Realistic Surgical Image Dataset Generation Based On 3D Gaussian Splatting

Tianle Zeng[1 [0009−0004−1659−6643]], Gerardo Loza Galindo[1 [0000−0003−2841−0506]], Junlei Hu[1 [0000−0001−7394−5580]], Pietro Valdastri[1 [0000−0002−2280−5438]], and Dominic Jones[1 [0000−0002−2961−8483]]

University of Leeds, Leeds, UK

Abstract. Computer vision technologies markedly enhance the automation capabilities of robotic-assisted minimally invasive surgery (RAMIS) through advanced tool tracking, detection, and localization. However, the limited availability of comprehensive surgical datasets for training represents a significant challenge in this field. This research introduces a novel method that employs 3D Gaussian Splatting to generate synthetic surgical datasets. We propose a method for extracting and combining 3D Gaussian representations of surgical instruments and background operating environments, transforming and combining them to generate high-fidelity synthetic surgical scenarios. We developed a data recording system capable of acquiring images alongside tool and camera poses in a surgical scene. Using this pose data, we synthetically replicate the scene, thereby enabling direct comparisons of the synthetic image quality (27.796±1.796 PSNR). As a further validation, we compared two YOLOv5 models trained on the synthetic and real data, respectively, and assessed their performance in an unseen real-world test dataset. Comparing the performances, we observe an improvement in neural network performance, with the synthetic-trained model outperforming the real-world trained model by 12%, testing both on real-world data.

Keywords: 3D Reconstruction · 3D Gaussian Splatting · Medical Imaging Processing.

1 Introduction

Detecting and tracking surgical instruments are crucial data sources used in the automation of Robotic-Assisted Minimally Invasive Surgery (RAMIS), providing essential proprioceptive data to the robot [13, 8, 11]. The robot's forward kinematics give a general measure of tool pose; however, the inherent compliance in cable-driven surgical robotic mechanisms, designed to ensure surgical safety and adaptability, can introduce positional inaccuracies, complicating the precise tracking and detection of instrument end-effectors [2]. Computer vision technologies offer a solution to this inaccuracy; however, the lack of high-quality labeled data available in surgical settings often complicates the training and supervision of learning-based methods.

Previous studies [3, 5] have explored creating artificial surgical images from real-world images to combat data scarcity. Using game engines [16, 21] offers a scalable, noise-free solution but fails to accurately replicate real surface properties and textures. Generative neural networks can produce fully synthetic datasets of surgical scenes, including tools, by training on surgical environments [1, 7, 9, 15, 6]. However, this method has limitations: the position and pose of instruments in generated images are fixed and uneditable, it lacks scalability as different scenes require retraining the whole network, and annotating datasets remains a time-consuming and complex task with few methods providing corresponding annotation information.

Neural Radiance Fields (NeRF) [12] construct an implicit 3D model of a scene from photographs taken at known positions, enabling the rendering of 2D images from new, unseen viewpoints, thus generating diverse image datasets. EndoNeRF [22] first applied NeRF in surgery, removing instruments from dynamic videos of soft tissue manipulation to reveal unobstructed tissue images. Psychogyios et al. [17] trained a light source location-conditioned NeRF to encapsulate a colon sequence's 3D and color information, generating new image datasets. NeRF-based methods surpass generative neural networks in image quality and dataset diversity by producing 2D images from various viewpoints. However, images generated by these methods do not include surgical instruments, as their removal is a prerequisite for operation. Consequently, such outputs are unsuitable for direct use in neural network training.

The 3D Gaussian Splatting [10] method advances the concept of scene representation and rendering by offering a novel approach to modeling 3D scenes from a collection of images. This method emphasizes explicit representation and high-quality real-time rendering, allowing for generating detailed and photo-realistic images from new viewpoints [4]. The high-quality image rendering, explicit scene representation, and rapid training times of 3D Gaussian Splatting position it as a potential method to overcome the drawbacks of image dataset generation methods based on NeRF.

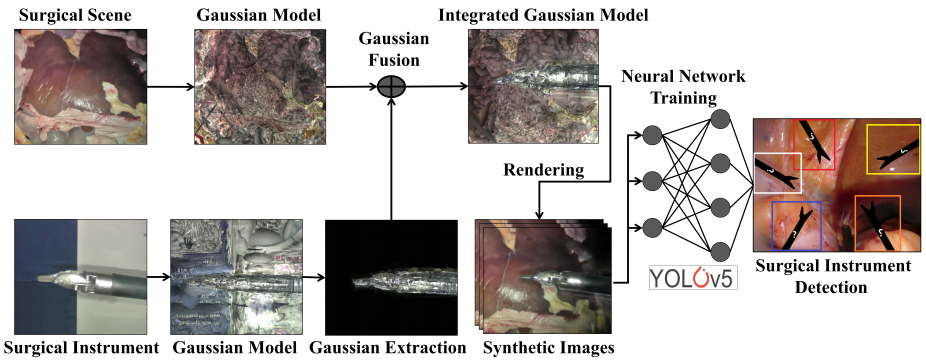

Fig. 1. Pipeline of proposed method.

