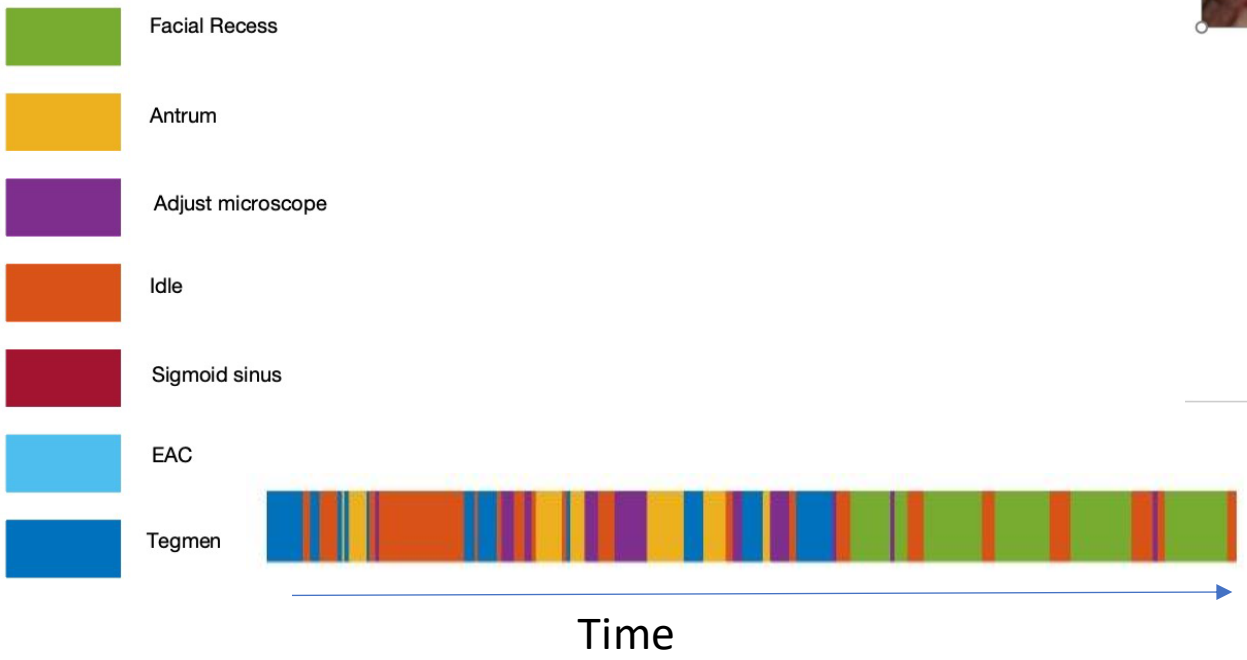
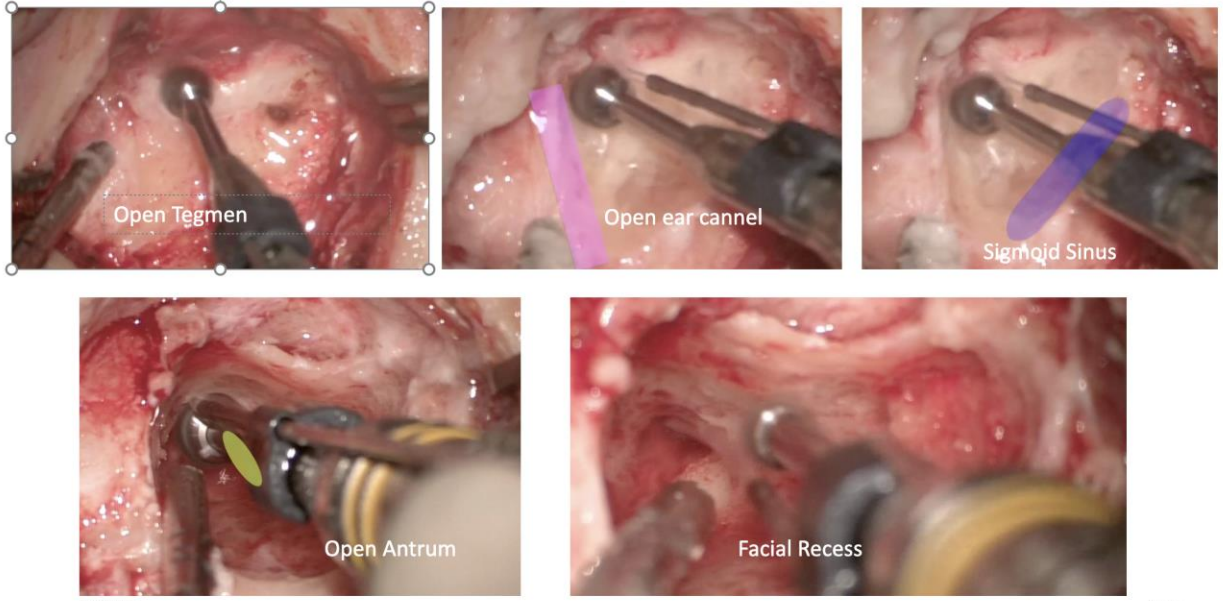
Brief Recap of our Project:

Our Task:

Mastoidectomy Surgical Video Segmentation using DL method.



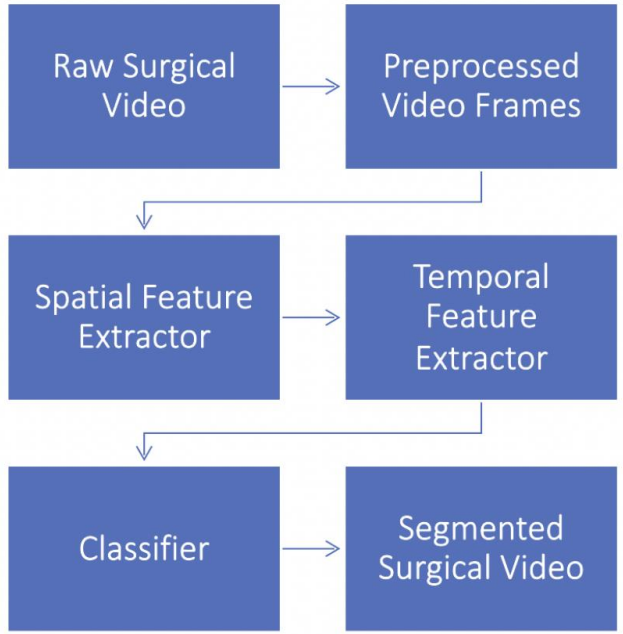
Phases:



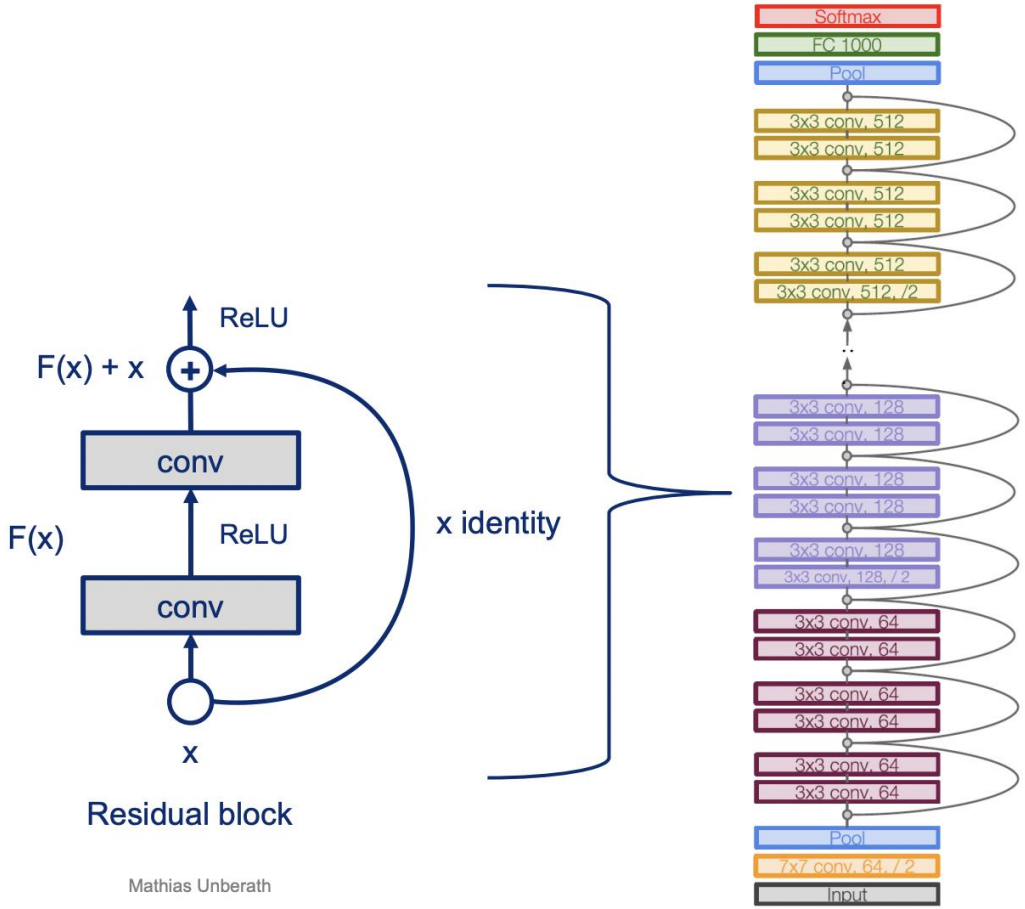
Recap Continue:

