Deep Learning in Surgical Workflow Analysis: A Review of Phase and Step Recognition

Kubilay Can Demir, Hannah Schieber, Tobias Weise, Daniel Roth, *Member, IEEE*, Matthias May, Andreas Maier, *Senior Member, IEEE*, and Seung Hee Yang

Abstract*—*Objective: **In the last two decades, there has been a growing interest in exploring surgical procedures with statistical models to analyze operations at different semantic levels. This information is necessary for developing context-aware intelligent systems, which can assist the physicians during operations, evaluate procedures afterward or help the management team to effectively utilize the operating room. The objective is to extract reliable patterns from surgical data for the robust estimation of surgical activities performed during operations. The purpose of this article is to review the state-of-the-art deep learning methods that have been published after 2018 for analyzing surgical workflows, with a focus on phase and step recognition.** ***Methods:*** **Three databases, IEEE Xplore, Scopus, and PubMed were searched, and additional studies are added through a manual search. After the database search, 343 studies were screened and a total of 44 studies are selected for this review.** ***Conclusion:*** **The use of temporal information is essential for identifying the next surgical action. Contemporary methods used mainly RNNs, hierarchical CNNs, and Transformers to preserve long-distance temporal relations. The lack of large publicly available datasets for various procedures is a great challenge for the development of new and robust models. As supervised learning strategies are used to show proof-of-concept, self-supervised, semi-supervised, or active learning methods are used to mitigate dependency on annotated data. Significance: The present study provides a comprehensive review of recent methods in surgical workflow analysis, summarizes commonly used architectures, datasets, and discusses challenges.**

***Index Terms*—Surgical workflow analysis, surgical data science, surgical phase recognition, deep learning.**

Manuscript received 13 February 2023; revised 21 July 2023; accepted 26 August 2023. Date of publication 4 September 2023; date of current version 7 November 2023. This work was supported in part by the U.S. Department of Commerce under Grant 123456, in part by Friedrich-Alexander-University Erlangen-Nuremberg, Medical Valley e.V., and in part by Siemens Healthineers AG within the d.hip framework. *(Corresponding author: Kubilay Can Demir.)*

Kubilay Can Demir, Hannah Schieber, Tobias Weise, and Seung Hee Yang are with the Department of Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91054 Erlangen, Germany (e-mail: kubilay.c.demir@fau.de; hannah.schieber@fau.de; tobias.weise@fau.de; seung.hee.yang@fau.de).

Daniel Roth is with the Technical University of Munich, School of Medicine and Health, Klinikum rechts der Isar, Orthopaedics and Sports Orthopaedics, 81675 Munich, Germany (e-mail: daniel.roth@tum.de).

Andreas Maier is with the Pattern Recognition Laboratory, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91054 Erlangen, Germany (e-mail: andreas.maier@fau.de).

Matthias May is with the Department of Radiology, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91054 Erlangen, Germany (e-mail: matthias.may@uk-erlangen.de).

Digital Object Identifier 10.1109/JBHI.2023.3311628

I. Introduction

TO PROVIDE better treatment to patients, increase the success rate in surgeries, and ensure cost-effective utilization of the Operating Room (OR), new types of equipment and functionalities are continuously being added to contemporary medical systems in hospitals [1]. ORs are evolving strongly towards digitalized environments and increasing the ability of physicians to carry out more complex and successful surgical procedures [2]. In addition to being better surgical instruments, these systems collect and display sensory data for navigation and monitoring purposes. As the ORs are evolving into more complex environments, the volume of data to be analyzed during or after the surgery is increasing rapidly [3], [4]. This data is essential for a successful surgery but it may also hinder the smooth execution of the procedure if it is not presented in the correct time and format. Integrating multiple sensors and orchestrating all data is an important aspect of future ORs [5]. In addition to the complexity and corresponding volume of medical data, the number of patients is also increasing, multiplying the workload in the OR. Consequently, new methods have to be considered to make the best possible use of the OR. Standardization of surgical routines and integration of intelligent systems into the surgical workflow is proposed by Herfarth et al. [6] to address this problem. Although surgeries are complex procedures, the same operation types often have similar patterns. These patterns can be analyzed by smart systems [7]. In that sense, an intelligent system, which can understand the actions in an OR, process the medical information accordingly, and represent it in a desirable way or trigger predefined events, would be very beneficial [8], [9], [10].

Standardization of surgical procedures can be beneficial for better execution of operations. It also contributes to the reliable analysis of surgical workflow systems [11]. The term Surgical Workflow Analysis (SWA) is used for referring to automatic methods to extract meaningful patterns for any semantic purpose from surgical procedures. A common method in surgical workflow analysis is the Surgical Process Model (SPM), which defines a surgical procedure with predefined smaller representations [12]. In this approach, a surgical process is hierarchically decomposed into predefined actions at different granularity levels, and classification algorithms are employed afterward for detecting these predefined blocks, i.e., surgical actions. It should be noted that starting and ending points of these surgical actions are also to be estimated together with the classification task. Given the surgical data, segmentation of the

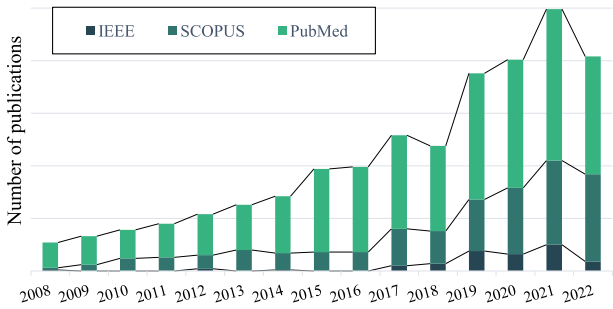

Fig. 1. Search results of surgical workflow analysis in different databases since 2008. There is a clear trend of increasing interest in surgical workflow analysis in recent years. The year bar 2022 shows studies presented before the term search, August 2022.

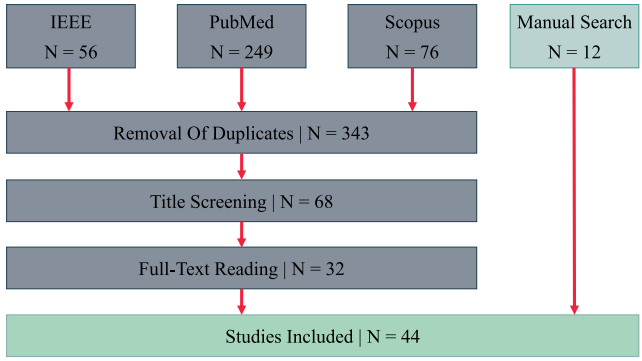

Fig. 2. Process of selecting the related studies for our review. The grey rectangles show the database search, the blue rectangle represents the manual search, and the green rectangle represents the total number of selected studies.

actions is not known a priori. The predefined representations aim to achieve specific objectives in different granularity levels and can be structured systematically [13]. *Surgical phases* and *surgical steps* are the terms commonly used to decompose an operation. Surgical phases are the main periods of intervention and refer to the highest-level actions in the OR such as anesthesia, sterilization, or cutting. Steps are actions needed to accomplish surgical objectives of phases, as they are more fine-grained units. Some examples are preparing instruments, setting, or removing covers [14], [15]. Recognition of these surgical actions provides semantic information about the surgical procedure and opens a way for various applications [16], [17]. We limit the scope of our review to phases, steps, or similarly defined surgical actions. For a specific application, more detailed information could be necessary. In that case, the decomposition can be further processed into finer-grained actions such as *activities*, *gestures*, and *dexemes* [16].

The earliest works on the analysis of surgical procedures by decomposing them into sub-parts are published in 2001 by *Jannin* et al. [12] and *MacKenzie* et al. [15]. Initial approaches used classical machine learning pipelines with statistical feature extraction methods and classifiers such as Support Vector Machine, Random Forest, or Hidden Markov Model [17]. However, these models resulted in limited success. Due to the tremendous classification and recognition capabilities of multi-layer neural networks, deep learning models are used progressively in surgical recognition tasks and their potential is confirmed by many studies [17]. The annual number of publications related to surgical workflow analysis since then is depicted in Fig. 1, showing a strongly increasing interest in the last few years. This trend is overlapping with the recent breakthroughs in deep learning. Therefore, we are focusing on deep learning methods in this review to provide a comprehensive overview of recent methods and data.

There are already reviews in the field of surgical workflow analysis. *Lalys and Jannin* [14] reviewed acquisition, modeling, analysis, application, and evaluation techniques for surgical process modeling in 2014. They provided the taxonomy and compared different approaches in their analysis. *Antunes* et al. [18] reviewed sensors applied for capturing the workflow of healthcare environments and analyzed gaps in this application area. *Garrow* et al. [19] have provided an overview of the latest algorithms and data sources for surgical phase recognition in 2018. *Junger* et al. [20] have investigated 58 studies published between 2010 and 2019 at different granularity levels with a focus on applicability and transferability. *Amsterdam* et al. [16] have focused on low-level granularity and reviewed gesture recognition techniques in robotic surgery.

The contribution of our review is fourfold: we review studies published after 2018 on high-level surgical workflow analysis such as phase and step recognition, we point out current challenges in the field, we provide a structured analysis of contemporary methods used for feature extraction and modeling distant temporal relations, we list publicly available datasets for surgical phase or step recognition, and we discuss challenges and what could be future research directions.

II. Methodology

The literature search for selecting studies is conducted following the PRISMA guidelines [21].

