3D Occupancy Reconstruction in Dynamic and Deforming Surgical Environments

Om Shah
Lakeside School
Seattle, US
oms25@lakesideschool.org

Yun-Hsuan Su
Mount Holyoke College
South Hadley, US
msu@mtholyoke.edu

2024 International Symposium on Medical Robotics (ISMR) | 979-8-3503-7711-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISMR63436.2024.10585941

***Abstract*—Vision dimensionality during minimally invasive surgery is a critical contributor to patient success. Traditional visualizations of the surgical scene are 2D camera streams that obfuscate depth perception inside the abdominal cavity. A lack of depth in surgical views cause surgeons to miss tissue targets, induce blood loss, and incorrectly assess deformation. 3D sensors, while offering key depth information, are expensive and often incompatible with current sterilization techniques. Furthermore, methods inferring a 3D space from stereoscopic video struggle with the inherent lack of unique features in the biological domain. We present an application of deep learning models that can assess simple binary occupancy from a single camera perspective to recreate the surgical scene in high-fidelity. Our quantitative results (IoU=0.82, log loss=0.346) indicate a strong representational capability for structure in surgical scenes, enabling surgeons to reduce patient injury during minimally invasive surgery.**

I. Introduction

Robotic minimally invasive surgeries have provided surgeons with fine motor control in laparoscopic environments while reducing physical damage in patients [1]. Surgical robots create smaller incision sites and result in faster patient recovery times compared to standard procedures [2], [3]. Despite the benefits of robotic surgery, observation of the surgical cavity is often restricted to 2D interfaces. For surgeons, 2D views lose critical depth information necessary for navigating the field-of-view, warranting a time-intensive training process for interpreting flat surgical scenes [4]. Studies of minimally invasive surgery report optical fatigue when viewing 2D laparoscopic video streams for extended periods of time [5]. To provide greater depth-encoded visibility, 3D reconstruction of the surgical cavity is required. 3D camera systems that capture RGB color and an additional depth channel are proven to help reduce fatigue and increase surgeon performance [6]. However, operating a 3D camera in surgery can be cumbersome due to the extra physical implement and requires additional hardware expense for medical institutions [7]. Infrared-based 3D cameras also struggle with specular reflectance and dark spots in the surgical view. While 3D capture systems are improving for a variety of general scene tasks, their use in a surgical environment is dependent on stringent sterilization standards and biocompatibility.

An alternative approach to observing surgical environments is through camera-based 3D reconstruction. Traditional 3D reconstruction algorithms mathematically build structure from either motion (SFM) or texture (SFT) [8]–[10]. These methods work well for reconstructing surfaces with informative features. The surgical environment by contrast contains dynamic, low-texture tissue and frequently deforming surfaces caused by organ palpitations, thus proving SFM difficult to conduct [11], [12]. Another approach, simultaneously locating and mapping (SLAM), calls for tracking camera pose and building a 3D environment at the same time [13], [14]. SLAM is scalable for video reconstruction, but decreases in accuracy when unable to extract distinctive features from frames [15]–[19]. Both SLAM and SFM require prior feature extraction to model 3D spaces and sometimes face difficulties in handling occlusion when presented with inconsistent feature sets.

For general object reconstruction outside this domain, deep learning has brought substantial improvements in accuracy and efficiency. These advtanges are most apparent in reconstruction of high contrast, feature dense objects in light-filled spaces [20], [21]. Object representations have also become more diverse, with models capable of generating voxels, point clouds, meshes, and hybrid mediums from singular images [22], [23]. One recent advancement addresses several long standing problems with object representations: occupancy networks, deep neural networks trained to classify 3D points as either occupied or vacant [24]–[26]. The surface of an object is thus implicitly represented between occupied and vacant points. In the context of surgical environment reconstruction, occupancy networks can be trained to dynamically handle transitions from feature-dense to feature-sparse scenes without the need for complex, manual feature extraction. We propose an deep learning network that can identify occupancy flow from sequences of single view laparoscopic images. This approach can increase the depth resolution of reconstructed surfaces by forgoing the explicit point cloud generation, which inherently discretizes textural resolution; we predict occupancy directly from images and sampled points which can be adapted to downstream representations such as the

This material is based upon work supported by the National Science Foundation under Grant No. IIS-2101107. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF.

