Hybrid(Transformer+CNN)-based Polyp Segmentation

Madan Baduwal
Mississippi State University
mb4239@msstate.edu
https://madanbaduwal.github.io/polyp-seg

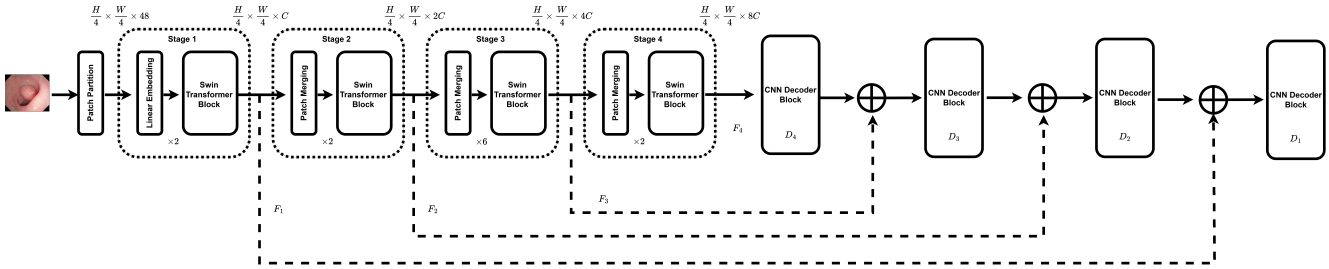

Figure 1. Hybrid(Transformer+CNN)-based Polyp Segmentation Architecture

Abstract

Colonoscopy is still the main method of detection and segmentation of colonic polyps, and recent advancements in deep learning networks such as U-Net, ResUNet, Swin-UNet, and PraNet have made outstanding performance in polyp segmentation. Yet, the problem is extremely challenging due to high variation in size, shape, endoscopy types, lighting, imaging protocols, and ill-defined boundaries(fluid, folds) of the polyps, rendering accurate segmentation a challenging and problematic task. To address these critical challenges in polyp segmentation, we introduce a hybrid (Transformer + CNN) model that is crafted to enhance robustness against evolving polyp characteristics. Our hybrid architecture demonstrates superior performance over existing solutions, particularly in addressing two critical challenges: (1) accurate segmentation of polyps with ill-defined margins through boundary-aware attention mechanisms, and (2) robust feature extraction in the presence of common endoscopic artifacts including specular highlights, motion blur, and fluid occlusions. Quantitative evaluations reveal significant improvements in segmentation accuracy (Recall improved by 1.76%, i.e., 0.9555, accuracy improved by 0.07 %, i.e., 0.9849) and artifact resilience compared to state-of-the-art polyp segmentation methods.

Keywords: Colorectal cancer(CRC), Polyp, Polyp segmentation, colonic polyp, colonoscopy, CNN, Transformer

1. Introduction

Colon polyps are neoplastic polyps developing from the lining epithelium of the colon or rectum. Incidence epidemiological data suggest a prevalence of approximately 30% among individuals over the age of 50 years [42]. Although, in most cases, they are benign, a proportion among them, especially adenomatous polyps, are liable to malignancy by the stepwise adenoma-carcinoma sequence. This shift carries critical clinical import, as CRC is still the third most diagnosed malignancy within the United States, with 38 per 100,000 cases and a commensurate rate of death being 13 per 100,000 annually [42]. Colonoscopy is the cornerstone of preventing CRC, with the ability to detect and treat concomitantly via eliminating pre-malignant polyps. Its use in the initial treatment decreases CRC mortality and incidence by quite a lot [39], showcasing its value in practice.

Colonoscopy is a minimally invasive but very effective diagnostic procedure for detecting colorectal polyps when performed by skilled endoscopists. While helpful, existing colonoscopy tests fail in the detection of 22-28 % of polyps [40], which could develop into advanced cancers and deteriorating clinical outcomes. The process entails using a colonoscope, a finger-thick, flexible tube containing a light source, and a video camera passed transanally to observe the colorectal mucosa. Integrated working channels allow concurrent removal of polyps or biopsy when suspicious lesions are encountered [40].

