OR-Action: Multi-Role Video Understanding with Fine-Grained Actions

Felix Tristram[1,2], Ege Özsoy[1,2], Christian Benz[3], Marcel Walch[3], Ghazal Ghazaei[3,*], and Nassir Navab[1,2,*]

[1] Technical University of Munich, Germany
[2] Munich Center for Machine Learning (MCML)
felix.tristram@tum.de
[3] Carl Zeiss AG, Germany

Abstract. Fine-grained understanding of operating room (OR) activity could enable workflow-aware assistance, yet remains difficult due to clutter, occlusions, and limited sensing. The prevailing approach to model this environment is scene graphs as an interpretable representation of OR interactions. Converting their frame-wise relational predictions into temporally extended, fine-grained actions however, is challenging without explicit temporal modeling. To enable a principled temporal evaluation of current OR understanding methods, we introduce the first action-centric benchmark built on a publicly available ego-exocentric OR dataset by defining a fine-grained, multi-role action taxonomy and generating dense action segments via distillation from ground-truth scene graph state changes. Experiments on this benchmark show that current scene graph prediction methods struggle to model temporal structure, even when adding explicit modeling through Graph Neural Networks. We therefore introduce a vision-only temporal model that outperforms graph-based methods significantly when using all available egocentric video as input. Building on this model we also introduce a novel multi- to single-view feature alignment strategy that improves single-view performance on multi-role action recognition, mitigating the need for extensive egocentric video capture. Benchmark and code will be released upon acceptance.

Keywords: Surgical Data Science · Workflow · Video Models

1 Introduction

Operating rooms (ORs) are dense, dynamic, and safety-critical environments in which multiple clinicians coordinate around a patient while interacting with tools, devices, and displays. Reliable perception of this external scene could enable context-aware safety checks, workflow-aware interfaces, automated documentation, and training feedback, aligning with the broader surgical data science vision [8]. However, compared to intra-body endoscopic and laparoscopic

* Equal advising.

video analysis—where progress has been driven by large-scale datasets and well-defined tasks such as workflow recognition, instrument presence, segmentation, keypoints, and skill assessment [16, 6, 17, 12]—external OR understanding remains underexplored due to privacy constraints, heavy occlusions and clutter, and the need to reason jointly about multi-actor interactions over time [15, 2, 13]. Recent work has begun to model the OR as a structured interaction space using semantic scene graphs that encode entities and relations [11, 9, 10], offering an interpretable and queryable abstraction of "who interacts with what." Yet many clinically meaningful activities, such as tool pick-ups, handovers, equipment positioning, or site preparation, are inherently temporal and defined by state changes rather than single timepoints; mapping per-timepoint relational snapshots to fine-grained actions is therefore brittle, particularly for long-tailed relations, under occlusion, and when action semantics depend on temporal ordering and persistence [7]. In this work, we introduce a dense, fine-grained action benchmark on top of a publicly available OR dataset with scene graph annotations [9]. Inspired by progress in general video understanding [5, 4, 14, 3], we define a taxonomy of atomic, role-specific OR actions and generate dense action segments by applying a rule-based mapper to ground-truth scene graphs, using temporal heuristics that convert relation and state transitions into events, yielding 78 action classes spanning 1295 segments over the full dataset. These annotations provide the OR-Action benchmark grounded in structured supervision and enable a principled evaluation of scene-graph-based pipelines as action recognition systems; we show that performance degrades substantially when extrapolating from *predicted* graphs to dense temporal multi-role actions, revealing a gap between relational parsing accuracy and robust temporal action recognition.

Finally, we show that strong performance is achievable without scene graphs by training a temporal pooling and multi-person classification framework on top of a video foundation model using all available egocentric views. Studying the impact of different viewpoints on action recognition performance of our vision-only model, we find that there is a large variability depending on which roles viewpoint is being observed. Some actions can only be observed by the main actor - e.g. surgeon cutting - while being occluded from other viewpoints. This motivated us to explore a novel feature alignment strategy that aligns video representations from a multi-view teacher to a single view student and enhances the students action recognition performance by distilling multi-view context into a single view representation.

