2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Microsurgical Tool Detection and Characterization in Intra-operative Neurosurgical Videos

Ajay Ramesh[1], Manish Beniwal[2], Alok Mohan Uppar[2], Vikas V[2], and Madhav Rao[1]

Abstract— **Brain surgery is complex and has evolved as a separate surgical specialty. Surgical procedures on the brain are performed using dedicated micro-instruments which are designed specifically for the requirements of operating with finesse in a confined space. The usage of these microsurgical tools in an operating environment defines the surgical skill of a surgeon. Video recordings of micro-surgical procedures are a rich source of information to develop automated surgical assessment tools that can offer continuous feedback for surgeons to improve their skills, effectively increase the outcome of the surgery, and make a positive impact on their patients. This work presents a novel deep learning system based on the Yolov5 algorithm to automatically detect, localize and characterize microsurgical tools from recorded intra-operative neurosurgical videos. The tool detection achieves a high 93.2%** ***mean average precision*****. The detected tools are then characterized by their on-off time, motion trajectory and usage time. Tool characterization from neurosurgical videos offers useful insight into the surgical methods employed by a surgeon and can aid in their improvement. Additionally, a new dataset of annotated neurosurgical videos is used to develop the robust model and is made available for the research community.**

Clinical relevance— **Tool detection and characterization in neurosurgery has several online and offline applications including skill assessment and outcome of the surgery. The development of automated tool characterization systems for intra-operative neurosurgery is expected to not only improve the surgical skills of the surgeon, but also leverage in training the neurosurgical workforce. Additionally, dedicated neurosurgical video based datasets will, in general, aid the research community to explore more automation in this field.**

I. INTRODUCTION

Human beings tend to make errors in their daily activities that are widely accepted provided that the errors do not have strong ramifications. The same notion applies even to healthcare and surgical fields, where doctors or surgeons offer their services in the form of consultation or surgical procedures based on their expertise and skill. Technological advances have played an important role in evaluating and assessing professionals in different fields [1], [2], [3], [4] as well as in preventing errors [5]. However, concerning surgical procedures, poor performance is often attributed to inadequate training and feedback [6], [7]. Thus, continuous positive improvements can be achieved by providing individualized objective feedback regarding a surgeon's skill [8]. A rich source of information regarding surgical procedures comes from recorded videos of surgeries. These recordings can capture the characteristics of a surgeon's operating skill such as the different kinds of tools that the surgeon employs, the tool holding time, tool operating frequency, number of operating hands, the types and precision of cuts, and various other traits. The videos also contain patient-specific information which is envisioned to be useful in standardizing and automating the overall surgical process for similarly diagnosed problems. The tool-specific parameters are known to be effective in assessing surgical skill [9], [10]. However, manual assessment of skill from videos is time consuming and not feasible. It is hence necessary and important to develop systems to automatically analyze surgical skills from videos.

