Improved-STCN Network with Enhanced Strategy for Sequence Polyp Segmentation

Quan He[1], Xiaobo Hu[1], Feng Sun[1], Lulu Zhou[1], Jing Wang[1] and Qiming Wan[1]

[1]*Hangzhou Hikvision Digital Technology Co.,ltd, Hangzhou, China*

Abstract

The detection of polyps is helpful to the diagnosis of early colorectal cancer. With the rapid development of deep learning, more and more researchers apply detection and segmentation technology to assist polyp detection. This work is our solution to the polyp segmentation subtask in the EndoCV2022 challenge. We come up with the idea from the semi-supervised video object segmentation and build on STCN [1] for this challenge. STCN is built for the task when the correct segmentation mask of the first frame of the video is given as input, then the model just tracks the target, no matter what it is. We modify STCN into a sequence polyp segmentation network named improved-STCN, which can not only segment the polyps but also track the polyps. As EndoCV2022 challenge [2] [3] is a sequence challenge, the images in the same sequence are very similar, which will lead to bad performance. Thus, we adopt semi-supervised learning to get more abundant data for training. We also carry out experiments on how to make the segmentation results more credible, that single frame detection and reverse sequence information will help in this part. Finally, on the round-II test, our system achieves a segmentation score of 0.7654 and ranked the second.

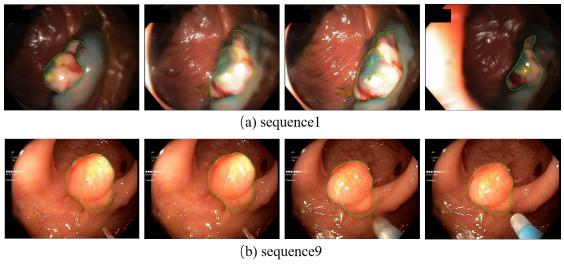
Keywords

Polyp segmentation, Sequence data, Deep learning, Semi-supervised learning, Improved-STCN

1. Introduction

Colorectal cancer (CRC) is a common malignant tumor in the gastrointestinal tract. Its incidence rate and mortality rate are the second most important in digestive system cancer, followed by gastric cancer, esophageal cancer and primary liver cancer. Polyp is considered a sign of precancerous lesions, thus, finding it at any time during precancerous lesions and blocking it not only reduce the mortality of colorectal cancer, but also reduce the incidence rate. Colorectal lesions are usually diagnosed by colonoscopy, but unfortunately, it is estimated that about 6-27% of pathological missed diagnosis in colonoscopy [4]. Colonoscopy image analysis and decision support system have shown great potential in improving examination efficiency and reducing the number of missed lesions [5]. Deep learning is more and more widely used in the field of medical images. Since MICCAI 2015 Automatic Polyp.

Detection in Colonoscopy Videos challenge, more and more datasets and challenges have been launched, which further promote the application of deep learning-based endoscopic vision [6]. Among them, the most widely used deep learning model is Unet [7] and its variants. The Unet consists of two paths. The first path is a compression path (also known as an encoder) that captures the context in the image. The encoder is just a traditional convolution and maximum pool layer stack. The second path is the symmetric spread path (also known as the decoder), which is used for precise positioning using transpose convolution. This structure has been proved to be able to segment medical images effectively. However, for sequence data in real scenes, this kind of method can not effectively model timing information. In the field of video object segmentation, the model is trained to extract the relationship between video frames to improve the performance of segmentation. Masktrack [8] is a typical network of video object segmentation. Taking the mask of the previous frame and the current frame as the model input, the trained model will outputs the mask of the current frame with high segmentation accuracy. However, the performance of this method often depends on the accuracy of the output of the previous frame, which has the risk of cumulative error. This work is our solution to the polyp segmentation subtask in the



(a) sequence1

(b) sequence9

Figure 1: Example of EndoCV2022 challenge sequence data

4th International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2022) in conjunction with the 19th IEEE International Symposium on Biomedical Imaging ISBI2022, March 28th, 2022, IC Royal Bengal, Kolkata, India

whut2014hq@163.com (Q. He)

