# Semi-supervised Spatial Temporal Attention Network for Video Polyp Segmentation

Xinkai Zhao[1], Zhenhua Wu[1], Shuangyi Tan[2,3], De-Jun Fan[4], Zhen Li[3], Xiang Wan[3,5], and Guanbin Li[1(✉)]

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
liguanbin@mail.sysu.edu.cn
[2] Shenzhen Research Institute of Big Data, Shenzhen, China
[3] The Chinese University of Hong Kong, Shenzhen, China
[4] The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China
[5] Pazhou Lab, Guangzhou, China

**Abstract.** Deep learning-based polyp segmentation approaches have achieved great success in image datasets. However, the frame-by-frame annotation of polyp videos requires a large amount of workload, which limits the application of polyp segmentation algorithms in clinical videos. In this paper, we address the semi-supervised video polyp segmentation task, which requires only sparsely annotated frames to train a video polyp segmentation network. We propose a novel spatial-temporal attention network which is composed of Temporal Local Context Attention (TLCA) module and Proximity Frame Time-Space Attention (PFTSA) module. Specifically, TLCA module is to refine the prediction of the current frame using the prediction results of the nearby frames in the video clip. PFTSA module utilizes a simple yet powerful hybrid transformer architecture to capture long-range dependencies in time and space efficiently. Combined with consistency constraints, the network fuses representations of proximity frames at different scales to generate pseudo-masks for unlabeled images. We further propose a pseudo-mask-based training method. Additionally, we re-masked a subset of LDPolypVideo and applied it as a semi-supervised polyp segmentation dataset for our experiments. Experimental results show that our proposed semi-supervised approach can outperform existing image-level semi-supervised and fully supervised methods with sparse annotation at a speed of 135 fps. The code is available at github.com/ShinkaiZ/SSTAN.

**Keywords:** Polyp segmentation · Semi-supervised learning · Medical image segmentation

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-16440-8_44.


L. Wang et al. (Eds.): MICCAI 2022, LNCS 13434, pp. 456–466, 2022.
https://doi.org/10.1007/978-3-031-16440-8_44

## 1 Introduction

Colorectal Cancer (CRC) has become a worldwide human health threat especially for people over fifty years old. In 2021, 147,000 people in United State were diagnosed with this disease, while 53,200 among them died from it [18]. Most CRCs develop from intestinal polyps (adenomas and serrated type), which means the detection and treatment of polyps with colonoscopy is exceedingly significant for the prevention and screening [6]. Clinically, the diagnosis of polyps was completed by an experienced endoscopist, which suffers from a high labor cost and may lead to the omission of diagnosis. With the development of artificial intelligence, many automatic polyp segmentation methods were proposed for ancillary diagnosis and made remarkable progress. Inspired by the great progress achieved by fully convolutional network (FCN) [1], UNet [20], ResUnet [2], UNet++ [31] and ResUNet++ [10] were firstly applied to the polyp segmentation task. Later with the development of attention and transformer [23], ACSNet [29], PraNet [9] and SANet [26] were presented. Inspired by vision transformer [8], which is a novel structure adopting the transformer to computer vision task, Polyp-PVT [7] was soon suggested. All these deep learning based medical segmentation methods mainly focus on the polyp segmentation at the image level, which means they ignore the temporal consistency in endoscopic videos. To better integrate temporal information, video polyp segmentation methods such as Hybrid CNN [19] and PNS-Net [12] were presented recently.

However, the outstanding performances achieved by above supervised models all depend on a large amount of image annotations. In reality, annotation of polyp images and videos would be labor-intensive and resource-intensive. Different from nature images, the labels of medical images require experts in related fields to be annotated and refined. Moreover, for each endoscopic video, many video frames describing a similar content are included, which causes repetitive work and the consistency of manual labels is hard to be guaranteed. Therefore, many semi-supervised polyp segmentation models are suggested, which reduce their requirements for the amount of labeled data and try to fully utilize the unlabelled data. For example, interpolation consistency training method [25] and its improvements [14,17,27,28,30] employ and predict the unlabelled data with the assumption that there is consistency between adjacent labeled data and unlabelled data. Nevertheless, all semi-supervised methods simply consider the consistency and complete the segmentation at the image-level, which ignores the consistency between consecutive frames in video clips.