Our approach leverages 3D Gaussian Splatting to produce a novel method that enables the creation of image datasets featuring various surgical instruments across different surgical scenes. The method offers capabilities for scene editing and novel scene synthesis to enhance dataset diversity. Our key contributions are as follows: 1) We are the first to apply 3D Gaussian Splatting for medical dataset generation, offering a new methodological route for creating a surgical image dataset. 2) We developed a technique for precise editing of 3D Gaussian models, allowing independent training and a flexible combination of surgical scenes and instrument models. 3) The proposed method can automatically generate accurate annotation information alongside image datasets. 4) We demonstrate the high quality of the synthetic image datasets produced by our method and their potential for neural network training.

2 Methods

As shown in Figure 1, our method first requires a set of images of the surgical scene and surgical instruments with known camera intrinsic and extrinsic parameters, obtainable through tracking the camera's motion path or using structure from motion (SFM) methods like COLMAP [19]. Then, we separately train 3D Gaussian representations for the surgical scene and instruments. Next, we extract the Gaussian representation of the surgical instruments from the background and perform necessary edits, such as translation and rotation. Following this, we fuse the instrument's Gaussian representation with that of the surgical scene, resulting in a scene that includes the surgical instruments. Utilizing the fused Gaussian scene enables the rendering of 2D images with varying poses from multiple viewpoints, facilitating the creation of an image dataset for neural network training.

2.1 Preliminary: Gaussian Splatting

3D Gaussian Splatting [10] is a technique for representing static 3D scenes, distinguished by its differentiability and the ease with which it can be projected into 2D splats. This feature enables efficient α-blending for rapid image rendering. The 3D scenes are represented by a collection of 3D Gaussians defined by a mean μ and covariance matrix $\mathbf{\Sigma}$, described in the equation:

$$\mathrm{G}(x) = \frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}\frac{1}{2}(x-\mu)} \tag{1}$$

Where Σ is decomposed into rotation matrix R and scaling matrix S, writing as $\mathbf{\Sigma} = \mathbf{RSS^TR^T}$. The 3D Gaussians are enhanced with opacity and spherical harmonic (SH) coefficients for color representation, enabling the depiction of anisotropic appearances. These Gaussians encapsulate the 3D spatial information of scenes through learned attributes, which are refined during the training process. Gaussian density control step is also implemented to interleave these Gaussians effectively. Our work is based on 3D Gaussian Splatting, wherein we

initially train two distinct Gaussian models: one for the surgical scene and another for the surgical instrument.

2.2 Gaussian Extraction and Labelling

After training, we obtain Gaussian models for both the surgical scene and instruments, with the instrument model capturing both the tool and its background. For Gaussian model fusion, we aim to isolate and use only those Gaussians representing the instruments, necessitating a method to segment these from the background representation. In our instrument Gaussian model, the tool Gaussians are densely centered with sparse background Gaussians distributed in the distance. This allows for extraction by selecting center-distributed Gaussians and filtering out others.

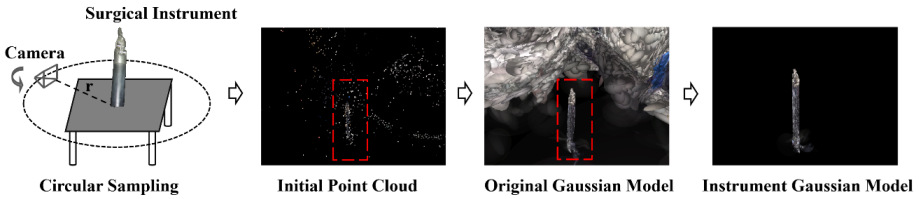

Fig. 2. Using circular sampling, we center the surgical tool in the scene, resulting in a dense distribution at the center (highlighted by a red rectangle).

Labeling Gaussian models representing surgical instruments is crucial for the subsequent fusion process, where adjustments to the Gaussians' orientations and positions are required. Although conceptually simple, this labeling necessitates significant changes to the program's data structures, making it complex and labor-intensive. An efficient alternative uses existing Gaussian properties, notably the color difference between instruments and scenes, as natural labels. In Gaussians, color is represented using the Spherical Harmonics (SH) function [10]. Therefore, instruments can be labeled by tagging the SH function's direct current component within the Gaussians, generating a secondary tool representation solely for data labeling.

