NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME

"Data Science and Machine Learning"

Methodological comparative study of machine learning techniques for the detection of polyps in endoscopy images

Postgraduate Diploma Thesis

Alexis Milionis

Supervisor: George Matsopoulos, Professor

Athens, October 2025

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME

"Data Science and Machine Learning"

Methodological comparative study of machine learning techniques for the detection of polyps in endoscopy images

Postgraduate Diploma Thesis

Alexis Milionis

Supervisor: George Matsopoulos, Professor

The postgraduate diploma thesis has been approved by the examination committee

1st member:	2nd member:	3rd member:
George Matsopoulos	Panagiotis Tsanakas	Athanasios Panagopoulos
Professor	Professor	Professor
School of Electrical & Computer Engineering	School of Electrical & Computer Engineering	School of Electrical & Computer Engineering
NTUA	NTUA	NTUA

Athens, October 2025

Alexis Milionis
Graduate of the Interdisciplinary Postgraduate Programme,
“Data Science & Machine Learning”,
Master of Science,
School of Electrical and Computer Engineering,
National Technical University of Athens

Contents

Abstract

This thesis investigates the problem of automatic polyp segmentation in gastrointestinal endoscopy using machine learning semantic segmentation approaches. In our study, we compare six different architectures that are common in medical imaging: UNet, AttentionUnet, SegResNet, EffiSegNet, UNETR, and SwinUNETR. We used PolypGen dataset's centres C1-C5 to train all compared models from scratch, and centre C6 for evaluation. A notable challenge in our study is the significant domain shift observed in every center, and especially the evaluation center C6, which includes different populations, lighting conditions, image resolutions and field of view.
The pipeline is created using Pytorch and Lightning frameworks for both the model architectures and the training loops. Hydra framework is used for configuration management, and MONAI for medical imaging components. Results are logged and visualized in TensorBoard.
Our pipeline evaluates the six beforementioned architectures, using a variety of metrics common in segmentation tasks, and then creates an ensemble method, that combines predictions from the three best performing architectures through a soft voting mechanism. The ensemble method seems to outperform all previously tested individual architectures, by a slight margin.

Keywords

gastrointestinal endoscopy, polyp segmentation, medical imaging, deep learning, domain shift, UNet, Attention UNet, SegResNet, EffiSegNet, UNetR, SwinUNetR, ensemble methods

Περίληψη

Η παρούσα διπλωματική εργασία εξετάζει το πρόβλημα της αυτόματης τμηματοποίησης πολυπόδων σε γαστρεντερική ενδοσκόπηση, χρησιμοποιώντας προσεγγίσεις τμηματοποίησης μεθόδων μηχανικής μάθησης. Συγκρίνονται έξι διαφορετικές αρχιτεκτονικές, οι οποίες είναι ευρέως διαδεδομένες στην ιατρική απεικόνιση: UNet, AttentionUNet, SegResNet, EffiSegNet, UNETR και SwinUNETR. Για την εκπαίδευση των μοντέλων εκ του μηδενός, χρησιμοποιήθηκαν τα κέντρα C1–C5 του συνόλου δεδομένων PolypGen, ενώ για την αξιολόγηση χρησιμοποιήθηκε το κέντρο C6. Ένα σημαντικό ζήτημα που αναδείχθηκε στη μελέτη είναι το έντονο domain shift μεταξύ των κέντρων, και ιδιαίτερα στο κέντρο αξιολόγησης C6, το οποίο περιλαμβάνει διαφορετικούς πληθυσμούς, συνθήκες φωτισμού, αναλύσεις εικόνας και οπτικά πεδία.
Η υλοποίηση πραγματοποιήθηκε με χρήση PyTorch και Lightning τόσο για τις αρχιτεκτονικές μοντέλων όσο και για την εκπαίδευση. Για τη διαχείριση των configuration αξιοποιήθηκε το Hydra framework, ενώ για τα στοιχεία ιατρικής απεικόνισης χρησιμοποιήθηκε το MONAI. Η καταγραφή και η οπτικοποίηση των αποτελεσμάτων πραγματοποιήθηκαν μέσω του TensorBoard.
Η προτεινόμενη υποδομή αξιολογεί τις προαναφερθείσες έξι αρχιτεκτονικές, χρησιμοποιώντας ένα σύνολο μετρικών που είναι καθιερωμένες σε segmentation, και στη συνέχεια δημιουργεί μία συνδυαστική μέθοδο, η οποία συμψηφίζει τα αποτελέσματα των τριών αποδοτικότερων αρχιτεκτονικών μέσω μηχανισμού soft voting. Η συνδυαστική μέθοδος φαίνεται να υπερτερεί όλων των μεμονωμένων αρχιτεκτονικών που δοκιμάστηκαν, με μικρή διαφορά.

Λέξεις-Κλειδιά

Γαστρεντερική ενδοσκόπηση, τμηματοποίηση πολύποδων, ιατρική απεικόνιση, βαθιά μάθηση, domain shift, UNet, Attention UNet, SegResNet, EffiSegNet, UNetR, SwinUNetR, συνδυαστικές μέθοδοι

Acknowledgements

I would like to express my deepest gratitude to all those who have supported me throughout the completion of this thesis. First and foremost, I am grateful to my advisors, Professor George Matsopoulos as my supervisor, as well as Ph.D. candidate Ioannis Vezakis, for their invaluable guidance and insightful feedback.

I extend my heartfelt thanks to the faculty and staff of TEAM Master of Science program at NTUA for providing a stimulating academic environment and for their assistance in various stages of my MSc journey.

Lastly, I owe an immense debt of gratitude to my family for their belief, support and encouragement.

Thank you all for your support and contributions to this work.

1 Introduction

Colorectal cancer (CRC) has emerged as a critical public health issue worldwide, ranking as the third most commonly diagnosed cancer and the second leading cause of cancer-related deaths globally. According to comprehensive data from the World Health Organization (WHO) and the International Agency for Research on Cancer (IARC), over 1.9 million new CRC cases and more than 930,000 deaths occurred in 2020. These figures are projected to escalate significantly, with estimates indicating a rise to 3.2 million annual cases and 1.6 million deaths by 2040, reflecting increases of 63% in incidence and 73% in mortality, principally driven by demographic changes such as population growth and aging [3]. The burden is not uniformly distributed; incidence rates are highest in regions with very high Human Development Index (HDI) such as Europe, Australia, and New Zealand, whereas mortality is disproportionately elevated in Eastern Europe and transitioning economies, highlighting health system disparities and variable access to screening and treatment. Notably, countries undergoing rapid economic transition experience rising CRC incidence, concurrent with lifestyle Westernization, obesity prevalence, and dietary shifts, which modulate risk factor exposure. In addition, there is a concerning global trend of increasing CRC incidence among younger adults under 50 years, often linked to hereditary predispositions and lifestyle factors such as low physical activity and unhealthy diets. These trends underscore the imperative for targeted cancer control strategies, including enhanced primary prevention focusing on modifiable risk factors, improved early detection via screening programs, and equitable access to high-quality treatment. Addressing regional and socioeconomic disparities is essential to reduce the future CRC burden and achieve meaningful reductions in morbidity and mortality worldwide. [40], [39]

Polyps are abnormal tissue growths arising from the mucosal lining of the colon or rectum and are classified based on their size, shape, morphology, and histological type. The Paris classification system, widely accepted for describing polyp morphology in vivo, categorizes lesions into polypoid types (pedunculated, sessile, and subpedunculated) and non-polypoid types (flat, superficial elevated, flat, and depressed), with additional categories for excavated or ulcerated lesions. Lateral spreading tumors (LSTs), which are superficially spreading flat lesions larger than 10 mm, are further subclassified by granularity and nodularity, reflecting their risk of malignancy—with nongranular pseudo-depressed types carrying the highest malignant potential. Histologically, polyps are broadly divided into non-neoplastic (e.g., hyperplastic) and neoplastic polyps, with adenomatous polyps (including tubular, tubulovillous, and villous adenomas) recognized as precursors to colorectal cancer due to their dysplastic potential. Serrated polyps, such as sessile serrated lesions, also contribute to carcinogenesis through distinct molecular pathways. The Narrow-Band Imaging In-

ternational Colorectal Endoscopic (NICE) classification further refines endoscopic differentiation among hyperplastic, adenomatous, and malignant lesions based on vascular and surface pattern characteristics without requiring optical magnification. Polyps generally occur sporadically but may also result from hereditary syndromes like familial adenomatous polyposis; their development is influenced by genetic mutations along with environmental and lifestyle risk factors including age, family history, diet, smoking, and obesity. Detecting and removing these polyps primarily via colonoscopy remains paramount, as timely intervention can prevent malignant transformation and significantly reduce colorectal cancer incidence and mortality. The detailed morphological and histological understanding aids clinicians in risk stratification and tailored management strategies to improve patient outcomes [28].

Colonoscopy is widely regarded as the current gold standard in the early detection and diagnosis of colorectal abnormalities, particularly for identifying colon polyps, which are potential precursors to colorectal cancer. This procedure enables direct and comprehensive visualization of the entire colon—from the rectum through the cecum—allowing clinicians not only to detect lesions with high sensitivity and specificity but also to perform immediate therapeutic interventions such as polypectomy during the same session. Compared to other screening methods like flexible sigmoidoscopy, CT colonography, or stool-based tests, colonoscopy demonstrates superior performance with the greatest sensitivity for advanced neoplasia, substantially reducing colorectal cancer incidence by up to 69% and mortality by up to 88% when used in routine screening programs. Its unique capacity to combine diagnosis with treatment contributes decisively to improving patient outcomes through early intervention. Recent advancements in the technique and technology of colonoscopy have further enhanced its effectiveness. Innovations including high-definition imaging, chromoendoscopy, electronic chromoendoscopy, patient positioning strategies during withdrawal, and adjunctive devices such as Endocuff and distal caps have optimized mucosal visualization and lesion detection. Additionally, the integration of computer-aided detection (CAD) systems employing deep learning algorithms shows promise in reducing polyp miss rates and augmenting the endoscopist's diagnostic accuracy. While emerging non-invasive methods like FDA-approved blood tests (e.g., the Shield test) offer more accessible but less sensitive alternatives, especially for precancerous lesions, colonoscopy remains the cornerstone for colorectal cancer screening, prevention, and diagnosis. Moreover, updated guidelines from leading health organizations recommend initiating colonoscopy screening at age 45 for average-risk populations, reflecting the rising incidence of early-onset colorectal cancer. The ability of colonoscopy to detect and remove precancerous polyps during screening is critical, given that early-stage colorectal cancers have substantially better survival rates than advanced disease. Despite challenges such as procedure-related risks and

the need for sedation, the overall risk-benefit profile firmly supports colonoscopy as the most effective strategy for colorectal cancer control to date. [47]

As medical imaging technologies advance, there is a growing demand for accurate and efficient tools to assist clinicians in polyp detection, given that manual colonoscopy is associated with substantial miss rates estimated between 14-30%, varying significantly with polyp type, size, and morphology. A meta-analysis of over 15,000 tandem colonoscopies revealed adenoma miss rates of approximately 26% and serrated polyp miss rates of 27%, with small, flat, or proximal lesions being particularly prone to oversight. This variability in detection rates among endoscopists—ranging widely from below 10% to over 50% adenoma detection rate (ADR)—poses a critical limitation in colorectal cancer prevention, as missed polyps substantially contribute to post-colonoscopy colorectal cancers. Each 1% increase in ADR translates to an estimated 3% reduction in colorectal cancer risk, underscoring the urgency of improving detection consistency. To address these challenges, computer-aided detection (CADe) systems powered by deep learning have demonstrated considerable promise. These systems serve as a real-time "second observer," analyzing colonoscopy video streams with high sensitivity and specificity, consistently improving ADRs by approximately 14-24% across multiple randomized controlled trials and meta-analyses. For instance, CADe-assisted colonoscopy has been shown to reduce adenoma miss rates by over 50%, significantly mitigating the human errors inherent in manual detection. Modern CNN architectures and foundation models fine-tuned for polyp detection achieve state-of-the-art accuracy, recall, and precision, including for subtle, flat, or small polyps that are otherwise challenging to detect. Furthermore, real-world evaluations using robust datasets with precise polyp size and type annotations validate the clinical feasibility of CADe systems, which operate effectively under typical endoscopic conditions, enhancing lesion detection without prolonging procedure times. The integration of CADe in routine colonoscopy workflows not only improves diagnostic quality but also holds promise for reducing colorectal cancer incidence and mortality through earlier and more consistent detection of premalignant lesions. While some heterogeneity in CADe performance exists depending on clinical context and baseline operator skill, the overall evidence supports its role as a transformative adjunct tool in colorectal cancer screening. [52], [20], [33].

This thesis addresses the complex challenge of training deep learning architectures from scratch and evaluating their performance using the PolypGen dataset [6], a comprehensive multi-center dataset specifically curated for polyp detection and segmentation tasks in colonoscopy images. Unlike many popular segmentation datasets used in academia, PolypGen introduces a substantial domain shift between samples as it incorporates data from six distinct medical centers across Europe and Africa, encom-

passing over 1,500 polyp images, 2,225 positive video sequences, and 4,275 negative frames from diverse populations, different endoscopic systems, and varying expert operators. This diversity introduces significant variability in image appearance, polyp size, morphology, and acquisition protocols, making generalization a critical but difficult goal for model development. Given these domain shifts, single-model architectures often struggle to maintain consistent performance across all dataset subsets, highlighting the importance of robust methods capable of mitigating overfitting to specific center distributions. To this end, this work proposes a final ensemble scheme that combines multiple complementary deep learning models to utilize their individual strengths and address their unique limitations. By ensembling the three best-performing architectures, EffiSegNet-B4 [57], Attention U-Net [45], and SegResNet [42], the approach aims to capitalize on their distinct feature extraction capabilities, attention mechanisms, and residual learning frameworks respectively, leading to a more reliable and accurate segmentation output. The ensemble method is designed to improve segmentation performance by reducing false negatives and false positives that arise from individual model weaknesses, thereby enhancing the robustness and generalization of the system across the challenging PolypGen dataset. This strategy aligns with recent advances in medical imaging, where multi-model ensembles have shown superior results in handling heterogeneous data and complex clinical scenarios. Ultimately, the proposed approach contributes to advancing automatic polyp segmentation techniques with potential clinical implications for improving colorectal cancer screening and early diagnosis.

2 Related Work

The field of automated polyp detection and segmentation in gastrointestinal endoscopy has experienced substantial advancement through the application of deep learning methodologies, driven by the need to address the significant miss rates inherent in conventional colonoscopic procedures. Research reveals that manual polyp detection suffers from considerable limitations, with adenoma miss rates ranging from 14-30% even among experienced endoscopists, and these rates can exceed 60% in procedures where multiple polyps are present. This variability in detection performance directly correlates with colorectal cancer prevention efficacy, as each 1% increase in adenoma detection rate corresponds to an estimated 3% reduction in colorectal cancer risk. [26], [49], [32], [34].

2.1 Computer-Aided Detection Systems

The development of computer-aided detection (CAD) systems has emerged as a promising solution to mitigate human limitations in polyp identification. Real-time CAD systems utilizing deep learning algorithms have demonstrated the capacity to function as a ”second observer” during colonoscopic procedures, with studies reporting sensitivity rates of 95-99% and specificity rates exceeding 95%. Notably, clinical evaluations of AI-based polyp detection systems have shown improvements in adenoma detection rates of approximately 14-24% when compared to unassisted colonoscopy procedures. Contemporary CAD systems achieve real-time performance processing capabilities of 180+ frames per second, substantially exceeding the 25-30 frames per second required for clinical deployment [49], [46]. The EndoVigilant system, representative of advanced CAD implementations, utilizes specialized single shot detector algorithms optimized for real-time performance, capable of simultaneously highlighting multiple polyps within a single video frame. Clinical validation studies demonstrate that such systems can detect 98.8% of polyps identified by endoscopists during routine procedures, while maintaining frame-based specificity rates of 97.2%. However, the clinical benefit of CAD systems appears to be modulated by operator experience, with limited improvement observed among highly experienced endoscopists who already maintain high baseline adenoma detection rates exceeding 37%. [46], [49].

2.2 UNet and Encoder-Decoder Architectures

The UNet architecture has established itself as the foundational framework for medical image segmentation tasks, providing the archetypal encoder-decoder structure

with skip connections that preserve spatial information across multiple resolution scales. The architecture's distinctive U-shaped design enables the combination of low-level spatial details with high-level semantic information through direct feature concatenation between corresponding encoder and decoder levels. Recent investigations into UNet's skip connection mechanisms reveal that these connections face fundamental limitations in multi-scale feature interaction and rely on simplistic concatenation operations that constrain efficient information integration [53], [63], [58], [27]. Subsequent developments have focused on enhancing the encoder components, with Attention U-Net introducing spatial attention mechanisms that selectively emphasize relevant anatomical structures while suppressing background regions. The attention gates compute spatial coefficients using additive attention formulations, enabling grid-based gating that focuses on local spatial regions without requiring external region-of-interest cropping. Clinical applications of Attention U-Net demonstrate improved sensitivity in polyp detection tasks, with particular effectiveness in handling complex tissue morphologies and varying imaging conditions. [36], [27].

2.3 Residual Learning Frameworks

The integration of residual learning principles has significantly advanced medical image segmentation capabilities through architectures such as SegResNet and ResUNet++ variants. These frameworks address the vanishing gradient problem inherent in deep networks by implementing identity skip connections that facilitate direct information propagation across network layers. SegResNet specifically incorporates a variational autoencoder branch during training for encoder regularization, improving generalization performance on limited medical datasets while maintaining computational efficiency during inference [13], [61], [5]. Contemporary residual architectures such as ESRNet demonstrate enhanced feature learning capabilities through strategic incorporation of efficient downsampling techniques and batch normalization, enabling in-depth hierarchical feature extraction while preserving gradient flow during training. The ResUNet++ architecture further extends these concepts by implementing Atrous Spatial Pyramid Pooling (ASPP) modules and attention mechanisms, achieving Jaccard indices exceeding 98% in brain tumor segmentation tasks [13], [10].

2.4 Vision Transformers in Medical Imaging

The adaptation of transformer architectures for medical image analysis represents a paradigmatic shift from purely convolutional approaches, leveraging self-attention mechanisms to capture long-range dependencies that are challenging for traditional

CNN architectures. Vision Transformers (ViTs) reformulate image segmentation as sequence-to-sequence prediction problems, wherein input images are partitioned into patches and processed as 1D embedding sequences [31], [54], [29]. Recent comparative studies demonstrate that transformer-based models exhibit superior performance in complex neuroimaging tasks requiring sophisticated feature analysis, with ViTs achieving average accuracies of 88.01% compared to 84% for CNN architectures across medical imaging classification tasks. However, pure transformer approaches often struggle with detailed local spatial information critical for precise boundary delineation in medical segmentation tasks [31], [54], [29].

