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. Introduction

Depth estimation in minimally invasive surgery underpins a wide range of image-guided tasks, including 3D navigation, soft-tissue tracking, and augmented-reality (AR) overlay. Although recent stereo and monocular methods achieve strong spatial accuracy, surgical scenes evolve rapidly due to endoscope motion, respiration, and tissue deformation. As a result, temporal stability becomes as important as spatial precision.

Recent work shows that surgical vision models often suffer from *temporal inconsistency*, producing depth fields whose ordering or surface geometry flicker across frames [1]. Even when per-frame depth is accurate, this instability distracts the surgeon and undermines trust. In parallel, studies on temporal modeling report that frame-based networks can lag behind the video stream, leading to predictions that trail the true scene motion [6]. Such delays manifest as visible drift in geometric overlays, making AR guidance unreliable.

These issues are amplified in AR systems: continuous soft-tissue motion means that even small misalignments grow over time, and registration methods report failure during rapid camera motion or deformation [8]. This highlights a core limitation of conventional depth estimation, its outputs remain tied to the timestamp of input frame, not state of the scene when depth is actually displayed.

Streaming perception studies formalize this mismatch. Because every vision pipeline incurs processing delay, the output computed at time t corresponds to an observation made at $t - \Delta t$. The framework argues that real-time systems must be *predictive*, estimating state that will exist when output is used rather than replicating a past scene [7].

Motivated by these observations, we investigate future depth forecasting for latency compensation in surgical video. Instead of predicting disparity for the current frame, we target the depth that will exist at display time $t+\eta$, where η represents system latency. Our goal is to stabilize geometry, reduce temporal drift, and ensure that the depth seen by the surgeon matches the surgical scene at the moment it is viewed.

2. Related Work

vy This section reviews prior work that is most relevant to our goal of forecasting depth for latency compensation. We first summarize recent endoscopic depth estimation methods, including approaches that improve robustness and temporal stability under surgical motion and deformation. We then summarize latency-aware work in streaming perception, which evaluates predictions under processing delay and motivates outputs that remain correct when they are displayed. Together, these works show why current-frame depth can drift in real-time systems and why forecasting is worth studying.

2.1. Endoscopic Depth Estimation Under Surgical Dynamics

Recent endoscopic depth methods often rely on self-supervision or weak supervision because dense ground-truth depth is difficult to obtain in surgery. M3Depth uses stereo pairs to learn depth with 3D geometric consistency during training, while keeping efficient monocular inference at test time [5]. More recent work adapts large depth priors or foundation models to endoscopy. EndoDAC adapts a foundation depth model for self-supervised endoscopic depth across different cameras [4], and Surgical-DINO adapts DINOv2 with parameter-efficient updates for surgical depth estimation [3]. Other approaches add explicit geometric cues to recover scale more reliably in monocular endoscopy, such as using instrument geometry for scale-aware depth [10]. Finally, some work directly targets temporal stability. Budd and Vercauteren use temporal consistency self-supervision to reduce flicker when transferring relative monocular depth to surgical video [1]. Overall, these methods improve depth quality and stability, but their outputs remain aligned to the input frame time rather than the display time.

2.2. Latency-Aware and Streaming Perception

Streaming perception work formalizes a key deployment issue: perception is delayed, and evaluation should reflect the time when the output becomes available. Li et al. argue that the world changes while the model is computing, and propose metrics that account for this mismatch [7]. Following this idea, latency-aware semantic segmentation explicitly trains models to predict the segmentation that will match the near-future frame at completion time [2]. In detection, StreamYOLO (CVPR) argues that predicting future object states is central for streaming settings and proposes modules and losses to support this goal [11]. The ASAP benchmark (CVPR) further shows that model rankings can change under constrained-computation streaming evaluation, encouraging designs that consider both accuracy and latency [9]. These works support a consistent message: when delay matters, it is often better to be slightly predictive than precisely "correct" about the past.