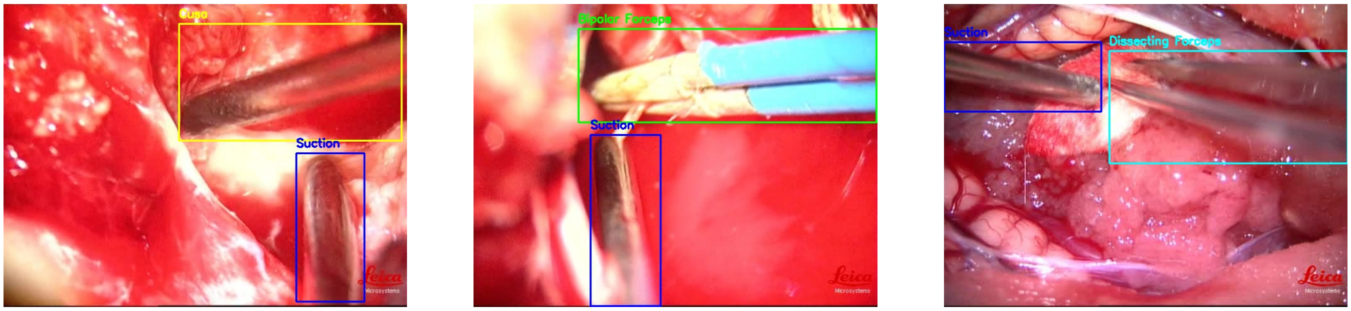
The first and foremost step, however, is to accurately characterize the surgical tools that are used throughout the procedure. This involves detection and localization of tools followed by the characterization of their usage. Several works exist to detect surgical tools. Early works include methods that use markers such as LEDs [11] and RFIDs [12], which raise concerns on the safety of the surgical procedure and the convenience of their usage. Moreover, they do not provide the flexibility and scope that image-based approaches offer. Kumar *et al.* [13] proposed to model surgical instruments using histogram of oriented gradients (HOG) features derived from the computer vision domain and the Lagrangian support vector machine (LSVM) classifier. Sznitman *et al.* [14] built upon the deformable detector proposed by [15] and integrated it with a gradient-based tracker to detect retinal microsurgical tools. Allan *et al.* [16] proposed a pixel-wise detection method using color, scale-invariant feature technique (SIFT), and HOG features on laparoscopic surgical sources. Since the rise of deep learning techniques, particularly convolutional neural networks (CNN) for computer vision tasks, research on surgical tool detection and localization has leveraged the immense advantages that deep learning provides. Choi *el al.* [17] proposed a detection model using a CNN and th YOLO algorithm for laproscopic robot-assisted surgeries. The introduction of the M2CAI dataset [18] served as an additional motivation. The authors proposed a multi-task network architecture that showed promising results on the dataset. The authors also remark that significant improvement is possible if more data is collected [19]. Kanakatte *et al.* [20] use a spatio-temporal deep network to segment and localize tools in laparoscopic cholecystectomy surgeries.

In general, tool localization and classification in laparoscopic and endoscopic surgical videos have played an important role in realizing automated tool detection. However, neurosurgical videos present different kinds of challenges

[1] Surgical and Assistive Robotics Lab, IIIT-Bangalore, Bangalore-560100, India. (e-mail: ajayramesh.ranganathan@iiitb.org)

[2] Department of Neurosurgery, NIMHANS, Bangalore-560029, India. (e-mail: vikas.drv@gmail.com)

978-1-7281-1178-0/21/$31.00 ©2021 IEEE



(a) Suction and Cusa (b) Suction and Bipolar Forceps (c) Suction and Dissecting Forceps

Fig. 1: Sample annotated frames from the dataset

such as a low field of view, a small operating region, low light conditions to view the operating space and the distinct nature of the tools used. Additionally, neurosurgery is a specialized field wherein the operating procedures differ from other studied surgeries and have relatively fewer practitioners in the world [21]. Training the neurosurgical workforce through automated systems offers significant scope in meeting the current demands. Hence, special attention needs to be invested for automating the assessment of neurosurgical operating skill to help in training and individualized feedback-based improvement. However, very few works exist that pertain to the detection and characterization of tools in neurosurgery.

This paper presents a neurosurgical tool detection and characterization system built using the latest YOLOv5 object detection algorithm [22], [23]. The localization and characterization of neurosurgical tools using deep learning and the object detection approach is a novel area of work. Also, a new and much-needed neurosurgical tool video dataset is introduced. The dataset consists of four major neurosurgical tools—Suction, Cusa, Bipolar Forceps, and Dissecting Forceps. This comprehensive dataset was used to train and test the tool detection model. The proposed detection model achieved a high 93.2% *mean average precision* on the test dataset of images. Extending further, frame-level tool detection was performed on the test videos. Bounding box matching and interpolation techniques [24] were allied to significantly boost detection performance. Post detection, the tools were characterized in three ways that are known to effectively indicate surgical skill—On-off heatmaps that show tool activity and no-activity, total tool usage time, and tool motion trajectory.

