Artificial Intelligence for Intraoperative Guidance

Using Semantic Segmentation to Identify Surgical Anatomy During Laparoscopic Cholecystectomy

Amin Madani, MD, PhD,✉ Babak Namazi, PhD,† Maria S. Altieri, MD, MS,‡
Daniel A. Hashimoto, MD, MS,§ Angela Maria Rivera, MD,* Philip H. Pucher, MD, PhD,||
Allison Navarrete-Welton,§ Ganesh Sankaranarayanan, PhD,† L. Michael Brunt, MD,¶
Allan Okrainec, MD, MHPE,* and Adnan Alseidi, MD, MEd#*

Objective: The aim of this study was to develop and evaluate the performance of artificial intelligence (AI) models that can identify safe and dangerous zones of dissection, and anatomical landmarks during laparoscopic cholecystectomy (LC).

Summary Background Data: Many adverse events during surgery occur due to errors in visual perception and judgment leading to misinterpretation of anatomy. Deep learning, a subfield of AI, can potentially be used to provide real-time guidance intraoperatively.

Methods: Deep learning models were developed and trained to identify safe (Go) and dangerous (No-Go) zones of dissection, liver, gallbladder, and hepatocystic triangle during LC. Annotations were performed by 4 high-volume surgeons. AI predictions were evaluated using 10-fold cross-validation against annotations by expert surgeons. Primary outcomes were intersection- over-union (IOU) and F1 score (validated spatial correlation indices), and secondary outcomes were pixel-wise accuracy, sensitivity, specificity, ± standard deviation.

Results: AI models were trained on 2627 random frames from 290 LC videos, procured from 37 countries, 136 institutions, and 153 surgeons. Mean IOU, F1 score, accuracy, sensitivity, and specificity for the AI to identify Go zones were 0.53 (±0.24), 0.70 (±0.28), 0.94 (±0.05), 0.69 (±0.20). and 0.94 (±0.03), respectively. For No-Go zones, these metrics were 0.71 (±0.29), 0.83 (±0.31), 0.95 (±0.06), 0.80 (±0.21), and 0.98 (±0.05), respectively. Mean IOU for identification of the liver, gallbladder, and hepatocystic triangle were: 0.86 (±0.12), 0.72 (±0.19), and 0.65 (±0.22), respectively.

Conclusions: AI can be used to identify anatomy within the surgical field. This technology may eventually be used to provide real-time guidance and minimize the risk of adverse events.

Keywords: artificial intelligence, bile duct injury, cholecystectomy, Convolutional neural network, deep learning, deep neural network, go zone, machine learning, no-go zone, patient safety

(*Ann Surg* 2022;276:363–369)

Operative complications are often blamed on technical mishaps. However, many adverse events tend to occur due to errors in human visual perception leading to errors in judgment that subsequently drive behaviors and actions that lead to adverse events.[1-4] For example, during laparoscopic cholecystectomy (LC), misinterpretation of the biliary anatomy is often caused by significant anatomic distortion from inflammation or aberrant anatomy, thus leading to major bile duct injuries.[4] Expert intraoperative performance that allows a surgeon to perform a safe dissection requires an ongoing process of interpreting the surgical field and making critical decision.[5] Theories of surgical expertise and empirical evidence suggest that expert surgeons have an ability to understand and visualize "safe" and "dangerous" zones of dissection within hostile territory and unknown anatomy.[6] This mental model provides the basis for exercising sound judgment and minimizing the risk of inadvertent injuries. Consequently, one potential method to improve performance during a procedure is by augmentation of the surgeon's mental model with real-time intraoperative guidance on the anatomy to improve quality and safety.

Artificial intelligence (AI), and its subfield Deep Learning, is a branch of computer science that uses algorithms to approximate human cognitive functions such as problem-solving, decision-making, object detection, and classification.[7] Algorithms such as deep neural networks can be trained without explicit programming using large quantities of data to learn to predict an outcome on new data. These methodologies have been

From the *Department of Surgery, University Health Network, Toronto, ON, Canada; †Center for Evidence-Based Simulation, Baylor University Medical Center, Dallas, TX; ‡Department of Surgery, East Carolina University Brody School of Medicine, Greenville, NC; §Department of Surgery, Surgical AI & Innovation Laboratory, Massachusetts General Hospital, Boston, MA; ||Department of Surgery, Imperial College London, London, UK; ¶Department of Surgery, Washington University School of Medicine, St. Louis, MO; and #Department of Surgery, University of California – San Francisco, San Francisco, CA.

✉amin.madani@uhn.ca.

Conflicts of Interest and Source of Funding: There were no sources of funding for this manuscript. D.A. is a consultant for Johnson & Johnson Institute, Verily Life Sciences, Worrell, and Mosaic Research Management. D.A. and A.N.W. have received grant funding from Olympus. A.O. has received honoraria for speaking and teaching from Medtronic, Ethicon, and Merck. A.A. is a consultant for Johnson & Johnson Institute.

The authors report no conflicts of interest.

Disclosures: A.M., B.N., M.S.A., A.M.R., P.P., G.S., L.M.B., and A.A. have no proprietary or commercial interest in any product mentioned or concept discussed in this article. D.H. is a consultant for Johnson & Johnson Institute, Verily Life Sciences, Worrell, and Mosaic Research Management. D.H. and A.N.W. have received grant funding from Olympus. Dr. Hashimoto has a pending patent on deep learning technology related to automated analysis of operative video that is unrelated to this work. A.O. has received honoraria for speaking and teaching from Medtronic, Ethicon, and Merck.

