SegmentOR: Obtaining Efficient Operating Room Semantics Through Temporal Propagation

Lennart Bastian(✉), Daniel Derkacz-Bogner, Tony D. Wang, Benjamin Busam, and Nassir Navab

Computer Aided Medical Procedures,
Technical University Munich, Munich, Germany
lennart.bastian@tum.de

Abstract. The digitization of surgical operating rooms (OR) has gained significant traction in the scientific and medical communities. However, existing deep-learning methods for operating room recognition tasks still require substantial quantities of annotated data. In this paper, we introduce a method for weakly-supervised semantic segmentation for surgical operating rooms. Our method operates directly on 4D point cloud sequences from multiple ceiling-mounted RGB-D sensors and requires less than 0.01% of annotated data. This is achieved by incorporating a self-supervised temporal prior, enforcing semantic consistency in 4D point cloud video recordings. We show how refining these priors with learned semantic features can increase segmentation mIoU to 10% above existing works, achieving higher segmentation scores than baselines that use four times the number of labels. Furthermore, the 3D semantic predictions from our method can be projected back into 2D images; we establish that these 2D predictions can be used to improve the performance of existing surgical phase recognition methods. Our method shows promise in automating 3D OR segmentation with a 20 times lower annotation cost than existing methods, demonstrating the potential to improve surgical scene understanding systems.

Keywords: Surgical Scene Understanding · Context-aware Systems · Surgical Phase Recognition · Surgical Data Science

1 Introduction

Automating systems to interpret complex behaviors in surgical operating rooms (OR) has seen a surge of interest in recent years [13,20]. Robot-assisted surgeries

Lennart Bastian and Daniel Derkacz-Bogner contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43996-4_6.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 57–67, 2023.
https://doi.org/10.1007/978-3-031-43996-4_6

have improved patient outcomes by reducing blood loss, recovery periods, and hospitalization times [26,27]. For robotic systems to autonomously interact with hospital staff, surgical tools, or patients, they must attain a sophisticated and detailed understanding of a highly complex environment. To achieve this, robotic detection systems must comprehend high-level surgical phases and granular 3D object semantics and interactions [16,25].

Recent works have established the necessity of combining data from multiple cameras to obtain better coverage of surgical procedures [25,26], as frequent occlusions and obstructions due to personnel and medical equipment obscure important events for individual cameras. Furthermore, a metric 3D semantic understanding is crucial for robotic systems operating and interacting with objects in an environment [31]. In surgical operating rooms, 3D segmentation has previously been approached by fusing semantic RGB predictions from multiple views in 3D under the full supervision of dense 2D labels [16].

However, to adequately represent the distribution of possible events in the surgical domain, large volumes of data must be acquired [27]. Training deep-learning models for automated recognition tasks thus induces an enormous annotation burden, particularly as the privacy-sensitive nature of such materials can prevent annotation outsourcing. Therefore, the surgical data science community actively seeks methods to alleviate this burden, particularly through means of domain-adaptation [24], as well as unsupervised and self-supervised learning [12]. While progress has been made in surgical workflow recognition, methods for 3D surgical scene understanding still require fine-grained labels [25], particularly for semantic segmentation [16].

To this end, we propose SegmentOR, a weakly-supervised indoor semantic segmentation method for 4D multi-view OR datasets. By leveraging the innate temporal consistency of 4D point cloud sequences, we reduce the annotation burden to only a single click per class (about 0.005% of points), decreasing average annotation time per surgical phase from 3 h to 9.6 min while achieving a higher segmentation mIoU than existing methods that use four times the amount of labels. Furthermore, we establish the soundness of semantic predictions from our model for surgical scene understanding by showing that surgical phase recognition performance can be improved using our segmentation predictions as input. Our main contributions can thus be summarized as follows:

- We propose the first 3D weakly-supervised semantic segmentation method for operating room environments and validate it on a manually annotated dataset of 3D point clouds from real surgical acquisitions.
- We demonstrate that various temporal priors can be used to exploit consistency in weakly-supervised semantic segmentation, improving performance to 10% mIoU above baseline methods.
- We show that the semantic outputs from our model can improve the performance of downstream surgical phase recognition methods, formally establishing the link between these two previously disjoint tasks.

