Deep Learning Based Image Processing for Robot Assisted Surgery: A Systematic Literature Survey

SARDAR MEHBOOB HUSSAIN1, ANTONIO BRUNETTI1,2, GIUSEPPE LUCARELLI3, RICCARDO MEMEO4, VITOANTONIO BEVILACQUA1,2, AND DOMENICO BUONGIORNO1,2

1Department of Electrical and Information Engineering (DEI), Polytechnic University of Bari, 70126 Bari, Italy
2Apulian Bioengineering s.r.l., 70026 Modugno, Italy
3Urology, Andrology and Kidney Transplantation Unit, Department of Emergency and Organ Transplantation, University of Bari, 70121 Bari, Italy
4Department of Hepatobiliary and Pancreatic Surgery, Miulli Hospital, 70021 Bari, Italy

Corresponding author: Vitoantonio Bevilacqua (vitoantonio.bevilacqua@poliba.it)

ABSTRACT The recent advancements in the surging field of Deep Learning (DL) have revolutionized every sphere of life, and the healthcare domain is no exception. The enormous success of DL models, particularly with image data, has led to the development of image-guided Robot Assisted Surgery (RAS) systems. By and large, the number of studies concerning image-driven computer assisted surgical systems using DL has increased exponentially. Additionally, the contemporary availability of surgical datasets has also boosted the DL applications in RAS. Inspired by the latest trends and contributions in surgery, this literature survey presents a summarized analysis of recent innovations of DL in image-guided RAS systems. After a thorough review, a sum of 184 articles are selected and grouped into four categories, based on the literature and the relevancy of the task in the articles, comprising 1) Surgical Tools, 2) Surgical Processes, 3) Surgical Surveillance, and 4) Surgical Performance. The survey also discusses publicly available surgical datasets and highlights the basics of the DL models. Furthermore, the legal, ethical, and technological challenges together with the intuitive predictions and recommendations related to the autonomous RAS systems are also presented. The study reveals that Convolutional Neural Network (CNN) is most widely adopted architecture, whereas, the JIGSAWS is most employed dataset in RAS. The study suggests fusing kinematic data along with image data, which produces better accuracy and precision, particularly in gesture and trajectory segmentation tasks. Additionally, CNN and long short term memory networks have shown remarkable performance, however, authors recommend employing these gigantic architectures only when simpler models have failed to produce satisfactory results. The simpler models, despite their limitations, are time and cost effective and yield considerable outcomes even on the smaller datasets.

INDEX TERMS Deep learning, convolutional neural network, minimally invasive surgery, computer-assisted intervention, robotic surgery, robot-assisted surgery, image-guided surgery, surgical robotics, healthcare.

I. INTRODUCTION

The introduction of Artificial Intelligence (AI) in the healthcare realm has drawn a tremendous amount of attention in recent years [1], [2], [3]. The subsequent rise of the Deep Learning (DL) has assisted the surgeons in the operating room in several different ways [4], [5]. This successful incorporation has paved the way for Robot Assisted Surgery (RAS) [6].

Unlike traditional surgery, a RAS system includes a camera arm and few other mechanical arms with surgical instruments attached. The surgeon controls the arms while seated at a computer console near the operating table. The console gives the surgeon a high-definition, magnified, 3D view of the surgical site. The purpose of RAS, as the name suggests, is not to replace the surgeons and physicians but to assist them, in order to achieve higher proficiency in security and safety of the undergoing patients in preoperative, intraoperative, and postoperative surgical procedures [7], [8].
Image driven DL methods for robotic surgery have already taken care of the instrument detection and segmentation [9], [10], gesture recognition [11], workflow analysis [12], skill assessment [13], and many more [14], [15], [16], [17] to facilitate the semi-autonomous RAS. Moreover, the development of a fully autonomous image-guided surgical system, where the direct involvement of the surgeon is seldom required, is foreseeable task for the DL models. The surgical procedures go through several complicated scenes and contain artefacts and performance variances [13]. Additionally, the blur images and videos generated by camera are often misinterpreted and mislabelled by physicians and AI systems, because of the presence of smoke, shade of tools, plasma stains and vessels [18], [19], [20], [21].

Before the advent of the modern image modality capturing systems, the surgeons mostly relied on simple cameras and naked eyes to study the internal behaviour of the organs. Today, the most relied imaging modalities include X-rays, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound (US), and Positron Emission Tomography (PET) [22], [23].

However, even the modern imaging modalities required intensive preprocessing and feature engineering [24]. Thanks to the DL, this laborious, time consuming, and cost intensive task is no more as tedious as heretofore. Moreover, the basic underlying principle of the DL mimics the (functionality of) biological neuron, connects with a complex layered structure, learns from generalization, and keeps the neuron-associated weights updated. One of the most powerful models of the DL is believed to be the Convolutional Neural Network (CNN). The introduction of the CNN can be traced back to early 1960s [25], which has led to the development of several highly efficient diagnostic systems [26].

The DL has ultimately proven the enormous success in Minimally Invasive Surgery (MIS) systems. The very first RAS system i.e. da Vinci surgical system, introduced in the year 2000, has successfully performed around 1, 594, 000 surgical procedures in 2021 [27] with an increase of 28% from the previous year (1, 243, 000 in 2020) and is expected to perform 12 − 15% more in the following part of the year.

The MIS reduces the post-surgery trauma, minimises the hospital stay, improves recovery, and avoids potential risk of contagion [28]. The extreme difficulty of indirect surgical operation leads to the development of instrument tracking, gaze estimation, gesture and trajectory recognition, hand-eye coordination, organ and smoke detection, and depth and pose estimation systems [29], [30], [31], [32], [33], [34], [35].

Furthermore, the research in the DL based image driven RAS systems is expanding and also the availability of recent datasets, i.e. Johns Hopkins University and Intuitive Surgical Inc. Gesture and Skill Assessment Working Set (JIGSAWS), Medical Image Computing and Computer-Assisted Intervention (MICCAI), Cholec80, and ATLAS Dione [36], [37], [38], [39] has boosted the interdisciplinary synergies of biomedical engineers and physicians.

Several recent survey articles span the medical domain [40], [41], [42], however lack the DL part in the technical aspects. All the reviewed technical surveys consider a specific application of deep learning and image processing in the robotic-assisted surgery, such as: surgical phase recognition [43], skill assessment [44], registration [45], tool tracking or segmentation and detection phases [46], [47]. For instance, the study by Rivas et al. [48], published in 2021, considered merely one article published post 2020, and mostly emphasized on available surgical datasets and future of robotics. Another article by Unberantha et al. [49] surveyed 2D/3D image registration in workflow analysis. The study was limited to the CNNs and has not incorporated robotic part and surveyed only one particular subdomain in the RAS. In authors’ opinion, an updated survey that deals with and encircles all the possible applications and aspects is required by the research and medical communities, especially for the new researchers in the field to have the possibility to see the big picture. Another aspect that encouraged the authors to perform a new survey study concerns with the exponentially growing number of publications in the field, as depicted in the Figure 1, since the existing survey articles are pretty old. Authors in [50] have included merely three articles published post 2019, and in last couple of years, a great number of worthy articles have contributed to the domain. Additionally, the main focus of the survey remained limited to tool tracking. From recent studies, it can clearly be observed that image and video guided DL based robotic surgery survey dates back quite a few years, and in the meantime, a huge number of studies have been published on the topic. Therefore, a comprehensive updated survey is missing that can accommodate the DL part and the clinical part, considering image and video driven robotic surgery in light of the recent advancements.

After the comprehensive analysis and thorough survey, the selected papers are classified into 4 different classes, i.e. Surgical Tools, Surgical Processes, Surgical Surveillance, and Skills/Performance Assessment. Each of these classes are further subdivided, and the details can be found in the Sections II and IV of the article. The full text analysis revealed that majority of the articles included in the survey are published in year 2020 and 2021 as shown in the Figure 1. The most frequently used DL method and dataset are CNN and...
JIGSAWS, whereas, the tool segmentation and detection are most studied subcategories within RAS.

The remainder of the article is organised as follows. The Section II consists of the search and review methodology adopted in this survey. A formal introduction of the DL methods is inscribed in the Section III. Image-guided interventions are provided in Section IV followed by the frequently used datasets in the Section V. The Section VI presents legal, ethical, and technological challenges towards the autonomous RAS. Lastly, the Section VII spans the discussion and enlists the prominent findings of the survey followed by the conclusive remarks in the Section VIII.

II. THE LITERATURE SEARCH AND SURVEY METHODOLOGY
The following section describes the literature survey methodology adopted in this study. Initially the literature search and inclusion and exclusion protocols are provided, followed by the article selection process. Moreover, the objectives and the results are also illustrated for the conducted review. Finally, the survey classification layout is presented.

A. LITERATURE SEARCH
A thorough literature search is performed on Scopus database to select the relevant articles for review and analysis. The conducted search is confined to the literature published in English language. To retrieve the optimised results, the combination of the keywords is used interchangeably with slight modifications over repetitive iterations of web search. The specific query used for the final search is: (“deep learning*” OR “deep-learning*” OR convolution* OR “deep networks*” OR “neural network*”) AND (surg*) AND (robot*). The survey study is conducted on the published articles (including those accepted and available online) until August 31, 2022.

B. INCLUSION AND EXCLUSION
The inclusion criteria span the image driven DL models used in any type of robotic surgery. The search query is constrained to computer science, engineering, biomedical engineering, and medical disciplines. Only the published articles are included without considering the books, seminars, doctoral symposiums, and talks. Any article that goes beyond the aforementioned limits, any study not tackling surgery or a part of surgery, the articles related to only engineering side of the robot, and the articles related to only medical side of the surgery (i.e. no intervention of DL) come under the exclusion criteria. The Figure 2 illustrates the stages of the inclusion and exclusion process flow with number of studies included and excluded at each phase.

C. ARTICLE SELECTION
Initially, the titles of the articles and the venues of the publications (i.e. publishing authority and domain) are used to decide the relevancy on a general scale. In the further stages, the abstracts are reviewed, and the contents of each study are skimmed to limit the number of articles to the decided realm for the survey study. Finally, the full-text review is performed, and the appropriate articles are selected for further proceedings. The Figure 2 illustrates the selection stages of the survey which is performed under the recommendations of Preferred Reporting Items for Systematic Review and Meta Analysis (PRISMA).

D. TARGETED OBJECTIVES
The primary objective of the review is to systematically analyse and summarize the recent contributions in the field of image-guided robotic surgery accounting the advancements of the DL. Generally, the review study is conducted on RAS systems and specifically on image based RAS systems. Additionally, the study aims to comprehensively state DL methods, the future of surgical robotics and the challenges to achieve the autonomous surgery. Finally, the secondary objectives include the introduction of currently available surgical datasets, the legal and ethical issues, and the limitations.

E. RESULTS
The aforementioned query resulted in a total number of 879 articles and the minor changes (upto date search) in the query showed 75 additional results. After the first check i.e. title, relevancy and venue, a sum of 482 articles are found appropriate. Another 211 articles are discarded amid irrelevancy to the scope of the survey. At each of the stages, a considerable number of the articles is rejected and at the final stage, 184 articles are tagged eligible to the purview of the study, therefore, 184 articles out of total 954 are appended in this study as shown in the Figures 1 and 2.
F. SURVEY CLASSIFICATION

After the thorough analysis, the articles are found to be greatly overlapping that can be organised in numerous different topologies. However, the careful inspection resulted into four groups, each of which is further classified into several subgroups as depicted in the Figure 3. This classification includes: a) Surgical Tools, b) Surgical Processes, c) Surgical Surveillance, and d) Surgical Performance/Assessment.

The Surgical Tool section is further subdivided into Tool Detection and Tool Segmentation sections, the Surgical Processes includes Gesture Segmentation, Trajectory Segmentation, and Tissue Segmentation categories. The Surgical Surveillance is segregated in Surgical Planning, Phase & State Estimation, and Activity Recognition phases, and the last but not least, Surgical Skill Assessment and Surgical Workflow Recognition come under the Surgical Performance/Assessment group.

III. DEEP LEARNING: A BROADER PICTURE

The section below encircles the fundamental concepts of the DL in the context of medical image analysis and inscribes the formal introduction of the methods and techniques that appear in the literature of the surveyed articles.

A. LEARNING PARADIGMS

ML algorithms are broadly categorized into four major subgroups, supervised, unsupervised, semi-supervised, and reinforcement learning algorithms [51], [52]. Although the literature suggests more subgroups, however, generally two groups are widely cited, namely supervised and unsupervised learning algorithms.

As the name suggests, supervised learning algorithms are fed with complete information under the managed supervision and the model makes decision based on the inputs. A supervised model is given input data along with the relevant labels and it learns by finding the relevant parameters. Whereas, in the unsupervised learning, only input data is provided without any corresponding additional information. The model by itself finds the patterns in the data, learns the meaningful concepts, and makes the prediction [53].

B. NEURAL NETWORKS

Most of the DL models are based on neural network architecture that is verily inspired by the complex structure of human brain. A deep neural network is composed of neurons that has an input and a bias which undertakes a series of processing units to understand, extract and learn the meaningful information from the data.

The DL has revolutionized the traditional ML by illuminating the manual feature extraction process. Furthermore, the introduction of the back propagation algorithm has enabled researchers to compute the impact each parameter imposes on the objective function [54]. The back propagation has further enriched the neural networks and made the computation faster, easier and better.

Until the introduction of layer over layer training of deep neural network, the training of the neural network was widely believed to be quite tedious and ineffective. However, Bengio
et al. [55] proposed a mixed of unsupervised training during layer over layer training and supervised training while fine tuning at two different stages which showed considerable results. The algorithms trained in this manner include autoencoders and belief networks, which are still considered complex because of hectic process to reach substantial results.

C. CONVOLUTIONAL NEURAL NETWORKS

The most widely used DL model is CNN which has proven its applicability in image processing applications [9], [26], [50]. The generally accepted common CNNs based models include VGG16-19, ResNet, Inception, Xception, MobileNet, EfficientNet and many more. The key differences between a Multilayer Perceptron (MLP) and a CNN model are the inclusion of pooling layer in the CNN, sparsely connected layers instead of fully connected, and small associated weights of the layers that particularly help in dealing with image data.

The main building blocks of the CNN are convolutional layer, pooling layer, normalization layer, dense layer, dropout layer and activation layer, which along with their nature and responsibilities are described in the Table 1. The optimal number of layers in a network depends upon the nature of the problem that a network has to deal with, however, in general, there is no fixed number of layers, and hence, it is a matter of search to figure out the optimal number given a certain problem. To avoid the possible trade-off between the computational complexity and the performance, different numbers of the layers and neurons can be considered over repetitive iterations.

The networks with fewer number of layers and trainable parameters take less time, however, at the expense of lower accuracy. These type of models may not reach to full potential by modeling all the required parameters. On the other side, an overly populated network will provide better accuracy results but can also learn unnecessary features which will result in overfitting of the network. This type of model will perform poor on unforeseen data. The solution of the above problems is provided by the pretrained networks [56].

D. MULTILAYER PERCEPTRON

The word perceptron comes from the human like functionality of perception, which was intended for image recognition tasks. The most fundamental category of the artificial neural network is MLP, which incorporates feed-forward neurons. The MLP comprises of several FC layers following the input layer. In MLP, each connected layer contains several neurons, the associated weights, and the bias information coming from the activation function of the previous layer.

The term MLP has largely been misunderstood and therefore used with arguably different definitions. In literature, numerous authors believe that any feed-forward neural network is a MLP in a general context, whereas, some authors remain strict to precise definition, i.e. a network composed of multiple layers of perceptrons.

The concept of MLP arose to tackle the limitation of non-linear mapping of the input to the output. Apart from having the certain input and output layers, an MLP consists of hidden layers along with the several neurons. An activation function is used to control the neurons by imposing a threshold.

E. RECURRENT NEURAL NETWORK

The neural network models appear to have another class known as Recurrent Neural Network (RNN) models. An RNN follows the sequential feeding of the input data. The RNN models are great improvement for time series sequential input data problems. The internal state in the RNN model, also called the memory of the neuron, saves the leading information coming from the previous computations.

The implementation of the RNN in the image driven computer-assisted methods has not been much appreciated in the literature, however, its successful adoption in the natural language processing tasks makes it standout. Another worth noting point of the RNN is its ability to work on variable length input data. The use of the RNN in the robotic systems driven by images and the kinematic data has been increasing over the time.

F. MAJOR BUILDING BLOCKS

The major building blocks of a DL architecture are presented hereunder.

1) LAYER

As the name suggests, the principal operation of CNN is convolution, however, a number of additional layers are added namely dense, dropout, pooling layers and few more to the model to improve performance. The types of layers, the relevant hyperparameters, and their work is summarized in the Table 1.

2) COST FUNCTION

The cost function or the loss function describes how well a model has performed with respect to the ground truth. A number of loss functions have been used in literature depending upon the operations to be performed by model over a specific data. The cross-entropy is most widely used loss function in classification problems [57].

3) PERFORMANCE MEASURING METRICS

The performance measuring metrics are important part of CNN which are the measuring scales that quantify the performance of the model.

The most commonly used metrics for the classification task include accuracy, precision, recall, and f1-score. The segmentation and the object detection tasks may have additional measuring parameters depending upon the nature and the definition of the problem.