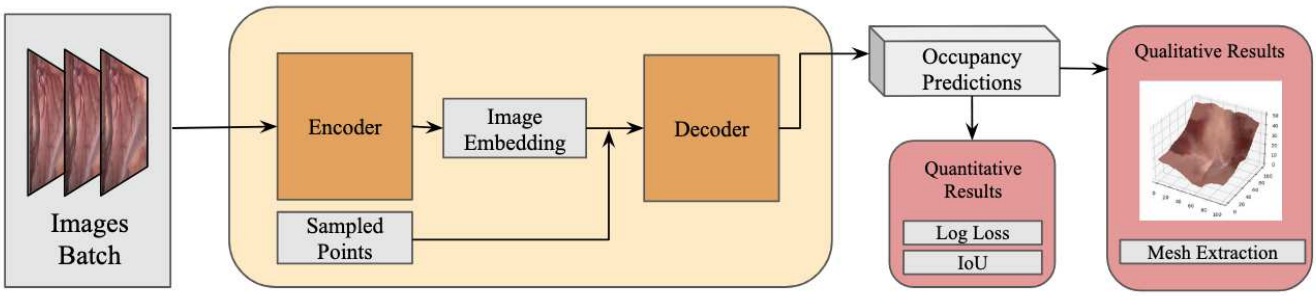

Fig. 1. High-level system architecture. Quantitative results are in the form of binary cross-entropy (log loss) and intersection over union. Qualitative results are visualized using extracted meshes.

mesh or voxel grid.

In summary, our contributions are as follows:

- We develop a deep learning model that learns the decision boundary for 3D surfaces in surgical environments.
- We demonstrate a representation that can be utilized for occupancy flow in dynamic, deforming scenes.

II. Related Work

A. Simultaneous Localization and Mapping

Approaches studying concurrent location and environmental tracking in surgical scenes focus on matching features between frames to comprehend unexplored areas. SLAM outputs successive point clouds for observed regions. Due to a heavy emphasis on feature extraction, point cloud densities are highly reliant on the quality of features in the scene [27]. Continual advancements in SLAM methodology led to the creation of the parallel tracking and mapping (PTAM) design paradigm: two separate threads were dedicated to the problems of agent tracking and mapping respectively [28]. A significant feature of SLAM and PTAM is their use of multiple camera angles for point cloud construction. However, this design is not always supported by surgical equipment – many operating rooms utilize a sole camera for minimally invasive surgery.

B. Deep Learning Approaches

Efforts to reconstruct the surgical environment using deep learning are sparse. One study attempts to translate SLAM to a deep learning model by employing neural networks to predict pixel-wise depth [29]. A truncated signed distance function applied to the predicted depth maps builds a volumetric reconstruction. However, this approach inherits the depth maps' difficulty predicting on occluded regions. Due to model output being restricted to a 2D depth image, the constructed volume will simply not be exposed to the existence of features behind occlusions. Similar approaches apply models to rectified stereo image pairs for depth prediction and volumetric tasks [30], [31]. The minimal number of studies for deep learning-based 3D reconstruction specifically for surgical scenes may be attributed to lack of large-scale surgical video datasets with an associated ground truth [11]. Studies that do employ deep learning often employ some variation of a self-supervised model to circumvent the present data void [29].

In the broader field of reconstructing objects using deep learning, models approach the problem through distinct representations. Early work introduced recurrent neural networks as a solution to creating intermittent 3D predictions when more rooted methods such as SFM and SLAM failed [20]. Voxels were also proposed as an early iteration of raw model output due to their standardized shape. To improve on memory-heavy voxel representations, new neural networks predicted point clouds [22]. However, extracting usable meshes requires intensive post-processing steps that can nullify the inference time reductions for point clouds. More recent increases in model complexity allow for direct prediction of meshes from image input [23]. This approach, like the voxel prediction network, is bound to a set prediction dimension (eg. 64^3) and often leads to self-intersecting meshes.

III. Method

We define occupancy in a 3D surgical space as points occupied by tissue and other biological matter, with vacancy referring to the lack thereof. The problem of identifying occupancy membership can thus be simplified to binary classification on neural networks, as demonstrated by previous work on occupancy networks in static spaces [24]. Sampled points in the 3D space are queried on the model and a 3D surface is formed from the network's decision boundary. We provide spatial understanding to the model through visual information from a monocular camera navigating the surgical space. An encoder-decoder network derives numerical feature vector for each point and converts the resulting low level features into a multi-resolution voxel grid.

