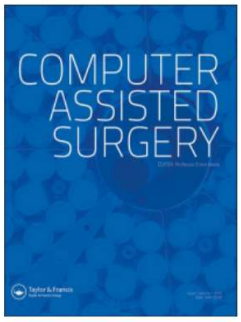

Computer Assisted Surgery

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/icsu21

Image-based laparoscopic tool detection and tracking using convolutional neural networks: a review of the literature

Congmin Yang, Zijian Zhao & Sanyuan Hu

To cite this article: Congmin Yang, Zijian Zhao & Sanyuan Hu (2020) Image-based laparoscopic tool detection and tracking using convolutional neural networks: a review of the literature, Computer Assisted Surgery, 25:1, 15-28, DOI: 10.1080/24699322.2020.1801842

To link to this article: https://doi.org/10.1080/24699322.2020.1801842

Published online: 04 Sep 2020.

COMPUTER ASSISTED SURGERY
2020, VOL. 25, NO. 1, 15–28
https://doi.org/10.1080/24699322.2020.1801842

Taylor & Francis
Taylor & Francis Group

REVIEW

Image-based laparoscopic tool detection and tracking using convolutional neural networks: a review of the literature

Congmin Yang[a], Zijian Zhao[a] and Sanyuan Hu[b]

[a]School of Control Science and Engineering, Shandong University, Jinan, China; [b]Department of General surgery, First Affiliated Hospital of Shandong First Medical University, Jinan, China

ABSTRACT

Intraoperative detection and tracking of minimally invasive instruments is a prerequisite for computer- and robotic-assisted surgery. Since additional hardware, such as tracking systems or the robot encoders, are cumbersome and lack accuracy, surgical vision is evolving as a promising technique to detect and track the instruments using only endoscopic images. The present paper presents a review of the literature regarding image-based laparoscopic tool detection and tracking using convolutional neural networks (CNNs) and consists of four primary parts: (1) fundamentals of CNN; (2) public datasets; (3) CNN-based methods for the detection and tracking of laparoscopic instruments; and (4) discussion and conclusion. To help researchers quickly understand the various existing CNN-based algorithms, some basic information and a quantitative estimation of several performances are analyzed and compared from the perspective of 'partial CNN approaches' and 'full CNN approaches'. Moreover, we highlight the challenges related to research of CNN-based detection algorithms and provide possible future developmental directions.

KEYWORDS

Tool detection; tool tracking; convolutional neural network; laparoscopic surgery

1. Introduction

Robot-assisted minimally invasive surgery (RMIS) systems have gained increasing attention in recent years. Instead of surgeons operating directly on the patient, RMIS allows operation by telemanipulation of dexterous robotic tools through small incisions. With RMIS systems, surgeons sit at a console near the operating table and utilize joysticks to perform complex procedures. Moreover, minimally invasive surgery (MIS) using cameras to observe the internal anatomy is the preferred approach for many surgical procedures [1], which can reduce operative trauma, speed up recovery, and shorten hospitalization. However, indirect operations are highly complex, and surgeons must deal with difficult hand-eye coordination, restricted mobility, and a narrow field of view, which necessitates acquisition of additional information to monitor the operational instruments moving within the body. Thus, it is necessary to detect and track surgical instruments with a view to providing information for operational navigation during MIS.

Nowadays, detection and tracking data of surgical tools are furnished by electromagnetic, optical, and image-based (also called vision-based) techniques in MIS operational navigation. The first two methods require expensive devices and hardware. Meanwhile, image-based methods directly estimate the tool position in the video frames of the observing camera (endoscope) using flexible software-based implementation with no need for modification of tools or surgical workflow [2], which have become state-of-the-art methods for tool detection tasks in MIS.

However, developing image-based methods for the detection and tracking of instruments in endoscopic videos is not an easy task since endoscopic data present several challenges. Firstly, there is a large variety of cases with high deformation or artifacts. Secondly, complicated surgical scenes lead to limited inter-phase and high intra-phase variance. Lastly, observed surgical scenes are often blurred due to the camera motion and the gas generated by tools, and are even sometimes completely occluded when blood stains the camera lens. Extra noise and artifacts introduced by the consequent lens cleaning process make the recognition tasks even harder [3]. Therefore, it has become an urgent problem to find an effective method robust

CONTACT Zijian Zhao zhaozijian@sdu.edu.cn School of Control Science and Engineering, Shandong University, Jinan, China

enough to deal with all the possible complicated situations described above.

With the continuous development of CNN, an increasing number of studies have begun to apply CNN to the detection and tracking of surgical instruments in endoscopic videos due to its high capability of feature extraction and expression. The present paper analyzes the CNN-based detection algorithms mainly from the perspective of 'partial CNN approaches' and 'full CNN approaches' to facilitate the systematic transfer of knowledge to researchers in related fields regarding the current research progress. Furthermore, we also provide a general direction for further research pertaining to the detection and tracking of laparoscopic tools based on CNN.

The remainder of the present paper is organized as follows: Section 2 introduces the fundamentals of CNN; several public datasets are described in Section 3; Section 4 briefly reviews CNN-based detection and tracking methods from the perspective of 'partial CNN approaches' and 'full CNN approaches'; and a discussion and possible further developmental directions are provided in Section 5.

2. Fundamentals of CNN

Detection of any object can be described quite generally as a parameter estimation problem over a set of image features [4]. The strategies used to address this problem are mainly separated into generative methods and discriminative methods. The generative methods (including particle filtering, mean shift, and optical flow) use the feature model to describe the appearance of the target, subsequently locating the target by minimizing the reconstruction error between the tracking target and the candidate target. The discriminative methods (including support vector machine, decision forests, and CNN) train the classifier about the target and the background with a view to identifying the target from candidate targets. Since the discrimination methods can effectively distinguish between background and target information and have a strong performance robustness, they have gradually become state-of-the-art methods to solve the problem of target detection and tracking.

Traditional feature extraction methods (generative methods and parts of discriminative methods) usually have many limitations. For example, the features need to be designed manually, which is inaccurate and time-consuming; it is difficult to build high-level semantic information; and it cannot be applied to complex scenes. However, deep learning [5] (especially CNN) allows deep neural networks to discover the representations from raw data for specific tasks such as classification [6] and detection [7]. It has outperformed other methods in many areas, especially when the dataset is sufficiently large. The model based on CNN is characterized by a strong learning ability, an efficient feature expression ability, and more advanced semantic features, which makes it much easier to detect and track laparoscopic instruments in MIS. Therefore, researchers have paid great attention to CNN in recent years.

Regarding the structure, classical CNN consist of an input layer, a convolution layer, a pooling layer, a fully connected layer, and an output layer. The input layer, convolution layer, and pooling layer constitute the feature extraction layer; and the fully connected layer and output layer constitute the classification layer. The architecture of the CNN is shown in Figure 1.