Polyp shape is highly heterogeneous during stages of de-

velopment. As seen in Fig. 2, variation of structural characteristics, size measurements, and color characteristics provides difficult diagnostic issues at a high level. These are the most demanding challenges for low-contrast small polyps (< 5 mm) in size. These constraints lead to low rates of detection, even being carried out by trained operators using standard image gathering methods.

The clinical need for accurate segmentation of polyps beyond detection is the need to precisely demarcate lesion boundaries to direct therapeutic interventions. Online segmentation of polyp architectures (e.g., discrimination between adenomatous and hyperplastic growth patterns) would directly influence intraoperative decisions, allowing for the complete excision of neoplastic lesions with sparing of normal tissue when polyps are not neoplastic. Despite convolutional and transformer-based models achieving $> 90\%$ Dice scores on offline testing, clinical adoption is strictly limited by some inherent limitations. Current architectures cannot maintain diagnostic-grade segmentations in true real-time endoscopy settings primarily because of the computational delay involved in the processing of 1080p video streams at 30fps. This is compounded by morphological nuances: indistinct boundaries on sessile or flat polyps decrease segmentation accuracy by 15–20% compared with pedunculated lesions, and motion artifacts and specular reflections also degrade boundary precision. This underlying conflict between model complexity (to provide pixel-level precision) and inference speed (to provide clinical usability) identifies the medical imperative for lightweight but robust segmentation algorithms in gastrointestinal endoscopy.

To address these challenges, our work presented in this paper contributes the following:

- We suggest a hybrid approach in which pre-trained vision transformers are utilized for global feature extraction and light-weight CNNs for spatial fine-tuning to achieve precise polyp segmentation and minimize computational complexity.
- Achieved the highest frames-per-second (FPS) alongside state-of-the-art (SOTA) results in performance metrics on various datasets (such as the Kvasir-SEG dataset [25]), compared to other SOTA models like NanoNet [28], ResUNet++ [27], and ResUNet++ + CRF [24].

2. Related work

2.1. Classic methods

Early works proposed methods to address the problem of polyp segmentation using classical image processing techniques [38]. However, these methods struggled to achieve satisfactory performance due to the similarity between the polyps and the surrounding background.

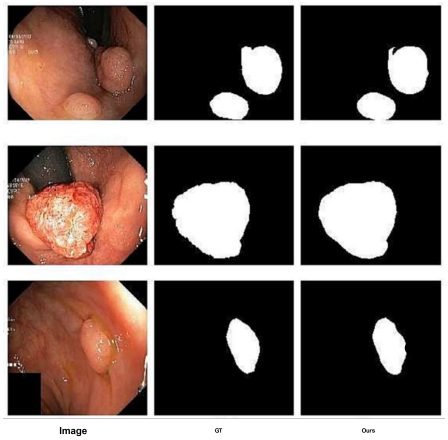

Figure 2. Comparative visualization of polyp segmentation results on the Kvasir-SEG dataset. From left to right: (a) Original endoscopic images, (b) Pixel-level ground truth annotations, and (c) Predicted segmentation masks generated by our proposed model.

2.2. Convolution networks

Deep learning techniques [50, 36, 19] have immensely enhanced the performance of polyp segmentation. U-Net model [47] with its encoder-decoder in generic form and skip connections is a smooth benchmark for segmentation of polyps. Its variations, such as U-Net++ [70], used nested skip connections to visualize fine details while ResUNet [68] and ResUNet++ [27] made use of residual blocks to achieve smooth gradient propagation and feature discovery. DoubleU-Net [23] further enhanced segmentation using a two-stage U-Net model, further advancing the detection of small polyps.

Recent advancements have concerned attention mechanism boost and contextual modeling. PraNet [17] proposed reverse attention and boundary refining and achieved SOTA performance in a number of polyp benchmarks. DDANet [33] also utilized dilated dual attention blocks to boost contextual modeling, and UACANet-S/L [43] utilized channel attention to concentrate on polyp regions. For addressing size variation, MSRF-Net and MSRFE-Net [63] utilized multi-scale residual fusion and boosted segmentation of various polyp sizes.