– Finally, we release all code and tools, as well as 2577 anonymized and annotated point clouds from the dataset, to advance progress in surgical scene understanding. https://bastianlb.github.io/segmentOR/

2 Related Work

Surgical Scene Understanding. Workflow recognition is pivotal for contextual awareness in operating room (OR) intelligent systems. Activity recognition has been achieved for single-frame [27,29] and multi-view acquisitions [26], including laparoscopic views and ceiling-mounted cameras [6,7].

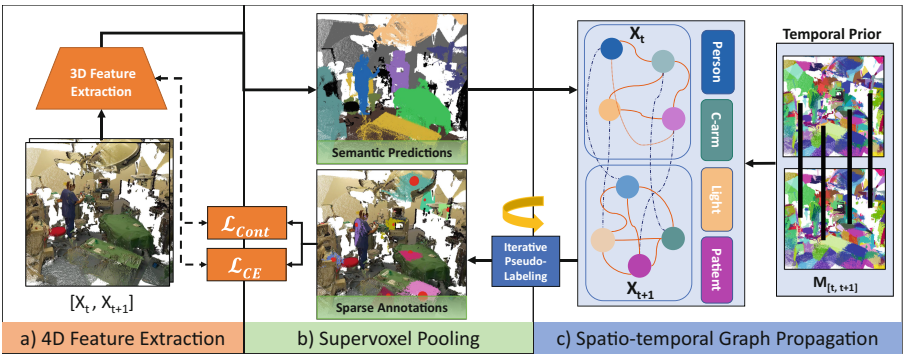

Fig. 1. Proposed architecture. SegmentOR extracts semantic and class-specific features from a sequence of point clouds (a). Sparse labels (in red) are expanded to their nearest supervoxel cluster (b) and used as supervision for learning the segmentation task. In contrast to previous works, we propose to incorporate a prior to establish a temporal consistency between the pooled semantic features in a point cloud sequence (c), enabling spatiotemporal pseudo-label propagation across timesteps. (Color figure online)

Modeling OR activities also requires semantic understanding [16]. Semantic scene graphs offer a detailed approach to surgical procedure modeling [25]. While future intelligent OR systems will employ semantics, manual labeling is time-consuming. However, recent progress in weakly-supervised semantic segmentation, such as one-thing-one-click (OTOC) [19], can decrease this burden, requiring only a single annotation per semantic class [18,19,30].

The effectiveness of such weakly-supervised segmentation methods in dynamic OR environments remains unclear. Unlike static indoor datasets like ScanNet [8], dynamic ORs blur the geometric class separation due to human-object interaction. Moreover, 3D surgical acquisitions, unlike static indoor reconstructions, are fragmented due to static sensor positions and severe occlusions. Imprecise 3D registration or temporal synchronization creates artifacts that further complicate 3D modeling, worsened by dynamic non-rigid movements (see suppl. for examples).

Temporal Modeling. Optical flow effectively extracts movement from image sequences [1,9]. Self-supervised methods have recently offered scene flow extraction from LiDAR point clouds [15,23], but few ground-truth flow annotated datasets exist, leading to potential generalization issues [21,22].

Few have combined temporal consistency with 3D semantic segmentation for dynamic point cloud sequences outside autonomous driving settings [4,10,14,28]. These methods typically rely on dense ground truth labels or require multi-stage label propagation methods and pre-training. Our approach guides temporal label propagation with unsupervised priors, resolving this cold-start problem cost-effectively.

3 Method

Problem Setting. Given a point cloud $X_t \in \mathbb{R}^{N\times 3}$ in a temporal sequence $t \in [T]$, we seek to predict a semantic label $\hat{y}_i \in Y_t$ for each point $p_i \in X_t$. In contrast to the supervised setting where dense ground truth labels $y_i \in Y_t$ are available for every point p_i, we infer dense semantic labels in an unseen test sequence by training on sparsely annotated point clouds. We refer to this setting as "weakly-supervised", meaning ground truth train annotations consist of a randomly chosen point per class (see Fig. 1b). This results in an average of 0.005% of ground truth labels compared to full supervision.

