Representation Matters: Rethinking Domain Generalization in Polyp Segmentation

Carla Monteiro[1,2,3⋆][0009−0003−4754−3133], Valentina Corbetta[2,3,4⋆][0000−0002−3445−3011], Regina Beets-Tan[3,5][0000−0002−8533−5090], Luís F. Teixeira[1][0000−0002−4050−7880], and Wilson Silva[2,3][0000−0002−4080−9328]

[1] INESC TEC, Faculty of Engineering, University of Porto, Porto, Portugal
carla.s.monteiro@inesctec.pt
[2] AI Technology for Life, Department of Information and Computing Sciences, Department of Biology, Utrecht University, Utrecht, The Netherlands
[3] Department of Radiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands
[4] GROW Research Institute for Oncology and Reproduction, Maastricht University, Maastricht, The Netherlands
[5] Maastro, Maastricht, The Netherlands

Abstract. Automatic polyp segmentation is crucial for improving the clinical identification of colorectal cancer. While pretrained representations are reshaping computer vision, their role in medical image segmentation under domain shift remains underexplored. We propose a representation-driven framework that leverages self-attention "key" features from DINO-pretrained Vision Transformers, which capture object-centric and spatially consistent representations, enabling more robust segmentation without relying on complex task-specific architectures. We evaluate our framework on a multi-center dataset with substantial inter-center variability, using Domain Generalization (DG) and Extreme Single Domain Generalization (ESDG) protocols to simulate realistic distribution shifts. Frozen DINO features achieve performance comparable to specialized architectures, while supervised fine-tuning consistently improves results, reaching state-of-the-art performance particularly in data-scarce settings. A systematic comparison of DINO variants (V1 to V3) further shows that advances in self-supervised representation learning translate into improved cross-domain generalization. These findings suggest that representation quality, rather than architectural complexity, is the primary driver of generalization in medical image segmentation. Code: `https://github.com/Trustworthy-AI-UU-NKI/RepMatters.git`.

Keywords: Polyp segmentation · representation learning · generalizability.

1 Introduction

Colorectal Cancer (CRC) is the third most common cancer type and the second leading cause of cancer-related deaths. Around 95% of cases arise from the

⋆ These authors contributed equally to this work.

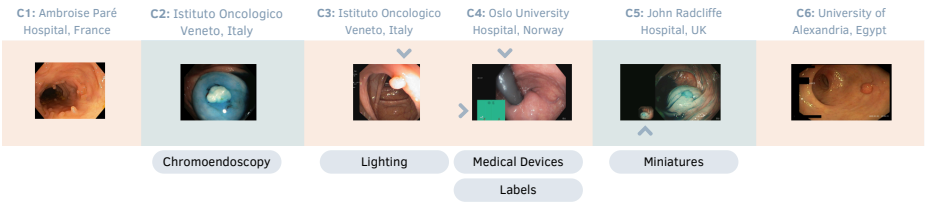

Fig. 1: Illustration of the variability of PolypGen [1] samples across different centers. Blue highlights chromoendoscopy samples, with arrows indicating other segmentation challenges, such as varying lighting and device-related labels.

development of polyps in the colon or rectum, which, if undetected, can lead to cancer [11]. Colonoscopy is the gold standard for polyp detection, but the resemblance of polyps to normal tissue makes diagnosis challenging, particularly for less experienced operators. Additionally, operator-dependent factors, such as colonoscope speed, can lead to missed detections [28]. As a result, automatic polyp segmentation has been extensively researched to enhance polyp detection and patient outcomes [29].

The introduction of U-Net [22] significantly advanced medical image segmentation and inspired numerous general [5,31] and polyp-specific [19,21,29] CNN-based approaches. To reduce the need for task-specific designs, nnU-Net [12] introduced a self-configuring framework applicable across datasets. Nevertheless, specialized architectures such as UM-Net [8], currently state-of-the-art for polyp segmentation, continue to be proposed to improve performance. In parallel, transformer-based models have gained popularity due to their ability to capture long-range contextual information [16,26].