G. COMMON FRAMEWORKS AND LIBRARIES

There exist numerous DL frameworks that facilitate to design, train, and validate neural networks using several interfaces. These frameworks and libraries include but not limited to
TensorFlow, PyTorch, MATLAB, NVIDIA Caffe, Chainer, Theano, and Keras.

These high level interfaces help researchers, mathematicians, scientists, and developers to implement the complex architectures of deep neural networks to solve the various real world problems. Few of the most commonly used frameworks and libraries are listed below.

1) TensorFlow
An end-to-end open source library developed by Google Brain team, supports the numerical computation and analysis, is extensively used library that works with both CPU and GPU. The programming interface of TensorFlow is limited to Python and C++ [58].

2) PyTorch
PyTorch has received a tremendous amount of attention by the researchers and the developers because of its ability of easily implement the complex architectures of DL models [59]. Additionally, it also supports the tensor manipulations, e.g. NumPy computations.

3) MATLAB
MATLAB is well-known mathematical framework which is highly regarded in scientific society. It offers great visualization tools and is not limited to DL and neural networks [60]. The high-level features in the MATLAB do not require high level of expertise to implement. The CUDA code is automatically generated by MATLAB from simple code.

4) NVIDIA CAFFE
It largely supports the GPU based computations. NVIDIA Caffe is a worthy contribution of Berkeley Vision and Learning Center to the developer community. The main aims behind the Caffe development were speed and modularity [61].

5) KERAS
Keras is another product by Google engineers which is deemed fruitful for beginners. The four main basic principles were considered during the development of Keras including modularity, minimalism, extensibility, and Python based [62].

IV. IMAGE GUIDED ROBOT ASSISTED SURGERY
In the previous decade, DL methods brought tremendous amount of success, especially in image-guided Computer Aided Diagnosis (CAD) systems as illustrated in the Figure 1. This enormous triumph gave birth to the idea of image driven DL models in the field of robotic surgery. The availability of large amount of image data, the less complicated operational facilities, and the DL algorithms’ performance on image dataset are major knocks towards autonomous surgery.

Based on the literature surveyed, the articles are classified into four categories, each of which are subdivided into further groups. These categories include Surgical Tools, Surgical Processes, Surgical Surveillance, and Surgical Performance. As the names suggest, the divisions are fundamental and encircle the most relevant parts of surgical scenarios in computer-assisted autonomous and semiautonomous surgical systems.

An additional fifth category named Others is provided for the applications that either do not fall in any of the aforementioned categories or the number of found articles were fewer. The section below inscribes all categories in detail.

A. SURGICAL TOOLS
Surgical tools are the most important actuators in surgery because they are responsible for performing interventions;
however, keeping track of surgical instruments requires real-time knowledge of the pose and the movement of the tool. Literature suggests numerous tool localisation techniques embracing electromagnetic tracking [63], kinematic [64], optical tracing [65], and image-guided detection [66] among others [67]. Unlike other approaches, image-driven surgical instrument localisation offers attractive benefits including the knowledge of pose and motion, and does not require instrument design modification [47].

The section below contains the literature surveyed in this study about image-based surgical tool detection and segmentation which are most studied areas in robotic surgery with an average of around 40% of the total publications encompassed in this study.

1) TOOL DETECTION
This section includes the articles that deal with the presence of surgical instruments in surgical videos. Among the articles studying surgical tools, 55% are about detection and/or recognition, whereas the other half of the articles belongs to the forthcoming subsection of tool segmentation. The CNN is the most applied DL method followed by Long Short Term Memory (LSTM), RNN and autoencoder architectures (Figure 4). The CNN model and the variants, with few modifications in the underlying architecture in some cases, yielded better performance in [29], [68], [69], [70], [73], [74], [75], [76], [78], [79], [81], [82], [83], [84], [85], [86], [88], [89], [90], [91], [92], [94], [97], [99], [100], whereas autoencoders, RNN, LSTM, and Generative Adversarial Network (GAN) formed another notable synergy [39], [71], [72], [77], [80], [87], [93], [95].

The reason behind CNN being the most applied architecture lies in the ability of multiple tool detection and localisation which traditional ML models have not been sufficiently successful at [67]. Among the tool detection, the articles employing public datasets revealed an accuracy range of 89-100%, whereas the precision and Dice values vary greatly. The in-house datasets are incorporated by 18 studies achieving an overall accuracy range of well above 90% except one study ( [29]) that managed to reach around 85%.

The Endoscopic Vision (EndoVis) challenge [37] and m2cai16-tool datasets [39] are most widely used followed by the ATLAS Dione for the task of tool detection. Furthermore, accuracy is top used performance measuring metric with precision and Area Under the Curve (AUC) being the other most important evaluation parameters. The more details about the year of publication, objective/s, data description and performance outcome can be found in the Table 2.

2) TOOL SEGMENTATION
Surgical instrument segmentation is different from surgical instrument detection in terms of binary, semantic, and instance segmentation. Generally, tool detection either looks for the presence of any tool (recognition) or the location of a particular tool (tool tip or landmark detection), whereas the segmentation distinguishes (i.e. segments) the tools from other organs and also differentiates among numerous tools. It involves the individual identification of each instrument within an image. As mentioned in the above subsection, tool segmentation is second most common researched field inside the image-guided RAS. Instead of only tool presence recognition, numerous articles focus on the type of tool available in the surgical procedure with semantic segmentation [46], [101], [102], [103], [104], [105], [106], [107], [108], [109], [110], [111], [112], [113], [114], [115], [116], [117], [118], [119], [120], [121], [122], [123], [124], [125], [126]. A noteworthy point arises when articles dealing with organ/object segmentation (see Section IV-E) also consider tool segmentation [127], therefore forming another interconnected relation between two different but relevant tasks.

Additionally, real-time instrument segmentation has gained ample amount of attention in recent studies [128], [129], [130], [131], [132], [133]. Numerous authors also consider the semantic segmentation of a part of a particular instrument, such as tool tip segmentation, guide-wire segmentation and needle segmentation [134], [135], [136].

Likewise, semantic segmentation by using unsupervised DL methods is another growing concept [137], [138]. The data provided by EndoVis robotic instrument segmentation challenge [39] is most frequently used dataset for segmentation, whereas the Dice score is common performance measuring metric. Out of total 38 studies, merely 4 studies incorporated in-house datasets and 2 assimilated both in-house and public datasets. The description of the input, results and other relevant information is provided in the Table 3.

B. SURGICAL PROCESSES
Surgical process is a nontechnical parent terminology induced to explain those sub-tasks of MIS which are not directly relevant to incision but lead to the understanding and developing the next generation autonomous medical robotic
| Objective                          | DL Model | Dataset | Data Description | Results          | Year | Ref |
|-----------------------------------|----------|---------|------------------|-------------------|------|-----|
| Microsurgery Tool Tracking       | LeNet    | RMIT    | 1171 Frames 480 x 640 Pixels 3 Surgeries | Accuracy: 99.13% | 2016 | [68]|
| Line Tracking                     | CNN      | In-house| 1000 Frames 2500 Training Images | Accuracy: 99.7%  | 2017 | [69]|
| Surgical Tool Detection          | CNN YOLO | M2CAI16-Tools Dataset | 10 Procedures 2532 Frames | Recall: 80.62% Precision: 84% mAP: 72.26 | 2017 | [70]|
| Tool Landmark Detection          | Encoder-Decoder CNN | In-house | 10 Sequences 1500 HD Images | RMSE: 25.479 μm | 2017 | [71]|
| Robotic Tool Detection           | Faster RPN | ATLAS Dione | 10 Surgeons 99 Videos total 854 x 480 22,467 Images | Precision: 91% | 2017 | [39]|
| Tool Joint Detection             | 3D FCNN | EndoVis | 10 Videos 1083 Frames 720 x 576 Resolution 8 Videos 3075 Frames | DSC: 88.6% DSC: 86.9% Dice: 85.1% | 2019 | [72]|
| Tool Localization & Detection    | ResNet-18 50-152 AlexNet VGG-16 | cataRACT | 50 Videos 10 Min & 56 Sec Duration | AUC: 0.65 0.68 0.64 0.58 | 2019 | [73]|
| Guidewire Tip Tracking           | U-Net    | In-house | 11 Videos 11268 Frames | Dice: 88.07% IoU: 85.07% | 2019 | [74]|
| Needle Localization              | ResNet-18 RetinaNet | In-house | 19,200 Images 512 x 1024 Resolution | Accuracy: 99.2% | 2019 | [75]|
| Surgical Tool Detection          | Hourglass VGG-16 | ATLAS Dione EndoVis | 99 Video 10 Surgeons 22467 Frames 1083 Frames 720 x 576 Resolution | mAP: 91.60% mAP: 100% | 2019 | [76]|
| Surgical Tool Detection          | CNN VGG-M | M2CAI16-Tools Dataset | 10 Procedures 2532 Frames | Accuracy: 89% | 2019 | [77]|
| Instrument Detection             | YOLO90000 CNN | M2CAI16-Tools Dataset | 10 Procedures 2532 Frames | mAP: 84.7 | 2019 | [78]|
| Surgical Tool Detection          | ResNet-18 ResNet-101 Hourglass-104 | ATLAS Dione EndoVis | 99 Video 10 Surgeons 22467, 1083 Frames 720 x 576 Resolution | mAP: 98.5% mAP: 100% | 2020 | [79]|
| Needle Detection                 | LSTM CNN | In-house | NA | 100% TPR | 2020 | [80]|
| Instrument Detection             | VGG-16 CNN | ATLAS Dione | 10 Surgeons 99 Videos total 854 x 480 22,467 Images | Precision: 90.08% | 2020 | [81]|
| Surgical Tool Navigation         | ResNet-18 ConvNet | In-house | 4500 Image | Accuracy: 137 μm | 2020 | [82]|
| Instrument Detection             | ResNet-18 ConvNet | EndoVis | 300 Frames 640 x 480 Size 400 Frames 640 x 480 Size | Accuracy: 91.2% Accuracy: 75% | 2020 | [83]|
| Needle Insertion Tracking        | U-Net    | In-house | 30 Porcine Eyeballs 300 Training Images 1024 x 640 Pixels | Errors: 7.4 μm 10.5 μm 3.6 μm | 2020 | [84]|
| Object Recognition               | CNN      | In-house | 5670 Images 3968 x 2976 Size | Accuracy: 98% | 2020 | [85]|
TABLE 2. (Continued.) The summarised results of the articles dealing with the tool detection task.

| Objective                        | DL Model             | Dataset        | Data Description                              | Results                      | Year | Ref |
|----------------------------------|----------------------|----------------|----------------------------------------------|------------------------------|------|-----|
| Needle Detection                 | Faster RCN           | In-house       | 27 Videos 9 Subsets                         | Precision: 89.2% IoU: 73.9% | 2020 | [86]|
| Tool Presence Analysis           | Multitask RCN        | Cholec80       | 80 Videos 13 Surgeons 854 x 480 Resolution  | mAP: 89.1% F1 Score: 87.4%  | 2020 | [87]|
| Tool Tracking                    | GAN                  | EndoVis15      | 3 Videos 446 Long                           | Accuracy: 95.2%             | 2020 | [88]|
| Surgical Tool Detection          | CNN                  | Cholec80       | 80 Videos 13 Surgeons 854 x 480 Resolution  | mAP: 91.6% mAP: 100%        | 2021 | [89]|
| Tool Tip Detection               | RetinaNet YOLOv2     | In-house       | 2310 Frames 9 Videos 640 x 480 Resolution   | Recall: 1.000 Precision: 0.733 F1 Score: 0.846 Recall: 0.864 Precision: 0.808 F1 Score: 0.835 | 2021 | [90]|
| Needle Detection & Segmentation  | CNN NN               | In-house       | 203 US Images Terson uSmart 3200 T.NexGen US system 22-gauge 0.7 mm diameter 80 mm length | Accuracy: 99.7% Precision: 96.2% Recall: 95.1% F1-score: 0.87 | 2021 | [91]|
| Instrument Tracking              | TernausNet-11        | In-house       | 1846 Images 640 x 480 Size                  | Accuracy: 85.87%            | 2021 | [29]|
| Object Detection                 | YOLOv4               | M2CA16-Tools Dataset | 10 Procedures 2532 Frames                  | Recall: 79.1% Precision: 96.7% | 2021 | [92]|
| Tool Detection & Segmentation    | YOLACT               | In-house       | 5,319 Frames 70 Videos 1920 x 1080 Resolution | Accuracy: 91.2% Precision: 56.5% Dice: 48.2% | 2021 | [93]|
| Needle Tracking                  | NN YOLO-v3           | In-house       | 778 & 332 Images                            | Accuracy: 1.98mm            | 2021 | [94]|
| Instrument Detection             | Faster R-CNN         | In-house       | 5085 Images                                 | IoU: 0.825 Recall: 0.950 Precision: 0.950 | 2022 | [95]|
| Instrument Tracking              | YOLO-v4              | In-house       | 6243 Images                                 | Accuracy: 95.12%            | 2022 | [96]|
| Tool Detection                   | YOLO-v5              | In-house       | 20 Videos 7500 Frames                       | Precision: 89.5-91.4%       | 2022 | [97]|
| Instrument Triplet               | SIR-Net              | EndoVis18      | 16 Videos 1280 x 1024 Pixels 8 Instruments  | Average Precision: 0.6515   | 2022 | [98]|
| Micro-robot Detection            | VGG Net              | In-house       | 15000 Ultrasound Images                      | Accuracy: 0.95 & 0.93       | 2022 | [99]|
| Object Detection                 | ResNet-101 Back Projection | Data Generation from Video               | 380 Images                     | Accuracy: 94.12% Recall 86.23% | 2022 | [100]|
| Object Detection                 | YOLO-v4 Faster-RCNN  | In-house       | 196.55 Minute Videos 870 Images              | mAP: 29.3, 22.2, 23.4, 33.6 F1 Score: 75.86, 82.34, 82.49, 93.50 | 2022 | [67]|

systems. The surgical process section is subdivided into three sections, i.e. Gesture Recognition, Trajectory Segmentation, and Tissue Segmentation, based on the contribution of the authors towards the field. In addition, numerous studies employ these terminologies interchangeably, however, considering the in-depth analysis, the suturing task cannot be confused with unique movement of surgical instrument. The former can be perceived as series...
TABLE 3. The summarised results of the articles dealing with the instrument segmentation task.

| Objective                                      | DL Model         | Dataset       | Data Description                                      | Results                      | Year  | Ref  |
|------------------------------------------------|------------------|---------------|-------------------------------------------------------|------------------------------|-------|------|
| Surgical Tool Segmentation                     | CNN RNN AutoEnc  | EndoVis16     | 4 Videos 45-seconds Each 720 x 576 Resolution 25 Frames | Accuracy: 93.3% IoU: 0.936 | 2017  | [103]|
| Automatic Instrument Segmentation              | U-Net TernausNet-11 LinkNet | EndoVis17 | 8 Sequences 225 Frames Test Set 8 x 75 Frames 2 x 300 Frames Videos | IoU: 0.836 Dice: 90.01 Time: 184 | 2018  | [46] |
| Guidewire Tip Segmentation                      | Faster R-CNN     | In-house COCO PASCAL VOC | 22 Sequences 2D X-ray Images 1080 x 1080 Pixels | Precision: 0.532 F1 Score: 0.939 | 2018  | [134]|
| Binary Segmentation                             | ResNet-18 FNN    | EndoVis17     | 8 Sequences 1280 x 1024 Resolution 225 Frames 8 Sequences 75 Frames | IoU: 0.764                  | 2019  | [101]|
| Instrument Segmentation                         | CNN              | EndoVis17     | 225 Frames 8 Surgeries                                | Dice: 0.916 Hausdorff: 11.11 Sensitivity: 0.928 | 2019  | [128]|
| Semantic & Instance Segmentation               | U-Net            | EndoVis17     | 8 Sequences 225 Frames Test Set 8 x 75 Frames 2 x 300 Frames Videos | Dice: 90.20%                | 2019  | [102]|
| Realtime Instrument Segmentation                | MobileNet-v2     | EndoVis17 Cata7 | 8 Sequences 225 Frames Test Set 8 x 75 Frames 2 x 300 Frames Videos | Dice: 96.91% IoU: 94.10% Dice: 58.30% | 2020  | [131]|
| Object Extraction for Instrument Tracking       | U-Net ResNet-18  | M2CAI16-Tools Dataset | 10 Procedures 2532 Frames 1280 x 720 Pixels | Accuracy: 100%              | 2018  | [108]|
| Surgical Instrument & Workflow Recognition      | Bayesian AlexNet LSTM | Cholec80    | 80 Videos 13 Surgeons 25 fps 854 x 480 Resolution | Bipolar: wMAE: 0.76 pMAE: 0.96 Scissors: wMAE: 0.51 pMAE: 0.76 | 2020  | [109]|
| Tool Segmentation                               | FCNN ResNet-50 U-Net | In-house       | 14 Videos 300 Frames 720 x 576 Pixels 4200 Annotations | IoU: 81.80/7.74             | 2019  | [132]|
| Synthetic Image Segmentation                    | U-Net            | In-house      | 160 Frames 5 Videos                                  | mIoU: 0.235 mIoU: 0.458 mIoU: 0.729 | 2020  | [110]|
| Multi-Angle Instrument Segmentation            | TernausNet-16 VGG-16 | Sinus-Surgery- C Dataset | 10 Videos 5-23 Minute 320 x 240 Resolution | mDSC: 90.2±0.0% mIoU: 85.6±1.2% | 2020  | [111]|
| Unsupervised Learning for Instrument Segmentation | CycleGAN DRN     | EndoVis17     | 8 Sequences 225 Frames 2 x 300 Frames Videos         | IoU: 0.732                  | 2020  | [137]|
| Ultrasound Needle Segmentation                  | LinkNet          | In-house      | 996 Images 102 Videos 3 fps                         | IoU: 0.410% Dice: 56.65% F1 Score: 36.61% RMS: 13.3 | 2020  | [135]|
| Instrument Segmentation                         | GAN EndoVis18 EndoVis17 | 8 Sequences 225 Frames 19 Sequences               | Accuracy: 76.29%             | 2020  | [112]|
| Tools Collision Avoidance using Segmentation    | U-Net            | EndoVis17     | 8 Sequences 225 Frames 2 x 300 Frames Videos        | MAE: 0.126 ± 0.08 mm        | 2020  | [116]|

