A Multi-Task Convolutional Neural Network for Semantic Segmentation and Event Detection in Laparoscopic Surgery

Journal of *Personalized Medicine*

MDPI

Article

A Multi-Task Convolutional Neural Network for Semantic Segmentation and Event Detection in Laparoscopic Surgery

Giorgia Marullo [1,*], Leonardo Tanzi [1], Luca Ulrich [1], Francesco Porpiglia [2] and Enrico Vezzetti [1]

[1] Department of Management, Production, and Design Engineering, Polytechnic University of Turin, 10129 Turin, Italy
[2] Division of Urology, Department of Oncology, School of Medicine, University of Turin, 10124 Turin, Italy
* Correspondence: giorgia.marullo@polito.it

Abstract: The current study presents a multi-task end-to-end deep learning model for real-time blood accumulation detection and tools semantic segmentation from a laparoscopic surgery video. Intraoperative bleeding is one of the most problematic aspects of laparoscopic surgery. It is challenging to control and limits the visibility of the surgical site. Consequently, prompt treatment is required to avoid undesirable outcomes. This system exploits a shared backbone based on the encoder of the U-Net architecture and two separate branches to classify the blood accumulation event and output the segmentation map, respectively. Our main contribution is an efficient multi-task approach that achieved satisfactory results during the test on surgical videos, although trained with only RGB images and no other additional information. The proposed multi-tasking convolutional neural network did not employ any pre- or postprocessing step. It achieved a Dice Score equal to 81.89% for the semantic segmentation task and an accuracy of 90.63% for the event detection task. The results demonstrated that the concurrent tasks were properly combined since the common backbone extracted features proved beneficial for tool segmentation and event detection. Indeed, active bleeding usually happens when one of the instruments closes or interacts with anatomical tissues, and it decreases when the aspirator begins to remove the accumulated blood. Even if different aspects of the presented methodology could be improved, this work represents a preliminary attempt toward an end-to-end multi-task deep learning model for real-time video understanding.

Keywords: multi-task convolutional neural network; CNN; semantic segmentation; bleeding detection; laparoscopic surgery

Citation: Marullo, G.; Tanzi, L.; Ulrich, L.; Porpiglia, F.; Vezzetti, E. A Multi-Task Convolutional Neural Network for Semantic Segmentation and Event Detection in Laparoscopic Surgery. *J. Pers. Med.* **2023**, *13*, 413. https://doi.org/10.3390/jpm13030413

Academic Editors: Dorit E. Zilberman and Sabina Tangaro

Received: 4 November 2022
Revised: 10 February 2023
Accepted: 24 February 2023
Published: 25 February 2023

1. Introduction

Laparoscopy, widely known as minimally invasive surgery, is a surgical procedure that allows a surgeon to see within the abdomen and pelvis, without creating significant cuts [1], by having trocars with attached instruments controlled from consoles by the main operating surgeon and the assistant operator [2]. With the advancement of this technology, robotic-assisted laparoscopy has grown in popularity during the last few decades, and it is now extensively and thoroughly used in surgical procedures, replacing most open surgeries. The advantages of performing robotic-assisted laparoscopy include a smaller incision [3], faster recovery [4,5] and, consequently, shorter postoperative hospital stay [6,7], better aesthetic outcomes [6], less discomfort [8,9], a decreased risk of infection [2], and no oncological drawbacks in cancer patients [10]. Furthermore, the laparoscope's enlarged vision allows surgeons to observe anatomical structures in detail and accurately dissect, suture, and repair them [11].

On the other hand, several drawbacks linked to this methodology have to be considered. The workspace during laparoscopic surgery is narrower [12] than for open surgery and the field of view is more limited [10]; the difficulty in maintaining hemostasis rises [13]; the extra personnel is costly [14,15]; more work is delegated to humans; and the

primary operator and the assistant operator may have communication blunders as they are not staring at the same console [2].

Dealing with intraoperative bleeding is one of the most difficult aspects of laparoscopic surgery [16] accounting for 23% of all adverse events [17]. Some efforts have been undertaken in recent years to speed up bleeding identification in endoscopic procedures. In particular, most of the methods detected bleeding by utilizing the RGB space parameters [18] or categorizing pixels into "blood" or "non-blood" using color features. These techniques can process and classify information [19–21] using a machine learning approach, such as the Support Vector Machine (SVM) [22], that tries to maximize the distance between elements belonging to different classes, or using deep learning [23] which aims at tackling challenging issues by breaking complex concepts down into smaller ones and portraying them as a nested hierarchy of tasks with various levels of abstraction.

Within the deep learning scenario, convolutional neural networks (CNNs) proved to be the most suitable option to automatically extract features and detect adverse events [24], segment bleeding sources and display them to the surgeon [16], classify the images into bleeding and non-bleeding [25], or for real-time bleeding point location, recognition, and tracking [13,26]. CNNs have a unique design that allows them to interact with images while also leveraging their spatial patterns and being quick to train. This efficiency enables us to train deep and multi-layer networks. As a result, these networks achieve exceptional picture categorization and identification outcomes [27]. When compared to other image classification methods, CNNs require very minimal pre-processing. This implies that, in contrast to traditional methods, the network learns to improve the filters (or kernels) through automatic learning. This freedom from past information and human interference in feature extraction is a significant benefit [28]. However, most of the mentioned research has limited use in real-time videos, or cannot be considered end-to-end, since they require time-consuming additional steps to obtain the final result. In other words, end-to-end CNN architecture allows us to retrieve a solution without implementing further steps, such as stabilizers to minimize jitter and smooth camera path, expensive multi-stage temporal CNNs, or Optical Flow input data so that features may be gathered over numerous frames while leveraging temporal information.