Considering the labeling difficulty and the utilization of the consistency between consecutive video frames, in this paper, we address the semi-supervised video polyp segmentation task with sparsely annotated frames as well as unlabelled frames and proposed Semi-Supervised Spatial Temporal Attention Network (SSTAN). In our work, we consider both the temporal consistency between video frames and the spatial information contained in each frame with semi-supervised transformer block and vision transformer block, respectively. For evaluation, we applied our model and other cutting-edge models on the subset of LDPolypVideo dataset [15] with masks re-annotated by us. In our experiments,

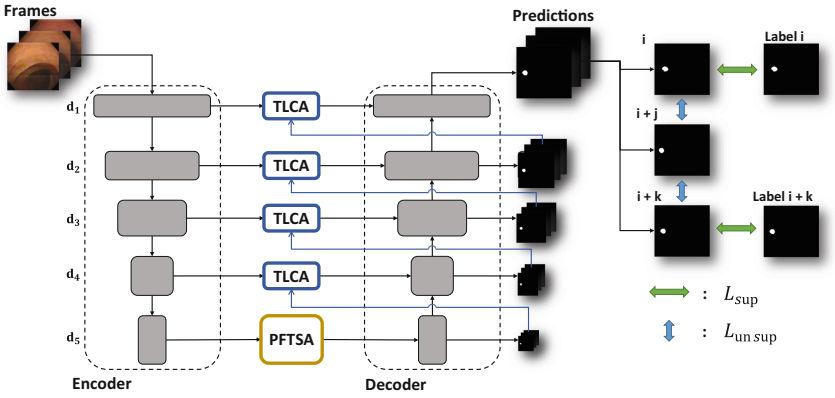


**Fig. 1.** The overview of our approach for semi-supervised video polyps segmentation, which consist of four Temporal Local Context Attention (TLCA) module in the skip connections with a Proximity Frame Temporal-Spatial Attention (PFTSA) module in the bottom layer.

SSTAN requires only 10% of sparse annotations even to outperform existing fully supervised methods on public benchmarks.

Generally, our contributions are four-folds: (1) We creatively address a semi-supervised video polyp segmentation task, which requires the model to be trained under the supervision of a few sparsely annotated video frames and a large number of unlabeled video frames. (2) To both exploit the temporal and spatial features, we propose a novel Semi-Supervised Spatial Temporal Attention Network (SSTAN) with Temporal Local Context Attention (TLCA) and Proximity Frame Time-Space Attention (PFTSA). Additionally, we suggest a corresponding guided training flow consisted with two stages, which allows the model to generate pseudo labels for unlabeled frames under the supervision of labeled data firstly and be finetuned with both true label and pseudo labels. (3) We relabelled and provided corresponding masks to partial video frames from the LDPolypVideo dataset, which was originally labeled with bounding boxes. With dense video frames, the partial re-annotated dataset could be served as one of the few semi-supervised video polyp segmentation datasets. (4) We evaluated and compared our model with both image and video polyp segmentation models and our SSTAN significantly outperformed existing state-of-the-art fully supervised methods with limited labels (e.g., 10% ground truth labels) (Fig. 1).

## 2 Method

This paper is targeted at tackling the semi-supervised video polyps segmentation task. Suppose we have a colonoscopy video clip which is constituted by $n$ frames $X = \{x_i\}_{i=1}^{n}$ for training, including $M$ frames with pixel-wise annotations, donated as $L$, and other $N - M$ frames without annotations, donated as

$U$. The goal of this task is to train the video segmentation model using $L$ and $U$, thus reducing the dependency on annotations in the training process. The framework of our approach is shown in Fig. 3, which is based on ResUnet [10] as the framework like ACSNet [29]. In order to fuse the proximity frame information at different layers, the TLCA module is placed in each skip link. Moreover, we utilize the PFTSA module to capture contextual information in both time and space at the bottom layer. Finally, we use consistency loss to constrain the unlabeled frames in the sparsely annotated video frames.