122636
TABLE 3. (Continued.) The summarised results of the articles dealing with the instrument segmentation task.

| Objective                                      | DL Model          | Dataset          | Data Description                     | Results                  | Year  | Ref  |
|------------------------------------------------|-------------------|------------------|--------------------------------------|--------------------------|-------|------|
| Instrument Segmentation                        | ResNet-18         | EndoVis17        | 8 Sequences 225 Frames 2 x 300 Frames Videos | IoU: 0.852 Time: 11.8 ms | 2020  | [117]|
| Instrument Segmentation                        | GAN CNN           | Sinus-Surgery-C  | 10 Videos 320 x 240 Resolution       | mIoU: 82.7              | 2020  | [115]|
| Unsupervised Instrument Segmentation           | Vanilla U-Net     | EndoVis17        | 8 Sequences 225 Frames 2 x 300 Frames Videos | IoU: 0.71 Dice: 0.81     | 2020  | [138]|
| Surgical Instrument Segmentation               | CycleGAN U-Net    | EndoVis17        | 8 Sequences 225 Frames Test Set 8 x 75 Frames 2 x 300 Frames Videos | Dice: 92.8% IoU: 84.7%   | 2021  | [104]|
| Real-Time Instrument Segmentation              | LSTM              | EndoVis18        | 8 Sequences 225 Frames Test Set 8 x 75 Frames 2 x 300 Frames Videos | mDSC: 61.03% mIoU: 53.89% mDice: 77.53% mIoU: 67.50% | 2021  | [129]|
| Multi-Instance Segmentation                    | Encoder-Decoder CNN | EndoVis17       | 8 Sequences 225 Frames Test Set 8 x 75 Frames 2 x 300 Frames Videos | mAP: 0.481 ± 0.099 mIoU: 0.657 | 2021  | [105]|
| Real-Time Instrument Segmentation              | VGG MobileNet ResNet | UW-Sinus-Surgery-C/L ROBUST-MIS | 10 Videos 5-23 Minute 320 x 240 Resolution 3 Videos 12-66 Minute 1920 x 1080 Resolution | mDSC: 3.1% 9.5% mIoU: 3.3% 10.7% | 2021  | [130]|
| Instrument Segmentation                        | U-Net             | ISBI2018         | 8 Sequences 225 Frames Test Set 8 x 75 Frames 2 x 300 Frames Videos | mAUC: 0.6819 IoU: 83.70% Dice: 90.24% | 2021  | [106]|
| Surgical Tool Segmentation                     | GAN U-Net         | EndoVis17        | 8 Sequences 225 Frames Test Set 8 x 75 Frames 2 x 300 Frames Videos | IoU: 0.867 Dice: 0.924 | 2021  | [107]|
| Robot Positioning Using Instrument Segmentation| YOLOv3 ResNet     | EndoVis17        | 8 Sequences 225 Frames 2 x 300 Frames Videos | IoU: 86.6%               | 2021  | [113]|
| Real-Time Instrument Semantic Segmentation     | MobileNet-v3      | EndoVis17        | 8 Sequences 225 Frames 2 x 300 Frames Videos | IoU: 69.74% Dice: 79.88% Hausdorff: 11.36 | 2021  | [133]|
| Guide Wire Segmentation                        | MobileNet U-Net   | In-house         | 1050 Images 1440 x 1560 Pixels | Accuracy: 97.81%        | 2021  | [136]|
| Instrument Segmentation                        | Modified CNN      | MICCAI 2018      | 15 Videos | IoU: 0.4354 Accuracy: 0.9638 | 2022  | [119]|
| Instrument Segmentation                        | Modified CNN      | EndoVis & In-house | 10 & 20 Videos 4 & 5 Videos 720 x 576 & 1920 x 1080 Pixels & 5000 Frames | Average Accuracy: 93.31% | 2022  | [119]|
| Instrument Segmentation                        | U-Net & VGG-16    | Hamlyn's & Proprietary | 1920 x 1080 Pixel 8 x 255 Frames | IoU: 0.708 & 0.826 | 2022  | [120]|
| Instrument Segmentation                        | SurgiNet & MobileNet-v2 | EndoVis 2017 & Catalyst | 10 & 9 Videos 3000 & 2671 Images 7 & 11 Instruments | Mean IoU: 89.14% & 63.30% | 2022  | [121]|

| Objective                                      | DL Model          | Dataset          | Data Description                     | Results                  | Year  | Ref  |
|------------------------------------------------|-------------------|------------------|--------------------------------------|--------------------------|-------|------|
| Instrument Segmentation                        | ResNet-18         | EndoVis17        | 8 Sequences 225 Frames 2 x 300 Frames Videos | IoU: 0.852 Time: 11.8 ms | 2020  | [117]|
| Image-to-Image Translation for Instrument Segmentation | GAN CNN           | Sinus-Surgery-C  | 10 Videos 320 x 240 Resolution       | mIoU: 82.7              | 2020  | [115]|
| Unsupervised Instrument Segmentation           | Vanilla U-Net     | EndoVis17        | 8 Sequences 225 Frames 2 x 300 Frames Videos | IoU: 0.71 Dice: 0.81     | 2020  | [138]|
| Instrument Segmentation                        | CycleGAN U-Net    | EndoVis17        | 8 Sequences 225 Frames Test Set 8 x 75 Frames 2 x 300 Frames Videos | Dice: 92.8% IoU: 84.7%   | 2021  | [104]|
| Surgical Instrument Segmentation               | LSTM              | EndoVis18        | 8 Sequences 225 Frames Test Set 8 x 75 Frames 2 x 300 Frames Videos | mDSC: 61.03% mIoU: 53.89% mDice: 77.53% mIoU: 67.50% | 2021  | [129]|
| Multi-Instance Segmentation                    | Encoder-Decoder CNN | EndoVis17       | 8 Sequences 225 Frames Test Set 8 x 75 Frames 2 x 300 Frames Videos | mAP: 0.481 ± 0.099 mIoU: 0.657 | 2021  | [105]|
| Real-Time Instrument Segmentation              | VGG MobileNet ResNet | UW-Sinus-Surgery-C/L ROBUST-MIS | 10 Videos 5-23 Minute 320 x 240 Resolution 3 Videos 12-66 Minute 1920 x 1080 Resolution | mDSC: 3.1% 9.5% mIoU: 3.3% 10.7% | 2021  | [130]|
| Instrument Segmentation                        | U-Net             | ISBI2018         | 8 Sequences 225 Frames Test Set 8 x 75 Frames 2 x 300 Frames Videos | mAUC: 0.6819 IoU: 83.70% Dice: 90.24% | 2021  | [106]|
| Surgical Tool Segmentation                     | GAN U-Net         | EndoVis17        | 8 Sequences 225 Frames Test Set 8 x 75 Frames 2 x 300 Frames Videos | IoU: 0.867 Dice: 0.924 | 2021  | [107]|
| Robot Positioning Using Instrument Segmentation| YOLOv3 ResNet     | EndoVis17        | 8 Sequences 225 Frames 2 x 300 Frames Videos | IoU: 86.6%               | 2021  | [113]|
| Real-Time Instrument Semantic Segmentation     | MobileNet-v3      | EndoVis17        | 8 Sequences 225 Frames 2 x 300 Frames Videos | IoU: 69.74% Dice: 79.88% Hausdorff: 11.36 | 2021  | [133]|
| Guide Wire Segmentation                        | MobileNet U-Net   | In-house         | 1050 Images 1440 x 1560 Pixels | Accuracy: 97.81%        | 2021  | [136]|
| Instrument Segmentation                        | Modified CNN      | MICCAI 2018      | 15 Videos | IoU: 0.4354 Accuracy: 0.9638 | 2022  | [119]|
| Instrument Segmentation                        | Modified CNN      | EndoVis & In-house | 10 & 20 Videos 4 & 5 Videos 720 x 576 & 1920 x 1080 Pixels & 5000 Frames | Average Accuracy: 93.31% | 2022  | [119]|
| Instrument Segmentation                        | U-Net & VGG-16    | Hamlyn's & Proprietary | 1920 x 1080 Pixel 8 x 255 Frames | IoU: 0.708 & 0.826 | 2022  | [120]|
| Instrument Segmentation                        | SurgiNet & MobileNet-v2 | EndoVis 2017 & Catalyst | 10 & 9 Videos 3000 & 2671 Images 7 & 11 Instruments | Mean IoU: 89.14% & 63.30% | 2022  | [121]|

VOLUME 10, 2022 122637
of analogous gestures, whilst latter can be described as the movement in and around a particular region.

1) GESTURE SEGMENTATION
Surgical gesture recognition has been perceived in several different overlapping contexts i.e. path planning, needle positioning, continuous tip detection, etc. and therefore, numerous DL methods have been applied in respective perspectives. The gesture segmentation is generally implemented for suturing tasks, therefore, it is a cumbersome job because of the similarity and repetition of analogous suturing steps.

The gesture recognition can be either live or in-vitro environments, however for suturing tasks, it is broadly available as in-vitro experiment in the literature [11], [16], [139], [140], [141], [142], [143], [144], [145]. The live suturing task involves risks and requires close consideration and high costs, therefore fewer studies adopt live suturing [146], [147], [148].

The LSTM model is adopted by five out of total thirteen studies, whereas, CNN and RCN are other most applied models by several authors. Moreover, all thirteen articles used accuracy to measure the performance of the DL models along with other metrics and nine out of thirteen studies incorporated JIGSAWS [36] dataset and three employed in-house datasets (among these two studies used both JIGSAWS and other datasets also).

However, the outcome of these research studies is evident that fusion of different types of data (i.e. video and kinematic data) yields better accuracy as compared to only image/video data. The further technical details extracted from the gesture segmentation works are enlisted in the Table 4.

2) TRAJECTORY SEGMENTATION
The task of trajectory segmentation also involves the motion analysis and pattern recognition of the involved surgical tools. Similar to the gesture segmentation task, the in-vitro experiments are generally used [150], [151], [152], [153], [154], [155], [156], [157], [158], [159]. To improve the results of segmentation, authors also incorporate the kinematic data along with the video and image data [150], [155], [156], [157], [158], [160]. The kinematic data has particular importance because it leads to the learning from demonstration.

Likewise, not only the thread detection but also knot tying and path planning are largely associated with trajectory segmentation tasks [151], [152], [154], [157], [158], [161], [162]. It is worth mentioning that instead of using one single architecture, authors used combination of DL architectures to produce better performance results (e.g. CNN and LSTM). The Table 5 highlights the salient features of the trajectory segmentation task and the employed DL models.

3) TISSUE SEGMENTATION
The tissue segmentation appears to be the third largest studied task in this survey comprising around 12% of the total articles (see Table 6). This section comprises all the studies performed on tissues including vessel segmentation, edge detection, healthy and cancerous tissue classification, uncertainty inference segmentation, and tissue retraction [163], [164], [165], [166], [167], [168], [169], [170], [171], [172], [173], [174].

The applicability of tissue segmentation spans the liver to the brain to the kidneys to several other organs [174], [175], [176], [177], [178], [179], [180], [181].
TABLE 4. The summarized results of the surgical gesture recognition and segmentation articles.

| Procedure        | Objective                  | DL Model      | Dataset        | Input Datatype | Input Description | Data            | Results                  | Ref  | Year |
|------------------|----------------------------|---------------|----------------|----------------|-------------------|-----------------|----------------------------|------|------|
| Laparoscopy      | Gesture Recognition        | NN            | In-house       | Video Data     | 7 Gestures, 2 Tools | Accuracy: 100% & 80% | 2006 | 143  |
| In-vitro Suturing| Suturing segmentation     | RCN, LSTM     | JIGSAWS        | Video & Kinematic | 11 Gestures, 14 Sequences | Accuracy: 71% & 67% | 2019 | 141  |
| In-vitro Suturing| Suturing recognition      | 3D CNN        | JIGSAWS        | Video & Kinematic | 39 Videos, 11 Gestures | Accuracy: 84.3%   | 2019 | 142  |
| In-vitro Suturing| Suturing classification   | CNN           | JIGSAWS        | Image Data     | 10 Gestures, 39 Sequences | Accuracy: 81.67% | 2020 | 139  |
| Live Suturing    | Suturing recognition      | RCN, LSTM     | JIGSAWS        | Video Data     | 10 Gestures, 39 Sequences | Accuracy: 85.5%   | 2020 | 146  |
| Live Suturing    | Suturing recognition      | CNN           | JIGSAWS        | Video Data     | 10 Gestures, 39 Sequences | Accuracy: 90.1%   | 2020 | 147  |
| In-vitro Suturing| Suturing segmentation     | 3D CNN, LSTM  | JIGSAWS        | Image Data     | 10 Gestures, 39 Sequences | Accuracy: 76.3%   | 2020 | 11   |
| Prostatectomy Suturing | Suturing in Needle Driving | AlexNet LSTM ConvLSTM | In-house | Video & Kinematic | 2395 & 511 Videos, 5 Gestures | Accuracy: 78% & 62% | 2021 | 16   |
| Suturing Tasks   | Suturing classification   | LSTM          | JIGSAWS, In-house | Video & Kinematic | 12 Gestures | Accuracy: 75% | 2021 | 148  |
| In-vitro Suturing| Suturing recognition      | SD-Net        | JIGSAWS        | Video Data     | 10 Gestures, 39 Sequences | Accuracy: 90.5% | 2021 | 140  |
| Prostatectomy Suturing | Suturing recognition       | TCN           | JIGSAWS, RARP-45 | Video & Kinematic | 39 & 45 Videos, 12 & 7 Gestures | Accuracy: 86.8% & 81% | 2022 | 149  |
| Hand Gesture     | Suturing recognition      | Modified MobileNet-v2 | In-house Jester | Video         | 30 Subjects, 2*30 Gesture, 148,092 Videos, 7 Gestures | mAP: 96.82% | 2022 | 144  |
| In-vitro Suturing| Suturing recognition      | ResNet-50 TCN | JIGSAWS        | Video & Kinematic | 39 Videos, 10 Gestures | Accuracy: 89.8% | 2022 | 145  |

It also includes the binary classification of healthy and cancerous tissues [169], [171], [182].

The aforementioned division of the segmentation phase is placed under the same category because of the overlapping interest and the main task involved in the article. The individual categories can be assumed in a study that only encircle the tissue segmentation task regardless of the input type. The basic reason behind including the smaller categories (even with fewer published papers found) is the significant contribution discussed in the robotic surgery field.

The U-Net architecture is the most applied network among all studies with being adopted by eleven out of seventeen articles, as evident in the Table 6. The U-Net is often applied in conjunction to the other networks including LSTM, GAN [180] and other variants of CNN [164], [166], [169], [170], [172], [177], [179]. Because of the unavailability of large scale public datasets, a significant majority of the studies (13 out of 18) incorporate in-house dataset. Considering the proprietary datasets, authors have used varying performance measuring metrics including accuracy, Intersection over Union (IoU), Dice and AUC. The further insights about the datasets, performance, and the DL techniques are provided in the Table 6.

C. SURGICAL SURVEILLANCE

The increasing introduction of biomedical images facilitates the surgical surveillance and the navigation during the surgical process. This section surveys the articles that monitor the surgical situation with respect to the patient and the ongoing procedure.

