Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

A Review on Deep Learning in Minimally Invasive Surgery

IRENE RIVAS-BLANCO[1], CARLOS J. PÉREZ-DEL-PULGAR[1], ISABEL GARCÍA-MORALES[1], AND VÍCTOR F. MUÑOZ[1]

[1]Department of Systems Engineering and Automation, University of Málaga, Málaga, Spain.

Corresponding author: Irene Rivas-Blanco (e-mail: irivas@uma.es).

This work was supported by the Andalusian Regional Government, under grant number UMA18-FEDERJA-18.

ABSTRACT In the last five years, deep learning has attracted great interest in computer-assisted systems for Minimally Invasive Surgery. The straightforward accessibility to images in surgical interventions makes deep neural networks enormously powerful for solving classification problems in complex surgical scenarios. The objective of this work is to provide readers a survey on deep learning models applied to minimally invasive surgery, identifying the different architectures used depending on the application, the results achieved until now, and the publicly available surgical datasets that can be used for validating new studies. A total of 85 publications have been extracted from manual research from four databases (IEEE Xplorer, Springer Link, Science Direct, and ACM Digital Library). After analyzing all these studies, they have been classified into four applications: surgical image analysis, surgical task analysis, surgical skill assessment, and automation of surgical tasks. This work provides a technical description of these works and a comparison among them. Finally, promising research directions to advance in this field are identified.

INDEX TERMS Deep Learning, Convolutional Neural Network, Deep Neural Network, Minimally Invasive Surgery, Laparoscopic Surgery, Robot-Assisted Surgery.

I. INTRODUCTION

Minimal invasive surgery (MIS), or laparoscopic surgery, has become the common practice in many surgical interventions with high benefits for patients. However, it introduces new challenges for surgeons such as the lack of direct vision, tactile sensation, and limitation in the motion of the instruments. Robot-assisted surgery (RAS) overcomes several issues of MIS and improves the surgeons' efficiency with a more accurate and intuitive movement of the instruments. The interest in surgical robots is undisputed if we have a look at the huge economic efforts that the major economies of the world are making to boost this market, which is expected to grow at a compound annual growth rate of 10.7% during the forecast period from 2019 to 2029, reaching a market of $15.43 billion by 2029 [1]. In the academic field, the scientific community is also showing great interest with thousands of publications in the last decades and numerous projects, which are being founded to advance in this field. The company Intuitive Surgical is also boosting the research in surgical robotics supporting the scientific community with research platforms of the da Vinci Surgical System, known as da Vinci Research Kit (dVRK), and facilitating cooperation among different research groups.

The efficiency of surgical robots as a tool to improve surgeons' skills has been widely demonstrated. However, at this moment these systems are not able to provide real assistance to the surgeon. They just limit to replicate the motions performed by the surgeon in a master console into a slave platform. Hence, researchers have addressed their efforts on developing automatic ways of assistance to reduce the surgeons' workload during the interventions. Being able to perform some autonomous tasks and to take decisions autonomously in real-time requires a deep understanding of the environment in which the system is working. Thus, recognizing what elements are in the scene and inferring what is happening at a particular time during an intervention is vital to advance in intelligent systems for MIS. Explicit modeling approaches for developing visual servoing techniques in a surgical scenario are inefficient given the large variability between people, organs, and tissue. In contrast, machine learning techniques that learn implicit models directly from raw data appear to be very suitable in these dynamic and complex scenarios [1].

[1]https://bisresearch.com/industry-report/global-surgical-robotics-market.html

Garrow et al. [2] provided a review of machine learning techniques for surgical phase recognition. In their study, the authors identified Hidden Markov Models (HMM) and artificial neural networks as the most frequent ML models for this application. They also pointed out that neural networks are becoming more popular due to their capability to learn important features from raw data, unlike HMMs, which require manual feature extraction. This is due to the technological advancements in surgery, especially in MIS and RAS, developed in the last decades, which have increased the quantity of data available during an intervention. Thus, deep learning (DL) techniques result very attractive in this area. Furthermore, the surgical scene represents a huge challenge in perception techniques due to the dynamic and complex nature of the human body, which leads to a wide field of new opportunities for investigation. Anteby et al. [3] presented an interesting study of deep learning techniques in laparoscopic surgery. The aim of their review is to familiarize clinicians with this new technique, so they focused their study on the clinical value of the reporting works.

In this work, we present a systematic review of deep learning models applied to minimally invasive surgery from a technical point of view. We analyze the different DL models used in this field and their main applications for surgical computer-assisted systems. The major contributions of this paper are the following:

- It presents a comprehensive review of deep learning publications in minimally invasive surgery. To enable a systematic analysis, the publications are categorized according to their application: surgical image analysis, surgical task analysis, surgical skill assessment, and automation of surgical tasks. This design aims to serve as a reference for researchers looking for studies related to their work.
- It provides a description of the publicly available surgical datasets that researchers can use to validate their DL models. To facilitate future researches of readers, links for downloading these public datasets are also provided.
- For each application, we present a technical comparison of the publications included in this survey. These comparisons offer the readers a classification of the studies with relevant information such as the DL model employed, the type of input data, the surgical procedure analyzed, or the dataset used to validate the models. To facilitate the comparison of the different approaches presented, performance metrics are also reported.
- This work also points out two promising research directions in deep learning for minimally invasive applications.

The rest of the paper is organized as follows. Section II provides an overview of deep learning as well as of the most common models used for minimally invasive surgery applications. Section III describes the review methodology followed to conduct this survey, the criteria used to select the relevant publications, and an analysis of the results. Section IV lists and describes the publicly available datasets of MIS, providing the links for downloading the data. In Section V, the studies included in this survey are described and grouped by the following categories: surgical image analysis, surgical tasks analysis, surgical skill assessment, and automation of surgical tasks. Finally, section VI suggests two promising research directions to advance in this field, and section VII outlines the conclusions

II. BRIEF OVERVIEW OF DEEP LEARNING

A Deep Neural Network (DNN) is a machine learning technique inspired in the human brain structure that provides computational systems with artificial intelligence. The network receives a set of inputs that undergo successively transformations through processing units, called hidden layers, to learn a high-level representation of the data useful to solve a particular problem. The units in each layer are connected to units in the adjacent layers with a particular weight and bias. The weighted sum of the inputs of each layer is transformed based on an activation function. The output of this function is then fed as input to the subsequent unit in the next layer [4]. The goal of deep learning techniques is to learn or adjust the network parameters (connection weights and bias) to minimize a loss function, which computes the distance between the prediction of the network and the objectives from the training data [5]. This iterative process is represented in Fig. 1. The main difference with earlier generation machine learning techniques is the automation of feature extractors without any manual design. The high advances of DL in the mid-1980s were possible thanks to the advent of the backpropagation learning algorithm, which allows the computation of the contribution of each parameter of the network to the final loss from the outer layers back to the bottom ones. This improvement was complemented by the proliferation of cheaper and powerful processing units and the explosion of big data in the last 10 years.

A. DL MODELS

Convolutional Neural Networks (CNNs) are the most known architectures of DL and they are the most used for image processing applications. Some well-known implementations of CNNs are AlexNet [6], VGGNet [7], GoogleNet [8], or ResNet [9]. A CNN is composed of a series of convolution and pooling layers followed by a fully connected layer. The role of each of these layers of the network is as follows:

- Convolution layers: the convolution operation learns local patterns to extract the high-level features of the input image. Usually, the first convolution layer captures low-level features such as edges or colors, while the outer layers provide a high-level understanding of the images.
- Pooling layers: the goal of these layers is to decrease the dimensionality of the feature maps through a function such as max-pooling or average-pooling.
- Fully-connected layers: these layers are responsible for the actual classification of the image by learning non-

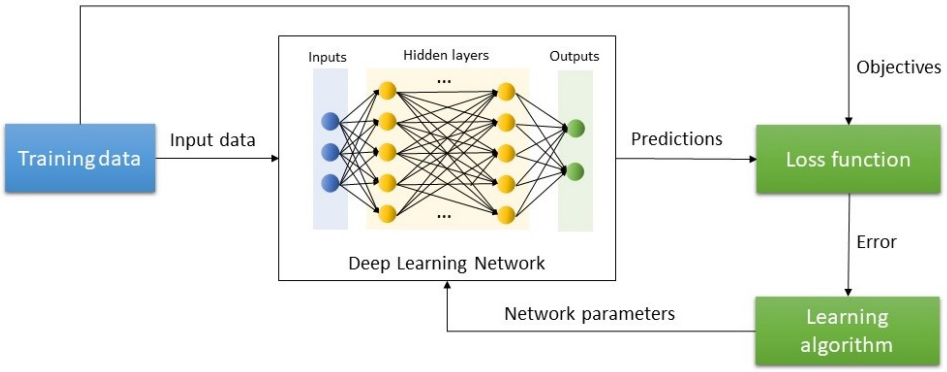

FIGURE 1. Training process in Deep Learning algorithms.

linear combinations of the high-level features extracted in the previous layers.

The designer of the network has to decide the optimal number of layers to achieve a trade-off among a good performance of the model, generalization with new data, and high computational speed to perform real-time inferences. Shallow networks model a few number of parameters and therefore they can perform predictions very fast with more generalization but less accuracy. In this case, the network has not yet modeled all the relevant parameters of the training data. In contrast, very deep networks model a high number of parameters and can provide high accuracy predictions for the training data, but it may lack generalization to new data due to the overfitting of the network, i.e., the network may be learning specific patterns of the training data, but which are not relevant for new data. Another key fact to design good predictive models is to have large amounts of labeled data for training. However, obtaining a sufficient amount of annotated data in specific domains such as surgery is difficult and costly. To alleviate this problem, most networks are pre-trained using labeled data coming from other domains, such as Imagenet [10].

One of the limitations of CNNs is that they cannot handle variable input image sizes. In contrast, Fully Convolutional Neural Networks (FCNNs) have the advantage over CNN of operating on inputs of any size, producing an output with reduced spatial dimensions. This makes them suitable for end-to-end pixel-level semantic labeling, as the spatial configuration of the image is preserved across the layers. However, they lack real-time capabilities and masks usually have holes or do not respect edges. Another limitation of CNNs is that they lack the ability of processing temporal information of data that come in sequences, as video data. To consider the temporal dependencies in the input data, we use Recurrent Neural Networks (RNNs). Unlike feedforward neural networks, the processing units in an RNN form a cycle. This allows the network to have memory about the previous states and to use that to influence the current output [4]. The main implementation of RNNs is Long Short Term Memory (LSTM) networks. A LSTM consists of blocks of memory cell state through which signal flows while being regulated by input, forget, and output gates [4], which allows to add or remove information to the cell state. The input gate decides which values will be updated, while the forget gate is used to discard information. Finally, the output gate retains the information that is not used in the current time step but can be useful in the future. To take advantage of both networks, many authors propose DL models that combine CNNs with RNNs connected in a serial configuration. These models use a CNN to extract spatial features from the input images, an their output is fed to a RNN to take into account the temporal context of the data.

B. DL FRAMEWORKS AND LIBRARIES

To facilitate the implementation of DL architectures, there exist several open-source frameworks and libraries that incorporate the complex mathematical functions, training algorithms, and statistical modeling required for developing DL applications. Most of these tools are located on GitHub in the form of repositories. GitHub itself keeps a lot of monitoring information about software development such as the number of stars, watches, or forks. The most popular DL frameworks and libraries, which main characteristics are summarized in Table 1, are listed below:

- TensorFlow: it is an end-to-end source platform for machine learning created by Google. It is by far the most popular DL library based on the number of GitHub stars, and it supports both CPU and GPUs. The programming interface includes APIs for Python and C++.
- Keras: Keras, also created by Google, is a high-level DL API perfect for beginner users. As low-level motor, Keras uses libraries like TensorFlow, Theano, or CNTK, wrapping them and hiding their complexity for the user. One of the strongest points of this framework is its modularity, which allows an easy combination of neural layers, cost functions, optimizers, initialization schemes, and activation functions.
- CNTK: the Microsoft Cognitive Toolkit (CNTK) im-

TABLE 1. Popular frameworks and libraries for Deep Learning [11].

Tool	Type	Creator	Written in	API	GitHub stars	Link
TensorFlow	Framework	Google	Python, C++	Python	153k	https://github.com/tensorflow/tensorflow
Keras	Library	F. Chollet	Python	Python	50.5k	https://github.com/keras-team/keras
CNTK	Framework	Microsoft	C++	Python, C++, ONNX	17k	https://github.com/microsoft/CNTK
Theano	Framework	University of Montreal	Python	Python	9.3k	https://github.com/Theano/Theano
Caffe2	Framework	Facebook	C++	Python, C++, ONNX	8.4k	https://github.com/facebookarchive/caffe2
PyTorch	Library	S. Chintala, G.Chanan	Python, C	Python, ONNX	45.6k	https://github.com/pytorch/pytorch

plements efficient DNNs training for speech, image, handwriting, and text data.

- Theano: it is a tool written in Python for creating networks using symbolic logic. Theano was started in 2007, but it is no longer under active development.
- Caffe2: it is a lightweight, modular and scalable DL framework created by Facebook. It is used at the production level at Facebook while development is done in PyTorch.
- PyTorch: it is a Python library for GPU-accelerated DL. It has become popular by allowing complex architectures to be built easily.

III. REVIEW METHODOLOGY

This section describes the review methodology followed for selecting the publications included in this review. First, the review protocol is described, followed by the selection criteria to discard the non-relevant papers for this survey. Finally, an analysis of the results considering the selected publications is presented.

A. REVIEW PROTOCOL

The first step to conduct this review is to search relevant publications in several databases. This search has been carried out in the following databases:

- IEEE Xplorer (https://ieeexplore.ieee.org).
- Springer Link (https://springerlink.com).
- Science Direct (https://sciencedirect.com).
- ACM Digital Library (https://dl.acm.org).