CNN approaches have been employed in minimally invasive surgery for software and hardware-based solutions since enhanced computing power and recent breakthroughs in this field enable a standard computer to comprehend the content of an image or video stream in real time. Among the possible neural network approaches, multi-task learning (MTL) is a strategy that may efficiently solve multiple learning tasks in a CNN unified model. MTL more closely matches the learning process of humans than single-task learning since integrating information across domains is a core principle of human intelligence [29]. MTL is an area of machine learning that adopts a training paradigm in which a shared model learns many tasks at the same time [30]. This strategy increases data efficiency, decreases overfitting by exploiting shared features from multiple tasks, and could improve learning speed by leveraging contextual information [31], hence alleviating deep learning's renowned drawbacks, i.e., high data availability and computation power [32,33]. On the other hand, learning concepts for numerous tasks introduce problems that are not present in single-task learning. Choosing which tasks to study together is difficult since various tasks may have competing requirements. In this situation, improving a model's performance on one job could damage performance on another with distinct requirements.

This study proposes an end-to-end encoder–decoder multi-tasking CNN for joint blood accumulation detection and tool segmentation in laparoscopic surgery to maintain the operating room as clean as possible and, consequently, improve the physicians' visibility. For this purpose, we employed a shared backbone based on the encoder of the U-Net architecture [34], the gold standard for semantic segmentation in medical images. Two separate branches were instead implemented to classify the blood accumulation event and output the segmentation map. Our main contribution is the introduction of an

efficient multi-tasking approach for real-time surgical videos, trained with only RGB images and no other additional information, commonly unavailable in real applications.

The paper is organized as follows: Section 2 describes the system's architecture (Section 2.1), the dataset (Section 2.2), and the training process and metrics (Section 2.3); Section 3 illustrates and discusses the results obtained; finally, Section 4 summarizes and concludes the study.

2. Materials and Methods

The challenge of detecting bleeding in laparoscopic recordings is particularly complex because it requires distinguishing between the presence of blood residual (Figure 1a), which is not an index of bleeding, and when the blood is actively flowing (active bleeding) as shown in Figure 1b, which is the event of interest since it requires surgeon's prompt intervention. Detecting adverse occurrences during laparoscopic surgeries could be defined as an issue of action detection and localization. Action recognition and object segmentation interactions can be mutually advantageous and increase total video understanding. For example, precise positional identification of the key actor participating in action may boost the robustness of action recognition, and vice versa [31].

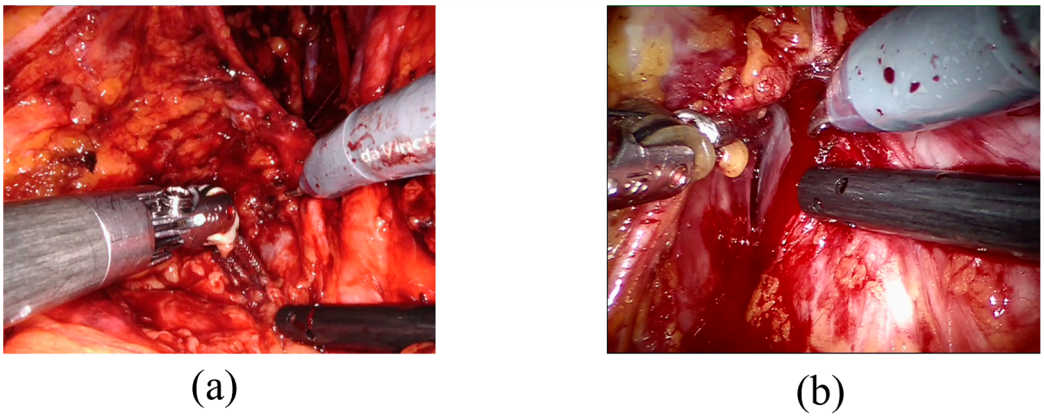


Figure 1. Examples of blood residue on tissues (**a**) and active bleeding (**b**).

The current investigation introduced a multi-task CNN, an architecture able to simultaneously learn multiple tasks. Particularly, in the examined case study, the model could jointly perform semantic segmentation and event detection, namely bleeding identification, in real-time during laparoscopic surgery. To this aim, a new architecture was implemented, fed with a properly manually labeled dataset. A shared trunk architecture was utilized for this purpose, which included a global feature extractor composed of convolutional layers shared by all tasks, followed by a different output branch for each output, performing the same computation for each input of the same task [29]. The weights for multiple tasks were pooled, such that each weight is trained to minimize several loss functions simultaneously. The CNN architecture, the dataset, and the training process are described in the following sections.

2.1. Neural Network Architecture

Figure 2 illustrates the network architecture. The backbone is a slightly modified version of the U-Net architecture [34].

