Surgical Phase Detection Using Deep Learning
Critical Review

Xiaorui Zhang, Wenkai Luo, Xucheng Ma

March 2022

1 Project Problem Statement & Background

Our project focuses on developing Deep Learning models which perform mastoidectomy video segmentation. The model **inputs** is a sequence of frames, and model **output** is a sequence of surgical phase labels. The main technical problems we need to solve is how to design an efficient and robust feature extractor for surgical videos, then we can perform classification in the feature space to obtain phase labels.

In general surgical phase segmentation workflow, both **spatial** features and **temporal** features are often considered, e.g.

- Spatial Features: anatomical structures, tool presence/positioning ...
- Temporal Features: anatomical changes, tool movement, and camera view changes ...

Mastoidectomy videos has some characteristic features:

- Anatomical features are mostly **rigid** since is a skull based surgery.
- Anatomical changes are mostly in depth direction as the overall surgery process is exposing the interior structures behind patient's ear.
- Camera view changes might imply transition between surgical phases.
- Tool positioning may contain more useful information about correct phase label than tool presence, because main tools involved are only drill and suction device and the are used across whole procedure.

Typical DL methods for surgical phase segmentation has following components:

- Spatial Feature Extractor
- Temporal Feature Extractor
- Spatial-Temporal Feature Fusion
- Classifier in Feature Space

Papers which we included in this review are all consisted of components above, but different approaches were chosen for each component. For each paper, we first introduce the problem that the specific network design aimed to address, then summarize the main contributions, and discuss key results and potential problems.

2 Critical Review of SV-RCNet

2.1 Background and problems with prior works

Before this work, several Vision-Based and Deep Learning oriented methods were proposed to extract spatial and temporal features and train a classifier to perform the segmentation task. However, there are three main drawbacks to the previous works.

- The previously used visual features, either hand-crafted or shallow CNN-based, are still far from sufficient to represent the complicated visual characteristics of the frames in surgical videos.
- When exploiting the temporal information, most traditional methods rely on linear statistical models with pre-defined dependencies, which are incapable of precisely representing motions in surgical videos.
- Most existing methods harness visual and temporal information separately, i.e., first using visual features with classifiers to predict each frame, and then using temporal dependencies to refine the results. In this way, visual features are unable to play a role in the temporal model.

2.2 Significance of this paper

There are three main contributions from this paper.

- First, this paper proposed for the first time to use a deep neural network to extract visual features from the video frames. The ResNet makes it possible to optimize a much deeper network by embedding the identity transformation into the network through residual blocks. Using the ResNet as a spatial feature extractor, the SV-RCNet is able to find more discriminative features compared with models of shallow CNN.
- Second, instead of dealing with spatial and temporal features separately, SV-RCNet integrates ResNet and LSTM to form a novel recurrent convolutional architecture in order to take full advantage of the complementary information of visual and temporal features. LSTM is a direct improvement on the top of recurrent neural network (RNN), LSTM uses gates to generate "cell states" to aid the gradient flow and alleviate vanishing gradient.
- SV-RCNet integrates the ResNet and the LSTM network, so that they are jointly trained in an end-to-end manner to generate high-level features that encode both spatial (visual) and temporal information. Particularly, the Spatio-temporal features learned by SV-RCNet are sensitive to motions in surgical videos and can precisely identify the phase transition frames.

2.3 Results, problems and relation to our Project

SV-RCNet was the state-of-the-art model in surgical video segmentation in terms of performance in both Cholec80 and MICCAI M2CAI workflow Challenge. However, there are still drawbacks inherited in LSTMs, which retain the memory of a limited sequence, that cannot span minutes or hours, which is the average duration of surgery. Thus, the temporal information must be present in a slow, sequential way and prohibits inference parallelization, which would be beneficial for integration in an online OR scenario.