2.3. Summary and gap

Across both themes, most endoscopic depth work focuses on improving depth for the current frame (sometimes with temporal regularization), while streaming perception work shows that real-time deployment fundamentally changes what "correct" means under latency. Our work connects these two lines: we forecast **future depth** at $t + \eta$ so that the depth shown to the surgeon is aligned with the scene at display time, rather than the scene at capture time. In this sense, prior methods remain current-frame aligned, whereas we explicitly predict the immediate future to compensate for inevitable system delay.

3. Methods

Our goal is to forecast the disparity map that will exist at display time $t + \eta$ given the most recent observed disparities. We describe the full pipeline, including data preparation, model design, and forecasting strategy. We also document several alternative architectures we investigated: RAFT-Stereo, SwinUNETR, and diffusion-based predictors, which informed the final approach but did not meet the computational or stability requirements of the task.

3.1. Data Preparation and Disparity Generation

Each SCARED keyframe provides rectified stereo video, calibration data, and per-frame 3D point clouds. We initially experimented with RAFT-Stereo for disparity generation, but full-resolution inference was too slow and fine-tuning yielded limited benefit.

We therefore used the official SCARED toolkit, which converts depth Z to disparity using camera intrinsics,

$$d = \frac{f_x B}{Z},$$

producing accurate and temporally stable disparity without training a stereo network. These maps form the basis for all forecasting experiments.

3.2. Backbone Model: UNet for Single-Horizon Forecasting

We began by establishing a deterministic forecasting baseline. A lightweight UNet was trained to predict d_{t+5} from the three most recent disparity frames $\{d_{t-2}, d_{t-1}, d_t\}$ using a masked L1 loss over valid pixels. We evaluated several horizon lengths and found that $t \to t+1$ provided almost no temporal change (and thus weak supervisory signal), while $t \to t + 10$ introduced motion too large for stable learning. The $t \to t + 5$ task offered a balanced degree of motion and served as a reliable supervisory target.

Across datasets, the UNet converged consistently, produced sharp and temporally stable disparity fields, and ultimately became the backbone for our multi-horizon forecasting pipeline.

3.3. Explored Alternatives: Swin-UNETR and Diffusion Heads

SwinUNETR. We evaluated SwinUNETR on the $t \to t+5$ task due to its strong results in medical segmentation. In practice, the model overfit rapidly: validation loss stopped improving early, and predictions collapsed to low-contrast disparity fields with weak geometry, even under heavier regularization. SwinUNETR did not generalize to full-resolution surgical frames, making it less suitable than UNet for high-resolution temporal forecasting.

Diffusion-based prediction. We also tested conditional diffusion models for future disparity synthesis, using cropped disparity patches and conditioning the diffusion head on either a UNet or SwinUNETR prior. Training remained unstable, noise-prediction loss decreased slowly, and generated samples were noisy or lacked coherent structure. In addition, the iterative sampling required by diffusion made inference far too slow for real-time surgical use.

These results motivated a lightweight deterministic alternative: a forecasting head that builds on the stable UNet prior while efficiently supporting multiple latency horizons.

3.4. Multi-Horizon Forecasting with HorizonHead

To compensate for unknown system latency η, we require a model that predicts disparity at several possible future horizons. We designed a compact convolutional module, *HorizonHead* which refines the UNet's $t + 5$ prior to arbitrary horizons $h \in \{3, 5, 7, 9\}$. The input is the concatenation of three signals:

$$x = \big[\, d_t,\ \mu_5,\ \text{enc}(h) \,\big],$$

where μ_5 is the UNet prediction for $t + 5$, and $\text{enc}(h)$ is a normalized scalar channel encoding the target horizon. The head is trained jointly across all horizons using masked L1 loss.

This design has two advantages: (1) it leverages a strong geometric prior μ_5 from the UNet, and (2) a single model handles multiple latency values without retraining.