Author contributions: Concept and design: All authors; Acquisition, analysis, or interpretation of data: A.M., B.N., M.S.A, D.H., M.R., P.P., A.N.W., G.S.; Drafting of the manuscript: A.M., B.N., M.R., D.H.; Critical revision of the manuscript for important intellectual content: All authors; Statistical analysis: A.M., B.N., D.H.; All authors have given final approval of the version to be published.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.annalsofsurgery.com).

ISSN: 0003-4932/22/27602-0363
DOI: 10.1097/SLA.0000000000004594

applied towards the identification and classification of objects in images and videos. The end-result is a new generation of computer vision algorithms that are capable of identifying digital patterns in pixelated data to achieve human-level object detection.[8] Although deep learning has shown promising results for various computer vision tasks in medicine (eg, cancer diagnosis from radiology images, identification of polyps during colonoscopy),[9–12] its application and value for real-time surgical guidance and decision-support are much more complex and has yet to be demonstrated. Unlike images and videos from diagnostic radiology, fundoscopy, or endoscopy, surgical videos have significantly more variability in terms of background noise, image quality, and objects within the field. Furthermore, surgical planes and anatomical structures are almost never clearly delineated, and are often hidden or partially visible under fatty and fibrous tissues. This is a major obstacle for using computer vision in the operating room to provide clinically meaningful data. Using LC as an index procedure, the purpose of this study was to train deep learning models to identify anatomical landmarks as well as safe and dangerous zones of dissection, and to assess their performance compared to expert annotations.

METHODS

Videos of LC were used to train deep neural networks to accomplish 2 objectives: identify the safe zone of dissection (Go zone) and the dangerous zone of dissection (No-Go zone), and identify target anatomy, including the gallbladder, liver, and hepatocystic triangle. The Go zone was defined as the area located within the hepatocystic triangle (closer to the inferior edge of the gallbladder) that is deemed safe to proceed with dissection with a low probability of causing a major bile duct injury. The No-Go zone was defined as the deeper region within the hepatocystic triangle, where further dissection was deemed to be unnecessary and dangerous with an unacceptable probability of causing a major bile duct injury. The No-Go zone also included the hepatoduodenal ligament, liver hilum, and all structures inferiorly.

Dataset

LC videos were procured from several readily available and preobtained datasets. These included a total of 308 anonymized videos from 37 countries (including all continents), 153 surgeons and 136 different institutions, with the top countries being the United States, France, Canada, UK, and India (Table 1). Videos were obtained between 2008 and 2019. Specifically, two of the datasets used in this study (*Cholec80*[13] and *M2CAI16-workflow Challenge*[13,14] datasets) included 117 open-source videos derived from a single institution (Institute of Image-Guided Surgery, Strasbourg, France). All duplicate videos were removed. Eligibility criteria included laparoscopic recordings where a cholecystectomy was performed (ie, subtotal cholecystectomy was excluded), and where the cholecystectomy was not performed using a top-down approach. Given the anonymity of the videos, demographic and other personal identifying information was not attached to this dataset. Ten randomly-selected frames were extracted from each video from the moment the gallbladder was grasped to begin dissection of the hepatocystic triangle until just before clipping of the cystic structures, using *ffmpeg 4.1* software (www.ffmpeg.org). Frames wherein the camera was outside the operating field were excluded. Frames were reformatted to an aspect ratio of 16:9.

TABLE 1. Total Number of Institutions, Surgeons, and Videos Procured From Each Continent in the Entire Dataset

Continent*	Institutions	Surgeons	Videos
North America	21 (15%)	25 (16%)	43 (15%)
Central and South America	18 (13%)	18 (12%)	18 (6%)
Europe	33 (24%)	45 (29%)	149 (51%)
Middle East and Africa	22 (16%)	22 (14%)	22 (8%)
Central Asia	36 (26%)	36 (24%)	36 (12%)
East Asia and Pacific	6 (4%)	7 (5%)	7 (2%)
Unknown	—	—	15 (5%)
Total	136	153	290

Data are displayed as n (%).

*Countries include: Argentina, Australia, Bangladesh, Belgium, Brazil, Bulgaria, Canada, Chile, Columbia, Ecuador, Egypt, El Salvador, France, Georgia, Germany, Greece, India, Indonesia, Italy, Japan, Malaysia, Mexico, Mongolia, Morocco, Nepal, Netherlands, Pakistan, Palestine, Romania, South Africa, Spain, Switzerland, Tunisia, Turkey, United Arab Emirates, United Kingdom, United States.

Annotations

Three acute care and minimally invasive surgeons (A.M., M.A., P.P.) made free-hand annotations on all extracted frames to describe the location of the Go zone, No-Go zone, liver, gallbladder, and hepatocystic triangle within each frame. All annotations were reviewed for accuracy by a fourth high-volume hepatobiliary surgeon (A.A.). All 4 annotators are members of the Society of American Gastrointestinal and Endoscopic Surgeons' Safe Cholecystectomy Task Force and underwent an annotation training protocol, which included a discussion and agreement on the definition of each structure as well as a 1-hour practice session for using the annotation software. Annotations were done using Think Like A Surgeon software (https://thinklikeasurgeon.ca).[15,16] While watching the source video for context, annotations are made on the still frames and selected pixels are submitted and mapped on the back-end of the platform. Details of the annotation process are provided in the Supplemental Digital Content (Appendix 1, http://links.lww.com/SLA/C727). This software was initially developed to objectively assess experts' conceptualization of anatomical structures in the surgical field, with evidence for validity as an accurate reflection of their mental model.[17,18] To help eliminate annotation errors and interannotator variability, only pixels that were annotated by all surgeons were used as both input for training the network and as the ground truth (ie, standard criterion).

