TopoOR: A Unified Topological Scene Representation for the Operating Room

Tony Danjun Wang[1,2⋆], Ka Young Kim[4], Tolga Birdal[3], Nassir Navab[1,2], and Lennart Bastian[1,2,3]

[1] Technical University of Munich, Munich, Germany
[2] Munich Center for Machine Learning, Germany
[3] Imperial College London, United Kingdom
[4] Kyung Hee University, South Korea

Abstract. Surgical Scene Graphs abstract the complexity of surgical operating rooms (OR) into entities and their relations, but existing paradigms suffer from structural limitations. Frameworks that predominantly rely on pairwise message passing or tokenized sequences flatten the manifold geometry inherent to the connectivity and lose information in the process. We introduce TopoOR, a new paradigm that models multimodal operating rooms as a higher-order topological structure, innately preserving pairwise and group relationships. By lifting interactions between entities into higher-order topological cells, TopoOR natively models complex dynamics and multimodality present in the OR. This topological representation subsumes traditional scene graphs, thereby offering strictly greater expressivity. We also propose a higher-order attention mechanism that explicitly preserves manifold structure and modality-specific features throughout hierarchical relational attention. In this way, we circumvent combining 3D geometry, audio, and robot kinematics into a single joint latent representation, preserving the precise multimodal structure required for safety-critical reasoning, unlike existing methods. Extensive experiments demonstrate that our approach outperforms traditional graph and LLM-based baselines across sterility breach detection, robot phase prediction, and next-action anticipation.

Keywords: Surgical Data Science · Topological Deep Learning · Surgical Domain Models · Workflow Recognition · Scene Graphs.

1 Introduction

The overarching goal of Surgical Data Science (SDS) is to improve patient outcomes and procedural efficiency by developing computational models of the surgical operating room (OR) [5,20,22]. Early work applied machine learning models [25,19] for tasks such as endoscopic phase recognition [29,10], skill assessment [31], and tool tracking [6]. Moving beyond these isolated tasks, action triplets model joint interactions between an instrument and target through an

⋆ Corresponding author: tony.wang@tum.de

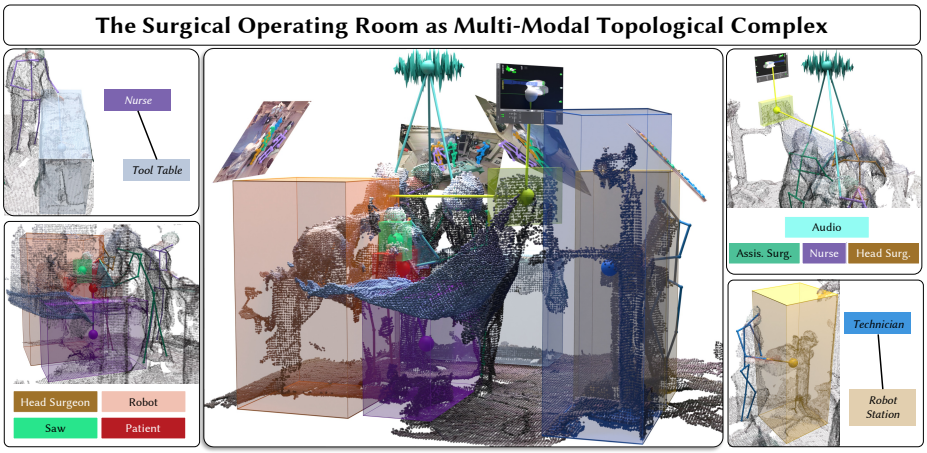

Fig. 1: We model surgical ORs as higher-order structures. Our framework explicitly instantiates *cells* anchored in 3D for physical entities, such as the nurse and surgical robot, as well as data sources from diverse modalities, including audio. Combining these elements into a higher-order structure enables modeling of the unified interaction among the head surgeon, robot, saw, and patient (bottom left), while preserving the complex, multi-actor dynamics of surgical procedures.

action (e.g., ⟨ grasper, retract, gallbladder ⟩) [18,28]. Surgical scene graphs (SSGs) were later conceived to formally structure the complexity of the entire operating theater by encoding relationships among humans, tools, and equipment into a structured, holistic representation [17,16,23]. However, a significant gap remains between their theoretical ambition and practical implementation.

