Joint Temporal Modeling for Surgical Phase Recognition and Schedule Prediction

Yuanzhe Chang

University College London

I acknowledge the use of Gemini and ChatGPT for brainstorming, plot generation,code debugging and report proofreading.

Abstract. Modelling surgical workflow from video is an important problem in computer-assisted surgery, with applications in progress monitoring and decision support. This coursework addresses two related tasks on laparoscopic cholecystectomy videos. In Task A, we investigate temporal structure prediction(TSP) by estimating both the remaining duration of the current surgical phase and the start and end times of all upcoming phases. A multi-stage temporal convolutional network (MS-TCN) is employed to jointly perform phase recognition and temporal regression, enabling the model to capture global procedural progress. In Task B, we examine whether the predicted temporal information can benefit a surgical phase recognition (PR) task. A baseline phase classification model using only visual features is compared with a timed model that incorporates the predicted temporal schedules from task A as auxiliary input. Experimental results demonstrate that incorporating predicted temporal information leads to an improved ability to recognise the phase 2 and 4. This suggests that temporal prior knowledge enhances the model's ability to recognise some specific phases, which are particularly relevant in clinical workflow monitoring.

Keywords: Surgical workflow analysis · Temporal Structure Prediction · Phase Recognition· MS-TCN

1 Introduction

Understanding and modelling surgical workflow from intraoperative video is a fundamental problem in computer-assisted surgery. Accurate recognition of surgical phases enables downstream applications such as progress monitoring, resource allocation, and context-aware decision support systems [1]. In laparoscopic procedures such as cholecystectomy, surgical workflows exhibit strong temporal regularities, including a consistent ordering of phases and characteristic phase durations [2].

Most existing approaches to surgical phase recognition rely on deep learning models that implicitly learn temporal context from visual features [3]. While temporal convolutional networks and recurrent architectures have shown strong

performance, they typically treat time as an implicit latent variable rather than an explicitly modelled quantity. As a result, information about remaining duration, phase progress, and upcoming phase transitions is not directly constrained during training.

From a clinical perspective, explicit temporal information is highly valuable. Surgeons and operating room staff are often interested not only in the current phase, but also in how much time remains in the ongoing phase and when subsequent phases are expected to occur [4]. Moreover, short or transitional phases, although brief, may correspond to critical procedural steps and should not be ignored by automated systems.In this work, we explore whether explicitly predicting temporal structure can improve surgical workflow modelling. Specifically, we address the following research questions:

- Can a deep temporal model jointly learn surgical phase recognition and explicit prediction of remaining time and future phase schedules?
- Does incorporating predicted temporal structure as auxiliary input improve the performance of surgical phase recognition?

To answer these questions, we design a framework based on MS-TCN network that performs both phase classification and temporal structure prediction. The predicted temporal information is subsequently used to augment visual features in phase recognition task. Through quantitative and qualitative experiments, we demonstrate that explicit temporal modelling provides meaningful inductive bias and improves recognition of short-duration phases, despite not increasing overall classification accuracy.

2 Methods

An overview of the proposed framework is shown in Fig. 1. Given a surgical video, frame-level visual features are extracted using a pre-trained vision encoder (DINOv2) and fed into an MS-TCN backbone [5], which is shared across all experiments. The MS-TCN architecture consists of multiple refinement stages. Each stage is composed of a stack of dilated temporal convolutional layers with residual connections. The first stage operates directly on the input feature sequence, while subsequent stages iteratively refine predictions using the softmax outputs from the previous stage, following a coarse-to-fine temporal refinement strategy. On top of the shared MS-TCN backbone, task-specific learning objectives are applied. For Task B, the network performs frame-wise phase classification supervised by focal loss. For Task A, additional prediction heads are attached and trained with regression-based supervision to model temporal structure explicitly. By fixing the temporal backbone and varying only the supervision strategy, the framework allows a controlled study of the impact of explicit temporal modelling in surgical workflow analysis.