Real-time segmentation effort has also been made. Jha et al. [28] employed Conditional Random Field (CRF) post-processing for enhancing contextual information, whereas Thambawita et al. [55] proposed pyramid-based augmentation for generalization. ColonSegNet [22] was explicitly designed for real-time segmentation on the Kvasir and CVC datasets, but with some loss of accuracy for the sake of speed. Jha et al. [28] then created NanoNet, which is a light model with three variants (NanoNet-A, NanoNet-B, and NanoNet-C) that balance speed and accuracy. Among these, NanoNet-A has greater accuracy with more param-

eters, whereas NanoNet-C puts greater emphasis on speed with fewer parameters.

2.3. Transformers networks

Transformers were originally proposed in natural language processing (NLP) and delivered outstanding performance [56]. Transformers use multi-head self-attention (MHSA) layers for capturing long-range dependencies. Dosovitskiy et al. [14] transferred transformers to computer vision, presenting the Vision Transformer (ViT), which represents images as a sequence of patch embeddings. Although ViT achieves good results in classification, low-resolution, single-scale feature maps are hard to use in dense prediction tasks such as segmentation and object detection.

Pyramid Vision Transformer (PVT)-based models [62, 61] overcome these shortcomings using fine-grained patches (4×4 per patch) and hierarchical pyramid architecture for high-resolution feature learning in a computationally less demanding process. Trailblazing PVT, Dong et al. [13] proposed Polyp-PVT, a polyp segmentation network that augments feature extraction using a transformer encoder and multi-level feature fusion.

Some other improvements include TransUNet [12], which combines ViT with U-Net to retain global context while maintaining precise localization. Swin-UNet [11] and Swin-UNETR [18] utilize better spatial efficiency through shifted window attention for enabling better scalability for high-resolution medical images. For colonoscopy segmentation, FCB-Former and its extension FCB-Former+SEP [67] combine convolutional and transformer blocks to improve feature representation. Lastly, NanoNet-A/C [60] provide light-weight architectures optimized for deployment on edge devices with an attempt to balance efficiency with accuracy.

3. Proposed Architecture

Existing hybrid Transformer-CNN segmentation models hardly reach the full potential of synergies between global attention and local feature extraction. We resolve this with a thoughtfully crafted Swin Transformer-CNN architecture with three innovations: (1) adaptive fusion modules to balance transformer and CNN features adaptively across different scales, (2) context-preserving skip connections that preserve spatial accuracy, and (3) a computationally efficient cross-attention bridge. The last structure illustrates enhanced performance in medical image segmentation, where current approaches fail to capture anatomical context and fine boundaries at the same time.

3.1. Encoder: Swin Transformer Backbone

We used a transformer-based backbone to perform multi-scale features extraction in encoding. It starts with an input RGB image $\mathbf{X}$ of size $\mathbb{R}^{H\times W\times 3}$. It conditions the image as it passes through the backbone to increasingly four stages at different resolutions. The model in every stage produces feature maps of certain spatial sizes and channel depths.

$$\begin{aligned}
&\mathbf{F}_1 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C} \quad \text{(Stage 1)}\\
&\mathbf{F}_2 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2C} \quad \text{(Stage 2)}\\
&\mathbf{F}_3 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 4C} \quad \text{(Stage 3)}\\
&\mathbf{F}_4 \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 8C} \quad \text{(Stage 4)}
\end{aligned} \tag{1}$$

where $\mathbf{F}_i$ denotes level i features with decreasing spatial resolutions and deeper channels. The Swin Transformer's *shifted window self-attention* supports efficient modeling of long-range dependencies without the loss of computational efficiency.

3.2. Decoder: Feature Fusion and Upsampling

The decoder generates high-resolution predictions via upsampling and feature fusion operations. Decoder block $\mathcal{D}_i$ is comprised of:

1. The features are first processed by a 33 convolutional layer with batch normalisation and ReLU activation.

$$\hat{\mathbf{F}}_i = \sigma(\text{BN}(\text{Conv}_{3\times 3}(\mathbf{F}_i))) \tag{2}$$

   where σ denotes the ReLU activation function.