II. DATASET

Few publicly available datasets of surgical tools have been used for tool presence detection in the past. These include the m2cai16-tool dataset [18] which was released as part of the M2CAI Tool Presence Detection Challenge, and the Cholec80 dataset [18]. The datasets contain videos of cholecystectomy and laparoscopic surgeries and are labeled with binary annotations to indicate tool presence. A dataset of retina laparoscopic videos available at [25], contains only one tool for the retina procedures and 1000 images from a single video containing two tools. A dataset of minimally invasive surgery [16] consisting of approximately 100 images taken from 6 videos has been used to detect and localize instruments. A neurosurgical tools dataset which contains 2476 frames was introduced and evaluated for tool detection and tool region segmentation [26]. Many of the above-mentioned datasets lack the quantity and diversity that is required for building robust automated tool detection systems. Moreover, very few datasets exist for neurosurgical instruments in particular. This served as a motivation to acquire and share our dataset publicly.

TABLE I: Number of train and test instances of each tool. The bounding box size is computed with respect to the image size

Number of instances				Bounding box size		
Tool	**Train**	**Test**	**Total**	**avg**	**min**	**max**
Suction	3807	949	4756	0.136	0.634	0.001
Cusa	2307	408	2715	0.105	0.593	0.005
Bipolar. F	56	184	240	0.239	0.545	0.019
Dissecting. F	464	111	575	0.188	0.634	0.0221
All	6634	1652	8286	0.132	0.634	0.001

Our dataset consists of 5641 annotated frames, at a resolution of 640x480, extracted at 1 frame per second (FPS) from 32 neurosurgical videos. The data was collected in accordance with the Helsinki Declaration of 1975, as revised in 2000. Every tool in an image is annotated by a bounding box and the tool category. The annotation was performed under the supervision of an experienced neurosurgeon. Presently, the dataset consists of four major tools used in neurosurgery—*Suction, Cusa, Bipolar Forceps, and Dissecting Forceps*. The dataset was split into 22 videos for training and 10 videos for testing. The split was made while ensuring independence between the training and testing videos. The distribution of annotated instances of each tool in the dataset for the training and testing phases is shown in Table I. Sample images from the dataset are shown in the Figure 1.

III. APPROACH

A. Neurosurgical Tool Detection

The localization of surgical tools is an object detection task, i.e., spatially detect the presence of every tool in an image as well as identify the kind of tool being used. Deep learning-based techniques have proved to be very successful for the detection of objects in images as well as videos. The R-CNN [27], Fast-RCNN [28], Faster-RCNN [29], YOLO [22], and SSD [30] are examples of well known object detection networks. The YOLO object detection algorithm was chosen for this work, primarily due to its prominent detection accuracy, better inference speed, and less training time over other algorithms.

1) The Object Detection Network: The YOLOv5 implementation by Ultralytics [23] was used to build a YOLOv5 network and fine-tune the same for the surgical tool detection application. Inherently, the YOLOv5 network is built on top of the YOLOv4 network which incorporates a "Bag of Freebies" [31] and "Bag of Specials" that can substantially improve the robustness of detection [32]. The network uses the YOLOv5 CNN-based backbone to extract visual features that are fed to the YOLO detection layers. Extensive data augmentation is applied during the training phase to prevent over-fitting and improve generalization [33]. Random scaling, rotations, x-flips, and y-flips were applied to incorporate variations in tool pose, tool orientation and microscope focus. Photometric distortions were applied by adjustments in hue, saturation, and brightness. The other techniques used were mix-up [34] in which objects were cropped out and pasted in random backgrounds and mosaic augmentation [32] that mixes four training images into one. The mosaic augmentation generates a mixed form of different contexts and enables the model to detect objects out of their normal setting.

Transfer learning was applied to train the network. Transfer learning has shown to improve localization results on surgical tool detection tasks as stated in [35]. Hence, the weights of the network pre-trained on the COCO dataset [36] were initialized for the training process. The network was trained and fine-tuned for 150 epochs. The Stochastic Gradient Descent with momentum and warm restarts algorithm [37] was used as the optimizer with a cosine annealing scheduler [38] to decay the learning rate.

