INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

PoCaPNet: A Novel Approach for Surgical Phase Recognition Using Speech and X-Ray Images

Kubilay Can Demir[1], *Tobias Weise*[1,2], *Matthias May*[3], *Axel Schmid*[3], *Andreas Maier*[2], *Seung Hee Yang*[1]

[1]SLU Lab. Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2]Pattern Recognition Lab. Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[3]Department of Radiology Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

kubilay.c.demir@fau.de, seung.hee.yang@fau.de

Abstract

Surgical phase recognition is a challenging and necessary task for the development of context-aware intelligent systems that can support medical personnel for better patient care and effective operating room management. In this paper, we present a surgical phase recognition framework that employs a Multi-Stage Temporal Convolution Network using speech and X-Ray images for the first time. We evaluate our proposed approach using our dataset that comprises 31 port-catheter placement operations and report 82.56 % frame-wise accuracy with eight surgical phases. Additionally, we investigate the design choices in the temporal model and solutions for the class-imbalance problem. Our experiments demonstrate that speech and X-Ray data can be effectively utilized for surgical phase recognition, providing a foundation for the development of speech assistants in operating rooms of the future.

Index Terms: surgical workflow, surgical phase recognition, speech assistant, port-catheter placement, TCN

1. Introduction

Modern operating rooms (ORs) are optimized for better patient care and the most effective utilization of medical resources. Advances in technology are presented in ORs with cutting-edge surgical tools, monitoring, and navigation systems. These systems enable physicians to perform more complex surgical procedures with a high success rate that was not possible before [1]. Simultaneous to these advances, the amount of data created by modern medical systems is increasing [2]. This data is necessary for the successful execution of the operation and needs to be processed by the medical personnel after or during operations. It has been proposed that intelligent systems processing this growing amount of information and projecting it in the correct time and format will be vital in the future of ORs [3, 4].

Surgical phase recognition (SPR) is a topic of automatically extracting semantic information from ongoing or recorded surgical operations by recognizing different predefined phases [5]. The highest level actions performed in the operating room, such as anesthesia, sterilizing, or cutting, are referred to as surgical phases. Robust estimation of these phases is a prerequisite for the development of the envisioned context-aware intelligent assistants in the OR.

The majority of studies in SPR focused on *laparoscopic cholecystectomy*, removal of the gallbladder, through endoscopic videos [6, 7, 8, 9, 10]. Microscopic videos are another popular source of information used mainly for cataract surgeries [11, 12, 13]. Sensory data from robotic surgeries are considered together with surgical videos in several studies [14, 10]. As these modalities are not used in every operation, it is not possible to cover all types of operations with endoscopic and microscopic videos. Moreover, these data are typically recorded inside or near the body and do not contain information about the environment of the OR. Despite the growing interest in SPR, the use of speech and audio data has received little attention [15]. *Guzmán-García et al.* [16] extracted Spanish transcriptions from 15 online education videos on *laparoscopic cholecystectomy* and achieved 82.95 % accuracy in surgical phase recognition. *Seibold et al.* [17] used discrete segments from the German audio dataset of five *total hip arthroplasty* operation for the phase recognition task and recognized seven phases with 95.60 % accuracy. We hypothesize that using speech and audio is a necessary direction in SPR as it can open the way for various applications based on natural language processing (NLP) and enable the development of interactive smart assistants in ORs.

In this study, we propose a novel approach for SPR using speech and X-Ray images collected during port-catheter placement operations. Our method utilizes Multi-Stage Temporal Convolutional Network (MS-TCN) [18] architecture and past estimations for temporal modeling, leverages wav2vec 2.0 XLSR-53 [19] representations for speech signals and TorchXRayVision [20] representations for X-Ray images.

The contribution of our work as follows:

- To best of our knowledge, it is the first approach in SPR that utilizes speech data from entire surgical operations and combination with X-Ray images.
- We analysed usage of positional encodings and previous estimates for integrating long term temporal information.
- We analysed class-imbalance problem with re-weighting and modified loss functions for recognizing short duration surgical phases.

