Realistic Surgical Simulation from Monocular Videos

Abstract

This paper tackles the challenge of automatically performing realistic surgical simulations from readily available surgical videos. Recent efforts have successfully integrated physically grounded dynamics within 3D Gaussians to perform high-fidelity simulations in well-reconstructed simulation environments from static scenes. However, they struggle with the geometric inconsistency in reconstructing simulation environments and unrealistic physical deformations in simulations of soft tissues when it comes to dynamic and complex surgical processes. In this paper, we propose SurgiSim, a novel automatic simulation system to overcome these limitations. To build a surgical simulation environment, we maintain a canonical 3D scene composed of 3D Gaussians coupled with a deformation field to represent a dynamic surgical scene. This process involves a multi-stage optimization with trajectory and anisotropic regularization, enhancing the geometry consistency of the canonical scene, which serves as the simulation environment. To achieve realistic physical simulations in this environment, we implement a Visco-Elastic deformation model based on the Maxwell model, effectively restoring the complex deformations of tissues. Additionally, we infer the physical parameters of tissues by minimizing the discrepancies between the input video and simulation results guided by estimated tissue motion, ensuring realistic simulation outcomes. Experiments on various surgical scenarios and interactions demonstrate SurgiSim's ability to perform realistic simulation of soft tissues among surgical procedures, showing its enormous potential for enhancing surgical training, planning, and robotic surgery systems.

1 Introduction

Realistic surgical simulation systems are pivotal in enhancing clinical training, offering substantial benefits for training surgeons, and advancing the development of robotic surgery systems (18; 35). Currently, mainstream surgical simulation systems such as LaparoS ™ from VIRTAMED are commercial products that primarily utilize mesh-based technology.

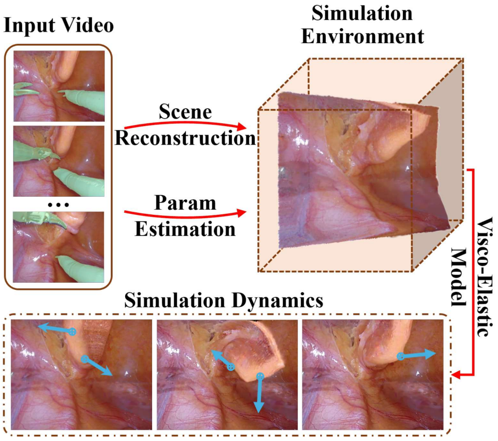

Figure 1: An overview of SurgiSim. From an input surgical video, we first reconstruct a 3D scene as the simulation environment, then estimate its physical parameters based on a Visco-Elastic deformation model, and finally perform realistic simulation dynamics in it. The green parts are surgical tool masks and the blue arrows indicate the directions of applied forces in simulations.

While these systems provide highly realistic simulations, their performance and reliability heavily depend on labor-intensive mesh building and physical parameter tuning. The dependence on the skills of modelers and animators limits the variety of possible simulation scenarios. In this paper, we aim to develop a high-quality and flexible simulation system capable of automatically converting real surgical videos into surgical simulation environments as well as performing high-fidelity and realistic simulations.

Building high-quality surgical simulation systems from readily available surgical videos poses significant challenges, including the reconstruction of simulation environments from monocular videos with complex rigid and non-rigid deformations and accurately modeling the soft tissues for realistic surgical simulation. Fortunately, recent advancements in scene reconstruction and physically grounded dynamics offer plausible solutions. Methods based on Neural radiance fields (NeRF) (47; 54; 53) and 3D Gaussian Splatting (3DGS) (46; 56; 34) have demonstrated impressive capabilities in reconstructing surgical scenarios. Besides, PhysGaussian (50) combines continuum mechanics with 3DGS and employs the Material Point Method (MPM) to simulate realistic motions in reconstructed models under simple point load. However, despite these technological strides, two significant challenges still obstruct the realization of the target system.

Firstly, current methods cannot **reconstruct geometrically consistent surgical simulation environments**. Recent advances in dynamic tissue reconstruction have shown promising capabilities by modeling tissues using canonical scenes with deformation fields (51; 34; 47; 54; 53). However, these methods typically treat dynamic tissue reconstruction as a task of synthesizing high-fidelity images at specific timestamps. Under this definition, the deformation field is optimized to produce visually plausible results for each individual frame rather than maintaining consistent geometry across time. Consequently, the geometry mapping between the canonical scene and different timestamps becomes physically implausible. In surgical simulations, these geometric inconsistencies lead to significant artifacts such as fragmentation and messy deformations. Thus, it is crucial to develop a method capable of efficiently transforming surgical videos into high-quality, geometrically consistent, and simulation-ready environments.

Secondly, current methods fail to **simulate complex tissue deformations**. Current representative simulation methods, such as PhysGaussian (50), primarily employ a Plastic-Elastic model with uniform simulation parameters assigned manually across the entire subject. This design falls short in capturing the complicated deformation behaviors inherent in real surgical scenarios. Moreover, the manual assignment of physical parameters to the simulation subject compromises the realism of the simulation. Recent research (60; 33) propose to leverage diffusion priors, either in the form of generated videos or through score distillation sampling (SDS) (41) to estimate physical parameters and avoid the need for manually assigned parameters. While these diffusion-based methods are effective in simple scenarios such as handling natural fluctuation, they fail in more complex surgical situations that involve external forces from operations like pulling or cutting. Moreover, current medical video generation models (27; 45) are still in the early stages of development (5), and the quality of their output is insufficient to provide guidance for simulations.

To address the aforementioned challenges, we propose SurgiSim, an automated system for realistic, high-quality surgical simulation leveraging readily available surgical videos. SurgiSim incorporates a novel module that employs 3DGS (23), an explicit 3D representation capable of real-time rendering, to construct a canonical scene from surgical video footage. This module utilizes multi-stage optimization with trajectory and anisotropic regularization to enhance geometry consistency, complemented by a surface thickening method to enrich reconstructed tissue content for improved simulation fidelity. To better model tissue deformations across various surgical scenarios, we implement a Visco-Elastic deformation model based on the Maxwell model (21). We further employ a physics-guided parameter estimation method to acquire physical parameters of tissues, leveraging observed physical effects in the input video. Specifically, we utilize point tracking (25) and depth estimation (55) to derive the tissue motions induced by external forces, with physical parameters optimized through differentiable simulation and rasterization. This method ensures high-quality simulation with realistic physical effects, closely mimicking the complexities of real surgical scenarios. Extensive experiments across diverse surgical procedures demonstrate SurgiSim's capability to realistically simulate soft tissue interactions, highlighting its potential for enhancing surgical training, planning, and robotic surgery systems.

We summarize our contributions as follows.

1. An automatic system that takes only a monocular surgical video as input to a) build a simulation-ready surgical environment with physics parameters estimated based on the input video and b) perform realistic physics-based simulation in it.
2. A novel multi-stage optimization strategy for 3DGS that ensures geometric consistency across temporal deformations, enabling reliable simulation environments.
3. A Visco-Elastic deformation model, coupled with automated physical parameter estimation from video observations using point tracking and depth estimation, facilitates high-fidelity and realistic surgical simulations.

2 Related Work

Dynamic Scene Representations. The rapidly advancing field of Neural Radiance Filed (NeRF (39)) and 3DGS (23) has garnered significant interest for dynamic scene reconstruction. Early works based on NeRF (1; 4; 11; 30) model scenes with implicit representations, which are difficult to interact with and therefore unsuitable for simulation. Later works based on 3DGS (49; 37; 36; 31; 57) use explicit representation, but require multi-view images or videos with severe movement of cameras, which is hard to obtain in the field of surgery. Recently, some other research (47; 54; 53; 56; 34; 51; 19; 61; 29) has focused on the reconstruction of dynamic surgical scenes. However, these studies either still use implicit representation, or primarily aim to reconstruct the dynamic scenes at specific timestamps, instead of the geometric continuity and consistency of the reconstructed canonical model.

