Predicting Remaining Cataract Surgery Duration

Group (CCSZ) :

Yuqing Cao (yc3998)
yc3998@columbia.edu

Chengbo Zang (cz2678)
cz2678@columbia.edu

Fengyang Shang (fs2752)
fs2752@columbia.edu

Abstract—This project reproduced the basic structure and workflow of *CataNet* for the prediction of remaining surgical durations (RSD) proposed by the original paper. Necessary modifications were made with respect to type of baseline feature extractors in order to compare the performances, and the surgery videos of hernia reduction from Assignment 2 were then added to test the generalization ability of the model. It can be observed from the experiments that while a CNN+LSTM structure performs well for RSD prediction, surgical phase segmentation can also be obtained as a reasonable auxiliary information of the model. The codes of this project can be found here.

Keywords—CataNet, SV-RCNet, Remaining Surgical Duration, Surgical phase recognition, DenseNet, LSTM

I. INTRODUCTION

In the paper [1], the authors introduced a novel network for Cataract surgery [3] to estimate the remaining surgical duration (RSD) jointly with other two crucial elements: the surgeon's experience and the phases of surgery. In the training stage, authors incorporated the surgeon's experience, surgical phases and elapsed time of the surgery, utilizing the temporal information to increase the performance. In this way, CataNet is reported to outweigh the previous networks [4-13] in predicting the remaining surgical durations and the surgical phases.

This project conducts various experiments based on the original paper to explore the applications of deep learning tools in surgery using videos. First, it is aimed to reproduce the main outputs of the original paper. In addition, different feature extractors are implemented to the model, in order to compare their performances with the original paper.

Furthermore, inspired by CataNet, a modified model is applied to the surgery of hernia reduction to infer its phases and remaining surgical duration. And quantitative evaluations are included to study whether it could outperform the network used in [2] previously. Moreover, improved data preprocessing techniques such as weighted sampling and video I/O are utilized for the revised model.

II. APPROACH

A. CataNet Architecture

In the original paper, CataNet consists of a CNN as the feature extractor which maps the input tensor x_t to a feature vector, seamlessly followed by a RNN that leverages the benefits of video datasets.

As the main improvement, the original paper incorporated three crucial elements in the training process: the surgeon's experience, the current surgical phase and the elapsed time of the surgery (Fig.1). And they are incorporated in different ways.

Elapsed time seen as the additional input information learnt by the CNN is appended directly as an additional input channel in the frames. And it is scaled to the range [0, 1] by dividing the max length of the expected videos. The surgeon's experience and the current surgical phase which cannot be inferred from the videos are the labels that need to be trained by models. Thus, they are connected with the label information and the model is trained to solve multi-task problems.

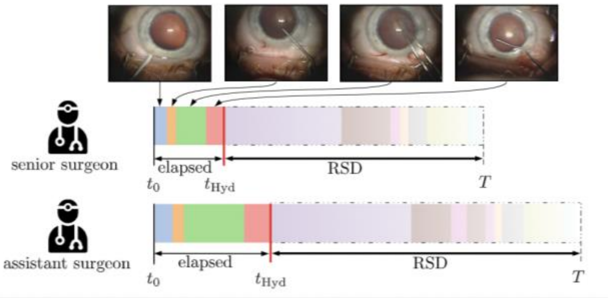


Fig.1 surgeon's experience, elapsed time and RSD.

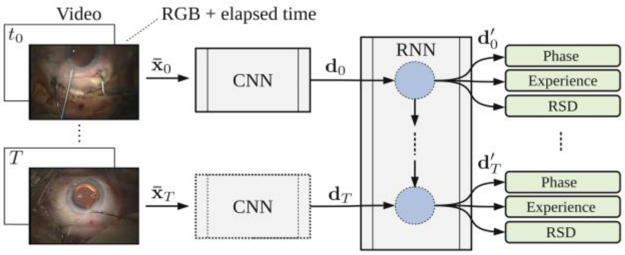


Fig.2 CataNet Architecture proposed in the paper [1]

Fig.2 illustrates the model and how to incorporate three elements into it. To be more specific, the input data of the model is like $x_t = [x_t, \frac{t}{T_{max}}]$, which contains both the original RGB channels and one additional channel of elapsed time of the frame at time step t. After being processed by CNN, the input vector x_t generates feature vector d_t and as the input data that is sent to the RNN. Also shown in Fig.2, the feature vector will be passed to the LSTM and become the descriptor vector $d_t{}'$ produced by LSTM. Such descriptor vectors are finally processed through three independent fully connected layers to predict the surgeon's experience, the surgical phase and the RSD.

