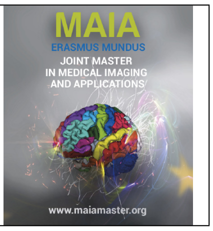
Medical Imaging and Applications

Master Thesis, July 2024



MAIA
ERASMUS MUNDUS
JOINT MASTER
IN MEDICAL IMAGING
AND APPLICATIONS

www.maiamaster.org

Deep Spatiotemporal Models for the Assessment of Operative Difficulty in Laparoscopic Cholecystectomy Videos

Leonardo Pestana Legori[a], Saurav Sharma[a], Mario Scaglia[b], Maria Vannucci[c,d], Giovanni Guglielmo Laracca[e], Sergio Alfieri[f,g], Pietro Mascagni[d,f,g], Nicolas Padoy[a,d]

[a]*ICube, University of Strasbourg, CNRS, France*
[b]*Università degli Studi di Milano*
[c]*General Surgery Department, University of Torino, Turin, Italy*
[d]*IHU Strasbourg, Strasbourg, France*
[e]*Department of Medical Surgical Science and Translational Medicine, Sant'Andrea Hospital, Sapienza University of Rome, Rome, Italy*
[f]*Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy*
[g]*Università Cattolica del Sacro Cuore, Rome, Italy*

Abstract

Gallstone disease is a common diagnosis that can lead to life-threatening conditions if untreated. Laparoscopic cholecystectomy (LC) is the gold standard procedure for gallbladder removal, and, despite being safer than open surgery, major complications can still arise, leading to decreased patient survival and significant healthcare costs. The risks of complications are correlated with preoperative and intraoperative findings and, ultimately, the operative difficulty. Prediction of the LC operative difficulty (LCOD) could reduce the risk of adverse events by stratifying patients and assigning surgeons with the relevant skills. At the same time, there is a need to identify objective, clinically meaningful, and operator-independent definitions of LCOD. With that in mind, this study investigates deep spatiotemporal models for predicting LCOD in surgical videos, utilizing a novel dataset of 100 LC surgeries annotated with intraoperative features indicative of operative difficulty. We introduce spatiotemporal prediction pipelines that employ state-of-the-art deep learning architectures for both spatial and temporal sequence modeling. Our results demonstrate that spatiotemporal models enhance prediction performance compared to spatial-only models. These findings underscore the importance of temporal context in surgical video analysis and highlight the potential for improved intraoperative decision-making.

Keywords: Surgical workflow analysis, Laparoscopic surgery, Cholecystectomy, Deep learning, Transfer learning, Spatiotemporal modeling

1. Introduction

Gallstone disease is a common diagnosis in developed countries, especially among female and older populations (Russo et al., 2004; Shaffer, 2005, 2006). While most patients with gallstones are asymptomatic, nearly one in ten may develop symptoms or complications that require treatment (Halldestam et al., 2004). If untreated, gallstones can develop into life-threatening conditions such as acute cholecystitis, cholangitis, obstructive jaundice, and acute pancreatitis (Kanakala et al., 2011).

Laparoscopic cholecystectomy (LC) is a minimally invasive surgical procedure for gallbladder removal, which replaced open surgery to become the standard treatment for symptomatic gallstone disease (Bittner, 2004). Compared to open surgery, LC reduces hospitalization, recovery time, and complications (Barkun et al., 1993). Despite its low operative morbidity and mortality rates (Pucher et al., 2018), major complications can still occur during an LC operation (Deziel et al., 1993; Pucher et al., 2018). One notorious example is bile duct injury, which is associated with adverse patient survival (Törnqvist et al., 2012, 2009) and significant costs to the healthcare system (Andersson et al., 2008; Halle-Smith et al., 2019).

The risk of complications in LC has been correlated to preoperative (Kanakala et al., 2011; Murphy et al., 2010; Terho et al., 2016) and intraoperative (Kanakala et al., 2011; Terho et al., 2016) factors, which can impact the operative difficulty (Hussain, 2011). Prediction of the operative difficulty in LC could reduce the risk of adverse events by stratifying patients and assigning surgeons with the relevant skills (Griffiths et al., 2019).

Recently, a systematic review by Vannucci et al. (2022) analyzed different definitions of LC operative difficulty (LCOD) in the literature and current statistical models used for the preoperative prediction of these attributes. The study observes that there is no clear consensus on the definition of LCOD and that some definitions, such as operating time and intraoperative events, are biased toward surgeons' skills. To remedy that, the study suggests using video-based assessment of LC videos annotated with scales defining intraoperative findings to identify objective, clinically meaningful, and operator-independent definitions of LCOD. More specifically, it suggests using artificial intelligence systems to automate, increase prediction performance, and facilitate using best-identified models.

To the best of our knowledge, this is still an unexplored area in literature. To address this gap, clinical partners have built a large dataset with 100 videos annotated with LC-specific intraoperative features based on difficulty scales available in the literature, with the goal of investigating meaningful and operator-independent definitions of LCOD.

Working with this novel dataset, the present study aims to predict intraoperative features related to operative difficulty in LC videos. To accomplish that, we implement deep learning models for LCOD prediction and explore approaches for incorporating time series information through spatiotemporal models to enhance prediction. This research aims to facilitate the development of clinical tools for intraoperative surgical analysis and planning in LC, ultimately improving patient outcomes and surgical efficiency.

2. State of the art

2.1. Surgical workflow analysis in LC

Surgical workflow analysis (SWA) research aims to develop context-aware systems to support surgeons' decision-making and improve the safety and effectiveness of surgical procedures. Significant progress has been made in the field in recent years, with the yearly number of publications in the field steadily increasing following breakthroughs in deep learning (Demir et al., 2023).

This has been addressed in LC by proposing methods for different tasks. Twinanda et al. (2017) introduces EndoNet, a convolutional neural network (CNN) specifically designed for phase recognition and tool presence detection in laparoscopic videos. Unlike previous methods that rely heavily on handcrafted features or tool usage signals obtained through manual annotation or additional equipment, EndoNet uniquely leverages visual information directly from surgical videos. The authors propose a multi-task CNN architecture that jointly performs phase recognition and tool presence detection, achieving state-of-the-art results in both tasks. To validate their approach, they created the Cholec80 dataset, consisting of 80 cholecystectomy videos annotated for surgical phases and tool presence.

More recent studies address the action triplet recognition task, an approach for providing fine-grained and comprehensive insights into surgical activities. The study by Nwoye et al. (2022) introduced a novel framework, Rendezvous (RDV), for the recognition of surgical action triplets, which include instrument, verb, and target combinations, from endoscopic video data. This model leverages a Class Activation Guided Attention Mechanism (CAGAM) to identify verbs and targets based on instrument activations and a Multi-Head of Mixed Attention (MHMA) to capture complex semantic relationships between the triplet components.

Building on top of RDV, Sharma et al. (2023a) proposes Rendezvous in Time (RiT), which leverages temporal cues from earlier frames to improve the recognition of surgical action triplets from videos. By focusing on the verb component of triplets and employing a temporal attention mechanism, RiT achieves smoother predictions and enhanced accuracy compared to previous approaches.

In Sharma et al. (2023b), the authors propose a novel two-stage network, MCIT-IG (Multi-Class Instrument-aware Transformer - Interaction Graph), which leverages instrument spatial information and target embeddings to improve triplet detection. The MCIT stage learns class-wise embeddings of the targets, while the IG stage constructs a bipartite dynamic graph to model interactions between instruments and targets. Their mixed-supervised learning strategy demonstrates improved performance on both instrument localization and triplet detection.

The study by Mascagni et al. (2022) presents a novel approach using deep learning algorithms to automatically segment hepatocystic anatomy and assess the achievement of the critical view of safety (CVS) during LC, which requires accurate identification of key anatomical structures and their geometric relationships. By analyzing still images from LC videos, the authors trained a deep neural network to recognize key anatomical structures, such as the gallbladder, cystic duct, and cystic artery. Additionally, the model predicts CVS criteria achievement, which is crucial for preventing bile duct injuries.

In Murali et al. (2024), the authors propose a novel method for CVS prediction that leverages a disentangled latent graph representation, explicitly encoding semantic information and visual features. Their approach outperforms several baseline methods, even when trained with bounding box annotations, and maintains state-of-the-art performance when trained with segmentation masks.

At the moment, though, the prediction of LCOD in SWA remains an unexplored area.

2.2. *Transfer learning*

Transfer learning is a machine learning technique where a pre-trained model, trained on a large dataset, is utilized as the foundation for a new, related task. This method significantly reduces training time and improves performance, especially when the new task has limited data. The pre-trained model's learned features are either fine-tuned or used for feature extraction on the new dataset, enabling better generalization and efficiency. Transfer learning is widely applied in domains such as image recognition, natural language processing, and speech recognition, demonstrating its efficacy in leveraging existing knowledge to enhance model performance on specific, data-constrained tasks.

