Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 56-2 (2023) 5626–5631

Surgical Tool Classification & Localisation Using Attention and Multi-feature Fusion Deep Learning Approach

N. A. Jalal *,** **T. Abdulbaki Alshirbaji** *,**
P. D. Docherty ***,* **H. Arabian** * **T. Neumuth** ** **K. Moeller** *

* *Institute of Technical Medicine (ITeM), Furtwangen University, Villingen-Schwenningen, 78054, Germany (e-mails: nour.a.jalal@hs-furtwangen.de, abd@hs-furtwangen.de, arah@hs-furtwangen.de, moe@hs-furtwangen.de).*
** *Innovation Center Computer Assisted Surgery (ICCAS), University of Leipzig, Leipzig, 04103, Germany (e-mail: thomas.neumuth@uni-leipzig.de)*
*** *Department of Mechanical Engineering, University of Canterbury, Christchurch, 8041, New Zealand (e-mail: paul.docherty@canterbury.ac.nz)*

Abstract: Analysing laparoscopic videos, particularly for surgical tool classification and localisation, has attained interest in the field of surgical data science since they represent an extensive information source. However, the difficulties for acquiring and labelling these videos have led to paucity of labelled datasets. Consequently, the progress of developing robust and generalised surgical tool detection models was slowed down, and translating these models into the medical field was hindered. In this work, supervised surgical tool classification and weakly-supervised tool localisation in laparoscopic videos were addressed. A base convolutional neural network (CNN) model was adapted to perform both tasks by incorporating multi-map localisation layers. Squeeze-and-excitation modules were added to the CNN to enhance the ability of the model to generate better focused informative features. Additionally, features at multiple stages of the CNN were combined and fused in a batch normalisation layer to enhance model generalisability. The proposed model was evaluated on the popular Cholec80 dataset. Experimental results of 94.1% mean average precision for tool classification and 70.1% F1-score for tool localisation revealed the ability of the model to learn better features for both tasks. The proposed approach showed the advantages of integrating attentions modules and multi-stage features fusion technique for surgical tool classification and localisation.

Keywords: Context-aware systems, Laparoscopic video, Surgical tool localisation, convolutional neural network (CNN), Weakly-supervised learning.

1. INTRODUCTION

Recent advancements in medical technologies inside the operating room (OR) have greatly improved surgical treatment in terms of patient safety, patient outcome, and surgical support (Maier-Hein et al. (2017)). However, surgical workflow complexity inside the OR has increased which inhibits the communication between the medical teams (e.g., anaesthesiologic and surgical teams) collaborating to perform the surgery. Additionally, the quality of the surgical treatment is still solely based on the clinicians' unique knowledge and experience (Maier-Hein et al. (2022)). Therefore, developing intelligent data-driven medical-support systems has gained momentum over recent years. These systems, known as context-aware systems (CAS), aim at enhancing awareness inside the OR by promoting real-time communication between clinicians. Therefore, possible complications could be potentially avoided. Moreover, CAS will provide medical teams with well-informative knowledge extracted by analysing and fusing available data streams of different perspectives (Jalal et al. (2021a)).

Various data streams have been analysed to establish CAS components, such as anaesthesiology data (Jalal et al. (2021b,a)), sensor-based data (Meißner and Neumuth (2012)), and surgical videos (Twinanda et al. (2017); Abdulbaki Alshirbaji et al. (2021); Jalal et al. (2021c)). In fact, laparoscopic surgical videos represent the dominant data source used in literature, since they represent an easily-acquired, dense dataset that captures the surgical process. Therefore, analysing laparoscopic video has become an increasingly active research area in the medical computer vision community. Recognising surgical activities (Twinanda et al. (2017); Jalal et al. (2019)) and detecting surgical tools (Abdulbaki Alshirbaji et al.

⋆ This work was supported by the German Federal Ministry of Research and Education (BMBF) under CoHMed/DigiMedOP grant no. 13FH5I05IA.

10.1016/j.ifacol.2023.10.473

(2021); Alshirbaji et al. (2020); Jin et al. (2020)) using laparoscopic videos have been investigated. Surgical tool detection and localisation is essential for many applications, such as recognising surgical phases and developing robotic-assisted surgical systems.