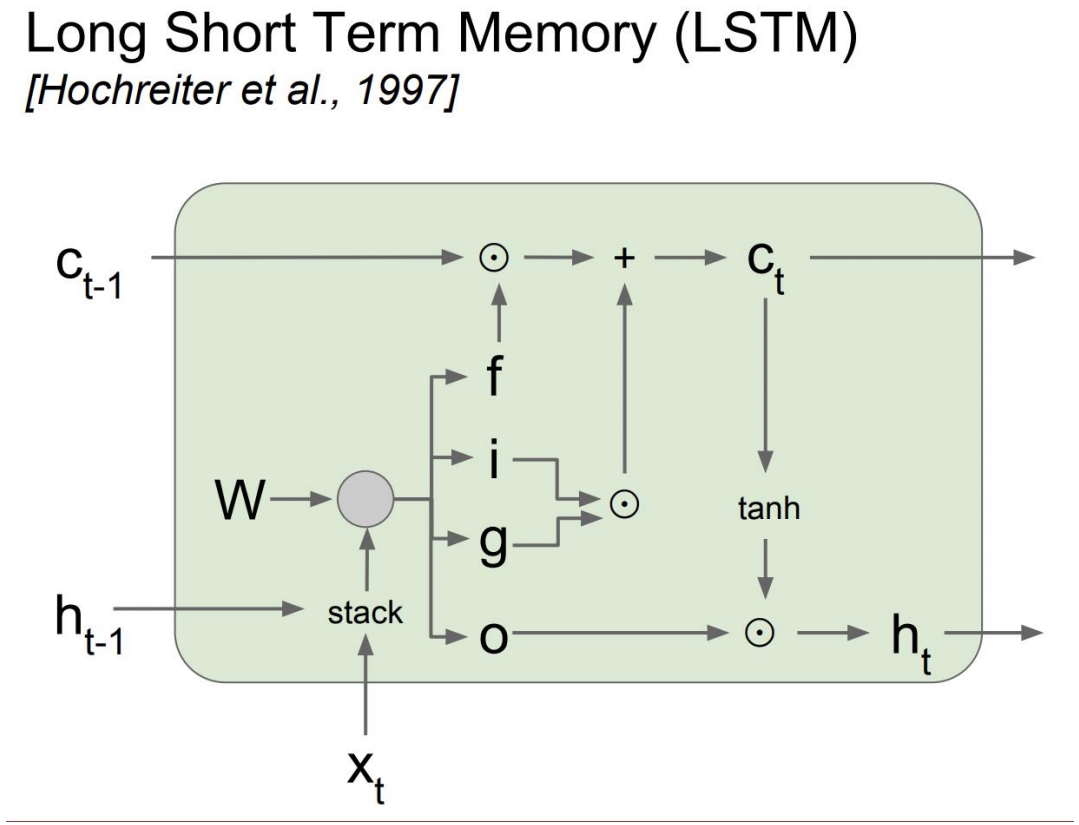
- **Project focus:**
  - Spatial Feature Extractor (anatomical structures (rigid), tool presence/positioning)
  - Temporal Feature Extractor (anatomical changes (depth), tool movement..)
  - Spatial-Temporal Feature Fusion
  - Classifier in Feature Space
- **Today's presentation:**