Despite these advances, generalization remains a major challenge. Colonoscopy data is inherently heterogeneous, varying across acquisition devices, patient populations, imaging protocols, and lighting conditions [27] (see Fig. 1). Thus, models often learn spurious correlations and rely on non-clinically relevant features, leading to performance degradation when applied to unseen domains [10,18]. While transfer learning and domain adaptation might mitigate these issues, they require large annotated datasets, which are costly to obtain [30].

In contrast, Self-Supervised Learning (SSL) has emerged as a promising alternative, enabling the learning of robust representations without labeled data [13,25]. In particular, Distillation with No Labels (DINO) [3], combined with Vision Transformers (ViTs) [7], has shown strong performance in representation learning. While most approaches rely on token representations for downstream tasks, the self-attention mechanism also produces auxiliary representations. Notably, the "key" features have been shown to capture more object-centric and spatially consistent information, suggesting potential benefits for generalization [2]. Recent foundation models for medical image segmentation, such as MedSAM [17], have demonstrated impressive generalization by leveraging large-scale pretraining. However, these approaches typically rely on geometric prompts (e.g., bound-

ing boxes or point clicks) at inference time, assuming some form of user interaction. In contrast, we focus on fully automatic segmentation under domain shift, where no prompts are available at test time. This setting is particularly relevant for scalable clinical deployment. We therefore investigate whether strong generalization can emerge from self-supervised representations alone, without relying on interactive inference or complex task-specific architectures, thereby isolating the role of representation quality in generalization.

Motivated by the spatial consistency and object-centric properties of DINO "key" features, we investigate their use for fully automatic polyp segmentation under domain shift. We show that frozen representations achieve performance similar to that of specialized architectures, while supervised fine-tuning enhances this further, yielding state-of-the-art results with strong generalization, especially in data-scarce scenarios. We analyze multiple DINO variants, showing that advances in self-supervised learning improve cross-domain performance.

The main contributions of this work are: (1) a simple segmentation framework based on DINO ViT "key" features, avoiding complex task-specific architectures; (2) a study of frozen and fine-tuned DINO representations under Domain Generalization (DG) and Extreme Single Domain Generalization (ESDG), demonstrating robustness in data-scarce settings; (3) a comparison of DINO variants, analyzing the impact of representation learning on downstream performance.

2 Materials and Methods

2.1 Data

PolypGen [1] is a multi-center colonoscopy dataset collected from six hospitals with distinct demographics. It contains 3,762 annotated images, both sequence and single-frame, all validated by six senior gastroenterologists. We focus on the single-frame subset, which contains 1,537 samples distributed across institutions as follows: **C1**: 256, **C2**: 301, **C3**: 457, **C4**: 227, **C5**: 208 and **C6**: 88. Within each center, data were partitioned at the patient level, ensuring that images from the same patient were not shared across splits.

This dataset enables evaluation of segmentation generalization across multiple centers, exhibiting substantial heterogeneity. As shown in Fig. 1, images contain device-related annotations and vary in lighting and blur, while the inclusion of chromoendoscopy samples [24] further increases variability. This makes it a challenging benchmark that more accurately represents real-world clinical deployment.

2.2 Self-Distillation with No Labels (DINO)

DINO [3] is a self-supervised learning framework that learns visual representations without labeled data through a teacher–student distillation paradigm. Given multiple views of the same image, the student network is trained to

match the teacher's output representations, enabling the emergence of semantically meaningful features. When applied to Vision Transformers (ViTs) [7], this approach has been shown to yield robust representations [3,20].

In Vision Transformers, images are represented as sequences of patch tokens that interact through self-attention. This mechanism produces intermediate representations—queries, keys, and values—that encode relationships between image regions and can be interpreted as patch-level descriptors for dense prediction tasks. Prior work has shown that self-supervised ViT representations exhibit strong semantic and spatial properties, capturing well-localized object parts at high resolution [2,3]. Among these attention representations, "keys" have been shown to produce cleaner patch-level similarity patterns, making them suitable for tasks requiring spatial correspondence, such as segmentation [2].

