SurgTPGS: Semantic 3D Surgical Scene Understanding with Text Promptable Gaussian Splatting

Yiming Huang[1,2] *, Long Bai[1,2,3] *, Beilei Cui[1,2] *, Kun Yuan[3,4], Guankun Wang[1,2], Mobarak I. Hoque[5], Nicolas Padoy[4], Nassir Navab[3], and Hongliang Ren[1,2] **

[1] The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Shenzhen Research Institute, CUHK, Shenzhen, China
[3] Technical University of Munich, Munich, Germany
[4] University of Strasbourg & IHU Strasbourg, Strasbourg, France
[5] University College London, London, United Kingdom
{yhuangdl, b.long, beileicui}@link.cuhk.edu.hk, hren@cuhk.edu.hk

Abstract. In contemporary surgical research and practice, accurately comprehending 3D surgical scenes with text-promptable capabilities is particularly crucial for surgical planning and real-time intra-operative guidance, where precisely identifying and interacting with surgical tools and anatomical structures is paramount. However, existing works focus on surgical vision-language model (VLM), 3D reconstruction, and segmentation separately, lacking support for real-time text-promptable 3D queries. In this paper, we present SurgTPGS, a novel text-promptable Gaussian Splatting method to fill this gap. We introduce a 3D semantics feature learning strategy incorporating the Segment Anything model and state-of-the-art vision-language models. We extract the segmented language features for 3D surgical scene reconstruction, enabling a more in-depth understanding of the complex surgical environment. We also propose semantic-aware deformation tracking to capture the seamless deformation of semantic features, providing a more precise reconstruction for both texture and semantic features. Furthermore, we present semantic region-aware optimization, which utilizes regional-based semantic information to supervise the training, particularly promoting the reconstruction quality and semantic smoothness. We conduct comprehensive experiments on two real-world surgical datasets to demonstrate the superiority of SurgTPGS over state-of-the-art methods, highlighting its potential to revolutionize surgical practices. Our code is available at: https://github.com/lastbasket/SurgTPGS.

Keywords: 3D Scene Understanding · Text-Promptable Segmentation · Robotic Surgery.

* Co-first authors.
** Corresponding author.

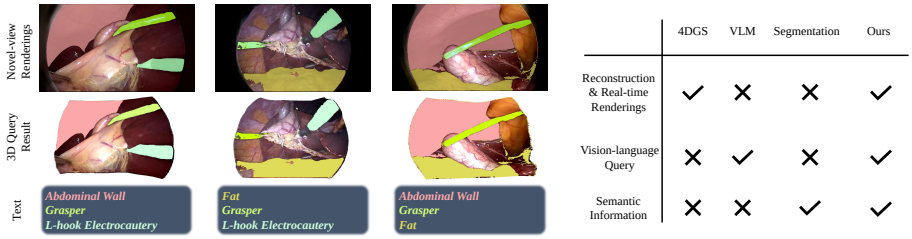

Fig. 1. SurgTPGS provides 3D semantic-segmentation with text-promptable queries, complementing the drawbacks of 4DGS, VLM, and segmentation networks.

1 Introduction

In modern computer-assisted interventions (CAI) procedures, accurately understanding 3D surgical scenes has become a cornerstone for successful surgical outcomes [16]. Previous studies [1, 5] have emphasized the significance of incorporating semantic segmentation as pixel-wise surgical guidance, providing context awareness of tissues and instruments for the surgeon [8, 14, 26]. Despite significant progress in related fields, the vision-language 3D understanding of surgical scenes remains largely unexplored. This knowledge gap critically hampers semantic-aware planning and navigation in robot-assisted surgery. Without a comprehensive vision-language 3D understanding, robotic systems struggle to interpret the surgical environment accurately.