In total our contributions include (i) the first fine-grained action-centric external OR video understanding benchmark (*OR-Action*), (ii) a principled evaluation of scene graph based methods and our vision only model, and (iii) a novel cross-view feature alignment strategy to improve single-view egocentric action prediction. Together, these contributions bring external OR understanding closer to mature intra-body surgical analysis and modern video understanding benchmarks [5, 14, 3].

2 Rule-Based Scene-Graph to Action Mapping

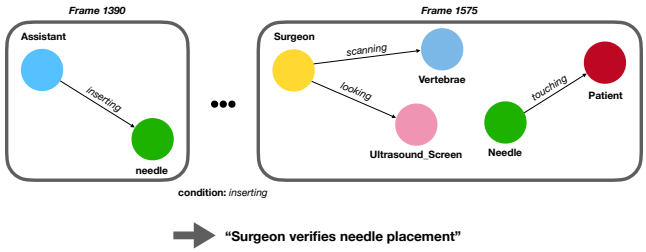

Fig. 1. Illustration of "*Surgeon verifies needle placement*" class mapping. From scene graphs this is indistinguishable from *"Surgeon scans for target vertebrae"*, so we define the heuristic that this action can only trigger after *"recent"* needle insertion event.

We deterministically map per-frame scene graphs from the publicly available EgoExOR dataset [9] to fine-grained, role-specific action labels and compress them into dense temporal segments. Each original annotated frame is represented by multiple relational triplets (*subject*, *predicate*, *object*). Our rule-based mapper ingests these frame-wise annotations for an entire video and uses them to query a set of predefined rules for an action label. Rules are evaluated independently for every egocentric role in the dataset, assigning either an idle or class label for every role at every time point.

There are two types of rules: **State rules**, that trigger whenever boolean conditions over triplets hold and capture persistent activities ((*anaesthetist, looking, health_ monitor*) $\Rightarrow$ "Anesthetist monitors vitals"). **Event rules**, that detect change points such as instrument pick-ups, handovers, and returns and convert them into short temporal segments: events are localized to anchor frames and extended into the past and future to reflect movement cues, e.g. the first frame of (*assistant, holding, scalpel*) is the anchor for scalpel pick-up, which then gets extended with past and future frames to reflect the motion of instrument pick-up. Fig. 1 illustrates an event rule in which a recent needle-insertion event disambiguates otherwise similar scene-graph states.

To improve robustness to annotation noise, we use temporal evidence aggregation - has something happened *"recently"* within a window, or *"ever-before"* in the video. Object handovers are detected from changes in *holding(object)* across roles within a short window; object returns are detected at the end of holding episodes but gated by (*subject,closeTo,instrument-table*) context and brief confirmed absence of the holding relation to suppress brief missing annotations. We prevent repeated pick-ups of the same role–object pair without an intervening return. The rule set is compositional (later decisions may reference earlier outputs) and evaluated in a fixed order; rule conflicts are resolved via priority settings to enforce one label per role per frame. Finally, we merge frame-wise outputs into contiguous segments and apply light smoothing by absorbing very short

segments into neighboring actions. This approach yields 1295 action segments over 78 unique classes and almost 100k active and 230k idle frames. Classes are defined from human observation over the entire dataset and refined manually to match behavior between videos; generated annotations are visually verified for correctness. We re-use the original train/val splits of EgoExOR to maintain compatibility and showcase a visual example with GT and predicted labels in the supplementary video.

3 Method

3.1 Multi-Role Framewise Action Recognition

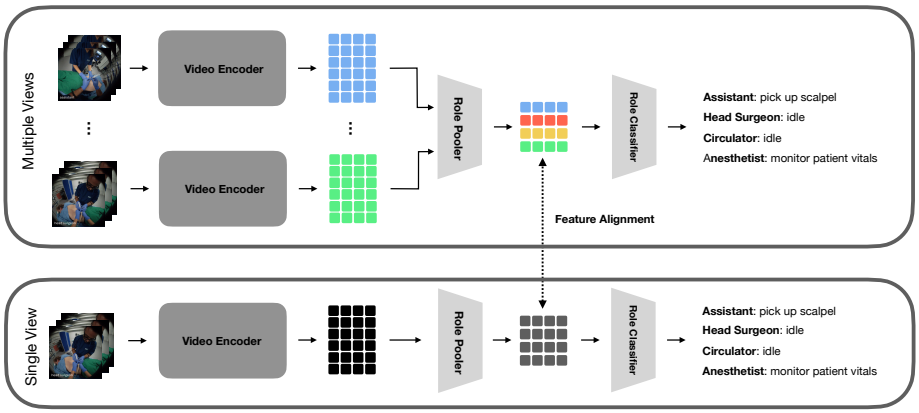

Fig. 2. Overview of our vision-only model and cross-view feature alignment strategy