The keywords used during the search were ("Deep Learning" OR "Deep Neural Network") AND ("Laparoscopic Surgery" OR "Minimally Invasive Surgery" OR "Robotic Surgery" OR "Robot Assisted Surgery"). In each database, the eighth combinations of these keywords were used. The search was restricted to publications in the last five years, i.e., the period between 2015 and 2020. Moreover, in Springer Link the search was limited to the disciplines 'Computer Science' and 'Engineering', and in Science Direct, only 'research articles' type were considered. This first search led to a total of 338 publications, distributed as followed: 104 publications in IEEE Xplorer, 134 in Springer Link, 81 in Science Direct, and 19 in ACM Library.

B. SELECTION CRITERIA

From this initial search, abstracts, summaries, reference book entries, tutorials, surveys, and doctoral symposiums were excluded. For the remaining publications, the following inclusion criteria were applied:

- Only technical studies were considered. Thus, medical studies with no technical content were excluded from this survey.
- Selected publications are only related to minimally invasive surgical applications. Thus, papers on the field of medicine related to rehabilitation or out-patient surgery were excluded.
- The type of input data of the papers included in this survey covers laparoscopic images (including simulation environments) of conventional and robotic surgery (using rigid robotic instruments). Thus, non-rigid instruments and studies using X-ray images, computed tomography (CT) scans, or magnetic resonance imaging were excluded from the review.
- After a comprehensive reading of the publications, they have been grouped into four main applications: surgical image analysis, surgical tasks analysis, surgical skill assessment, or automation of surgical tasks. To focus the research, publications out of these categories have been excluded from this survey.

After applying these inclusion and exclusion criteria, a total of 85 publications were considered for exhaustive reading and analysis. As shown in Table 2, most of the papers are from the databases IEEE Xplorer and Springer Link. This table also shows the number of journal and conference papers found in each database.

TABLE 2. Number of publications after applying the inclusion criteria from the initial search.

Database	Journals	Conferences	Total
IEEE Explorer	19	30	49
Springer Link	6	19	25
Science Direct	10	0	10
ACM Digital Library	0	1	1
Total	35	50	85

C. ANALYSIS OF THE RESULTS

Figure 2 shows the growing interest in DL techniques in the field of minimally invasive surgery in the last years, from 3 publications in 2016 to 24 and 25 in 2019 and 2020, respectively (no publications were found in 2015). From the graph, we can see that the explosion of DL in MIS started in 2017, and it has been growing until the present.

Tables 3 and 4 contain the journals and conferences, respectively, identified as the most relevant according to the

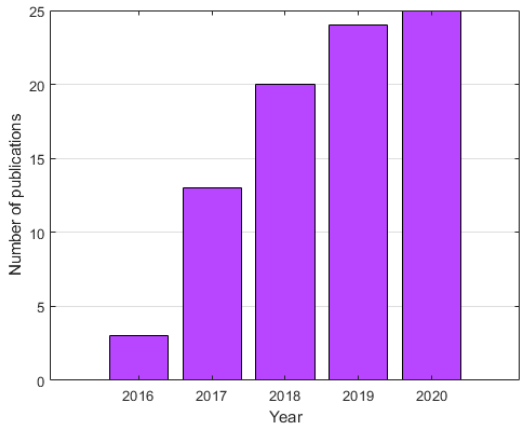

FIGURE 2. Number of publications per year for the last five years.

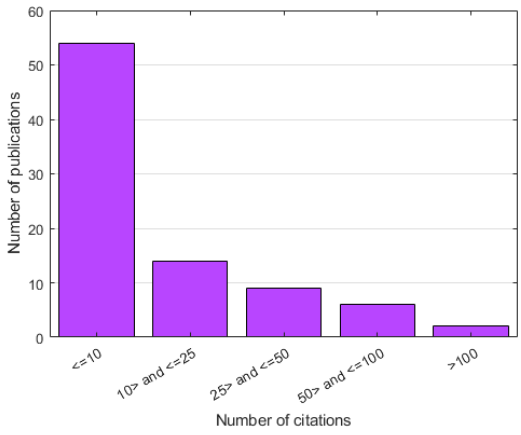

FIGURE 3. Number of citations of the publications from 2015 to 2020.

number of publications. On the one hand, IEEE Transactions on Medical Imaging and IEEE Robotics and Automation Letters have been identified as the most pertinent journals with 6 and 5 publications, respectively, followed by Medical Image Analysis and IEEE Access with 3 publications each. On the other hand, the Medical Image Computing and Computer-Assisted Intervention (MICCAI) conference has been identified as the most pertinent international conference with 12 publications, followed by the IEEE International Conference on Robotics and Automation (ICRA) with 7 publications and the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) with 3 publications.

TABLE 3. Journals identified as pertinent and number of publications.

Name of the journal	Publications
IEEE Transactions on Medical Imaging	6
IEEE Robotics and Automation Letters	5
Medical Image Analysis	3
IEEE Access	3

TABLE 4. Conferences identified as pertinent and number of publications.

Name of the conference	Publications
MICCAI	12
ICRA	7
IROS	3

The impact of the research of DL for MIS applications can be measured with the number of citations of the publications, which are represented in Figure 3. Although most of the publications are under 10 cites, it is quite relevant that there are 9 publications with more than 25 and less than 50 cites, 6 with more than 50 and less than 100, and 2 publications with more than 100 citations. These are Twinanda et al. [12] with 246 citations and Shvets et al. [13] with 124.

IV. PUBLICLY AVAILABLE SURGICAL DATASETS

The availability of large datasets is essential to advance in the field of deep learning in order to train neural networks. In the last years, there have been great efforts to develop large public datasets of surgical procedures annotated by experts. Public datasets are important, not only for giving the possibility of developing DL algorithms to researchers that do not have the possibility of acquiring their own data but also for comparing the performance of the different algorithms proposed in the literature.

The Medical Image Computing and Computer Assisted Intervention (MICCAI) society, formed in 2004, hosts annual challenges in their events to promote and facilitate the research, education, and practice of computer vision in medical interventions. These challenges are international competitions that aim the benchmarking of multiple algorithms on publicly released datasets. This facilitates the comparison of the different computer vision solutions developed by researchers all around the world. In 2020, the MICCAI challenge working group elaborated a guideline to standardize the writing and reviewing process of biomedical image analysis challenges and help researchers to interpret and reproduce the results [14]. Yearly accepted challenges datasets are available online on the MICCAI website, and past events challenges can be found on the Grand Challenge website[2], a platform for end-to-end development of machine learning solutions in biomedical imaging.

The complete list of the publicly available datasets found during our search is presented in Table 5. This table provides the name of each dataset, the year of publication, the number of videos or images available, the surgical procedure used to collect the data, the type of instruments used (rigid or robotic), and a brief description of the annotations of the dataset. The most common procedures used for collecting data are cholecystectomy, colorectal surgery, and gynecologic surgery. This is because these are easy procedures that are mainly performed using laparoscopy instead of open surgery. Most datasets include video data, but only two incorporate kinematic data, which provides a big amount

[2]https://grand-challenge.org/challenges/

TABLE 5. Publicly released datasets.

Name	Year	Data	Procedure	Instruments	Annotations
JIGSAWS	2014	103 videos + kinematic data	In-vitro experiments	Robotic	Gestures labels and global rating score
EndoVis 2015	2015	Images and videos	Colorectal surgery	Rigid and robotic	Tools segmentation[1]
M2CAI16	2016	41 videos	Cholecystectomy	Rigid	Phases[2]
		15 videos			Tools presence[3]
					Tools spatial bounds[4]
Cholec80	2016	80 videos	Cholecystectomy	Rigid	Phases and tools
EndoVis 2017	2017	8 videos	Porcine procedures	Robotic	Tools segmentation[5]
		30 videos	Colorectal surgery	Rigid	Phase and tools[6]
ATLAS Dione	2017	86 videos	In-vitro experiments	Robotic	Tools, actions, and expertise levels
SurgicalActions160	2017	160 videos	Gynecologic surgery	Rigid	Surgical action
EndoVis 2018	2018	16 videos	Nephrectomy	Robotic	Scene segmentation[7]
		30 videos	Colorectal surgery	Rigid	Phases and tools[8]
LapGyn4	2018	55K images	Gynecologic surgery	Rigid	Not annotated
EndoVis 2019	2019	30 videos	Cholecystectomy	Rigid	Phases, actions, tools categories, and skills[9]
		10 seconds video	Colorectal surgery		Tools segmentation[10]
UCL dVRK	2020	14 videos + kinematic data	Ex-vivo experiments	Robotic	Tools segmentation
FlapNet	2020	62 minutes video	Lobectomy	Robotic	Tissue flap and tools
LapSig300	2020	300 videos	Colorectal surgery	Rigid	Phases, actions and tools

[1] Instrument segmentation and tracking sub-challenge.
[2] m2cai16-workflow dataset.
[3] m2cai16-tool dataset.
[4] m2cai16-tool-location dataset.
[5] Robotic Instrument Segmentation sub-challenge.
[6] Surgical Workflow Analysis in the SensorOR sub-challenge.
[7] Robotic Scene Segmentation sub-challenge.
[8] Surgical Workflow Analysis in the SensorOR sub-challenge.
[9] Surgical Workflow and Skill Analysis sub-challenge.
[10] Robust Medical Instrument Segmentation (ROBUST-MIS) sub-challenge.

of useful information for analyzing metrics related to the motion of the tools. Moreover, datasets including kinematic parameters are collected on in-vivo and ex-vivo experiments, but not on real surgeries. This is because kinematic data is straightforward to acquire from the dVRK, used in research environments, but not from the commercial versions of the da Vinci, used in real surgeries. Finally, annotations are mainly for tool classification and segmentation, and phase recognition, but there is only one dataset that offers annotations of the complete surgical scene. Next, these surgical datasets are further described:

- EndoVis sub-Challenges: these are yearly sub-challenges launched by the MICCAI society under the endoscopic vision challenge. These sub-challenges include rigid and robotic instruments segmentation and tracking as well as surgical workflow analysis for different procedures.
- M2CAI16: This event included two sub-challenges: the surgical workflow challenge, and the surgical tool detection challenge. For the first challenge, they introduced eight surgical phases for cholecystectomy procedures. The challenge consists of identifying the phase at a particular time using only visual information. For this challenge, they created the m2cai16-workflow dataset[3], which has 41 videos (27 for training and 14 for testing) with ground truth annotations of the phases (they defined eight surgical phases) [15], [16]. In the second challenge, the objective is to identify all surgical tools that are present in an image. For this challenge, they created the m2cai16-tool dataset [15], consisting of 15 cholecystectomy videos (10 for training and 5 for testing) with binary annotations of the present tools (seven different tools). The m2cai16-tool dataset has been extended with annotations of the spatial bounds of the tools, named m2cai16-tool-locations[4].
- Cholec80[5]: it is a large dataset containing 80 videos, recorded at 25 fps, of cholecystectomy surgeries performed by 13 surgeons at the University Hospital of Strasbourg [12]. The whole dataset is annotated with the surgical phase (at 25fps) and tool presence (at 1 fps). The dataset is divided into two subsets of 40 videos each, for training and testing. In the training subset, 10 videos have also been fully annotated with the bounding boxes of tools. The Cholec80 dataset has been increased with 40 additional annotated with the surgical phases [17]. This dataset, named Cholec120[6], accumulate over 75h of recordings. Cholec80 has also been used to generate a new dataset for smoke removal applications. This dataset, named ITEC Smoke_Cholec80 Image, contains 100K frames from Cholec80 annotated with classes smoke and non-smoke.
- JIGSAWS[7]: the JHU-ISI gesture and skill assessment working set (JIGSAWS) includes data on three elemen-

[3] http://camma.u-strasbg.fr/m2cai2016/index.php/program-challenge/
[4] https://ai.stanford.edu/~syyeung/tooldetection.html
[5] http://camma.u-strasbg.fr/datasets
[6] http://ftp.itec.aau.at/datasets/Smoke_cholec80/
[7] http://cirl.lcsr.jhu.edu/jigsaws

tary surgical tasks (suturing, knot-tying, and needle-passing) performed by eight surgeons on the da Vinci surgical system. Data consists of three components: kinematic data (19 kinematic variables divided into 76-dimensional kinematic data), video data, and manual annotations (15 surgical gesture labels and global rating scoring of each task) [18].

- ATLAS Dione[8]: this dataset consists of 99 action videos of 10 surgeons from the Roswell Park Cancer Institute (Buffalo, NY) performing 6 different surgical tasks on the da Vinci Surgical System. These tasks include basic skill tasks which are part of the Fundamental Skills of Robotic Surgery (FSRS) curriculum, and also specific skills required for the Robotic Anastomosis Competency Evaluation (RACE). The dataset contains annotations of the robotic tools, action labels and their timestamps, and surgeon expertise levels [19].
- UCL dVRK dataset[9]: this dataset consists of 14 videos of 300 frames each using the da Vinci Research Kit on 5 different kinds of animal tissue. Each video is produced following four steps: (1) a movement is performed with the dVRK and the kinematic data is recorded; (2) the movement is reproduced using the recorded data with animal tissue background to collect image frames; (3) the same movement is reproduced a second time on a green screen to obtain tools ground truth segmentation masks; (4) for each frame, an associated image of the virtual tools is produced using a dVRK simulator [20].
- FlapNet[10]: this dataset contains videos of lobectomy surgery performed on embalmed cadaver by experienced surgeons using the da Vinci Xi. Images are labeled with the tissue flap to be retracted and the instrument visible in the scene [21].
- LapGyn4[11]: it is a four-part gynecologic laparoscopic dataset of over 500 interventions with over 55K images. It contains a collection of images of (1) surgical actions, (2) anatomical structures, (3) conducted actions on specific anatomy, and (4) examples of visible instruments [22].
- SurgicalActions160[12]: this dataset contains 10 examples of 16 classes of surgical actions (160 videos in total). These actions are subject to surgical errors so that they constitute good testing data in the context of surgical skill assessment [23].
- LapSig300: it is the first large-scale dataset of laparoscopic colorectal surgery. It contains 300 videos obtained from 19 high-volume endoscopic centers in Japan. The surgeries were performed by numerous different surgeons over 10 years. Frames have annotations of the surgical phase and action, and tools' presence and location. This data is made available by the authors on request [24].

