Hisham ElMoaqet*, Rami Janini, Tamer Abdulbaki Alshirbaji, Nour Aldeen Jalal, and Knut Möller

Using Vision Transformers for Classifying Surgical Tools in Computer Aided Surgeries

https://doi.org/10.1515/cdbme-2024-2056

Abstract: Automated laparoscopic video analysis is essential for assisting surgeons during computer aided medical procedures. Nevertheless, it faces challenges due to complex surgical scenes and limited annotated data. Most of the existing methods for classifying surgical tools in laparoscopic surgeries rely on conventional deep learning methods such as convolutional and recurrent neural networks. This paper explores the use of pure self-attention based models—Vision Transformers for classifying both single-label (SL) and multi-label (ML) frames in Laparoscopic surgeries. The proposed SL and ML models were comprehensively evaluated on the Cholec80 surgical workflow dataset using 5-fold cross validation. Experimental results showed an excellent classification performance with a mean average precision mAP=95.8% that outperforms conventional deep learning multi-label models developed in previous studies. Our results open new avenues for further research on the use of deep transformer models for surgical tool detection in modern operating theaters.

Keywords: Computer aided surgeries, Surgical tool classification, Vision Transformers, laparoscopic video analysis

1 Introduction

Laparoscopic surgery is a surgical procedure where a surgeon operates on the patient internal organs inside the abdomen without having to make large incisions in the skin. This offers benefits to patients by leading to decreased discomfort, quicker healing periods, and minimized scarring in comparison to conventional open surgical procedures. The avoidance of large incisions decreases the chances of complications like infections and excessive bleeding. In these procedures, surgeons use a laparoscope–a thin tube with a camera and other surgical instruments. Analyzing videos captured during these surgeries has potential intra- and post-surgery applications such as warning systems, decision-making support, operating room resource management, surgical report documentation, video database indexing, surgeon training, and skill assessment [1]. Recognising surgical tools in laparoscopic videos is a key to develop such applications.

Different analysis tasks have emerged in recent years for analyzing various aspects of laparoscopic videos. Most importantly the classification of surgical tools in each frame of the laparoscopic videos. Initially, the general approach for this task involved a combination of hand-crafted features like color, texture, shape, and motion-based features along with traditional machine learning algorithms such as support vector machines [2]. Recent approaches have employed convolutional neural networks (CNNs) [3], a specific category of neural network capable of acquiring visual features from images. However, despite the advancements made with CNNs, there is still room for exploration in improving the classification of surgical tools using self-attention-based architectures. In particular, transformers [4] have become the de-facto standard for natural language processing tasks but are relatively new in computer vision applications. Inspired by different works that combine CNNs with attention modules [5], this paper explores the use of pure self-attention based models—Vision Transformers (ViT) [6]—for classifying surgical tools in both single label and multi-label classification scenarios.

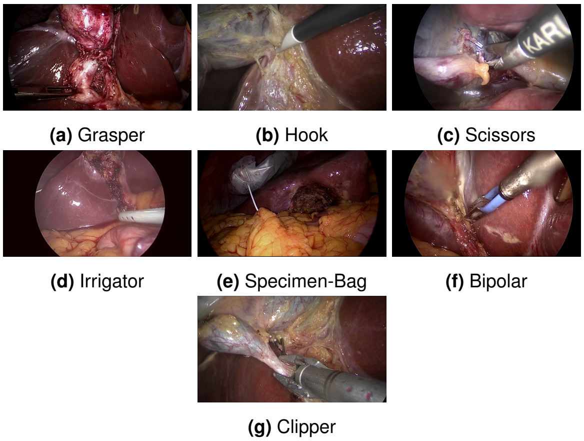
(a) Grasper **(b)** Hook **(c)** Scissors
(d) Irrigator **(e)** Specimen-Bag **(f)** Bipolar
(g) Clipper

Fig. 1: Surgical tools in Cholec80 dataset.