We observe synchronized video clips from a set of camera streams associated with OR *roles* (e.g., different staff members and/or fixed external cameras) as showcased in Figure 2. Let $\mathcal{R}$ denote the target roles for which we predict actions, and $\mathcal{R}_{\text{in}}$ the set of available input streams for a given clip. For each target role $r \in \mathcal{R}$ and frame index $t \in \{1, \dots, T\}$, we are given (i) a fine-grained action label $y_t^{(r)} \in \{1, \dots, K\}$ and (ii) a binary activity label $a_t^{(r)} \in \{0, 1\}$ indicating whether the role is active or idle.

Each input stream $r \in \mathcal{R}_{\text{in}}$ provides a clip $\mathbf{x}_{1:T}^{(r)}$. Spatiotemporal tokens are extracted with a frozen video foundation model E:

$$\mathbf{z}^{(r)} = E\left(\mathbf{x}_{1:T}^{(r)}\right), \tag{1}$$

where $\mathbf{z}^{(r)}$ is a sequence of token embeddings. VJEPA2 [1] is used as encoder due to its impressive performance representing general motion and semantic cues.

Token Compression. The encoder produces a long token sequence from which we extract the most informative representations using an attentive Role Pooler $\mathrm{RP}_Q(\cdot)$: a set of Q learnable query tokens that attend to the input token set via cross-attention. For each role we obtain a compact sequence

$$\mathbf{u}^{(r)} = \mathrm{RP}_{Q_f}\left(\mathbf{z}^{(r)}\right), \tag{2}$$

and interpret $\mathbf{u}^{(r)}$ as *temporal tokens*. We select Q to obtain exactly T tokens per role, yielding per-frame embeddings $\mathbf{u}_t^{(r)}$.

Framewise cross-role fusion. To exploit interactions, shared context and mitigate occlusions, we fuse pooled representations between all observed input streams. For a fixed t, we stack available role embeddings $\{\mathbf{u}_t^{(r)}\}_{r\in\mathcal{R}_{\text{in}}}$ and apply a second attentive pooler with $|\mathcal{R}|$ queries as a Role Classifier RC, producing one fused embedding per *target* role:

$$\mathbf{h}_t = \mathrm{RC}_{|\mathcal{R}|}\left(\{\mathbf{u}_t^{(r)}\}_{r\in\mathcal{R}_{\text{in}}}\right), \tag{3}$$

where $\mathbf{h}_t = \left[\mathbf{h}_t^{(1)}, \ldots, \mathbf{h}_t^{(|\mathcal{R}|)}\right]$ and $\mathbf{h}_t^{(r)}$ is the role-conditioned representation used for prediction. Missing streams are handled by zeroing their tokens and masking losses.

Action and activity heads. We predict per-frame logits using linear heads:

$$\boldsymbol{\ell}_t^{(r)} = W_{\text{act}}\,\mathbf{h}_t^{(r)}, \qquad \boldsymbol{s}_t^{(r)} = W_{\text{aux}}\,\mathbf{h}_t^{(r)}, \tag{4}$$

where $\boldsymbol{\ell}_t^{(r)}$ is a K-way action logit vector and $\boldsymbol{s}_t^{(r)}$ is a 2-way activity logit vector.

We use a masked multi-task cross-entropy over roles and frames,

$$\mathcal{L}_{\text{frame}} = \sum_{r\in\mathcal{R}}\sum_{t=1}^{T} m_t^{(r)}\Big(\mathrm{CE}(\mathrm{softmax}(\boldsymbol{\ell}_t^{(r)}), y_t^{(r)}) + \mathrm{CE}(\mathrm{softmax}(\boldsymbol{s}_t^{(r)}), a_t^{(r)})\Big), \tag{5}$$

where $m_t^{(r)}$ masks roles that are not available in a procedure. We choose separate action and activity heads due to large class imbalance between active and idle frames.

