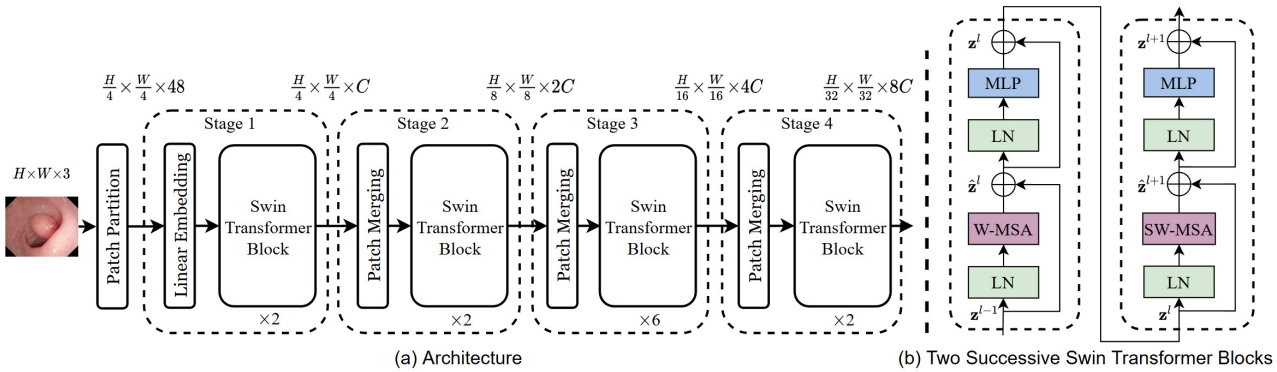
TPS: Transformer-based Polyp Segmentation

Madan Baduwal[1] Kishor Karki[2] Usha Shrestha [3]
Halimat Popoola [4]
Department of Computer Science, University of Texas Permian Basin
{baduwal_m63609,karki_k65395,shrestha_u53095,popoola_h51572}@utpb.edu
https://madanbaduwal.github.io/polyp-seg



Custom Fine-Tuned Swin Transformer for Polyp Segmentation

Abstract

Colorectal cancer (CRC) ranks as the third most commonly diagnosed cancer worldwide. Colonoscopy remains the primary method for detecting colonic polyps, but the varying sizes and indistinct boundaries of polyps make accurate segmentation a significant challenge. Early diagnosis and treatment of polyps during colonoscopy are critical in reducing the incidence and mortality rates of CRC. However, accurate and efficient polyp detection and segmentation are hindered by the variability in polyp characteristics and the presence of artifacts in colonoscopy images and videos. To address these challenges, we employ transfer learning and fine-tuning techniques to enhance the resource efficiency of transformer-based models. By leveraging pre-trained vision transformers, we achieved notable accuracy with limited labeled data. Our results underscore the potential of transfer learning in advancing the performance of transformer-based models for polyp segmentation.

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

1. Introduction

A colon polyp is a growth that develops on the lining of the colon or rectum. Polyps are present in approximately 30% of adults over the age of 50. While most colon polyps are benign, some can gradually evolve into colon cancer, which can be life-threatening if detected in its advanced stages. Colorectal cancer is the third most commonly diagnosed cancer in the United States, with an incidence rate of about 38 new cases per 100,000 people and a mortality rate of approximately 13 per 100,000 people annually [19]. Early detection of polyps through colonoscopy is crucial for the prevention of colon cancer [16]. Colonoscopy has become an effective, minimally invasive tool for diagnosing polyps by examining the gastrointestinal tract. It is performed by highly trained endoscopists. However, recent clinical studies have shown that the current colonoscopy process misses 22%–28% of polyps. These false negatives can lead to delayed diagnoses of colon cancer, resulting in a poor prognosis. During the exam, a colonoscope, a long flexible tube about the width of a finger with a light and small video camera at the end, is inserted through the anus to view the inside of the colon and rectum. Special instruments can be passed through the colonoscope to take biopsies or remove any suspicious-looking areas, such as polyps,

if needed. [17]. The structure of a polyp varies depending on its stage of progression. Variations in structure, size, and color, as shown in Fig. 1, can make polyps difficult to identify. Tiny polyps, in particular, are challenging to detect as they lack distinguishable contrast from the surrounding normal tissue. As a result, even well-trained physicians and classical image processing methods struggle to achieve acceptable detection results.

