Probabilistic 3D Tissue Motion Forecasting from Stereo Surgical Video

Bhaskar Nikhil Sunkara Pallavi Sharma Kesav Nagendra
University of Wisconsin-Madison
Madison, WI
{bsunkara3, pallavisharm, nagendra2}@wisc.edu

1. Problem Definition

Modern surgical vision estimates depth for the current frame, but does not anticipate how the scene will change a split second later, so overlays and trackers drift when tissue moves. We pose a focused computer-vision task: given a short window of recent rectified stereo frames and calibration, predict the next left-view depth/disparity map at a short horizon ($\delta = 200$ ms) turning "see now" into "see next." Our approach frames forecasting as a compact, real-time problem: stereo input, left-view future depth/disparity as output, a few recent frames as context, and a small prediction gap for actionability. We use two simple signals from the video: how things look and how they move. From the last few frames, the model learns small shifts and deformations to guess the 3D surface a fraction of a second ahead. Because the inputs are only images and known calibration (no external sensors or pre-op models), the setup is reproducible and model-agnostic, enabling fair comparisons across architectures. If successful, the prediction stabilizes image-fusion roadmaps and keeps trackers locked during endoscopic navigation, supporting catheter and stent alignment where a steady near-future 3D view reduces manual corrections and improves guidance reliability.

2. Motivation

Recent lightweight stereo methods improve depth-now in endoscopy by sharpening boundaries, sustaining real-time throughput at HD, and showing encouraging cross-dataset signs (for example, SERV-CT). Yet they still answer "where is the surface now?" rather than "where will it be a split second later?" when tissue moves with breathing, heartbeat, or tool motion, per-frame depth forces overlays and trackers to chase the scene and introduces small but important lags. This motivates a short-horizon forecasting step. This direction is motivated by three observations:

(i)Stereo endoscopy is fast and informative, but current methods still optimize instantaneous disparity and struggle with boundary ambiguity, heavy runtime at 1024×1280, and generalization. A naive forecast that only copies last or warps last inherits the same issues [4].

(ii) Monocular self-supervision uses longer temporal windows to handle occlusions and small pose changes, showing that motion and visibility matter, but the goal is still current-frame depth, not next-frame depth [7].

(iii) Current benchmarks such as SCARED and SERV-CT focus on per-frame correspondence and do not test whether a lightweight temporal head reduces short-horizon drift [5][1]. A small, reproducible future-depth benchmark would ask whether see-now models can become see-next and keep overlays and trackers steady without extra hardware or exotic supervision [2].

3. Tentative Approach and Evaluation Plans

3.1. Approach

Train a small cost-volume stereo backbone (e.g., RAFT-Stereo-small) on SCARED for disparity at time t, then attach a conditional diffusion forecast head that, from the last three rectified stereo pairs and calibration, samples K plausible disparities at (t+200) ms. For robustness, we add LoRA adapters on late layers and adapt them lightly at test time under lighting/smoke shifts with a retention guard to prevent drift.

3.2. Evaluation

We adopt a citation-backed metric set. Future-disparity EPE at t+200 ms on SCARED is primary average pixel error of the predicted future disparity, directly testing whether we anticipated the short-horizon 3D shift [3]. For depth accuracy, we report AbsRel (mean absolute error relative to true depth) and RMSE (root-mean-squared error), the standard pair used in recent endoscopic depth studies [2]. Because our forecasts are probabilistic, we include ECE (expected calibration error) to check whether predicted confidence aligns with realized errors (only when diffusion is enabled) [6].

References

[1] Max Allan, Jonathan Mcleod, Congcong Wang, Jean Claude Rosenthal, Zhenglei Hu, Niklas Gard, Peter Eisert, Ke Xue Fu, Trevor Zeffiro, Wenyao Xia, Zhanshi Zhu, Huoling Luo, Fucang Jia, Xiran Zhang, Xiaohong Li, Lalith Sharan, Tom Kurmann, Sebastian Schmid, Raphael Sznitman, Dimitris Psychogyios, Mahdi Azizian, Danail Stoyanov, Lena Maier-Hein, and Stefanie Speidel. Stereo correspondence and reconstruction of endoscopic data challenge, 2021. 1

[2] Beilei Cui, Mobarakol Islam, Long Bai, An Wang, and Hongliang Ren. Endodac: Efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera, 2024. 1

[3] Rema Daher, Francisco Vasconcelos, and Danail Stoyanov. A temporal learning approach to inpainting endoscopic specularities and its effect on image correspondence. *Medical Image Analysis*, 90:102994, 2023. 1

[4] Yang Ding, Can Han, Sijia Du, Yaqi Wang, and Dahong Qian. Lightendostereo: A real-time lightweight stereo matching method for endoscopy images, 2025. 1

[5] P.J. Eddie Edwards, Dimitris Psychogyios, Stefanie Speidel, Lena Maier-Hein, and Danail Stoyanov. Serv-ct: A disparity dataset from cone-beam ct for validation of endoscopic 3d reconstruction. *Medical Image Analysis*, 76:102302, 2022. 1

[6] Alexander Kurz, Katja Hauser, Hendrik Mehrtens, Eva Krieghoff-Henning, Achim Hekler, Jakob Kather, Stefan Frohling, Christof Kalle, and Titus Brinker. Uncertainty estimation in medical image classification: Systematic review. *JMIR Medical Informatics*, 10:e36427, 2022. 1

[7] Xiaowei Shi, Beilei Cui, Matthew J. Clarkson, and Mobarakol Islam. Long-term reprojection loss for self-supervised monocular depth estimation in endoscopic surgery. *Artificial Intelligence Surgery*, 4(3), 2024. 1