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

Spatio-Temporal Transformer for Surgical Instrument Recognition in Computer Aided Surgeries

https://doi.org/10.1515/cdbme-2025-0236

Abstract: Artificial Intelligence (AI) continues to transform medical imaging and surgical assistance systems, particularly in laparoscopic surgeries. Accurate recognition of surgical instruments across space and time is vital for workflow analysis and real-time decision support. This study presents a spatio-temporal transformer-based approach for multi-label surgical tool recognition in laparoscopic videos. We fine-tune the TimeSFormer network to capture spatial and temporal dependencies across video frames. To address challenges like class imbalance and visual occlusion, we incorporate a targeted data augmentation pipeline, balanced batch sampling, and Focal Loss. A new background masking technique further enhances model focus on tool regions by blurring irrelevant textures. Evaluated on the Cholec80 benchmark, our model achieves a mean Average Precision (mAP) of 96.3%, outperforming prior baselines. Attention heatmaps confirm effective tool tracking, underscoring the promise of spatio-temporal transformers in surgical AI.

Keywords: Laparoscopies video analysis; Computer aided surgeries; Surgical tool classification; Tool localization

1 Introduction

Minimally Invasive Surgery (MIS) enables surgeons to access internal organs with minimal external incisions. A common form of MIS is Laparoscopic Surgery (LS), widely used for diagnosing and treating conditions of the gallbladder, gastrointestinal tract, and pancreas. Compared to traditional open procedures, LS offers several patient benefits, including reduced post operative pain, faster recovery, and minimal scarring. In LS, specialized surgical instruments are inserted through small incisions in the abdominal wall, and the procedure is guided using a laparoscope, a slender instrument equipped with a camera and light source. The laparoscope transmits real-time video of the internal surgical site to external monitors, which serve as the primary data source for this work. Analyzing these videos offers value both intra operatively and post operatively. In real time, automatic detection of surgical tools can assist in identifying potential risks, such as instrument collisions or unintended tissue contact. Post-surgery, tool analysis supports surgical report generation, operating room resource management, video indexing, and surgeon training [1]. In this paper, we propose a spatio temporal transformer based model for surgical tool classification in laparoscopic videos by fine-tuning the TimeSFormer network [2]. The model applies divided space-time attention to a sequence of eight video frames, enabling it to capture both spatial and temporal dependencies relevant to tool presence. This work extends our previous research using Vision Transformers (ViT) for frame level tool classification [3], which demonstrated the effectiveness of transformer-based models without relying on convolutional architectures. By leveraging temporal information, our model can better handle scenarios where tools are only partially visible or momentarily occluded due to motion blur, blood, tissue, smoke, or fog. Temporal modeling allows the network to infer tool presence even when individual frames lack clear visual cues, offering improved robustness under challenging surgical conditions.

2 Methodology

2.1 Dataset

We use the Cholec80 dataset [4], a widely used benchmark for laparoscopic surgical video analysis. It contains 80 cholecystectomy procedures performed by 13 surgeons, recorded at 25 frames per second. Each frame is annotated with the presence of seven tools: *Grasper*, *Hook*, *Scissors*, *Irrigator*, *Specimen Bag*, *Bipolar*, and *Clipper*. A tool is labeled as present if at least half of its tip is visible. Tool identification is often complicated by challenges such as blood and tissue occlusion,

***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

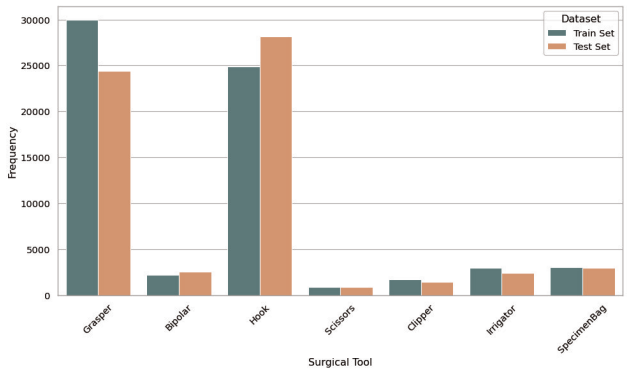

Fig. 1: Distribution of surgical tools in the training and testing set.