2.5 Hierarchical Transformer Architectures

SwinUNETR represents a significant advancement in transformer-based medical image segmentation by implementing hierarchical Swin Transformers with shifted window self-attention mechanisms. The architecture achieves state-of-the-art performance on medical segmentation benchmarks, with Dice coefficients exceeding 82% on multi-organ segmentation tasks while maintaining computational efficiency through localized attention computations. [21], [56]. UNETR architectures reformulate 3D medical image segmentation by replacing traditional convolutional encoders with pure transformer encoders, enabling global context modeling across volumetric data. However, empirical evaluations reveal that UNETR models often exhibit training instability and reduced performance compared to hybrid approaches when applied to challenging datasets with significant domain [27].

2.6 CNN-Transformer

The convergence of convolutional and transformer paradigms has yielded hybrid architectures that synergistically combine local feature extraction capabilities of CNNs with global context modeling of transformers. These hybrid approaches address the inherent limitations of both architectural families, with CNNs providing detailed spatial information and transformers contributing long-range dependency modeling [54], [62], [12]. Recent implementations such as CTHP (CNN-Transformer Hybrid model for Polyp segmentation) demonstrate superior performance over individual architectures, achieving mDice scores exceeding 0.85 on polyp segmentation benchmarks while maintaining computational efficiency suitable for clinical deployment. The integration typically involves parallel processing paradigms where convolutional and transformer pathways operate simultaneously before feature fusion [62].

2.7 Multi-Scale Feature Fusion Methodologies

Advanced feature fusion methodologies have emerged to address the semantic gap between low-level spatial details and high-level contextual information. TransCeption introduces inception-like modules into transformer encoders, enabling multi-scale representation capture within single stages through ResInception Patch Merging and Multi-Branch transformer architectures. The Three-Branch Feature Fusion Network (TBSFF) approach demonstrates enhanced segmentation performance by dynamically selecting semantic information from multiple levels through attention-based mechanisms [5], [8].Contemporary research indicates that traditional skip connections in UNet architectures may reintroduce domain-specific information that impedes generalization performance. Empirical studies reveal that removing uppermost skip connections can improve segmentation performance by up to 13% in cross-domain scenarios, suggesting that shallow layer features may be more susceptible to domain shifts than deeper representations [58].

2.8 Diversity-Promoting Ensemble Approaches

Ensemble methodologies have demonstrated consistent improvements in medical image segmentation accuracy through the strategic combination of complementary model architectures. Diversity-promoting ensemble (DiPE) strategies utilize Dice coefficient correlations between model pairs to estimate output correlation, selecting models with low inter-prediction similarity to maximize ensemble diversity. These approaches surpass both individual model performance and conventional top-scoring model selection strategies by leveraging architectural diversity rather than pure performance rankings [14], [18].

2.9 Soft Voting and Probabilistic Aggregation

Contemporary ensemble implementations employ soft voting mechanisms that preserve continuous probability distributions rather than discrete binary predictions, enabling confidence-weighted aggregation strategies. The EnsembleEdgeFusion technique demonstrates superior performance by combining DeepLabv3+, U-Net, DANet, and FastFCN architectures, achieving Mean Intersection over Union scores of 77.73% through strategic architectural complementarity. These probabilistic ensemble approaches effectively address individual model limitations such as target pixel mixing, imprecise boundary delineation, and insufficient feature information extraction [14], [27].

2.10 Multi-Center Dataset Variability

The generalizability challenge in medical image segmentation is exemplified by multi-center datasets such as PolypGen, which incorporates colonoscopy data from six distinct medical centers across Europe and Africa. This dataset design introduces substantial domain shifts arising from different endoscopic systems, imaging protocols, patient populations, and operator expertise levels. Empirical evaluations reveal that models trained on single-center data often demonstrate significant performance degradation when applied to external validation sets, emphasizing the critical importance of domain-aware training strategies [5], [7], [19], [41].

2.11 Domain Adaptation Strategies

Domain shift mitigation approaches encompass dataset alignment methodologies, dataset enlargement through physics-inspired augmentations, and representation alignment techniques such as Domain Adversarial Neural Networks (DANN). Physical imaging parameter variations constitute a primary driver of domain shift effects, with studies demonstrating that residual domain signature information persists even after state-of-the-art preprocessing and normalization procedures. Federated learning approaches represent an emerging paradigm for addressing multi-center variability while preserving patient privacy, with comparative studies indicating performance comparable to centralized learning approaches despite smaller sample [19], [41], [30], [38].

2.12 Performance Evaluation and Benchmarking

Contemporary polyp segmentation research employs standardized evaluation metrics including Dice coefficient, Intersection over Union (IoU), precision, recall, and F-score variants to ensure reproducible performance comparisons. The F2 score has gained particular prominence in medical applications due to its emphasis on recall over precision, reflecting the clinical priority of minimizing false negative detections over false positive identifications. Recent benchmarking studies reveal that compound loss functions combining Dice and Cross-Entropy losses provide superior optimization characteristics compared to individual loss formulations [23], [9], [2], [26].

2.13 Architectures Selection

The six architectures evaluated in this study were chosen to represent the families of modern segmentation networks and to assess how their distinct design principles address the challenges of polyp segmentation under strong domain shift. The classical UNet serves as the foundational encoder–decoder baseline, providing a refer-

ence for subsequent variants. Attention U-Net augments this design with spatial attention gates to improve localization in cluttered endoscopic scenes. SegResNet incorporates residual learning and VAE regularization to enhance feature propagation and generalization on limited data. EffiSegNet-B4 leverages compound scaling and squeeze-and-excitation blocks of EfficientNet to achieve high representational capacity with computational efficiency. UNETR replaces the convolutional encoder with a pure transformer to capture long-range dependencies across the image. Swin-UNETR combines hierarchical Swin Transformers with a convolutional decoder to balance global context modeling and precise boundary delineation. Together, these architectures span convolutional, attention-augmented, residual, transformer, and hybrid paradigms.

3 Materials & Methods

3.1 Dataset

The machine learning models compared in this study were trained using PolypGen, a Multicenter Segmentation dataset [6]. PolypGen represents one of the most comprehensive publicly available polyp detection and segmentation datasets specifically designed for generalizability assessment. It was created by a collaborative team of computational scientists and expert gastroenterologists from six different medical centers in Europe and Africa, incorporating data from more than 300 unique patients. The dataset comprises 8,037 frames in total, including 3,762 positive frames containing polyps and 4,275 negative sequence frames without polyps. Of the positive samples, 1,537 are static polyp images with pixel-level segmentation annotations, while the remainder consists of 2,225 positive video sequences. For this study we use explicitely the 1537 static images, from which, centers C1-C5 are utilized for training and C6 for evaluation.

Several well-known polyp segmentation datasets exist, including Kvasir-SEG [25], CVC-ClinicDB, and ETIS-Larib [51]. While these datasets have been widely adopted in benchmarking, they are limited by single-center origins, relatively small sample sizes, or restrictive imaging conditions. The fact that distinguishes PolypGen from other polyp segmentation datasets is its unique multi-center design. The dataset includes colonoscopy data from six unique centers spanning over different geographical regions, populations, and endoscopic systems, making it particularly suitable for testing model robustness across different clinical settings. This multi-center approach is different from most existing datasets that typically originate from a single center or limited institutions. The dataset includes images from varied endoscopic equipment, imaging conditions and resolutions, polyp sizes and patient populations, providing significant domain shift challenges.

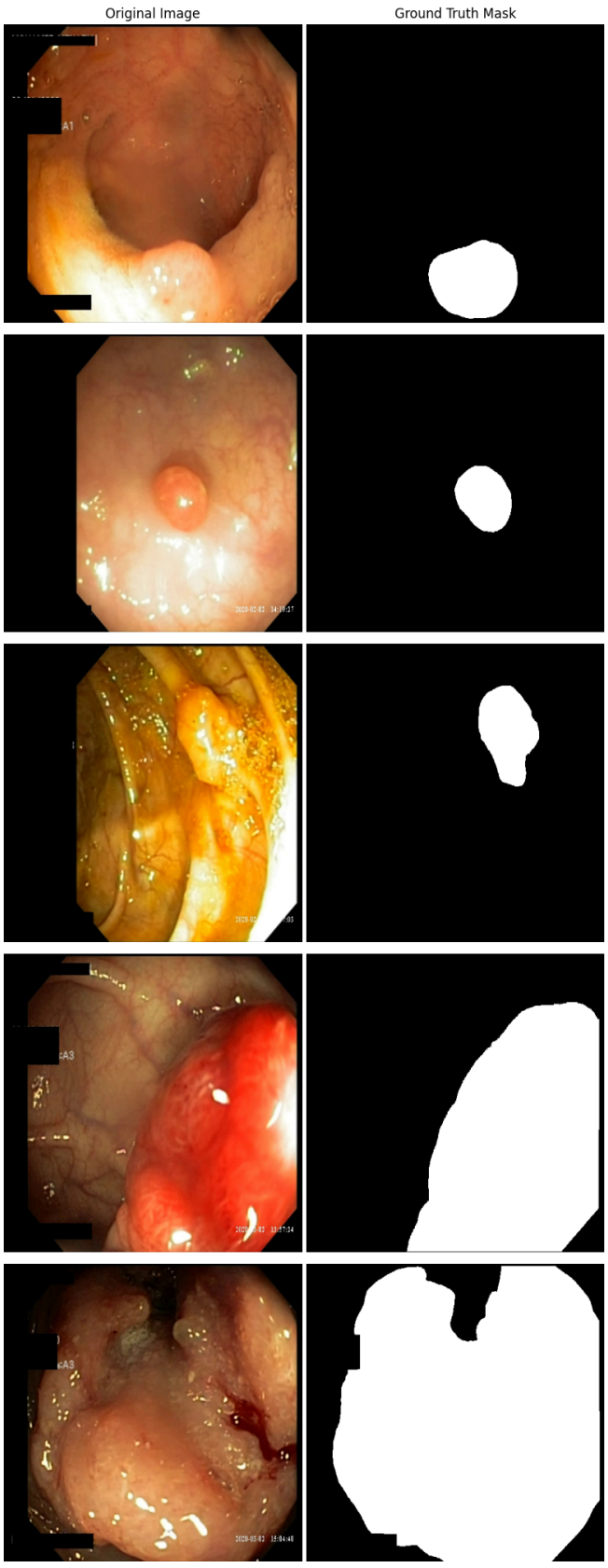


Figure 1: Subset of PolypGen dataset, with varied polyp sizes and shapes

3.2 UNet Architecture

The UNet architecture, introduced by Ronneberger et al. [48], has emerged as the foundational framework for medical image segmentation tasks, demonstrating exceptional performance in scenarios with limited training data through its innovative design that combines contextual understanding with precise spatial localization capabilities. This U-shaped architecture represents a paradigm shift in segmentation approaches, establishing itself as the gold standard for biomedical applications where accurate boundary delineation is paramount for clinical decision-making.

The architecture derives its distinctive from its characteristic U-shaped structure, which is a symmetric encoder-decoder configuration that systematically reduces spatial resolution while capturing increasingly abstract semantic representations, followed by a corresponding expansion phase that reconstructs the original spatial dimensions while preserving fine-grained detail. This design philosophy addresses the fundamental challenge in segmentation tasks: the need to simultaneously capture global contextual information through hierarchical feature extraction while maintaining the spatial precision required for accurate pixel-level classification.

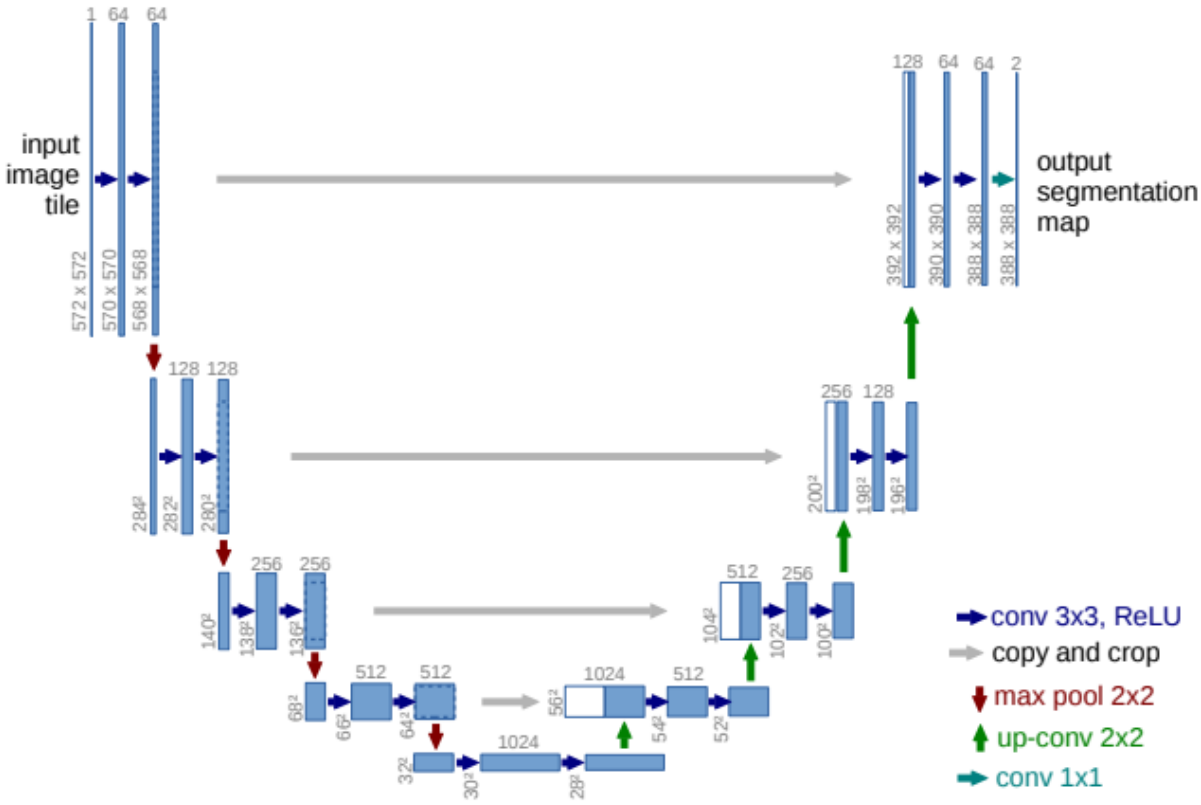


Figure 2: UNet architecture

The encoder pathway, constituting the contracting portion of the U-shape, functions as a feature extraction mechanism that progressively compresses the input representation through a series of carefully orchestrated operations. Each encoder block im-

plements a standardized computational sequence comprising two successive 3×3 convolutional layers, each followed by batch normalization operations that stabilize the learning process by normalizing activation distributions across mini-batches, thereby mitigating internal covariate shift and accelerating convergence. The rectified linear unit (ReLU) activation function is applied after each batch normalization step, introducing essential non-linearities that enable the network to learn complex, non-linear mappings between input features and target outputs. To provide regularization and prevent overfitting, particularly crucial in medical imaging applications where training data may be limited, a 10% dropout rate is applied within each encoder block. The spatial downsampling is achieved through 2×2 max pooling operations with stride 2, which systematically reduce the spatial dimensions by half while preserving the most salient features through the maximum operation.

A fundamental architectural principle of UNet involves the systematic doubling of feature channels at each downsampling step, creating a hierarchical feature representation that progresses from fine-grained textural details to increasingly abstract semantic concepts. This channel progression, typically following the pattern [1024], enables the network to maintain representational capacity despite the reduction in spatial resolution, ensuring that important discriminative features are preserved and enhanced through the encoding process. This design decision reflects the understanding that as spatial resolution decreases, the network must compensate by increasing the depth of feature representation to capture the complex semantic relationships necessary for accurate segmentation.

The bottleneck layer represents the deepest level of the network architecture, operating at the most compressed spatial resolution while maintaining the highest number of feature channels. This critical component serves as the bridge between the encoder and decoder pathways, containing the most abstract and contextually rich feature representations that encapsulate the global understanding of the input image. In typical UNet implementations, the bottleneck operates at 32×32 spatial resolution with 1024 channels, representing the maximum compression point where spatial information is most condensed while semantic understanding is most comprehensive.

The decoder pathway orchestrates the systematic reconstruction of spatial resolution through a series of upsampling operations that progressively restore the original image dimensions while integrating hierarchical features from the encoder through skip connections. Each decoder block initiates with a 2×2 transposed convolution (also termed up-convolution or deconvolution) that doubles the spatial dimensions while halving the number of feature channels. This upsampling operation is followed by feature concatenation with the corresponding encoder level through skip connections,

creating a merged representation that combines high-level semantic information from the decoder pathway with fine-grained spatial details preserved from the encoder. The concatenated features are subsequently processed through two 3×3 convolutional layers with batch normalization and ReLU activation, similar to the encoder blocks, allowing the network to refine and integrate the multi-scale information. Consistent with the encoder design, 10% dropout is applied for regularization throughout the decoder pathway.

Skip connections constitute the defining innovation of the UNet architecture, establishing direct pathways between corresponding encoder and decoder levels that bypass the bottleneck compression. These connections address the fundamental challenge of information loss during the encoding process by providing alternative routes for fine-grained spatial information to reach the decoder without degradation. The concatenation operation at each decoder level creates feature maps that combine the upsampled semantic information from the previous decoder stage with the high-resolution details from the corresponding encoder stage, enabling the network to achieve both global context understanding and precise spatial localization. This mechanism is particularly crucial for medical image segmentation, where accurate boundary delineation often depends on subtle textural and structural details that might otherwise be lost during the compression-decompression process.

The output layer implements a final 1×1 convolution that reduces the multi-channel feature representation to a single-channel probability map, effectively mapping the learned feature vectors to the desired segmentation classes. For binary segmentation tasks, this layer is typically followed by a sigmoid activation function that constrains the output values to the range , providing interpretable probability estimates for each pixel's class membership. The sigmoid activation is particularly well-suited for binary classification scenarios, as it provides a natural threshold at 0.5 for converting continuous probability values to discrete binary predictions.

In the context of our MONAI implementation, the UNet configuration is specifically tailored for medical image segmentation with a 2D architecture that processes 512×512 pixel RGB images through three input channels, generating single-channel probability maps for binary segmentation tasks. The network implements the standard channel progression of [1024], following the canonical UNet design where feature channels systematically double at each encoder level, creating a balanced hierarchy of feature representations from fine-grained textures to abstract semantic concepts. The uniform stride configuration ensures consistent 2× spatial reduction at each pooling operation, maintaining a balanced receptive field expansion across all network levels and enabling effective multi-scale feature learning. Our implemen-

tation employs the basic UNet architecture without residual connections, preserving the original design philosophy while acknowledging that MONAI supports residual variants for applications requiring deeper network architectures.