We posit, for the first time, that surgical procedures are irreducibly *polyadic*. For instance, during a robotic-assisted bone resection [24], the head surgeon dynamically guides a robotic arm and surgical saw to resect the patient's anatomy while continuously adjusting based on visual feedback from a navigation monitor (see fig. 1). Standard graphs artificially fragment this unified loop into isolated *dyadic* links (e.g., Surgeon–Robot, Surgeon–Monitor, Saw–Patient), stripping away the joint spatial and kinematic constraints that collectively describe the interaction [33]. The resulting representation is strictly less expressive than one that models it natively [14], with potential downstream impacts on workflow understanding [10,2] and ultimately patient safety.

Another limitation is semantic. The multimodal data that characterize the OR—including articulated 3D human motion in SE(3) [34,3], robot joint kinematics [1], audio spectrograms [9], and RGB appearance features—each reside on geometric manifolds with distinct characteristics. The relationships among these entities inherit this geometry: the space of valid surgical scene configurations forms a high-dimensional, coupled data manifold that *cannot* be faithfully represented in a single latent space while preserving its manifold structure. VLM-based

approaches attempt to sidestep explicit graph construction by mapping heterogeneous data to a shared latent representation [24], but tokenizing forces this inherently non-Euclidean data through a semantic bottleneck that discards its metric and topological structure. As a result, these models flatten the manifold geometry of the surgical scene, stripping away structure essential to downstream clinical tasks [4] and surgical safety [15].

To overcome these limitations, we propose a framework for surgical reasoning rooted in algebraic topology, designed to model the complexity of the operating room. Rather than relying on pairwise interactions, we model the scene as a *combinatorial complex* (CC) [14,12], preserving the structural and semantic meaning of higher-order relationships. To reason over these structures, we employ higher-order attention networks (HAT) which distribute and aggregate messages directly across the higher-order incidence structure of the complex, establishing a learned hierarchical representation of the surgical scene. We evaluate our framework on the multimodal MM-OR dataset, jointly optimizing next-action anticipation and robot-phase prediction while enabling zero-shot, rule-based sterility breach detection. We show that our topological representation subsumes traditional approaches: when supervised with flattened scene graph labels, it predicts the conventional tokenized format more effectively than existing baselines. Our primary contributions can be summarized as follows:

- **TopoOR**: We introduce a unified topological framework for surgical scenes, modeling the operating room as a higher-order structure. Our higher-order attention mechanism preserves multimodal geometry and captures dynamics between entities without losing structure and semantic meaning.
- **Empirical Performance**: TopoOR outperforms existing graph methods on the MM-OR dataset through a higher-order attention (HAT) mechanism that preserves structure essential for downstream clinical tasks.
- **Expressiveness**: We demonstrate that our higher-order representation subsumes traditional scene graphs by optimizing directly for downstream tasks while also decoding flattened tokenized relationships.

2 Methodology

To overcome the limitations of graph-based surgical domain models, we model the collective relationships between individuals and entities in the OR as a higher-order topological structure that encapsulates heterogeneous multi-actor interactions. To this end, we formulate the preliminaries of our approach (section 2.1), define our Higher-Order Attention Network (HAT) (section 2.2), detail the construction of the multimodal combinatorial complex (section 2.3), and finally outline the implementation details and downstream multi-task inference (section 2.4). A visual overview is provided in fig. 2.

2.1 Preliminaries

To model the heterogeneous nature of the OR in a unified manner, we abstract the surgical scene using the topological deep learning (TDL) framework [14,26].