Development of the Model

The models used in this study were designed to perform semantic segmentation, which is a form of computer vision task whereby an object is identified and highlighted along its exact boundaries pixel-by-pixel as an overlay on the original video (Fig. 1). Details of the model are provided in the Supplemental Digital Content (Appendix 2, http://links.lww.com/SLA/C728). Three models were developed. The first 2 models separately identify the location of the Go zone and No-Go zone in each frame, where each pixel was classified as either the presence or absence of each zone. These models will be collectively referred to as *GoNoGoNet*. The third model simultaneously mapped the location of the anatomical structures in each frame, where each pixel was classified as either gallbladder, liver, hepatocystic triangle, or none of the above. This model will be referred to as *CholeNet*. All training and validation of GoNoGoNet and CholeNet was done using an Nvidia 1080Ti graphics processing unit (GPU).

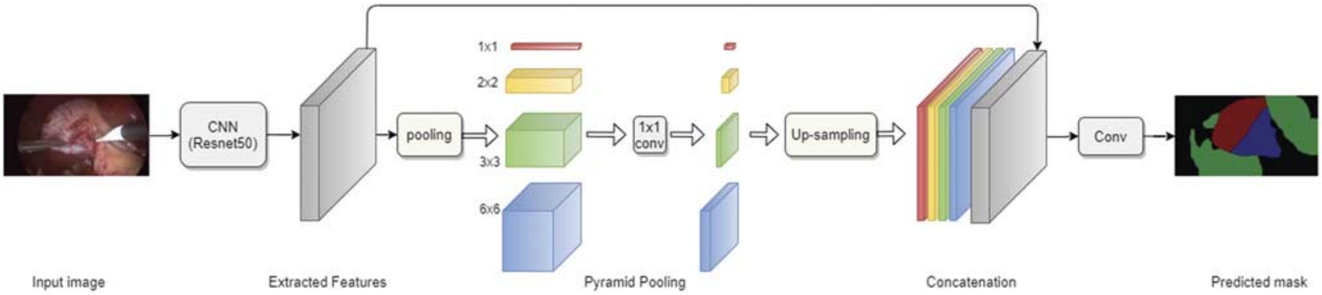

FIGURE 1. A Pyramid Scene Parsing Network (PSPNet) was used for pixel-wise semantic segmentation.[28] The architecture consists of a deep convolutional neural network (CNN; ResNet50) followed by a multiscale pyramid pooling module, which aggregates the feature maps from the CNN at 4 different scales (1 × 1, 2 × 2, 3 × 3, and 6 × 6) using the extracted frames. Details of the model are described in the Supplemental Digital Content (Appendix 2, http://links.lww.com/SLA/C728).

Evaluation of Model Performance

A 10-fold cross-validation technique was used to evaluate the performance of GoNoGoNet and CholeNet. The entire dataset of videos was randomly split into 10 partitions, whereby 9 partitions made up the development set and the 10th partition was the validation set. This development-validation sequence was executed 10 times to evaluate the models' performance based on the average of all 10 tests. The data were split randomly to include videos from all institutions into the various partitions on a per-case level rather than at a perframe level to ensure that frames from a particular video in the development set did not appear in the validation set.[19]

The primary outcomes were Intersection-Over-Union (IOU) and Dice/F1 spatial correlation index, which are commonly used metrics in computer vision to measure the overlap between the actual object location area (in this case, concordance of expert annotations, also known as the *ground truth)* and the AI's predicted area of a segmented object.[20,21] Secondary outcomes included the ability to correctly or incorrectly classify each pixel within an image with respect to the reference standard (ground truth), including mean accuracy, sensitivity, specificity, negative predicted value, and positive predicted value for every frame in the validation set. Since a 10fold cross-validation methodology was used, these metrics were averaged for all 10 sequences. Model outputs are displayed as overlay segmentations of the target structure. For GoNoGoNet, model outputs are demonstrated as both binary format (eg, each pixel that is predicted to be a Go zone is highlighted and each pixel that is predicted not to be a Go zone is not highlighted), and as a topographical heat map where each pixel is denoted as a probability of being part of a zone. Examples of these formats are shown in Figure 2.

Subsampling Experiments

Additional experiments were conducted to determine whether the results are reproducible both within an institution and between institutions. First, the models were trained and tested using data within the same institution after removal of all duplicate videos (development set: *Cholec80* dataset; validation set: *M2CAI16-work- flow Challenge* dataset). Subsequently, to determine whether a model can be trained on a large dataset and subsequently applied to an outside institution, training was performed using videos from all institutions outside the Institute of Image-Guided Surgery (Strasbourg, France) and validation was performed using videos from the *Cholec80* and *M2CAI16-workflow Challenge* datasets.