Current approaches often employ convolutional neural networks (CNN) to perform surgical tool detection and localisation. For instance, Twinanda et al. introduced a multi-task CNN that jointly performed tool classification and phase recognition (Twinanda et al. (2017)). To alleviate the effect of imbalanced data problem, loss-sensitive and resambling techniques were applied in subsequent studies (Alshirbaji et al. (2018)). In recent approaches, modeling temporal information across video sequences proved to achieve great improvements over spatial models. Recurrent neural networks, such as long short-term memory (LSTM) (Abdulbaki Alshirbaji et al. (2021)), convolutional LSTM (Nwoye et al. (2019)), and graph convolutional networks (GCN) (Wang et al. (2019)) were utilised for temporal modeling. The main drawback of these approaches, especially tool localisation methods, is the need for large and fully labelled datasets, e.g., by tool bounding boxes. Weakly-supervised learning of CNN (WILDCAT) for object localisation that relies only on object binary labels represents a potential solution for the afore-mentioned obstacle. Durand et al. proposed a WILDCAT approach by incorporating a multi-map localisation layer to the base CNN and introducing a new spatial pooling strategy (Durand et al. (2017)). Vardazaryan et al. (2018) and Nwoye et al. (2019) transferred the previous method into the localisation of surgical tools in laparoscopic videos.

In this work, a weakly-supervised CNN for surgical tool classification and localisation in laparoscopic videos was introduced. We built upon the successful weakly-supervised CNN model for object localisation (Durand et al., 2017) and the follow-up approaches (Vardazaryan et al., 2018; Nwoye et al., 2019) for surgical tool localisation, but introduced the following modifications: First, based on our feasibility study on the advantages of attention modules for surgical tool localisation (Jalal et al., 2022), four squeeze-and-excitation (SE) (Hu et al., 2018) modules were added to the CNN architecture; Second, features from lower and top layers of the CNN base model were combined and fused to generate better representation of content of images. Third, loss-sensitive training was implemented to prevent the bias of the CNN towards highly-representative tools in the data.

2. MODEL ARCHITECTURE

The overall model architecture (see Fig. 1) is based on a base CNN model, four squeeze-and-excitation attention modules integrated to the CNN, a multi-stage features combined in a batch normalisation layer, a multi-map convolutional layer generating localisation maps associated with tool classes, and tool-wise spatial pooling layer.

2.1 Base CNN

The CNN model ResNet-50 (He et al., 2016) was utilised in this study because of the high performance of this model for surgical tool classification that exceeded other base CNN models. The ResNet-50 is a residual network composed of five convolutional blocks followed by a global average pooling layer (GAP) and a fully connected layer (FC). The input image size of the ResNet-50 is 224×224×3. The model architecture was modulated to preserve spatial information. Hence, the spatial resolution of the input was increased to 375×300 (instead of 224×224), the last fully-connected and global average pooling layer were removed, and the stride of the convolutional layers in the last two blocks were set to 1×1, similar to Vardazaryan et al. (2018).

2.2 Squeeze-and-Excitation Attention Modules

Squeeze-and-Excitation (SE) attention modules (Hu et al., 2018) aim to enhance the feature representation of the CNN by explicitly modelling the channel relationship. SE attention modules were integrated into the convolutional architecture to boost the model performance. Four SE modules were added after the second, third, fourth, and fifth convolutional blocks. The reduction ratio was set to 16 in all SE modules.

2.3 Multi-stage Feature Fusion

Conventional deep learning approaches use features learnt by the last layers. However, recent studies have shown improvements in the performance of CNN by combining features at multiple stages (Alshirbaji et al., 2022). Therefore, features of intermediate layers in the convolutional model were combined. The output of the second, third and fourth SE modules were concatenated and passed to a batch normalization layer and then to the multi-map convolutional layer (see Fig. 1).

2.4 Multi-map Convolutional Layer

A convolutional layer was added on top of the batch normalisation layer to encode the information learnt at the different stages into tool-related feature maps. The convolutional layer has M×N filters (where N=7 is the number of tool classes, and M=4 is tool-related feature maps) and a kernel size of 3×3. The stride was set to 1×1 to preserve the spatial resolution. The output of this layer is four feature maps for each tools, where these features are learnt by the model with only binary tool labels.