***Corresponding author: Hisham ElMoaqet,** Department of Mechatronics Engineering, German Jordanian University, 11118 Amman, Jordan, e-mail: hisham.elmoaqet@gju.edu.jo
Rami Janini, Department of Electrical Engineering, German Jordanian University, 11118 Amman, Jordan
Tamer Abdulbaki Alshirbaji, Nour Aldeen Jalal, Knut Möller, Institute of Technical Medicine (ITeM), Furtwangen University, Villingen-Schwenningen, 78054, Germany
Tamer Abdulbaki Alshirbaji, Nour Aldeen Jalal, Innovation Centre Computer Assisted Surgery (ICCAS), University of Leipzig, Leipzig, 04103, Germany

2 Dataset

The dataset used in this study is the Cholec80 dataset [7] which is a widely used benchmark for evaluating computer vision algorithms in laparoscopic surgical setting. This dataset includes 80 cholecystectomy surgery videos performed by 13 surgeons. The videos were recorded at 25 frames per second (fps) and downsampled to 1 fps for processing. Each video in the dataset was fully annotated to detect the presence of seven surgical tools. However, identifying these tools can be challenging due to factors such as blood and tissue occlusion, motion blur, fluctuating lighting conditions, cluttered backgrounds, smoke and fog. A tool is considered present if at least half of its tip is visible. The identifiable tools in the Cholec80 dataset are Grasper, Hook, Scissors, Irrigator, Specimen-Bag, Bipolar and Clipper as shown in Figure 1.

3 Methodology

3.1 Model Architecture

Vision Transformers begin with the systematic process of Patch Embedding. In this initial step, the input (e.g., an image) is partitioned into fixed-size square patches, usually of sizes 16, 32, or 64. These patches are then converted into vectors through a trainable linear projection. The resulting sequence of patch embeddings serves as input tokens for subsequent layers. Following Patch Embedding is the integration of Positional Embedding, which plays a crucial role in capturing spatial relationships. Typically, these embeddings are learned and incorporated into the patch embeddings during the initial stage [6].

The core architecture of the Vision Transformer includes multiple Transformer encoder layers [4]. Each layer consists of two primary sub-layers: multi-head self-attention and feedforward neural networks. The self-attention mechanism captures relationships between different patches within the input sequence by computing a weighted sum of all patch embeddings for each patch embedding, with the weights depending on the relevance to the current one. This enables prioritization of significant patches while considering both local and global contexts. Multi-head attention leverages multiple sets (attention heads) to capture diverse types of relationships [6].

After the self-attention process, each patch's output is processed through a feedforward neural network. This network usually includes a fully connected layer followed by an activation function like ReLU. The main purpose of the feedforward network is to introduce non-linearity, allowing the model to discern complex relationships between patches [6].

Layer Normalization and Residual Connections are then incorporated into the process. Both the self-attention mechanism and the feedforward network outputs undergo layer normalization and residual connections. Layer normalization plays a crucial role in stabilizing and speeding up training by standardizing the inputs to each sub-layer. Residual connections, also known as skip connections, involve adding the original input embeddings to the output of each sub-layer. This helps facilitate gradient flow during training and alleviates issues such as the vanishing gradient problem [6].

Training Transformer models to achieve optimal results requires a large amount of data. Therefore, it is common to pre-train large Transformer-based models on a substantial data before fine-tuning them for specific tasks. This paper evaluates two classification scenarios based on the ViT-Base model [6]. The first modeling scenario handled single-label classification (SL-Model), by adding a Softmax activation layer and categorical cross-entropy loss function. The second modeling scenario was designed for multi-label classification model (ML-Model) with a Sigmoid activation layer and binary cross-entropy loss function.

Tab. 1: Frequency of used frames in SL and ML Models.

Tool	SL-Model Frequency	ML-Model Frequency
Grasper	23494	102569
Hook	44886	103099
Scissors	1483	3254
Irrigator	2899	9814
Specimen-Bag	1545	11462
Bipolar	3222	8876
Clipper	2647	5986
Total	**80176**	**245060**

3.2 Handling Class Imbalance

In the initial single-label classification model, frames that contains multiple visible tools and those without any visible tools were excluded. For the multi-label classification model, the entire dataset was used. Table 1 displays the frequency of each tool and highlights a substantial imbalance in the data. This imbalance poses a challenge for training the models as it can lead to biased predictions and reduced performance on underrepresented classes. To tackle this challenge of imbalanced data, we involved class weights to be used during training. Class weights implement a flexible weighting system to address the disparity between majority and minority classes. Class weights were computed as follows (Eq. (1))

$$\mathrm{CW}_k = \frac{N}{K \times n_k} \tag{1}$$

where: N represents the total number of training data samples, K represents the number of classes, and n_k represents the number of samples n in class k.

Tab. 2: SL-Model classification performance with $k = 5$ folds.