RESULTS

Of 308 available videos, 290 met the inclusion criteria. Resolution ranged between 854 × 480 pixels (480p) and 1920 × 1080 pixels (1080p) with an aspect ratio of either 16:9 or 4:3. A total of 2900 frames were extracted, 273 of which were excluded from the analysis and a total of 2627 frames were used for training and testing the AI models. From the included videos, 127 (44%) showed signs of acute or chronic cholecystitis, 63 (22%) required lysis of adhesions to reach the infundibulum of

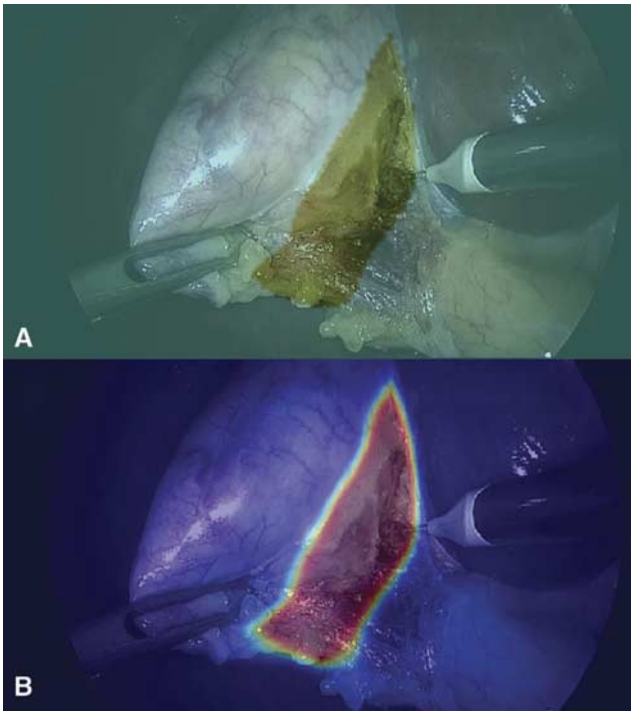

FIGURE 2. Model outputs are displayed as overlay segmentations of the target anatomical structure(s). For GoNoGoNet, model outputs are demonstrated as both binary format and topographical heat map. In the binary format example (A), each pixel is predicted to be either part of the Go zone (highlighted) or not part of the Go zone (not highlighted). In the probability heat map example (B), each pixel is denoted as a probability of being part of the Go zone (red region: highest probability; blue region: lowest probability).

TABLE 2. Annotation Concordance Among Annotators

Target Structure	Concordance*
Go Zone	0.89 (0.08)
No-Go Zone	0.91 (0.07)
Liver	0.96 (0.05)
Gallbladder	0.98 (0.04)
Hepatocystic triangle	0.86 (0.07)

Data are displayed as the mean value (Standard Deviation).

*Concordance was calculated using a validated visual concordance test methodology[14–17] as the average per-pixel agreement amongst annotators for each frame. The proportion of pixels that have 100% agreement by annotators is calculated for every frame, and the mean value is calculated for each specific target structure.

the gallbladder, 71 (24%) encountered major or minor bleeding, 6 (2%) encountered bile leak, and 3 (1%) encountered other events. A total of 13,135 annotation files were generated for GoNoGoNet and CholeNet. Annotation concordance between annotators for each target structure is summarized in Table 2.

Performance metrics for all models are shown in Table 3. For the entire dataset, mean IOU and F1 scores were >0.5 and >0.7 respectively showing good spatial overlap compared to ground truths. Accuracy for pixel-wise identification was consistently greater than 90% for all structures. Examples of model outputs for GoNoGoNet and CholeNet are shown in Figures 3 and 4, respectively. Inference time for all models was <0.01 seconds, satisfying real-time requirements. Examples of active segmentation during videos of LC are shown in the Supplemental Digital Content [Video 1 (Active segmentation of Go and No-Go zones during a LC. Model outputs are derived from GoNoGoNet. Green overlay: Go zone; Red overlay: No-Go zone.), http://links.lww.com/SLA/C721; Video 2 (Active segmentation of Go zone during a LC. Model outputs are derived from GoNoGoNet with segmentation as a probability heat map. Each pixel is denoted as a probability of being part of the Go zone (red region: highest probability; blue region: lowest probability), http://links.lww.com/SLA/C723; Video 3 (Active segmentation of No-Go zone during a LC. Model outputs are derived from GoNoGoNet with segmentation as a probability heat map. Each pixel is denoted as a probability of being part of the Go zone (red region: highest probability; blue region: lowest probability), http://links. lww.com/SLA/C725]. Performance metrics for GoNoGoNet and CholeNet were also segregated according to the presence of cholecystitis, need for lysis of adhesions, and the presence of bleeding or bile leak. These are summarized in the Supplemental Digital Content (eTable 1, http://links.lww.com/SLA/C726).

For the subset analyses, No-Go zone identification maintained a mean F1 score greater than 0.80, mean IOU of 0.68, and mean accuracy $>94\%$ for pixel-wise identification. For Go zone identification, mean F1 score and IOU decreased to 0.63 and 0.46, respectively, whereas accuracy remained $>90\%$.

DISCUSSION

AI is a general-purpose technology which has begun to permeate all facets of society, and although early success of deep learning has been highly encouraging in many clinical environments, its value in a surgical theater has been limited thus far. Given that most adverse events among surgical patients have root causes that can be traced back to intraoperative events,[1–3,22,23] there is an opportunity to utilize innovations in machine learning to design quality-improvement tools that directly address these unmet needs. In this study, we explored the use of deep learning in computer vision to identify complex and poorly defined anatomy within the surgical field. Specifically, deep learning models were developed and shown to perform semantic segmentation functions for the identification of safe and dangerous zones of dissection and other anatomical structures during LC with a high level of performance.