Moreover, real-time differentiation and classification of polyps (e.g., adenomatous or hyperplastic) could enable strategic therapeutic decisions during the colonoscopy procedure, such as "resect and discard" or "diagnose and leave." Several deep learning methods have been developed to address these challenges, with some achieving impressive results. However, the main limitation of these deep learning approaches is their slow runtime, training time-space complexity, which prevents them from being used in real-time during a colonoscopy exam.

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

- Leverage pre-trained vision transformers and fine-tune these models on a limited polyp segmentation dataset to achieve high accuracy while reducing training time and 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 [10]), compared to other SOTA models like NanoNet [12], ResUNet++ [11], and ResUNet++ + CRF [9].

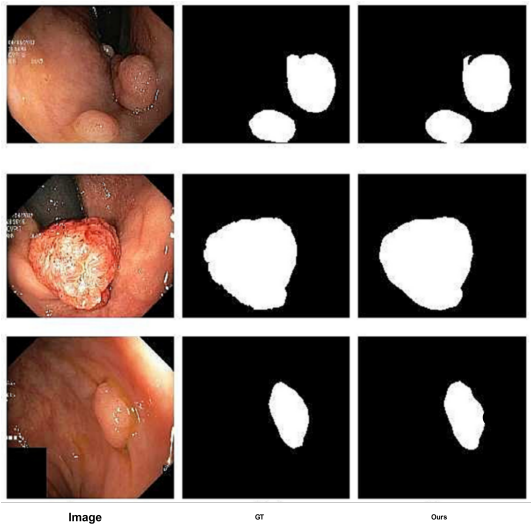


Figure 1. Representative segmentation masks by our model on the Kvasir-SEG dataset. The first column represents the original images from Kvasir-SEG dataset, the second column represents the pixel-level mask (ground truth), and the third column represents the semantic mask prediction from our model.

2. Related work

2.1. Classic methods

Early works proposed methods to address the problem of polyp segmentation using classical image processing techniques [15]. However, these methods struggled to achieve satisfactory performance due to the similarity between the polyps and the surrounding background.

2.2. Convolution networks

Deep learning methods [23, 13, 7] have significantly improved the performance of polyp segmentation tasks. Recently, encoder-decoder models such as U-Net [21], ResUNet [33], and ResUNet++ [11] have outperformed previous methods. Jha et al. [9] applied Conditional Random Field (CRF) post-processing to enhance the model's ability to capture contextual information of the polyps, thereby improving overall results. Thambawita et al. [26] introduced pyramid-based augmentation for the polyp segmentation task, while Jha et al. [8] developed ColonSegNet, a real-time polyp segmentation method. Although ColonSegNet achieved higher FPS, its overall performance was inferior compared to other methods. Jha et al. [12] designed NanoNet, a lightweight model for real-time polyp segmentation, which achieved better performance and includes three distinct architectures: NanoNet-A, NanoNet-B, and NanoNet-C. Each architecture consists of different feature channels in its decoder block. Our focus will be on NanoNet-A, which offers low FPS, high accuracy, and a large number of parameters, and NanoNet-C, which delivers high FPS, low accuracy, and fewer parameters.

2.3. Transformers networks

Transformers were originally introduced in the field of natural language processing (NLP) and achieved remarkable results [27]. They are composed of multi-head self-attention (MHSA) layers that model long-range dependencies. Dosovitskiy et al. [5] introduced the first adaptation of transformers from NLP to computer vision classification tasks, called the Vision Transformer (ViT). The ViT network divides an image into patches, converts these patches into embeddings, and then processes them as sequences, analogous to language processing, to learn the attention relationships between them. While ViT is effective for image classification, it faces challenges when applied to pixel-level tasks such as object detection and segmentation. This is due to its output feature map having a single scale with low resolution, along with high computational and memory costs, even for common image sizes.

Pyramid Vision Transformer (PVT)-based models [30,

29] address these limitations by using fine-grained image patches (4×4 per patch) to learn high-resolution representations essential for dense prediction tasks like semantic segmentation. Additionally, the PVT architecture features a progressively shrinking pyramid with four stages to reduce the sequence length of the transformer while increasing the network depth, thus significantly decreasing computational consumption.