Recent breakthroughs in neural rendering [9, 17] enable high-quality reconstruction and real-time rendering, revolutionizing the surgical scene reconstruction and rendering [6,7,15,23,28,30]. Subsequent extensions [10,18,24] enable the 3D-language interaction by incorporating the vision-language model (VLM) [19]. In surgical understanding, VLMs [20,27] provide textual guidance but are limited to 2D analysis, while visual question localized-answering (VQLA) [2, 13] adds anatomical grounding. The recent works [22, 29] integrate the vision-language model with surgical tool segmentation, achieving diverse surgical instruments in minimally invasive surgeries. However, current 3D surgical scene reconstruction methods [7, 15, 25] lack language-aligned anatomical understanding, critically limiting their ability to interpret clinician intents. While surgical VLM [2, 20] and segmentation networks [8,14,26] remain confined to 2D frames, failing at the spatial level in 3D scenes. This deficiency means they cannot meet the real-time demands of intraoperative navigation, where surgeons need immediate access to accurate 3D information to make informed decisions during procedures.

Although general-domain 3D-language models [10,12,18,24] achieve high accuracy in semantic queries for rigid objects, they fail in surgical scenes due to bleeding, occlusion, and soft-tissue motion. Besides, surgical anatomy's dynamic and deformable nature introduces unique challenges for semantic Gaussian training, causing inconsistent feature alignment during the training progress. This inconsistent alignment disrupts the spatial coherence of the appearance and the

semantic feature for the 3D Gaussian representation, leading to blurred and corrupted training results. The sub-optimal semantic features cause the mismatch between vision and textual embedding, degrading the final query accuracy. To complement this research gap, we introduce SurgTPGS, a novel text-promptable Gaussian Splatting for 3D surgical semantic query. We present the *3D semantics feature learning* strategy, integrating the Segment Anything model (SAM) [11] with vision-language surgery features. To account for the dynamic nature of surgical scenes, we propose *semantic-aware deformation tracking*, capturing the semantic-level deformation for more high-quality and precise reconstruction for both appearance and semantic features. Furthermore, we introduce *semantic region-aware optimization* with regional-based semantic supervision, improving the reconstruction quality semantic smoothness for object identification. As shown in Fig 1, SurgTPGS could offer surgeons real-time, semantic-aware assistance with language queries, improving surgical outcomes and patient care. Specifically, the contributions of this paper are as follows:

- We present SurgTPGS, the **first** text-promptable Gaussian Splatting approach for 3D surgical scene understanding.
- We introduce a 3D semantic feature learning strategy integrating the SAM with VLM for in-depth surgical scene understanding.
- We propose semantic-aware deformation tracking to capture semantic deformation, enabling more precise texture and semantic feature reconstruction.
- We propose semantic region-aware optimization, which uses regional-based semantic information to supervise training and improve reconstruction quality and semantic smoothness.
- We show that SurgTPGS achieves state-of-the-art results through comprehensive experiments on real-world surgical datasets.

2 Methodology

As illustrated in Fig 2, our method first extracts semantic embedding with SAM [11] and CAT-Seg [3], then we train the deformable Gaussian with the semantic features by *semantic-aware deformation tracking* and *semantic region-aware optimization*. We achieve text-promptable 3D query by mapping the rendered semantic features with the text embedding.

2.1 Preliminaries

3D Gaussian Splatting (3DGS) represents scenes using a 3D Gaussian point cloud with learnable attributes, enabling efficient real-time rendering and inspiring various applications. The 3DGS is represented as $\mathcal{G} = \{\mu, r, s, o, \mathbf{SH}\}$, where $\mu, r, s, o, \mathbf{SH}$ are the mean, rotation, scale, opacity and spherical harmonic for color representation. To apply 3DGS in surgical scene reconstruction, Deform3DGS [25] proposed the flexible deformation modeling scheme (FDM) to

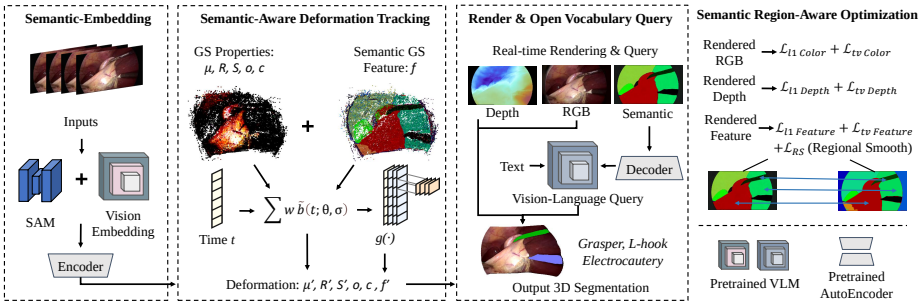

Fig. 2. Overview of SurgTPGS. We first extract semantic embedding from SAM [11] and VLM [3]; then, we train the deformable Gaussians with semantic-aware deformation tracking and semantic-region-aware optimization. SurgTPGS supports real-time semantic 3D query and novel-view rendering simultaneously.