Relation to our project:
LSTM, as one of the most effective temporal feature extractors prior to the transformer, we believe it's still worthwhile to benchmark its performance in our task. The main focus of this benchmarking will be on the LSTM itself. We will test out the segmentation task using different input sequence lengths, tune and find the best hyperparameters to get the most out of the LSTM. The authors of the paper [1] claimed, training the LSTM and the ResNet in an end-to-end manner gives the best performance, but they have not done any experiment on training them separately. Moreover, in paper [2] we learned that training the ResNet separately, and feeding spatial features directly into the temporal model can yield promising results since much fewer parameters need to be optimized at one time. In our benchmarking task, we will investigate both scenarios and field out whether training both models in an end-to-end manner is indeed better than training them separately.

3 Critical Review of TeCNO

3.1 Problem to be addressed & Proposed methods

Once LSTM has been proven to be state-of-the-art in dealing with the temporal task, lots of research aims to adapt the LSTM model to the surgery workflow analysis. The promising results from LSTM-related models prove its capability on time-series data. Meanwhile, more experiments and analyses on the surgery phase segmentation indicated that the variability of patient anatomy and surgeon style are the bottleneck for more accurate segmentation. Recent research found that the long-range temporal dependencies are beneficial for compensating these challenges, while LSTM has been proved to be unable to capture the long-range temporal pattern. The weakness of LSTM is the motivation why the author of Temporal Convolutional Network for the OPerating room (TeCNO) introduces a temporal convolution neural network instead of LSTM.

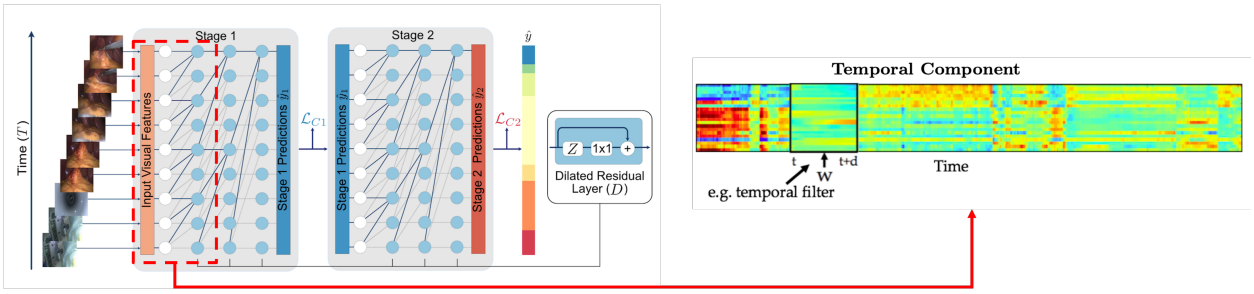

Figure 1: TeCNO architecture [2]

In terms of the detailed architecture, TeCNO has a similar structure to the previous model except for the temporal feature extractor. TeCNO leverages transfer learning using a pre-trained ResNet50 model as a spatial feature extractor. Instead of using the LSTM model, a multi-stage temporal convolutional network(MS-TCN) is deployed. The extracted spatial features will then be fed into MS-TCN to detect the desired temporal pattern and refine the segmentation result. Each stage consists of multiple dilated temporal convolution layers. The essential part of dilated temporal convolution layers is the one-dimensional convolution against the time stamp, shown in the figure1. The predictions are refined by stacking multiple stages together. The number of stages is one hyperparameter for this network. By ablation study, two stages are enough to produce accurate results.

3.2 Significances and experiment results

There are three Significances of this model:

1. **Receptive field**: Since dilated convolutions are implemented and multiple layers are stacked together, MS-TCN has a much larger receptive field than LSTM, compensating for the issue.

2. **Segmentation refinement**: Compared with the original temporal convolution network, the author leverage the idea of the stacking stage for better segmentation refinement.

3. **Amount of parameter**: Compared with LSTM, MS-TCN has less amount of parameter. By using the dilated convolution layer, TeCNO can achieve slightly better performance with fewer parameters, which is more suitable for fast inference.