B. End-to-End Learning

In the whole training state, CataNet is trained using an End-to-End manner so as to make full use of the temporal information. The parameters of CNN and LSTM are trained partly together, in order to combine the visual information and temporal information improving the performance. Different from previous networks TimeLSTM and RSDNet, the CNN of CataNet is trained for predicting phase recognition and the surgeon's experience in independent frames extracted from raw video frames concatenated with elapsed time in 3 epochs.

CNN model appended two temporary fully connected layers is used to make predictions without RNN. Given the predictions of the CNN and the ground truth labels of cataract surgery, the loss which needs to be minimized combined with cross-entropies is formulated as below:

$$\ell_{\text{cnn}} = \mathrm{H}(\hat{\mathbf{y}}_{\text{cnn},t}^{\text{phase}}, y_t^{\text{phase}}) + \mathrm{H}(\hat{\mathbf{y}}_{\text{cnn},t}^{\text{exp}}, y_t^{\text{exp}})$$

In the original paper, authors notice that the number of frames of each phase in the cataract surgery videos is imbalanced. So in the loss function, the original paper uses weights parameters to tackle such imbalance. Inspired by this notion, both smooth sampling of the frames in the reduction hernia dataset and weight sampling in the loss function were experimented.

After pre-training CNN, the model needs to learn from the temporal information and make predictions of RSD by training RNN without training CNN. The loss of RNN is combined with cross-entropies for both the experience of surgeons and the phase recognition and l_1 norm for RSD predictions. It is formulated as below:

$$\ell_{\text{rnn}} = \alpha \left| \hat{y}_t^{\text{rsd}} - y_t^{\text{rsd}} \right| + \mathrm{H}(\hat{\mathbf{y}}_t^{\text{phase}}, y_t^{\text{phase}}) + \mathrm{H}(\hat{\mathbf{y}}_t^{\text{exp}}, y_t^{\text{exp}})$$

With the pertained CNN and RNN, the whole model was then trained in end-to-end manner. And in the project, truncated back-propagation on sub-sequences for 10 epochs were adopted as the original paper did.

Finally, freezing the parameters of CNN again and do another 20 epochs training for RNN. In this part, the loss function of RNN applies l_2 norm to reduce overfitting problems, and set α = 1.

III. METHODOLOGY

A. Data Preprocessing

It is proposed in the original paper that time elapsed should be considered a key factor in surgical phase classification and RSD prediction. This can be simply deduced by frame number and frame rate, which is given by

$$\bar{x}_t{}^{W\times H\times C} = [x_t, 1\frac{t}{T_{max}}]^{W\times H\times (C+1)}$$

where T_{max} is the total length of the video, normalizing the feature to the range of [0,1] for the consistency with other channels (RGB).

Labels should then include surgical phase, surgeon's experience and RSD, which reads

$$\bar{y}_t = [y_t{}^{phase}, y_t{}^{exp}, y_t{}^{rsd}]$$

B. Dataset Sampling

According to the experiments of Assignment 2, class-wise precision-recall can be significantly low when it comes to short phases that don't have enough training samples. During the training of CNN backbone, all frames are sampled with respect to their corresponding phases before being put to the model. The frequency of each class is defined by

$$f_i = \frac{n_i}{N}$$

where n_i represents the number of frames in class i, $N = \sum n_i$ is the total number of frames in all videos. The sampling weight follows as $w_i = f_i{}^{-1}$ for the i-th class.