2.5 Tool-wise and Spatial Pooling

Tool-wise pooling was applied to transfer the M feature maps obtained for each tool into one localisation map. Here, max pooling operation was applied across the M maps. The output dimension of this tool-wise pooling is W×H×N, where W and H are the spatial dimension of the tool-related feature maps.

Spatial pooling operation was then implemented to transfer the feature maps of the tools into confidences. Similar to (Vardazaryan et al., 2018), the spatial pooling introduced by (Durand et al., 2017) was applied as in 1. Giving $\mathbf{M}$ the output of the class-wise pooling, $\tilde{\mathbf{M}}_{max}$ the top maximum K_{max} elements of $\mathbf{M}$, and $\tilde{\mathbf{M}}_{min}$ the lowest

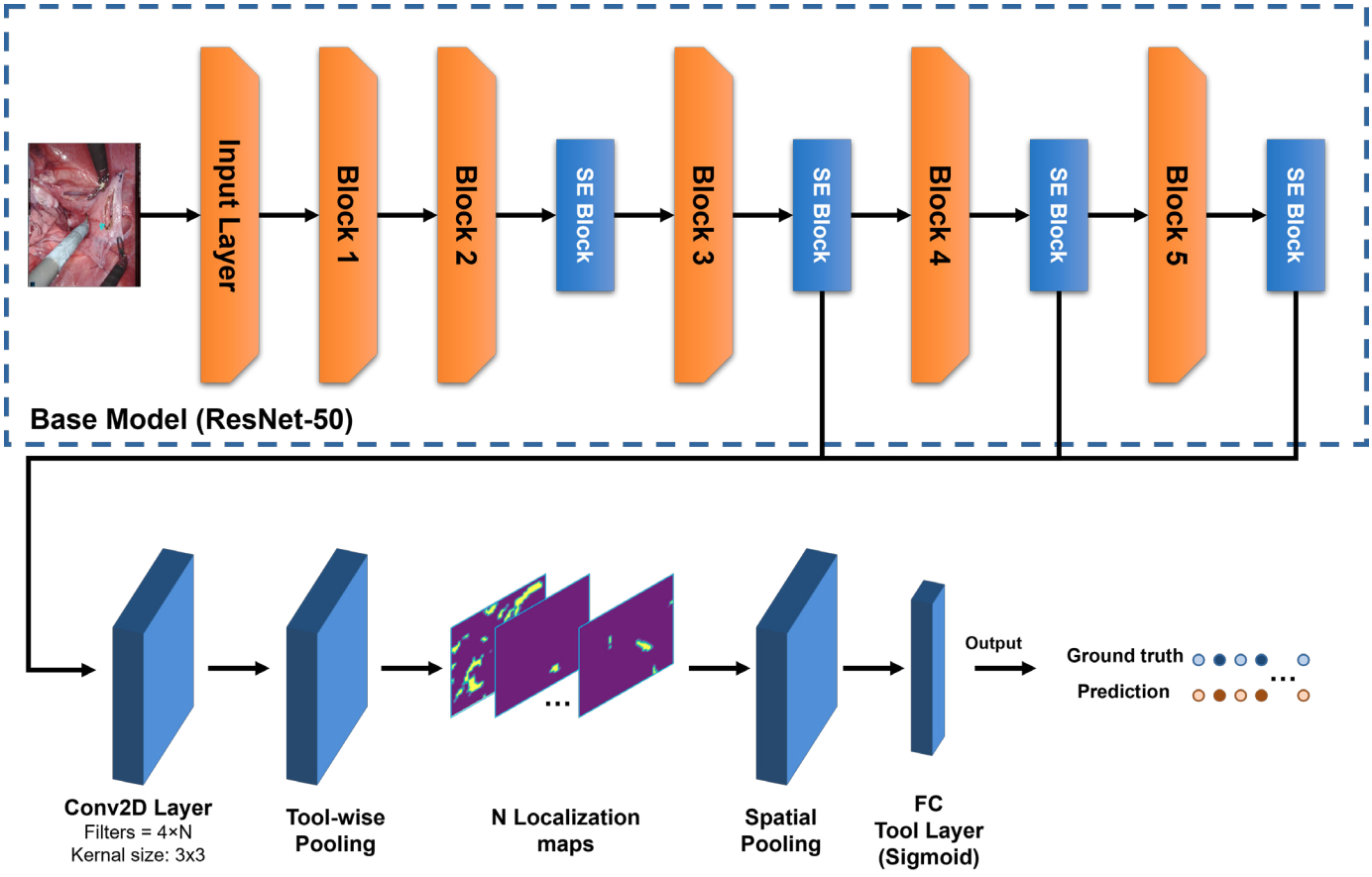

Fig. 1. The proposed architecture for surgical tool classification and localisation. N = 7 is the number of tool classes in the dataset