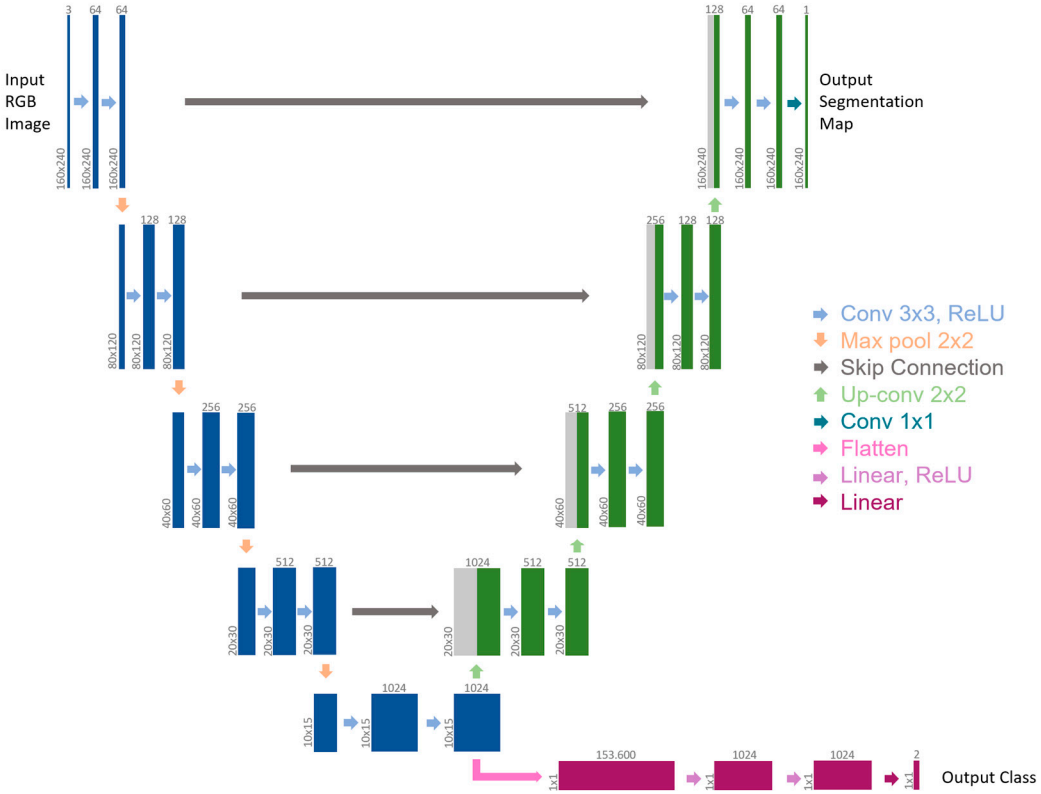


Figure 2. Multi-tasking CNN architecture (example for 160 × 240 input image). Each box represents a multi-channel feature map. The number of channels is provided on top of the box, while the x-y size is shown at the lower left edge. Blue boxes belong to the backbone, green boxes refer to the first branch for semantic segmentation, gray boxes denote copied feature maps (skip connections), and purple boxes represent the second branch for event detection. The arrows refer to the different operations.

Because of how it is designed, medical data can be analyzed in great detail, making the U-Net a model that is often used in the literature. To deal with RGB images, the number of input channels has been increased in comparison to the original architecture. The contracting path (left side) represents a global feature extraction, and it is shared by all tasks. It comprises a four-time repetition of the same sequence, meaning: two 3 × 3 convolutions (padded convolutions) to double the number of feature channels, each followed by a rectified linear unit (ReLU) and a down-sampling step, made of a 2 × 2 max pooling operation with stride 2, to halve the x-y image size. The last max pooling operation is followed by two 3 × 3 convolutions, which generate the bottleneck of the network. Differently from the original architecture of U-Net, two separate output branches derive from this bottleneck, each addressing a distinct task. The first branch (right-top side) is adapted from the expansive path of U-Net. Here, every step is symmetrical to the related contracting part. It includes an upsampling of the feature map, a 2 × 2 convolution ("up-convolution") to halve the number of feature channels, a concatenation with the symmetrical feature map of the contracting path (skip connection), and two 3 × 3 convolutions that double the x-y image size, each followed by a ReLU. The final layer of this branch applies a 1 × 1 convolution to map each 64-component feature vector to the desired number of classes in the output segmentation map. Unlike U-Net, padded convolutions were employed so that the output segmentation map and the input RGB image had the same size.

The second branch is connected to the encoder output, which is the U-Net architecture's bottleneck. As a result, it uses the U-Net encoder as the backbone for feature extraction and, from a flattened version of the bottleneck as input, it tackles event detection as

a classification problem. This branch is based on a sequence of fully connected layers: two Linear layers with 1024 features, each followed by a ReLU and a Dropout Layer, and a final Linear layer that maps each 1024-component feature vector to the desired number of classes. In this instance, two classes could appear in the output: 0 for "*no blood accumulation*" and 1 for "*blood accumulation*".

2.2. Dataset

The images were acquired from 26 endoscopic videos recording Robotic Assisted Radical Prostatectomy (RARP), a laparoscopic procedure conducted to remove the prostate gland and tissues surrounding it in case of prostate cancer. These data were provided by the "Division of Urology, University of Turin, San Luigi Gonzaga Hospital, Orbassano (Turin), Italy". The length of the procedures' videos ranged from a few seconds to 15 min. The recordings were initially edited, omitting sections when the endoscope was not within the abdominal cavity and the operational phases where there was no bleeding. This preliminary skimming produced 104 fragments with a length of less than one minute in 70% of cases or a few minutes in the remaining instances. Then, frames were extracted from these surgical pieces, considering a rate of around one frame per second. Finally, samples that may affect CNN training due to poor resolution, or the absence of surgical equipment were removed. As a result, only high-quality frames were included in the final dataset, which comprised 318 images. The dataset was then divided into train, validation, and test sets which contained 200, 32, and 86 images, respectively. Among them, four videos were kept apart for subsequently testing the network in real time.