Motivated by these findings, we leverage "key" representations from the final attention layer of the ViT as patch-level descriptors for polyp segmentation. To study the effect of representation choice and pretraining, we evaluate multiple representation extraction strategies across different DINO variants, including DINOv1 [3], DINOv2 [20], and DINOv3 [23]. For DINOv2, we additionally consider variants with register tokens [6], which improve representation quality and stability. This enables us to analyze how both representation type and pretraining advances influence downstream segmentation performance.

2.3 Proposed Model

Fig. 2 illustrates the proposed architecture, which consists of a DINO-pretrained Vision Transformer (ViT-S) feature extractor followed by a CNN-based decoder.

Following prior work on self-supervised ViT representations [2], we extract the "key" descriptors from the self-attention module of the final transformer layer. These representations provide patch-level descriptors that capture semantic structure and spatial relationships. The extracted features are then passed through a projection layer, mapping the 384-dimensional embeddings to 1024 channels to bridge the transformer and convolutional decoder.

The decoder consists of five blocks. The first four blocks perform double convolution, progressively refining the feature maps. To generate the segmentation mask, the last block uses pointwise convolution and bilinear upsampling.

We intentionally adopt a simple architecture, using a small ViT backbone and a lightweight decoder, to focus on the impact of the extracted representations.

2.4 Baselines

To validate the effectiveness of our approach, we compare it with **U-Net** [22], **nnU-Net** [12], and **UM-Net** [8]. **U-Net** is a lightweight reference baseline with similar CNN-based blocks to our decoder. **nnU-Net** is a self-configuring framework that automatically adapts its architecture, preprocessing, and training to the dataset, and is widely regarded as a strong benchmark across medical segmentation tasks [9,12]. **UM-Net** is a recent state-of-the-art method for polyp

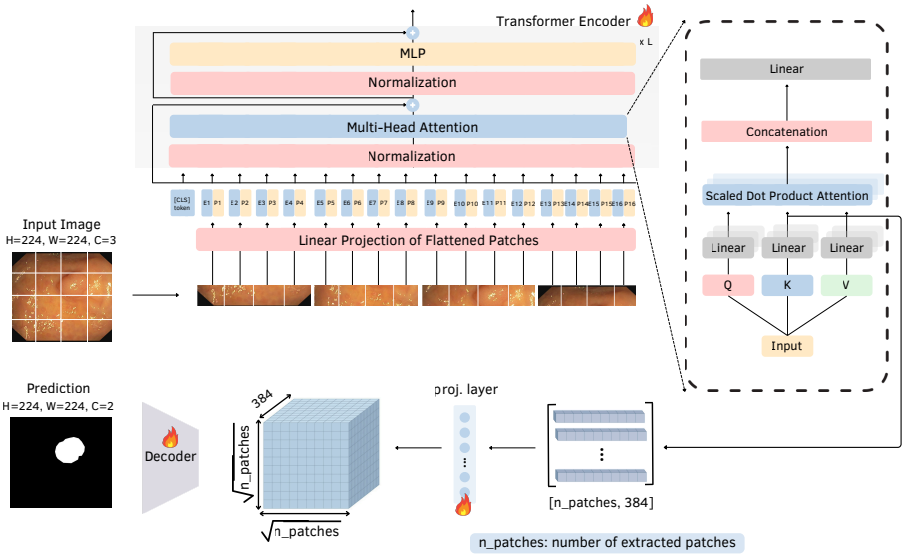

Fig. 2: Model architecture for polyp segmentation using DINO ViT "key" features. H, W and C represent the height, width and number of channels, respectively. The input image is passed through a ViT encoder, and "key" features are extracted from the multi-head attention block in the last layer. These features are reshaped and projected to match the spatial dimensions required by the decoder, which generates the final segmentation mask.

segmentation that combines multi-scale context modeling, boundary refinement, and task-specific training strategies, including specialized loss functions and a color-based augmentation. Due to its architectural complexity and strong performance, it represents a particularly challenging baseline for this task.