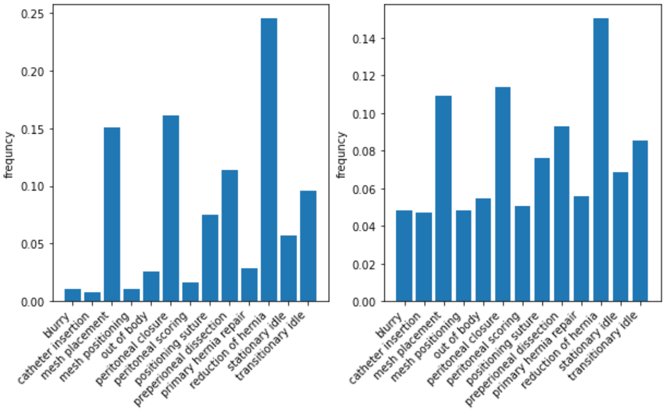


Fig.3 Phase distribution for hernia dataset

Fig.3 shows the distribution of frame number across each phase where the contrast is observed to be drastic (having x frames in phase x while x in x). This results in moderately increased performance in minor classes and dramatic decrease in major classes. Laplace smoothing was conducted to all phases using

$$w_i{}' = (f_i{}')^{-1} = (\frac{f_i + \beta}{N + \alpha K})^{-1}$$

where K denotes the number of classes, which preserves the relative proportions of all phases controlled by a smoothing factor β (chosen to as $\beta = 0.1$ in the following experiments). The figure below shows the weights of each phase after smoothing.

C. Weighted loss

The class-wise imbalance of frame number would also impair backpropagation, for the loss of majority phases can easily dominate the overall loss, leading to insufficient training of minor phases. In the experiments, all cross-entropy losses are weighted by the number of samples in corresponding classes, which is given by

$$H(y, \hat{y}, p) = -\sum p_i y_i \, log \, \hat{y}_i$$

The training of the entire network structure (CNN and RNN) requires both cross entropy loss for surgical phases and L_1 loss for RSD. The latter is multiplied with a constant to form the final loss which reads

$$l_{hernia} = H(y_{phase}, \hat{y}_{phase}, p) + \alpha|y_{rsd} - \hat{y}_{rsd}|$$

D. Training stages

Considering the variations between CNN and RNN structure, the training process was conducted in several stages following the methodologies of the original paper:

a) Sample datasets, train CNN (3 epochs) with l_{cnn}

b) Freeze CNN, train RNN (50 epochs) with l_{rnn}

c) Train entire model (10 epochs) with l_{rnn}

d) Freeze CNN, fine-tune RNN (20 epochs) with l_{rnn}

IV. EXPERIMENTS

A. Datasets

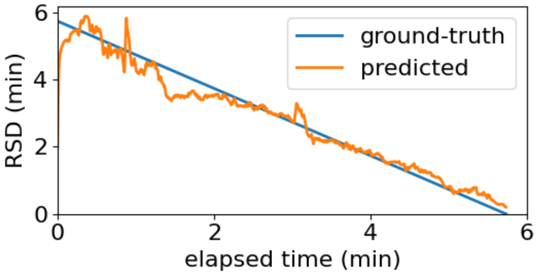
The experiments used two different datasets. Firstly, CataNet was used on the cataract-101 dataset which contains 101 videos with a resolution of 720 * 540 pixels acquired at 25 fps. All the frames extracted from the videos can be categorized into 10 phases. Additionally, the videos can be split into two parts, 56 senior surgeons and 45 assistant surgeons. The dataset is also randomly split into a training set with 81 videos and a test set with 20 videos. Also in the training set, frames are split randomly to a training set and a validation set.

The other dataset is surgery of hernia reduction supported by the class. The dataset consists of 70 videos recording the hernia reduction. These videos are acquired at 1 frame per second and the resolution of each frame is 854 × 480. And all these images can be categorized into 14 phases. Such labeled videos were divided into the training set (60 videos) and the validation set (10 videos).