The training and validation samples were labeled under the supervision of specialized medical personnel, and the results were assessed in the same manner.

The images were tagged using two different labels according to the specific task, namely, semantic segmentation and event detection, as shown in Figure 3.

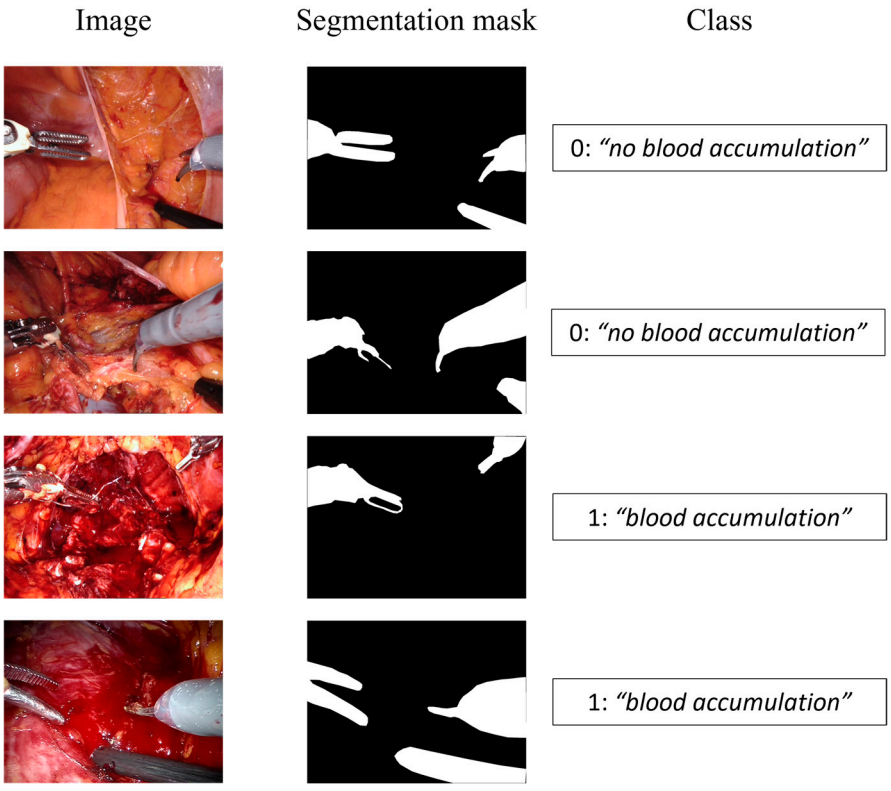


Figure 3. Dataset samples. The first column displays the input image, the second shows the segmentation mask for the semantic segmentation branch, and the third represents the class for the event detection branch.

The semantic segmentation consists in identifying the pixels of the image belonging to the surgical tool, labeling the region of interest (ROI) as "tool" and the other pixels as "background". On the other hand, the event detection task aims at classifying if the blood is accumulating and an intervention by the surgeon is required, or if the surgical field of view is not affected by a too copious accumulation of blood. To this aim, two possible classes were considered, 0 for "*no blood accumulation*" and 1 for "*blood accumulation*".

Data augmentation was added during the network training, to improve the numerosity and transformation invariance of the medical image dataset. Particularly, train samples were rotated by a random factor in the range (−35, 35), and randomly flipped, including vertically, horizontally, or both flips.

2.3. Training and Metrics

The multi-tasking architecture was trained for 30 epochs using a batch size of 32, and an Adam optimizer with a learning rate of 0.0001. A multi-task loss was chosen for parameter optimization:

$$loss = seg_{loss} + cls_{loss}, \quad (1)$$

where seg_{loss} is a Binary Cross Entropy Loss function followed by a *Sigmoid* activation function, and cls_{loss} is a Cross Entropy Loss function followed by a *Softmax* activation function. The model ran on an NVIDIA Quadro P4000 GPU, adopting the open-source *PyTorch* machine learning framework, written in Python, and based on the Torch library. The semantic segmentation branch accuracy was assessed by the Dice Coefficient (F1 Score) metric, a diffuse metric for semantic segmentation, defined as:

$$Dice\ Coefficient = \frac{2 \times Overlap\ Area}{Total\ pixels\ combined}, \quad (2)$$

where the overlap area represents the intersection between the pixels belonging to the predicted segmentation masks and those belonging to the ground truth one, and the total pixels parameter represents the total number of pixels in both images. The Dice Coefficient ranges from 0 to 1, where the edge values mean completely wrong and perfectly correct predictions, respectively.

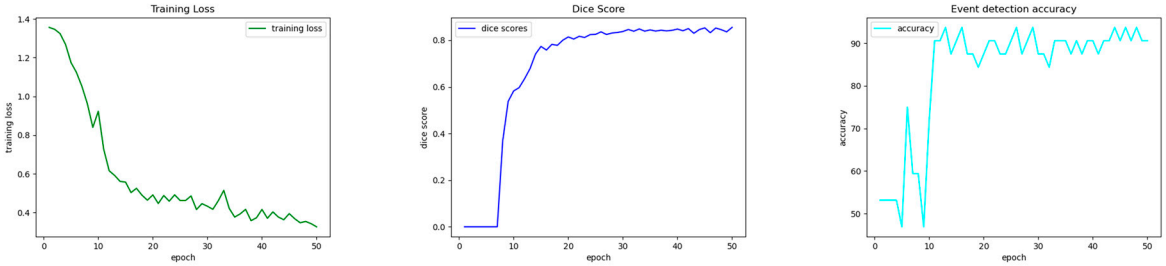
The event detection branch accuracy was calculated as follows:

$$Classification\ Accuracy = \frac{\#correct\_predictions}{\#samples}. \quad (3)$$

This value was monitored both to examine the accuracy of the entire branch, and the accuracy of each class separately.

