SurgCUT3R: Surgical Scene-Aware Continuous Understanding of Temporal 3D Representation

Kaiyuan Xu[1], Fangzhou Hong[2], Daniel Elson[1] and Baoru Huang[1,3]

***Abstract*— Reconstructing surgical scenes from monocular endoscopic video is critical for advancing robotic-assisted surgery. However, the application of state-of-the-art general-purpose reconstruction models is constrained by two key challenges: the lack of supervised training data and performance degradation over long video sequences. To overcome these limitations, we propose SurgCUT3R, a systematic framework that adapts unified 3D reconstruction models to the surgical domain. Our contributions are threefold. First, we develop a data generation pipeline that exploits public stereo surgical datasets to produce large-scale, metric-scale pseudo-ground-truth depth maps, effectively bridging the data gap. Second, we propose a hybrid supervision strategy that couples our pseudo-ground-truth with geometric self-correction to enhance robustness against inherent data imperfections. Third, we introduce a hierarchical inference framework that employs two specialized models to effectively mitigate accumulated pose drift over long surgical videos: one for global stability and one for local accuracy. Experiments on the SCARED and StereoMIS datasets demonstrate that our method achieves a competitive balance between accuracy and efficiency, delivering near state-of-the-art but substantially faster pose estimation and offering a practical and effective solution for robust reconstruction in surgical environments.**

I. INTRODUCTION

Reconstruction of the surgical scene from monocular endoscopic video is a crucial task in advancing robotic-assisted surgery [1], [2]. By creating a dense Reconstruction model of the observed tissues and instruments, it enables a range of downstream applications, including intraoperative navigation [3]–[5], robotic surgery automation [6], [7], and virtual reality simulation [8]–[10]. This problem has been long studied at the intersection of computer vision and medical imaging, building upon foundational areas such as monocular depth estimation [11], [12], Structure-from-Motion (SfM) [13], [14], and Simultaneous Localization and Mapping (SLAM) [15], [16].

Despite decades of progress, achieving robust, scale-consistent reconstruction from monocular endoscopic video remains an open challenge. Classical SfM or SLAM pipelines [13], [14], [16], while mature in rigid and well-textured environments, often break down in surgical settings due to non-rigid tissue deformation, frequent occlusions by instruments, and texture-poor surfaces. Meanwhile, general-purpose learning-based methods for monocular depth estimation [17]–[19] have made rapid strides, but they suffer from a

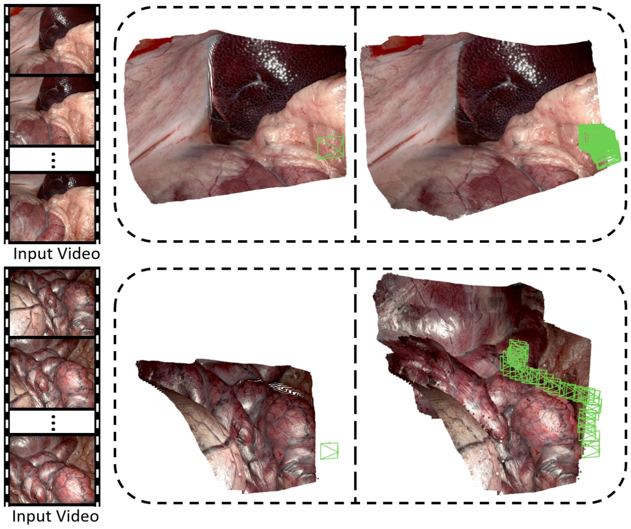

Fig. 1: **Qualitative results of 3D reconstruction.** With videos (small images) as input, this figure shows the reconstruction from the first frame (large images left) and the accumulated 3D model from multiple frames (large images right). This alignment between the single-frame and multi-frame reconstruction results highlights the geometric consistency of our method.

severe domain gap when applied to surgical data. As a result, they typically fail to deliver the geometrically-consistent and scale-consistent reconstructions that meet the requirements of clinical applications.

In parallel, recent advances in large-scale learning-based models for general-purpose 3D reconstruction, such as those based-on DUSt3R [20], have shown transformative potential. Unified frameworks like CUT3R [21] have demonstrated remarkable success in reconstructing diverse scenes from online monocular videos. However, adapting this state-of-the-art (SOTA) technology to the surgical domain is impeded by two fundamental barriers. First, the success of these models is based on vast datasets with high-quality, ground-truth (GT) 3D data for supervision, creating a significant data gap for the surgical domain where such data is difficult to acquire. Second, the autoregressive architecture of these models (e.g. CUT3R [21]), though effective for short clips, degrade over the long, continuous surgical video streams, leading to a scalability gap manifested as accumulated pose drift.

To address these challenges, we propose SurgCUT3R, a systematic framework that adapts an SOTA unified reconstruction model to the surgical domain. This work specifi-

[1] The Hamlyn Centre for Robotic Surgery, Imperial College London, SW7 2AZ, UK. Baoru.Huang18@imperial.ac.uk

[2] S-Lab, College of Computing and Data Science, Nanyang Technological University, Singapore 639798.

[3] Department of Computer Science, University of Liverpool, L69 7ZX, UK.

cally addresses two fundamental challenges: (i) How can we effectively train a supervised model in the absence of real GT data? (ii) How can we architect a long-term inference pipeline that preserves the model's accuracy while mitigating pose drift?

Our main contributions are summarized as follows:

- We develop a scalable data generation pipeline that leverages public stereo surgical videos to produce large-scale pseudo-GT depth maps with scale-consistency, bridging the data gap for supervised model training.
- We introduce a hybrid supervision strategy that combines direct supervision from our pseudo-GT with a comprehensive geometric self-correction mechanism, enhancing robustness against inherent data imperfections.
- We design a hierarchical inference framework that employs two specialized models to effectively mitigate pose drift, enabling stable camera tracking over long surgical video sequences.