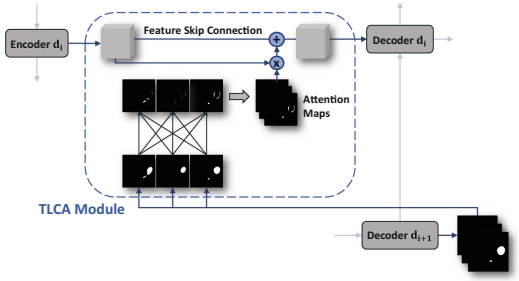


**Fig. 2.** Temporal local context attention module

### 2.1 Temporal Local Context Attention

The goal of Temporal Local Context Attention(TLCA) module is to exploit the prediction differences of adjacent frames to focus the network attention on regions that are harder to predict accurately, thus refining the decoding results. As shown in Fig. 2, for the outputs of encoder layer $d$, donated as $\{\mathcal{E}_d(x_t)\}_{t=1}^n$, we leverage the predictions $\{\mathcal{P}_{d+1}(x_t)\}_{t=1}^n$ of the decoder layer $d+1$ to calculate the attention map for each frame. Specifically, the attention map of frame $x_i$ is donated as follow:

$$\mathcal{M}_i^d = \frac{1}{n-1}\sum_{t\neq i}(|\mathcal{P}_{d+1}(x_t) - \mathcal{P}_{d+1}(x_i)|) \tag{1}$$

where $d$ represents the depth, $t$ stands for the frames except the current frame $i$. We calculate the absolute difference between the prediction of the current frame and the prediction of nearby frames. For each pixel in the image, when the prediction of different frames is similar, the attention map is close to 0. Conversely, when the prediction differs significantly between frames, the attention map is close to 1, representing that this position needs to be better refined in the next decoder layer. Finally, the attention enhanced feature is used as input to the previous layer, to optimize the output of the higher resolution mask.

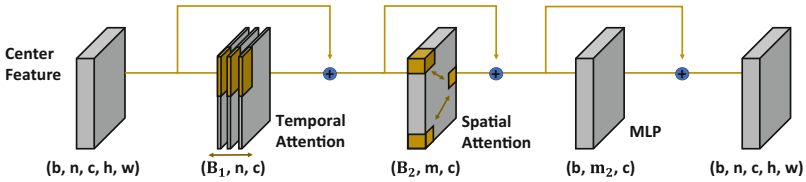


**Fig. 3.** Proximity Frame Temporal-Spatial Attention Module. The b, n, c, h, w stands for batch size, frames number, channel number, height and width, respectively.

### 2.2 Proximity Frame Temporal-Spatial Attention

Although the TLCA module could fuse local information at the same position across frames, the network still lacks the ability to capture long-term contexts. Motivated by the rapid application of the transformer, we applied multi-head attention in two different dimensions, temporal and spatial, respectively. Specifically, for the last layer output of the encoder, we regard each pixel in the feature map as the embedding of a patch in the original image. In contrast to the vanilla transformer, which consists of a multi-head attention module and an MLP, we borrow the idea from [5] to use two multi-head attention modules to capture spatial and temporal contextual dependencies. For feature $\mathcal{E}_5(x) \in \mathbb{R}^{b*n*c*h*w}$, we firstly reshape it to $\mathbb{R}^{B_1*n*c}$, where $B_1 = b*h*w$, then calculate the multi-head attention across $n$ frames. Consequently, the feature is re-arranged to $\mathbb{R}^{B_2*m*c}$, where $B_2 = b*n, m = h*w$. And another multi-head attention is calculated within each image. Finally, after the MLP module, the output is used as the input of decoder.

### 2.3 Loss Function

Our loss function is divided into supervised and unsupervised parts. The supervised loss function is formulated as follow:

$$\mathcal{L}_{sup} = \frac{1}{2*|L|*|D|} \sum_{x_n \in L} \sum_{d \in D} \left( Dice\left(\mathcal{P}_d(x_n), y_d\right) + CE\left(\mathcal{P}_d(x_n), y_d\right)\right), \quad (2)$$

where $y_d$ is the ground truth of the labeled image which is down-scaled to the feature size of the corresponding layer $d$. $Dice(\cdot)$ is dice loss and $CE(\cdot)$ is binary cross entropy loss. $L_{sup}$ is used to calculate between prediction $\mathcal{P}_d(x_n)$ and ground truth $y_d$ of frames which have been labeled.

The unsupervised loss function can be formulated as follow:

$$\mathcal{L}_{unsup} = \frac{1}{|X|*|D|} \sum_{x_n \in X} \sum_{d \in D} SmoothL1\left(\mathcal{P}_d(x_n), \mathcal{P}_d(x_{n+1})\right) \quad (3)$$

To compute consistency, $L_{unsup}$ is calculated between the current frame $x_n$ and the next frame $x_{n+1}$ except the last frame in a video clip.