Our study is organized as follows: we explain our proposed framework in Section 2; we report our findings in Section 3, and we give our conclusion in Section 4.

2. Proposed Method

2.1. Intervention

Port-catheter placement is a frequently applied minimally-invasive procedure in the radiology department. The main purpose of this operation is to place a port under the skin of the chest, which is connected via a catheter to the large veins emptying into the heart. This operation prevents injuries to small vessels after repetitive infusions during treatments such as chemotherapy [21]. The procedure is typically performed by a single physician and an assistant. The duration of an operation varies between half an hour to three hours depending on the experience level of the medical personnel, complications during the operation or the general condition of the patient.

10.21437/Interspeech.2023-753

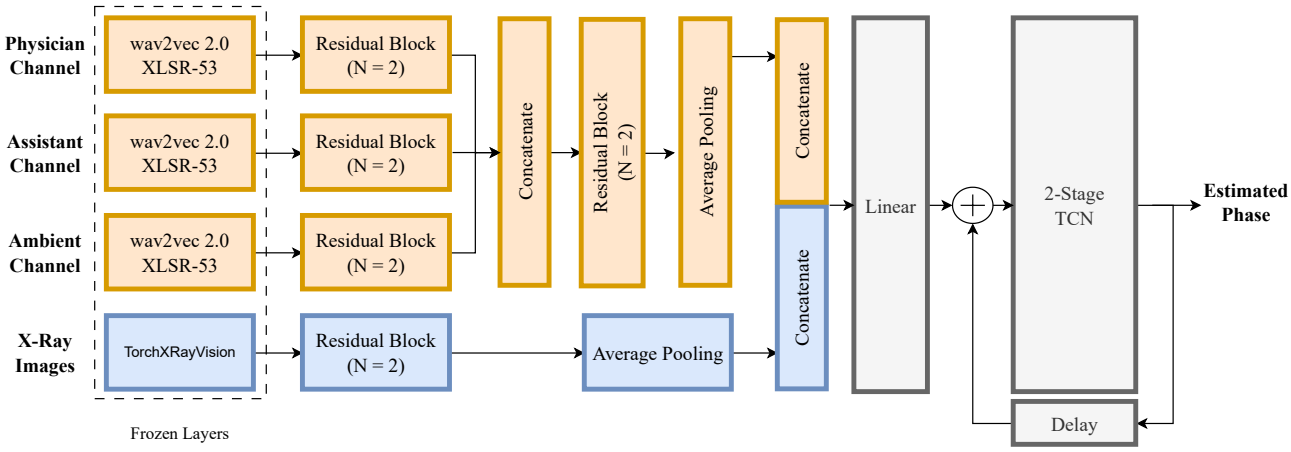

Figure 1: *The framework of the proposed model. The orange-colored blocks represent the audio branch, blue-colored blocks represent the visual branch and grey-colored blocks represent the temporal model. N shows the number of residual blocks.*

2.2. Feature Extraction Backbone

For speech and audio signals, self-supervised approaches such as wav2vec 2.0 [22] have been widely used. To create better representations for languages with little available data, the wav2vec 2.0 model extended to multi-language setting in XLSR [19]. A large variant XLSR-53 is trained with 50k hours of public data from 53 languages, including German which is the language of the experiment corpus. Therefore, we chose XLSR-53 as the backbone model for our framework. In our study, we used the output of the final Transformer layer. The motivation for this choice is to include maximum temporal information in the feature vectors. In the windowing step, we used seven-seconds casual Hann window with a one-second hop length. The seven-second long audio data of each channel is used as the input to the feature extraction backbone.

