Journal of Pathology Informatics 14 (2023) 100197

Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

Segmentation of polyps based on pyramid vision transformers and residual block for real-time endoscopy imaging

Roi Nachmani [a], Issa Nidal [b], Dror Robinson [c], Mustafa Yassin [c], David Abookasis [a,d,*]

[a] Department of Electrical and Electronics Engineering, Ariel University, Ariel 407000, Israel
[b] Department of Surgery, Hasharon Hospital, Rabin Medical Center, affiliated with Tel Aviv, University School of Medicine, Petah Tikva, Israel
[c] Department of Orthopedics, Hasharon Hospital, Rabin Medical Center, affiliated with Tel Aviv, University School of Medicine, Petah Tikva, Israel
[d] Ariel Photonics Center, Ariel University, Ariel 407000, Israel

ARTICLE INFO

Keywords:
Colorectal cancer
Semantic segmentation
Convolutional neural network
Deep learning
Pyramid vision transformers
Computer vision

ABSTRACT

Polyp segmentation is an important task in early identification of colon polyps for prevention of colorectal cancer. Numerous methods of machine learning have been utilized in an attempt to solve this task with varying levels of success. A successful polyp segmentation method which is both accurate and fast could make a huge impact on colonoscopy exams, aiding in real-time detection, as well as enabling faster and cheaper offline analysis. Thus, recent studies have worked to produce networks that are more accurate and faster than the previous generation of networks (e.g., NanoNet). Here, we propose ResPVT architecture for polyp segmentation. This platform uses transformers as a backbone and far surpasses all previous networks not only in accuracy but also with a much higher frame rate which may drastically reduce costs in both real time and offline analysis and enable the widespread application of this technology.

Introduction

A colon polyp is a growth that forms on the lining of the colon and rectum. Polyps are found in about 30% of the adult population over the age of 50. Most colon polyps are harmless. But over time, some colon polyps progress into colon cancer, which may be fatal when discovered at its later stages. Colorectal cancer is the third most common cancer diagnosed in the USA, with a rate of about 38 new cases per 100 000 people and the death rate of about 13 per 100 000 people per year.[1] Early detection of polyps by colonoscopy exam is critical for the prevention of colon cancer.[2] The colonoscopy exam has emerged as an effective, minimally invasive tool for diagnosing polyps by examining the gastrointestinal tract and is performed by highly trained endoscopists. Still, recent clinical investigations have shown that the current colonoscopy process misses 22%–28% of polyps. These false negatives can lead to late diagnosis of colon cancer, resulting in a poor prognosis. During the exam, the doctor uses a colonoscope, a long flexible tube about the width of a finger with a light and small video camera on the end, inserted through the anus to view the inside of the colon and rectum. Special instruments can be passed through the colonoscope to take a biopsy or remove any suspicious-looking areas such as polyps, if needed.[3] The structure of a polyp varies depending on its stage of progression. Variations in structure, size, and color of the polyp, as shown in Fig. 1, may make them difficult to identify. Tiny polyps are particularly challenging as they don't have distinguishable contrast from the normal surrounding lining, and thus even a well-trained physician and even classical image processing methods cannot achieve acceptable detection results.

Furthermore, real-time differentiation and classification of polyps (i.e., adenomatous or hyperplastic) may allow for strategic therapeutic decisions during the colonoscopy procedure (such as "resect and discard" or "diagnose and leave"). A number of deep learning methods have been developed in order to address these issues, and some have achieved impressive results. The main shortcoming of the deep learning network solutions is their slow run time, and as a result, they cannot be run in real time during a colonoscopy exam. To overcome the above issues, our work presented in this paper contributes the following:

- Construction of a new polyp segmentation architecture, named ResPVT, that contains a pyramid vision transformer (PVT) as an encoder to extract more powerful and robust features, a fusion module for the high-level features (semantic cues and location), and ResBlock for the low-level features (color, edges, etc.).
- Achievement of the highest frames-per-second (FPS) along with state-of-the-art (SOTA) results in performance metrics on different datasets (such as Kvasir-SEG dataset[4]), Compared to other SOTA such as NanoNet[5] ResUNet++,[7] ResUNet+++CRF.[8]

* Corresponding author.
E-mail address: davida@ariel.ac.il (D. Abookasis).

http://dx.doi.org/10.1016/j.jpi.2023.100197
Received 16 November 2022; Received in revised form 22 January 2023; Accepted 22 January 2023
Available online 26 January 2023

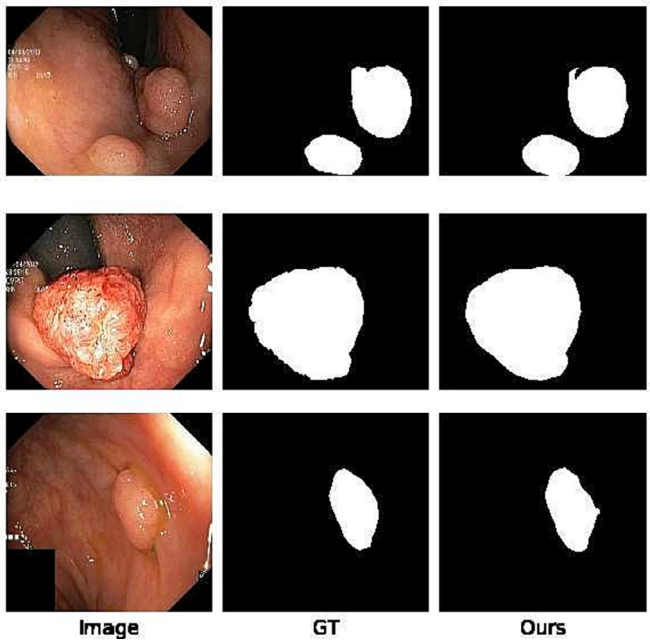

Fig. 1. Representative segmentation masks by our model on the Kvasir-SEG dataset. The first column represents the original images from Kvasir-SEG dataset, the second column represents the pixel-level mask (ground truth), and the third column represents the semantic mask prediction from our model.