2. Bilinear up-sampling (×2) to achieve spatial resolution restore:

$$\tilde{\mathbf{F}}_i = \text{Upsample}_{\times 2}(\hat{\mathbf{F}}_i) \tag{3}$$

The decoder increasingly integrates features through skip connections:

$$\begin{aligned}
\mathbf{D}_4 &= \mathcal{D}_4(\mathbf{F}_4),\\
\mathbf{D}_3 &= \mathcal{D}_3(\text{Concat}(\mathbf{D}_4, \mathbf{F}_3)),\\
\mathbf{D}_2 &= \mathcal{D}_2(\text{Concat}(\mathbf{D}_3, \mathbf{F}_2)),\\
\mathbf{D}_1 &= \mathcal{D}_1(\text{Concat}(\mathbf{D}_2, \mathbf{F}_1)).
\end{aligned} \tag{4}$$

3.3. Final Prediction Layer

The decoder output $\mathbf{D}_1 \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 64}$ is mapped to the target mask space through a 1×1 convolution:

$$\mathbf{Y} = \text{Conv}_{1\times 1}(\mathbf{D}_1), \quad \mathbf{Y} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times K}, \tag{5}$$

where K is the number of classes. Another bilinear interpolation generates the prediction in the input resolution:

$$\hat{\mathbf{Y}} = \text{Upsample}(\mathbf{Y}, \text{size} = (H, W)). \tag{6}$$

Table 1. Comparison of publicly available polyp detection/segmentation datasets.

Dataset Name (Year, Country)	Ground Truth	Images	Resolution
CVC-ColonDB (2013, Spain) [7]	Binary Mask	380	500×574
ETIS-LaribPolypDB (2014, France) [48]	Binary Mask	196	1225×966
CVC-ClinicDB (2015, Spain) [5]	Binary Mask	612	576×768
ASU-Mayo (2016, USA) [54]	Binary Mask + BBox	18,781	512×512
GI Lesions (2016, France) [3]	Annotated File + BBox	30 videos	768×576
EndoScene (2016) [44]	Binary Mask	912	224×224
CVC-ClinicVideoDB (2017) [8]	Binary Mask	11,954 frames	384×288
Kvasir-SEG (2019, Norway) [26]	Binary Mask + BBox	1,000	320×320
KvasirCapsule-SEG (2019, Norway) [52]	BBox	47,238	Varies
NBIPolyp-Ucdb (2019, Portugal) [45]	Binary Mask	86	576×720
WLPolyp-UCdb (2019, Portugal) [45]	Annotated File	3,040	726×576
KUMC (2020, Korea) [29]	BBox	4,856	224×224
SUN (2020, Japan) [31]	BBox	49,136	416×416
PICCOLO (2020, Spain) [21]	Binary Mask	3,433	854×480
CP-CHILD (2020, China) [32]	Annotated File	9,500	256×256
EDD2020 (2020, International) [1]	BBox + Binary Mask	386 videos	Varies
HyperKvasir (2020, Norway) [10]	Binary Mask	10,662	224×224
Kvasir-Capsule (2021, Norway) [20]	BBox	47,238	Varies
LD Polyp Video (2021, China) [34]	BBox	40,187 frames	560×480
SUN-SEG (2022, Japan) [16]	Multiple types	158,690	416×416
PolypGen (2022, Multi-center) [2]	Binary Mask + BBox	6,282	Various

4. Implementation details

We implemented our hybrid model in PyTorch framework and evaluated on an NVIDIA GeForce RTX 4090 GPU with 24GB VRAM. For handling variations in polyp image sizes, a multi-scale approach was employed during training. AdamW optimizer, which is appropriate for transformer-based models, was utilized with learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. The loss function combined binary cross-entropy (BCE) and Intersection over Union (IoU) to optimize segmentation accuracy.