3.3 Attention UNet Architecture

The Attention U-Net, introduced by Oktay et al. [45], represents a sophisticated enhancement of the standard U-Net architecture through the strategic integration of Attention Gates (AGs) that automatically focus on target structures of varying shapes and sizes in medical imaging applications. This architectural innovation was specifically engineered to eliminate the dependency on external tissue or organ localization modules while simultaneously improving model sensitivity and prediction accuracy with minimal computational overhead, addressing fundamental limitations in conventional segmentation approaches where irrelevant background regions often interfere with precise anatomical structure identification.

The Attention U-Net maintains the foundational encoder-decoder structure of the standard U-Net while introducing intelligent feature filtering mechanisms that enhance the network's ability to discriminate between relevant anatomical structures and irrelevant background information. The encoder pathway follows the conventional U-Net design with progressive downsampling by a factor of 2 at each hierarchical scale, implementing multiple convolutional blocks featuring 3×3×3 convolutions followed by ReLU activation functions. The spatial dimension reduction is achieved through max-pooling layers with 2×2×2 kernels, while feature channel doubling occurs after each pooling operation to maintain representational capacity despite spatial compression. The decoder pathway orchestrates upsampling operations to restore spatial resolution while utilizing skip connections to merge features from corresponding encoder levels, enhanced by convolutional layers for feature refinement. The bottleneck layer serves as the bridge between encoder and decoder pathways, containing the most compressed yet semantically rich feature representations.

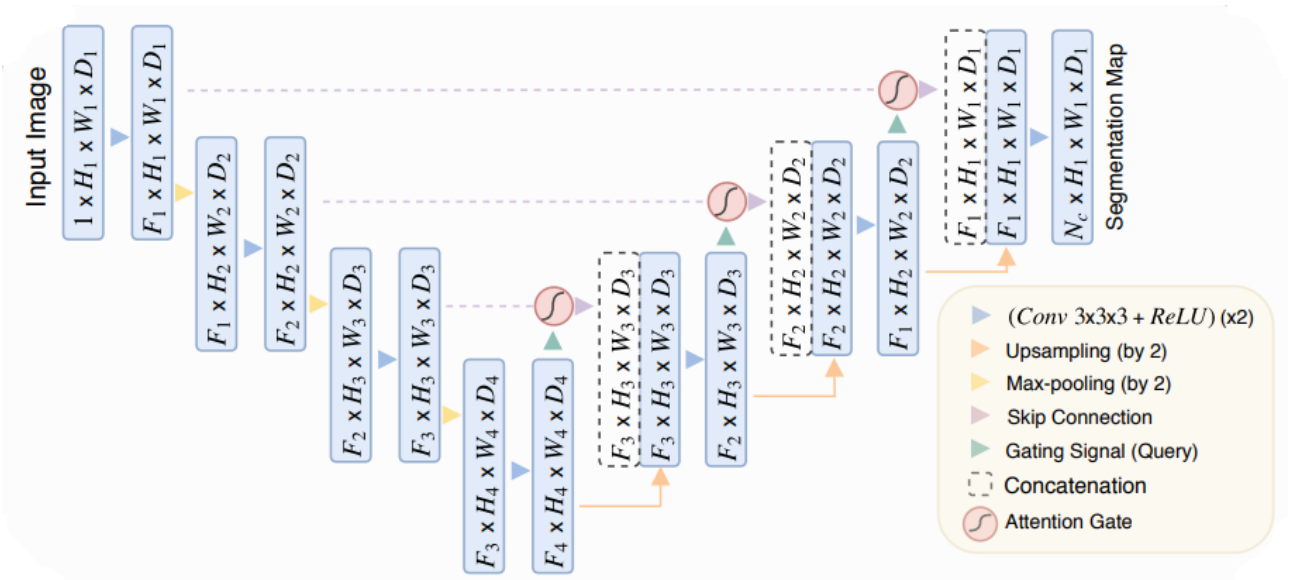


Figure 3: Attention UNet architecture

The defining innovation of Attention U-Net lies in the integration of Attention Gates that intelligently filter features propagated through skip connections, transforming the passive information transfer mechanism of standard U-Net into an active, context-aware feature selection process. These attention mechanisms employ a sophisticated computational framework that dynamically determines the relevance of encoder features for the current decoding task, effectively suppressing irrelevant background regions while amplifying salient anatomical structures without requiring explicit region-of-interest cropping.

The attention mechanism operates through a mathematically elegant formulation involving attention coefficient computation using additive attention principles. The attention gates compute coefficients α_i^l through the equation $q_{att}^l = \psi^T\big(\sigma_1\big(W_x^T x_i^l + W_g^T g_i + b_g\big)\big)$ b_ψ, where x_i^l represents input features from skip connections carrying fine-grained spatial details, while g_i serves as the gating signal from coarser scales providing essential contextual information about what anatomical structures are currently relevant for segmentation. The sigmoid activation function σ_2 is deliberately chosen over softmax for superior convergence properties in medical imaging applications, while Θ_{att} encompasses all learnable parameters including transformation matrices W_x, W_g, projection vector ψ and associated bias terms.

The feature gating process implements the computed attention coefficients through element-wise multiplication expressed as $\hat{x}_i^l = x_i^l \cdot \alpha_i^l$, where this operation scales input features by their corresponding attention coefficients, effectively suppressing irrelevant spatial regions while highlighting salient anatomical features that contribute to accurate segmentation boundaries. This multiplicative gating mechanism ensures that only the most relevant encoder features are propagated to the decoder, dramatically improving the signal-to-noise ratio in the feature maps and enabling more precise localization of target anatomical structures.

The strategic placement of attention gates immediately before concatenation operations in skip connections ensures optimal information filtering at the critical junctions where encoder and decoder pathways merge. This positioning guarantees that only relevant activations are combined between encoder and decoder paths, while irrelevant background regions are systematically suppressed without requiring explicit region-of-interest preprocessing. During backpropagation, the attention mechanism down-weights gradients from background regions, naturally focusing the learning process on anatomically relevant areas and improving convergence stability.

Unlike global attention mechanisms that apply uniform attention across entire feature maps, Attention U-Net employs sophisticated grid-based gating where attention coef-

ficients are computed for specific local spatial regions, enabling fine-grained control over feature selection at the pixel level. The gating signals aggregate information from multiple imaging scales, creating a hierarchical understanding of anatomical context that guides attention allocation. Each attention gate can focus on different subsets of target structures simultaneously, allowing the network to handle complex multi-organ segmentation tasks where different anatomical regions require distinct attention patterns.

The computational efficiency of Attention U-Net represents a remarkable achievement in architectural design, adding only 8% additional parameters compared to standard U-Net while delivering substantial performance improvements in medical image segmentation tasks. This minimal parameter overhead is achieved through the efficient design of attention gates that utilize lightweight 1×1 convolutions for feature transformation and sigmoid activations for coefficient generation, avoiding computationally expensive operations while maintaining attention quality.

The architecture successfully combines the proven effectiveness of U-Net's encoder-decoder framework with selective attention mechanisms, creating a powerful computational framework for medical image segmentation that automatically learns to focus on relevant anatomical structures without requiring external localization modules or manual region-of-interest annotation. This capability is particularly valuable in clinical applications where anatomical structures exhibit significant inter-patient variability in size, shape, and spatial location, requiring robust attention mechanisms that can adapt to diverse pathological presentations.

In our MONAI implementation, the Attention U-Net configuration employs a 2D architecture optimized for medical image processing with 3 input channels to accommodate standard RGB or multi-modal imaging data, generating a single output class suitable for binary segmentation tasks commonly encountered in medical applications. The network implements a five-level encoder-decoder hierarchy with channel depths progressing to 1024 at the bottleneck level, utilizing four stride-2 downsampling steps that create a balanced multi-scale representation while maintaining computational efficiency. This configuration strikes an optimal balance between representational capacity and computational requirements, making it particularly suitable for clinical deployment where both accuracy and processing speed are critical considerations.

3.4 SegResNet Architecture

SegResNet, introduced by Myronenko [42], represents a sophisticated 3D encoder-decoder convolutional network architecture that has been augmented with a variational autoencoder (VAE) branch specifically designed for regularization purposes in medical image segmentation applications. This innovative architecture employs an asymmetrically large ResNet-based encoder that extracts comprehensive volumetric features from multimodal MRI data, while utilizing a compact decoder pathway to reconstruct precise subregion segmentation masks. The fundamental design philosophy underlying SegResNet involves the joint regularization of the shared encoder through the VAE branch during training, which significantly improves generalization performance on limited datasets, a common challenge in medical imaging applications where annotated data is scarce.

The architectural framework of SegResNet comprises two parallel computational pathways that share a common encoder backbone, creating a dual-purpose system that simultaneously performs segmentation and reconstruction tasks. The primary segmentation branch implements a traditional encoder-decoder network structure that generates multi-channel segmentation maps through sigmoid activation functions, enabling the precise delineation of anatomical structures such as whole tumor regions, tumor cores, and enhancing tumor components in brain imaging applications. The auxiliary variational autoencoder branch operates exclusively during the training phase, reconstructing the original 4-channel MRI input from the encoder's deepest feature representations, thereby enforcing additional constraints on the encoder's latent representations and promoting more generalizable feature learning. During inference, only the segmentation branch is utilized, ensuring computational efficiency while maintaining the benefits of the regularization learned during training.

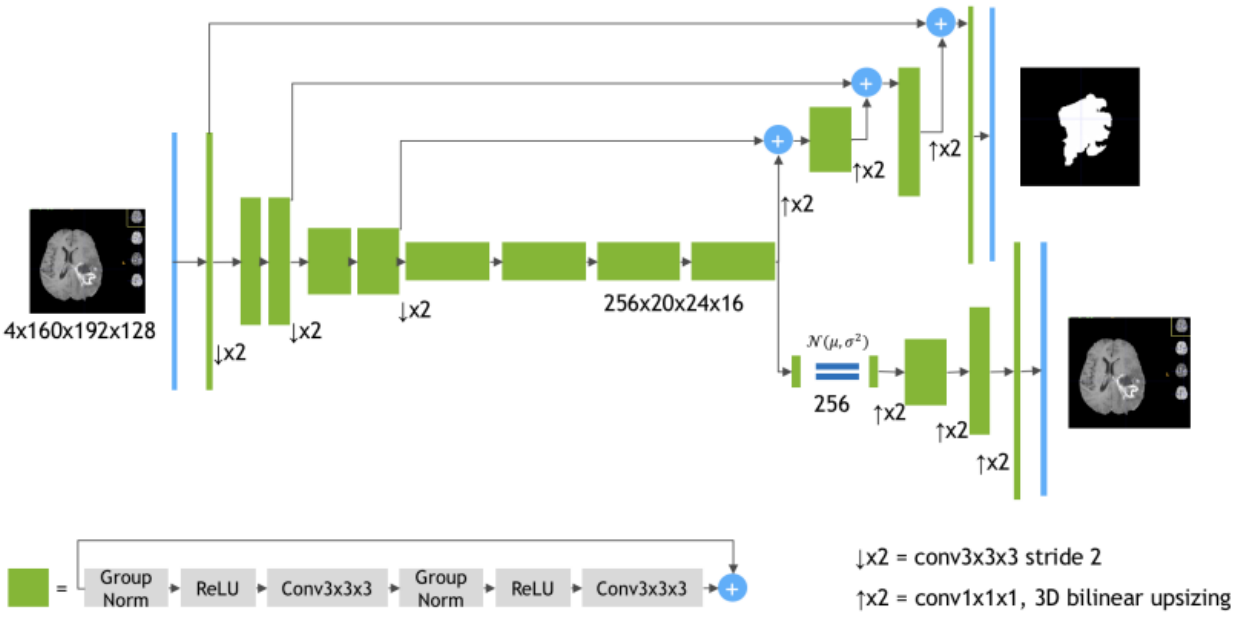


Figure 4: Schematic visualization of SegResNet architecture

The encoder component of SegResNet implements a ResNet-like feature extraction architecture that systematically processes four-channel 3D MRI data through a hierarchical sequence of convolutional operations. The initial processing begins with a 3×3×3 convolutional layer utilizing 32 filters, followed by Group Normalization and ReLU activation functions, establishing the foundation for subsequent feature extraction. Group Normalization is specifically chosen over Batch Normalization due to its superior performance characteristics when dealing with small batch sizes, which is common in 3D medical imaging applications where memory constraints limit the number of samples that can be processed simultaneously.

The encoder incorporates four hierarchical levels of residual downsampling blocks that progressively reduce spatial dimensions while increasing feature depth. Each residual block contains two sequences of Group Normalization, ReLU activation, and 3×3×3 convolution operations, connected by identity skip connections that facilitate gradient flow and enable the training of deeper networks without degradation. The downsampling between encoder levels is achieved through 3×3×3 convolutions with stride 2 applied at the first convolution layer of each block, systematically reducing spatial dimensions by half at each level. The feature channel progression follows a systematic doubling pattern of 32, 64, 128, 256, and 512 channels, ensuring that as spatial resolution decreases, the representational capacity increases to maintain discriminative power throughout the hierarchical feature extraction process.

The segmentation decoder pathway reconstructs the original spatial resolution through four levels of upsampling blocks that reverse the encoding process while integrating

multi-scale features. Each upsampling block initiates with a 1×1×1 convolution to reduce channel dimensionality, followed by 3D linear upsampling that doubles the spatial dimensions in each direction. The upsampled features are subsequently processed through residual blocks that mirror the encoder's structure, implementing Group Normalization, ReLU activation, and convolutional operations to refine the reconstructed features. Skip connections from corresponding encoder levels are additively combined with decoder features at equivalent spatial scales, providing direct pathways for fine-grained spatial information to bypass the bottleneck compression and contribute to precise boundary delineation. The final output layer employs a 1×1×1 convolution to produce three channels corresponding to different tumor subregions, followed by sigmoid activations that generate probability maps for each anatomical structure of interest.

The VAE regularization branch implements a sophisticated probabilistic framework that constrains the encoder's latent space representation during training. The latent projection component begins with a downsampling convolution applied to the encoder's deepest feature map, producing 256 compressed features that capture the most abstract representations of the input volume. A subsequent dense layer generates two 128-dimensional vectors representing the mean and log-variance parameters of a Gaussian latent distribution, following the standard VAE formulation that enables probabilistic sampling in the latent space. The latent sample is drawn from the normal distribution $z \sim \mathcal{N}(\mu, \sigma^2)$, where the reparameterization trick ensures differentiability for backpropagation.

The VAE decoder pathway mirrors the segmentation decoder architecture while operating independently of the skip connections, reconstructing the original 4-channel input resolution through a series of upsampling operations. A dense layer initially expands the sampled latent vector back to 256×20×24×16 dimensions, followed by four upsampling blocks that each apply 1×1×1 convolutions, spatial upsampling, and residual processing to progressively reconstruct the original input dimensions. This reconstruction process forces the encoder to learn meaningful latent representations that preserve essential anatomical information necessary for both segmentation and reconstruction tasks.

The loss function of SegResNet combines multiple objectives to balance segmentation accuracy with regularization constraints. The primary Dice loss focuses on maximizing overlap between predicted and ground truth segmentation masks, while the reconstruction loss (typically L2/MSE) ensures that the VAE branch can accurately reconstruct the input images from the learned latent representations. The Kullback-Leibler (KL) divergence term enforces the latent distribution to approximate a stan-

dard normal distribution, promoting regularity and preventing overfitting in the latent space. The combined loss function is expressed as $L = L_{\text{Dice}} + 0.1 \cdot (L_{\text{Rec}} + L_{\text{KL}})$, where the 0.1 weighting factor balances the regularization terms against the primary segmentation objective.

This architectural design successfully utilizes a deep, high-capacity ResNet-style encoder that benefits from VAE regularization, yielding state-of-the-art performance on challenging 3D brain tumor segmentation tasks while maintaining computational efficiency. The regularization effect of the VAE branch is particularly valuable in medical imaging scenarios where limited training data can lead to overfitting, as the reconstruction constraint forces the encoder to learn more generalizable feature representations that capture essential anatomical characteristics.

For our specific 2D polyp segmentation application, the SegResNet architecture was adapted with carefully configured network-specific hyperparameters to accommodate the different dimensionality and task requirements. The spatial dimensions were set to 2 for 2D image processing, while maintaining 32 initial filters to determine the starting number of feature channels in the encoder pathway. The input channels were configured to 3 to accommodate standard RGB images commonly used in endoscopic polyp imaging, with the output classes set to 1 for binary polyp segmentation tasks. Batch normalization was employed as the normalization strategy to stabilize training dynamics in the 2D implementation, while a dropout probability of 0.1 was applied throughout the network for regularization to prevent overfitting on the polyp segmentation dataset. These adaptations ensure that the proven effectiveness of SegResNet's encoder-decoder architecture with VAE regularization is successfully transferred to the 2D polyp segmentation domain while maintaining computational efficiency and segmentation accuracy.

3.5 EffiSegNet-B4 Architecture

EffiSegNet-B4 builds upon the encoder–decoder paradigm by integrating a pre-trained EfficientNet-B4 backbone [55] with a streamlined decoder, achieving state-of-the-art polyp segmentation while maintaining remarkable computational efficiency [57]. The EfficientNet-B4 encoder, originally optimized for ImageNet classification at a native input resolution of 380×380 pixels, employs Mobile Inverted Bottleneck Convolution (MBConv) blocks enhanced by squeeze-and-excitation modules. These MBConv blocks efficiently balance network depth, width, and resolution through compound scaling, resulting in 17.5 million parameters dedicated to rich, hierarchical feature extraction across eight main stages. Each stage progressively increases channel dimensionality while reducing spatial resolution, producing multi-scale feature maps that capture textural, structural, and contextual information essential for precise polyp delineation.