deal with deformable tissue challenges:

$$\psi\left(t;\Theta\right)=\sum_{j=1}^{B}\omega_j\tilde{b}\left(t;\theta_j,\sigma_j\right),\ |\ \tilde{b}(t;\theta,\sigma)=\exp\left(-\frac{1}{2\sigma^2}(t-\theta)^2\right), \tag{1}$$

where $\tilde{b}(\cdot)$ is the Fourier and polynomial basis function, $\{\omega,\theta,\sigma\}\in\Theta$ are learnable weight, center, and variance. For mean, rotation, and scale, the deformation is represented as $\mu'=\mu+\psi(t;\Theta^{\mu})$, $r'=r+\psi(t;\Theta^{r})$, and $s'=s+\psi(t;\Theta^{s})$. The mean and covariance [31] are projected into 2D with the given camera parameters and rendered into color C and depth D with alpha blending [9].

2.2 Proposed Methodology

3D Semantic Feature Learning in Surgical Scenes. Inspired by [18], we introduce 3D Semantics feature learning with SAM [11] as the segmentation generator and CAT-Seg [3] as our vision-language baseline. We use SAM to obtain a segmentation mask M and then gather the vision-language feature from CAT-Seg within the segmented areas. Similar to [18], we train an autoencoder $\mathbf{\Phi}$ for compressing the feature dimension for more efficient training. For pixel x, the pixel-aligned semantic feature can be obtained as follows:

$$\hat{F}(x)=\mathbf{\Phi}_{encode}(V_{CAT-Seg}\left(I\odot M(x)\right), \tag{2}$$

where I is the input image, $\odot$ is the element-wise multiplication, $\mathbf{\Phi}_{encode}$ is the encoding function, and $V_{CAT-Seg}$ is the feature extraction function of CAT-Seg.

Semantic-Aware Deformation Tracking. Instead of using the deformation model as the mean, rotation, and scale from [25], our proposed method utilizes a convolution-based network to enhance the spatial consistency for semantic

feature representation. Building on the semantic features obtained from our 3D semantics feature learning, we track the deformation of semantic regions with a lightweight spatial-aware network $g = (\text{Conv1d}(\cdot))$. For a feature f associate with Gaussian, the deformed f' is computed as:

$$f' = f + g(\text{Concat}(f, t)) \cdot \beta, \ | \ \beta = 1/(1 + \exp(-\delta\psi(t; \Theta^f))), \tag{3}$$

where β is the probability map from the sloped sigmoid. t is the time, and δ is the sigmoid slope, where 2.5 is applied. Similar to the color rendering [9], the final semantic feature map F is rendered from f' using alpha blending. With the proposed semantic-aware deformation tracking, we achieve spatially consistent modeling of semantic feature deformation.

Text-Promptable Surgical Scene Querying. By obtaining the pixel-aligned semantic feature F from rendering, we query the 3D language feature by the language embedding ϕ_{text}. The rendered feature is first decoded back into the image embedding $\phi_{img} = \mathbf{\Phi}_{decode}(F)$, and then following [10], we calculate the relevancy score as:

$$score = \min_i \frac{\exp(\phi_{img} \cdot \phi_{text})}{\exp(\phi_{img} \cdot \phi_{text}) + \exp(\phi_{img} \cdot \phi^i_{canon})}, \tag{4}$$

where ϕ^i_{canon} is a set of CLIP embeddings from the canonical phrase *"object"*, *"things"*, *"stuff"*, and *"texture"*. After computing the score, we threshold the segmentation area with $\mathbf{e} = 0.4$, and the final mask output is defined by $M_{query} = score \geq \mathbf{e}$. With text-promptable 3D querying, surgeons can use natural language to obtain the 3D semantic information of surgical instruments and anatomical structures in real-time.