With these two significances, the TeCNO outperforms the state-of-the-art model, which can be seen in the evaluation table2. Meanwhile, the multi-stage structure also smooths the segmentation result by removing the ripples. As we can see on the bar chart2, TeCNO has a smoother result than the model using LSTM.

	Cholec80			Cholec51		
	Acc	Prec	Rec	Acc	Prec	Rec
PhaseLSTM [12]	79.68 ± 0.07	72.85 ± 0.10	73.45 ± 0.12	81.94 ± 0.20	68.84 ± 0.11	68.05 ± 0.79
EndoLSTM [23]	80.85 ± 0.17	76.81 ± 2.62	72.07 ± 0.64	—	—	—
MTRCNet [17]	82.76 ± 0.01	76.08 ± 0.01	78.02 ± 0.13	—	—	—
ResNetLSTM [15]	86.58 ± 1.01	80.53 ± 1.59	79.94 ± 1.79	86.15 ± 0.60	70.45 ± 2.85	67.42 ± 1.43
TeCNO	$\mathbf{88.56 \pm 0.27}$	$\mathbf{81.64 \pm 0.41}$	$\mathbf{85.24 \pm 1.06}$	$\mathbf{87.34 \pm 0.66}$	$\mathbf{75.87 \pm 0.58}$	$\mathbf{77.17 \pm 0.73}$

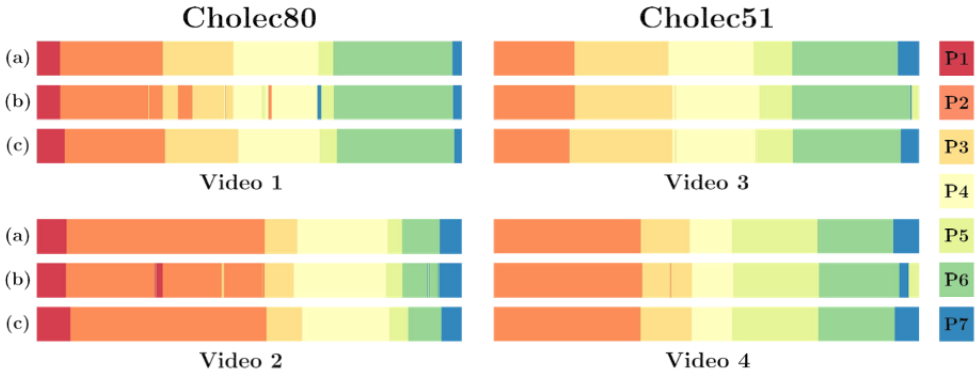

Fig. 2. Qualitative Results regarding quality of phase recognition for Cholec80 and Cholec51. (a) Ground Truth (b) ResNetLSTM Predictions (c) TeCNO Predictions. P1 to P7 indicate the phase label.

Figure 2: TeCNO evaluation result [2]

3.3 Assessment and future work

As for the assessment of this paper, several pros and cons of TeCNO will be presented in this review report.This paper provides two perspectives, which might be beneficial for our project. The first one is introducing a temporal convolutional network into this task. TCN provides another potential to have better segmentation results instead of using Recurrent neural networks(RNN). The benefits

brought from TCN include the larger receptive field and fewer parameters. This paper inspires us to find out if a better TCN model can address the problem we have in the mastoidectomy surgery dataset. The second one is how the author considers the temporal convolutional network. Some previous papers consider the TCN a spatial feature extractor, while TeCNO considers MS-TCN a prediction refinement mechanism. This assumption somehow brings a thought to our project: how can we utilize the spatial feature efficiently. There are two shortcomings of this paper where the author doesn't provide any experimental results to support his conclusion of the over-fitting issue, while it doesn't provide any comparison between using different TCN models.
Based on this, we will work on this project to explore the potential of TCN in mastoidectomy surgery video and how we can utilize the spatial features efficiently to achieve online usage.