  Three benchmarking papers on Cholecystectomy
  - SV-RCNet (ResNet + LSTM)
  - TeCNO (ResNet + TCN)
  - Trans-SVNet (ResNet + TCN + Transformer)



1. SV-RCNet (2018) :ResNet and LSTM





1. SV-RCNet (2018) : ResNet + LSTM

Comparing to previous methods, why this paper is significant:

	Previous Networks	SV-RCNet
Spatial Features:	Shallow CNN (AlexNet)	The first paper uses a **Deep** CNN (ResNet) to extract discriminative visual features from the video frames.
Temporal Information:	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.	learn the temporal dependencies by utilizing the long short term memory (LSTM) network.
Spatial-Temporal relation:	Visual features are unable to play a role in the temporal model and therefore such a scheme hardly benefits from the spatial-temporal information.	Integrates the ResNet and the LSTM network, so that they are jointly trained in an end-to-end manner

1. SV-RCNet (2018) :Result and Problems:

State of the art performance on Cholec80 dataset. (2018)

TABLE VI
PHASE RECOGNITION RESULTS OF DIFFERENT METHODS ON CHOLEC80 DATASET

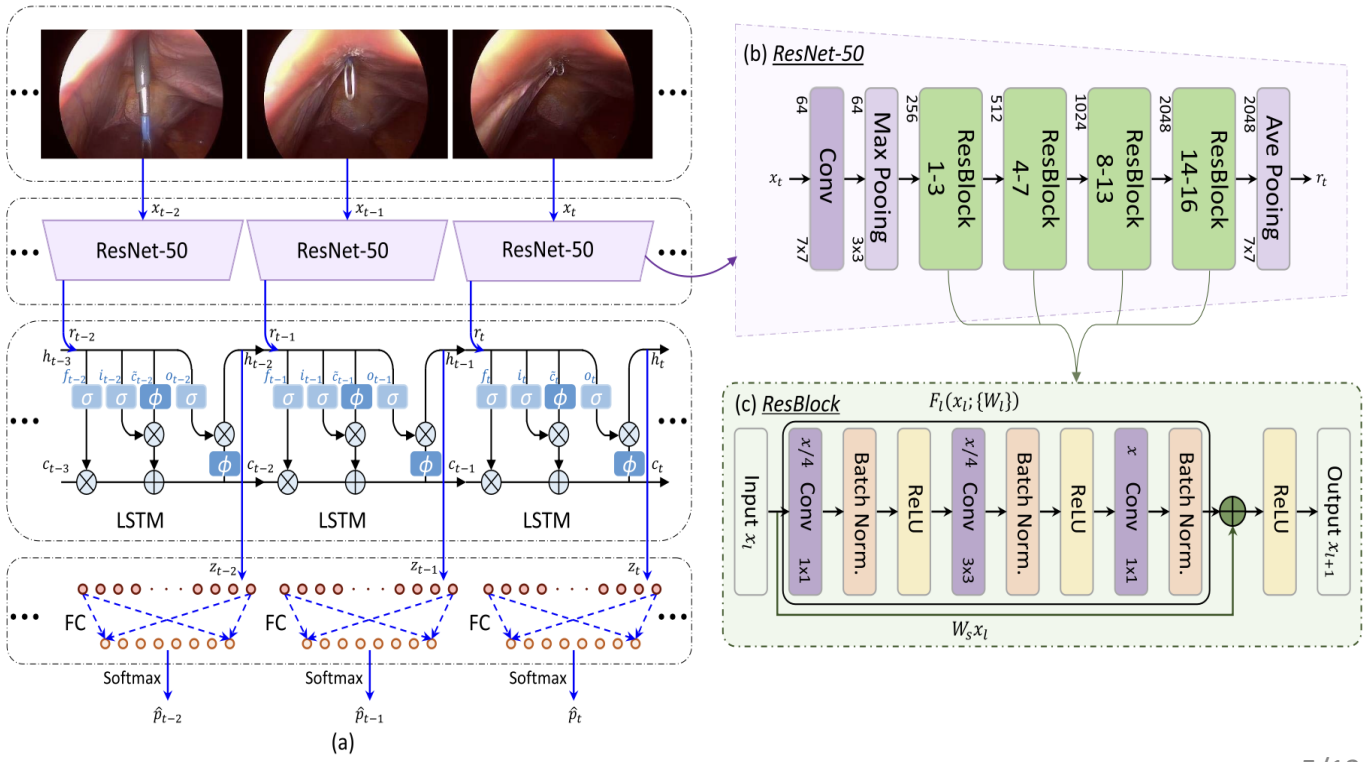
Methods	Precision (%)	Recall (%)	Accuracy (%)
SV-RCNet+PKI	$\mathbf{90.6 \pm 8.1}$	$\mathbf{86.2 \pm 15.3}$	$\mathbf{92.4 \pm 5.2}$
SV-RCNet	80.7 ± 7.0	83.5 ± 7.5	85.3 ± 7.3
EndoNet [11]	73.7 ± 16.1	79.6 ± 7.9	81.7 ± 4.2
PhaseNet [11]	71.3 ± 15.6	76.6 ± 16.6	78.8 ± 4.7