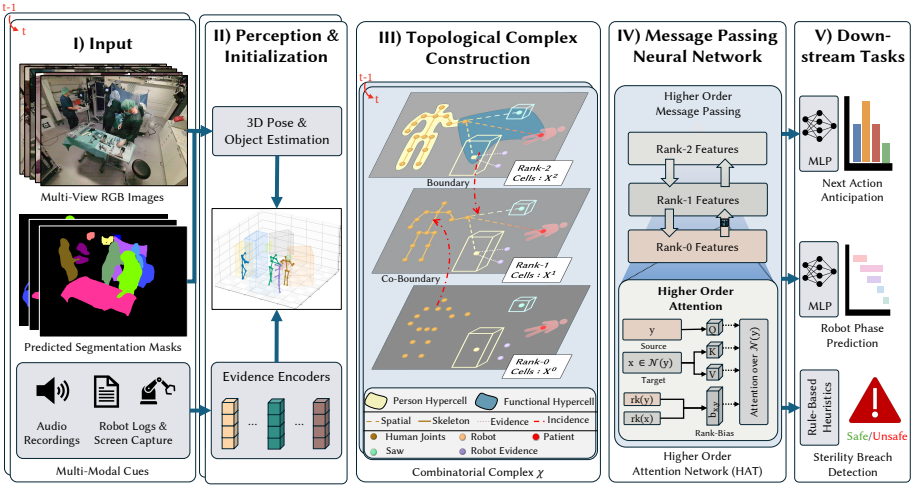

Fig. 2: Overview of TopoOR. Given multi-modal sensory inputs over a temporal window (I), 3D entities and evidence features are initialized (II) and abstracted into a CC $\mathcal{X}$ (III), where rank-2 cells encode the higher-order (polyadic) group relations. Higher-order attention is computed across incidence neighborhoods $x \in \mathcal{N}(y)$, using a learnable rank-bias to preserve structural heterogeneity (IV). Pooled representations are routed to downstream tasks, enabling, e.g., simultaneous next-action anticipation and robot-phase prediction (V).

Definition 1 (Combinatorial Complex [12,14]). *A combinatorial complex (CC) consists of a finite set of cells X and a rank function $\mathrm{rk} : X \to \mathbb{Z}_{\geq 0}$ which partitions the complex into k-dimensional sets X^k. The CC is endowed with a partial ordering $\preceq$ denoting the face relation, where $x \preceq y$ indicates that cell $x \in X^p$ is on the boundary of cell $y \in X^q$ with $p < q$.*

CCs generalize graphs and other higher-order structures by combining features of cell complexes (hierarchy among relationships) and hypergraphs (arbitrary set-type relations) [12,11]. Consider, for instance, that a graph is a CC with rank-0 nodes X^0 connected by rank-1 edges X^1. To realize this, let us define how cells of different dimensions interact through incidence relations:

Definition 2 (Incidence Neighborhoods). *For any cell $y \in X^k$, we define its boundary neighborhood $\mathcal{B}(y) = \{x \in X^p \mid x \preceq y,\ p < k\}$ and its co-boundary neighborhood $\mathcal{C}(y) = \{z \in X^q \mid y \preceq z,\ q > k\}$. The full incidence neighborhood is their union: $\mathcal{N}(y) = \mathcal{B}(y) \cup \mathcal{C}(y)$.*

2.2 Higher-Order Attention Network (HAT)

Next, we introduce Higher-Order Attention Networks (HAT), which generalize GAT [32] from graph neighborhoods to the incidence structure of combinatorial complexes [7,13,14,11], as an inductive bias over higher-order surgical scenes.