1) SURGICAL PLANNING

Surgical planning is largely considered as preoperative planning, where the steps are performed in advance in order to pre-visualise the intervention. The application has a large benefit in emergency situations and war field areas where reaching the hospital requires time. The vessel detection and needle insertion are prominent and attractive applications, and surgical robots are made to achieve timely fashioned aid [183]. The recent trends in surgical planning have shown great interest in the 2D, 3D, and 4D model construction for the interventional guidance [183], [184], [185], [186], [187], [188], [189], [190], [191]. Due to the unavailability of the large amount of data for preoperative planning, majority of the articles rely upon in-house data [184], [185], [186], [187], [189], [190], [191] which includes CT scans and MRI scans [184], [185], [186], [188], [189].
TABLE 5. The summary of the results of the trajectory segmentation publications.

| Objective                  | DL Model                                      | Dataset          | Input Datatype  | Data Description                  | Results                                | Year | Ref |
|----------------------------|-----------------------------------------------|------------------|-----------------|-----------------------------------|----------------------------------------|------|-----|
| Trajectory Segmentation    | Deep CNN, VGG, AlexNet                        | JIGSAWS          | Video & kinematic | 10 Gestures 39 Sequences          | Silhouette Score: 0.733±0.056, 0.716±0.097 | 2016 | [155] |
| Pattern Cutting            | Deep Reinforcement Learning                  | In-house         | Images          | NA                                | IoU: 0.833                             | 2017 | [154] |
| Trajectory Planning        | CNN                                           | In-house         | Video & Kinematic | NA                                | Accuracy: 91.25%                        | 2017 | [157] |
| Trajectory Segmentation    | Convolutional Auto-Encoder                   | JIGSAWS          | Video & kinetic  | 10 Gestures 39 Sequences          | Accuracy: 78.2% Accuracy: 92.1%         | 2018 | [156] |
| Trajectory Segmentation    | Dense Convolutional Encoder-Decoder Network   | JIGSAWS          | Video & kinetic  | 28 Videos 8 Video 11 Voices 38 Video 10 & 19 Voices | Accuracy: 70.8% Accuracy: 62.1%       | 2018 | [150] |
| Trajectory Generation      | CNN                                           | In-house         | Images          | 60 & 10 Cable Images              | IoU: 0.754 IoU: 0.583                  | 2018 | [160] |
| Thread Tip Detection       | CNN, LSTM with RNN                            | NA               | Images          | 1278 & 1215 Labeled Images        | Precision: 99.63 Recall: 98.89        | 2019 | [151] |
| Pedicle Screw Path Planning| CNN based 3D U-Net                            | NA               | CT Images       | 21 Spinal CT Images               | Dice: 95.55 Jaccard: 91.92 MSE: 1.340  | 2019 | [152] |
| Trajectory Planning        | NN                                            | LumSeg, SpiSeg, xVertSeg | CT Images    | 105 CT Scans                      | Positioning Error ± Std: 2.37±0.97    | 2020 | [161] |
| Trajectory Generation      | DNN, FC DenseNet                              | LumSeg, SpiSeg, xVertSeg | CT Images    | 105 CT Scans                      | Positioning Error ± Std: 2.37±0.97    | 2021 | [153] |
| Path Planning              | 3D U-Net                                      | In-house         | CT Images       | 15 & 8 CT Scans                   | Accuracy: 93%                          | 2022 | [162] |
| Path Planning              | GAN CNN LSTM                                  | In-house         | Image & Kinematic | NA                                | Accuracy: 72.94%                       | 2022 | [158] |
| Motion Prediction          | TCN Attention Module                          | In-house         | Ultrasound Images | 33 + 25 Subjects 30 - 45Hz        | RMSE: 1.02                             | 2022 | [159] |

As can be seen in the Table 7, the majority of the authors use Dice performance measure. Along with the surgical planning, the articles [184], [190], [191] also deal with the depth estimation, motion detection and path planning, which are also crucial parts of RAS. The articles are enlisted under the surgical planning section considering their major contribution towards the category.

2) PHASE AND STATE ESTIMATION

The surgical phase and state estimation subgroup a particular surgical process into several chunks of phases and establish a system that recognises the phase or state of the surgical process on a given set of inputs. The task is largely applied in several domains including cholecystectomy, endovascular, and esophagectomy procedures [87], [192], [193], [194], [195], [196], [197], [198], [199], however, it appears to have less generalization for other types of surgical procedures.

Dissimilar to the surgical planning, the availability of the data for phase recognition makes majority of the articles rely upon public datasets [87], [192], [194], [195], [196], [199] which include kinematic data with images and videos. The CNN and LSTM are two most frequently applied methods with accuracy being the top performance measuring parameter. Phase and state estimation are directly related to identifying the status of the process at certain time, therefore, the accuracy is used by almost all studies to evaluate the performance of employed DL models. The Table 8 provides further details on the phase and state estimation studies.

3) ACTIVITY RECOGNITION

The automatic surgical activity recognition before the surgical procedure and in the operating room during the surgical intervention gained considerable attention in the recent past [201], [202]. The surgical activity recognition involves the real-time followup of the procedure under consideration. The integration of CNN and LSTM networks helped surgeons draw reasonable conclusions, however, the unavailability of large datasets has led to use the pretrained DL models [203], [204], [205], [206]. The pretrained DL models are highly trained on ImageNet dataset [207].
### TABLE 6. The concise results of the tissue segmentation papers.

| Procedure | Objective | DL Model | Dataset | Data Description | Results | Year | Ref |
|-----------|-----------|----------|---------|------------------|---------|------|-----|
| **In-vitro Experiment** | Brain Tumor Tissues | FFNN | In-house | 144 Samples | IoU: 0.911 | 2016 | [178] |
| **Laparoscopy** | Liver Tissue Segmentation | CNN | In-house | 2050 Video Frames | Dice ≥ 0.95 | 2017 | [175] |
| **Ex-vivo Experiment** | Classify Tumours | AlexNet VGG19 Inception-v3 | In-house | 250000 Frames | Accuracy %: 99.47, 99.82, 99.71 | 2018 | [169] |
| **Spondylectomy** | Tissue Reconstruction | GAN, DCNN U-Net, VNet | DeepLesion | 282 Images 3863 Images | IoU: 0.9584 | 2019 | [180] |
| **Laparoscopy** | Organs & Tissue Segmentation | U-Net, LinkNet SegNet, FCN | EndoVis19 | Videos 1 & 2 EndoVis19 12 Videos | IoU: 78.31 | 2020 | [170] |
| **Arthroscopy** | Segment Femoral Cartilage | U-Net | In-house | 18278 2D Images | Dice: 0.87% | 2020 | [171] |
| **Arthroplasty** | Multi Tissue Segmentation | U-Net U-Net++ | In-house | 3368 Images | Dice: 0.64% | 2020 | [164] |
| **Lobectomy** | Soft Tissue Retraction | DNN U-Net | In-house | 1080 Images 62 Minutes Video | Accuracy: 83.4% ± 3.3% | 2020 | [166] |
| **Arthroscopy** | Inference Segmentation | Bayesian CNN | In-house | 16973 17944 Images | AUC: 90.0 AUC: 89.2 | 2020 | [167] |
| **In-vivo Experiments** | Tissue Edge Detection | U-Net | EndoVis17 | 150 Images from da Vinci | IoU: 0.3 | 2020 | [176] |
| **Arthroplasty** | Tissue Classification | GoogLeNet | In-house | 500 Images | Accuracy: 97.8% | 2020 | [181] |
| **In-vivo & Ex-vivo Experiments** | Healthy Tissues vs Cancerous | SVM RF | In-house | 53 Patients 67893 In-vivo 89695 Ex-vivo | ROC-AUC: 0.88 | 2020 | [182] |
| **Nephrectomy** | Vessel Segmentation | FCNN 2D U-Net 3D U-Net NephCNN | Nephrec9 | 8 RAPN videos 1871 Frames | Dice: 71.76% | 2021 | [177] |
| **Prostatectomy** | Tissue Semantic Segmentation | U-Net RezNet MobileNet | In-house | 5 Videos 15570 Images | IoU: 0.894 | 2021 | [179] |
| **Gastrectomy** | Tissue Segmentation | U-Net | In-house | 33 Videos 30 fps | Mean Recall: 0.606 Mean Dice: 0.549 | 2021 | [172] |
| **Abdominal Surgery** | Segmenting Soft Tissue | U-Net LSTM | FlapNet | 2736 Sequences | Accuracy: 83.77% ± 2.18% | 2021 | [163] |
| **Neurosurgery** | Vessel Segmentation | Modified U-Net | In-house Proprietary | 25 fps 40, 34, 41 Frames 224x224 Pixels | Dice: 0.97, 0.86 0.87 0.77 | 2022 | [173] |
| **In-vivo Experiments** | Tissue Classification | CNN | In-house | 9059 Images 17777 Annotations | Accuracy: 0.95 | 2022 | [174] |

Recently, surgical activity recognition has been considered the essential part of surgical planning and state estimation, therefore, as can be observed from the above two subsections, the activity recognition comes under the umbrella of surgical procedure planning. Apart from the video and image data, the kinematic data is integrated to generate better results, whereas, the accuracy is considered the reasonable performance measure. The Table 9 highlights the main concepts, methods, datasets, and other relevant information about the articles addressing the surgical activity recognition tasks.

### D. SURGICAL PERFORMANCE/ASSESSMENT

The performance and skill assessment of surgeon in a surgical procedure is one of the most crucial aspects of autonomous robotic surgery because it lays down the steppingstone for computer-aided autonomous systems. The skill evaluation of the surgeons along with the workflow recognition from a surgical video allows to compute the dexterity and precision. Therefore, building a robotic surgical system requires the beforehand understanding of the skill assessment and flow during the procedure.
TABLE 7. The summary of the studies related to the surgical planning.

| Objective                          | DL Model           | Dataset         | Data Description                              | Results                                      | Year | Ref  |
|-----------------------------------|--------------------|-----------------|-----------------------------------------------|----------------------------------------------|------|------|
| Preoperative Model Tract          | ANN                | In-house        | 10 Patients 3 Months                          | Kapa Index: 0.78                            | 2016 | [186]|
| Syndrome Planning                 | CNN, LSTM          | In-house        | MRI & 3D US 12 Monochorionic Twin Pregnancies | Dice: 0.55 ± 0.06 & 0.79 ± 0.05 Jaccard: 0.73 ± 0.10 & 0.66 ± 0.08 AUC: 0.88 ± 0.06 & 0.84 ± 0.03 Sensitivity: 0.77 ± 0.10 & 0.73 ± 0.07 Specificity: 1.0 & 0.99 | 2018 | [189]|
| Stent Graft Modeling & Planning   | U-Net              | In-house        | 78 Images                                     | Distance Error: 1–3 mm                       | 2018 | [190]|
| Intraoperative Liver View         | CNN VGG16          | In-house        | 2016 Liver View 9504 Live Liver View 100 Patients | mAP: 85.9%                                   | 2020 | [191]|
| 4D Guidance & Construction        | U-Net              | In-house        | 600 Scans                                     | Dice: 0.5 mm Precision: 0.794 ± 0.065 Recall: 0.803 ± 0.047 Z-coverage: 0.790 ± 0.087 | 2021 | [185]|
| 3D Reconstruction of Wound Edge   | ANN                | LumSeg, xVertSeg, SpiSeg | Camera Images, Kinematic Info | MSE: 0.67                                    | 2021 | [187]|
| Vascular Access Planning          | YOLO-v3            | In-house        | 19000 Images                                  | Mean Time: 53 ± 36s                          | 2021 | [183]|
| Laminae Planning                  | SegReNet, DenseSeg, FC-DenseNet | In-house | 10 Scans 15 Scans 10 Scans | Dice: 96.38%                                 | 2021 | [188]|
| Automatic Ablation Planning       | LeNet-5            | In-house        | 20 OCT Volumes                                | Precision: 1.16 mm Error 0.74 mm             | 2021 | [200]|

1) SKILL ASSESSMENT
The manual skill assessment and skill development monitoring of the doctors, surgeons, and trainees is burdensome tasks and requires a great deal of time and expertise. This usual task comes under the responsibility of expert doctors, which is not only arduous but also prone to errors. The automatic surgical skill evaluation for the RAS is indispensable.

The surgical skill assessment and skill level assessment are widely realised using CNN models driven by video and kinematic data [208], [209], [210], [211], [212], [213], [214]. The JIGSAWS [36] is most common dataset for skill evaluation since it provides both video frames and kinematic data (Table 10). All the studies included in this section evaluated the models on accuracy and/or AUC. The skill assessment also handles the instrument tracking in some cases where the dexterity of surgeons carries extreme importance. The other performance measures, dataset description, and DL models are provided in the Table 10.

2) WORKFLOW RECOGNITION
Surgical workflow analysis describes the steps involved during the surgical interventions. Automatizing the surgical workflow has great importance in the modern operating room. Autonomous workflow recognition is vital in developing computer-aided autonomous and semiautonomous surgical frameworks. These systems have the ability to supervise the surgery within operating room by scheduling the tasks and resources and providing the seamless assistance to clinicians.

The interpretation of the recorded video of the surgical procedures requires expertise, focus, and huge amount of time. The technological advances automatically extract the valuable information by analysing the videos. The cholecystectomy procedure videos are largely used in literature to understand the workflow and surgery type recognition [109], [215], [216], [217], [218], followed by nephrectomy [202], [219], [220], [221]. The CNN and the LSTM are common methods used to study the workflow in the videos [109], [202], [215], [216], [217], [218], [219], [220], [221], [222], [223] because of their high performance. Eight out of ten studies used accuracy as a performance measuring metric along with others. The workflow recognition also overlaps with the future state prediction and phase recognition [221], [224]. The Table 11 contains the other necessary and relevant details about the section of the study.
TABLE 8. The summary of the studies dealing with the task of phase and state estimation using DL models.

| Objective                        | DL Model        | Dataset      | Data Description          | Results                  | Year | Ref |
|----------------------------------|-----------------|--------------|---------------------------|--------------------------|------|-----|
| State Perception                 | CNN             | JIGSAWS      | 39 Sequences 8 Users 10 Gesture Classes | Accuracy: 96%           | 2019 | [193]|
| Action Recognition               | ResNet Encoder-decoder | JIGSAWS | 39 Sequences 8 Users 10 Gesture Classes | Accuracy: 81.71% Edit Score: 91.74 F1: 80.08 | 2020 | [192]|
| Tool Presence Analysis           | RCN LSTM        | Cholec80     | 80 Videos                 | mAP: 89.1% F1 Score: 87.4% | 2020 | [87] |
| Pull State Recognition Needle Detection | YOLOv3 CNN | In-house     | 15505 Images              | Accuracy: 72.4% Accuracy: 93.2% | 2020 | [197]|
| Phase Recognition                | CNN RCN         | Bypass40     | 40 Surgical Procedures    | Accuracy: 90.9 ± 3.2 Precision: 85.6 ± 4.5 Recall: 84.0 ± 4 F1 Score: 84.2 ± 4.2 | 2021 | [196]|
| State Estimation                 | DNN U-Net LSTM  | HERNIA 20    | 20 Inguinal Hernia Repair Surgeries on da Vinci | Accuracy: 80.4%         | 2021 | [194]|
| State Estimation                 | LSTM            | HERNIA 20 RIOUS+ JIGSAWS | 20 Inguinal Hernia Repair Surgeries on da Vinci 40 Uson da Vinci 39 Sequences 8 Users 10 Gesture Classes | Accuracy: 82.7% Accuracy: 89.5% Accuracy: 85.6% | 2021 | [195]|
| Phase Recognition                | Temporal CN     | In-house     | 31 Patients’ Videos       | Accuracy: 84%           | 2022 | [198]|
| Phase Recognition                | Temporal CN ResNet-50 | M2CAI16 Cholec80 | 10 Procedures 2532 Frames 1280 x 720 Pixels 80 Videos | Accuracy: 91.8 ± 8.1 Precision: 90.3 ± 6.4 Recall: 90.0 ± 6.4 Jaccard: 81.2 ± 5.5 | 2022 | [199]|

TABLE 9. The summary of the articles incorporating DL methods for the task of activity recognition.

| Objective                        | DL Model        | Dataset      | Data Description          | Results                  | Year | Ref |
|----------------------------------|-----------------|--------------|---------------------------|--------------------------|------|-----|
| Action Segmentation              | VGG & AlexNet   | JIGSAWS & 50 Salad | 39 Sequences 8 Users 10 Classes 50 Instances 25 Users 2 Trials | Accuracy: 74.22% Accuracy: 72% | 2016 | [205]|
| Activity Recognition             | RP-Net Inception-v3 | In-house | 100 Videos 12 Tasks Each | Precision: 80.9% Recall: 76.7% | 2018 | [206]|
| Activity Recognition             | 3D ConvNet LSTM | In-house     | 400 Surgical Videos 103 Procedures 8 Surgeons | Precision: 88%           | 2020 | [201]|
| Surgery Type Recognition         | CNN LSTM        | Laparo425    | 425 Videos 9 Surgeries    | Accuracy: 75%           | 2020 | [202]|
| Surgical Action Recognition      | Deep CNN pretrained CNN | Lapgyn4DS | 30,682 Frames 8 Actions 500 Surgeries | Accuracy: 99.20% AUC: 99.12% | 2020 | [203]|
| Surgical Activity Recognition    | Multitask CNN ResNet-18 | CholecT40 | 40 Videos Cholec80 128 Triplets | Accuracy: 89.7% Action Triplet Recognition: 24.78% | 2020 | [204]|

E. OTHERS
This section describes the articles that do not come under the major categories, however their contribution towards the image-guided RAS is not negligible. The other reason of this distinctive but amalgam section formation is the small number of found publications for the relevant subclass.
Therefore, this section highlights the objective, contribution, DL methods adopted, and other significant details of each study.

In [225], the authors proposed dual neural network based models for organ recognition and presence or absence of internal organ in endoscopy image data. The second neural network model testifies the presence of the organ on series of images on the live screen. The in-house generation of small dataset resulted in 92% of the accuracy with only 200 randomly selected images for testing.