Given the importance of pattern recognition and creating an accurate mental model that reflects true surgical anatomy throughout an operation, we sought to utilize computer vision to segment the surgical field and to display these data as an overlay on the surgical field. This novel application of deep learning is a significant step given the many limitations and complexities of using surgical data, including frequent camera motions, variations in surgical anatomy, tissues characteristics, instruments, lighting, camera angle and surgical approach, presence of artefacts such as smoke and fluids, limited videos to use for training, and difficulty recruiting content experts

TABLE 3. Summary of Performance Metrics for GoNoGoNet and CholeNet

Target Structure	F1/Dice Score	IOU	Accuracy	Sensitivity	Specificity	NPV	PPV
Development set: complete dataset							
Validation set: complete dataset							
Go Zone	0.70 (0.28)	0.53 (0.24)	0.94 (0.05)	0.69 (0.20)	0.94 (0.03)	0.96 (0.05)	0.74 (0.28)
No-Go Zone	0.83 (0.31)	0.71 (0.29)	0.95 (0.06)	0.80 (0.21)	0.98 (0.05)	0.97 (0.05)	0.86 (0.30)
Liver	0.92 (0.10)	0.86 (0.12)	0.95 (0.04)	0.93 (0.10)	0.96 (0.04)	0.96 (0.04)	0.92 (0.11)
Gallbladder	0.84 (0.14)	0.72 (0.19)	0.95 (0.04)	0.83 (0.15)	0.97 (0.03)	0.97 (0.02)	0.84 (0.14)
Hepatocystic triangle	0.79 (0.18)	0.65 (0.22)	0.93 (0.05)	0.80 (0.17)	0.95 (0.04)	0.96 (0.03)	0.78 (0.19)
Development set: Institute of Image-Guided Surgery (Strasbourg, France)*							
Validation set: Institute of Image-Guided Surgery (Strasbourg, France)*							
Go Zone	0.63 (0.26)	0.46 (0.21)	0.94 (0.06)	0.59 (0.25)	0.97 (0.02)	0.96 (0.07)	0.67 (0.28)
No-Go Zone	0.80 (0.23)	0.68 (0.24)	0.94 (0.05)	0.74 (0.26)	0.98 (0.02)	0.95 (0.06)	0.89 (0.21)
Development set: all institutions except the Institute of Image-Guided Surgery (Strasbourg, France)*							
Validation set: Institute of Image-Guided Surgery (Strasbourg, France)*							
Go Zone	0.63 (0.30)	0.46 (0.24)	0.94 (0.04)	0.61 (0.32)	0.97 (0.04)	0.96 (0.03)	0.64 (0.29)
No-Go Zone	0.81 (0.29)	0.68 (0.28)	0.95 (0.05)	0.79 (0.30)	0.98 (0.02)	0.97 (0.05)	0.83 (0.03)

IOU, F1/Dice Spatial Correlation Index, Accuracy, Sensitivity, Specificity, NPV, and PPV are reported as the mean value (Standard Deviation) for every frame. Experimental results are shown for: the entire dataset using 10-fold cross-validation (ie, 10 repetitions of random 90/10 development-validation set split), the training and validation performed using the same institutional data, and validation performed on a single institution's data after training on the data from every other institutions.

NPV indicates negative predictive value; IOU, intersection-over-union; PPV, positive predictive value.

*Data from the Institute of Image-Guided Surgery (Strasbourg, France) was acquired from open-source videos from the *Cholec80* and *M2CAI16-workflow Challenge* datasets.

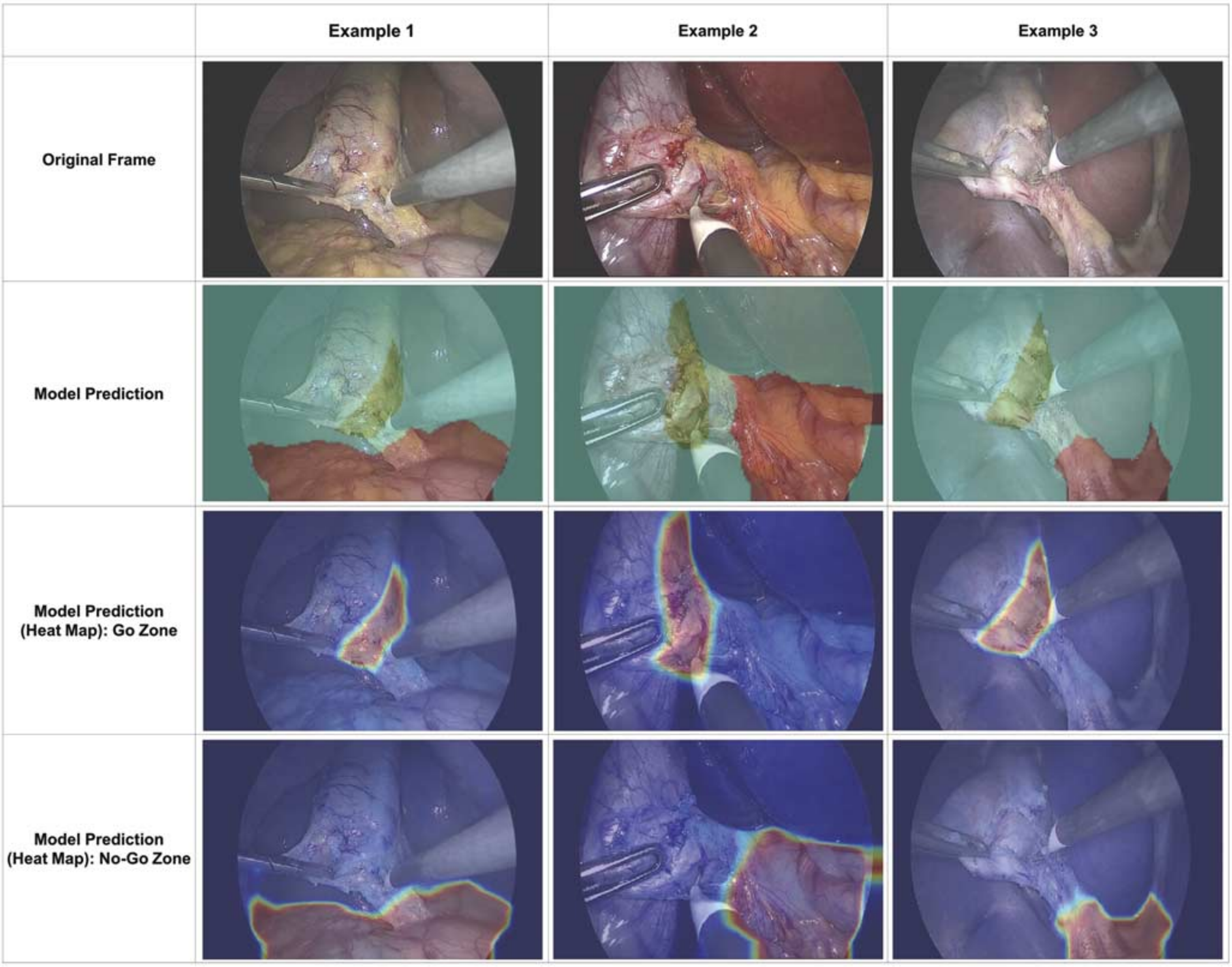
Example 1
Example 2
Example 3
Original Frame
Model Prediction
Model Prediction (Heat Map): Go Zone
Model Prediction (Heat Map): No-Go Zone