For X-Ray images, we used Densenet121 [23] model pre-trained on all publicly available chest X-Ray datasets. The details of the pre-trained model and the corresponding TorchXRayVision library are released in [20]. The pretraining setting and datasets are closely relevant to the port-catheter placement intervention as chest X-Rays are used. To have the same temporal resolution as the speech and audio features, we extracted features from seven X-Ray images, which correspond to seven seconds at a $1fps$ rate, and shifted a single image at each time step. Both wav2vec 2.0 XLSR-53 and Densenet121 models held frozen during training.

2.3. Temporal Model

Modeling temporal relations is a vital component of SPR architectures. To have a large receptive field, we used a two-stage TCN network in our temporal model. Although Transformer networks became very popular and achieved state-of-the-art results in many applications, we chose TCN architecture in our network due to limited available training data. We used past estimates in an auto-regressive manner to further increase the receptive field of the network. The effect of our design choices are experimented in Section 3.2.

Our model runs at a rate of one second, i.e. an estimation is made every second utilizing features from a window of seven seconds as explained in the previous section. The overview of our whole proposed framework is depicted in Figure 1. Our source code is available at: `https://github.com/kubicndmr/PoCaPNet`.

3. Experiments

3.1. Dataset

In our study, we used the PoCaP Corpus containing 31 port-catheter placement operations recorded in the Radiology Department of University Hospital Erlangen, Germany [24]. In this dataset, the physician and the medical assistant wore Sennheiser XSW 2 ME3-E wireless headsets for the audio recording. All conversations in the dataset are held in German. Additionally, the internal microphone of a single GoPro Hero 8 camera, which is initially set up for easing annotation work, is also included in the dataset as the ambient microphone channel. All channels are aligned and then re-sampled at $16kHz$. X-Ray images are captured from the output of the X-Ray machine at a $1fps$ rate. The data set is unfortunately not publicly available due to local laws for protecting the data privacy of the patients and the medical personnel. To the best of our knowledge, any other multi-modal data set of any operation type containing full-length speech signals is not existing for repeating our experiments. Thus, we can only provide our results with our in-house dataset.

The recordings suffer significant data loss during the six operations. Data loss happened during data recording for a variety of causes, such as software failure, a change in operating personnel, or a recording error. Thus, we excluded these recordings and used the remaining 25 operations for our experiments. For the training, validation, and test set separation, we employed the conventional $60-20-20$ percent random split.

The dataset contains eight surgical phases and a transition phase. These are: *Preparation*, *Puncture*, *Positioning of the Guide Wire*, *Pouch Preparation and Catheter Placement*, *Catheter Positioning*, *Catheter Adjustment*, *Catheter Control*, and *Closing*. Transition phases are defined at phase borders as instant stops, breaks, talks, or behaviors that could belong to both surgical phases and are not considered during the training and testing. These few seconds long relaxation periods are proposed to address ambiguity in phase annotations.

3.2. Temporal Relation

Problem: For successful and robust phase recognition, long-distance temporal information must be aggregated to the estimation step at the current time frame. Even for humans, trying to estimate a surgical phase from a short data window without a prior knowledge would be challenging or even may not be possible in some cases. With our choice of TCN architecture, we aimed to mitigate this problem. However, convolutional layers cannot access previous frames outside of the mini-batch, preventing them from leveraging this crucial information. This factor limits the receptive field of the proposed models to batch size. Previous estimation steps can contain relevant information for the current estimation step.

Proposed Solution: To address this problem, we designed an experiment with three settings: (1) Two-Stage TCN. In this setting, the output of the audio and visual branches are concatenated and used as input to the linear layer. The Two-Stage TCN network uses this input for surgical phase recognition; (2) Positional Encodings [25]. In addition to the previous setting, we added positional encodings to the output of the linear layer via summation. We aimed to provide timely information to the current estimation, thus, distinguishing similar-looking inputs via their time order. For example, the beginning and the end of each operation have similar audio and visual characteristics and belong to different classes, e.g. *Preparation* and *Closing*. We claim that it would be easier to differentiate these classes by knowing their position within the ongoing operation; (3) Auto-regressive delayed estimations. In this setting, we used the phase estimations from the previous mini-batch to create an additional memory-like feature instead of positional encodings. We added this vector to the output of the linear layer via summation, see Figure 1. As a result, we attempted to keep track of both position and phase order simultaneously. We claim that using the previous estimation includes necessary time order information as positional encodings and additional phase order information. Similar to the previous analogy, it would be easier to classify an input at the end of the operation as *Closing* by knowing that the previously estimated phase was the *Catheter Control*.