Tool	Precision	Recall	F1-Score
Grasper	97.52	98.57	96.02
Bipolar	99.06	99.06	96.92
Hook	99.2	99.67	98.04
Scissors	91.10	88.08	96.49
Clipper	98.44	93.70	94.82
Irrigator	96.98	96.01	99.43
Specimen-Bag	98.05	91.52	89.56
Mean	**97.19**	**95.23**	**95.89**

3.3 Model Training and Evaluation

The training procedure involves resizing input images to the size of $224 \times 224 \times 3$ as required by the pretrained vision transformer model (ViT) and normalizing pixels in the range 0 to 1. Adam optimizer [8] was used with a learning rate of 5e-5, while the training batch size was set at 32 with 1000 warmup steps. Additionally, weight decay is adjusted to reduce overfitting, set at 0.01. The model was trained on four Nvidia GeForce GTX 1080 Ti GPU for 8 epochs.

For evaluation purposes, previous studies considered a 50-50% split for training and testing, using 40 surgical videos for each purpose. This method offers simplicity and speed, which is advantageous under limited computational resources or time constraints. Nevertheless, it may not provide the most robust estimate of model performance due to variability introduced by randomness in selection. To obtain a more accurate assessment of the proposed models' performance, we implemented a k-fold cross-validation technique on the dataset. This approach involved dividing the dataset into k folds and iteratively using $k-1$ folds for training and one fold for testing. For our study, we selected a value of $k = 5$ for conducting cross-validation resulting in each group consisting of 16 videos; thus utilizing four groups (64 videos) for training and leaving one group containing 16 videos for testing. This method is repeated k times to ensure that each group serves as the test set exactly once and the average value across these tests are computed to ensure rigorous evaluation of the model's generalizability across various sections of data. Finally, we evaluated our proposed models using 50-50% split for training and testing in order to compare our results with previously published studies that used similar evaluation criteria.

Tab. 3: ML-Model classification performance with $k = 5$ folds.

Tool	AP	Precision	Recall	F1-Score
Grasper	96.30	88.92	91.81	92.62
Bipolar	96.39	96.13	93.57	93.05
Hook	99.78	98.28	92.93	98.41
Scissors	90.34	92.24	81.21	84.36
Clipper	97.24	94.48	91.50	95.20
Irrigator	94.42	92.38	88.88	90.22
Specimen-Bag	96.56	93.61	91.16	94.86
Mean	**95.86**	**93.72**	**90.15**	**92.67**

4 Results

The single-label classification model (SL-model) achieved accuracy of 98.43%. Moreover, precision, recall and F1-score were computed for each surgical tool as shown in Table 2). For the multi-label model (ML-Model), the precision, recall and F1-Score were calculated as well as the average precision to provide a comprehensive evaluation of the model's effectiveness as shown in Table 3). Average precision (AP) is a popular metric for assessing the performance of object classification models. AP was first calculated by determining the area under precision-recall curve for each individual tool. After that, the mean-average-precision (mAP) is calculated by averaging AP across all 7 tool classes. mAP gives a holistic view of the object detection model's performance, making it a widely used metric for evaluating previous studies for tool classification in laparoscopic surgeries. Table 4) compares the mAP of the proposed approach with previously published studies considering the multi label scenario and 50-50% split for training and testing similar to previous methods in this field. Finally, heatmaps were generated to visually evaluate the performance of the proposed approach by calculating the mean of attention weights, normalizing them, and iteratively computing joint attentions. This process helps us understand how attention functions and assess its performance in this model. Figure 2 illustrates some examples for visualization of attention heatmaps that were generated from single and multi label frames.

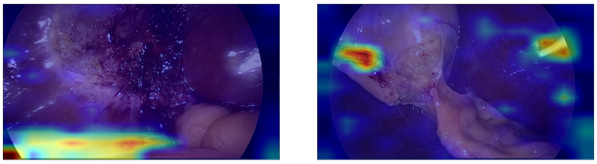

Fig. 2: Surgical tools attention visualisation.