2.5 Implementation details

For all experiments, we use DINO-pretrained ViT-S backbones (DINOv1, DINOv2, and DINOv3) with a stride of 50% of the patch size. We evaluate both frozen and fine-tuned configurations of the ViT extractor. Input images are resized to 224×224 and normalized. Training augmentations include random flips, rotations (up to 45°), and color jittering. UM-Net additionally uses the authors' color transfer, while nnU-Net follows its own preprocessing pipeline. Models are trained on an NVIDIA RTX8000 for 300 epochs with Adam [15] (batch size 4, learning rate 1×10^{-5}), selecting checkpoints by maximum validation DSC. All models use the Dice loss, except UM-Net$_{\text{loss}}$, which follows [8].

Model	DG			ESDG		
	DSC	Recall	Acc	DSC	Recall	Acc
U-Net	$60.75_{12.78}$	$59.98_{15.73}$	$95.84_{0.80}$	$38.15_{6.78}$	$48.81_{11.01}$	$89.48_{5.80}$
nnU-Net	$65.09_{14.75}$	$65.68_{15.15}$	$95.50_{1.64}$	$52.36_{7.97}$	$57.40_{10.63}$	$93.21_{1.65}$
UM-Net$_{\text{base}}$	$69.69_{14.16}$	$69.41_{15.92}$	$96.82_{1.33}$	$61.04_{3.38}$	$66.07_{2.43}$	$95.19_{0.43}$
UM-Net	$71.44_{12.81}$	$68.46_{17.39}$	$97.13_{0.78}$	$60.77_{4.05}$	$63.73_{5.10}$	$95.07_{0.87}$
UM-Net$_{\text{loss}}$	$73.26_{13.42}$	$70.88_{16.02}$	$97.42_{0.88}$	$63.22_{3.23}$	$65.03_{4.05}$	$95.60_{0.31}$
Proposed$_{\text{V1(A)}}$	$67.92_{13.71}$	$66.00_{17.63}$	$96.98_{0.55}$	$60.50_{3.82}$	$62.21_{3.45}$	$95.50_{0.84}$
Proposed$_{\text{V2(A)}}$	$72.71_{11.98}$	$69.48_{15.67}$	$97.30_{0.60}$	$66.00_{2.88}$	$66.60_{1.63}$	$96.19_{0.46}$
Proposed$_{\text{V2R(A)}}$	$71.98_{13.02}$	$69.22_{14.43}$	$97.25_{0.76}$	$64.31_{0.96}$	$67.95_{4.25}$	$95.80_{0.48}$
Proposed$_{\text{V3(A)}}$	$70.42_{13.51}$	$67.60_{17.76}$	$97.38_{0.68}$	$63.53_{3.92}$	$69.50_{3.49}$	$95.42_{1.13}$
Proposed$_{\text{V1(B)}}$	$71.85_{13.04}$	$69.32_{17.28}$	$97.10_{0.86}$	$64.28_{4.17}$	$68.50_{2.29}$	$95.79_{0.69}$
Proposed$_{\text{V2(B)}}$	$71.88_{13.04}$	$71.38_{16.94}$	$97.07_{1.26}$	$66.19_{5.16}$	$70.61_{1.92}$	$95.79_{0.88}$
Proposed$_{\text{V2R(B)}}$	$\underline{76.34}_{12.13}$	$\underline{72.06}_{17.13}$	$\underline{97.71}_{0.79}$	$\underline{70.92}_{2.83}$	$\underline{73.17}_{2.93}$	$\underline{96.58}_{0.49}$
Proposed$_{\text{V3(B)}}$	$\mathbf{79.88}_{8.98}$	$\mathbf{75.71}_{14.88}$	$\mathbf{97.91}_{0.78}$	$\mathbf{73.47}_{3.27}$	$\mathbf{76.51}_{3.93}$	$\mathbf{96.90}_{0.37}$

Table 1: Domain Generalization (DG) and Extreme Single Domain Generalization (ESDG) test results represented as $(100 \times \text{avg})_{100\times\text{std}}$. Best performance per column is depicted in **bold**, and second best is underlined.