3. Results and Discussion

The implemented multi-task CNN was tested both on images and videos. For each epoch the training loss (Figure 4a) and the validation metrics, namely Dice Score (Figure 4b) and event detection accuracy (Figure 4c), were plotted to show their trend. Following that, epoch 30 was picked for the final model since it was deemed the optimal tradeoff between the two branches of the network according to the experimental tests on images and videos.



(a) (b) (c)

Figure 4. Training and validation metrics trends. Multi-task Training Loss (**a**), Validation Dice Score for semantic segmentation branch (**b**), and Validation accuracy for event detection branch (**c**).

Concerning the tests on images, the multi-task CNN achieved a Dice Score equal to 81.89% for the semantic segmentation task, and an accuracy equal to 90.63% for the event detection task without any pre- or postprocessing step. Furthermore, the accuracy for each class was detailed, obtaining an accuracy of 86.67% and 94.12% for classes "*no blood accumulation*" and "*blood accumulation*", respectively.

Afterward, the network was also tested on videos to assess its real-time performance. Starting from test videos with a resolution of 1280 × 720 and a frame rate of 30 frames-per-second, the model output a processed video stream with 15 frames-per-second, when the prediction was performed for each frame. Comparative real-time and accuracy tests were carried on, and the network output was experimentally observed at different prediction frequencies, namely, the number of frames between one prediction and the next one. A test input video stream with a frame rate of 30 frames per second was employed to accomplish this test. For the real-time comparison tests, ten distinct values were investigated, as given in Table 1, with each indicating the prediction frequency.

Table 1. Comparison between prediction rate and frame rate during video tests.

Prediction Frequency	Frames-per-Second
1	15
2	21
3	22
4	22
5	23
10	26
15	28
20	29
25	29
30	30

The condition under which the prediction is made on all frames of the video stream was chosen as the minimal value, which is one frame. Instead, 30 frames were determined as the largest possible value, assuming one prediction each second. Higher values were not examined because, as previously stated, active bleeding necessitates immediate action; hence, an update rate of one second was deemed a limiting number. As seen in the table, just one prediction per second is required for the processed video stream to have the same frame rate as the input stream. However, it has been observed experimentally that the accuracy reduces considerably, particularly for segmentation masks, due to the rapid movement of the surgical instruments. In contrast, when the prediction frequency is lowered, there is a wider tolerance in terms of the percentage of blood accumulation predicted. The experimental findings show that lowering the forecasts by 50% and making

a prediction every two frames delivers an increase in frame rate without impacting prediction accuracy; hence, it was regarded as the ideal threshold, as shown in Figure 5.

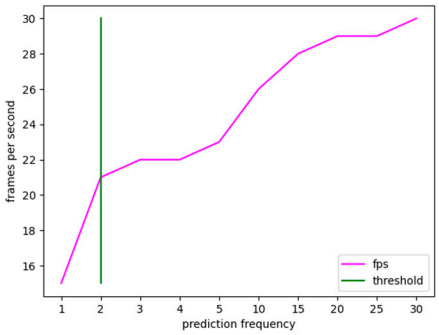


Figure 5. Frame rate trend versus prediction refresh rate calculated on a test video stream. The value 2 was experimentally chosen as the optimal tradeoff between real-time and accuracy.

As a result, the 21-frames-per-second limitation was assessed by the medical equipment to be the upper limit as the trade-off between real-time and accuracy. In other words, further increasing the frame rate at the expense of accuracy was considered unacceptable.

Figure 6 shows some examples of output frames, containing the output segmentation map overlapped on the real counterpart, and the predicted percentage of the "*blood accumulation*" class, if it is greater than 50%. As can be seen from the figure, both branches provided satisfactory results. The CNN succeeded, although with some marginal imperfections, in correctly recognizing the surgical tools and overlaying the mask. In addition, the event related to blood accumulation was also properly detected, as the percentage goes above 50% when bleeding is effectively visible.

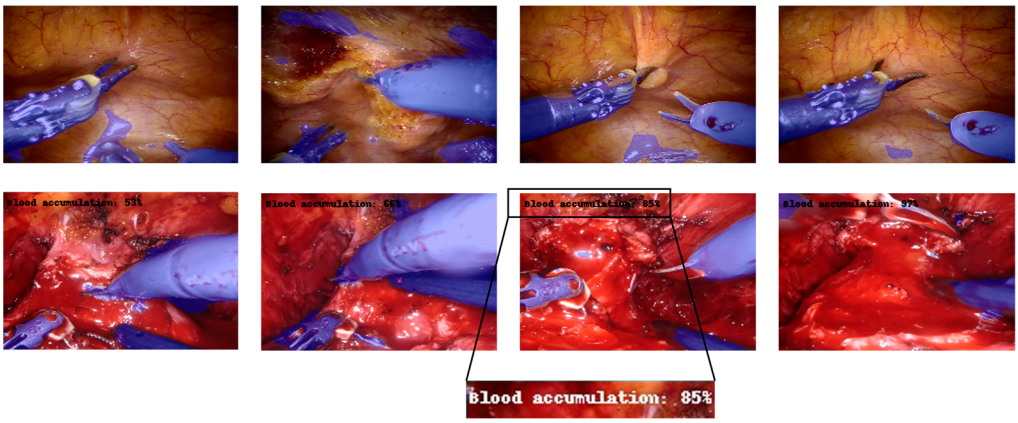


Figure 6. Multi-tasking CNN architecture (example for 160 × 240 input image). Each box represents a multi-channel feature map. Tools detected are blue-colored, while the blood accumulation percentage is indicated in the top-left corner of the image. The first row contains samples for which the percentage of blood accumulation is not displayed because the model predicted a value of less than 50%. The second row, in contrast, displays samples with a blood accumulation percentage greater than 50%.

