HHS Public Access

Author manuscript

Surg Endosc. Author manuscript; available in PMC 2023 January 01.

Published in final edited form as:

Surg Endosc. 2022 January ; 36(1): 679–688. doi:10.1007/s00464-021-08336-x.

A Contextual Detector of Surgical Tools in Laparoscopic Videos Using Deep Learning

Babak Namazi, Ph.D.[1], **Ganesh Sankaranarayanan, Ph.D.**[2], **Venkat Devarajan, Ph.D.**[3]

[1]Baylor Scott & White Research Institute, Dallas, Texas

[2]Department of Surgery, Baylor University Medical Center, Dallas, Texas

[3]Electrical Engineering Department, University of Texas at Arlington, Texas

Abstract

Background—The complexity of laparoscopy requires special training and assessment. Analyzing the streaming videos during the surgery can potentially improve surgical education. The tedium and cost of such an analysis can be dramatically reduced using an automated tool detection system, among other things. We propose a new multilabel classifier, called LapTool-Net to detect the presence of surgical tools in each frame of a laparoscopic video.

Methods—The novelty of LapTool-Net is the exploitation of the correlations among the usage of different tools and, the tools and tasks - i.e., the context of the tools' usage. Towards this goal, the pattern in the co-occurrence of the tools is utilized for designing a decision policy for the multilabel classifier based on a Recurrent Convolutional Neural Network (RCNN), which is trained in an end-to-end manner. In the post-processing step, the predictions are corrected by modeling the long-term tasks' order with an RNN.

Results—LapTool-Net was trained using publicly available datasets of laparoscopic cholecystectomy, viz., M2CAI16 and Cholec80. For M2CAI16, our exact match accuracy (when all the tools in one frame are predicted correctly) in online and offline modes were 80.95% and 81.84% with per-class F1-score of 88.29% and 90.53%. For Cholec80, the accuracies were 85.77% and 91.92% with F1-scores if 93.10% and 96.11% for online and offline respectively.

Conclusions—The results show LapTool-Net outperformed state-of-the-art methods significantly, even while using fewer training samples and a shallower architecture. Our context-aware model does not require expert's domain-specific knowledge and the simple architecture can potentially improve all existing methods.

Keywords

Convolutional Neural Networks; Recurrent Neural Networks; Tool detection; Laparoscopic surgery; Label power-set

Corresponding Author: Ganesh Sankaranarayanan, 3500 Gaston Ave, Dallas, TX 75246, Tel.: +1-214-820-6755, ganesh.sankaranarayanan@bswhealth.org.

Presented as poster at SAGES 2017.

Disclosure

Babak Namazi, Ganesh Sankaranarayanan and Venkat Devarajan have no conflicts of interest or financial ties to disclose.

Introduction

Numerous advantages of minimally invasive surgery such as shorter recovery time, less pain and blood loss, and better cosmetic results, make it the preferred choice over conventional open surgeries [1]. In laparoscopy, the surgical instruments are inserted through small incisions in the abdominal wall and the procedure is monitored using a laparoscope. The special way of manipulating the surgical instruments and the indirect observation of the surgical scene introduce more challenges in performing laparoscopic procedures [2]. The complexity of laparoscopy requires special training and assessment for the surgery residents to gain the required bi-manual dexterity. Analyzing the streaming videos during the surgery and the recorded videos from previously accomplished procedures can potentially improve the outcomes. The tedium and cost of such an analysis can be dramatically reduced using an automated tool detection system, among other things and is, therefore, the focus of this paper.

Tracking surgical tools is essential in understanding the workflow of a procedure and in the assessment and rating of the videos. For example, it has been shown that experts have a better economy of motion compared to novice or less experienced surgeons [3, 4, 5]. Also, by detecting the tools, we can check for wrong tool usage, monitor activation time of electro-surgical tools, and the use of proper technique (how a needle is positioned and moved with a needle driver during suturing) etc.

Manual annotation of long videos from surgeries is a time-consuming and expensive task. A vision-based algorithm for automated detection of the presence, location or movement of surgical tools is indispensable in designing a fast and objective surgical evaluation system. A well-annotated database of surgical videos can also be used in information retrieval and is a reliable source for education and training of the future surgeons.

During surgery, monitoring the usage of surgical tools can provide real-time feedback to the surgeons and operating room staff. Furthermore, in computer-aided intervention, the surgical tools are controlled by a surgeon with the aid of a specially designed robot [6], which requires a real-time understanding of the current task. Therefore, detecting the presence, location or pose of the surgical instruments is useful in robotic surgeries as well [7, 8, 9]. Finally, an automated tool usage detector can help to generate an operative summary.

Recent years have witnessed great advances in deep learning techniques in various computer vision areas such as image classification, object detection, and segmentation etc., and in medical imaging [10]. Therefore, there is a trend towards using these methods in analyzing the videos taken from laparoscopic operations.

Endonet [11] was the first deep learning model designed for detecting the presence of surgical instruments in laparoscopic videos, wherein Alexnet [12] was used as a Convolutional Neural network (CNN), for feature extraction and was trained for the simultaneous detection of surgical phases and instruments. Inspired by this work, other researchers used different CNN architectures [13, 14] to classify the frames based on the visual features. For example, in [15], three CNN architectures were used, and [16] proposed an ensemble of two deep CNNs.

Sahu et al. [17] were the first to address the imbalance in the classes in a Multi-Label (ML) classification of video frames. They balanced the training set according to the combinations of the instruments. The data were re-sampled to have a uniform distribution in label-set space and, class re-weighting was used to balance the data. Despite the improvement gained by considering the co-occurrence in balancing the training set, the correlation of the tools' usage was not considered directly in the classifier and the decision was made solely based on the presence of single tools. Alshirbaji et al. [18] used class weights and re-sampling together to deal with the imbalance issue.

In order to consider the temporal features of the videos, Twinanda et al. employed a hidden Markov model (HMM) in [11] and Recurrent Neural Network (RNN) in [19]. Sahu et.al utilized a Gaussian distribution fitting method in [13] and a temporal smoothing method based on a moving average in [17] to improve the classification results, after the CNN was trained. Mishra et al. [20] were the first to apply a Long Short-Term Memory model (LSTM) [21], as an RNN to a short sequence of frames, to simultaneously extract both spatial and temporal features for detecting the presence of the tools by end-to-end training.

A variety of different approaches were as following. Hu et al. [22] proposed an attention guided method using two deep CNNs to extract local and global spatial features. In [23], a boosting mechanism was employed to combine different CNNs and RNNs. In [24], the tools were localized, after labeling the dataset with bounding boxes containing the surgical tools.

It should be noted that none of the previous methods takes advantage of any knowledge regarding the order of the tasks and, the correlations of the tools are not directly utilized in identifying different surgical instruments. In this paper, we propose a novel context-aware model called LapTool-Net to detect the presence of surgical instruments in laparoscopic videos. The uniqueness of our approach is based on the following three original ideas:

- A novel ML classifier is proposed as a part of LapTool-Net, to take advantage of the co-occurrence of different tools in each frame – in other words, the context is taken into account in the detection process. To accomplish this objective, each combination of tools is considered as a separate class during training and testing and, is further used as a decision model for the ML classifier. To the best of our knowledge, this is the first attempt at directly using the information about the co-occurrence of tools in laparoscopic videos in the classifier's decision-making.
- The ML classifier and the decision model are trained in an end-to-end fashion. For this purpose, the training is performed by jointly optimizing the loss functions for the ML classifier and the decision model using a multitask learning approach.
- At the post-processing step, the trained model's prediction for each video is sent to another RNN to consider the order of the usage of different tools/tool combinations and long-term temporal dependencies; yet another consideration for the context.

The pre-print version of this paper with more results and detailed discussions can be found in [25]. The preliminary results were presented at the SAGES 2017 Annual Meeting.

Materials and Methods

The overview of the proposed model is illustrated in Fig. 1. Let $D = \{(x_{ij}, Y_{ij}) | 0 \leq i < m, 0 < j < n\}$ be a ML dataset, where $x_{ij} \in \mathscr{R}^d$ is the ith frame of the jth video and $Y_{ij} \subseteq \mathscr{Y}$ is the corresponding surgical instruments and $\mathscr{Y} \triangleq \{y_1, y_2, \ldots, y_K\}$ is the set of all possible tools. Each subset of $\mathscr{Y}$ is called a label-set and each frame can have a different number of labels $|Y_{ij}|$. The tools associations can also be represented as a K dimensional binary vector $y_{ij} = (y_1, y_2, \ldots, y_K) = \{0, 1\}^K$, where each element is a 1 if the tool is present and a 0 otherwise. The goal is to design a classifier $F(x)$ that maps the frames of surgical videos, to the tools in the observed scene.

To take advantage of the combination of the surgical tools in a laparoscopic video, the well-known label power-set (LP) method is adopted in a novel way. The output of $F(x_{ij})$ is a label-set $\widehat{Y_{ij}} \subseteq \widehat{\mathscr{Y}}$ (also called a superclass) of size $\left|\widehat{Y_{ij}}\right| \leq K$, where $\widehat{\mathscr{Y}}$ is the set of all possible subsets of $\mathscr{Y}$.