Transfer learning has been extensively used in SWA to improve model performance (Mascagni et al., 2022; Murali et al., 2024; Nwoye et al., 2022; Sharma et al., 2023a; Twinanda et al., 2017). Pre-trained models, such as those trained on large image datasets like ImageNet (Deng et al., 2009), are fine-tuned on surgical data to enhance their capability to recognize and classify surgical phases and actions.

2.3. *Self-supervised learning*

Self-supervised learning (SSL) is a subset of machine learning where the model learns to generate labels from the data itself rather than relying on manually labeled datasets. This approach leverages inherent structures and patterns within the data to create proxy tasks, enabling the model to learn useful representations. SSL is especially useful in scenarios where labeled data is scarce or expensive to obtain, as it exploits large amounts of unlabeled data to pre-train models, which can then be fine-tuned on smaller labeled datasets. Applications of SSL span various domains, including computer vision, where it aids in tasks like image classification and object detection, and natural language processing, where it enhances language models by learning from raw text. This method significantly reduces the dependency on extensive labeled datasets, making it a valuable tool for advancing machine learning capabilities in data-constrained environments.

Momentum Contrast (MoCo), introduced by He et al. (2020), is a framework for unsupervised visual representation learning. MoCo formulates contrastive learning as a dictionary look-up problem, maintaining a dynamic dictionary through a queue and a momentum-updated encoder. The queue enables the use of a large set of negative samples, decoupling the dictionary size from the mini-batch size, and the momentum update ensures that the encoder evolves slowly, maintaining consistency in the representations. MoCo demonstrates competitive results on ImageNet classification and superior transferability to downstream tasks such as object detection and segmentation, often outperforming supervised pre-training counterparts.

Building on this, Chen et al. (2020b) proposed MoCo v2, which integrates a multilayer perceptron (MLP) projection head and stronger data augmentation into the MoCo framework. These enhancements are inspired by the SimCLR framework (Chen et al., 2020a). These modifications, while simple, significantly enhance the quality of the learned representations. MoCo v2 achieves superior results across various tasks without the need for large training batches, making it more computationally efficient and accessible. With these improvements, MoCo v2 establishes new baselines in unsupervised learning, outperforming SimCLR in image classification and object detection tasks.

In recent years, the application of self-supervised learning (SSL) methods in SWA has gained significant interest due to their potential to improve performance in label-scarce scenarios. In the study by Ramesh et al. (2023), various SSL techniques were evaluated on the Cholec80 dataset, focusing on tasks such as surgical phase recognition and tool presence detection. Among the methods tested, MoCo v2 consistently outperformed other SSL approaches, particularly in low-label settings, demonstrating up to 6.1% improvement in single-frame phase recognition and 14.7% in tool presence detection with minimal labeled data. This indicates MoCo v2's robustness and adaptability to surgical contexts, reinforcing its utility in scenarios where labeled data is limited.

2.4. *Temporal modeling*

Temporal modeling is essential for analyzing sequential data, such as surgical videos, where the temporal context provides critical information about the evolution and transitions of surgical tasks. By capturing the sequence and timing of events, temporal modeling helps in understanding the dynamic interactions and changes that occur throughout the procedure. This approach allows for a more accurate identification of phase transitions and the recognition of patterns that are crucial for effective surgical analysis and intervention planning.

Twinanda (2017) proposes a two-step approach that integrates long short-term memory (LSTM) networks for surgical phase recognition. This approach extends the EndoNet framework by first extracting visual features using CNNs and then employing LSTM networks to capture temporal dependencies across these features. This method aims to address the limitations of hierarchical hidden Markov models (HHMM) by leveraging LSTMs to model long-term dependencies more effectively. The experimental results, particularly on the Cholec80 dataset, demonstrate that EndoLSTM outperforms HHMM-based EndoNet in both offline and online phase recognition scenarios, highlighting its efficacy in improving surgical phase detection accuracy and reliability.

Czempiel et al. (2020) introduced TeCNO, a novel approach leveraging Multi-Stage Temporal Convolutional Networks (MS-TCNs) (Farha and Gall, 2019) for hierarchical prediction refinement in surgical workflow analysis. Their method utilizes causal, dilated convolution refined in a multi-stage architecture to enable large receptive fields and smooth online inference during ambiguous transitions, outperforming various LSTM models on laparoscopic cholecystectomy datasets. This approach, which separates feature extraction and temporal refinement stages, addresses the limitations of capturing long-term dependencies inherent in LSTM-based methods, paving the way for more robust and efficient surgical phase recognition.

In recent years, the integration of Transformers in computer vision tasks has garnered significant attention, particularly for surgical workflow analysis. Jin et al. (2022) introduced Trans-SVNet, a hybrid embedding aggregation Transformer designed to enhance real-time surgical workflow analysis by effectively integrating spatial and temporal features. Unlike conventional methods that sequentially encode spatiotemporal features, Trans-SVNet utilizes a Transformer-based approach to jointly consider spatial and temporal embeddings, thereby preserving critical intermediate features and improving workflow recognition and anticipation accuracy. Evaluated on three large surgical video datasets, the proposed method demonstrated superior performance in both recognition and anticipation tasks.

2.5. Contributions

Building on the foundation of advancements in SWA for LC, this study's contributions are:

- *Study of operative difficulty in LC:* we perform a study of deep learning architectures on the assessment of operative difficulty in LC videos (LCOD prediction).

- *Exploration of spatial-only models:* we investigate spatial-only models for LCOD prediction, focusing on methods that rely on visual information from surgical videos. These models will serve as a baseline for understanding the capacity of spatial features alone in predicting operative difficulty.

- *Integration of temporal information*: recognizing the importance of temporal context in surgical video analysis, we explore various approaches for incorporating temporal information into our models.

- *Transfer learning with SSL*: we employ transfer learning techniques using SSL pre-trained models within the same domain. Specifically, we leverage models from Ramesh et al. (2023), which were pre-trained on the Cholec80 dataset using MoCo v2, which has shown superior performance in low-label settings for tasks such as surgical phase recognition and tool presence detection.

3. Material and methods

This section presents the details of our dataset, the methodology employed, the deep learning models used, the experiments planned, the evaluation metrics, and the implementation details.

3.1. Dataset

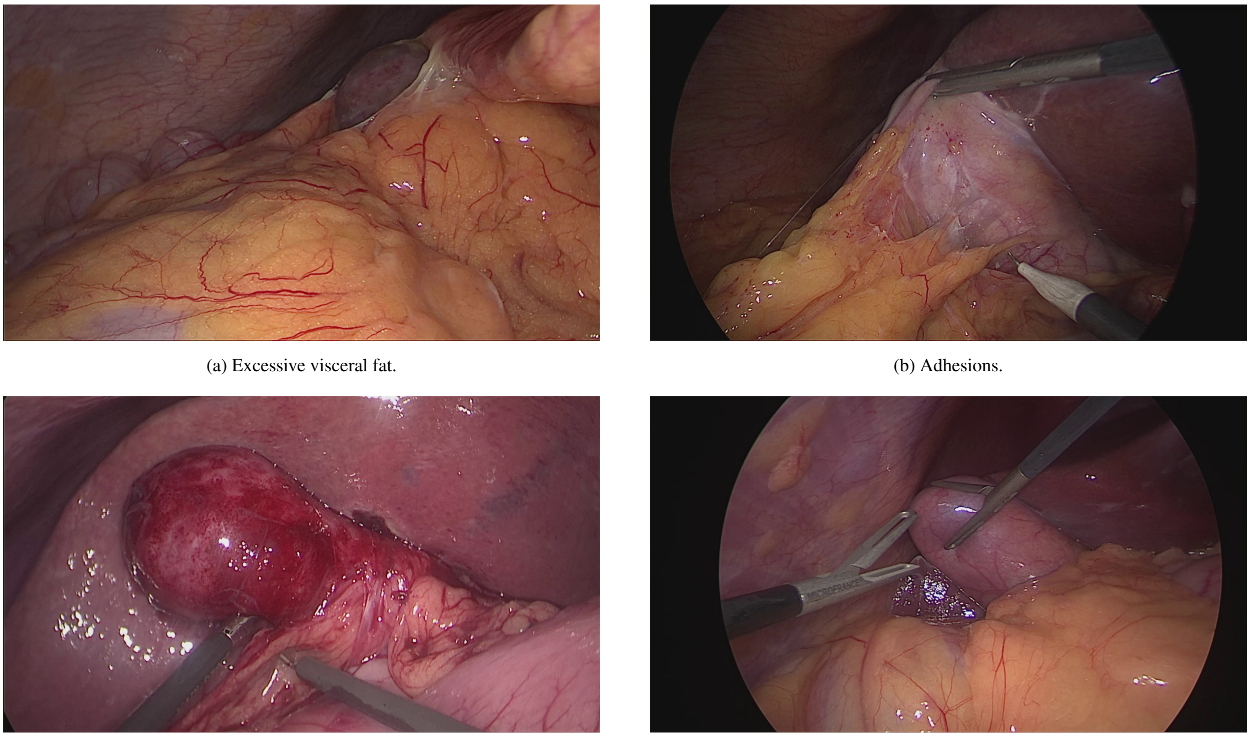
For this work, we use a novel dataset containing 100 videos of LC surgeries of varying difficulty performed at IHU Strasbourg. The dataset was constructed by selecting 25 videos per quartile from a database of 326 LC videos ranked according to operation duration, intended to act as a surrogate of LCOD. The videos were captured at 25 frames per second (FPS) and downsampled to 1 FPS for processing.