motion blur, lighting variations, cluttered scenes, smoke, and fog. For preprocessing, every 50 frames are extracted from each video. Each 50-frame clip is split into two 25-frame blocks, and a one-hot encoded tool label vector is generated for each. These are merged into a single vector representing tool presence across the 50-frame span. This approach offers better temporal coverage than using only 25-frame clips, which may capture limited motion. From each 50-frame segment, 8 frames are selected using linear spacing to ensure uniform temporal distribution. This sampling preserves spatio temporal context while maintaining consistent labels, enabling the model to leverage visual and temporal cues during fine tuning.

2.2 Data Imbalance

As illustrated in Figure 1, certain tool classes—specifically scissors and clipper—are significantly underrepresented in the dataset. This class imbalance can lead to biased model predictions, where the model tends to favor majority classes while neglecting minority ones. To mitigate this is, three complementary strategies were applied. First, data augmentation was performed at the preprocessing stage to increase the number of training samples for the underrepresented classes. Second, a class-balanced sampler was used during training to ensure that each batch contains a more uniform distribution of classes, thereby improving the model's exposure to minority categories. Lastly, Focal Loss was applied as the objective function to further reduce the influence of well-classified examples and encourage the model to focus on harder, less frequent samples.

2.2.1 Data Augmentation

To address class imbalance and enhance generalization, we apply a set of data augmentation techniques to the training set. These augmentations introduce appearance variability while preserving semantic content. Spatial transformations include random short side scaling (resizing the shorter side to 256-320 pixels), horizontal flipping (p=0.5), and random rotations (up to 15°) to simulate viewpoint variation. A final random resized crop adjusts the scale and aspect ratio to emulate different framing conditions. During development, attention heatmaps revealed that the model often focused on background textures instead of surgical tools likely due to tools being partially visible and background occupying most of the frame. To counteract this, we introduce a targeted augmentation step aimed at suppressing background salience. Each frame is converted from RGB to HSV color space, where two color thresholds (targeting red and pink hues typical of surgical backgrounds) are applied to generate a binary mask. Gaussian blur is selectively applied to masked background regions, preserving sharpness in likely tool containing areas. This enhances the visual prominence of the tools and guides the model's attention toward more informative regions. Figure 2 illustrates the complete augmentation pipeline.

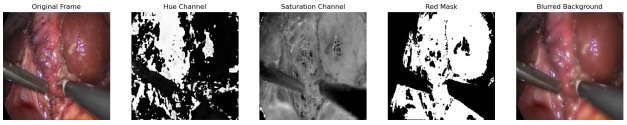

Fig. 2: Background masking pipeline.

2.2.2 Balanced Batch Sampling

We also implemented a custom balanced batch sampler. This approach ensures that each mini-batch contains a more uniform representation of all classes, rather than relying on the natural distribution of the dataset. Specifically, the sampler precomputes the set of indices corresponding to each class using existing annotations. During training, it selects an equal number of samples from each class to construct a batch, with the number of samples per class determined by the batch size and the total number of classes. If the batch size is not perfectly divisible, the remaining slots are filled with randomly sampled instances from the entire dataset to maintain batch consistency. This method improves the frequency at which the model encounters underrepresented classes, helping it to learn more balanced feature representations and reducing the bias toward majority classes.

2.2.3 Focal Loss

For training we used the Focal Loss function, which is designed to focus more on misclassified examples while reducing the loss contribution from well-classified samples. This is particularly useful in imbalanced datasets, where the model may otherwise become overly confident on the major-

ity classes, ignoring the minority classes. Focal Loss modifies the standard binary cross-entropy loss by introducing a modulating factor that downweights easy examples.