3.5. Probabilistic Forecasting via Monte Carlo Noise

The HorizonHead is trained deterministically, but we introduce a lightweight mechanism for estimating predictive uncertainty at inference time. During testing, we inject small Gaussian perturbations into the UNet prior μ_5 and evaluate the forecasting head multiple times:

$$d_{t+h}^{(k)} = H\big(d_t,\ \mu_5 + \epsilon_k,\ \mathrm{enc}(h)\big), \qquad \epsilon_k \sim \mathcal{N}(0, \sigma^2).$$

Using a small number of samples (K = 3-5) consistent with our inference pipeline, we compute

$$\hat{d}_{t+h} = \frac{1}{K}\sum_{k=1}^{K} d_{t+h}^{(k)}, \qquad \hat{\sigma}^2 = \mathrm{Var}\Big(d_{t+h}^{(k)}\Big).$$

The mean $\hat{d}_{t+h}$ serves as the final forecast, while the variance $\hat{\sigma}^2$ highlights spatial regions where the model is uncertain. In practice, uncertainty concentrates near occlusion boundaries, fast deformations, and specular regions. This information is valuable for AR overlays and tracking modules, which can down-weight or stabilize rendered content in regions where future geometry is unreliable.

3.6. Latency-Aware Streaming Evaluation

Following the Streaming Perception formulation, we simulate real-time use by measuring the horizon implied by system latency:

$$h^\star = \mathrm{round}\left(\frac{\eta}{\Delta t}\right),$$

where η is display delay and Δt is the frame interval. During streaming replay, the model receives disparities up to time t and predicts $d_{t+h^\star}$, which is compared against the true future frame. This provides a faithful estimate of real-world performance under latency.

4. Experiments and Results

We trained all models on three SCARED datasets using keyframes 1–5, and evaluated streaming performance on a held-out dataset using three keyframes. This setup exposes the forecasting model to a broad range of tissue motion while ensuring that all reported results come from unseen surgical scenes.

4.1. Model Comparison

Table 1 summarizes the behaviour of all forecasting architectures. The UNet backbone trained for the $t \to t+5$ task converged smoothly (train loss $2.72 \to 1.39$) and reached a validation MAE of 1.21 pixels. Predictions were sharp and stable across datasets, making UNet a strong geometric prior for future forecasting.

SwinUNETR trained without instability, but consistently overfit: although the validation loss improved from 27.97 to 2.56, the model collapsed to low-contrast disparity fields and failed to generalize. This mirrors prior observations that transformer-based architectures struggle when resolution is high and temporal context is limited.

Diffusion heads, whether paired with UNet or SwinUNETR, reduced their noise-prediction loss only slightly ($0.74 \to 0.69$) and never produced coherent samples. Sampled disparities were noisy or blank, and multi-step inference was far slower than real-time. As a result, diffusion was unsuitable for surgical forecasting.

In contrast, the HorizonHead delivered consistent improvements across latency horizons $h \in \{3, 5, 7, 9\}$. Each horizon-specific MAE increased smoothly with prediction distance (1.03, 1.20, 1.42, 1.60 px), indicating stable refinement of the UNet prior. This combination yielded the best overall accuracy and the only model capable of supporting multiple latency settings.