FIGURE 3. Model predictions for GoNoGoNet. Three separate examples of model predictions for GoNoGoNet compared to original frames are shown, displayed as overlays of Go zone (green overlay) and No-go zone (red overlay), and probability heat maps for Go and No-Go zones.

(surgeons) to annotate data for supervised machine learning. More importantly, identifying ill-defined anatomical structures that are covered with fatty and fibrous tissue and that have no clear border (eg, Go and No-Go zones) is a task that has more inherent obstacles for machine learning compared to structures on radiographic images or endoscopy, where borders are more delineated and the data have more consistency. This is a problem that is compounded by the fact that experts have significant variations in their annotations, which makes it difficult to establish a consensus reference standard. Furthermore, the common use of *bounding boxes* (rectangular box) for object localization and tracking in AI applications have a limited role in the surgical field where structures have more complex geometric configurations and blend with their background. For these tasks, semantic segmentation provides more valuable information.

Despite these challenges, GoNoGoNet and CholeNet showed high degrees of spatial correlation, as well as high accuracy between ground truth annotations and annotations predicted by the models. More importantly, when these predictions are coalesced to run throughout videos of LC, the deep learning models produced a consistent yet dynamic overlay regardless of the angle and perspective (Video 1, http://links.lww.com/SLA/C721; Video 2, http://links.lww.com/SLA/C723; Video 3, http://links.lww.com/SLA/C725). Such an overlay could provide augmented reality feedback to surgeons. Another important finding is the use of probability heat maps to represent uncertainty in the exact boundary of zones. These maps tend to fit more consistently with surgeons' mental models by representing Go and No-Go zones as probabilistic predictions of a generalized region that is safe or dangerous, as opposed to exact boundaries of a binary output, which are less realistic. As our previous qualitative studies have shown,[6] expert surgeon mental models are almost never binary. Instead, there will be indeterminate regions and even in the most routine cholecystectomies, there will seldom be an exact boundary between safe and dangerous zones of dissection. Most surgeons think in terms of probabilistic predictions and this study attempted to replicate this by using heat maps. In fact, as Figure 2 suggests, when Go and No-Go zones are simultaneously shown, there is a buffer zone (ie, an unlabeled area) between the 2, which