Implementation Details: After our initial experiments in the first setting with single, two, and three-stage TCN models, we observed the best results with the two-stage model and used this model for all experiments. The two-stage model has 2.8 million parameters. We performed our experiments using class-weighted cross-entropy loss and Adam [26] optimizer with weight decay $1E-6$, learning rate $9E-6$, and batch size 512. Our method was implemented in PyTorch and our models were trained on a single NVIDIA RTX 3090 Ti 24 GB GPU.

Evaluation & Discussion: In the evaluation, we used frame-wise accuracy and weighted F1-score. Our results are presented in Table 1. We observed lowest accuracy in the first setting with the vanilla two-stage TCN network. In parallel to our hypothesis, we achieved significant performance improvement with the addition of temporal connections. By using positional encodings, we reached approximately 8.5 points higher accuracy. By using auto-regressive delayed connection, we achieved 3.5 points further increase in the accuracy. In F1-score results, similar results can be seen.

3.3. Class-Imbalance

Problem: Although it is equally important from clinical perspective to recognize all phases robustly, the durations of surgical phases of the port-catheter placement procedure and the corresponding data are not evenly distributed. This class-imbalance problem possesses a challenge during both training and testing. Figure 2 shows distribution of durations via Gaussian density estimation for each surgical phase and cumulative means in the PoCaP Corpus. The phases *Guide Wire* (green), *Catheter Positioning* (brown), and *Catheter Control* (grey) have considerably shorter durations when compared to *Puncture* (orange), *Catheter Placement* (red), and *Catheter Adjustment* (pink) phases. Thus, it is difficult to recognize these phases.

Table 1: *Phase recognition results of three settings of the temporal model: (1) Two-Stage TCN, (2) Addition of Positional Encodings (3) Addition of past delayed estimation.*

Temporal Connection	Accuracy	F1-Score
Two-Stage TCN	67.94 ± 8.67	68.62 ± 9.22
+ Positional Encoding	76.48 ± 4.92	76.15 ± 5.91
+ Delayed Estimation	79.99 ± 7.57	80.74 ± 6.47

Proposed Solution: Several strategies are proposed to mitigate class-imbalance problem in various research topics, including re-weighting of classes and modified loss functions. In this experiment, we considered these techniques with our framework and tested following settings: (1) Cross-entropy (CE) loss. Although an intuitive decision given the strong class-imbalance problem would be the class-weighted cross-entropy loss, we used this setting to illustrate the effects of the re-weighting method in the next step; (2) Class-weighted cross-entropy. In this case, class weights are calculated as inverse frequencies; (3) Focal loss [27]. In this setting, we aimed to give less importance to well-classified phases; (4) Label-distribution-aware margin (LDAM) loss [28]. We attempted to stimulate larger margins for short duration phases in this setting.

Implementation Details: In all experiments in this section, we used the two-stage TCN model with the delayed estimation connection. We kept all other variables and hyperparameters except loss functions the same as in the previous section. In the focal loss experiment, we tested different γ values and reported the best results with $\gamma = 2$.