Definition 3 (Higher-Order Attention Layer). *Let $\mathcal{X}$ be a CC with cell features $\{\mathbf{h}_y^{(l)}\}_{y\in X}$ at layer l. For each cell $y \in X^k$, HAT computes:*

$$\mathbf{h}_y^{(l+1)} = \mathbf{h}_y^{(l)} + \sum_{x\in\mathcal{N}(y)\cup\{y\}} \alpha_{yx}^{(l)}\, \mathbf{W}_V^{(l)}\, \mathbf{h}_x^{(l)}, \tag{1}$$

where the attention coefficients are given by

$$\alpha_{yx}^{(l)} = \underset{x'\in\mathcal{N}(y)\cup\{y\}}{\mathrm{Softmax}} \left(\frac{\langle\, \mathbf{W}_Q^{(l)}\mathbf{h}_y^{(l)},\ \mathbf{W}_K^{(l)}\mathbf{h}_x^{(l)} \,\rangle}{\sqrt{d_k}} + b_{\mathrm{rk}(y),\,\mathrm{rk}(x)} \right), \tag{2}$$

with rank-pair bias

$$b_{\mathrm{rk}(y),\,\mathrm{rk}(x)} = \phi\big(\, \mathbf{e}_{\mathrm{rk}(y)} \odot \mathbf{e}_{\mathrm{rk}(x)} \,\big), \tag{3}$$

where $\mathbf{e}_r \in \mathbb{R}^{d_r}$ are learnable rank embeddings and $\phi : \mathbb{R}^{d_r} \to \mathbb{R}^H$ is a linear projection producing per-head biases.

Information in HAT flows primarily along the incidence structure of $\mathcal{X}$: boundary cells ($\mathrm{rk}(x) < \mathrm{rk}(y)$) propagate entity-level features upward, while co-boundary cells ($\mathrm{rk}(x) > \mathrm{rk}(y)$) distribute aggregated group context downward. The rank-pair bias $b_{\mathrm{rk}(y),\,\mathrm{rk}(x)}$ modulates information flow based on the topological relationship between source and target cells. For nodes of distinct rank in $\mathcal{X}$, we can use features from various modalities, detectors, or sensors. By exchanging information between these heterogeneous spaces based on the rank function $\mathrm{rk}(x)$, HAT retains the structural origin and information of each feature. This allows the model to differentiate between, e.g., human kinematics modeled on X^0 and aggregated multi-actor behavior in X^2.

2.3 Multimodal Graph Construction

To construct our higher-order surgical scene representation, we build the CC $\mathcal{X}$ by using spatial thresholding and imposing clinical priors.

Rank-0 Cells (X^0) for Physical Entities and Evidence. These atomic nodes represent human anatomical joints (acquired via 3D pose estimation with semantic embeddings) and 3D-localized object entities (initialized via 3D spatial boundaries). We also establish *auxiliary evidence nodes* to integrate additional modalities, including processing robot logs with a language model [27], monitoring screens with a small CNN, and spatial audio with an audio encoder [9].

Rank-1 Cells (X^1) for Interactions. We construct intra-entity skeleton edges using predefined human kinematic trees, and inter-entity spatial edges formed dynamically when physical proximity falls below a threshold. We further enforce predefined domain-specific semantic links (e.g., Technician–Robot) regardless of spatial distance, and connect all physical entities to their corresponding auxiliary evidence nodes.

Rank-2 Cells (X^2) for Higher-Order Behavior. To capture irreducible group dynamics, we build rank-2 cells incident to the underlying rank-1 edges.

Table 1: Performance evaluation on MM-OR [24] measured in Macro F1-Score. Best results are **emphasized**.

Method	F1-Score (↑)				
	Sterility Breach	Next Action	Robot Phase		
LLM-Based					
MM2SG [24]	55.00	35.40	56.90		
Latent & Relational Models					
Vanilla Transformer	**76.83**	34.80	65.29		
SurgLatentGraph [8]	**76.83**	37.46	64.61		
TopoOR (Ours)	**76.83**	**41.10**	**73.53**		

Table 2: Scene graph relation prediction F1 when reducing our topological representation to a string-based format [24].