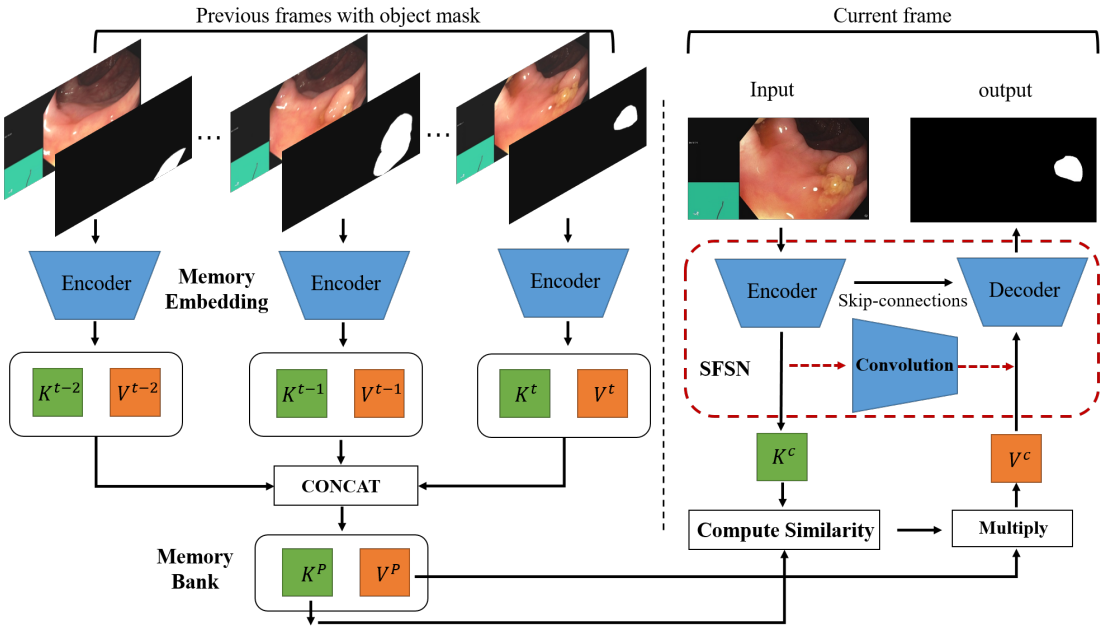


Figure 2: Overview of the improved-STCN

EndoCV2022 challenge. The proposed approach is built on STCN, a semi-supervised video object segmentation network. In particular, we modify STCN into a sequence polyp segmentation network ,which can not only segment the polyps but also track the polyps. In short, our main contribution for this work are as follows:

- We modify STCN into a sequence polyp segmentation network, which will no need the first frame's mask to predict like it used to be. And we also practice the experiment on training strategy to find a stronger model.
- We learn from semi-supervised learning to generate more training data, as the image of the same sequence have great similarity, which is not conducive to the improvement of network generalization and feature extraction ability.
- We propose an enhanced scheme to make the segmentations results more credible. Overall, our method is proved to be effective in the EndoCV2022 challenge round-I and round-II.

2. Method

2.1. Overview of the framework

Figure 2 shows the overall process of the improved-STCN. The network use ResNet50 and ResNet18 to build a key encoder and a value encoder respectively. The key encoder encodes the images into the key feature space and the value encoder encode both the images and mask into the value feature space. The key correspond with value one by one will be stored in the memory bank. Then, when a new frame in the video sequence is collected, the frame will be encoded into the key feature space firstly, and then calculate the similarity with the key features of the previous frame stored in the memory bank. The most similar features will be combined into the feature space of the current frame for model outputs. Here, the negative square Euclidean distance is used as similarity functions, which is defined as follows:

$$S = -||K^P - K^C||_2^2 \tag{1}$$

where K^P represents the previous frames' key features, K^C represents the current frames' key feature. Then the aggregated readout feature V^C for the current frame can be computed as a weighted sum of the memory features with an efficient matrix multiplication:

$$V^C = C^P.S \tag{2}$$

which is then passed to the decoder for mask generation [1].

STCN is used to meet the semi-supervised video object segmentation task where the first frame of the video is needed. We have specially improved the STCN's structure named improved-STCN for EndoCV2022 challenge.