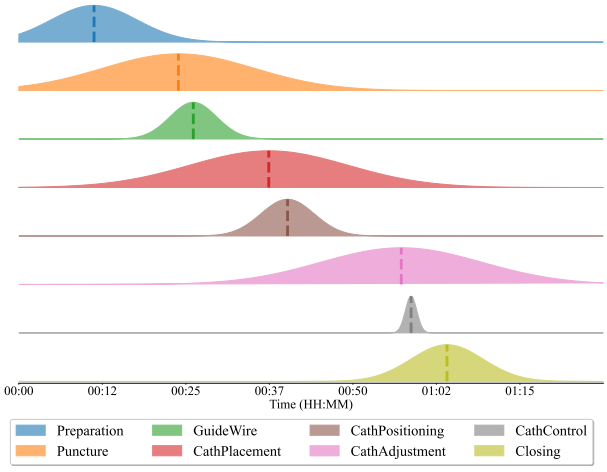

Figure 2: *Cumulative mean (dashed line) and Gaussian density estimation (shaded area) of the durations for eight surgical phases in the PoCaP Corpus.*

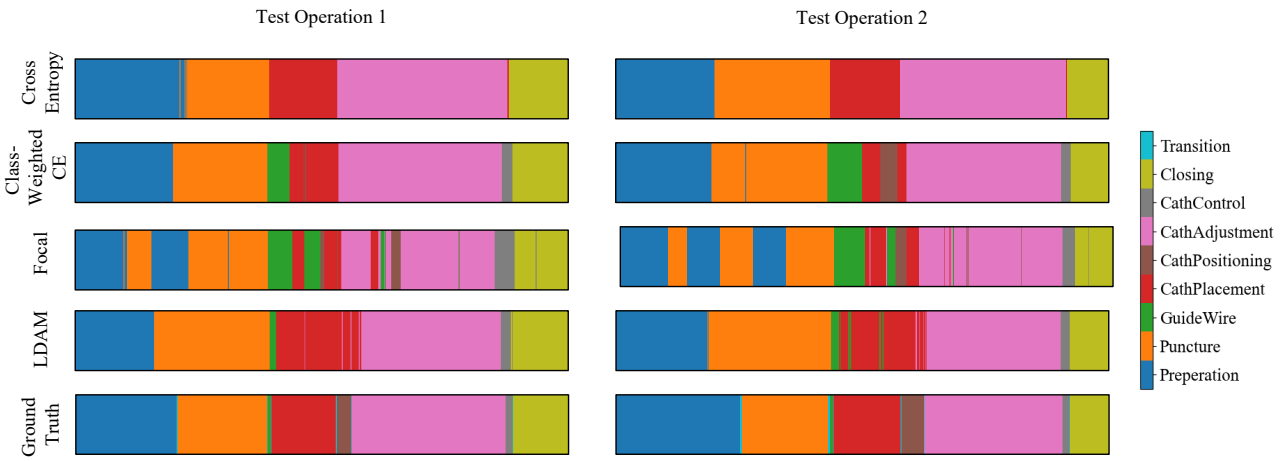

Figure 3: *Visualization of estimated surgical phases and ground truth labels of two port-catheter placement operations.*

Evaluation & Discussion: In the evaluation, we used frame-wise accuracy, weighted F1-score, and ribbon plots. Our results are presented in Table 2 and Figure 3. We achieved the highest frame-wise accuracy with cross-entropy loss, however, as depicted in the first row of the Figure 3, this setting could only correctly classify dominant phases *Preparation* (blue), *Puncture* (orange), *Catheter Placement* (red), *Catheter Adjustment* (pink), and *Closing* (yellow). This was an expected behavior since we handled all phases equally without considering the label distribution of the dataset. Although the overall classification rate is high, this is not a clinically desired output. In the second setting, we used class-weighted cross-entropy and could recognize *Guide Wire* (green) and *Catheter Control* (grey) phases additionally. *Catheter Positioning* (brown) is either misclassified or entirely missed and *Guide Wire* (green) is over-emphasized in all test data. The second row of the Figure 3 shows the results for this setting. In the third experiment, we observed similarly over-emphasized short phases *Guide Wire* (green) and *Catheter Control* (grey) phases and many phase shifts with the focal loss function. In addition, we obtained the lowest accuracy and F1-score results in this experiment. This could be resulted from probabilities of class estimations being not significantly different for easy and hard samples. Example results are illustrated in the third row of Figure 3. In the final setting, we achieved the most consistent results with LDAM loss. In this case, we could recognize all phases except *Catheter Positioning* (brown) consistently and achieved 2.5 points better accuracy than class-weighted cross-entropy loss. Results are shown in the fourth row of the Figure 3. We think that this setting provides the most robust results for possible clinical applications. *Catheter Positioning* (brown) phase is consistently misclassified in all experiments. We claim that this is caused by this phase being both very difficult to distinguish from neighboring phases and having a short duration. In contrast to this phase, another short phase *Catheter Control* (grey) phase has a very distinctive X-Ray setting, which makes it easier to recognize, thus, it is recognized better in all experiments.