Furthermore, it is noteworthy that during the test on the videos, the network adequately recognized the reduction in accumulated blood in the operating scene as the percentage decreased when the laparoscopic aspirator was removing the accumulated blood (Figure 7).

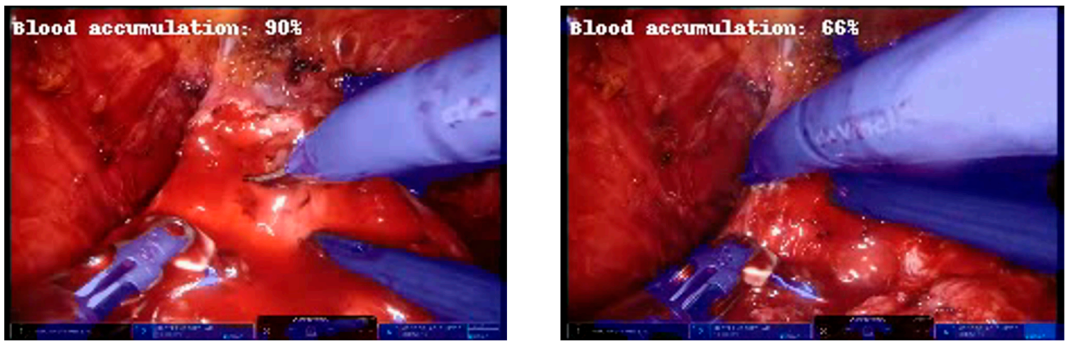


Figure 7. Example of frames at about 3 s distance in which the predicted percentage of blood accumulation decreases as the aspirator removes excess blood.

It can be inferred that tasks to be solved simultaneously were chosen properly since the features extracted from the shared backbone for tool segmentation also proved worthwhile for event detection since the two tasks are related. In fact, in most cases, active bleeding occurs at the time when one of the tools closes or interacts with anatomical structures such as the prostate. On the other hand, there is a reduction in accumulated bleeding when the aspirator starts to extract the blood from the surgical field. However, it was noted that this mutual benefit was lost when the number of epochs increased. In this situation, there was an improvement in accuracy relative to semantic segmentation, but the event detection task degenerated. This situation according to which the accuracy related to one task increases at the expense of the accuracy-related to the other task is not uncommon and, as already mentioned, is a known issue in the literature.

It was also possible to make assumptions about the reasons for the network's flaws because frames taken from recordings of actual interventions were investigated. Particularly when the light is changing, it could be challenging to distinguish between surgical instruments. In those circumstances, active bleeding might be misinterpreted with passive bleeding on the walls or on other anatomical parts within the surgical field.

The achieved results fulfilled the aim of using the same CNN architecture to simultaneously identify surgical tools in the field of view and detect the bleeding in real time. Future work is going to be planned to proceed in four different directions:

1. Neural Network Architecture. The architecture of the network should be extended to consider temporal information extracted from sequential images. This improvement will likely enhance the branch related to event detection in terms of accuracy and reliability. In this research context, for example, it may be easier to distinguish passive bleeding, namely, blood residue on tissues, and active bleeding, which is the surgeons' object of interest. Moreover, the semantic segmentation task should distinguish tools with different labels to improve the prediction, while the detection task should be extended by adding the localization of the origin of bleeding, which may provide remarkable clinical advantages when the human eye cannot instantly catch it [35].
2. Dataset. An improved dataset in terms of numerosity and variance could be beneficial to increase the accuracy of prediction. Furthermore, the sequences of images that do not contain any structure of interest (for instance, the external view of the operating room, and the images inside the trocar) in a limited-size dataset might improve the knowledge of the network about the studied environment [35]. Alternatively, from an algorithmic perspective, it could be advantageous to provide depth information as well as RGB information, to improve the tools' tracking accuracy. To accomplish this advancement, 3D acquisition cameras should be integrated with the RGB cameras employed during surgical interventions.
3. Testing. The model should be evaluated in the operating room to determine its practical limitations in the setting of a real-time application and then enhanced accordingly by adding new features.

4. Research field. Extending the algorithm into different domains would be of interest, to perform further analysis both in terms of actor segmentation and tracking and from the point of view of event detection and classification.

4. Conclusions

This study implemented an end-to-end encoder-decoder multi-tasking CNN for joint blood accumulation detection and tool segmentation in laparoscopic surgery. One of the most problematic aspects of laparoscopic surgery is dealing with intraoperative bleeding. Since the operation area is constantly limited and blood rapidly fills the bleeding site, it is difficult to control bleeding during laparoscopic procedures, indeed any sort of surgical manipulation, including suction, grabbing, retraction, cutting, and dissection might result in rapid bleeding and immediate treatment is required to avoid significant consequences.