minimum K_{min} of $\mathbf{M}$, the spatial pooling equation can be described as

$$S = \frac{1}{K_{max}} \sum_{i,j} \tilde{\mathbf{M}}_{max} + \alpha \left(\frac{1}{K_{min}} \sum_{i,j} \tilde{\mathbf{M}}_{min} \right) \quad (1)$$

where K_{max} and K_{min} were chosen to be equal to 50.

3. EXPERIMENTAL SETUP

3.1 Data Description

The publicly available Cholec80 (Twinanda et al., 2017) dataset was utilised for evaluating model performance. Cholec80 consists of 80 cholecystectomy videos labelled at 1 fps with surgical tools and phases. In accordance with previous works, the data of the first 40 videos was used for training, while the last 40 videos were utilised as a test set. The first five videos of the test set were additionally labelled with the bounding boxes of the surgical tools and utilised for evaluating the localisation performance of the proposed model. The labelled bounding boxes surrounded only the characteristic tip of the surgical tools but not the shaft.

3.2 Training Process

The layers of the ResNet-50 model were initialised with ImageNet weights. The weights of the new added layers were randomly initialised. The binary cross-entropy function was used to compute the loss of each tool class, as in Eq. 2. The loss-sensitive technique was employed by multiplying the loss of each tool by a weighting factor that was calculated based on the distribution of each surgical tool in the training set.

$$loss = \frac{-1}{B} \sum_{n=1}^{B} [l_n \log(C_n) + (1 - l_n) \log(1 - C_n)] \quad (2)$$

where *loss* is the computed loss of a specific tool, B is the batch size, l_n is the tool binary label, and C_n is the tool confidence obtained from the spatial pooling operation. Adam optimiser was used with an initial learning rate of 0.01 for the new layers and 10^{-4} for the transferred layers. A batch size of 50 images was chosen, and the training images were shuffled every epoch. The models were run with Keras framework, on an NVIDIA RTX A6000 graphics processing unit (GPU).

3.3 Evaluation Criteria

The average precision (AP) metric was utilised to evaluate the performance of the model for tool presence detection. The AP represents the area under the precision-recall curve. For surgical tool localisation, F1-score was utilised for evaluation and computed as in (3). The predicted bounding box was considered as true positive (T_P) detection if the presence tool confidence was greater than a

specific threshold ($T_C = 0.5$), and the intersection over union (IoU) between the predicted bounding box and ground-truth box was greater than a specific threshold ($T_{IoU} = 0.5$). False positive predictions represented the predicted bounding boxes with confidence score and IoU lower than the T_C and T_{IoU}, respectively. Bounding boxes with confidences lower than T_C, and bounding boxes with confidences greater than T_C but IoU lower than T_{IoU} were considered as false negative detections.

$$\begin{aligned} F1 &= 2\frac{Precision * Recall}{Precision + Recall} \\ Precision &= \frac{T_P}{T_P + F_P} \\ Recall &= \frac{T_P}{T_P + F_N} \end{aligned} \tag{3}$$

where T_P represents the true positive detections, F_P is the false positive detections, and F_N is the false negative detections.

4. RESULTS

To highlight the advantage of incorporating the attention blocks and the multi-stage feature fusion, two approaches were evaluated. The first approach was the ResNet-50 with the multi-map convolutional layer (termed CNN_MMC), while the second approach was the complete proposed pipeline (termed CNN_SE_MF). Table 1 shows the average precision for tool classification task achieved for each tool and the mean average precision (mAP) over the seven tools. F1-score for tool localisation at T_{IoU} are presented in table 2. For qualitative evaluation of the two models' performances, feature class activation maps for the 7 tools are visualised in Fig. 2.