The dataset consists of 90 LCOD features annotated with binary presence information per surgical phase based on five intraoperative difficulty scales: Cuschieri (Hanna et al., 1998), Nassar (Nassar et al., 1995), Sugrue (Sugrue et al., 2015), Parkland grading scale (PGS) (Madni et al., 2018), and Iwashita (Iwashita et al., 2017). The annotation was performed independently by three raters with different levels of surgical expertise. The annotations from three raters are uniformly averaged to obtain the final value. The annotations are multi-label in nature, where a phase may have multiple LCOD features present simultaneously.

Table 1: Surgical phases observed in the dataset and their mean duration.

ID	Phase	Duration (s)
P1	Trocar Placement and Preparation	798 ± 1059
P2	Hepatocystic Triangle Dissection	1085 ± 967
P3	Clipping and Cutting	255 ± 617
P4	Gallbladder Bed Dissection	621 ± 455
P5	Gallbladder Packaging, Extraction, Cleaning, and Coagulation	748 ± 550
P6	Subtotal Cholecystectomy	726 ± 369

Table 1 reports the phases observed in the 100 videos and their average duration computed across videos. The



(a) Excessive visceral fat.

(b) Adhesions.

(c) Hyperemia/inflammation of the gallbladder.

(d) Distended gallbladder.

Figure 1: Representative video frames showcasing examples of annotated LCOD features from the dataset.

subtotal cholecystectomy phase (P6), a bailout procedure, is observed in only four videos.

In this study, we include only LCOD features that appear in at least 8 videos, resulting in 20 features covering the first four phases. To balance the proportion of LCOD features in training, validation, and testing sets, we apply multi-label stratified sampling (Sechidis et al., 2011) to the dataset, resulting in 52 training, 10 validation, and 38 testing videos.

Table 2 contains the final list of LCOD features and their number of occurrences in each set. Figures 1a to 1d show examples of LCOD features and their representative frames, highlighting the anatomical variations observed for the features.

We concentrate on the first two phases, as the identification of LCOD assessment features in these phases can assist surgeons in managing the complexity of the surgical procedure, such as by assigning a more experienced surgeon. Additionally, most of the LCOD features are annotated in the first two phases compared to other phases.

3.2. Methodology

We formulate LCOD prediction as a multi-label classification task performed for each surgical phase separately, ensuring that each phase is independently modeled for better prediction. We employ two main deep learning-based pipelines: the first utilizes state-of-the-art models that analyze individual frames for LCOD classification, and the second leverages temporal models to enhance predictions by capturing temporal signals. Additionally, we conduct multi-class phase recognition using our dataset and compare the performance across various model initializations. Figure 2 illustrates these pipelines.

Spatial-only pipeline. We implement the architecture described in Figure 2a. Each video frame is input to a deep learning model for visual feature extraction, then to a fully connected layer, followed by sigmoid activation to generate a multi-label probability score. To produce a single result for the phase, we average all predictions for each video to generate a video-wise prediction score.

In addition to LCOD feature prediction, we train and evaluate spatial models for phase recognition in our dataset, as shown in Figure 2b. We use the same method described above, with a few key differences. After the features go through the fully connected layer, we perform a softmax operation to obtain multi-class probabilities for each class. In this case, the video-wise averaging of the probabilities is not performed.

In this study, the term spatial-only model is used interchangeably to refer to both the deep learning model used for feature extraction and the whole pipeline described above.

Spatiotemporal pipeline. We implement the method described in Figure 2c. To predict LCOD features, we want to use all the frames in each phase to cap-

Table 2: List of annotated LCOD features and their occurrence across training, validation, and testing sets.

Phase ID	Feature	# Occurrences			
		Train	Validation	Testing	Total
P1	Excessive visceral fat	39	5	29	73
	Fat-laden falciform ligament	37	7	32	76
	Adhesions	37	5	29	71
	Adhesions limiting access to the abdominal cavity	9	2	8	19
	Liver hypertrophy preventing gallbladder retraction	4	1	3	8
	Floppy, non adherent gallbladder	25	5	13	43
	Hyperemia/inflammation of the gallbladder	10	1	8	19
	Distended gallbladder	12	1	8	21
P2	Hepatocystic triangle visible	19	5	16	40
	Clear, thin cystic plate	8	4	9	21
	Fat-laden cystic plate	36	5	27	68
	Dilated cystic plate	9	3	8	20
	Cystic duct seen with retraction	12	5	11	28
	Critical view of safety: two structures	34	5	26	65
	Critical view of safety: hepatocystic triangle	39	8	31	78
	Critical view of safety: cystic plate	20	5	25	50
P3	Gallbladder packed with stones	5	1	4	10
P4	Easy bleeding at dissection of gallbladder fossa	6	1	5	12
	Edematous changes in gallbladder fossa	14	2	12	28
	Fibrotic gallbladder fossa	11	3	10	24
	All features	**386**	**74**	**314**	**774**

ture temporal information. However, using the original frame images as inputs is unfeasible due to memory constraints. As a workaround, we implement a two-stage approach. First, we generate and store a spatially pooled representation of each frame using a pre-trained model as a visual feature extractor. This work uses pre-trained models on phase recognition tasks for feature extraction. In the next stage, the features of all the video frames are concatenated together in sequential order and input to a temporal model for refinement. The refined features are passed to a fully connected layer, followed by sigmoid activation, to generate a final multi-label prediction score for a given video.

In this study, the term spatiotemporal model is used interchangeably to refer to both the model architecture used for temporal refinement and the whole pipeline described above.

3.3. Spatial-only deep learning architectures

This section describes the deep learning architectures utilized for feature extraction in the spatial-only pipeline. These models are designed to capture spatial features from individual video frames.

ResNet. Introduced by He et al. (2016), the residual network (ResNet) is a CNN that utilizes residual skip connections to mitigate the vanishing gradient problem, enabling the construction of deeper networks. ResNet's architecture comprises multiple residual blocks, each containing two or more convolutional layers and a skip connection. This skip connection adds the input of the block to its output, effectively enabling the network to learn identity mappings that preserve gradient flow. The key innovation here is the ability to train deep networks that retain high performance on complex tasks. Our experiments use ResNet-18 and ResNet-50 variants, with 18 and 50 layers, respectively.

EfficientNetV2. EfficientNetV2, introduced by Tan and Le (2021), represents a family of CNNs that optimize both accuracy and efficiency through a systematic scaling approach. Unlike traditional models that arbitrarily scale dimensions (depth, width, and resolution), EfficientNetV2 uses a compound scaling method. This method uniformly scales all dimensions in a principled manner, balancing the trade-offs between network size and performance. EfficientNetV2 builds on the success of its predecessor, EfficientNet (Tan and Le, 2019), by introducing several enhancements, such as improved training speed and parameter efficiency. The network employs advanced techniques like Fused-MBConv, which merges depthwise and pointwise convolutions for faster training, and progressive learning, which gradually increases image size during training to stabilize the learning process. We employ the EfficientNetV2-XL (EffNetV2-XL) variant in our ex-