Input images were downsized to 352×352 pixels, with a mini-batch of size 8, for 100 epochs. Training took about 1 hour, and best coscusative performance was at epoch 63. In an effort to avoid overfitting, an early stopping condition was used, inspecting the Dice score on the test set after every epoch. Training was stopped if no progress had been made in the last 37 epochs, which happened at epoch 63.

During training, the data augmentation processes like random rotation, flip left-right, and flip top-bottom were utilized. Images were rescaled to 352×352 during test time without any extra post-processing or optimization techniques. This method possessed strong performance with low computational cost.

5. Experiments

5.1. Datasets

We used our approach to the Kvasir-SEG dataset, which contains 1000 polyp images that are subclasses of the Kvasir polyp class. We divided the dataset into 900 images for training purposes and 100 images for the test set. Training was carried out once while cross-validation was applied solely on the test set within the Kvasir dataset. As can be seen from Table **??**, we achieved a test set accuracy of 0.987.

For further testing, we have tested the model on four test datasets, i.e., CVC-ClinicDB, ETIS, CVC-ColonDB, and Endotect. The test datasets are of a variety of polyp images for which generalizability estimation is possible with reliability. ETIS contains 196 images, CVC-ColonDB contains 380 images, CVC-ClinicDB contains 612 images, and Endotect contains 1000 images. Multi-dataset testing provides a certain realization about method performance in various clinical conditions. The list of all the datasets is as depicted in figure 1.

Datasets:https://bit.ly/polyp-datasets

5.2. Evaluation metrics

For the purpose of measuring our model, our chosen metrics are: Dice Score Coefficient (DSC), mean Intersection over Union (mIoU), Precision, Recall, F2, Accuracy, and Frames-per-Second (FPS).

Figure 3. Evaluation of our model in the Kvasir-SEG dataset, fifty epochs

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{DSC} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{Recall/Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{11}$$

$$\text{F2-score} = \frac{5 \times \text{Precision} \times \text{Recall}}{4 \times \text{Precision} + \text{Recall}} \tag{12}$$

$$\text{FPS} = \frac{\text{Frames Processed}}{\text{Time (seconds)}} \tag{13}$$

6. Results

We evaluated our model and compared its performance against recent state-of-the-art (SOTA) polyp segmentation models. The evaluation metrics, as shown in table 2, were used for benchmarking. On the Kvasir-SEG dataset, our method achieved a recall score of 0.9555, which is 1.76 % higher than the existing real-time SOTA method DUCK-Net. Similarly, the accuracy reached 0.9849, reflecting a little improvement in the existing DUCK-Net.

7. Conclusion

In this paper, we present high accurate image polyp segmentation, called Hybrid(Transformer + CNN), which incorporates a vision transformer backbone for efficient feature extraction with CNN skip connection layers. Experimental results on various endoscopy datasets demonstrate that our model achieves state-of-the-art performance across key metrics, including DSC, IoU, precision, recall, F2-score, and, crucially, FPS with resonalbale model parameters and inference speed.

We believe hybrid architecture offers significant potential for detecting pathological and abnormal tissues within the colon lining. One of its key advantages is the ability to identify flat polyps in challenging regions like variation in size, shape, endoscopy types, lighting, imaging protocols of the colon and detect small lesions that might be overlooked during standard endoscopy. Besides this, it can also help differentiate residual tissue after polyp removal during colonoscopy, ensuring total removal and reducing the recurrence risk.

We hope our work inspires other researchers to tackle real-time polyp segmentation tasks using hybrid-based networks. Beyond endoscopy, we envision hybrid architecture being applied in other medical fields. For example, it can assist with early diagnosis of polyp. By enhancing outcomes in cases that are under subjective clinical judgments, our technique might maximize patient therapy.

We believe our proposed method has broad implications and can contribute to advancing medical imaging and intervention techniques across multiple specialties.