Material Point Method. Material Point Method is a hybrid Lagrangian/Eulerian discretization scheme for solid mechanics (16). The MPM system's inherent ability to handle topology changes and frictional interactions makes it well-suited for simulating a wide range of materials (50). including elastic objects, sand, cloth, hair, snow and lava (9; 7; 12; 14; 15; 20; 48). In addition, modern implementations of MPM utilize the parallel ability of GPUs to achieve advanced efficiency(10). Recently, Xie et al. (50) uses GPU-based MPM to efficiently incorporate dynamics into different scenarios using a unified particle representation within the Gaussian Splatting framework. However, their methods rely on manually assigned physical parameters, and assume the parameters to be the same over large regions. Follow-up works by Zhang et al. (60); Liu et al. (33) further utilize the differentiable ability of the simulation process to estimate the parameters in MPM guided by Diffusion Models (2). These methods extract the physical prior from generative models, which is unreliable in the surgery field.

Surgical Simulation Systems. Medical simulation systems are based on the deformation models used to handle tissue motion under the interaction of external forces (38). Early systems apply mainly heuristic approaches like deformable spines, spring-mass models or linked volumes (6; 22; 13; 8). These systems are limited by the computing hardware and only perform simple simulations. With the development of computer graphics technology, later methods and products either base the simulation on manually built mesh models and predefined animation or replay real videos captured for VR simulators (28; 26). They require expensive human labor or advanced hardware or VR video capture but provide limited interaction. Recently, Yang et al. (58) proposed SimEndoGS, a data-driven system based on 3DGS and MPM. However, they use an elastic model to capture tissue motion and only support minor interactions like tiny force impulses.

3 Method

We propose SurgiSim (as illustrated in Fig. 1), an automatic system for surgical simulation, which constructs realistic surgical scenes as well as performs realistic simulations. The input of the system is a monocular surgical video with obvious tissue motion caused naturally or by surgical instruments. The system will reconstruct a simulation environment with physical parameters inferred from the video, and then perform high-fidelity simulations. The following sections detail the methodology: Sec. 3.1 details foundational techniques. Sec. 3.2 describes the reconstruction of a high-quality canonical simulation environment from a surgical video using a multi-stage optimization process with trajectory and anisotropic regularization. Sec. 3.3 demonstrates how we model the complex deformations of tissues, thereby facilitating high-quality simulations.

3.1 Preliminaries

3D Gaussian Splatting. We use 3D Gaussian Splatting (23), a novel differentiable rendering method which represents scenes with collections of anisotropic 3D Gaussian Kernels $\mathcal{G} = \{G_i : \mu_i, o_i, \Sigma_i, C_i\}_{i=1}^{N}$, where $\mu_i, o_i, \Sigma_i, C_i$ represents the position, opacity, covariance matrix, and spherical harmonic (SH) coefficients of the i_{th} Gaussian from all N Gaussian kernels. The covariance matrix Σ is further decomposed into rotation matrix R and scaling matrix S. To render an image through the differentiable rasterization of 3DGS, 3D Gaussian kernels will be projected onto the image plane, and the RGB color is computed as

$$\hat{\mathbf{C}} = \sum_{i \in \{N\}} \alpha_i \, \mathrm{SH}(d_i, C_i) \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{1}$$

where SH denotes the computation of color values based on the given view and SH coefficients, d_i is the view direction from the camera to G_i, and α_i represents the effective opacity ordered by z-depth, calculated by multiplying the 2D Gaussian weight with each point's inherent opacity o_i. To integrate 3DGS into the simulation pipeline, we view Gaussian kernels as particles carrying properties.

Continuum Mechanics. Continuum mechanics models motion with a transformation map $\mathbf{x} = \phi(\mathbf{X}, t)$, where $\mathbf{x}$ represents a material point in the world space Ω^t at time t, deformed from point $\mathbf{X}$ in the undeformed material space Ω^0. The deformation gradient, $\mathbf{F} = \frac{\partial \phi}{\partial \mathbf{X}}$, describes local motion and strain (3). In continuum mechanics, the two primary constraints are mass conservation and momentum conservation, given by

$$\int_{\Omega^t_\epsilon} \rho(\mathbf{x}, t) = \int_{\Omega^0_\epsilon} \rho(\mathbf{X}, 0), \ \rho(\mathbf{x}, t)\dot{\mathbf{v}}(\mathbf{x}, t) = \nabla \cdot \boldsymbol{\sigma}(\mathbf{x}, t) + \mathbf{f}^{\text{ext}}, \tag{2}$$

where $\Omega^t_\epsilon \in \Omega^t$ is an infinitesimal region, ρ and $\mathbf{v}$ denote the density and velocity filed respectively, and $\mathbf{f}^{\text{ext}}$ is the external force. $\boldsymbol{\sigma}$ is the Cauchy stress tensor, usually related to a given energy function Ψ.

Marerial Point Method. Material Point Method (MPM) discretizes the continuum into a collection of Lagrangian particles. The mass conservation of each particle ensures the overall mass conservation. Following Stomakhin et al. (44), we use particle-to-grid (P2G) and grid-to-particle (G2P) to transfer properties between these particles and an Eulerian grid. The momentum conservation is ensured on the grid where the calculation is simpler and more natural. In each simulation step, the mass and momentum are first transferred from particles onto the grid. The stress tensor is used to update the grid velocities, and the velocities will be transferred back to update particle states. The velocities on the grid are updated as

$$\mathbf{v}_i^{n+1} = \mathbf{v}_i^n - \frac{\Delta t}{m_i} \cdot \mathrm{P2G}(\{\boldsymbol{\sigma}\}), \ \mathbf{v}_j^{n+1} = \mathrm{G2P}(\{\mathbf{v}_i^{n+1}\}) \tag{3}$$

where P2G is the transfer of stress from particles to the grid and G2P is the reverse transfer. n denotes the n-th simulation step, each lasts for Δt, and $\mathbf{v}_i$, $\mathbf{v}_j$ denote velocity vectors on the i-th grid node and j-th particle respectively. Gravity is ignored in our implementation. Please refer to the supplementary materials for further details on MPM.

3.2 Simulation environment setup

In this section, we first describe the process of constructing a static simulation environment (canonical 3D scene) from a dynamic surgical video. Subsequently, we detail the regularization techniques designed to enhance geometric consistency, along with the surface thickening methods used to complete the canonical scene.

Data Preparation. Given a monocular RGB surgical video $\{\mathbf{C}_o^t\}_{t=1}^T$ with T frames, we first use SAM (24) segment frames to generate tissue masks $\{\mathbf{M}^t\}_{t=1}^T$. Then we employ a video inpainting method (32) to inpaint the RGB frames on the areas covered by surgical instruments with $\{\mathbf{M}^t\}_{t=1}^T$. These inpainted images $\{\mathbf{C}^t\}_{t=1}^T$ are input to Depth Anything v2 model (55) to estimate depth $\{\hat{\mathbf{D}}^t\}_{t=1}^T$.