II. Related Work

A.Learning-Based Dense Correspondence for Reconstruction

Recent advances in learning-based dense correspondence have been largely driven by transformer architectures, such as CroCo [22], which perform the task of cross-view completion. This pre-training paradigm enabled the development of DUSt3R [20], a seminal work that recasts the problem by directly regressing dense 3D pointmaps from an image pair. By predicting two pointmaps in one coordinate frame, DUSt3R [20] implicitly solves for relative pose and geometry but requires an offline, optimization-based global alignment for multi-view scenes. The DUSt3R [20] framework inspired several specialized extensions. MASt3R [23] improved matching precision by adding a head to predict dense features, improving performance on correspondence tasks like visual localization. Concurrently, MonST3R [24] adapted the model to handle dynamic scenes by fine-tuning it on video data with moving objects. To overcome the offline, pairwise nature of these models, Spann3R [25] introduced an online and incremental approach. It incorporates an external spatial memory to store past geometric information, allowing each new frame's pointmap to be regressed directly into a consistent global coordinate system without optimization. Building on this concept of continuous reconstruction, CUT3R [21] proposed a model with a persistent state that is not simply a cache of past observations. This learned state representation captures powerful scene priors, enabling the model to not only reconstruct observed regions online but also to infer and generate geometry for unobserved areas by querying the state with virtual camera views.

B.SLAM-based Methods for Long-Sequence Consistency

To mitigate the inherent drift of reconstruction models over long sequences, several works have integrated them into SLAM-based systems to ensure global consistency. MASt3R-SLAM [26] pioneered this direction by building a real-time, keyframe-based SLAM system upon the MASt3R [23] prior, treating the two-view model as a robust front-end for a classic SLAM back-end with pose graph optimization. SLAM3R [27] pursued a fully neural approach, avoiding explicit camera parameters and using separate networks to generate local reconstructions and register them into a global model. To address the challenge of casual dynamic videos, MegaSaM [28] extended a deep visual SLAM framework by integrating monocular depth priors and motion probability maps into a differentiable bundle adjustment layer and using excellent optimization methods to enhance robustness for low-parallax scenarios.

C.Reconstruction for Surgical Scenes

The 3D reconstruction of surgical scenes has rapidly developed. Early self-supervised methods like AF-SfMLearner [29] addressed specific challenges such as brightness inconsistency but relied on separate networks for depth and motion estimation. The advent of foundation models shifted the focus to efficient adaptation, as seen in EndoDAC [30], which uses parameter-efficient techniques like LoRA [31] to fine-tune large pre-trained models for surgical data. While effective, this maintained a non-unified structure. More recently, the pursuit of geometrically coherent online systems led to unified frameworks like Endo3R [32]. It extended pairwise reconstruction to long, dynamic videos using an uncertainty-aware memory mechanism and output a single, globally aligned pointmap from which both scale-consistent depth and camera poses are derived, ensuring inherent consistency. Our work specifically adapts a SOTA general-domain model, CUT3R [21], to the unique challenges of the surgical environment.

III. Methodology

Our proposed method, SurgCUT3R, establishes a systematic framework for adapting a unified reconstruction model to the challenging domain of monocular surgical video, with the main objective of achieving geometrically consistent results. To this end, we introduce a solution built on three key contributions: (1) To address the scarcity of supervised training data, we develop a pseudo-GT generation pipeline that leverages existing stereo datasets to produce high-quality depth labels. (2) To ensure robustness against inherent data imperfections, we employ a hybrid supervision strategy that couples our pseudo-GT data with a geometric self-correction mechanism. (3) To tackle the challenge of long-sequence inference, we propose a hierarchical framework that effectively suppresses pose drift. An overview of our method is illustrated in Fig. 2.

A.Preliminaries: CUT3R [21]

CUT3R [21] is a unified online reconstruction framework. Its core feature is the use of a continuously updated state to process an incoming stream of images. The model generates scale-consistent 3D pointmaps and camera parameters for each new image frame, progressively accumulating these pointmaps to build a coherent and dense 3D scene reconstruction.