Related work

Classic methods

Early works proposed methods for solving the problem of polyp segmentation using classical methods of image processing.[9] These methods did not perform well because of the similarity between the polyp and the surrounding background.

Convolution networks

Deep learning methods[10–12] improved the performance of polyp segmentation tasks. Recently, encoder–decoder models such as U-Net,[13] ResUNet,[14] and ResUNet++,[7] have achieved better performance compared to previous methods. Jha et al.[8] applied Conditional Random Field (CRF) post-processing to improve the model's ability to capture contextual information of the polyps and thus improve overall results. Thambawita et al. applied the first pyramid-based augmentation to the polyp segmentation task[15] while Jha et al. designed a real-time polyp segmentation method called ColonSegNet.[16] Although it achieved higher FPS, its overall performance was inferior when compared to other methods. Jha et al.[5] designed a lightweight model for real-time polyp segmentation called NanoNet which achieved better performance and includes 3 different architectures: NanoNet-A, NanoNet-B, and NanoNet-C. Each architecture consists of different feature channels in its decoder block. We will focus on NanoNet-A, feature with low FPS, high accuracy, and a large number of parameters, and NanoNet-C (high FPS, low accuracy, small number of parameters).

Transformers networks

Transformers were first proposed in the area of natural language process (NLP) and achieved notably good results.[17] They are made up of multi-head self-attention (MHSA) layers to model long-term dependencies. Dosovitskiy et al.[18] proposed the first shifted transformer method from NLP to computer vision classification tasks, called vision transformer (ViT). The ViT network divides an image into patches, converts these patches to embedding, and then feeds them as sequences equivalent to the embedding in language processing to find the attentions between each other. Although ViT is applicable to image classification, it is challenging to directly adapt it to pixel-level density predictions such as object detection and segmentation because its output feature map has only a single scale with low resolution and its computations and memory cost are relatively high even for a common input image size. Pyramid Vision Transformer (PVT)-based models[19,20] overcome the difficulties of ViT by taking fine-grained image patches (4×4 per patch) as input to learn high-resolution representation which is essential for dense prediction tasks such as semantic segmentation. Furthermore, the PVT architecture includes a progressively shrinking pyramid with 4 stages to reduce the sequence length of the transformer while the depth of the network is increased, significantly reducing the computational consumption. Two years ago, Dong et al. proposed a new image polyp segmentation framework, named Polyp-PVT, which utilizes a pyramid vision transformer backbone as the encoder to explicitly extract more powerful and robust features.[21] In addition, it includes 3 modules that extract high- and low-level cues separately and effectively fuse them for the final output.

Network architectures

Overall architecture

The ResPVT architecture is depicted in Fig. 2. The model's architecture contains 3 main blocks: pyramid vision transformer (PVT) encoder, fusion module (FM), and ResBlock. Specifically, the PVT encoder is used to extract multi-scale features from the input image. As shown in the result section, different versions of PVT encoder were compered (Table 1). The FM is used to collect the semantic cues and locate polyps by aggregating high-level features, and the ResBlock is used for extracting low-level features such as color, edges, etc. Given an input image $I \in R^{H \times W \times 3}$, the image is fed into the pre-trained encoder (PVT) to extract 4 pyramid features $X_i \in R^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i}$, where $C_i \in \{32,64,128,256\}$ and $i \in \{1,2,3,4\}$. Then, the channel is reduced from the last 3 feature maps F_2, F_3, and F_4 to 32 with convolution units and inserted into the FM block that produces $O_1 \in R^{\frac{H}{4} \times \frac{W}{4} \times 32}$. At the same time, F_1 is fed into the ResBlock that produces $O_2 \in R^{\frac{H}{4} \times \frac{W}{4} \times 32}$. Finally, O_1 and O_2 are concatenated and 2 more convolution layers are given as the final polyp segmentation mask.

PVT encoder

Recent works have shown that the vision transformer is stronger than well-designed CNN backbones.[22,23] Inspired by that, a pyramid vision transformer (PVT) was used here as the backbone network to extract more robust and powerful features for polyp segmentation. PVT is a progressive shrinking pyramid and a spatial-reduction attention (SRA) layer to obtain multi-scale feature maps under limited computation or memory resources.[19] The entire model is divided into 4 stages, each of which is comprised of a patch embedding layer and a Linear-layer Transformer encoder. Following a pyramid structure, the output resolution of the 4 stages progressively shrinks from high (4-stride) to low (32-stride). In the first stage, given an input image of size $H \times W \times 3$, the image is first divided into $HW/4^2$ patches, each of size $4 \times 4 \times 3$. Then, the flattened patches are fed to a linear projection and embedded patches of size $HW/4^2 \times C_1$ are obtained. Next, the embedded patches along with a position embedding are passed through a Transformer encoder with L1 layers, and the output is reshaped to a feature map F_1 of size $\frac{H}{4} \times \frac{W}{4} \times C_1$. In the same way, using the feature map from the previous stage as input, the following feature maps are obtained: F_2, F_3, and F_4, whose strides are 8, 16, and 32 pixels with respect to the input image. Since PVT needs to process high-resolution (4-stride) feature maps, a SRA layer is proposed to replace the traditional multi-head attention (MHA) layer in the encoder. Similar to MHA, the proposed SRA receives a query Q, key K, and value V as input, and outputs a refined feature. The difference is that the proposed SRA reduces the spatial scale of K and V before the attention operation, which largely reduces the