Canonical Scene Reconstruction. The frames and their corresponding depths from data preparation are used to build a canonical 3D scene composed of 3D Gaussians, which later serves as a simulation environment. Inspired by Wu et al. (49); Liu et al. (34), we use a deformation field $\mathbf{D}$ to model the 4D deformation of a canonical 3D Gaussian model $\mathcal{G}^0$. To achieve this, we start by initializing a coarse 3D Gaussian model using RGBD projection using the inpainted first frame and the corresponding estimated depth. Pixels in other frames will be projected into the coarse 3D Gaussian model if they belong to the mask area of all previous frames. The deformation field $\mathbf{D}$ is composed of a set of multi-resolution feature planes $\{\mathbf{V}_{ij}\} \subset \mathcal{R}^{h \times lN_i \times lN_j}$ and an MLP $\boldsymbol{\theta}$, where h is the hidden feature size, l is the upsampling scale and N is the resolution parameter. To query the deformation of a Gaussian kernel $G_k : \mu_k, \alpha_k, \Sigma_k, C_k$, we first calculate the voxel feature:

$$f_v = \bigcup_l \mathrm{lerp}(\mathbf{V}_{ij}, \mu_k), \ ij \in \{xy, xz, xt, yz, yt, zt\}. \tag{4}$$

Here lerp denotes 4-nearest bilinear interpolation. Then the feature is decoded by the MLP,

$$\Delta\mu, \Delta o, \Delta R, \Delta S = \boldsymbol{\theta}(f_v), \tag{5}$$

and the attributes of the deformed Gaussian can be computed as

$$G'_k = (\mu_k + \Delta\mu, o'_k + \Delta o, \Sigma(R_k + \Delta R, S_k + \Delta S), C_k). \tag{6}$$

Different from previous methods (49; 34), we allow the opacity to change during deformation. The opacity of tissues would change significantly when we pull it hard, unlike common objects. To render the deformed Gaussian model at a certain timestamp for optimization, we use the deformed attributes calculated above for rasterization in Eq. 1.

Multi-Stage Optimization. Our key goal is to create a surgical scene that allows for high-quality surgical simulation rather than generating high-fidelity images at specific timestamps, which all previous tissue reconstruction methods (54; 53; 34; 51) aimed for. Thus, our design focuses on maintaining the geometric consistency of the canonical Gaussian model $\mathcal{G}^0$ during the deformation process. Specifically, we design an explicit trajectory regularization strategy to prevent positional interleaving during deformation, which otherwise leads to faulty optimization of the canonical model. This regularization requires that the trajectories of Gaussian points within any small region tend to be parallel within a short period of time. For a Gaussian model $\mathcal{G}$, we first find the k-nearest neighbor point set N_i for each Gaussian G_i. For a certain timestamp t, the regularization is given by

$$\begin{aligned}\mathcal{L}_{traj} &= \sum_{G_j, G_k \in N_i} \frac{\Delta\boldsymbol{\mu}_j^t \cdot \Delta\boldsymbol{\mu}_k^t}{\|\Delta\boldsymbol{\mu}_j^t\| \cdot \|\Delta\boldsymbol{\mu}_k^t\|} \cdot \|\Delta\boldsymbol{\mu}_j^t\| \cdot \|\Delta\boldsymbol{\mu}_k^t\| \\ &= \sum_{G_j, G_k \in N_i} \Delta\boldsymbol{\mu}_j^t \cdot \Delta\boldsymbol{\mu}_k^t. \end{aligned} \tag{7}$$

where $\Delta\boldsymbol{\mu}_i^t$ means the deformation vector from the previous timestamp, i.e., $\mu_i^t - \mu_i^{t-1}$. This regularization consists of two main parts. The first part requires the deformation direction to be parallel to avoid tangential misalignment in the direction of motion. The second part requires the movement length to be small to avoid radial misalignment.

The multi-stage trajectory optimization contains three stages: 1) Optimizing only the canonical model $\mathcal{G}^0$ and the deformation field $\mathbf{D}$ (the feature planes and the MLP) to estimate the coarse trajectory. 2) Adding the trajectory regularization to refine the trajectory of all Gaussian kernels. 3) Freezing the deformation module and only optimizing the attributes of the canonical model $\mathcal{G}^0$.

Additionally, an anisotropic regularization proposed to prevent excessively large or extremely anisotropic Gaussian kernels is employed across all the stages mentioned above. It is defined as:

$$\begin{aligned}\mathcal{L}_{geo} = \sum_{i \in \{N\}} &(\mathrm{ReLU}\left(\max(S_i) - r_m\right) \\ &+ \mathrm{ReLU}\left(\max(S_i)/\min(S_i) - r_{ani}\right)) \end{aligned} \tag{8}$$

where $r_m = 1$ is the max scaling limit, and $r_{ani} = 3$ is the anisotropic factor.

Surface Thickening. Due to the limited camera view, the reconstructed canonical model $\mathcal{G}^0$ is a single surface with invalid thickness and volume for simulation. To address this, we apply a surface thickening method, pushing the Gaussian kernels in $\mathcal{G}^0$ along the z-axis with a certain probability. The thickening algorithm is described in Algo. 1.

Algorithm 1: Surface Thickening Algorithm

```
Input  : The canonical Gaussian model $\mathcal{G}^0$
Output: A thickened model for simulation $\mathcal{G}^0_{\mathrm{d}}$
Initialize $\mathcal{G}^0_{\mathrm{d}} \leftarrow \mathcal{G}^0$;
$\{x_{\mathrm{m}}, x_{\mathrm{M}}, y_{\mathrm{m}}, y_{\mathrm{M}}, z_{\mathrm{m}}, z_{\mathrm{M}}\} \leftarrow$ bounding box of $\mathcal{G}^0$;
for Layer $l$ in range(1, 1000) do
    for Each Gaussian kernel $G_i^0$ in $\mathcal{G}^0$ do
        $G_i^l \leftarrow G_i^0$;
        The position of $G_i^l$: $\mu_i^l \leftarrow \mu_i^0 \cdot (\mathrm{rand}(3) + l)/1000$;
        if $G_i^l$ in $\{x_{\mathrm{m}}, x_{\mathrm{M}}, y_{\mathrm{m}}, y_{\mathrm{M}}, z_{\mathrm{m}}, (1+0.25)z_{\mathrm{M}}\}$ then
            $\mathcal{G}^0_{\mathrm{d}} \leftarrow \mathcal{G}^0_{\mathrm{d}} \cup \{G_i^l\}$ ;
return $\mathcal{G}^0_{\mathrm{d}}$;
```

3.3 Physics-based Realistic Simulation

In this section, we first introduce how we model the complex physical effects of tissues with the Visco-Elastic model. Then we demonstrate how we use this model and input video to infer the physical parameters. For better understanding, we employ conventional notations from physics. This may lead to some repetition of the previously defined notations. These corrupted definitions are only used in this section, and their meaning will be re-defined.

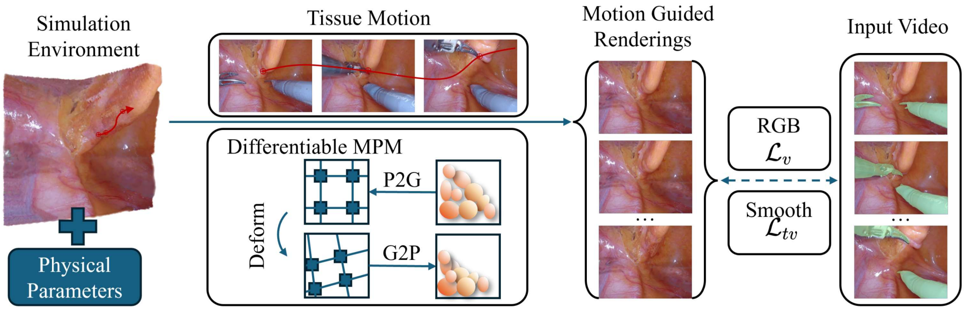

Figure 2: Illustration of our physical parameter estimation. SurgiSim automatically infers physical parameters by minimizing the discrepancy between rendered simulation results and the input video through differentiable MPM and rasterization.