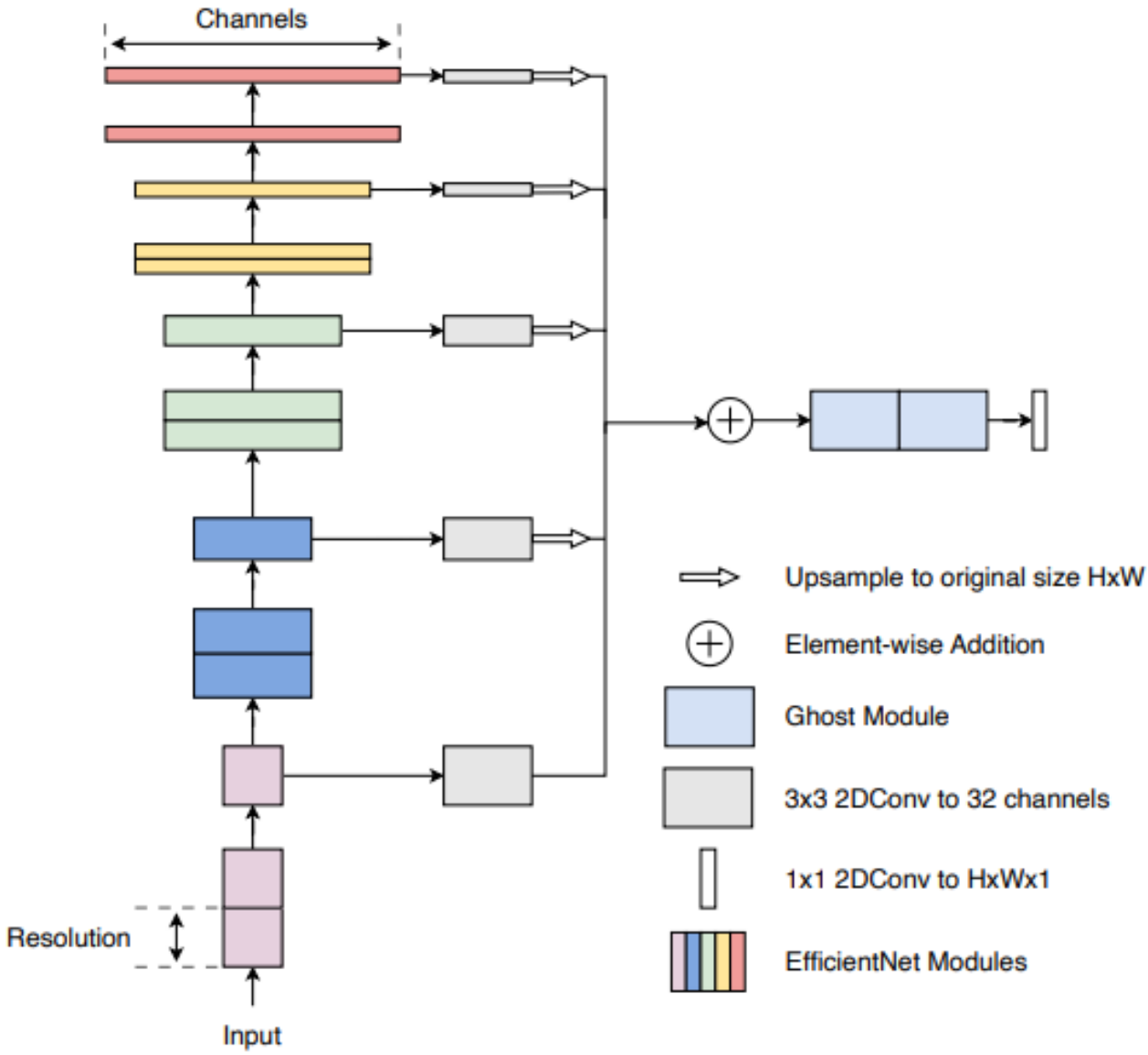


Figure 5: EffiSegNet architecture

The EffiSegNet decoder eschews traditional U-Net concatenation in favor of element-wise addition, merging upsampled feature maps from each of the five selected encoder stages with the stem-layer features. Prior to fusion, each feature map is passed through a 1×1 convolutional layer that reduces its channel count to 32, followed by batch normalization to stabilize activations. Nearest-neighbor upsampling then resizes these feature maps to the original input dimensions, ensuring spatial alignment for additive fusion. This approach not only simplifies the decoder architecture—adding a mere 0.21 million parameters—but also preserves salient multi-scale information without incurring the overhead of concatenation and additional convolutions.

To further enhance efficiency, the decoder incorporates Ghost modules, which generate supplementary feature maps through cheap linear operations rather than conventional convolutions. These modules replicate intrinsic redundancy in feature repre-

sentations, extending channel capacity with minimal parameter increment. The fused feature map, now enriched by both the EfficientNet backbone and Ghost-augmented decoding, is passed through a final 1×1 convolution to produce a single-channel probability map. A sigmoid activation transforms raw logits into pixel-wise probabilities for binary polyp versus background segmentation.

In our implementation, input images are resized to 512×512 pixels to accommodate endoscopic imaging resolutions, and the decoder fusion layers consistently operate at a channel dimension of 32. The network is trained from scratch using a batch size of 16, with all encoder weights unfrozen to allow joint fine-tuning of both backbone and decoder. This configuration leverages EfficientNet-B4's pre-training advantages while adapting its feature extraction capabilities to the specific characteristics of polyp imagery, yielding a high-precision, low-complexity segmentation model suitable for real-time clinical workflows.

3.6 UNETR Architecture

The UNETR architecture, introduced by Hatamizadeh et al. [22], reimagines three-dimensional medical image segmentation by leveraging a pure Vision Transformer (ViT) as the encoder within the canonical U-shaped paradigm, thereby unifying global context modeling and localized feature reconstruction in a single end-to-end framework. In this design, the input volumetric image is first partitioned into uniform, non-overlapping 3D patches of size $p \times p \times p$. Each patch is then flattened into a vector of dimension $p^3 \times C$, where C denotes the number of input channels, and linearly projected to a D-dimensional embedding space via a learnable weight matrix $E \in R^{(p^3C)\times D}$. Positional embeddings are added to these patch embeddings to encode spatial locality, resulting in a sequence of $N = \frac{H}{p} \cdot \frac{W}{p} \cdot \frac{L}{p}$ tokens, which forms the input to a standard transformer encoder comprising L layers of multi-head self-attention and feed-forward networks.

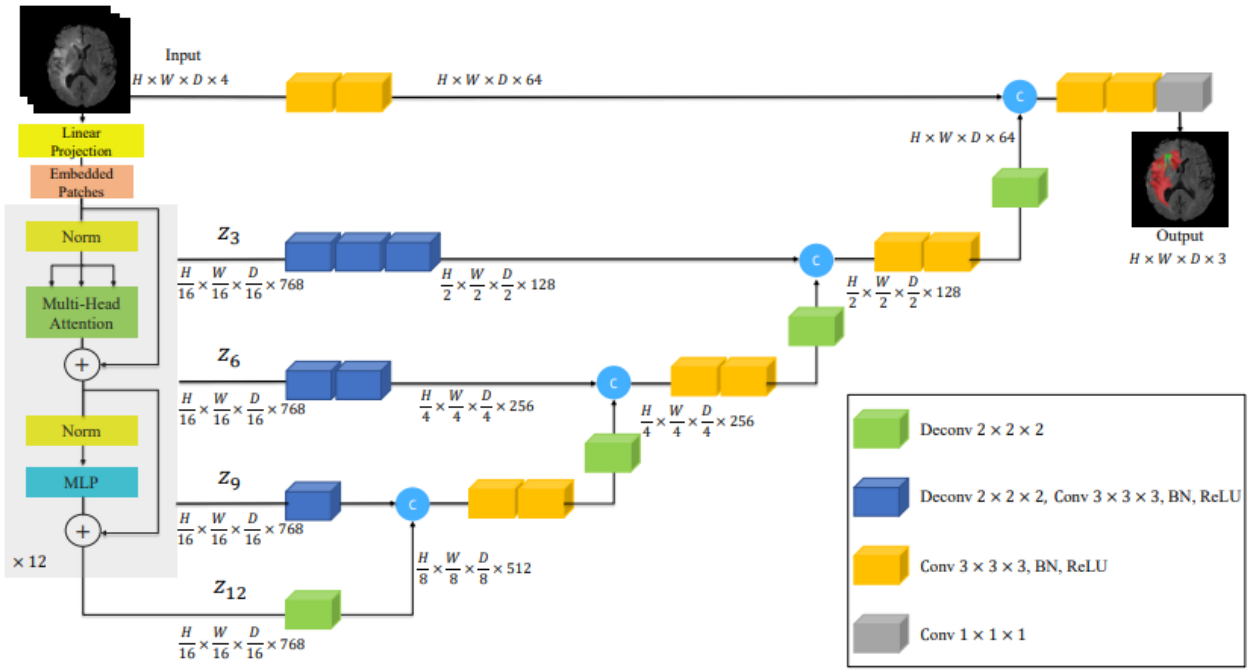


Figure 6: UNETR architecture

Through its self-attention mechanism, the transformer encoder captures long-range dependencies across the entire volumetric field, enabling holistic contextual reasoning that transcends the limited receptive fields of convolutional operations. Importantly, UNETR extracts intermediate representations from select transformer layers corresponding to multiple resolution levels. Specifically, transformer outputs at depths $\{l_1, l_3, l_3, l_4\}$ are reshaped back into volumetric feature maps of spatial dimensions $\{\frac{H}{2}, \frac{W}{2}, \frac{L}{2}\}$, $\{\frac{H}{4}, \frac{W}{4}, \frac{L}{4}\}$, $\{\frac{H}{8}, \frac{W}{8}, \frac{L}{8}\}$, $\{\frac{H}{16}, \frac{W}{16}, \frac{L}{16}\}$, respectively, thus furnishing multi-scale feature hierarchies that bridge global semantic information with local anatomical detail.

These multi-scale transformer features are transmitted to a convolutional decoder via skip connections akin to those in classical U-Net architectures. The decoder performs successive upsampling by factors of two through transposed convolutions of kernel size $2 \times 2 \times 2$, interleaved with $3 \times 3 \times 3$ convolutions, batch normalization, and ReLU activations. At each scale, the upsampled feature map is concatenated with the corresponding transformer-derived feature map to integrate global context with fine grain spatial cues. This fusion restores the original resolution $H \times W \times L$ and refines voxel-wise predictions through final $1 \times 1 \times 1$ convolutions that map to the desired number of segmentation classes, followed by sigmoid or softmax activations for probability estimation.

By supplanting the conventional convolutional encoder with a transformer, UNETR excels in capturing inter-slice and intra-slice contextual relationships, significantly improving segmentation accuracy for structures exhibiting complex shapes or dispersed spatial distributions [22]. Empirical evaluations on the BTCV and MSD benchmarks demonstrate that UNETR achieves superior delineation of small anatomical regions, reduces boundary ambiguity, and exhibits robust generalization across modalities and organ systems.

In our MONAI adaptation for two-dimensional polyp segmentation, the UNETR encoder is configured to process 16×16 patches extracted from 512×512 images, projected into a $D = 16$-dimensional embedding space and processed through $L = 12$ transformer layers. The decoder mirrors the original 3D design with 2D transposed convolutions and skip connections at four scales. A dropout rate of 0.1 is applied within both encoder and decoder layers to mitigate overfitting, and batch normalization ensures stable feature distributions during training. This tailored configuration harnesses UNETR's global attention strengths while accommodating the computational constraints and spatial characteristics of 2D endoscopic imagery.

3.7 SwinUNetR Architecture

The SwinUNetR architecture, introduced by Hatamizadeh et al. [21], extends the U-shaped segmentation paradigm to three-dimensional medical imaging by integrating a hierarchical Swin Transformer encoder with a convolutional decoder in a seamless end-to-end framework. In this design, multi-modal MRI volumes are first partitioned into non-overlapping 3D patches and projected into a one-dimensional embedding space via a linear layer. Positional encodings are added to these patch embeddings to retain spatial context, and the resulting sequence is processed through a series of Swin Transformer blocks organized into four stages, each comprising two transformer layers. The Swin Transformer's shifted window self-attention mechanism enables efficient capture of long-range dependencies within local windows while allowing cross-window interactions at each stage, producing rich hierarchical feature representations at five spatial resolutions corresponding to the original volume and its successive patch-merged reductions.

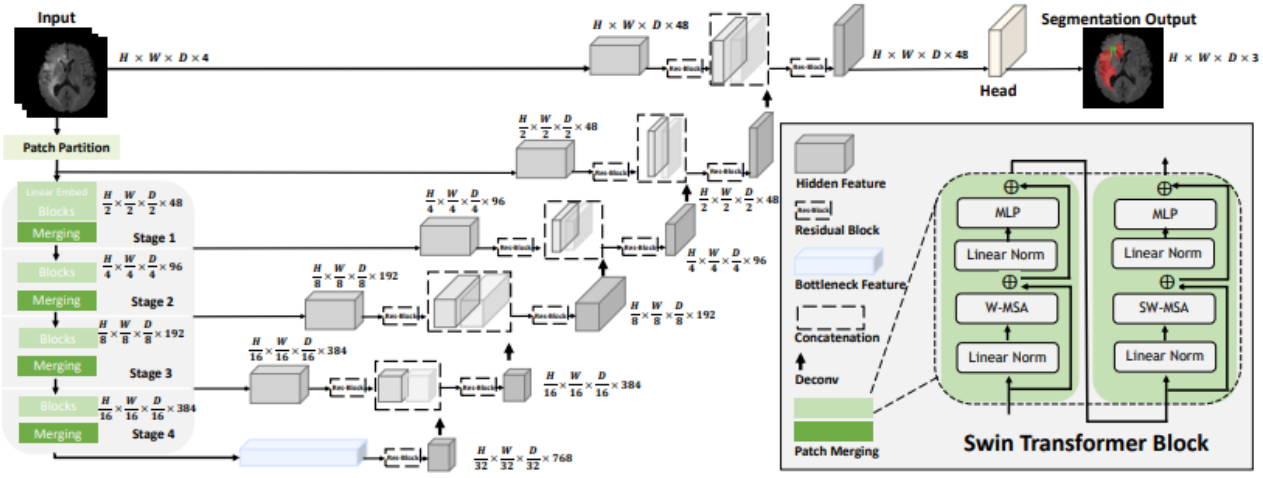


Figure 7: SwinUNETR architecture

At the conclusion of each transformer stage, the encoder outputs volumetric feature maps that are reshaped to their respective spatial dimensions. These multi-scale transformer features are then transmitted to a convolutional decoder through skip connections that mirror the U-shaped topology. The decoder employs residual convolutional blocks composed of two consecutive $3 \times 3 \times 3$ convolutions followed by instance normalization and ReLU activations, refining the fused feature maps at each resolution. Deconvolutional layers perform upsampling operations that progressively restore the original spatial dimensions, integrating both global context from the transformer encoder and local spatial detail captured by the convolutional blocks.

The final segmentation mask is produced by a $1 \times 1 \times 1$ convolution that maps the

decoder's feature representation to the desired number of output channels, followed by a sigmoid activation to generate voxel-wise probability estimates for binary segmentation. This architecture capitalizes on the Swin Transformer's ability to model long-range interactions efficiently, while the convolutional decoder ensures precise recovery of fine anatomical boundaries, resulting in improved delineation of complex tumor regions in multi-modal brain MRI.

In our MONAI implementation for two-dimensional polyp segmentation, the SwinUNetR model is adapted by employing a 2D Swin Transformer encoder configured with an initial feature dimension of 48 and four stages of shifted window self-attention. Input images of size 512×512 pixels with three channels are divided into 4×4 patches, embedded, and processed through the transformer encoder. The convolutional decoder mirrors the 3D design with 2D residual blocks and transposed convolutions to reconstruct the segmentation map at full resolution. Batch normalization stabilizes feature distributions across both encoder and decoder layers, while stochastic dropout with a rate of 0.1 is applied to prevent overfitting. To accommodate the high memory demands of the transformer encoder, gradient checkpointing is enabled, ensuring efficient training without sacrificing model capacity or segmentation accuracy.

3.8 Ensemble Architecture

In this study, we developed an ensemble framework to enhance the performance of binary semantic segmentation for polyp detection. The ensemble combines the 3 convolutional neural network architectures with the best performances in medical image segmentation, taken from our training: an EffiSegNet-B4 based model, an Attention U-Net, and a SegResNet. Each model is independently trained and configured through Hydra, enabling flexible and modular management of model specifications and training hyperparameters. The corresponding pretrained weights are loaded from checkpoint files to initialize each network.

For each sample in the evaluation set, the predicted output probabilities are extracted from all models in the ensemble. These probability maps are concatenated across batches and subsequently stacked into a multidimensional tensor of shape (number of models, total samples, channels, height, width). The ensemble prediction is obtained by computing the element-wise average of the probability maps across the models, simulating a soft voting mechanism. This probabilistic averaging smooths out individual model uncertainties and biases, improving overall segmentation quality.

The ensemble approach achieves improvements over individual models in key segmentation metrics, highlighting the effectiveness of soft probabilistic voting.

4 Experimental Setup

4.1 Hardware and Software Infrastructure

All experiments were conducted in the Google Colaboratory virtual environment leveraging an NVIDIA A100 GPU with 40 GB of VRAM and 64G GB RAM. The deep learning workflows were implemented in Python, utilizing PyTorch 1.13.0 as the primary framework for model development. Training, validation, and testing loops were managed using PyTorch Lightning 2.0.0. Medical imaging-specific components and metrics were implemented through MONAI 1.1.0, while configuration management was facilitated by Hydra 1.3.1. The image augmentation pipeline was constructed using Albumentations 1.3.0. Real-time visualization and experiment tracking were performed with TensorBoard.

4.2 Dataset Configuration

We trained the six network architectures on PolypGen [6], an open-access segmentation dataset. The set used for train and validation contains 1449 static frame endoscopic images of gastrointestinal polyps and their corresponding ground truth delineations from centers C1-C5 in a 80:20 split, ensuring equal representation from all centers in the train and validation sets. After completing training, the finalized model was evaluated on the held-out C6 center, including 88 data points. This strict separation simulates deployment in an unseen clinical environment, providing an unbiased assessment of each model's capacity to generalize across different endoscopy systems, patient populations, and image acquisition protocols. Metrics reported for C6 reflect true external validity, demonstrating the robustness of the segmentation pipeline under domain shift conditions.

4.3 Data Augmentations

All input images were standardized to 512×512 pixels using Lanczos interpolation, regardless of their original dimensions, which varied significantly across centers. We followed the augmentation techniques used in [15], with the addition of elastic deformation. More specifically, during training we applied:

- Random horizontal and vertical flip.
- Color jitter with the brightness chosen randomly between 0.6 and 1.6, a contrast factor of 0.2, saturation factor 0.1 and hue factor 0.01.
- Affine transformation with scale value uniformly sampled between 0.5 and 1.5, translation up to 12.5% of the image height and width, and rotation between -90 and 90 degrees.

- Elastic deformation with the Gaussian filter sigma set to 50, alpha value of 1, and Lanczos interpolation.
- Finally, all images were normalized using the calculated channel mean and standard deviation of the PolypGen dataset: [0.5543, 0.3644, 0.2777] and [0.2840, 0.2101, 0.1770] for each of RGB channels, respectively.