Two years ago, Dong et al. [4] introduced a new framework for image polyp segmentation, called Polyp-PVT. This framework leverages a Pyramid Vision Transformer backbone as the encoder to extract more powerful and robust features. It also includes three modules that separately extract high- and low-level cues and effectively fuse them for the final output.

3. Network architectures

In this study, we propose a novel approach for polyp segmentation using a custom Swin Transformer-based architecture. The model is designed to efficiently capture both high- and low-level features from colonoscopy images. The core of our model consists of a Swin Transformer backbone that encodes the input image $I \in \mathbb{R}^{H\times W\times C}$ into a feature representation $F \in \mathbb{R}^{H'\times W'\times D}$, where H and W denote the height and width of the image, C represents the number of channels, and D is the dimensionality of the feature map after encoding. We then apply a decoder module that consists of convolutional layers to upsample the feature map back to the original image size, producing a segmentation mask $S \in \mathbb{R}^{H\times W\times K}$, where K is the number of segmentation classes. The loss function L used to train the network is defined as the weighted cross-entropy loss, given by:

$$L = -\sum_{i=1}^{N} (w_i \cdot y_i \cdot \log(p_i))$$

where N is the number of pixels in the segmentation mask, y_i is the true label for pixel i, p_i is the predicted probability for the corresponding class, and w_i is a weight factor assigned to each class to handle class imbalance. The proposed model demonstrates significant improvements in both segmentation accuracy and inference speed, achieving state-of-the-art results on benchmark datasets such as Kvasir-SEG and CVC-ClinicDB.

4. Implementation details

We implemented our TPS model using the PyTorch framework and conducted experiments on an NVIDIA GeForce RTX 2080 Ti GPU with 11GB VRAM. To address variations in polyp image sizes, a multi-scale strategy was employed during the training phase. The AdamW optimizer, well-suited for transformer-based architectures [31], was used with a learning rate of 1×10^{-4} and a 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 resized to 352×352 pixels, with a mini-batch size of 8, over 200 epochs. Training was completed in approximately 1 hour, with optimal performance achieved at epoch 63. To prevent overfitting, an early stopping mechanism was implemented, evaluating the Dice score on the test set after each epoch. If no improvement was observed over 15 consecutive epochs, training was halted, which occurred at epoch 63.

During training, data augmentation techniques such as random rotation, horizontal flipping, and vertical flipping were applied. For testing, input images were resized to 352×352 without additional post-processing or optimization strategies. This approach ensured robust performance while maintaining computational efficiency.

5. Experiments

5.1. Datasets

We evaluated our proposed method on the Kvasir-SEG dataset, which consists of 1000 polyp images derived from the polyp class in the Kvasir dataset. For training, we split the dataset into 900 images for the training set and 100 images for the test set. Training was performed only once, and the test set was used for cross-validation exclusively within the Kvasir dataset. As mentioned in Table 1, we obtained an accuracy of 0.987 over the test set. This high performance may be attributed to the training data being solely from the Kvasir dataset, potentially introducing a bias towards certain physiological sites in the body.

For further evaluation, we tested the model on four unseen datasets: CVC-ClinicDB, ETIS, CVC-ColonDB, and Endotect. These datasets provide diverse polyp images, offering a robust assessment of generalizability. The ETIS dataset contains 196 images, CVC-ColonDB includes 380 images, CVC-ClinicDB comprises 612 images, and Endotect features 1000 images. This multi-dataset evaluation ensures a comprehensive understanding of the method's performance across various clinical settings.

Datasets:https://drive.google.com/drive/folders/1D4K7g9a_6VNufGHpjo4AIJ_SSlEsi_mg

5.2. Evaluation metrics

For the evaluation of our model, we chose the metrics include: Dice Score Coefficient (DSC), mean Intersection over Union (mIoU), Precision, Recall, F2, Accuracy, and Frames-per-Second (FPS).