. In t

V. DEEP LEARNING APPLICATIONS IN MINIMALLY INVASIVE SURGERY

In the latest years, DL techniques are having a huge impact on minimally invasive surgery research. On the one hand, deep learning was devised to manage complex data such as raw images. In MIS, images are a straightforward source, which is always available in any intervention. On the other hand, the number of publicly available datasets greatly facilitates to advance in this field even for researchers that do not have access to clinicians. Moreover, robot-assisted surgery augments the possibilities of DL techniques with large kinematic data of the instruments and the surgeons' gestures. As shown in Figure 4, the main applications of DL models in the field of minimally invasive surgery are: surgical image analysis, surgical tasks analysis, surgical skill assessment, and automation of surgical tasks. These tasks provide computer-assisted surgical systems information for understanding the surgical scenario, and therefore, they are the basis for developing autonomous systems able to make decisions, to collaborate with surgeons during the interventions, or to provide useful feedback to surgeons during trainee or on-line surgery. This section describes the studies included in this survey, classified into the previous categories.

A. SURGICAL IMAGE ANALYSIS

Analyzing the surgical image is essential for understanding the surgical scenario, and therefore, for being able to reason and take-decisions about it. The methods for objects recognition in an image can be divided into the following, depending on the expected output of the model:

- Classification: given an input image, the network outputs a class or label mask (Figure 5(a)). This mask may be binary, in case the aim of the network is to detect the presence of an object, or it can be a multi-class classification problem, in which there are several masks, one for each type of object. For example, an image can be labeled as "Grasper" or "Scissors", if we are performing instruments classification.
- Detection: these algorithms are used to locate the objects in the image and to represent them with bounding boxes (Figure 5(b)). The bounding boxes are always rectangular, so this method cannot detect shapes or edges.
- Segmentation: in segmentation tasks, the image is marked with pixel-wise masks for each object in the image. It can be divided into binary segmentation, every pixel in an image is labeled as an instrument or background, as in Figure 5(c), or multi-class segmentation, in which different colors represent different object categories.

Technical characteristics of the publications included in this survey for surgical image analysis are shown in Table

[8]https://datasets.bifrost.ai/info/848

[9]https://www.ucl.ac.uk/interventional-surgical-sciences/ex-vivo-dvrk-segmentation-dataset-kinematic-data

[10]https://github.com/Stormlabuk/FlapNet

[11]https://zenodo.org/record/1219280#.X6E4O4hKiUk

[12]http://ftp.itec.aau.at/datasets/SurgicalActions160/

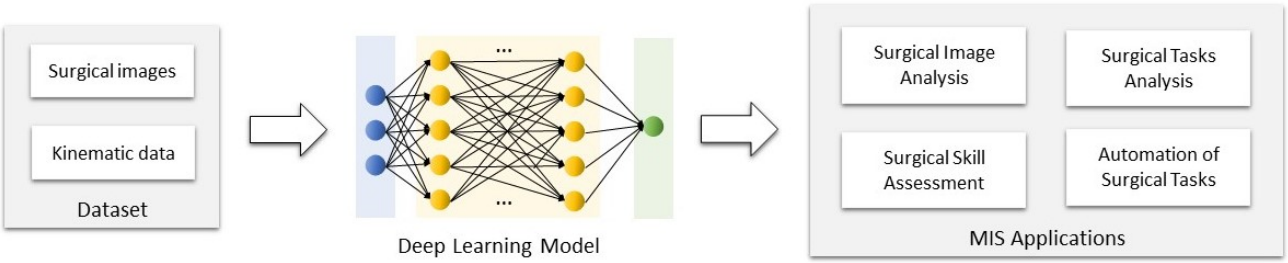

FIGURE 4. Main applications of DL models in Minimally Invasive Surgery. The DL models input surgical images and/or kinematic data from a dataset. Then, the DL models are trained to perform one of the following applications: surgical image analysis, surgical tasks analysis, surgical skill assessment, or automation of surgical tasks.

(a) (b) (c)

FIGURE 5. Methods for objects recognition in a surgical image (particularized to the task of instruments recognition): (a) Classification ("Grasper" label) [25]; (b) Detection [19]; and (c) Segmentation [26].

6. These studies are ordered by the particular application, and the following information is given: the year of the publication, the type of tool (rigid or robotic), the application of the study, the DL model used, the surgical procedure of the input data, the dataset used in the study, and the metric that evaluates the performance of the model. The type of input data is not included in this table because all these studies, except [20] that combines images with kinematic data, uses only image data to train the networks.

Surgical instrument recognition is the most studied application in surgical image analysis, as instruments represent the interaction mechanism between the surgeon and the surgical scenario. Moreover, the type of surgical instrument used at a given time provides key information of what is happening in the surgery. Thus, being able to recognize surgical instruments in the scene is vital for context-aware surgical systems. Recognition of anatomical structures is also an important task in a surgical scenario and has many applications for understanding the surgical scene as well as for computer-assisted diagnosis systems. Recognition of other surgical material, such as the suturing thread, provides useful information for the automation of surgical tasks. Recognition of the suture thread is a challenging task, as it is high deformable and suffers from frequent occlusion. Hu et al. [27] proposed a multi-stage framework for suture segmentation based on predicting directly the curvilinear structural information of the thread, instead of modeling the task. Similarly, Lu et al. [28] addressed the problem of suture thread segmentation. They proposed a DL model for accurately detect the suture's tip. Then, they used a numerical method to segment and compute the 3D coordinates of the thread. They achieved good results in in-vitro experiments. Next, the rest of the studies included in the surgical image analysis category are further described.

1) Classification and detection of surgical instruments

In laparoscopic videos, each image usually contains more than one instrument at once, thus multi-label classification is more interesting than binary classification. In this type of algorithms, each instance can belong to more than one class. Wang et. al [25] presented a multi-class classification method combining two CNN models, VGGNet and GoogLeNet, to produce the final result. Each network is trained separately, and the prediction of each of them is averaged to compute the final classification. The main limitation of this work is that it does not consider the temporal information of the videos. However, temporal context is important to distinguish surgical tools, to overcome the problem of the high similarity to one another. Mishra et al. [29] propose to incorporate spatio-temporal information into the tools classification problem using a deep LSTM network. In the first stage, a CNN is trained to detect tool presence in individual frames. Then,

TABLE 6. Comparison of the surgical image analysis publications using DL architectures.

Ref.	Year	Tools	Application	DL model	Procedure	Dataset	Results
[25]	2017	Rigid	Instruments classification	VGGNet+GoogLeNet	Cholecystectomy	m2cai16-tool	63.8% (mAP)
[12]	2017	Rigid	Instruments classification	EndoNet	Cholecystectomy	Cholec80	81% (mAP)
[29]	2017	Rigid	Instruments classification	ResNet50+LSTM	Cholecystectomy	m2cai16-tool	88.7% (mAP)
[30]	2018	Rigid	Instruments classification	CNN+RNN	Cholecystectomy	Cholec80	97.9% (mAP)
[31]	2019	Rigid	Instruments classification	3D DenseNet+GCN	Cholecystectomy	m2cai16-tool	90.2% (mAP)
						Cholec80	90.13% (mAP)
[32]	2017	Robotic	Instruments classification	U-Net	Colorectal surgery	EndoVis 2015	99.9% (mAP)
[33]	2020	Rigid	Instruments classification	VGG-50+LSTM	Cholecystectomy	Cholec80	89.1% (acc)
[19]	2017	Robotic	Instruments detection	RPN + Fast R-CNN	In-vitro experiments	Atlas Dione	90% (mAP)
[34]	2018	Robotic	Instruments detection	FCNN	Colorectal surgery	EndoVis 2015	83.7% (AP)
[15]	2018	Rigid	Instruments detection	VGG16	Cholecystectomy	m2cai16-tool	81.8% (mAP)
[35]	2019	Robotic	Instruments detection	3D FCNN	Colorectal surgery	EndoVis 2015	85.1% (dice)
[36]	2017	Robotic	Instruments detection	CNN+STC	In-vivo experiments	NPA	93.2% (AP)[1]
[37]	2019	Robotic	Instruments detection	Hourglass+VGG-16	In-vitro experiments	Atlas Dione	91.6% (mAP)
					Colorectal surgery	EndoVis 2015	100% (mAP)
[38]	2020	Robotic	Instruments detection	Stacked Hourglass	In-vitro experiments	Atlas Dione	98.5% (mAP)
						EndoVis 2015	100% (mAP)
[39]	2020	Robotic	Instruments detection	VGG16	In-vitro experiments	Atlas Dione	90.01% (mAP)
[40]	2017	Rigid	Binary segmentation (instruments)	ResNet-50	Colorectal surgery	EndoVis 2015	88.9% (dice)
[41]	2017	Robotic	Binary segmentation (instruments)	ToolNet	Colorectal surgery	EndoVis 2015	81% (mAP)
[42]	2018	Robotic	Binary segmentation (instruments)	ResNet+LSTM	Colorectal surgery	EndoVis 2015	92.6% (mAP)
			Multi-class segmentation (instruments)				92.4% (mAP)
[13]	2018	Robotic	Binary segmentation (instruments)	TernausNet-16	Porcine procedures	EndoVis 2017	90.1% (dice)
			Parts segmentation (instruments)	TernausNet-16			76% (dice)
			Multi-class segmentation (instruments)	TernausNet-11			45.9% (dice)
[43]	2020	Robotic	Binary segmentation (instruments)	ResNet-18	Porcine procedures	EndoVis 2017	89.6% (dice)
			Parts segmentation (instruments)				76.4% (dice)
[26]	2019	Robotic	Binary segmentation (instruments)	CNN	Porcine procedures	EndoVis 2017	91.6% (dice)
			Parts segmentation (instruments)				73.8% (dice)
			Multi-class segmentation (instruments)				34.7% (dice)
[44]	2019	Robotic	Binary segmentation (instruments)	U-NetPlus	Porcine procedures	EndoVis 2017	90.2% (dice)
			Parts segmentation (instruments)				76.6% (dice)
			Multi-class segmentation (instruments)				46.07% (dice)
[45]	2020	Robotic	Multi-class segmentation (instruments)	MobileNetV2	Porcine procedures	EndoVis 2017	58.3% (IoU)
[46]	2020	Robotic	Binary segmentation (instruments)	ResNet-18+LSTM	Porcine procedures	EndoVis 2017	91% (dice)
			Multi-class segmentation (instruments)				64% (dice)
[47]	2020	Robotic	Binary segmentation (instruments)	VGG16	Porcine procedures	EndoVis 2017	81.15% (dice)
[48]	2019	Robotic	Multi-class segmentation (instruments)	ResNet-10	Various surgeries	NPA	81.4% (mAP)
			Instruments detection				83.1% (mAP)
[49]	2019	Rigid	Binary segmentation (instruments)	LinkNet-152	Phantom and porcine tissue	NPA	88.9% (dice)
[50]	2020	Rigid	Binary segmentation (instruments)	Ternaus-11	Cholecystectomy	Cholec80	93% (AP)
[51]	2017	Rigid	Multi-class segmentation (instruments)	ResNet-101+RNN	Cholecystechtomy	m2cai16-tool	93.3% (mAP)
[20]	2020	Robotic	Binary segmentation (instruments)	FCNN	Ex-vivo experiments	UCL dVRK	95.16% (IoU)
[27]	2018	Robotic	Suture thread detection	U-Net	In-vitro experiments	NPA	1.33 (p-e)
[28]	2019	Robotic	Suture thread segmentation	GoogleNet+LSTM	Artificial tissue	NPA	99.63% (acc)
					Porcine tissue		97.93% (acc)
[52]	2018	NA	Classification of anatomical structures	GoogLeNet	Gynechologic surgeries	NPA	78.1% (acc)
[53]	2019	NA	Liver segmentation	U-Net	Liver surgery	NPA	90.32% (dice)
[54]	2020	NA	Liver detection	VGG16	Liver surgery	NPA	85.9% (acc)
[55]	2019	NA	Nerve and dura mater detection	YOLOv3	Spinal endoscopy	NPA	95.12% (AP)
[56]	2019	NA	Surgical scenario segmentation	Xception	Sleeve gastrectomy	EndoVis 2015	98.44% (dice)
[57]	2019	NA	Polyp detection	VGG16	Colonoscopy	EndoVis 2015	80% (dice)

* When more than one result is presented in the study, the one with the best performance is reported in this table.
** AP = average precision; mAP = mean average precision;p-e = pixels-error; NPA = not publicly available; NA = not applicable.

the features learned by the CNN are used to learn a temporal model using a LSTM network, providing a higher accuracy of classification. Similarly, Al-Hajj et al. [30] propose monitoring tool usage during surgery using convolutional and recurrent neural networks. The proposed framework consists of several CNNs that extract visual features of the videos and RNNs for analyzing the temporal sequence throughout the entire surgery, based on the outputs of the CNNs. With this approach, they augmented the model performance to around 98%. The temporal dimension is also considered in [31]. In this work, the authors use a Graph Convolutional Network (GCN) to learn better features by considering the relationship between continuous video frames. Kurman et al. [32] proposed a modified U-Net architecture for semantic segmentation, with a high performance score. However, this study only considers three different tools versus the seventh classes of the previous works. The code of this work is available at GitHub [13].

[13]https://github.com/aimi-lab/instrument-pose

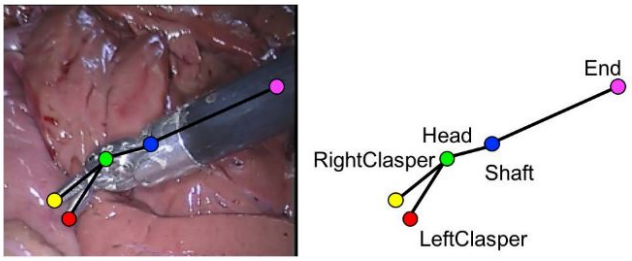

FIGURE 6. Skeleton of the EndoWrist Needle Driver instrument divided into 5 joints and 4 connections [34].