2.6 Generalization Assessment and Statistical Analysis

We evaluate all models under two protocols designed to assess generalization across clinical centers with varying distribution shifts [4]. In the **Domain Generalization (DG)** setting, models are trained on five centers and evaluated on the held-out center, following a leave-one-domain-out protocol. In the **Extreme Single Domain Generalization (ESDG)** setting, models are trained on a single center and evaluated on the remaining five, simulating data-scarce deployment scenarios with strong distribution shifts. To ensure rigorous evaluation, we assess the statistical significance of DSC differences across all experiments. We apply the Friedman test to compare all models and the Wilcoxon signed-rank test for pairwise comparisons, using the per-fold average DSC values. For both DG and ESDG, six-fold-level observations per model were used.

3 Results

We present the results in Table 1, using **Accuracy**, **Recall**, and **DSC** as evaluation metrics. Proposed variants are denoted by their DINO version (V1, V2, or V3), the presence of register tokens (R), and whether the ViT extractor is kept frozen (A) or fine-tuned (B). Note that R is omitted for V3, as register tokens are included by default in this version.

The proposed model achieves the best overall performance across both DG and ESDG settings, with the largest gains in the more challenging ESDG setting.

Notably, our approach maintains strong performance even in highly data-scarce scenarios (e.g., center 6 with 88 samples), highlighting its robustness under severe distribution shifts. We also observe consistent improvements across DINO variants, with the largest gains associated with the introduction of registers in DINOv2. Among the baselines, U-Net shows a substantial performance drop from DG to ESDG, indicating limited generalization. Surprisingly, nnU-Net does not substantially surpass U-Net in the DG tests; however, it does demonstrate superior generalization. Out of the baselines, UM-Net variants demonstrate the strongest generalization. Importantly, of the proposed versions, the frozen variant already performs comparably to UM-Net. Fine-tuning further improves performance in later DINO versions, consistently achieving the best results across all settings. The proposed model also exhibits the smallest relative performance drop from DG to ESDG ($\approx$ 7–11%), compared to $\approx$ 37% for U-Net, $\approx$ 20% for nnU-Net and $\approx$ 12–15% for UM-Net variants, highlighting its improved robustness to domain shifts.

The Friedman test ($\alpha = 0.05$) performed on the DSC across DG and ESDG folds yields p-values of 3.99×10^{-8} and 1.44×10^{-8}, respectively, suggesting that there is at least one model showing statistically significant improvements. Pairwise Wilcoxon tests further confirm the statistical significance of most improvements of the proposed variants over baseline methods.

While nnU-Net does not show significant improvements over U-Net in the DG setting, gains are observed in ESDG. For UM-Net, no substantial differences are found between the base model and its standard variant, whereas incorporating the full loss formulation leads to improved performance in DG tests.

For the proposed approach, the strongest variants V2R(B) and V3(B) consistently outperform all baselines and other variants with significant gains. Earlier variants, V1 and V2 improve over U-Net and nnU-Net but do not consistently surpass UM-Net. While V2 does not yield clear improvements over V1, the inclusion of register tokens leads to noticeable gains in the fine-tuned setting. Overall, advances in DINO pretraining, combined with register tokens and fine-tuning, contribute to the best performance.

For most models, fine-tuning improves performance in DG tests, while it does not lead to meaningful gains in ESDG tests, suggesting that self-supervised pretraining yields representations that are highly generalizable and well-aligned with the segmentation task. V2R and V3 are exceptions, demonstrating significant improvements with fine-tuning in both evaluations.

To isolate the effect of representation choice, Table 2 compares attention components across DINO variants in the fine-tuned setting on a single-center dataset. "Key" features achieve the highest DSC in three of four variants, with V2 being the only exception, where "value" performs marginally better. This consistent pattern supports the use of "key" representations as patch-level descriptors, in line with prior work [2].