References

[1] Sharib Ali et al. Edd2020: A comprehensive dataset for endoscopic artifact detection. *Endoscopy*, 52:S1–S7, 2020. 4

[2] Sharib Ali et al. Polypgen: A multi-center polyp detection and segmentation dataset. *Nature Scientific Data*, 9(1):1–12, 2022. 4

[3] Sharib Ali, Feng Zhou, Barbara Braden, Adam Bailey, Suhui Yang, Guangtao Cheng, Pengyi Zhang, Xiao Li, Max

Table 2. Performance Comparison of Polyp Segmentation Models in Kavasir-SEG dataset

Model	F1 Score	mDice	mIoU	Precision	Recall	Accuracy
U-Net† [47]	0.8655	–	0.7629	0.8593	0.8718	-
ResUNet [68]	0.7878	–	0.7778	–	–	
ResUNet++ [27]	0.8133	–	0.7927	0.7064	0.8774	-
Li-SegPNet [30]	0.9058	–	0.8800	0.9424	0.9254	-
PraNet [17]	0.9094	–	0.8339	0.9599	0.8640	-
ColonFormer [66]	–	0.927	0.877	–	–	-
DUCK-Net [59]	0.9502	–	0.9051	0.9628	0.9379	0.9842
Hybrid(Our)	0.9499	-	-	0.9422	**0.955**	**0.9849**

Kayser, Roger Soberanis-Mukul, et al. Automatic detection and classification of gi lesions for capsule endoscopy. In *IEEE International Symposium on Biomedical Imaging*, pages 516–519, 2016. 4

[4] Christophe Batailler, Jeremy Shatrov, Emmanuel Sappey-Marinier, Elvire Servien, Sebastien Parratte, and Sébastien Lustig. Artificial intelligence in knee arthroplasty: Current concept of the available clinical applications. *Arthroplasty*, 4:1–16, 2022.

[5] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 48(11):3166–3182, 2015. 4

[6] Jorge Bernal, Francisco J. Sanchez, G. Fernandez-Esparrach, D. Gil, C. Rodrıguez, and F. Vilarino. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computers in Medical Imaging and Graphics*, 43:99–111, 2015.

[7] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Wm-dova maps for accurate polyp highlighting in colonoscopy. *Medical Image Analysis*, 17(7):703–718, 2013. 4

[8] Jorge Bernal, Nima Tajkbaksh, Francisco Javier Sánchez, Bogdan J Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjørn Rustad, Ilangko Balasingham, et al. Comparative validation of polyp detection methods in video colonoscopy. *IEEE Transactions on Medical Imaging*, 36(2):496–507, 2017. 4

[9] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10211–10221, 2021.

[10] Rune Borgli et al. Hyperkvasir: A comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1):1–14, 2020. 4

[11] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021. 3

[12] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 3

[13] Bing Dong, Wenhai Wang, Deng-Ping Fan, Jian Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *ArXiv preprint*, 2021. 3

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv preprint*, 2021. 3

[15] Marco G. Espinosa, Gaston A. Otarola, Jerry C. Hu, and Kyriacos A. Athanasiou. Cartilage assessment requires a surface characterization protocol: Roughness, friction, and function. *Tissue Engineering Part C: Methods*, 27:276–286, 2021.

[16] Deng-Ping Fan et al. Sun-seg: A large-scale dataset for surgical scene segmentation. *IEEE Transactions on Medical Imaging*, 41(3):786–796, 2022. 4

[17] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 263–273. Springer, 2020. 2, 6

[18] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 272–284. Springer, 2022. 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2

[20] Steven A Hicks et al. Kvasir-capsule: A video capsule endoscopy dataset. *Nature Scientific Reports*, 11:15671, 2021. 4

[21] Alejandro Jaquéz et al. Piccolo dataset for endoscopic polyp segmentation. *Medical Physics*, 47:e123–e124, 2020. 4

[22] Debesh Jha, Safdar Ali, Nikhil Kumar Tomar, et al. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. *IEEE Access*, 9:40496–40510, 2021. 2

[23] Debesh Jha, Sharib Ali, Naman Kumar Tomar, Håvard D. Johansen, Dag Johansen, Jens Rittscher, Michael A. Riegler, and Pål Halvorsen. Doubleu-net: A deep convolutional neural network for medical image segmentation. *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 558–564, 2020. 2