Table 1. Surgical tool presence average precision (AP (%)) on the Cholec80 dataset

Tool	CNN_MMC	CNN_SE_MF
Grasper	81.9	90.6
Bipolar	94.8	95.3
Hook	98.8	99.4
Scissors	82.3	86.1
Clipper	94.9	96.6
Irrigator	88.8	92.8
Specimen Bag	92.9	96.5
mean AP	90.6	94.1

Table 2. Surgical tool localisation F1-Score (%) on the Cholec80-Boxes dataset

Tool	CNN_MMC	CNN_SE_MF
Grasper	56.4	72.5
Bipolar	66.7	74.9
Hook	42.3	60.4
Scissors	37.6	51.8
Clipper	79.8	83.1
Irrigator	62.7	71.5
Specimen Bag	68.6	75.6
mean F1	59.2	70.1

5. DISCUSSION

This study presents a weakly-supervised deep learning approach for surgical tool classification and localisation. The proposed approach is composed of a base CNN model (ResNet-50), multi-stage feature fusion, four SE attention modules, and a multi-map convolutional layer. The proposed approach was evaluated on the Cholec80 dataset.

Table 3. A Comparison of mAP results with the state-of-the-art methods

Approach	mAP
Twinanda et al. (2017)	81.02
Jin et al. (2020)	89.1
Nwoye et al. (2019).	92.9
Abdulbaki Alshirbaji et al. (2021)	94.57
CNN_MMC	90.6
CNN_SE_MF	94.1

Quantitative and qualitative results obtained in this paper show the improvements gained by combining the multi-stage feature fusion and the SE attention modules with the multi-map convolutional layer for tool classification and localisation in laparoscopic images (see tables 1, 2 and Fig. 2). The CNN_SE_MF approach yielded mAP of 94.1% for surgical tool classification. This value enhanced on the CNN_MMC model mAP value of 90.6%. Similarly, the incorporation of the multi-stage feature fusion and the SE attention modules allowed the CNN_SE_MF approach to exceed the CNN_MMC model by more than 10% F1-Score for surgical tool localisation.

The achieved tool classification performance exceeded the leading methods except ResNet-LC-LV (Abdulbaki Alshirbaji et al., 2021) (see table 3). However, temporal information across the video sequence was modelled using an LSTM network in the ResNet-LC-LV method, while only spatial information was considered in the CNN_SE_MF approach. Therefore, further improvement can be potentially added to the CNN_SE_MF approach by modelling temporal dependencies.

The obtained localisation performance of the CNN_SE_MF approach showed a good performance with mean F1-Score of 70.1% considering that only binary tool labels were used for training the model. Fig. 2 highlights the improvement acquired after adding the attention modules and the multi-feature fusion to the CNN_MMC model. As can be seen from Fig. 2, feature maps for all tools were refined, and the model was better able to focus on tool-related information in the image. This, in turn, led to better localisation performance, where the IoU was enhanced by a large margin over the CNN_MMC model. It is worth mentioning that the CNN_SE_MF approach was able to detect multiple instances of same tool, even though all instances shared the same localisation map. This was done through a post-processing step.

The attention, multi-feature fusion model described in this study performed well on the Cholec80 dataset for tool classification and localisation. Nevertheless, this study had some limitations. The results was obtained for a single data split for training and testing. Therefore, Monte-Carlo cross validation should be considered to generate more statistically relevant results. Temporal dependencies between laparoscopic video frames was not considered. It is possible that future work could adapt the current models to learn temporal information and improve the results. Finally, only one dataset of one surgical procedure type