3.2 Multi- to Single-View Feature Alignment

In realistic OR deployments, we may only observe a single wearable stream (or a subset of streams), yet we still want to predict multi-role actions. We therefore experiment with transferring multi-view context from a *multi-role teacher* to a *single-view student*.

Teacher representation. We first train the multi-role model from Sec. 3.1 using all available streams, and then freeze it. Given a clip with streams $\mathcal{R}_{\text{in}}$ (typically $\mathcal{R}_{\text{in}} = \mathcal{R}$), the teacher produces a role-major sequence by pooling each role independently and concatenating:

$$\mathbf{t} = \left[\,\mathrm{RP}_{Q_T}^{T}(\mathbf{z}^{(r)})\,\right]_{r\in\mathcal{R}_{\text{in}}}. \tag{6}$$

This sequence encodes OR context distributed across roles (what other staff members are doing, shared tool state, etc.).

Single-view student and alignment loss. The student observes only a fixed input role r_0 and produces a sequence of the same length, $\mathbf{s} = \mathrm{RP}^{S}_{Q_S}(\mathbf{z}^{(r_0)})$, where Q_S is chosen so that $\mathbf{s}$ matches the teacher sequence length $(R_{in} \cdot Q)$. We align tokens using a per-token ℓ_1 loss:

$$\mathcal{L}_{\text{align}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \left\| \hat{\mathbf{t}}_i - \hat{\mathbf{s}}_i \right\|_1 , \tag{7}$$

where Ω indexes tokens that correspond to available input streams and $\hat{\cdot}$ denotes ℓ_2 normalization. Intuitively, the student is encouraged to extract more contextual cues from the dense output $\mathbf{z}^{(r_0)}$ that would otherwise be missed using $\mathcal{L}_{\text{frame}}$ alone.

Joint supervised training. To ensure the aligned representation remains predictive of actions, we optimize both alignment and frame-based losses simultaneously. We reshape the student sequence into role-by-time tokens and apply the same per-frame cross-role fusion and action/activity heads as in Sec. 3.1, yielding $\boldsymbol{\ell}_t^{(r)}$ and $\boldsymbol{s}_t^{(r)}$ from the single input stream. The final objective is a weighted combination of alignment and supervised losses:

$$\mathcal{L} = \lambda_{\text{align}} \mathcal{L}_{\text{align}} + \lambda_{\text{sup}} \mathcal{L}_{\text{frame}}. \tag{8}$$

4 Experiments

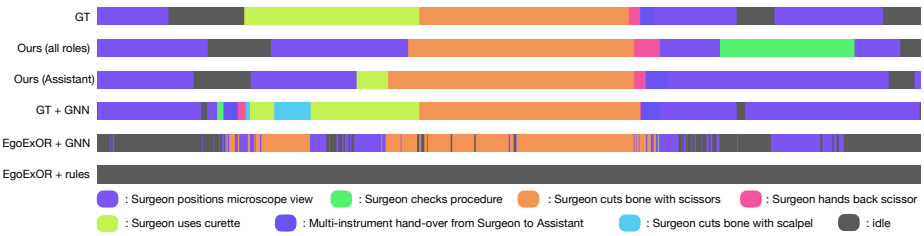

Fig. 3. Qualitatively example on (*MISS/3/take/1*) from validation set.

Implementation Details. We sample $T = 64$ frames at 4 fps (16s windows) to align with VJEPA2's training [1], setting $Q = 64$ pooling tokens per role. This keeps streaming inference feasible, since larger temporal windows quickly become infeasible due to quadratic self-attention scaling. For alignment we set $\lambda_{align} = 10.0, \lambda_{sup} = 0.1$ to roughly match loss magnitudes. The OR-Action benchmark comprises $K = 78$ classes. Due to class imbalance—*idle* classes account for 69% of frames—we use separate activity and action prediction heads. The GNN baseline initializes nodes from (*subject, object*) with stable embeddings

Table 1. Benchmarking OR temporal action understanding. Metrics are computed without idle class. GT + GNN is evaluated on GT scene graphs as an upper bound, whereas EgoExOR + GNN is using predicted scene graphs.