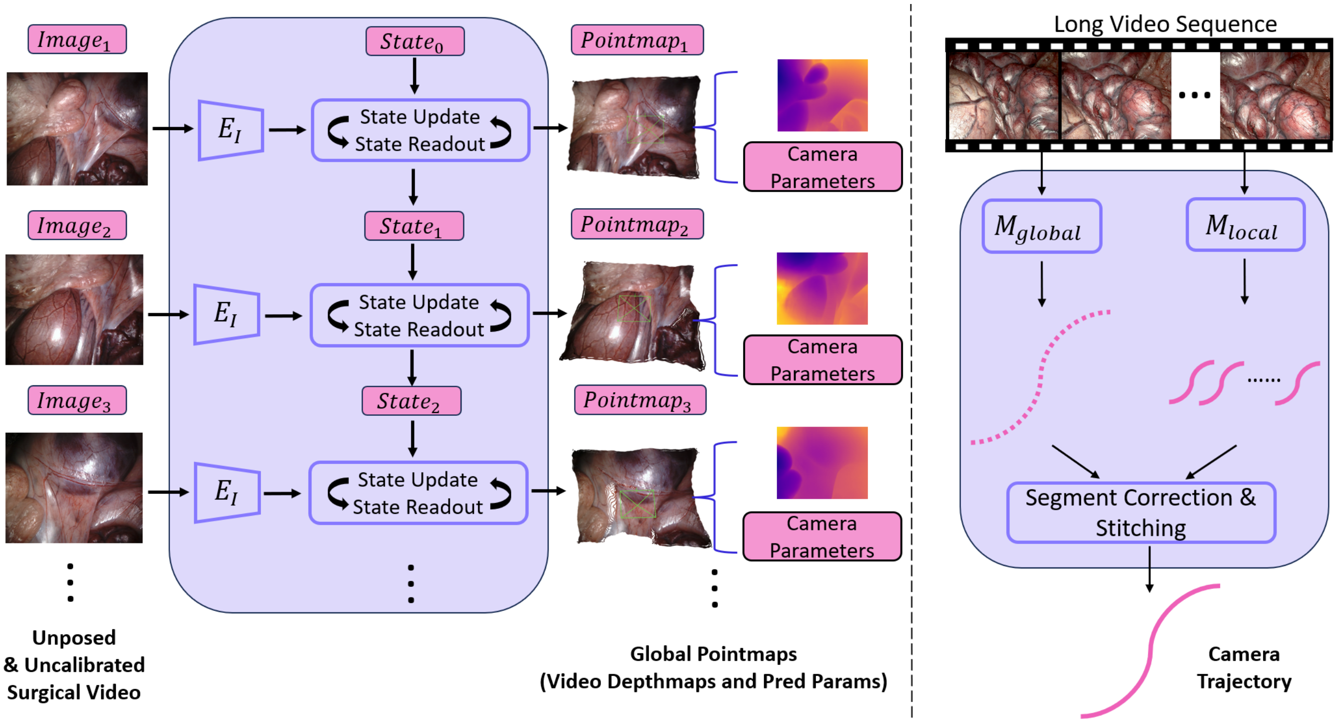

Fig. 2: **Overview of SurgCUT3R. Left:** The unified reconstruction pipeline. Streaming video frames are encoded via a ViT encoder and interact with a persistent state, which is continuously updated to sequentially output the pointmap and camera parameter for each frame. **Right:** Our hierarchical framework for long-sequence inference. The pink lines represent camera trajectories. A sparse but globally stable trajectory from a global model (M_{global}) provides anchor points to correct and stitch the dense but locally drifting trajectories from a local model (M_{local}), producing a final, drift-corrected trajectory.

Its core module is a State-Input interaction mechanism, where the model maintains a set of tokens as a latent state, denoted by s. Before processing any image, this state is initialized as a set of learnable tokens. For each input image I_t at timestep t, a Vision Transformer encoder first converts it into image tokens F_t.

These tokens then enter the main interaction process, which is visualized in the left panel of Fig. 2. Here, the tokens F_t engage in a bidirectional interaction with the previous state, s_{t-1}, within the module labeled as 'State Update State Readout'. This module performs two simultaneous operations: the 'state-update' operation integrates new information from I_t to produce the updated state s_t. Concurrently, the 'state-readout' operation enriches the image tokens F_t with historical context from s_{t-1} to create enhanced tokens F'_t. While not explicitly shown in the figure, a learnable 'pose token' (z) is processed alongside F_t, and its output (z'_t) is used to capture global scene information for predicting the camera pose. The original CUT3R [21] paper formulates this entire interaction as:

$$[z'_t, F'_t], s_t = \text{Decoder}([z, F_t], s_{t-1}) \tag{1}$$

Following the interaction, the model extracts explicit 3D representations from the enhanced tokens using several heads. Specifically, the model predicts pointmaps and their corresponding confidence maps in two coordinate frames: $(\hat{X}_t^{\text{self}}, C_t^{\text{self}})$ in the input image's camera coordinate system, and $(\hat{X}_t^{\text{world}}, C_t^{\text{world}})$ in the world coordinate system (defined by the first frame's camera). Concurrently, the model predicts the camera pose $\hat{P}_t$, which represents the transformation from the current frame to the world frame:

$$(\hat{X}_t^{\text{self}}, C_t^{\text{self}}) = \text{Head}_{\text{self}}(F'_t) \tag{2}$$

$$(\hat{X}_t^{\text{world}}, C_t^{\text{world}}) = \text{Head}_{\text{world}}(F'_t, z'_t) \tag{3}$$

$$\hat{P}_t = \text{Head}_{\text{pose}}(z'_t) \tag{4}$$

where $\text{Head}_{\text{self}}$ and $\text{Head}_{\text{world}}$ adopt the DPT [33] architecture, and $\text{Head}_{\text{pose}}$ is an MLP network.

During the training phase, the model is optimized using a composite loss function. For pointmap regression, a confidence-aware regression loss is applied:

$$\mathcal{L}_{\text{conf}} = \sum_{(\hat{x},c)\in(\hat{\mathcal{X}},C)} \left(c \cdot \left\| \frac{\hat{x}}{\hat{s}} - \frac{x}{s} \right\|_2 - \alpha \log c \right) \tag{5}$$

where $\hat{s}$ and s are scale normalization factors for the predicted point set $\hat{\mathcal{X}}$ and GT $\mathcal{X}$ respectively. For the camera pose, the model minimizes the L2 norm between the prediction and the GT:

$$\mathcal{L}_{\text{pose}} = \sum_{t=1}^{N} \left(\|\hat{q}_t - q_t\|_2 + \left\| \frac{\hat{\tau}_t}{\hat{s}} - \frac{\tau_t}{s} \right\|_2 \right) \tag{6}$$

where the pose $\hat{P}_t$ is parameterized by a quaternion $\hat{q}_t$ and a translation vector $\hat{\tau}_t$.

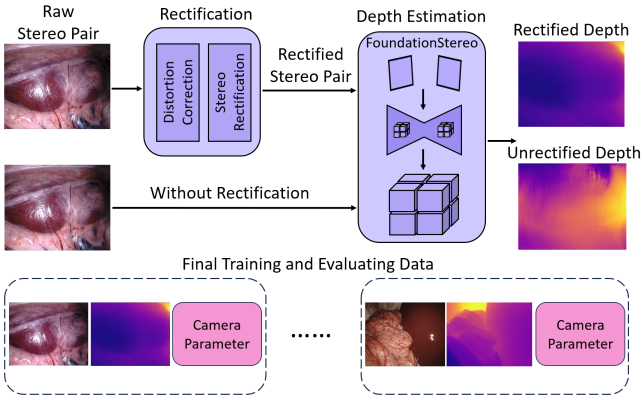

Fig. 3: **Pipeline of pseudo GT depth generation.** The process rectifies the raw stereo pair before feeding it into FoundationStereo [36] to generate a geometrically correct and metric-scale depth map.

B.Proposed Method: SurgCUT3R

1) Pseudo-GT Generation for Surgical Scenes

A primary challenge in applying supervised 3D reconstruction models like CUT3R [21] to specialized medical field is the scarcity of suitable training data. These models require large-scale datasets with per-frame GT for both camera pose and dense depth. While existing surgical datasets such as SCARED [34] and StereoMIS [35] provide stereo video sequences with GT camera parameters, they critically lack dense and reliable depth information for every frame. The SCARED [34] dataset only contains sparse structured-light depth for the initial frame of each sequence, while StereoMIS [35] offers no GT depth at all. This data gap presents a fundamental barrier to fine-tuning SOTA unified reconstruction models for the surgical domain.