4.4 Model Training Protocol

We evaluated the six segmentation architectures previously described, on PolypGen dataset. All models were trained from scratch using identical hyperparameters to ensure fair comparison. We trained every architecture with batch size of 16 for 300 epochs. The training process employed the AdamW optimizer. AdamW was chosen for its better performance in deep learning tasks due to its decoupled weight decay regularization mechanism, which prevents weight decay from interfering with gradient-based updates. Furthermore, a cosine annealing learning rate scheduler was used, with initial learning rate of 1×10^{-4}, gradually decreasing the learning rate to a minimum of 1×10^{-5} to achieve convergence. Dropout rate of 0.1 was used across all architectures, to enforce the network's generalization ability to the out of sample evaluation set. A hybrid loss function combining Dice loss and Cross-Entropy loss was used. This hybrid approach utilizes the complementary strengths of both loss components to address the specific challenges in medical image segmentation tasks. The DiceCELoss computes a weighted combination of Dice loss and Cross-Entropy loss according to the following formulation:

$$L_{total} = \lambda_{dice} \cdot L_{Dice} + \lambda_{ce} \cdot L_{CE}$$

where λ_{dice} and λ_{ce} are weighting coefficients that control the relative contribution of each loss component. In our implementation, both coefficients were set to their default values of 1.0, providing equal weighting between the two loss terms.
The Dice loss component optimizes directly for segmentation overlap quality, which is particularly crucial for medical imaging applications where precise boundary delineation is important. The Dice coefficient measures the similarity between predicted and ground truth segmentations. The dice part of the loss is derived from:

$$\text{Dice} = \frac{2|X \cap Y|}{|X| + |Y|}$$

The Cross-Entropy component ensures accurate pixel-wise classification, particularly important for boundary precision. While Dice loss is inherently robust to class imbalance (common in medical segmentation where pathological regions are typically small). The Cross-Entropy component provides consistent gradients even when Dice gradients become small, particularly in early training phases or for small objects. The Cross Entropy part of the loss is derived from:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[g_i \log(\sigma(p_i)) + (1 - g_i)\log(1 - \sigma(p_i))]$$

where σ denotes the sigmoid function, p_i represents the raw logits before activation, and g_i denotes the ground truth for pixel i.

4.5 Evaluation Metrics

Model performance was assessed using multiple evaluation metrics, including Mean Dice, Mean IoU, precision, recall, F1 and F2 scores.

The Dice coefficient, also known as the Sørensen-Dice coefficient, quantifies the overlap between predicted segmentation masks and ground truth annotations, making it very suitable for medical image segmentation tasks where region overlap is very critical. The Dice coefficient is defined as:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|}$$

where P represents the predicted segmentation mask and G denotes the ground truth mask. The coefficient ranges from 0 (no overlap) to 1 (perfect overlap). $|P \cap G|$ represents the number of common elements in the intersection of sets X and Y, while |X| and |Y| represent the number of elements in sets X and Y, respectively. The Dice coefficient is particularly valuable in medical image segmentation for several reasons:

- Overlap Focus: It directly measures the spatial overlap between predicted and actual polyp regions, which correlates with clinical utility.
- Size Invariance: The metric is relatively insensitive to absolute polyp size, making it suitable for evaluating performance across polyps of varying dimensions.
- Boundary Sensitivity: While primarily measuring overlap, the Dice coefficient is sensitive to boundary accuracy, as boundary errors directly affect the intersection and union calculations.

In our study, the Mean Dice Coefficient was used, computed as the arithmetic mean of individual Dice scores across all images in the evaluation set. The Dice coefficient for each image was calculated as the overlap between predicted and ground truth segmentation masks, with the final metric representing the average performance across the entire dataset.

The Mean Intersection over Union (mIoU), also known as the Jaccard Index provides an alternative perspective on segmentation overlap quality that is particularly sensitive to prediction accuracy and boundary precision. This metric is more sensitive to

prediction errors than Dice coefficient, as it normalizes by the union rather than the sum of areas. The Intersection over Union for a single image is defined as:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} = \frac{|P \cap G|}{|P| + |G| - |P \cap G|}$$

where P represents the predicted segmentation mask, G denotes the ground truth mask, $|P \cap G|$ is the intersection (true positive pixels), and $|P \cup G|$ is the union of both sets.
In our study, the Mean IoU was used, by averaging all IoU scores across all images in the evaluation set, providing a single representative metric value per epoch. The Mean IoU is then computed as:

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \text{IoU}_i$$

where N is the total number of images in the evaluation set, and IoU_i is the IoU score for the i-th image.

Moreover, the following metrics were computed for evaluating the test set:

- Accuracy: Overall pixel-wise classification correctness. Accuracy formula:

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

- Precision: Positive predictive value, measuring the proportion of predicted polyp pixels that are actually polyps. Precision formula:

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

- Recall/Sensitivity: True positive rate, measuring the proportion of actual polyp pixels that are correctly identified. Recall formula:

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

- F_1 Score: Harmonic mean of precision and recall, providing balanced assessment of both metrics. This metric is calculated by formula:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

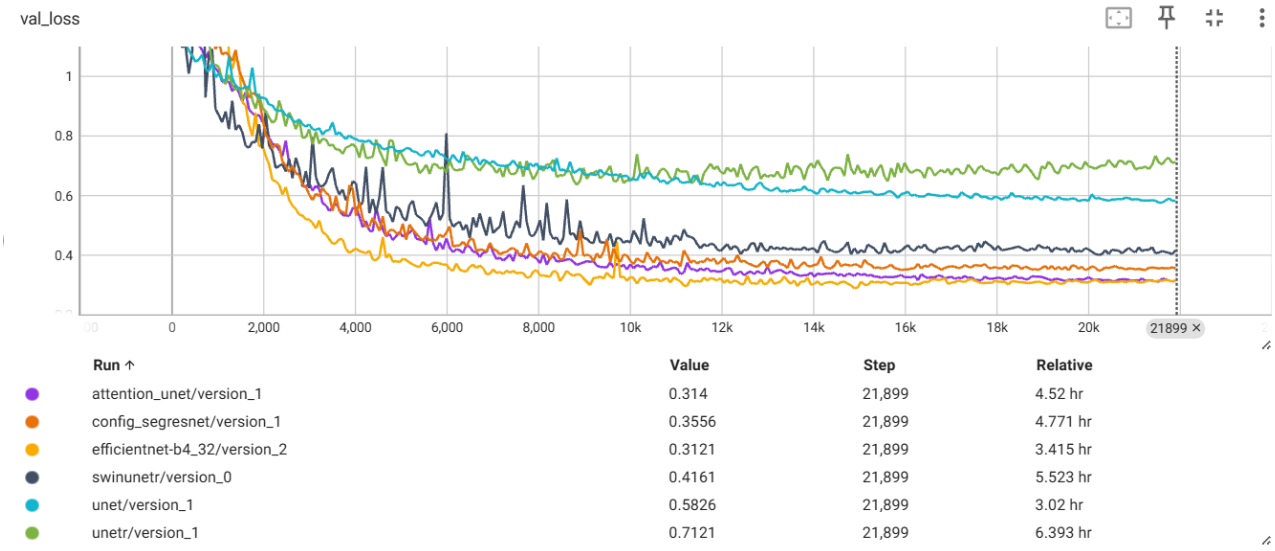
- F_2 Score: Weighted F-score emphasizing recall over precision, which is particularly relevant in medical applications where missing pathological regions (false

negatives) is generally more concerning than false positives. This metric is calculated by formula:

$$F_2 = 5 \cdot \frac{\text{Precision} \cdot \text{Recall}}{4 \cdot \text{Precision} + \text{Recall}} = \frac{5 \cdot TP}{5 \cdot TP + 4 \cdot FN + FP}$$

5 Results

We trained PolypGen dataset from scratch for 300 epochs, and we logged the validation metrics, based on the unseen validation subset per epoch. We also log the metrics that are based on the evaluation set, the C6 center.
The validation loss trajectories for the six architectures: Attention U-Net, SegResNet, EffiSegNet-B4-32, SwinUNETR, UNet, and UNETR are shown in Figure 8. All models demonstrate an initial rapid decline in validation loss, corresponding to the learning of fundamental low-level features during the early training epochs.
EffiSegNet-B4-32 achieves the lowest minimum validation loss, converging and maintaining a stable plateau throughout the training process. This finding highlights its strong generalization capacity and effective adaptation to the domain-specific characteristics of the PolypGen dataset. Attention U-Net closely follows, also exhibiting highly stable loss patterns with minimal oscillation.
SegResNet achieves an intermediate minimum validation loss, maintaining generally stable curve with minor oscillations. SwinUNETR and UNETR validation loss curves are characterized by greater variance and noticeable fluctuations, especially in later training epochs. SwinUNETR shows improved losses relative to the standard UNet, while the UNETR is performing worse than UNet baseline.
The standard UNet, serving as the classical encoder-decoder baseline, records the second highest and most fluctuating validation losses, reflecting limitations in extracting complex and contextually relevant features for binary polyp segmentation compared to contemporary architectures.

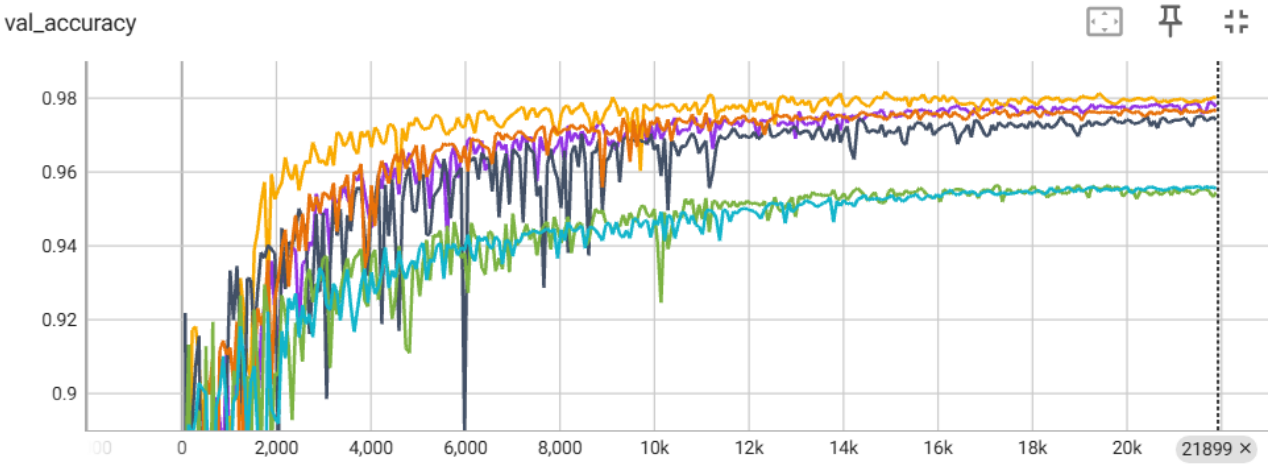


Run	Value	Step	Relative
attention_unet/version_1	0.314	21,899	4.52 hr
config_segresnet/version_1	0.3556	21,899	4.771 hr
efficientnet-b4_32/version_2	0.3121	21,899	3.415 hr
swinunetr/version_0	0.4161	21,899	5.523 hr
unet/version_1	0.5826	21,899	3.02 hr
unetr/version_1	0.7121	21,899	6.393 hr

Figure 8: Validation Loss curves

The validation accuracy curves presented in the figure 9 reveal trends in the generalization performance of the evaluated architectures over the 300 epochs. EffiSegNet-B4-32 achieves the highest validation accuracy, demonstrating both rapid conver-

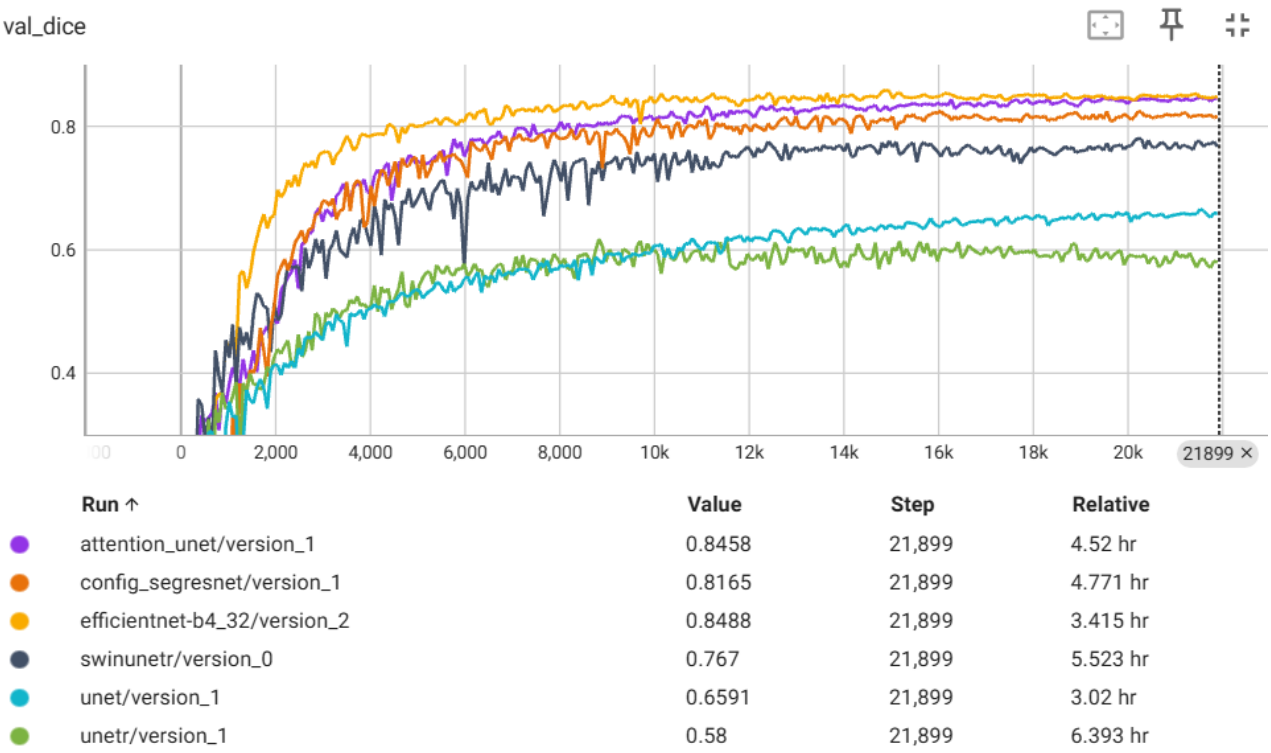
gence and better stability in performance across epochs. Attention U-Net, Config SegResNet, and SwinUNETR also achieve high accuracies, exhibiting consistent upward movement with only minor oscillations after initial convergence. In contrast, both UNet and UNETR show lower accuracy plateaus, indicating comparatively reduced generalization capacity on the validation set.



Run ↑	Value	Step	Relative
attention_unet/version_1	0.9776	21,899	4.52 hr
config_segresnet/version_1	0.9766	21,899	4.771 hr
efficientnet-b4_32/version_2	0.9802	21,899	3.415 hr
swinunetr/version_0	0.9741	21,899	5.523 hr
unet/version_1	0.9555	21,899	3.02 hr
unetr/version_1	0.9544	21,899	6.393 hr

Figure 9: Validation Accuracy curves

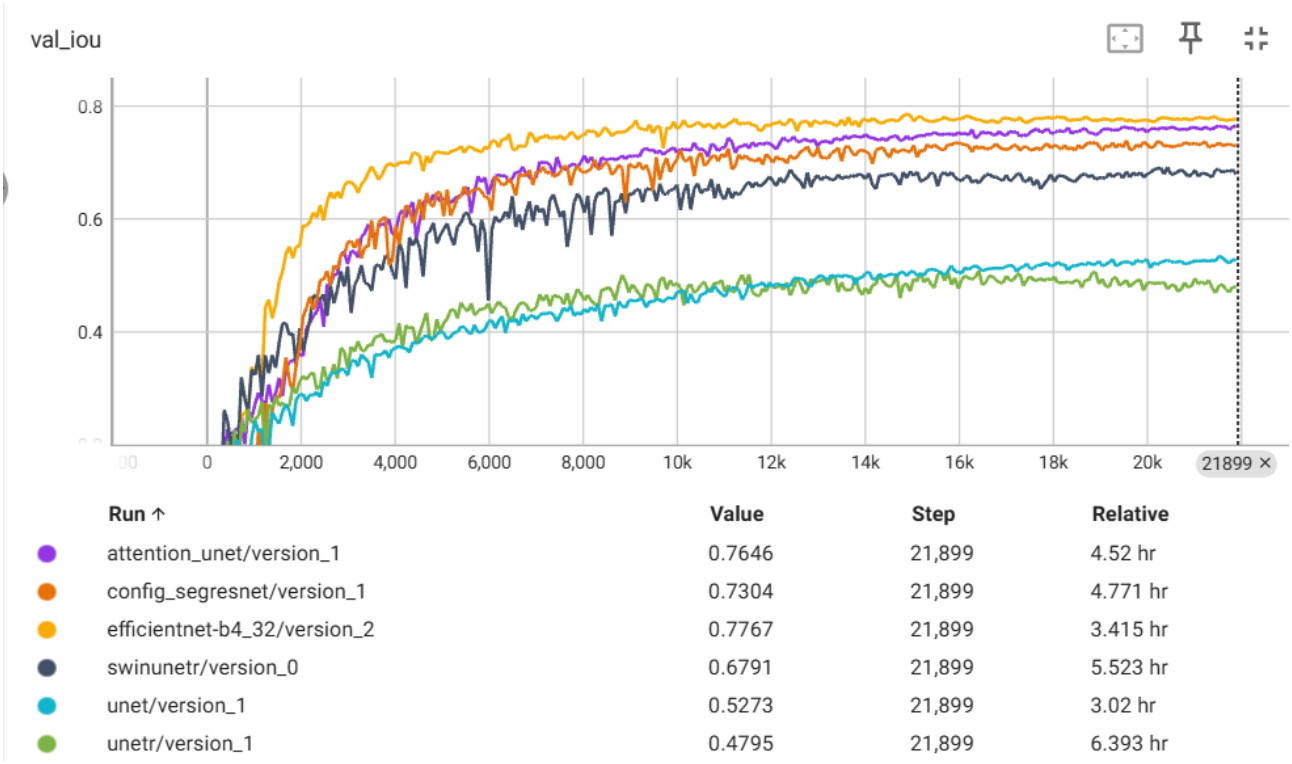
Figure 10 presents the evolution of the mean Dice coefficient on the validation set for all examined architectures. EffiSegNet-B4-32 and Attention U-Net achieve the highest final Dice scores, indicating big overlap between predicted and ground-truth segmentations. SegResNet also shows strong performance, while SwinUNETR attains a moderate Dice score. Both UNet and UNETR yield substantially lower Dice coefficients, reflecting their comparatively limited segmentation accuracy.