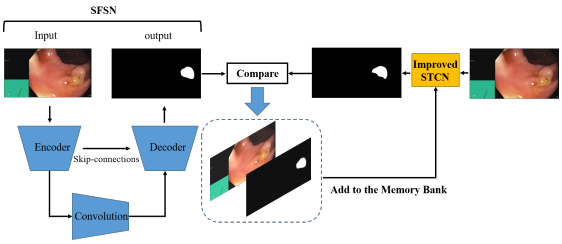


Figure 3: Overview of the enhanced scheme

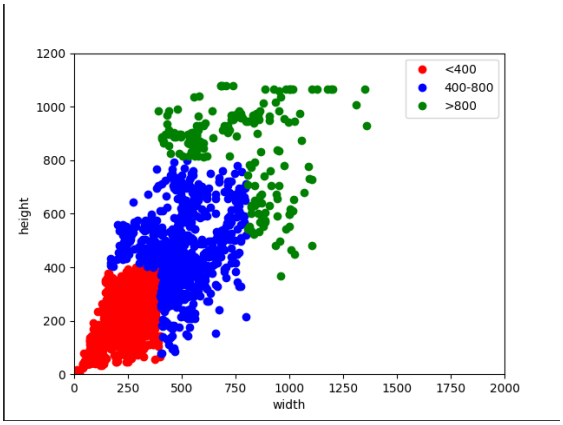


Figure 4: EndoCV2022 challenge Dataset statistical

In particular, we firstly hidden memory bank and affinity compute module, then add a convolution module to get the single frame segmentations network (SFSN), as shown in the red dashed box in the figure 2. In the training phase, we train the SFSN only to make the encoder and decoder strong. Then the parameter of SFSN will be the pre-training parameters for STCN's training. In the inference phase, for the first frame, SFSN will outputs the result firstly, then STCN will track the mask and complete the predictions of all subsequent sequences. In this way, improved-STCN build the ability of single frames' segmentation without the help of other frames. Finally, the improved-STCN can not only segment the polyps but also track the polyps that appear in the previous frame.

2.2. Semi-supervised learning

Due to the small field of vision of the endoscope and the slow movement during endoscopy, the sequence data collected over a period of time are highly approximate, as figure 1 shows. These approximate data are not conducive to the improvement of network generalization ability and feature extraction ability. We learn from semi-supervised learning to generate more training data. In practice, firstly, we use all the EndoCV2022 challenge Dataset and STCN to train the polyp tracking model. Then we manually annotate the first frame of the Hyper-Kvasir videos [9], and the polyp tracking model will generate the pseudo labels. In this way, we get more abundant sequence data with labels, which is helpful for our model's learning.

2.3. Enhanced scheme

Although the model mentioned in the Subsection 2.1 has the ability to segment and track the polyps, we find that train two models to segment and track polyps separately will get better results. As figure 3 shows, SFSN that change from STCN is used to segment the polyps in the first few frames of the sequence data. Meanwhile, STCN will also outputs the segmentation results. The results of the two models will use the same calculation method to obtain confidence, which is defined as the average value of the network output response in the segmentation target area. Then the key encoder and value encoder of STCN will encode the segmentation results with higher confidence and store the coding results in the memory bank. The prediction of all subsequent sequences will be completed next.

Sequence information is helpful for model segmentation. Usually, we use forward sequence information. As for offline diagnosis, such as capsule endoscopy diagnosis, we can take advantage of backward sequence information. Thus, we reverse the input sequence data and make the model to predict. Then, fuse the forward sequence data results and the backward sequence data results as the final output of the network. Here, fuse method is the same as the above, that is comparing the confidence in the segmentation result and select the one with higher confidence as the final result.

3. EXPERIMENTAL RESULTS

The experimental part is mainly composed of two partsnamed baseline experiments and experiments used for the challenge. In part one, the baseline experiments were used to find the suitable hyper-parameters and data augmentation strategy for the training of improved-STCN. Besides, we carried out the semi-supervised learning mentioned in the Subsection 2.2. We also explored the effects of illumination and size on model's performance. In part two, we used the same train strategy as the part one to train model with all the dataset we have, and tested model with the Endocv2022 challenge unseen dataset. The enhanced scheme was adopted to get the more credible segmentation results.