Similarly, the segmentation of organs [127], [226], [227] and tissues is also well studied task in RAS [170]. The Mask-RCN and CNN based YOLO, U-Net, TernausNet, LinkNet, and SegNet are applied on famous EndoVis Challenge and in-house datasets. The aforementioned articles coincide with the tool detection and segmentation category because of the segmentation of tool during organ detection.

In another similar study, the volume of organ segmentation in intraoperative guidance is studied using U-Net and V-Net architectures [228]. Two publicly available datasets namely VISCERAL and SLiver07 datasets are used in this study and 12.6% and 6.2% IoU for the aortic and liver segmentation are achieved.

The gaze estimation has an important role in tele-robotic surgery, however fewer studies have incorporated image data to support autonomous surgical systems. The [229] proposed a dense CNN architecture to control the surgical robot using gaze estimation. The local dataset generated from camera images is manually labelled in nine different gaze directions. The accuracy of the direction based gaze estimation reached 90%.

Smoke detection is another less common but growing application within the robotic surgery domain. The authors in [230], [231], and [232] use intraoperative images to detect and remove smoke using U-Net, GoogLeNet, and CycleGAN. The obtained results are considerable and detailed in the Table 12 for further reading.

An image based gauze detection and segmentation approach is proposed by Sanchez et al. [233] where pretrained models are employed to test the novel dataset of 4003 video provided by authors. The InceptionV3, MobileNetV2, and ResNet-50 managed to reach an accuracy value of 77.68%, 75.67%, 90.16%, respectively.

V. PUBLICLY AVAILABLE DATASETS

The impressive outcomes of the DL are underscored by the large amount of available datasets because of the intrinsic nature of neural networks. Neural network based models require training which follows the principle of more the data is available, better the results are. During this survey, a decent number of articles is found to have used publicly available datasets, however, numerous studies are conducted on proprietary datasets. The publicly available datasets do not only provide benchmark for DL model development but also support in evaluation and comparison of several models.

This section illustrates the datasets employed by the articles surveyed in this study. Furthermore, around 35 datasets are adopted by the surveyed articles, however, for the purpose of concision, only 10 datasets are described. The Table 13 contains the information of publicly available datasets including name, year of publication, modalities, and a short description. According to our study, the JIGSAWS is the most
TABLE 11. The summarised results of the flow recognition articles.

| Objective                        | DL Model     | Dataset          | Data Description                       | Results                      | Year  | Ref   |
|----------------------------------|--------------|------------------|----------------------------------------|------------------------------|-------|-------|
| Workflow Recognition             | RNN, CNN, ResNet, LSTM | Cholec80, MICCAI 2016 | 80 Videos 13 Surgeons 25 fps 854 × 480 Resolution 41 Videos 25 fps 1920 × 1080 Resolution | Accuracy: 90.7% Accuracy: 92.4% | 2018  | [217] |
| Self Supervised Workflow Analysis| ResNet-50    | Cholec80          | 80 Videos 13 Surgeons 25 fps 854 × 480 Resolution | Accuracy: 92.7 ± 4.3 Recall: 87.0 ± 4.0 Precision: 87.6 ± 5.3 F1 Score: 84.6 ± 5.4 | 2018  | [218] |
| Surgery Type Recognition         | CNN, LSTM    | Laparo425         | 425 Videos 9 Surgeries                | Accuracy: 75%                | 2019  | [202] |
| Automatic Workflow Analysis      | CNN, LSTM    | In-house          | 9 Videos 24 Hz 82.49 ± 37.54 Minutes Length | Accuracy: 100% Precision: 100% Recall: 100% | 2019  | [220] |
| Flow & Context Recognition       | RNN, LSTM    | In-house          | 9 Videos 24 Hz 82.49 ± 37.54 Minutes Length | Accuracy: 74.29%              | 2019  | [219] |
| Surgical Workflow Recognition    | ResNet-50    | In-house          | 8 Videos 1920 × 1080 Resolution 30 fps | Accuracy: 0.9482% Loss: 0.0765 | 2020  | [215] |
| Optical Flow Prediction           | U-Net        | Inhouse RIDE      | 700 Patients 1920 × 1080 Resolution 25 fps | s-EPE: 2.5 (2.6) l-EPE: 14.7 (7.9) Grid EPE 15.8 (7.9) | 2020  | [222] |
| Surgical Workflow Analysis       | RNN, ResNet-50 | Cholec80 | 80 Videos 13 Surgeons 25 fps 854 × 480 1920 × 1080 Resolution | Accuracy: 85.3 ± 6.96 Precision: 82.94 ± 6.20 Recall: 85.04 ± 5.15 Jaccard: 69.96 ± 8.83 F1 Score: 82.08 ± 6.45 | 2020  | [216] |
| Instrument and Workflow Recognition | Bayesian, AlexNet, LSTM | Cholec80 | 80 Videos 13 Surgeons 25 fps 854 × 480 1920 × 1080 Resolution | Bipolar: wMAE: 0.76 pMAE: 0.96 Scissors: wMAE: 0.51 pMAE: 0.76 | 2020  | [109] |
| 3D Workflow Analysis             | U-Net, ResNet-50 | In-house | 9 Videos 20 Minutes Length | Mean IoU: 80% | 2022  | [224] |
| Workflow Detection               | CNN, Mask R-CNN | M2CAI2016 | 10 Procedures 2532 Frames 1280 × 720 Pixels | mAP: 96.8 | 2022  | [221] |

employed dataset in the RAS that encompasses video and kinematic data (Figure 5).

A. JIGSAWS
The well-known gesture and skill assessment dataset JIGSAWS comprises of kinematic and video data for suturing, needle passing and knot-tying tasks. This dataset was recorded by 8 experts on da Vinci surgical system performing the five repetitions of suturing, knot-tying and needle passing procedures [36]. The Figure 5 shows that the JIGSAWS is most extensively applied dataset by the studies included in this survey, however, not all the authors have utilized the complete dataset but a part of the dataset. The JIGSAWS dataset contains 163 videos and kinematic data. The brief description of the dataset is provided in the Table 13.

B. MICCAI DATASETS
The Medical Image Computing and Computer Assisted Intervention (MICCAI), officially known as The MICCAI Society was established in July 2004 [37] as a non-profit organisation. The MICCAI Society holds the annual competitions with the
TABLE 12. The summary of the articles incorporating DL methods for miscellaneous tasks in robotic surgery realm.

| Objective                                | DL Model  | Dataset | Data Description                                                                 | Results           | Year   | Ref  |
|------------------------------------------|-----------|---------|-----------------------------------------------------------------------------------|-------------------|--------|------|
| Gaze Estimation to Control Robot         | Dense CNN | In-house| 9 Gaze Direction                                                                  | Accuracy: 90%     | 2018   | [229]|
| Organ Recognition                        | NN        | In-house| 200 Images Endoscopy Records                                                       | Accuracy: 92%     | 2019   | [230]|
| Volume Segmentation in CT Images         | 3D DCNN   | VISCERAL| 20 CT Volumes                                                                      | IoU: 12.6%        | 2020   | [225]|
|                                          | U-Net     | SLiver07| 20 CT Volumes                                                                      | IoU: 6.2%         |        |      |
|                                          | V-Net     |         |                                                                                   |                   |        |      |
| Pixel-wise Smoke Detection               | U-Net     | Hamlyn  | 21000 Images Cholec80 m2csai6-workflow                                             | MSE: 0.002 ± 0.001| 2020   | [170]|
| Smoke Detection                          | GoogLeNet | In-house| 4500 Images 8 Surgeries 30000 Images                                               | ROC: 0.98         | 2020   | [228]|
|                                          |           |         |                                                                                   | AUC: 0.92-0.97    |        |      |
| Organs Segmentation                      | U-Net     | EndoVis | 12 Videos 25 Hz 960 × 540 Resolution                                              | IoU: 78.31%       | 2021   | [231]|
|                                          | TenaurusNet|         |                                                                                   | Time: 28.07 ms    |        |      |
|                                          | LinkNet   |         |                                                                                   |                   |        |      |
|                                          | SegNet    |         |                                                                                   |                   |        |      |
|                                          | FCN       |         |                                                                                   |                   |        |      |
| Object Detection                         | YOLO-v4   | In-house| 5 Videos 398 Images                                                                | Accuracy: 90%     | 2022   | [127]|
| Organ Segmentation                       | Mask R-CNN| In-house| 55 Videos 8 Hospitals 1578 Images                                                  | Dice Score: 97.65%| 2022   | [226]|
| Smoke Removal                            | CycleGAN  | In-house| 10 Videos 6000 Images                                                              | Accuracy: 93%     | 2022   | [232]|
| Organ Segmentation                       | U-Net     | In-house| 20 Subjects 506 Images                                                             | mDSC: 0.90        | 2022   | [227]|
|                                          | EfficientNet-b5|     |                                                                                   |                   |        |      |
| Gaze Detection & Segmentation            | YOLOv3    | In-house| 4003 Videos                                                                       | Accuracy: 77.68, 75.67, 90.16 | 2022   | [233]|
|                                          | U-Net     |         |                                                                                   | IoU: 0.85         |        |      |
|                                          | InceptionV3|         |                                                                                   |                   |        |      |
|                                          | MobileNetV2|         |                                                                                   |                   |        |      |
|                                          | ResNet-50 |         |                                                                                   |                   |        |      |

purpose to bring the researchers, clinicians, and engineers together to advance in the field of biomedical engineering. The MICCAI offers a wide range of datasets each year with different objectives and holds the competitions. A worth considering number of articles take advantage of datasets provided by challenges for biomedical analysis.

The most common challenge of MICCAI is EndoVis challenge, which is organised every year since 2015. The sub-challenges of EndoVis contain datasets for wide variety of tasks including instrument detection, segmentation, boundary detection, workflow analysis, skill assessment, etc. Further information about the MICCAI and its EndoVis challenge can be found in the Table 13.

C. Cholec80

The Cholec80 is another famous cholecystectomy dataset. It comprises of 80 cholecystectomy surgery videos recorded during the surgical interventions performed by 13 experts at 25 Frames Per Second (FPS) [38]. The Cholec80 data is widely used for tool presence detection and phase recognition as evident by the Figure 5 and Table 13. The original data contains the videos where the tool cannot be easily visualised by naked eye which makes the detection and segmentation a challenging task.

The Cholec80 dataset is further split into two groups of 40 videos each, namely for training and testing. Several different versions of Cholec80 are also available in literature.

D. ATLAS DIONE

The ATLAS Dione dataset contains the video data of ten surgeons having different expertise level and record six different surgical procedures on da Vinci surgical system at Roswell Park Cancer Institute, Buffalo, New York, USA [39]. The surgical procedures include basic robotic surgery task to high-level surgical processes. The motivation behind this data generation was unavailability of annotations in JIGSAWS.
TABLE 13. The summary of the publicly available datasets used by the articles surveyed in this study.

| Name               | Data                  | Data Description                                                                 | Procedure                                      | Purpose                                    | Year  | Ref  |
|--------------------|-----------------------|----------------------------------------------------------------------------------|------------------------------------------------|--------------------------------------------|-------|------|
| MICCAI             | Video                 | 41 Videos                                                                        | Subject to Change                               | Detection Recognition                     | 2004  | [37] |
|                    |                       | 25 fps                                                                           |                                                | Recognition Segmentation                  |       |      |
|                    |                       | 1920 × 1080 Resolution                                                           |                                                | Gesture Recognition Skill Assessment       |       |      |
| JIGSAWS            | Video                 | 103 Videos                                                                       | Suturing                                       |                                            |       |      |
|                    | Kinematic Data        | 39 Sequences                                                                     | Knot Tying                                     |                                            |       |      |
|                    |                       | 8 Users                                                                           | Needle Passing                                 |                                            |       |      |
|                    |                       | 10 Gesture Classes                                                               |                                                |                                            |       |      |
| Cholec80           | Video                 | 80 Videos                                                                        | Cholecystectomy Surgery                        |                                            | 2016  | [234]|
|                    |                       | 13 Surgeons                                                                       |                                                | Phase Recognition Tool Detection          |       |      |
|                    |                       | 25 fps                                                                           |                                                |                                            |       |      |
|                    |                       | 854 × 480 Resolution                                                             |                                                |                                            |       |      |
| M2CAI16 Tool       | Video                 | 15 Procedures                                                                    | Cholecystectomy Surgery                        | Tool Detection                             | 2016  | [235]|
|                    |                       | 2532 Frames                                                                      |                                                |                                            |       |      |
| M2CAI16            | Video                 | 41 Videos                                                                        | Cholecystectomy Surgery                        | Work Flow Analysis                        | 2016  | [236]|
| Workflow           |                       |                                                                                  |                                                |                                            |       |      |
| M2CAI16 Location   | Video                 | 3141 Annotations                                                                 | Cholecystectomy Surgery                        | Tool Detection                             | 2016  | [235]|
|                    |                       | 7 Instrument                                                                     |                                                | Skill Assessment                           |       |      |
|                    |                       | 10 Videos                                                                        |                                                |                                            |       |      |
| ATLAS Dione        | Video                 | 10 Surgeons                                                                       | 6 Different Surgeries                          | Tool Detection                             | 2017  | [39] |
|                    |                       | 6 Tasks on daVinci                                                               |                                                | Action Recognition                         |       |      |
|                    |                       | 99 Videos                                                                        |                                                |                                            |       |      |
|                    |                       | 854 × 480 Pixels                                                                 |                                                |                                            |       |      |
|                    |                       | 22467 Images                                                                      |                                                |                                            |       |      |
|                    |                       | 32000 Images                                                                      |                                                |                                            |       |      |
|                    |                       | 4400 Images                                                                       |                                                |                                            |       |      |
| DeepLision         | CT Images             | Whole Body                                                                        | Lesion Detection                               |                                            | 2018  | [237]|
|                    |                       |                                                                                  | Semantic Segmentation                          |                                            |       |      |
| Lapro425           | Video                 | 425 Videos                                                                        | Laparoscopy                                    |                                            | 2019  | [202]|
|                    |                       | 9 Surgeries                                                                       | Flow Analysis                                  |                                            |       |      |
| UCL dVRK           | Video                 | 14+6 Videos                                                                       | Different Surgeries                            | Tool Segmentation                          | 2020  | [132]|
|                    | Kinematic Data        | 300 Frames Each                                                                  |                                                |                                            |       |      |
|                    |                       | 720 × 576 Pixels                                                                 |                                                |                                            |       |      |
| CholecT40          | Video                 | 40 Videos                                                                        | Cholecystectomy Surgery                        | Action Recognition                         | 2020  | [204]|
|                    |                       | 13 Surgeons                                                                       |                                                |                                            |       |      |
|                    |                       | 25 fps                                                                           |                                                |                                            |       |      |
|                    |                       | 854 × 480 Resolution                                                             |                                                |                                            |       |      |
| Sinus Surgery-C    | Image & Video         | 10 Videos                                                                        | Sinus Endoscopy                                | Smoke Detection                            | 2020  | [238]|
|                    |                       | 5-32 Minute                                                                       | Instrument Segmentation                        | Tool Shadow                                |       |      |
|                    |                       | 320 × 240 Resolution                                                             |                                                | Shadow Instrument                          |       |      |
|                    |                       | 30 fps                                                                           |                                                | Segmentations                              |       |      |
|                    |                       |                                                                                  |                                                | Tissue Segmentation                        |       |      |
|                    |                       |                                                                                  |                                                | Tool Segmentation                          |       |      |
| FlapNet            | Video                 | 62 Minute Videos                                                                  | Lobectomy                                      |                                            | 2020  | [166]|

FIGURE 5. Frequency of the used datasets by the number of studies. Articles that adopted two or more datasets are counted respectively.

dataset. The videos in the dataset are annotated in the frames of 854 × 480 pixels each and the annotations are provided in XML format. Further details about the ATLAS Dione dataset can be found in the Table 13.

E. UCL dVRK DATASET
The UCL da Vinci Research Kit (dVRK) comprises 14 videos of 300 frames each of segmentation task. It also contains six videos of robotic kinematic recorded at 300 frames each. All the recorded frames are 720 × 576 pixels. The frames contained the camera artefacts which are later cropped at 720 × 576 pixel resolution. The dataset is recorded using four consecutive steps that are repeated to record the video and kinematic data. For the further information on the setup and the data acquisition methods, interested reads are referred to the [132].

F. M2CAI16 DATASET
The M2CAI16 challenge is a satellite event of MICCAI offering two different datasets including workflow and tool detection, which are enlisted hereunder as well as in the Table 13.

1) M2CAI16-TOOLS DATASET
This dataset was generated with the collaboration of University Hospital of Strasbourg and the Hospital Klinikum Rechts
der Isar in Munich, Germany [39]. The dataset contains 41 laparoscopic procedural videos of the cholecystectomy with eight distinct phases. Out of the total 41 videos, 27 and 14 videos are for training and testing purposes, respectively.

2) M2CAI16-WORKFLOW DATASET
The M2CAI16-Workflow dataset was generated at University Hospital of Strasbourg [234]. Similar to the previous data, this dataset also contains laparoscopic videos of cholecystectomy. A total of 15 procedures were recorded, ten and five for training and testing purposes, respectively.

3) M2CAI16-LOCATION DATASET
The M2CAI16-Location dataset is extension of the aforementioned M2CAI16-Tool dataset [235]. It additionally contains the annotations of the tools and locations.