CNN have the characteristics of both receptive field and weight sharing, thus multi-dimensional images can be used as an input to the network, which can reduce the complexity of data reconstruction during

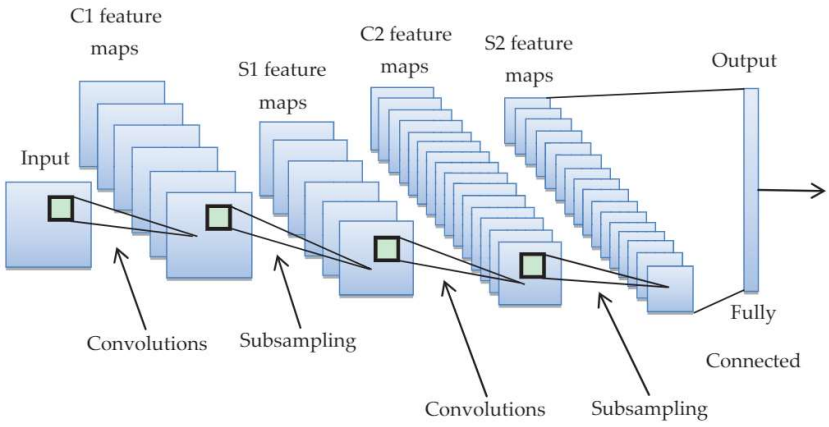

Figure 1. The architecture of the convolutional neural network.

the process of feature extraction, and restrain the influence of translation and scaling. Therefore, in the field of computer vision, CNN-based methods for feature extraction have exceeded the traditional manual feature extraction methods and become start-of-the-art methods.

3. Public datasets

Data play an important role in the detection and tracking of endoscopic instruments using CNN. The following subsections provide a description of the public datasets that are widely used by researchers in related fields.

3.1. Cholec 80

This dataset consists of 80 videos of cholecystectomy surgeries performed by 13 surgeons at the University Hospital of Strasbourg. The videos are recorded to 1 fps by taking the first frame of every 25 frames. The entire dataset is labeled with the phase and tool presence annotations. While most of the videos are recorded at a resolution of 854×480 pixels, a few are 1920×1080 pixels with the same aspect ratio [8]. There are seven types of surgical tools in total, as shown in Figure 2: grasper, hook, clipper, bipolar, irrigator, scissors, and specimen bag.

Some of the videos in Cholec 80 are included in the M2CAI16-tool dataset, which consists of 15 cholecystectomy videos with ground truth binary annotations of the present tools. The dataset is split into two parts: a training subset (10 videos) and a testing subset (5 videos) (http://camma.u-strasbg.fr/M2CAI2016/). Furthermore, Jin et al. [9] introduced a new dataset, M2CAI16-tool-locations, which extends the M2CAI16-tool dataset with spatial bounds of tools.

3.2. Endovis challenge dataset

This multi-instrument dataset is separated into training and test data: the training data includes four 45-s *ex vivo* video sequences of interventions; and the test set is composed of additional 15-s video sequences for each of the training sequences and two additional 1-min recorded interventions [10]. All sequences have a resolution of 720×576 pixels and include one or two surgical instruments. (https://endovissub-instrument.grand-challenge.org/)

3.3. Atlas dione

ATLAS Dione was firstly introduced in [11] and consists of 99 action video clips of 10 surgeons from the Roswell Park Cancer Institute (RPCI) (Buffalo, NY) performing six surgical tasks (subject study) on the da Vinci Surgical System (dVSS). The resolution of each frame is 854×480 with the surgical tool annotations.

4. The CNN-based methods for detection and tracking laparoscopic instruments

In recent years, scholars have proposed various CNN-based methods to detect and track laparoscopic instruments in MIS. This section briefly reviews the basic working principles of the different methods from the perspective of 'partial CNN approaches' and 'full CNN approaches.'

4.1. Partial CNN approaches

4.1.1. Methods based on line detection

In general, the first step to tracking surgical instruments is finding their position. In recent work [12], Chen et al. proposed to obtain the positions and widths of lines using a linear-time line segment detector (LSD) that yields subpixel accurate results. Moreover, based on the positions of the selected lines,

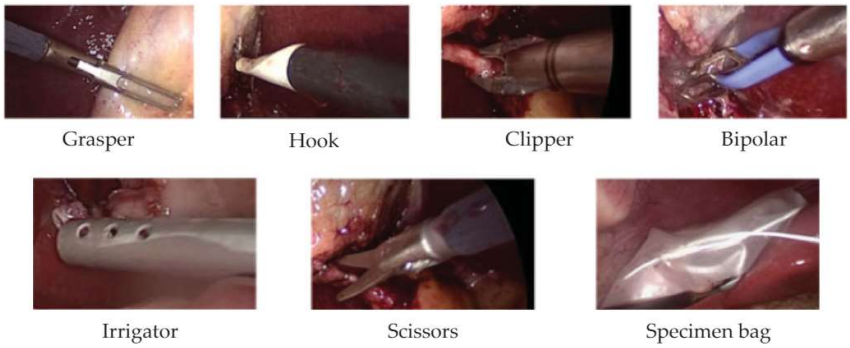

Figure 2. The seven surgical tools used in the Cholec80 dataset.

the tip of a tool is quickly found by the CNN (consisting of 5 convolutional layers, 2 maximum pooling layers, 1 average pooling layer, and a softmax-loss layer). Furthermore, as described by [13], a spatiotemporal context learning algorithm is employed to track the tip using Fast Fourier Transform (FFT) and Inverse Fast Fourier Transform (IFFT) to compute the spatial and temporal contextual information of the current frame and the next frame to quickly locate the tip. The full pipeline of this method is shown in Figure 3. The experimental results of the Endovis 2015 challenge showed that the detection accuracy (the percentage of successful tool identifications) reached 100% but the speed was only 29.6333 fps.

4.1.2. Methods based on the hierarchical hidden markov model

Padoy et al. [14] proposed the Hidden Markov Model (HMM) to detect surgical tool phases online by combining the tool usage signals and two visual cues from the laparoscopic images. Inspired by this, Twinanda et al. [15] presented a surgical instrument detection method based on the Hierarchical Hidden Markov Model (HHMM), an extension of HMM. In the proposed approach, the EndoNet network is trained *via* a fine-tuning process, which consists of an input layer, five convolutional layers, and two fully connected layers. The confidence given by the network is directly used to detect tool presence. Meanwhile, the visual features extracted by EndoNet are passed to the Support Vector Machine (SVM) and HHMM to detect the phase. The full pipeline of this method is shown in Figure 4. The experimental results of the tool presence detection for the large dataset, Cholec80, showed that the detection accuracy was 81% and the method performed very well in detecting all the phases.

4.1.3. Methods based on random forests

Surgical tool detection in a surgical workflow is usually modeled as a classification problem. Meanwhile, random forests are inherently suited for multi-class detection [16]; therefore, Sahu et al. [17] applied a modified AlexNet and Random Forest (RF) to surgical tool and phase detection. For tool detection, the CNN features are directly fed in a random forest classifier for prediction; and for tool phase detection, the random forest classifier is combined with a time series, and hard negative mining is applied for final prediction of the surgical phase. The full pipeline of this method is shown in Figure 5. The experimental results of the tool presence detection for the M2CAI16 challenge training dataset demonstrated that the detection accuracy reached 61.5% and the mean F1-score was 53.13. To improve the performance of this method, it is important to explore further feature fusion techniques to enhance the contextual features.