[24] Debesh Jha, Pia H. Smedsrud, Daniel Johansen, et al. A comprehensive study on colorectal polyp segmentation with

resunet++, conditional random field and test-time augmentation. *IEEE Journal of Biomedical and Health Informatics*, 25:2029–2040, 2021. 2

[25] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, et al. Kvasir-seg: a segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462. Springer, Cham, 2020. 2

[26] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. *arXiv preprint arXiv:1911.07069*, 2019. 4

[27] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Daniel T. Johansen de Lange, Pål Halvorsen, and Håvard D. Johansen. Resunet++: an advanced architecture for medical image segmentation. In *Proceedings of IEEE International Symposium on Multimedia (ISM)*, pages 225–255, 2019. 2, 6

[28] Debesh Jha, Nikhil Kumar Tomar, Safdar Ali, et al. Nanonet: real-time polyp segmentation in video capsule endoscopy and colonoscopy. In *Proceedings of IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, pages 37–43, 2021. 2

[29] Tae Kyoung Kim, Chan Ho Park, Chang Min Lee, Min Ju Kang, and Hyun Gun Kim. Deep learning-based detection of polyps in colonoscopy. *Scientific Reports*, 10(1):1–9, 2020. 4

[30] Yuan Li, Hao Chen, Manning Wang, and Xiaolong Zhang. Li-segpnet: A lightweight pyramid network for real-time polyp segmentation. *IEEE Transactions on Medical Imaging*, 41(5):1124–1135, 2022. 6

[31] Daochang Liu, Ziyu Jiang, Yizhou Wang, Qiang Wang, Fei Wang, Ziyu Li, and Changhu Wang. Sun: A large-scale dataset for surgical understanding. *IEEE Transactions on Medical Robotics and Bionics*, 2(1):41–48, 2020. 4

[32] Jiang Liu et al. Cp-child: A pediatric colon polyp dataset. *Scientific Data*, 7(1):1–8, 2020. 4

[33] Jiang Liu, Tengfei Song, Meng Li, Linlin Qiao, Yitian Zhao, and Yunde Jia. Ddanet: Dual decoder attention network for automatic polyp segmentation. *Pattern Recognition*, 122:108318, 2022. 2

[34] Xiaoyu Liu et al. Ld polyp video: A large-scale colonoscopy video dataset. *Medical Image Analysis*, 70:101987, 2021. 4

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *ICCV*, 2021.

[36] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 2

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[38] Alexander V. Mamonov, Isabel N. Figueiredo, Pedro N. Figueiredo, and Yu-Hsiang R. Tsai. Automated polyp detection in colon capsule endoscopy. *IEEE Transactions on Medical Imaging*, 33:1488–1502, 2014. 2

[39] T. Matsuda, A. Ono, M. Sekiguchi, T. Fujii, and Y. Saito. Advances in image enhancement in colonoscopy for detection of adenomas. *Nature Reviews Gastroenterology Hepatology*, 14:305–314, 2017. 1

[40] The American Cancer Society Medical and Editorial Content Team. Understanding your diagnosis: Colonoscopy. https://www.cancer.org/treatment/understanding-your-diagnosis/tests/endoscopy/colonoscopy.html, n.d. Accessed: YYYY-MM-DD. 1

[41] Khairul Munadi, Khairunnisa Saddami, Masayu Oktiana, et al. A deep learning method for early detection of diabetic foot using decision fusion and thermal images. *Applied Sciences*, 12:7524–7545, 2022.

[42] NIH. Cancer stat facts: Colorectal cancer. http://www.seer.cancer.gov/statfacts/html/colorect.html, n.d. Accessed: YYYY-MM-DD. 1

[43] Seung-Jun Oh, Hyun Kim, Sang-Hoon Oh, and Seung-Won Lee. Uacanet: Unified adaptive context-aware network for polyp segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1762–1770. ACM, 2021. 2

[44] Konstantin Pogorelov, Klaus R Randel, Carsten Griwodz, Sigrun L Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter T Schmidt, et al. Deep learning and hand-crafted features for automatic polyp detection. In *IEEE International Conference on Multimedia & Expo Workshops*, pages 1–6, 2017. 4

[45] Sara Reis, Guilherme Macedo, Cláudia Bahia, and Miguel T Coimbra. Nbi and wl ucdb databases for computer-assisted detection of ulcerative lesions. *IEEE Access*, 7:74293–74302, 2019. 4

[46] Eduardo C. Rodríguez-Merchán and Pilar Gómez-Cardero. The outerbridge classification predicts the need for patellar resurfacing in tka. *Clinical Orthopaedics and Related Research*, 468:1254–1257, 2010.