[1]

Main limitation comes from LSTMs:

1. Still cannot deal with videos that span minutes or hours.

2. Thus, 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.



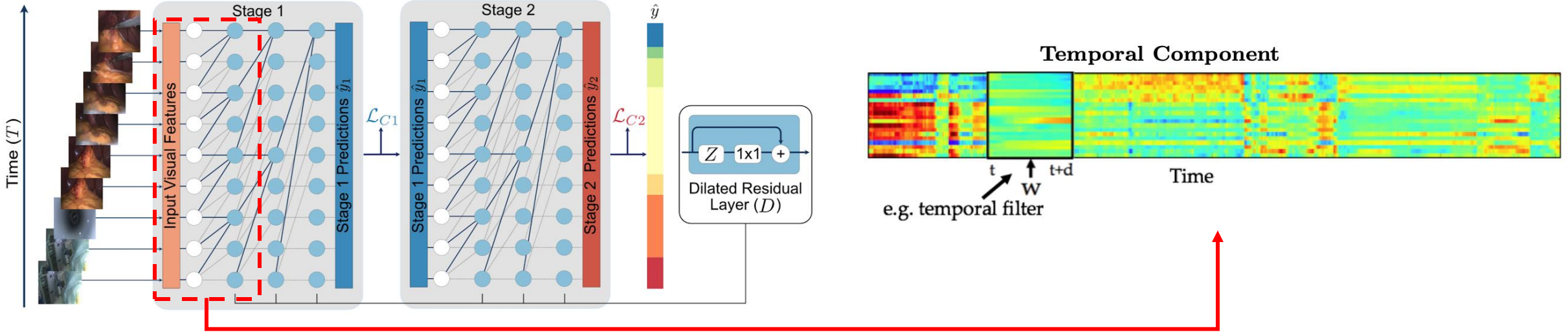
2. TeCNO (2020): Problem & Method

Problem to be addressed:

- Long-range temporal dependencies might be crucial for accurate segmentation.
- LSTM has difficulties capturing long-term temporal patterns.

Proposed Solution:

- Use Multi-Stage Temporal Convolutional Network as temporal feature extractor.

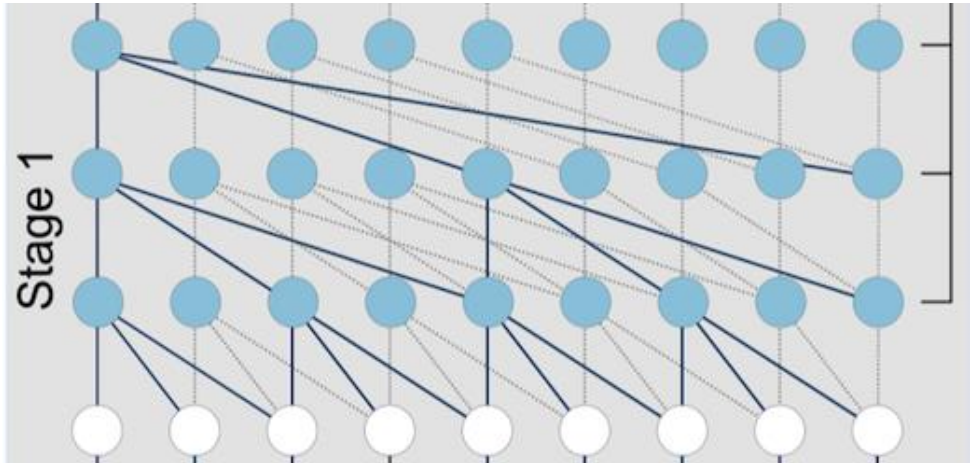


[2] Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S. T., & Navab, N. (2020, October). Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In International conference on medical image computing and computer-assisted intervention (pp. 343-352).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

2. TeCNO (2020): Significances & Results

Significances:

1. Larger receptive field than LSTM
2. Multistage TCN for better segmentation refinement.
3. Less amount of parameter than LSTM, which is suitable for online application.