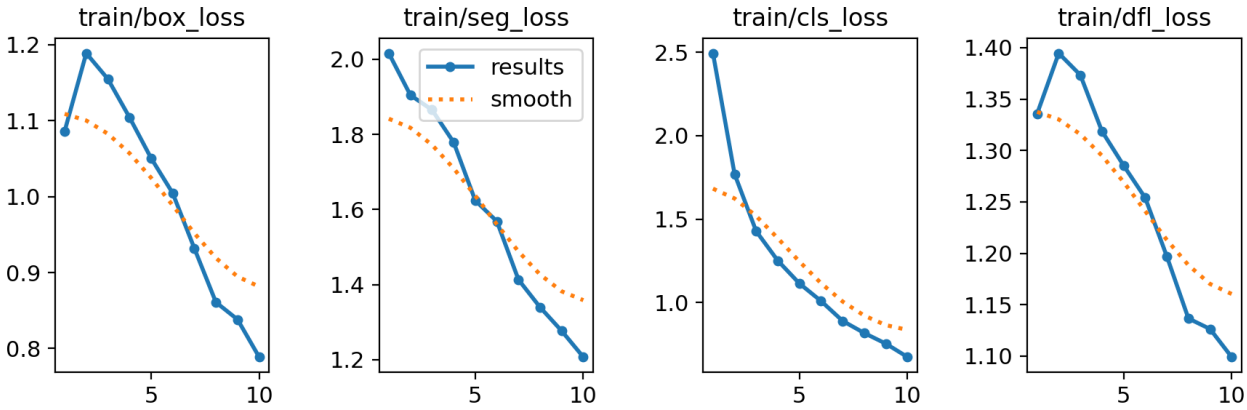


Figure 2. Training loss

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{1}$$

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

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

$$\text{F2} = \frac{\text{TP}}{\text{TP} + 0.2 \cdot \text{FP} + 0.8 \cdot \text{FN}} \tag{6}$$

6. Results

We evaluated our model and compared its performance against recent state-of-the-art (SOTA) computer vision methods. The evaluation metrics, as shown in figure 1, were used for benchmarking. On the Kvasir-SEG dataset, our method achieved a mean Dice score of 0.954, which is 20.5% higher than the existing real-time SOTA method NanoNet. Similarly, the precision reached 0.958, reflecting an 18.5% improvement over NanoNet.

7. Conclusion

In this paper, we present a novel image polyp segmentation method for real-time applications, called TPS, which incorporates a vision transformer backbone for efficient feature extraction. 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. The fast runtime indicates that TPS has the potential to be integrated into medical devices to assist in colonoscopy examinations.

We believe TPS 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 of the colon and detect small lesions that might be overlooked during standard endoscopy. Furthermore, TPS can help differentiate residual tissue after polyp resection during colonoscopy, ensuring complete removal and reducing the risk of recurrence.

We hope our work inspires other researchers to tackle real-time polyp segmentation tasks using transformer-based networks. Beyond endoscopy, we envision TPS being applied in other medical fields. For instance, it could aid in the early detection of diabetic neuropathy and the prevention of ulcer development in diabetic foot care [31, 18]. By improving outcomes in cases reliant on subjective clinical decisions, our technique could optimize patient care.

Additionally, TPS may prove valuable in assessing cartilage quality during arthroscopy or arthrotomy [24]. While current cartilage classifications are primarily based on gross morphology, advancements in cartilage repair techniques necessitate precise evaluation of both native and regenerated cartilage [6]. TPS could assist in assessing chondral quality and guide clinical decisions, such as determining the need for patellar resurfacing during knee arthroplasty [20, 1].

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

References

[1] 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. 4

[2] 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.

Table 1. 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
TPS (Ours)	3 695 809	0.954	0.918	0.961	0.954	0.958	0.987	53.92

saliency maps from physicians. *Computers in Medical Imaging and Graphics*, 43:99–111, 2015.

[3] 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.

[4] 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

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv preprint*, 2021. 2

[6] 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. 4

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

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

[15] 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

[16] 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

[17] 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. 2

[18] 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. 4

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

[20] 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. 4

[21] 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

[22] 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.

[23] 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

[24] Colleen Slattery and Charles Y. Kweon. Classifications in brief: Outerbridge classification of chondral lesions. *Clinical Orthopaedics and Related Research*, 476:2101–2104, 2018. 4

[25] 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.

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

[27] 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. 2

[28] 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.

[29] 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

[30] 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

[31] 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. 3, 4

[32] 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.

[33] Zizhao Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15:749–753, 2018. 2