Method	F1-Score (↑)
MM2SG [24]	52.90
Ours (Decision Tree)	43.72
Ours (Learned Head)	**61.30**

This includes *person cells* that topologically aggregate a single individual's skeleton, and *functional cells* that encapsulate fine-grained multi-actor events (e.g., {Surgeon, Robot, Saw, Patient} complexes) guided by clinical priors. This explicit modeling creates a shared topological neighborhood, allowing involved entities to update their features simultaneously based on a group state.

Visual Context and Temporal Linking. We ground geometric cells visually by projecting 3D nodes onto multi-view 2D planes and fusing sampled features [36] via attention gating. To capture the temporal evolution of the surgical scene, we combine consecutive frames into a spatio-temporal complex, establishing bidirectional temporal edges between identical entities across frames.

2.4 Implementation Details and Multi-Task Learning

Entity Initialization. To instantiate our entity nodes without requiring laborious manual annotations, we leverage frozen perception modules. We use COMPOSE [34] to acquire training-free 3D human pose estimations. To extract object parameters and human semantics, we align 2D semantic segmentation predictions generated by a pre-trained model provided with the dataset with depth estimates derived from DepthAnythingv3 [21]. Using the camera parameters, we back-project the 2D segmentations into a unified 3D space, which yields 3D bounding boxes and allows us to assign definitive semantic class labels (e.g., tools, roles) to our entity nodes.

Multi-Task Learning. Our model is trained end-to-end to infer high-level clinical objectives from the topological representation. We attach MLP heads to the pooled CC embeddings to perform simultaneous multi-task learning for *Next Action Anticipation* and *Robot Phase Prediction*. For sterility breach detection, we apply rule-based spatial heuristics directly over the 3D entities in our topological structure, flagging a violation when a non-sterile entity (e.g., technician) enters within a critical proximity threshold of a sterile entity (e.g., patient). This protocol is independent of any learned head, so all 3D-grounded methods share the same evaluation and produce identical performance on this task.

3 Experiments and Results

Dataset and Evaluation Metrics. To evaluate our proposed framework, we conduct all experiments on the multimodal MM-OR dataset [24] using its standard train/val/test splits. We utilize the macro F1-score across all experiments to maintain consistency with established evaluation protocols [24].

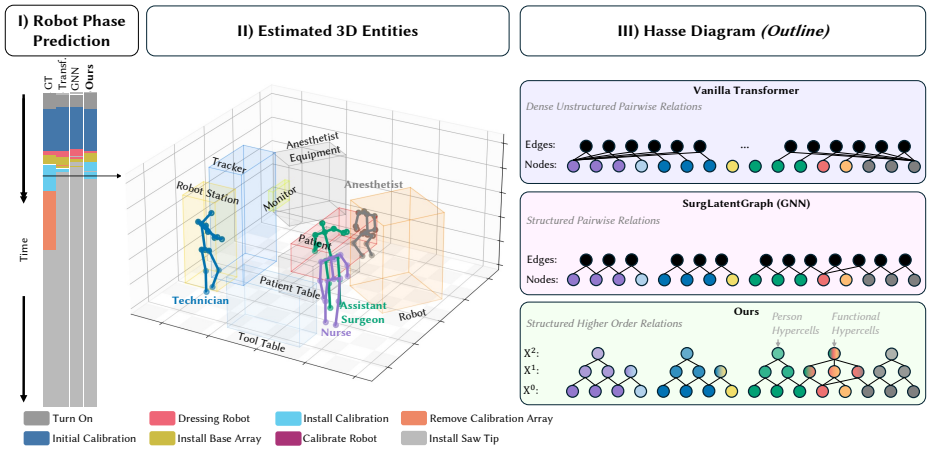

Fig. 3: Qualitative results and scene abstraction. We demonstrate improved performance in (I) robot phase prediction over baseline models. While all methods operate on the same (II) explicit 3D input entities, (III) illustrates the distinct structural formulation of each approach. Unlike the flattened relations of standard networks, our higher-order representation explicitly models the hierarchical incidence of the surgical environment through rank-0, rank-1, and rank-2 cells.