4.2. Streaming Evaluation at Inferred Latency

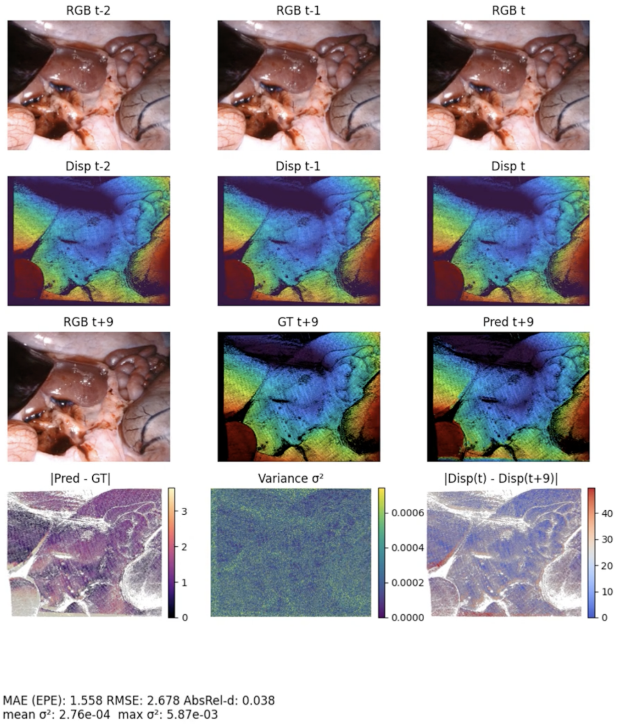
To approximate real surgical delay, we run a streaming replay experiment in which the model predicts disparity at time $t + H^\star$, where $H^\star$ corresponds to the effective display latency. We evaluate forecasting accuracy using MAE (our primary pixelwise end-point error), RMSE (which penalizes large deviations), and AbsRel-d (error relative to local disparity magnitude). We also compute per-pixel variance $\hat{\sigma}^2$ from Monte Carlo sampling to quantify predictive uncertainty. Figure 1 shows a representative result from **dataset 4, keyframe 1, with $H^\star = 9$.**

Forecasting accuracy. As illustrated in Figure 1, the predicted future disparity aligns well with the ground-truth frame at $t+9$: the major tissue surfaces, folds, and overall geometry match closely. The corresponding $|\mathrm{Pred} - \mathrm{GT}|$ map is mostly dark except near steep boundaries, indicating low absolute error. Quantitatively, MAE (EPE) across sampled frames ranged from 0.6 to 3.4 pixels (mean ≈ 1.5 px). RMSE varied between 1.2 and 5.0 pixels (mean ≈ 2.6 px), and AbsRel-d stayed between 0.017 and 0.082. Most frames remained under $\sim$ 1.7 px MAE, with only rapid-motion frames exceeding 2 px.

Improvement over a latency baseline. The right-most panel in Figure 1 visualizes $|d_t - d_{t+9}|$, showing how much the scene truly changes over a 9-frame latency. Under typical soft-tissue and camera motion, this difference often reaches 40–50 disparity pixels. In contrast, our forecast error (middle-bottom panel) stays within 3–5 px over most of the surface, reducing the effective drift caused by latency by an order of magnitude.

Model	Training Loss Trend	Best Val MAE (px)	Observations
UNet (t→t+5)	2.72 → 1.39	**1.21**	Stable; sharp predictions
SwinUNETR	27.97 → 2.56	2.56	Overfit at full res; blurry outputs
UNet + Diffusion Head	0.74 → 0.69 (noise-MSE)	—	Unstable samples; unusable predictions
UNet + HorizonHead (3,5,7,9)	1.12 → 0.94 (per-horizon batches)	**1.03**, **1.20**, **1.42**, **1.60**	Accurate across horizons; best overall

Table 1. Comparison of forecasting architectures across all experiments.



```
MAE (EPE): 1.558 RMSE: 2.678 AbsRel-d: 0.038
mean σ²: 2.76e-04  max σ²: 5.87e-03
```

Figure 1. Streaming evaluation example from SCARED dataset 4, keyframe 1, horizon $H^\star = 9$. Top rows show past RGB and disparity inputs; middle rows show ground-truth and predicted disparity at t+9; bottom row visualizes absolute error, variance $\hat{\sigma}^2$, and the latency baseline $|d_t - d_{t+9}|$.

Failure modes. Higher errors occur in frames with abrupt endoscope motion or sharp deformation. In such cases, the error concentrates around depth discontinuities, specular highlights, and occluded regions—the same areas that challenge conventional stereo algorithms. Even in these settings, the interior tissue surfaces remain reasonably stable.