Modeling Tissues with Visco-Elasticity Deformation. Previous methods (50; 60) primarily employ elastic models for simulation, which restricts their performance to jelly-like effects that oscillate back and forth. However, in surgical settings, due to the complex composition of tissues, their deformations are usually very complicated in physical behaviors, which significantly differ from these jelly-like ones. To better model these specific effects, we involve a Visco-Elastic model (42), which consists of an elastic part and a viscous part to model the complex deformations.

For the elastic part, we use fixed corotated elasticity. The energy function is defined as:

$$\Psi(\mathbf{F_E}) = \Psi(\mathbf{\Sigma_E}) = \mu_\mathbf{E} \sum_i (\sigma_{\mathbf{E},i} - 1)^2 + \frac{\lambda_\mathbf{E}}{2}(\det \mathbf{F_E} - 1)^2, \tag{9}$$

where the elastic deformation gradient $\mathbf{F_E}$ is decomposed into $\mathbf{U_E \Sigma_E V_E^\top}$ through SVD, $\sigma_{\mathbf{E},i}$ are the singular values of $\mathbf{F_E}$. $\mu_\mathbf{E} = \frac{E}{2(1+\nu)}$ and $\lambda_\mathbf{E} = \frac{E\nu}{(1+\nu)(1-2\nu)}$ are physical parameters, namely Shear modulus and Lamé modulus, computed from the Young's modulus E and Poisson's ratio ν. The Cauchy stress tensor for the elastic part is $\boldsymbol{\sigma}_\mathbf{E} = \frac{1}{\det \mathbf{F}} \frac{\partial \Psi}{\partial \mathbf{F_E}}(\mathbf{F_E})\mathbf{F_E^\top}$. The update rule of $\mathbf{F_E}$ is

$$\mathbf{F}_{\mathbf{E},j}^{n+1} = \mathbf{F}_{\mathbf{E},j}^{n}(\mathbf{I} + \Delta t \cdot \nabla \mathbf{v}_j), \tag{10}$$

and j denotes the j-th particle.

As for the viscous part, we get inspiration from Simo & Miehe (43), Johnson & Quigley (21) and propose a simple model regarding viscous energy dissipation to integrate it into the MPM. The viscous energy dissipation is described through a dissipation potential, given by

$$\Psi(\frac{\partial \mathbf{F_v}}{\partial t}) = \frac{1}{2}\eta_\mathbf{v} \operatorname{tr}\left(\left(\frac{\partial \mathbf{F_v}}{\partial t}\right)^\top \frac{\partial \mathbf{F_v}}{\partial t}\right), \tag{11}$$

where $\eta_\mathbf{v}$ is the viscosity coefficient. The gradient of the viscous deformation gradient $\mathbf{F_v}$ can be updated by $\frac{\partial \mathbf{F_v}}{\partial t} = \gamma \cdot \mathbf{D}$, and $\mathbf{D} = \frac{1}{2}(\nabla \mathbf{v}_j + \nabla \mathbf{v}_j^\top)$ is the symmetric part of the grid velocity. The viscous Cauchy tensor is $\boldsymbol{\sigma}_\mathbf{v} = \det(\mathbf{F_v}) \cdot 2\eta \mathbf{D}$, and $\mathbf{F_v}$ can be updated by

$$\mathbf{F}_{\mathbf{v},j}^{n+1} = \mathbf{F}_{\mathbf{v},j}^{n}(\mathbf{I} + \gamma_\mathbf{v} \Delta t \cdot \mathbf{D}). \tag{12}$$

The overall stress tensor $\boldsymbol{\sigma} = \boldsymbol{\sigma}_\mathbf{E} + \boldsymbol{\sigma}_\mathbf{v}$ is used to update the grid velocity.

Tissue Motion Estimation. To estimate physical parameters, we involve extracting the physical priors of tissues from surgical videos. This process comprises two main steps: 1) recovering the tissue motion induced by external forces from instruments as captured in the input video, and 2) refining physical parameters by minimizing the discrepancies between the simulated dynamics guided by the tissue motion and input video.

One straightforward way to recover motion in MPM simulation is to estimate the forces involved. However, due to uncertainties in system dynamics, accurately estimating forces is a challenging task (40). Instead, we turn to estimating the 3D trajectory of tissues. We start by selecting pixels on the tissues near the contact point with the surgical instruments and employ a 2D dense optical tracking method (25) to capture the 2D trajectories of tissue movements directly influenced by the external forces. We then augment these 2D trajectories with estimated depth values, $\{\hat{\mathbf{D}}^t\}_{t=1}^T$, to derive 3D trajectories, $\{\mathbf{p}^t\}_{t=1}^T$.

To accurately recover motion in the video, we update the velocity during MPM at time t. Specifically, we adjust the grid velocity in a small region $\mathcal{B}^0 \in \Omega^0$ around the starting point of the 3D trajectory $\mathbf{p}^0$. The velocity is updated according to the following formula:

$$\mathbf{v}_{\mathcal{B}^0}^t = \frac{p^{n_t+1} - p^{n_t}}{\Delta T}, \tag{13}$$

where ΔT is the video frame duration and n_t is the frame index that $n_t \Delta T \leq t < (n_t + 1)\Delta T$.

Physical Parameter Estimation. We then use the input video to estimate the physical parameters of tissues. This estimation is achieved by minimizing the discrepancies between frames from the input video and the simulation results driven by estimated motion at the time of the frames, as shown in Fig. 2.

Firstly, we run $\frac{t}{\Delta t}$ simulation steps to get the simulated model $\mathcal{G}_d^t$ deformed from the dense model $\mathcal{G}_d^0$, and then rasterize the model as $\hat{\mathbf{C}}_s^t$ with Eq. 1. In this way, we align the simulation results with the input video. We then optimize the parameters using

$$\mathcal{L}_v = \|\mathbf{C}_o^t - \mathbf{C}_s^t\|_1 \cdot \mathbf{M}^t. \tag{14}$$

Due to the influence of previous simulation steps on subsequent ones, we implement training in a rolling manner. Specifically, we begin by optimizing with the first k video frames. For each new round, we start from the first guide frame and add additional k frames until the guide length surpasses the video length. A parameter smoothing is conducted every k frame. For each neighbor group N_i for each Gaussian $G_{d,i}$, we apply a total variation loss:

$$\mathcal{L}_{tv} = \underset{G_{d,j} \in N_i}{\mathrm{MSE}} (\xi_j - \xi_i), \tag{15}$$

where ξ is one of the physical parameters, namely $\mu_{\mathbf{E}}$, $\eta_{\mathbf{v}}$ and $\gamma_{\mathbf{v}}$. We set $\nu_{\mathbf{E}}$ to a constant value of 0.45, because we have found that variations within the valid range have little impact on the results.

4 Experiments

4.1 Implementation Details.

Dataset. We evaluate our method on the EndoNeRF dataset (47), which comprises several surgical video clips totaling 807 frames. Each clip, captured by stereo cameras from a single viewpoint, spans 4-8 seconds and shows typical soft tissue scenarios encountered in robotic surgery, including complex non-rigid deformations. We utilize 5 of these clips to establish our simulation environments, excluding one clip that involves significant tissue cutting. Though EndoNeRF is a binocular dataset, we use only the left image and no binocular depth. Please refer to supplement material for demos on more clips beyond EndoNeRF.