2.3 Gaussian Scene Fusion

Once we have obtained 3D Gaussian representations of the surgical scene and the segmented surgical instruments, we can then combine the two to produce a fully synthetic scenario with a specified tool pose. The representation of 3D Gaussians is independent; as long as the properties of an individual Gaussian are not altered, adding new Gaussians to a Gaussian model does not affect its representation [10]. Therefore, we can achieve Gaussian scene fusion by incorporating the extracted surgical instrument into the surgical scene. However, the

fused model at this stage is not yet suitable for generating synthetic images. For each synthetic image, we must transform the instrument Gaussian to match its intended location with respect to the camera and background. 3D Gaussians are represented by their mean μ and covariance $\mathbf{\Sigma}$, indicating their position and orientation [10], which can be adjusted to correct the Gaussians' pose in the fused scene as:

$$\begin{aligned} &G'(x) = \frac{1}{(2\pi)^{3/2}|\Sigma I|^{1/2}} e^{-\frac{1}{2}\left(x-\mu'\right)^T \Sigma'^{-1}\left(x-\mu'\right)} \\ &where: \mu' = \mu + \Delta\mu; \Sigma' = R\Sigma R^T \end{aligned} \tag{2}$$

Where $\Delta\mu$ denotes the translation of Gaussians, R denotes the rotation of Gaussians. By editing the fused Gaussians, we can generate synthetic images in arbitrary orientations and positions of the camera and instrument.

2.4 Automatic Annotation Generation

As an extension of the fused Gaussian representation, we can generate a pixel-wise segmentation mask that clearly delineates tool boundaries in synthetic images. By exclusively rendering the tool Gaussian, labeled accordingly, we utilize the differential Gaussian rasterization pipeline introduced by [10]. These 3D Gaussians are projected into 2D using the covariance matrix Σ':

$$\Sigma' = JW\Sigma W^T J^T, \tag{3}$$

where J is the Jacobian from the affine approximation of the projective transformation, W is the view matrix for world-to-camera coordinates, and Σ is the 3D covariance matrix. Pixel colors on the image plane, denoted by C, are computed by α-blending the contributions of Gaussians ordered from nearest to farthest:

$$C = \sum_{i \in N} \alpha_i c_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{4}$$

$$\alpha_i = \sigma_i e^{-\frac{1}{2}(\mu - u_i)^T \Sigma'(\mu - u_i)}, \tag{5}$$

where c_i is the color of each Gaussian, and u_i represents their projected uv coordinates. Rendering only the tool Gaussians, the background appears black ($C = 0$). In the resulting 2D image, only the surgical instrument regions are colored, enabling clear segmentation. By setting a contrast threshold, we differentiate the black background (background) from the colored instrument (foreground) to generate the 2D mask. This mask allows for the application of contour detection algorithms to define the contours of the foreground. The pixel coordinates of these bounding boxes facilitate the automated creation of annotation files.

2.5 Experimental Data Recording

In order to assess our synthetic images, we present a novel methodology for acquiring images and instrument poses (camera and tools) from real scenarios. Images with tools are used for validation, but realistic backgrounds and tools are not mixed during the generation of Gaussian representation. Fresh ex-vivo lamb liver, kidney, and fat placed within a laparoscopic trainer platform (POP Trainer, Optimist GMBH, Austria) to mimic the surgical scene. We use a clinical laparoscopic camera from the daVinci Surgical System (HD-2 Stereoendoscope Module, Intuitive Surgical) for image collection, and a daVinci Large Needle Driver as our tool. Given the current requirement for a rigid tool, we fixed all joints in a neutral position. We utilise the NDI Aurora electromagnetic tracking system with 6DoF sensors for our Ground Truth pose of the tool and camera. Image and pose data were collected synchronously. The setup for surgical data acquisition is depicted in Figure 3.

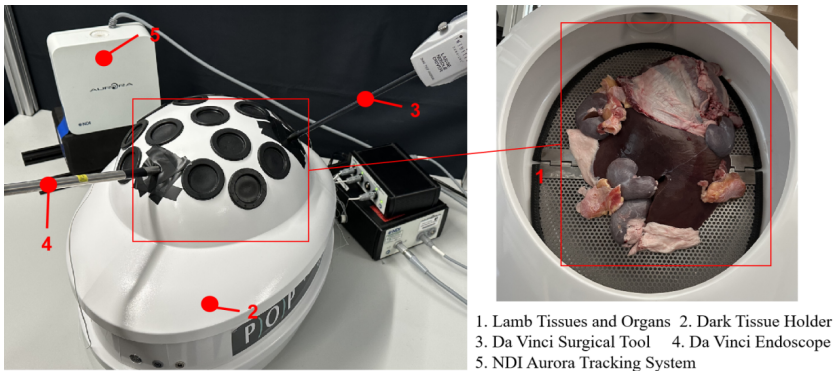

Fig. 3. Our dataset recording platform.