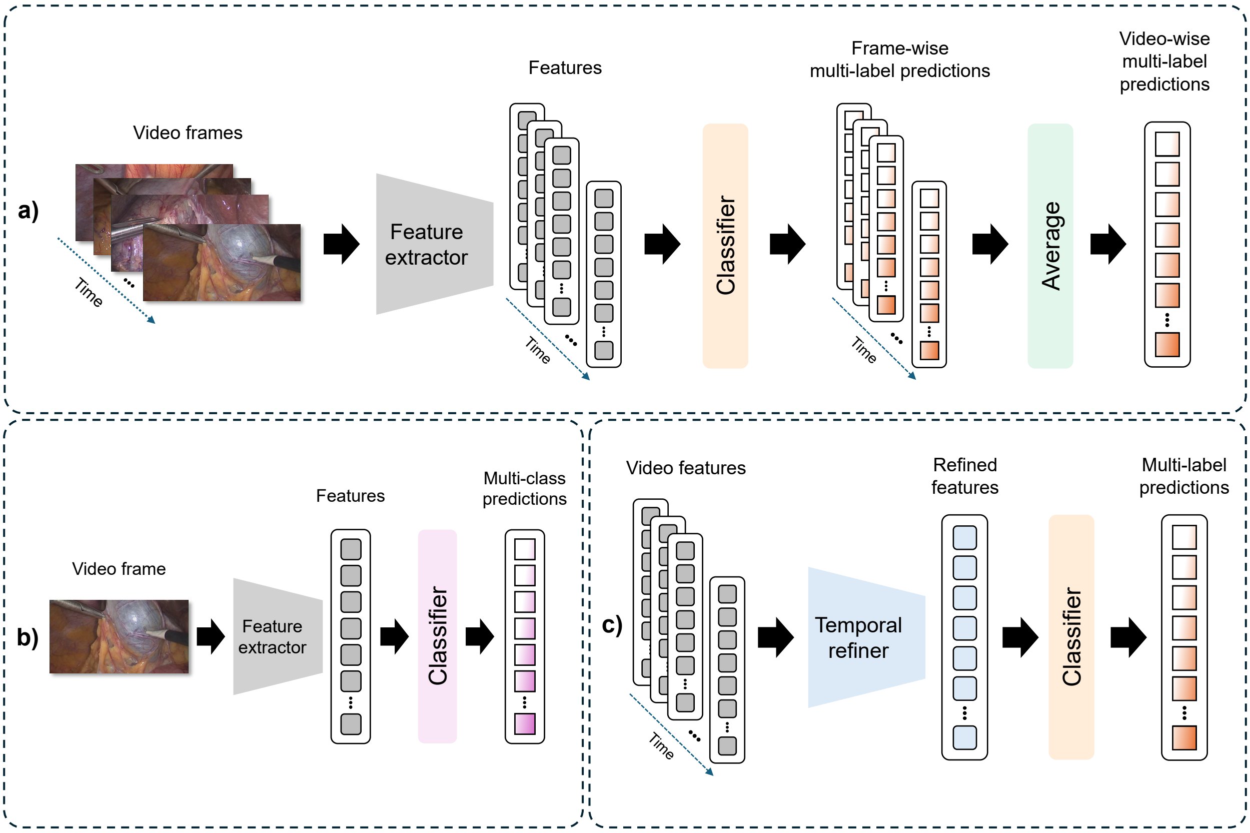


Figure 2: Illustration of methods explored. a) Pipeline using spatial-only pre-trained models for LCOD prediction. b) We modify the pipeline in a) to perform phase recognition. c) Pipeline for LCOD prediction using features extracted in b) for temporal refinement using temporal models.

periments, the largest model in the EfficientNetV2 family.

Swin Transformer. The Swin Transformer, proposed by Liu et al. (2021), introduces a novel hierarchical vision transformer architecture that efficiently processes high-resolution images by computing self-attention within local windows. Unlike traditional transformers that operate on global self-attention, the Swin Transformer restricts self-attention to non-overlapping windows, significantly reducing computational complexity and memory usage. One of the key innovations of this architecture is its ability to shift the windows across different layers. This shift mechanism allows cross-window connections, enabling the model to capture long-range dependencies and contextual information effectively. The hierarchical structure of the Swin Transformer consists of stages, each reducing the spatial resolution and increasing the number of feature channels, similar to CNNs. Our experiments use the Swin-S variant, a smaller variant that balances model complexity and performance.

3.4. Spatiotemporal deep learning architectures

This section describes the deep learning architectures utilized in the spatiotemporal pipeline. We use those models to refine the spatial features extracted from individual video frames.

RNN-based models. Recurrent neural networks (RNNs) are designed to handle sequential data by maintaining a hidden state that captures information from previous time steps. These models are effective in capturing short-term dependencies but may struggle with longer sequences due to issues like vanishing gradients (Bengio et al., 1994). Long short-term memory (LSTM), introduced by Hochreiter and Schmidhuber (1997), is a type of RNN architecture designed to address the vanishing gradient problem.

LSTMs achieve this by incorporating a memory cell capable of maintaining its state over long periods. Three gates regulate this memory cell: the input gate, which controls the flow of new information into the cell; the forget gate, which determines what portion of the past information to discard; and the output gate, which controls how much of the cell's state is used to compute the output. These gates are crucial in ensuring that the network can learn to retain information for long durations and decide what to forget and retain, thereby making LSTMs highly effective for tasks involving long-range temporal dependencies, such as language modeling and time-series prediction.

In contrast, gated recurrent units (GRUs), introduced by Cho et al. (2014), offer a simplified alternative to LSTMs by combining the input and forget gates into a single update gate and merging the cell state and hidden state. This streamlined design reduces the computational complexity and number of parameters, making GRUs more efficient to train while still effectively capturing dependencies in sequential data. The GRU architecture's update and reset gates work together to determine how much of the previous information to retain and how much of the new input to integrate. Although GRUs and LSTMs perform similarly on various tasks, GRUs are often preferred for their simplicity and faster training times.

In our study, we evaluate both LSTM and GRU models.

MS-TCN. The Multi-Stage Temporal Convolutional Network (MS-TCN), introduced by Farha and Gall (2019), is an advanced architecture for temporal action segmentation that leverages the power of dilated temporal convolutions across multiple stages to effectively capture and classify actions in long, untrimmed video sequences. Each stage within the MS-TCN comprises a series of dilated temporal convolutional layers specifically designed to handle video frames' temporal dependencies and sequential nature. These layers use dilation to expand the receptive field without increasing the number of parameters, allowing the network to encompass long-range temporal dependencies efficiently. At each stage, the network generates an initial prediction for action segments, which is subsequently refined by the next stage. This iterative refinement process helps progressively improve the accuracy of the action segmentation, ensuring that the final output is a highly refined representation of the action segments within the video. The MS-TCN can iteratively enhance its predictions by stacking multiple stages, effectively capturing complex temporal patterns and nuances within the video data.

Transformer. The Transformer model, introduced by Vaswani et al. (2017), transforms sequence transduction tasks by using a fully attention-based mechanism, eliminating the need for recurrent structures. The core of its architecture is the multi-head self-attention mechanism, which computes attention scores across all positions in an input sequence by projecting inputs into query, key, and value vectors and then calculating the weighted sum of values based on scaled dot-product attention. This allows the model to focus on different parts of the sequence simultaneously. Both the encoder and decoder consist of stacks of identical layers; encoder layers include multi-head self-attention followed by position-wise feed-forward networks, while decoder layers add another multi-head attention mechanism to attend to the encoder's output. Residual connections and layer normalization around each sub-layer help maintain stable training. To incorporate sequential information, the Transformer adds positional encodings to the input embeddings, using sine and cosine functions of varying frequencies. This design not only enhances parallelization but also achieves state-of-the-art performance on various tasks, such as machine translation, demonstrating its effectiveness and computational efficiency.

In our work, we leverage only the encoder part of the Transformer, implementing a similar architecture to the Vision Transformer (ViT) introduced by Dosovitskiy et al. (2020). Our model uses sequential features extracted from video frames as our input, fed directly to the Transformer encoder without positional encoding. Like the ViT, we include a classification token at the start of the sequence that interacts with the self-attention mechanism, enabling the model to aggregate information from the entire sequence for final classification. Using the encoder structure, we benefit from its powerful attention-based mechanism to capture dependencies across the video frames.

3.5. Experiments

Phase recognition. We evaluate spatial-only models on the phase recognition task. We use models pre-trained on the ImageNet (Deng et al., 2009) and one (ResNet-50) on the Cholec80 dataset using the MoCo v2 SSL method (Ramesh et al., 2023). For the latter, we evaluate linear probing and fine-tuning protocols.

LCOD prediction with spatial models. We evaluate spatial-only models on the LCOD prediction task. We use models pre-trained on the ImageNet and one (ResNet-50) on the Cholec80 dataset using the MoCo v2 SSL method. For the former, we evaluate linear probing and fine-tuning protocols.

LCOD prediction with spatiotemporal models. We evaluate spatiotemporal models on the LCOD prediction task. As a preliminary step, we extract compact features of all the video frames using the best-performing models in the phase recognition task. For this set of experiments, we use the optimal parameters determined in our ablation studies (Section 4.4).

3.6. Evaluation

The performance of all models is measured by average precision (AP), mean AP (mAP), and F_1 score metrics averaged per class.