In order to calculate the confidence scores for each tool, along with the final decision, which is the class index in $\widehat{\mathscr{Y}}$, the classifier F is decomposed into $F(.)=g(f(.))$, where $g(f(x_{ij})): \mathscr{R}^K \rightarrow \mathscr{R}^{\hat{K}}$ is the decision model, which maps the confidence scores of the frame i of the video j to the label-set $\widehat{Y_{ij}}$. The model f takes the video frames as input and produces the confidence scores $P = (p_1, p_2, \ldots, p_k) = [0,1]^K$, where each element is the probability of the presence of one tool from the set $\mathscr{Y}$. The output of the decision model g for all the frames of each video forms a larger sequence $\bar{C}$ of the model's predictions. The sequence is used as the input to another RNN, g' to exploit the long-term order of the tool usage.

The overall system is described based on a the dataset from M2CAI16[1] tool detection challenge, which is a subset of Cholec80 dataset [11]. We chose the smaller dataset to highlight the improvements caused by the main contributions of this paper. The dataset contains 15 videos from cholecystectomy procedure, which is the surgery for removing the gallbladder. All the videos are labeled with seven tools for every 25 frames. The tools are Bipolar, Clipper, Grasper, Hook, Irrigator, Scissors, and Specimen bags. There are ten videos for training and five videos for validation. The type and shape of all seven tools remain the same for the training and validation sets.

Since the publicly available Cholec80 dataset was used in this study to train and test our deep learning model, an Institutional Review Board (IRB) approval is not required for this study.

Spatio-temporal Features:

To detect the presence of surgical instruments in laparoscopic videos, the visual features (intra-frame spatial and inter-frame temporal features) need to be extracted. We use CNN to extract spatial features. As shown in Figure 1, the input frame x_{ij} is sent through the trained

[1] http://camma.u-strasbg.fr/m2cai2016/index.php/program-challenge

CNN and the output of the last convolutional layer (after pooling) forms a fixed size spatial feature vector v_{ij}.

Since there is a high correlation among video frames, it can be exploited by an RNN to improve the performance of the tool detection algorithm. An RNN uses its internal memory (states) to process a sequence of inputs for time series and videos processing tasks [26]. This helps the model identify the tools even when they are occluded or not clear due to motion blur. For this purpose, short sequences of frames (say 5 frames) are selected.

For each frame x_{ij}, the sequence of the spatial features is the input for the RNN. The total length of the input is no longer than one second, which ensures that the tools remain visible during that time interval. We selected Gated Recurrent Unit (GRU) [27] as our RNN for its simplicity. The final hidden state h_{ij} is the output of the GRU and is the input to a fully connected neural network FC1.

Decision Model:

One of the main challenges in ML classification is effectively utilizing the correlation among different classes. In LP, multiple classes are combined into one superclass and the problem is transformed into an MC classification. The advantage of LP is that the class dependencies are automatically considered. Also, by eliminating uncommon combinations from the outputs, the classifier's attention is directed towards the more possible combinations.

In a laparoscopic surgery, not all the 2^K combinations are possible as the total number of incisions are typically 3 or 4. Figure 2 shows the percentage of the most likely combinations in the M2CAI dataset. The first 15 classes out of a possible maximum of 128 span more than 99.5% of the frames in both the training and the validation sets, and the tools combinations have almost the same distribution in both cases.

Since an LP classifier is MC, training with Softmax loss requires the classes to be mutually exclusive. In other words, each superclass is treated as a separate class, i.e. separate features activate a superclass. This causes performance degradation in the classifier and therefore, more data is required for training. We address this issue by a novel use of LP as the decision model g, which we apply to the ML classifier f. Our method helps the classifier to consider our superclasses as the combinations of classes rather than separate mutually exclusive classes.

The decision model is a fully connected neural network (FC2), which takes the confidence scores of f and maps them to the corresponding superclass. When the Softmax function is applied, the output of g(.) is the probability of each. The final prediction c_{ij} of the tool detector F is the index of the superclass with the highest probability.

Class Imbalance:

It is known that in skewed datasets, the classifier's decision is inclined towards the majority classes. Therefore, it is always beneficial to have a uniform distribution for the classes during training. This can be accomplished using over-sampling for the minority classes and

under-sampling for the majority classes. However, in ML classification, finding a balancing criterion for re-sampling is challenging [28].

To overcome imbalance, we perform under-sampling to have a uniform distribution of the combination of the classes. The main advantage of under-sampling over other re-sampling methods is that it can also be applied to avoid overfitting caused by the high correlation between the neighboring frames of a laparoscopic video. Therefore, we try different under-sampling rate to find the smallest training set without sacrificing the performance.

Figure 3 shows the relationship among the tools after re-sampling. It can be seen that the LP-based balancing method not only tends to a uniform distribution in the superclass space, it also improves the balance of the dataset in the single class space (with the exception of Grasper, which can be used with all the tools).

Since this approach will not guarantee balance, a cost-sensitive weighting approach can be used along with an ML loss, prior to the LP decision layer; nonetheless, we empirically found that this does not affect the performance of the ML classifier.

Training:

Having the vector of the confidence scores P, the ML loss L_f is the sigmoid cross-entropy (CE) and the Softmax CE loss function L_g is used for training the decision model. We use the joint training paradigm for optimizing the ML and MC losses as a multitask learning approach. Two optimizers are defined based on the two losses with separate hyper-parameters such as learning rate and trainable weights θ. Using Stochastic Gradient Descent (SGD), the weights update at iteration l and the minibatch b can be written as:

$$\theta^{(l)} = \theta^{(l-1)} - \sum_{x \in D_b} \left[n_f \nabla_\theta L_f\left(\theta^{(l-1)}\right) + \beta \cdot n_g \nabla_\theta L_g\left(\theta^{(l-1)}\right) \right]$$

where β is a constant weight for adjusting the impact of the two loss functions, and η_f and η_g are the learning rates for the ML and MC loss functions respectively. The training is performed in an end-to-end fashion and the gradients ($\nabla_\theta L$) are calculated using the back propagation through time (BPTT) method.

The trainable weights for the ML optimizer are all the weights in the CNN, the weights in the RNN, and FC1. On the other hand, for the MC optimizer, the CNN, RNN, and FC2 are trainable. Note that the shared weights between the two optimizers are the RCNN weights. By keeping the FC1 layer untouched by the MC optimizer, the spatio-temporal features are extracted by the RCNN, considering both the presence of each tool and the combination of them, and FC2 is solely trained as a decision model.

Post-processing:

To smooth the RCNN prediction and consider the long-term ordering of the tools, we model the order in the usage of the tools with an RNN over all the frames of each video [29]. Due to memory constraints, the final predictions of the RCNN, $\bar{C}(j) = [c_{0j}, c_{1j}, \ldots, c_{mj}]$ for 0<j<n, is selected as the input for the post-processing RNN.

In the online mode, only the past frames are available for classifying the current frame. In the offline mode, future frames can also be used along with past frames to improve the classification results of the current frame. To accomplish this, a bi-directional RNN is employed. The post-processing RNN is a two-layer GRU with 128 and 32 units in each layer.

The post-processing method described in this section is similar to [23] in extracting the long-term temporal features using RNNs. However, in contrast to these researchers, we used the final predictions of the RCNN model instead of the vector of confidence scores of the tools. Besides containing the information about the co-occurrences, training RNNs can be accomplished easier with a single scalar versus the vector of the size of the total number of tools or the tools' combinations. With the aid of the shorter size input, we were able to train larger sequences, even after performing the temporal data augmentation (to be explained later).

Results

In this section, the performance of the different parts of the proposed tool detection model is validated through numerous experiments using the appropriate metrics. We selected Tensorflow [30] for all of the experiments. The CNN in all the experiments was Inception V1 [31]. To have better generalization, extensive data augmentation, such as random cropping, horizontal and vertical flipping, rotation and a random change in brightness, contrast, saturation, and hue were performed during training. The initial learning rate was 0.001 with a decay rate of 0.7 after 5 epochs and the results were taken after 100 epochs. The batch size was 32 for training the CNN models and, 40 for the RNN-based models. All the experiments were conducted using an Nvidia TITAN XP GPU.

LapTool-Net Results on M2CAI Dataset:

Since the dataset was labeled only for one frame per second (out of 25 frames/sec), there was a possibility of using the unlabeled frames for training, as long as the tools remain the same between two consecutive labeled frames. We used this unlabeled data to balance the training set, according to the LPs.

To balance the datasets, 15 superclasses were selected and the original frames were re-sampled to have a uniform distribution. The numbers of frames for each superclass were randomly selected to be 400, forming a training set of 6000 frames. In other words, under-sampling was performed based on the tool combinations.

We tested the model before and after adding the decision model. For training the RCNN model, we used 5 frames at a time (current frame and 4 previous frames) with an inter-frame interval of 5, which resulted in a total distance of 20 frames between the first and the last frame. The RCNN model was trained with a Stochastic Gradient Descent (SGD) optimizer. The value of β in the joint training paradigm is selected to be 0.1. The data augmentation for the post-processing model includes adding random noise to the input and randomly dropping frames to change the duration of the sequences; the final predictions of the RCNN model are

saved every 20 frames, and the frames are dropped with the probability of 10 to 30%. Table 1 shows the results of the proposed RCNN and LapTool-Net.