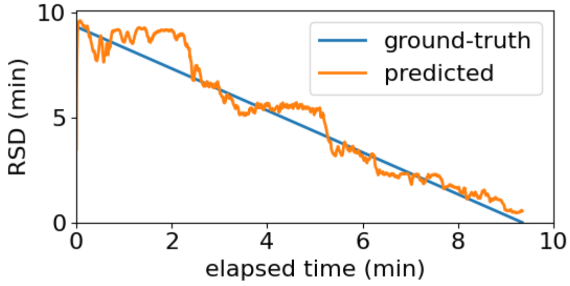
B. Results

1) Cataract RSD Inference

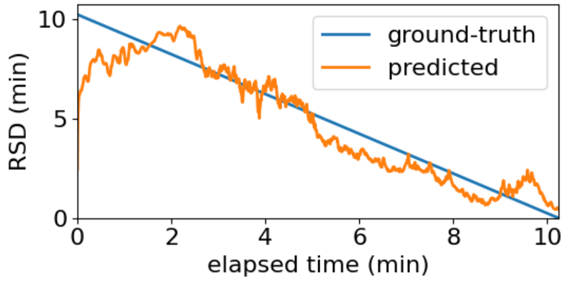
As is shown in Fig. 4, the model of CataNet has been successfully reproduced to achieve a relevant good performance in view of the task for inferring RSD. For the purpose of demonstration, only some examples from the results were chosen in this report. Fig.4 (a), (b), (c) and (d) present the plots of RSD for patient ID as 830, 889, 899 and 929 respectively. It could be seen that for the majority of the surgery workflow, the model could predict a precise RSD (shown as the orange line), which is quite close to its ground truth (shown as the blue line).



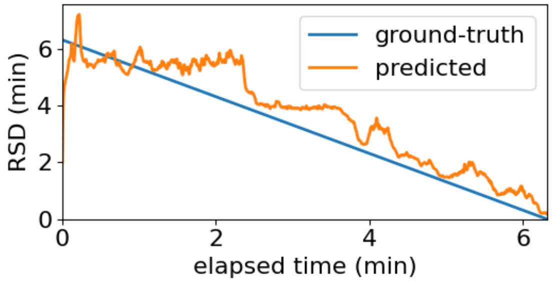
(a)



(b)



(c)



(d)

Fig.4 Example plots for relations of RSD and elapsed time

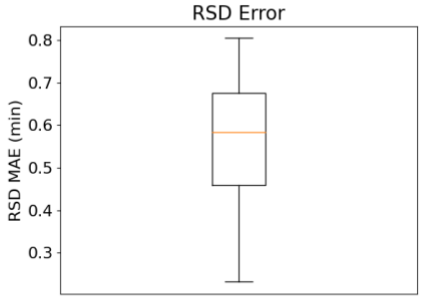


Fig.5 The RSD error over all datasets

In addition to several specific plots of part of the dataset, it is also necessary to evaluate the model over the entire dataset. Table 1 gives some insight into the evaluation of the overall results. The inference results are evaluated sequentially for RSD As is shown in Table 1, the mathematical mean and standard deviation of the absolute eros of RSD are computed for workflow when the remaining duration is below 2.0, below 5.0 and all (shown in the second, third and fourth rows). And the fifth row of Table 1 records the mean and standard deviation of the workflow duration. It could be calculated that the mean relevant error of the RSD is around 5%, which is similar to the original paper of 1.11 ± 0.62, indicating a good capability of inferring RSD over all patients. And Fig.5 acts as a visualization of the performance

Table 1. Macro Average RSD

	Mean	Std
RSD 2	0.421632	0.237285
RSD 5	0.521985	0.257030
RSD All	0.551066	0.240639
Duration	7.906667	2.217286

2) Cataract Phase Recognition

Experiments were conducted on segmenting surgical phases for cataract surgery as a part of RSD prediction.

Fig.6 shows four different examples of phase prediction with regard to patient ID as 830, 889, 899 and 929, which also corresponds to the confusion matrices given in Fig.7 (a) - (d). The x-coordinates in Fig.6 correspond to total time elapsed, and the color represents different surgical phases as illustrated in the color bar on the right. Fig.7 provides a visualized result of classification by specifying the frequency of assigning a frame with some prediction (x-axis) that is supposed to have some ground truth (y-axis).

It can be observed that the majority of the samples appear along the diagonal, which corresponds to a correct phase assignment. The “Phaco” step accounts for most of the samples for being the longest lasting phase during the entire surgery.

Experiments revealed that even if phase classification is not the first priority of this model, it is still reasonable to expect this auxiliary information to be useful in reality.