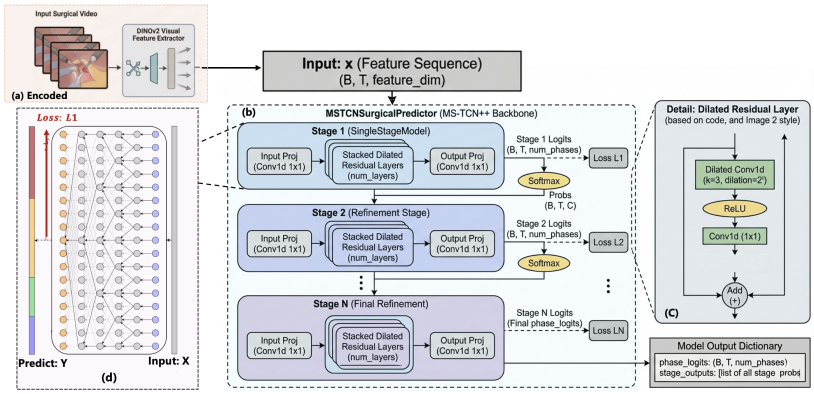

Fig. 1. Overview of MS-TCN.(a) Feature extraction using DINOv2. (b) Global structure of the MS-TCN. (c) Overview of the dilated residual layer. (d) Detailed view of a single refinement stage [5].

2.1 Task A: Temporal Structure Prediction(TSP)

As shown in Fig. 2, the MS-TCN processes the sequence and produces frame-wise predictions through multiple refinement stages. At each time step, the network outputs three heads: a phase classification head, a time position head estimating procedural progress, and a regression head predicting the temporal structure of all surgical phases.

The inclusion of the classification head is motivated by the requirements of real-world clinical deployment: to reliably estimate the remaining time of a specific phase, the model must first possess the visual intelligence to recognize the ongoing surgical task. This multi-task design ensures that temporal regressions are semantically grounded in the current surgical context. The classification head is supervised with focal loss to address class imbalance(e.g. phase 0 covers only 4.68%, but phase 1 covers 41.91% in one video.), while the regression head adopts a Huber loss for its robustness to outliers and differentiablity around zero.

2.2 Task B: Phase Recognition(PR)

To evaluate whether explicit temporal schedule prediction benefits phase classification, we incorporate the regression output from Task A as auxiliary input with visual feature to the MS-TCN model shown in Fig. 1. A visual-only baseline is trained using the same MS-TCN architecture but with visual features as the sole input. All other training settings are kept identical between the two models to ensure a fair comparison. Both models are supervised using the focal loss. Notably, the temporal information is not provided as ground truth but is estimated by the temporal prediction model described in Fig. 2.

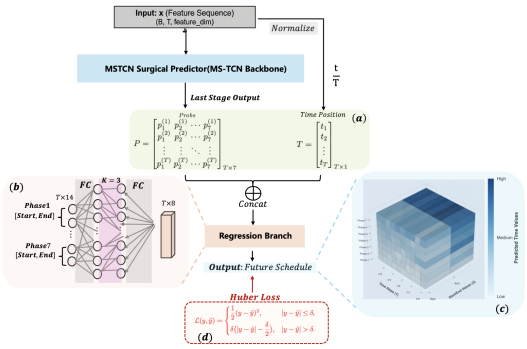

Fig. 2. Overview of Task A. (a) The time position is concatenated to the probability matrix. (b) The regression branch consists of three 1D convolutional layers, where the middle layer uses a kernel size of 3, while the other two layers use kernel size 1. (c) The predicted output tensor $\mathbb{R}^{T\times 7\times 2}$. (d) The Huber loss is adopted to supervise the learning.

3 Experiments

3.1 Dataset and Experimental Setup

Experiments were conducted on the Cholec80 dataset [6], which comprises 80 laparoscopic cholecystectomy videos that are down-sampled to 1 Hz and provided with frame-level phase annotations. The dataset was randomly partitioned into training, validation, and test sets with a 60:10:10 split, respectively. During training, the models shared the same training parameters: epochs = 100, batch size =1,Huber loss function with delta = 1.0, learning rate = 0.0005, dropout=0.3. After training and fine-tuning, test dataset will be only used once.