Sarikaya et al. [19] presented in 2017 the first approach that incorporates DNNs for tools detection and localization in robot-assisted surgery. They applied a Region Proposal Network (RPN) jointly with a multimodal convolutional network for localization and a Fast R-CNN for object detection. In this work, they also introduced the ATLAS Dione dataset, the first public set of data of robot-assisted surgery videos with tool annotations. In [34] and [35], the authors focused on articulation detection for robotic instruments. They model each tool as a set of joints and connections between joints (Figure 6). Then, they used a Fully Convolutional Neural Network (FCNN) to detect the joint pairs, which output is used to estimate the pose of the instruments. Real-time tracking of the surgical tools is addressed in [36]. In this work, the authors use a CNN with line segment detector (LSD) to detect the lines of the tools and spatio-temporal context (STC) for tracking the tools' frame by frame in real-time. In [37], a cascading CNN is proposed to recognize and localize robotic surgical tools. The network consists of an hourglass network, which outputs the heatmaps of the instruments' tip area, and a modified VGG-16 network that performs bounding-box regression on these heatmaps. Advancing in this work, Liu et al. [38] propose an anchor-free CNN, modeling the surgical tools as a single point. These works exhibit better results both in speed and accuracy. Yu et al. [39] focused on the detection of small surgical instruments. They combined an attention map created from high-level features with low-level features to enrich the low semantic information.

2) Segmentation of surgical instruments

Binary segmentation of surgical instruments is studied in Laiana et al. [40], who proposed a novel method for real-time instrument tracking that takes advantage of the interdependency between localization and segmentation by carrying out these two tasks simultaneously in a unified CNN. For the same task, Garcia-Peraza-Herrera et al. [41] proposed a lightweight architecture, called ToolNet, which feature one order of magnitude fewer parameters than the state-of-the-art, requiring less memory and allowing for real-time inference. They encoded the multi-scale constraint inside the network architecture to improve the performance of the CNN. Milletari et al. [42] proposed an encoder-decoder architecture in which the encoder is a very deep network based on residual learning, and the decoding is implemented using LSTM cells. This model achieves an accuracy of over 92% both for binary and multi-class segmentation. The code of this work is available at GitHub[14].

The EndoVis 2017 sub-challenge Robotic Instrument Segmentation proposed segmentation tasks: binary segmentation, segmentation of different parts of the instruments, and multi-class segmentation for recognizing different surgical instruments (Figure 7). A comparison of different approaches used in this challenge is shown in Table 6. The lower performance in multi-class segmentation in all the studies may be due to the relatively small dataset size. There are 7 instruments classes, and several of them appear just few times in the training set. Shvets et al. [13] evaluated 4 different DL architectures for instruments segmentation: a modification of U-Net, two modifications of TernausNet and a modification of LinkNet. They achieved the best performance using modified TernausNet architectures for binary and multi-class classification, although in terms of computational efficiency, LinkNet-34 is the fastest model due to the lighter encoder. Their solution is publicly available at GitHub[15]. Pakhomov et al. [43] used a FCNN built upon a ResNet-18 for the same task, getting better performance results. This light-weight deep residual network allows real-time segmentation with a seed of up to 125 fps on high-resolution images.The code of this work is available at [16]. A similar approach is presented in [26], but using auxiliary supervised deep adversarial learning, which allows outperforms previous works in inference speed.

In [44], the authors propose a modified U-Net architecture, named U-NetPlus, for surgical tool segmentation that uses a pre-trained model as the encoder with batch-normalization. In the decoder part, they substitute the deconvolution layer with an upsampling layer that uses nearest-neighbor interpolation, followed by two convolution layers. In [45], robotic instruments segmentation is addressed using a encoder-decoder architecture. The lightweight architecture MobileNetV2 is used as the encoder, and a custom lightweight attention decoder is used to recover the location details. Similar speed but with a better segmentation performance is achieved by Islam et al. [46] using ResNet-18 network with LSTM.

Lee et al. [49] present a weakly supervised framework for surgical tools tracking and segmentation based on a hybrid sensor system that integrates electromagnetic tracking with processing of visual data. This way, it is possible to generate semantic labelling of surgical tools without the need of manual annotations. Another method for instruments segmentation that does not require labeled data is proposed in [50]. In this work, the authors merge supervised learning using simulated images (automatically labeled) and unsuper-

[14]http://github.com/faustomilletari/CFCM-2D
[15]https://github.com/ternaus/robot-surgery-segmentation
[16]https://github.com/warmspringwinds/pytorch-segmentation-detection

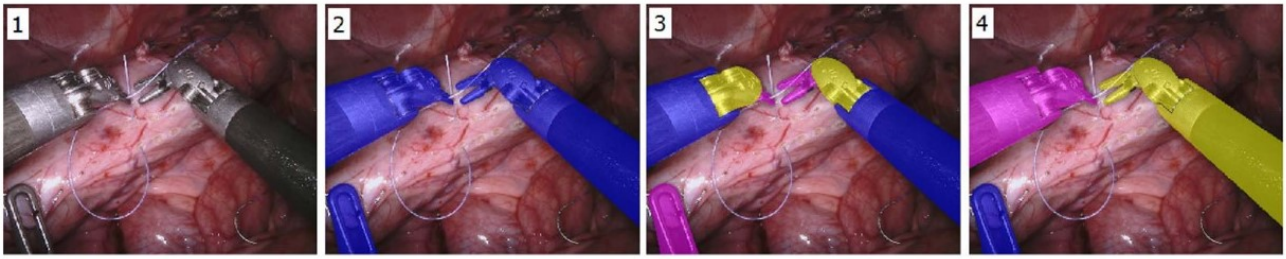

FIGURE 7. Robotic instruments segmentation: (1) original video frame; (2) binary segmentation; (3) parts segmentation where each class corresponds to a different part of the instrument; and (4) multi-class segmentation where each class corresponds to a different instrument [13].

vised learning using real images in a joint learning scheme. Similarly, Lui et al. [47] propose an unsupervised framework for instruments segmentation based on generating anchors to provide initial training supervision, and augmenting the supervision by a semantic diffusion loss. For the anchor generation, they encode the knowledge about surgical instruments into hand-designed cues and generate pseudo labels for training. Then, a semantic diffusion loss is proposed to resolve the ambiguity in the generated anchors exploiting the temporal correlation in the surgical videos.

Previous works use surgical images as the input source for the deep learning models. Surgical robots allow an easy access to an additional key information of the motion of surgical instruments: the kinematic data of the surgical system. Kinematic data include the 3D position and velocity of the instruments, the rotation and also force and torques. Thus, combining kinematic data with laparoscopic images could improve the performance of the models for instruments recognition, and could deal with the problem of large annotated data. This approach is addressed by Colleoni et al. [20] using a FCNN model. The model inputs images recorded with a dVRK and segmentation masks produced using a simulator, which shares the same kinematic values with the real robot. For this work, they generated the UCL dVRK dataset, which contains annotated images with both segmentation ground truth and kinematic information.

3) Segmentation of anatomical structures

Segmentation of surgical instruments has grabbed the attention of many researchers in the field of machine learning applied to surgical procedures. However, laparoscopy is a complex scenario that involves many other objects which recognition is essential for a deep analysis and understanding of a surgical scene, such as other surgical material as the suture thread or anatomical structures. Liver segmentation is a another particularly challenging task as this organ suffers many deformations during a surgery, and it is usually overlapped by other organs. Nazir et al. [54] proposed a method to search a part of the liver view from its respective full view at runtime. The idea is to construct an image pyramid using different sizes from a full view input image for scale-invariance. Liver segmentation has also been studied by Fu et al. [53]. In this work, they study the effect of adding more labelled or unlabelled data for improving segmentation tasks, particularizing in liver segmentation. They concluded that although adding more labelled data improves the segmentation, using more unlabelled data in a semi-supervised learning can achieve a comparable level of segmentation accuracy.

Cui et al. [55] particularized the surgical image recognition problem to nerve and dura mater in spinal endoscopy videos using a YOLOv3 architecture. They collected videos from 15 patients, and three senior surgeons labelled the images. Hwang et al. [57] addressed the segmentation of colonoscopic images for automatic detection of polyps using a cascaded structure of encoder-decoders. Finally, Kadkhodamohammadi et al. [56] created a custom laparoscopic sleeve gastrectomy dataset labeled with surgical instruments and the anatomical structures present in the scene. For segmentation, they propose an AE framework with the Xception architecture as the encoder and a simple feature aggregation decoder.

B. SURGICAL TASKS ANALYSIS

Surgical tasks analysis is an important task in the field of minimally invasive surgery due to its many potential applications, ranging from development of context-aware systems, automatic indexing of surgical video databases, autonomous robotic applications, etc. Within this field, surgical phases recognition has been the most studied task, as it allows a computer-assisted system to follow the workflow of a procedure. This tasks consists on dividing a procedure into a set of phases, and training the system to identify what phase corresponds to a given image. The main limitation of this task is that to be effective it requires a general consensus about the surgical procedure. Thus, it has been widely studied for cholecystectomy and gynecologic surgeries, but it has a poor generalization to other procedures.

Others authors analysis the surgical tasks in a deeper level by analyzing surgical gestures instead of phases. In this case, a particular task such as suturing is divided into a set of gestures. This is a more challenging problem as gestures are more similar to one another compared to overall phases. Until now, gestures segmentation has been analyzed only for the suturing task. Most attempts have been addressed on in-vitro environments with quite good results, and only one work [71] perform it in a live suturing.

Trajectory segmentation is another approach to deeply

TABLE 7. Comparison of the surgical tasks analysis publications using DL models.

Ref.	Year	Method	Procedure	DL model	Input data	Dataset	Results
[58]	2016	Phases recognition	Gynecology	AlexNet	Images	NPA	48.67% (acc)
[52]	2018	Phases recognition	Gynecology	GoogleNet	Images	NPA	59% (acc)
[59]	2018	Phases recognition	Gynecology	GoogleNet	Images	NPA	79.6% (AP)
[12]	2017	Phases recognition	Cholecystechtomy	EndoNet	Images	Cholec80	92% (acc)
			Colorectal surgery			EndoVis 2015	86% (acc)
[33]	2020	Phases recognition	Cholecystechtomy	VGG-50+LSTM	Images	Cholec80	89.2% (acc)
[60]	2018	Phases recognition	Cholecystechtomy	ResNet+LSTM	Images	m2cai16-workflow	90.7% (acc)
						Cholec80	92.4% (acc)
[61]	2020	Phases recognition	Cholecystechtomy	ResNet50	Images	NPA	93%
[62]	2018	Phases recognition	Cholecystechtomy	GAN+LSTM	Images	m2cai16-workflow	85.8% (acc)
[63]	2018	Phases recognition	Cholecystechtomy	ResNet-50+LSTM	Images	Cholec80	92.7% (acc)
[24]	2020	Phases recognition	Colorectal surgery	XCeption	Images	LapSig300	81% (acc)
[64]	2018	Phases recognition	Prostatectomy	InceptionV3	Images+kinematic data+events	NPA	80.9% (AP)
[65]	2019	Phases boundaries detection	Cholecystechtomy	LSTM	Images	Cholec80	48 s (MAE)
[66]	2020	Surgery type recognition	9 types of surgeries	VGG16+LSTM	Images	NPA	75% (acc)
[67]	2019	Similar frames detection	Cholecystechtomy	ResNet50	Images	Cholec80	99.1% (acc)
[68]	2019	Gestures segmentation	In-vitro experiments	3D CNN	Images	JIGSAWS	84.3% (acc)
[69]	2020	Gestures segmentation	In-vitro experiments	Deep RL	Images	JIGSAWS	81.7% (acc)
[70]	2020	Gestures segmentation	In-vitro experiments	VGG16	Images+kinematic data	JIGSAWS	86.3% (acc)
					Images+kinematic data+events	NPA	89.4% (acc)
[71]	2020	Gestures identification	Live suturing tasks	AlexNet+LSTM	Images	NPA	81% (acc)
		Gestures segmentation					63% (acc)
[72]	2020	Gestures segmentation	Gynecology	Shallow CNN	Images	LapGyn4	99.2% (acc)
[73]	2020	Fine-grained activities	Cholecystechtomy	ResNet-18	Images	Cholec80	24.8% (acc)
[74]	2019	Objects state detection	In-vitro experiments	VGG16	Images	NPA	89% (IoU)
[75]	2016	Trajectory segmentation	In-vitro experiments	AlexNet+ VGGNet	Images+kinematic data	JIGSAWS	0.806 (NMI)
[76]	2018	Trajectory segmentation	In-vitro experiments	Dense CNN	Images+kinematic data	JIGSAWS	70.6% (mAP)
[77]	2018	Trajectory segmentation	In-vitro experiments	Stacked AE	Images+kinematic data	JIGSAWS	79.1% (mAP)
[78]	2019	Trajectory segmentation	In-vitro experiments	RNN	kinematic data	JIGSAWS	71% (acc)
[79]	2017	Trajectory segmentation	In-vitro experiments	VGG16+LSTM	Images+kinematic data	JIGSAWS	
[17]	2017	Surgery time prediction	Cholecystechtomy	ResNet-152	Images	Cholec120	460 s (MAE)
[80]	2019	Surgery time prediction	Cholecystechtomy	ResNet-152+LSTM	Images	Cholec120	
[81]	2020	Anticipating instruments usage	Cholecystechtomy		Images	Cholec80	

* When more than one result is presented in the study, the one with the best performance is reported in this table.
** NR = not reported; AP = average precision; mAP = mean average precision. NMI = normalized mutual information; MAE = mean aboslute error (in seconds); NPA = not publicly available.

analysis the motion of the surgical instruments. It consists on splitting trajectories into sub-trajectories. This task can facilitate learning from demonstration, skill assessment, phase recognition, etc. To perform trajectory segmentation, authors leverage the kinematics information provided by surgical robots, which merged with video data provide better accuracy results. Finally, surgery time estimation is another important task that may be useful for optimizing clinical resources or to predict instruments usage for context-aware assistance.