In the dataset, we recorded three distinct videos. 1) A training dataset for representing the background scene with no tools present; 2) An isolated acquisition of the surgical tools; and 3) A ground truth dataset containing both background and tool, designated below as the Ground Truth (GT) dataset. An additional test dataset mimicking the GT dataset was also recorded.

3 Experiments and Results

After acquiring all three data sets (Background, Tool, GT), we utilized the tool and camera positions within the GT dataset as direct inputs to the image synthesis. This approach guarantees that generating images that align with the GT images regarding the same camera and surgical instrument positions is feasible, replicating the GT image with a synthesized version. Consequently, real-world images can be a test case for synthetic images under such conditions.

3.1 Synthetic Image Quality Evaluation

A key challenge in image generation is obtaining precise GT images for direct comparison. We utilize Gaussian editing and tool pose tracking to acquire accurate GT images, a capability that sets our method apart from other generative approaches, which often lack precise GT data. For evaluation, we focus on comparing our results with these GT images to accurately assess our method's effectiveness. We evaluate the quality of our synthesized images using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). We also overlay synthetic and GT images to identify discrepancies. The results, illustrated in Figure 4, demonstrate the superiority of our method in producing high-quality images in GT scenes compared to alternatives. We conducted rigorous comparisons with two state-of-the-art (SOTA) Nerf-based methods [12, 14, 20], chosen for their comparable training durations. The PSNR, SSIM, and LPIPS scores are presented in Table 1.

Table 1. Comparative Analysis of Image Quality of GT Scene (Mean and standard deviation).

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Instant-NGP	16.603±0.782	0.741±0.015	0.758±0.047
Nerfacto	22.736±1.435	0.796±0.016	0.394±0.035
Ours	27.796±1.796	0.912±0.029	0.287±0.022

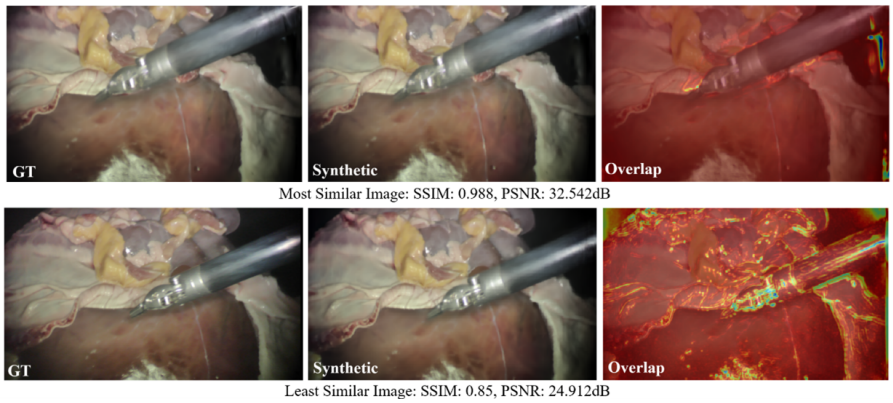

Fig. 4. Illutsration of GT and Synthetic pairs for the most and least similar image, indicating the regions of increased difference

3.2 Neural Network Training Experiment

Next, we evaluated the efficacy of our synthesized images for training neural networks, specifically using YOLOv5 [18] for object detection. We trained two models: one on the GT Dataset and one on our synthetic data, both using 1568 images. For consistency, images were resized to 640x640 pixels, and training was conducted for 100 epochs using YOLOv5's default parameters. All training was performed on an NVIDIA GeForce RTX 4050 (6G). Our test set comprised 300 images, each with unique combinations of camera and instrument poses. We assessed precision and recall to determine the viability of training with synthesized data and conducted multi-fold experiments for reliability. Table 2 shows our results, indicating that models trained on generated images outperform those trained on GT images in both precision and recall. This improvement is attributed to our method's ability to augment data by rendering images from new viewpoints and instrument poses, providing stronger priors for detection.