Inspired by recent works, we assume semantic instances in a point cloud X_t adhere to geometric boundaries and partition the point cloud into a set of supervoxels $\mathcal{S}_t$ [18,19]. For $S_j, S_{j'} \in \mathcal{S}_t$ we have $S_j \cap S_{j'} = \emptyset$, and $\cup_{j=1}^{N} S_j = X_t$. This allows us to represent a group of points with a single feature, increasing the level of supervision obtained from a single "click", and enabling efficient label propagation. Due to the sparse nature of the click annotations, most supervoxels in each point cloud are unlabeled (see Fig. 1c). By propagating learned information into unlabeled supervoxels, the level of supervision can be drastically increased through pseudo-labeling. For clarity, we use indices i for points and j for supervoxels [19]. Indices t are used to indicate timesteps.

Proposed Method. Following the approach of OTOC [19], we train a sparse 3D-UNet [5] F_Θ to model the semantic class of each point p_i, $Y_t = F_\Theta(X_t)$ with a cross-entropy loss $\mathcal{L}_{CE}$, and an identically structured relation network R_Ψ to predict a category specific embedding $R_t = R_\Psi(X_t)$ with a contrastive loss $\mathcal{L}_{Cont}$. The point-wise embeddings obtained from both networks are accumulated using mean-pooling over each supervoxel S_j. The pooled feature embeddings f_j, r_j can then be used to construct a fully-connected graph G_i to propagate labels to unlabeled supervoxels. This is achieved by maximizing the expectation E of unlabeled supervoxels given the complete supervoxel set $\mathcal{S}$:

$$E(Y|\mathcal{S}) = \sum_{j'} \psi_u(y_{j'}|\mathcal{S}, F_\Theta) + \sum_{j'} \psi_p(y_j, y_{j'}|\mathcal{S}, R_\Psi, F_\Theta) \tag{1}$$

where ψ_u represents the pooled class predictions, and ψ_p is a pairwise similarity between supervoxels $S_j, S_{j'}$ [19]. By measuring the similarity between supervoxels in this manner, semantic information in the two networks F_Θ and R_Ψ can be

propagated to unlabeled supervoxels iteratively through pseudo-labeling, using a likelihood threshold of, e.g., $E(Y|S_j) \geq 0.90$. Setting a high confidence threshold reduces incorrect pseudo-labels, which would negatively impact subsequent training iterations.

OTOC [19] considers all supervoxel pairs in a single static acquisition as candidates during this expansion. An intuitive way to extend the graph propagation in the temporal dimension would be to pool all supervoxels over a pair of frames X_t and X_{t+1}, creating a fully connected graph over both supervoxel sets. We refer to this method as $OTOC{+}T$, as the original method uses only spatial context. Aside from being computationally expensive, this naive approach does not consider that the nearest supervoxel in an adjacent timestep is highly likely to describe a similar region in the point cloud.

To further improve upon this idea, we propose to enforce temporal consistency through the use of a supervoxel matching matrix $M_{[t,t+1]} \in \mathbf{R}^{m \times n}$ where $|\mathcal{S}_t| = m$ and $|\mathcal{S}_{t+1}| = n$ are the dimensions of the respective supervoxel sets, reducing computational complexity to an additional comparison per supervoxel instead of n as with the $OTOC{+}T$. An entry $m_{j,j'} \in M_{[t,t+1]}$ indicates the probability that supervoxels S_j and $S_{j'}$ describe a similar region across time steps. Intuitively, this can establish consistency between the pair of point clouds by considering matched supervoxels from a different timestep $X_{t'}$ as pseudo-label candidates. To initialize the matching matrix, we explore how nearest neighbor, unsupervised optical [9] and scene flow [23] priors can improve temporal pseudo-label propagation (see suppl. sec. 1 for mathematical details).