Technical characteristics of the works included in this survey performing analysis of surgical tasks using DL models are shown in Table 7. Next, these works are further described.

1) Surgical phases recognition

Petscharnig and Schöffmann [58] explored the single-frame model for semantic surgery shot classification in gynecologic surgery videos manually annotated with 14 semantic classes. In [52], the authors extended the previous work with anatomical structure annotations and with more images, improving the performance with respect to the previous work. In later studies [59], they investigated the impact of early and late fusion of temporal information in surgical phases classification. Early fusion refers to extracting motion information from two consecutive video frames, and fuses it to the RGB image. With this approach, they outperforms previous work by more than 10%.

Twinanda et al. [12] presented a novel CNN architecture, called EndoNet, that performs phase recognition and tool presence detection in a multi-task manner using only visual information. EndoNet is an extension of the AlexNet architecture, in which the last layer is connected to a fully-connected layer which performs tool detection. The output of this layer is then concatenated with the output of the AlexNet to extract visual features from the images. Then, these features are used to estimate the current phase using Support Vector Machine and Hierarchical Hidden Markov Models. The authors validated this approach with the Cholec80 dataset, and they demonstrated the generalization of the results with the EndoVis workflow challenge at MICCAI 2015. Jin et al. [33] also exploit the relatedness between

tools detection and phase recognition using a multi-task deep learning network. The proposal is an end-to-end architectures with two branches: a CNN module for tools detection and a RNN for phase recognition. They designed a correlation loss to model the relatedness between this two tasks and to minimize the divergence of the predictions of the two branches. The source code is available at [17].

Another custom framework is presented in [60]. This network, called SV-RCNet, merges visual and temporal information in an end-to-end architecture using a ResNet-50 architecture to extract visual features and a LSTM network to model the temporal information of sequential frames. They demonstrated the effectiveness of the spatio-temporal joint learning versus separate training. They performed experiments on two surgical datasets, m2cai16-workflow and Cholec80. Segmentation of cholecystectomy is also addressed in [61] using a ResNet550 with a high accuracy. However, in this work they use a dataset annotated with only 4 phases, two of them out-of-the-body (preparation and trocar placement), which simplifies the problem compared with m2cai16-workflow and Cholec80 datasets.

The previous supervised methods requires large amount of annotated data. Chen et al. [62] proposed a semi-supervised method based on a spatio-temporal CNN. First, they use a Generative Adversarial Network (GAN) to extract spatial features from the images. Then, they use a LSTM network to distinguish frames based on their temporal context. Finally, they use a semi-supervised learning method to integrate the spatial and temporal information to fine-tune the network. Funke et al. [63] proposed a self-supervised method that uses a ResNet-50 CNN initialized with the general database ImageNet and fine-tuned with unlabelled videos of laparoscopic surgery using temporal coherence. They achieved better results compared to non-pretrained networks. Their work is available at GitLab[18]. A self-supervised method is also proposed by Chittajallu et al. [67], but applied to extraction of video content descriptors to find similar segments in laparoscopic videos. In the medical domain, recordings of surgical procedures are commonly used off-line for teaching inexperienced surgeons or to check and learn from errors that occurred during the interventions. But manual checking of particular events is very cumbersome and time-consuming. Thus, these authors propose to train a ResNet50 model to extract semantic image descriptors to facilitate the searching in large databases.

Previous works can successfully perform surgical phases classification at a frame level. However, they are not able to explicitly determine the transition time (frame) between two consecutive phases. Namazi et al. [65] proposed a deep learning method to detect the transition time of different phases by learning the beginning and end frames of each phase. Kannan et al. [66] address the problem of video classification for the early recognition of the type of surgery using a CNN+LSTM architecture. The CNN captures spatial information within the video frames, while the LSTM captures temporal information related to the evolution of the surgery. As the aim of this work is early recognition, during the training, the CNN is fed with samples of the complete video, emphasizing the initial frames with higher weights. They introduce a novel framework with a teacher LSTM model to predict future events, which improves the early recognition performance. They used the Laparo425 dataset, consisting of 425 videos of 9 types of laparoscopic surgeries performed at the University Hospital of Strasbourg/IHU. They obtained an accuracy of 75% after 10 minutes of the surgery.

2) Gestures segmentation

Gao et al. [69] address the problem of gesture recognition in surgical videos as a path searching problem, proposing a framework based on Deep Reinforcement Learning (RL) and tree search, taking advantage of predictions of future frames to make decisions on the current time step. Funke et al. [68] addressed the same problem with better results, but proposing a 3D CNN to learn spatio-temporal features from consecutive video frames. They demonstrated the superiority of this approach compared to 2D CNNs on JIGSAWS dataset. Source code of this work can be accessed at GitHub[19]. Khatibi et al. [72] evaluated a shallow CNN for performing surgical action recognition from single frames and multiple frames. The designed CNN can be trained faster because it requires to tune fewer parameters. They achieved the best performance using multiple frames for training the network (the first, the middle and the last video frames of each surgical frames). The main limitation of this work is that they do not consider video frames without surgical instruments in the scene.

The same problem is addressed by Qin et al. [70], but using multiple input data to train the DL model. For the JIGSAWS dataset, they augmented the accuracy score incorporating the kinematic data to the recognition. They also created a dataset, called RIOUS, that incorporates system events as an additional source to train the model. The system events include camera and instruments follow, surgeon head in/out of the console, master clutch for the hand controller, and two ultrasound probe events. They demonstrated the better performance when fusing multiple data compared to using only video data, kinematic data or video+kinematics. Luongo et al. [71] studied the problem of gestures identification (identifying when a gesture is happening) and gestures segmentation (identifying what gesture is happening) on live suturing clips. This work achieved an accuracy over 63% identifying 5 different gestures on live suturing.

In an attempt to deeply analyze the surgical actions, Nwoye et al. [73] presented a method for fine-grained surgical actions recognition based on modeling each phase as action triplets <*instrument, verb, target*> representing the tool activity. They annotated 40 videos from the Cholec80

[17]https://github.com/YuemingJin/MTRCNet-CL

[18]https://gitlab.com/nct_tso_public/pretrain_tc

[19]https://gitlab.com/nct_tso_public/surgical_gesture_recognition

dataset with 6 instruments, 8 verbs, and 19 target classes. To recognize the instruments-tissue interactions they used a multitask deep learning network with three branches (instrument, verb, and target). The performance score of this work is low, but it outperforms previous models addressing actions recognition. This shows the challenging nature of fine-grained action recognition. Peng et al. [74] proposed a method to detect the state of an object, defined as the location of the object and the interaction among objects. First, the object state is modeled as a semantic object, which contains the target object class and the interaction with other objects. Then, a DL method is applied to locate and detect these semantic objects in the image. This methodology can be applied to surgical training simulators and to other context-aware computer-assisted systems.

3) Trajectory segmentation

Murali et al. [75] presented an algorithm called Transition State Clustering with Deep Learning (TSC-DL) for surgical tasks segmentation, which is an unsupervised method that merges video and kinematic data. The key of this work is that they use pre-trained CNNs (AlexNet and VGGNet) for extracting relevant features from videos, and then they create an augmented state-space with the visual features and the kinematic data. Their results reveal that using both kinematic and visual information results in better performance over just using kinematics. The code of this work is available at GitHub[20]. Kinematic data is also used for the task of instruments trajectory segmentation. Zhao et al. [76] presented an unsupervised network for tools trajectory segmentation based on laparoscopic image and kinematic data. This work is based on a structure of dense connection, in which the first half of the network, the dense block, is an encoder that performs feature extraction, the transition layer performs the trajectory segmentation, and the up-sampling layer is used for image reconstruction. A similar approach is presented in [77], but using a compact stacking convolutional auto-encoder model and wavelet transform based filtering. Marban et al. [79] propose a method to estimate the position and velocity of the instruments in 3D from monocular videos using a regression model based on CNN+LSTM.

In contrast to previous works that merges video and kinematic data, Itzkovich et al. [78] proposed a DL model for trajectory segmentation which relies only on kinematic data, in order to use their model in online segmentation in the future. In this work, they deal with the problem of the lack of generalization of the models trained with the JIGSAWS dataset to real surgery. They demonstrated the poor generalization to rotation of the data when trained on a small and not sufficiently diversified dataset. Thus, they augmented the original dataset generating new data by rotating the images about different axes and rotation angles.

4) Surgery time prediction

Aksamentov et al. [17] proposed a deep learning pipeline with a CNN and a LSTM network to estimate the remaining duration of a surgical procedure. They connected the visual features coming from a ResNet network, pre-trained on the ImageNet dataset, to the LSTM network to extract phase information. Then, they used regression to estimate the remaining time of the surgery. Twinanda et al. [80] proposed to eliminate the need for manual annotations for the training process by training the CNN to perform progress estimation instead of surgical phase recognition, i.e., predicting how long the surgery has progressed with respect to its expected duration. At training time, the CNN architecture is fed with the video frames along with their progress label, which is automatically generated. They demonstrated the generalization of their approach using two datasets: Cholec80 and Bypass170, which contains 170 bypass videos performed by 28 surgeons.

Rivoir et al. [81] addressed the problem of the anticipation of surgical instruments by predicting the remaining time until the occurrence of sparse events rather than dense action segmentations. This way, only annotations related to instrument occurrence are required. The code of this work has been shared at GitLab[21].

C. SURGICAL SKILL ASSESSMENT

A major task in medical training is the assessment of surgical skills to grade the trainees' performance and to monitor their development during the training process. This evaluation is usually performed manually by experts, which is not only very time-consuming but also subjective and lacks consistency and reliability. To solve these problems, many authors have addressed the task of automatic skill assessment through descriptive analysis of the instruments motion, which requires high manual feature engineering, or using predictive modeling such as Hidden Markov Models, achieving high accuracy ranging from 94.4 to 100% [90]- [91]. However, these methods requires large amount of time and computational effort for tuning and modeling the parameters. In contrast, deep learning models can process raw data and can perform feature self-learning to discover abstract representations during the training process. Table 8 shows the technical characteristics of the studies performing surgical skill assessment using DL models included in this survey.

Most of the works found in the literature performing surgical skill assessment segmentation using DL models divide the levels of expertise into three categories: novice (N), intermediate (I), and expert (E). Thus, given a task performance, the DL algorithms are trained to provide the probability of the input data to belong to one of these classes. This is the case of the work developed by Fawaz el at. [82]. They designed a one dimensional CNN dedicated to surgical skill classification, and achieved very competitive results with 100% accuracy on the suturing and needle-passing tasks of the JIGSAWS

[20]https://github.com/BerkeleyAutomation/tsc-dl

[21]https://gitlab.com/nct_tso_public/ins_ant

TABLE 8. Comparison of the surgical skill assessment publications using DL models.

Ref.	Year	Method	DL model	Input data	Dataset	Results
[82]	2018	Level of expertise	CNN	Kinematic data	JIGSAWS	100% (acc)
[83]	2018	Level of expertise	CNN	Kinematic data	JIGSAWS	95.4% (acc)
[84]	2018	Level of expertise	SATR-DL	Kinematic data	JIGSAWS	96% (acc)
[85]	2019	Level of expertise	3D ConvNet	Images	JIGSAWS	95% (acc)
[86]	2019	Level of expertise	CNN+LSTM	Kinematic data	JIGSAWS	98.4% (acc)
[87]	2020	Level of expertise	CNN	Kinematic data	JIGSAWS	99.1% (acc)
[88]	2020	Detecting similar levels of expertise	SNN	Kinematic data	NPA	83.4% (acc)
[89]	2019	Pairwise ranking	LSTM	Kinematic data	JIGSAWS	75.1% (acc)
[15]	2018	Performance score (GOALS)	R-CNN (VGG16)	Images	m2cai16-tool-location	

* When more than one result is presented in the study, the one with the best performance is reported in this table.
** acc = accuracy.

dataset. Their source code is publicly available[22]. A similar approach is presented in [83], which model is able to reliably interpret skills within a 1-3 second window, without needing an observation of the entire training trial. Wang et al. [84] proposed a multi-output model, SATR-DL, for online trainee skill analysis and task recognition, achieving accuracies of 96% and 100% for these two tasks.

Other studies perform the task of automatic skill assessment using only video data. Funke et al. [85] used a 3D ConvNet achieving an accuracy of 95% on the JIGSAWS dataset. The source code of this work is available at GitLab[23]. Nguyen et al. [86] extended automatic skill assessment to open surgery procedures, using inertial measurement units to get the participants' hands motion. They achieved an accuracy of 98.2% on in-vitro experiments. They also perform experiments in the well-known robotic surgery dataset JIGSAWS to demonstrate the generalization of their approach, with competitive results. Zhang et al. [87] proposed an automatic microsurgical skill assessment for robot-assisted microsurgery based on cross-domain transfer learning. The pretrained model is obtained via the JIGSAWS dataset and then transfered for microsurgical skill assessment. The idea is to transfer the knowledge gained from JIGSAWS to accelerate learning in the new domain.

The previous works have demonstrated to be able to separate between experts, intermediate, and novices surgeons. However, it still remains to be shown if deep learning techniques are able to distinguished trainees with similar expertise from one another. In this way, Getty et al. [88] propose a Spiking Neural Network (SNN) to detect surgeons of similar level using only kinematic data. The purpose of this approach is to be able to offer adaptive assistance during surgery and training. Similarly, Oğul et al. [89] address the problem of surgical skill assessment as a pairwise ranking task in which two input actions are compared to identify the better surgical performance.

Other works addressing surgical skill assessment are based on analyzing the motion of the tools to extract key metrics for analyzing the performance of the surgeon. This allows, not only to classify the performance into a level of expertise, but to provide a performance score to a given demonstration, which is very useful objectively evaluate trainees. Jin et al. [15] developed an approach leveraging region-based convolutional neural networks (R-CNN) to perform spatial detection of tools, and then they used this information to analyze the movement of the tools. This way, they are able to extract tool usage patterns, movement range, and economy of motion metrics to analyze surgical skills. In this work, they use a modified version of the GOALS assessment rubric to provide a performance score.