To overcome this obstacle, we propose a pseudo-GT generation pipeline designed to synthesize a high-fidelity dataset for supervised training. The core principle of our pipeline is to leverage the inherent geometric constraints of the available stereo video data to generate dense and metric-scale pseudo-GT depth maps for monocular training.

Our pipeline transforms the raw stereo sequences into a collection of (image, pseudo-GT depth, and GT pose) triplets through a systematic two-step process, as illustrated in Fig. 3:

a) Stereo Preprocessing and Rectification.

We utilize the stereo video sequences from the SCARED [34] and StereoMIS [35] datasets. Some of the raw endoscopic images in these datasets suffer from non-linear lens distortions and are not co-planar, resulting in skewed epipolar lines that complicate stereo matching. To address this, we follow the preprocessing procedures outlined in MSDESIS [37] and Endo-4DGS [38]. The process, guided by the provided camera calibration files, involves two main steps: first, we perform distortion correction to remove lens-induced artifacts. Second, we apply stereo rectification to align the image planes. This procedure yields the final distortion-free and row-aligned stereo image pairs, I_L and I_R, required for reliable stereo matching.

b) Metric-Scale Depth Synthesis and Dataset Assembly.

We generate dense depth maps from the rectified stereo pairs, I_L and I_R, using FoundationStereo [36]. The resulting disparity map , denoted as d, is converted to a metric-scale depth map D using the known camera baseline b and focal length f which are provided by the dataset, according to the equation 7:

$$D(u, v) = \frac{b \cdot f}{d(u, v)}. \tag{7}$$

Finally, for each frame at time t, we assemble the left image $I_L^{(t)}$, the synthesized depth map $D^{(t)}$, and the GT camera pose and intrinsics $P_{\text{GT}}^{(t)}$ to form the training dataset. This pipeline produces a large metric-scale dataset suitable for supervised training.

2) Hybrid Supervision for Robust Training

While the pseudo-GT pipeline described in Section B.1 provides high-quality metric-scale supervision, the synthesized depth maps are not entirely free from imperfections. Surgical scenes frequently contain challenging elements such as specular reflections from wet tissue surfaces, smoke from electrocautery, and low-texture regions. These factors can introduce local noise and inaccuracy into the depth maps generated by the stereo matching model. Directly training a network with a purely supervised loss on this imperfect data risks overfitting to label noise, which could lead the model to learn incorrect geometric priors.

To mitigate the effect of label noise and enhance the model's geometric understanding, we introduce a hybrid supervision strategy for training. This strategy combines the direct supervision from our pseudo-GT data with a geometric consistency self-supervision term. The core idea is to use the pseudo-GT to guide the overall learning while leveraging multi-view self-consistency as a powerful regularizer to refine the geometric structure and promote robustness against noisy labels.

Our total training objective, $\mathcal{L}_{\text{total}}$, is a weighted sum of the supervised losses from the baseline model and our self-supervised consistency loss:

$$\mathcal{L}_{\text{total}} = (\mathcal{L}_{\text{conf}} + \mathcal{L}_{\text{pose}}) + \lambda_{\text{consist}} \cdot \mathcal{L}_{\text{consistency}}, \tag{8}$$

where λ_{consist} is a hyperparameter that balances the supervised and self-supervised terms.

a) Supervised Terms ($\mathcal{L}_{conf}$ and $\mathcal{L}_{pose}$).

The primary component of our loss is the supervised objective inherited from the baseline CUT3R [21] model, comprising two parts. The first is a confidence-weighted regression loss, $\mathcal{L}_{\text{conf}}$ (Eq. 5), which minimizes the error between the predicted and pseudo-GT point clouds. The second is a pose regression loss, $\mathcal{L}_{\text{pose}}$ (Eq. 6), which minimizes the error between the predicted and GT camera pose. These terms anchor the model's predictions to the largely accurate and metric-scale pseudo-GT we have generated, ensuring the model learns the overall correct scale and structure of the surgical scene.

TABLE I
QUANTITATIVE EVALUATION OF OUR SURGCUT3R METHOD AGAINST EXISTING METHODS IN ENDOSCOPIC SCENE RECONSTRUCTION. THE OPTIMAL AND SUBOPTIMAL RESULTS ARE SHOWN IN BOLD AND UNDERLINED RESPECTIVELY.

Dataset	Category	Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE LOG↓	$\delta < 1.25$ ↑	ATE↓	RTE↓	FPS↑
SCARED [34]	Optimization-based	MonST3R(w/ Opt) [24]	$\underline{0.098}$	$\underline{1.237}$	$\underline{7.979}$	$\underline{0.132}$	$\underline{0.904}$	$\underline{21.774}$	$\underline{1.582}$	$\underline{0.3}$
		MegaSaM [28]	**0.056**	**0.392**	**4.586**	**0.074**	**0.978**	**2.002**	**0.315**	**0.7**
	Feed-forward	Spann3R [25]	0.119	2.524	10.218	0.148	0.867	10.258	1.260	19.2
		AF-SfMLearner [29]	0.073	0.534	5.028	0.081	0.964	10.312	0.971	3.6
		EndoDAC [30]	$\underline{0.059}$	$\underline{0.443}$	$\underline{4.833}$	$\underline{0.079}$	$\underline{0.973}$	$\underline{10.225}$	$\underline{0.963}$	**36.3**
		SurgCUT3R (Ours)	**0.057**	**0.410**	**4.647**	**0.077**	**0.977**	**5.514**	**0.752**	$\underline{19.7}$
StereoMIS [35]	Optimization-based	MonST3R(w/ Opt) [24]	$\underline{0.187}$	$\underline{4.342}$	$\underline{15.671}$	$\underline{0.225}$	$\underline{0.672}$	$\underline{41.183}$	$\underline{1.412}$	$\underline{0.3}$
		MegaSaM [28]	**0.061**	**0.506**	**4.615**	**0.084**	**0.976**	**19.705**	**0.877**	**0.7**
	Feed-forward	Spann3R [25]	0.213	7.075	17.535	0.318	0.693	29.842	1.391	19.2
		AF-SfMLearner [29]	0.102	0.872	7.410	0.110	0.896	$\underline{25.128}$	1.197	3.6
		EndoDAC [30]	$\underline{0.075}$	$\underline{0.672}$	$\underline{6.047}$	$\underline{0.094}$	$\underline{0.957}$	**24.264**	$\underline{1.121}$	**36.3**
		SurgCUT3R (Ours)	**0.070**	**0.637**	**5.732**	**0.091**	**0.965**	25.939	**0.902**	$\underline{19.7}$

b) Self-Supervised Term ($\mathcal{L}_{consistency}$).