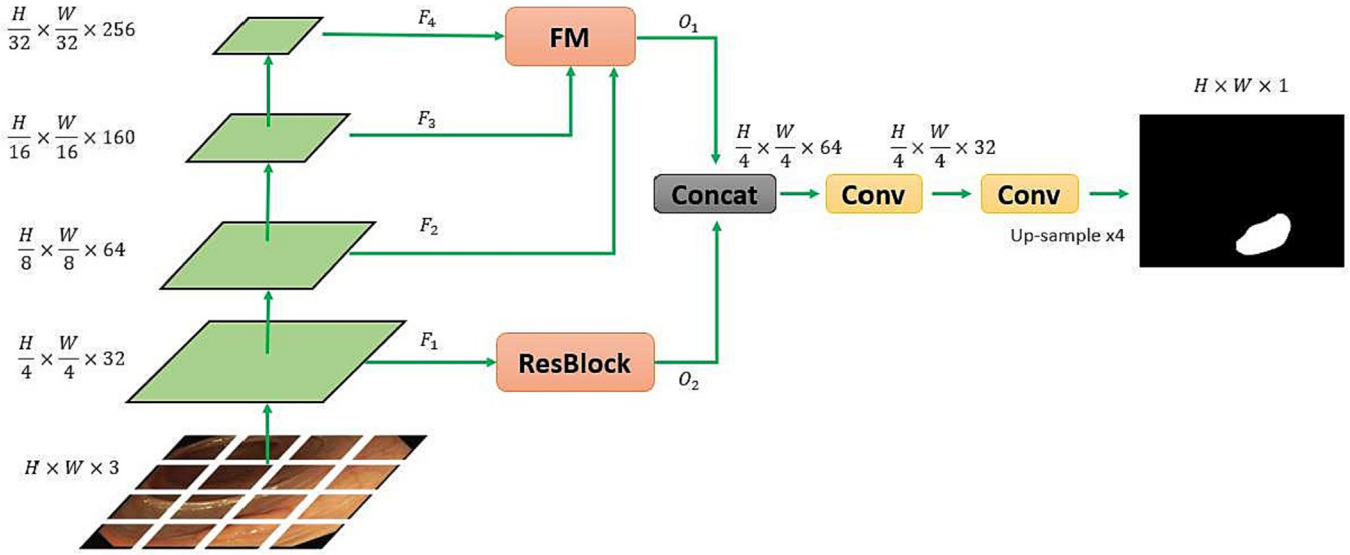

Fig. 2. Overview of the proposed ResPVT architecture, which contains PVT encoder (a), fusion module (FM) for fusing the high-level features (b), residual block (ResBlock) for extracting low-level features (c), and ResPVT head for fusion between high- and low-level features to produce the final mask (d).

computational and memory overhead. We adopt the PVTv2 which is significantly improves PVTv1 on the 3 tasks, classification, detection, and segmentation.[20] We extract from the PVT encoder 4 multi-scale feature maps (F_1, F_2, F_3, and F_4). Among these feature maps, F_1 gives detailed appearance information of polyps, and F_2, F_3, and F_4 provide high-level semantic cues.

Fusion module

Inspired by the work of Dong et al.,[21] we implement the fusion module (FM). As shown in Fig. 3, we define C as a convolution layer that contains 3 × 3 convolution layers with padding of 1, batch normalization and ReLU activation. First, we reduce the channel to 32 for F_2, F_3, and F_4. Then, we fuse the feature maps of F_3 and F_4 as follows: we up-sample (denote as Up_2) F_4 by 2 and feed the result through 2 separate convolution units. One of the results is multiplied (denoted by $\otimes$) by F_3 and then concatenated (denoted by Concat) with the result from the second convolution. Then, we feed the result in 2 convolution units and obtain M_2. Mathematically,

$$M_2 = \mathrm{C}(\mathrm{C}(\mathrm{Concat}(F_3 \otimes \mathrm{C}(\mathrm{Up}_2(F_4)), \mathrm{C}(\mathrm{Up}_2(F_4))))) \quad (1)$$

Subsequently, we fuse the feature maps of F_2, F_3, and F_4 as follows: we up-sample F_4 by 4 and F_3 by 2 to get the same feature map size of F_2. Then, we feed each of them in separated convolution units and multiply by F_2 to obtain M_1,

$$M_1 = F_2 \otimes \mathrm{C}(\mathrm{Up}_2(F_3)) \otimes \mathrm{C}(\mathrm{Up}_4(F_4)) \quad (2)$$

Table 1
PVT Encoder versions.