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351, pages 234–241, 2015. 2, 6

[48] João Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. In *IEEE International Conference on Robotics and Biomimetics*, pages 1790–1795, 2014. 4

[49] Jorge Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery*, 9:283–293, 2014.

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. https://arxiv.org/abs/1409.1556. 2

[51] Colleen Slattery and Charles Y. Kweon. Classifications in brief: Outerbridge classification of chondral lesions. *Clinical Orthopaedics and Related Research*, 476:2101–2104, 2018.

[52] Pia H Smedsrud, Vajira Thambawita, Steven A Hicks, Håvard L Gjestang, Oda O Nedrejord, Espen Næss, Rune Borgli, Debesh Jha, Torunn Berstad, Sigrun L Eskeland, et al. Kvasir-capsule, a video capsule endoscopy dataset. *arXiv preprint arXiv:1907.05719*, 2019. 4

[53] Nima Tajbakhsh, Suryakanth R. Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging*, 35:630–644, 2015.

[54] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging*, 35(2):630–644, 2016. 4

[55] Vajira Thambawita, Steven Hicks, Pål Halvorsen, and Michael A. Riegler. Pyramid-focus-augmentation: Medical image segmentation with step-wise focus. *ArXiv preprint*, 2020. 2

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6000–6010, 2017. 3

[57] David Vazquez, Jorge Bernal, Francisco J. Sanchez, et al. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, 2017:1–9, 2017.

[58] Ioannis A. Vezakis, Konstantinos Georgas, Dimitrios Fotiadis, and George K. Matsopoulos. Effisegnet: Gastrointestinal polyp segmentation through a pre-trained efficientnet-based network with a simplified decoder, 2024.

[59] Cheng Wang, Zheng Li, Xiaolong Zhang, Minghui Qiu, and Guang Yang. Duck-net: Dense u-shaped convolutional kernel network for polyp segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1029–1040, 2023. 6

[60] Cheng Wang, Zheng Li, Xiaolong Zhang, Minghui Qiu, and Guang Yang. Nanonet: Real-time polyp segmentation with ultra-lightweight models for edge devices. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1234–1241. IEEE, 2023. 3

[61] Wenhai Wang, Enze Xie, Xiang Li, et al. Pvt v2: Improved baselines with pyramid vision transformer. *Computer Vision and Image Understanding (CVIU)*, 8:415–424, 2021. 3

[62] Wenhai Wang, Enze Xie, Xiang Li, et al. A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 548–558, 2021. 3

[63] Yujia Wang, Yu Zhang, Jindong Tian, Cheng Zhong, Yongdong Zhang, and Yiyan Shi. Msrfe-net: Multi-scale residual feature enhancement network for polyp segmentation. *IEEE Transactions on Medical Imaging*, 41(12):3721–3734, 2022. 2

[64] Ben M. Williams, Domenico Borroni, Renqiang Liu, 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*, 63:419–430, 2020.

[65] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[66] Wei Zhang, Qiang Li, Zhou Yu, Yang Chen, and Jingdong Wang. Colonformer: An efficient transformer based method for colon polyp segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(8):3970–3980, 2022. 6

[67] Yu Zhang, Jiang Liu, Qinghua Hu, and Manning Wang. Fcbformer: A fully convolutional bridge transformer with pyramid squeeze-excitation for colonoscopy segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 352–363. Springer, 2023. 3

[68] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018. 2, 6

[69] Zizhao Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15:749–753, 2018.

[70] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. *arXiv preprint arXiv:1807.10165*, 2018. 2