Precision measures the proportion of true positive (TP) predictions from the total predicted positives. It is calculated as the ratio of true positives to the sum of true positives and false positives (FP), as seen in Equation 1.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

Recall, also known as sensitivity, measures the proportion of true positive predictions out of the actual positives. It is calculated as the ratio of true positives to the sum of true positives and false negatives (FN), as seen in Equation 2.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

The F_1 score, given by Equation 3, is the harmonic mean of precision and recall, providing a single score that balances both precision and recall.

$$F_1 = 2\frac{\text{Precision} \cdot \text{Recall}}{(\text{Precision}) + \text{Recall}} \tag{3}$$

The AP is a metric used to summarize the precision-recall curve, which shows the tradeoff between precision and recall for different thresholds. It calculates the weighted mean of precisions achieved at each threshold, using the increase in recall from the previous threshold as the weight. It is given by Equation 4, where P_n and R_n are the precision and recall at the nth threshold.

$$\text{AP} = \sum_n (R_n - R_{n-1})P_n \tag{4}$$

mAP, given by Equation 5, is a performance metric widely used in various machine learning tasks, particularly in object detection and multi-label classification. It calculates the average precision for each class and then computes the mean of those average precisions.

$$\text{mAP} = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{5}$$

mAP captures the model's performance across different thresholds, offering a detailed understanding of its behavior in terms of precision and recall. This may be particularly important in clinical settings, where the costs of false positives and false negatives can vary significantly. Unlike metrics that depend on a fixed threshold, mAP considers performance across all possible thresholds. This robustness is crucial in medical applications, where the optimal threshold might vary depending on the clinical scenario.

As such, mAP is selected as the main metric for this work due to its ability to effectively handle class imbalance, evaluate multi-label classifications, provide fine-grained performance insights, and offer robustness to threshold selection.

3.7. Implementation details

All the code is implemented in Python using PyTorch and MMEngine to handle the training, validation, and testing pipeline. We use the implementation of ResNet and Swin models provided by the MMDetetection framework and the implementation of EfficientNet by the MMPretrain framework. We use the implementation of GRU and LSTM models provided by PyTorch. The MS-TCN implementation is adapted from Czempiel et al. (2020). The Transformer encoder architecture used in this work is implemented from scratch.

All the experiments were performed on Nvidia V100 and RTX6000 GPUs provided by the University of Strasbourg.

3.7.1. Data preprocessing

For all the spatial experiments, a single data preprocessing pipeline is used for training, comprising of:

1. Resizing to 384x384 for CNN-based models and 224x224 for Swin-S model.
2. Random horizontal and vertical flip.
3. Random brightness, contrast, and hue modification by a factor chosen uniformly from the range $[0.7, 1.3]$.
4. Random hue modification by a factor chosen uniformly from the range $[-0.1, 0.1]$.

Only the resizing operation is applied for validation and testing. The same is true for extracting the features used in the spatiotemporal models.

3.7.2. Hyperparameters

Spatial-only models: phase recognition task. For phase recognition, we configured ResNet-18, ResNet-50, EffNetV2-XL, and Swin-S models with the AdamW optimizer (Loshchilov and Hutter, 2017). The base learning rate was set to 10^{-5} for ResNet-18, ResNet-50, and Swin-S, while a higher learning rate of 10^{-4} was used for EffNetV2-XL. The weight decay for all models was maintained at 10^{-6}. We utilized a batch size of 32 for ResNet-18, ResNet-50, and Swin-S, whereas EffNetV2-XL was trained with a batch size of 128. The learning rate schedule followed the one-cycle policy (Smith and Topin, 2017). The maximum number of epochs was set to 50, with early stopping based on the F_1 metric and a patience of 10 epochs.

Spatial-only models: LCOD prediction task. For LCOD prediction, the same set of spatial-only models was configured similarly with the AdamW optimizer. The base learning rates remained the same as for the phase recognition task, with 10^{-5} for ResNet-18, ResNet-50, and Swin-S, and 10^{-4} for EffNetV2-XL. The weight decay was 10^{-6} across all models. A batch size of 32 was used for ResNet-18, ResNet-50, and Swin-S, and 128 for EffNetV2-XL. The one-cycle learning rate schedule was applied with a reduced maximum epoch count of 30. Early stopping was guided by the mAP metric with a patience of 6 epochs.

Spatiotemporal models. All models used the AdamW optimizer. The GRU and LSTM models were trained with a base learning rate of 10^{-4}, whereas the MS-TCN and Transformer models utilized a slightly lower learning rate of 7×10^{-5}. Consistent with the spatial-only models, the weight decay was set at 10^{-6}. The one-cycle learning rate schedule was implemented, and the maximum number of epochs was 50 for all spatiotemporal models. The early stopping mechanism was based on the mAP metric, with a patience of 10 epochs.

Loss function: phase recognition task. For the training, we use weighted cross-entropy loss for the phase recognition task, which is given by the following formula:

$$\mathcal{L}_{\text{WCE}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} w_c\, y_{nc} \cdot \log(\hat{y}_{nc}) \tag{6}$$

where N is the number of samples, C is the number of classes, y_{nc} is the true label for the c-th class of the n-th sample (one-hot encoded), $\hat{y}_{nc}$ is the predicted probability for the c-th class of the n-th sample and w_c is the weight for the c-th class. We compute the class weights based on the inverse frequency of each class in the training subset. Let N be the total number of training samples, and n_c be the number of samples belonging to class c. The weight w_c for class c is calculated as:

$$w_c = \frac{N}{n_c} \tag{7}$$

Loss function: LCOD prediction task. For the training of the LCOD prediction task, we use the weighted binary cross-entropy with logits loss, which is given by:

$$\mathcal{L}_{\text{WBCE}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}\Big(p_c\, y_{nc} \cdot \log(\sigma(\hat{y}_{nc})) + (1 - y_{nc}) \cdot \log(1 - \sigma(\hat{y}_{nc}))\Big) \tag{8}$$

where N is the number of samples, C is the number of labels (classes), y_{nc} is the true label for the c-th class of the n-th sample (either 0 or 1), $\hat{y}_{nc}$ is the raw predicted logit for the c-th class of the n-th sample, $\sigma(\hat{y}_{nc})$ is the sigmoid function applied to the raw logit $\hat{y}_{nc}$, p_c is the positive class weight for the c-th class, used to handle class imbalance. The positive class weight p_c is calculated as the ratio of the number of points in the negative class to the number of points in the positive class in the training subset. Let n_{neg} denote the number of negative samples and n_{pos} denote the number of positive samples. The weight p_c is computed as:

$$p_c = \frac{n_{\text{neg}}}{n_{\text{pos}}} \tag{9}$$

4. Results

4.1. Phase recognition

We compare the spatial models' performance on the phase recognition task. In Table 3, we show that Swin-S is the best spatial model, with an F_1 score of 0.7075 and mAP of 0.7719.

We also show that the ResNet-50 model's performance increases when using fine-tuned SSL pre-trained weights compared with the one using fine-tuned ImageNet weights.

Table 3: Phase recognition results. LP: linear probing, FT: fine-tunning.

Pretraining	Model	F_1	mAP
	Random	0.1400	0.1667
ImageNet	ResNet-18	0.6575	0.7179
	ResNet-50	0.6806	0.7509
	EffNetV2-XL	0.6798	0.7464
	Swin-S	**0.7075**	**0.7719**
Cholec80 MoCo v2 SSL	ResNet-50 (LP)	0.5598	0.6007
	ResNet-50 (FT)	0.7004	0.7695

4.2. LCOD prediction in phase 1: Trocar Placement and Preparation

Table 4: Results for phase 1, for the spacial-only models (a) and the spatiotemporal models (b), showing their performance measured by F_1 and mAP metrics. LP: linear probing, FT: fine-tunning.

Pretraining	Model	F_1	mAP
	Random	0.4160	0.4434
	a) Spatial-only models		
ImageNet	ResNet-18	0.5071	0.5831
	ResNet-50	0.4855	0.5662
	EffNetV2-XL	0.5421	0.6335
	Swin-S	0.4613	0.6061
Cholec80 MoCo v2 SSL	ResNet-50 (LP)	0.5641	0.6153
	ResNet-50 (FT)	0.4968	0.5730
	b) Spatiotemporal models		
	ResNet-50 features		
ImageNet	GRU	0.3042	0.5479
	LSTM	0.3504	0.5561
	MS-TCN	0.3048	0.5935
	Transformer	0.2716	0.5929
Cholec80 MoCo v2 SSL	GRU	0.3853	0.5473
	LSTM	0.2899	0.5600
	MS-TCN	0.1922	0.6221
	Transformer	0.5342	0.6364
	EffNetV2-XL features		
ImageNet	GRU	0.4085	0.5945
	LSTM	0.5259	0.6305
	MS-TCN	0.2634	0.5093
	Transformer	**0.5698**	**0.6780**
	Swin-S features		
ImageNet	GRU	0.5043	0.5867
	LSTM	0.4524	0.5851
	MS-TCN	0.2690	0.6351
	Transformer	0.3186	0.5853

We compare the spatial-only and spatiotemporal models' performance on the LCOD prediction task for the Trocar Placement and Preparation phase, present in Table 4.

Spatial-only prediction. The EffNetV2-XL model shows remarkable performance, being the best of all ImageNet pre-trained models when looking at both metrics. The ResNet-50 pre-trained on the Cholec80 dataset also shows increased results. Unlike the phase recognition results, the linear probing experiment performs better in this one.