### 2.4 Training Flow

The whole network is trained following the end-to-end scheme in two stages: *i) Pretraining phase* We used the training data for semi-supervised training of our model with $L_{pretrain} = \frac{1}{2}(\mathcal{L}_{sup} + \mathcal{L}_{unsup})$. *ii) Finetuning phase* The model pretrained in the first stage was applied to generate pseudo labels for the unlabeled frames in the training set. With training data as well as both true labels and pseudo labels of all frames, the model was supervised finetuned with loss function $L_{sup}$ subsequently.

## 3 Experiments

### 3.1 Datasets and Implementation

**Datasets.** Commonly used polyp segmentation datasets including five benchmarks (Kvasir [11], CVC-ClinicDB [3], EndoScene [24], ETIS-Larib Polyp DB [21] and CVC-ColonDB [4]) are image-based, which contain selected frames from video clips. For video polyp segmentation task, due to the expensive cost of video annotation, the only currently knowable video polyp segmentation dataset, i.e., ASU-Mayo, contains video with dense frames and is annotated with masks [22]. However, ASU-Mayo is not publicly accessible, which means other datasets are desired for training models. Meanwhile, LDPolypVideo [15] and SUN Colonoscopy Video Database [16] are two recent video polyp dectection datasets fully annotated with bounding boxes. LDPolypVideo contains 160 video clips with 15397 dense video frames describing polyps in more variety under different bowel environments. To adapt LDPolypVideo to our task, we re-masked 60 videos out of 160 videos in LDPolypVideo for training and testing. The details of our re-masked dataset are described in supplementary material.

**Training and Testing.** In our experiment, the partial masked 36 videos and the following fully annotated 12 videos in re-masked LDPolypVideo were applied as training data and validation data, respectively. The initial learning rate, batch size and optimizer applied in our model training is $1e^{-4}$, 4, and AdamW [13], respectively. Every 11 frames with the first frame and the last frames masked in each video were resized to $256 \times 256$ as a single input. The model was pretrained for 100 epochs and finetuned for 50 epochs. Same parameters and the last 12 fully annotated videos were used for evaluation. When testing, our approach achieves a speed of about 135fps on a single Nvidia Tesla V100 GPU.

**State-of-the-Art Models.** We compared our model with other state-of-the-art models mainly in three types: (1) Image-based Supervised Model (ACSNet [29], SANet [26], PVT [7]); (2) Video-based Supervised Model (PNS [12]) and (3) Image-based Semi-Supervised Model (URPC [14], CLCC [30]). We trained the Image-based Semi-Supervised Model in an end-to-end way under their default settings. Models in other types were retrained in two stages similarly to training

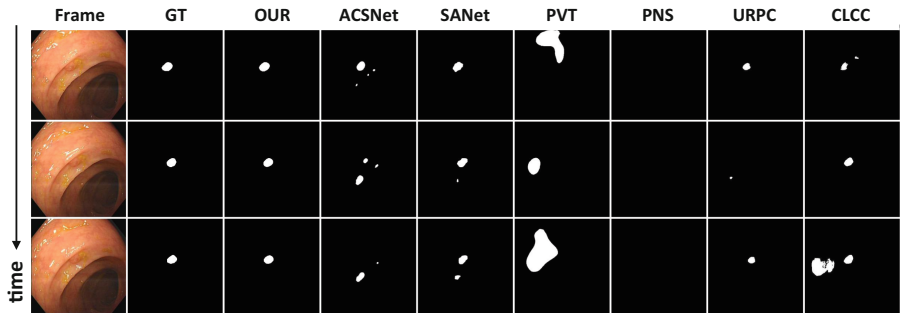


**Fig. 4.** Qualitative results of different models on LDPolypVideo testing set.

**Table 1.** The results and comparison with other state-of-the-art methods. The highest score is highlighted in black bold.

| Model | Labeled | Unlabeled | Accuracy | MAE | F1-Score | F2-Score | mIoU |
|---|---|---|---|---|---|---|---|
| ACSNet | 10% | | 0.984 | 0.016 | 0.396 | 0.411 | 0.658 |
| | 10% | 90% | 0.983 | 0.017 | 0.332 | 0.338 | 0.631 |
| SANet | 10% | | 0.986 | 0.014 | 0.405 | 0.399 | 0.665 |
| | 10% | 90% | 0.982 | 0.018 | 0.396 | 0.391 | 0.665 |
| PVT | 10% | | 0.966 | 0.034 | 0.149 | 0.177 | 0.538 |
| | 10% | 90% | 0.953 | 0.047 | 0.165 | 0.214 | 0.531 |
| PNSNet | 10% | 90% | 0.989 | 0.011 | 0.314 | 0.296 | 0.628 |
| URPC | 10% | 90% | 0.984 | 0.016 | 0.370 | 0.389 | 0.648 |
| CLCC | 10% | 90% | 0.987 | 0.013 | 0.366 | 0.367 | 0.657 |
| Ours | 10% | 90% | **0.990** | **0.010** | **0.482** | **0.486** | **0.700** |