To mitigate the effect of label noise, we introduce a comprehensive self-supervised consistency loss $\mathcal{L}_{\text{consistency}}$, inspired by the depth optimization objective in MegaSaM [28]. While MegaSaM [28] employs this loss for post-process refinement, we adapt it as a self-supervised signal during training. The necessary inputs for this loss, such as per-frame depth maps and relative poses, are taken directly from our model's own predictions within the training batch. The loss is a composite of three main terms:

$$\mathcal{L}_{\text{consistency}} = w_{\text{flow}}\mathcal{C}_{\text{flow}} + w_{\text{temp}}\mathcal{C}_{\text{temp}} + w_{\text{prior}}\mathcal{C}_{\text{prior}}, \qquad (9)$$

where w are weighting coefficients. The components are defined as follows:

- **Optical Flow Consistency ($\mathcal{C}_{\text{flow}}$):** This term enforces consistency between the 2D optical flow $\text{flow}_{i\to j}$, which is computed by a pre-trained RAFT model [39] and the 2D motion field induced by the model's predicted depth $\hat{D}_i$ and relative pose $\hat{G}_{ij}$. For a pixel $\mathbf{p}$ in frame i, its corresponding point in frame j is $\mathbf{p}' = \mathbf{p} + \text{flow}_{i\to j}(\mathbf{p})$. The reprojection of $\mathbf{p}$ using the model's prediction is $u_{ij}(\mathbf{p})$. The loss is explained as follows:

$$\mathcal{C}_{\text{flow}} = \|u_{ij}(\mathbf{p}) - \mathbf{p}'\|_1. \qquad (10)$$

- **Temporal Geometric Consistency ($\mathcal{C}_{\text{temp}}$):** This loss encourages the 3D structure to be consistent over time. It measures the scale-invariant error between a point's depth $\hat{D}_j(\mathbf{p}')$ in the target frame j at the flow-warped coordinate $\mathbf{p}'$, and its depth $P_z^{i\to j}$ projected from the source frame i. The loss is explained as follows:

$$\mathcal{C}_{\text{temp}} = \left\| \max\left(\frac{P_z^{i\to j}(\mathbf{p})}{\hat{D}_j(\mathbf{p}')}, \frac{\hat{D}_j(\mathbf{p}')}{P_z^{i\to j}(\mathbf{p})} \right) - 1 \right\|_1. \qquad (11)$$

- **Prior Regularization ($\mathcal{C}_{\text{prior}}$):** This term regularizes the geometry to prevent drift and maintain surface properties. It is a composite of three sub-terms from MegaSaM [28]: a scale-invariant loss $\mathcal{C}_{\text{si}}$ to maintain overall shape, a multi-scale gradient matching loss $\mathcal{C}_{\text{grad}}$ to preserve geometric details, and a surface normal consistency loss $\mathcal{C}_{\text{normal}}$ to encourage locally smooth surfaces.

It is worth noting that the original MegaSaM [28] utilizes a predicted uncertainty map to apply per-pixel weights within the photometric and temporal consistency losses, which down-weights contributions from dynamic regions. Since our training is conducted on static surgical scenes, we do not predict this map and instead use a constant value for this uncertainty weights.

The supervised loss provides the strong and scale-consistent signal necessary for accuracy, while the self-supervision loss acts as a geometric regularizer, enabling the model to self-correct from imperfections in the pseudo-GT data. This combination enhances the model's robustness, leading to a more accurate and geometrically coherent final reconstruction.

3) Hierarchical Framework for Long-Sequence Inference

A significant challenge for autoregressive reconstruction models like CUT3R [21] is their performance degradation over long video sequences. As the model processes frames sequentially, small prediction errors in pose inevitably accumulate. This accumulation of errors manifests itself as pose drift, where the estimated camera trajectory gradually deviates from the true path. This issue severely limits the applicability of such models for tracking entire surgical procedures, which are often lengthy.

To address the challenge of long-sequence tracking, we propose a hierarchical framework for pose estimation. The core idea is to establish a globally consistent camera trajectory by correcting the drift of a dense, short-term tracker using a sparse, long-term tracker as a stable reference. Our method achieves this through per-segment alignment and an error correction module, effectively suppressing error accumulation.

This hierarchical framework features a global model, a local model, and a fusion pipeline that integrates their outputs to generate a drift-corrected camera trajectory.

a) Dual-Model Specialization.

We train two instances of our model with different temporal sampling strategies to specialize them for different tasks:

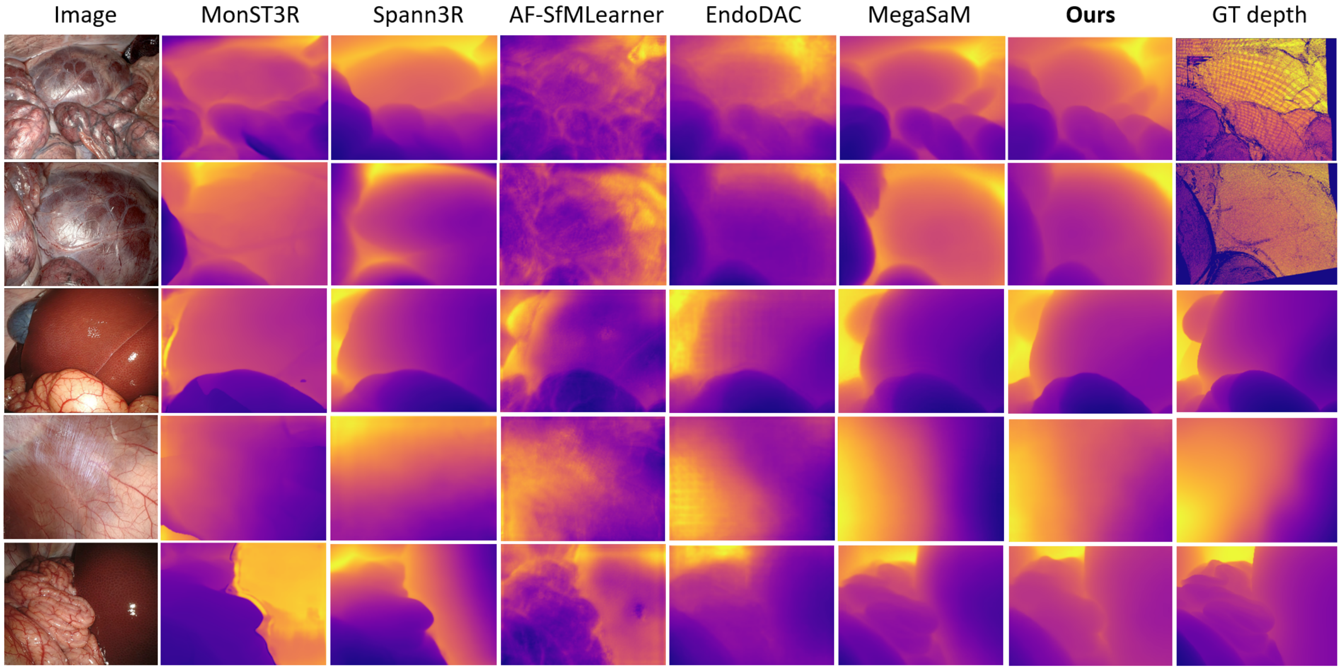

Fig. 4: **Qualitative results of monocular depth estimation.** We compare our method with MonST3R [24], Spann3R [25], AF-SfMLearner [29], EndoDAC [30] and MegaSaM [28] on SCARED [34] and StereoMIS [35] datasets. Our method achieves the best qualitative results in feed-forward methods.

- **Global Model ($\mathcal{M}_{\text{global}}$):** This model is trained on sparsely sampled video frames (e.g., with a max interval of 12 frames). Its purpose is to learn robust long-range motion estimation, focusing on global trajectory consistency.
- **Local Model ($\mathcal{M}_{\text{local}}$):** This model is trained on densely sampled video frames (e.g., with a max interval of 3 frames). It is specialized for capturing accurate relative camera motion within a short time window.

b) Trajectory Correction and Stitching.

Our pipeline corrects the trajectory on a per-segment basis as follows:

- **Pose Generation:** We first generate two sets of poses: a sparse but globally stable ”anchor” trajectory using $\mathcal{M}_{\text{global}}$, and multiple dense but locally drifting pose sequences for each segment between anchors using $\mathcal{M}_{\text{local}}$.
- **Per-Segment Correction:** For each local segment, we first align the segment's starting pose to its corresponding global anchor. Then, we calculate the drift error, which is the discrepancy between the segment's aligned end pose and the next global anchor. This error is then distributed across all frames within the segment by interpolating the rotational and translational components using spherical linear interpolation and linear interpolation, respectively. This process ultimately yields the complete drift-corrected trajectory for the entire long sequence.

This hierarchical framework effectively mitigates pose drift in long-sequence tracking. By combining the long-term stability of a global model with the local accuracy of a dense model, our method generates a complete drift-corrected trajectory.

IV. Experiments

A.Experiment Setting

1) Datasets

We train and evaluate our method on two public surgical datasets: SCARED [34] and StereoMIS [35]. For training, we use the SCARED [34] dataset, following the official split in SCARED [34] challenge (i.e., Datasets 1–7 for training and Datasets 8–9 for testing). However, a prior study [40] noted significant calibration errors in Dataset4 and Dataset5. To ensure that our model learns from high-quality data, we exclude these two subsets from our training dataset. To evaluate the generalization ability, we test on 4 unseen StereoMIS [35] sequences for cross-dataset validation: Sequence 1 (frames 6900–7600 in P2-1), Sequence 2 (frames 1200–1600 in P2-4), Sequence 3 (frames 9300–10200 in P2-5), and Sequence 4 (frames 1000–1300 in P2-8). These sequences were selected from our generated pseudo-GT data for their significant scene dynamics and camera motion, as well as for their high reconstruction quality with minimal visual artifacts like smoke or glare.

2) Evaluation Metrics

We evaluate performance across three aspects: pose accuracy, depth estimation quality, and efficiency. For pose accuracy, we use Absolute Trajectory Error (ATE) for global consistency and Relative Trajectory Error (RTE) with a window size of 16 for local accuracy. The unit for both ATE and RTE is millimeter (mm). For depth quality, we follow EndoDAC [30] and report five standard metrics: Absolute Relative Error (Abs Rel), Square Relative Error

(Sq Rel), Root Mean Square Error (RMSE), RMSE log, and the accuracy metric $\delta < 1.25$. The evaluation is performed at a resolution of 256x192. GT depth for SCARED [34] is from the dataset itself, while for StereoMIS [35], we use our generated pseudo-GT depth. For efficiency, we report the inference speed in Frames Per Second (FPS).

3) Implementation Details

The experiments were conducted on an NVIDIA GPU 4090 with PyTorch framework. We use the AdamW optimizer with a learning rate of 1.0×10^{-5}, a weight decay of 0.05, and a batch size of 8. Our training follows a two-stage strategy. First, we fine-tune the publicly available CUT3R [21] weight for 5 epochs using only the supervised losses ($\mathcal{L}_{\text{conf}} + \mathcal{L}_{\text{pose}}$). Then, we introduce the self-supervised term $\mathcal{L}_{\text{consistency}}$ and continue training for two more epochs to refine the model. For the hierarchical inference framework, both the global and local models share the same training configuration, but are differentiated by their maximum temporal sampling interval during training, which is set to 12 for the global model and 3 for the local model.

B.Qualitative and Quantitative Evaluation

We evaluate our method by comparing its performance with several SOTA methods on both the SCARED [34] and StereoMIS [35] datasets.

a) Quantitative Comparison.