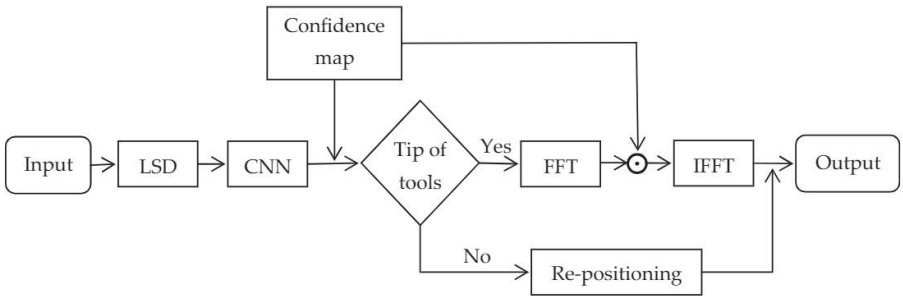

Figure 3. Full pipeline of the method based on LSD.

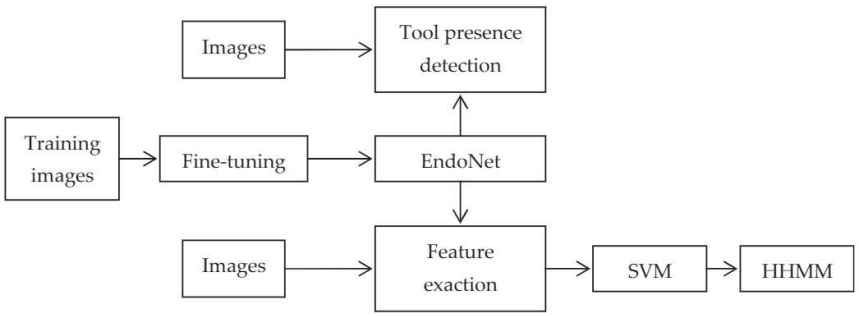

Figure 4. Full pipeline of the method based on the Hierarchical Hidden Markov Model.

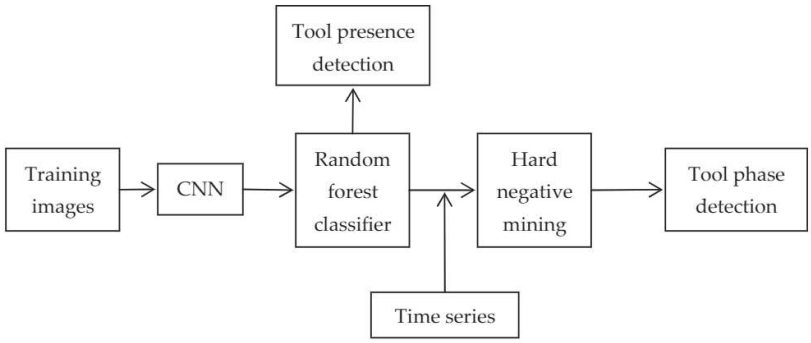

Figure 5. Full pipeline of the method based on Random Forests.

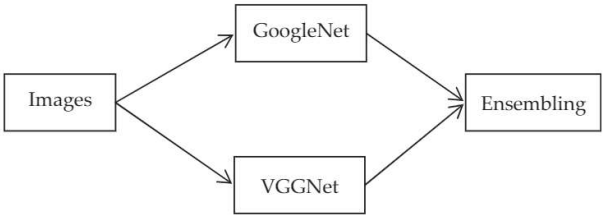

Figure 6. Full pipeline of the method based on Model Ensembling.

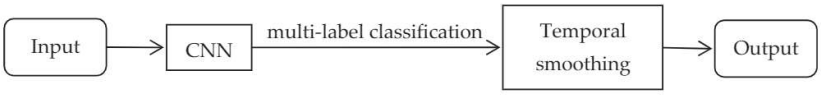

Figure 7. Full pipeline of the method based on Temporal Smoothing.

4.1.4. Methods based on model ensembling

Due to the co-occurrence of multiple surgical tools in different endoscopic video frames, rather than treating surgical tool detection as a supervised multi-class classification task, it should be considered a multi-label classification problem. Wang et al. [18] combined two deep neural networks, VGGNet [19] and GoogleNet [20], and used ensemble learning to detect tool presence in laparoscopic videos. After training each of the two networks, an average ensembling technique is applied to avoid overfitting and obtain better performance. The full pipeline of this method is shown in Figure 6. This algorithm was evaluated in the dataset from M2CAI surgical tool presence detection, and the results showed that the detection accuracy was 63.8%, yielding a better tool presence detection performance than the other methods in the M2CAI challenge.

4.1.5. Methods based on temporal smoothing

As mentioned in [18], it could be regarded as a multi-label classification problem; however, false detections occur during the testing step due to the stochastic nature of the classification process. Aiming to reduce such false detections, Sahu et al. [21] adopted a temporal smoothing (TS) approach as an online post-processing step. Moreover, a ZIBNet architecture (similar to AlexNet) is introduced with novel design choices that are incorporated during learning. The full pipeline of this method is shown in Figure 7. The experimental results of the tool presence detection in the M2CAI tool detection challenge demonstrated that the detection accuracy reached 65%. In particular, the simple stratification benefited more from TS, with an overall precision increase of 8%.

4.1.6. Methods based on optical flow

Fully Convolutional Networks (FCNs) are a particular type of CNN proposed by Long et al. [22], which play an important role in detecting and tracking surgical instruments; however, it is difficult to run in real time using fine-tuned FCN. To address this problem, García-Peraza-Herrera et al. [23] adapted the fined-tuned FCN-8S [22] in combination with optical flow, forming a widely successful tracking framework used for temporal constraints to satisfy the real-time requirement. The full pipeline of this method is shown in Figure 8.

The experimental results showed that this method yielded a detection accuracy of 78.2% and reached a real-time speed of 30 fps in the EndoVis dataset.

4.1.7. Methods based on 'coarse to fine'

Recently, with the aim of achieving good performance in detecting and tracking surgical instruments in laparoscopic videos, the 'coarse to fine' method was proposed, which uses a coarse CNN to provide the rough location for the fine CNN.

In 2018, Du et al. [10] proposed a fully convolutional detection-regression network for articulated multi-instrument 2-D pose detection. Surgical instruments are located by a detection subnetwork and subsequently refined through a regression subnetwork. To obtain the final poses of all the instruments in an image, association probabilities are used as a measurement to connect joint pairs for each instrument by maximum bipartite matching. The full pipeline of the proposed framework is shown in Figure 9. The experimental results in the multi-instrument EndoVis dataset showed that the detection accuracy reached 90.68%. Furthermore, the model exhibits some generalizability to new unseen instruments, and has a good robustness under smoke simulation.