Semantic Region-Aware Optimization. Given a semantic feature map F, we introduce a region-based loss to enforce a more accurate and consistent semantic feature for surgical scenes. We perform a specific optimization for each unique semantic area l greater than 1000 pixels within the ground truth $\hat{F}$. Our proposed region-aware smoothness loss is defined as:

$$\mathcal{L}_{RS} = \sum_{l \in \hat{F}} \sum_{(i,j) \in l} |F_{ij} - \text{mean}(F_l)|, \ | \ |l| > 1000 \tag{5}$$

The region-aware loss ensures that the semantic features within large regions are more consistent and accurate, benefiting both the reconstruction of rendering color and semantic information with clearer object boundaries. Additionally, we apply the L1 loss and total variation (TV) loss L_{tv} to supervise color, depth, and semantic features. Our final loss is presented as follows:

$$\mathcal{L} = |C - \hat{C}| + |\frac{D - \hat{D}}{D\hat{D}}| + |F - \hat{F}| + \lambda(\mathcal{L}_{tv}(C) + \mathcal{L}_{tv}(D) + \mathcal{L}_{tv}(F) + \mathcal{L}_{RS}), \tag{6}$$

where λ is the weight for the TV loss and the region-aware smoothness. $\hat{C}$ is the ground truth color, $\hat{D}$ is the metric depth obtained from pre-trained model [4].

Table 1. Quantative results on the CholecSeg8K [5] dataset. We compare the mIoU scores (%), and highlighted the first, second, and third. Our method outperforms state-of-the-art baseline methods.

Method	01_00080	01_00240		01_15019		12_15750		17_01803	
	Liver	Liver	Grasper	Abdominal Wall	Grasper	Fat	L-hook Electrocautery	Abdominal Wall	Grasper
LangSplat [18]	64.87	57.91	4.5	23.84	4.13	25.63	5.97	43.38	4.13
LangSplat (SurgVLP [27])	65.03	57.14	4.31	32.79	5.61	4.1	5.97	42.97	4.53
LangSplat (CAT-Seg [3])	77.51	73.69	4.96	91.05	58.29	75.83	23.81	72.93	14.46
OpenGaussian [24]	2.64	0.08	0	0	0	0	8.05	0	0
OpenGaussian (SurgVLP [27])	1.76	0	0	0	0	0.19	0	13.02	2.24
OpenGaussian (CAT-Seg [3])	2.21	1.9	12.79	14.89	12.82	1.26	7.49	0	0
DGD [12]	13.98	42.43	3.83	9.68	2.9	14.79	0	4.09	2.67
Ours (CLIP [19])	64.96	57.52	4.69	15.94	4.1	26.59	5.34	23.55	7.18
Ours (SurgVLP [27])	65.07	57.04	6.01	32.99	5.91	8.03	6.14	42.94	4.56
Ours (CAT-Seg [3])	89.82	79.06	65.78	97.24	68.83	74.84	73.29	92.12	55.04

Table 2. Quantative results on the EndoVis18 [1] dataset. We highlighted first, second, and third values. We compare the mIoU scores (%) for segmentation, and query speed (FPS), training time for model efficiency. The performance of our proposed method surpasses state-of-the-art baselines.

Method	Seq_5		Seq_9			Query FPS↑	Training Time↓
	instrument -wrist	kidney -parenchyma	instrument -shaft	instrument -wrist	instrument -clasper		
LangSplat [18]	5.17	59.16	16.47	7.89	6.15	67.5	8 min
LangSplat (SurgVLP [27])	7.36	52.94	9.87	11.72	7.02	23.8	9 min
LangSplat (CAT-Seg [3])	17.16	54.42	38.49	21.74	0.11	67.5	8 min
OpenGaussian [24]	0	1.62	0	0	0	378.9	15 min
OpenGaussian (SurgVLP [27])	1.13	0.54	0	0	0	118.5	16 min
OpenGaussian (CAT-Seg [3])	0	0.06	29.38	12.82	14.89	378.9	15 min
DGD [12]	1.14	21.68	8.22	5.04	7.37	4.43	26 hrs
Ours (CLIP [19])	5.17	59.17	15.55	4.1	15.94	67.5	4 min
Ours (SurgVLP [27])	11.2	21.26	25.3	4.73	3.53	23.8	5 min
Ours (CAT-Seg [3])	43.29	71.98	20.79	24.34	38.94	67.5	4 min