As listed in Table I, our method demonstrates a highly competitive balance between accuracy and efficiency. On the SCARED [34] dataset, the multi-stage optimization method MegaSaM [28] achieves the best overall accuracy in both depth and pose estimation. However, this high precision comes at a significant computational cost, limiting its reconstruction speed to 0.7 FPS. In contrast, our method achieves the second-best pose estimation results, and maintains depth quality nearly identical to the SOTA, while operating at a much faster 19.7 FPS. When tested in the StereoMIS [35] dataset, our method continues to deliver competitive performance in depth estimation, proving its effectiveness beyond the training domain. The results highlight our approach's primary strength: it provides a practical and efficient solution that achieves near-SOTA accuracy at a relatively high speed.

b) Qualitative Comparison.

Fig. 4 visually validates our depth estimation, showing that SurgCUT3R achieves precise depth maps with superior relative scale. And as shown in Fig. 1, The high quality of the resulting 3D reconstructions and pose estimations is mainly attributed to our accurate depth and pose predictions.

C.Ablations and Analysis

To validate the effectiveness of our proposed components, we conduct two ablation studies on the SCARED [34] dataset.

a) Effect of Hybrid Supervision.

We first evaluate the impact of our self-supervised consistency loss. As shown in Table II, our consistency loss (w/ $L_{\text{consistency}}$) yields a marginal but consistent improvement across depth estimation metrics compared to using only supervised losses (w/o $L_{\text{consistency}}$).

TABLE II
ABLATION STUDY ON THE LOSS FUNCTION

Configuration	**Sq Rel**↓	**RMSE**↓	$\boldsymbol{\delta < 1.25} \uparrow$
w/o $L_{\text{consistency}}$	0.423	4.763	0.975
w/ $L_{\text{consistency}}$ (Ours)	**0.410**	**4.647**	**0.977**

b) Effect of Hierarchical Framework.

Next, we analyze the contribution of our dual-model hierarchical framework for long-sequence pose estimation. Table III shows that using our dual-architecture significantly reduces the ATE compared to using a single model. This result quantitatively demonstrates the effectiveness of our approach in mitigating accumulated pose drift. The qualitative improvement in trajectory stability is further visualized in Fig. 5.

TABLE III
ABLATION STUDY ON THE ARCHITECTURE

Configuration	**ATE**↓
CUT3R [21] Only	9.361
Dual-Arch (Ours)	**5.514**

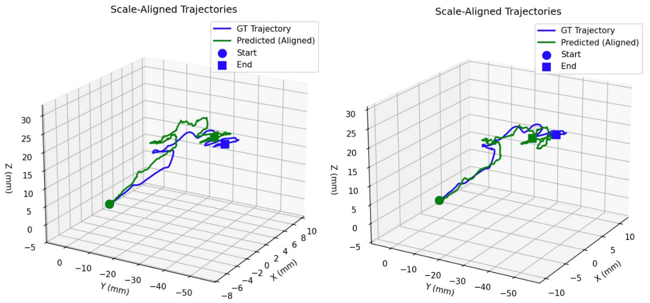

Fig. 5: Qualitative comparison of camera trajectories. **Left:** Without the hierarchical inference framework. **Right:** With our hierarchical inference framework (Ours).

V. Conclusion

We present SurgCUT3R, a unified framework for monocular surgical 3D reconstruction. To overcome data scarcity and inherent label noise, we develop a metric-scale pseudo-GT generation pipeline from stereo datasets, coupled with a hybrid supervision strategy for geometric self-correction. Furthermore, a hierarchical inference framework is introduced to effectively mitigate long-term pose drift. Unlike slow offline methods such as MegaSaM [28], SurgCUT3R strikes a clinically practical balance, delivering competitive accuracy at 19.7 FPS. This efficiency makes it a robust solution for surgical navigation. Although our current pseudo-GT is effective, future work will take advantage of offline optimization frameworks such as MegaSaM [28] to mitigate artifact-induced depth misalignments and construct more accurate training data.

References

[1] A. Moglia, K. Georgiou, E. Georgiou, R. M. Satava, and A. Cuschieri, "A systematic review on artificial intelligence in robot-assisted surgery," *International Journal of Surgery*, vol. 95, p. 106151, 2021.

[2] L. Qian, J. Y. Wu, S. P. DiMaio, N. Navab, and P. Kazanzides, "A review of augmented reality in robotic-assisted surgery," *IEEE Transactions on Medical Robotics and Bionics*, vol. 2, no. 1, pp. 1–16, 2019.

[3] S. C. Overley, S. K. Cho, A. I. Mehta, and P. M. Arnold, "Navigation and robotics in spinal surgery: where are we now?" *Neurosurgery*, vol. 80, no. 3S, pp. S86–S99, 2017.

[4] A. J. Karkenny, J. R. Mendelis, D. S. Geller, and J. A. Gomez, "The role of intraoperative navigation in orthopaedic surgery," *JAAOS-Journal of the American Academy of Orthopaedic Surgeons*, vol. 27, no. 19, pp. e849–e858, 2019.

[5] R. B. Bell, "Computer planning and intraoperative navigation in cranio-maxillofacial surgery," *Oral and Maxillofacial Surgery Clinics*, vol. 22, no. 1, pp. 135–156, 2010.

[6] V. Penza, E. De Momi, N. Enayati, T. Chupin, J. Ortiz, and L. S. Mattos, "Envisors: Enhanced vision system for robotic surgery. a user-defined safety volume tracking to minimize the risk of intraoperative bleeding," *Frontiers in Robotics and AI*, vol. 4, p. 15, 2017.

[7] Z. Wang, T. Li, J.-Q. Zheng, and B. Huang, "When cnn meet with vit: Towards semi-supervised learning for multi-class medical image semantic segmentation," in *ECCV*. Springer, 2022, pp. 424–441.

[8] A. Ayoub and Y. Pulijala, "The application of virtual reality and augmented reality in oral & maxillofacial surgery," *BMC oral health*, vol. 19, no. 1, p. 238, 2019.

[9] E. Yiannakopoulou, N. Nikiteas, D. Perrea, and C. Tsigris, "Virtual reality simulators and training in laparoscopic surgery," *International Journal of Surgery*, vol. 13, pp. 60–64, 2015.