After initialization through any of these priors, we propose to update the matching iteratively during training, establishing a link between temporal consistency and semantic understanding. To strengthen this dynamic and account for potentially incorrect matches, we additionally incorporate relation net R_Ψ features to refine the supervoxel matching matrix M. Formally, we can define the matching update for a single entry $m_{j,j'} \in M_{[t,t+1]}$ as follows:

$$p(S_{j'}|S_j) = \lambda\, p(S_{j'}|S_j, M_{[t,t+1]}) + (1-\lambda)\, p(S_{j'}|S_j, R_\Psi) \tag{2}$$

The updated matching is the supervoxel with the highest matching probability, i.e., $\hat{m}_{j,j'} = \arg\max_{S_j \in Y_{t+1}} p(S_{j'}|S_j, M_{[t,t+1]})$. We can then additionally regularize the graph propagation (Eq. 1) using the updated matching matrix for the merged supervoxel sets $\hat{\mathcal{S}} = \mathcal{S}_t \cup \mathcal{S}_{t+1}$:

$$E(Y|\hat{\mathcal{S}}) = \sum_{j'} \psi_u(y_{j'}|\hat{\mathcal{S}}, F_\Theta) + \sum_{j'} \psi_p(y_j, y_{j'}|\hat{\mathcal{S}}, M_{[t,t+1]}) \tag{3}$$

where ψ_u describes the probability of S_j being assigned label $y_{j'}$ based on the prediction $f_j = F_\Theta(S_j)$. ψ_p describes the pairwise similarity between two supervoxels additionally based on $r_j = R_\Psi(S_j)$, mean supervoxel color c_j, and mean coordinate p_j.

$$\psi_u(y_{j'}|S_j, F_\Theta) = -\log P(y_{j'}|S_j, F_\Theta) \tag{4}$$

$$\psi_p(y_j, y_{j'}|S_j, M_{[t,t+1]}) = m_{j,j'} \cdot e^{\{-\|c_j - c_{j'}\|_2^2 - \|p_j - p_{j'}\|_2^2 - \|f_j - f_{j'}\|_2^2 - \|r_j - r_{j'}\|_2^2\}} \tag{5}$$

4 Experiments

Dataset Description. All experiments are carried out on an existing dataset of 18 laparoscopic surgeries [2,3]. The full dataset consists of RGB-D video acquisitions from four co-registered ceiling-mounted Azure Kinect cameras, containing 582,000 images per camera, with video annotations for 8 surgical workflow phases. We uniformly split a subset of the fused point cloud data into training and validation sets, ensuring similar distributions of non-overlapping surgeries, camera calibrations, and class distributions. We use 1500 point clouds for training and 1077 for validation. We additionally manually annotate these point clouds with segmentation labels covering 12 medically relevant classes.

Each training split contains around 500 sparsely annotated point clouds, with 5436 click annotations on average per split. The densely labeled validation annotations comprise approximately 93% of the on average one million points per point cloud. To reduce the bias from a subjective annotation, we employed four different data annotators for a total annotation time of ∼115 h. To measure the robustness of our method, we split the train and validation sets into 3 non-overlapping splits, referring to this as "cross-validation" despite the train and validation splits having different types of labels (sparse click and dense, respectively). For more details on data annotation and splits, as well as qualitative examples, please refer to the supplementary materials.

Experimental Setup. In contrast to OTOC [19], which relies on additional ground truth instance segmentation as input [8], we use a heuristic over-segmentation approach to generate supervoxels [17]. All experiments use the same hyperparameters unless otherwise noted. The relation and feature networks were pre-trained on ScanNet [8]. We then fine-tune on our dataset for four iterations, with pseudo-label propagation occurring after each iteration. We report standardized segmentation metrics. The average training time on an NVIDIA A40 GPU was 7.45 h, using PyTorch 1.13.1 with CUDA 11.7.

5 Results and Discussion

Experiment 1: Baseline Comparisons and Color Ablation. To quantify the overall impact of temporal guidance in the weakly-supervised setting, we compare SegmentOR with the baseline OTOC [19] over three-fold cross-validation. As color information may be unavailable in OR datasets due to privacy concerns, we perform additional experiments without RGB features to contextualize the performance in such a setting. Furthermore, we explore how different unsupervised priors can improve semantic predictions, namely (i) nearest neighbor matching, (ii) optical flow matching, and (iii) scene flow matching. For each prior, the matching matrix $M_{[t,t+1]}$ is then initialized based on obtained

Table 1. Main Results. Comparison of our best proposed model, using a nearest neighbor temporal prior with a learned update, with OTOC [19] in a three-fold cross-validation.