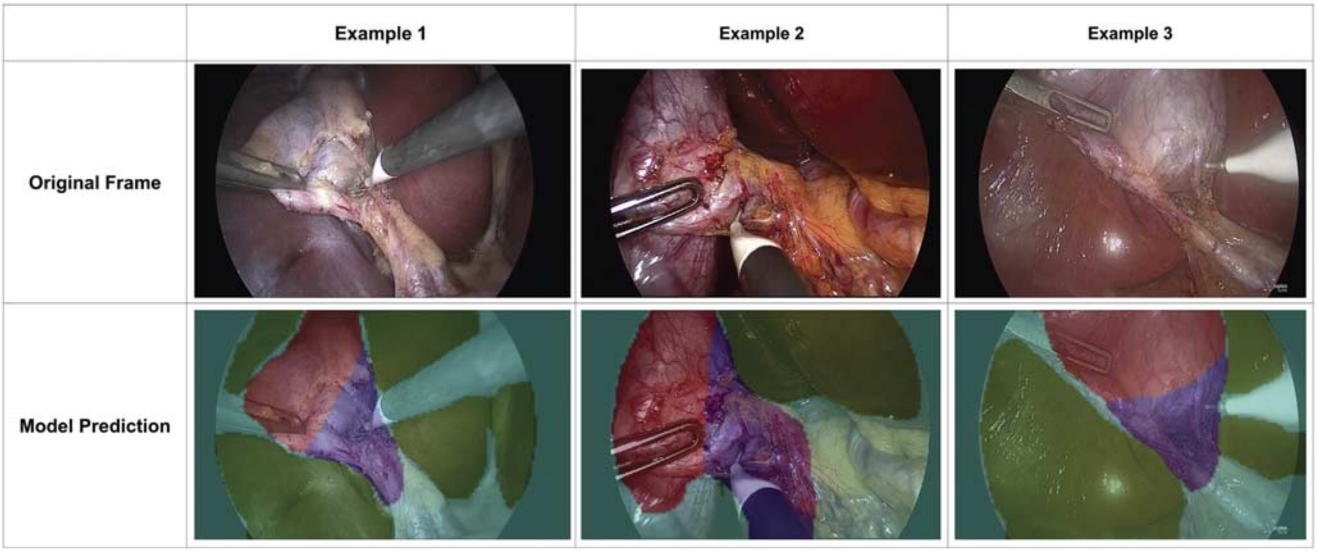

FIGURE 4. Model predictions for CholeNet. Three separate examples of model predictions for CholeNet compared to original frames are shown, displayed as simultaneous overlays of the liver (green overlay), gallbladder (red overlay), and hepatocystic triangle (purple overlay).

further supports GoNoGoNet's ability to approximate how expert surgeons perceive safe and unsafe areas of dissection.

Due to the very short inference time necessary for real-time predictions, this form of augmented reality has tremendous potential for quality-improvement. By providing real-time display of where to dissect and where not to dissect, this technology could potentially be incorporated in the future as part of integrated endoscopic platforms for real-time decision-support. This concept is analogous to current motor-vehicle navigation systems that are guided by computer vision and Global Positioning System (GPS) data to provide guidance on the optimal trajectory. However, additional trials will need to be conducted to demonstrate its generalizability, safety and effectiveness for deployment in real-time to improve surgical performance. Various design issues also need to be considered, including which data to provide (eg, Go zone, No-Go zone, or both), how to provide it, and when to provide it during a case. This will require additional survey and qualitative studies amongst end-users to evaluate the clinical and educational value of such a system and to make sure it is assimilated into the operative environment for optimal effectiveness. Similar to the evolution of automated vehicles, these data could eventually be incorporated into advanced robotic surgery platforms that increasingly aim to automate surgical tasks. Another important application of these results is in surgical education, performance assessment, and the emerging field of automated coaching.[24,25] Theories of professional expertise suggest that skill development in any task is developed through focused and deliberate practice of specific competencies – a principle that is universal for many professions, including surgery.[26] This requires specific, measurable and reproducible metrics for assessment and immediate feedback with ample opportunities for repetition. The data output from these models can be potentially used on-demand to assess intraoperative decision-making either as feedback on performance or deliberate practice of cognitive behaviors (such as accurate identification of safe and danger zones).

Although these preliminary results show promise, there are many limitations to consider, and its translation into the operating room as a routine surgical tool requires considerably more validity evidence. Although videos were obtained from a large range of sources and recording platforms, the models' performance on videos from a single institution diminished slightly when the development set did not include any videos from that institution. This suggests that larger datasets from a wide breadth of sources are required to develop a more robust model whose output is generalizable, has less inherent bias, and avoids other problems such as overfitting. Additional instances are crucial as one of the biggest limitations to AI is that the model can only make inferences and predictions based on the dataset that was used for training. In other words, pathology or anatomical configurations that were outside the scope of training may have lower performance on this model. It is also important to note that, although the aim of these models was to identify the Go zone, No-Go zone, hepatocystic triangle, gallbladder and liver, the models were not trained to identify variant biliary anatomy whose presence increases the risk of inadvertent major bile duct injury. Future iterations of these models will aim to detect and flag these aberrant anatomical variants as well as other intraoperative factors that may create a high-risk scenario (eg, significant inflammation, short cystic duct, suboptimal retraction and exposure of the hepatocystic triangle). In this study, data were obtained from many countries, patients with different ethnicities, videos with different qualities, and surgeons with different techniques and instruments. Half of the data also included patients with either acute or chronic inflammatory changes as opposed to only routine non-inflamed cases with very similar anatomy. Recently, there has been increasing momentum by several groups to create a central repository of surgical videos that can be used to coalesce many datasets and overcome these limitations.[27]

As the field of computer vision progresses at an exponential rate, newer and more advanced algorithms for semantic segmentation are becoming increasingly available to yield higher performance and better prediction abilities. Furthermore, most of the videos used in this study only utilized a random sample of ten frames, which represents a very small proportion of the total number of frames from each video. Although it is not feasible for a group of expert annotators to annotate every single frame for every video, it is possible to use semisupervised or unsupervised machine learning methodologies to automatically annotate many frames with subsequent review by a trained human annotator, increasing the amount of data that could be utilized for training. Finally, with the advent of Generative Adversarial Networks (ie, *deep fakes*), it is now possible to develop synthetic datasets which can potentially act as additional datasets to train deep neural networks.