It can be seen that by considering the temporal features through the RCNN model, the exact match accuracy and F1-macro were improved by 3.15% and 7.52% respectively. Also, the F1-macro improves by 2.94% after adding the LP decision model.

The higher performance of the LapTool-Net, shown in Table 1 is due to consideration of the long-term order of the usage of the tools. In the offline mode, the utilization of the frames from both the past and the future of the current frame causes the improvements over the online model in accuracy and F1 scores.

To check the effectiveness of the multi-task approach used for the end-to-end training of the RCNN-LP model, we took the output of the ML classifier, after removing the decision model from the trained RCNN-LP. In other words, we replaced the LP-based decision layer of the trained model with the threshold-based decision method. The results are shown in Table 2.

Compared with the RCNN results in Table \ref{rnn0}, we can see 1.55% improvement in F1-macro after training with the proposed method. The reason is that with the help of the joint training strategy, the presence of the tools is detected based on the pattern of the tool combinations and therefore, richer extracted features. The precision for the Grasper is an indicator of the remaining imbalance in the training set. We believe this could have been better by re-weighting the loss function, especially for Grasper.

It is worth mentioning that the results from Table 2 show that the RCNN model without the LP decision can be taken for making prediction for all the combinations including the rare combinations that were originally excluded during training. In this setting, the role of the LP decision model can be interpreted as an auxiliary training task to improve the performance of the ML classifier. The importance of the ML classifier can be adjusted by the value of β, in case the rare combinations are of more significance in our application.

In order to localize the predicted tools, the attention maps were visualized using grad-CAM method [32]. The results for some of the frames are shown in figure \ref{visualization}. In order to avoid confusion with frames that multiple tools, only the class activation map of a single tool is shown based on the prediction of the model. The results show that the visualization of the attention of the proposed model can also be used in reliably identifying the location of each tool without any additional annotations for the location and shape of the tools.

Comparison with Current Work:

To validate the proposed model, we compared it with previously published research on the M2CAI dataset. The result is shown in Table 3. We show that our model out-performed previous methods by a significant margin even when choosing a relatively shallower model (Inception V1) and while using less than 25% of the labeled images.

It is worth mentioning that a fair comparison with previous work on the same dataset is not feasible, since the evaluation metrics might not be the same. Nevertheless, we compared our ML classifier f, which is the RCNN model, along with the final models to show the superiority of our balancing and temporal consideration methods. Regardless of the choice of the CNN architecture, which is the most dominant component that can affect the results, the superiority of our model over the works in Table 3 is due to the end-to-end temporal consideration and the inclusion of the context such as the co-occurrence and tasks ordering, which are the main contributions of this paper.

LapTool-Net Results on Cholec80 Dataset:

In this section, the performance of our model is evaluated on a larger dataset of laparoscopic cholecystectomy videos called Cholec80. We used the first 40 videos for training and the remaining 40 videos for testing our model.

The total number of tool combinations in Cholec80 dataset is 32, out of which 20 combinations are present in over 99.5% of the duration of videos. Compared with M2CAI dataset, the higher number of tool combinations are due to the more diversity in the larger dataset. Nonetheless, the extra five superclasses in Cholec80 dataset contain less than 0.4% of all frames. For each of the 20 tool combinations, 1500 samples were selected, forming a uniform class distribution on 30K frames.

We used the same model as for M2CAI dataset for extracting the spatio-temporal features, the decision policy, and the post-processing step, as well as the training strategy. The results for the different parts of the model are shown in Table 4.

Compared with the M2CAI results in Table 1, we can see significant improvement in accuracy and F1-scores. For example, the F1-macro of the CNN on the balanced Cholec80 is 9.19% higher than M2CAI dataset. Also, the mAP of the RCNN is 92.44%, which is 4.56% higher than M2CAI dataset (shown in Table 3). The main reason is the higher variations of the visual features in the input of the larger dataset.

As was to be expected, the accuracy and F1-scores increase after adding the LP-based decision layer. However, the improvements are relatively smaller compared with the M2CAI results. For instance, the F1-macro of the RCNN-LP is less than one percent higher than RCNN. Similarly, the increase in the F1-macro for the CNN and RCNN is less compared with M2CAI dataset (less than 5% versus over 10% in M2CAI). The reason behind this observation is likely due to the fact that while the end-to-end training of the CNN, RNN, and LP layer results in the richer discriminating features, considering the co-occurrence and temporal coherence, the performance is dominated and bounded by the capacity of the CNN.

Discussion

In this paper, we proposed a novel system called LapTool-Net, for automatically detecting the presence of tools in every frame of a laparoscopic video. The main feature of the proposed RCNN model is the context-awareness, i.e. the model learns the short-term and long-term patterns of the usage of the tools by utilizing the correlation between the usage

of the tools with each other and, with the surgical steps. Our method outperformed all previously published results on M2CAI dataset, while using less than 1% of the total frames in the training set.

While our model is designed based on the previous knowledge of the cholecystectomy procedure, it does not require any domain-specific knowledge from experts and can be effectively applied to any video captured from laparoscopic or even other forms of surgeries. Also, the relatively small training set after under-sampling suggests that the labeling process can be accomplished faster by using fewer frames (e.g. one frame every 5 seconds). Moreover, the simple architecture of the proposed LP-based classifier makes it easy to use it with other proposed models such as [23] and [22], or with weakly supervised models [33, 34] to localize the tools in the frames. To accomplish that, the threshold mechanism of the ML classifier in all these papers can be simply replaced by our combination-aware decision model.

Though the proposed end-to-end model for considering the tools combination pattern in laparoscopic videos resulted in a significant performance gain, while capturing the dependencies in different tools' usages, it requires all the possible combinations of tools to be present in the training set in order to make the correct prediction on all of them. For instance, in Cholec80 dataset, we found four unique combinations in the test set, without having the corresponding training samples. Furthermore, the tool combinations might slightly vary for each operation. For example, we found 20 combinations in 80 videos and 15 combinations in 15 videos of the same procedure in 99.5% of the frames, which suggests that the combination patterns can be surgeon/ patient dependent. Therefore, the LP-based decision policy might not be suitable if detecting the rare combinations is of priority. One possible solution to address this issue is using graph neural networks [35] for learning the tool dependencies.

Acknowledgements

The authors would like to thank NVIDIA Inc. for donating the TITAN XP GPU through the GPU grant program.

Funding:

This work was supported by Joseph Seeger Surgical Foundation award from the Baylor University Medical Center at Dallas.

References

[1]. Velanovich V, "Laparoscopic vs open surgery," Surgical Endoscopy, vol. 14, no. 1, pp. 16–21, 8 2000. [PubMed: 10653229]

[2]. Ballantyne GH, "The pitfalls of laparoscopic surgery: challenges for robotics and telerobotic surgery.," Surgical laparoscopy, endoscopy & percutaneous techniques, vol. 12, no. 1, pp. 1–5, 8 2002.

[3]. A. P. C. M B. T. C George Evalyn I and Skinner, "Performance Assessment in Minimally Invasive Surgery," in Surgeons as Educators : A Guide for Academic Development and Teaching Excellence, Köhler B Tobias S and Schwartz, Ed., Cham, Springer International Publishing, 2018, pp. 53–91.

[4]. L. S S. D K. R F. G. M Sherman V and Feldman, “Assessing the learning curve for the acquisition of laparoscopic skills on a virtual reality simulator,” Surgical Endoscopy And Other Interventional Techniques, vol. 19, no. 5, pp. 678–682, 5 2005. [PubMed: 15776208]

[5]. M. T. N J. J-P F. J B. L H. J Perrenot Cyril and Perez, “The virtual reality simulator dV-Trainer® is a valid assessment tool for robotic surgical skills,” Surgical Endoscopy, vol. 26, no. 9, pp. 2587–2593, 9 2012. [PubMed: 22476836]

[6]. Antico M, Sasazawa F, Wu L, Jaiprakash A, Roberts J, Crawford R, Pandey AK and Fontanarosa D, “Ultrasound guidance in minimally invasive robotic procedures,” Medical Image Analysis, vol. 54, pp. 149–167, 8 2019. [PubMed: 30928829]

[7]. Du X, Allan M, Dore A, Ourselin S, Hawkes D, Kelly JD and Stoyanov D, “Combined 2D and 3D tracking of surgical instruments for minimally invasive and robotic-assisted surgery,” International Journal of Computer Assisted Radiology and Surgery, vol. 11, no. 6, pp. 1109–1119, 8 2016. [PubMed: 27038963]

[8]. Allan M, Ourselin S, Thompson S, Hawkes DJ, Kelly J and Stoyanov D, “Toward Detection and Localization of Instruments in Minimally Invasive Surgery,” IEEE Transactions on Biomedical Engineering, vol. 60, no. 4, pp. 1050–1058, 8 2013. [PubMed: 23192482]

[9]. Allan M, Ourselin S, Hawkes DJ, Kelly JD and Stoyanov D, “3-D Pose Estimation of Articulated Instruments in Robotic Minimally Invasive Surgery,” IEEE Transactions on Medical Imaging, vol. 37, no. 5, pp. 1204–1213, 8 2018. [PubMed: 29727283]

[10]. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B and Sánchez CI, “A survey on deep learning in medical image analysis,” Medical Image Analysis, vol. 42, pp. 60–88, 8 2017. [PubMed: 28778026]

[11]. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M and Padoy N, “EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos,” IEEE Transactions on Medical Imaging, vol. 36, no. 1, pp. 86–97, 8 2017. [PubMed: 27455522]

[12]. Krizhevsky A, Sutskever I and Hinton GE, “ImageNet Classification with Deep Convolutional Neural Networks,” in Advances In Neural Information Processing Systems, 2012.