Spatiotemporal prediction. The Transformer model using visual features extracted with EffNetV2-XL performs the best. This particular model configuration also outperforms the best spatial-only model, with a 5.11% and 7.02% increase in the F_1 score and mAP, respectively. We also observe a correlation between the best spatial-only models' results and the quality of their visual features for temporal modeling, leading to improved spatiotemporal results.

4.2.1. Feature-wise analysis in phase 1

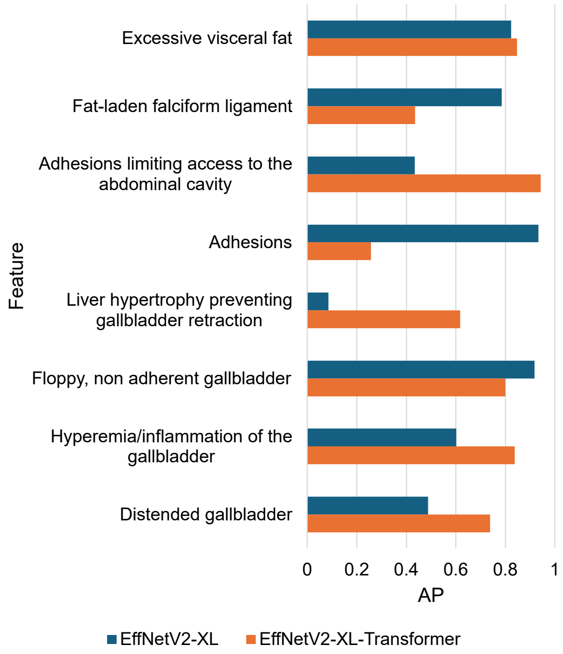


Figure 3: Feature-wise results of phase 1. The performance is evaluated using average precision (AP).

We compare the feature-wise results for the top-performing spatial-only and spatiotemporal models for the features in surgical phase 1. The results are in Figure 3, showing each feature's AP performance.

The results show that, of all eight features, the spatial model performs better when predicting the presence of fat-laden falciform ligaments, floppy, non-adherent gallbladder, and adhesions, the last of which have a predictive score many times higher than that of the best spatiotemporal model. In contrast, excessive visceral fat, adhesions limiting access to the abdominal cavity, liver hypertrophy preventing gallbladder retraction, inflammation of the gallbladder, and distended gallbladder are predicted better by the spatiotemporal model.

Overall, the spatiotemporal models show improved results in most features. Notably, a much better prediction of adhesions limiting access to the abdominal cavity and liver hypertrophy preventing gallbladder retraction is observed.

4.3. LCOD prediction in phase 2: Hepatocystic Triangle Dissection

Table 5: Results for phase 2, for the spacial-only models (a) and the spatiotemporal models (b), showing their performance measured by F_1 and mAP metrics. LP: linear probing, FT: fine-tunning.

Pretraining	Model	F_1	mAP
	Random	0.4393	0.4556
	a) Spatial-only models		
ImageNet	ResNet-18	0.5168	0.6622
	ResNet-50	0.4713	0.6472
	EffNetV2-XL	0.6310	0.7011
	Swin-S	0.6015	0.7089
Cholec80 MoCo v2 SSL	ResNet-50 (LP)	0.6394	0.7555
	ResNet-50 (FT)	0.5289	0.6877
	b) Spatiotemporal models		
	ResNet-50 features		
ImageNet	GRU	0.5798	0.7685
	LSTM	0.4192	0.7301
	MS-TCN	0.6187	0.7463
	Transformer	0.5571	0.7457
Cholec80 MoCo v2 SSL	GRU	0.6141	0.7687
	LSTM	**0.6526**	**0.7721**
	MS-TCN	0.5899	0.7507
	Transformer	0.5768	0.7169
	EffNetV2-XL features		
ImageNet	GRU	0.6324	0.7078
	LSTM	0.5996	0.7484
	MS-TCN	0.6015	0.7282
	Transformer	0.6093	0.6844
	Swin-S features		
ImageNet	GRU	0.5786	0.7556
	LSTM	0.5799	0.7630
	MS-TCN	0.5708	0.7325
	Transformer	0.4884	0.7041

We compare the spatial and temporal models' performance on the LCOD prediction task for the Hepatocystic Triangle Dissection phase, present in Table 5.

Spatial-only prediction. The best performant model in this phase is the ResNet-50 pre-trained on Cholec80 using a linear probing protocol. It outperforms the second-best model in terms of mAP (Swin-S), with a 6.57% increase. Compared to the ResNet-50 model pre-trained on ImageNet, we see a significant improvement in both metrics for the Chole80 MoCo v2 SSL pre-trained models. The increase is, in particular, higher for the linear probing experiment, with an increase of 35.67% in the F_1 score and 16.73% in the mAP.

Spatiotemporal prediction. We observe a similar situation as in the phase 1 results, with the best spatiotemporal model using the visual features extracted using the model with the best spatial performance. In this case, the temporal refinement using the LSTM-based model using visual features extracted by the Cholec80 pre-trained ResNet-50 model performs best, with an increase of 2.06% and 2.20% in the F_1 score and mAP, respectively.

4.3.1. Feature-wise analysis in phase 2

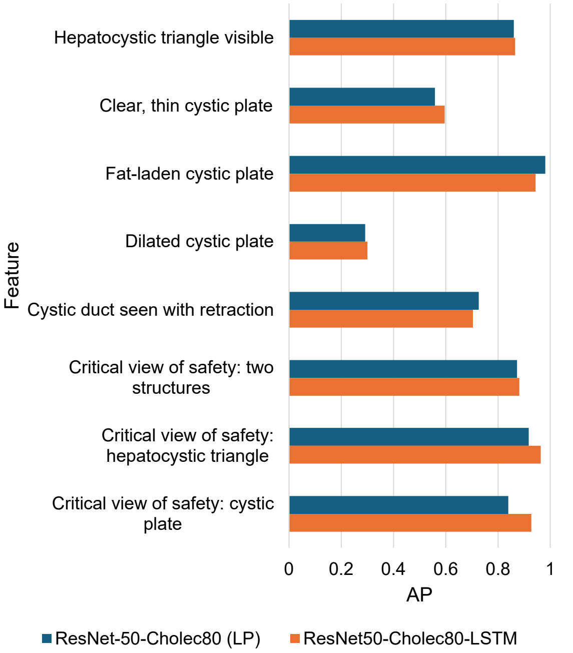


Figure 4: Feature-wise results of phase 2. The performance is evaluated using average precision (AP).

We compare the feature-wise results for the top-performing spatial-only and spatiotemporal models for the features in the Hepatocystic Triangle Dissection phase. The Figure 4 shows each feature's AP performance.

The results show similar overall performance when comparing the spatial-only and spatiotemporal models. Of all eight LCOD features, the spatiotemporal model performs better in six features, while the spatial-only model predicts the remaining two features better.

The spatial-only model is better at predicting the presence of fat-laden cystic plate and cystic duct seen with retraction. In contrast, the spatiotemporal model predicts better the visibility of the hepatocystic triangle, clear, thin cystic plate, dilated cystic plate, and features related to the critical view of safety.

4.4. Ablation studies

We ablate the parameters of the spatiotemporal models to find their optimal configuration. All experiments use visual features extracted by the Swin-S model pre-trained on ImageNet.

RNN-based models. We study the impact when using different numbers of layers, hidden size dimensions, and bi-directionality. The Table 6a shows that the GRU model performs best in a bi-directional configuration, with 2 layers and a hidden size of 600. Likewise, as the Table 6b shows, the same configuration is the best for the LSTM model.

MS-TCN. We study the impact when using different numbers of layers, stages, and hidden size dimensions. The Table 6c shows that the best configuration for the MS-TCN model is 1 stage, 8 layers, and a hidden size of 32.

Transformer. We study the impact when using different numbers of heads, layers, and hidden size dimensions. The Table 6d shows that the best configuration for the Transformer model has 4 heads, 2 layers, and a hidden size of 4096.

5. Discussions

This study presents and evaluates many spatial-only and spatiotemporal models using laparoscopic cholecystectomy videos to predict intraoperative features related to operative difficulty.

Our findings indicate that incorporating temporal information improves the models' LCOD prediction compared to using spatial information alone. In phase 1, Trocar Placement and Preparation, the best-performing spatial model is the EfficientNetV2-XL. However, the Transformer-based spatiotemporal model using visual features from EfficientNetV2-XL outperformed those results, demonstrating the added value of temporal context for the prediction. The same is true for phase 2, Hepatocystic Triangle Dissection, where the best overall model is the LSTM-based spatiotemporal model using visual features from the Cholec80 SSL pre-trained ResNet-50 model, which is also the best spatial-only model.