Literature Search: The literature search is carried out by using IEEE Xplore, Scopus, and PubMed databases. Since the terminology for surgical workflow analysis is not necessarily the same in every study, we have decided to include synonyms that might relate to our search in order to achieve a better exploration of the available literature. For our database search, we included the terms: *surgical workflow analysis*, *surgical workflow recognition*, *surgical workflow segmentation*, *surgical workflow detection*, *surgical phase recognition*, and *surgical step recognition*. Only studies between 2018 and July 2022, the time of writing this article, were considered.

The initial search yielded 381 studies, which were reduced to 343 unique studies after removing duplicates. During the following title screening process, review papers and unrelated studies were eliminated. A full-text reading for the remaining 68 studies followed and we ultimately selected 32 studies. In addition to the database search, a manual search was conducted afterward using Google Scholar and PubMed, which was guided by references and the above-mentioned terms. Here, we selected 12 studies, and thus the total number of studies for our review increased to 44. The overall study selection procedure is illustrated in Fig. 2.

Inclusion and Exclusion Criteria: In the full-text reading part, we included studies explicitly explaining their model architecture, and reporting their results on surgical phase recognition, step recognition, or with a similar level of detail. In this review, we have constrained our search to deep learning methods, thus, in this full-text reading part, we have selected only studies that utilized deep learning techniques. Furthermore, we included only peer-reviewed studies and excluded studies that were not published in English, used only synthetic data, or focused on

animal surgeries. We applied these exclusion criteria since such types of data and surgeries may not be representative enough for human surgeries and thus potentially negatively affect model comparisons.

Study Selection: Literature search, removal of duplicates, and title screening were performed by a single reviewer. The full-text reading inclusion and exclusion criteria were validated by a second reviewer. Disagreements were resolved by agreement after a discussion or by consulting other reviewers' opinions. The list of selected studies is given in Table I.

III. Challenges

Achieving accurate surgical phase and step recognition poses significant challenges which can be broadly categorized into two main areas: model and data-based challenges. Model-based challenges arise during the design and development of deep learning models. Data-based challenges pertain to the availability and quality of the datasets.

A. Model-Based

1) Temporal Relation Modeling: A well-known problem in surgical workflow analysis is aggregating the necessary long-term temporal information into an estimation at the current time step. This information is vital when the data window used for estimation does not contain sufficient distinctive information to generate reliable outputs. An example in Laparoscopic Cholecystectomy (LC) operations is the *limited inter-phase and high intra-phase variance* among frames, which refers to weak dissimilarities of signals in different classes and substantial differences in the same classes [42], [58]. These frames can be recognized correctly by aggregating related temporal information. The temporal-relation models in selected studies are analyzed in Section IV-B.

2) Class-Imbalance: Although recognizing all phases is equally important for clinicians, the distribution of samples in each phase may vary significantly. The *class-imbalance problem* hinders the learning capabilities of the models [28], challenges correct performance evaluation [60], and results in missing recognition of short phases or steps [55]. The problem originates from the fact that certain phases or steps may consist of many more actions and hence naturally require more time. In an example study, *Zhang* et al. [57] reported varying phase duration's ranging from one minute to four hours in Laparoscopic Sacrocolpopexy (LS). The execution time of each action can also differ depending on the operating personnel and condition of the patient, or optional phases can even be entirely omitted under certain conditions [48], [54].

For cataract surgeries, *Primus* et al. [26] undersampled large phases and augmented minor phases to have an equal number of samples in all phases. In a similar sense, *Zhang* et al. [44] used Synthetic Minority Oversampling Technique (SMOTE) [66] in Sleeve Gastrectomy (SG) operations. However, these resampling techniques do not consider temporal information and may even degrade the performance of the model in some cases [67]. Modified loss functions, such as class-weighted cross-entropy or focal loss are proposed to mitigate this problem and shown to be effective [44], [48].

3) Input Noise: Noisy inputs distort short-time temporal information and make the classification tasks more challenging [27], [33], [47]. In endoscopic video data of LC, these frames occur when the camera lens is covered with blood or moisture, moved out of the body for cleaning, not correctly focused on the operating area, or created blurry images. In [32] and [40], authors initially detected these frames and classified them separately without disturbing the temporal information flow of other frames. Similarly, an out-of-topic conversation with the medical staff can create noise with speech modality in a similar manner [65]. Filtering these conversations before the workflow analysis task could be beneficial.

4) Modality: Different modalities present particular challenges and require specialized techniques for feature extraction and model design. Video is the most commonly used modality in our review, especially endoscopic video in LC and microscopic video in cataract surgery. Limited inter-phase and high intra-phase variance, noisy input frames, and subtle details being of key importance are reported as the main challenges. Tool usage information is another data source that indicates which tools are being used and when. Obtaining these data is often done by manual labeling, thus it is not possible to use in real-world applications. Solutions to this problem are offered by modifying surgical instruments physically, e.g., adding tool-tracking devices [68] or RFID tags [69]. Therefore, this modality is not widely used in the latest studies. *Sahu* et al. [36] initially recognized tools in videos and estimated surgical phases with this information. Besides visual information acquired from videos another important aspect is the audio data. In the speech modality, using multiple expressions for similar meanings, different languages, off-topic conversations, and background noise are reported as possible challenges [45].

5) Online/Offline Setting: An online algorithm refers to a method that can perform the SWA task as the new data arrives sequentially, hence the algorithm utilizes only current and past data. That translates to real-time decision-making capability, in contrast to offline algorithms which require access to data from the entire operation. In the online setting, algorithms must optimize efficiency and reliability at the same time while utilizing limited hardware capacity [56]. Conversely, offline algorithms have the advantage of accessing the entire data set at once, enabling them to produce more accurate results [51]. These properties are utilized in *Yu* et al. [25] in which a teacher network operates offline and is trained on a small set to predict the annotations for all training data. The training data is then used in the student network operating online.

Online recognition algorithms can help physicians to reduce surgical errors by following surgical procedure steps or administration to plan OR schedule more effectively. Offline recognition algorithms can be used for educational purposes in the analysis and assessment of the surgery. Table I shows the online/offline settings of all selected studies.

6) Evaluation: Correct evaluation of models and comparing them poses a great challenge. Commonly, frame-wise performance analysis of algorithms is done with the following metrics: Precision (PR), Recall (RE), Accuracy (AC), F1 score, and Jaccard index (J). Using the phase-wise set of Ground Truth (GT) sample-label pairs and set of Prediction (P) sample-label pairs, these metrics are computed as:

$$PR = \frac{|GT \cap P|}{|P|}, \quad RE = \frac{|GT \cap P|}{|GT|}, \quad J = \frac{|GT \cap P|}{|GT \cup P|}.$$

The AC shows the percentage of frames correctly classified in the ground truth labels. The F1 score is the harmonic mean of the precision and recall. Except for AC, phase-wise classification results are then *macro-averaged* over an operation, and average

TABLE I
SUMMARY OF OUR INCLUDED STUDIES GROUPED BY YEAR

Method	Algorithm	Online/Offline	Modality	Dataset	Procedure	Classes	Code Available
2018							
Chen et al. [22]	RNN	-	Video	M2CAI16	LC	8	No
Chen et al. [23]	RNN	Online	Video	Cholec80	LC	7	No
Funke et al. [24]	RNN	Online	Video	Cholec80	LC	7	Yes
Yu et al. [25]	RNN	Both	Video	Cholec120	LC	7	Yes
Primus et al. [26]	CNN	Online	Video	21 Recordings	Cataract	11	No
Jin et al. [27]	RNN	Online	Video	M2CAI16, Cholec80	LC	7,8	Yes
Zisimopoulos et al. [28]	RNN	Offline	Video	CATARACTS	Cataract	14	No
Zia et al. [29]	CNN	Both	Video& Sensor	100 Recordings	RAPD	12	No
2019							
Qi et al. [30]	CNN	Online	Video	cataract-101	Cataract	10	No
Bodenstedt et al. [31]	RNN	-	Video	Cholec80	LC	7	No
Yi & Jiang [32]	RNN	Online	Video	M2CAI16, Cholec80	LC	7,8	Yes
2020							
Ding et al. [33]	Non-Local	Online	Video	M2CAI16	LC	8	No
Shi et al. [34]	RNN+ Non Local	Online	Video	Cholec80	LC	7	Yes
Jin et al. [35]	RNN	Online	Video	Cholec80	LC	7	Yes
Sahu et al. [36]	RNN	Online	Video	Cholec80	LC	7	No
Czempiel et al. [37]	CNN	Online	Video	Cholec80, 51 Recordings	LC	7	Yes
Kitaguchi et al. [38]	CNN	Online	Video	71 Recordings	LS	11	No
Nwoye et al. [39]	CNN	-	Video	CholecT40	LC	6, 8, 19	No
2021							
Li et al. [40]	RNN	Online	Video	M2CAI16	LC	8	No
Ban et al. [41]	RNN	Both	Video	Cholec80, 100 Recordings	LC	7, 13	No
Jin et al. [42]	Non-local	Online	Video	M2CAI16, Cholec80	LC	7, 8	Yes
Pradeep & Sinha [43]	CNN	Online	Video	Cholec80	LC	7	Yes
Zhang et al. [44]	CNN	Both	Video	461 Recordings	SG	8	No
Guzmán-García et al. [45]		Offline	Speech	15 Online Videos	LC	7	No
Xia & Jia [46]	RNN	Both	Video	cataract-101	Cataract	10	No
Shi et al. [47]	RNN+ Non Local	Online	Video	M2CAI16, Cholec80	LC	7, 8	No
Ramesh et al. [48]	TCN	Online	Video	40 Recordings	GB	11	Yes
Gao et al. [49]	Transformer	Online	Video	M2CAI16, Cholec80	LC	7, 8	Yes
Czempiel et al. [50]	Transformer	Online	Video	Cholec80, 85 Recordings	LC	7,8	Yes
Zhang et al. [51]	TCN	Both	Video	461 Recordings	SG	8	No
Paysan et al. [52]	HMM	-	Video& Sensor	38 Recordings	HM	2	Upon Request
Kadkhodamohammadi et al. [53]	RNN	Offline	Ambient Video	18 Recordings	TKR	18	No
Zhang et al. [54]	Transformer	Offline	Video	337 Recordings	GB	11	No
Ward et al. [55]	RNN	Online	Video	50 Recordings	PEM	5	No
2022							
Shi et al. [56]	RNN+ Non Local	Online	Video	Cholec80	LC	7	No
Zhang et al. [57]	RNN+ Transformer	-	Video	Cholec80, 14 Recordings	LC, LSp	7, 6	Yes
Ding & Li [58]	TCN+ Transformer	Online	Video	M2CAI16, Cholec80	LC	7, 8	Yes
Ban et al. [59]	RNN	-	Video	Cholec80, 200 recordings	LC	7, 12	No
Kadkhodamohammadi et al. [60]	GNN	Online	Video	Cholec80	LC	7	No
Nwoye et al. [61]	Transformer	Online	Video	CholecT50	LC	6,8,19	Yes
Valderrama et al. [62]	Transformer	-	Video	PSI-AVA	RP	11	Yes
Ding et al. [63]	TCN, CNN	-	Video	M2CAI16, Cholec80	LC	7, 8	Yes
Zhang et al. [64]	RNN	Offline	Video	Cholec80, 38 Recordings	LC	7	Yes
Seibold et al. [65]	CNN	-	Speech	THADataset	THA	6	Yes

