Polypsense3d: A multi-source benchmark dataset for depth-aware polyp size measurement in endoscopy. In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2025)
22. Liu, Y., Yu, D., Liu, L., Xiao, Y., He, T.: Metriccol: Metric depth and pose estimation in colonoscopy via geometric consistency and domain adaptation. In: *2025 10th International Conference on Image, Vision and Computing (ICIVC)*. pp. 364–369. IEEE (2025)

23. Misawa, M., Kudo, S.e., Mori, Y., Hotta, K., Ohtsuka, K., Matsuda, T., Saito, S., Kudo, T., Baba, T., Ishida, F., et al.: Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointestinal endoscopy* **93**(4), 960–967 (2021)
24. Paruchuri, A., Ehrenstein, S., Wang, S., Fried, I., Pizer, S.M., Niethammer, M., Sengupta, R.: Leveraging near-field lighting for monocular depth estimation from endoscopy videos. In: *European Conference on Computer Vision*. pp. 473–491. Springer (2024)
25. Roffo, G., Biffi, C., Salvagnini, P., Cherubini, A.: Feature selection gates with gradient routing for endoscopic image computing. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 339–349. Springer (2024)
26. Ruano, J., Bravo, D., Giraldo, D., Gómez, M., González, F.A., Manzanera, A., Romero, E.: Estimating polyp size from a single colonoscopy image using a shape-from-shading model. In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. pp. 1–5. IEEE (2024)
27. Siegel, R.L., Giaquinto, A.N., Jemal, A.: Cancer statistics, 2024. *CA: a cancer journal for clinicians* **74**(1), 12–49 (2024)
28. Song, Y., Du, S., Wang, R., Liu, F., Lin, X., Chen, J., Li, Z., Li, Z., Yang, L., Zhang, Z., et al.: Polyp-size: a precise endoscopic dataset for ai-driven polyp sizing. *Scientific Data* **12**(1), 918 (2025)
29. Sudarevic, B., Sodmann, P., Kafetzis, I., Troya, J., Lux, T.J., Saßmannshausen, Z., Herlod, K., Schmidt, S.A., Brand, M., Schöttker, K., et al.: Artificial intelligence-based polyp size measurement in gastrointestinal endoscopy using the auxiliary waterjet as a reference. *Endoscopy* **55**(09), 871–876 (2023)