Another major challenge is how to interpret metrics of performance. Although IOU and F1 score can objectively assess spatial overlap on a global level, it does not place weight on more important regions of interest. For instance, the value of the No-Go zone is in suggesting where not to dissect in the deeper regions of the hepatocystic triangle. However, its performance was also evaluated based on its ability to predict other less relevant regions in the No-Go zone (eg, porta hepatis, duodenum), where visual misinterpretation of biliary anatomy would be very unlikely. It is also interesting to note that sensitivity was consistently lower than specificity in this study. Whereas specificity and positive predictive value are more important for Go zones (ie, AI makes the suggestion to dissect in an area that has a high likelihood of being safe), sensitivity and negative predictive value are more applicable for No-Go zones, where a user can develop a sense of security to dissect in areas not identified as "dangerous". As the field expands and better metrics are developed for computer vision segmentation tasks and other applications (eg, automated robotic surgery), additional studies will be necessary to demonstrate its value and impact on patient care.

CONCLUSIONS

This study suggests that deep learning can be used to identify safe and dangerous zones of dissection and other anatomical structures in the surgical field during LC with a high degree of performance. As additional evidence emerges on the safety and effectiveness of using AI in the operating room, these automated computer vision tasks have the potential to augment performance and eventually be used for real-time decision-support and other qualityimprovement initiatives in the future.

ACKNOWLEDGMENTS

The authors acknowledge the work of Mr. Robert Messina for developing the Think Like A Surgeon annotation tool.

REFERENCES

1. Gawande AA, Thomas EJ, Zinner MJ, et al. The incidence and nature of surgical adverse events in Colorado and Utah in 1992. *Surgery*. 1999;126:66–75.
2. Rogers SO Jr., Gawande AA, Kwaan M, et al. Analysis of surgical errors in closed malpractice claims at 4 liability insurers. *Surgery*. 2006;140:25–33.
3. Guru V, Tu JV, Etchells E, et al. Relationship between preventability of death after coronary artery bypass graft surgery and all-cause risk-adjusted mortality rates. *Circulation*. 2008;117:2969–2976.
4. Way LW, Stewart L, Gantert W, et al. Causes and prevention of laparoscopic bile duct injuries: analysis of 252 cases from a human factors and cognitive psychology perspective. *Ann Surg*. 2003;237:460–469.
5. Madani A, Watanabe Y, Feldman LS, et al. Expert intraoperative judgment and decision-making: defining the cognitive competencies for safe laparoscopic cholecystectomy. *J Am Coll Surg*. 2015;221:931–940. e8.
6. Madani A, Vassiliou MC, Watanabe Y, et al. What are the principles that guide behaviors in the operating room? *Ann Surg*. 2017;265: 255–267.
7. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018;2:719–731.
8. Hashimoto DA, Rosman G, Rus D, et al. Artificial intelligence in surgery: promises and perils. *Ann Surg*. 2018;268:70–76.
9. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577:89–94.
10. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115–118.
11. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402.
12. Berzin TM, Topol EJ. Adding artificial intelligence to gastrointestinal endoscopy. *Lancet*. 2020;395:485.
13. Twinanda AP, Shehata S, Mutter D, et al. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans Med Imag*. 2017;36:86–97.
14. Stauder R, Ostler D, Kranzfelder M, Koller S, Feußner H, Navab N. The TUM LapChole dataset for the M2CAI 2016 workflow challenge. Published online October 28, 2016. Accessed June 24, 2020. Available at: http://arxiv.org/abs/1610.09278.
15. Madani A, Watanabe Y, Bilgic E, et al. Measuring intra-operative decisionmaking during laparoscopic cholecystectomy: validity evidence for a novel interactive Web-based assessment tool. *Surg Endosc*. 2017;31:1203–1212.
16. Madani A, Gornitsky J, Watanabe Y, et al. Measuring decision-making during thyroidectomy: validity evidence for a web-based assessment tool. *World J Surg*. 2018;42:376–383.
17. Madani A, Grover K, Watanabe Y. Measuring and teaching intra-operative decision-making using the visual concordance test. *JAMA Surg*. 2020;155:78.
18. Madani A, Keller DS. Assessing and improving intraoperative judgement. *Br J Surg*. 2019;106:1723–1725.
19. Liu Y, Chen P.-H.C., Krause J, et al. How to read articles that use machine learning: users' guides to the medical literature. *JAMA*. 2019;322: 1806–1816.
20. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Published online 2019. DOI: 10.1109/cvpr.2019.00075.
21. Sudre CH, Li W, Vercauteren T, et al. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support Published online*. 2017;240–248. DOI: 10.1007/978-3-319-67558-9_28.
22. Gawande AA, Zinner MJ, Studdert DM, et al. Analysis of errors reported by surgeons at three teaching hospitals. *Surgery*. 2003;133: 614–621.
23. Kable AK, Gibberd RW, Spigelman AD. Adverse events in surgical patients in Australia. *Int J Qual Health Care*. 2002;14:269–276.
24. Winkler-Schwartz A, Yilmaz R, Mirchi N, et al. Machine learning identification of surgical and operative factors associated with surgical expertise in virtual reality simulation. *JAMA Netw Open*. 2019;2: e198363.
25. Malpani A, Vedula SS, Lin HC, et al. Effect of real-time virtual reality-based teaching cues on learning needle passing for robot-assisted minimally invasive surgery: a randomized controlled trial. *Int J Comput Assist Radiol Surg*. 2020;15:1187–1194.
26. Anders Ericsson K, Hoffman RR, Kozbelt A, et al. The Cambridge Handbook of Expertise and Expert Performance. Cambridge Handbooks in Psychol; 2018.
27. Maier-Hein L, Vedula SS, Speidel S, et al. Surgical data science for nextgeneration interventions. *Nat Biomed Eng*. 2017;1:691–696.
28. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid Scene Parsing Network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Published online 2017. DOI: 10.1109/cvpr.2017.660.