For each study, the main temporal model is given in the algorithm column. Online/offline setting shows real-time working capability. If a publicly available dataset is used, the name of the dataset, otherwise, the number of recordings is given. Type of surgery is denoted by abbreviations: RAPD robot-assisted radical prostatectomy; C cataract; LC laparoscopic cholecystectomy; LS laparoscopic sigmoidectomy; SG sleeve gastrectomy, GB laparoscopic gastric bypass, HM hysteroscopic myomectomy; TKR total knee replacement; PEM peroral endoscopic myotomy; LSP laparoscopic sacrocolpopexy; RP radical prostate tomy; THA total hip arthroplasty. The modality, the number of classes per dataset, and the availability of the source code for each selected study are reported.

results of all operations are commonly reported as the final metrics. Especially in strongly imbalanced datasets, the AC score under-represents the errors in short phases. That can result in a high accuracy even if the model performs poorly on the minority class. Macro-averaging assumes equal importance of all phases. Thus, metrics such as Jaccard and F1 score give better insight into the model performance. The precision assesses the rate of false-positive predictions, indicating phases recognized erroneously. The recall checks for false-negative predictions, evaluating whether parts of a phase are missed.

In addition to frame-wise metrics, *Padoy* et al. [7] proposed two custom metrics to calculate classification errors within phases or completely miss-classified phases. Similarly, *Dergachyova* et al. [70] proposed using three new metrics to track the consistency of the estimations, positive and negative time delays, and *over-segmentation* errors. Among selected studies, *Zhang* et al. [57] leveraged an event-based *Ward* [71] metric for the evaluation of their model in a highly imbalanced dataset. Moreover, *Ban* et al. [59] used *Levenshtein distance* [72] and *Zhang* et al. [54] used *segmental edit distance* [73] and *segmental F1 score* [74]. Confusion matrices and color ribbons showing ground truth and predictions are also used widely. Although these tools can not provide numerical conclusions, they can give insights into the performance of the proposed models.

B. Data-Based

1) Dataset Size: There is a great effort in the research community to create large annotated datasets for surgical workflow analysis. However, limited available surgical data is still the key limitation for designing robust systems. Publicly available datasets for surgical workflow analysis and their properties are summarized in Section V. According to *Maier-Hein* et al. [75], regularity constraints, incompatibility of different data sources, insufficient and unstructured data storage, and hardware limitations are the current main problems for establishing large surgical data sources. In response to legal constraints for patient data protection, *Kadkhodamohammadi* et al. [53] anonymized ambient videos via face detection and subsequent masking algorithm. Anonymization algorithms can help protect patient privacy and ease the establishment of new public datasets. Moreover, synthetic data generation can also be an alternative for creating large annotated data automatically [76], [77]. However, there is a sim-to-real gap between synthetic and real-world data [78]. Thus synthetic data can not alone solve the issue of limited dataset size.

2) Operation Type: In SWA, each operation type might necessitate the design of different methods considering the unique characteristics of the procedure to solve the recognition task. Even though a particular method may be effective for one type of operation, it may not be suitable for use in another. While limited inter-phase and high intra-phase variance problems are reported in LC [27], [42], [58], the strong similarity of frames needs to be considered in cataract surgeries [46]. In the SG procedure, an extreme class-imbalance problem is reported [44]. In open surgeries, the usage of head-mounted cameras and the necessity of privacy-preserving methods are mentioned as challenges [53]. An inspiring work on this topic is presented by *Neimark* et al. [79]. The authors performed step recognition tasks on four different laparoscopic surgeries.

3) Annotation: In our study, the majority of the selected studies focus on supervised learning. However, creating a large annotated dataset for fully supervised learning is a laborious and costly task. Collaborating medical experts define surgical phases or steps of interested operation type. Following annotation work in an entire dataset is either directly done or validated by medical experts. However, these definitions or annotations may not be universally accepted and can vary between experts [48]. When multiple experts are contributing to the study, a strong correlation of their annotation or validation work is required to ensure a minimum degree of ambiguity and label noise [55]. An example of the potential differences in definitions of the same procedure type can be seen between M2CAI16 [80], [81] and the Choec80 datasets. M2CAI16 has an additional *Placement of Trocars* phase which is included within the *Preparation* phase of Cholec80. To mitigate the dependency on annotations and lessen the costly process, unsupervised, semi-supervised, self-supervised learning, and active learning methods are proposed. These methods are summarized with selected studies in Section IV-C.

IV. Step and Phase Recognition Approaches

In this section, we examine selected studies for their feature extraction and temporal relation modeling methods, see Fig. 3. Additionally, annotated data for workflow analysis is limited. In the last section, we review strategies developed to address the challenge of limited data.

A. Feature Extraction

In early studies, Convolutional Neural Networks (CNNs) were shown to be more effective for feature extraction than state-of-the-art statistical methods. In surgical workflow recognition, *Twinanda* et al. [80] showed that using features extracted via a CNN-based architecture yields significantly better results than using handcrafted features. In our review, all of the selected studies using videos preferred CNN-based architectures for feature extraction. In the speech domain, *Seibold* et al. [65] used log-spectrograms and *Guzmán-García* et al. [45] used the Word2Vec [82] model for creating features from the transcription of speech uttered during surgery.

To improve the shallow architecture used in [80], *Jin* et al. [27] compared a 22-layer GoogLeNet [83], and a ResNet [84] model with 35, 50 and 101 layers respectively. *Jin* et al. achieved the best results with ResNet-101 and showed that the depth of the CNN module has a positive impact on its performance. However, using very deep networks increases the risk of overfitting when training data is not abundant. Thus, they used ResNet50 in their study. *Czempiel* et al. [37] compared AlexNet [85] with ResNet50 and reported an increase in accuracy up to 8% by using the ResNet architecture. ResNet pre-trained on ImageNet [86] is used extensively as a backbone architecture, i.e. in 20 of the reviewed studies. *Jin* et al. [42] compared their model with ResNet and ResNeSt [87] backbones and reported improved results with ResNeSt. *Zhang* et al. [54] made an ablation study for feature extraction with a CNN-based network $R(2+1)D$ [88], a BERT [89] based transformer network and a hybrid model with their combinations. They reported very close results with their CNN-based network and hybrid model, and worse results with only transformer-based architecture.

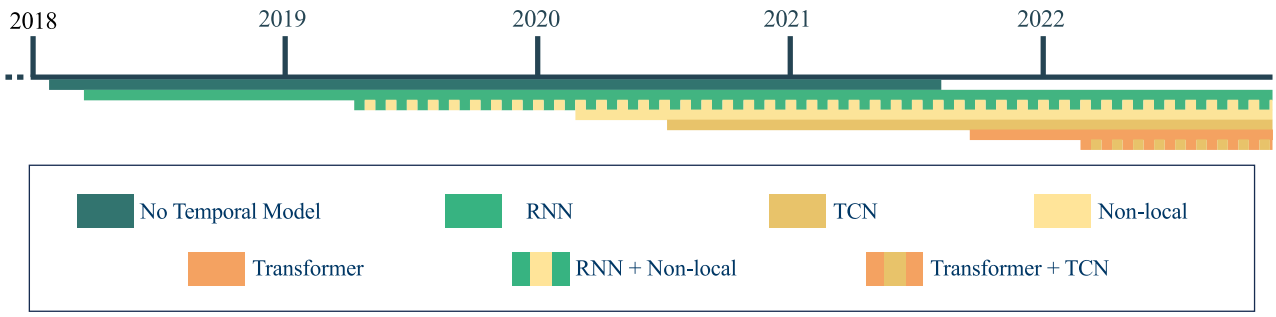

Fig. 3. Temporal relations are crucial to analyze surgical workflows. While a few studies rely on no temporal connections, RNNs have been clearly the dominant algorithm in the beginning. Lately, non-local networks, TCNs, or transformer-based approaches are used more frequently.