To the best of our knowledge, there are no other systems that address simultaneously both the surgical tools identification and the event detection, namely the bleeding detection, in an end-to-end fashion and with only RGB images as input. The current study's results suggested that multi-task learning may be a remarkable strategy for improving efficiency and performance by employing shared features from multiple tasks. In this sense, maintaining a high level of accuracy and preserving the real-time is of utmost importance to make the methodology suitable to be used in the surgical room and support surgeons during the interventions.

The obtained findings allow us to deal with real-time data performing more tasks at the same time and achieving a noteworthy trade-off between accuracy and performance. Future research is indeed aimed at enhancing the system in the following aspects: temporal information of sequential pictures could be taken into consideration to improve the accuracy (a new architecture adaptation will be needed), the dataset could be expanded to increase the variability of the data and improve the neural network generalization, and the test should be performed on live surgeries to tune the CNN parameters, although considerable changes are not expected since the current work has already used data provided by real interventions. Moreover, it would be desirable to involve other domains to produce a generalizable real-time framework helpful for applications in several disciplines.

Author Contributions: Conceptualization, G.M., L.T. and L.U.; methodology, G.M. and L.T.; software, G.M. and L.T.; validation, G.M. and L.T.; formal analysis, G.M., L.T. and L.U.; investigation, G.M., L.T. and L.U.; resources, F.P. and E.V.; data curation, G.M.; writing—original draft preparation, G.M. and L.T.; writing—review and editing, L.U. and E.V.; visualization, G.M. and L.T.; supervision, L.U., F.P., and E.V.; project administration, F.P. and E.V.; funding acquisition, F.P. and E.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: No new data were created in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Laparoscopy (Keyhole Surgery). 2018. Available online: https://www.nhs.uk/conditions/laparoscopy/ (accessed on 30 September 2022).
2. Kaping'a, F. Deep learning for action and event detection in endoscopic videos for robotic assisted laparoscopy. *Comput. Sci.* **2018**, 1–6.
3. Shah, A.; Palmer, A.J.R.; Klein, A.A. Strategies to minimize intraoperative blood loss during major surgery. *Br. J. Surg.* **2020**, *107*, e26–e38. https://doi.org/10.1002/bjs.11393.