G. Laparo425
The Laparo425 dataset contains the laparoscopy videos of nine distinctive classes. The dataset was recorded at the University Hospital of Strasbourg. As its name suggests, it contains 425 procedures recorded at 25 FPS and down-sampled at one FPS for experiments.

H. CholecT40
The CholecT40 datasets come from the Cholec80 dataset. It contains 40 videos from the Cholec80 dataset and annotates them to action triplets. The 128 different action triplets are introduced and recognised using video data [204].

I. SINUS-SURGERY-C DATASET
This dataset contains the endoscopic sinus surgery images which are annotated for surgical tool segmentation task. The dataset comes from the BioRobotics Lab at the University of Washington, USA [238]. The segmentation of the dataset is not easy since it contains smoke and shadows of the instruments. As soon as the movement occurs, the blue images generate that make the dataset more challenging for instrument segmentation tasks.

J. DeepLesion
The DeepLesion dataset is large dataset of National Institute of Health of USA that contains 32000 lesions of the CT images. The total number of 4400 unique patients were involved during the data generation [237]. This dataset offers a great diversity because of its diverse collection of images from liver, lungs, lymph, and many more human body organs. The further details about all datasets are also provided in the Table 13.

VI. CHALLENGES AND RISKS
This section describes the potential pitfalls that hinder the development of autonomous robotic surgical systems. The first part of the section illustrates the technical and technological challenges that are either under development phases or soon will be. However, the second part demonstrates the legal and ethical concerns of the DL guided computer-aided interventions.

A big thanks to the advancement in the AI, the healthcare industry has not only improved the diagnostic methods but also moved to surgical robots. However, the other side of the DL in medical domain is still darker. The technical risks include the design of the robotic components, the precision in complex scenarios, and unpredictability of the amount of surgical procedure assignment [239] among numerous others. The success rate of surgical systems is increasing day by day [239], nevertheless the high associated costs are not negligible [240]. These costs can be traced back to component design and continue through the operating room to the maintenance expenditures.

Additionally, the surgical systems are designed with limited allowed movement and the dexterity, which have a positive and a negative side. The controlled movement will fail if any unscheduled task originates during the surgery because of no or less adaptability expertise [241]. In the meantime, a robotic system can physically harm the patient and damage the involved components. The safety of the patient under the surgery is at risk, not only by hardware part of the robotic system but also by the software [242].

In a semi-autonomous surgical system, a minute human error can put the life of patient in danger [243]. The training and the testing capabilities of human are not as precise as machine, however completely relying on machine with present day technology could produce less efficient outcomes in general [244], [245]. Therefore, all these concerns require detailed studies and considerations.

Similarly, all the perks DL offers are appreciable, however, there are few things beyond technology. The legal and ethical issues concern the privacy of the medical data of the patients [246], [247], [248]. Another potential threat is cybercrime involving the genetic information of the patient and potential hack of the robotic system [249]. The prospective cyber attacks to seize the control of surgical system can lead to devastating results including lethal physical harm to patients, as discussed by Bonaci et al. [250]. The software designer and the developers should come forward and certify that the delivered products are not vulnerable to attacks. An automatic rescue service should be activated in cases of emergency i.e. power cuts, jamming, transmission, etc. Unlike autonomous driving, there exist no legal standard methods to define the level of degree to which a robotic system can be autonomous [251]. The National Health Service (NHS) of the United Kingdom came under similar attack back in 2017 which affected medical devices nationwide [252]. Moreover, the transparency is another worrisome issue because of the blackbox decision making nature of the DL models. The recent advent of eXplainable AI has taken care of transparency and interpretability problems, however, the field is growing and requires further exploration. Finally, all the aforementioned applications are accelerating the dread of physical harm by the AI systems and eventually reduce the human involvement. The additional intervention of government and giant companies spreads fear among the undergoing patient that demands and requires a code of conduct.
The General Data Protection Regulation (GDPR) by European Union also states the concise and transparent information provision and privacy protection of users [253]. This ensures the maximal control, however the full enforcement of the law across the board is yet to be experienced.

Furthermore, the fully autonomous robotic surgical system should first go through several security checks to answer the concerns of physicians, engineers, patients and general public. Who will be responsible in case of harm to the undergoing patient or patient’s data? Is it the robot? Or the engineer developing the robot? Or anyone else? Successfully building the autonomous robotic systems will leave nothing to human but the responsibility which brings culpability and liability. The authors, in no case, are against the development of fully autonomous robots, rather demand a proper regulation before the train is missed.

Conclusively, the cyber-security of the robotics and particularly the surgical robotics is big market where researcher need to explore because a medical robotic system can not be judged based on average results but best possible outcomes are indispensable (better than human), otherwise, a medical robot looses the whole point to be developed for surgical applications.

VII. DISCUSSION

During the recent decade, the DL models have made substantial impact in healthcare domain (e.g. CAD systems). The large scale availability of surgical datasets and straightforward data acquisition protocols encourage cross-domain research synergies in order to reach fully autonomous robotic surgery. Recently, numerous DL models have been applied on medical images to capture the relevant information and provide to the RAS system for surgical procedures (Figure 1).

This survey analyses and summarises the contribution of the DL architectures in the field of image-driven surgical robots. The analysis reveals the current interests of researchers in image-guided DL based RAS is increasing over time. Inside the RAS, the majority of articles contribute to the tool detection and segmentation tasks. One reason behind detection and segmentation task selection is the serene accessibility to enormous surgical tool data. The other reason can be the working mechanism of DL models. Since the DL models learn from patterns, the tools segmentation tasks are arduous because of the presence of smoke, reflection of instruments, and vessels in the surgical dataset.

During this study, the authors found that CNN is most commonly employed DL method because of its successful history with image data. The Figure 4 and Table 2 through Table 11 are evident of the aforementioned claim. Not only the CNN is most applied method, but also it outperforms other techniques. However, as depicted in the Table 2, the combination of several different models yields improved results. Most of the articles dealing with action recognition applications (e.g. surgical gesture recognition and segmentation, trajectory segmentation, phase and state estimation, etc.) consider the combination of traditional static CNN architectures along with the RNN/LSTM models. The combination of CNN and RNN lies in the fact that CNN are able to automatically extract spatial features within images, whereas RNN have been designed to capture the temporal information. However, it is worth noting that not all the articles related to action segmentation employ a RNN. In fact, some authors choose 3D CNNs that are able to simultaneously capture both spatial and temporal features of the action/motion behaviour. These kind of solutions can directly extract information from multiple video frames through 3D convolutions. Comparing the two different approaches it is not easy to determine which approach is the best, a future work might investigate deeper this aspect. However, it can surely stated that the main limitation of 3D CNN is related to the increased dimension due to 3D convolutions, thus a bigger amount of data and time is needed for training such models.

Similarly, the most widely used dataset is JIGSAWS that contains video and kinematic data (Figure 5 and Table 13). Authors proved that the fusion of video data into the kinematic data leads to better performance outcomes. The annual competition organised by MICCAI for several varying imaging tasks also brings new datasets. Although huge number of studies employ in-house datasets, a total number of 35 different publicly available datasets are used either partially or fully.

This fact reveals that the amount of available data is significant and has attracted scientists and boosted the research. Another benefit of extensive data availability is the possibility of comparison between the novel conducted studies.

A considerable number of papers have taken benefit of pretrained networks, as can be seen through the Table 2 and Table 11. These models are heavily trained on ImageNet dataset [207]. The biggest reason of using pretrained nets is the unavailability of sufficient amount of data to train a DL architecture from the scratch. In the surgical workflow analysis task, numerous authors adopted in-house datasets which led them to use pretrained models. Since these models are pretrained, therefore, they come with the added benefit of lower time consumption. The most commonly used pretrained net remained VGG architecture followed by the ResNet and MobileNet.

With the increasing development of DL models, and the massive success with image modalities, the research in the field of image-guided robotic surgery is inevitably growing. Firstly, image data acquisition methods are smooth with least harm (tolerable) and do not require colossal acquisition cost. Secondly, image data contains bulk of information about the patient and helps understand internal state of patient without incision or physical injection.

Moreover, analysis and preprocessing on the image modalities is relatively easy and also, the DL methods work better with the image data. Therefore, the future of MIS, RAS and computer-aided interventions relies mainly on the image and video data, and of course DL.

Finally, based on the exhaustive analysis, the authors point out to the two major concerns, first for medical experts
and other for DL engineers. The recorded image modalities and videos contain smoke, instrument reflections, and surgeon’s movements (i.e. hands, reflections), therefore, the DL algorithms learn from the unnecessary features which may bias the results. Secondly, not only the DL methods require huge amount of time for training purposes that needs to be optimized, but also DL is blackbox in decision making mechanism which concerns the clinicians and the patients.

VIII. CONCLUSION

In this study, technical articles concerning image driven computer-assisted interventions incorporating DL models are selected for survey from Scopus database. The intensive text assessment indicated that the selected 184 articles can be grouped into four categories including: 1) Surgical Tools, 2) Surgical Processes, 3) Surgical Surveillance, and 4) Surgical Performance/Assessment. The key findings include: a) Surgical Tools is most studied topic which comprises Surgical Tool Detection and Surgical Tool Segmentation (45% of total articles), b) CNN is most widely applied DL topology (roughly 54% of total articles), c) the gesture recognition articles incorporate JIGSAWS dataset (around 77% of articles in relevant subcategory), whereas MICCAI datasets are top consideration for detection and segmentation tasks (around 60% of articles in relevant subcategory), d) VGG remains the widely accepted pretrained network especially when available dataset was not large enough, f) the most studied applications appear to be cholecystectomy and prostatectomy, g) for gesture and trajectory applications, suturing task is frequently studied application area, h) the fusion of kinematic data with image data yields better results. Considering the characteristics of the proposed survey, the authors believe that the main limitation of the study concerns the lack of deep details of the pre-processing and processing approaches proposed in the reviewed paper to solve the problems related to each application category. However, this was not the original scope of the review, since providing the general overview of the topics under discussion required gigantic efforts. In the context of future direction, the development of fully autonomous RAS system appears highly promising and fascinating research topic. Additionally, self-supervised learning based models can greatly improve the environment in operating room. Finally, a steady walk from the weak AI to strong AI and to the super AI can lead to the notable breakthroughs.

REFERENCES

[1] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, “Artificial intelligence in healthcare: Past, present and future,” Stroke Vascular Neurol., vol. 2, no. 4, pp. 1–14, 2017.
[2] A. Panesar, Machine Learning and AI for Healthcare. Cham, Switzerland: Springer, 2019.
[3] M. Y. Shaheen, “Applications of artificial intelligence (AI) in healthcare: A review,” ScienceOpen, Burlington, MA, USA, Tech. Rep. PPVRY3K-v1, 2021.
[4] D. C. Birkhoff, A. S. H. V. Dalen, and M. P. Schijven, “A review on the current applications of artificial intelligence in the operating room,” Surgical Innov., vol. 28, no. 5, 2021, Art. no. 1553350621996061.
[5] Y.-W. Chen and L.-C. Jain, Deep Learning in Healthcare. Cham, Switzerland: Springer, 2020.
[6] T. M. Ward, P. Mascagni, Y. Ban, G. Rosman, N. Padoy, O. Meireles, and D. A. Hashimoto, “Computer vision in surgery,” Surgery, vol. 169, no. 5, pp. 1253–1266, May 2021.
[7] M. C. E. Simsekler, C. Rodrigues, A. Qazi, S. Ellahham, and A. Ozonoff, “A comparative study of patient and staff safety evaluation using tree-based machine learning algorithms,” Rel. Eng. Syst. Saf., vol. 208, Apr. 2021, Art. no. 107416.
[8] J. L. Fenc, C. Willoughby, and K. Jackson, “Just culture: The foundation of staff safety in the perioperative environment,” AORN J., vol. 113, no. 4, pp. 329–336, Apr. 2021.
[9] A. Rehman, M. A. Butt, and M. Zaman, “A survey of medical image analysis using deep learning approaches,” in Proc. 5th Int. Conf. Comput. Methodologies Commun. (ICCMC), Apr. 2021, pp. 1334–1342.
[10] N. Kumar and M. Raubal, “Applications of deep learning in congestion detection, prediction and alleviation: A survey,” Transp. Res. C. Emerg. Technol., vol. 133, Dec. 2021, Art. no. 103432.
[11] H. Huyhn Nguyen and U. A. Buy, “Toward gesture recognition in robot-assisted surgical procedures,” in Proc. 2nd Int. Conf. Societal Autom., May 2021, pp. 1–4.
[12] Z. Akkus, J. Cai, A. Boomrod, A. Zeinoddini, A. D. Weston, E. Schilbrink, et al., “Survey of deep-learning applications in ultrasound: Artificial intelligence-powered ultrasound for improving clinical workflow,” J. Amer. College Radiol., vol. 16, no. 9, pp. 1318–1328, 2019.
[13] Y. Ming, Y. Cheng, Y. Jing, L. Liangzhe, Z. Guang, and C. Feng, “Surgical skills assessment from robot assisted surgery video data,” in Proc. IEEE Int. Conf. Power Electron., Comput. Appl. (ICPECA), Jan. 2021, pp. 392–396.
[14] Y. Liu, J. Jiang, and J. Sun, “Hand pose estimation from RGB images based on deep learning: A survey,” in Proc. IEEE 7th Int. Conf. Virtual Reality (ICVR), May 2021, pp. 82–89.
[15] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, “Deep learning for sensor-based activity recognition: A survey,” Pattern Recognit. Lett., vol. 119, pp. 3–11, Mar. 2019.
[16] 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, vol. 169, no. 5, pp. 1240–1244, May 2021.
[17] Z. Chua, A. M. Jac, and A. M. Okamara, “Toward force estimation in robot-assisted surgery using deep learning with vision and robot state,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2021, pp. 12335–12341.
[18] S. Ashiquzzaman, S. M. Oh, D. Lee, J. Lee, and J. Kim, “Context-aware deep convolutional neural network application for fire and smoke detection in virtual environment for surveillance video analysis,” in Smart Trends in Computing and Communications: Proceedings of SmartCom 2020. Singapore: Springer, 2021, pp. 459–467.
[19] J. Tang, S. Li, and P. Liu, “A review of lane detection methods based on deep learning,” Pattern Recognit., vol. 111, Mar. 2021, Art. no. 107623.
[20] R. Magno, L. Rocchi, R. Dainelli, A. Mateo, S. F. Di Gennaro, C.-F. Chen, N.-T. Son, and P. Toscano, “AgroShadow: A new Sentinel-2 cloud shadow detection tool for precision agriculture,” Remote Sens., vol. 13, no. 6, p. 1219, Mar. 2021.
[21] H. Boudegga, Y. Elloumi, M. Akil, M. H. Bedoui, R. Rachouri, and A. B. Abdallah, “Fast and efficient retinal blood vessel segmentation method based on deep learning network,” Computerized Med. Imaging Graph., vol. 90, Jun. 2021, Art. no. 101902.
[22] G. Wadhwa and A. Kaur, “Various image modalities used in computer-aided diagnosis system for detection of breast cancer using machine learning techniques: A systematic review,” in Soft Computing and Signal Processing, Singapore: Springer, 2022, pp. 281–292.
[23] S. V. M. Sagheer and S. N. George, “A review on medical image denoising algorithms,” Biomed. Signal Process. Control, vol. 61, Aug. 2020, Art. no. 102306.
[24] Y. Tian and S. Fu, “A descriptive framework for the field of deep learning applications in medical images,” Knowl.-Based Syst., vol. 210, Dec. 2020, Art. no. 106445.
[25] S. Albawi, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network,” in Proc. Int. Conf. Eng. Technol. (ICET), Aug. 2017, pp. 1–6.
[26] W. Wang, Y. Yang, X. Wang, W. Wang, and J. Li, “Development of a convolutional neural network and its application in image classification: A survey,” Opt. Eng., vol. 58, no. 4, 2019, Art. no. 040901.
S. M. Hussain et al.: DL Based Image Processing for Robot Assisted Surgery: A Systematic Literature Survey

VOLUME 10, 2022

J. minimal [27]
J. Li, Y. Fang, Z. Jin, Y. Wang, and M. Yu, "The impact of robot-assisted spine surgeries on clinical outcomes: A systematic review and meta-analysis," Int. J. Med. Robot. Comput. Assist. Surg., vol. 16, no. 6, pp. 1–14, Dec. 2020.

C. R. Garrow, K.-F. Kowaleski, L. Li, M. Wagner, M. W. Schmidt, S. Engelhard, D. A. Hashimoto, H. G. Kengott, S. Bodenstein, S. Speidel, B. P. Müller-Stich, and F. Nickel, "Machine learning for surgical phase recognition: A systematic review," Ann. Surg., vol. 273, no. 4, pp. 684–693, Jan. 2021.

E. Yanik, X. Intes, U. Kruger, P. Yan, D. Diller, B. Van Voorst, B. Makled, J. Norfleet, and S. De, "Deep neural networks for the assessment of surgical skills: A systematic review," J. Defense Model. Simul., Appl., Methodol., Technol., vol. 19, no. 2, pp. 159–171, Apr. 2022.

G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: A survey," Mach. Vis. Appl., vol. 31, nos. 1–2, pp. 1–18, Feb. 2020.