Adding auxiliary features to boost the performance of neural networks is considered by several studies. *Qi* et al. [30] added manually extracted edge information from original images to ResNet features and reported improved accuracy in phase recognition. However, the reason why edge information is not already captured by ResNet is not discussed in this study. Additionally, tools could be informative complementary sources for workflow analysis, as they can be used for recognizing specific phases or steps [17]. *Zisimopoulos* et al. [28] used ResNet to estimate binary tool usage information and added this information to classification layers. Moreover, using multi-task learning with closely correlated tasks can improve feature extraction. *Jin* et al. [35] showed consistent improvement in both phase and tool recognition tasks with multi-task learning. In addition to recognizing surgical phases and tools, recognizing surgical actions at different granularity levels is another technique for multi-task learning. *Ramesh* et al. [48] reported improved results when the network trained simultaneously for recognizing phases and steps.

B. Temporal Relation Modeling

1) Frame-Wise Models: A naïve approach for surgical workflow analysis is performing frame-wise classification directly based on extracted features from a single image. Even for experts, this is almost an impossible task, and the implicit fact of ignoring temporal relations causes a significant drop in performance [27]. *Zhang* et al. [44] improved this approach using a 3D CNN to extract features. This includes temporal information from the set of past frames. *Pradeep and Sinha* [43] extracted spatio-temporal features from 64 images together and used them for classification directly. In this way, they achieved $86.07 \pm 0.04\%$ accuracy on Cholec80 with a significantly smaller-sized network than standard networks, i.e. with $4.7\,M$ parameters. It shows that this approach may be useful for hardware-limited applications. However, this aspect is not considered a priority in other studies.

2) Recurrent Neural Networks: Recurrent Neural Networks (RNNs) have been employed successfully with sequential data in many applications. The RNN can carry the memory state h_{t-1} from previous states to the computation of the current prediction y_t, t indicating the time step. Commonly, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are used instead of the general RNN model. The key idea in this model is forgetting non-useful information and memorizing new information in separate steps. In the LSTM, for the input signal x and learnable weights, W_f, W_c, W_i, W_o, the first step is computing the vector f in the *forget gate* to control which part of the information to de-emphasize:

$$f_t = \sigma(W_f[h_{t-1}, x_t])$$

where σ is an activation function, and $[\cdot]$ is the concatenation operation. In the next step, new information to be added is memorized in *input gate* using input data and hidden state:

$$i_t = \sigma(W_i[h_{t-1}, x_t])$$

$$\tilde{C}_t = tanh(W_C[h_{t-1}, x_t]).$$

The cell state C is updated afterward with candidate vectors:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

using the element-wise multiplication operator $\odot$. Finally, the current hidden state h_t and the output vector y_t are computed with the following equations:

$$h_t = \sigma(W_o[h_{t-1}, x_t]) \odot tanh(C_t)$$

$$y_t = \sigma(h_t).$$

The GRU operates following the same logic, differing only in the fact that the memory operates directly via a hidden state and does not use an additional cell state. This recursive characteristic makes RNNs an appropriate candidate for analyzing sequential surgical workflow. *Jin* et al. [27] made an ablation study comparing the ResNet model with and without additional LSTM layers for phase recognition and reported improved accuracy. 21 studies in our review used RNN as a part of their architectures, including CNN-RNN models and using solely RNNs for the temporal modeling. *Zisimopoulos* et al. [28] compared LSTM and GRU [90] architectures and reported that the LSTM-based version of their model showed better performance. Bi-directional RNNs (BRNNs) can also further enhance performance by processing the data first in the forward direction and in the backward direction afterward. *Zia* et al. [29] achieved the best results in their study with single-layer bi-directional RNNs. A limitation of the BRNN is that it can only be used in an offline application.

RNNs have drawbacks as they are slow to train with large datasets and have a limited receptive field in practice [91]. That is similarly reported in several studies in our review [37], [41]. The duration of a surgery can be multiple hours and the short receptive field of LSTMs can be a limiting factor [37].

3) Temporal Convolution Networks: Temporal Convolutional Network (TCN) [73] and multi-stage TCNs (MS-TCN) [92] are proposed to hierarchically capture long-range

spatio-temporal relationships from sequential input data. TCNs aim to capture low-level and high-level features via a stack of multiple dilated convolutional layers, which has a larger filter derived by dilating the original convolution filter with zeros. This helps the convolution layer to operate on a coarser level. A stage is constructed by stacking multiple dilated convolutional layers with increasing dilation factors. For online algorithms, *casual* dilated convolution layers are used in stages, which ensures that kernel computations are made only using previous data points.

Five studies in our review used TCN. *Czempiel* et al. [37] replaced RNN layers of a CNN-LSTM model with a multi-stage TCN for the first time and reported an increase of 2% accuracy. They compared their model with one, two, and three stages and observed improvements from the one-stage model to the two-stage model, but decreases in performance in the three-stage model. The performance drop could be explained by overfitting. *Ramesh* et al. [48] compared one and two-stage TCN and LSTM with the same ResNet backbone and confirmed the best results with two-stage TCN. *Zhang* et al. [51] used four-stage non-causal TCN, however, the effect of this choice is not additionally experimented with. Similarly to other approaches, Ding et al. [63] used ResNet50 and an MS-TCN in a contrastive learning setting.

4) Non-Local Networks: A non-local block aims to compute the response of the input signal at position *j* to the input signal in all other positions in time or space. This way, interactions between any two positions can be represented [93]. In that sense, the non-local block can be seen as a more general version of self-attention, encompassing broader space and space-time relationships. The non-local operation is constructed as follows:

$$\mathbf{y}_j = \frac{1}{\mathcal{C}(x)} \sum_{\forall k} \mathcal{F}(x_j, x_k)\mathcal{G}(x_k), \tag{1}$$

where j is the position index, k enumerates all other indexes, x is the input signal, $\mathcal{F}$ is a pairwise function outputting a scalar for representing relationship, $\mathcal{G}$ computes the representation of the input signal, and $\mathcal{C}$ is the normalization factor. A non-local block can be plugged into existing architectures to learn the relationship between frames in space. In our review, five studies used non-local blocks. *Shi* et al. [56] compared their network with and without non-local block on top and reported an increase in accuracy with non-local block. *Jin* et al. [42] used this scheme to incorporate the current feature vector and memory vector. They examined the effectiveness of a non-local block by comparing it with a weighted average operation in surgical workflow analysis and reported better results with the integrated non-local block. In [34], *Shi* et al. used non-local blocks to learn dependencies between frames and used this dependency information to find the most informative frames in an active learning setting. *Ding* et al. [33] extracted relationships between all pixels and frames simultaneously by using non-local blocks embedded in 3D CNNs. Computing these relationships, however, significantly increases the computational load of the network.

5) Transformers: Transformers aim to efficiently model temporal relations and extract global features using the self-attention mechanism [94]. The Transformer uses stacked self-attention and point-wise, fully connected layers in the encoder-decoder structure. The self-attention layer input consists of a *Query* matrix Q, a *Key* matrix K with dimension d_k, and a *Value* matrix V. It is represented as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{2}$$

The attention mechanism calculates the similarity between queries and keys via dot product operation, resulting in a similarity score. The dot product is scaled by the square root of the dimension of the key vectors to have more stable gradients. The result is utilized to assign weights to the corresponding values. The weighted values are combined through summation, ultimately generating the output of the attention mechanism.

In our review, seven studies used Transformers. Similar to common CNN-LSTM models, *Czempiel* et al. [50] used ResNet50 for frame-wise feature extraction but they preferred an 11-layer transformer network for temporal relation modeling. They compared a CNN-LSTM model, a CNN-TCN model [37], and their proposed CNN-Transformer architectures using two different datasets. They reported best results with CNN-Transformer model achieving 1–2% increase in accuracy compared to CNN-TCN model and 4–6% increase compared to CNN-LSTM model. *Gao* et al. [49] presented a TCN-Transformer network that uses TCN layers to extract temporal features from spatial features and fuses them with Transformer layers. *Zhang* et al. [54] compared temporal modeling methods with an MS-TCN model, MS-Longformer [95], and a hybrid model with their combination. They achieved the best results with their hybrid model. *Ding and Li* [58] used Transformers to learn the relationship between segments in different temporal resolutions. *Valderrama* et al. [62] leverage Transformers to detect actions, phases, instruments, and steps. With their architecture, improved results on the PSI-AVA dataset can be achieved compared to CNN-based methods.

Applications using Transformers are frequently trained on vast amounts of data, which is not available for tasks analyzing surgical workflows. When trained in the absence of a sufficient amount of data and proper regularization, Transformers do not generalize well unlike CNNs due to the lack of *inductive bias*. Another drawback is that Transformers require substantial power for processing, and GPU memory increases quadratically with the input size. Utilizing Transformers can be particularly difficult given these factors.

6) Other: Kadkhodamohammadi et al. [60] used Graph Neural Networks (GNNs) to incorporate temporal information. They represented each feature vector extracted from a single image with a node in the graph and processed the whole video through several layers by aggregating information from neighboring frames. *Jin* et al. [42] used feature vectors extracted from short video clips to create a memory bank, which stores information about the distant past. In their memory bank, they can access the last 30 stored feature vectors. Later, they use the past temporal information from the memory bank during the current surgical phase estimation. *Ban* et al. [41] used hidden states of the LSTM layer to create a memory. Then, they leveraged this memory via statistical models and added it to the current phase estimation. Hierarchical learning or multi-step learning methods are also considered in several studies. *Ji and Jiang* [32] designed a framework with the goal of classifying frames that are hard to detect first, in order to then process these frames separately for the final estimation. *Ding and Li* [58] used segment-level features for estimating the current phase of a frame. In addition, *Guzmán-García* et al. [45] extracted word vectors and compared

HMMs together with SVM, random forest, logistic regression, and shallow linear layers for classification. *Paysan* et al. [52] used *hidden semi-Markov model* for modeling temporal relations. Zhang et al. [64] present an offline approach using a multi-agent network to predict the transition of phases. The network contains LSTM and a Deep Q-Learning Network (DQN) which segments a single phase. Different phases are merged via a Gaussian composition operator.

