ST(OR)^2: Spatio-Temporal Object Level Reasoning for Activity Recognition in the Operating Room

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Modern operating theaters are highly-complex environments in which clinicians have to manage these intricate surroundings while conducting high-skill-requiring tasks. This over-complication of the workflow is even amplified for robotic-assisted surgery (RAS) and could affect Operating room (OR) efficiency. Human factor-focused activity recognition in surgical videos is a critical research area for next-generation OR workflow monitoring tools. Indeed compared to traditional surgery, RAS poses new challenges in terms of interactions between clinicians and their environment; it also involves more stereotyped manipulation of the surgical machines.

We propose a new object-focused approach to exploit the compositionality of specific activities in the OR. Our method only focuses on the geometric arrangements between clinicians and surgical devices, thus utilizing the significant object interaction dynamics in the OR.

We exhibit the value of our method as an efficient low-data learner by conducting experiments with an increasing amount of annotated data. We also benchmark our method against other object-centric approaches on both clip-level action classification and long video action recognition.

Keywords: OR Workflow Analysis, Surgical Activity Recognition, Object Level Reasoning, Human-object interaction

1. Introduction

Robotic-assisted surgery (RAS) is diffusing across a broad range of surgical procedures as it has been shown to improve patient recovery time (Sheetz KH, 2020). Still, some barriers are limiting its overall adoption. Among those restricting factors, we can point out the communication (Schiff et al., 2016) and Operating room (OR) size issues (Kanji et al., 2021) that are linked to the new spatial distribution of clinicians and surgical devices brought by RAS.

Those spatial distribution and communication issues can lead to workflow disruptions that could hamper the outcome of procedures and increase turnover time for hospital administrators (Catchpole et al., 2015). Consequently, a burgeoning research field has been working towards a better automatic assessment of OR efficiency.

One of the core technologies developed in this direction is surgical action recognition using computer vision and deep learning, which aims to temporally segment surgical videos into phases. This coarse temporal segmentation is a first step towards measuring OR efficiency and will allow us to estimate turnover time for an improved allocation of the operating room.

© 2023 Anonymous.

Computer vision researchers have extensively studied action recognition. Deep networks with temporal connections for action recognition by using Recurrent Neural Networks with 2D Convolutions (He et al., 2015; Hochreiter and Schmidhuber, 1997), 3D ConvNets (Carreira and Zisserman, 2017; Feichtenhofer, 2020; Feichtenhofer et al., 2018) and Video Transformers (Bertasius et al., 2021; Patrick et al., 2021) were developed in this direction. However, although powerful those methods do not explicitly embed object information as the information is de-localized throughout the representation at the pixel level.

Most of the developmental psychology theories agree on object-centric reasoning being a core constituent of human-level common sense (Tenenbaum et al., 2011). It has been shown that decomposing a scene into meaningful units including foreground and background objects, can benefit from numerous advantages of abstract symbolic representation. These include interpretability, sample efficiency, the ability to reason over semantically defined objects, and compositionality for better generalization. Objects being the atomic units of physical interactions, it seems reasonable to discretize video representations at the object level.

In OR videos we need to deal with cluttered environments, depending on the camera's point of view certain objects will not be visible. This clutter can both yield expertise movement and visibility, and might also hinder the performance of any tracking method. The contribution of this paper is two-fold. First, we introduce a new tracking agnostic object-centric approach for surgical activity recognition (SAR) on long videos. Secondly, we show that our method leads to better data-efficiency and we benchmark other state-of-the-art object-centric methods on SAR for the first time.