2) Bounding Box Matching and Interpolation: Frame-wise tool detection in videos is prone to errors caused by motion blur, occlusions, etc. leading to missed or false detection. Thus, a post-processing method is required that improves the detection performance of frame-wise tool detection. The Tubulet-level Bbox linking method proposed by Belhassen *et al.* [24] was used to design a robust post-processing detection step. The method involves minimal distance box matching across video frames to form tubulets and a sequential matching of tubulets from the start to the end frame within a specific time window. This helps to infer missed detections and correct false detections in intermediate frames. The time window was configured to $1s$, keeping in mind that neurosurgeons switch tools in approximately $2s$ on average. A bounding box at $t = i$ is defined by the Equation 1, where x_i, y_i are the coordinates of the center of the bounding box and w_i, h_i are its width and height respectively. The bounding boxes for the newly inferred detections at a given time t_j are estimated by linear interpolation [39] as described by Equation 2.

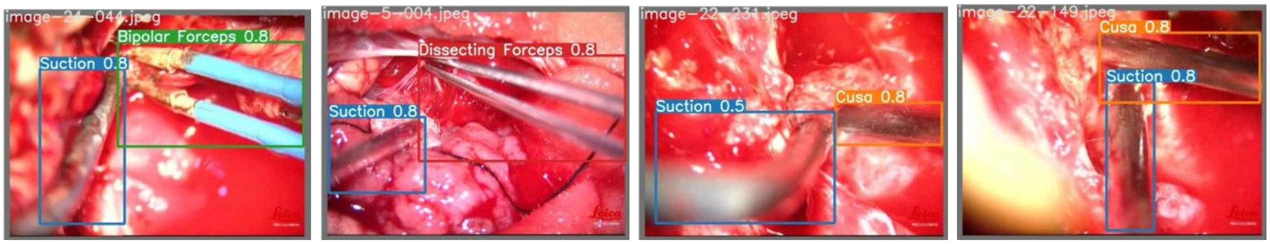
$$b_i = [x_i, y_i, w_i, h_i] \tag{1}$$

$$b_j = \frac{t_{i+1} - t_j}{t_{i+1-t_i}} b_i + \frac{t_j - t_i}{t_{i+1} - t_i} b_{i+1} \tag{2}$$

B. Tool Characterization

The characterization of tools was conducted in the following manner -

- Tool on-off time: The on-off time heatmap for each tool shows the different tools that were used in various parts of the surgery. It also represents the frequency and duration of a tool or a combination of tools that were used during the procedure. Such statistics are useful not only for understanding the different phases of a surgery [9], but also for detecting unwarranted errors made by a surgeon (e.g. using incorrect tools). The on-off heatmap also offers quantification of no-activity periods and its frequency in a surgical procedure. This can be used to directly characterize a surgeon's skill given the kind of surgical procedure being performed. For instance, a less skilled surgeon is likely to have a greater frequency of no-activity during a surgery [40].
- Tool usage time: The total usage time of each tool in the surgical procedure is also an indicator of skill since the usage time varies as a surgeon acquires knowledge of the tool's handling and orientation. For instance, as



(a) (b) (c) (d)

Fig. 2: Tool detection results showing the confidence of detection alongside the tool label.

learnt from [40], the Cusa's tool usage was observed to increase with experience for tumor decompression procedures whereas, the Suction's usage was observed to decrease.

- Tool motion trajectory: The centroids of the detected tools were tracked throughout the video segments that had a constant surgical field of view acquired from a stationary microscope. Tracking the motion of tools can aid in differentiating an expert surgeon from a novice in terms of dexterity [10].

IV. RESULTS AND OBSERVATIONS

A. *Tool Detection*

The detection model was evaluated by performing inference on the test image-dataset that was mentioned in Table I. Few sample frames of the detection results are shown in Figure 2. The model was able to detect tools with high confidence scores, and also detect tools in different orientations. For example, the Suction tool appearing in different orientations is successfully detected by the model as shown in Figures 2a, 2b, 2c and 2d. In addition, the developed detection model is insensitive to slight motion blur caused by the movement of the tool by the surgeon. Overall the developed model generalizes adequately and detects tools with varying orientations, scales, blurriness and tools that are partially visible in the frame.