Main Quantitative Results. As shown in table 1, we compare TopoOR against MM2SG [24] (VLM paradigm), a Vanilla Transformer [30] (structureless sequences), and SurgLatentGraph [8] (flattened pairwise graphs). For sterility breach detection, all 3D-grounded methods achieve 76% F1, leveraging explicit spatial inputs to significantly outperform the text-reliant MM2SG (55% F1). In next action anticipation, TopoOR achieves 41%, outperforming the Transformer and SurgLatentGraph by natively preserving irreducible multi-agent dynamics within our rank-2 hypercells. Furthermore, TopoOR reaches a state-of-the-art 73% in robot phase prediction. Rather than flattening high-frequency 3D poses and textual robot logs into a uniform feature space, TopoOR processes heterogeneous OR cues strictly through topological incidence boundaries, underscoring the advantages of our explicit structural design over flattened architectures.

Qualitative Results. In fig. 3, we demonstrate our model's superior performance in robot phase prediction. While shown baselines process identical explicit 3D input entities, they organize and fuse this information differently. Unlike Vanilla

Table 3: Ablation study of incremental input modalities on Next Action Anticipation and Robot Phase Prediction.

Input Modalities						F1-Score (↑)	
Object Nodes	**Human Skel.**	**RGB Image**	**Robot Logs**	**Audio**	**Temporality**	**Next Action**	**Robot Phase**
✓						25.82	21.08
✓	✓					26.26	20.25
✓	✓	✓				36.35	60.71
✓	✓	✓	✓			38.26	67.06
✓	✓	✓	✓	✓		40.57	69.63
✓	✓	✓	✓	✓	✓	**41.10**	**73.53**

Transformers [30] (unstructured attention) and standard GNNs [8] (pairwise edges), our approach structures the scene as a hierarchical complex, natively capturing the polyadic dynamics required to align with the ground truth.

Ablation Over Modalities. To evaluate the contribution of each modality, we conduct an incremental ablation study, beginning from a purely geometric baseline and systematically introducing additional modalities (table 3). Relying solely on *estimated* geometric inputs (objects and human skeletons) yields limited performance, with both tasks below 27% F1. Grounding these spatial estimates with RGB visual context produces significant improvement, increasing robot phase prediction by over 40%. Incrementally fusing heterogeneous non-spatial cues (robot logs, audio) provides consistent additional gains, suggesting that TopoOR can effectively integrate diverse modalities without degradation. Finally, temporal edges yield a further improvement, particularly for robot phase prediction, consistent with the expectation that surgical phases are duration-dependent and benefit from temporal context (8 frames).

Representational Power and Graph Reduction. To evaluate whether our topological representation subsumes traditional scene graphs, we reduce our continuous combinatorial complex into the discrete string-based format used by [24]. As shown in table 2, a simple decision tree (depth 16) trained exclusively on rank-0 spatial entities already achieves 43% F1, indicating that the estimated 3D localization encodes meaningful relational structure. Training a relation classification head on the full topological complex yields 61% F1, outperforming the 53% F1 of the LLM baseline [24], suggesting that the structured multimodal representation retains richer relational information than VLM-generated graphs.

Runtime Efficiency. We compare parameter count and inference-time efficiency on a single NVIDIA A40 GPU. TopoOR (12M parameters) requires $59^{\pm 0.45}$ ms per forward pass, while MM2SG [24] (7B parameters, 4-bit quantized) requires $194.11^{\pm 2.62}$ ms. The resulting latency advantage makes TopoOR more amenable to intra-operative use, where real-time deployment is essential.

4 Concluding Remarks

We introduced a novel topological framework that redefines the representation of surgical scenes. By modeling ORs as multimodal combinatorial complexes, we explicitly preserve the environment's geometric structure and physical reality. Unlike standard graph networks or vision-language models that flatten complex interactions, our higher-order attention network (HAT) successfully bridges the gap between atomic entities and collective multi-actor dynamics [35]. We show that this structural integrity translates to superior representational expressiveness on the MM-OR dataset. Code will be released at https://github.com/wngTn/TopoOR.