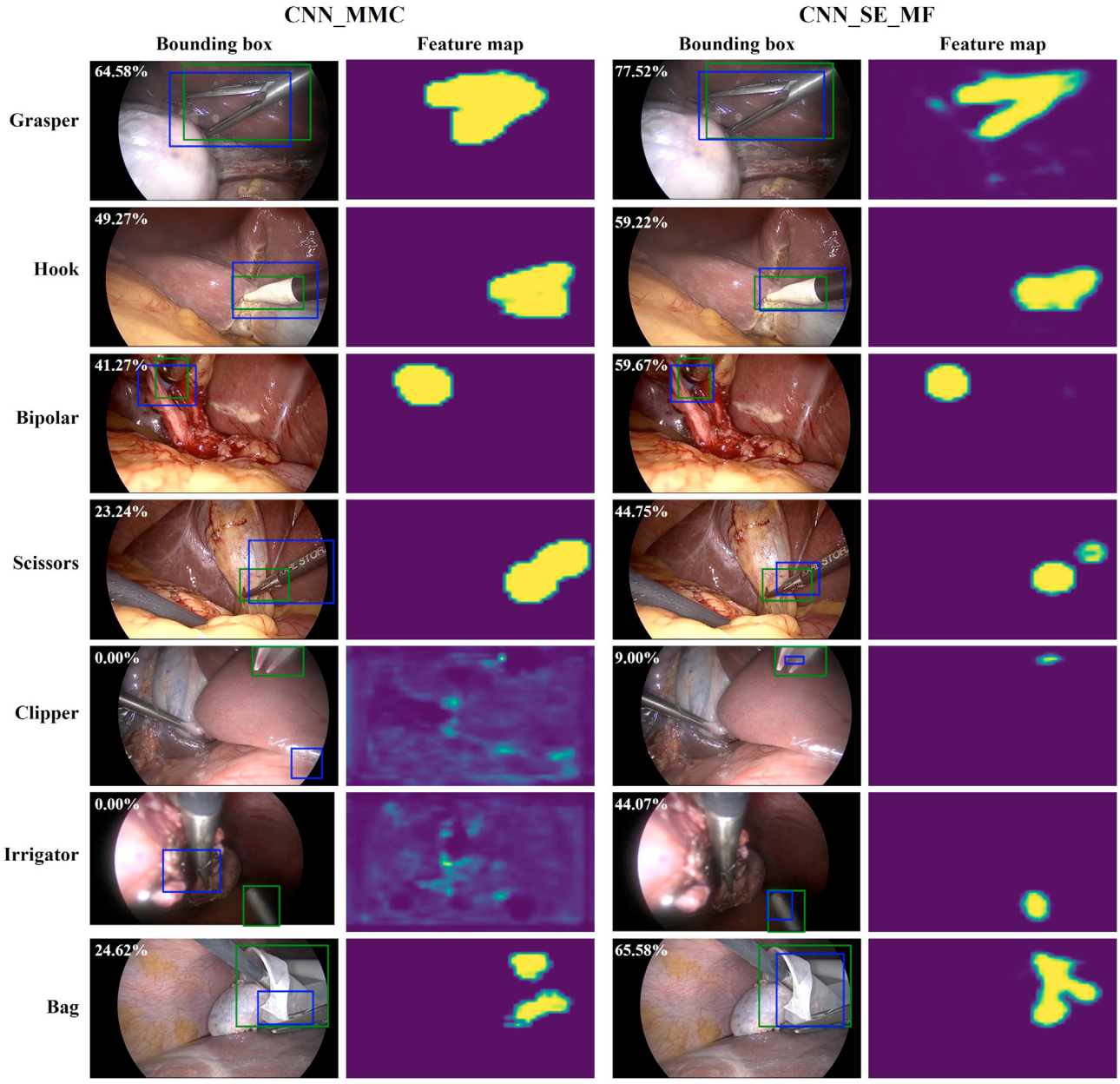

Fig. 2. Visualisation of feature localisation maps of the CNN_MMC and the CNN_SE_MF approach. Green and blue boxes represent the ground-truth and the predicted bounding boxes, respectively. IoU scores are presented on the top-left corner of images.

was utilised for evaluation. To evaluate the robustness and generalisation capability of the proposed approach, an extensive evaluation on other datasets is still required.

6. CONCLUSION

This study proposed a weakly-supervised deep learning approach dedicated to perform surgical tool classification and localisation by using only tool binary labels. The proposed approach relies on attention modules, multi-feature fusion, and multi-map localisation layer. Experimental results showed that this preliminary implementation of the method performed well among the stat-of-the-art methods for tool classification. Moreover, obtained tool localisation results showed that this approach is very promising for developing weakly-supervised surgical tool localisation systems.