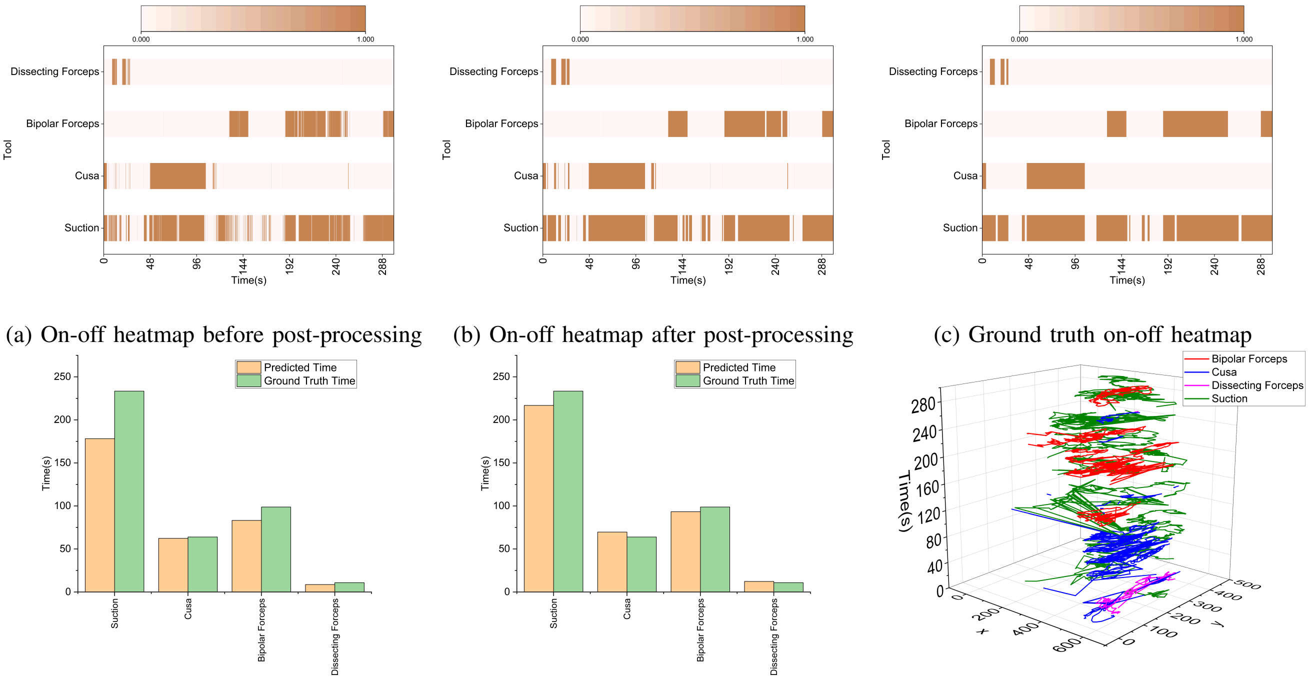
The tool detector is evaluated using the mean average precision (*map*) metric. A prediction that has an Intersection over Union (IoU) greater than 0.5 with the ground truth is considered a correct detection. The performance evaluation for each tool is shown in Table II. The model detects Suction and Cusa tool the best, followed by the Bipolar and Dissecting forceps. The *map* of the Dissecting and Bipolar Forceps were expected to be lower compared to the Suction and the Cusa. This is attributed to lesser training data available for these tools when compared to the other two since Dissecting and Bipolar Forceps are less frequently used in surgeries.

TABLE II: Results of tool detection on image test dataset

Tool	**Precision**	**Recall**	**map@0.5IoU**
Suction	0.896	0.93	0.96
Cusa	0.777	0.966	0.958
Bipolar Forceps	0.877	0.918	0.931
Dissecting Forceps	0.426	0.928	0.883
Total	0.744	0.936	0.932