[10] V. F. Bielsa, "Virtual reality simulation in plastic surgery training. literature review," *Journal of Plastic, Reconstructive & Aesthetic Surgery*, vol. 74, no. 9, pp. 2372–2378, 2021.

[11] J. J. Han, A. Acar, C. Henry, and J. Y. Wu, "Depth anything in medical images: A comparative study," *arXiv preprint arXiv:2401.16600*, 2024.

[12] Q. He, G. Feng, S. Bano, D. Stoyanov, and S. Zuo, "Monolot: Self-supervised monocular depth estimation in low-texture scenes for automatic robotic endoscopy," *JBHI*, vol. 28, no. 10, pp. 6078–6091, 2024.

[13] Z. Wang, C. Liu, S. Zhang, and Q. Dou, "Foundation model for endoscopy video analysis via large-scale self-supervised pre-train," in *MICCAI*. Springer, 2023, pp. 101–111.

[14] D. Recasens, J. Lamarca, J. M. Fácil, J. M. Montiel, and J. Civera, "Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints," *RAL*, vol. 6, no. 4, pp. 7225–7232, 2021.

[15] H. Zhou and J. Jayender, "Emdq-slam: Real-time high-resolution reconstruction of soft tissue surface from stereo laparoscopy videos," in *MICCAI*. Springer, 2021, pp. 331–340.

[16] K. B. Ozyoruk, G. I. Gokceler, T. L. Bobrow, G. Coskun, K. Incetan, Y. Almalioglu, F. Mahmood, E. Curto, L. Perdigoto, M. Oliveira, *et al.*, "Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos," *MIA*, vol. 71, p. 102058, 2021.

[17] S. Chen, H. Guo, S. Zhu, F. Zhang, Z. Huang, J. Feng, and B. Kang, "Video depth anything: Consistent depth estimation for super-long videos," in *CVPR*, 2025, pp. 22 831–22 840.

[18] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *NeuIPS*, vol. 37, pp. 21 875–21 911, 2024.

[19] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 10 371–10 381.

[20] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *CVPR*, 2024, pp. 20 697–20 709.

[21] Q. Wang, Y. Zhang, A. Holynski, A. A. Efros, and A. Kanazawa, "Continuous 3d perception model with persistent state," in *CVPR*, 2025, pp. 10 510–10 522.

[22] P. Weinzaepfel, V. Leroy, T. Lucas, R. Brégier, Y. Cabon, V. Arora, L. Antsfeld, B. Chidlovskii, G. Csurka, and J. Revaud, "Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion," *NeuIPS*, vol. 35, pp. 3502–3516, 2022.

[23] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3d with mast3r," in *ECCV*. Springer, 2024, pp. 71–91.

[24] J. Zhang, C. Herrmann, J. Hur, V. Jampani, T. Darrell, F. Cole, D. Sun, and M.-H. Yang, "Monst3r: A simple approach for estimating geometry in the presence of motion," *arXiv preprint arXiv:2410.03825*, 2024.

[25] H. Wang and L. Agapito, "3d reconstruction with spatial memory," *arXiv preprint arXiv:2408.16061*, 2024.

[26] R. Murai, E. Dexheimer, and A. J. Davison, "Mast3r-slam: Real-time dense slam with 3d reconstruction priors," in *CVPR*, 2025, pp. 16 695–16 705.

[27] Y. Liu, S. Dong, S. Wang, Y. Yin, Y. Yang, Q. Fan, and B. Chen, "Slam3r: Real-time dense scene reconstruction from monocular rgb videos," in *CVPR*, 2025, pp. 16 651–16 662.

[28] Z. Li, R. Tucker, F. Cole, Q. Wang, L. Jin, V. Ye, A. Kanazawa, A. Holynski, and N. Snavely, "Megasam: Accurate, fast and robust structure and motion from casual dynamic videos," in *CVPR*, 2025, pp. 10 486–10 496.

[29] S. Shao, Z. Pei, W. Chen, W. Zhu, X. Wu, D. Sun, and B. Zhang, "Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue," *MIA*, vol. 77, p. 102338, 2022.

[30] B. Cui, M. Islam, L. Bai, A. Wang, and H. Ren, "Endodac: Efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera," in *MICCAI*. Springer, 2024, pp. 208–218.

[31] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[32] J. Guo, W. Dong, T. Huang, H. Ding, Z. Wang, H. Kuang, Q. Dou, and Y.-H. Liu, "Endo3r: Unified online reconstruction from dynamic monocular endoscopic video," *arXiv preprint arXiv:2504.03198*, 2025.

[33] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *CVPR*, 2021, pp. 12 179–12 188.

[34] M. Allan, J. Mcleod, C. Wang, J. C. Rosenthal, Z. Hu, N. Gard, P. Eisert, K. X. Fu, T. Zeffiro, W. Xia, *et al.*, "Stereo correspondence and reconstruction of endoscopic data challenge," *arXiv preprint arXiv:2101.01133*, 2021.

[35] M. Hayoz, C. Hahne, M. Gallardo, D. Candinas, T. Kurmann, M. Allan, and R. Sznitman, "Learning how to robustly estimate camera pose in endoscopic videos," *IJCARS*, vol. 18, no. 7, pp. 1185–1192, 2023.

[36] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," in *CVPR*, 2025, pp. 5249–5260.

[37] D. Psychogyios, E. Mazomenos, F. Vasconcelos, and D. Stoyanov, "Msdesis: Multitask stereo disparity estimation and surgical instrument segmentation," *TMI*, vol. 41, no. 11, pp. 3218–3230, 2022.

[38] Y. Huang, B. Cui, L. Bai, Z. Guo, M. Xu, M. Islam, and H. Ren, "Endo-4dgs: Endoscopic monocular scene reconstruction with 4d gaussian splatting," in *MICCAI*. Springer, 2024, pp. 197–207.

[39] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *ECCV*. Springer, 2020, pp. 402–419.

[40] Z. Li, X. Liu, N. Drenkow, A. Ding, F. X. Creighton, R. H. Taylor, and M. Unberath, "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *CVPR*, 2021, pp. 6197–6206.