3.1. Dataset

The EndoCV2022's organizing committee provided a total of 46 sequence data for all participants. According to the statistics, the EndoCV2022 challenge Dataset consists 3348 frames sampled in the real-world clinical scenario. As figure 4 shows, most polyps are around 400 in size while a few polyps are larger than 800. Due to the different sizes of polyps and images, we need to pay attention to using some strategies to reduce the sensitivity of the network to resolution, such as Multi-scale training. Although polyps have different shapes and sizes, the image of the same sequence data have great similarity, which is not conducive to the improvement of network's generalization and feature extraction ability. Thus, in baseline experiments, we split the EndoCV2022 challenge Dataset into 80% for training and 20% for validation in sequence. To enhance the generalization and feature extraction ability of our model, we also utilized three well-known publicly endoscopy sequence datasets, ETIS-Larib Polyp [10], CVC-Clinic [11], and Hyper-Kvasir dataset. ETIS-Larib Polyp DB were used directly as a training set. CVC-Clinic were used as validation set as more data can better evaluate the generalization of the model. As HyperKvasir dataset has only video data and no labels, we adopted the method mentioned in the subsection 2.2 to generate labels. Then, these sequence data with pseudo labels were also used as a training set. In the experiments for challenge, we used the same train strategy as the baseline experiments, and trained model with all the dataset we have

3.2. Evaluation Metrics

The EndoCV2022's organizing committee provided participants a toolbox to calculate the scores between the predicted mask and the ground truth mask at github [12, 13]. There are seven metrics in the toolbox: Jaccard (Jac), Dice, F2-score, Precision (Positive Predictive Value, PPV), Recall (Rec), Accuracy (Acc), and Hausdorff distance (Hdf). As these metrics are similar, and to make experiments more efficient, we chose the most commonly used metrics for the medical image segmentation, the Jaccard and the Dice coefficient. The Jaccard is defined as follows:

$$Jac = \frac{TP}{2 * TP + FP + FN} \tag{3}$$

Where TP represents true positive "polyp", while FP and FN represents false positive and false negative respectively. Similarly, the Dice coefficient is calculated as follows:

$$Dice = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

EndoCV2022 leaderboard also chosen the Dice coefficient as the scores to evaluate the performance of the model.

3.3. Training Details

We chose PyTorch to train our model, and both the train and inference were run on the NVIDIA TESLA V100 GPU. Here, we minimized the cross-entropy loss using Adam optimizer with default momentum $\beta 1$ = 0.9, $\beta 2$ = 0.999. The learning rate lr=0.0001 and the batch size was set to 16. The input image size of the model was 384 × 384 pixels As it was an sequential learning task, the maximum temporal distance between frames was set to be [5,10,15,20,25,5] at the corresponding iterations of [0%,10%,20%,30%,40%,90%] of the total 20000 training iterations We also adopted the strategy to make the model pay more attention to the learning of difficult pixels. After 15000 iterations, only the top-20% pixels that had the highest loss would be selected to compute gradients. As we describes in the subsection 3.1, we added multi-scale training strategy to train model. The initial input image size of the model was 384 × 384 pixels, the model would be trained with multi-scale training parameters 0.75, 1, 1.25.

3.4. Experimental Results

Table 1 shows the Ablation study result of Endocv2022 validation and CVC-Clinic datasets. Firstly, we see that when we use semi-supervised learning, the dice coefficient of the model in the Validation Set (EndoCV2022 validation + CVC-Clinic) has increased by 3%. It proves that adding more sequence data for model to learn does help .Secondly, colonoscopy is a product of a combined light source, thus, the collected images are either very bright or very dark. We set color jitter of (brightness=0.5, contrast=0.03, saturation=0.03) to simulated light change. In this way, the dice coefficient improves to 0.7694. Figure 5 shows that images cases which the base model can not segment benefit from this approach. Lastly, we see that the scale of images will affect the performance of the model. The multi-scale training strategy reduces the sensitivity of the model to image resolution, as the dice coefficient of the model improves to 0.7800.

Table 2 provides our model's segmentation results on EndoCV2022 challenge segmentation task. Firstly, the improved-STCN model we have trained for polyp segmentation have an excellent performance on the unseen dataset while the dice coefficient is up to 0.7423.This result already make us ranked the top5 on the leaderboards. When we adopt the two methods mentioned in the subsection 2.3, the dice coefficient has increased by 2% and by 3% respectively. From the results, we see that our enhance scheme mentioned above does help. Unfortunately,

Table 1
Ablation study result of Endocv2022 validation combined with CVC-Clinic datasets

Method	Dice	IOU
base	0.7338	0.6701
semi-supervised learning	0.7613	0.6894
semi-supervised learning + Light Change	0.7694	0.7058
semi-supervised learning + Light Change + Multi-scale training	0.7800	0.7237

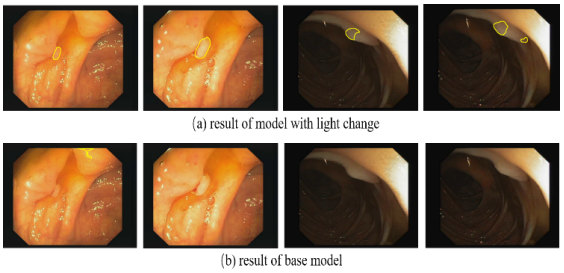


Figure 5: Comparison of model segmentation under strong light and low light (a) shows model trained with light change strategy has better performance, as (b) can not distinguish the target.