4. Kurian, E.; Kizhakethottam, J.J.; Mathew, J. Deep learning based Surgical Workflow Recognition from Laparoscopic Videos. In Proceedings of the 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 10–12 June 2020; pp. 928–931.
5. Kimmig, R.; Buderath, P.; Heubner, M.; Aktas, B. Robot-assisted hysterectomy: A critical evaluation. *Robot. Surg. Res. Rev.* **2015**, *2*, 51–58. https://doi.org/10.2147/RSRR.S50267.
6. Basunbul, L.I.; Alhazmi, L.S.S.; Almughamisi, S.A.; Aljuaid, N.M.; Rizk, H.; Moshref, R.; Basunbul, L.I.; Alhazmi, L.; Almughamisi, S.A.; Aljuaid, N.M.; et al. Recent Technical Developments in the Field of Laparoscopic Surgery: A Literature Review. *Cureus* **2022**, *14*, e22246. https://doi.org/10.7759/cureus.22246.
7. Casella, A.; Moccia, S.; Carlini, C.; Frontoni, E.; De Momi, E.; Mattos, L.S. NephCNN: A deep-learning framework for vessel segmentation in nephrectomy laparoscopic videos. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 6144–6149.
8. Kaushik, R. Bleeding complications in laparoscopic cholecystectomy: Incidence, mechanisms, prevention and management. *J. Minimal Access Surg.* **2010**, *6*, 59. https://doi.org/10.4103/0972-9941.68579.
9. Smith, J.A.; Herrell, S.D. Robotic-Assisted Laparoscopic Prostatectomy: Do Minimally Invasive Approaches Offer Significant Advantages? *J. Clin. Oncol.* **2005**, *23*, 8170–8175. https://doi.org/10.1200/JCO.2005.03.1963.
10. Tomimaru, Y.; Noguchi, K.; Morita, S.; Imamura, H.; Iwazawa, T.; Dono, K. Is Intraoperative Blood Loss Underestimated in Patients Undergoing Laparoscopic Hepatectomy? *World J. Surg.* **2018**, *42*, 3685–3691. https://doi.org/10.1007/s00268-018-4655-1.
11. Guillonneau, B.; Vallancien, G. Laparoscopic radical prostatectomy: The montsouris technique. *J. Urol.* **2000**, *163*, 1643–1649. https://doi.org/10.1016/S0022-5347(05)67512-X.
12. Fuchs, H.; Livingston, M.A.; Raskar, R.; Colucci, D.; Keller, K.; State, A.; Crawford, J.R.; Rademacher, P.; Drake, S.H.; Meyer, A.A. Augmented reality visualization for laparoscopic surgery. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI'98*; Wells, W.M., Colchester, A., Delp, S., Eds.; Springer: Berlin/Heidelberg, Germany, 1998; pp. 934–943.
13. Hua, S.; Gao, J.; Wang, Z.; Yeerkenbieke, P.; Li, J.; Wang, J.; He, G.; Jiang, J.; Lu, Y.; Yu, Q.; et al. Automatic bleeding detection in laparoscopic surgery based on a faster region-based convolutional neural network. *Ann. Transl. Med.* **2022**, *10*, 546. https://doi.org/10.21037/atm-22-1914.
14. Rawlings, A.L.; Woodland, J.H.; Vegunta, R.K.; Crawford, D.L. Robotic versus laparoscopic colectomy. *Surg. Endosc.* **2007**, *21*, 1701–1708. https://doi.org/10.1007/s00464-007-9231-y.
15. Schroeck, F.R.; Jacobs, B.L.; Bhayani, S.B.; Nguyen, P.L.; Penson, D.; Hu, J. Cost of New Technologies in Prostate Cancer Treatment: Systematic Review of Costs and Cost Effectiveness of Robotic-assisted Laparoscopic Prostatectomy, Intensity-modulated Radiotherapy, and Proton Beam Therapy. *Eur. Urol.* **2017**, *72*, 712–735. https://doi.org/10.1016/j.eururo.2017.03.028.
16. Rabbani, N.; Seve, C.; Bourdel, N.; Bartoli, A. Video-Based Computer-Aided Laparoscopic Bleeding Management: A Space-Time Memory Neural Network with Positional Encoding and Adversarial Domain Adaptation. In Proceedings of The 5th International Conference on Medical Imaging with Deep Learning, Zurich, Switzerland, 6–8 July 2022.
17. Zegers, M.; de Bruijne, M.C.; de Keizer, B.; Merten, H.; Groenewegen, P.P.; van der Wal, G.; Wagner, C. The incidence, root-causes, and outcomes of adverse events in surgical units: Implication for potential prevention strategies. *Patient Saf. Surg.* **2011**, *5*, 13. https://doi.org/10.1186/1754-9493-5-13.
18. Garcia-Martinez, A.; Vicente-Samper, J.M.; Sabater-Navarro, J.M. Automatic detection of surgical haemorrhage using computer vision. *Artif. Intell. Med.* **2017**, *78*, 55–60. https://doi.org/10.1016/j.artmed.2017.06.002.
19. Fu, Y.; Mandal, M.; Guo, G. Bleeding region detection in WCE images based on color features and neural network. In Proceedings of the 2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS), Seoul, Republic of Korea, 7–10 August 2011; pp. 1–4.
20. Fu, Y.; Zhang, W.; Mandal, M.; Meng, M.Q.-H. Computer-Aided Bleeding Detection in WCE Video. *IEEE J. Biomed. Health Inform.* **2014**, *18*, 636–642. https://doi.org/10.1109/JBHI.2013.2257819.
21. Okamoto, T.; Ohnishi, T.; Kawahira, H.; Dergachyava, O.; Jannin, P.; Haneishi, H. Real-time identification of blood regions for hemostasis support in laparoscopic surgery. *Signal Image Video Process.* **2019**, *13*, 405–412. https://doi.org/10.1007/s11760-018-1369-7.
22. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. https://doi.org/10.1038/nbt1206-1565.
23. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. https://doi.org/10.1038/nature14539.
24. Wei, H.; Rudzicz, F.; Fleet, D.; Grantcharov, T.; Taati, B. Intraoperative Adverse Event Detection in Laparoscopic Surgery: Stabilized Multi-Stage Temporal Convolutional Network with Focal-Uncertainty Loss. In Proceedings of the 6th Machine Learning for Healthcare Conference, Virtual, 6–7 August 2021; pp. 283–307.
25. Jia, X.; Meng, M.Q.-H. A deep convolutional neural network for bleeding detection in Wireless Capsule Endoscopy images. In Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 16–20 August 2016; pp. 639–642.
26. Richter, F.; Shen, S.; Liu, F.; Huang, J.; Funk, E.K.; Orosco, R.K.; Yip, M.C. Autonomous Robotic Suction to Clear the Surgical Field for Hemostasis Using Image-Based Blood Flow Detection. *IEEE Robot. Autom. Lett.* **2021**, *6*, 1383–1390. https://doi.org/10.1109/LRA.2021.3056057.
27. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. https://doi.org/10.1109/5.726791.

28. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827.
29. Crawshaw, M. Multi-Task Learning with Deep Neural Networks: A Survey 2020. *arXiv* **2021**. https://doi.org/10.48550/arXiv.2009.09796.
30. Zhang, Y.; Yang, Q. A Survey on Multi-Task Learning. *IEEE Trans. Knowl. Data Eng.* **2021**, *34*, 5586–5609. https://doi.org/10.1109/TKDE.2021.3070203.
31. Ji, J.; Buch, S.; Soto, A.; Niebles, J.C. End-to-End Joint Semantic Segmentation of Actors and Actions in Video. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 734–749.
32. Hou, R.; Chen, C.; Shah, M. An End-to-end 3D Convolutional Neural Network for Action Detection and Segmentation in Videos 2017. *arXiv* **2017**. https://doi.org/10.48550/arXiv.1712.01111.
33. Goodman, E.D.; Patel, K.K.; Zhang, Y.; Locke, W.; Kennedy, C.J.; Mehrotra, R.; Ren, S.; Guan, M.Y.; Downing, M.; Chen, H.W.; et al. A real-time spatiotemporal AI model analyzes skill in open surgical videos. *arXiv* **2021**, arXiv:2112.07219.
34. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
35. Madad Zadeh, S.; Francois, T.; Calvet, L.; Chauvet, P.; Canis, M.; Bartoli, A.; Bourdel, N. SurgAI: Deep learning for computerized laparoscopic image understanding in gynaecology. *Surg. Endosc.* **2020**, *34*, 5377–5383. https://doi.org/10.1007/s00464-019-07330-8.