D. AUTOMATION OF SURGICAL TASKS

In the last decades, much progress has been made in the automation of complete surgical tasks or the semi-automation of particular parts so robotic systems can collaborate with surgeons with specific maneuvers. A classification of the publications using DL models for surgical tasks automation is shown in Table 9.

Reinforcement learning (RL) is a popular control method in uncertain scenarios and when dealing with complex dynamics systems. The recent fusion between RL and DNNs opened a new field, known as Deep Reinforcement Learning (Deep RL), that leverages new opportunities to control non-linear systems. In the RAS domain, Thananjeyan et al. [92] employed Deep RL techniques to develop a tensioning planner for pattern cutting tasks. The input of the planner is a desired cutting contour, and then the planner selects a tensioning point and the sequence of tensioning actions as the surgical scissor follows a pre-planned trajectory. The optimal tensioning policy is learned using Deep RL, trying to minimize the error from the cutting trajectory to the marked contour. For this, they modeled the tensioning problem as a Markov Decision Process, where the actions are the movements of the tensioning arm, and the state space is a tuple consisting of the time index of the trajectory, the displacement vector from the original pinch point, and the location fiducial points of the cutting sheet. They implemented this approach using a dVRK in both simulated and physical scenarios. In later works, this method is improved with a multiple pinch point Deep RL algorithm that exhibits better results [93]- [94].

Another task that has been the focus of automation is

[22]https://germain-forestier.info/src/miccai2018/
[23]https://gitlab.com/nct_tso_public/surgical_skill_classification

TABLE 9. Classification of the surgical tasks automation publications using DL models.

Ref.	Year	Autonomous task	DL method	Experiments
[92]	2017	Tensioning planning for cutting tasks	Deep RL	Simulator and phantom experiments with the dVRK
[93]	2019	Tensioning planning for cutting tasks	Deep RL	Simulator experiments
[94]	2019	Tensioning planning for cutting tasks	Deep RL	Simulator and phantom experiments with the dVRK
[95]	2018	Autonomous surgical debridement	CNN	phantom experiments with the dVRK
[21]	2020	Tissue removal	U-Net	Phantom experiments with the dVRK
[96]	2020	Cooperative control for suturing	YOLOv3	Phantom experiments with the dVRK
[97]	2020	Autonomous palpation for tumors detection	CNN+LSTM	Phantom experiments with an UR3 robot
[98]	2016	Force estimation	Deep-Neuro-Fuzzy	Phantom experiments with a Staubli robot
[99]	2017	Force estimation	LSTM	Phantom experiments with a Staubli robot
[100]	2018	Force estimation	LSTM	Phantom experiments with a Staubli robot
[101]	2019	Force estimation	CNN+LSTM	Phantom experiments with a Staubli robot
[102]	2018	Force estimation	TCN	Ex-vivo experiments with the da Vinci Surgical System

autonomous surgical debridement. Seita et al. [95] propose the use of DNNs to automate the debridement task using a dVRK. First, the robot collects data to train a DNN by automatically exploring trajectories in the workspace with random targets. The DNN inputs are the tool position relative to the camera frame and rotations relative to the base frame, and the output is the tool position relative to the base frame. In the second phase, the robot moves to target locations, and a human directly corrects the positions to generate a small amount of high-quality data. Similarly, Attanasio et al. [21] propose a method to remove tissue from the surgical area autonomously to expose the underlying anatomical structures. First, a CNN (U-Net) is used to detect candidate tissue to be retracted, and then a planning algorithm based on experts interviews is used to perform the retracting task. They perform a set of experiments on the dVRK and release a specific dataset, named FlapNet, dedicated to retraction.

Mikada et al. [96] propose a human cooperative control for suturing in which a human operator inserts the needle with an instrument and a robot automatically pulls it out with another instrument. This method uses two CNNs, one to detect the needle, and another one to estimate its state. For autonomous robotic palpation tasks, Xio et al. [97] proposed a CNN+LSTM architecture to detect tumorous areas and to estimate the depth of the inclusion.

Force estimation based on visual data is also a problem that has been addressed using DL strategies. Aviles et al. [98] proposed using deep-neuro-fuzzy strategies for the force estimation task, which is a fusion between fuzzy systems and deep neural networks. On the one hand, the Fuzzy theory enables handling uncertainties of the visual information and allows to increase the accuracy of the DL model. Adding Fuzzy resulted in a reduction of the absolute error form 2mm to as little as 0.1025. In [99], the authors proposed a solution based on visual geometric information, using a learning system to find the nonlinear relationship between tissue deformation from the images and geometric data provided by the robot. The proposed solution starts with extracting the geometry of motion of the heart's surface by minimizing an energy functional to recover its 3D deformable structure. Then, a LSTM-RNN architecture is used to learn the relationship between the extracted visual-geometric information and the applied force, and to find accurate mapping between the two. They achieved a root-mean square error of 0.02N. Advancing in this work, Marban et al. [100] proposed a semi-supervised learning approach consisting of an encoder network serially connected with an LSTM network. First, unlabelled video sequences are used to train the encoder network to extract visual features from the images. Then, these feature vectors are used to train the LSTM network. In [101], these authors investigated different input data to the network, and concluded that force estimation is better when both video and tools data is processed.

In [102], the authors proposed a force estimation based on visual cues to infer tissue deformation, using a Temporal Convolutional Network (TCN). The input of the network are RGB and depth images collected using a Kinetic2 camera. Then, a spatial block encodes 2D and 3D features, and temporal block models force changes over time. They achieved an absolute error of 0.814N in an ex-vivo experiment using the da Vinci Surgical System.

VI. OPPORTUNITIES

Despite the complexities of minimally invasive scenarios, significant progress has been made in computer-assisted systems thanks to the advances in machine learning techniques, especially in deep learning approaches in the last years. Based on our review, we point out two promising research directions to augment the capabilities of current surgical systems and to develop new tools to benefit medical personnel and patients: the development of autonomous surgical systems and intelligent surgical training systems.

A. DEVELOPMENT OF AUTONOMOUS SURGICAL SYSTEMS

The scientific community has made huge advances in the line of extracting semantic information from surgical images to provide context-awareness in real complex scenarios. To date, recognition of surgical instruments and segmentation of some surgical tasks has been successfully addressed. Segmentation of some specific organs, such as the liver, and surgical material such as the suture thread have also been focused of study. However, there is still much to do to provide computer systems a real understanding of the

surgical scene. This is the idea of the European project *Smart Autonomous Robotic Assistant Surgeon*, which aims to go a step further developing a cognitive robotic system able to autonomously understand the present and future surgical situation to autonomously collaborate with the main surgeon supplying the functions of the assistant [103]. To achieve this goal, recognition of the complete surgical scene, including anatomical structures, and segmentation of fine-grained gestures are essential tasks that must be addressed.

Surgical scene understanding is essential to apply current strategies for performing autonomous surgical tasks to real dynamic scenarios with a priori unknown conditions and to perform tasks on-line in collaboration with humans during a real surgical intervention. Deciding when the system should take a particular action is also vital for computer-assisted tools such as virtual reality or haptic guidance systems. Recently, Islam et al. [104] proposed an enhanced graph neural network to perform spatial reasoning to infer the tool-tissue interaction graph structure in a surgical scene. Based on the scene segmentation of the MICCAI Challenge 2018 dataset, they generated a graph-based tissue-tool interaction dataset with new annotations. Moccia et al. [105] propose to use instrument segmentation to develop shared control techniques based on virtual fixtures to avoid instruments collision during surgery.

B. INTELLIGENT SURGICAL TRAINING SYSTEMS

Current methods for objective and autonomous surgical skill assessment have demonstrated to be effective to quantitative and qualitative assess a surgical performance. Thanks to these advances, commercial virtual reality training systems used for novice surgeons to get skilled in laparoscopy (either traditional or robotic) usually provide a performance report to the user with the overall score of the task. However, these systems lack the ability to assist a trainee during the performance of the exercises, which would allow to online teach the user how to perform a task or how to improve his/her performance. In this sense, there is a research opportunity in the field of surgical training systems to study approaches for autonomously coaching novices during their training process. This would reduce the long learning curve, especially in robot-assisted surgery, and would facilitate self-sufficiency of the trainees, which could make a difference in surgical training, taking into account the scarce availability of expert surgeons for teaching purposes.

Some authors are proposing new methods that go in this line. Fawaz et al. [82] use the class activation map technique to visualize which parts of a performance contribute the most to a certain skill level classification, which could serve to guide novices surgeons to improve their skills. Similarly, Zia et. al [106] propose a method for generating *task highlights* to give surgeons more direct feedback about their performance. Tan et al. [107] presented a robot-assisted laparoscopy training system for improving surgeons' skills based on experts' demonstration and reinforcement learning. This approach combines the latent patterns from experts' trajectories and objective-constrained trajectories generated by the RL agent. Engelhardt et al. [108] propose the use of conditional adversarial networks to provide a more realistic visual appearance to phantoms used for surgical training.

VII. CONCLUSION

Many scientific fields have been transformed in the latest years thanks to the advances in deep learning algorithms, including predicting movie ratings, decision to approve loan applications, time taken by car delivery, diagnostic of diseases, discovery of new drugs, prediction of natural disasters, etc. This has been possible thanks to the proliferation of cheaper and powerful processing units and the explosion of big data. Thus, in the latest years the research community has made a huge effort on publishing large annotated data of minimally invasive surgery interventions, which has boosted great advances in this area.

This work presents a rigorous systematic literature review of DL methods in the field of minimally invasive surgery. This survey has been conducted with a total of 85 publications from the last five years. After a comprehensive reading of each of them, we have classified these works into four applications: surgical image analysis, surgical task analysis, surgical skill assessment, and automation of surgical tasks. The most studied application in surgical image analysis is instruments recognition. This task can be classified into instruments classification, instruments detection and instruments segmentation. In the literature we can find many works proposing deep learning architectures that exhibit high instruments recognition performance. In addition to instruments, recognition of anatomical structures and other surgical material such as the suture thread or the needle, have also attracted the interest of researchers. Surgical task analysis is another important task in the field of minimally invasive surgery. The most studied application is surgical phases recognition for cholecystectomy and gynecologic surgeries, although there are also studies that offer a deeper analysis of the task performing gesture and trajectory segmentation. On the other hand, surgical skill assessment applications allows to classify a surgical performance into different levels of expertise, generally novice surgeons, intermediate and experts. Most of these works use kinematic data to train the networks. Finally, automation of surgical tasks has also been addressed using deep algorithms, such as deep reinforcement learning. This algorithm has been used for developing an automatic tensioning planner for pattern cutting tasks. Force estimation in robot-assisted surgery has also been addressed using deep learning architectures.

Based on the review presented in this work, we point out two promising research directions: the development of autonomous surgical systems and of intelligent surgical training systems. We believe that these research lines are key to augment the capabilities of current surgical systems and to develop new tools that will benefit medical personnel and patients.

REFERENCES

[1] Y. Kassahun, B. Yu, A. T. Tibebu, D. Stoyanov, S. Giannarou, J. H. Metzen, and E. Vander Poorten, "Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions," 4 2016.

[2] C. R. Garrow, K.-F. Kowalewski, L. Li, M. Wagner, M. W. Schmidt, S. Engelhardt, D. A. Hashimoto, H. G. Kenngott, S. Bodenstedt, S. Speidel, B. P. Müller-Stich, and F. Nickel, "Machine Learning for Surgical Phase Recognition," Annals of Surgery, 2020.

[3] R. Anteby, N. Horesh, S. Soffer, Y. Zager, Y. Barash, I. Amiel, D. Rosin, M. Gutman, and E. Klang, "Deep learning visual analysis in laparoscopic surgery: a systematic review and diagnostic test accuracy meta-analysis," 1 2021.

[4] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," IEEE Access, vol. 7, pp. 53040–53065, 2019.

[5] François Chollet, Deep Learning with Python. Manning Publications Co., 2017.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, pp. 84–90, 6 2017.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, International Conference on Learning Representations, ICLR, 9 2015.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 07-12-June, pp. 1–9, IEEE Computer Society, 10 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-Decem, pp. 770–778, IEEE Computer Society, 12 2016.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, Institute of Electrical and Electronics Engineers (IEEE), 3 2010.

[11] G. Nguyen, S. Dlugolinsky, M. Bobak, V. Tran, A. Lopez Garcia, I. Heredia, P. Malik, and L. Hluchy, "Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey," Artificial Intelligence Review, vol. 52, pp. 77–124, 6 2019.

[12] 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, pp. 86–97, 1 2017.

[13] A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov, "Automatic Instrument Segmentation in Robot-Assisted Surgery using Deep Learning," in Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018, pp. 624–628, Institute of Electrical and Electronics Engineers Inc., 1 2019.

[14] L. Maier-Hein, A. Reinke, M. Kozubek, A. L. Martel, T. Arbel, M. Eisenmann, A. Hanbury, P. Jannin, H. Müller, S. Onogur, J. Saez-Rodriguez, B. van Ginneken, A. Kopp-Schneider, and B. A. Landman, "BIAS: Transparent reporting of biomedical image analysis challenges," Medical Image Analysis, vol. 66, p. 101796, 12 2020.

[15] A. Jin, S. Yeung, J. Jopling, J. Krause, D. Azagury, A. Milstein, and L. Fei-Fei, "Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks," in Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, vol. 2018-Janua, pp. 691–699, Institute of Electrical and Electronics Engineers Inc., 5 2018.

[16] "Workshop and Challenges on Modeling and Monitoring of Computer Assisted Interventions."

[17] I. Aksamentov, A. P. Twinanda, D. Mutter, J. Marescaux, and N. Padoy, "Deep neural networks predict remaining surgery duration from cholecystectomy videos," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10434 LNCS, pp. 586–593, Springer Verlag, 9 2017.

[18] Y. Gao, S. Swaroop Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh, C. Chiung, G. Chen, R. Vidal, S. Khudanpur, and G. D. Hager, "JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling," in MICCAI Workshop: Modeling and Monitoring of Computer Assisted Interventions (M2CAI), (Boston, MA), 2014.