Model	**Acc**	**Edit Score**	**F1**@(*0.1*	*0.25*	*0.5*)
Validation					
GT + GNN	65.19	48.33	32.08	30.42	25.42
EgoExOR + rule mapping	16.54	*38.24*	21.21	16.16	13.13
EgoExOR + GNN	36.29	30.01	3.56	2.43	1.13
Ours (all roles)	**61.65**	**38.24**	**39.90**	**35.47**	**25.12**
Test					
GT + GNN	69.54	48.63	39.34	36.53	29.98
EgoExOR + rule mapping	17.37	38.28	21.59	20.45	18.18
EgoExOR + GNN	43.79	37.29	4.20	3.00	1.80
Ours (all roles)	**67.50**	**45.47**	**40.64**	**37.88**	**29.10**

across timepoints, while also connecting node IDs to adjacent frames $[t-1; t+1]$. Scene graph predictions are obtained using the publicly available model checkpoint of EgoExOR [9]. At inference, all learned models run in a sliding window fashion: the activity head gates idle frames, after which the action head classifies active frames. Windows overlap by 50%, with predictions merged via a Hanning-weighted sum to reduce boundary artifacts.

Metrics. We report traditional temporal action-segmentation metrics: Frame Accuracy (**Acc** *w/idle*) (percentage of correctly classified frames, including idle classes); Frame Accuracy without idle (**Acc**) for evaluation without class imbalance; **Edit Score** (normalized edit distance on the predicted action sequence, penalizing ordering errors and over/under-segmentation); and Segmental F1 (**F1**@k) (segment-level F1 at overlap thresholds of 10%, 25%, and 50%, requiring both class and boundary agreement). Higher is better for all metrics.

Discussion of Results. Table 1 summarizes the main findings of the OR-Action benchmark, with SG denoting scene graph. As a simple baseline, we apply the rule-based mapper from Sec. 2 to EgoExOR *predicted* SGs. Because these rules are brittle and tuned to ground-truth annotations, we also introduce EgoExOR+GNN, a learned baseline trained on ground-truth SGs but evaluated on EgoExOR *predictions*. The rule-based mapper performs very poorly, underscoring both the brittleness of the rules and the gap between predicted and ground-truth SGs. EgoExOR+GNN improves on this baseline, but still fails to produce coherent temporal segments, as shown by its very low F1-scores and the temporal jitter in Figure 3.

To assess scene graphs as a representation independently, we also report an upper bound with GT+GNN, trained *and* evaluated on ground-truth SGs. Its strong performance suggests that scene graphs can be informative for temporal action understanding when reliable relational states are available, while the drop

Table 2. Ablations on viewpoints, feature alignment strategy and split prediction head design. All experiments use our vision-only model and report test-set results.

Input Roles	**Acc***(w/idle)*	**Acc**	**Edit**	**F1@**(*0.1*	*0.25*	*0.5*)
all roles (*ego*)	62.13	67.50	45.47	40.64	37.88	29.10
all roles (*no activity head*)	11.33	38.57	17.67	24.92	19.27	10.63
only exo views	50.05	44.94	30.99	26.60	23.50	13.30
Surgeon	57.44	57.81	30.82	**32.46**	**27.70**	16.88
Surgeon*(align)*	**62.11**	**58.16**	**42.93**	31.29	26.64	**19.45**
Assistant	60.39	64.79	40.14	38.84	36.16	24.11
Assistant*(align)*	**67.89**	**64.90**	**44.84**	**41.03**	**36.36**	**26.10**
Circulator	**36.02**	**40.28**	24.12	22.58	**17.74**	**9.68**
Circulator*(align)*	35.49	38.31	**30.61**	**23.53**	15.68	8.62

in the predictive setting points to the difficulty of converting class-imbalanced, frame-wise scene-graph predictions into temporally coherent dense action segments.

Our vision-only model reaches accuracy and Edit Score close to the oracle GT+GNN, while outperforming it in F1-score at all but the strictest threshold. By explicitly modeling temporal structure and leveraging strong video foundation model features, it captures even brief actions such as the scissor handover in Figure 3. Its main limitation remains visually similar actions, such as *surgeon positions microscope* and *surgeon checks procedure*, which are hard to distinguish from vision alone.