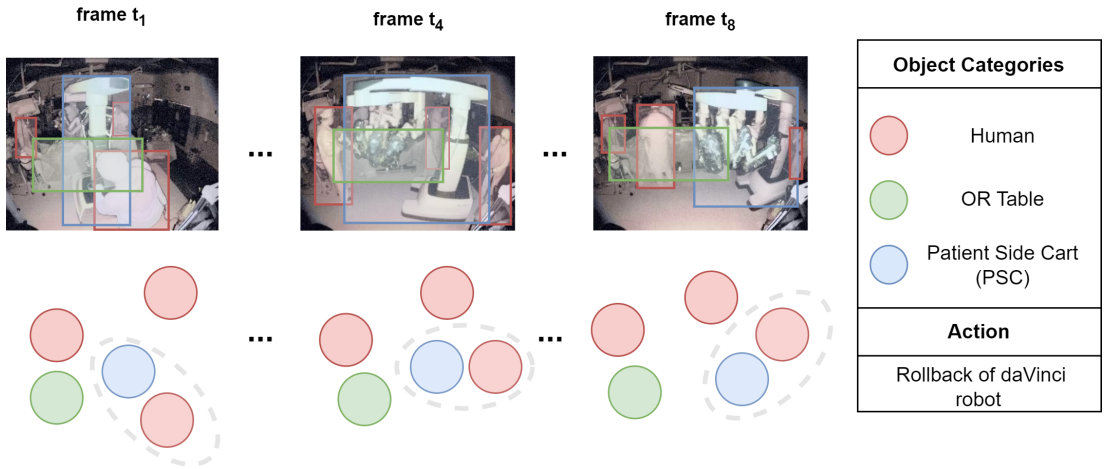

Figure 1: Example video of an action class, in this case "Rollback of the surgical robot". Our method focuses on geometric interactions of semantically identified objects. Here the proximity of one of the clinician with the PSC and the fact they are both getting away from the OR table is a good clue to predict the correct action.

2. Related Work

2.1. Surgical Video Understanding

Surgical activity recognition has been widely studied in laparoscopic videos and ophthalmological videos, mainly due to the availability of well-explored datasets such as Cholec80 (Twinanda et al., 2017) or Cataracts100 (Hajj et al., 2018). State of the art approaches on those full-length procedure videos all adopt two-stage training. In TECNO (Czempiel et al., 2020) proposed to use a TCN (Lea et al., 2016) on top of ResNet18 (He et al., 2015) features for each frame to temporally segment videos of cholecystectomy.

Our work falls into the scope of holistic operating room video understanding (Srivastav et al., 2019, 2018; Luo et al., 2018; Sharghi et al., 2020; Li et al., 2020; Chakraborty et al., 2013; Issenhuth et al., 2019). The pioneering works on this type of data have been proposed for hand hygiene compliance and human pose estimation for clinicians in the operating room. An important aspect of those datasets is privacy preservation through a low resolution of the images (Srivastav et al., 2019) or using depth images.

In the work proposed by (Sharghi et al., 2020), authors introduced the first OR Activity Recognition dataset for workflow monitoring, they proposed a two-stage approach with a 3D CNN coupled with an LSTM. Their work was later on extended by (Schmidt et al., 2021), who proposed to use an attention layer to merge information from multiple views of the OR. Recently, (He et al., 2022) proposed an extensive benchmark of temporal models on this same dataset.

In 4D-OR, (Özsoy et al., 2022) introduced a new dataset with densely annotated 3D scene graphs in an orthopedic OR. They exploit geometric interactions and annotated interactions to determine clinicians' roles.

2.2. Object-centric Video Models

Video understanding benchmarks such as action recognition on large publicly available datasets have recently seen a significant improvement with the development of 3D CNN (Carreira and Zisserman, 2017; Feichtenhofer, 2020; Feichtenhofer et al., 2018) and Video Transformer (Bertasius et al., 2021; Patrick et al., 2021) based approaches. Those methods albeit powerful also often suffer from spatial and temporal biases and do not explicitly exploit spatio-temporal human-object interaction cues.

Recently, many approaches to making use of relational reasoning have been proposed in the video-understanding community. To spatially ground features from a 3d CNN authors of STRG (Wang and Gupta, 2018) propose a class-agnostic approach using a temporal graph over densely sampled regions of interest in video frames. In STIN (Materzynska et al., 2019), the authors build an object graph using both position and category information of tracked objects to recognize actions in a compositional setting. In ORViT (Herzig et al., 2022), authors proposed a fully transformer-based approach to reason over object information in two ways: by attending to pooled object features similarly to STRG (Wang and Gupta, 2018) and by encoding object trajectories at different intermediate layers using tracked bounding boxes of objects. A detailed study of the transferability of object-centric action recognition to other downstream tasks was lately proposed by (Zhang et al., 2022).

3. Methods

We propose Spatio-Temporal Object-level Reasoning in the Operating Room (ST(OR)^2) for activity recognition in the OR. We use a specific object detector trained on a small fraction of the entire OR-AR dataset to detect OR-specific objects. We use the bounding box geometric information and the semantic information about each object class to build our graph. Our semantically rich spatio-temporal graph can then be used to reason over the objects' relative locations to recognize clip actions based on interaction dynamics. This backbone serves as the first stage of our full-time video activity recognition, as it allows us to extract reliable features.