Run ↑	Value	Step	Relative
attention_unet/version_1	0.8458	21,899	4.52 hr
config_segresnet/version_1	0.8165	21,899	4.771 hr
efficientnet-b4_32/version_2	0.8488	21,899	3.415 hr
swinunetr/version_0	0.767	21,899	5.523 hr
unet/version_1	0.6591	21,899	3.02 hr
unetr/version_1	0.58	21,899	6.393 hr

Figure 10: Validation mean Dice curves

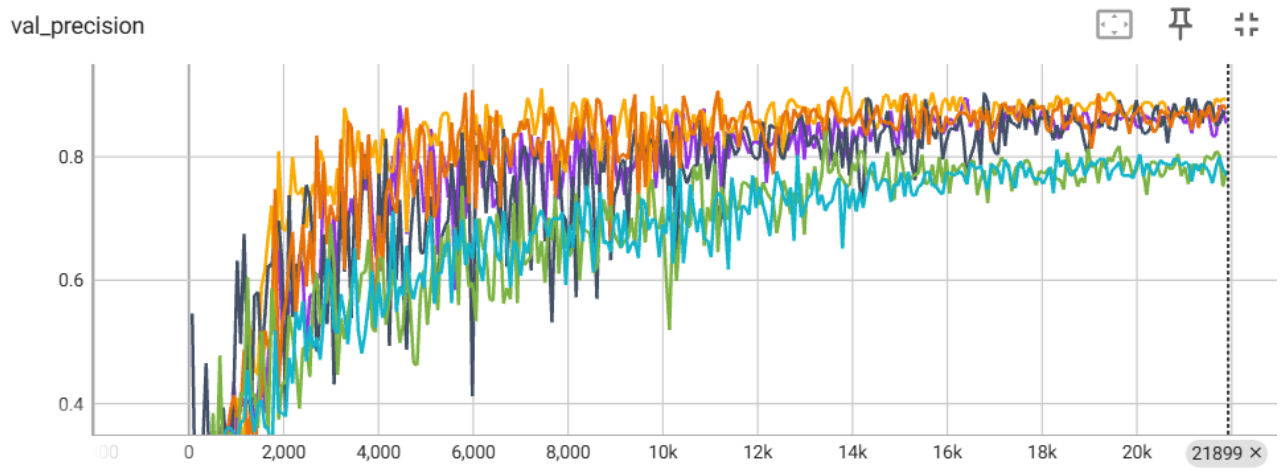
Figure 11 shows the progression of mean Intersection over Union (IoU) across training epochs for the evaluated architectures. EffiSegNet-B4-32 achieves the highest final mean IoU, followed closely by Attention U-Net and SegResNet, each displaying consistently strong overlap between predicted and true segmentation masks. Swin-UNETR shows moderate performance, while both UNet and UNETR lag considerably behind. These results showcase the superior region-level segmentation capabilities of advanced architectures, highlighting their enhanced effectiveness in accurate polyp delineation compared to traditional UNet and transformer-only models.



Run	Value	Step	Relative
attention_unet/version_1	0.7646	21,899	4.52 hr
config_segresnet/version_1	0.7304	21,899	4.771 hr
efficientnet-b4_32/version_2	0.7767	21,899	3.415 hr
swinunetr/version_0	0.6791	21,899	5.523 hr
unet/version_1	0.5273	21,899	3.02 hr
unetr/version_1	0.4795	21,899	6.393 hr

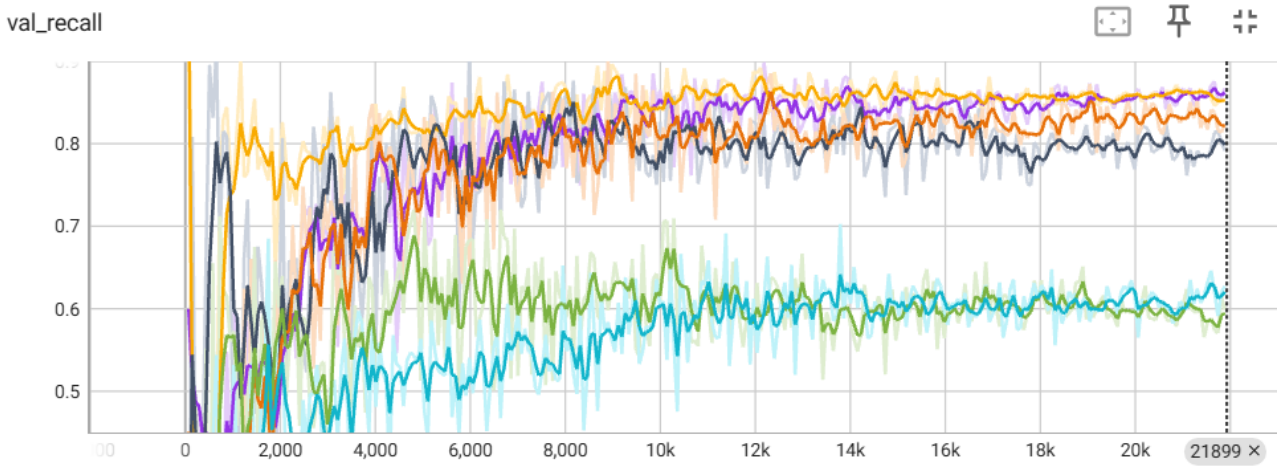
Figure 11: Validation mean IoU curves

Figures 12, 13 present the evolution of validation precision and recall, respectively, for all compared architectures. EffiSegNet-B4-32 achieves the highest final precision, followed closely by SwinUNETR, SegResNet, and Attention U-Net. These high precision values reflect each model's ability to minimize false positive predictions in polyp segmentation. For the recall metric, Attention U-Net achieves the highest value, with EffiSegNet-B4-32 and SegResNet also performing strongly, indicating a greater capacity for capturing true positives and limiting false negatives. SwinUNETR demonstrates balanced performance, while both UNet and UNETR show consistently lower values across both metrics, reflecting their limited segmentation sensitivity and specificity relative to modern architectures.



Run	Value	Step	Relative
attention_unet/version_1	0.8538	21,899	4.52 hr
config_segresnet/version_1	0.876	21,899	4.771 hr
efficientnet-b4_32/version_2	0.8924	21,899	3.415 hr
swinunetr/version_0	0.8708	21,899	5.523 hr
unet/version_1	0.7705	21,899	3.02 hr
unetr/version_1	0.7775	21,899	6.393 hr

Figure 12: Validation precision curves

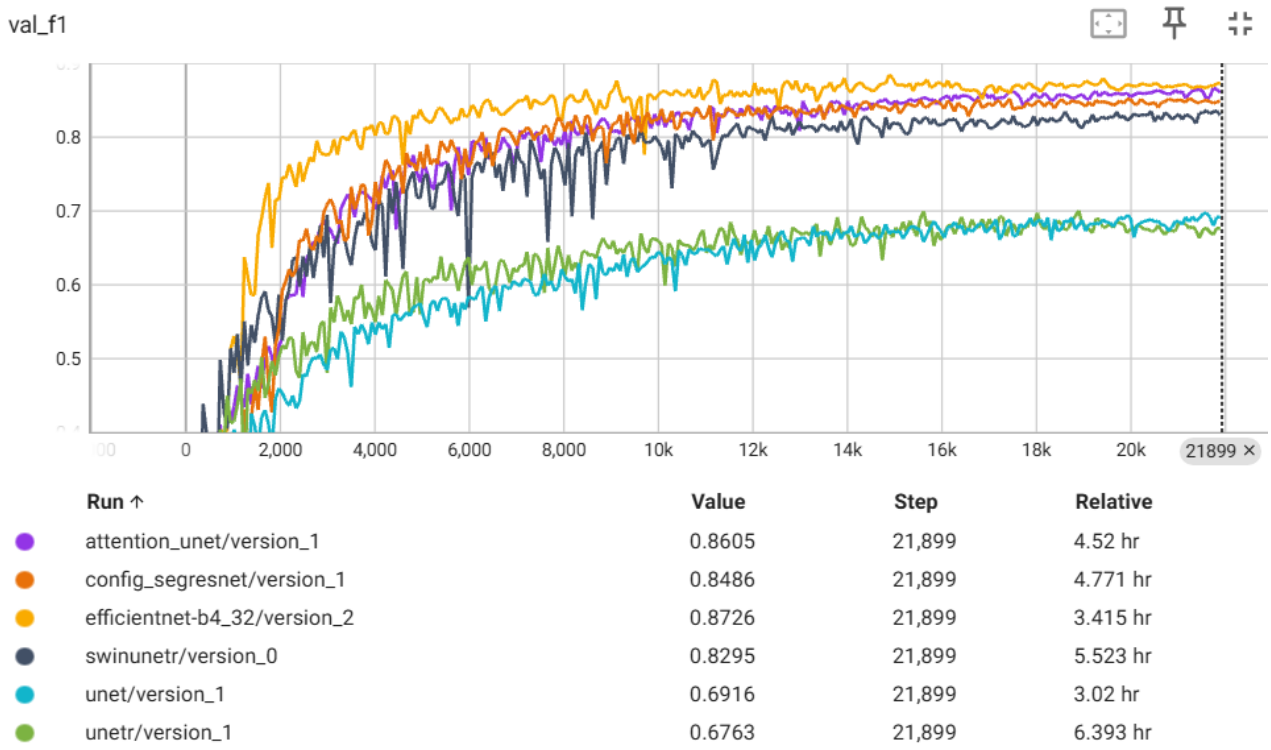


Run	Smoothed	Value	Step	Relative
attention_unet/version_1	0.8621	0.8672	21,899	4.52 hr
config_segresnet/version_1	0.8223	0.8229	21,899	4.771 hr
efficientnet-b4_32/version_2	0.8528	0.8537	21,899	3.415 hr
swinunetr/version_0	0.7989	0.7919	21,899	5.523 hr
unet/version_1	0.6204	0.6275	21,899	3.02 hr
unetr/version_1	0.5947	0.5984	21,899	6.393 hr

Figure 13: Validation recall curves

The F_1 score, represents the harmonic mean of precision and recall. EffiSegNet-B4-32 and Attention U-Net, with both high precision and recall, achieve the highest F_1 scores, indicating robust balanced performance in both error types. Similarly, Seg-

ResNet and SwinUNETR maintain intermediate F_1 scores due to their consistent alignment of precision and recall. The F_2 score, which places additional emphasis on recall, favors architectures with superior sensitivity to true positives. Here, Attention U-Net and EffiSegNet-B4-32 again lead, confirming their efficiency in reducing missed polyp pixel detections. The notably lower F_1 and F_2 scores of UNet and UNETR imply a higher degree of error either in missed segmentations or in incorrect positive segmentations.



Run	Value	Step	Relative
attention_unet/version_1	0.8605	21,899	4.52 hr
config_segresnet/version_1	0.8486	21,899	4.771 hr
efficientnet-b4_32/version_2	0.8726	21,899	3.415 hr
swinunetr/version_0	0.8295	21,899	5.523 hr
unet/version_1	0.6916	21,899	3.02 hr
unetr/version_1	0.6763	21,899	6.393 hr

Figure 14: Validation F_1 curves

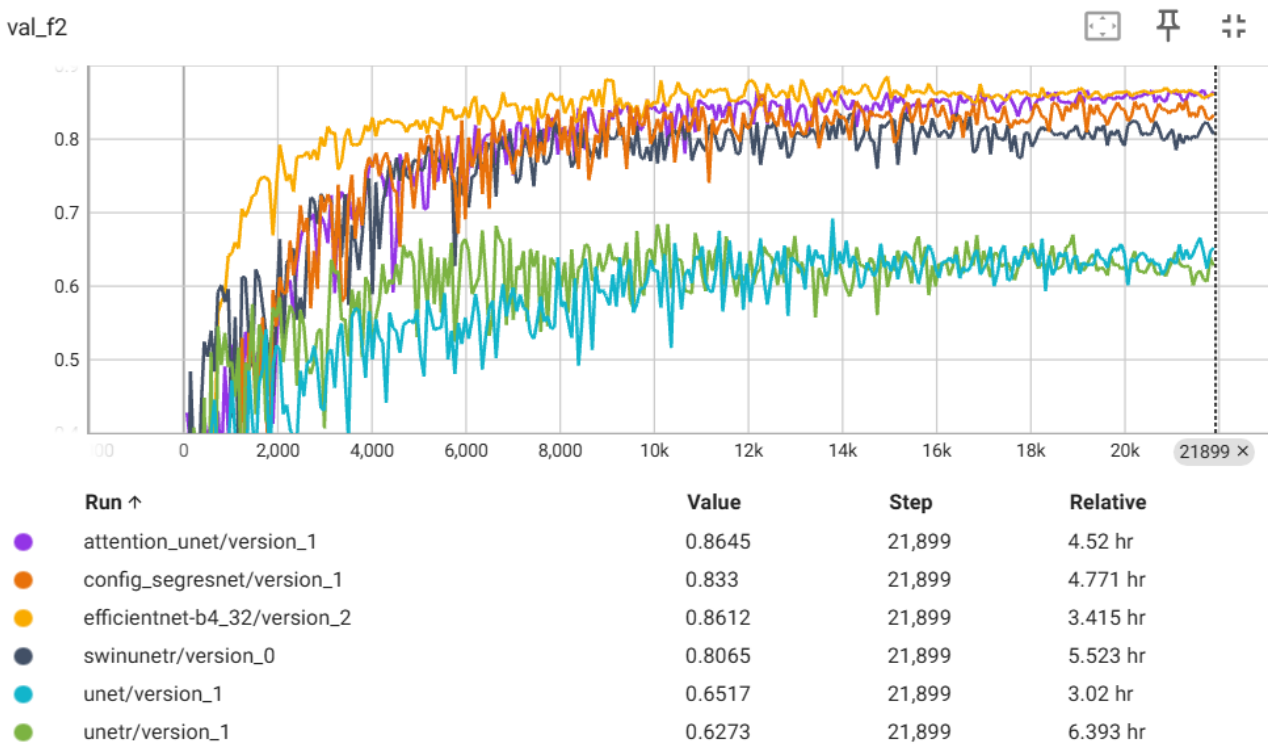


Figure 15: Validation F_2 curves

Table 1 presents a comparison of segmentation models evaluated on the PolypGen test set (C6 center), reporting performance across loss, accuracy, mean Dice (mDice), mean Intersection-over-Union (mIoU), precision, recall, F_1, and F_2 scores. The table includes the from scratch trained architectures, as well as the ensemble method, thus offering insight into both single-model and voting-based segmentation strategies.

	Loss	Accuracy	mDice	mIoU	Precision	Recall	F1	F2
Unet [48]	0.713	0.9429	0.5495	0.4234	0.665	0.5666	0.6119	0.5839
Attention Unet [45]	0.3931	0.9697	0.7672	0.6863	0.8007	0.8232	0.8118	0.8186
SegResNet [24]	0.4649	0.9673	0.7351	0.6458	0.8245	0.7476	0.7842	0.7618
EffiSegNet [57]	**0.3724**	**0.9721**	**0.786**	**0.7115**	0.8057	**0.8546**	**0.8294**	**0.8444**
UnetR [22]	0.8848	0.9399	0.4984	0.4074	0.6744	0.4717	0.5551	0.5018
SwinUnetR [21]	0.4815	0.9663	0.7229	0.6377	**0.8301**	0.7241	0.7734	0.743
Ensemble	-	0.9763	0.7933	0.7223	0.8791	0.8136	0.845	0.8259

Table 1: Evaluation Metrics results for all architectures

EffiSegNet achieves the best overall individual performance. It showcases the lowest loss (0.3724), highest test accuracy (0.9721), as well as the top mDice (0.786), mIoU (0.7115) and recall (0.8546) among the single models. Its precision (0.8057) is also among the highest, resulting in F_1 and F_2 scores of 0.8294 and 0.8444, respectively, which are also the highest values. These results underscore EffiSegNet's ability to produce highly accurate and consistent polyp segmentations, likely because of its efficient feature extraction and deep representational capacity.

Attention U-Net and SegResNet also display strong performance. Attention U-Net

stands out for high recall (0.8232) and overall balance, achieving an F_1 of 0.8118 and F_2 of 0.8186, emphasizing its strength in correctly identifying positive segmentation cases, crucial in clinical applications to minimize missed detections. SegResNet provides competitive accuracy (0.9673), mDice (0.7351), and mIoU (0.6458), with balanced precision and recall, which contributes to robust F_1/F_2 scores (0.7842/0.7618). SwinUNetR displays the highest precision among all models (0.8301), indicating a greater ability to avoid false positive predictions. However, its recall is lower (0.7241), suggesting a trade-off where the model may miss more true polyp regions than EffiSegNet or Attention U-Net. Its F_1 (0.7734) and F_2 (0.743) still demonstrate competitive, but not leading, overall segmentation quality.

Baseline UNet and UNetR lag behind the advanced models, reflected by considerably lower mDice (0.5495 and 0.4984) and mIoU (0.4234 and 0.4074). They also exhibit the lowest recall (0.5666 for UNet, 0.4717 for UNetR), F_1 (0.6119 and 0.5551), and F_2 (0.5839 and 0.5018), confirming the limitations of traditional encoder-decoder structures and transformer-only architectures in extracting the complex features necessary for robust polyp segmentation in this challenging dataset.

The ensemble methodology, combining EffisegNet, Attention U-Net, and SegResNet via soft voting on predicted probabilities, delivers the top aggregate performance across nearly all metrics. Notably, it achieves the highest mDice (0.7933), mIoU (0.7223), and accuracy (0.9763). Precision (0.8791) is also elevated, and recall remains high (0.8136), resulting in outstanding F_1 (0.845) and F_2 (0.8259) scores. This superior test-time performance can be attributed to the ensemble's ability to aggregate complementary strengths and compensate for individual model weaknesses, particularly by increasing robustness and generalization to unseen data.