[13]. Sahu M, Mukhopadhyay A, Szengel A and Zachow S, “Tool and Phase recognition using contextual CNN features,” 8 2016.

[14]. Prellberg J and Kramer O, “Multi-label Classification of Surgical Tools with Convolutional Neural Networks,” 2018.

[15]. Zia A, Castro D and Essa I, “Fine-tuning Deep Architectures for Surgical Tool Detection,” in Workshop and Challenges on Modeling and Monitoring of Computer Assisted Interventions (M2CAI), 2016.

[16]. Wang S, Raju A and Huang J, “Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos,” in Proceedings - International Symposium on Biomedical Imaging, 2017.

[17]. Sahu M, Mukhopadhyay A, Szengel A and Zachow S, “Addressing multi-label imbalance problem of surgical tool detection using CNN,” International Journal of Computer Assisted Radiology and Surgery, vol. 12, no. 6, pp. 1013–1020, 8 2017. [PubMed: 28357628]

[18]. Abdulbaki Alshirbaji T, Jalal NA and Möller K, “Surgical Tool Classification in Laparoscopic Videos Using Convolutional Neural Network,” Current Directions in Biomedical Engineering, vol. 4, no. 1, pp. 407–410, 8 2018.

[19]. Twinanda AP, Padoy N, Troccaz MJ and Hager G, “Vision-based Approaches for Surgical Activity Recognition Using Laparoscopic and RBGD Videos,” 2017.

[20]. Mishra K, Sathish R and Sheet D, “Learning Latent Temporal Connectionism of Deep Residual Visual Abstractions for Identifying Surgical Tools in Laparoscopy Procedures,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

[21]. Hochreiter S and Schmidhuber J, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 8 1997. [PubMed: 9377276]

[22]. Hu X, Yu L, Chen H, Qin J and Heng P-A, “AGNet: Attention-Guided Network for Surgical Tool Presence Detection,” in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, Cham, 2017, pp. 186–194.

[23]. Al Hajj H, Lamard M, Conze P-H, Cochener B and Quellec G, “Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks,” Medical Image Analysis, vol. 47, pp. 203–218, 8 2018. [PubMed: 29778931]

[24]. Jin A, Yeung S, Jopling J, Krause J, Azagury D, Milstein A and Fei-Fei L, “Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks,” in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018.

[25]. Namazi B, Sankaranarayanan S and Devarajan V, “LapTool-Net: A Contextual Detector of Surgical Tools in Laparoscopic Videos Based on Recurrent Convolutional Neural Networks,” Arxiv Pre-print, 2019.

[26]. Jin Y, Dou Q, Chen H, Yu L, Qin J, Fu C-W and Heng P-A, “SV-RCNet: Workflow Recognition From Surgical Videos Using Recurrent Convolutional Network,” IEEE Transactions on Medical Imaging, vol. 37, no. 5, pp. 1114–1126, 8 2018. [PubMed: 29727275]

[27]. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H and Bengio Y, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

[28]. Charte F, Rivera AJ, del Jesus MJ and Herrera F, “Addressing imbalance in multilabel classification: Measures and random resampling algorithms,” Neurocomputing, vol. 163, pp. 3-16, 8 2015.

[29]. Namazi B, Sankaranarayanan G and Devarajan V, “Automatic Detection of Surgical Phases in Laparoscopic Videos,” in Proceedings on the International Conference in Artificial Intelligence (ICAI), 2018.

[30]. Abadi e. a. , “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems”.

[31]. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A, Going Deeper With Convolutions, 2015, pp. 1–9.

[32]. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D and Batra D, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in 2017 IEEE International Conference on Computer Vision (ICCV), 2017.

[33]. Nwoye CI, Mutter D, Marescaux J and Padoy N, “Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos,” International Journal of Computer Assisted Radiology and Surgery, vol. 14, no. 6, pp. 1059–1067, 8 2019. [PubMed: 30968356]

[34]. Vardazaryan A, Mutter D, Marescaux J and Padoy N, “Weakly-Supervised Learning for Tool Localization in Laparoscopic Videos,” in Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, Springer, Cham, 2018, pp. 169–179.

[35]. Chen Z-M, Wei X-S, Wang P and Guo Y, “Multi-Label Image Recognition With Graph Convolutional Networks,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[36]. He K, Zhang X, Ren S and Sun J, “Deep Residual Learning for Image Recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[37]. Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z, “Rethinking the Inception Architecture for Computer Vision,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[38]. Twinanda AP, Mutter D, Marescaux J, de Mathelin M and Padoy N, “Single- and Multi-Task Architectures for Tool Presence Detection Challenge at M2CAI 2016,” 8 2016.

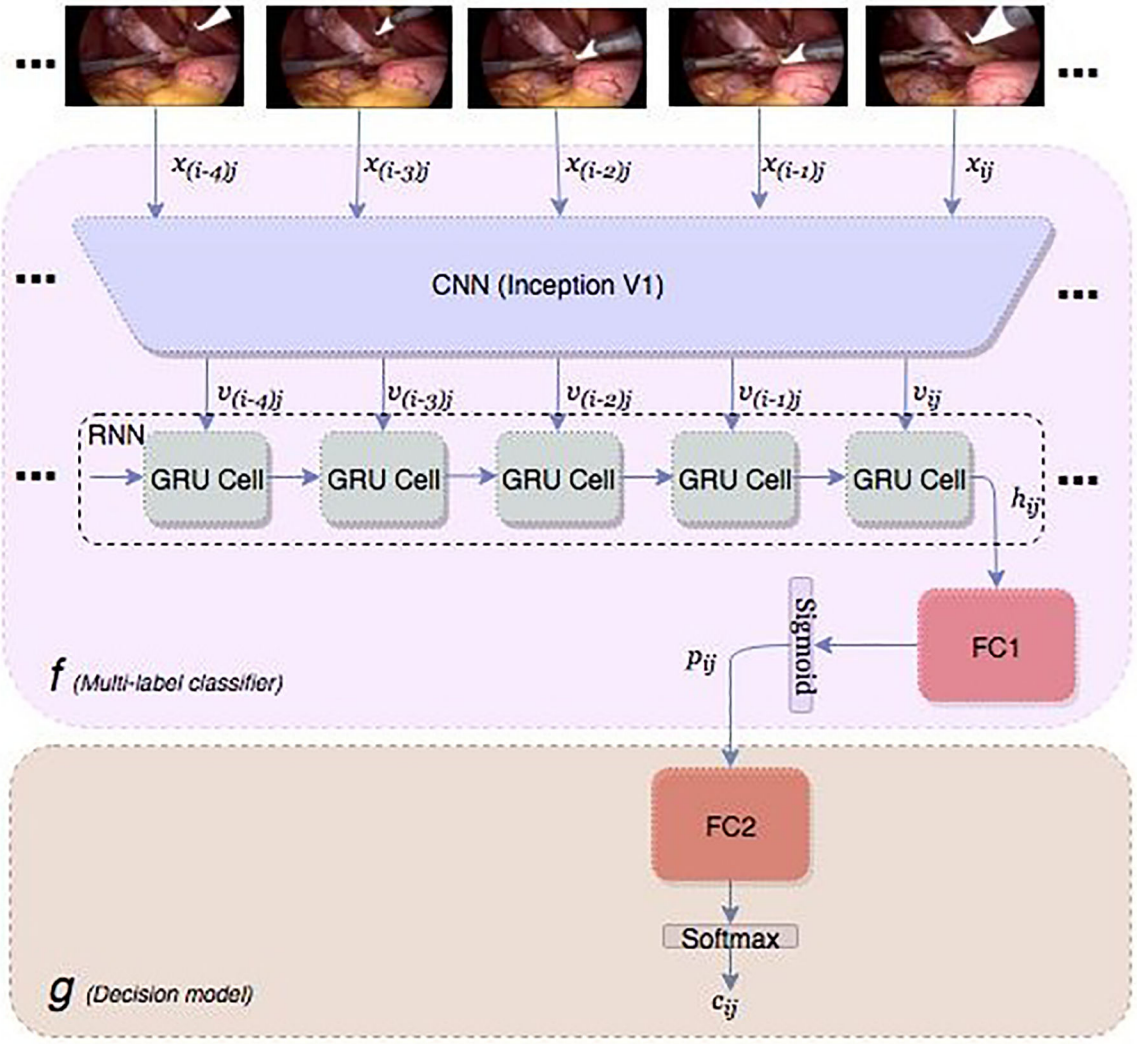


Figure 1.

