Is Medical Pretraining Enough When the Modality Is Different? A Study on Endoscopic Polyp Segmentation

Dipika Boro[1] dipika_boro@uml.edu
Yu Cao[1] yu_cao@uml.edu
Benyuan Liu[1] benyuan_liu@uml.edu
Qilei Chen[1] qilei_chen1@uml.edu
[1] *University of Massachusetts, Lowell.*

Editors: Under Review for MIDL 2025

Abstract

Using pretrained models for fine-tuning is a widely adopted strategy in medical imaging, where labeled data is scarce. ImageNet remains the standard for pretraining in computer vision tasks, including medical imaging. RadImageNet, a medical-specific alternative trained on radiological data, has shown promising results in radiology-focused applications; however, its effectiveness in non-radiological modalities, such as endoscopy, remains unexplored. In this study, we conduct a focused evaluation of how transfer learning from ImageNet and RadImageNet affects performance in endoscopic segmentation. We compare two backbone architectures—ResNet-50 and ViT-Small—each integrated into a DeepLabV3+ decoder, and evaluate their performance on three public polyp segmentation datasets: CVC-ClinicDB, Kvasir-SEG and SUN-SEG. Our results show that ImageNet-pretrained models consistently outperform those pretrained on RadImageNet. These findings challenge the notion that medical-domain pretraining is universally beneficial and underscore the importance of modality alignment when selecting pretrained models for medical image analysis. Github - https://github.com/dipikaboro2/med-pretraining

Keywords: Transfer learning, Pretraining, Polyp segmentation, ViT, ResNet

1. Introduction

Transfer learning with pretrained models is a widely adopted strategy in computer vision, offering improved generalization, faster convergence, and more efficient training. This approach is especially valuable in domains where annotated data is limited, such as medical image analysis. ImageNet (Deng et al., 2009) remains the default pretraining dataset due to its large scale and strong generalization capabilities across diverse tasks, including many in the medical domain. However, ImageNet comprises natural images that differ substantially in structure and appearance from medical images.

RadImageNet (Mei et al., 2022) was proposed as a radiology-specific alternative to ImageNet, comprising of over a million medical images from CT, MRI, and ultrasound modalities. It has shown performance gains in several radiology-focused tasks. However, medical imaging spans a variety of modalities beyond those in RadImageNet, such as endoscopy and histopathology, which differ significantly in structure, color, and visual semantics. Whether radiology-specific pretraining—including from CT, MRI, and ultrasound—generalizes to these non-radiological tasks, such as polyp segmentation in endoscopic images, remains unexplored.

This study presents a systematic evaluation of ImageNet and RadImageNet as pretraining sources for endoscopic image segmentation. We investigate two representative encoder architectures—ResNet-50 (He et al., 2016), a CNN based architecture, and ViT-Small (Dosovitskiy et al., 2021), a transformer based architecture—each integrated into a unified DeepLabV3+ (Chen et al., 2018) decoder. In order to get a comprehensive analysis of performance across the two pretraining domains, these models are evaluated on three publicly available polyp segmentation benchmarks: CVC-ClinicDB (Bernal et al., 2015), Kvasir-SEG (Jha et al., 2020), and SUN-SEG (Fan et al., 2020; Ji et al., 2021, 2022).

2. Method

2.1. Framework

We follow an encoder-decoder framework, where ResNet-50 or ViT-Small are chosen as the backbone encoder as shown in Figure 1. Each encoder is pretrained on either ImageNet or RadImageNet. A DeepLabV3+ decoder is used to generate the segmentation output.

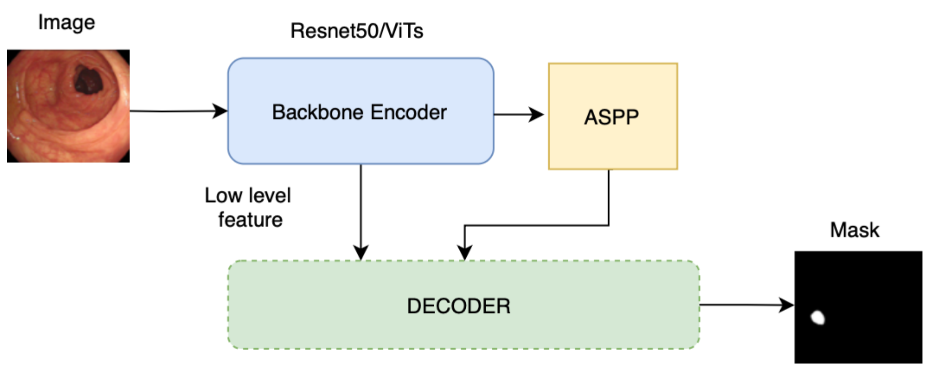

Figure 1: Architectural overview of the encoder-decoder framework.