Both spatial-only and spatiotemporal models show substantial improvement when using visual features extracted from the model pre-trained on the Cholec80 dataset using the MoCo v2 SSL method. This suggests that domain-specific SSL pre-training enhances feature extraction quality, improving the temporal refinement of the features and the model's overall performance.

The feature-wise analysis revealed that spatiotemporal models excel in predicting dynamic features that evolve over the surgical phases, such as adhesions limiting access to the abdominal cavity and liver hypertrophy preventing gallbladder retraction, underscoring the importance of temporal dynamics in surgical video analysis. On the other hand, spatial-only models proved to be particularly effective for predicting visually distinct features consistently present across many frames with minimal temporal variation. This was the case for the fat-laden falciform ligament and adhesions features, which performed better with the spatial-only models.

The improved prediction accuracy of LCOD has several practical implications. Accurate and early prediction of operative difficulty can help in intraoperative decision-making, leading to better patient outcomes.

A major limitation of this study is that the dataset only contains sparse annotations, meaning that the ex-

Table 6: Ablation experiments for the spatiotemporal models.

Hidden size	Layers	Bi-directional	mAP		
			Phase 1	Phase 2	Average
300	2	True	0.5741	0.7079	0.6410
600	2	True	0.5867	0.7556	**0.6712**
1200	2	True	0.5440	0.7168	0.6304
600	1	True	0.5517	0.7431	0.6474
600	4	True	0.5504	0.7525	0.6515
600	2	False	0.5477	0.7523	0.6500

(a) GRU

Hidden size	Layers	Bi-directional	mAP		
			Phase 1	Phase 2	Average
300	2	True	0.5461	0.7446	0.6454
600	2	True	0.5851	0.7630	**0.6741**
1200	2	True	0.5351	0.7483	0.6417
600	1	True	0.5593	0.7443	0.6518
600	4	True	0.5147	0.6680	0.5914
600	2	False	0.6106	0.7131	0.6619

(b) LSTM

Hidden size	Layers	Stages	mAP		
			Phase 1	Phase 2	Average
16	8	1	0.6009	0.7272	0.6641
32	8	1	0.6351	0.7325	**0.6838**
64	8	1	0.6000	0.7478	0.6739
32	4	1	0.6105	0.7488	0.6797
32	16	1	0.5946	0.7464	0.6705
32	8	3	0.5585	0.7241	0.6413
32	8	2	0.5372	0.7308	0.6340

(c) MS-TCN

Hidden size	Layers	Heads	mAP		
			Phase 1	Phase 2	Average
2048	6	8	0.5519	0.7039	0.6279
3072	6	8	0.5855	0.6406	0.6131
4096	6	8	0.5680	0.7106	0.6393
5120	6	8	0.6261	0.6449	0.6355
4096	1	4	0.5901	0.7021	0.6461
4096	2	4	0.6201	0.7265	**0.6733**
4096	3	4	0.6086	0.6948	0.6517
4096	6	2	0.5416	0.7156	0.6286
4096	6	4	0.6170	0.6771	0.6471
4096	6	16	0.6005	0.6820	0.6413

(d) Transformer

act locations of the features within the frames are not known. Given that a surgical phase can contain thousands of frames, this lack of precise annotations presents a significant challenge. Consequently, this may limit the models' generalizability and effectiveness. Future work should explore methods to address this limitation, such as attention-based approaches to identify frames or image features contributing to the prediction.

A possible extension of these models can be deployed to provide real-time predictive feedback during surgery, which could further enhance their practical utility in clinical settings. Real-time prediction during surgery can provide valuable feedback to surgeons, potentially reducing the risk of complications.

6. Conclusions

We explore multiple spatial-only and spatiotemporal models for the assessment of operative difficulty in LC videos.

We demonstrate the potential of deep spatiotemporal models for predicting LCOD. These models achieve higher prediction accuracy by leveraging both spatial and temporal features, offering a promising tool for enhancing surgical planning and intraoperative decision-making.

We also demonstrate that transfer learning of SSL pre-trained models in domain-specific datasets, such as Cholec80, is a viable option for enhancing the prediction results.

Acknowledgments

This work was supported by French state funds managed by the ANR within the National AI Chair program under Grant ANR-20-CHIA-0029-01 (ChairAI4ORSafety) and within the Investments for the Future program under Grant ANR-10-IAHU-02 (IHU Strasbourg). It was granted access to the HPC resources of Unistra Mesocentre.

References

Andersson, R., Eriksson, K., Blind, P.J., Tingstedt, B., 2008. Iatrogenic bile duct injury – a cost analysis. HPB 10, 416–419. doi:10.1080/13651820802140745.

Barkun, J.S., Barkun, A.N., Meakins, J.L., 1993. Laparoscopic versus open cholecystectomy: The canadian experience. The American Journal of Surgery 165, 455–458. doi:10.1016/S0002-9610(05)80940-7.

Bengio, Y., Simard, P., Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks 5, 157–166. doi:10.1109/72.279181.

Bittner, R., 2004. The standard of laparoscopic cholecystectomy. Langenbeck's Archives of Surgery 389, 157–163. doi:10.1007/S00423-004-0471-1/TABLES/3.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations. 37th International Conference on Machine Learning, ICML 2020 PartF168147-3, 1575–1585. URL: `https://arxiv.org/abs/2002.05709v3`.

Chen, X., Fan, H., Girshick, R.B., He, K., 2020b. Improved baselines with momentum contrastive learning. CoRR abs/2003.04297. URL: https://arxiv.org/abs/2003.04297. arXiv 2003.04297.

Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y., 2014. On the properties of neural machine translation: Encoder-decoder approaches. Proceedings of SSST 2014 - 8th Workshop on Syntax, Semantics and Structure in Statistical Translation , 103–111URL: https://arxiv.org/abs/1409.1259v2, doi:10.3115/v1/w14-4012.

Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N., 2020. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12263 LNCS, 343–352. doi:10.1007/978-3-030-59716-0_33.

Demir, K.C., Schieber, H., Weise, T., Roth, D., May, M., Maier, A., Yang, S.H., 2023. Deep learning in surgical workflow analysis: A review of phase and step recognition. IEEE Journal of Biomedical and Health Informatics 27, 5405–5417. doi:10.1109/JBHI.2023.3311628.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

Deziel, D.J., Millikan, K.W., Economou, S.G., Doolas, A., Ko, S.T., Airan, M.C., 1993. Complications of laparoscopic cholecystectomy: A national survey of 4,292 hospitals and an analysis of 77,604 cases. The American Journal of Surgery 165, 9–14. doi:10.1016/S0002-9610(05)80397-6.

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. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021 - 9th International Conference on Learning Representations URL: https://arxiv.org/abs/2010.11929v2.

Farha, Y.A., Gall, J., 2019. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June, 3570–3579. doi:10.1109/CVPR.2019.00369.

Griffiths, E.A., Hodson, J., Vohra, R.S., Marriott, P., Katbeh, T., Zino, S., Nassar, A.H., 2019. Utilisation of an operative difficulty grading scale for laparoscopic cholecystectomy. Surgical Endoscopy 33, 110–121. doi:10.1007/S00464-018-6281-2/TABLES/6.

Halldestam, I., Enell, E.L., Kullman, E., Borch, K., 2004. Development of symptoms and complications in individuals with asymptomatic gallstones. British Journal of Surgery 91, 734–738. doi:10.1002/BJS.4547.

Halle-Smith, J.M., Hodson, J., Stevens, L.G., Dasari, B., Marudanayagam, R., Perera, T., Sutcliffe, R.P., Muiesan, P., Isaac, J., Mirza, D.F., Roberts, K.J., 2019. A comprehensive evaluation of the long-term economic impact of major bile duct injury. HPB 21, 1312–1321. doi:10.1016/J.HPB.2019.01.018.

Hanna, G.B., Shimi, S.M., Cuschieri, A., 1998. Randomised study of influence of two-dimensional versus three-dimensional imaging on performance of laparoscopic cholecystectomy. Lancet 351, 248–251. doi:10.1016/S0140-6736(97)08005-7.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition , 9726–9735doi:10.1109/CVPR42600.2020.00975.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-December, 770–778. doi:10.1109/CVPR.2016.90.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation 9, 1735–1780. doi:10.1162/NECO.1997.9.8.1735.

Hussain, A., 2011. Difficult laparoscopic cholecystectomy: Current evidence and strategies of management. Surgical Laparoscopy, Endoscopy and Percutaneous Techniques 21, 211–217. doi:10.1097/SLE.0B013E318220F1B1.