setting: *i) Pretraining phase* We used the annotated 10% data for supervised training of the model under their default settings. *ii) Finetuning phase* We used the model obtained in the previous step to predict the remaining 90% of our datasets and got the corresponding masks. These masks were used as pseudo-labels for 90% of the data, and then we trained the model using both data with pseudo-labels and data with ground truth. For fair comparison, we trained the first stage over 100 epochs and the second stage over 50 epochs and the result of all models in two stages have been tested except PNS, which is a video-based model while the first stage of its default training process utilizing images.

### 3.2 Qualitative Evaluation

In Fig. 4, we provide the visualization results of our model and other compared models on the testing set of re-annotated LDPolypVideo. We selected three adjacent frames for visualization. Our model has two main advantages: (1) our model has the ability to locate and segment the polyps in many conditions,

**Table 2.** The results of ablation study. The highest score is highlighted in black bold.

| Model | Accuracy | MAE | F1-Score | F2-Score | mIoU |
|---|---|---|---|---|---|
| Baseline | 0.979 | 0.021 | 0.306 | 0.300 | 0.619 |
| Baseline+PFTSA | 0.975 | 0.015 | 0.410 | 0.432 | 0.665 |
| Baseline+PFTSA+TLCA | 0.988 | 0.012 | 0.432 | 0.431 | 0.686 |
| Baseline+PFTSA+TLCA+Finetuning | **0.990** | **0.010** | **0.482** | **0.486** | **0.700** |

such as motion blur, different lighting, complex environment with reflections and bubbles, *etc.*. (2) Our model can consistently predict polyps among consecutive frames because the information of adjacent frames is taken into account. More visualization results is shown in supplementary material.

### 3.3 Quantitative Evaluation

For quantitative evaluation, we selected six metrics: Accuracy, MAE, F1-Score (Dice), F2-Score and mean IoU (mIoU). The results of our model and other state-of-the-art models are shown in Table 1. Our model outperformed all three types of models under the same data setting over all metrics. Specially, our model improves the Dice, F2-Score and mIoU achieved by other models by 8.6%, 7.5% and 4.2%, respectively. This result indicates that our model utilizes the 90% unlabeled data better than other image-based models as well as the video-based supervised model by considering the consistency between consecutive frames. Additionally, two notable results are worth mentioning. One is that the performance of PVT is remarkably worse than other convolution-based models, which demonstrates the perspective shown in [8] that the performance of vision transformer highly depends on the size of training data. The other is that F1-score, F2-score and mIoU are unsatisfactory as the original LDPolypVideo is a challenging dataset that contains various polyps under complex colonial environment [15]. For more experimental results, see supplementary materials.

### 3.4 Ablation Study

In order to verify the effectiveness of our proposed modules, we conducted ablation experiments on the same testing dataset. The baseline model is the ResUNet framework, and we evaluated module effectiveness by adding components. Specifically, we gradually added PFTSA at the bottom layer, TLCA modules at the skip links, and finetuning at the training phase.

**Effectiveness of PFTSA.** We trained the baseline both with PFTSA and without PFTSA. The results are shown in the first and second line of Table 2. We found that results with PFTSA performed better. The improvements suggest that PFTSA improves performance by using spatial and temporal information.

**Effectiveness of TLCA.** Similarly, we investigated the contribution of TLCA by introducing the module additionally. The results are shown in the third line of

Table 2. Compared to the model with PFTSA, F1-score and mIoU of the model were increased by 2.2% and 2.1% respectively, which indicates the attention mechanism can enable the model to focus on the hard regions.

**Effectiveness of Finetuning.** Notably, the above experiments were only trained in pretraining phase. To analyze the effectiveness of our training process, we additionally performed finetuning on the model with both PFTSA and TLCA. The improvement suggests that introducing pseudo labels and performing supervised training are necessary for increasing performance.