PVT version	Parameters
B0	3 695 809
B1	13 868 321
B2	25 222 177
B3	45 098 017
B4	62 415 393
B5	81 815 329

Finally, we concatenate between M_1 and M_2 and feed the result in 2 convolution units and obtain,

$$\mathrm{O}_1 = \mathrm{C}(\mathrm{C}(\mathrm{Concat}(M_1, M_2))) \quad (3)$$

Residual block

We define 2 types of convolution layers. The first is C_1 that contains 1 × 1 convolution layers without padding, and the second is C_3 that contains 3 × 3 convolution layers with padding of 1 and stride of 2. Both contain a batch normalization and ReLU activation. First, we feed the low-level feature map F_1 into the main root that contains C_1 to C_3 and C_1 convolution layers. Additionally, F_1 is feed into C_1 convolution layer in the skip connection (left branch in Fig. 4). An element-wise addition is performed between the skip connection and main root results. Finally, we up-sample by 2 followed by C_3 convolution layer and obtain O_2,

$$\mathrm{O}_2 = \mathrm{C}_3(\mathrm{Up}_2(\mathrm{C}_3(\mathrm{F}_1) \otimes \mathrm{C}_1(\mathrm{C}_3(\mathrm{C}_1(\mathrm{F}_1))))) \quad (4)$$

ResPVT head

For fusion between high- and low-level features we implement aggregation on O_1 and O_2. Given the feature maps O_1, which contains high-level semantic information, and O_2, which contains low-level semantic information, we concatenate them and follow by 2 convolution layers for channel reduction. Finally, we up-sample by 4 and obtain the final output,

$$\mathrm{Output} = \mathrm{Up}_4(\mathrm{C}_1(\mathrm{C}_3(\mathrm{Concat}(\mathrm{O}_1, \mathrm{O}_2)) \quad (5)$$

Implementation details

We implement our Res-PVT with the PyTorch framework and use a NVIDIA GeForce RTX 2080 Ti machine with 11GB VRAM. Considering the differences in the sizes of each polyp image, we used a multi-scale strategy in the training stage. We use the AdamW optimizer, which is widely used in transformer networks.[24] The learning rate and the weight decay are set to 1e-4. Our loss function is a combination of binary cross-entropy (BCE) and IoU. Further, we resize the input images to 352 × 352 with a mini-batch size of 8 for 200 epochs. The total training time is nearly 7 h

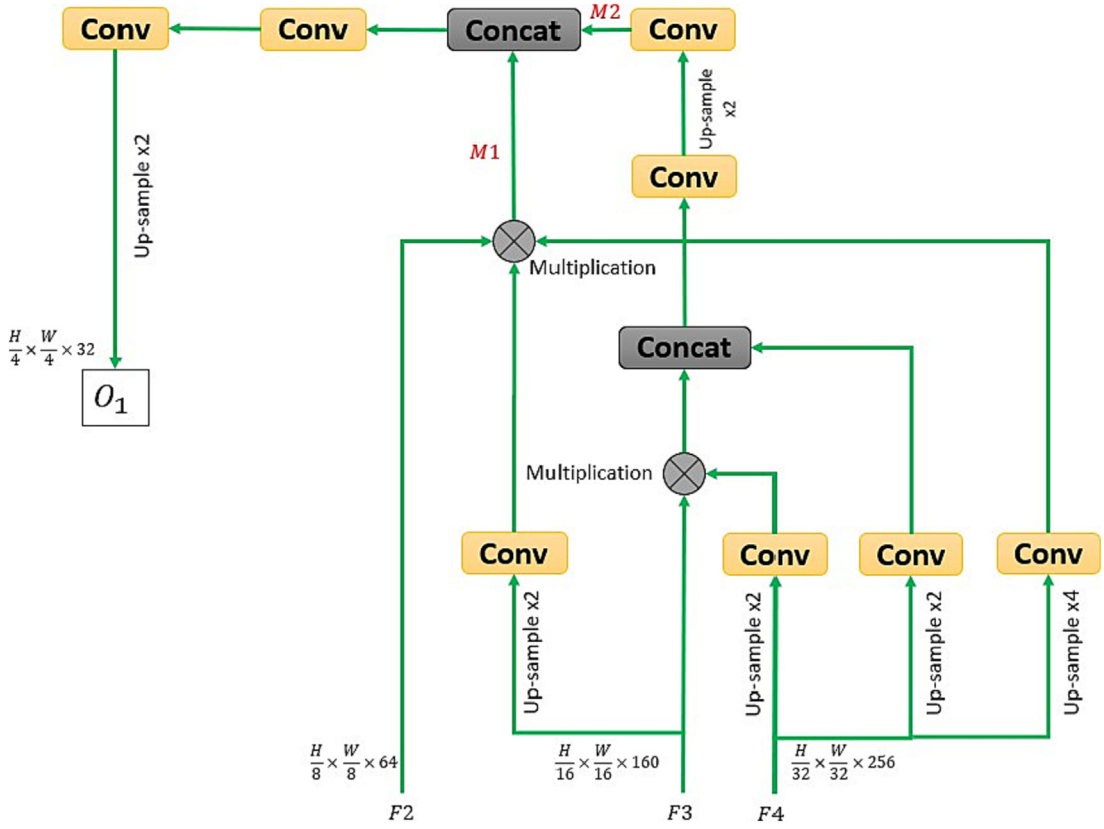

Fig. 3. Details of the fusion module (FM) architecture for fusing the high-level features.

to achieve the best performance (63 epochs). Additionally, we used an early stopping mechanism to prevent over-fitting. This stopping was obtained by measuring the dice score over the test set after each epoch; if improvement in 15 stride epochs was not achieved, the training was stopped. In our training, the stop occurred at epoch 63. For the training stage, we used simple augmentation such as random rotation, horizontal flipping, and vertical flipping. For testing, we only resize the images to 352×352 without any post-processing optimization strategies.