Uncertainty behaviour. The variance map $\hat{\sigma}^2$ (center-bottom panel) is low and uniform across the majority of the tissue surface (mean $\sim$ 2.7–3.0 $\times 10^{-4}$). Higher variance appears exactly where prediction error is largest—around edges, occlusions, and specularities. This alignment suggests that the Monte Carlo mechanism provides meaningful confidence estimates that could inform downstream AR or tracking modules.

5. Conclusion

In this project, we explored future depth forecasting to reduce latency-induced drift in surgical depth pipelines. Instead of estimating depth for the current frame, we predict the depth expected at display time so the geometry is better aligned with what the surgeon sees. We built a lightweight pipeline on stereo surgical video. A base model produces a depth prior, and a horizon-conditioned head adapts it to multiple future time offsets. We also estimate uncertainty to highlight regions where the forecast may be less reliable. In our experiments, forecasting reduced the mismatch caused by delay in many sequences. Errors were higher during abrupt camera motion, strong tissue deformation, and near depth boundaries, and uncertainty tended to increase in the same regions. These results suggest that forecasting can help with latency compensation for endoscopic AR, but it still needs careful handling of challenging cases.

For future work, we plan to test on more procedures and latency settings, and to integrate the method into an AR pipeline to measure overlay drift directly. We also want to improve behavior under occlusions and rapid motion, and use uncertainty to limit updates in unreliable areas.

6. Contributions

Kesav Nagendra contributed RAFT-Stereo preprocessing and disparity generation. He also implemented the UNet-based disparity model and ran its evaluation and also did the dataset curation.

Nikhil integrated the SCARED toolkit for disparity generation. He implemented and evaluated the SwinUNETR-based disparity model, designed the latency-aware streaming evaluation protocol, and assembled the final pipeline and visualizations.

Pallavi implemented diffusion-based forecasting experiments. She also developed the multi-horizon refinement head and the uncertainty modeling component.

References

[1] Charlie Budd and Tom Vercauteren. *Transferring Relative Monocular Depth to Surgical Vision with Temporal Consistency*, page 692–702. Springer Nature Switzerland, 2024. 1

[2] Evann Courdier and Francois Fleuret. Real-time segmentation networks should be latency aware, 2022. 2

[3] Beilei Cui, Mobarak Hoque, and Long Bai. Surgical-dino: adapter learning of foundation models for depth estimation in endoscopic surgery. *International Journal of Computer Assisted Radiology and Surgery*, 19, 2024. 1

[4] 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

[5] Baoru Huang, Jian-Qing Zheng, Anh Nguyen, Chi Xu, Ioannis Gkouzionis, Kunal Vyas, David Tuch, Stamatia Giannarou, and Daniel S. Elson. Self-supervised depth estimation in laparoscopic image using 3d geometric consistency, 2023. 1

[6] Matthew Lee, Felix John Samuel Bragman, Ricardo Sanchez-Matilla, Imanol Luengo, and Danail Stoyanov. Spatial-temporal nas for fast surgical segmentation. In *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA, 2024. 1

[7] Mengtian Li, Yu-Xiong Wang, and Deva Ramanan. Towards streaming perception, 2020. 1, 2

[8] Erica Padovan, Giorgia Marullo, Leonardo Tanzi, Pietro Piazzolla, Sandro Moos, Francesco Porpiglia, and Enrico Vezzetti. A deep learning framework for real-time 3d model registration in robot-assisted laparoscopic surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 18, 2022. 1

[9] Xiaofeng Wang, Zheng Zhu, Yunpeng Zhang, Guan Huang, Yun Ye, Wenbo Xu, Ziwei Chen, and Xingang Wang. Are we ready for vision-centric driving streaming perception? the asap benchmark, 2022. 2

[10] Ruofeng Wei, Bin Li, Kai Chen, Yiyao Ma, Yunhui Liu, and Qi Dou. Enhanced scale-aware depth estimation for monocular endoscopic scenes with geometric modeling, 2024. 1

[11] Jinrong Yang, Songtao Liu, Zeming Li, Xiaoping Li, and Jian Sun. Real-time object detection for streaming perception, 2022. 2