A. Dataset

A video dataset with ground truth depth is procured from a previous surgical scene depth mapping study [29], [32]. This study manually generated depth maps from stereo vision and associated intrinsic and extrinsic camera

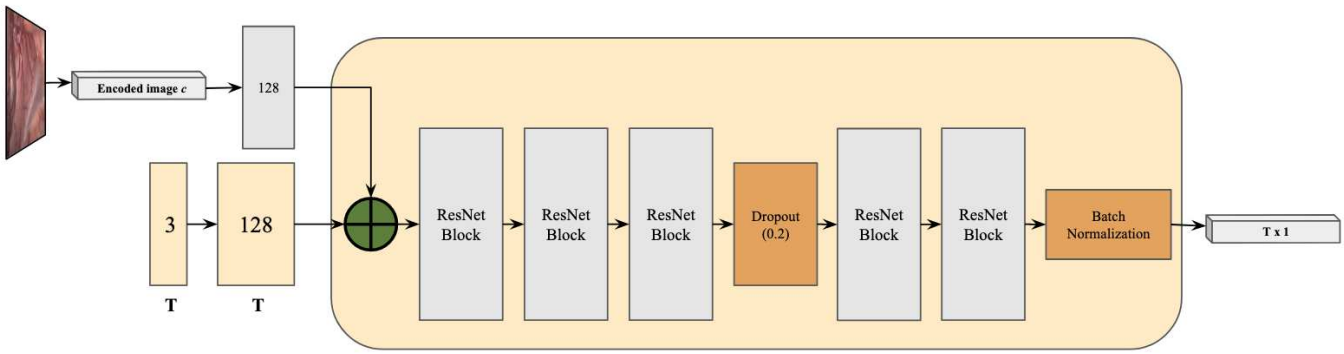

Fig. 2. Occupancy network decoder with image and point input. Five ResNet blocks consecutively decode the concatenated input into a single $T \times 1$ occupancy prediction.

information. Across multiple frames in the provided videos, surfaces are in motion and occasionally deform due to periodic biological functions. The dataset contains a total of 21 rectified videos. To compensate for the time constraints of this study, we utilize a single video with 1057 frames. Each frame's corresponding depth image has a saturation of 300mm represented by grayscale. To extract 3D points, we first build point clouds from the depth images. Each point cloud relies on a pinhole camera intrinsic to lift the 2D grayscale points into a 3D space. We perform voxel downsampling on the point cloud with a voxel size of 0.005 to represent the 3D view while minimizing downstream system load. A smooth, surface mesh is then derived using poisson surface reconstruction. The poisson reconstruction traverses an octree with a depth of 8, which bounds the resolution of the underlying grid to a maximum of 2^8. Reconstruction artifacts such as disconnected and self intersecting triangles are filtered out. We finally use the mesh to extract strictly surface point samples and translate them into a 3D space with the (x, y, z) bounds determined by point cloud bounds across the entire dataset: $(100, 100, 50)$. Points with a smaller z than the initial surface samples are labeled as occupied while points with a greater z are labelled as vacant. Through this process, we sample a total of $32,000$ points.