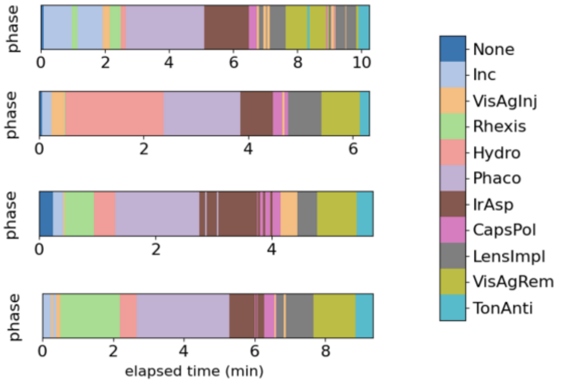
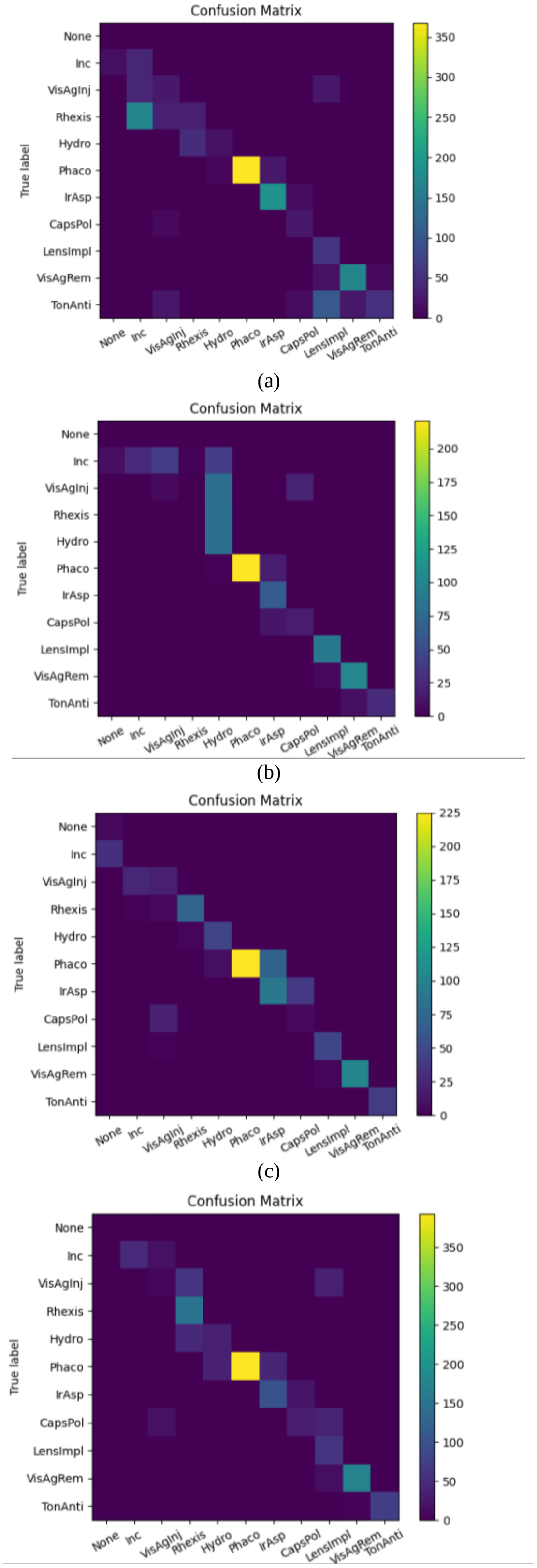


Fig.6 Examples of the prediction results of phases



(a)

(b)

(c)

(d)

Fig. 7 Phase confusion matrix for different patients (Patient ID as 830, 889, 899 and 929)

3) *Hernia RSD Inference*

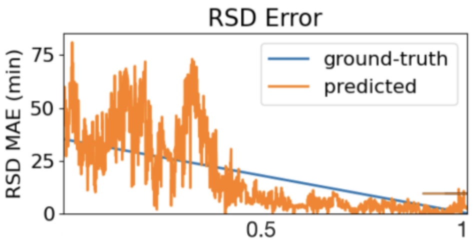
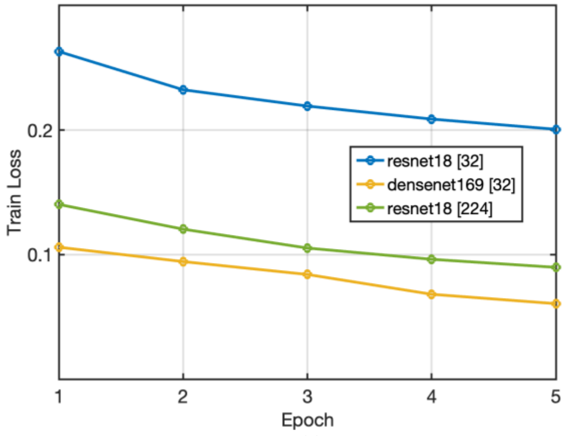