In future studies, we would like to focus on *Catheter Positioning* (brown) phase specifically. Moreover, we would like to experiment with the individual effects of each microphone channel and X-Ray input. Physicians and assistants have different tasks during an intervention and their contributions would be different. An ambient microphone channel typically captures all background sounds in the OR and has a noisy input. The X-ray channel provides very sparse but informative data. Thus, understanding the contribution of multi-modal data is a promising research direction. Finally, we would like to test our approach with different interventions, medical institutes and languages.

Table 2: *Phase recognition results of four settings with different loss functions proposed for class-imbalance problem.*

Loss Function	**Accuracy**	**F1-Score**
Cross Entropy	84.82 ± 6.76	82.24 ± 6.58
Class-Weighted CE	79.99 ± 7.57	80.74 ± 6.47
Focal	70.19 ± 4.90	58.35 ± 3.39
LDAM	82.56 ± 3.21	81.30 ± 3.89

4. Conclusion

In this work, we introduce the PoCaPNet, a framework for surgical phase recognition using speech and X-Ray images. Our study is based on audio features extracted from three different microphone channels using the wav2vec 2.0 XLSR-53 model and visual features extracted from X-Ray images using the TorchXRayVision model. To the best of our knowledge, we are the first to employ speech data from the entire intervention and X-Ray data in combination for the surgical phase recognition task. Aggregating long-term temporal information and learning with class-imbalanced data were the two biggest problems in our study. We proposed using delayed estimation in an auto-regressive manner to integrate past temporal information into the current classification step and LDAM loss to address the class-imbalance problem. Our experimental results show significant performance improvement with these additions. This study shows proof-of-concept for using speech data in an SPR task. Our results encourage the development of many new applications such as interactive intelligent assistants in ORs.

5. Acknowledgements

We gratefully acknowledge funding for this study by Friedrich-Alexander-University Erlangen-Nuremberg, Medical Valley e.V. and Siemens Healthineers AG within the d.hip framework.

6. References

[1] L. Maier-Hein, S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou *et al.*, "Surgical data science for next-generation interventions," *Nature Biomedical Engineering*, vol. 1, no. 9, pp. 691–696, 2017.

[2] T. Vercauteren, M. Unberath, N. Padoy, and N. Navab, "Cai4cai: the rise of contextual artificial intelligence in computer-assisted interventions," *Proceedings of the IEEE*, vol. 108, no. 1, pp. 198–214, 2019.

[3] K. R. Cleary, "Or2020: the operating room of the future," GEORGETOWN UNIV WASHINGTON DC MEDICAL CENTER, Tech. Rep., 2004.

[4] D. W. Rattner and A. Park, "Advanced devices for the operating room of the future," in *Seminars in laparoscopic surgery*, vol. 10, no. 2. Sage Publications Sage CA: Thousand Oaks, CA, 2003, pp. 85–89.

[5] F. Lalys and P. Jannin, "Surgical process modelling: a review," *International journal of computer assisted radiology and surgery*, vol. 9, pp. 495–511, 2014.

[6] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "Endonet: a deep architecture for recognition tasks on laparoscopic videos," *IEEE transactions on medical imaging*, vol. 36, no. 1, pp. 86–97, 2016.