C. Learning Strategies

Assessing different learning strategies in deep learning is quite common. The most frequently used strategy in surgical workflow analysis is supervised learning. However, the field of supervised methods evolved to address the challenges in complex and cost-intensive data annotation processes. While purely supervised methods often need a huge amount of data to generalize, semi-supervised learning uses only a small amount of annotated data. Annotating OR videos requires expert knowledge, as the phases are even hard to distinguish by individuals [96]. Therefore, the strategy of semi-supervised learning [22] is of interest for such complex tasks. *Chen* et al. [22] uses semi-supervised learning to combine temporal and spatial information. Considering the learning strategies, classical supervised methods, they for example use a CNN for feature extraction followed by temporal modeling [28]. Semi-supervised approaches require more individual steps. *Chen* et al. [22] use a spatial CNN based on unsupervised generative adversarial learning followed by a connection between low-level surgical video features and high-level surgical workflow semantics. In their last step, they use semi-supervised learning to integrate the spatial and the temporal models for finetuning, respectively the CNN and LSTM.

For missing data annotations, another strategy is self-supervised learning. Self-supervised learning can be categorized as a method between supervised and unsupervised learning. In the first place, pseudo labels are used to initialize the network weights. Afterward, either a supervised or unsupervised method is applied to fit the network toward the final task. *Yengera* et al. [97] used readily available time-stamps of endoscopic videos to train their network for remaining surgery duration estimation before fully performing surgical phase recognition. For surgical workflow analysis among selected studies, *Funke* et al. [24] proposes several self-supervised pre-training strategies using temporal coherence. Similar to many supervised approaches [23], [32], [38], [41], [48], *Funke* et al. [24] rely on a ResNet50 backbone. The backbone is pre-trained using a contrastive loss and a ranking loss. For supervised fine-tuning, the CNN is extended using LSTM. Contrastive learning is a sub-strategy of self-supervised learning. A contrastive learning algorithm tries to map features from similar data instances closely while mapping features from distinct data instances into separated points in the feature space. The model learns which points in the input are similar to one another and which ones differ. *Xia* et al. [46] introduced a contrastive branch in their CNN to learn spatio-temporal features. They used this technique to to handle the limited inter-phase and high intra-phase variance. Ding et al. [63] use contrastive learning to transfer knowledge in a teacher-student manner from publicly available datasets to the surgical domain.

Moreover, a weakly-supervised learning strategy can be suitable for a huge amount of data where only a subset is annotated or when the data contains noisy labels. *Nwoye* et al. [39] use weak supervision to support the action recognition. Zhang et al. [64] utilized the method of reinforcement learning for surgical phase recognition. While previous methods predict frame-wise and potentially add the time axis via e.g. RNN. The approach of Zhang et al. predicts the start and end frame of every phase.

Furthermore, active learning evolved which also aims to tackle the cost-intensive data annotation processes. In active learning, a random subset of data is selected and a teacher, mostly a human is asked to annotate this subset which is then used to train the network. Afterward, the trained network is used to annotate the rest of the data. This process is repeated and can be seen as an iterative supervised approach. For surgical workflow analysis, *Bodenstedt* et al. [31] and *Shi* et al. [34] apply active learning to find important data points in videos of the surgical workflow.

Besides the challenge of missing annotations or only a small amount of data, the models used in the deep learning domain are growing rapidly and often contain a high computational cost. The teacher-student strategy applies two networks where the student network contains less computational cost than the teacher network. The student learns the intermediate feature maps of the teacher and aims for convergence with a simpler architecture. *Yu* et al. [25] approach this for surgical workflow analysis and further address the annotation problem. Their teacher network operates offline and is trained on a small set of data to predict the annotations for all training data. This data is then used to train the student network operates online.

V. Datasets

In this section, publicly available datasets for surgical workflow analysis and their properties are summarized, see Table II.

M2CAI16: This dataset contains two separate sub-datasets [80], [81]. The first dataset is *m2cai16-workflow*, contains endoscopic videos from 41 procedures for phase recognition. In this challenge 27 videos are separated for training and remaining videos are used for testing. Videos are recorded at 25 fps with 1920×1080 resolution. Operations in these videos are annotated with eight phases.

Cholec80 & CholecT50: The dataset contains endoscopic videos from 80 laparoscopic cholecystectomies, performed by 13 surgeons at the University Hospital of Strasbourg, France [80]. Videos are recorded at 25 fps with 1920×1080 resolution. The dataset contains annotations for phases and tool usage. Surgeries are decomposed into seven phases by a senior surgeon in the same hospital.

HeiChole: The dataset contains 33 videos with seven surgical phases similar to Cholec80 [98]. In addition to previous datasets, the cameras and fps vary. HeiChole is recorded in three different Hospitals in Germany. Among all, 15 videos captured at Heidelberg University Hospital have a resolution of 960×540 and 25 fps. Other 15 videos were recorded at Salem Hospital and the remaining three videos at GRN-hospital Sinsheim. These are recorded with a resolution of 1920×1080 and 50 fps and three videos at Salem Hospital are recorded with a resolution of 720×576 and $25 fps$. The dataset is annotated by medical experts. It contains annotations of seven surgical phases, four actions, and 21 tools.

HeiCo: The dataset contains ten laparoscopic videos from each proctocolectomy, rectal resection, and sigmoid resection procedure [99]. All 30 operations are recorded at Heidelberg University Hospital and annotated for surgical phase recognition, binary and multi-instance segmentation tasks. Video frames are originally recorded at 1920×1080 resolution and

TABLE II
SUMMARY OF PUBLICLY AVAILABLE DATASETS FOR SURGICAL WORKFLOW ANALYSIS

Dataset	Procedure	Annotation	Modality	Details
M2CAI16 [80], [81]	Cholecystectomy	8 Phases	Video	25 fps, 1920 x 1080 resolution, 41 recordings
Cholec80 [80]	Cholecystectomy	7 Phases, 7 Tools	Video	25 fps with 1920 x 1080 resolution, 80 Recordings
CholecT50 [61]	Cholecystectomy	6 Tools, 10 Verbs, 15 Target	Video	25 fps with 1920 x 1080 resolution, 50 Recordings
HeiChole [98]	Cholecystectomy	7 Phases, 4 Actions, 21 Tools	Video	25, 50, 25 fps with (960 x 540), (1920 x 1080), (720 x 576) resolutions, 33 Recordings
HeiCo [99]	Proctocolectomy, Rectal Resection, Sigmoid Resection	14 Phases, Tool Segmentation Masks	Video & Sensor	(960 x 540) resolution, 14 sensor data, 30 recordings
Cataract-101 [100]	Cataract	10 Phases	Video	$25fps$ with 720 x 540 resolution, 101 Recordings
CATARACTS [28], [101]	Cataract	14 Phases	Video	$30fps$ with $1920x1080$ resolution, 50 Recordings
PSI-AVA [62]	Prostatectomy	11 Phases, 21 Steps 7 Tools, 16 Actions	Video	$1280x800$ resolution, 8 recordings
THADataset [65]	Total Hip Arthroplasty	6 Phases	Speech	$44.1kHz$ sampling rate, 568 recordings with durations $1-30s$

downsampled to 960 $\times$ 540. For the anonymization of videos, frames captured outside of the operated body area are manually replaced with blue frames. In addition to video frames, 14 sensor data streams, annotations for surgical phases, and segmentation masks for tool segmentation tasks are given. All operation types are covered with total 14 surgical phases.

Cataract-101: The dataset contains 101 microscope videos from cataract surgeries performed by four surgeons with two different experience levels in Klinikum Klagenfurt, Austria [100]. Videos are recorded $25fps$ and with 720 $\times$ 540 resolution. Average duration of a surgery is 8.3 minutes, maximum and minimum durations are 17 and 4 minutes respectively. Surgeries are defined by ten phases.

CATARACTS: The dataset contains 50 cataract surgery videos from Brest University Hospital, France [101]. For each surgery microscope and surgical tray videos together with surgical action and 21 binary tool annotations are available. Tools are considered in use when they touch the eyeball. Phase annotations are not provided in this dataset, authors in [28] created annotations with medical experts using the 14 phases.

PSI-AVA: The dataset contains eight radical prostatectomy operations performed in Fundación Santafé de Bogotá University Hospital, Colombia [62]. Annotations are structured for high-level analysis in phase and step recognition levels and for low-level analysis in instrument detection and action recognition levels. Total duration of dataset is 20.45 hours and the operation is defined with 11 phases and 21 steps.

THADataset: The dataset contains recordings from five total hip arthroplasty operations performed in Balgrist University Hosiptal, Switzerland [65]. During the data collection an airborne shutgun microphone is used. The recordings are then manually cut to have recordings without overlapping classes and background talks. Resulting 568 recordings are then labelled with six classes.

VI. DISCUSSION

In this section, the progress in the contemporary SWA field is discussed under model and data based topics.