In the second stage, we use the features extracted from our clip-based model to train a temporal sequence model to capture long-range dependencies in the videos.

3.1. Object-level Representation

To train our backbone we need to sample clips from each segment of our full-time procedure. Those clips are 16 frames long and we uniformly sample 8 frames to define our object graph. We use two specific pretrained detectors to extract human and RAS-specific objects bounding boxes.

Each object appearing in those frames will serve as a node of our graph. Inspired by STIN (Materzynska et al., 2019), each node will be associated with specific features grounded on its position and object category.

Spatial Position Embedding We represent object position using their bounding box coordinates as a 4-dimension vector containing center coordinates and the width and height of the box. This 4-dimension vector is then forwarded to a Multi-Layer Perceptron (MLP), to obtain a d-dimension embedding.

$$\sigma_{i,t} = MLP_{SPE}(box_{i,t}) \quad for\ i\ in\ [\![1, N]\!]\ and\ t\ in\ [\![1, 8]\!] \tag{1}$$

Category Embedding The annotated semantic classes include human, patient side cart (PSC), vision side cart (VSC), OR table, Non-OR table, and gurney. Each of those classes will be associated with a d-dimension learnable embedding, that is randomly initialized.

$$\kappa_{i,t} = MLP_{CE}(class_{i,t}) \quad for\ i\ in\ [\![1, N]\!]\ and\ t\ in\ [\![1, 8]\!] \tag{2}$$

Both spatial position embedding and category embedding are concatenated and passed through an MLP to obtain a fused representation for each node of the graph.

$$x_{i,t} = MLP_{Fusion}(\sigma_{i,t}||\kappa_{i,t}) \quad for\ i\ in\ [\![1, N]\!]\ and\ t\ in\ [\![1, 8]\!] \tag{3}$$

3.2. Spatio-Temporal Reasoning

Category-wise aggregation We aggregate features of objects belonging to the same category by summing them together. This alleviates any need for tracking different instances of the same class across frames.

$$\phi_{c,t} = \sum_{class_{i,t}=c} x_{i,t} \quad for\ c\ in\ [\![1, 6]\!]\ and\ t\ in\ [\![1, 8]\!] \tag{4}$$

Temporal-Category interaction module Using the aggregated features for each object category, we first carry out temporal reasoning across categories after concatenating the $\phi_{c,t}$ features for each frame t of the clip.

$$\phi_c = MLP_{Temp}(\phi_{c,1}||...||\phi_{c,8}) \quad for\ c\ in\ [\![1,6]\!] \tag{5}$$

Once we obtain a feature vector representing the temporal evolution of each object category, we perform category-wise reasoning over the concatenated features of each category to obtain a clip-level representation that is fed to a classifier to predict the action of the clip.

$$\phi_{clip} = MLP_{Category}(\phi_{c=1}||...||\phi_{c=6}) \tag{6}$$

Our object-level representation can also easily be combined with video appearance features extracted from a 3D CNN, in our experiments we will be using I3D (Carreira and Zisserman, 2017).

3.3. Temporal Sequence Modeling

Following our clip-based feature extraction, each video is then represented as $v_i = \{\phi_1, ..., \phi_T\}$ with ϕ_t is the feature extracted from the tth clip. Those features can then be concatenated with the I3D features and be fed to UniGRU (Cho et al., 2014). This allows us to deal with long-range temporal dependencies and obtain a more robust temporal segmentation for long videos.