REFERENCES

Abdulbaki Alshirbaji, T., Jalal, N.A., Docherty, P.D., Neumuth, T., and Möller, K. (2021). A deep learning spatial-temporal framework for detecting surgical tools in laparoscopic videos. *Biomedical Signal Processing and Control*, 68, 102801. doi:10.1016/j.bspc.2021.102801.

Alshirbaji, T.A., Jalal, N.A., Docherty, P.D., Neumuth, P.T., and Möller, K. (2022). Improving the Generalisability of Deep CNNs by Combining Multi-stage Features for Surgical Tool Classification. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 533–536. IEEE.

Alshirbaji, T.A., Jalal, N.A., and Möller, K. (2018). Surgical Tool Classification in Laparoscopic Videos Using Convolutional Neural Network. *Current Directions in Biomedical Engineering*, 4(1), 407–410. doi: 10.1515/cdbme-2018-0097.

Alshirbaji, T.A., Jalal, N.A., and Möller, K. (2020). A convolutional neural network with a two-stage LSTM model for tool presence detection in laparoscopic videos. *Current Directions in Biomedical Engineering*, 6(1).

Durand, T., Mordan, T., Thome, N., and Cord, M. (2017). Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 642–651.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Jalal, N.A., Alshirbaji, T.A., Laufer, B., Docherty, P.D., Russo, S.G., Neumuth, T., and Möller, K. (2021a). Effects of Intra-Abdominal Pressure on Lung Mechanics during Laparoscopic Gynaecology. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2091–2094. IEEE.

Jalal, N.A., Alshirbaji, T.A., Docherty, P.D., Neumuth, T., Laufer, B., and Moeller, K. (2021b). Changes of Physiological parameters of the patient during laparoscopic gynaecology. *Current Directions in Biomedical Engineering*, 7(2), 500–503.

Jalal, N.A., Alshirbaji, T.A., Docherty, P.D., Neumuth, T., and Moeller, K. (2021c). A Deep Learning Framework for Recognising Surgical Phases in Laparoscopic Videos. *IFAC-PapersOnLine*, 54(15), 334–339. doi: 10.1016/j.ifacol.2021.10.278.

Jalal, N.A., Alshirbaji, T.A., and Möller, K. (2019). Predicting Surgical Phases using CNN-NARX Neural Network. *Current Directions in Biomedical Engineering*, 5(1), 405–407. doi:10.1515/cdbme-2019-0102.

Jalal, N.A., Arabian, H., Alshirbaji, T.A., Docherty, P.D., Neumuth, T., and Moeller, K. (2022). Analysing attention convolutional neural network for surgical tool localisation: A feasibility study. *Current Directions in Biomedical Engineering*, 8(2), 548–551.

Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C.W., and Heng, P.A. (2020). Multi-task recurrent convolutional network with correlation loss for surgical video analysis. *Medical Image Analysis*, 59, 101572. doi: 10.1016/j.media.2019.101572.

Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., and Mascagni, P. (2022). Surgical data science–from concepts toward clinical translation. *Medical image analysis*, 76, 102306.

Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., Hashizume, M., Katic, D., Kenngott, H., Kranzfelder, M., Malpani, A., März, K., Neumuth, T., Padoy, N., Pugh, C., Schoch, N., Stoyanov, D., Taylor, R., Wagner, M., Hager, G.D., and Jannin, P. (2017). Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9), 691–696. doi:10.1038/s41551-017-0132-7.

Meißner, C. and Neumuth, T. (2012). RFID-based surgical instrument detection using Hidden Markov models. *Biomedical Engineering/Biomedizinische Technik*, 57, 689–692.

Nwoye, C.I., Mutter, D., Marescaux, J., and Padoy, N. (2019). Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos. *International journal of computer assisted radiology and surgery*, 14(6), 1059–1067.

Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., and Padoy, N. (2017). EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *IEEE Transactions on Medical Imaging*, 36(1), 86–97. doi:10.1109/TMI.2016.2593957.

Vardazaryan, A., Mutter, D., Marescaux, J., and Padoy, N. (2018). 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*, 169–179. Springer.

Wang, S., Xu, Z., Yan, C., and Huang, J. (2019). Graph convolutional nets for tool presence detection in surgical videos. In *International Conference on Information Processing in Medical Imaging*, 467–478. Springer.