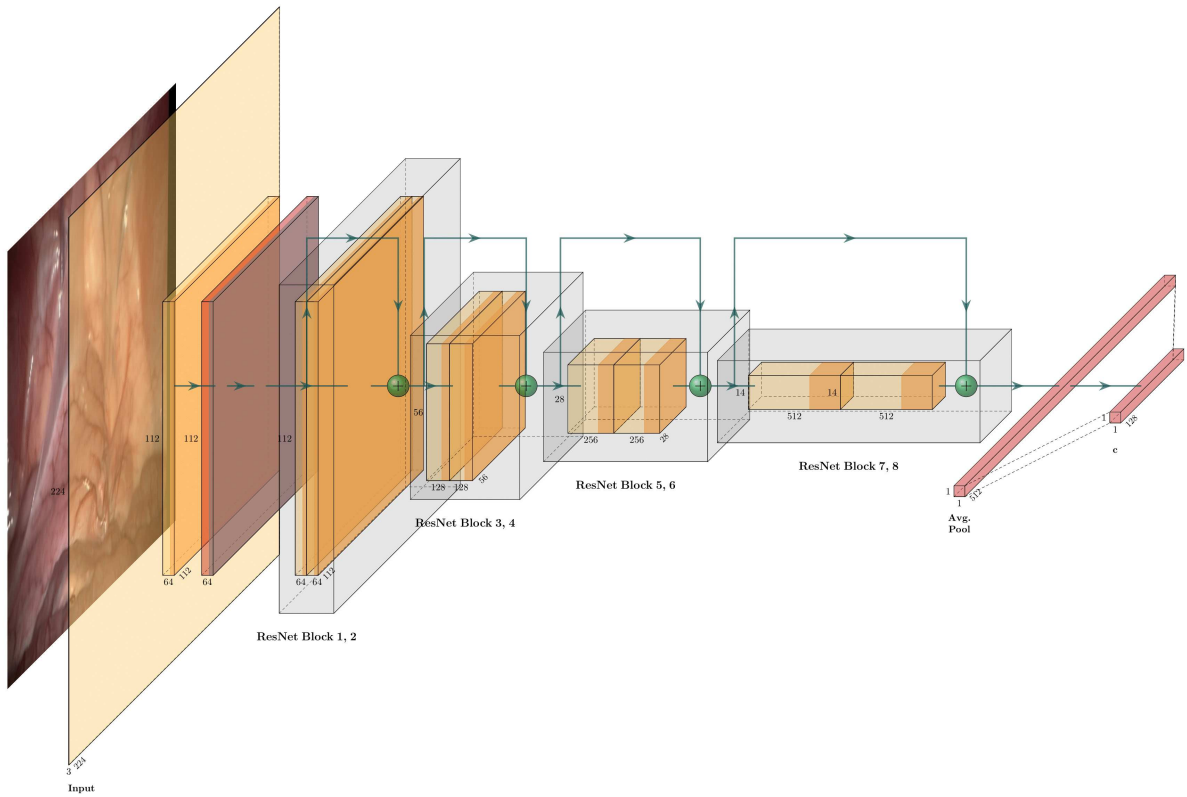

Fig. 3. ResNet-18 encoder for monocular camera input. The embedding c is concatenated to the sampled points before the decoder.

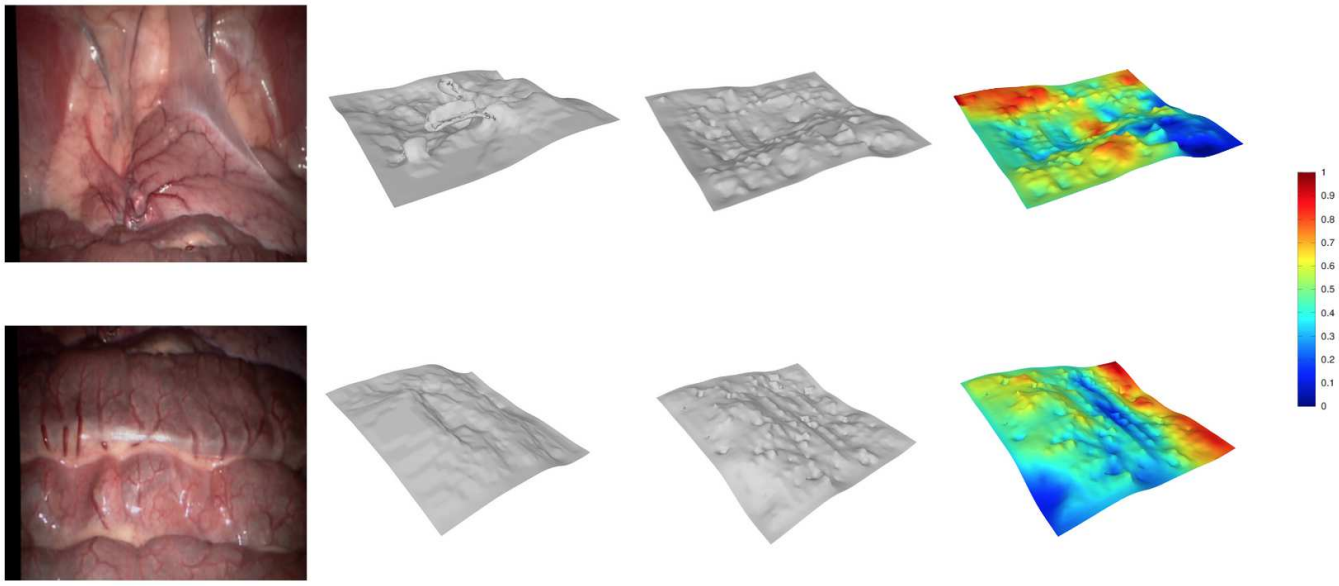

Fig. 4. Occupancy predictions visualized: (far left) input image; (left mesh) ground truth; (middle mesh) prediction, (right mesh) normalized distance error between meshes

B. Modeling

We build the deep learning network on the basis of individual point occupancy classification. This allows for fixed resolution during training and scalable resolution during inference. We encode the image input x of the surgical scene using ResNet-18, an industry-standard model that prevents vanishing gradients, pretrained on ImageNet (images are standardized channel-wise by ImageNet mean and standard deviation) [33]. The model's final 512 dimension vector is then converted into a 128 dimension vector. This relatively smaller dimension contributes to an overall smaller number of model parameters, accelerating the training process. The set of sampled 3D points T are first projected into a $T \times 128$ dimensional space prior to a concatenation with the image encoding. The concatenation result is passed through five basic ResNet blocks with each layer consisting of 128 nodes. Furthermore, each block contains a skip connection that concatenates the input vector to the output vector. The final output is transformed to a $T \times 1$ dimensional occupancy classification. We train the model with a Adam optimizer that minimize a binary cross entropy loss objective (trained on A100 GPU running CUDA). The learning rate was experimentally chosen from $\{0.1, 0.01, 0.001\}$ as 0.01 (no learning rate scheduling utilized) and manually chose the number of epochs as 3 before validation loss began increasing. The training dataset was split into mini-batches with a size of 2.