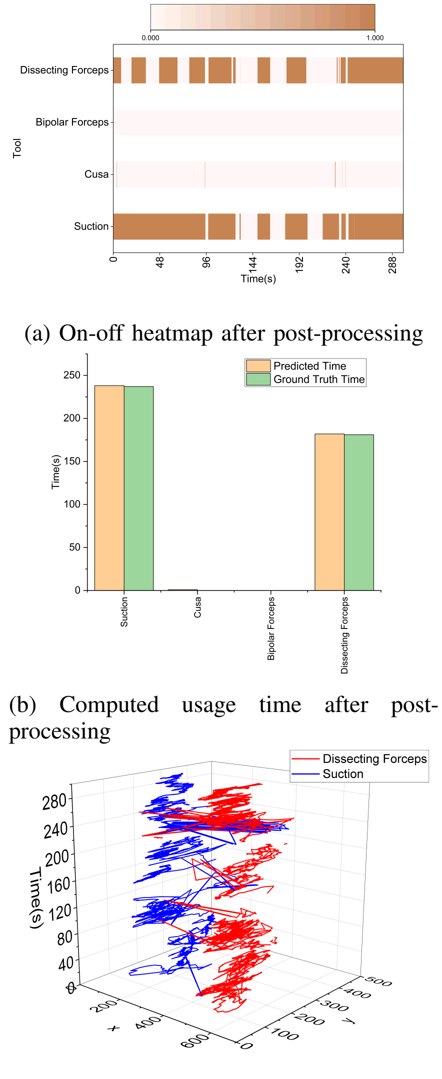
B. *Frame-Wise Video Tool Detection and Characterization*

Tool detection was performed on the test videos at $25fps$. Bounding box matching and interpolation across frames was performed as mentioned in Section III-A.2, and each tool was characterized as discussed in the Section III-B. Figure 3 presents the results on a test video during which the surgeon used all four tools in a span of 300 seconds. It was observed that the post-processing interpolation technique significantly improves the detection of tools throughout the video, as shown in the on-off heatmap in Figures 3a, and 3b, when compared to the ground truth characterization that is shown in Figure 3c. The dark regions in the heatmap



(a) On-off heatmap before post-processing
(b) On-off heatmap after post-processing
(c) Ground truth on-off heatmap
(d) Usage time before post-processing
(e) Usage time after post-processing
(f) Tool trajectories

Fig. 3: Tool characterization on a test video



(a) On-off heatmap after post-processing

(b) Computed usage time after post-processing

(c) Tool trajectories

Fig. 4: Tool characterization on a test video

represent tool usage. The characterization post interpolation is more accurate. Notice that the intermediate frames which are undetected in Figure 3a are successfully detected after interpolation in Figure 3b. Similar improvement is also observed in the tool usage time, where the predicted usage time for each tool is much more accurate post applying the interpolation technique as shown in Figure 3e.

Figure 3f shows the motion trajectory of each tool which is generated by tracking the centroid of the predicted bounding box. The X and Y axis are in pixel coordinates, and the Z axis represents time. The motion trajectory, as expected, complies with the heat-map in Figure 3b. The surgeon first uses the Suction and Cusa for a short duration followed by the Suction and Dissecting Forceps. Then, the surgeon uses the Bipolar Forceps and the Suction. The motion variations that are visible in the trajectory are an important indicator of the surgeon's dexterity; experienced surgeons are known to execute more focused movements, leading to better motion economy [8]. Another test video involving the usage of only two tools—Suction and Dissecting Forceps is characterized by tool usage, tool on-off time and tool trajectories as shown in the Figure 4. The predicted tool usage times for Suction and Dissecting forceps are in close agreement to the ground truth values as shown in Figure 4b. Figure 4a, Figure 4b and Figure 4c are the outcomes of characterizing the tools used by the surgeon.

CONCLUSION

A novel neurosurgical tool detection and characterization system based on the YOLOv5 algorithm was developed for detecting and localizing four primary microsurgical tools commonly employed in neurosurgery. The model was trained on an original custom dataset of intra-operative neurosurgical videos. The dataset offered a rich source of tool information which was effectively used to characterize the surgeon's tool usage. The characterization was based on three parameters: Tool on-off time, tool usage time and tool trajectory. Various data augmentation strategies, transfer learning, and tubulet-level bounding box linking methods were incorporated to design a robust detection and localization model. The model showcased a 93.2% *map* for all the four tools used, with a high accuracy reported for the Suction and Cusa tools when compared to the Dissection and Bipolar Forceps. Additionally, tools in different orientations, scales, with slight motion blur, and that are partially visible in frames were also detected successfully. The frame-wise tool detection and characterization were consistent with the reported ground truth. The development of a robust neurosurgical tool detection and characterization model using videos is a novel and significant step towards automating and characterizing neurosurgery in terms of the outcome of the surgery and assessing surgical skills.