	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	**88.56 ± 0.27**	**81.64 ± 0.41**	**85.24 ± 1.06**	**87.34 ± 0.66**	**75.87 ± 0.58**	**77.17 ± 0.73**

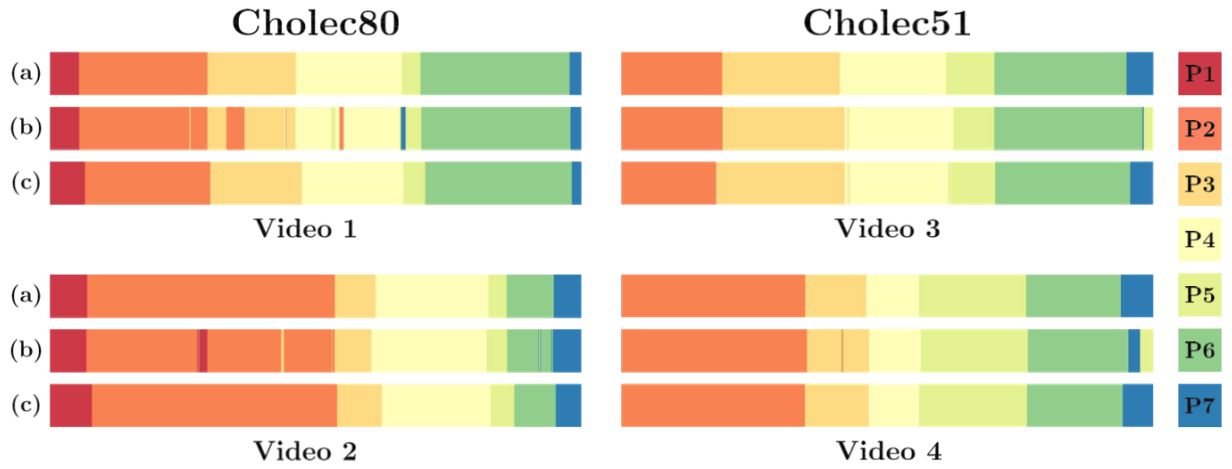


Fig. 2. Qualitative Results regarding quality of phase recognition for Cholec80 and Cholec51. (a) Ground Truth (b) ResNetLSTM Predictions (c) TeCNO Predictions. P1 to P7 indicate the phase label.

[2] Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S. T., & Navab, N. (2020, October). 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, Cham.

2. TeCNO (2020): Assessment & Future work

Good points:

- Introduce a new temporal convolution network into this task and explore the potential of TCN.
- The paper consider the TCN as **segmentation refinement** rather than temporal feature extractor.

Debatable points:

- Author consider the over-fitting issue as the main reason for slight improvement.
- No comparison with other TCN model

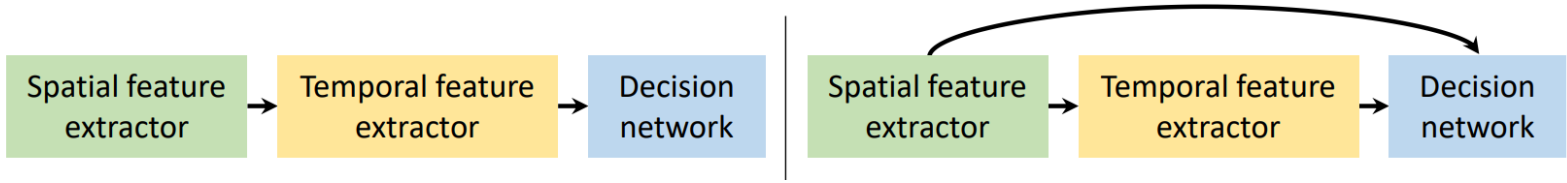
Future work:

- Potential of temporal convolution network. (Is there another reasonable TCN architecture)
- Is there another way to utilize the spatial feature and temporal feature?