Block diagram of the proposed classifier for detecting the presence of surgical tools in each frames of a laparoscopic video

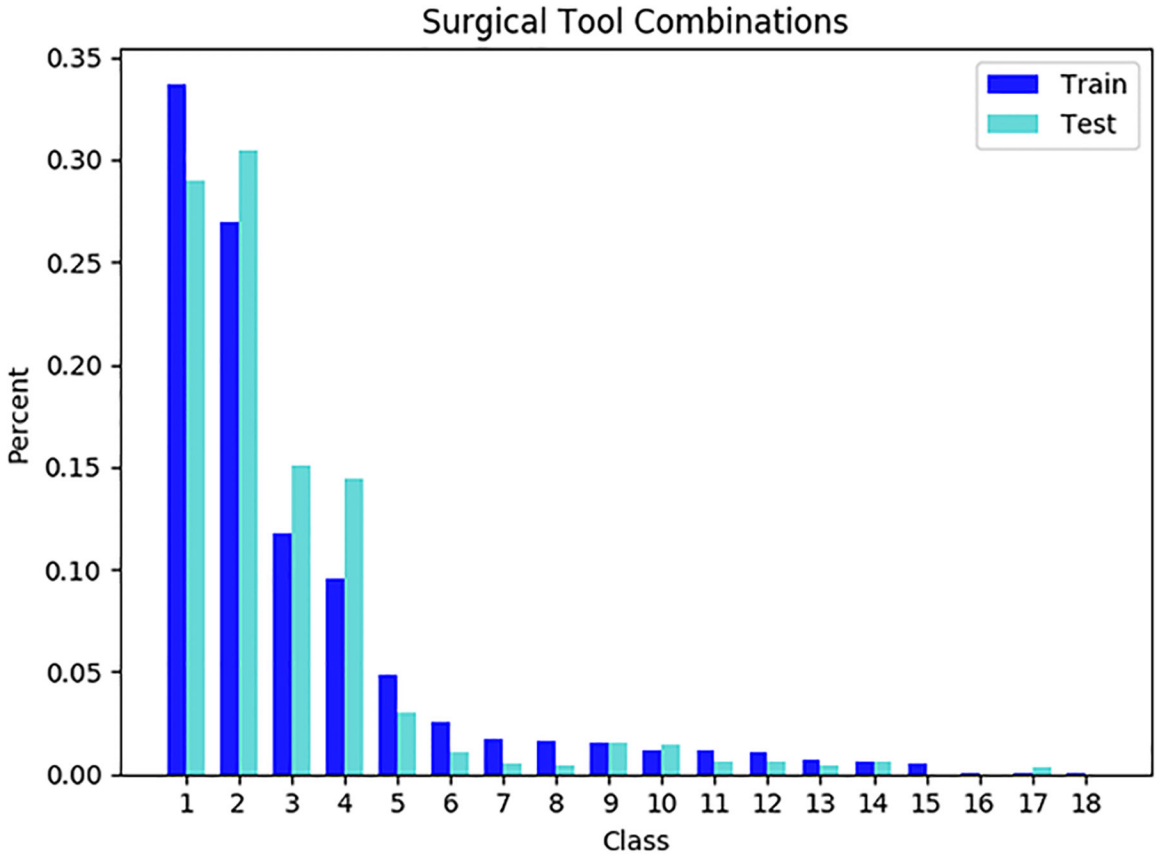
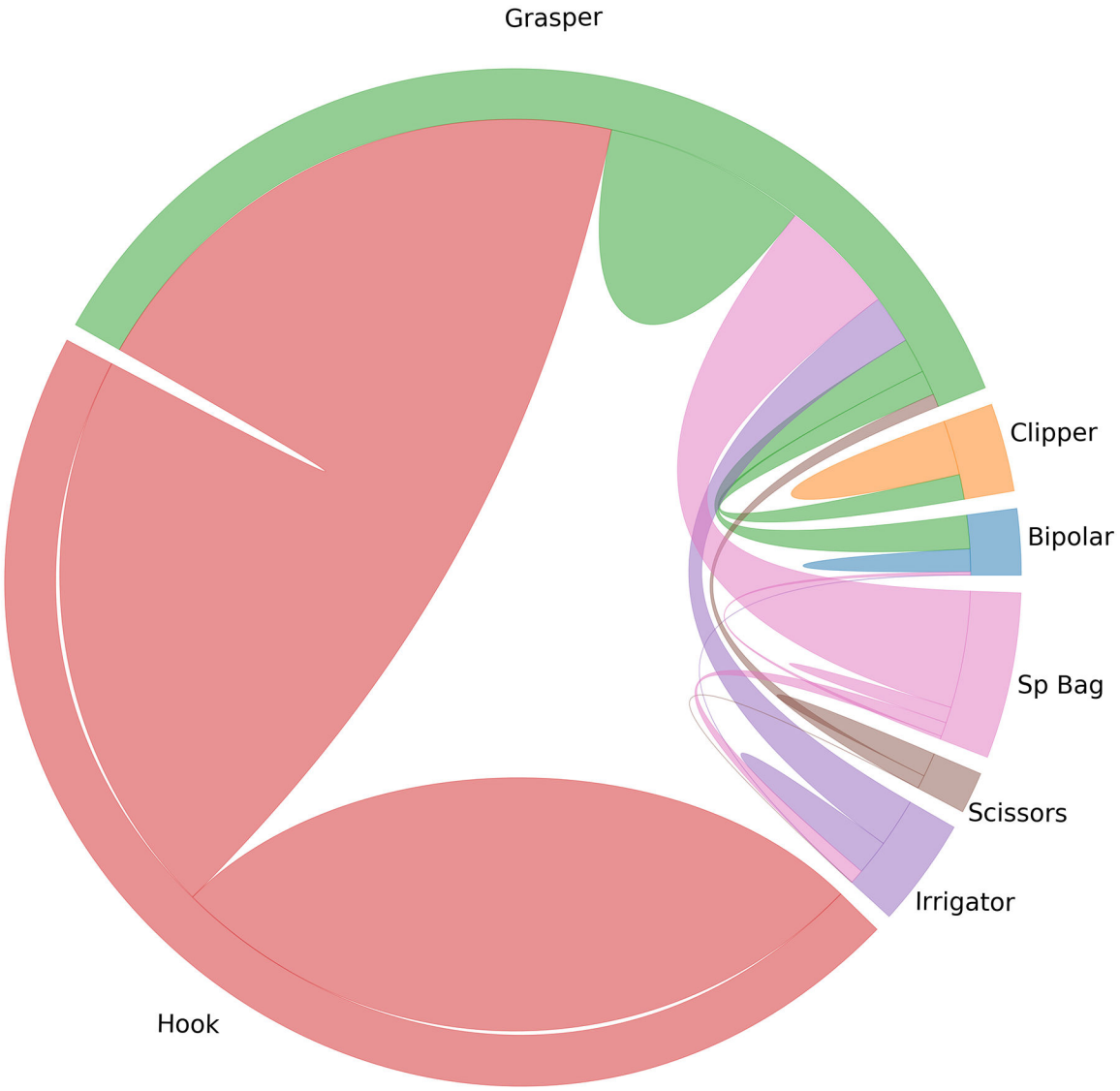