Experiments

Datasets

We evaluated our method on the Kvasir-SEG[4] dataset collected from the polyp class in the Kvasir dataset. The Kvasir-SEG includes 1000 polyp images. For training, we split the Kvasir-SEG to 900 as the training set, and the remaining 100 images as the test set; The training was preformed

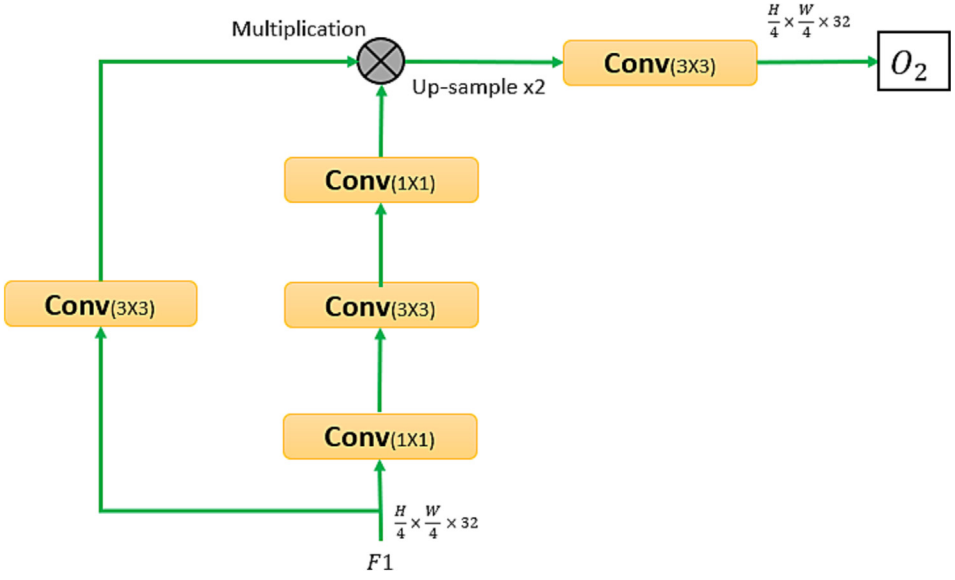

Fig. 4. Details of the ResBlock architecture for extracting low-level features.

Table 2
Performance evaluation of the SOTA methods on Kvasir-SEG.

Method	Parameters	DSC	mIoU	Recall	Precision	F2	Accuracy	FPS
ResUNet	8 227 393	0.720	0.610	0.760	0.762	0.732	0.925	17.72
ResUNet++	4 070 385	0.731	0.636	0.792	0.793	0.747	0.922	19.79
NanoNet-A	235 425	0.822	0.728	0.858	0.836	0.835	0.945	26.13
NanoNet-C	36 561	0.749	0.636	0.808	0.773	0.771	0.929	32.17
ResPVT-B0(Ours)	3 695 809	0.954	0.918	0.961	0.954	0.958	0.987	53.92

once, and the test set was used as a cross-validation (only on the Kvasir dataset). As mentioned in Table 2, we obtained an accuracy of 0.987 over the test set. The reason for the high performance could be because we trained the network only on data from Kvasir which may contain a sample of images that is biased towards a certain physiological site in the body. For testing, we used four unseen datasets, CVC-ClinicDB,[25] ETIS,[26] CVC-ColonDB,[27] and Endotect.[6] There are 196 images in ETIS, 380 images in CVC-ColonDB, 612 images in CVC-ClinicDB and 1000 images in Endotect.

Evaluation metrics

For the evaluation of our model, we chose the same metrics as used by NanoNet for comparison between the methods. Those metrics include: Dice Score Coefficient (DSC), mean Intersection over Union (mIoU), Precision, Recall, F2, Accuracy, and Frame-per-second (FPS).

$$IoU = \frac{TP}{TP + FP + FN} \tag{6}$$

$$Dice = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$F2 = \frac{TP}{TP + 0.2 \cdot FP + 0.8 \cdot FN} \tag{11}$$

Results

We evaluated our model and compared it to the recent SOTA computer vision methods. For evaluation, we used performance metrics as described above (Evaluation metrics section). On the Kvasir-SEG dataset,[4] our method achieves a mean Dice of 0.954, which is 20.5% higher than existing real-time SOTA method NanoNet[5] and precision of 0.958 which is 18.5% higher than it, as well. In addition, on the Endotect dataset,[6] our model achieves a mean Dice of 0.891, which is 19% higher than NanoNet and precision of 0.905 which is 19% higher than it, as well. In the FPS measurement, our model achieves around 54 FPS, which is 22 FPS higher than NanoNet and around 36 FPS from other method (ResUNet++,[7] ResUNet+++CRF,[8] etc.). Table 4 shows the results of ResPVT compared to other SOTA methods on CVC-ClinicDB,[25] CVC-ColonDB,[27] and ETIS-Larib[26] datasets. On CVC-ClinicDB dataset,[25] our method achieves a mean Dice of 0.833, which is 18.7% higher than ResUNet++[7] and precision of 0.857 which is 20.6% higher than it, as well. On ETIS-Larib dataset,[26] our method achieves a mean Dice of 0.755, which is 35.4% higher than ResUNet++,[7] and precision of 0.722 which is 33% higher than ResUNet++.[7] In addition, on the CVC-ColonDB dataset,[27] our model achieves a mean Dice of 0.763, which is 15% higher than ResUNet++ and precision of 0.794 which is 24.8% higher than ResUNet++. In the FPS measurement, our model achieves around 54 FPS, which is higher than the other methods (ResUNet++ and ResUNet+++) by ~36 FPS.