[19] D. Sarikaya, J. J. Corso, and K. A. Guru, "Detection and Localization of Robotic Tools in Robot-Assisted Surgery Videos Using Deep Neural Networks for Region Proposal and Detection," IEEE Transactions on Medical Imaging, vol. 36, pp. 1542–1549, 7 2017.

[20] E. Colleoni, P. Edwards, and D. Stoyanov, "Synthetic and Real Inputs for Tool Segmentation in Robotic Surgery," in International Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2020, (Lima, Peru), pp. 700–710, Springer, Cham, 10 2020.

[21] A. Attanasio, B. Scaglioni, M. Leonetti, A. F. Frangi, W. Cross, C. S. Biyani, and P. Valdastri, "Autonomous Tissue Retraction in Robotic Assisted Minimally Invasive Surgery - A Feasibility Study," IEEE Robotics and Automation Letters, vol. 5, pp. 6528–6535, 10 2020.

[22] A. Leibetseder, S. Petscharnig, M. J. Primus, S. Kletz, B. Münzer, K. Schoeffmann, and J. Keckstein, "LapGyn4: A dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology," in Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, vol. 18, (New York, NY, USA), pp. 357–362, Association for Computing Machinery, Inc, 6 2018.

[23] K. Schoeffmann, H. Husslein, S. Kletz, S. Petscharnig, B. Muenzer, and C. Beecks, "Video retrieval in laparoscopic video recordings with dynamic content descriptors," Multimedia Tools and Applications, vol. 77, pp. 16813–16832, 7 2018.

[24] D. Kitaguchi, N. Takeshita, H. Matsuzaki, T. Oda, M. Watanabe, K. Mori, E. Kobayashi, and M. Ito, "Automated laparoscopic colorectal surgery workflow recognition using artificial intelligence: Experimental research," International Journal of Surgery, vol. 79, pp. 88–94, 7 2020.

[25] S. Wang, A. Raju, and J. Huang, "Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos," Proceedings - International Symposium on Biomedical Imaging, pp. 620–623, 6 2017.

[26] M. Islam, D. A. Atputharuban, R. Ramesh, and H. Ren, "Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning," IEEE Robotics and Automation Letters, vol. 4, pp. 2188–2195, 4 2019.

[27] Y. Hu, Y. Gu, J. Yang, and G. Z. Yang, "Multi-stage suture detection for robot assisted anastomosis based on deep learning," in Proceedings - IEEE International Conference on Robotics and Automation, pp. 4826–4833, Institute of Electrical and Electronics Engineers Inc., 9 2018.

[28] B. Lu, X. B. Yu, J. W. Lai, K. C. Huang, K. C. Chan, and H. K. Chu, "A Learning Approach for Suture Thread Detection with Feature Enhancement and Segmentation for 3-D Shape Reconstruction," IEEE Transactions on Automation Science and Engineering, vol. 17, pp. 858–870, 4 2020.

[29] K. Mishra, R. Sathish, and D. Sheet, "Learning Latent Temporal Connectionism of Deep Residual Visual Abstractions for Identifying Surgical Tools in Laparoscopy Procedures," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, vol. 2017-July, pp. 2233–2240, IEEE Computer Society, 8 2017.

[30] H. Al Hajj, M. Lamard, P. H. Conze, B. Cochener, and G. Quellec, "Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks," Medical Image Analysis, vol. 47, pp. 203–218, 7 2018.

[31] S. Wang, Z. Xu, C. Yan, and J. Huang, "Graph Convolutional Nets for Tool Presence Detection in Surgical Videos," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11492 LNCS, pp. 467–478, Springer Verlag, 6 2019.

[32] T. Kurmann, P. Marquez Neila, X. Du, P. Fua, D. Stoyanov, S. Wolf, and R. Sznitman, "Simultaneous recognition and pose estimation of instruments in minimally invasive surgery," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10434 LNCS, pp. 505–513, Springer Verlag, 9 2017.

[33] 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, 1 2020.

[34] X. Du, T. Kurmann, P. L. Chang, M. Allan, S. Ourselin, R. Sznitman, J. D. Kelly, and D. Stoyanov, "Articulated multi-instrument 2-d pose estimation using fully convolutional networks," IEEE Transactions on Medical Imaging, vol. 37, pp. 1276–1287, 5 2018.

[35] E. Colleoni, S. Moccia, X. Du, E. De Momi, and D. Stoyanov, "Deep Learning Based Robotic Tool Detection and Articulation Estimation with Spatio-Temporal Layers," IEEE Robotics and Automation Letters, vol. 4, pp. 2714–2721, 7 2019.

[36] Z. Chen, Z. Zhao, and X. Cheng, "Surgical instruments tracking based on deep learning with lines detection and spatio-temporal context," in Proceedings - 2017 Chinese Automation Congress, CAC 2017, vol. 2017-Janua, pp. 2711–2714, Institute of Electrical and Electronics Engineers Inc., 12 2017.

[37] Z. Zhao, T. Cai, F. Chang, and X. Cheng, "Real-time surgical instrument detection in robot-assisted surgery using a convolutional neural network cascade," Healthcare Technology Letters, vol. 6, no. 6, pp. 275–279, 2019.

[38] Y. Liu, Z. Zhao, F. Chang, and S. Hu, "An anchor-free convolutional neural network for real-time surgical tool detection in robot-assisted surgery," IEEE Access, vol. 8, pp. 78193–78201, 2020.

[39] L. Yu, P. Wang, Y. Yan, Y. Xia, and W. Cao, "MASSD: Multi-scale attention single shot detector for surgical instruments," Computers in Biology and Medicine, vol. 123, p. 103867, 8 2020.

[40] I. Laina, N. Rieke, C. Rupprecht, J. P. Vizcaíno, A. Eslami, F. Tombari, and N. Navab, "Concurrent segmentation and localization for tracking of surgical instruments," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10434 LNCS, pp. 664–672, Springer Verlag, 9 2017.

[41] L. C. Garcia-Peraza-Herrera, W. Li, L. Fidon, C. Gruijthuijsen, A. Devreker, G. Attilakos, J. Deprest, E. V. Poorten, D. Stoyanov, T. Vercauteren, and S. Ourselin, "ToolNet: Holistically-nested real-time segmentation of robotic surgical tools," in IEEE International Conference on Intelligent Robots and Systems, vol. 2017-Septe, pp. 5717–5722, Institute of Electrical and Electronics Engineers Inc., 12 2017.

[42] F. Milletari, N. Rieke, M. Baust, M. Esposito, and N. Navab, "CFCM: Segmentation via Coarse to Fine Context Memory," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11073 LNCS, pp. 667–674, Springer Verlag, 9 2018.

[43] D. Pakhomov and N. Navab, "Searching for Efficient Architecture for Instrument Segmentation in Robotic Surgery," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pp. 648–656, Springer, Cham, 10 2020.

[44] S. M. Kamrul Hasan and C. A. Linte, "U-NetPlus: A Modified Encoder-Decoder U-Net Architecture for Semantic and Instance Segmentation of Surgical Instruments from Laparoscopic Images," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, pp. 7205–7211, Institute of Electrical and Electronics Engineers Inc., 7 2019.

[45] Z.-L. Ni, G.-B. Bian, Z.-G. Hou, X.-H. Zhou, X.-L. Xie, and Z. Li, "Attention-Guided Lightweight Network for Real-Time Segmentation of Robotic Surgical Instruments," in 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 9939–9945, Institute of Electrical and Electronics Engineers (IEEE), 9 2020.

[46] M. Islam, V. VS, C. M. Lim, and H. Ren, "ST-MTL: Spatio-Temporal Multitask Learning Model to Predict Scanpath While Tracking Instruments in Robotic Surgery," Medical Image Analysis, p. 101837, 10 2020.

[47] D. Liu, Y. Wei, T. Jiang, Y. Wang, R. Miao, F. Shan, and Z. Li, "Unsupervised Surgical Instrument Segmentation via Anchor Generation and Semantic Diffusion," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pp. 657–667, Springer, Cham, 10 2020.

[48] S. Kletz, K. Schoeffmann, J. Benois-Pineau, and H. Husslein, "Identifying Surgical Instruments in Laparoscopy Using Deep Learning Instance Segmentation," in Proceedings - International Workshop on Content-Based Multimedia Indexing, vol. 2019-Septe, IEEE Computer Society, 9 2019.

[49] E. J. Lee, W. Plishker, X. Liu, S. S. Bhattacharyya, and R. Shekhar, "Weakly supervised segmentation for real-time surgical tool tracking," Healthcare Technology Letters, vol. 6, no. 6, pp. 231–236, 2019.

[50] M. Sahu, R. Strömsdörfer, A. Mukhopadhyay, and S. Zachow, "Endo-Sim2Real: Consistency Learning-Based Domain Adaptation for Instrument Segmentation," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pp. 784–794, Springer, Cham, 10 2020.

[51] M. Attia, M. Hossny, S. Nahavandi, and H. Asadi, "Surgical tool segmentation using a hybrid deep CNN-RNN auto encoder-decoder," in 2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2017, vol. 2017-Janua, pp. 3373–3378, Institute of Electrical and Electronics Engineers Inc., 11 2017.

[52] S. Petscharnig and K. Schöffmann, "Learning laparoscopic video shot classification for gynecological surgery," Multimedia Tools and Applications, vol. 77, pp. 8061–8079, 4 2018.

[53] Y. Fu, M. R. Robu, B. Koo, C. Schneider, S. van Laarhoven, D. Stoyanov, B. Davidson, M. J. Clarkson, and Y. Hu, "More unlabelled data or label more data? a study on semi-supervised laparoscopic image segmentation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11795 LNCS, pp. 173–180, Springer, 10 2019.

[54] A. Nazir, M. N. Cheema, B. Sheng, P. Li, H. Li, P. Yang, Y. Jung, J. Qin, and D. D. Feng, "SPST-CNN: Spatial pyramid based searching and tagging of liver's intraoperative live views via CNN for minimal invasive surgery," Journal of biomedical informatics, vol. 106, p. 103430, 6 2020.

[55] P. Cui, Z. Guo, J. Xu, T. Li, Y. Shi, W. Chen, T. Shu, and J. Lei, "Tissue Recognition in Spinal Endoscopic Surgery Using Deep Learning," in 2019 IEEE 10th International Conference on Awareness Science and Technology, iCAST 2019 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 10 2019.

[56] A. Kadkhodamohammadi, I. Luengo, S. Barbarisi, H. Taleb, E. Flouty, and D. Stoyanov, "Feature Aggregation Decoder for Segmenting Laparoscopic Scenes," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11796 LNCS, pp. 3–11, Springer, 10 2019.

[57] M. Hwang, D. Wang, W. C. Jiang, X. Pan, D. Fu, K. S. Hwang, and K. Ding, "An Adaptive Regularization Approach to Colonoscopic Polyp Detection Using a Cascaded Structure of Encoder–Decoders," International Journal of Fuzzy Systems, vol. 21, pp. 2091–2101, 10 2019.

[58] S. Petscharnig and K. Schöffmann, "Deep learning for shot classification in gynecologic surgery videos," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10132 LNCS, pp. 702–713, Springer Verlag, 2017.

[59] S. Petscharnig, K. Schoffmann, J. Benois-Pineau, S. Chaabouni, and J. Keckstein, "Early and Late Fusion of Temporal Information for Classification of Surgical Actions in Laparoscopic Gynecology," in Proceedings - IEEE Symposium on Computer-Based Medical Systems, vol. 2018-June, pp. 369–374, Institute of Electrical and Electronics Engineers Inc., 7 2018.

[60] Y. Jin, Q. Dou, H. Chen, L. Yu, J. Qin, C. W. Fu, and P. A. Heng, "SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network," IEEE Transactions on Medical Imaging, vol. 37, pp. 1114–1126, 5 2018.

[61] E. Kurian, J. J. Kizhakethottam, and J. Mathew, "Deep learning based Surgical Workflow Recognition from Laparoscopic Videos," in Proceedings of the Fifth International Conference on Communication and Electronics Systems (ICCES 2020), pp. 928–931, Institute of Electrical and Electronics Engineers (IEEE), 7 2020.

[62] Y. Chen, Q. L. Sun, and K. Zhong, "Semi-supervised spatio-temporal CNN for recognition of surgical workflow," Eurasip Journal on Image and Video Processing, vol. 2018, pp. 1–9, 12 2018.

[63] 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 Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11041 LNCS, pp. 85–93, Springer Verlag, 9 2018.

[64] A. Zia, A. Hung, I. Essa, and A. Jarc, "Surgical Activity Recognition in Robot-Assisted Radical Prostatectomy Using Deep Learning," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11073 LNCS, pp. 273–280, Springer Verlag, 9 2018.

[65] B. Namazi, G. Sankaranarayanan, and V. Devarajan, "Attention-based surgical phase boundaries detection in laparoscopic videos," in Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019, pp. 577–583, Institute of Electrical and Electronics Engineers Inc., 12 2019.

[66] S. Kannan, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "Future-State Predicting LSTM for Early Surgery Type Recognition," IEEE Transactions on Medical Imaging, vol. 39, pp. 556–566, 3 2020.

[67] D. R. Chittajallu, B. Dong, P. Tunison, R. Collins, K. Wells, J. Fleshman, G. Sankaranarayanan, S. Schwaitzberg, L. Cavuoto, and A. Enquobahrie,

"XAI-CBIR: Explainable ai system for content based retrieval of video frames from minimally invasive surgery videos," in Proceedings - International Symposium on Biomedical Imaging, vol. 2019-April, pp. 66–69, IEEE Computer Society, 4 2019.

[68] I. Funke, S. Bodenstedt, F. Oehme, F. von Bechtolsheim, J. Weitz, and S. Speidel, "Using 3D Convolutional Neural Networks to Learn Spatiotemporal Features for Automatic Surgical Gesture Recognition in Video," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11768 LNCS, pp. 467–475, Springer, 10 2019.