$$L_{BCE} = -(y \log(p) + (1 - y) \log(1 - p)) \tag{1}$$

Model performance is measured by adapting the common Intersection over Union (IoU) metric to the occupancy space. Prior to calculation, the occupancy predictions are rounded to the nearest integer (either 0 or 1) to determine strict binary occupancy. Then, the IoU is derived from the resulting intersection and union.

$$IoU = \frac{|M_{\text{preds}} \cap M_{\text{GT}}|}{|M_{\text{preds}} \cup M_{\text{GT}}|} \tag{2}$$

C. Inference

For inferring the model on novel samples, we adopt more freedom in point sampling and object representation. Through training, the model implicitly builds a three-dimensional understanding of the sampling space and a sense of varying depth ranges across frames. We can iteratively build higher resolution 3D reconstructions by sampling additional points as needed. A 3D Reconstruction is initialized on 2048 points within (100, 100, 40) and subsequently converted into a mesh representation. Meshes allow occupancy to inherit the tissue characteristics such as continuity and texture. We extract topological meshes by sampling the highest (on the z-axis) predicted points that are occupied.

IV. Results

A. Representative Capacity

The current state of representation in surgical spaces is confined to either 2D views or expensive 3D measurements. We therefore expect a deep learning approach to effectively translate the spatial attributes of surgical video streams into 3D structures. Our model predicts volumetric occupancy in the operative field and from this, we extract a topological mesh that represents the biological surface. Quantitatively, the model achieved an IoU score of 0.82 and a binary cross-entropy loss of 0.332 on the test set. Furthermore, we tested

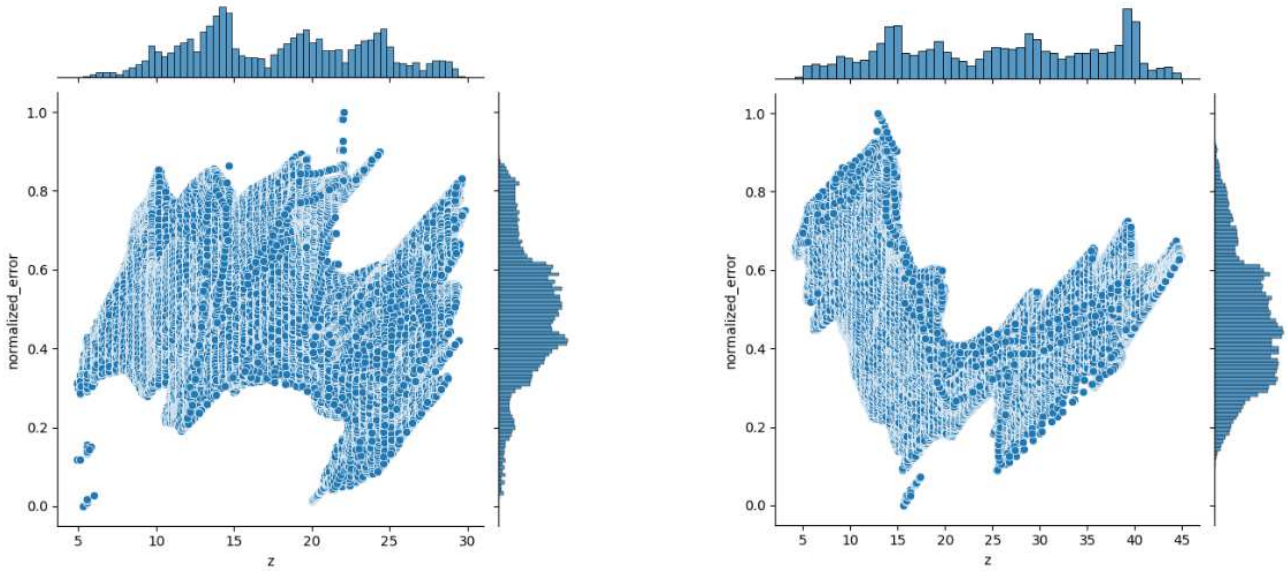

Fig. 5. Normalized distance error versus z (height): (left) plot for top row image in Figure 4; (right) plot for bottom row image in Figure 4

the trained occupancy network on a video with a completely different surgical environment featuring organ palpitation. On this hidden environment, the network achieved an IoU of 0.677 and a binary cross-entropy loss of 0.373.