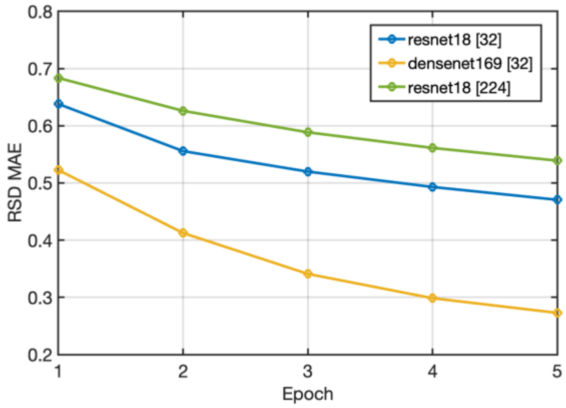
Fig.7 Hernia RSD Inference Example

Subsequently, to further explore the efficiency of the CataNet model, such as its generalization ability, our model is modified to be implemented on Hernia Dataset in Assignment 2 for the task of RSD inference. The overall model structure follows the network construction of CataNet. Inspired by [1], the original input vector of Hernia dataset has been contacted with a dimension of elapsed time (scaled to [0,1]) to enhance the model ability for predictions of RSD. And several useful techniques mentioned in III, such as weighted loss, sampling and has also been applied to leverage the model performance. Similar to CataNet training stages, the model is first pret-rained on its feature extractor, using the loss of phases to backpropagate. And then the entire network is trained with the loss of phases and the mean squared error of RSD. After some fine-tuning steps for hyperparameters, one inferred example could be seen from Fig.7. It indicates that the model could give a fair inference result for RSD on Hernia dataset, although the performance is not so good as it on the Cataract dataset. This also makes sense for reasons that Hernia surgery task is more difficult than the Cataract one, and that the quantity of Hernia dataset is not as sufficient as it of Cataract. It could be expected to achieve better results with more effort in hyperparameter tuning and training in the future.

4) *Different Feature Extractors for Hernia*



(a)



(b)

Fig.8 (a) Plots of training loss of RSD and phases for models with different feature extractors based on Hernia dataset. (b) Plots of RSD MAE for different models based on Hernia dataset.

To reduce the load of computation and time consumption in the training stage, ResNet-18 was chosen as a simpler baseline feature extractors in part of the experiments. This section serves to analyze the effect of different feature extractors and input image size on model performance in terms of overall training loss and MAE specifically for RSD prediction.

Fig.8 (a) depicts the change of overall training loss defined in Section III (C). The training losses were calculated on the same training videos with identical batch size of 32 and learning rate of 1e-4 respectively on ResNet-18 (input size of 32 and 224) and DenseNet-169 (input size of 32). There's a distinguishable decrease in loss value both for increased model capacity and input size, indicating that increasing the complexity of baseline model as well as visual clues has the potential of yielding better performance even under insufficient training.

However, a similar pattern with input size was not observed in terms of MAE of RSD prediction, where the performance of ResNet-18 with input size of 32 and 224 produced results of high resemblance, while model capacity still shows positive correlation with better prediction accuracy.

IV. DISCUSSIONS & CONCLUSIONS

This project reproduced the model structure of CataNet, conducted necessary modifications on model structures as well as hyper parameters, and conducted experiments under different model capacities competing their performances. In addition, extended experiments into the Hernia dataset in Assignment 2 were performed to find out the potential of generalization of the model.

From the experiments conducted, several insights can be gained to figure out further improvements both on the model structures and data processing pipelines.

Firstly, considering the elapsed time and the experience as the input information and ground truth serves to improve the performance of predicting RSD and phase recognition based on the limited training with both Cataract and Hernia dataset. It is worth noting that normalization is crucial when it comes to incorporating different types of data, as has been observed in experiments that poorly combined data would most likely leads to disastrous outcomes.