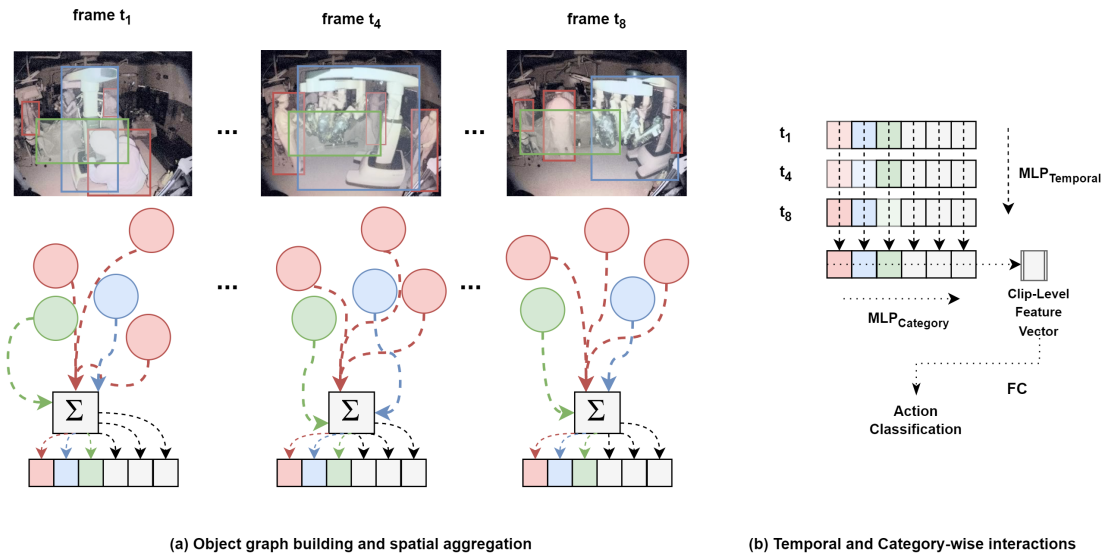

Figure 2: Architecture of our method, we first build our object graph and aggregate features category-wise. In the second step, we reason over time for each category and then we reason over categories to obtain a clip-level feature vector for action classification.

4. Experiments

4.1. OR SAR Dataset

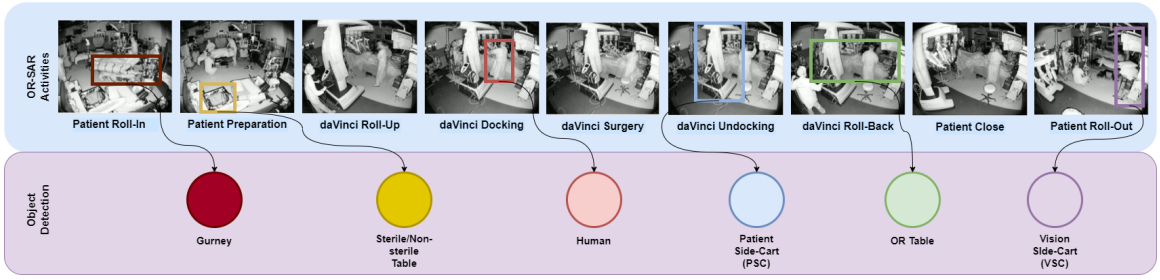

Figure 3: Overview of the OR SAR dataset, with ten clinically relevant activities and six OR-specific objects annotated.

We demonstrate the performance of our method on the Surgical Action Recognition (SAR) dataset that was first introduced by (Sharghi et al., 2020). This dataset contains 400 full-length videos extracted from 103 surgical procedures. Acquisition of the videos is performed by 4 ToF cameras positioned strategically on two different carts to properly capture the full OR. Ten activities relative to robotic system utilization are annotated on the entire dataset. This dataset was later extended with object bounding box annotations, humans and 5 OR-specific objects are annotated on around 9K frames across 20 full-length videos. To conduct our experiments we infer bounding boxes for all videos using two Cascade Mask RCNN (Cai and Vasconcelos, 2018) pretrained on the annotated bounding boxes for both humans and objects.

4.2. Data Efficient Experiments

We use UniGRU (Cho et al., 2014) as a temporal model for those experiments. We evaluate our method using an increasing amount of annotated data to demonstrate that our method can learn in a more efficient manner.

We show that ST(OR)^2 performs better than the I3D (Carreira and Zisserman, 2017) baseline for low data regime up to 10% of data, without using any visual features. The performance of ST(OR)^2 saturates when we increase the number of annotations.

4.3. Clip Classification Experiments

We evaluate our methods against both object-agnostic and object-aware approaches. The two object-aware methods we use as baselines are STRG (Wang and Gupta, 2018) and ORViT (Herzig et al., 2022) and as object-agnostic methods, we compare against I3D (Carreira and Zisserman, 2017), Timesformer (Bertasius et al., 2021) and Motionformer (Patrick et al., 2021).