A. Model-Based

Like other computer vision approaches, CNNs are the prominent methods for feature extraction. They put significant representation capability to use and became the standard choice in video-based models. The importance of aggregating long-time temporal information for better recognition performance is shown to be vital [23], [27], [33], [41], [42] and considered in almost all selected studies. The intuitive CNN-RNN models are the most frequently used method to leverage temporal relations in our review and these models achieved decent results. Their usage is limited by the fact that RNNs are non-parallelizable and can not preserve memory from a distant time in practice. This problem is approached by using CNN-based 3D-CNNs or TCN models, attention-based non-local networks or Transformers, memory networks, or hierarchical segmentation designs. In similar computer vision fields such as temporal action segmentation, temporal action detection, or sequence segmentation, these models are successfully utilized [102]. However, their effectiveness in the SWA can be affected by the limited amount of available data. Although impressive results are observed with these architectures as in Table III, difficulties in the comparison of different proposed models make it challenging to draw concrete conclusions.

Table III presents studies utilizing the most common publicly available dataset Cholec80. Comparing studies on this dataset shows that the respective training/testing strategies vary. Training and test data split, online/offline settings, and post-processing algorithms have an effect on the overall performance. In Table III, a *40:40* data split is most commonly used following previous work [80]. Moreover, if a cross-validation method is used, averaging is generally performed over average results of each training run, thus, standard deviation shows consistency of the model with different data splits rather than standard deviation over operations. Also, not all studies in Table III reported standard deviations in their results. During the evaluation, used metrics, their calculation steps, and reasons for their choices should be explicitly reported. Finally, relatively high standard deviation results in several studies indicate that proposed models perform differently among some operations in the test set. This could be a result of overfitting.

Difficulties of the class-imbalance problem arise due to the unequal phase distribution. Recognizing longer phases is easier for learning algorithms due to the large number of data samples and may lead to higher recognition accuracy easily. Similarly, the misclassification of short phases may have little effect on the overall performance. Because all phases of an operation are vital, such systems would not be clinically preferred. Moreover, the unequal distribution of phases might obstruct the learning capabilities of the model. The misclassification of major phases would contribute to the loss function significantly, and the

TABLE III
PHASE RECOGNITION RESULTS OF DIFFERENT STUDIES UNDER SUPERVISED LEARNING REGIME ON CHOLEC80 DATASET

Year	Study	Feature Extraction	Temporal Model	Cross-Validation	Online/Offline	Precision (%)	Recall (%)	Accuracy (%)
Datasplit: 40 Training, 40 Testing								
2018	*Jin et al.* [27]	ResNet50	LSTM	-	Online	80.7±7.0	83.5±7.5	85.3±7.3
2018	*Chen et al.* [23]	C3D [103]	LSTM	4-Fold	Online	81.3	87.7	91.2
2019	*Yi & Jiang* [32]	ResNet50	LSTM	10-Fold	Online	-	-	87.3±5.7
2020	*Jin et al.* [35]	ResNet50	LSTM	-	Online	86.9±4.3	88±6.9	89.2±7.6
2021	*Jin et al.* [42]	ResNet50	LSTM, Memory, Non-Local	-	Online	90.3±3.3	89.5±5	90.1±7.6
2021	*Gao et al.* [49]	ResNet50	Transformer	-	Online	88.8±7.4	90.7±5	90.3±7.1
2021	*Ban et al.* [41]	ResNet	LSTM, Memory	-	Online	87	83	90
2021	*Ban et al.* [41]	ResNet	LSTM, Memory	-	Offline	85.3	82.7	90.8
2022	*Shi et al.* [56]	ResNet50	LSTM, Non-Local	-	Online	87.8	89.5	89.8
2022	*Ding et al.* [58]	ResNet50	TCN, Transformer		Online	90.3±6.4	90±6.4	91.8±8.1
2022	*Kadkhodamohammadi et al.* [60]	SEResNet50 [104]	GNN	-	Online	86.8	84	91.4
Datasplit: 40 Training, 20 Validation, 20 Testing								
2022	*Zhang et al.* [64]	ResNet50	DQN	-	Offline	84.5±5.9	85.1±8.2	90.1±5.7
Datasplit: 40 Training, 8 Validation, 32 Testing								
2020	*Czempiel et al.* [37]	ResNet50	TCN	-	Online	81.6±0.4	85.2±1.1	88.6±0.3
Datasplit: 64 Training, 16 Testing								
2021	*Pradeep & Sinha* [43]	CNN	CNN	-	Online	77.5	72.2	86.1
Datasplit: 48 Training, 12 Validation, 20 Testing								
2021	*Czempiel et al.* [50]	ResNet50	Transformer	5-Fold	Online	82.2±0.7	86.9±0.8	91.3±0.6
2022	*Kadkhodamohammadi et al.* [60]	SEResNet50 [104]	GNN	5-Fold	Online	89.8±0.8	89.1±0.7	93.8±0.4

Results are retrieved from references and post-processing techniques are ignored.

model would mainly optimize for these phases. Depending on the operation type and phase distribution characteristics, the class-imbalance problem should be considered during model design, training, and evaluation processes.

The video modality is the major source of information in SWA encompassing the strong majority of all selected studies. Only two studies used additional sensor data and two studies used the speech modality. Every modality presents unique challenges and opportunities for surgical workflow analysis. Hence, endoscopic videos are captured from inside body cameras, thus, do not provide any information about the operating room [17]. Similarly, microscopic videos provide a view only from a limited aspect of the eyes. These might be limiting factors for recognizing the occurrence of out-of-sequence events. Furthermore, it is not possible to cover all surgeries with these modalities. Therefore, it is necessary to investigate other possible data sources in future research to extend the usability of surgical workflow analysis. The new data modalities can be utilized solely or together with existing sources. Including an additional modality can enhance algorithm performance and robustness.

Common medical imaging methods for surgeries or interventions like ultrasound, X-Ray, or Computed Tomography may provide useful information and these data can be captured similarly to endoscopic videos. Ambient monitoring of the OR with RGB or depth cameras could provide similar or even richer information [17]. Pose estimation of medical personnel, object recognition, or activity recognition algorithms can be employed to extract further information with this modality. Zia et al. [29] use kinematic data generated by the surgeon console during robot-assisted surgery. They report inferior results compared to image-based models. Investigating the combination of two modalities is pointed out as future work. Additionally, devices such as insufflators, irrigation pumps, or light sources can be utilized similar to kinematic data. Speech and audio recordings are the final types of data used for phase recognition among selected studies and are only utilized in [45], [65]. Despite the growing interest in surgical phase recognition, the use of speech and audio in these models has received little attention. Thus, the challenges and opportunities of using speech and audio sources can only be fully understood after more research is conducted. The main reasons for the limited usage of speech and audio can be difficulties in data acquisition and processing. Using Automatic Speech Recognition (ASR) and language understanding systems to extract simple clinical knowledge is already discussed [105]. In the same way, speech and audio data combined with language models can be used for phase recognition and the development of interactive surgical devices in the future.

B. Data-Based

The key obstacle in the SWA field is that publicly available data are limited by annotation, amount, variation, data type, and operation type. Operations in OR are complex procedures, and automatically analyzing these procedures is a difficult task. Therefore, large, diverse, and representative data is necessary for robustly solving this task [2]. Collection of high-quality large annotated surgical datasets with many variables is an expensive process and sharing these data with all researchers in the field is often challenged by local laws, patient data security concerns, or physical limitations. This bottleneck may cause developed models to perform at the desired level within only the same dataset and the same operation type.

Generalization problems stem from inadequate variable representation in datasets and require further investigation. Proposed models are often trained with data obtained from a single or few medical centers. The variable parameters such as used instruments, lightning conditions, complications, style of physicians, or their experience levels could have an effect on the algorithms' performance [26], [37], [57]. *Bar* et al. [106] investigated generalization with their private laparoscopic cholecystectomy dataset. Using 1243 videos from four medical centers, they demonstrated successful phase recognition in a new medical center by fine-tuning their network with 200 videos. *Neimark* et al. [79] used a pretrained network from *Bar* et al. [106] and reported high frame-wise accuracies when their network

trained on 100–200 videos from a new surgery type. The authors have included three other laparoscopic surgeries in their dataset: Right Hemicolectomy, Sleeve Gastrectomy, and Appendectomy.

An ideal intelligent system implemented in an OR should be able to assist physicians, not in a specific procedure but in all possible operation types regularly performed by them. Currently, publicly available datasets represent only the most common operation types in respective medical departments. The majority of selected studies focused on LC and cataract procedures belonging to the most common operation types in general surgery and ophthalmology [28], [107]. New difficulties, challenges, and opportunities will arise when operations in other branches are considered. Even using the same modalities, different surgery types would posses various challenges. Operational variances such as working area size, natural duration of predefined phases, used surgical tools, phase transition order, frequency, or the number of participating medical personnel would require dedicated approaches to achieve desired performance. It is impossible to fully foresee the unique challenges of each procedure. Thus, it is necessary to experiment with various procedure types. Table I shows that distinct operation types are considered more frequently, pointing a trend in this direction. We, therefore, suggest that future studies should expand to different types of surgeries. Moreover, the data collection process will be more expensive and difficult when rare operation procedures are investigated.

Specific to each procedure, clinical knowledge-based statistical models can be designed and fused to existing models to boost performance. That is considered in [27], [35], [41], [44], [51], and considerable improvements in accuracy are reported. For example, *Jin* et al. [27] used the order of phases to calibrate misclassified frames. *Zhang* et al. [44], [51] used order, duration, and incidence of phases. In future applications, it is reasonable to consider operation-specific prior knowledge and medical experience to design more robust architectures. In the design of such domain-specific models, the online/offline setting of the model should be considered.