Environment and Simulation Setup. Our environment setup employs a multi-stage optimization process consisting of 5000 iterations. The initial stage, comprising the first 50% of iterations, focuses on optimizing the canonical model $\mathcal{G}^0$ and the deformation field $\mathbf{D}$. This is followed by a trajectory regularization stage from 50% to 70% of the total iterations. A refinement period then occurs from 70% to 90%, allowing for further optimization of the deformation field after its correction by trajectory optimization. In the final stage, occupying the last 10% of iterations, we freeze the deformation field and fine-tune the canonical model. For the simulation, we adapt the MPM framework from Xie et al. (50); Zong et al. (62). The thickened model $\mathcal{G}_{\mathrm{d}}^0$ is normalized into a 2-unit cube, overlaid with an Eulerian grid of resolution $50 \times 50 \times 50$. We established a conversion rate where 10,000 simulation steps correspond to one second of video at 25 fps. For each operation in the simulation, we conduct 80k simulation steps to form an 8-second video. To perform a simulation operation, physical parameters estimated from previous simulations are loaded onto the simulation scene for initialization.

User Study. To evaluate the fidelity of our simulations, we conducted a user study involving 68 participants including both surgeons and laypersons. For More details, please refer to the supplementary material.

4.2 Comparisons on Surgical Simulations

We first provide qualitative comparisons on surgical simulation in Fig. 3. We primarily compare with three simulation baselines: 1) A baseline using our simulation environment and native PhysGaussian (50) as simulation

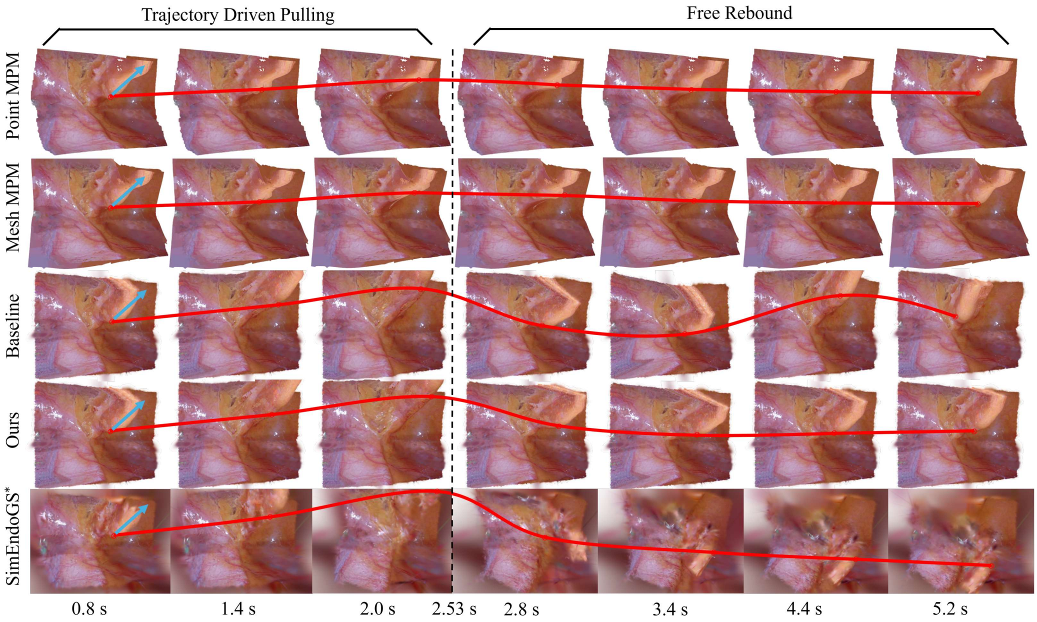

Figure 3: Visualization of simulations. We show the trajectory direction with a blue arrow and the motion of the tissues with a red line. The external force caused by the driving trajectory ends at 2.52s (63 frames), after which the tissue rebounds freely. SurgiSim consistently produces the most realistic simulation dynamics. Please refer to the supplementary material for more simulation results.

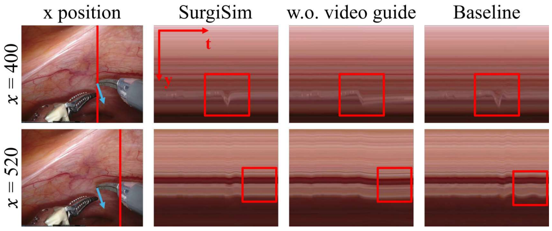

Figure 4: Y-T slices of simulation dynamics. The slices at $x = 400$ capture the motion of tissues being lifted and then released. The slices at $x = 520$ capture the oscillations after the rebound.

backend, without our Visco-Elastic model and estimated parameters; 2) A pointcloud-based MPM approach, *Pcd* for short; 3) A triangle mesh simulation with the same MPM, *Mesh* for short. The MPM is built on the string-mass model from Taichi (17) framework. All methods share the same manipulation during simulation. In real world, due to the viscous properties of tissues (42), under this manipulation, the tissue will quickly return to a stationary state rather than oscillating back and forth. The simulation results demonstrate that SurgiSim achieves superior simulation results compared to baseline in both visual quality and realism. Our approach generates more realistic tissue deformation responses and interaction dynamics, closely mimicking the behavior observed in actual surgical scenarios. In contrast, the baseline often produces unrealistic elastic behavior that deviates from true tissue properties. Pcd and Mesh suffer from fragmentation, artifacts and unnatural motion during simulation. We also provide results of SimEndoGS*, our own implementation of a SOTA method SimEndoGS (58), which is not yet open-source. However, it does not perform proper simulation scene preparation as we do in Sec. 3.2, which causes it to fail under large movement in all cases for evaluation, and thus we left this method for further quantitative comparison.

Besides, we show Y-T slices across all simulation steps in one case in Fig. 4. The epithelial tissue is lifted and then released. Results show that our simulated tissue rapidly reverts to a quiescent state following a quick rebound. Conversely, results from baseline show that the tissue continues to exhibit substantial oscillations post-rebound.

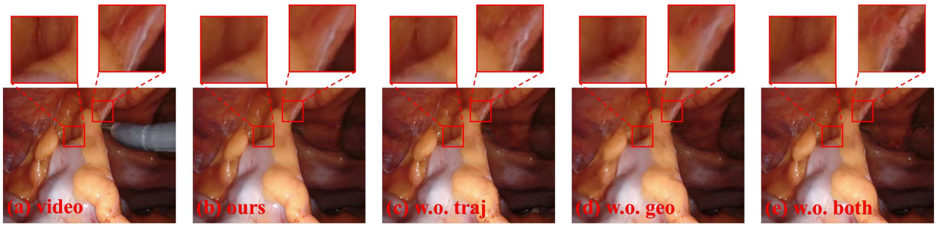

Figure 5: Ablation on the trajectory and geometric regulation. (a) is the input video as a reference. (b) is the result of multi-stage optimization. (c), (d) and (e) are the results of optimization without trajectory regularization, geometric regularization, and both, respectively.

This discrepancy is attributed to the proposed Visco-Elastic model which effectively damps out oscillations by simulating the inherent viscoelastic properties of the tissue.

We further provide quantitative comparisons in Tab. 1. We reproduce the operation in the input videos and compare the render results with the ground truth. Given the same manipulation, our method shows the best reproduction quality for our physics model behaves more like real tissue, driven by external forces.

Metric	Ours	Baseline	Pcd	Mesh
PSNR↑	**22.455**	<u>22.367</u>	20.446	21.102
SSIM↑	**0.7315**	<u>0.7235</u>	0.7068	0.6869
LPIPS↓	**0.1735**	<u>0.1759</u>	0.2295	0.2819

Table 1: Quantitative results in terms of simulation quality.