To complement the quantitative evaluation, we provide a subset of segmentation predictions on the evaluation set, from each architecture alongside the original images and the corresponding ground truth masks. These qualitative results illustrate the models' ability to detect polyp regions under diverse conditions and varied domain characteristics. Below we present a figure for each architecture:

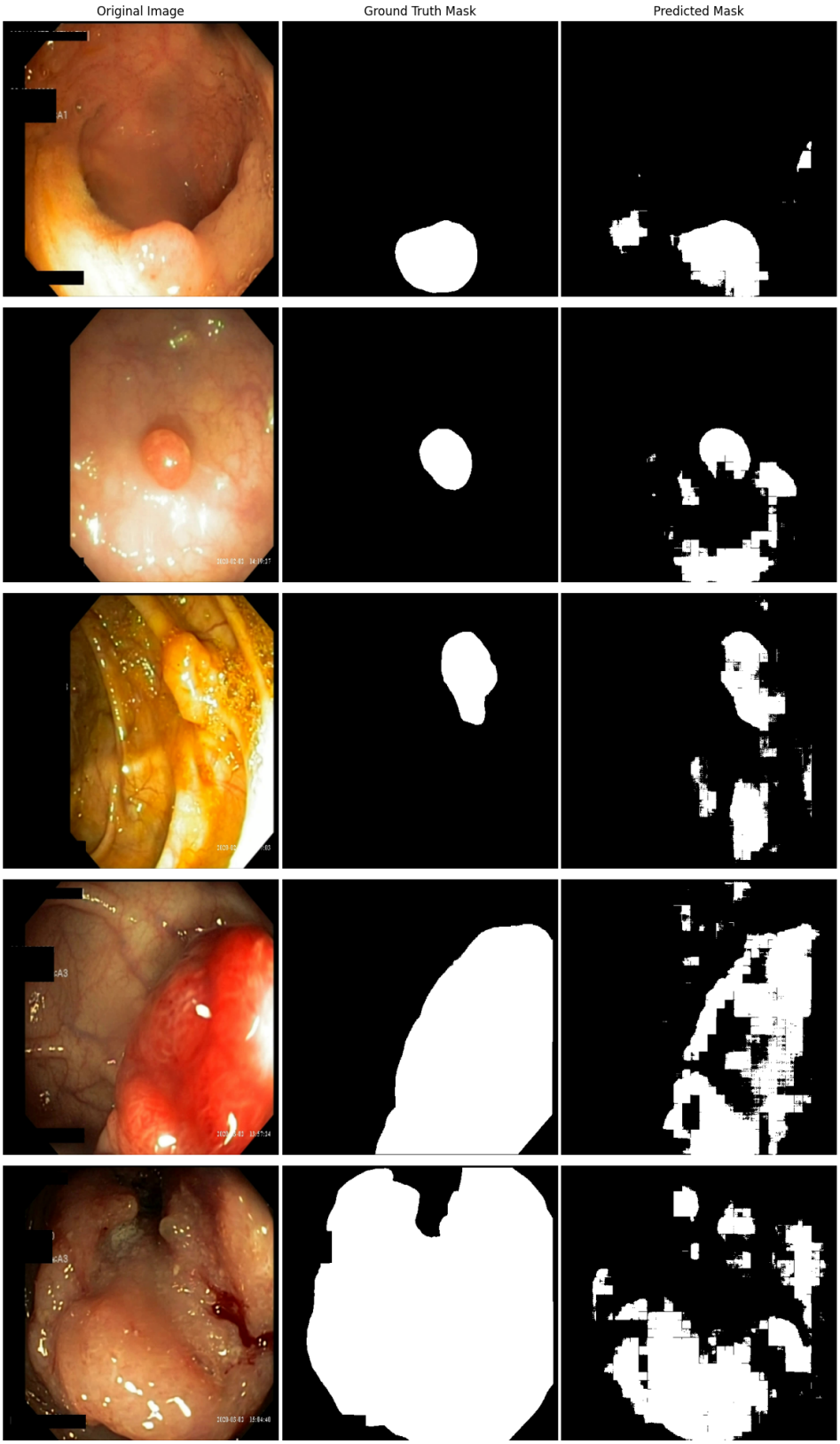


Figure 16: Predicted masks using UNet

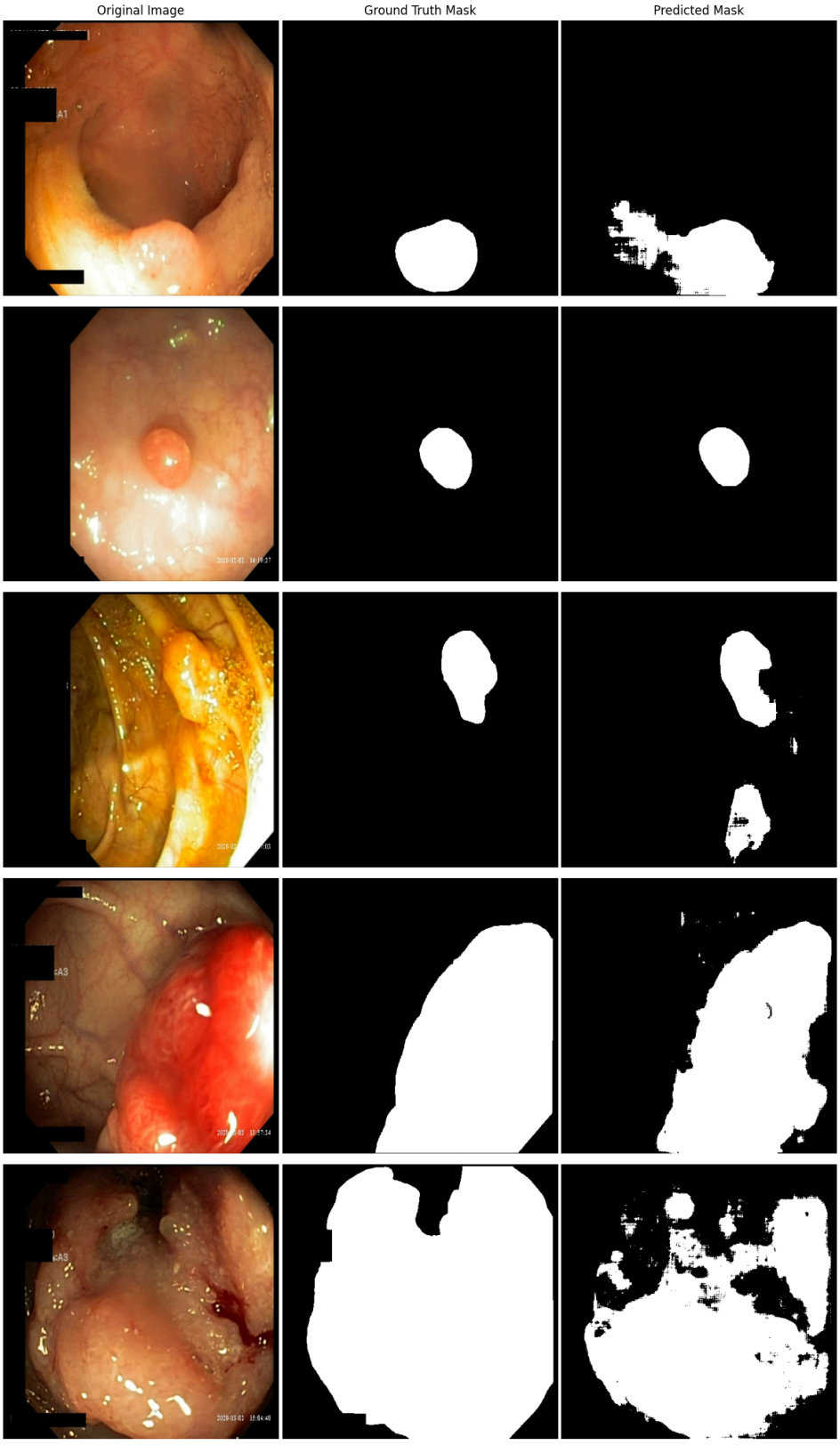


Figure 17: Predicted masks using Attention UNet

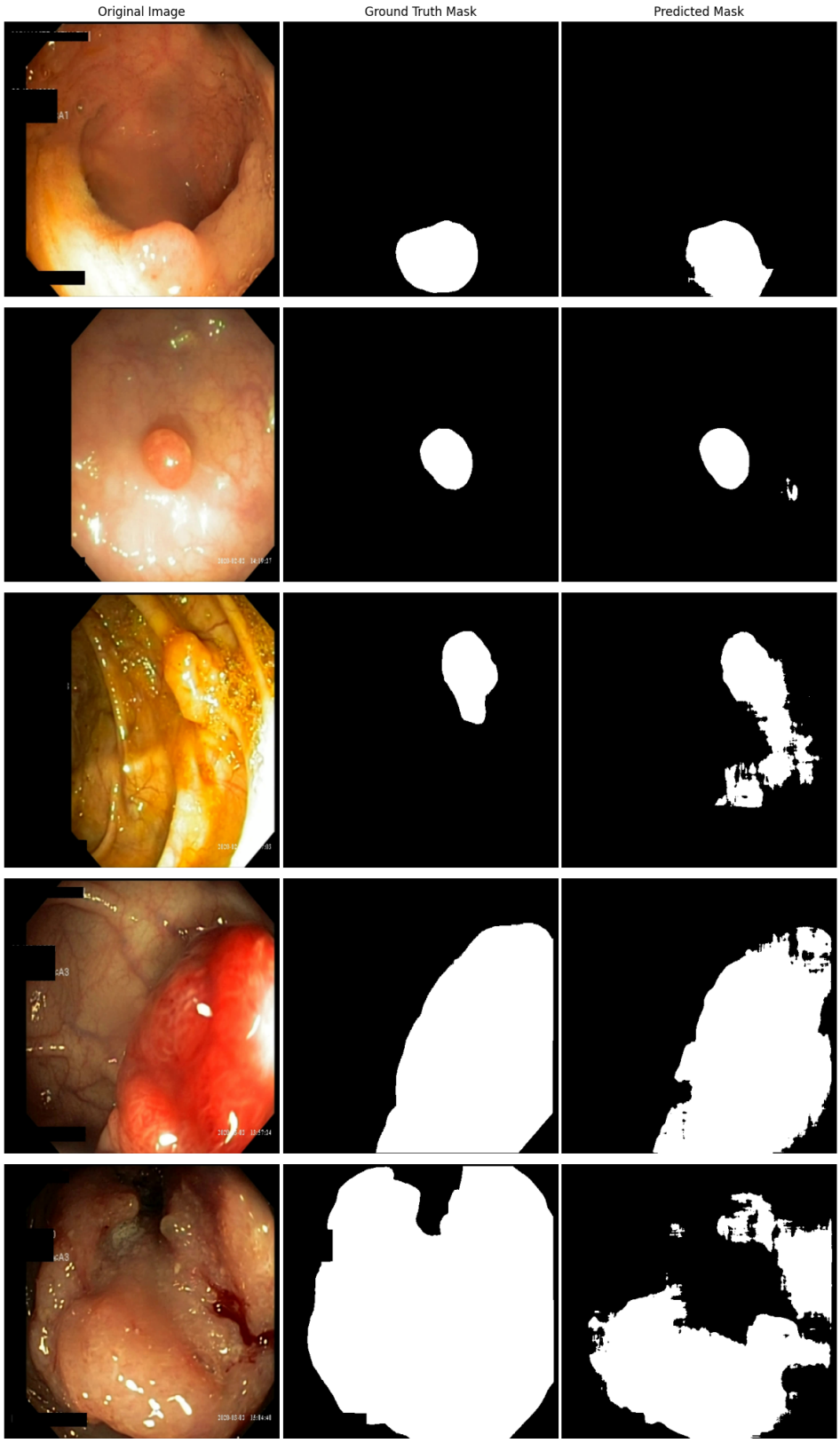


Figure 18: Predicted masks using SegResNet

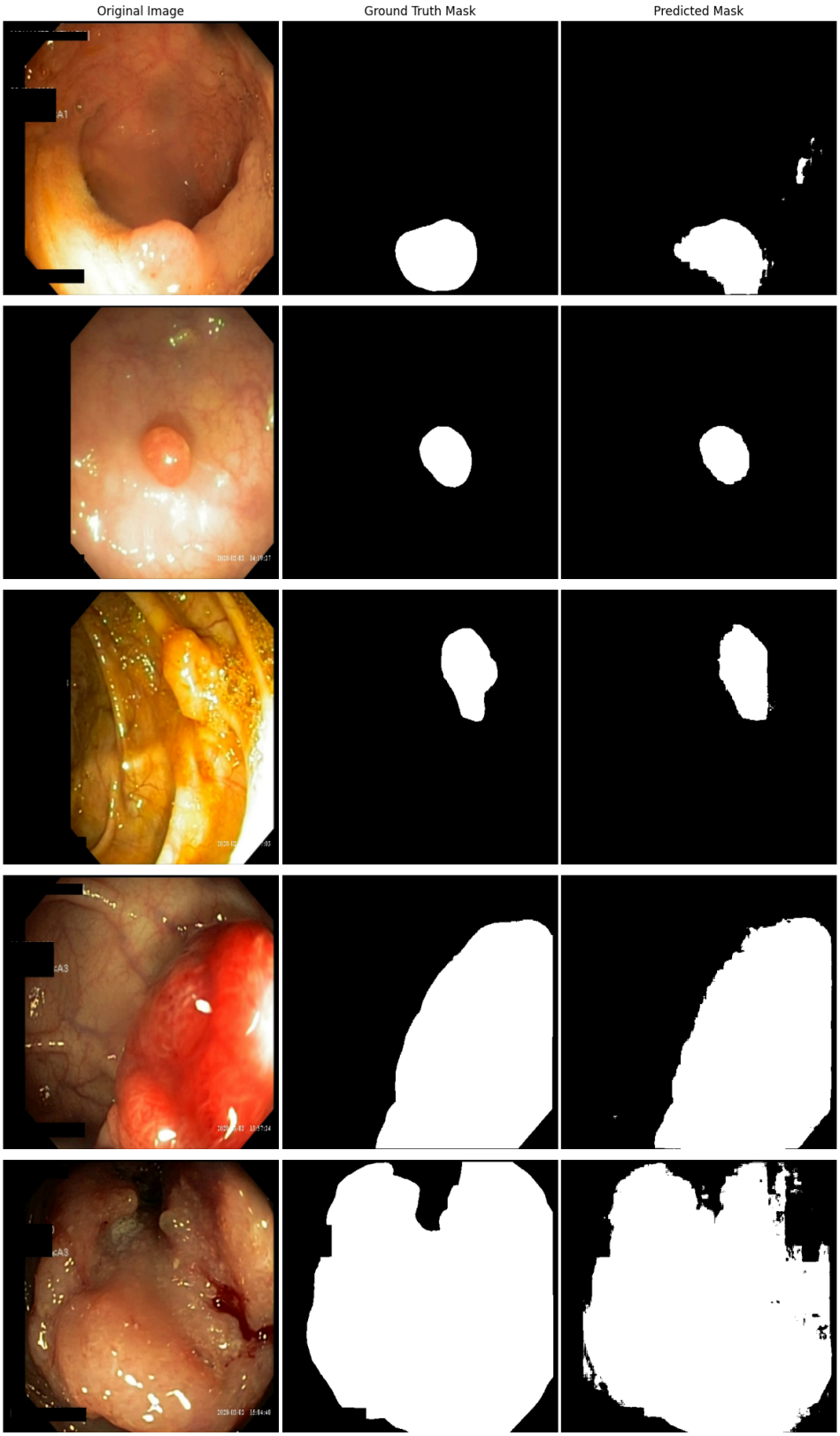


Figure 19: Predicted masks using EffiSegNet-B4

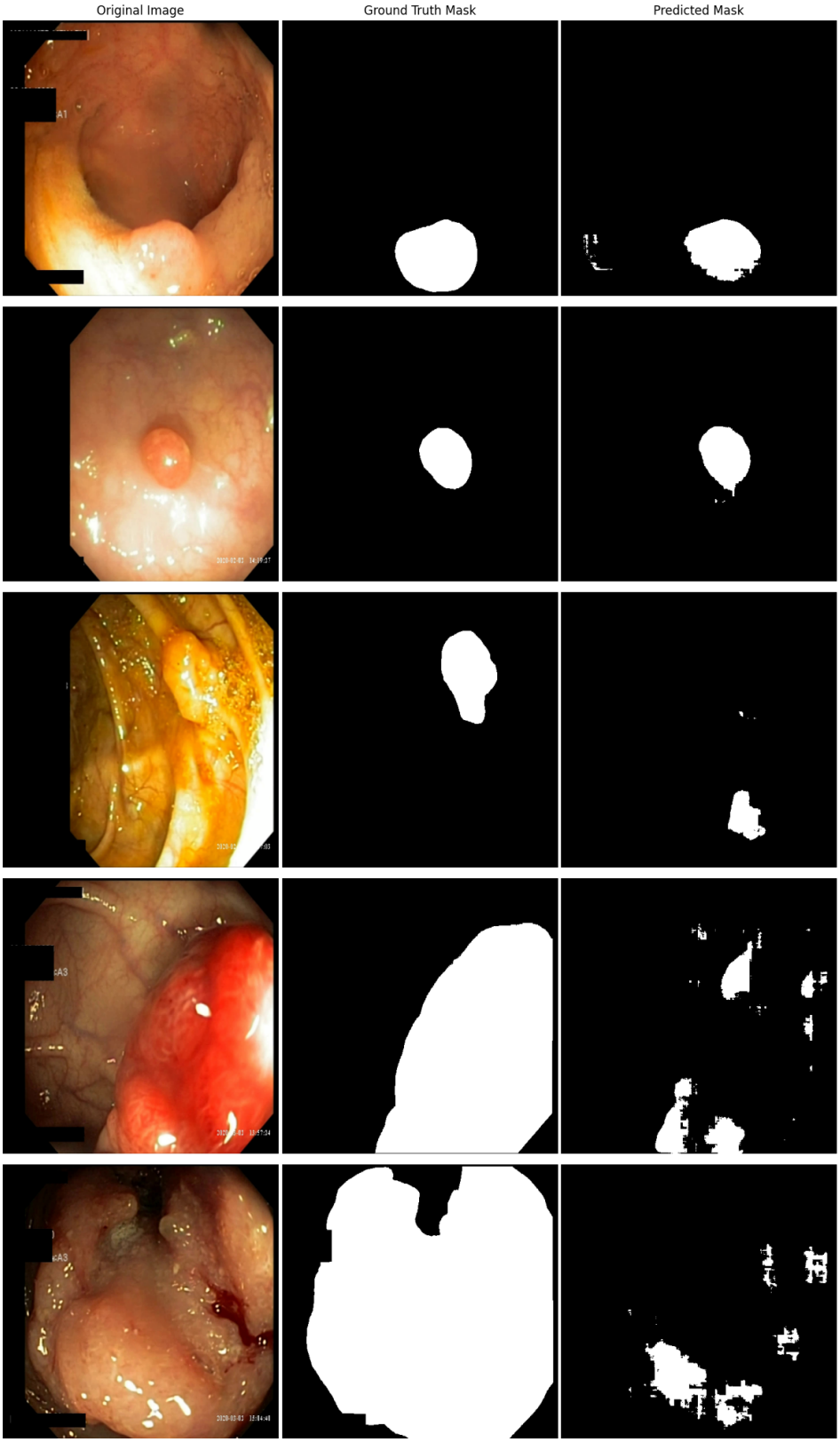


Figure 20: Predicted masks using UNETR

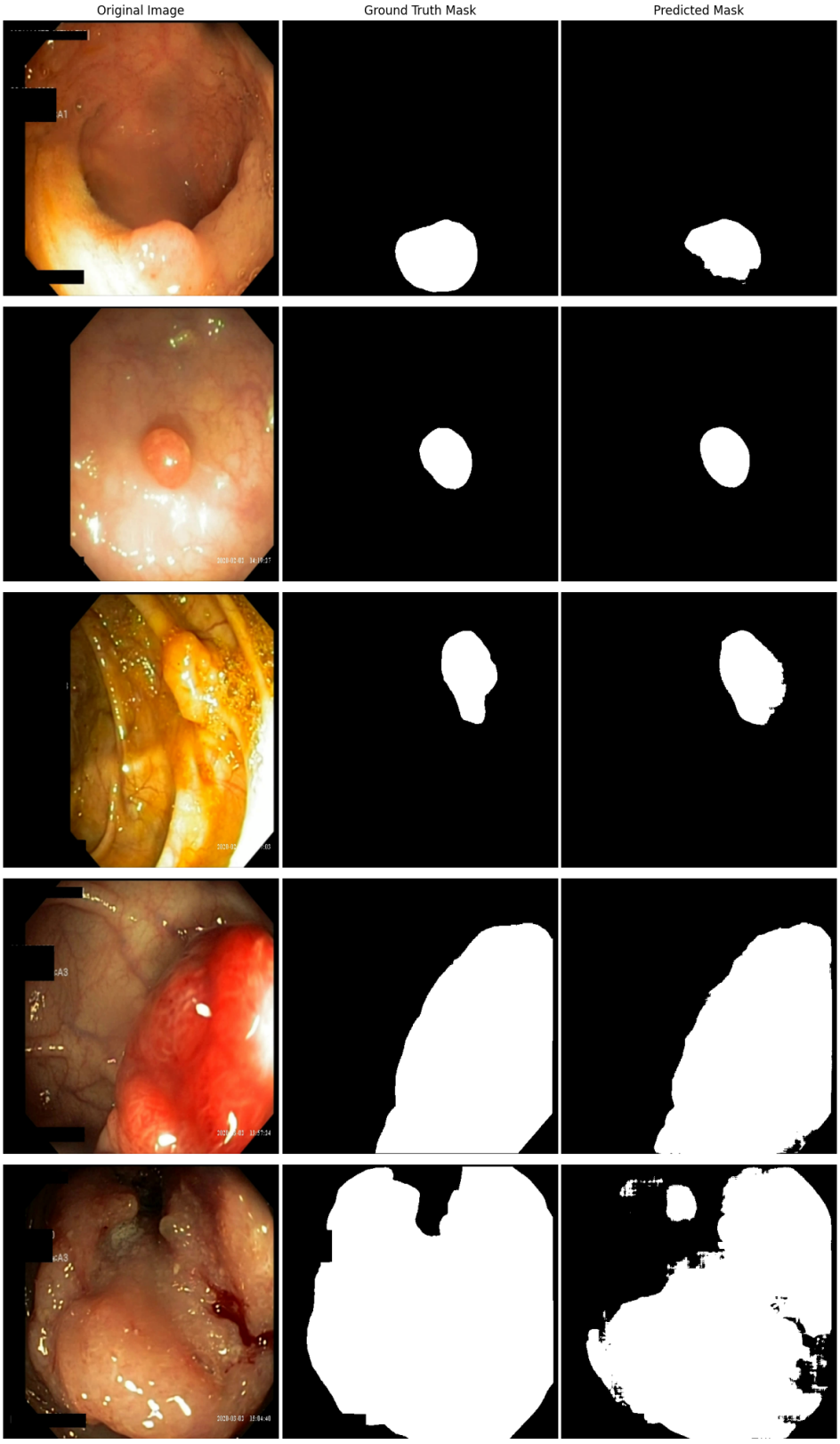


Figure 21: Predicted masks using SwinUNETR

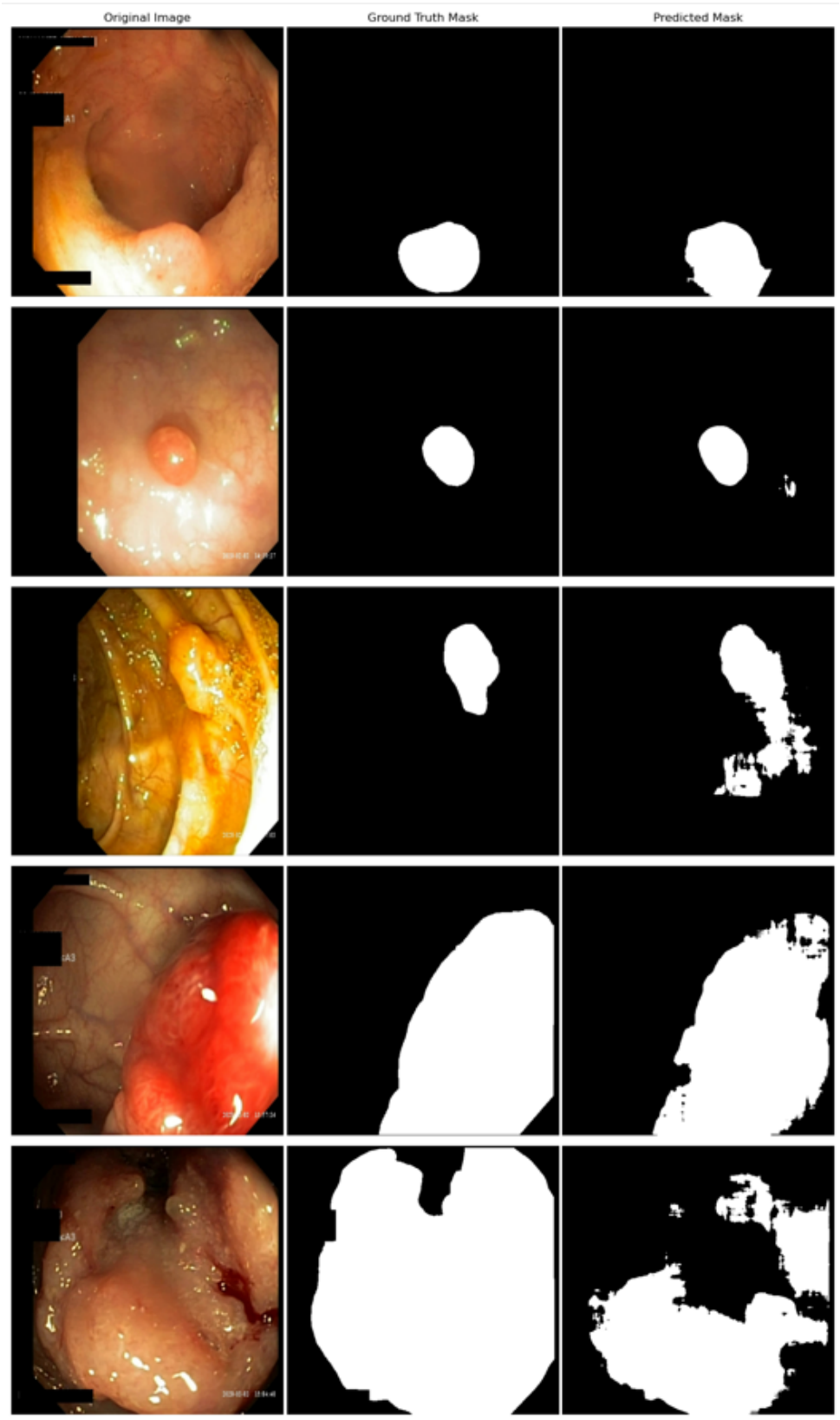


Figure 22: Predicted masks using the Ensemble method

UNet's segmentation masks show several common failure patterns: for small and regular polyps, the results often present smaller than ground truth, reflecting under-segmentation. For larger and more complex shapes, the predicted masks tend to include extraneous areas (false positives), with artifacts particularly evident in low-

contrast or cluttered backgrounds. Boundaries are often imprecise, and the segmentation may break up into several smaller regions. It is obvious that the model has difficulty distinguishing polyps from challenging tissue structures.

Attention UNet excels in focusing on the core polyp region, often producing segmentation masks that overlap very well with ground truth. In most cases, the architecture captures both the size and shape of the polyp accurately, with only minor misclassifications at the boundaries or in the presence of specular highlights. This demonstrates the effectiveness of spatial attention in suppressing irrelevant features and highlighting the polyp structure.

SegResNet produces segmentation masks that usually follow the main outline of the ground truth, but compared to attention-based models, it sometimes leaves out subtle portions of the polyp or includes some areas in the background, especially when polyps have irregular shapes or appear in challenging contexts. Overall, the architecture handles most test cases with solid consistency while showing some vulnerability in edge definition for complex boundaries.

EffiSegNet demonstrates strong segmentation performance across various polyp sizes and challenging imaging conditions. The predicted masks closely align with the ground truth, especially for small and moderately sized polyps, with only slight oversegmentation or undersegmentation in some areas. For larger polyps, the network accurately captures the main shape but occasionally introduces false positive fragments in ambiguous regions.

The UNETR model shows variable segmentation performance across examples. For small and well-contrasted polyps, it can approximate the shape and location of the ground truth mask, but the contours are often less precise, and predicted masks can be noticeably rough or contain holes. In cases with larger or irregular polyps and complex tissue background the mask predictions suffer from sparse or patchy activations, missing substantial portions of the target lesion. The segmentation output occasionally includes false positive blobs in background regions, and the model generally underestimates the full extent of the polyp, reflecting its difficulty in capturing fine boundary details and contextual information when compared to more established convolutional or attention-based architectures. This is consistent with limitations often observed in transformer-only models when applied to medical image segmentation tasks with limited data or strong domain shift.

SwinUNETR exhibits strong suppression of false positives but may suffer from fragmented or incomplete mask predictions. In many cases, only core sections of the

polyp are segmented, and the output mask can appear patchy, especially for larger or more irregular lesions. This suggests a tendency to be conservative, possibly prioritizing precision at the cost of recall. Background noise is mostly absent, but the model sometimes fails to capture the full extent of polyps, leaving visible gaps compared to the ground truth mask. This conservative behavior is typical for some transformer-based models trained on limited data.

The ensemble method integrates strengths from all its individual models, resulting in segmentation masks that best reflect the core polyp regions with improved boundary accuracy and reduced false positives. The predicted masks display a high degree of agreement with ground truth, even for polyps with challenging morphology or in cluttered backgrounds. Occasional small errors persist, but both oversegmentation and undersegmentation are less frequent than in single model outputs. Ensemble soft voting increases generalization on unseen cases.

6 Discussion

6.1 Research Outcomes

The evaluation of six segmentation architectures on the PolypGen dataset reveals clear patterns between model design and performance in clinical polyp segmentation. The results show how different architectural choices affect learning ability, generalization performance, and computational efficiency in each model framework.

The validation loss and accuracy curves demonstrate significant performance differences among the evaluated architectures. EffiSegNet-B4-32 shows the best performance characteristics, achieving the lowest validation loss and highest accuracy while remaining stable throughout 300 training epochs. EffiSegNet-B4 delivers the best single model performance by using the EfficientNet family's compound scaling strategy and Squeeze & Excitation (SE) modules. Compound scaling optimizes network depth, width, and input resolution together to maximize learning capacity per parameter. This allows EffiSegNet to capture features at multiple scales, from small protrusions to large flat surfaces in polyp regions. The SE blocks improve feature quality by adjusting channel responses, highlighting useful feature maps while reducing irrelevant ones. For polyp segmentation, these mechanisms help detect lesions of varying sizes and shapes, explaining the model's stable training, low validation loss, and strong performance on unseen center C6 images. These findings match the work of Vezakis et al [57], who showed that EffiSegNet-B4 achieved excellent results on the Kvasir-SEG dataset with an F1 score of 0.9552, mean Dice coefficient of 0.9483, and mean IoU of 0.9056, representing the highest reported scores for this dataset. Similarly, Abdelrahman and Viriri [1] confirmed that EfficientNet-based architectures work well for medical image segmentation, with EfficientNet models achieving mean IoU scores ranging from 0.976 to 0.980 across different variants, with EfficientNet-B4 showing excellent tumor detection capabilities.

Attention U-Net shows similarly strong performance, achieving the second highest performance metrics. Attention U-Net performs much better than standard U-Net mainly because it uses sophisticated Attention Blocks that calculate feature weights through a gating mechanism. This allows the model to focus on relevant polyp features while suppressing irrelevant background information like mucosal folds, bright reflections, and instrument shadows that can interfere with accurate segmentation in endoscopy images. Unlike standard U-Net's fixed skip connections that simply combine encoder and decoder features, potentially including noisy information, Attention U-Net learns to selectively control this information flow. By computing attention weights that emphasize encoder features based on their relevance to the current decoder state, the model filters out distracting signals and focuses on regions