Table 2 summarizes our experiments with the feature alignment strategy from Sec. 3.2, together with ablations on split prediction heads and ego- versus exocentric inputs. Using a single prediction head (*no activity head*) yields poor performance because the *idle* class dominates the data, accounting for 69% of all labeled frames. Relying only on exocentric inputs (e.g., ceiling-mounted cameras) also leaves a large gap to egocentric views, as these cameras are fixed, far from the active regions, and often occluded by staff or equipment, making fine movements and small instruments difficult to capture. The strong gain from egocentric inputs highlights the importance of close-up, human-viewpoint observations for reliable OR action understanding.

But not all egocentric viewpoints are equally informative. When using only the input stream from the Circulator, which is mostly inactive and observing the procedure from far away, performance is generally weak and, except for Edit Score, does not benefit from cross-view feature alignment. This is not unexpected, since the feature alignment relies on picking up contextual cues from its single view input, which are simply not observable in this role.

The opposite is true for the head surgeon and assistants viewpoints. Here we can observe large gains when using the feature alignment strategy, especially in Acc(*w/idle*) and Edit Score, indicating improved awareness of all roles idle/active states and superior action ordering, respectively.

5 Conclusion

We present OR-Action, the first multi-role OR benchmark for temporal action understanding, heuristically derived from structured scene graph annotations in the publicly available EgoExOR [9] dataset. Experiments show that current scene graph prediction models, while strong at scene graph prediction itself, struggle to represent temporal actions. Motivated by this, we introduce a vision-only model that clearly outperforms predicted scene graphs, as well as a novel multi- to single-view egocentric feature alignment strategy that reduces the need for extensive multi-person video capture. We hope OR-Action will accelerate research in external OR understanding and support the development of temporally aware methods in this domain.

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

References

1. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
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. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European conference on computer vision (ECCV). pp. 720–736 (2018)
4. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18995–19012 (2022)
5. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19383–19400 (2024)
6. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.: Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. arXiv preprint arXiv:2012.12453 (2020)
7. Ji, J., Krishna, R., Fei-Fei, L., Niebles, J.C.: Action genome: Actions as compositions of spatio-temporal scene graphs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10236–10247 (2020)
8. 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 **1**(9), 691–696 (2017)

9. Özsoy, E., Mamur, A., Tristram, F., Pellegrini, C., Wysocki, M., Busam, B., Navab, N.: Egoexor: An ego-exo-centric operating room dataset for surgical activity understanding. arXiv preprint arXiv:2505.24287 (2025)
10. Özsoy, E., Örnek, E.P., Eck, U., Czempiel, T., Tombari, F., Navab, N.: 4d-or: Semantic scene graphs for or domain modeling. In: International conference on medical image computing and computer-assisted intervention. pp. 475–485. Springer (2022)
11. 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: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19378–19389 (2025)
12. Rueckert, T., Maerkl, R., Rauber, D., Klausmann, L., Gutbrod, M., Rueckert, D., Feussner, H., Wilhelm, D., Palm, C.: Video dataset for surgical phase, keypoint, and instrument recognition in laparoscopic surgery (phakir). arXiv preprint arXiv:2511.06549 (2025)
13. Schmidt, A., Sharghi, A., Haugerud, H., Oh, D., Mohareri, O.: Multi-view surgical video action detection via mixed global view attention. In: International conference on medical image computing and computer-assisted intervention. pp. 626–635. Springer (2021)
14. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21096–21106 (2022)
15. Srivastav, V., Issenhuth, T., Kadkhodamohammadi, A., de Mathelin, M., Gangi, A., Padoy, N.: Mvor: A multi-view rgb-d operating room dataset for 2d and 3d human pose estimation. arXiv preprint arXiv:1808.08180 (2018)
16. 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)
17. Wagner, M., Müller-Stich, B.P., Kisilenko, A., Tran, D., Heger, P., Mündermann, L., Lubotsky, D.M., Müller, B., Davitashvili, T., Capek, M., et al.: Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. Medical image analysis **86**, 102770 (2023)