Figure 2.
The distribution for the combination of the tools in M2CAI dataset



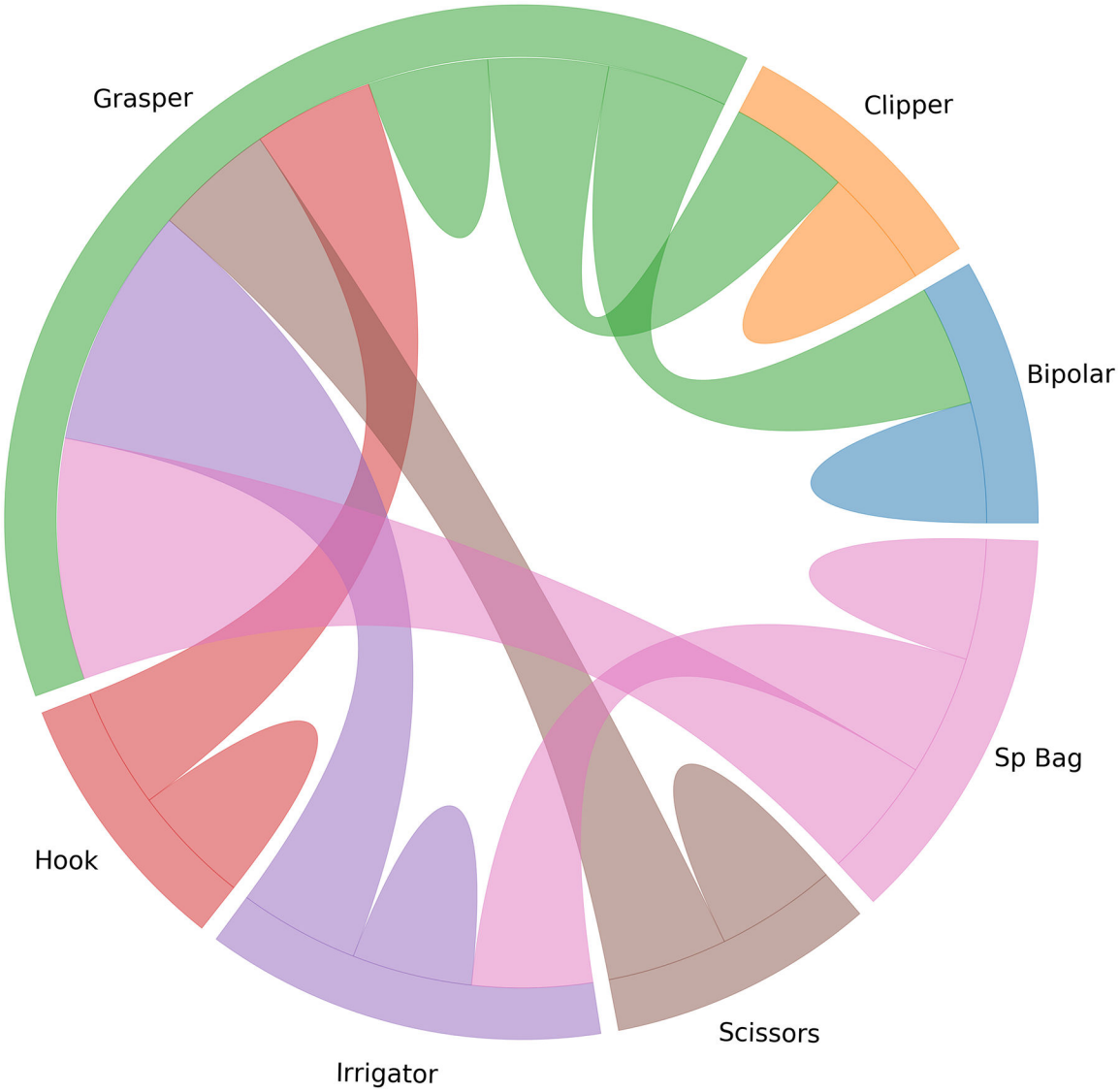
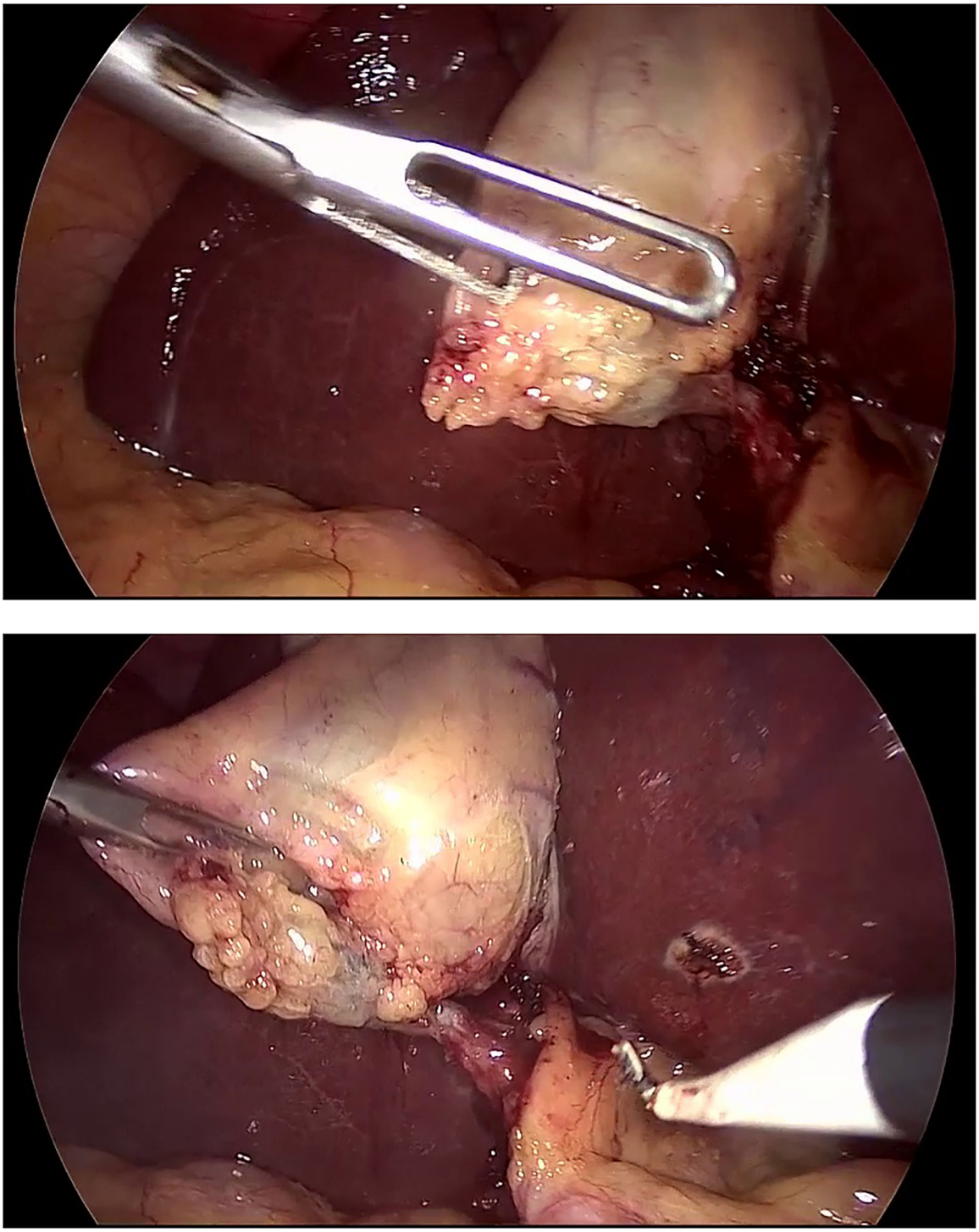
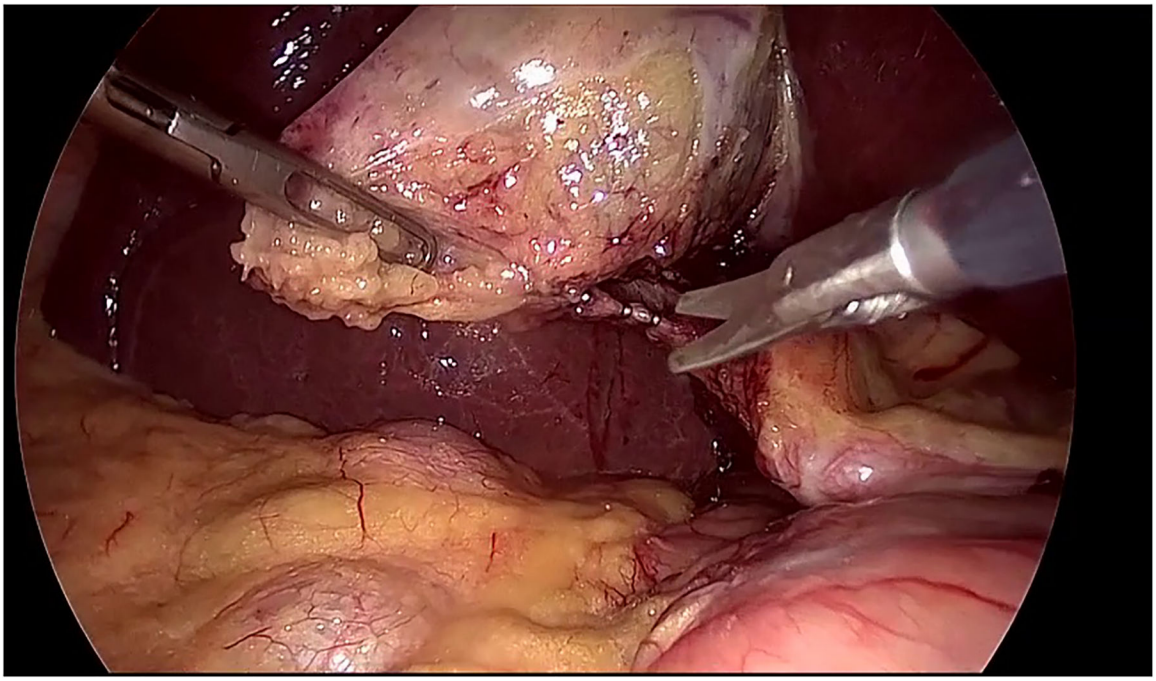
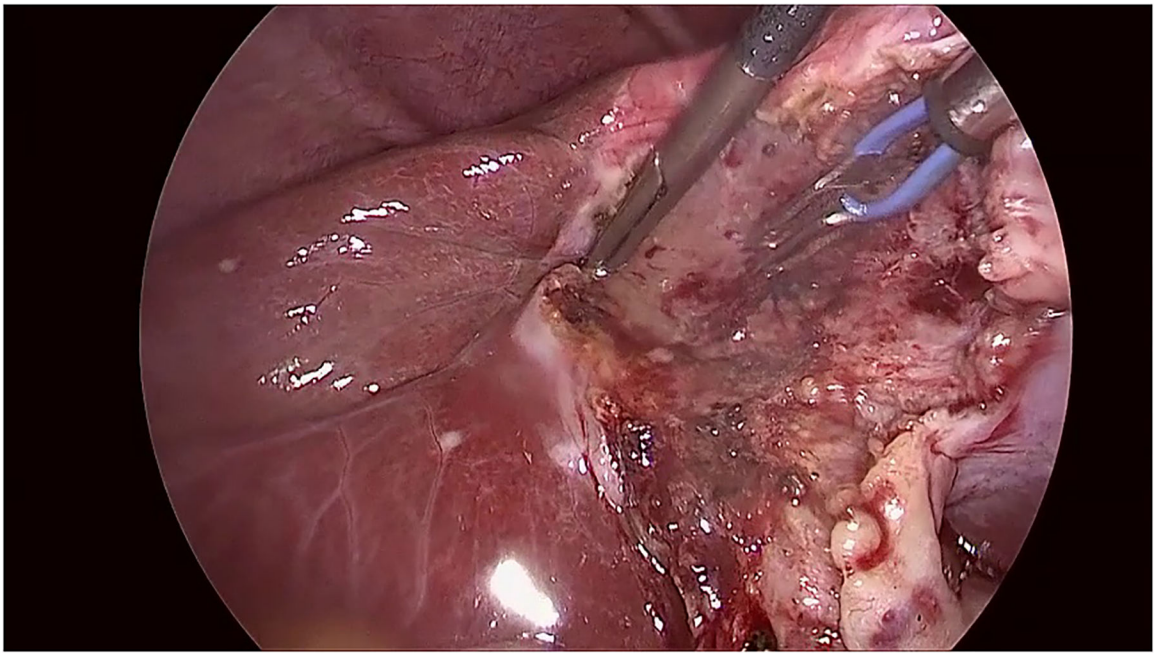
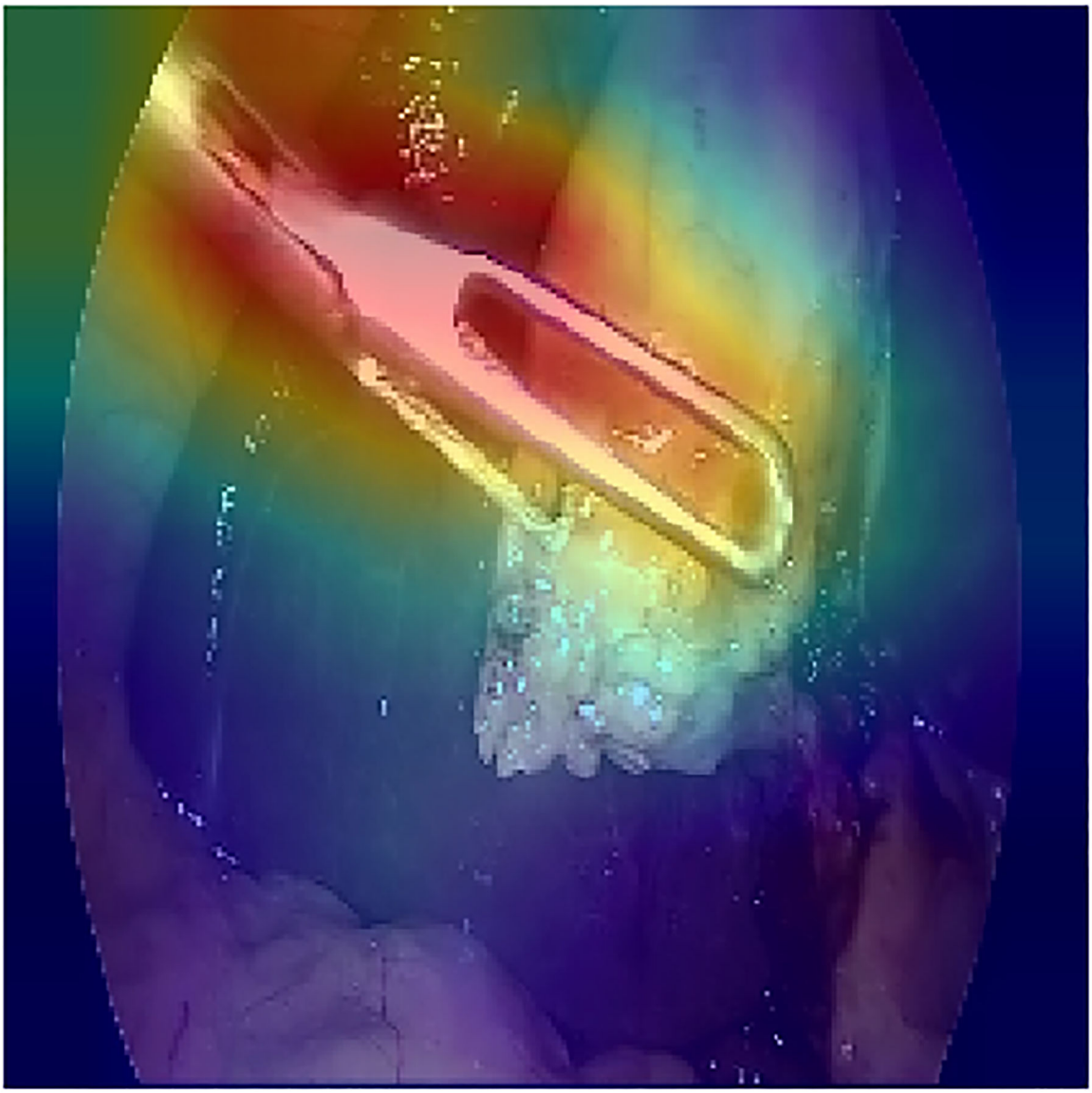