3.2 Evaluation Metrics

For task A, the TSP performance was evaluated by MAE as proposed in [7]. Given a surgical video sequence with N time steps, the MAE is defined as:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{1}$$

where $\hat{y}_i$ represents the predicted remaining time (in seconds) at frame i, and y_i denotes the corresponding ground-truth remaining time. It provides an intuitive and clinically meaningful assessment of temporal prediction accuracy, reflecting timing errors that may impact intraoperative planning and resource coordination. For Task B, PR performance is evaluated using the macro-averaged F1 score, defined as

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{C}\sum_{c=1}^{C}\frac{2\cdot \mathrm{Precision}_c \cdot \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c}, \tag{2}$$

where C denotes the number of surgical phases, and Precision_c and Recall_c are computed independently for each phase c. Macro-averaging assigns equal importance to all phases, making it well suited for imbalanced phase distributions and short-duration surgical phases. Moreover, the F1 score jointly considers precision and recall, providing a balanced evaluation that avoids over-emphasising either false positives or false negatives that can lead to incorrect phase transitions that are undesirable in clinical decision-making.

4 Results

4.1 Task A: TSP Performance

The result from (a) to (c) in Fig. 3 is based on the best model performing on the video 68. The overall MAE of ten test videos is 198.92 ± 83.54s. Fig. 3(d) shows that P0 and P3 have the largest variance(std= 163.0s and 145.8s) in MAE. Despite its relatively narrow duration range (mean = 84s, std = 43.20s), P0 exhibits a strong correlation with MAE (Pearson's $r \approx 0.83$, $p < 0.01$), indicating a high sensitivity of prediction error to temporal variation. Phase 3 spans a substantially wider duration range (mean = 773.10s, std = 377.28s) and demonstrates an even stronger duration–MAE correlation ($r \approx 0.94$, $p < 0.001$). In contrast, P1 and P2 show weak correlations between duration and MAE (P1: $r \approx -0.01$, $p \approx 0.97$; P2: $r \approx 0.14$, $p \approx 0.64$), despite their relatively large duration variability, suggesting that prediction difficulty in these phases may stem from noisy or less distinctive visual features rather than temporal length.

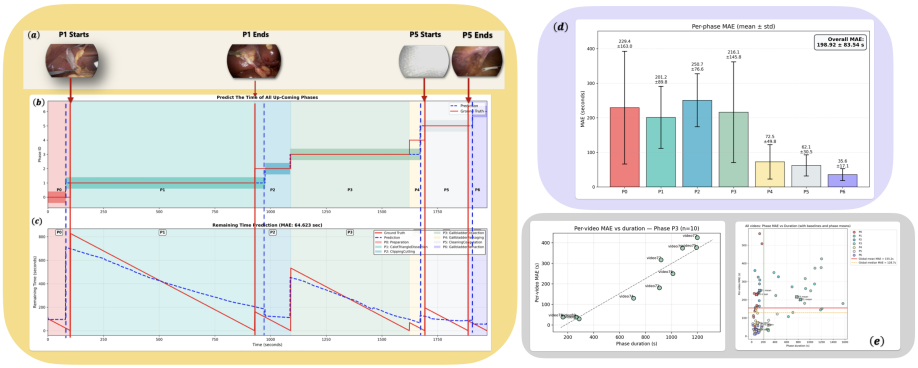

Fig. 3. Visualization of Temporal Structure Prediction.(a) Representative key-frames highlighting the start and end of specific phases (b) Coming Phases Prediction: Comparison between ground-truth (red) and predicted (blue) phase sequences. Overlapped colored regions represent predicted phase durations.(c) Remaining Time Prediction: Ground-truth remaining seconds(red) and the model's predicted remaining seconds(blue) are plotted over time.(d) Per-phase MAE (mean ± std) to quantify the model's error for each surgical phase.(e) Correlation between phase duration and MAE.

4.2 Task B: Phase Recognition Performance

The model with auxiliary temporal information achieves an overall F1 score of $49.78\% \pm 0.72\%$, representing a marginal 0.77% improvement over the baseline ($49.01\% \pm 0.75\%$). However, this overall metric masks an enhancement in the model's ability to recognize P2 and P4. As shown in Fig. 4(a) and (b), the baseline model fails to detect P2 and P4 in the test video (in the blue and yellow dashed boxes), whereas the model with auxiliary successfully identifies these transitions. This performance gain is consistent across the entire test set: as shown in the solid red box—which represents the per-phase F1 scores—the baseline model consistently fails to capture P2 and P4. Interestingly, while the confusion matrix in Fig. 4 suggests that the baseline can achieve a higher overall accuracy, this is primarily due to the "accuracy paradox" in imbalanced datasets. The baseline achieves its score by biasedly predicting dominant, long-duration phases while neglecting infrequent but clinically critical ones like P2 and P4.