Secondly, the input shape of the model was cut from 224 to 32 due to insufficiency of computing resources. This influenced the performance of both the phase and RSD prediction across all datasets (as has been studied in Section IV (B-4) for details). In general, improvement of model performance is believed to be correlated to reasonably larger input size and more complex models, indicating that the results obtained are far from touching the limit of the current model.

Finally, different data preprocessing techniques have been implemented especially including weighted sampling. No sufficient experiments were conducted to validate the effectiveness of this technique. Additionally, converting videos to frames before model training has been considered a standard procedure for problems of this type. This leads to a more straightforward preprocessing pipeline while being significantly more memory consuming at the same time. An implementation of extracting and sampling frames directly from the videos with the help of OpenCV was provided by this project. However, inconsistency issues with multi-processing hinders the actual application, leaving the problem for further study.

REFERENCES

1. Marafioti, A., Hayoz, M., Gallardo, M., Márquez Neila, P., Wolf, S., Zinkernagel, M., & Sznitman, R. (2021, September). CataNet: Predicting remaining cataract surgery duration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 426-435). Springer, Cham.
2. Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C. W., & Heng, P. A. (2017). SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network. *IEEE transactions on medical imaging*, *37*(5), 1114-1126.
3. K. Schoeffmann, M. Taschwer, S. Sarny, B. Mu¨nzer, M. J. Primus, and D. Putz- gruber, "Cataract-101 - Video dataset of 101 cataract surgeries," Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, pp. 421–425, 2018.
4. A.Achiron,F.Haddad,M.Gerra,E.Bartov,andZ.Burgansky-Eliash,"Predicting cataract surgery time based on preoperative risk assessment," European Journal of Ophthalmology, vol. 26, no. 3, 2016.
5. S. P. Devi, K. S. Rao, and S. S. Sangeetha, "Prediction of surgery times and scheduling of operation theaters in ophthalmology department," Journal of Medical Systems, vol. 36, no. 2, pp. 415–430, 2012.
6. M. Lanza, R. Koprowski, R. Boccia, K. Krysik, S. Sbordone, A. Tartaglione, A. Ruggiero, and F. Simonelli, "Application of artificial intelligence in the anal- ysis of features affecting cataract surgery complications in a teaching hospital," Frontiers in Medicine, vol. 7, 2020.
7. N. Padoy, T. Blum, H. Feussner, M. O. Berger, and N. Navab, "On-line recognition of surgical activity for monitoring in the operating room," in Proceedings of the National Conference on Artificial Intelligence, vol. 3, 2008.
8. S. Franke, J. Meixensberger, and T. Neumuth, "Intervention time prediction from surgical low-level tasks," Journal of Biomedical Informatics, vol. 46, no. 1, 2013.
9. A. C. Gu´edon, M. Paalvast, F. C. Meeuwsen, D. M. Tax, A. P. van Dijke, L. S. Wauben, M. van der Elst, J. Dankelman, and J. J. van den Dobbelsteen, "'It is Time to Prepare the Next patient' real-time prediction of procedure duration in laparoscopic cholecystectomies," Journal of Medical Systems, vol. 40, no. 12, 2016.
10. N. Spangenberg, M. Wilke, and B. Franczyk, "A big data architecture for intra- surgical remaining time predictions," in Procedia Computer Science, vol. 113, 2017.
11. M. Maktabi and T. Neumuth, "Online time and resource management based on surgical workflow time series analysis," International Journal of Computer Assisted Radiology and Surgery, vol. 12, no. 2, 2017.
12. S. Bodenstedt, M. Wagner, L. Mu¨ndermann, H. Kenngott, B. Mu¨ller-Stich, M. Breucha, S. T. Mees, J. Weitz, and S. Speidel, "Prediction of laparoscopic pro- cedure duration using unlabeled, multimodal sensor data," International Journal of Computer Assisted Radiology and Surgery, vol. 14, no. 6, 2019.
13. I. Aksamentov, A. P. Twinanda, D. Mutter, J. Marescaux, and N. Padoy, "Deep neural networks predict remaining surgery duration from cholecystectomy videos," in Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne,Eds. Cham: Springer International Publishing, 2017, pp. 586–593.

MEMBERS

Name	UNI	Task
Yuqing Cao	yc3998	Training process Model transfer
Fengyang Shang	fs2752	Model construction Fine tuning
Chengbo Zang	cz2678	Data processing Model transfer