The annotation work has two discussion points for future work: granularity and ambiguity. By using finer granularity levels, studies aim to have a more detailed understanding of an operation. In *Nwoye* et al. [39], [61], the authors annotated part of the Cholec80 dataset with *surgical action triplet*. The action triplet includes a combination of instrument usage, performed verb, and target anatomy. Similarly, *Valderrama* et al. [62] provided annotations in RAPD surgery for phases, steps, instrument usage, and *atomic actions* which refers to the finest body movements or object manipulation. The second point is the possible ambiguity in phase transitions. The annotation is always prepared or validated by medical experts. However, expecting them to choose the exact same time point as the phase transition is unrealistic and ground truth annotations of multiple medical experts can show a variance [96]. Instant stops, breaks, talks, or behaviors that can occur during neighboring surgical phases contribute to ambiguity. The problem is often addressed by finding a common point with personal discussion or averaging individual annotations. Moreover, varying length relaxation periods between phases are introduced recently with an additional *Not a Phase* label to take ambiguities into account [44]. In this case, relation periods should be reported during the evaluation of the models. That is relevant to a valid comparison of models.

VII. Conclusion

In this article, we presented a comprehensive review of recent research works in surgical workflow analysis, focusing on surgical phases and steps. That is an essential topic for building intelligent context-aware systems for ORs, and there has been an increasing interest in recent years. We analyzed challenges in the field of SWA in two groups: challenges in the design of algorithms and the creation of datasets. During the design of algorithms, modeling temporal relations, facilitating effects of noisy inputs, and achieving desired online/offline settings are considered in most of the studies, and impressive improvements are reported. Class-imbalance problem is focused in several studies, and strong variations in different operation types are observed. The effects of using different modalities are under-investigated and could be an interesting future direction. Details of the evaluation steps should be explicitly given, and comparisons should consider these details. A standardized and comprehensive evaluation method would help compare models and correctly assess advancements. Within the data-based challenges, limited size and variation of the datasets is the current bottleneck in the SWA. Collection and publication of large datasets are restricted by necessary laborious recording processes and legal procedures. Generalization and challenges in different operation types are still open for more research.

References

[1] M. Maktabi and T. Neumuth, "Online time and resource management based on surgical workflow time series analysis," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 2, pp. 325–338, 2017.

[2] L. Maier-Hein et al., "Surgical data science for next-generation interventions," *Nature Biomed. Eng.*, vol. 1, no. 9, pp. 691–696, 2017.

[3] M. Kranzfelder et al., "Reliability of sensor-based real-time workflow recognition in laparoscopic cholecystectomy," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 9, pp. 941–948, 2014.

[4] T. Vercauteren, M. Unberath, N. Padoy, and N. Navab, "CAI4CAI: The rise of contextual artificial intelligence in computer-assisted interventions," in *Proc. IEEE*, vol. 108, no. 1, pp. 198–214, Jan. 2020.

[5] V. Facco Rodrigues, R. da Rosa Righi, C. André da Costa, B. Eskofier, and A. Maier, "On providing multi-level quality of service for operating rooms of the future," *Sensors*, vol. 19, no. 10, 2019, Art. no. 2303.

[6] C. Herfarth, "'lean'surgery through changes in surgical work flow," *J. Brit. Surg.*, vol. 90, no. 5, pp. 513–514, 2003.

[7] N. Padoy, T. Blum, H. Feussner, M.-O. Berger, and N. Navab, "On-line recognition of surgical activity for monitoring in the operating room," in *Proc. AAAI Conf. Artif. Intell.*, 2008, pp. 1718–1724.

[8] A. P. Twinanda, "Vision-based approaches for surgical activity recognition using laparoscopic and RBGD videos," Ph.D. dissertation, ICube - Laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie, Strasbourg, France, 2017.

[9] K. Cleary, H. Y. Chung, and S. K. Mun, "Or 2020: The operating room of the future," *Laparoendoscopic Adv. Surg. Techn.*, vol. 15, no. 5, pp. 495–500, 2005.

[10] D. W. Rattner and A. Park, "Advanced devices for the operating room of the future," *Seminars Laparoscopic Surg.*, vol. 10, no. 2, pp. 85–89, 2003.

[11] O. Weede et al., "Workflow analysis and surgical phase recognition in minimally invasive surgery," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, 2012, pp. 1080–1074.

[12] P. Jannin, M. Raimbault, X. Morandi, and B. Gibaud, "Modeling surgical procedures for multimodal image-guided neurosurgery," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2001, pp. 565–572.

[13] T. Neumuth, G. Strauß, J. Meixensberger, H. U. Lemke, and O. Burgert, "Acquisition of process descriptions from surgical interventions," in *Proc. Int. Conf. Database Expert Syst. Appl.*, 2006, pp. 602–611.

[14] F. Lalys and P. Jannin, "Surgical process modelling: A review," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 9, no. 3, pp. 495–511, 2014.

[15] L. MacKenzie, J. Ibbotson, C. Cao, and A. Lomax, "Hierarchical decomposition of laparoscopic surgery: A human factors approach to investigating the operating room environment," *Minimally Invasive Ther. Allied Technol.*, vol. 10, no. 3, pp. 121–127, 2001.

[16] B. van Amsterdam, M. J. Clarkson, and D. Stoyanov, "Gesture recognition in robotic surgery: A review," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 6, pp. 2021–2035, Jun. 2021.

[17] N. Padoy, "Machine and deep learning for workflow recognition during surgery," *Minimally Invasive Ther. Allied Technol.*, vol. 28, no. 2, pp. 82–90, 2019.

[18] R. S. Antunes et al., "A survey of sensors in healthcare workflow monitoring," *ACM Comput. Surv.*, vol. 51, no. 2, pp. 1–37, 2018.

[19] C. R. Garrow et al., "Machine learning for surgical phase recognition: A systematic review," *Ann. Surg.*, vol. 273, no. 4, pp. 684–693, 2021.

[20] D. Junger, S. Frommer, and O. Burgert, "State-of-the-art of situation recognition systems for intraoperative procedures," *Med. Biol. Eng. Comput.*, vol. 60, no. 4, pp. 921–939, 2022.

[21] A. Liberati et al., "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration," *J. Clin. Epidemiol.*, vol. 62, no. 10, pp. e1–e34, 2009.

[22] Y. Chen, Q. L. Sun, and K. Zhong, "Semi-supervised spatio-temporal CNN for recognition of surgical workflow," *EURASIP J. Image Video Process.*, vol. 2018, no. 1, pp. 1–9, 2018.

[23] W. Chen, J. Feng, J. Lu, and J. Zhou, "Endo3d: Online workflow analysis for endoscopic surgeries based on 3D CNN and LSTM," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Berlin, Germany: Springer, 2018, pp. 97–107.

[24] I. Funke, A. Jenke, S. T. Mees, J. Weitz, S. Speidel, and S. Bodenstedt, "Temporal coherence-based self-supervised learning for laparoscopic workflow analysis," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Berlin, Germany: Springer, 2018, pp. 85–93.

[25] T. Yu, D. Mutter, J. Marescaux, and N. Padoy, "Learning from a tiny dataset of manual annotations: A teacher/student approach for surgical phase recognition," 2018, *arXiv:1812.00033*.

[26] M. J. Primus et al., "Frame-based classification of operation phases in cataract surgery videos," in *Proc. Int. Conf. Multimedia Model.*, 2018, pp. 241–253.

[27] Y. Jin et al., "SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network," *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1114–1126, May 2018.

[28] O. Zisimopoulos et al., "Deepphase: Surgical phase recognition in cataracts videos," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2018, pp. 265–272.

[29] A. Zia, A. Hung, I. Essa, and A. Jarc, "Surgical activity recognition in robot-assisted radical prostatectomy using deep learning," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2018, pp. 273–280.

[30] B. Qi, X. Qin, J. Liu, Y. Xu, and Y. Chen, "A deep architecture for surgical workflow recognition with edge information," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2019, pp. 1358–1364.

[31] S. Bodenstedt et al., "Active learning using deep bayesian networks for surgical workflow analysis," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 6, pp. 1079–1087, 2019.

[32] F. Yi and T. Jiang, "Hard frame detection and online mapping for surgical phase recognition," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2019, pp. 449–457.

[33] Y. Ding et al., "Surgical workflow recognition using two-stream mixed convolution network," in *Proc. IEEE 3rd Int. Conf. Adv. Electron. Mater., Comput. Softw. Eng.*, 2020, pp. 264–269.

[34] X. Shi, Y. Jin, Q. Dou, and P.-A. Heng, "LRTD: Long-range temporal dependency based active learning for surgical workflow recognition," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, no. 9, pp. 1573–1584, 2020.

[35] Y. Jin et al., "Multi-task recurrent convolutional network with correlation loss for surgical video analysis," *Med. Image Anal.*, vol. 59, 2020, Art. no. 101572.

[36] M. Sahu, A. Szengel, A. Mukhopadhyay, and S. Zachow, "Surgical phase recognition by learning phase transitions," *Curr. Directions Biomed. Eng.*, vol. 6, no. 1, 2020, Art. no. 20200037.

[37] T. Czempiel et al., "TeCNO: Surgical phase recognition with multi-stage temporal convolutional networks," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2020, pp. 343–352.

[38] D. Kitaguchi et al., "Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach," *Surg. Endoscopy*, vol. 34, no. 11, pp. 4924–4931, 2020.

[39] C. I. Nwoye et al., "Recognition of instrument-tissue interactions in endoscopic videos via action triplets," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2020, pp. 364–374.

[40] Y. Li, Y. Li, W. He, W. Shi, T. Wang, and Y. Li, "SE-OHFM: A surgical phase recognition network with SE attention module," in *Proc. IEEE Int. Conf. Electron. Inf. Eng. Comput. Sci.*, 2021, pp. 608–611.