The user study results are shown in Tab. 2. SurgiSim receives a significant preference across both groups, with a 77.8% preference rate among ordinary viewers and 57.8% among surgeons, significantly outperforming baseline methods. Interestingly, surgeons show a more balanced evaluation, suggesting their professionalism for technical nuances, yet still strongly favored SurgiSim.

Additionally, we provide the render quality of our canonical model compared with SOTA dynamic tissue reconstruction methods (47; 47; 54; 34) on EndoNeRF in the supplementary material. Note that our canonical model is for simulation environment preparation instead of reconstruction.

Group	Ours	Baseline	Pcd	Mesh	Ours	w.o. Guide
S	**57.8%**	25.6%	10.0%	6.7%	**65.6%**	34.4%
O	**77.8%**	21.3%	0.5%	0.4%	**82.9%**	17.1%

Table 2: **User Study**. S stands for surgeon and O for ordinary. The values are the mean percentages of each choice.

4.3 Ablation and Analysis

Multi-Stage Optimization. In the multi-stage optimization, we employed trajectory regularization and geometric regularization to enhance the geometric consistency. To assess their individual contributions, we conduct ablation studies, and results are shown in Fig. 5. The result without trajectory regularization (c) achieves overall sound quality, but the surfaces and edges of tissues are rough due to the lack of geometric consistency. The result without geometric regularization (d) contains Gaussian kernels with severe anisotropy, which results in burrs on the surface. These spiked kernels would result in unrealistic protrusions during surgical simulations.

We also provide a quantitative comparison in Tab. 3. Since the canonical model corresponds to the initial timestamp, rendering metrics are calculated over the first 25 frames. Without geometric regularization, the quality of results is hindered by burrs on the surface. Omitting trajectory regularization doesn't affect the metrics much. As previously discussed, this regularization primarily influences the geometry of the canonical model rather than render quality.

Physical Parameter Estimation. Fig. 4 shows the results of SurgiSim without video-guided parameter estimation. While the tissue can rebound quickly because of good elastic parameters in the result of SurgiSim, the tissue result without estimation would rebound very slowly with severe damping due to faulty viscous parameters. We

Metric	Ours	w.o. traj	w.o. geo	w.o. both
PSNR↑	**37.114** ± 3.7019	37.036 ± 4.3512	36.582 ± 3.6153	36.750 ± 3.6725
SSIM↑	**0.9772** ± 0.0147	0.9770 ± 0.0191	0.9730 ± 0.0143	0.9754 ± 0.0155
LPIPS↓	**0.0436** ± 0.0245	0.0442 ± 0.0359	0.0505 ± 0.0230	0.0465 ± 0.0268

Table 3: Quantitative results of quality of simulation environment.

also report the participants' preferences on the results with and without physical parameter estimation in Table 2. In both groups, participants show an obvious preference for the results with the inferred physical parameters.

5 Conclusion

In this paper, we present SurgiSim, an automated and flexible method for transforming monocular surgical videos into simulation-ready scenes, and performing realistic physics surgical simulations within these environments. SurgiSim employs multi-stage optimization with trajectory and anisotropic regularization to construct a geometrically consistent simulation environment. By incorporating a Visco-Elastic deformation model and precise physical parameters estimated from real videos, SurgiSim shows highly realistic tissue deformations during simulation. We hope that SurgiSim could contribute to the development of more diverse and realistic surgical simulations.

References

[1] Aayush Bansal, Minh Vo, Yaser Sheikh, Deva Ramanan, and Srinivasa Narasimhan. 4d visualization of dynamic events from unconstrained multi-view videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5366–5375, 2020.

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127, 2023.

[3] Javier Bonet and Richard D Wood. Nonlinear continuum mechanics for finite element analysis. Cambridge university press, 1997.

[4] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 130–141, 2023.

[5] Joseph Cho, Samuel Schmidgall, Cyril Zakka, Mrudang Mathur, Rohan Shad, and William Hiesinger. Surgen: Text-guided diffusion model for surgical video generation. arXiv preprint arXiv:2408.14028, 2024.

[6] Steven A Cover, Norberto F Ezquerra, James F O'Brien, Richard Rowe, Thomas Gadacz, and Ellen Palm. Interactively deformable models for surgery simulation. IEEE Computer Graphics and Applications, 13(6): 68–75, 1993.

[7] Gilles Daviet and Florence Bertails-Descoubes. A semi-implicit material point method for the continuum simulation of granular materials. ACM Transactions on Graphics (TOG), 35(4):1–13, 2016.

[8] F Boux De Casson and Christian Laugier. Modeling the dynamics of a human liver for a minimally invasive surgery simulator. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 1156–1165. Springer, 1999.

[9] Alban De Vaucorbeil, Vinh Phu Nguyen, Sina Sinaie, and Jian Ying Wu. Material point method after 25 years: theory, implementation, and applications. Advances in applied mechanics, 53:185–398, 2020.

[10] Youkou Dong and Jürgen Grabe. Large scale parallelisation of the material point method with multiple gpus. Computers and Geotechnics, 101:149–158, 2018.

[11] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12479–12488, 2023.

[12] Chuyuan Fu, Qi Guo, Theodore Gast, Chenfanfu Jiang, and Joseph Teran. A polynomial particle-in-cell method. ACM Transactions on Graphics (TOG), 36(6):1–12, 2017.

[13] Sarah Gibson, Joe Samosky, Andrew Mor, Christina Fyock, Eric Grimson, Takeo Kanade, Ron Kikinis, Hugh Lauer, Neil McKenzie, Shin Nakajima, et al. Simulating arthroscopic knee surgery using volumetric object representations, real-time volume rendering and haptic feedback. In International Conference on Computer Vision, Virtual Reality, and Robotics in Medicine, pp. 367–378. Springer, 1997.

[14] X. Han, B.B. Bai, C.J. Wang, S. Zhao, and Y. Chen. Risk factors for recurrent thrombosis in patients with polycythemia vera and essential thrombocythemia. Zhonghua xue ye xue za zhi = Zhonghua xueyexue zazhi, 40(1):17 – 23, 2019. Cited by: 1.

[15] Ying Hu, Xue Yan, Yun Shen, Mingxiao Di, and Jun Wang. Antibiotics in surface water and sediments from hanjiang river, central china: Occurrence, behavior and risk assessment. Ecotoxicology and Environmental Safety, 157:150–158, 2018. ISSN 0147-6513.

[16] Yuanming Hu, Yu Fang, Ziheng Ge, Ziyin Qu, Yixin Zhu, Andre Pradhana, and Chenfanfu Jiang. A moving least squares material point method with displacement discontinuity and two-way rigid body coupling. ACM Transactions on Graphics (TOG), 37(4):1–14, 2018.

[17] Yuanming Hu, Tzu-Mao Li, Luke Anderson, Jonathan Ragan-Kelley, and Frédo Durand. Taichi: a language for high-performance computation on spatially sparse data structures. ACM Transactions on Graphics (TOG), 38(6):201, 2019.

[18] Tao Huang, Kai Chen, Bin Li, Yun-Hui Liu, and Qi Dou. Guided reinforcement learning with efficient exploration for task automation of surgical robot. In 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 4640–4647. IEEE, 2023.

[19] Yiming Huang, Beilei Cui, Long Bai, Ziqi Guo, Mengya Xu, Mobarakol Islam, and Hongliang Ren. Endo-4dgs: Endoscopic monocular scene reconstruction with 4d gaussian splatting. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 197–207. Springer, 2024.