Beyond segmentation accuracy, computational efficiency is critical for clinical integration. All proposed variants achieve inference times below 21 ms, well within the 33 ms threshold required for real-time colonoscopy video processing

Facet	V1(B)	V2(B)	V2R(B)	V3(B)
Token	87.38	91.08	92.83	91.31
Query	87.85	91.20	92.44	91.89
Key	**88.39**	90.56	**93.15**	**92.48**
Value	87.56	**91.50**	92.14	91.89

Table 2: Facet ablation DSC results. Best per version in **bold**.

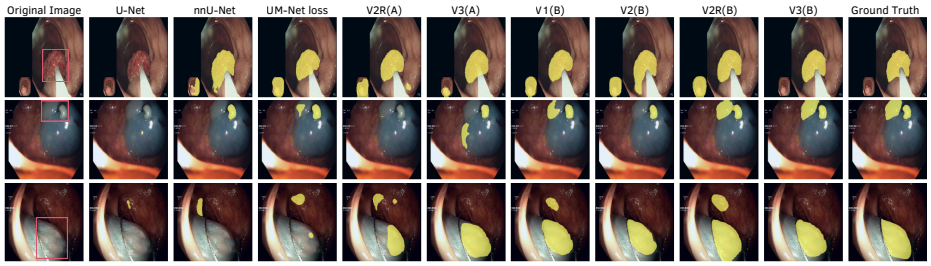

Fig. 3: Qualitative ESDG results of the baselines and variations of the proposed pipeline. Predictions and ground truth (GT) are overlaid in yellow.

at 30 FPS [14]. In contrast, UM-Net and UM-Net$_{\text{loss}}$ require over 27 ms per inference. Training times are also favorable: our variants consistently train faster than the baselines, despite achieving superior performance. This efficiency, combined with state-of-the-art generalization, makes the proposed framework particularly well-suited for deployment in resource-constrained clinical settings.

Qualitative results from the ESDG setting (Fig. 3) highlight these differences under challenging conditions, including device-related artifacts, chromoendoscopy, and miniature polyps. U-Net performs the worst, often missing polyps. While nnU-Net improves detection, it struggles with more complex cases. UM-Net is more robust but yields less precise segmentations. In contrast, the proposed variants yield more accurate and consistent predictions, particularly for small and challenging lesions, with V2R(B) and V3(B) demonstrating the most robust performance. These observations align with the quantitative results.

4 Conclusions

We investigate DINO ViT "key" features for polyp segmentation across heterogeneous clinical settings. Our results demonstrate state-of-the-art generalization, with the strongest variants achieving DSC improvements of up to 10.25 DSC points over UM-Net$_{\text{loss}}$ in the challenging ESDG setting, while exhibiting the smallest performance drop across domains (≈7–11% vs. ≈37%, ≈20% and ≈12–15% for U-Net, nnU-Net and UM-Net, respectively). Frozen features alone are already competitive with task-specific architectures, highlighting that representational quality, rather than architectural complexity, is the primary driver of

generalization. Within most DINO versions in data-scarce scenarios, the performance gap between frozen and fine-tuned variants is not statistically significant, with V2R and V3 being a notable exception. These findings establish a case for simple, pretrained-representation-based frameworks as a robust and computationally efficient alternative to specialized architectures in medical image segmentation.

Future work will focus on extending this approach to video segmentation, where real-time inference is required. Notably, all proposed variants already meet the computational constraints for colonoscopy videos, achieving inference times below 21 ms, well within the 33 ms threshold for real-time processing at 30 FPS.