Limitations. TopoOR depends on upstream perception (pose, depth, segmentation), so detection errors propagate into the cell structure; this constraint will ease as OR perception models improve. Moreover, the small set of semantic links is procedure-specific and must be re-specified for new settings (e.g., laparoscopy). More broadly, current benchmarks for surgical scene understanding focus on downstream regression and do not adequately capture clinical utility; future evaluations should prioritize clinically actionable metrics such as intraoperative risk mitigation, team cognitive load, and context-aware error prevention.

Acknowledgements. The authors acknowledge support from the UK AI Research Resource (AIRR) through grant 0251-4584-0945-1 and from the Excellence Strategy of local and state governments in Bavaria, Germany. T. B. was supported by a UKRI Future Leaders Fellowship (MR/Y018818/1) as well as a Royal Society Research Grant (RG/R1/241402). L.B. acknowledges support of the UK Royal Society through grant NIF/R1/254128.

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

References

1. Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Haro, B.B., Zappella, L., Khudanpur, S., Vidal, R., Hager, G.D.: A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. IEEE Transactions on Biomedical Engineering **64**(9), 2025–2041 (2017)
2. Bastian, L., Czempiel, T., Heiliger, C., Karcz, K., Eck, U., Busam, B., Navab, N.: Know your sensors—a modality study for surgical action classification. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **11**(4), 1113–1121 (2023)
3. Bastian, L., Rashed, M., Navab, N., Birdal, T.: Forecasting continuous non-conservative dynamical systems in so (3). In: ICCV (2025)
4. Bastian, L., Wang, T.D., Czempiel, T., Busam, B., Navab, N.: Disguisor: holistic face anonymization for the operating room. International Journal of Computer Assisted Radiology and Surgery **18**(7), 1209–1215 (2023)
5. Bigdelou, A., Sterner, T., Wiesner, S., Wendler, T., Matthes, F., Navab, N.: Or specific domain model for usability evaluations of intra-operative systems. In: IPCAI (2011)
6. Bouget, D., Allan, M., Stoyanov, D., Jannin, P.: Vision-based and marker-less surgical tool detection and tracking: a review of the literature. Medical image analysis **35**, 633–654 (2017)