3 Experiments

3.1 Dataset

Our evaluation utilizes two public datasets, CholecSeg8K [5] and EndoVis18 [1]. CholecSeg8K is a comprehensive semantic segmentation dataset derived from the Cholec80 [21] dataset for laparoscopic cholecystectomy procedures, containing various annotations. EndoVis18 is a robotic scene segmentation dataset involving surgical instruments and anatomical classes. In our evaluation, we utilize five video sequences (01_00080, 01_00240, 01_15019, 12_15750, 17_01803) from CholecSeg8K and two subsets of video sequences (Seq_5, Seq_9) from EndoVis18. We split the training and testing images following [28] and evaluate the segmentation with mean intersection over union (mIoU). We also evaluate the training time and query speed to compare the model efficiency.

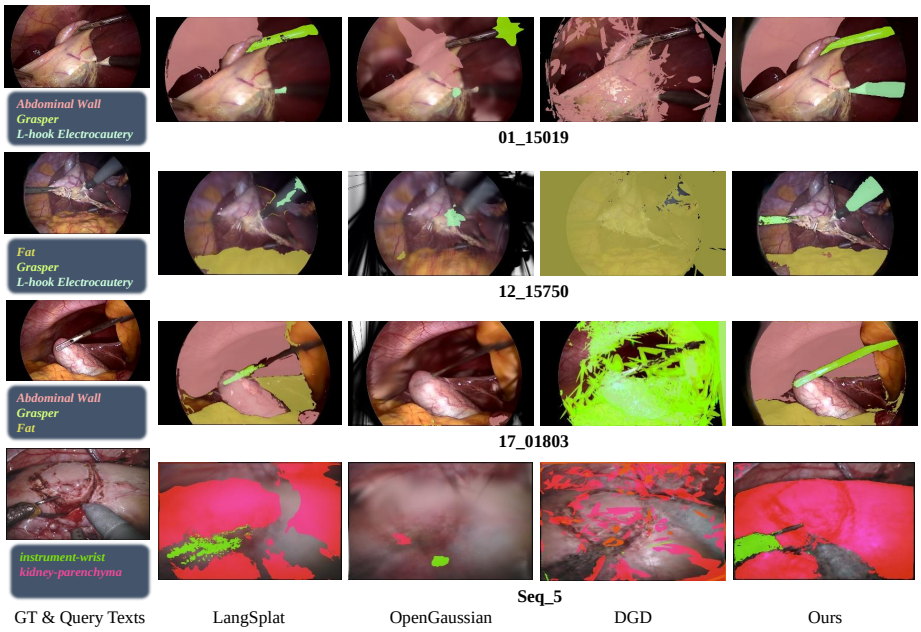

Fig. 3. Qualitative result on CholecSeg8K [5] and EndoVis18 [1] dataset. We demonstrate the results of LangSplat [18], OpenGaussian [24], and ours with CAT-Seg [3] as VLM. Our method shows more consistent and precise segmentation.

3.2 Implementation Details

We run all experiments on a single RTX4090 GPU. We utilize Adam optimizer with an initial learning rate of 1.6×10^{-3}. We train CAT-Seg [3] with the CholecSeg8K [5] and EndoVis18 [1], excluding all sequences in our experiments. For EndoVis18, we use half resolution for efficiency. The weight λ is set to 0.01.

3.3 Results

We compare our proposed SurgTPGS with state-of-the-art Gaussian Splatting methods for open-vocabulary query: LangSplat [18], OpenGaussian [24], DGD [12]. Furthermore, we reimplement LangSplat and OpenGaussian by replacing the vision-language model in their baseline with the SurgVLP [27] for surgical VQA and CAT-Seg [3]. In addition to the proposed method with CAT-Seg, we also provide the results of our methods in the CLIP and SurgVLP versions.