A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Gligovikov, "Automatic instrument segmentation in robot-assisted surgery using deep learning," in Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Dec. 2018, pp. 624–628.

D. Bouget, M. Allan, D. Stoyanov, and P. Jamin, "Vision-based and marker-less surgical tool detection and tracking: A review of the literature," Med. Image Anal., vol. 35, pp. 633–654, Jan. 2017.

I. Rivas-Blanco, C. J. Perez-Del-Pulgar, I. Garcia-Morales, and V. F. Munoz, "A review on deep learning in minimally invasive surgery," IEEE Access, vol. 9, pp. 48658–48678, 2021.

M. Unberath, C. Gao, Y. Hu, M. Judish, R. H. Taylor, M. Armand, and R. Grupp, "The impact of machine learning on 2D/3D registration for image-guided interventions: A systematic review and perspective," Frontiers Robot. AI, vol. 8, p. 260, Aug. 2021.

C. Yang, Z. Zhao, and S. Hu, "Image-based laparoscopic tool detection and tracking using convolutional neural networks: A review of the literature," Comput. Assist. Surg., vol. 25, no. 1, pp. 15–28, Jan. 2020.

T. O. Ayodele, "Types of machine learning algorithms," in New Advances in Machine Learning, vol. 3. Rijeeka, Croatia: InTech, 2010, pp. 19–48.

I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," Softw. Pract. Exper., vol. 2, no. 3, pp. 1–21, May 2021.

X. Li and Q. Tang, "Introduction to artificial intelligence and deep learning with a case study in analyzing electronic health records for drug development," in Real-World Evidence in Drug Development and Evaluation, London, U.K.: Chapman-Hall, 2021, pp. 151–172.

L. Zajmi, F. Y. H. Ahmed, and A. A. Jaharadak, "Concepts, methods, and performances of particle swarm optimization, backpropagation, and neural networks," Appl. Comput. Intell. Soft. Comput., vol. 2018, pp. 1–7, Sep. 2018.

Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in Proc. Adv. Neural Inf. Process. Syst., vol. 19, pp. 1–15, 2006.

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

L. Li, M. Doroslovački, and M. H. Loew, "Approximating the gradient of cross-entropy loss function," IEEE Access, vol. 8, pp. 111626–111635, 2020.

Open Source Machine Learning Framework by Google Brain, Accessed: May 16, 2022. [Online]. Available: https://www.tensorflow.org/

Open Source Machine Learning Framework by Facebook, Accessed: May 16, 2022. [Online]. Available: https://pytorch.org/

Numeric Computing Environment Developed by Mathworks, Accessed: May 16, 2022. [Online]. Available: https://www.mathworks.com/products/matlab.html

GPU based NVIDIA Caffe of Berkeley Vision and Learning Center, Accessed: May 16, 2022. [Online]. Available: https://ngc.nvidia.com/catalog/categories/nvidia/caffe

Google Inc. Keras Library, Accessed: May 16, 2022. [Online]. Available: https://keras.io/

M. P. Fried, J. Kleeﬁeld, H. Gopal, E. Reardon, B. T. Ho, and F. A. Kuhn, "Image-guided endoscopic surgery: Results of accuracy and performance in a multicenter clinical study using an electromagnetic tracking system," Laryngoscope, vol. 107, no. 5, pp. 594–601, 1997.

A. Reiter, P. K. Allen, and T. Zhao, "Articulated surgical tool detection using virtually-rendered templates," Comput. Assist. Radiol. Surg., vol. 10, pp. 1–8, Jun. 2012.

R. Elfring, M. de la Fuente, and K. Radermacher, "Assessment of optical localizer accuracy for computer aided surgery systems," Comput. Aided Surg., vol. 15, nos. 1–3, pp. 1–12, Feb. 2010.

Y. Hu, H. U. Ahmed, C. Allen, D. Pendsè, M. Sahu, M. Emberton, I. Hawkes, and D. Barratt, "MR to ultrasound image registration for guiding prostate biopsy and interventions," in Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. Cham, Switzerland: Springer, 2009, pp. 787–794.

A. Boonkong, D. Hornmee, S. Sonsilphong, and K. Khampitak, "Surgical instrument detection for laparoscopic surgery using deep learning," in Proc. 19th Int. Conf. Electr. Electron./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON), May 2014, pp. 1–15.

K. Mishra, R. Sathish, and D. Sheet, "Tracking of retinal microsurgery updates," in Proc. Int. Conf. Electr. Electron./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON), May 2014, pp. 358–366.

Z. Chen, Z. Zhao, and X. Cheng, "Surgical instruments tracking based on deep learning with lines detection and spatio-temporal context," in Proc. Clin. Autom. Congr. (CAC), Oct. 2017, pp. 2711–2714.
B. Choi, K. Jo, S. Choi, and J. Choi, “Surgical-tools detection based on convolutional neural network in laparoscopic robot-assisted surgery,” in Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2017, pp. 1756–1759.

T. Probst, K.-K. Maninis, A. Chhatkuli, M. Ourak, E. V. Poorten, and L. Van Gool, “Automatic tool landmark detection for stereo vision in robot-assisted retinal surgery,” IEEE Robot. Autom. Lett., vol. 3, no. 1, pp. 612–619, Jan. 2018.

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 Robot. Autom. Lett., vol. 4, no. 3, pp. 2714–2721, Jul. 2019.

N. Banerjee, R. Sathish, and D. Sheet, “Deep neural architecture for localization and tracking of surgical tools in cataract surgery,” in Computer Aided Intervention and Diagnostics in Clinical and Medical Images, Cham, Switzerland: Springer, 2019, pp. 31–38.

I. Ullah, P. Chikontwe, and S. H. Park, “Guidewire tip tracking using U-Net with shape and motion constraints,” in Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAINC), Feb. 2019, pp. 215–217.

M. Zhou, X. Wang, J. Weiss, A. Eslami, K. Huang, M. Maier, C. P. Lohmann, N. Navab, F. Knoll, and M. A. Nasserri, “Needle localization for robot-assisted subretinal injection based on deep learning,” in Proc. Int. Conf. Robot. Autom. (ICRA), May 2019, pp. 8727–8732.

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 Technol. Lett., vol. 6, no. 6, p. 275, 2019.

L. Qiu, C. Li, and H. Ren, “Real-time surgical instrument tracking in robot-assisted surgery using multi-domain deep convolutional neural network,” Healthcare Technol. Lett., vol. 6, no. 6, pp. 159–164, Dec. 2019.

K. Jo, Y. Choi, J. Choi, and J. W. Chung, “Robust real-time detection of laparoscopic instruments in robot surgery using convolutional neural networks with motion vector prediction,” Appl. Sci., vol. 9, no. 14, p. 2865, Jul. 2019.

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.

L. Guinot, R. Tsumura, S. Inoue, and H. Iwata, “Development of a needle deflection detection system for a CT guided robot,” in Proc. IEEE/SICE Int. Symp. Syst. Int. (SII), Jan. 2020, pp. 34–38.

L. Yu, P. Wang, Y. Yan, Y. Xia, and W. Cao, “MASSD: Multi-scale attention single shot detector for surgical instruments,” Comput. Biol. Med., vol. 123, Aug. 2020, Art. no. 103657.

J. Kim, C. Ham, P. C. Bhattacharjee, P. Gendebien, G. D. Hager, I. Jordachita, and M. Kobilarov, “Autonomously navigating a surgical tool inside the eye by learning from demonstration,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2020, pp. 7351–7357.

T. Cai and Z. Zhao, “Convolutional neural network-based surgical instrument detection,” Technol. Health Care, vol. 28, pp. 81–88, Jun. 2020.

I. Park, H. K. Kim, W. K. Chung, and K. Kim, “Deep learning based real-time OCT image segmentation and correction for robotic needle insertion systems,” IEEE Robot. Autom. Lett., vol. 5, no. 3, pp. 4517–4524, Jul. 2020.

S. Yin and A. Yuxchenko, “Object recognition of the robot system with using a parallel convolutional neural network,” in Robotics: Industry 4.0 Issues & New Intelligent Control Paradigms, Cham, Switzerland: Springer, 2020, pp. 3–11.

A. Nakazawa, K. Harada, M. Mitsushi, and P. Jannin, “Real-time surgical needle detection using region-based convolutional neural networks,” Int. J. Comput. Assist. Radiol. Surg., vol. 15, no. 1, pp. 41–47, Jan. 2020.

J. Yin, 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,” Med. Image Anal., vol. 59, Jan. 2020, Art. no. 101572.

N. Sachdeva, M. Klopukh, R. S. Clair, and W. E. Hahn, “Using conditional generative adversarial networks to reduce the effects of latency in robotic telesurgery,” J. Robot. Surg., vol. 15, no. 4, pp. 1–7, 2020.

Y. Yang, Z. Zhao, P. Shi, and S. Hu, “An efficient one-stage detector for real-time surgical tools detection in robot-assisted surgery,” in Proc. Ann. Conf. Med. Image Understand. Anal. Cham, Switzerland: Springer, 2021, pp. 18–29.

S. M. Cho, Y.-G. Kim, J. Jeong, I. Kim, H.-J. Lee, and N. Kim, “Automatic tip detection of surgical instruments in biportal endoscopic spine surgery,” Comput. Biol. Med., vol. 133, Jun. 2021, Art. no. 104384.
L. Seenivasan, S. Mitheran, M. Islam, and H. Ren, “Global-reasoned multi-frame feature aggregation for real-time instrument segmentation in endoscopic video,” IEEE Robot. Autom. Lett., vol. 6, no. 4, pp. 6773–6780, Oct. 2021.

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 Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2020, pp. 9939–9945.

E. Colleoni, P. Edwards, and D. Stoyanov, “Synthetic and real inputs for tool segmentation in robotic surgery,” in Proc. Int. Conf. Med. Image Comput. Assist. Intervent. Cham, Switzerland: Springer, 2020, pp. 700–710.

S. M. Hussain et al.: DL Based Image Processing for Robot Assisted Surgery: A Systematic Literature Survey

S. Lin, F. Qin, H. Peng, R. A. Bly, K. S. Moe, and B. Hannaford, “Multi-frame feature aggregation for real-time instrument segmentation in endoscopic video,” IEEE Robot. Autom. Lett., vol. 6, no. 4, pp. 6773–6780, Oct. 2021.

M. Islam, D. A. Atputharuban, R. Ramesh, and H. Ren, “Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning,” IEEE Robot. Autom. Lett., vol. 4, no. 2, pp. 2188–2195, Apr. 2019.

J. Wang, Y. Jin, L. Wang, S. Cai, P.-A. Heng, and A. Qin, “Efficient global-local memory for real-time instrument segmentation of robotic surgical video,” in Proc. Int. Conf. Med. Image Comput. Assist. Intervent. Cham, Switzerland: Springer, 2021, pp. 341–351.

S. Lin, F. Qin, H. Peng, R. A. Bly, K. S. Moe, and B. Hannaford, “Multi-frame feature aggregation for real-time instrument segmentation in endoscopic video,” IEEE Robot. Autom. Lett., vol. 6, no. 4, pp. 6773–6780, Oct. 2021.

S. Lin, F. Qin, H. Peng, R. A. Bly, K. S. Moe, and B. Hannaford, “Multi-frame feature aggregation for real-time instrument segmentation in endoscopic video,” in Proc. Int. Conf. Med. Image Comput. Assist. Intervent. Cham, Switzerland: Springer, 2021, pp. 341–351.

X. Sun, Y. Bu, and Y. Fu, “Lightweight deep neural network for real-time instrument semantic segmentation in robot assisted minimally invasive surgery,” IEEE Robot. Autom. Lett., vol. 6, no. 2, pp. 3870–3877, Apr. 2021.

Y. D. Wu, X.-L. Xie, G.-B. Bian, Z.-G. Hou, X.-R. Cheng, S. Chen, S.-Q. Liu, and Q.-L. Wang, “Automatic guidewire tip segmentation in 2D X-ray fluoroscopy using convolution neural networks,” in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2018, pp. 1–7.

J. Y. Lee, M. Islam, J. R. Woh, T. S. M. Washeem, L. Y. C. Ngoh, W. K. Wong, and H. Ren, “Ultrasound needle segmentation and trajectory prediction using excitation network,” Int. J. Comput. Assist. Radiol. Surg., vol. 15, no. 4, pp. 6639–6646, Oct. 2020.

W. Li and D. Que, “Research on master-slave isomorphism design and guide wire segmentation of robot for vascular intervention,” in Proc. Int. Conf. Robot. Control Eng., Apr. 2021, pp. 7–12.

D. Pakhomov, W. Shen, and N. Navab, “Towards unsupervised learning for instrument segmentation in robotic surgery with cycle-consistent adversarial networks,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Oct. 2020, pp. 8499–8504.

J. Zhang, Y. Nie, Y. Lyu, X. Yang, J. Chang, and J. J. Zhang, “SD-Net: Joint surgical gesture recognition and skill assessment,” Int. J. Comput. Assist. Radiol. Surg., vol. 15, no. 4, pp. 6639–6646, Oct. 2020.

J. Zhang, Y. Nie, Y. Lyu, X. Yang, J. Chang, and J. J. Zhang, “SD-Net: Joint surgical gesture recognition and skill assessment,” Int. J. Comput. Assist. Radiol. Surg., vol. 16, no. 10, pp. 1675–1682, Oct. 2021.

I. Funke, S. Bodenstedt, F. Oehme, F. V. Bechtolsheim, J. Weitz, and S. Speidel, “Using 3D convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video," in Proc. Int. Conf. Robot. Autom. (ICRA), May 2020, pp. 5068–5075.

I. Funke, S. Bodenstedt, F. Oehme, F. V. Bechtolsheim, J. Weitz, and S. Speidel, “Using 3D convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video," in Proc. Int. Conf. Robot. Autom. (ICRA), May 2020, pp. 5068–5075.

J. Hsu and S. Payandeh, “Toward tool gesture and motion recognition on a novel minimally invasive surgical robotic system,” in Proc. IEEE Int. Conf. Robot. Autom., Oct. 2006, pp. 631–636.

W. Wang, M. He, X. Wang, J. Ma, and H. Song, “Medical gesture recognition method based on improved lightweight network,” Appl. Sci., vol. 12, no. 13, p. 6414, Jun. 2022.

S. Agarwal, C. S. Pradeep, and N. Sinha, “Temporal surgical gesture segmentation and classification in multi-gesture robotic surgery using fine-tuned features and calibrated MS-TCN,” in Proc. IEEE Int. Conf. Signal Process. Commun. (SPCOM), Jul. 2022, pp. 1–5.

B. van Amstel, M. J. Clarkson, and D. Stoyanov, “Multi-task recurrent neural network for surgical gesture recognition and progress prediction,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2020, pp. 1380–1386.

J. Zhang, Y. Nie, Y. Lyu, H. Li, J. Chang, X. Yang, and A. J. Zhang, “Symmetric dilated convolution for surgical gesture recognition,” in Proc. Int. Conf. Med. Image Comput. Assist. Intervent. Cham, Switzerland: Springer, 2020, pp. 409–418.
D. Itzkovich, Y. Sharon, A. Jarc, Y. Refaely, and J. Nisky, “Generalization of deep learning gesture classification in robotic-assisted surgical data: From dry lab to clinical-like data,” IEEE J. Biomed. Health Inform., vol. 26, no. 3, pp. 1329–1340, Mar. 2022.

B. van Amsterdam, I. Funke, E. Edwards, S. Speidel, J. Collins, A. Sridhar, J. Kelly, M. J. Clarkson, and D. Stoyanov, “Gesture recognition in robotic surgery with multimodal attention,” IEEE Trans. Med. Imag., vol. 41, no. 7, pp. 1677–1687, Jul. 2022.

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.

B. Lu, X. B. Yu, J. W. Lai, K. C. Huang, K. C. C. Chan, and H. K. Chu, “A learning approach for suture thread detection with feature enhancement-modal segmentation for 3-D suture thread reconstruction,” IEEE Trans. Autom. Sci. Eng., vol. 17, no. 2, pp. 858–870, Apr. 2020.

D. Cai, Z. Wang, Y. Liu, Q. Zhang, X. Han, and W. Liu, “Automatic path planning for navigated pedicle screw surgery based on deep neural network,” in Proc. WRC Symp. Adv. Robot. Autom. (WRC SARA), Aug. 2019, pp. 62–67.

Q. Li, Z. Du, and H. Yu, “Grinding trajectory generator in robot-assisted laminectomy surgery,” Int. J. Comput. Assist. Radiol. Surg., vol. 16, no. 3, pp. 485–494, Mar. 2021.

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 Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2017, pp. 2371–2378.

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,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2016, pp. 4150–4157.

Z. Shao, H. Zhao, J. Xie, Y. Qu, Y. Guan, and J. Tan, “Unsupervised trajectory segmentation and promoting of multi-modal surgical demonstrations,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Oct. 2018, pp. 777–782.

J. Liang, J. Mahler, M. Laskey, P. Li, and K. Goldberg, “Using dVRK tele-operation to facilitate deep learning of automation tasks for an industrial robot,” in Proc. 13th IEEE Int. Conf. Robot. Autom. Sci. Eng. (CASE), Aug. 2017, pp. 1–8.