[69] X. Gao, Y. Jin, Q. Dou, and P.-A. Heng, "Automatic Gesture Recognition in Robot-assisted Surgery with Reinforcement Learning and Tree Search," in 2020 IEEE International Conference on Robotics and Automation (ICRA), (IE63), pp. 8440–8446, Institute of Electrical and Electronics Engineers (IEEE), 9 2020.

[70] Y. Qin, S. A. Pedram, S. Feyzabadi, M. Allan, A. J. McLeod, J. W. Burdick, and M. Azizian, "Temporal Segmentation of Surgical Sub-tasks through Deep Learning with Multiple Data Sources," in 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 371–377, Institute of Electrical and Electronics Engineers (IEEE), 9 2020.

[71] F. Luongo, R. Hakim, J. H. Nguyen, A. Anandkumar, and A. J. Hung, "Deep learning-based computer vision to recognize and classify suturing gestures in robot-assisted surgery," Surgery, 9 2020.

[72] T. Khatibi and P. Dezyani, "Proposing novel methods for gynecologic surgical action recognition on laparoscopic videos," Multimedia Tools and Applications, vol. 79, pp. 30111–30133, 8 2020.

[73] C. I. Nwoye, C. Gonzalez, T. Yu, P. Mascagni, D. Mutter, J. Marescaux, and N. Padoy, "Recognition of Instrument-Tissue Interactions in Endoscopic Videos via Action Triplets," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pp. 364–374, Springer, Cham, 10 2020.

[74] K. S. Peng, M. Hong, J. Rozenblit, and A. J. Hamilton, "Single shot state detection in simulation-based laparoscopy training," Simulation Series, vol. 51, no. 5, 2019.

[75] A. Murali, A. Garg, S. Krishnan, F. T. Pokorny, P. Abbeel, T. Darrell, and K. Goldberg, "TSC-DL: Unsupervised trajectory segmentation of multi-modal surgical demonstrations with Deep Learning," Proceedings - IEEE International Conference on Robotics and Automation, vol. 2016-June, pp. 4150–4157, 6 2016.

[76] H. Zhao, J. Xie, Z. Shao, Y. Qu, Y. Guan, and J. Tan, "A fast unsupervised approach for multi-modality surgical trajectory segmentation," IEEE Access, vol. 6, pp. 56411–56422, 2018.

[77] Z. Shao, H. Zhao, J. Xie, Y. Qu, Y. Guan, and J. Tan, "Unsupervised Trajectory Segmentation and Promoting of Multi-Modal Surgical Demonstrations," in IEEE International Conference on Intelligent Robots and Systems, pp. 777–782, Institute of Electrical and Electronics Engineers Inc., 12 2018.

[78] D. Itzkovich, Y. Sharon, A. Jarc, Y. Refaely, and I. Nisky, "Using augmentation to improve the robustness to rotation of deep learning segmentation in robotic-assisted surgical data," in Proceedings - IEEE International Conference on Robotics and Automation, vol. 2019-May, pp. 5068–5075, Institute of Electrical and Electronics Engineers Inc., 5 2019.

[79] A. Marban, V. Srinivasan, W. Samek, J. Fernández, and A. Casals, "Estimating Position & Velocity in 3D Space from Monocular Video Sequences Using a Deep Neural Network," in Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017, vol. 2018-Janua, pp. 1460–1469, Institute of Electrical and Electronics Engineers Inc., 7 2017.

[80] A. P. Twinanda, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "RSDNet: Learning to Predict Remaining Surgery Duration from Laparoscopic Videos Without Manual Annotations," IEEE Transactions on Medical Imaging, vol. 38, pp. 1069–1078, 4 2019.

[81] D. Rivoir, S. Bodenstedt, I. Funke, F. von Bechtolsheim, M. Distler, J. Weitz, and S. Speidel, "Rethinking Anticipation Tasks: Uncertainty-Aware Anticipation of Sparse Surgical Instrument Usage for Context-Aware Assistance," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pp. 752–762, Springer, Cham, 10 2020.

[82] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. A. Muller, "Evaluating Surgical Skills from Kinematic Data Using Convolutional Neural Networks," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11073 LNCS, pp. 214–221, Springer Verlag, 9 2018.

[83] Z. Wang and A. Majewicz Fey, "Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery," International Journal of Computer Assisted Radiology and Surgery, vol. 13, pp. 1959–1970, 12 2018.

[84] Z. Wang and A. M. Fey, "SATR-DL: Improving Surgical Skill Assessment and Task Recognition in Robot-Assisted Surgery with Deep Neural Networks," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, vol. 2018-July, pp. 1793–1796, Institute of Electrical and Electronics Engineers Inc., 10 2018.

[85] I. Funke, S. T. Mees, J. Weitz, and S. Speidel, "Video-based surgical skill assessment using 3D convolutional neural networks," International Journal of Computer Assisted Radiology and Surgery, vol. 14, pp. 1217–1225, 7 2019.

[86] X. A. Nguyen, D. Ljuhar, M. Pacilli, R. M. Nataraja, and S. Chauhan, "Surgical skill levels: Classification and analysis using deep neural network model and motion signals," Computer Methods and Programs in Biomedicine, vol. 177, pp. 1–8, 8 2019.

[87] D. Zhang, Z. Wu, J. Chen, A. Gao, X. Chen, P. Li, Z. Wang, G. Yang, B. Lo, and G. Z. Yang, "Automatic Microsurgical Skill Assessment Based on Cross-Domain Transfer Learning," IEEE Robotics and Automation Letters, vol. 5, pp. 4148–4155, 7 2020.

[88] N. Getty, Z. Zhao, S. Gruessner, L. Chen, and F. Xia, "Recurrent and Spiking Modeling of Sparse Surgical Kinematics," in ACM International Conference Proceeding Series, (New York, NY, USA), pp. 1–5, Association for Computing Machinery, 7 2020.

[89] B. B. Oğul, M. F. Gilgien, and P. D. Şahin, "Ranking robot-assisted surgery skills using kinematic sensors," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11912 LNCS, pp. 330–336, Springer, 11 2019.

[90] C. E. Reiley and G. D. Hager, "Task versus subtask surgical skill evaluation of robotic minimally invasive surgery," in International Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2009, pp. 435–442, Springer, Berlin, Heidelberg, 2009.

[91] L. Tao, E. Elhamifar, S. Khudanpur, G. D. Hager, and R. Vidal, "Sparse hidden Markov models for surgical gesture classification and skill evaluation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7330 LNCS, pp. 167–177, Springer, Berlin, Heidelberg, 2012.

[92] B. Thananjeyan, A. Garg, S. Krishnan, C. Chen, L. Miller, and K. Goldberg, "Multilateral surgical pattern cutting in 2D orthotropic gauze with deep reinforcement learning policies for tensioning," in Proceedings - IEEE International Conference on Robotics and Automation, pp. 2371–2378, Institute of Electrical and Electronics Engineers Inc., 7 2017.

[93] T. Nguyen, N. D. Nguyen, F. Bello, and S. Nahavandi, "A new tensioning method using deep reinforcement learning for surgical pattern cutting," in Proceedings of the IEEE International Conference on Industrial Technology, vol. 2019-Febru, pp. 1339–1344, Institute of Electrical and Electronics Engineers Inc., 2 2019.

[94] N. D. Nguyen, T. Nguyen, S. Nahavandi, A. Bhatti, and G. Guest, "Manipulating soft tissues by deep reinforcement learning for autonomous robotic surgery," in SysCon 2019 - 13th Annual IEEE International Systems Conference, Proceedings, Institute of Electrical and Electronics Engineers Inc., 4 2019.

[95] D. Seita, S. Krishnan, R. Fox, S. McKinley, J. Canny, and K. Goldberg, "Fast and Reliable Autonomous Surgical Debridement with Cable-Driven Robots Using a Two-Phase Calibration Procedure," in Proceedings - IEEE International Conference on Robotics and Automation, pp. 6651–6658, Institute of Electrical and Electronics Engineers Inc., 9 2018.

[96] T. Mikada, T. Kanno, T. Kawase, T. Miyazaki, and K. Kawashima, "Suturing Support by Human Cooperative Robot Control Using Deep Learning," IEEE Access, vol. 8, pp. 167739–167746, 9 2020.

[97] B. Xiao, W. Xu, J. Guo, H.-K. Lam, G. Jia, W. Hong, and H. Ren, "Depth Estimation of Hard Inclusions in Soft Tissue by Autonomous Robotic Palpation Using Deep Recurrent Neural Network," IEEE Transactions on Automation Science and Engineering, pp. 1–9, 3 2020.

[98] A. I. Aviles, S. M. Alsaleh, E. Montseny, P. Sobrevilla, and A. Casals, "A deep-neuro-fuzzy approach for estimating the interaction forces in robotic surgery," in 2016 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2016, pp. 1113–1119, Institute of Electrical and Electronics Engineers Inc., 11 2016.

[99] A. I. Aviles, S. M. Alsaleh, J. K. Hahn, and A. Casals, "Towards Retrieving Force Feedback in Robotic-Assisted Surgery: A Supervised Neuro-Recurrent-Vision Approach," IEEE Transactions on Haptics, vol. 10, pp. 431–443, 7 2017.
[100] A. Marban, V. Srinivasan, W. Samek, J. Fernandez, and A. Casals, "Estimation of Interaction Forces in Robotic Surgery using a Semi-Supervised Deep Neural Network Model," in IEEE International Conference on Intelligent Robots and Systems, pp. 761–768, Institute of Electrical and Electronics Engineers Inc., 12 2018.
[101] A. Marban, V. Srinivasan, W. Samek, J. Fernández, and A. Casals, "A recurrent convolutional neural network approach for sensorless force estimation in robotic surgery," Biomedical Signal Processing and Control, vol. 50, pp. 134–150, 4 2019.
[102] C. Gao, X. Liu, M. Peven, M. Unberath, and A. Reiter, "Learning to see forces: surgical force prediction with RGB-point cloud temporal convolutional networks," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11041 LNCS, pp. 118–127, Springer Verlag, 9 2018.
[103] F. Setti, E. Oleari, A. Leporini, D. Trojaniello, A. Sanna, U. Capitanio, F. Montorsi, A. Salonia, and R. Muradore, "A Multirobots Teleoperated Platform for Artificial Intelligence Training Data Collection in Minimally Invasive Surgery," in 2019 International Symposium on Medical Robotics, ISMR 2019, Institute of Electrical and Electronics Engineers Inc., 5 2019.
[104] M. Islam, L. Seenivasan, L. C. Ming, and H. Ren, "Learning and Reasoning with the Graph Structure Representation in Robotic Surgery," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pp. 627–636, Springer, Cham, 10 2020.
[105] R. Moccia, C. Iacono, B. Siciliano, and F. Ficuciello, "Vision-Based Dynamic Virtual Fixtures for Tools Collision Avoidance in Robotic Surgery," IEEE Robotics and Automation Letters, vol. 5, pp. 1650–1655, 4 2020.
[106] A. Zia and I. Essa, "Automated surgical skill assessment in RMIS training," International Journal of Computer Assisted Radiology and Surgery, vol. 13, pp. 731–739, 5 2018.
[107] X. Tan, C. B. Chng, Y. Su, K. B. Lim, and C. K. Chui, "Robot-Assisted Training in Laparoscopy Using Deep Reinforcement Learning," IEEE Robotics and Automation Letters, vol. 4, pp. 485–492, 4 2019.
[108] S. Engelhardt, L. Sharan, M. Karck, R. D. Simone, and I. Wolf, "Cross-Domain Conditional Generative Adversarial Networks for Stereoscopic Hyperrealism in Surgical Training," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11768 LNCS, pp. 155–163, Springer, 10 2019.

IRENE RIVAS-BLANCO received her M.S. degree in industrial engineering and her Ph.D. degree in mechatronics from the University of Malaga, Spain, in 2010 and 2017, respectively. In 2011, she joined the Medical Robotics Research Group of the University of Malaga, where she has participated in a number of projects. In 2013, she was a Visiting Researcher at The Biorobotics Institute of the Scuola Superiore Sant'Anna. Her main research interests include robotic systems for minimally invasive surgery and cognitive strategies for medical robotics. She has more than 20 publications on these topics, including journal papers, conference papers, and book chapters.

CARLOS J. PÉREZ-DEL-PULGAR received his M.Sc. and Ph.D. degrees in computer science from the University of Malaga, Malaga, Spain, in 2004 and 2016, respectively. In 2004, he was given a permanent position on the research support staff at the University of Malaga. Additionally, since 2010, he has been a Part-Time Assistant Lecturer in the Electrical Engineering Faculty, where he is responsible for various subjects related to automation and robotics. In 2014, he was a Visiting Researcher in the Telerobotics and Haptics Laboratory at the European Space Agency, European Space Research and Technology Centre. His research interests include machine learning, robotics, haptics, control, and automation. He has more than 20 publications on these topics and has been involved in more than ten Spanish and European projects.

ISABEL GARCÍA-MORALES received the M.S. and Ph.D. degrees in industrial electrical engineering from the University of Malaga in 2000 and 2006, respectively. She is currently an Associate Professor and is responsible for a variety of subjects related to robotics. She has authored or coauthored more than 50 journal articles, conference papers, and book chapters and has been involved in more than ten Spanish and European projects. Her research interests include process automation, control techniques, collaborative robotics, and surgical robotics.

VÍCTOR F. MUÑOZ (M'94) received his M.Sc. degree and his Ph.D. degree in computer science from the University of Malaga, Spain, in 1990 and 1995, respectively. Through a research fellowship, he joined the Research Group of Systems Engineering and Automation, University of Malaga, where he began his research in 1991 on the problem of mobile robot navigation. In December 1996, he became an Associate Professor at the University of Malaga. In 1997, he started his research on the application of robots in surgery. He completed this research in 2004 with the construction and use of a minimally invasive surgery robot assistant in human clinical trials. He is currently a Full Professor in the Faculty of Electrical Engineering, University of Malaga.