## 4 Conclusion

In this paper, we defined the semi-supervised polyp video segmentation task and proposed an accurate and novel network SSTAN, which exploits the spatial and temporal information from the proximity frames in endoscope videos with PFTSA and explores the hard regions with TLCA. Additionally, we produced and applied a re-masked sub-dataset of LDPolypVideo, which could be served as the first challenging dataset for semi-supervised polyp video segmentation task. Experiment results demonstrate that our SSTAN outperformed other state-of-the-art methods including image-based supervised model, image-based semi-supervised model and video-based supervised model under the same data setting with real time speed (135 fps). In future work, we will further explore a better performance of SSTAN on semi-supervised tasks for video polyp segmentation.

**Acknowledgment.** This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation (No. 2020B1515020048), in part by the National Natural Science Foundation of China (No. 61976250), in part by the Guangzhou Science and technology project (No. 202102020633), in part by the Chinese Key-Area Research and Development Program of Guangdong Province (2020B0101350001), and in part by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen.

## References

1. Akbari, M., et al.: Polyp segmentation in colonoscopy images using fully convolutional network. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 69–72. IEEE (2018)
2. Alam, S., Tomar, N.K., Thakur, A., Jha, D., Rauniyar, A.: Automatic polyp segmentation using u-net-resnet50. arXiv preprint arXiv:2012.15247 (2020)
3. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Med. Imaging Graph. **43**, 99–111 (2015)
4. Bernal, J., Sánchez, J., Vilarino, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recogn. **45**(9), 3166–3182 (2012)
5. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding. arXiv preprint arXiv:2102.05095 2(3), 4 (2021)

6. Buskermolen, M., et al.: Impact of surgical versus endoscopic management of complex nonmalignant polyps in a colorectal cancer screening program. Endoscopy (2022)
7. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-pvt: polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932 (2021)
8. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Fan, D.-P., Ji, G.-P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: PraNet: parallel reverse attention network for polyp segmentation. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 263–273. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_26
10. Jha, D., Smedsrud, P.H., Johansen, D., de Lange, T., Johansen, H.D., Halvorsen, P., Riegler, M.A.: A comprehensive study on colorectal polyp segmentation with resunet++, conditional random field and test-time augmentation. IEEE J. Biomed. Health Inform. **25**(6), 2029–2040 (2021)
11. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37
12. Ji, G.-P., et al.: Progressively normalized self-attention network for video polyp segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 142–152. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_14
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
14. Luo, X., et al.: Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 318–329. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_30
15. Ma, Y., Chen, X., Cheng, K., Li, Y., Sun, B.: LDPolypVideo benchmark: a large-scale colonoscopy video dataset of diverse polyps. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 387–396. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_37
16. Misawa, M., et al.: Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointestinal Endoscopy **93**(4), 960–967 (2021)
17. Pandey, P., Pai, A., Bhatt, N., Das, P., Makharia, G., AP, P., et al.: Contrastive semi-supervised learning for 2d medical image segmentation. arXiv preprint arXiv:2106.06801 (2021)
18. Patel, S.G., et al.: Updates on age to start and stop colorectal cancer screening: recommendations from the us multi-society task force on colorectal cancer. Gastroenterology **162**(1), 285–299 (2022)
19. Puyal, J.G.-B., et al.: Endoscopic polyp segmentation using a hybrid 2D/3D CNN. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 295–305. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_29
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. Int. J. Comput. Assist. Radiol. Surg. **9**(2), 283–293 (2014)

22. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans. Med. Imaging **35**(2), 630–644 (2015)
23. Vaswani, A., et al.: Attention is all you need. Advances in neural information processing systems 30 (2017)
24. Vázquez, D., et al.: A benchmark for endoluminal scene segmentation of colonoscopy images. J. Healthcare Eng. 2017 (2017)
25. Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. arXiv preprint arXiv:1903.03825 (2019)
26. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 699–708. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_66
27. Xiang, J., Li, Z., Wang, W., Xia, Q., Zhang, S.: Self-ensembling contrastive learning for semi-supervised medical image segmentation. arXiv preprint arXiv:2105.12924 (2021)
28. You, C., Zhou, Y., Zhao, R., Staib, L., Duncan, J.S.: Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. arXiv preprint arXiv:2108.06227 (2021)
29. Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., Yu, Y.: Adaptive context selection for polyp segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 253–262. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_25
30. Zhao, X., Fang, C., Fan, D.J., Lin, X., Gao, F., Li, G.: Cross-level contrastive learning and consistency constraint for semi-supervised medical image segmentation. arXiv preprint arXiv:2202.04074 (2022)
31. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-Net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1