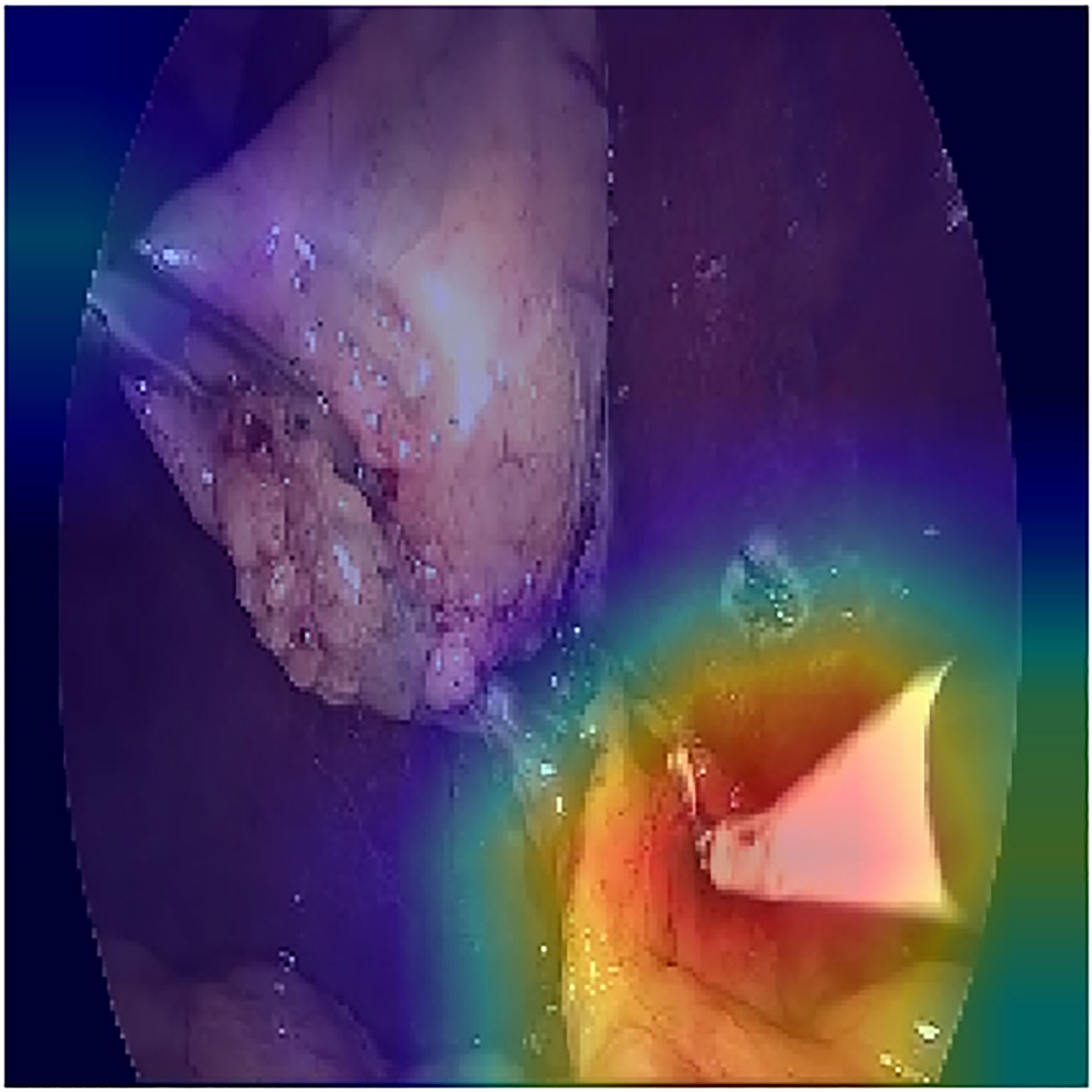
Figure 3.
The chord diagram for the relationship between the tools before and after balancing based on the tools' co-occurrences