[7] T. Czempiel, M. Paschali, M. Keicher, W. Simson, H. Feussner, S. T. Kim, and N. Navab, "Tecno: Surgical phase recognition with multi-stage temporal convolutional networks," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer, 2020, pp. 343–352.

[8] Y. Jin, H. Li, Q. Dou, H. Chen, J. Qin, C.-W. Fu, and P.-A. Heng, "Multi-task recurrent convolutional network with correlation loss for surgical video analysis," *Medical image analysis*, vol. 59, p. 101572, 2020.

[9] X. Gao, Y. Jin, Y. Long, Q. Dou, and P.-A. Heng, "Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*. Springer, 2021, pp. 593–603.

[10] C. I. Nwoye, T. Yu, C. Gonzalez, B. Seeliger, P. Mascagni, D. Mutter, J. Marescaux, and N. Padoy, "Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos," *Medical Image Analysis*, vol. 78, p. 102433, 2022.

[11] M. J. Primus, D. Putzgruber-Adamitsch, M. Taschwer, B. Münzer, Y. El-Shabrawi, L. Böszörmenyi, and K. Schoeffmann, "Frame-based classification of operation phases in cataract surgery videos," in *MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I 24*. Springer, 2018, pp. 241–253.

[12] T. Xia and F. Jia, "Against spatial–temporal discrepancy: Contrastive learning-based network for surgical workflow recognition," *International journal of computer assisted radiology and surgery*, vol. 16, no. 5, pp. 839–848, 2021.

[13] B. Qi, X. Qin, J. Liu, Y. Xu, and Y. Chen, "A deep architecture for surgical workflow recognition with edge information," in *2019 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE, 2019, pp. 1358–1364.

[14] D. Paysan, L. Haug, M. Bajka, M. Oelhafen, and J. M. Buhmann, "Self-supervised representation learning for surgical activity recognition," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, pp. 2037–2044, 2021.

[15] K. C. Demir, H. Schieber, T. Weise, D. Roth, A. Maier, and S. H. Yang, "Deep Learning in Surgical Workflow Analysis: A review," 10 2022. [Online]. Available: https://www.techrxiv.org/articles/preprint/Surgical_Phase_Recognition_A_Review_and_Evaluation_of_Current_Approaches/19665717

[16] C. Guzmán-García, M. Gómez-Tome, P. Sánchez-González, I. Oropesa, and E. J. Gómez, "Speech-based surgical phase recognition for non-intrusive surgical skills' assessment in educational contexts," *Sensors*, vol. 21, no. 4, p. 1330, 2021.

[17] M. Seibold, A. Hoch, M. Farshad, N. Navab, and P. Fürnstahl, "Conditional generative data augmentation for clinical audio datasets," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII*. Springer, 2022, pp. 345–354.

[18] Y. A. Farha and J. Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3575–3584.

[19] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[20] J. P. Cohen, J. D. Viviano, P. Bertin, P. Morrison, P. Torabian, M. Guarrera, M. P. Lungren, A. Chaudhari, R. Brooks, M. Hashir, and H. Bertrand, "TorchXRayVision: A library of chest X-ray datasets and models," in *Medical Imaging with Deep Learning*, 2022. [Online]. Available: https://github.com/mlmed/torchxrayvision

[21] S. J. Gonda and R. Li, "Principles of subcutaneous port placement," *Techniques in vascular and interventional radiology*, vol. 14, no. 4, pp. 198–203, 2011.

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[24] K. C. Demir, M. May, A. Schmid, M. Uder, K. Breininger, T. Weise, A. Maier, and S. H. Yang, "Pocap corpus: A multimodal dataset for smart operating room speech assistant using interventional radiology workflow analysis," in *Text, Speech, and Dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic, September 6–9, 2022, Proceedings*. Springer, 2022, pp. 464–475.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[28] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *Advances in neural information processing systems*, vol. 32, 2019.