Y. Zhao, Y. Wang, J. Zhang, X. Liu, Y. Li, S. Guo, X. Yang, and S. Hong, “Surgical GAN: Towards real-time path planning for passive flexible tools in endovascular surgeries,” Neurocomputing, vol. 500, pp. 567–580, Aug. 2022.

C. Yao, J. He, H. Che, Y. Huang, and J. Wu, “Feature pyramid self-attention network for respiratory motion prediction in ultrasound image guided surgery,” Int. J. Comput. Assist. Radiol. Surg., vol. 2022, pp. 1–8, Jan. 2022.

D. D. Gregorio, G. Palli, and L. D. Stefano, “Let’s take a walk on superpixels graphs: Deformable linear objects segmentation and model estimation,” in Proc. Asian Conf. Comput. Vis. Cham, Switzerland: Springer, 2018, pp. 662–677.

Q. Li, Z. Du, and H. Yu, “Trajectory planning for robot-assisted laminctomy decompensation based on ct images,” IOP Conf. Ser., Mater. Sci. Eng., vol. 768, no. 4, 2020, Art. no. 042037.

X. Qi, J. Meng, M. Li, Y. Yang, Y. Hu, B. Li, J. Zhang, and W. Tian, “An automatic path planning method of pedicle screw placement based on preoperative CT images,” IEEE Trans. Med. Robotics, Bionics, vol. 4, no. 2, pp. 403–413, May 2022.

A. Attanasi, C. Alberi, B. Scaglioni, N. Marahrens, A. F. Frangi, M. Leonetti, C. S. Biyani, E. De Momi, and P. Valdastri, “A comparative study of spatio-temporal U-Nets for tissue segmentation in surgical robotics,” IEEE Trans. Med. Robot. Bionics, vol. 3, no. 1, pp. 53–63, Feb. 2021.

Y. Jonmohamadi, “Automatic segmentation of multiple structures in knee arthroscopy using deep learning,” IEEE Access, vol. 8, pp. 51853–51861, 2020.

C. Shin, P. W. Ferguson, S. A. Pedram, J. Ma, E. P. Dutson, and J. Rosen, “Autonomous tissue manipulation via surgical robot using learning based model predictive control,” in Proc. Int. Conf. Robot. Autom. (ICRA), May 2019, pp. 3875–3881.

A. Attanasi, 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 Robot. Autom. Lett., vol. 5, no. 4, pp. 6528–6535, Oct. 2020.

D. C. Barratt, D. A. M. Hendry, J. S. M. Morris, A. K. D. S. L. Kam, and J. A. D. Aerts, “Intraoperative margin assessment in oral and oropharyngeal cancer using label-free fluorescence lifetime imaging and machine learning,” IEEE Trans. Med. Imaging, vol. 48, no. 11, pp. 2619–2624, Nov. 2019.

S. M. Hussain et al.: DL Based Image Processing for Robot Assisted Surgery: A Systematic Literature Survey
Z. Li, X. Zhang, L. Ding, K. Du, J. Yan, M. T. V. Chan, W. K. K. Wu, and S. Li, “Deep learning approach for guiding three-dimensional computed tomography reconstruction of lower limbs for robotically-assisted total knee arthroplasty,” Int. J. Med. Robot. Comput. Assist. Surg., vol. 17, no. 5, p. e2390, Oct. 2021.

A. Sharghi, J. Maeder, N. R. Bennett, K. Hörndler, A. S. Wang, and M. Kachelrieß, “Deep learning-based reconstruction of interventional tools and devices from four X-ray projections for tomographic interventional guidance,” Med. Phys., vol. 48, no. 10, pp. 5837–5850, Oct. 2021.

C. Baumgarten, Y. Zhao, P. Sauleau, C. Malrain, P. Jannin, and A. Haegeland, “Image-guided preoperative prediction of pyramidal tract side effect in deep brain stimulation,” Proc. SPIE, vol. 9786, Mar. 2016, Art. no. 978601.

Y. M. Zhao, E. H. Currie, L. Kavoussi, and S. Y. Rabbany, “Laser scanner for 3D reconstruction of a wound’s edge and topology,” Int. J. Comput. Assist. Radiol. Surg., vol. 16, no. 10, pp. 1761–1773, Oct. 2021.

Q. Li, Z. Du, and H. Yu, “Precise lanamization based on neural network for robot-assisted decompressive laminectomy,” Comput. Methods Programs Biomed., vol. 209, Sep. 2021, Art. no. 106333.

J. Torrellas-Barrena, R. López-Velázquez, N. Masoller, B. Valenzuela-Alcaraz, E. Gratacós, E. Eixarch, M. Ceresa, and M. G. Ballester, “Preoperative planning and simulation framework for twin-to-twin transfusion syndrome fetal surgery,” in OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis. Cham, Switzerland: Springer, 2018, pp. 184–193.

X.-Y. Zhou, J. Lin, C. Riga, G.-Z. Yang, and S.-L. Lee, “Real-time 3-D shape instantiation from single fluorescence projection for fenerated stent graft deployment,” IEEE Robot. Autom. Lett., vol. 3, no. 2, pp. 1314–1321, Apr. 2018.

A. Nazir, M. N. Cheema, B. Sheng, P. Li, H. Li, P. Yang, J. 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,” J. Biomed. Inform., vol. 106, Jun. 2020, Art. no. 103430.

G. De Rossi, S. Roin, F. Setti, and R. Muradore, “A multi-modal learning system for on-line surgical action segmentation,” in Proc. Int. Symp. Med. Robot. (ISMRR), Nov. 2020, pp. 132–138.

Y. Zhao, S. Guo, Y. Wang, J. Cui, Y. Ma, Y. Zeng, X. Liu, Y. Jiang, Y. Li, L. Shi, and N. Xiao, “A CNN-based prototype method of unstructured surgical state perception and navigation for an endovascular surgery robot,” Med. Biol. Eng. Comput., vol. 57, no. 9, pp. 1875–1887, 2019.

Y. Qin, M. Allan, J. W. Burdick, and M. Azizian, “Autonomous hierarchial surgical state estimation during robot-assisted surgery through deep neural networks,” IEEE Robot. Autom. Lett., vol. 6, no. 4, pp. 6220–6227, Oct. 2021.

Y. Qin, M. Allan, Y. Yue, J. W. Burdick, and M. Azizian, “Learning invariant representation of tasks for robust surgical state estimation,” IEEE Robot. Autom. Lett., vol. 6, no. 2, pp. 3208–3215, Apr. 2021.

S. Ramesh, D. Dal’Alba, C. Gonzalez, T. Yu, P. Mascagni, D. Mutter, J. Marescaux, P. Fiorinini, and N. Padov, “Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures,” Int. J. Comput. Assist. Radiol. Surg., vol. 2021, pp. 1–9, May 2021.

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, 2020.

M. Takeuchi, H. Kawakubo, K. Saito, Y. Maeda, S. Matsuda, K. Fukuda, R. Nakamura, and Y. Kitagawa, “Automated surgical-phase recognition for robot-assisted minimally invasive esophagectomy using artificial intelligence,” Ann. Surgical Oncol., vol. 29, pp. 1–9, Oct. 2022.

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

Y. Li, Y. Fan, C. Hu, F. Mao, X. Zhang, and H. Liao, “Intelligent optical diagnosis and treatment system for automated image-guided laser ablation of tumors,” Int. J. Comput. Assist. Radiol. Surg., vol. 16, no. 12, pp. 2147–2157, Dec. 2021.

A. Sharghi, H. Haug erud, D. Oh, and O. Mohareri, “Automatic operating room surgical activity recognition for robot-assisted surgery,” in Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. Cham, Switzerland: Springer, 2020, pp. 385–395.

S. Kannan, G. Yengerla, D. Mutter, J. Marescaux, and N. Padov, “Future-state predicting LSTM for early surgery type recognition,” IEEE Trans. Med. Imag., vol. 39, no. 3, pp. 556–566, Mar. 2020.
[233] R. Stauder, D. Ostler, T. Vogel, D. Wilhelm, S. Koller, M. Kranzfelder, and N. Navab, “Surgical data processing for smart intraoperative assistance systems,” Innov. Surg. Sci., vol. 2, no. 3, pp. 145–152, Sep. 2017.
[234] E. Padovan, G. Marullo, L. Tanzi, P. Piazzolla, S. Moos, F. Porriglia, and E. Vezzetti, “A deep learning framework for real-time 3D model registration in robot-assisted laparoscopic surgery,” Int. J. Med. Robot. Comput. Assist. Surg., vol. 18, no. 3, p. e2387, Jun. 2022.
[235] I. Artemchuk, E. Petlenkov, and F. Miyawaki, “Neural network based system for real-time organ recognition during surgical operation,” IFAC Proc. Vol., vol. 44, no. 1, pp. 6478–6483, Jan. 2011.
[236] L. Li, P. Peng, H. Ding, and G. Wang, “A preliminary exploration to make stereotactic surgery robots aware of the semantic 2D/3D working scene,” IEEE Trans. Med. Robot. Biom. Vol., vol. 4, no. 1, pp. 17–27, Feb. 2022.
[237] S. Seiditz, J. Sellner, J. Odenhal, B. Özdemir, A. Studier-Fischer, S. Knödler, L. Ayala, T. J. Adler, H. G. Kengnott, M. Tizabi, M. Wagner, F. Nickel, B. P. Müller-Stich, and L. Maier-Hein, “Robust deep learning-based semantic organ segmentation in hyperspectral images,” Med. Image Anal., vol. 80, Aug. 2022, Art. no. 102488.
[238] P. Li, X.-Y. Zhou, Z.-Y. Wang, and G.-Z. Yang, “Z-Net: An anisotropic 3D DCNN for medical CT volume segmentation,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Oct. 2020, pp. 2906–2913.
[239] P. Li, X. Hou, L. Wei, G. Song, and X. Duan, “Efficient and low-cost deep-learning-based gaze estimator for surgical robot control,” in IEEE Int. Conf. Real-time Comput. Robot. (RCAR), Aug. 2018, pp. 58–63.
[240] L. Chen, W. Tang, N. W. John, T. R. Wan, and J. J. Zhang, “De-smokeGCN: Generative cooperative networks for joint surgical smoke detection and removal,” IEEE Trans. Med. Imag., vol. 39, no. 5, pp. 1615–1625, May 2020.
[241] A. Leibetseder, M. J. Primus, S. Petscharnig, and K. Schoeffmann, “Image-based smoke detection in laparoscopic videos,” in Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures, Cham, Switzerland: Springer, 2017, pp. 70–87.
[242] Y. Pan, S. Bano, F. Vasconcellos, H. Park, T. T. Jeong, and D. Stoyanov, “DeSmoke-LAP: Improved unpaired image-to-image translation for desmoking in laparoscopic surgery,” Int. J. Comput. Assist. Radiol. Surgery, vol. 17, no. 5, pp. 885–893, May 2022.
[243] G. Sánchez-Brizuela, F.-J. Santos-Criado, D. Sanz-Gobernado, E. de la Fuente-López, J.-C. Fraile, J. Pérez-Turiel, and A. Cisneros, “Gaze detection and segmentation in minimally invasive surgery video using convolutional neural networks,” Sensors, vol. 22, no. 14, p. 5180, Jul. 2022.
[244] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescua, M. de Mathelin, and N. Padhy, “EndoNet: A deep architecture for recognition tasks on laparoscopic videos,” IEEE Trans. Med. Imag., vol. 36, no. 1, pp. 86–97, Jan. 2016.
[245] 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 Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2018, pp. 691–699.
[246] R. Stauder, D. Ostler, M. Kranzfelder, S. Koller, H. Feußner, and R. Belk, “Ethical issues in service robotics and artificial intelligence,” Service Industries J., vol. 41, nos. 13–14, pp. 860–876, Oct. 2021.
[247] D. Schönberger, “Artificial intelligence in healthcare: A critical analysis of the legal and ethical implications,” Int. J. Law Inf. Technol., vol. 27, no. 2, pp. 171–203, 2019.
[248] J. W. Collins, H. J. Marcus, A. Ghazi, A. Sridhar, D. Hashimoto, G. Hager, and A. Arezzo, “Ethical implications of AI in robotic surgical training: A Delphi consensus statement,” Eur. Urol. Focus, vol. 2021, pp. 1–22, Apr. 2021.
[249] S. Gerke, T. Minnen, and G. Cohen, “Ethical and legal challenges of artificial intelligence-driven healthcare,” in Artificial Intelligence in Healthcare, Amsterdam, The Netherlands: Elsevier, 2020, pp. 295–336.
[250] T. Bonaci, J. Herron, T. Yusuf, J. Yan, T. Kohn, and H. J. Chizeck, “To make a robot secure: An experimental analysis of cyber security threats against teleoperated surgical robots,” 2015, arXiv:1504.04339.
[251] S. O’Sullivan, N. Nevejans, C. Allen, A. Blyth, S. Leonard, U. Pagallo, K. Holzinger, A. Holzinger, M. I. Sajid, and H. Ashrafian, “Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery,” Int. J. Med. Robot. Comput. Assist. Surg., vol. 15, no. 1, p. e1968, Feb. 2019.
[252] R. Collier, “NHS ransomware attack spreads worldwide,” Can. Med. Assoc. J., 2017.
[253] C. J. Hoofnagle, B. van der Sloot, and F. Z. Borgesius, “The European union general data protection regulation: What it is and what it means,” Inf. Commun. Technol. Law, vol. 28, no. 1, pp. 65–98, Jan. 2019.
GIUSEPPE LUCARELLI is currently an Associate Professor of urology with the Department of Emergency and Organ Transplantation, University of Bari, Italy. He is a Clinician-Scientist, a Urologist, and a Transplant Surgeon. His primary research interests include urologic oncology and kidney transplantation. In particular, his research investigates the biology of prostate and renal cancer using high throughput omics technologies to obtain a detailed molecular characterization of cancer, and also to identify novel biomarkers that allow a more accurate prognostic stratification.

RICCARDO MEMEO is the Chief of the Hepatobiliary and Pancreatic Unit with Mülli Hospital, Bari, Italy. He attended the Medical School at Università degli Studi di Bari and then, during residency, he moved for the last two years of residency improving his knowledge in hepatobiliary, pancreatic surgery, and liver transplant at Paul Brousse Hospital, Paris and Henri Mondor Hospital. He pursued for other five years in France, improving his knowledge in minimally invasive and robotic surgery in Strasbourg, France. His clinical interests include oncologic surgery of liver and pancreas, focusing on minimally invasive access and enhanced recovery after surgery.

VITOANTONIO BEVILACQUA received the Laurea degree in electronic engineering, the Ph.D. degree in electrical engineering, and the postdoctoral degree in industrial informatics from Polytechnic University of Bari. He is a Full Professor of electronic and information bioengineering with the Electrical and Information Engineering Department, Polytechnic University of Bari. In March 2019, he founded Apulian Bioengineering s.r.l.—a start-up and spin-off company of Polytechnic University of Bari, where he is currently the Chief Executive Officer. In September 2019, he joined as an Affiliate Professor the BioRobotics Institute, Scuola Superiore Sant’Anna di Pisa. Currently, he is the Head of Industrial Informatics Laboratory. He has published more than 220 scientific articles. Since 1996, he has been working and investigating in the field of computer vision and image processing, bioengineering, human-machine interaction based on machine learning and soft computing techniques (neural networks, evolutionary algorithms, hybrid expert systems, and deep learning). The main applications of his research are in medicine, in biometry, in bioinformatics in ambient assisted living and industry. In July 2000, he was involved as a Visiting Researcher in an EC funded Trans-Mobility of Researchers (TMR) network (ERB FMRX-CT97-0127) called CAD Modeling Environment from Range Images (CAMERA) and worked at UK Robotics Ltd., Manchester, U.K., in the field of geometric feature extraction and 3D objects reconstruction.

DOMENICO BUONGIORNO received the B.Sc. and M.Sc. (cum laude) degrees in automation and control theory engineering from Polytechnic University of Bari, Bari, Italy, in 2011 and 2014, respectively, and the Ph.D. degree in emerging digital technologies from the PERCeptual RObotics Laboratory (PERCRO), Scuola Superiore Sant’Anna, Pisa, Italy, in 2017. He is an Assistant Professor with the Department of Electrical and Information Engineering, Polytechnic University of Bari. His bachelor’s and master’s theses were supervised by the Prof. Vitoantonio Bevilacqua, concerned machine learning-based optimization techniques for energy consumption optimization and human neuro-musculoskeletal modeling, respectively. His Ph.D. thesis titled “Advanced Control Strategies for Natural Human-Exoskeleton Interaction” concerns the control of robotic interfaces for interaction with virtual environments, robot-aided neurorehabilitation (EMG based control-Myocontrol), and bilateral multi-DoF teleoperation in disaster scenario. Since 2017, he has been a Visiting Ph.D. Student with the Biorobotics Laboratory, UCI Irvine, California, under the supervision of the Prof. David Reinkensmeyer. In 2019, he received the Italian National Bioengineering Group Award for best doctoral thesis. His research interests include robotic-assisted surgery, medical data processing, and biomedical engineering. He teaches biomedical instrumentation at the Polytechnic University of Bari and electronic and information bioengineering at the Medical School of the University of Bari. In March 2019, he founded Apulian Bioengineering s.r.l.—a spin-off company of the Polytechnic University of Bari.

* * *