Table 4
Performance evaluation of the SOTA methods on CVC-ClinicDB, CVC-ColonDB and ETIS-Larib.

Test set	Method	DSC	mIoU	Recall	Precision	FPS
CVC-ClinicDB	ResUNet++	0.646	0.731	0.698	0.6510	17.72
	ResUNet+++CRF	0.645	0.732	0.695	0.642	16.42
	ResPVT-B0(Ours)	0.833	0.757	0.866	0.857	54.60
ETIS-Larib	ResUNet++	0.401	0.641	0.441	0.392	17.72
	ResUNet+++CRF	0.401	0.642	0.437	0.375	16.42
	ResPVT-B0(Ours)	0.755	0.671	0.873	0.722	52.58
CVC-ColonDB	ResUNet++	0.513	0.674	0.539	0.546	17.72
	ResUNet+++CRF	0.512	0.674	0.536	0.528	16.42
	ResPVT-B0(Ours)	0.763	0.679	0.805	0.794	54.62

Discussion

The quantitative results show that ResPVT achieves the best results, both in terms of speed (FPS) and in terms of segmentation results. The quantitative results in Tables 2, 3, and 4 show that ResPVT achieves a real-time segmentation network with the highest scores of all the metrics without any post-processing methods. Specifically, we achieve the highest FPS results, despite the high values of model parameters. We also compared different versions of PVT encoder (see Table 1), as shown in Fig. 5, to observe the effect of the FPS and Dice score on several datasets. As seen, as the number of PVT encoder parameters increases the FPS decreases significantly and without significant effect on the dice score. Moreover, between version B0 and B1, there are almost no differences in the results despite a difference of almost 10-fold in the number of parameters in the models. In contrast, there is a significant difference in FPS measure between version B1 and version B2 even though the difference between the parameters of the models is only about 2-fold. This is probably due to the utilization of the capacity of the GPU device. We assume that in version B0 the machine's

Table 3
Performance evaluation of the SOTA methods on Endotect.

Method	Parameters	DSC	mIoU	Recall	Precision	F2	Accuracy	FPS
ResUNet	8 227 393	0.664	0.540	0.751	0.684	0.694	0.907	17.72
ResUNet++	4 070 385	0.664	0.583	0.879	0.659	0.759	0.884	18.58
NanoNet-A	235 425	0.750	0.646	0.823	0.774	0.777	0.925	27.19
NanoNet-C	36 561	0.701	0.579	0.801	0.715	0.738	0.909	32.98
ResPVT-B0(Ours)	3 695 809	0.891	0.830	0.916	0.905	0.899	0.966	54.07

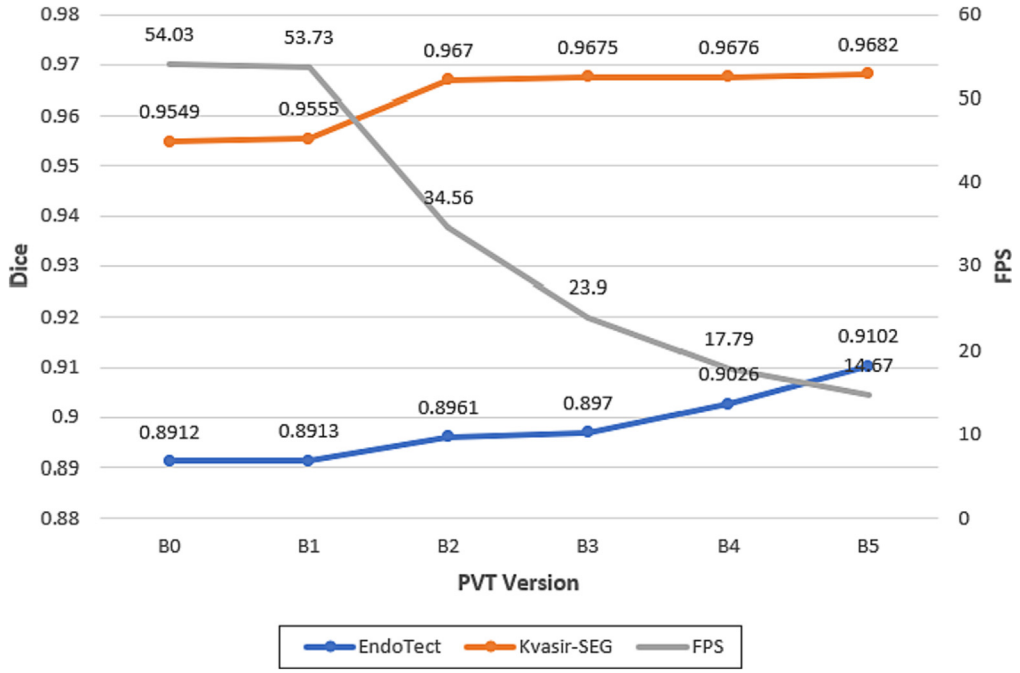

Fig. 5. Evaluation of several versions of PVT encoder (B0–B5) to observe the variation between the FPS and Dice scores achieved by different PVT encoders used to analyze several datasets.