TABLE I
METRICS BY DATASET

Dataset	Samples	IoU	Log Loss (BCE)
Training	739	0.823	0.346
Validation	211	0.785	0.432
Testing	106	0.821	0.332
Hidden Environment	252	0.677	0.373

B. Qualitative Performance

The model clearly captures underlying depth information across the spatial field, as seen in Figure 4. At a low level, varying depths are represented throughout each mesh, while continuity clearly delineates the 3D gradient. Minor tissue details with extreme local changes in depth were not found in some reconstruction samples. The source of the model smoothing over these depth changes is likely the use of a single video from which frames were fed into the model. While frames were shuffled prior to training, originally consecutive frames likely encoded to similar vectors and led the model to find a more generalized occupancy prediction for similar frames. Comparing errors for each vertex in the predicted mesh to the ground truth mesh in Figure 5, we can see that the model does not discriminate across a range of z-values. Higher z-values are often predicted with a higher accuracy than lower z-values due to proximity to the camera in 2D depth models. Our model is able to predict topology with similar error even in areas that are further away from the camera and more susceptible to the loss of depth perception effect.

V. Discussion

A. Accuracy and Scalable Resolution

The model performs well on frames from the same video that the model was trained on. Performance on the hidden environment indicates that the model was able to identify definitive features for prediction outside the training space, but not enough to reconstruct the scene as accurately. This is likely due to a lack of feature diversity – w.r.t. color, lighting, and tissue type – in the training frames that may have proved useful for extra-set prediction. Compared to previous 3D inference models designed for the surgical space, we present a system that can expand to any arbitrary resolution regardless of image resolution and memory requirements of the prediction. We demonstrate that resolution can be applied as needed in areas with higher densities of biological features. From the surgical perspective, resolution can be independently tuned to create highly detailed representations of tissue or organs as needed by the type of operation. This can enable surgeons to make real-time decisions based on the conditions inside the surgical cavity, resulting in lower blood loss, smaller operative times, and overall more successful patient outcomes.

B. Occupancy Flow

Reconstructing occupancy from each frame from video streams allows for depth flow to be represented temporally. In our work, we build singular mesh representations that capture frame-dependent structural features such as tissue and organs. While this study did not have access to hardware capable of occupancy inference, prior work demonstrates occupancy can be reevaluated well within the

time between consecutive video frames [25]. Within this temporal flow, dynamic tissue movement and deformation can be monitored as a proxy for human interaction within the surgical space. We believe that time series occupancy representations are fundamental to solving adjacent problems in the field of minimally invasive surgery. Due to the robotic interface between the surgeon and patient, important force information is lost. Occupancy deformation around regions of interest containing the tool tip may provide insight into surgeon applied force which can subsequently be delivered back to the surgeon via haptic devices. Force estimation is one of many various obstacles remaining in robot-assisted minimally invasive surgery that our occupancy predictions can help solve.

VI. Future Work

A. Tool Tracking

3D reconstruction in surgical scenes is first and foremost a lens through which surgeons examine patients for a variety of separate tasks. Tasks such as exploration and tool use in traditional open surgery center around the surgeon's field of view. In our study, we maintain the field of view of the camera as a constant. However in practical applications, fields of view are dynamic and often take a third person perspective of the tool in order to view the interactions between tool and biological matter. Older surgical reconstruction methods such as SLAM integrate tool tracking with building structure concurrently. While SLAM does not produce ideal reconstructions, the premise of the algorithm can be pursued in deep learning. A future model would optimize both 3D predictions and tool position in the surgical space to provide greater relevance to medical professionals.

B. Viewpoint Generalization

Previous approaches to 3D reconstruction replicate human stereopsis by utilizing stereo camera systems. While artificial stereoscopic vision falls short of true depth perception due to the Vergence-accommodation conflict, large portions of depth can still be inferred [34]. In the context of a deep learning network, stereoscopic view is identical to simply two independent views. Therefore, we can incorporate multiple camera angles, irrespective of stereo pairing, into the model input to provide more feature points. This generalized approach to collecting visual data on the surgical scene is a logical next step for future work.

VII. Conclusion

Our work approaches the problem of 3D reconstruction in surgical scenes with binary representation. We demonstrate a model that robustly evaluates frame by frame occupancy – capturing both features local to frames in temporally consistent bounds. Quantitative results indicates that the model is capable of not only reconstructing surgical scenes as outlined in this study, but impacting downstream problems in RMIS such as real-time tissue segmentation, tool tracking, and force sensing. Binary occupancy prediction proves a novel, but effective, representational foundation in surgical fields and can further elicit human-robot compatibility in surgery.