4 Critical Review of Trans-SVNet

4.1 Problem with previous work and proposed method

As shown in figure 3 (a), previous methods perform spatial and temporal feature extractions in serial, where CNN-based methods serves as spatial feature extractor, and the resulting features are then fed into temporal feature extractor; there's no specific module in the network that fuse the spatial and temporal features. The spatial-temporal representations obtained by successive spatial and temporal feature extractions overlook the complementary effects of spatial and temporal features. [3].

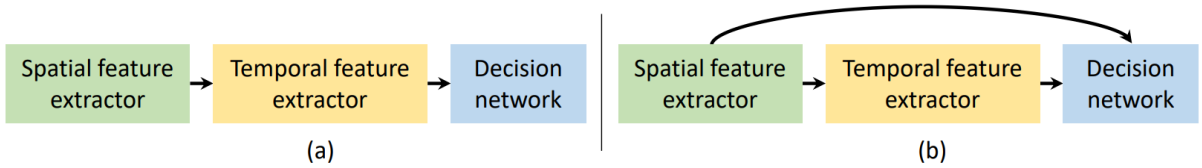

Figure 3: Structure of methods for surgical phase segmentation: (a) previous works (b) Trans-SVNet [3]

In Trans-SVNet, transformer layer is used as an aggregation model which fuses spatial and temporal features, and it serves as the decision network mentioned in figure 3. Figure 4 shows the network architecture.

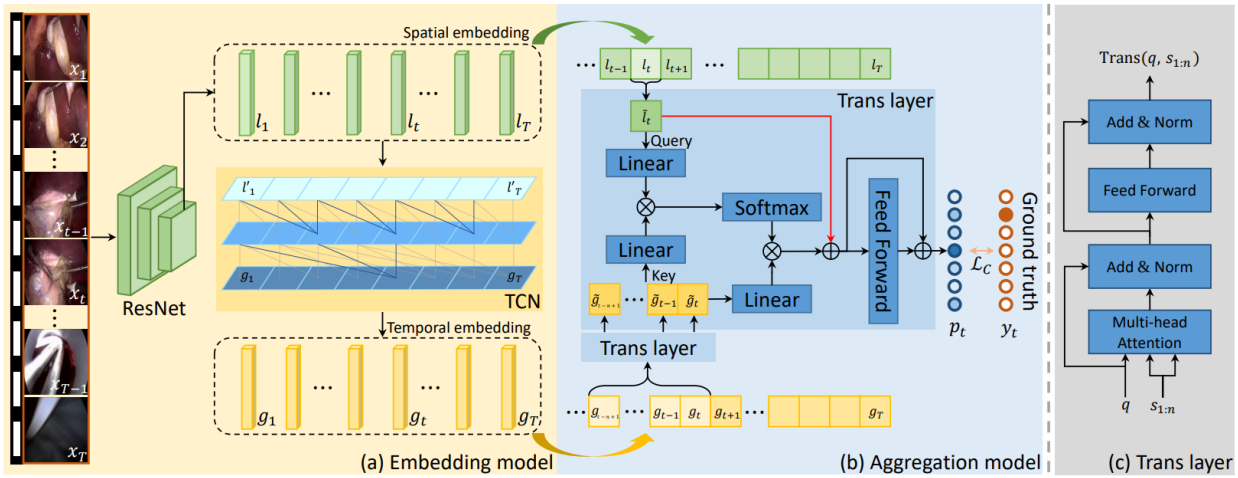

Figure 4: Trans-SVNet architecture[3]

The network use ResNet as the spatial feature extractor; temporal features are then extracted with TeCNO from spatial features. The first transformer layer of the aggregation model (not shown in figure 4) performs self-attention on sequence of temporal features , then the second transformer layer performs a cross-attention where dimensionally reduced spatial features are queries, and the resulting temporal features from previous transformer layer are both keys and values. The output features is a weighted sum of temporal features in the sequence, where the weights contains information of the relations between corresponding spatial feature and other temporal features in the sequence.