running ability is not fully utilized, compared to version B2 which has passed the machine's ability to run in parallel causing a decrease in the running time capacity. The qualitative results displayed in Fig. 6 indicate that our method presents stable segmentation ability despite different imaging environments such as lighting, contrast, etc. and has more accurately predicted edges. Furthermore, these results demonstrate high precision performance with low false negative that is crucial in medical diagnostics. Several challenges are associated with segmenting polyps, such as bowel preparation quality at the time of colonoscopy, camera angle, etc., which can affect the overall performance of a deep-learning model. For some images, there may even be disagreement in the interpretation among endoscopists. The quality of a colonoscopy examination is largely determined by the experience and skill of the endoscopist. Our proposed model can be used to assist in segmenting a detected polyp, providing an 'extra pair of eyes' to the

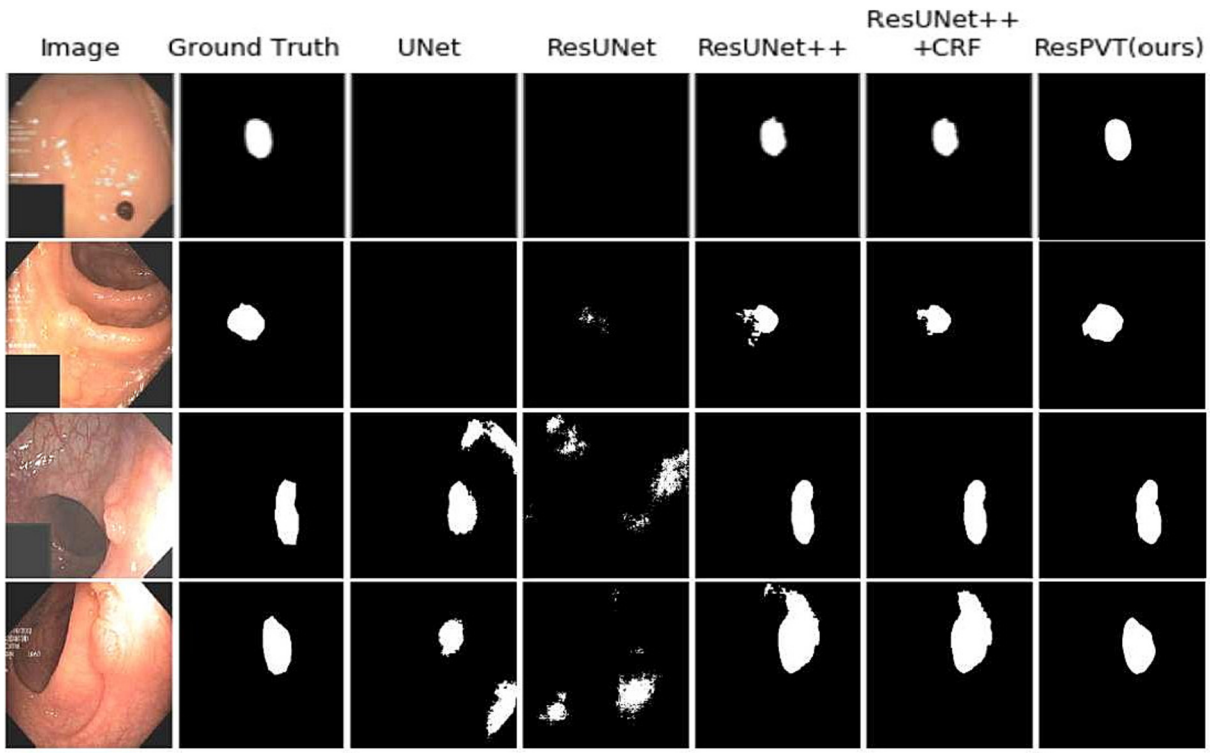

Fig. 6. Qualitative comparison of the results of ResPVT to UNet, ResUNet,ResUNet++, and ResUNet++ +CRF. The first column represents the original images from Kvasir-SEG dataset, the second column represents the pixel-level mask (ground truth), other columns represent the semantic mask predictions by the SOTA model and ours.

endoscopist for additional objective diagnostic information in real-time during the colonoscopy examination.

Conclusion

In this paper, we propose a novel image polyp segmentation method for real-time application, named ResPVT, which contains a pyramid vision transformer backbone for rapid feature extraction. The experimental results on a variety of endoscopy datasets show that our model has the highest result metrics of DSC, IoU, precision, recall, F2-score and, most importantly, FPS. This fast run time implies that our model has potential to be implemented in medical devices to aid colonoscopy examination. We believe that ResPVT has potential to be used in the detection of pathological and abnormal tissues in the lining of the colon. A significant advantage is the ability of the proposed method to identify flat polyps in difficult areas in the colon and tiny lesions that can be easily missed during normal endoscopy. The surrounding remaining tissue after polyp resection by colonoscopy can also be differentiated by the polyp segmentation ResPVT to assure completeness of the resection. We hope that our work will inspire other teams to trying to solve the real-time polyp segmentation task with transformer networks. We also envision implementation of our work in other fields of specialization such as in the care of diabetic feet. There is an interest in early detection of diabetic neuropathy and in particular susceptibility to ulcer development.[28,29] Our technique could be used to improve the outcomes of cases given to subjective surgeon decision-making and optimize the care of those patients. We heuristically believe that our method could also be applied to the assessment of cartilage quality during arthroscopy and arthrotomy.[30] To date, the accepted cartilage classification is based on gross morphology, however as new and advanced methods of cartilage repair are entering the clinical arena, it is important to be able to define the quality of both original cartilage and newly formed cartilage.[31] Finally, the described method could prove useful in assessing chondral quality and deciding on clinical interventions such as patellar resurfacing during knee arthroplasty.[32,33]