[41] Y. Ban et al., "Aggregating long-term context for learning laparoscopic and robot-assisted surgical workflows," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 14531–14538.

[42] Y. Jin, Y. Long, C. Chen, Z. Zhao, Q. Dou, and P.-A. Heng, "Temporal memory relation network for workflow recognition from surgical video," *IEEE Trans. Med. Imag.*, vol. 40, no. 7, pp. 1911–1923, Jul. 2021.

[43] C. S. Pradeep and N. Sinha, "Spatio-temporal features based surgical phase classification using CNNs," in *Proc. IEEE 43rd Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2021, pp. 3332–3335.

[44] B. Zhang, A. Ghanem, A. Simes, H. Choi, and A. Yoo, "Surgical workflow recognition with 3DCNN for sleeve gastrectomy," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, no. 11, pp. 2029–2036, 2021.

[45] 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, 2021, Art. no. 1330.

[46] T. Xia and F. Jia, "Against spatial–temporal discrepancy: Contrastive learning-based network for surgical workflow recognition," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, no. 5, pp. 839–848, 2021.

[47] X. Shi, Y. Jin, Q. Dou, and P.-A. Heng, "Semi-supervised learning with progressive unlabeled data excavation for label-efficient surgical workflow recognition," *Med. Image Anal.*, vol. 73, 2021, Art. no. 102158.

[48] S. Ramesh et al., "Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, no. 7, pp. 1111–1119, 2021.

[49] 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 *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2021, pp. 593–603.

[50] T. Czempiel, M. Paschali, D. Ostler, S. T. Kim, B. Busam, and N. Navab, "Opera: Attention-regularized transformers for surgical phase recognition," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2021, pp. 604–614.

[51] B. Zhang, A. Ghanem, A. Simes, H. Choi, A. Yoo, and A. Min, "Swnet: Surgical workflow recognition with deep convolutional network," in *Proc. Med. Imag. Deep Learn.*, 2021, pp. 855–869.

[52] D. Paysan, L. Haug, M. Bajka, M. Oelhafen, and J. M. Buhmann, "Self-supervised representation learning for surgical activity recognition," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, no. 11, pp. 2037–2044, 2021.

[53] A. Kadkhodamohammadi et al., "Towards video-based surgical workflow understanding in open orthopaedic surgery," *Comput. Methods Biomech. Biomed. Eng.: Imag. Visual.*, vol. 9, no. 3, pp. 286–293, 2021.

[54] B. Zhang et al., "Towards accurate surgical workflow recognition with convolutional networks and transformers," *Comput. Methods Biomech. Biomed. Eng.: Imag. Visual.*, vol. 10, no. 4, pp. 349–356, 2021.

[55] T. M. Ward et al., "Automated operative phase identification in peroral endoscopic myotomy," *Surg. Endoscopy*, vol. 35, no. 7, pp. 4008–4015, 2021.

[56] P. Shi, Z. Zhao, K. Liu, and F. Li, "Attention-based spatial–temporal neural network for accurate phase recognition in minimally invasive surgery: Feasibility and efficiency verification," *J. Comput. Des. Eng.*, vol. 9, no. 2, pp. 406–416, 2022.

[57] Y. Zhang, S. Bano, A.-S. Page, J. Deprest, D. Stoyanov, and F. Vasconcelos, "Large-scale surgical workflow segmentation for laparoscopic sacrocolpopexy," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 17, no. 3, pp. 467–477, 2022.

[58] X. Ding and X. Li, "Exploring segment-level semantics for online phase recognition from surgical videos," *IEEE Trans. Med. Imag.*, vol. 41, no. 11, pp. 3309–3319, Nov. 2022.

[59] Y. Ban et al., "SUPR-GAN: Surgical prediction GAN for event anticipation in laparoscopic and robotic surgery," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 5741–5748, Apr. 2022.

[60] A. Kadkhodamohammadi, I. Luengo, and D. Stoyanov, "PATG: Position-aware temporal graph networks for surgical phase recognition on laparoscopic videos," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 17, no. 5, pp. 849–856, 2022.

[61] C. I. Nwoye et al., "Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos," *Med. Image Anal.*, vol. 78, 2022, Art. no. 102433.

[62] N. Valderrama et al., "Towards holistic surgical scene understanding," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2022, pp. 442–452.

[63] X. Ding, Z. Liu, and X. Li, "Free lunch for surgical video understanding by distilling self-supervisions," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2022, pp. 365–375.

[64] Y. Zhang, S. Bano, A.-S. Page, J. Deprest, D. Stoyanov, and F. Vasconcelos, "Retrieval of surgical phase transitions using reinforcement learning," in *Proc. Int. Conf. Med. Image Comput. Comput-Assist. Interv.*, 2022, pp. 497–506.

[65] M. Seibold, A. Hoch, M. Farshad, N. Navab, and P. Fürnstahl, "Conditional generative data augmentation for clinical audio datasets," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2022, pp. 345–354.

[66] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[67] P. Zhao et al., "T-smote: Temporal-oriented synthetic minority oversampling technique for imbalanced time series classification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 2406–2412.

[68] M. S. Holden et al., "Feasibility of real-time workflow segmentation for tracked needle interventions," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 6, pp. 1720–1728, Jun. 2014.

[69] J. E. Bardram, A. Doryab, R. M. Jensen, P. M. Lange, K. L. Nielsen, and S. T. Petersen, "Phase recognition during surgical procedures using embedded and body-worn sensors," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2011, pp. 45–53.

[70] O. Dergachyova, D. Bouget, A. Huaulmé, X. Morandi, and P. Jannin, "Automatic data-driven real-time segmentation and recognition of surgical workflow," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 11, pp. 1081–1089, 2016.

[71] J. A. Ward, P. Lukowicz, and H. W. Gellersen, "Performance metrics for activity recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 1, pp. 1–23, 2011.

[72] V. I. Levenshtein et al., "Binary codes capable of correcting deletions, insertions, and reversals," *Sov. Phys. Doklady*, vol. 10, no. 8., pp. 707–710, 1966.

[73] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 47–54.

[74] C. Lea, R. Vidal, and G. D. Hager, "Learning convolutional action primitives for fine-grained action recognition," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2016, pp. 1642–1649.

[75] L. Maier-Hein et al., "Surgical data science–from concepts toward clinical translation," *Med. Image Anal.*, vol. 76, 2022, Art. no. 102306.

[76] I. Luengo, E. Flouty, P. Giataganas, P. Wisanuvej, J. Nehme, and D. Stoyanov, "SurReal: Enhancing surgical simulation realism using style transfer," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 1–12.

[77] M. Pfeiffer et al., "Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2019, pp. 119–127.

[78] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 5, pp. 1–46, 2020.

[79] D. Neimark, O. Bar, M. Zohar, G. D. Hager, and D. Asselmann, ""Train one, classify one, teach one"-cross-surgery transfer learning for surgical step recognition," in *Proc. Med. Imag. With Deep Learn.*, 2021, pp. 532–544.

[80] 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 Trans. Med. Imag.*, vol. 36, no. 1, pp. 86–97, Jan. 2017.

[81] R. Stauder, D. Ostler, M. Kranzfelder, S. Koller, H. Feußner, and N. Navab, "The TUM LapChole dataset for the M2CAI 2016 workflow challenge," 2016, *arXiv:1610.09278*.

[82] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[83] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[84] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[85] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf Process. Syst.*, 2012, pp. 84–90.

[86] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[87] H. Zhang et al., "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2736–2746.

[88] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6450–6459.

[89] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[90] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014, *arXiv:1409.1259*.

[91] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[92] Y. A. Farha and J. Gall, "MS-TCN: Multi-stage temporal convolutional network for action segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3575–3584.

[93] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[94] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[95] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.

[96] T. M. Ward et al., "Computer vision in surgery," *Surgery*, vol. 169, no. 5, pp. 1253–1256, 2021.

[97] G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "Less is more: Surgical phase recognition with less annotations through self-supervised pre-training of CNN-LSTM networks," 2018, *arXiv:1805.08569*.

[98] M. Wagner et al., "Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark," *Med. Image Anal.*, vol. 86, 2021, Art. no. 102770.

[99] L. Maier-Hein et al., "Heidelberg colorectal data set for surgical data science in the sensor operating room," *Sci. Data*, vol. 8, no. 1, 2021, Art. no. 101.

[100] K. Schoeffmann, M. Taschwer, S. Sarny, B. Münzer, M. J. Primus, and D. Putzgruber, "Cataract-101: Video dataset of 101 cataract surgeries," in *Proc. 9th ACM Multimedia Syst. Conf.*, 2018, pp. 421–425.

[101] H. Al Hajj et al., "CATARACTS: Challenge on automatic tool annotation for cataRACT surgery," *Med. Image Anal.*, vol. 52, pp. 24–41, 2019.

[102] G. Ding, F. Sener, and A. Yao, "Temporal action segmentation: An analysis of modern technique," 2022, *arXiv:2210.10352*.

[103] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[104] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[105] D. A. Hashimoto, G. Rosman, D. Rus, and O. R. Meireles, "Artificial intelligence in surgery: Promises and perils," *Ann. Surg.*, vol. 268, no. 1, 2018, Art. no. 70.

[106] O. Bar et al., "Impact of data on generalization of AI for surgical intelligence applications," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020.

[107] A. Boddy, J. Bennett, S. Ranka, and M. Rhodes, "Who should perform laparoscopic cholecystectomy? A 10-year audit," *Surg. Endoscopy*, vol. 21, pp. 1492–1497, 2007.