As shown in Table 1, our method demonstrates superiority over the baseline approaches on the CholecSeg8K dataset. For example, our method with CAT-Seg achieves a mIoU of 89.82% for *abdominal wall* in 01_00080, outperforming LangSplat (CAT-Seg) with a mIoU of 77.51% and other methods by a large margin. Our method also performed excellently in the EndoVis18 dataset in

Table 3. Ablation results on 01_00240 of CholecSeg8K [5]. We compare the mIoU and the PSNR of rendering results. The first and second are highlighted.

Semantic-Aware Deformation Tracking	Region-Aware Smoothness	mIoU↑ Liver	mIoU↑ Grasper	PSNR↑
✗	✗	79.43	64.02	18.72
✓	✗	78.96	62.79	20.54
✗	✓	78.85	64.78	19.38
✓	✓	79.06	65.78	24.06

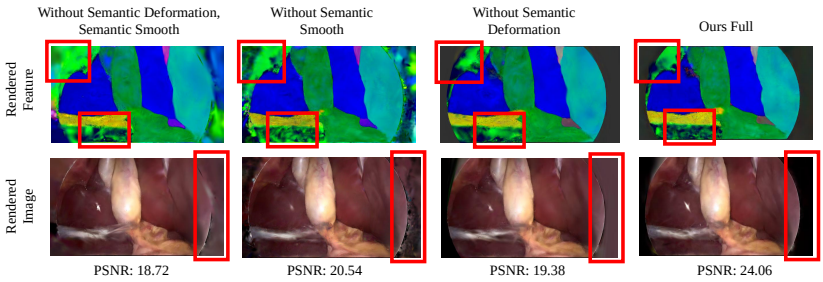

Fig. 4. Visualization of the ablation results. Results from our full model present smoother boundaries and clear visualization quality.

Table 2. For the kidney-parenchyma segmentation, our method with CAT-Seg achieves a mIoU of 71.98%, exceeding LangSplat (CAT-Seg), which had 59.17% and other baselines, while maintaining a high frame rate of 67.5 FPS. Additionally, our method achieves training with only 4 minutes, significantly less than the 8 - 16 minutes required by other methods. This indicates that our method can not only segment surgical objects accurately but also respond quickly to queries, while others fail due to the challenging deformation of surgical scenes. The choice of pre-trained VLM also significantly affects the results. With CAT-Seg, our method can precisely segment the objects in surgical scenes, while with SurgVLP [27] and CLIP [19] the performance drops due to incorrect mapping from VLM. As shown in Fig 2, the quantitative results from both datasets show that our method provides more accurate segmentation with clearer boundaries.

We conduct the ablation study on the CholecSeg8K dataset, as shown in Table 3 and Fig. 4. By removing the proposed components, the mIoU and the PSNR decrease, and the visual quality for rendering drops significantly, showing that both components are crucial for accurate semantic feature reconstruction.

4 Conclusions

In conclusion, SurgTPGS represents a significant advancement in understanding 3D surgical scenes. By integrating a 3D semantics feature learning strategy, semantic-aware deformation tracking, and semantic region-aware optimization,

our method addresses the limitations of existing methods and outperforms state-of-the-art techniques on real-world surgical datasets in terms of text-promptable segmentation accuracy, maintaining efficient training time and query speed. SurgTPGS can revolutionize robot-assisted surgery by enabling robots to better understand and interact with the surgical environment, leading to more precise and safer robot-assisted surgery procedures.