More recently, Zhao et al. [2] presented a 'coarse to fine' framework for surgical tool tracking, where the coarse CNN is a classification network of 10 classes, and the fine CNN is a regression network for the tool tip area. In the entire tracking process of surgical tools, the two CNNs cooperate together by updating the spatial and temporal context. The full pipeline of the proposed framework is shown in Figure 10. The experimental results in the Endo-Vis 2015 challenge demonstrated a good performance in surgical instrument tracking, with a high detection accuracy of 95% at 20 pixels and a speed of 23.69 fps. However, the method currently focuses on single tool tracking.

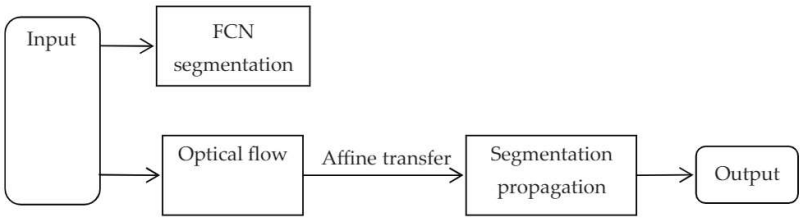

Figure 8. Full pipeline of the method based on Optical Flow.

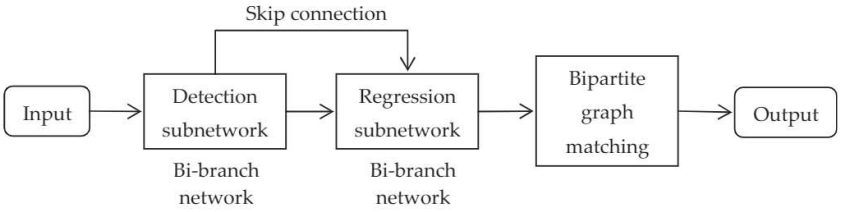

Figure 9. Full pipeline of the method based on a detection-regression network.

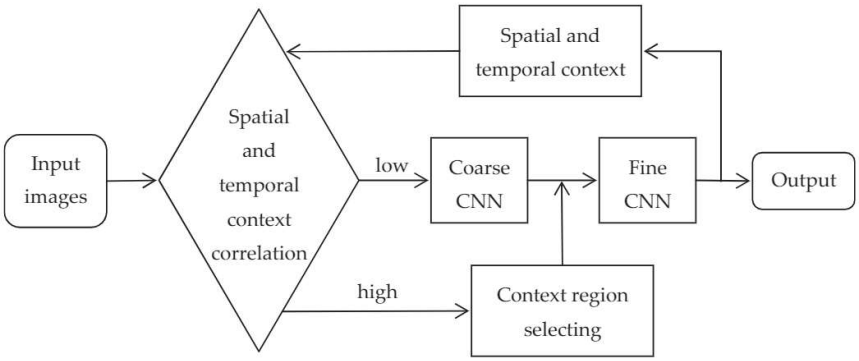

Figure 10. Full pipeline of the method based on 'coarse to fine.'

4.2. Full CNN approaches

4.2.1. Methods based on region proposal

The Region Proposal Network (RPN) was firstly shown in public as part of a Faster Region-based CNN (Faster R-CNN) in [24], with the aim of extracting candidate boxes. To generate region proposals, a small network is slid over the conv feature map output by the last shared conv layer. This network is fully connected to an $n \times n$ spatial window of the input conv feature map. Each sliding window is mapped to a lower-dimensional vector, which is fed into two fully connected sibling layers: a box-regression layer (reg) and a box-classification layer (cls). This mini network is shown in Figure 11. The architecture is naturally implemented with an $n \times n$ conv layer followed by two sibling 1×1 conv layers (for reg and cls, respectively). A Region Proposal Network (RPN) takes an image (of any size) as input and outputs a set of rectangular object proposals, each with an objectness score.

In a recent study [11], Sarikaya et al. proposed an end-to-end deep learning approach for fast tool detection and localization in endoscopic videos. The architecture applied an RPN and a multi-modal convolutional network (inspired by [25]) to jointly predict objectness and localization on a fusion of images and temporal motion cues. The full pipeline of this method is shown in Figure 12. Moreover, this method was evaluated in the ATLAS Dione dataset, which provides video data of 10 subjects performing six different surgical tasks on the dVSS with proper tool annotations. The experimental results reached a precision of 91% and a mean computation time of 0.1 s per test frame detection, emphasizing the benefits of using RPN for precision.

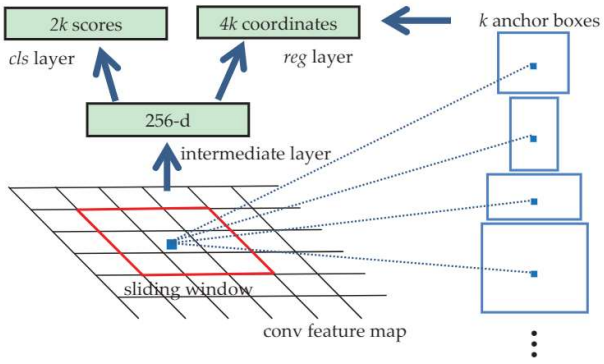

Figure 11. Region Proposal Network (RPN).

It is worth mentioning that RPN was also used by Jin et al. in [9], which was the first time that surgical tool localization was performed in real-world laparoscopic surgical videos, setting the stage for a richer analysis of surgical performance, such as tool usage patterns, movement range, and economy of motion. Moreover, in [26], Nakazawa et al. applied region-based CNNs to real-time surgical needle detection for the first time.

4.2.2. Methods based on 'You Only Look Once'

Different from the series of methods for R-CNN, 'You Only Look Once' (YOLO) is a special algorithm that frames object detection as a regression problem to spatially separated bounding boxes and associated class probabilities, which was first introduced by Redmon et al. in [27]. It is famous for being a simple model with a fast detection speed.

To detect and track surgical instruments in real-time, Choi et al. [25] proposed a CNN model with a structure based on YOLO. The network architecture consists of 24 convolutional layers and 3 fully connected layers by adding a fully connected layer based on the structure of YOLO. Dropout is used in a completely connected layer to prevent overfitting of the training data. Furthermore, batch normalization is also used to improve the performance of the learning. The full pipeline of this method is shown in Figure 13. The experimental results in the M2CAI16-tool dataset showed that the detection accuracy reached 72.6% and the frame rate was measured as 48.9 fps.