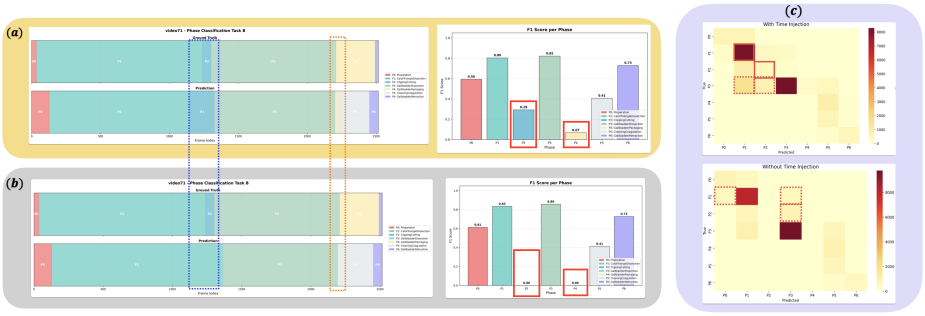

Fig. 4. Performance Comparison for Task B.(a) Phase recognition results and per-phase F1 scores for the model integrated with temporal structure information. (b) Baseline results using visual features only.(c) Confusion Matrix Comparison: the top matrix is the model with temporal injection compared to the baseline below. Solid boxes highlight correctly identified phases, while dashed boxes pinpoint specific errors.

5 Discussion and Conclusion

Regarding the questions raised in the Introduction, Task A's results align with [8]. However, the lack of an ablation group excluding phase recognition makes it difficult to determine whether such classification explicitly benefits future schedule prediction. For Task B, although the predicted temporal structure enhanced short-phase recognition, reusing prior knowledge from Task A on the same training set may have induced overfitting, despite outperforming the baseline. Furthermore, the limited test size of only 10 videos may reduce the statistical significance of the comparison. Future work should employ rigorous ablation studies, shuffled datasets, and larger test cohorts to further validate the efficacy of joint learning.

References

1. E. Travis *et al.*, "Operating theatre time, where does it all go? A prospective observational study," *BMJ*, vol. 349, 2014, doi:10.1136/bmj.g7182.
2. Yuniartha DR, Masruroh NA, Herliansyah MK. An evaluation of a simple model for predicting surgery duration using a set of surgical procedure parameters. *Inform Med Unlocked.* 2021;25:100633. https://doi.org/10.1016/j.imu.2021.100633
3. Lee SG, et al. Adaptive undersampling and short clip-based two-stream CNN-LSTM model for surgical phase recognition on cholecystectomy videos. *Biomedical Signal Processing and Control.* 2024;88:105637. https://doi.org/10.1016/j.bspc.2023.105637
4. V. Riahi, H. Hassanzadeh, S. Khanna *et al.*, "Improving preoperative prediction of surgery duration," *BMC Health Services Research*, vol. 23, no. 1, p. 1343, 2023, doi:10.1186/s12913-023-10264-6.
5. Farha YA, Jurgen G. MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2019:3575-3584. https://doi.org/10.1109/CVPR.2019.00369
6. 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, 2016, doi:10.1109/TMI.2016.2593957.
7. A. P. Twinanda, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "RSDNet: Learning to predict remaining surgery duration from laparoscopic videos without manual annotations," *IEEE Transactions on Medical Imaging*, vol. 38, no. 4, pp. 1066–1075, 2019, doi:10.1109/TMI.2018.2878056.
8. Loukas C, Prevezanou K, Seimenis I, Schizas D. Prediction of remaining surgery duration in laparoscopic videos based on visual saliency and the transformer network. *The International Journal of Medical Robotics and Computer Assisted Surgery.* 2024;20(3):e2632. https://doi.org/10.1002/rcs.2632