2.3 Model Architecture

The backbone of our approach is the TimeSformer model, which was originally introduced and pre-trained by researchers at Facebook for video classification tasks. The original version of TimeSformer was designed for single-label classification. Initially, the This model pre-trained on the Kinetics-600 (K600) dataset, a large-scale video benchmark that comprises over 500,000 video clips across 600 human action classes. Pre-training on such a vast dataset enables the model to learn robust general-domain features and representations including common visual patterns and object characteristics from sequences of images. To adapt it to our use case, we modified the model to support multi-label, multi-class classification.

TimeSformer is built upon the Vision Transformer (ViT) architecture [5], and it processes video inputs in the form of multiple image frames. In our setup, we input a sequence of 8 frames, each of spatial resolution 224×224. These frames are first divided into non-overlapping patches, which are then flattened into a sequence of tokens. Each token corresponds to a specific spatial location within a frame and is embedded using a learnable linear projection. Additionally, positional embeddings are added to retain spatial information across the sequence.

Among the several attention mechanisms proposed in the original TimeSformer paper, we adopt the Divided Space-Time Attention variant, where attention is applied sequentially—first across temporal dimensions and then across spatial dimensions. This separation enhances computational efficiency while still allowing the model to capture spatiotemporal dynamics effectively. The model employs 12 self-attention heads and consists of a stack of transformer encoder blocks.

To adapt the architecture for our multi-label classification task, we removed the original final classification layer and replaced it with a custom classification head. Specifically, the 768-dimensional feature vector output from the transformer is passed through a fully connected layer that reduces the dimensionality to 512, followed by a ReLU activation and dropout. Finally, another fully connected layer maps the 512-dimensional vector to a 7-dimensional output, corresponding to our 7 target classes:

2.4 Implementation Details

All components of the training pipeline, including data augmentation and model adaptation, were implemented using the PyTorch and Torchvision libraries. Fine-tuning of the TimeSformer model was conducted over a total of 5 epochs, starting with an initial learning rate of 1×10^{-5} on four Nvidia GTX 1080 Ti GPUs. Initially, all layers of the TimeSformer backbone were frozen, with only the newly added fully connected classification head being trainable. This setup was maintained for the first two epochs. Gradually, layers of the backbone were unfrozen in subsequent epochs. This strategy—common to transfer learning—helps to preserve useful pre-trained representations while avoiding large updates to sensitive early layers before the classification head has stabilized.

2.5 Evaluation Metrics

To evaluate the performance of our multi-label classification model, we report several key metrics: overall accuracy, macro-averaged precision, recall, and mean average precision (mAP). Macro-averaging treats all classes equally. In addition to these global metrics, we compute per-class Average Precision (AP). The mAP is obtained by averaging the AP scores across all classes, offering a comprehensive indicator of performance across the label space.

3 Results & Discussion

The final model achieved a macro-averaged precision of 98.9% and a recall of 97.5%. These scores indicate that the model is highly effective at identifying relevant labels (high precision), while also maintaining a strong ability to retrieve most of the true positive instances across all classes (high recall). Among all tool classes, the highest AP was observed for the *Hook* class, suggesting that this class is consistently and confidently identified by the model. Notably, minority classes such as *Scissors* and *Specimen-Bag* also performed well, with AP scores of 96.5% and 96.2% respectively. These results highlight the effectiveness of our data augmentation strategy and the use of balanced sampling, which helped mitigate the challenges of class imbalance.

Table 1 presents a comparative evaluation of our model against other baseline architectures. Transformer-based models that incorporate spatial and spatio-temporal attention mechanisms outperform traditional convolutional networks in some classes. Moreover, improvements in minority class performance further support the contribution of our augmentation pipeline and the integration of spatial features, demonstrating

that the model benefits not only from temporal context but also from spatial structure and balanced data exposure.