The method based on YOLO introduced in [28] performed well with respect to detection speed, but the

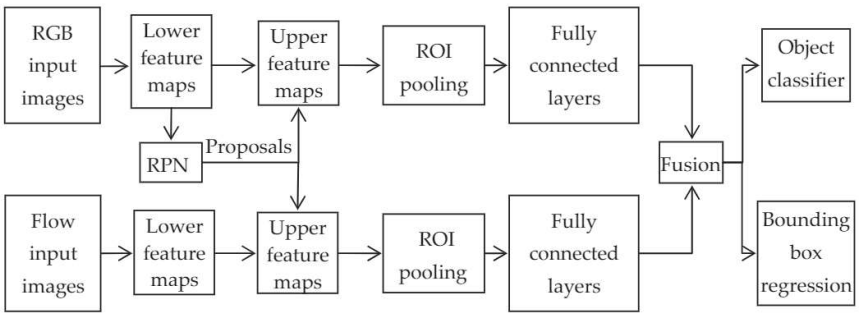

Figure 12. Full pipeline of the method based on a Region Proposal Network.

mean average precision was not high. To improve detection precision, Jo et al. [29] proposed a new real-time detection algorithm for the detection of surgical instruments in laparoscopic images based on YOLO9000 [30] (the second of the three versions of YOLO) and motion vector prediction detecting missing surgical tools. The experimental results in the M2CAI16-tool-locations and M2CAI16-tool datasets showed that the detection accuracy reached 84.7% with a real-time speed of 38 fps. Furthermore, this method exhibits a constant performance, irrespective of surgical instrument class.

4.2.3. Methods based on a recurrent neural network

Temporal information plays a vital role in the process of detecting and tracking surgical instruments, which can be used to learn connectionism across neighboring frames. Therefore, some researchers have recently applied a Recurrent Neural Network (RNN) to the detection and tracking of surgical instruments in laparoscopy videos.

In [31], Mishra et al. proposed a framework based on a long short-term memory (LSTM) to incorporate temporal information. The algorithm uses a residual network (ResNet) to extract high-level visual features from the frames of laparoscopy videos. Based on the feature extracted by the CNN, an LSTM (including three LSTM blocks) is trained to accurately detect tool presence by incorporating the temporal dimension of the information. The full pipeline of the proposed framework is shown in Figure 14. The experimental results in the M2CAI16-tool dataset showed a detection accuracy of 88.75% in detecting tool presence.

The RNN performed well in the process of surgical tool presence detection as shown by the results in [28], but it is difficult to train 'CNN + RNN' in an end-to-end manner for the computational complexity. To address this problem, a novel boosting strategy was proposed by Hajj et al. [32] to achieve the goal that the CNN and RNN are simultaneously enriched by progressively adding weak classifiers (either CNNs or RNNs) trained to improve the overall classification accuracy. Moreover, another novelty lies in the proposed temporal sequence augmentation strategy, which has proven quite effective despite being simple. The experimental results in Cholec80 showed that this method obtained a high detection accuracy of 97.89%.

4.2.4. Methods based on U-Net

In 2015, a new network architecture called U-Net was proposed by Ronnebeger et al. [33], which consists of a contracting path to capture context and a symmetrical expanding path that enables precise localization. Such a network was shown to be trained end-to-end from very few images and performed well in many visual tasks.

In recent work, Kurmann et al. [34] applied U-Net in the process of detecting and tracking surgical instruments in laparoscopic videos, avoiding the need for scale-dependent window sliding evaluation, which allowed the approach to be relatively parameter-free at test time. The experimental results showed an

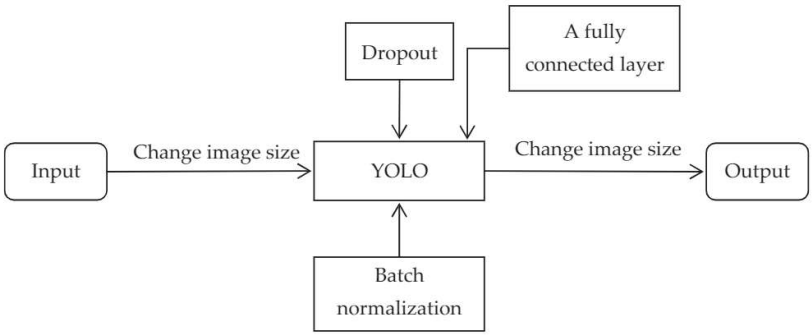

Figure 13. Full pipeline of the method based on YOLO.

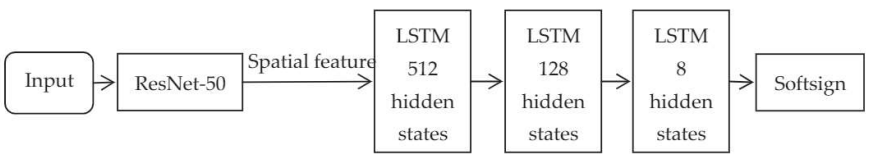

Figure 14. Full pipeline of the method based on LSTM.

extremely high detection accuracy in simultaneously detecting multiple instruments and their poses.

Different from [30] and [31], 3D convolutional neural networks (3D FCNN) could be employed to directly extract spatiotemporal information, which has been shown to be effective for action [35] and object [36] recognition, as well as for surgical skill assessment [37]. Following this paradigm, Colleoni et al. [38] proposed a modular encoder-decoder structure (similar to U-Net) together with a 3D FCNN for surgical instrument joint and location detection. The experimental results in the existing and new contribution datasets showed that this method performs better than the methods based on single-frame processing. The full pipeline of the proposed framework is shown in Figure 15. Furthermore, the good performance of 3D FCNN directs researchers toward a better framework for understanding surgical scenes and can lead to applications of CAI in both robotic systems and surgical data science.

4.2.5. Methods based on a multi-task learning network

Currently, most CNN-based detection tasks are single-task learning, which means that the learning of each task is independent of that of the others and does not make full use of their relatedness. With the aim of better detection and tracking of surgical instruments in complex laparoscopic videos, some researchers have regarded it as a multi-task problem to combine related sub-tasks and solve them simultaneously. In [15], Twinanda et al. implemented a multi-task framework with shared early layers and incorporated tool information in the feature learning process, which firstly achieves joint tool and phase detection; and secondly, the promising performance demonstrates effective leveraging such as relatedness, playing an essential role in improving both tasks.

In 2017, Laina at el. [39] proposed a novel method that takes advantage of the interdependency between localization and segmentation of the surgical tool. Moreover, to obtain a concurrent, robust, and near real-time regression of both tasks, the 2D instrument pose estimation is reformulated as a heatmap regression to ensure good performance in detecting and tracking surgical instruments. It is noteworthy that ResNet-50 [40] is employed in this architecture, which runs in an end-to-end manner to predict the instrument segmentation and its articulated 2D pose by modeling the localization of surgical instrument landmarks as a heatmap regression. The full pipeline of this method is shown in Figure 16. The experimental results in the MICCAI EndoVis Challenge 2015 showed that the detection accuracy reached 92.6% and the processing speed was 52 ms per frame.