References

[1] T. A. Rockall and A. W. Darzi, "Tele-manipulator robots in surgery," *British Journal of Surgery*, vol. 90, no. 6, pp. 641–643, Jun. 2003. [Online]. Available: https://academic.oup.com/bjs/article/90/6/641/6143280

[2] J. H. Palep, "Robotic assisted minimally invasive surgery," *Journal of Minimal Access Surgery*, vol. 5, no. 1, pp. 1–7, Jan. 2009.

[3] J. M. Sackier and Y. Wang, "Robotically assisted laparoscopic surgery: From concept to development," *Surgical Endoscopy*, vol. 8, no. 1, pp. 63–66, Jan. 1994. [Online]. Available: http://link.springer.com/10.1007/BF02909496

[4] G. H. Ballantyne, "The Pitfalls of Laparoscopic Surgery: Challenges for Robotics and Telerobotic Surgery:," *Surgical Laparoscopy, Endoscopy & Percutaneous Techniques*, vol. 12, no. 1, pp. 1–5, Feb. 2002. [Online]. Available: http://journals.lww.com/00129689-200202000-00001

[5] Y.-M. Dion and F. Gaillard, "Visual integration of data and basic motor skills under laparoscopy: Influence of 2-D and 3-D video-camera systems," *Surgical Endoscopy*, vol. 11, no. 10, pp. 995–1000, Oct. 1997. [Online]. Available: http://link.springer.com/10.1007/s004649900510

[6] S. Baum, M. Sillem, J. Ney, A. Baum, M. Friedrich, J. Radosa, K. Kramer, B. Gronwald, S. Gottschling, E. Solomayer, A. Rody, and R. Joukhadar, "What Are the Advantages of 3D Cameras in Gynaecological Laparoscopy?" *Geburtshilfe und Frauenheilkunde*, vol. 77, no. 01, pp. 45–51, Jan. 2017. [Online]. Available: http://www.thieme-connect.de/DOI/DOI?10.1055/s-0042-120845

[7] M. Silvestri, T. Ranzani, A. Argiolas, M. Vatteroni, and A. Menciassi, "A Multi-Point of View 3D Camera System for Minimally Invasive Surgery," *Procedia Engineering*, vol. 47, pp. 1211–1214, 2012. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877705812044335

[8] D. Stoyanov, M. V. Scarzanella, P. Pratt, and G.-Z. Yang, "Real-time stereo reconstruction in robotically assisted minimally invasive surgery," *Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 13, no. Pt 1, pp. 275–282, 2010.

[9] S. Giannarou and G. . Z. Yang, "Tissue deformation recovery with gaussian mixture model based structure from motion," in *Augmented Environments for Computer-Assisted Interventions*, C. A. Linte, J. T. Moore, E. C. S. Chen, and D. R. Holmes, Eds. Heidelberg: Springer, 2012. [Online]. Available: https://doi.org/10.1007/978-3-642-32630-1

[10] X. Lladó, A. D. Bue, A. Oliver, J. Salvi, and L. Agapito, "Reconstruction of non-rigid 3d shapes from stereo-motion," *Pattern Recognition Letters*, vol. 32, 2011. [Online]. Available: https://doi.org/10.1016/j.patrec.2011.02.010

[11] B. Lin, Y. Sun, X. Qian, D. Goldgof, R. Gitlin, and Y. You, "Video-based 3D reconstruction, laparoscope localization and deformation recovery for abdominal minimally invasive surgery: a survey," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 12, no. 2, pp. 158–178, Jun. 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/rcs.1661

[12] P. Mountney, D. Stoyanov, and G.-Z. Yang, "Three-Dimensional Tissue Deformation Recovery and Tracking," *IEEE Signal Processing Magazine*, vol. 27, no. 4, pp. 14–24, Jul. 2010. [Online]. Available: http://ieeexplore.ieee.org/document/5484176/

[13] P. Mountney and G. . Z. Yang, "Motion compensated SLAM for image guided surgery," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*, T. Jiang, N. Navab, J. P. W. Pluim, and M. A. Viergever, Eds. Heidelberg: Springer, 2010. [Online]. Available: https://doi.org/10.1007/978-3-642-15745-5

[14] P. Mountney, D. Stoyanov, A. J. Davison, and G. . Z. Yang, "Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*, R. Larsen, M. Nielsen, and J. Sporring, Eds. Heidelberg: Springer, 2006. [Online]. Available: https://doi.org/10.1007/11866565

[15] O. G. Grasa, J. Civera, and J. M. M. Montiel, "EKF monocular SLAM with relocalization for laparoscopic sequences," in *2011 IEEE International Conference on Robotics and Automation*. Shanghai, China: IEEE, May 2011, pp. 4816–4821. [Online]. Available: http://ieeexplore.ieee.org/document/5980059/

[16] E. Guerra, R. Munguia, and A. Grau, "Monocular SLAM for autonomous robots with enhanced features initialization," *Sensors (Basel, Switzerland)*, vol. 14, no. 4, pp. 6317–6337, Apr. 2014.