[20] Chenfanfu Jiang, Theodore Gast, and Joseph Teran. Anisotropic elastoplasticity for cloth, knit and hair frictional contact. ACM Transactions on Graphics, 36(4), 2017. Cited by: 130; All Open Access, Bronze Open Access.

[21] AR Johnson and CJ Quigley. A viscohyperelastic maxwell model for rubber viscoelasticity. Rubber chemistry and technology, 65(1):137–153, 1992.

[22] Erwin Keeve, Sabine Girod, and Bernd Girod. Craniofacial surgery simulation. In International Conference on Visualization in Biomedical Computing, pp. 541–546. Springer, 1996.

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph., 42(4):139–1, 2023.

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026, 2023.

[25] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. Dense optical tracking: connecting the dots. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19187–19197, 2024.

[26] Heather Lesch, Evan Johnson, Jörg Peters, and Juan C Cendán. Vr simulation leads to enhanced procedural confidence for surgical trainees. Journal of surgical education, 77(1):213–218, 2020.

[27] Chenxin Li, Hengyu Liu, Yifan Liu, Brandon Y Feng, Wuyang Li, Xinyu Liu, Zhen Chen, Jing Shao, and Yixuan Yuan. Endora: Video generation models as endoscopy simulators. arXiv preprint arXiv:2403.11050, 2024.

[28] Liang Li and Tingting Li. Animation of virtual medical system under the background of virtual reality technology. Computational Intelligence, 38(1):88–105, 2022.

[29] Qian Li, Shuojue Yang, Daiyun Shen, and Yueming Jin. Free-dygs: Camera-pose-free scene reconstruction based on gaussian splatting for dynamic surgical videos. arXiv preprint arXiv:2409.01003, 2024.

[30] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5521–5531, 2022.

[31] Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8508–8520, 2024.

[32] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 17562–17571, 2022.

[33] Fangfu Liu, Hanyang Wang, Shunyu Yao, Shengjun Zhang, Jie Zhou, and Yueqi Duan. Physics3d: Learning physical properties of 3d gaussians via video diffusion. arXiv preprint arXiv:2406.04338, 2024.

[34] Yifan Liu, Chenxin Li, Chen Yang, and Yixuan Yuan. Endogaussian: Gaussian splatting for deformable surgical scene reconstruction. arXiv preprint arXiv:2401.12561, 2024.

[35] Yonghao Long, Wang Wei, Tao Huang, Yuehao Wang, and Qi Dou. Human-in-the-loop embodied intelligence with interactive simulation environment for surgical robot learning. IEEE Robotics and Automation Letters, 8(8):4441–4448, 2023.

[36] Zhicheng Lu, Xiang Guo, Le Hui, Tianrui Chen, Min Yang, Xiao Tang, Feng Zhu, and Yuchao Dai. 3d geometry-aware deformable gaussian splatting for dynamic view synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8900–8910, 2024.

[37] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. arXiv preprint arXiv:2308.09713, 2023.

[38] Ullrich Meier, Oscar López, Carlos Monserrat, M-Carmen Juan, and M Alcaniz. Real-time deformable models for surgery simulation: a survey. Computer methods and programs in biomedicine, 77(3):183–197, 2005.

[39] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1): 99–106, 2021.

[40] Guangzhu Peng, Chenguang Yang, Wei He, and CL Philip Chen. Force sensorless admittance control with neural learning for robots with actuator saturation. IEEE Transactions on Industrial Electronics, 67(4): 3138–3148, 2019.

[41] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988, 2022.

[42] Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Nr-slam: Non-rigid monocular slam. IEEE Transactions on Robotics, 2024.

[43] Juan C Simo and Christian Miehe. Associative coupled thermoplasticity at finite strains: Formulation, numerical analysis and implementation. Computer Methods in Applied Mechanics and Engineering, 98(1): 41–104, 1992.

[44] Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, and Andrew Selle. A material point method for snow simulation. ACM Transactions on Graphics (TOG), 32(4):1–10, 2013.

[45] Weixiang Sun, Xiaocao You, Ruizhe Zheng, Zhengqing Yuan, Xiang Li, Lifang He, Quanzheng Li, and Lichao Sun. Bora: Biomedical generalist video generation model. arXiv preprint arXiv:2407.08944, 2024.

[46] Kailing Wang, Chen Yang, Yuehao Wang, Sikuang Li, Yan Wang, Qi Dou, Xiaokang Yang, and Wei Shen. Endogslam: Real-time dense reconstruction and tracking in endoscopic surgeries using gaussian splatting, 2024. URL `https://arxiv.org/abs/2403.15124`.

[47] Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In International conference on medical image computing and computer-assisted intervention, pp. 431–441. Springer, 2022.

[48] Joshuah Wolper, Yu Fang, Minchen Li, Jiecong Lu, Ming Gao, and Chenfanfu Jiang. Cd-mpm: Continuum damage material point methods for dynamic fracture animation. ACM Transactions on Graphics, 38(4), 2019. Cited by: 70; All Open Access, Bronze Open Access.

[49] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20310–20320, 2024.

[50] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4389–4398, 2024.

[51] Weixing Xie, Junfeng Yao, Xianpeng Cao, Qiqin Lin, Zerui Tang, Xiao Dong, and Xiaohu Guo. Surgicalgaussian: Deformable 3d gaussians for high-fidelity surgical scene reconstruction. arXiv preprint arXiv:2407.05023, 2024.

[52] Mengya Xu, Ziqi Guo, An Wang, Long Bai, and Hongliang Ren. A review of 3d reconstruction techniques for deformable tissues in robotic surgery. arXiv preprint arXiv:2408.04426, 2024.

[53] Chen Yang, Kailing Wang, Yuehao Wang, Qi Dou, Xiaokang Yang, and Wei Shen. Efficient deformable tissue reconstruction via orthogonal neural plane. arXiv preprint arXiv:2312.15253, 2023.

[54] Chen Yang, Kailing Wang, Yuehao Wang, Xiaokang Yang, and Wei Shen. Neural lerplane representations for fast 4d reconstruction of deformable tissues. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 46–56. Springer, 2023.

[55] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024. URL `https://arxiv.org/abs/2406.09414`.

[56] Shuojue Yang, Qian Li, Daiyun Shen, Bingchen Gong, Qi Dou, and Yueming Jin. Deform3dgs: Flexible deformation for fast surgical scene reconstruction with gaussian splatting. arXiv preprint arXiv:2405.17835, 2024.

[57] Zeyu Yang, Hongye Yang, Zijie Pan, Xiatian Zhu, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. arXiv preprint arXiv:2310.10642, 2023.

[58] Zhenya Yang, Kai Chen, Yonghao Long, and Qi Dou. Efficient data-driven scene simulation using robotic surgery videos via physics-embedded 3d gaussians. arXiv preprint arXiv:2405.00956, 2024.

[59] Ruyi Zha, Xuelian Cheng, Hongdong Li, Mehrtash Harandi, and Zongyuan Ge. Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In International conference on medical image computing and computer-assisted intervention, pp. 13–23. Springer, 2023.

[60] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. arXiv preprint arXiv:2404.13026, 2024.

[61] Lingting Zhu, Zhao Wang, Zhenchao Jin, Guying Lin, and Lequan Yu. Deformable endoscopic tissues reconstruction with gaussian splatting. arXiv preprint arXiv:2401.11535, 2024.

[62] Zeshun Zong, Xuan Li, Minchen Li, Maurizio M Chiaramonte, Wojciech Matusik, Eitan Grinspun, Kevin Carlberg, Chenfanfu Jiang, and Peter Yichen Chen. Neural stress fields for reduced-order elastoplasticity and fracture. In SIGGRAPH Asia 2023 Conference Papers, pp. 1–11, 2023.