Table 2. Performance Comparison of Neural Networks Trained with Synthetic vs. Ground Truth Images on the 300 image real-world Test Dataset

Model	Precision ↑	Recall ↑
Synthetic Training Input	0.801	0.901
GT Training Input	0.703	0.804

4 Conclusion

This paper introduces a novel surgical image dataset generation method based on 3D Gaussian Splatting, aiming to address the challenge of insufficient surgical image datasets. We first trained Gaussian models representing surgical scenes and instruments separately to achieve this. We adopted a circular sampling strategy for the surgical scene Gaussian models, enabling accurate extraction and labeling of surgical instrument Gaussians. We created new scene models by fusing the extracted surgical instrument Gaussians with those from the surgical scene, allowing for image rendering of surgical instruments in any pose. This process also auto-generates annotation information for surgical instruments. Our experiments confirmed the high quality of images generated by our method, achieving a PSNR of 29.592. Our generated datasets have been proven effective for training neural networks, resulting in a 12% improvement in performance when models are trained on generated images compared to those trained on ground truth images. Currently, our method can generate and edit static image data of surgical tools within static scenes. This work hopes to alleviate the data scarcity in the surgical domain and inspire further enhancement of 3D Gaussian Splatting techniques for data generation.

Disclosure of Interests. The authors have no relevant financial or non-financial interests to disclose.

References

1. Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., Yang, B.: Medgan: Medical image translation using gans. Computerized medical imaging and graphics **79**, 101684 (2020)
2. Attanasio, A., Scaglioni, B., De Momi, E., Fiorini, P., Valdastri, P.: Autonomy in surgical robotics. Annual Review of Control, Robotics, and Autonomous Systems **4**, 651–679 (2021)
3. Azagra, P., Sostres, C., Ferrández, Á., Riazuelo, L., Tomasini, C., Barbed, O.L., Morlana, J., Recasens, D., Batlle, V.M., Gómez-Rodríguez, J.J., et al.: Endomapper dataset of complete calibrated endoscopy procedures. Scientific Data **10**(1), 671 (2023)
4. Chen, G., Wang, W.: A survey on 3d gaussian splatting. arXiv preprint arXiv:2401.03890 (2024)
5. Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A.: A review of medical image data augmentation techniques for deep learning applications. Journal of Medical Imaging and Radiation Oncology **65**(5), 545–563 (2021)
6. Colleoni, E., Psychogyios, D., Van Amsterdam, B., Vasconcelos, F., Stoyanov, D.: Ssis-seg: Simulation-supervised image synthesis for surgical instrument segmentation. IEEE Transactions on Medical Imaging **41**(11), 3074–3086 (2022)
7. Colleoni, E., Stoyanov, D.: Robotic instrument segmentation with image-to-image translation. IEEE Robotics and Automation Letters **6**(2), 935–942 (2021)
8. Hasan, M.K., Calvet, L., Rabbani, N., Bartoli, A.: Detection, segmentation, and 3d pose estimation of surgical tools using convolutional neural networks and algebraic geometry. Medical Image Analysis **70**, 101994 (2021)
9. Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., Mukhopadhyay, A.: Gans for medical image analysis. Artificial Intelligence in Medicine **109**, 101938 (2020)
10. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
11. Lee, D., Yu, H.W., Kwon, H., Kong, H.J., Lee, K.E., Kim, H.C.: Evaluation of surgical skills during robotic surgery by deep learning-based multiple surgical instrument tracking in training and actual operations. Journal of clinical medicine **9**(6), 1964 (2020)
12. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
13. Moccia, S., Romeo, L., Migliorelli, L., Frontoni, E., Zingaretti, P.: Supervised cnn strategies for optical image segmentation and classification in interventional medicine. Deep Learners and Deep Learner Descriptors for Medical Applications pp. 213–236 (2020)
14. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
15. Ozawa, T., Hayashi, Y., Oda, H., Oda, M., Kitasaka, T., Takeshita, N., Ito, M., Mori, K.: Synthetic laparoscopic video generation for machine learning-based surgical instrument segmentation from real laparoscopic video and virtual surgical

instruments. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **9**(3), 225–232 (2021)
16. Öztürk, A.E., Erçelebi, E.: Real uav-bird image classification using cnn with a synthetic dataset. Applied Sciences **11**(9), 3863 (2021)
17. Psychogyios, D., Vasconcelos, F., Stoyanov, D.: Realistic endoscopic illumination modeling for nerf-based data generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 535–544. Springer (2023)
18. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
19. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
20. Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., et al.: Nerfstudio: A modular framework for neural radiance field development. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–12 (2023)
21. Tsirikoglou, A., Eilertsen, G., Unger, J.: A survey of image synthesis methods for visual machine learning. In: Computer Graphics Forum. vol. 39, pp. 426–451. Wiley Online Library (2020)
22. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: 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)