Tool	MTRC Net [6]	Nwoye [7]	Jalal et al. [8]	ViT [3]	Our Model
Grasper	84.7	**99.7**	91.0	91.6	94.4
Bipolar	90.1	95.6	97.3	**99.7**	96.8
Hook	95.6	99.8	**99.8**	97.3	97.3
Scissors	86.7	86.9	90.3	92.4	**96.5**
Clipper	89.8	**97.5**	97.4	95.8	96.7
Irrigator	88.2	74.7	95.6	**96.3**	96.1
Specimen-Bag	88.9	96.1	**98.3**	97.7	96.2
Mean (mAP)	89.1	92.9	95.6	95.8	**96.3**

Tab. 1: Tool performance comparison (mAP values)

To further interpret the model's predictions, we visualized attention heatmaps for all tool classes. As shown in Figure 3, the attention consistently follows the movement of the tool across the 8-frame input sequence, indicating effective spatio-temporal modeling. This visualization provides insight into how the model distributes its focus over time and space, and supports the effectiveness of our augmentation and training strategies in guiding attention toward the relevant tool regions.

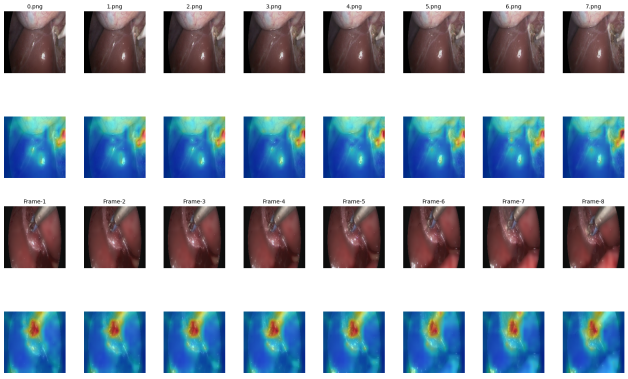

Fig. 3: Visualization of attention heatmaps.

4 Conclusion

Building on our previous research with vision transformers for laparoscopic surgical tool classification, this work explores the extension of such models into the spatio-temporal domain. We investigated the use of a spatio-temporal vision transformer architecture to capture both spatial and temporal features critical for accurate tool recognition in surgical videos. To address class imbalance, we incorporated several strategies, including targeted data augmentation, balanced batch sampling, and the use of focal loss. Our model achieved a mean Average Precision (mAP) of 96.3%, outperforming both conventional approaches and spatial-only vision transformer baselines. future work will explore the potential use of this model for unsupervised localization of surgical tools, which could further enhance its utility in real-time surgical assistance systems.

References

[1] Y. Jin et al., "Multi-task recurrent convolutional network with correlation loss for surgical video analysis," Medical Image Analysis, vol. 59, p. 101572, Oct. 2019, doi: 10.1016/j.media.2019.101572.

[2] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention all you need for video understanding?," International Conference on Machine Learning, pp. 813–824, Jul. 2021, [Online]. Available: http://proceedings.mlr.press/v139/bertasius21a/bertasius21a.pdf

[3] H. E. Moaqet, R. Janini, T. A. Alshirbaji, N. A. Jalal, and K. Möller, "Using vision transformers for classifying surgical tools in computer aided surgeries," Current Directions in Biomedical Engineering, vol. 10, no. 4, pp. 232–235, Dec. 2024, doi: 10.1515/cdbme-2024-2056.

[4] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "EndoNet: a deep architecture for recognition tasks on laparoscopic videos," IEEE Transactions on Medical Imaging, vol. 36, no. 1, pp. 86–97, Jul. 2016, doi: 10.1109/tmi.2016.2593957.

[5] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv.org, Oct. 22, 2020. https://arxiv.org/abs/2010.11929

[6] Y. Jin et al., "Multi-task recurrent convolutional network with correlation loss for surgical video analysis," Medical Image Analysis, vol. 59, p. 101572, Oct. 2019, doi: 10.1016/j.media.2019.101572.

[7] C. I. Nwoye, D. Mutter, J. Marescaux, and N. Padoy, "Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos," International Journal of Computer Assisted Radiology and Surgery, vol. 14, no. 6, pp. 1059–1067, Apr. 2019, doi: 10.1007/s11548-019-01958-6.

[8] N. A. Jalal et al., "Laparoscopic video analysis using Temporal, Attention, and Multi-Feature Fusion Based-Approaches," Sensors, vol. 23, no. 4, p. 1958, Feb. 2023, doi: 10.3390/s23041958.