Acknowledgements. This work was supported by Hong Kong RGC CRF C4026-21G, RIF R4020-22, GRF 14211420, 14216020 & 14203323).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)
2. Bai, L., Islam, M., Seenivasan, L., Ren, H.: Surgical-vqla: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 6859–6865. IEEE (2023)
3. Cho, S., Shin, H., Hong, S., Arnab, A., Seo, P.H., Kim, S.: Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4113–4123 (2024)
4. Cui, B., Islam, M., Bai, L., Wang, A., Ren, H.: Endodac: Efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 208–218. Springer (2024)
5. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.: Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. arXiv preprint arXiv:2012.12453 (2020)
6. Huang, Y., Cui, B., Bai, L., Chen, Z., Wu, J., Li, Z., Liu, H., Ren, H.: Advancing dense endoscopic reconstruction with gaussian splatting-driven surface normal-aware tracking and mapping. arXiv preprint arXiv:2501.19319 (2025)
7. Huang, Y., Cui, B., Bai, L., Guo, Z., Xu, M., Islam, M., Ren, H.: 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)
8. Jin, Y., Yu, Y., Chen, C., Zhao, Z., Heng, P.A., Stoyanov, D.: Exploring intra-and inter-video relation for surgical semantic scene segmentation. IEEE Transactions on Medical Imaging **41**(11), 2991–3002 (2022)
9. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
10. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19729–19739 (2023)

11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
12. Labe, I., Issachar, N., Lang, I., Benaim, S.: Dgd: Dynamic 3d gaussians distillation. In: European Conference on Computer Vision. pp. 361–378. Springer (2024)
13. Li, J., Skinner, G., Yang, G., Quaranto, B.R., Schwaitzberg, S.D., Kim, P.C., Xiong, J.: Llava-surg: towards multimodal surgical assistant via structured surgical video learning. arXiv preprint arXiv:2408.07981 (2024)
14. Liu, M., Han, Y., Wang, J., Wang, C., Wang, Y., Meijering, E.: Lskanet: Long strip kernel attention network for robotic surgical scene segmentation. IEEE Transactions on Medical Imaging **43**(4), 1308–1322 (2023)
15. Liu, Y., Li, C., Yang, C., Yuan, Y.: Endogaussian: Gaussian splatting for deformable surgical scene reconstruction. arXiv preprint arXiv:2401.12561 (2024)
16. McLachlan, G.: From 2d to 3d: the future of surgery? The Lancet **378**(9800), 1368 (2011)
17. 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)
18. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20051–20060 (2024)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
20. Seenivasan, L., Islam, M., Krishna, A.K., Ren, H.: Surgical-vqa: Visual question answering in surgical scenes using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 33–43. Springer (2022)
21. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1), 86–97 (2016)
22. Wang, H., Yang, G., Zhang, S., Qin, J., Guo, Y., Xu, B., Jin, Y., Zhu, L.: Video-instrument synergistic network for referring video instrument segmentation in robotic surgery. IEEE Transactions on Medical Imaging (2024)
23. 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)
24. Wu, Y., Meng, J., Li, H., Wu, C., Shi, Y., Cheng, X., Zhao, C., Feng, H., Ding, E., Wang, J., et al.: Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. arXiv preprint arXiv:2406.02058 (2024)
25. Yang, S., Li, Q., Shen, D., Gong, B., Dou, Q., Jin, Y.: Deform3dgs: Flexible deformation for fast surgical scene reconstruction with gaussian splatting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 132–142. Springer (2024)
26. Yoon, J., Hong, S., Hong, S., Lee, J., Shin, S., Park, B., Sung, N., Yu, H., Kim, S., Park, S., et al.: Surgical scene segmentation using semantic image synthesis with a virtual surgery environment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 551–561. Springer (2022)

27. Yuan, K., Srivastav, V., Yu, T., Lavanchy, J., Mascagni, P., Navab, N., Padoy, N.: Learning multi-modal representations by watching hundreds of surgical video lectures (2023)
28. Zha, R., Cheng, X., Li, H., Harandi, M., Ge, Z.: 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)
29. Zhou, Z., Alabi, O., Wei, M., Vercauteren, T., Shi, M.: Text promptable surgical instrument segmentation with vision-language models. Advances in Neural Information Processing Systems **36**, 28611–28623 (2023)
30. Zhu, L., Wang, Z., Jin, Z., Lin, G., Yu, L.: Deformable endoscopic tissues reconstruction with gaussian splatting. arXiv preprint arXiv:2401.11535 (2024)
31. Zwicker, M., Pfister, H., Van Baar, J., Gross, M.: Surface splatting. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 371–378 (2001)