More recently, with the aim of exploiting the natural relatedness of tool presence and surgical phase detection to simultaneously boost the performance of both tasks, Jin et al. [3] presented a novel method by developing a multi-task recurrent convolutional network with correlation loss (MTRCNet-CL), which has an end-to-end architecture with two branches, sharing the features in the early layers and holding the respective higher layers for corresponding tasks. In particular, the effective correlation loss is proposed to model the relatedness between tool presence and phase presence by minimizing the divergence in predictions from the two branches. The full pipeline of the proposed framework is shown in Figure 17. Extensive experiments in a large surgical video dataset (Cholec80) demonstrated outstanding performance of the method, with a precision of 89.1% for tool presence detection and an F1 score of 87.4% for phase detection. Furthermore, the speed reached 0.3 s per frame with one GPU, which could be applied in the

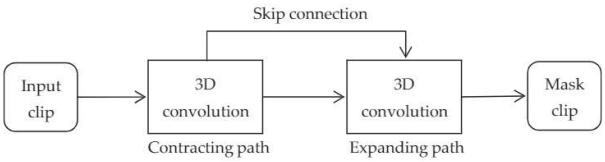

Figure 15. Full pipeline of the method based on U-Net.

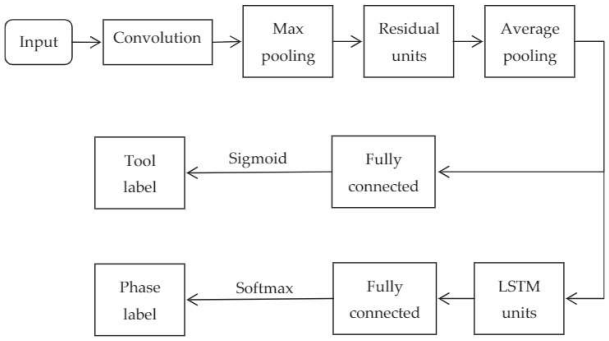

Figure 17. Full pipeline of the Multi-task Learning Network based on LSTM Units.

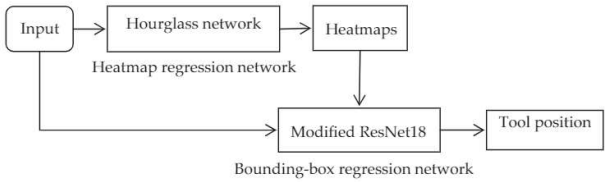

Figure 18. Full pipeline of the method based on an Hourglass Network.

real-time context-aware system and real-world surgical operation.

It is worth mentioning that Modal et al. [41] proposed a multi-task deep learning framework comprised of ResNet-50 and the bi-directional long short-term memory network (Bi-LSTM) [42] with a weighted joint distribution loss function to capture long-term dependencies, both in the past and the future. This method detects tool presence and phase simultaneously and reached an extremely high detection accuracy score of 99% and 86% for tool and phase detection in the Cholec80 dataset.

4.2.6. Methods based on an hourglass network

To capture information at every scale, the hourglass network was proposed by Newell et al. [43] for human pose estimation, which has the capacity to capture all of these features and bring them together to output pixel-wise predictions based on the successive steps of pooling and upsampling.

Inspired by [43], Zhao et al. [44] proposed a frame-by-frame detection method for real-time multi-instrument detection and location using a cascading CNN which consists of an hourglass network and a modified VGG-16 network. The hourglass network is applied to detect a heatmap of each instrument, which is composed of five maximum pooling layers, four upsample layers, and thirteen convolutional blocks of which each consists of several residual modules [43]; and the modified VGG contains six convolutional blocks, six pooling layers and three fully connected layers, which is responsible for bounding-box regression. The full pipeline of this method is shown in Figure 18. Experimental results showed the method achieved a detection accuracy of 91.6% and 100% at 43.5 fps on the ATLAS Dione and Endovis Challenge datasets.

More recently, Liu et al. [45] proposed an anchor-free CNN architecture using a compact stacked hourglass network, which models the surgical tool as a single point: the center point of its bounding box. The lightweight hourglass backbone consisted of two hourglass modules, and the residual modules were replaced with the more effective fire modules [46] to predict the heatmap at the center point of all instances of the surgical tools. This frame-by-frame method eliminated the need to design a set of anchor boxes, and was end-to-end differentiable, simpler, more accurate, and more efficient than anchor-box-based methods. Experimental results showed the method achieved a detection accuracy of 98.5% and 100% at 37.0 fps on the ATLAS Dione and Endovis Challenge datasets, respectively, and truly realizes real-time surgical tool detection in RAS videos.

4.2.7. Methods based on weakly supervised learning

Supervised learning technology builds a prediction model by learning a large number of training data, in which each training sample has a corresponding truth value output. Despite the great success of existing technologies, the lack of spatially annotated surgical data and the high cost of the data annotation process are noteworthy. To circumvent these problems, a new deep learning object tracking method with weak

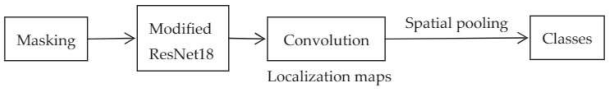

Figure 19. Full pipeline of the method based on Modified ResNet18.

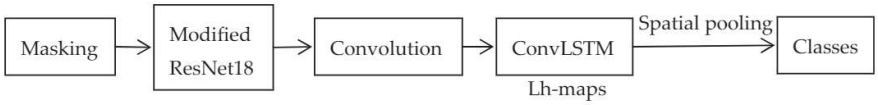

Figure 20. Full pipeline of the method based on ConvLSTM.

supervision on binary presence labels has been recently proposed and employed in the medical computer vision community; for instance, in cancerous region detection [47] and object instance segmentation [48].

Following the same trend, Vardazaryan et al. [49] proposed a deep architecture that relies on ResNet18 [40], trained solely on image level annotations in an end-to-end manner, which was used for both tool presence detection and localization in surgical videos. For the network architecture, the fully connected layer and average pooling are removed from the end of the network to preserve relative spatial information throughout the network, and the stride in the last two banks of ResNet is changed from 2 to 1 pixel to obtain localization maps with a higher resolution. The full pipeline of the proposed framework is shown in Figure 19. Several variants of the network were evaluated in a large public dataset, Cholec80, obtaining a very promising detection accuracy of 88%.

Building on [49], Nwoye et al. [8] presented a weakly supervised approach to tool tracking in laparoscopic videos, where a convolutional LSTM (ConvLSTM) is integrated to model the temporal dependencies in the motion of the surgical tools and to leverage its spatiotemporal ability to smooth the class peak activations in the localization heatmaps (Lh-maps). The full pipeline of the proposed framework is shown in Figure 20. This approach was evaluated in the Cholec80 dataset and yielded a detection accuracy of 92.9%, which showed that the proposed approach could be integrated into a surgical video-labeling software to initialize the tool annotations, such as bounding boxes and segmentation masks.

5. Discussion and conclusion

5.1. Discussion

In recent years, various CNN-based methods of detecting and tracking laparoscopic instruments have been proposed. In the present paper, we summarized these methods from the perspective of 'partial CNN approaches' and 'full CNN approaches.' The summary of the comparison of several CNN-based algorithms is shown in Table 1.