Acknowledgments. Research at the Netherlands Cancer Institute is supported by grants from the Dutch Cancer Society and the Dutch Ministry of Health, Welfare and Sport. The authors acknowledge the Research High Performance Computing (RHPC) facility of the Netherlands Cancer Institute (NKI).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ali, S., Jha, D., Ghatwary, N., Realdon, S., Cannizzaro, R., Salem, O.E., Lamarque, D., Daul, C., Riegler, M.A., Anonsen, K.V., et al.: A multi-centre polyp detection and segmentation dataset for generalisability assessment. Scientific Data **10**(1), 75 (2023)
2. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep vit features as dense visual descriptors. arXiv preprint arXiv:2112.05814 **2**(3), 4 (2021)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Che, H., Cheng, Y., Jin, H., Chen, H.: Towards generalizable diabetic retinopathy grading in unseen domains. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 430–440. Springer (2023)
5. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
6. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: International conference on learning representations. vol. 2024, pp. 2632–2652 (2024)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
8. Du, X., Xu, X., Chen, J., Zhang, X., Li, L., Liu, H., Li, S.: Um-net: Rethinking icgnet for polyp segmentation with uncertainty modeling. Medical Image Analysis **99**, 103347 (2025)
9. Ferrante, M., Rinaldi, L., Botta, F., Hu, X., Dolp, A., Minotti, M., De Piano, F., Funicelli, G., Volpe, S., Bellerba, F., et al.: Application of nnu-net for automatic segmentation of lung lesions on ct images and its implication for radiomic models. Journal of clinical medicine **11**(24), 7334 (2022)

10. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
11. Gupta, M., Mishra, A.: A systematic review of deep learning based image segmentation to detect polyp. Artificial Intelligence Review **57**(1), 7 (2024)
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
13. Jing, L., Tian, Y.: Self-supervised visual feature learning with deep neural networks: A survey. IEEE transactions on pattern analysis and machine intelligence **43**(11), 4037–4058 (2020)
14. Kim, B.S., Cho, M., Chung, G.E., Lee, J., Kang, H.Y., Yoon, D., Cho, W.S., Lee, J.C., Bae, J.H., Kong, H.J., et al.: Density clustering-based automatic anatomical section recognition in colonoscopy video using deep learning. Scientific Reports **14**(1), 872 (2024)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015), `https://arxiv.org/abs/1412.6980`
16. Ling, T., Wu, C., Yu, H., Cai, T., Wang, D., Zhou, Y., Chen, M., Ding, K.: Probabilistic modeling ensemble vision transformer improves complex polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 572–581. Springer (2023)
17. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)
18. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)
19. Nguyen-Mau, T.H., Trinh, Q.H., Bui, N.T., Thi, P.T.V., Nguyen, M.V., Cao, X.N., Tran, M.T., Nguyen, H.D.: Pefnet: Positional embedding feature for polyp segmentation. In: International Conference on Multimedia Modeling. pp. 240–251. Springer (2023)
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
21. Patel, K., Bur, A.M., Wang, G.: Enhanced u-net: A feature enhancement network for polyp segmentation. In: 2021 18th conference on robots and vision (CRV). pp. 181–188. IEEE (2021)
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
23. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
24. Song, L.M.W.K., Adler, D.G., Chand, B., Conway, J.D., Croffie, J.M., DiSario, J.A., Mishkin, D.S., Shah, R.J., Somogyi, L., Tierney, W.M., et al.: Chromoendoscopy. Gastrointestinal endoscopy **66**(4), 639–649 (2007)
25. Tendle, A., Hasan, M.R.: A study of the generalizability of self-supervised representations. Machine Learning with Applications **6**, 100124 (2021)
26. Wang, A., Wu, M., Qi, H., Shi, H., Chen, J., Chen, Y., Luo, X.: Pyramid transformer driven multibranch fusion for polyp segmentation in colonoscopic video

images. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 2350–2354. IEEE (2023)
27. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 699–708. Springer (2021)
28. Wu, Z., Lv, F., Chen, C., Hao, A., Li, S.: Colorectal polyp segmentation in the deep learning era: A comprehensive survey. arXiv preprint arXiv:2401.11734 (2024)
29. Yeung, M., Sala, E., Schönlieb, C.B., Rundo, L.: Focus u-net: A novel dual attention-gated cnn for polyp segmentation during colonoscopy. Computers in biology and medicine **137**, 104815 (2021)
30. Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B.J., Roth, H., Myronenko, A., Xu, D., et al.: Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE transactions on medical imaging **39**(7), 2531–2540 (2020)
31. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: International workshop on deep learning in medical image analysis. pp. 3–11. Springer (2018)