REFERENCES

[1] S. Martin, E. Lopez-Martín, A. Lopez-Rey, J. Cubillo, A. Moreno-Pulido, and M. Castro, "Analysis of new technology trends in education: 2010–2015," *IEEE Access*, vol. 6, pp. 36 840–36 848, 2018.

[2] S. Martin, E. Lopez-Martin, A. Moreno-Pulido, R. Meier, and M. Castro, "A comparative analysis of worldwide trends in the use of information and communications technology in engineering education," *IEEE Access*, vol. 7, pp. 113 161–113 170, 2019.

[3] U. Shafi, R. Mumtaz, N. Iqbal, S. M. H. Zaidi, S. A. R. Zaidi, I. Hussain, and Z. Mahmood, "A multi-modal approach for crop health mapping using low altitude remote sensing, internet of things (iot) and machine learning," *IEEE Access*, vol. 8, pp. 112 708–112 724, 2020.

[4] Y. Han, J. Kim, and K. Lee, "Deep convolutional neural networks for predominant instrument recognition in polyphonic music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 208–221, 2017.

[5] L. Kohn, J. Corrigan, and M. Donaldson, *To Err is Human: Building a Safer Health System*, 01 2000, vol. 6.

[6] T. E. J. Z. M. J. . B. T. A. Gawande, A. A., "The incidence and nature of surgical adverse events in colorado and utah in 1992," *Surgery*, vol. 126, no. 1, p. 66–75, 1999.

[7] O. T. R. F. B. E. Healey MA, Shackford SR, "Complications in surgical patients," *Archives of surgery (Chicago, Ill. : 1960)*, vol. 137, no. 5, p. 611–618, 2002.

[8] A. Jin, S. Yeung, J. Jopling, J. Krause, D. Azagury, A. Milstein, and L. Fei-Fei, "Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks," 2018.

[9] F. Lalys, L. Riffaud, X. Morandi, and P. Jannin, “Surgical phases detection from microscope videos by combining svm and hmm,” in *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, B. Menze, G. Langs, Z. Tu, and A. Criminisi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 54–62.

[10] H. Lin, I. Shafran, D. Yuh, and G. Hager, “Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions,” *Computer aided surgery : official journal of the International Society for Computer Aided Surgery*, vol. 11, pp. 220–30, 10 2006.

[11] A. Krupa, J. Gangloff, C. Doignon, M. F. de Mathelin, G. Morel, J. Leroy, L. Soler, and J. Marescaux, “Autonomous 3-d positioning of surgical instruments in robotized laparoscopic surgery using visual servoing,” *IEEE Transactions on Robotics and Automation*, vol. 19, no. 5, pp. 842–853, 2003.

[12] F. A. e. a. Kranzfelder M, Schneider A, “Real-time instrument detection in minimally invasive surgery using radiofrequency identification technology,” *The Journal of Surgical Research*, vol. 185, no. 2, pp. 704–10, 2013.

[13] S. Kumar, M. Sathia narayanan, S. Misra, S. Garimella, P. Singhal, J. Corso, and V. Krovi, “Videobased framework for safer and smarter computer aided surgery,” *The Hamlyn Symposium on Medical Robotics*, pp. 107–108, 01 2013.

[14] R. Sznitman, K. Ali, R. Richa, R. H. Taylor, G. D. Hager, and P. Fua, “Data-driven visual tracking in retinal microsurgery,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 568–575.

[15] K. Ali, F. Fleuret, D. Hasler, and P. Fua, “A real-time deformable detector,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 225–239, 2012.

[16] M. Allan, S. Ourselin, S. Thompson, D. J. Hawkes, J. Kelly, and D. Stoyanov, “Toward detection and localization of instruments in minimally invasive surgery,” *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 4, pp. 1050–1058, 2013.