Table 1. Summary of the comparison of several CNN-based algorithms.

Dataset	Reference	Type	Detection accuracy (%)	Speed
Cholec80	Twinanda [15]	Partial CNN	81	–
	Vardazaryan [49]	Full CNN	88	–
	Jin [3]	Full CNN	89.1	3.33 fps
	Du [10]	Full CNN	90.68	–
	Nwoye [8]	Full CNN	92.9	–
M2CAI16	Sahu [21]	Partial CNN	61.5	–
	Wang [18]	Partial CNN	63.8	–
	Sahu [17]	Partial CNN	65	48.9 fps
	Choi [28]	Full CNN	72.6	38 fps
	Simonyan [19]	Full CNN	84.7	–
	Al Hajj [32]	Full CNN	97.89	–
Endovis	Laina [39]	Full CNN	92.6	19.23 fps
	Zhao [2]	Full CNN	95	–
	Zhao [44]	Full CNN	100	43.5 fps
	Liu [45]	Full CNN	100	37.0 fps
ATLAS Dione	Sarikaya [11]	Full CNN	91	10 fps
	Zhao [44]	Full CNN	91.6	43.5 fps
	Liu [45]	Full CNN	98.5	37.0 fps

Based on the results of the present study, we arrived at the conclusion that the 'full CNN approaches' learning methods based on CNN outperform other methods in instrument detection and tracking tasks, which has made them become state-of-the-art methods in this field.

Real-time tool segmentation is also an essential component of computer-assisted surgical systems. Surgical tool segmentation is used for detection, tracking, and pose estimation of the tools in the vicinity of surgical scenes [50]. It is considered an essential task in surgical phase recognition and flow identification. Semantic segmentation is used for the accurate delineation of surgical tools from the background; each label is assigned to a class as a tool or a background.

With the growing need for MIS techniques in surgery, surgical tool detection, segmentation, and tracking tasks have become essential components of various applications in modern operating rooms,

which are highly instrumental in analyzing and understanding surgical activities. Firstly, real-time automated surgical video analysis could facilitate objective and efficient assessment of surgical skills and provide feedback on surgical performance. Secondly, if we generate timelines displaying tool usage during a surgery, we could retrieve data regarding which surgical tools are being used at any given moment. Finally, automatic analysis of surgery monitoring can help to optimize the surgical workflow and plan surgical procedures automatically.

5.2. Conclusion

There is a growing need for CAI systems in surgery, with the ever increasing use of MIS techniques. In the present paper, we reviewed the state-of-the-art methods in this field. We discussed the CNN-based approach of detecting, localizing, and tracking instruments without the requirement for modifying instrument design or interfering with the surgical work-flow.

Although many CNN-based algorithms have been proposed, there are numerous outstanding technical difficulties that must be addressed in the presence of challenging conditions such as occlusion, blood, smoke, other instruments, and the many other hazards that routinely occur within the field of view. Therefore, elucidating the manner by which robustness can be enhanced in order to improve the performance of detecting and tracking laparoscopic instruments remains the ongoing focus of research. Our results indicate that while 'full CNN approaches' outperform other methods in instrument detection and tracking tasks, the results are still not optimal. There remain many areas to improve. Firstly, data play an important role in training neural networks; therefore, the existence of well-established and publicly available datasets and methodologies can allow proper measurement of the progress made in detecting and tracking surgical instruments. In addition, temporal information plays a vital role in the process of detecting and tracking surgical instruments, which can be used to learn connectionism across neighboring frames. From [31, 32], we can see RNN and HMM play an important role in modeling temporal dynamics. However, RNN-based methods are not mature enough and need further improvement. Furthermore, weakly supervised learning circumventing the lack of annotated data is an emerging research method and will become a trend for future research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Key Research and Development Program of China under Grant 2019YFB1311300.

References

[1] Bodenstedt S, Allan M, Agustinos A, et al. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. arXiv: Comput Vis Pattern Recogn. 2018.
[2] Zhao Z, Voros S, Chen Z, et al. Surgical tool tracking based on two CNNs: from coarse to fine. J Engg-Joe. 2019;2019(14):467–472.
[3] Jin Y, Li H, Dou Q, et al. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Med Image Anal. 2020;59:1–14.
[4] Bouget D, Allan M, Stoyanov D, et al. Vision-based and marker-less surgical tool detection and tracking: a review of the literature. Med Image Anal. 2017;35:633–654.
[5] Lecun Y, Bengio Y, Hinton GE. Deep learning. Nature. 2015;521(7553):436–444.
[6] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, et al, editors. Neural information processing systems. Red Hook (NY): Curran Associates Inc.; 2012. p. 1097–1105.
[7] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Mortensen E, Fidler S, editors. Computer Vision and Pattern Recognition. Washington (DC): IEEE Computer Society; 2014. p. 580–587.
[8] Nwoye CI, Mutter D, Marescaux J, et al. Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos. Int J Comput Assist Radiol Surg. 2019;14(6):1059–1067.
[9] Jin A, Yeung S, Jopling J, et al. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In: Medioni G, Hoogs A, McCloskey S, editors. IEEE Winter Conference on Applications of Computer Vision. Washington (DC): IEEE Computer Society; 2018. p. 691–699.
[10] Du X, Kurmann T, Chang P, et al. Articulated multi-instrument 2-D pose estimation using fully convolutional networks. IEEE Trans Med Imaging. 2018;37(5):1276–1287.
[11] Sarikaya D, Corso JJ, Guru KA. Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection. IEEE Trans Med Imaging. 2017;36(7):1542–1549.