For fairness, we experimented with an RPN pretrained on our dataset and one pretrained on MS-COCO (Lin et al., 2014) for STRG. We used the detected human bounding boxes in

Table 1: Results and comparison against baselines for full-time surgical activity recognition. mAP (%) is reported across three different splits for all data fractions. We provide the mean and the standard deviation averaged across splits.

Surgical Activity Recognition							
Temporal Model	**Backbone**	**Visual Features**	**2 %**	**5 %**	**10 %**	**20 %**	**100 %**
UniGRU	I3D	✓	19.7	39.2	53.5	79.5	90.7
	ST(OR)^2	×	27.3	48.8	58.2	68.3	73.6
	I3D + ST(OR)^2	✓	**29.5**	**54.2**	**60.1**	**82.3**	**91.8**

Table 2: Results and comparison against baselines for clip-based surgical action classification. top-1 and top-2 accuracies (%) are reported to deal with symmetric action confusions.

Surgical Action Classification				
Backbone	**Object-Centric Model**	**Pretrained Detector**	**Top-1 Accuracy**	**Top-2** Accuracy
I3D	×	×	87.9±	None
TimesFormer	×	×	85.7	None
MotionFormer	×	×	84.5±	None
I3D	STRG	MS-COCO	87.6	**None**
I3D	STRG	OR-SAR-H	**89.7±0.34**	**None**
I3D	STRG	OR-SAR-HO	88.4±	None
MotionFormer	ORViT	OR-SAR-H	83.9	None
MotionFormer	ORViT	OR-SAR-HO	84.1±	None
×	ST(OR)^2	OR-SAR-HO	47.3±	None
I3D	ST(OR)^2	OR-SAR-HO	89.4±	None

our ORViT implementation, with a Timesformer backbone. By combining our method with I3D features we get a 1.5% accuracy improvement against the I3D baseline. We overperform all other methods except STRG using an RPN pretrained on human bounding boxes.

4.4. Ablations

5. Conclusion and Discussion

As the field of human factor-focused activity recognition in the OR is growing, more interpretable and data-efficient methods are required. To deal with this issue we proposed an object-centric approach using OR-specific object annotations. The extra effort required to annotate object bounding boxes is compensated by the boost we are getting in a low-data regime. We manage to match baseline results on full-video activity recognition using only half the annotations for 5% annotations. This work only uses a rough geometric representation of objects, i.e., bounding boxes, to induce object awareness in visual representations; in the future, more spatially rich features like human pose for clinicians could be introduced to deal with ambiguous actions like daVinci docking and undocking.

Informed consent Data has been collected within an Institutional Review Board (IRB) approved the study and all participants informed consent has been obtained.

Acknowledgments

Acknowledgments withheld.

References

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR, 2021. URL http://proceedings.mlr.press/v139/bertasius21a.html.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018. doi: 10.1109/CVPR.2018.00644.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. doi: 10.1109/CVPR.2017.502.

Kenneth Catchpole, Colby E Perkins, Catherine Bresee, M. Jonathon Solnik, Benjamin Sherman, John L Fritch, Bruno Gross, Samantha Jagannathan, Niv Hakami-Majd, Raymund M. Avenido, and Jennifer T. Anger. Safety, efficiency and learning curves in robotic surgery: a human factors analysis. *Surgical Endoscopy*, 30:3749–3761, 2015.

Ishani Chakraborty, Ahmed Elgammal, and Randall S. Burd. Video based activity recognition in trauma resuscitation. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP*, abs/1406.1078, 2014. URL http://arxiv.org/abs/1406.1078.

Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. TeCNO: Surgical phase recognition with multi-stage temporal convolutional networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 343–352. Springer International Publishing, 2020. doi: 10.1007/978-3-030-59716-0_33. URL https://doi.org/10.1007%2F978-3-030-59716-0_33.

Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 200–210, 2020.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210, 2018.

Hassan Al Hajj, Mathieu Lamard, Pierre-Henri Conze, Béatrice Cochener, and Gwenolé Quellec. Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks. *MedIA*, 47:203–218, July 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Zhuohong He, Ali Mottaghi, Aidean Sharghi, Muhammad Abdullah Jamal, and Omid Mohareri. An empirical study on activity recognition in long surgical videos. In Antonio Parziale, Monica Agrawal, Shalmali Joshi, Irene Y. Chen, Shengpu Tang, Luis Oala, and Adarsh Subbaswamy, editors, *Proceedings of the 2nd Machine Learning for Health symposium*, volume 193 of *Proceedings of Machine Learning Research*, pages 356–372. PMLR, 28 Nov 2022. URL https://proceedings.mlr.press/v193/he22a.html.

Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3138–3149. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00315. URL https://doi.org/10.1109/CVPR52688.2022.00315.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997. doi: 10.1162/neco.1997.9.8.1735.

Thibaut Issenhuth, Vinkle Kumar Srivastav, Afshin Gangi, and Nicolas Padoy. Face detection in the operating room: comparison of state-of-the-art methods and a self-supervised approach. *IJCARS*, 2019.

Falisha Kanji, Tara Cohen, Myrtede Alfred, Ashley Caron, Samuel Lawton, Stephen Savage, Daniel Shouhed, Jennifer T Anger, and Ken Catchpole. Room size influences flow in robotic-assisted surgery, Jul 2021. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8345669/.

Colin Lea, René Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In Gang Hua and Hervé Jégou, editors, *ECCV 2016 Workshops*, pages 47–54, Cham, 2016. Springer International Publishing. ISBN 978-3-319-49409-8.

Zhaoshuo Li, Amirreza Shaban, Jean-Gabriel Simard, Dinesh Rabindran, Simon P. DiMaio, and Omid Mohareri. A robotic 3d perception system for operating room environment awareness. *IPCAI*, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

Zelun Luo, Jun-Ting Hsieh, Niranjan Balachandar, Serena Yeung, Guido Pusiol, Jay S. Luxenberg, Grace Li, Li-Jia Li, Arnold Milstein, and Li Fei-Fei. Vision-based descriptive analytics of seniors ' daily activities for long-term health monitoring. 2018.

Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. *CVPR*, abs/1912.09930, 2019. URL http://arxiv.org/abs/1912.09930.

Ege Özsoy, Evin Pinar Örnek, Ulrich Eck, Tobias Czempiel, Federico Tombari, and Nassir Navab. 4d-or: Semantic scene graphs for OR domain modeling. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part VII*, volume 13437 of *Lecture Notes in Computer Science*, pages 475–485. Springer, 2022. doi: 10.1007/978-3-031-16449-1_45. URL https://doi.org/10.1007/978-3-031-16449-1_45.

Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12493–12506, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/67f7fb873eaf29526a11a9b7ac33bfac-Abstract.html.

Lauren Schiff, Ziv Tsafrir, Joe Aoun, Andrew R. Taylor, Evan Theoharis, and David Eisenstein. Quality of communication in robotic surgery and surgical outcomes. *JSLS : Journal of the Society of Laparoendoscopic Surgeons*, 20, 2016.

Adam Schmidt, Aidean Sharghi, Helene Haugerud, Daniel Oh, and Omid Mohareri. Multi-view surgical video action detection via mixed global view attention. In *MICCAI*, 2021.

Aidean Sharghi, Helene Haugerud, Daniel Oh, and Omid Mohareri. Automatic operating room surgical activity recognition for robot-assisted surgery. In *MICCAI*, 2020.

Dimick JB. Sheetz KH, Claflin J. Trends in the adoption of robotic surgery for common surgical procedures. *JAMA Network Open*, 2020.

Vinkle Kumar Srivastav, Thibaut Issenhuth, Abdolrahim Kadkhodamohammadi, Michel de Mathelin, Afshin Gangi, and Nicolas Padoy. Mvor: A multi-view rgb-d operating room dataset for 2d and 3d human pose estimation. *ArXiv*, abs/1808.08180, 2018.

Vinkle Kumar Srivastav, Afshin Gangi, and Nicolas Padoy. Human pose estimation on privacy-preserving low-resolution depth images. *MICCAI*, abs/2007.08340, 2019.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331:1279 – 1285, 2011.

Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. EndoNet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE TMI*, 36(1):86–97, January 2017.

Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. *ECCV*, abs/1806.01810, 2018. URL http://arxiv.org/abs/1806.01810.

Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Is an object-centric video representation beneficial for transfer? *ACCV*, abs/2207.10075, 2022. doi: 10.48550/arXiv.2207.10075. URL https://doi.org/10.48550/arXiv.2207.10075.

Appendix A. Implementation Details