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. NIH. Cancer stat facts: Colorectal cancer. URL: http://www.seer.cancer.gov/statfacts/html/colorect.html.
2. Matsuda T, Ono A, Sekiguchi M, Fujii T, Saito Y. Advances in image enhancement in colonoscopy for detection of adenomas. Nat Rev Gastroenter Hepatol 2017;14:305–314.
3. The American cancer society medical and editorial content team. URL: https://www.cancer.org/treatment/understanding-your-diagnosis/tests/endoscopy/colonoscopy.html.
4. Jha D, Smedsrud PH, Riegler MA, et al. Kvasir-seg: a segmented polyp dataset. Int. Conf Multimedia Modeling. Cham Switzerland: Springer; 2020. p. 451–462.
5. Jha D, Tomar NK, Ali S, et al. NanoNet: real-time polyp segmentation in video capsule endoscopy and colonoscopy. Proc IEEE Int Computer-Based Medical Systems 2021:37–43.
6. Vazquez D, Bernal J, Sanchez FJ, et al. A benchmark for endoluminal scene segmentation of colonoscopy images. J Healthcare Eng 2017;2017:1–9. https://doi.org/10.1155/2017/4037190. 4037190.
7. Jha D, Smedsrud PH, Riegler MA, Johansen de Lange DT, Halvorsen P, Johansen HD. ResUNet++: an advanced architecture for medical image segmentation. Proc IEEE Int. Multimedia (ISM); 2019. p. 225–255.
8. Jha D, Smedsrud PH, Johansen D, et al. A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation. IEEE J Biomed Health Info 2021;25:2029–2040.
9. Mamonov AV, Figueiredo IN, Figueiredo PN, Tsai Y-HR. Automated polyp detection in colon capsule endoscopy. IEEE Trans Med Imaging 2014;33:1488–1502.
10. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Int Conf Learning Representations; 2015.
11. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Conf Comp Vis. Pattern Recog; 2015.
12. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. IEEE Conf. Comp. Vis. Pattern. Recog; 2016. p. 770–778.
13. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Med Image Comp Assisted Intervention 2015;9351:234–241.
14. Zhang Z, Liu Q, Wang Y. Road extraction by deep residual unet. IEEE Geosci Remote Sensing Lett 2018;15:749–753.
15. Thambawita V, Hicks S, Halvorsen P, Riegler MA. Pyramid-focus-augmentation: medical image segmentation with step-wise focus. ArXiv 2020. https://arxiv.org/pdf/2012.07430v1.pdf.
16. Jha D, Ali S, Tomar NK, et al. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. IEEE Access 2021;9:40496–40510.
17. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. NeurIPS 2017:6000–6010.
18. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. ArXiv 2021:1-21. https://arxiv.org/pdf/2010.11929.pdf.
19. Wang W, Xie E, Li X, et al. A versatile backbone for dense prediction without convolutions. IEEE Int Conf Comp Vis 2021:548–558.
20. Wang W, Xie E, Li X, et al. PVT v2: improved baselines with pyramid vision transformer. Comp Vis Med 2021;8:415–424.
21. Dong B, Wang W, Fan DP, Li J, Fu H, Shao L. Polyp-PVT: polyp segmentation with pyramid vision transformers. ArXiv 2021. https://arxiv.org/pdf/2108.06932.pdf.
22. Bhojanapalli S, Chakrabarti A, Glasner D, Li D, Unterthiner T, Veit A. Understanding robustness of transformers for image classification. IEEE Int Conf Comp Vis 2021:10211–10221.
23. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: simple and efficient design for semantic segmentation with transformers. 35th Conf Neural Inf Proc Sys; 2021.
24. Loshchilov I, Hutter F. Decoupled weight decay regularization. Int Conf Learning Rep; 2019.
25. Bernal J, Sanchez FJ, Fernandez-Esparrach G, Gil D, Rodrıguez C, Vilarino F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput Med Imaging Graph 2015;43:99-111.
26. 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 2014;9:283–293.
27. Tajbakhsh N, Gurudu SR, Liang J. Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans Med Imaging 2015;35:630–644.
28. Williams BM, Borroni D, Liu R, et al. An artificial intelligence-based deep learning algorithm for the diagnosis of diabetic neuropathy using corneal confocal microscopy: a development and validation study. Diabetologia 2020;63:419–430.
29. Munadi K, Saddami K, Oktiana M, et al. A deep learning method for early detection of diabetic foot using decision fusion and thermal images. Appl Sci 2022;12:7524–7545.
30. Slattery C, Kweon CY. Classifications in brief: outerbridge classification of chondral lesions. Clin Orthop Relat Res 2018;476:2101–2104.
31. Espinosa MG, Otarola GA, Hu JC, Athanasiou KA. Cartilage assessment requires a surface characterization protocol: roughness, friction, and function. Tissue Eng Part C Methods 2021;27:276–286.
32. Rodríguez-Merchán EC, Gómez-Cardero P. The outerbridge classification predicts the need for patellar resurfacing in TKA. Clin Orthop Relat Res 2010;468:1254–1257.
33. Batailler C, Shatrov J, Sappey-Marinier E, Servien E, Parratte S, Lustig S. Artificial intelligence in knee arthroplasty: current concept of the available clinical applications. Arthroplasty 2022;4:1-16.