[12] Chen ZR, Zhao ZJ, Cheng XL. Surgical instruments tracking based on deep learning with lines detection and spatio-temporal context. Proceedings of 2017 Chinese Automation Congress. New York (NY): IEEE; 2017. p. 2711–2714.
[13] Zhang K, Zhang L, Liu Q, et al. Fast visual tracking via dense spatio-temporal context learning. In: Fleet D, Pajdla T, Schiele B, et al., editors. Lecture notes in computer science. Proceedings of 2014 European Conference on Computer Vision. Cham-Heidelberg: Springer; 2014. p. 127–141.
[14] Padoy N, Blum T, Feussner H, et al. On-line recognition of surgical activity for monitoring in the operating room. In: Goker MH, editor. IAAI'08: Proceedings of the 20th National Conference on Innovative Aplications of Artificial Intelligence. Palo Alto (CA): AAAI Press; 2008. p. 1718–1724.
[15] Twinanda AP, Shehata S, Mutter D, et al. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imaging. 2017;36(1): 86–97.
[16] Stauder R, Okur A, Peter L, et al. Random forests for phase detection in surgical workflow analysis. In: Stoyanov D, Collins DL, Sakuma I, Abolmaesumi P, Jannin P, editors. Information processing in computer-assisted interventions. IPCAI 2014. Lecture notes in computer science. Vol 8498. Cham; Heidelberg: Springer; 2014. p. 148–157.
[17] Sahu M, Mukhopadhyay A, Szengel A, et al. Tool and Phase recognition using contextual CNN features. arXiv: Computer Vision and Pattern Recognition. 2016.
[18] Wang S, Raju A, Huang J. Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos. Proceedings of 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). Melbourne (VIC): IEEE; 2017. p. 620–623.
[19] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington (DC): IEEE Computer Society; 2014.
[20] Szegedy C, Liu W, Jia Y, et al.Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Boston, MA. Washington (DC): IEEE Computer Society; 2015. p. 1–9.
[21] Sahu M, Mukhopadhyay A, Szengel A, et al. Addressing multi-label imbalance problem of surgical tool detection using CNN. Int J Comput Assist Radiol Surg. 2017;12(6):1013–1020.
[22] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Mortensen E, Fidler S editors. IEEE Conference on Computer Vision and Pattern Recognition. Washington (DC): IEEE Computer Society; 2015. p. 3431–3440.
[23] Garcia-Peraza-Herrera LC, Li W, Gruijthuijsen C, et al. Real-time segmentation of non-rigid surgical tools based on deep learning and tracking. In: Peters T, Yang GZ, Navab N, Mori K, Luo X, Reichl T, et al. editors. Lecture notes in computer science. Cham; Heidelberg: Springer; 2017. p. 84–95.
[24] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017; 39(6):1137–1149.
[25] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Lecture notes in computer science. Cham; Heidelberg: Springer; 2014. p. 818–833.
[26] Nakazawa A, Harada K, Mitsuishi M, et al. Real-time surgical needle detection using region-based convolutional neural networks. Int J Comput Assist Radiol Surg. 2020;15(1):41–47.
[27] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection. In: Russakovsky O, editor. IEEE Conference on Computer Vision and Pattern Recognition. Washington (DC): IEEE Computer Society; 2016. p. 779–788.
[28] Choi B, Jo K, Choi S, et al. Surgical-tools Detection based on Convolutional Neural Network in Laparoscopic Robot-assisted Surgery. In: Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society. New York (NY): IEEE; 2017. p. 1756–1759.
[29] Jo K, Choi Y, Choi J, et al. Robust real-time detection of laparoscopic instruments in robot surgery using convolutional neural networks with motion vector prediction. Appl Sci-Basel. 2019;9(14). DOI:10.3390/app9142865
[30] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: Mortensen E, editor. IEEE Conference on Computer Vision and Pattern Recognition. Washington (DC): IEEE Computer Society; 2017. p. 6517–6525.
[31] Mishra K, Sathish R, Sheet D. Learning latent temporal connectionism of deep residual visual abstractions for identifying surgical tools in laparoscopy procedures. In: Mortensen E, editor. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Washington (DC): IEEE Computer Society; 2017. p. 2233–2240.
[32] Al Hajj H, Lamard M, Conze P, et al. Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks. Med Image Anal. 2018;47:203–218.
[33] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Lecture notes in computer science. Cham; Heidelberg: Springer; 2015. p. 234–241.
[34] Kurmann T, Marquez Neila P, Xiaofei D, et al. Simultaneous recognition and pose estimation of instruments in minimally invasive surgery. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S, editors. Medical image computing and computer assisted intervention - MICCAI 2017. Lecture notes in computer science. Vol.10434. Cham; Heidelberg: Springer; 2017. p. 505–513.
[35] Hou R, Chen C, Shah M. An end-to-end 3D convolutional neural network for action detection and segmentation in videos. arXiv: Computer Vision and Pattern Recognition. 2017.
[36] Maturana D, Scherer S. VoxNet: a 3D convolutional neural network for real-time object recognition. In: Wang Z, Papanikolopoulos N, editors. IEEE

International Conference on Intelligent Robots and Systems. New York (NY): IEEE; 2015. p. 922–928.

[37] Funke I, Mees ST, Weitz J, et al. Video-based surgical skill assessment using 3D convolutional neural networks. Int J Comput Assist Radiol Surg. 2019;14(7): 1217–1225.

[38] Colleoni E, Moccia S, Du X, et al. Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. IEEE Robot Autom Lett. 2019;4(3):2714–2721.

[39] Laina I, Rieke N, Rupprecht C, et al. Concurrent segmentation and localization for tracking of surgical instruments. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S, editors. Medical image computing and computer-assisted intervention – MICCAI 2017. MICCAI 2017. Lecture notes in computer Science. Vol 10434. Cham; Heidelberg: Springer; 2017. p. 664–672.

[40] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Mortensen E, Saenko K, editors. IEEE Conference on Computer Vision and Pattern Recognition. Washington (DC): IEEE Computer Society; 2016. p. 770–778.

[41] Mondal SS, Sathish R, Sheet D. Multitask Learning of Temporal Connectionism in Convolutional Networks using a Joint Distribution Loss Function to Simultaneously Identify Tools and Phase in Surgical Videos [arXiv]. arXiv. 2019. p. 15.

[42] Ma X, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Erk K, Smith NA, editors. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin (Germany): Association for Computational Linguistics; 2016. p. 1064–1074.

[43] Newell A, Yang K, Deng J. Stacked Hourglass networks for human pose estimation. In: Leibe B, Matas J, Sebe N, Welling M, editors. Lecture notes in computer science. Cham; Heidelberg: Springer; 2016. p. 483–499.

[44] Zhao Z, Cai T, Chang F, et al. Real-time surgical instrument detection in robot-assisted surgery using a convolutional neural network cascade. Healthc Technol Lett. 2019;6(6):275–279.

[45] Liu Y, Zhao Z, Chang F, et al. An anchor-free convolutional neural network for real-time surgical tool detection in robot-assisted surgery. IEEE Access. 2020;8: 78193–78201.

[46] Howard AG, Menglong Z, Bo C, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. [arXiv]. arXiv. 2017. p. 9.

[47] Jia Z, Huang X, Chang EI, et al. Constrained deep weak supervision for histopathology image segmentation. IEEE Trans Med Imaging. 2017;36(11): 2376–2388.

[48] Zhou Y, Zhu Y, Ye Q, et al. Weakly supervised instance segmentation using class peak response. In: Mortensen E, Brendel W, editors. IEEE Conference on Computer Vision and Pattern Recognition. Washington (DC): IEEE Computer Society; 2018. p. 3791–3800.

[49] Vardazaryan A, Mutter D, Marescaux J, et al. Weakly-supervised learning for tool localization in laparoscopic videos. In: Stoyanov D, Taylor Z, Balocco S, Sznitman R, editors. Intravascular imaging and computer assisted stenting and large-scale annotation of biomedical data and expert label synthesis. LABELS 2018, CVII 2018, STENT 2018. Lecture notes in computer science. Vol 11043. Cham; Heidelberg: Springer; 2018. p. 169–179.

[50] Attia M, Hossny M, Nahavandi S, et al. Surgical tool segmentation using a hybrid deep CNN-RNN auto encoder-decoder. Proceedings of 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC). New York (NY): IEEE; 2017. p. 3373–3378.