7. Chen, C., Cheng, Z., Li, Z., Wang, M.: Hypergraph attention networks. In: 2020 IEEE TrustCom. pp. 1560–1565. IEEE (2020)
8. Dhamo, H., Farshad, A., Laina, I., Navab, N., Hager, G.D., Tombari, F., Rupprecht, C.: Semantic image manipulation using scene graphs. In: CVPR (2020)
9. Elizalde, B., Deshmukh, S., Al Ismail, M., Wang, H.: Clap learning audio concepts from natural language supervision. In: ICASSP. IEEE (2023)
10. Garrow, C.R., Kowalewski, K.F., Li, L., Wagner, M., Schmidt, M.W., Engelhardt, S., Hashimoto, D.A., Kenngott, H.G., Bodenstedt, S., Speidel, S., et al.: Machine learning for surgical phase recognition: a systematic review. Annals of surgery **273**(4), 684–693 (2021)
11. Hajij, M., Bastian, L., Osentoski, S., Kabaria, H., et al.: Copresheaf topological neural networks: A generalized deep learning framework. In: NeurIPS (2025)
12. Hajij, M., Zamzmi, G., Papamarkou, T., Guzman-Saenz, A., Birdal, T., Schaub, M.T.: Combinatorial complexes: bridging the gap between cell complexes and hypergraphs. In: 2023 57th Asilomar Conference on Signals, Systems, and Computers. pp. 799–803. IEEE (2023)
13. Hajij, M., Zamzmi, G., Papamarkou, T., Miolane, N., Guzmán-Sáenz, A., Ramamurthy, K.N.: Higher-order attention networks. arXiv:2206.00606 (2022)
14. Hajij, M., Zamzmi, G., Papamarkou, T., Miolane, N., Guzmán-Sáenz, A., Ramamurthy, K.N., Birdal, T., Dey, T.K., Mukherjee, S., Samaga, S.N., et al.: Topological deep learning: Going beyond graph data. arXiv:2206.00606 (2022)
15. Hashimoto, D.A., Rosman, G., Rus, D., Meireles, O.R.: Artificial intelligence in surgery: promises and perils. Annals of surgery **268**(1), 70–76 (2018)
16. Islam, M., Seenivasan, L., Ming, L.C., Ren, H.: Learning and reasoning with the graph structure representation in robotic surgery. In: MICCAI. pp. 627–636. Springer (2020)
17. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3668–3678 (2015)
18. Katić, D., Wekerle, A.L., Gärtner, F., Kenngott, H., Müller-Stich, B.P., Dillmann, R., Speidel, S.: Knowledge-driven formalization of laparoscopic surgeries for rule-based intraoperative context-aware assistance. In: IPCAI (2014)
19. Kennedy-Metz, L.R., Mascagni, P., Torralba, A., Dias, R.D., Perona, P., Shah, J.A., Padoy, N., Zenati, M.A.: Computer vision in the operating room: Opportunities and caveats. IEEE transactions on medical robotics and bionics (2020)
20. Lalys, F., Jannin, P.: Surgical process modelling: a review. International journal of computer assisted radiology and surgery **9**(3), 495–511 (2014)
21. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv:2511.10647 (2025)
22. Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al.: Surgical data science for next-generation interventions. Nature Biomedical Engineering pp. 691–696 (2017)
23. Özsoy, E., Örnek, E.P., Eck, U., Tombari, F., Navab, N.: Multimodal semantic scene graphs for holistic modeling of surgical procedures. arXiv preprint arXiv:2106.15309 (2021)
24. Özsoy, E., Pellegrini, C., Czempiel, T., Tristram, F., Yuan, K., Bani-Harouni, D., Eck, U., Busam, B., Keicher, M., Navab, N.: Mm-or: A large multimodal operating room dataset for semantic understanding of high-intensity surgical environments. In: CVPR (2025)
25. Padoy, N.: Machine and deep learning for workflow recognition during surgery. Minimally Invasive Therapy & Allied Technologies **28**(2), 82–90 (2019)

26. Papamarkou, T., Birdal, T., Bronstein, M., Carlsson, G., Curry, J., Gao, Y., Hajij, M., Kwitt, R., Lio, P., Di Lorenzo, P., et al.: Position: Topological deep learning is the new frontier for relational learning. Proceedings of machine learning research **235**, 39529 (2024)
27. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: EMNLP-IJCNLP (2019)
28. Sharma, S., Nwoye, C.I., Mutter, D., Padoy, N.: Rendezvous in time: an attention-based temporal fusion approach for surgical triplet recognition. IJCARS (2023)
29. 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)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
31. Vedula, S.S., Malpani, A.O., Tao, L., Chen, G., Gao, Y., Poddar, P., Ahmidi, N., Paxton, C., Vidal, R., Khudanpur, S., et al.: Analysis of the structure of surgical activity for a suturing and knot-tying task. PloS one **11**(3), e0149174 (2016)
32. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. ICLR (2018)
33. Wang, T.D., Bastian, L., Czempiel, T., Heiliger, C., Navab, N.: Beyond role-based surgical domain modeling: Generalizable re-identification in the operating room. Medical Image Analysis (2025)
34. Wang, T.D., Birdal, T., Navab, N., Bastian, L.: Compose: Hypergraph cover optimization for multi-view 3d human pose estimation. arXiv:2601.09698 (2026)
35. Wang, T.D., Heiliger, C., Navab, N., Bastian, L.: Trackor: Towards personalized intelligent operating rooms through robust tracking. In: MICCAIW COLAS (2025)
36. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021)