Input images are resized to 224×224 and passed through the backbone encoder to extract hierarchical features. For ResNet-50, the final convolutional layer output is fed into an Atrous Spatial Pyramid Pooling (ASPP) module to capture multi-scale context, while a low-level feature map from an earlier layer (layer1) is used in a skip connection to the decoder. For ViT-Small, intermediate feature maps are obtained. The final transformer block output is processed by ASPP, and the earliest feature map serves as the decoder skip connection. The decoder combines ASPP output with low-level features and produces a segmentation map, which is then upsampled to the original resolution.

2.2. Experimental Setup

We evaluate our models on three publicly available polyp segmentation datasets: CVC-ClinicDB (612 image-mask pairs), Kvasir-SEG (1000 image-mask pais), and SUN-SEG (49,136 frames with expert annotations). Each dataset is randomly split into 80% training and 20% validation using a fixed seed for reproducibility.

All images and masks are resized to 224×224 and normalized. We use binary cross-entropy loss for training and the Adam optimizer with a learning rate of 0.0001. Models are trained for 20 epochs with a batch size of 32 on a single NVIDIA TITAN RTX GPU. Performance is evaluated using the Dice coefficient and Intersection over Union (IoU), computed on the validation set. Final results are reported using the best model checkpoint.

3. Results and Conclusion

Table 1 presents segmentation performance across the three datasets for backbones pretrained on ImageNet and RadImageNet. ImageNet-pretrained models consistently outperform RadImageNet-pretrained ones, with the largest differences on smaller datasets. This pattern holds for both CNN and transformer architectures. Sample results are shown in Figure 2.

Table 1: Dice and IoU scores of ImageNet and RadImageNet pretrained models.

Model	**Dataset**	**ImageNet**		**RadImageNet**	
		Dice	IoU	Dice	IoU
ResNet50	CVC-DB	0.8230	0.7389	0.5820	0.4634
	Kvasir-SEG	0.8244	0.7340	0.5279	0.4006
	SUN-SEG	0.9353	0.8862	0.9209	0.8657
ViT-S	CVC-DB	0.8705	0.7913	0.5033	0.4024
	Kvasir-SEG	0.8706	0.7986	0.5228	0.3954
	SUN-SEG	0.9000	0.8334	0.8411	0.7580

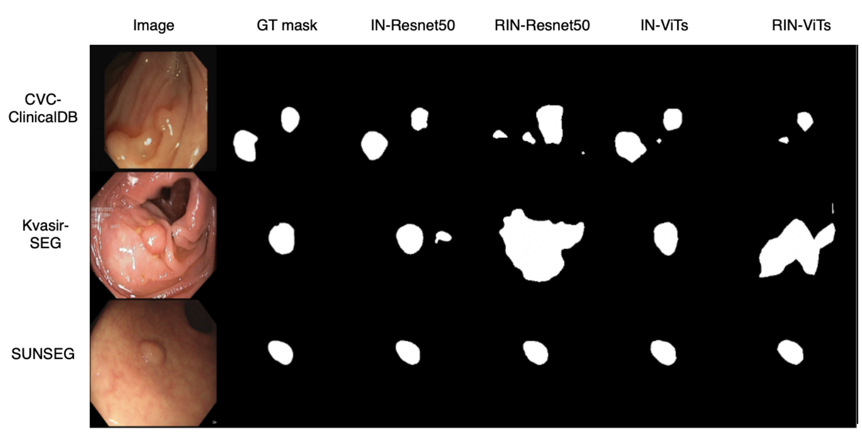

Figure 2: Predicted masks from ImageNet (IN) and RadImageNet (RIN) models.

The performance gap narrows on SUN-SEG, the largest dataset in our evaluation. However, RadImageNet-pretrained models continue to underperform slightly, suggesting that while larger datasets can mitigate the impact of pretraining misalignment, features learned from ImageNet remain more transferable to non-radiological modalities such as endoscopy. These results emphasize that medical-domain pretraining is not universally advantageous, and that modality-specific characteristics should guide the selection of pretrained models in medical imaging tasks.

References

Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. CVC-ClinicDB, 2015. URL https://polyp.grand-challenge.org/CVCClinicDB/.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.11929.

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.

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.

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. In *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020.

Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Huazhu Fu, Debesh Jha, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 142–152. Springer, 2021.

Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, 19(6):531–549, 2022.

Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. RadImageNet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5):e210315, 2022.