[17] B. Lin, A. Johnson, X. Qian, J. Sanchez, and Y. Sun, "Simultaneous Tracking, 3D Reconstruction and Deforming Point Detection for Stereoscope Guided Surgery," in *Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, H. Liao, C. A. Linte, K. Masamune, T. M. Peters, and G. Zheng, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, vol. 8090, pp. 35–44, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-642-40843-4

[18] D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, D. Stoyanov, M. V. Scarzanella, P. Pratt, and G.-Z. Yang, "Real-Time Stereo Reconstruction in Robotically Assisted Minimally Invasive Surgery," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*, T. Jiang, N. Navab, J. P. W. Pluim, and M. A. Viergever, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, vol. 6361, pp. 275–282, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-642-15705-9

[19] C. M. Mateo, J. A. Corrales, and Y. Mezouar, "Hierarchical, Dense and Dynamic 3D Reconstruction Based on VDB Data Structure for Robotic Manipulation Tasks," *Frontiers in Robotics and AI*, vol. 7, p. 600387, Feb. 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2020.600387/full

[20] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction," Apr. 2016, arXiv:1604.00449 [cs]. [Online]. Available: http://arxiv.org/abs/1604.00449

[21] E. Palazzolo, J. Behley, P. Lottes, P. Giguère, and C. Stachniss, "ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals," Aug. 2019, arXiv:1905.02082 [cs]. [Online]. Available: http://arxiv.org/abs/1905.02082

[22] H. Fan, H. Su, and L. Guibas, "A Point Set Generation Network for 3D Object Reconstruction from a Single Image," Dec. 2016, arXiv:1612.00603 [cs]. [Online]. Available: http://arxiv.org/abs/1612.00603

[23] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images," Aug. 2018, arXiv:1804.01654 [cs]. [Online]. Available: http://arxiv.org/abs/1804.01654

[24] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy Networks: Learning 3D Reconstruction in Function Space," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[25] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction," Mar. 2023, arXiv:2302.07817 [cs]. [Online]. Available: http://arxiv.org/abs/2302.07817

[26] Y. Zhang, Z. Zhu, and D. Du, "OccFormer: Dual-path Transformer for Vision-based 3D Semantic Occupancy Prediction," *arXiv preprint arXiv:2304.05316*, 2023.

[27] L. Chen, W. Tang, N. W. John, T. R. Wan, and J. J. Zhang, "Slam-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality," *Computer Methods and Programs in Biomedicine*, vol. 158, p. 135–146, 2018.

[28] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*. Nara, Japan: IEEE, Nov 2007, p. 1–10. [Online]. Available: http://ieeexplore.ieee.org/document/4538852/

[29] D. Recasens, J. Lamarca, J. M. Fácil, J. M. M. Montiel, and J. Civera, "Endo-Depth-and-Motion: Reconstruction and Tracking in Endoscopic Videos using Depth Networks and Photometric Constraints," 2021, publisher: arXiv Version Number: 2. [Online]. Available: https://arxiv.org/abs/2103.16525

[30] Z. Chen, A. Marzullo, D. Alberti, E. Lievore, M. Fontana, O. D. Cobelli, G. Musi, G. Ferrigno, and E. D. Momi, "FRSR: Framework for real-time scene reconstruction in robot-assisted minimally invasive surgery," *Computers in Biology and Medicine*, vol. 163, p. 107121, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482523005863

[31] Y. Wang, Y. Long, S. H. Fan, and Q. Dou, "Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery," no. arXiv:2206.15255, Jun 2022, arXiv:2206.15255 [cs]. [Online]. Available: http://arxiv.org/abs/2206.15255

[32] "Hamlyn Centre Laparoscopic/Endoscopic video datasets," 2012.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015, arXiv:1512.03385 [cs]. [Online]. Available: http://arxiv.org/abs/1512.03385

[34] D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks, "Vergence–accommodation conflicts hinder visual performance and cause visual fatigue," *Journal of Vision*, vol. 8, no. 3, pp. 33–33, Mar. 2008. [Online]. Available: https://doi.org/10.1167/8.3.33