[17] B. Choi, K. Jo, S. Choi, and J. Choi, “Surgical-tools detection based on convolutional neural network in laparoscopic robot-assisted surgery,” in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017, pp. 1756–1759.

[18] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, “Endonet: A deep architecture for recognition tasks on laparoscopic videos,” *IEEE Transactions on Medical Imaging*, vol. 36, no. 1, pp. 86–97, 2017.

[19] A. P. Twinanda, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, “Single- and multi-task architectures for tool presence detection challenge at m2cai 2016,” 2016.

[20] A. Kanakatte, A. Ramaswamy, J. Gubbi, A. Ghose, and B. Purushothaman, “Surgical tool segmentation and localization using spatio-temporal deep network,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2020, pp. 1658–1661.

[21] S. Mukhopadhyay, M. Punchak, A. Rattani, Y.-C. Hung, J. Dahm, S. Faruque, M. C. Dewan, S. Peeters, S. Sachdev, and K. B. Park, “The global neurosurgical workforce: a mixed-methods assessment of density and growth,” *Journal of Neurosurgery JNS*, vol. 130, no. 4, pp. 1142 – 1148, 01 Apr. 2019. [Online]. Available: https://thejns.org/view/journals/j-neurosurg/130/4/article-p1142.xml

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[23] G. Jocher, A. Stoken, J. Borovec, NanoCode012, ChristopherSTAN, L. Changyu, Laughing, A. Hogan, lorenzomammana, tkianai, yxNONG, AlexWang1900, L. Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Hatovix, J. Poznanski, L. Y. , changyu98, P. Rai, R. Ferriday, T. Sullivan, W. Xinyu, YuriRibeiro, E. R. Claramunt, hopesala, pritul dave, and yzchen, “ultralytics/yolov5: v3.0,” Aug. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3983579

[24] H. Belhassen, H. Zhang, V. Fresse, and E. Bourennane, “Improving video object detection by seq-bbox matching,” in *VISIGRAPP*, 2019.

[25] R. Sznitman, K. Ali, R. Richa, R. H. Taylor, G. D. Hager, and P. Fua, “Data-driven visual tracking in retinal microsurgery,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 568–575.

[26] D. Bouget, R. Benenson, M. Omran, L. Riffaud, B. Schiele, and P. Jannin, “Detecting surgical tools by modelling local appearance and global shape,” *IEEE Transactions on Medical Imaging*, vol. 34, pp. 1–1, 12 2015.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[28] R. Girshick, “Fast r-cnn,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[29] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, “SSD: single shot multibox detector,” *CoRR*, vol. abs/1512.02325, 2015. [Online]. Available: http://arxiv.org/abs/1512.02325

[31] Z. Zhang, T. He, H. Zhang, Z. Zhang, J. Xie, and M. Li, “Bag of freebies for training object detection neural networks,” 2019.

[32] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “Yolov4: Optimal speed and accuracy of object detection,” 2020.

[33] B. Zoph, E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le, “Learning data augmentation strategies for object detection,” 2019.

[34] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” 2018.

[35] M. Sahu, A. Mukhopadhyay, A. Szengel, and S. Zachow, “Tool and phase recognition using contextual CNN features,” *CoRR*, vol. abs/1610.08854, 2016. [Online]. Available: http://arxiv.org/abs/1610.08854

[36] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, “Microsoft coco: Common objects in context,” 2015.

[37] I. Loshchilov and F. Hutter, “Sgdr: Stochastic gradient descent with warm restarts,” 2017.

[38] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, “Bag of tricks for image classification with convolutional neural networks,” 2018.

[39] P. Gil-Jiménez, H. Gómez-Moreno, R. López-Sastre, and S. Maldonado-Bascón, “Geometric bounding box interpolation: an alternative for efficient video annotation,” *EURASIP Journal on Image and Video Processing*, vol. 2016, 12 2016.

[40] Gurupadappa, “Paradigm development for intraoperarive skill assessment by video analysis,” Ph.D. dissertation, National Institute of Mental health and Neuro Sciences, Department of Neurosurgery, 2018.