4.2 Contributions and experiment results

The main contributions of this paper are summarized as followings:

- It's the first time that transformer is used to reconsider the Complementary Effects of spatial and temporal features.
- A significant boost in inference speed is achieved by processing hybrid embeddings in parallel

The experiment results on Cholec80 and M2CAI16 are shown in table 5, and we can see that Trans-SVNet outperforms other state-of-arts methods while keeping a relative low number of parameters.

Method	Cholec80				M2CAI16				#param
	Accuracy	Precision	Recall	Jaccard	Accuracy	Precision	Recall	Jaccard	
EndoNet* [26]	81.7 ± 4.2	73.7 ± 16.1	79.6 ± 7.9	—	—	—	—	—	58.3M
EndoNet+LSTM* [27]	88.6 ± 9.6	84.4 ± 7.9	84.7 ± 7.9	—	—	—	—	—	68.8M
MTRCNet-CL* [14]	89.2 ± 7.6	86.9 ± 4.3	88.0 ± 6.9	—	—	—	—	—	29.0M
PhaseNet [24,26]	78.8 ± 4.7	71.3 ± 15.6	76.6 ± 16.6	—	79.5 ± 12.1	—	—	64.1 ± 10.3	58.3M
SV-RCNet [13]	85.3 ± 7.3	80.7 ± 7.0	83.5 ± 7.5	—	81.7 ± 8.1	81.0 ± 8.3	81.6 ± 7.2	65.4 ± 8.9	28.8M
OHFM [30]	87.3 ± 5.7	—	—	67.0 ± 13.3	85.2 ± 7.5	—	—	68.8 ± 10.5	47.1M
TeCNO [4]	88.6 ± 7.8	86.5 ± 7.0	87.6 ± 6.7	75.1 ± 6.9	86.1 ± 10.0	85.7 ± 7.7	$\mathbf{88.9 \pm 4.5}$	74.4 ± 7.2	24.7M
Trans-SVNet (ours)	$\mathbf{90.3 \pm 7.1}$	$\mathbf{90.7 \pm 5.0}$	$\mathbf{88.8 \pm 7.4}$	$\mathbf{79.3 \pm 6.6}$	$\mathbf{87.2 \pm 9.3}$	$\mathbf{88.0 \pm 6.7}$	87.5 ± 5.5	$\mathbf{74.7 \pm 7.7}$	24.7M

Figure 5: Experiment Results on Cholec80 and M2CAI16

4.3 Assessment and future work

As for our project, this paper inspires us that attention mechanism provides a good way to fuse spatial and temporal features; moreover, the attention mechanism can also be use to fuse any sequence of features. Another valuable knowledge we learn from this paper is that low-dimensional video embeddings result in significant inference speed boost, but we should be aware that this is under the assumption that image frames with same surgical labels do cluster in the spatial feature space.

On the other hands, there are some debatable points in this paper. The embedding model (2 feature extractors) and aggregation model (transformer layer) are trained separately, and there is no comparison between the proposed method and results from running transformer-based method directly on spatial features. To address these two concerns and build our network based on Trans-SVNet, we are planning to first compare Trans-SVNet with end-to-end trained models and pure-spatial features model, and then treat "idle" and "camera adjustment" as different tasks instead of labels and use cross-attention to exploring the complementary effects of "idle"/"camera adjustment" and the main surgical phase segmentation task.

References

[1] Y. Jin, Q. Dou, H. Chen, L. Yu, J. Qin, C.-W. Fu, and P.-A. Heng, "Sv-rcnet: Workflow recognition from surgical videos using recurrent convolutional network," *IEEE Transactions on Medical Imaging*, vol. 37, no. 5, pp. 1114–1126, 2018.

[2] T. Czempiel, M. Paschali, M. Keicher, W. Simson, H. Feussner, S. T. Kim, and N. Navab, "Tecno: Surgical phase recognition with multi-stage temporal convolutional networks," in *International conference on medical image computing and computer-assisted intervention*, pp. 343–352, Springer, 2020.

[3] X. Gao, Y. Jin, Y. Long, Q. Dou, and P. Heng, "Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer," *CoRR*, vol. abs/2103.09712, 2021.