Table 2
Results on EndoCV2022 segmentation task round II test set

Method	Dice	std
STCN	0.7423	0.3756
STCN + SFSN	0.7613	0.3571
STCN + Reverse Sequence	0.7694	0.3543

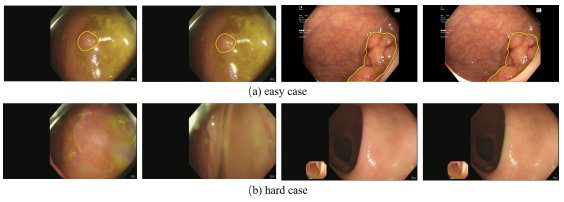


Figure 6: Example of model segmentation results on EndoCV 2022 round-II.(a) shows the easy case for model and (b) shows the hard case in complex scenarios.

as figure 6 shows, our model does not recognize objects in complex scenarios, such as dim and dark scenes.

4. Conclusion

In this work, we have detailed our solution for the polyp segmentation subtask in the EndoCV2022 challenge. We have proposed improved-STCN network with a semi-supervised learning method to improve model's generalization and an enhanced scheme to make model output more credible results. Limited experimental results show that our method achieves consistently high Dice scores at very low standard deviations, suggesting its suitability for polyp segmentation on endoscopic sequence data.

References

[1] H. K. Cheng, Y.-W. Tai, C.-K. Tang, Rethinking space-time networks with improved memory coverage for efficient video object segmentation, Advances in Neural Information Processing Systems 34 (2021).

[2] S. Ali, N. Ghatwary, D. Jha, E. Isik-Polat, G. Polat, C. Yang, W. Li, A. Galdran, M.-Á. G. Ballester, V. Thambawita, et al., Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge, arXiv preprint arXiv:2202.12031 (2022). doi:10.48550/arXiv.2202.12031.

[3] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, K. V. Anonsen, M. A. Riegler, et al., Polypgen: A multi-center polyp detection and segmentation dataset for generalisability assessment, arXiv preprint arXiv:2106.04463 (2021). doi:10.48550/arXiv.2106.04463.

[4] S. B. Ahn, D. S. Han, J. H. Bae, T. J. Byun, J. P. Kim, C. S. Eun, The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies, Gut and liver 6 (2012) 64.

[5] T. K. Lui, C. K. Hui, V. W. Tsui, K. S. Cheung, M. K. Ko, D. C. Foo, L. Y. Mak, C. K. Yeung, T. H. Lui, S. Y. Wong, et al., New insights on missed colonic lesions during colonoscopy through artificial intelligence–assisted real-time detection (with video), Gastrointestinal Endoscopy 93 (2021) 193–200.

[6] C. Yua, J. Yana, X. Lia, Parallel res2net-based network with reverse attention for polyp segmentation (2021).

[7] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[8] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, A. Sorkine-Hornung, Learning video object segmentation from static images, in: Proceedings of

the IEEE conference on computer vision and pattern recognition, 2017, pp. 2663–2672.

[9] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. d. Lange, D. Johansen, H. D. Johansen, Kvasir-seg: A segmented polyp dataset, in: International Conference on Multimedia Modeling, Springer, 2020, pp. 451–462.

[10] J. Silva, A. Histace, O. Romain, X. Dray, B. Granado, Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer, International journal of computer assisted radiology and surgery 9 (2014) 283–293.

[11] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. Vilariño, Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians, Computerized Medical Imaging and Graphics 43 (2015) 99–111.

[12] S. Ali, F. Zhou, B. Braden, A. Bailey, S. Yang, G. Cheng, P. Zhang, X. Li, M. Kayser, R. D. Soberanis-Mukul, et al., An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy, Scientific reports 10 (2020) 1–15.

[13] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. Matuszewski, et al., Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, Medical image analysis 70 (2021) 102002. doi:10.1016/j.media.2021.102002.