Method	Color	mIoU (%)	F1 (%)	Recall (%)	Precision (%)	Accuracy (%)
OTOC	✗	63.29 ± 3.49	75.51 ± 3.76	75.88 ± 4.07	76.80 ± 3.88	75.88 ± 4.07
	✓	62.40 ± 0.01	74.59 ± 0.01	75.02 ± 0.54	76.19 ± 0.20	75.02 ± 0.54
SegmentOR	✗	69.32 ± 2.05	70.57 ± 1.06	79.88 ± 2.06	79.96 ± 2.49	79.88 ± 2.05
	✓	$\mathbf{72.99 \pm 0.19}$	$\mathbf{82.77 \pm 0.13}$	$\mathbf{83.25 \pm 0.71}$	$\mathbf{83.31 \pm 1.08}$	$\mathbf{83.25 \pm 0.71}$

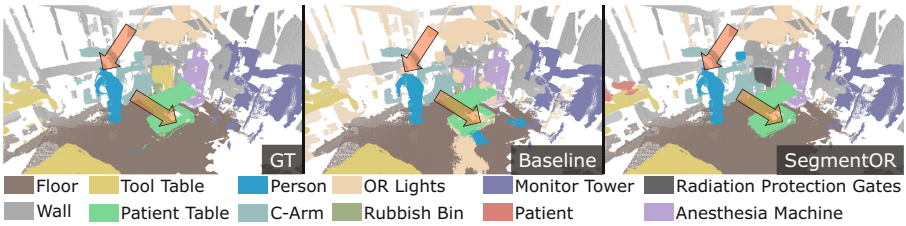

Fig. 2. Qualitative Results. Comparison between the ground truth annotation (GT), the OTOC baseline, and SegmentOR. SegmentOR demonstrates improved segmentation mIOU for moving classes such as Person (red arrow) but also for static classes such as Patient Table (orange arrow). Best viewed digitally.

flow. Interestingly, initializing the matching matrix with the nearest neighbor prior outperformed optical and scene flow initialization. Using any single prior resulted in improvements over *OTOC+T*. We refer to the supplementary materials for a detailed description and ablation of the flow priors.

Results: Table 1 shows that SegmentOR consistently outperforms the baseline both with and without RGB features when initialized with a nearest neighbor temporal prior and a learned update according to Eq. 2. The addition of RGB features marginally impacts the baseline, with a difference of only 0.9% mIoU. Without color, SegmentOR outperforms OTOC by ∼6% mIoU, making it more suitable for privacy-constrained setups. This improves to ∼10% for colored point clouds, indicating that the proposed learned temporal matching can leverage color similarities across time steps more effectively. The most significantly moving entities in ORs are surgical staff, who tend to wear gowns with specific and consistent colors. The presence of color features could influence the ability to distinguish humans from other objects more consistently (Fig. 2). This is supported by a segmentation mIoU increase of over 15% concerning the human class for SegmentOR (see supplementary for all class distributions).

Experiment 2: Number of Clicks. A model's performance should theoretically increase with the level of supervision. We thus quantify the impact of adding up to three additional clicks on OTOC's performance. This experiment is performed on the first training and validation split, using a varying number of click annotations per class.

Table 2. To assess the trade-off between annotation % and performance, we train the baseline OTOC method with up to four times the amount of "click" annotations.

# clicks	annotated (%)	mIoU (%)	F1 (%)	Recall (%)	Precision (%)	Accuracy (%)
1	0.0054	62.60	74.91	74.45	77.86	74.40
2	0.0108	69.01	79.92	79.86	81.13	79.82
3	0.0163	**70.83**	**81.45**	**81.57**	**82.12**	**81.56**
4	0.0217	70.56	80.01	79.13	81.75	79.17

Results: Increasing annotations by three or four times lead to an improvement of approximately 8% mIoU (see Table 2). The performance saturates with three clicks. Notably, the baseline does not achieve the 73% mIoU of the proposed method, even with increased supervision. This could suggest that temporal consistency not only increases the supervision signal but enables a more robust overall feature representation.

Experiment 3: Application to Surgical Phase Recognition. To further assess the quality of our semantic predictions, we evaluate their impact on surgical workflow analysis (see Table 3). We use a ResNet50 [11] backbone and perform four-fold cross-validation using random splits over different surgeries of the complete, larger RGB-D video dataset. We use our best-performing segmentation model to infer the semantic predictions for each of the 582k fused point clouds (inference @12.5 fps), projecting them back into the two cameras. We then compare the performance of raw RGB, depth, and semantic labels inputs against fusing the latter two via late fusion [29].