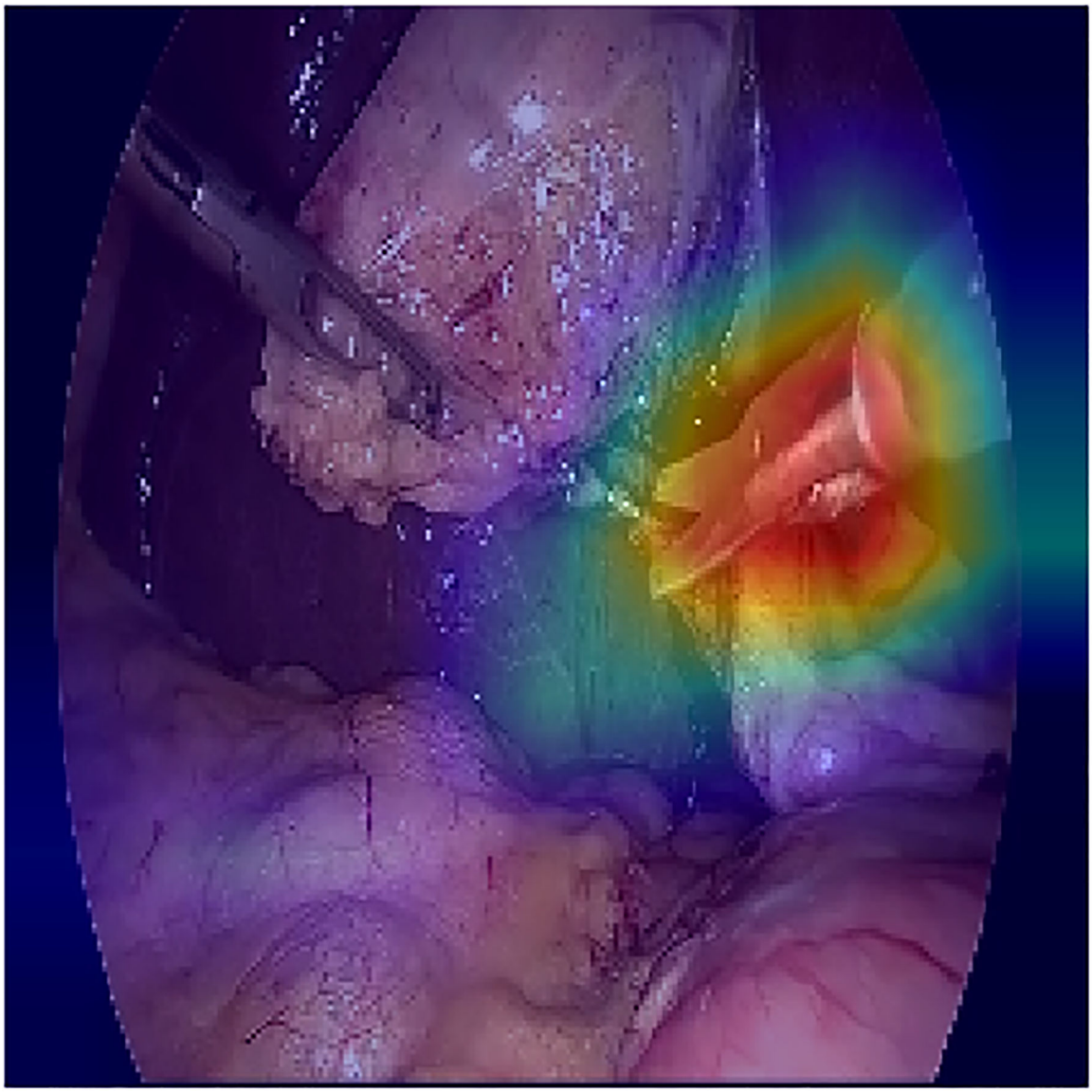












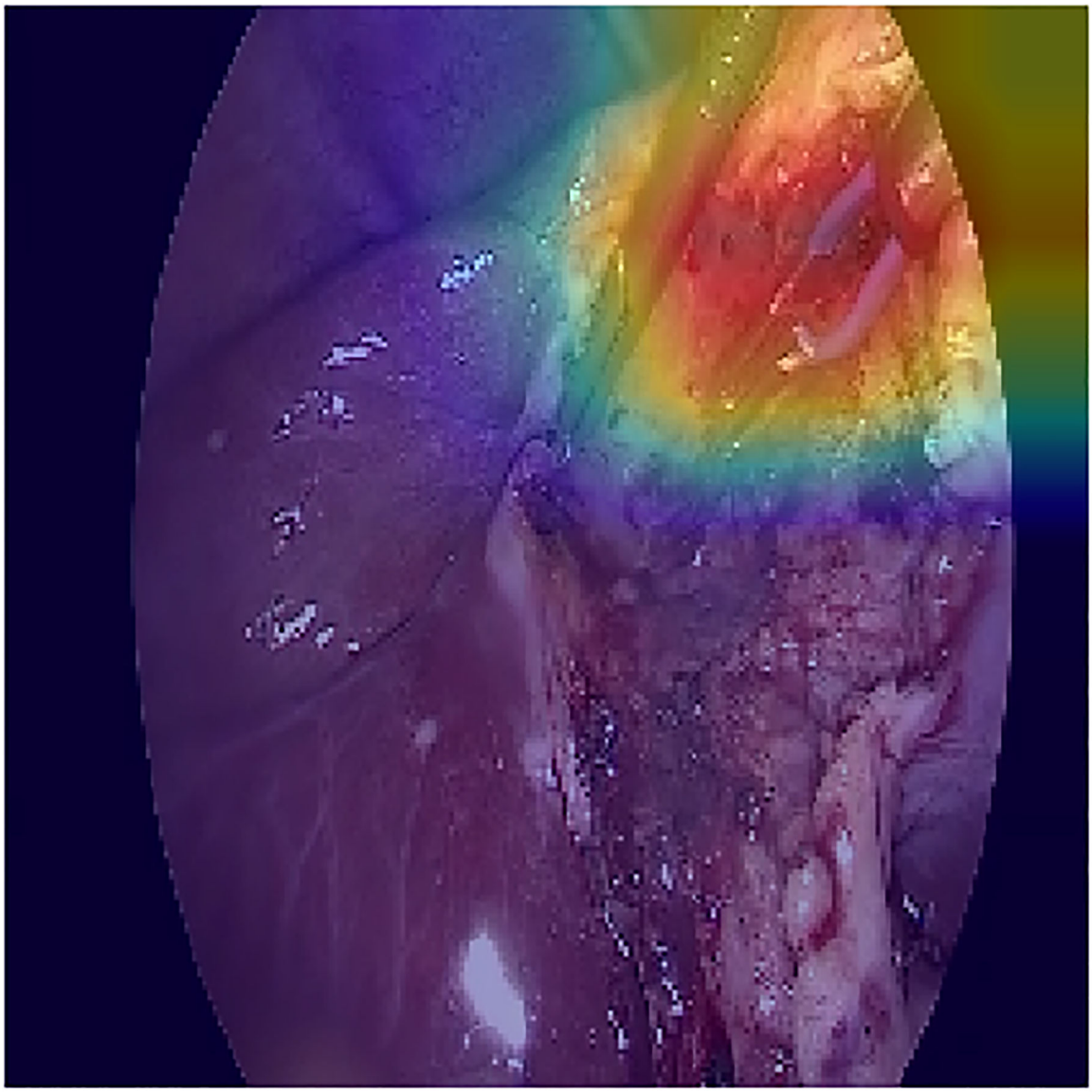


Figure 4.
The visualization of the class activation maps for some examples, based on the prediction of the model

Table 1.

Final results for the proposed model on M2CAI dataset

	ACC(%)	F1-MACRO(%)	F1-MICRO(%)
CNN	74.36	74.43	87.70
CNN-LP	76.31	78.32	88.53
RCNN	77.51	81.95	89.54
RCNN-LP	78.58	84.89	89.79
LAPTOOL-NET(ONLINE)	80.95	88.29	91.24
LAPTOOL-NET(OFFLINE)	81.84	90.53	91.77

Table 2.

The precision, recall and F1 score of each tool for the ML classifier in RCNN-LP after removing the decision model

TOOL	PRECISION (%)	RECALL (%)	F1 (%)
BIPOLAR	77.62	83.57	80.49
CLIPPER	83.22	81.90	82.56
GRASPER	69.99	90.28	78.85
HOOK	95.33	93.43	94.37
IRRIGATOR	77.27	83.60	80.31
SCISSORS	82.91	82.91	82.91
SPECIMEN BAG	76.96	94.91	85.00
MEAN	80.55	87.22	83.50

Table 3.

Comparison of tool presence detection methods on M2CAI

METHOD	CNN	MAP(%)	F1-MACRO(%)
LAPTOOL-NET(OFFLINE)	**Inception-V1**	**-**	**90.53**
LAPTOOL-NET(ONLINE)	**Inception-V1**	**-**	**88.29**
RCNN(OURS)	**Inception-V1**	**89.88**	**81.95**
[22]	Resnet-101 [36]	86.9	-
[24]	VGG	81.8	-
[17]	Alexnet [12]	65	-
[16]	Inception-v3 [37]	63.8	-
[13]	Alexnet	61.5	-
[38]	Alexnet	52.5	-

Table 4.

Final results for the proposed model on Cholec80 dataset

	ACC(%)	F1-MACRO(%)	F1-MICRO(%)
CNN	75.41	83.62	89.05
CNN-LP	76.30	86.16	89.56
RCNN	77.77	88.39	90.41
RCNN-LP	79.95	89.17	91.21
LAPTOOL-NET(ONLINE)	85.77	93.10	93.71
LAPTOOL-NET(OFFLINE)	91.92	96.11	96.40