Tab. 4: Comparison of AP Metrics (50-50% split)

Tool	**EndoNet [9]**	**MTRC Net [10]**	**Nwoye [11]**	**Jalal et al. [12]**	**Our Model**
Grasper	84.8	84.7	**99.7**	91.0	91.6
Bipolar	86.9	90.1	95.6	97.3	**99.7**
Hook	95.6	95.6	99.8	**99.8**	97.3
Scissors	58.6	86.7	86.9	90.3	**92.4**
Clipper	80.1	89.8	**97.5**	97.4	95.8
Irrigator	74.4	88.2	74.7	95.6	**96.3**
Specimen-Bag	86.8	88.9	96.1	**98.3**	97.7
Mean (mAP)	81.02	89.1	92.9	95.6	**95.8**

5 Discussion and Conclusion

This paper proposes a new deep learning approach for detecting the presence of surgical tools in Laparoscopic surgery videos. The proposed method leverages the state-of-the-art Vision Transformer models for classifying both single-label and multi-label frames in Laparoscopic surgeries. The models were fine-tuned on the Cholec80 dataset while incorporating class weights for handling class imbalance in the dataset.

The achieved results show high classification performance with F1-score of 95%. All surgical tools were classified with a precision and recall greater than 96%, except for the scissors and specimen-bag. Those tools have low number of samples in the training data, however, the SL-model was able to detect them with an F1-score of 96% and 89%, respectively.

The ML-model outperforms the previous approaches with an mAP of 95.8%. The self-attention mechanism helps to enhance model focus on discriminative areas of surgical tools in the images. On the other hand, the class weighting technique alleviates model biasing to tools with the majority of images. Thus, the ML-Model demonstrates high capability for classifying the underrepresented tools such as the scissors and irrigator with an AP of 92.4% and 96.3%, respectively. The conducted 5-fold evaluation emphasizes the robustness of both the SL- and ML-models, as the data were randomly partitioned into five folds. This ensures that the achieved high results are not due to special characteristics in training or testing data.

For future work, we plan to explore other deep transformer models for surgical tool classification, address issues with real-time object detection, as well as to explore in more depth surgical tool localization.

Author Statement

This work was supported by the German Jordanian University (GJU)–Study Group Mobility Program and the German Federal Ministry of Research and Education (BMBF) under project CoHMed/DigiMedOP grant no. 13FH5I05IA & project PersonaMed grant no. 13FH5I06IA.

References

[1] Jin, Yueming et al. “Multi-task recurrent convolutional network with correlation loss for surgical video analysis.” Medical image analysis vol. 59 (2020): 101572. doi:10.1016/j.media.2019.101572

[2] Primus, Manfred & Schoeffmann, Klaus & Böszörmenyi, Laszlo. (2015). Instrument Classification in laparoscopic Videos. Proceedings - International Workshop on Content-Based Multimedia Indexing. 2015. doi: 10.1109/CBMI.2015.7153616.

[3] LeCun, Yann et al. “Deep learning.” Nature vol. 521,7553 (2015): 436-44. doi:10.1038/nature14539

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017, June 12). Attention is all you need. arXiv.org. https://arxiv.org/abs/1706.03762

[5] Jalal, N. A., Alshirbaji, T. A., Docherty, P., Arabian, H., Laufer, B., Krueger-Ziolek, S., Neumuth, T., & Möller, K. (2023). Laparoscopic video analysis using Temporal, Attention, and Multi-Feature Fusion Based-Approaches. Sensors, 23(4), 1958. doi: 10.3390/s23041958

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020, October 22). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv.org. https://arxiv.org/abs/2010.11929

[7] Twinanda AP, Shehata S, Mutter D, Marescaux J, De Mathelin M, Padoy N. 2016. Endonet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imaging. 36(1):86–97. doi:10.1109/TMI.2016.2593957.

[8] Kingma, D. P., & Ba, J. (2014, December 22). Adam: A method for stochastic optimization. arXiv.org. https://arxiv.org/abs/1412.6980

[9] Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., & Padoy, N. (2017). EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Transactions on Medical Imaging, 36(1), 86–97. doi: 10.1109/tmi.2016.2593957

[10] Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C., & Heng, P. (2020). Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Medical Image Analysis (Print), 59, 101572. https://doi.org/10.1016/j.media.2019.101572

[11] Nwoye, C. I., Mutter, D., Marescaux, J., & Padoy, N. (2019). Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos. International Journal of Computer Assisted Radiology and Surgery (Print), 14(6), 1059–1067. https://doi.org/10.1007/s11548-019-01958-6

[12] Jalal, N. A., Alshirbaji, T. A., Docherty, P., Arabian, H., Laufer, B., Krueger-Ziolek, S., Neumuth, T., & Möller, K. (2023). laparoscopic video analysis using Temporal, Attention, and Multi-Feature Fusion Based-Approaches. Sensors, 23(4), 1958. https://doi.org/10.3390/s23041958