A Overview

The contents of this appendix include:

1. Attachment Descriptions (Sec. B).
2. Details of Material Point Method (Sec. C).
3. More Experimental Results (Sec. D).
4. More Implementation Details (Sec. E).

B Attachment Descriptions

We strongly suggest reviewing our attached **HTML page**, which contains the following materials:

1. **More Visual Results:** The video page includes demo videos showing different surgical operations across various real surgical scenes. We also provide the complete videos referenced in Fig. 3 of the paper, alongside quantitative comparisons between four methods (Pcd, Mesh, Baseline, SurgiSim) and ablation results without Video Guide.
2. **User Study Page:** The original interface used in our user study, containing 9 sets of videos for qualitative comparison and Video Guide ablation analysis.

Please click on 'index.html' and select 'Videos' to view all attachments.

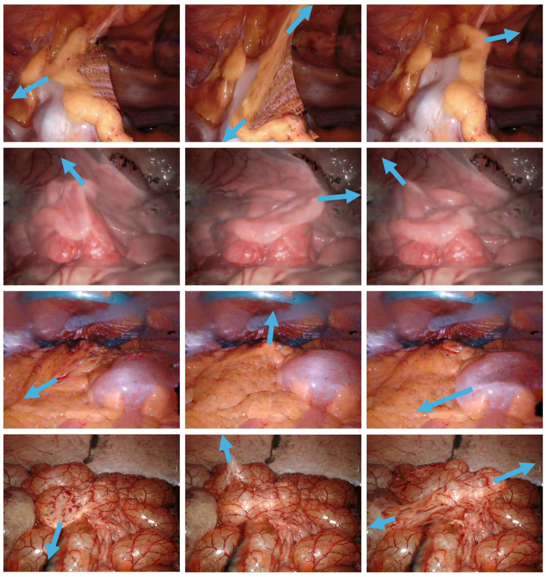

Figure 6: Visualization of more simulation cases on different scenes using SurgiSim.

C Details of Material Point Method

The full MPM methods include particle-to-grid (P2G) and grid-to-particle (G2P) to transfer properties between these particles and an Eulerian grid. Following Stomakhin et al. (44); Xie et al. (50), we use C^1 continuous B-spline kernels for two-way transfer. The mass and momentum are transferred from particles to grid nodes:

$$m_i = \sum_p m_p \, w_{ip}, \tag{16}$$

Metric	SurgiSim	EndoNeRF	EndoSurf	LerPlane	4D-GS	EndoGaussian
PSNR↑	35.490	27.077	34.795	34.643	22.832	36.31
SSIM↑	0.966	0.900	0.945	0.922	0.827	0.971
LPIPS↓	0.056	0.107	0.119	0.072	0.368	0.050

Table 4: Quantitative comparison of reconstruction quality on the EndoNeRF dataset. While our method is not specifically designed for novel-timestamp synthesis, it achieves competitive performance across all metrics.

$$m_i \mathbf{v}_i = \sum_p m_p \left(\mathbf{v}_p + \mathbf{C}_p (\mathbf{x}_i - \mathbf{x}_p) \right) w_{ip}, \tag{17}$$

where m_i is the mass at grid node i, $m_p \mathbf{v}_p$ are the mass and velocity of particle p and $\mathbf{v}_i$ is the velocity at grid node i. $\mathbf{x}_i$ and $\mathbf{x}_p$ are the positions of grid node i and particle p, respectively, w_{ip} is the B-spline weighting function between particle p and grid node i and $\mathbf{C}_p$ is the affine velocity matrix of particle p, capturing local velocity gradients.

After grid velocities are updated, particle velocities and affine matrices are interpolated from the grid:

$$\mathbf{v}_p^{n+1} = \sum_i w_{ip} \mathbf{v}_i^{n+1}, \tag{18}$$

$$\mathbf{C}_p^{n+1} = \frac{12}{\Delta x^2 (b+1)} \sum_i w_{ip} \mathbf{v}_i^{n+1} (\mathbf{x}_i - \mathbf{x}_p)^\top, \tag{19}$$

where $\mathbf{v}_p^{n+1}$ is the updated velocity of particle p, $\mathbf{v}_i^{n+1}$ is the updated velocity at grid node i, $\mathbf{C}_p^{n+1}$ is the updated affine velocity matrix for particle p, Δx is the grid spacing and The term $(\mathbf{x}_i - \mathbf{x}_p)^T$ represents the transpose of the position difference vector.

D More Experimental Results

D.1 Visualization of Surgical Simulation

Fig. 6 presents additional simulation cases, performing different operations on each of the different surgical scenes. The results demonstrate SurgiSim's advanced capability to adapt to diverse new scenes and perform various kinds of operations, including severe pulling and cutting. See the videos for Fig. 6 and more demos in index.html.

D.2 Reconstruction Quality Analysis

We evaluate SurgiSim's reconstruction capabilities against state-of-the-art dynamic tissue reconstruction methods on the EndoNeRF dataset (47). For a fair comparison following the protocol in (52), we utilize the original dataset masks and stereo depth information rather than our SAM-refined masks and estimated monocular depth.

Tab. 4 presents quantitative results comparing SurgiSim with EndoNeRF (47), EndoSurf (59), LerPlane (54), 4D-GS (49), and EndoGaussian (34). While these metrics primarily evaluate novel-timestamp synthesis capabilities—which is not the primary objective of SurgiSim—our method still achieves compelling results, consistently ranking second best across all metrics. This strong reconstruction performance, achieved as a byproduct of our focus on creating realistic surgical simulation environments, demonstrates SurgiSim's capability to accurately model tissue deformations across video frames, proving its powerful ability to extract high-quality canonical scenes from dynamic inputs.

E More Implementation Details

E.1 Training and Performance

The training is per scene. For each input video, it takes less than 4 minutes to build a canonical scene from the input video with Surface Thickening (Sec. 3.2) done. For physical parameter estimation, because we train in a rolling manner (Sec. 3.3), the time complexity concerning the length of the operation is $O(n^2)$. The mean optimization time is 10 minutes, but the time can reach 20 minutes for long sequences. When performing novel simulations after optimization, the simulation speed can reach 7 fps by setting the simulation step duration 10 times longer for fewer steps.

All experiments were conducted on a machine equipped with a Core i7-13700K CPU and a single NVIDIA RTX 4090 GPU, running Ubuntu 24.04. The code will be made public to promote the virtual surgery.

E.2 Details on User Study

To evaluate the fidelity of our simulations, we conducted a user study involving 68 participants including both surgeons and laypersons. These participants were categorized into two distinct groups: the Surgeons group, consisting of 44 board-certificated surgeons, and the Ordinary group, comprising 24 laypersons with no medical background.

The study was structured into two parts. In the first part, participants were tasked with selecting one most realistic simulation from four options, results from four different methods. In the second part, they were required to choose one superior simulation between the two presented results. Each participant was exposed to nine sets of simulation results, with the order of the sets randomized to prevent order bias. Before analysis, we executed a basic data-cleaning process to remove any invalid or outlier responses.

We provide the page used for our user study in the supplement materials. To view the page, just click on the ‘index.html’ and select ‘User Study’.

E.3 Future Work

In the simulation environment setup, our method struggled to handle the invisible portions on the sides of the tissues, leading to imperfections in the texture generated by our thickening approach. Similarly, the texture on the exposed areas after cutting still lacked sufficient realism and required manual correction. In the future, we hope to use diffusion or the large reconstruction model to fix these textures. Our method does not yet support more complicated operations like topological inversions of structure. In the future, we hope to build a geometry-aware MPM method based on 3D scene understanding techniques.