Iwashita, Y., Hibi, T., Ohyama, T., Honda, G., Yoshida, M., Miura, F., Takada, T., Han, H.S., Hwang, T.L., Shinya, S., Suzuki, K., Umezawa, A., Yoon, Y.S., Choi, I.S., Huang, W.S.W., Chen, K.H., Watanabe, M., Abe, Y., Misawa, T., Nagakawa, Y., Yoon, D.S., Jang, J.Y., Yu, H.C., Ahn, K.S., Kim, S.C., Song, I.S., Kim, J.H., Yun, S.S., Choi, S.H., Jan, Y.Y., Shan, Y.S., Ker, C.G., Chan, D.C., Wu, C.C., Lee, K.T., Toyota, N., Higuchi, R., Nakamura, Y., Mizuguchi, Y., Takeda, Y., Ito, M., Norimizu, S., Yamada, S., Matsumura, N., Shindoh, J., Sunagawa, H., Gocho, T., Hasegawa, H., Rikiyama, T., Sata, N., Kano, N., Kitano, S., Tokumura, H., Yamashita, Y., Watanabe, G., Nakagawa, K., Kimura, T., Yamakawa, T., Wakabayashi, G., Mori, R., Endo, I., Miyazaki, M., Yamamoto, M., 2017. An opportunity in difficulty: Japan-Korea-Taiwan expert delphi consensus on surgical difficulty during laparoscopic cholecystectomy. Journal of Hepato-Biliary-Pancreatic Sciences 24, 191–198. doi:10.1002/JHBP.440.

Jin, Y., Long, Y., Gao, X., Stoyanov, D., Dou, Q., Heng, P.A., 2022. Trans-svnet: hybrid embedding aggregation transformer for surgical workflow analysis. International Journal of Computer Assisted Radiology and Surgery 17, 2193–2202. doi:10.1007/S11548-022-02743-8/FIGURES/5.

Kanakala, V., Borowski, D.W., Pellen, M.G., Dronamraju, S.S., Woodcock, S.A., Seymour, K., Attwood, S.E., Horgan, L.F., 2011. Risk factors in laparoscopic cholecystectomy: A multivariate analysis. International Journal of Surgery 9, 318–323. doi:10.1016/J.IJSU.2011.02.003.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE International Conference on Computer Vision , 9992–10002URL: https://arxiv.org/abs/2103.14030v2, doi:10.1109/ICCV48922.2021.00986.

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. 7th International Conference on Learning Representations, ICLR 2019 URL: https://arxiv.org/abs/1711.05101v3.

Madni, T.D., Leshikar, D.E., Minshall, C.T., Nakonezny, P.A., Cornelius, C.C., Imran, J.B., Clark, A.T., Williams, B.H., Eastman, A.L., Minei, J.P., Phelan, H.A., Cripps, M.W., 2018. The parkland grading scale for cholecystitis. The American Journal of Surgery 215, 625–630. doi:10.1016/J.AMJSURG.2017.05.017.

Mascagni, P., Vardazaryan, A., Alapatt, D., Urade, T., Emre, T., Fiorillo, C., Pessaux, P., Mutter, D., Marescaux, J., Costamagna, G., Dallemagne, B., Padoy, N., 2022. Artificial intelligence for surgical safety automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. Annals of Surgery 275, 955–961. doi:10.1097/SLA.0000000000004351.

Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Mutter, D., Padoy, N., 2024. Latent graph representations for critical view of safety assessment. IEEE Transactions on Medical Imaging 43, 1247–1258. doi:10.1109/TMI.2023.3333034.

Murphy, M.M., Ng, S.C., Simons, J.P., Csikesz, N.G., Shah, S.A., Tseng, J.F., 2010. Predictors of major complications after laparoscopic cholecystectomy: Surgeon, hospital, or patient? Journal of the American College of Surgeons 211, 73–80. doi:10.1016/J.JAMCOLLSURG.2010.02.050.

Nassar, A.H., Ashkar, K.A., Mohamed, A.Y., Hafiz, A.A., 1995. Is laparoscopic cholecystectomy possible without video technology? Minimally Invasive Therapy 4, 63–65. doi:10.3109/13645709509152757.

Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N., 2022. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Medical Image Analysis 78, 102433. doi:10.1016/J.MEDIA.2022.102433.

Pucher, P.H., Brunt, L.M., Davies, N., Linsk, A., Munshi, A., Rodriguez, H.A., Fingerhut, A., Fanelli, R.D., Asbun, H., Aggarwal, R., 2018. Outcome trends and safety measures after

30 years of laparoscopic cholecystectomy: a systematic review and pooled data analysis. Surgical Endoscopy 32, 2175–2183. doi:10.1007/S00464-017-5974-2/FIGURES/4.

Ramesh, S., Srivastav, V., Alapatt, D., Yu, T., Murali, A., Sestini, L., Nwoye, C.I., Hamoud, I., Sharma, S., Fleurentin, A., Exarchakis, G., Karargyris, A., Padoy, N., 2023. Dissecting self-supervised learning methods for surgical computer vision. Medical Image Analysis 88, 102844. doi:10.1016/J.MEDIA.2023.102844.

Russo, M.W., Wei, J.T., Thiny, M.T., Gangarosa, L.M., Brown, A., Ringel, Y., Shaheen, N.J., Sandler, R.S., 2004. Digestive and liver diseases statistics, 2004. Gastroenterology 126, 1448–1453. doi:10.1053/j.gastro.2004.01.025.

Sechidis, K., Tsoumakas, G., Vlahavas, I., 2011. On the stratification of multi-label data. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 6913 LNAI, 145–158. doi:10.1007/978-3-642-23808-6_10.

Shaffer, E.A., 2005. Epidemiology and risk factors for gallstone disease: Has the paradigm changed in the 21st century? Current Gastroenterology Reports 7, 132–140. doi:10.1007/S11894-005-0051-8/METRICS.

Shaffer, E.A., 2006. Epidemiology of gallbladder stone disease. Best Practice & Research Clinical Gastroenterology 20, 981–996. doi:10.1016/J.BPG.2006.05.004.

Sharma, S., Nwoye, C.I., Mutter, D., Padoy, N., 2023a. Rendezvous in time: an attention-based temporal fusion approach for surgical triplet recognition. International Journal of Computer Assisted Radiology and Surgery 18, 1053–1059. doi:10.1007/S11548-023-02914-1/TABLES/5.

Sharma, S., Nwoye, C.I., Mutter, D., Padoy, N., 2023b. Surgical action triplet detection by mixed supervised learning of instrument-tissue interactions. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 14228 LNCS, 505–514. doi:10.1007/978-3-031-43996-4_48/TABLES/4.

Smith, L.N., Topin, N., 2017. Super-convergence: Very fast training of neural networks using large learning rates. URL: https://arxiv.org/abs/1708.07120v3, doi:10.1117/12.2520589. 1708.07120v3.

Sugrue, M., Sahebally, S.M., Ansaloni, L., Zielinski, M.D., 2015. Grading operative findings at laparoscopic cholecystectomy- a new scoring system. World Journal of Emergency Surgery 10, 1–8. doi:10.1186/S13017-015-0005-X/FIGURES/5.

Tan, M., Le, Q.V., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. 36th International Conference on Machine Learning, ICML 2019 2019-June, 10691–10700. URL: https://arxiv.org/abs/1905.11946v5.

Tan, M., Le, Q.V., 2021. Efficientnetv2: Smaller models and faster training. Proceedings of Machine Learning Research 139, 10096–10106.

Terho, P.M., Leppäniemi, A.K., Mentula, P.J., 2016. Laparoscopic cholecystectomy for acute calculous cholecystitis: A retrospective study assessing risk factors for conversion and complications. World Journal of Emergency Surgery 11, 1–9. doi:10.1186/S13017-016-0111-4/TABLES/5.

Twinanda, A.P., 2017. Vision-based approaches for surgical activity recognition using laparoscopic and RBGD videos. Theses. Université de Strasbourg. URL: https://theses.hal.science/tel-01557522.

Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., Mathelin, M.D., Padoy, N., 2017. Endonet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Transactions on Medical Imaging 36, 86–97. doi:10.1109/TMI.2016.2593957.

Törnqvist, B., Strömberg, C., Persson, G., Nilsson, M., 2012. Effect of intended intraoperative cholangiography and early detection of bile duct injury on survival after cholecystectomy: population based cohort study. BMJ 345. doi:10.1136/BMJ.E6457.

Törnqvist, B., Zheng, Z., Ye, W., Waage, A., Nilsson, M., 2009. Long-term effects of iatrogenic bile duct injury during cholecystectomy. Clinical Gastroenterology and Hepatology 7, 1013–1018. doi:10.1016/J.CGH.2009.05.014.

Vannucci, M., Laracca, G.G., Mercantini, P., Perretta, S., Padoy, N., Dallemagne, B., Mascagni, P., 2022. Statistical models to preoperatively predict operative difficulty in laparoscopic cholecystectomy: A systematic review. Surgery 171, 1158–1167. doi:10.1016/J.SURG.2021.10.001.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Łukasz Kaiser, Polosukhin, I., 2017. Attention is all you need. Advances in Neural Information Processing Systems 2017-December, 5999–6009. URL: https://arxiv.org/abs/1706.03762v7.