Table 3. Downstream Surgical Phase Recognition. We evaluate the capability of a ResNet50 [11] as a surgical phase recognition backbone based on 3 modalities. Accuracy and mAP are reported for two differently placed surgical workflow cameras

Input	Camera 01		Camera 02	
	Accuracy	mAP	Accuracy	mAP
Semantic	61.98 ± 7.89	70.19 ± 8.44	52.48 ± 5.19	58.34 ± 6.74
Depth	69.19 ± 1.46	77.95 ± 2.82	63.62 ± 6.45	67.56 ± 5.98
RGB	$\mathbf{76.04 \pm 4.27}$	$\mathbf{88.71 \pm 3.61}$	$\mathbf{73.74 \pm 1.85}$	$\mathbf{85.44 \pm 3.84}$
Depth + Semantic	74.98 ± 5.62	83.34 ± 4.08	64.29 ± 5.35	71.76 ± 4.76

Results: Consistent with previous works [2,29], RGB achieves the best performance across both cameras. Semantic predictions alone yield a performance well below RGB or depth. However, the fact that noisy semantic predictions from our network (achieving 73% mIoU) can be used for this challenging task demonstrates the benefits of our segmentation outputs for surgical scene understanding. Furthermore, when combined with depth through late fusion, results

are improved by nearly 6% and 0.6% accuracy over depth for the surgical and workflow cameras, respectively. This suggests that segmentation maps could substitute RGB features when unavailable due to privacy reasons [2,27].

Conclusion. This work presents a novel semantic segmentation method for surgical scene understanding that significantly reduces the annotation burden by leveraging the temporal consistency of point cloud sequences. We demonstrate the effectiveness of our approach on point clouds from a surgical phase recognition dataset, which we enrich with manual 3D annotations. By incorporating self-supervised temporal priors, our method achieves a high segmentation mIoU of 73.10% using only 0.005% of annotated points. Furthermore, we establish a formal link between semantic segmentation and workflow analysis by demonstrating that our semantic predictions benefit downstream surgical phase recognition methods. Finally, we release all anonymized point clouds, annotations, and code used to ease the deployment of context-aware systems in surgical environments.

Acknowledgements. This work was funded by the German Federal Ministry of Education and Research (BMBF), No.: 16SV8088 and 13GW0236B. We additionally thank the J&J Robotics & Digital Solutions team for their support. Furthermore, we thank Ruiyang Li for supporting the point cloud annotation. Code and data can be found at: https://bastianlb.github.io/segmentOR/.

References

1. Baker, S., Matthews, I.: Lucas-Kanade 20 years on: a unifying framework. Int. J. Comput. Vision **56**, 221–255 (2004)
2. Bastian, L., et al.: Know your sensors-a modality study for surgical action classification. Comput. Methods Biomech. Biomed. Eng.: Imaging Vis. **11**, 1–9 (2022)
3. Bastian, L., Wang, T.D., Czempiel, T., Busam, B., Navab, N.: DisguisOR: holistic face anonymization for the operating room. Int. J. Comput. Assist. Radiol. Surg. 1–7 (2023)
4. Choy, C., Gwak, J., Savarese, S.: 4D spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3075–3084 (2019)
5. Spconv Contributors: Spconv: spatially sparse convolution library (2022). https://github.com/traveller59/spconv
6. Czempiel, T., et al.: TeCNO: surgical phase recognition with multi-stage temporal convolutional networks. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 343–352. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_33
7. Czempiel, T., Sharghi, A., Paschali, M., Navab, N., Mohareri, O.: Surgical workflow recognition: from analysis of challenges to architectural study. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022, Part III. LNCS, vol. 13803, pp. 556–568. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-25066-8_32
8. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: richly-annotated 3D reconstructions of indoor scenes. In: CVPR, pp. 5828–5839 (2017)

9. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Bigun, J., Gustavsson, T. (eds.) SCIA 2003. LNCS, vol. 2749, pp. 363–370. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-45103-x_50
10. Hanyu, S., Jiacheng, W., Hao, W., Fayao, L., Guosheng, L.: Learning spatial and temporal variations for 4D point cloud segmentation. arXiv preprint arXiv:2207.04673 (2022)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
12. Liu, M., Zhou, Y., Qi, C.R., Gong, B., Su, H., Anguelov, D.: Less: Label-efficient semantic segmentation for lidar point clouds. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022, Part VII. LNCS, vol. 13699, pp. 70–89. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19842-7_5
13. Kennedy-Metz, L.R., et al.: Computer vision in the operating room: opportunities and caveats. IEEE Trans. Med. Robot. Bionics **3**(1), 2–10 (2020)
14. Kochanov, D., Ošep, A., Stückler, J., Leibe, B.: Scene flow propagation for semantic mapping and object discovery in dynamic street scenes. In: IROS, pp. 1785–1792. IEEE (2016)
15. Li, R., Zhang, C., Lin, G., Wang, Z., Shen, C.: RigidFlow: self-supervised scene flow learning on point clouds by local rigidity prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16959–16968 (2022)
16. Li, Z., Shaban, A., Simard, J.G., Rabindran, D., DiMaio, S., Mohareri, O.: A robotic 3D perception system for operating room environment awareness. arXiv:2003.09487 [cs] (2020)
17. Lin, Y., Wang, C., Zhai, D., Li, W., Li, J.: Toward better boundary preserved supervoxel segmentation for 3D point clouds. ISPRS J. Photogram. Remote Sens. **143**, 39–47 (2018). https://www.sciencedirect.com/science/article/pii/S0924271618301370. iSPRS Journal of Photogrammetry and Remote Sensing Theme Issue "Point Cloud Processing"
18. Liu, M., Zhou, Y., Qi, C.R., Gong, B., Su, H., Anguelov, D.: LESS: label-efficient semantic segmentation for lidar point clouds. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13699, pp. 70–89. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19842-7_5
19. Liu, Z., Qi, X., Fu, C.W.: One thing one click: a self-training approach for weakly supervised 3d semantic segmentation. In: CVPR, pp. 1726–1736 (2021)
20. Maier-Hein, L., et al.: Surgical data science-from concepts toward clinical translation. Med. Image Anal. **76**, 102306 (2022)
21. Mayer, N., et al.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: CVPR, pp. 4040–4048 (2016)
22. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR (2015)
23. Mittal, H., Okorn, B., Held, D.: Just go with the flow: self-supervised scene flow estimation. In: CVPR, pp. 11177–11185 (2020)
24. Mottaghi, A., Sharghi, A., Yeung, S., Mohareri, O.: Adaptation of surgical activity recognition models across operating rooms. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022, Part VII. LNCS, vol. 13437, pp. 530–540. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_51
25. Özsoy, E., Örnek, E.P., Eck, U., Czempiel, T., Tombari, F., Navab, N.: 4D-OR: semantic scene graphs for or domain modeling. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13437, pp. 475–485. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_45

26. Schmidt, A., Sharghi, A., Haugerud, H., Oh, D., Mohareri, O.: Multi-view surgical video action detection via mixed global view attention. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 626–635. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_60
27. Sharghi, A., Haugerud, H., Oh, D., Mohareri, O.: Automatic operating room surgical activity recognition for robot-assisted surgery. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 385–395. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_37
28. Shi, H., Wei, J., Li, R., Liu, F., Lin, G.: Weakly supervised segmentation on outdoor 4D point clouds with temporal matching and spatial graph propagation. In: CVPR, pp. 11840–11849 (2022)
29. Twinanda, A.P., Winata, P., Gangi, A., Mathelin, M., Padoy, N.: Multi-stream deep architecture for surgical phase recognition on multi-view RGBD videos. In: Proceedings of the M2CAI Workshop MICCAI, pp. 1–8 (2016)
30. Yang, C.K., Wu, J.J., Chen, K.S., Chuang, Y.Y., Lin, Y.Y.: An mil-derived transformer for weakly supervised point cloud segmentation. In: CVPR, pp. 11830–11839 (2022)
31. Yousif, K., Bab-Hadiashar, A., Hoseinnezhad, R.: An overview to visual odometry and visual SLAM: applications to mobile robotics. Intell. Industr. Syst. **1**(4), 289–311 (2015)