3. Trans-SVNet(2021): Problem & Method

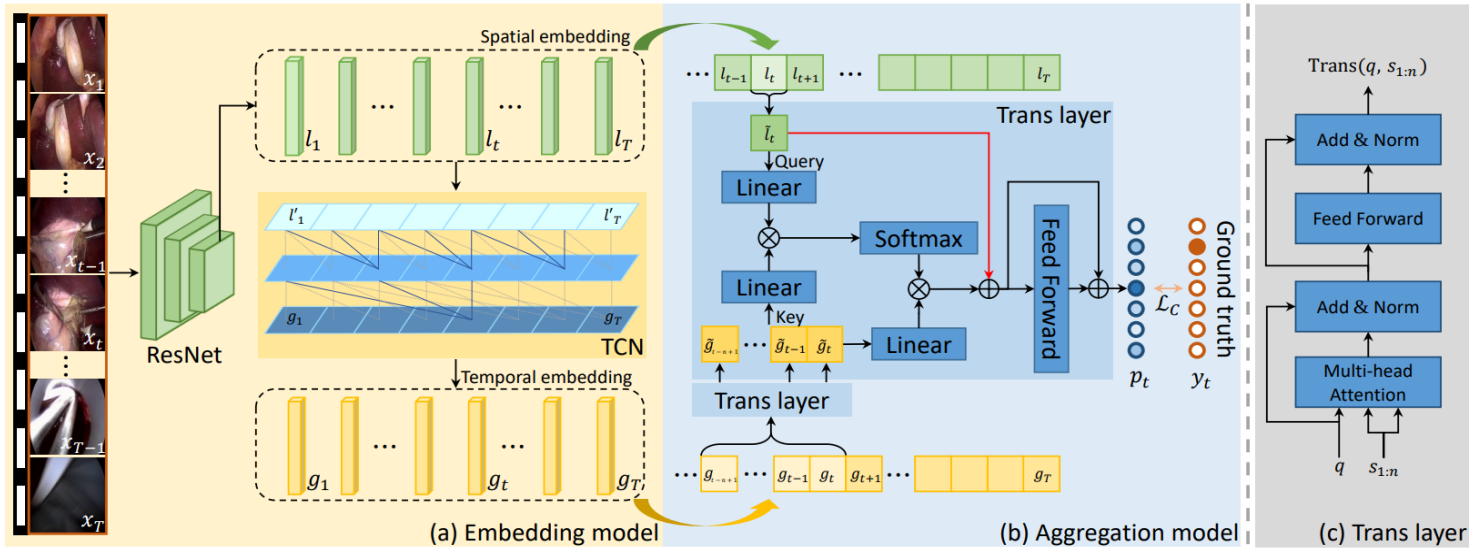
Problem Addressed:

Spatial and Temporal feature extractors are run **in serial**, only Temporal feature are used in decision



Left: previous works Right: Trans-SVNet[3]

Trans-SVNet:



Trans-SVNet architecture[3]

3. Trans-SVNet(2021): Contributions & Results

Contributions

- Brought in **Transformer** to reconsider **the Complementary Effects** of spatial and temporal features
- Achieved a boost in **Inference Speed** by processing hybrid embeddings **in parallel**

Experiment Results on Cholec80 and M2CAI16

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
1. 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
2. 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

Trans-SVNet experiment results[3]

3. Trans-SVNet(2021): Assessment & Future Work

Good points:

- **Attention mechanism** provides a good way to **fuse** spatial and temporal features
- **Low-dimensional** video embeddings result in significant inference speed boost

Debatable points:

- Embedding model (2 feature extractors) and aggregation model (transformer layer) are trained **separately**
- No comparison between results from **same** architecture with **different spatial and temporal feature extractors**

Future work:

- Compare Trans-SVNet with **end-to-end** trained models and test other feature extractors
- Treat "idle" and "camera adjustment" as different **tasks** instead of **labels**, use **transformer** and **cross-attention** to improve surgical phase labeling

Thank you!

Q & A

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] Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S. T., & Navab, N. (2020, October). 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, Cham.

[3] Xiaojie Gao, Yueming Jin, Yong-Hao Long, Qi Dou, and Pheng-Ann Heng. Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer.CoRR, abs/2103.09712, 2021.