important for polyp identification. This selective enhancement explains the significant improvement in recall (0.8232 vs. 0.5666) and boundary precision observed in our results, as the network better captures subtle tissue variations against complex backgrounds. These findings match evaluations by Al Qurri and Almekkawy [4], who showed that attention enhanced U-Net architectures consistently improved prediction performance across different medical imaging datasets. The effectiveness of attention mechanisms is further supported by Xie et al., who systematically reviewed over 300 articles and found that attention mechanisms help distinguish important anatomical features from irrelevant background information, which is particularly important for accurate polyp boundary detection. Additionally, Cai and Wang [11] confirmed that attention gates reduce confusion in skip connections while modeling long-range feature dependencies, directly addressing the challenges seen in complex polyp shapes.

SegResNet achieves competitive middle-range performance due to its residual framework. SegResNet uses ResBlock modules throughout its encoder-decoder structure, with each block implementing skip connections that help gradient flow and enable training of deeper networks. This design allows effective learning of features at different levels while maintaining computational efficiency.

The transformer-based models SwinUNETR and UNETR show very different results. SwinUNETR achieves moderate success (accuracy 0.9663) but shows training instability, while UNETR demonstrates poor performance (accuracy 0.9399, Dice 0.4984). SwinUNETR uses Window Attention mechanisms that divide input features into local windows and compute attention within these spatial regions. The shifted window mechanism alternates between different window configurations across layers, enabling connections between windows while maintaining computational efficiency. However, the training instability observed in our validation curves suggests that this localized attention may not be optimal for polyp segmentation, where understanding the full image context is crucial for distinguishing polyps from surrounding tissue. UNETR's poor performance despite its sophisticated Vision Transformer (ViT) encoder architecture is notable. These results match findings by researchers who evaluated transformer-based architectures in polyp segmentation tasks. The FCB-SwinV2 Transformer study [16] reported that while SwinV2-based models can achieve competitive performance on general segmentation benchmarks like ADE20K, their application to medical imaging requires careful modifications to maintain stability and accuracy. Similarly, PolySegNet research [37] showed that transformer-based approaches for polyp segmentation, while promising, require substantial computational resources and careful parameter tuning to achieve good performance, confirming our observations about training instability in SwinUNETR.

The poor performance metrics suggest that pure global attention mechanisms, while theoretically good for capturing long-range relationships, may lack the local feature detection mechanisms needed for accurate polyp boundary identification. The UNETR and SwinUNETR models replace standard convolutional encoders with Vision Transformer (ViT) blocks or hierarchical shifted-window attention mechanisms and were expected to excel at capturing long-range dependencies. However, their poor performance (UNETR Dice 0.4984, SwinUNETR Dice 0.7229) reveals two key limitations. First, the fixed patch division (e.g. 16×16 grid) and localized attention windows prevent the fine-grain boundary refinement needed for accurate polyp identification, causing both missed small or flat polyps and fragmented predictions on irregular lesions. Second, transformers generally need very large, consistent training datasets to avoid overfitting their position information and attention weights. The relatively small size (1,537 images) and variability of PolypGen causes training instability, as seen in the volatile validation loss curves of both models. Therefore, while global context is theoretically valuable, the lack of dedicated local feature extraction mechanisms hurts practical segmentation accuracy in small data, high domain shift scenarios. This assessment is supported by comprehensive analysis from medical imaging transformer studies [35], which noted that while Swin UNETR achieved excellent performance on large-scale datasets like Medical Segmentation Decathlon, performance drops significantly when applied to smaller, specialized datasets, confirming our observations with the PolypGen dataset.

The strong performance of attention-enhanced models (EffiSegNet-B4-32 and Attention U-Net) compared to traditional architectures demonstrates the importance of selective feature enhancement in medical image segmentation applications. Unlike natural image segmentation tasks, polyp detection requires distinguishing subtle tissue variations against complex, varied anatomical backgrounds. The excellent performance of EffiSegNet-B4-32 matches findings in the literature. Vezakis et al. [57] showed that EfficientNet-based architectures achieve state-of-the-art performance on the Kvasir-SEG dataset, with their EffiSegNet-B4 achieving F_1 scores of 0.9552. Our results agree with these findings, with EffiSegNet-B4-32 achieving superior test metrics across all evaluation criteria (F_1: 0.8294, Dice: 0.786, IoU: 0.7115).

The standard UNet performance characteristics (accuracy 0.9429, Dice 0.5495) show the limitations of traditional encoder-decoder architectures when applied to challenging medical segmentation tasks. While this design has worked well for many segmentation applications, the fixed skip connection mechanism lacks the adaptive feature selection capabilities shown by attention-based approaches. The substantial performance difference between UNet and Attention U-Net demonstrates the benefit of

incorporating learned attention mechanisms.

Our ensemble approach, combining EffiSegNet-B4-32, Attention U-Net, and SegResNet through a soft voting mechanism, achieves the best performance across all evaluation metrics (accuracy: 0.9763, Dice: 0.7933, F_1: 0.845). The ensemble method uses soft voting, a strategy that uses the continuous probability distributions generated by each individual model rather than simple binary predictions. Soft voting preserves the probability distributions from each model, keeping the uncertainty information that is lost when only using discrete predictions. Models with higher confidence naturally receive greater influence in the final prediction, creating an automatic confidence weighting mechanism. By averaging probability distributions, soft voting typically produces more stable and robust predictions compared to simple majority voting schemes. These findings are strongly supported by ensemble learning research in medical image analysis. Sherazi et al. [50] showed that soft voting ensemble classifiers consistently outperformed individual machine learning models, achieving 99.61% AUC compared to individual model performance ranging from 98.15-99.41%. Similarly, comprehensive ensemble studies in polyp segmentation [43] confirmed that combining diverse architectures including CNNs and transformers through ensemble learning significantly improves segmentation accuracy and robustness. The dual ensemble system research [60] further validated that multi-model approaches addressing architectural diversity lead to substantial performance improvements in polyp segmentation tasks, with ensemble methods showing superior stability across different datasets compared to individual models. EffiSegNet-B4-32 provides superior feature extraction through its compound scaling methodology, Attention U-Net provides enhanced feature localization through its attention mechanisms, and SegResNet offers robust residual learning characteristics. The ensemble's superior precision (0.8791) and balanced F_1 score demonstrate the value of architectural diversity in achieving robust segmentation performance. This approach aligns with diversity-promoting ensemble strategies, which emphasize that combining models with complementary strengths while maintaining independence between ensemble components leads to optimal segmentation performance in medical imaging applications.

In summary, these findings show that effective medical image segmentation under domain variability requires architectures that combine strong local feature detection through attention or SE mechanisms with stable training dynamics. Pure transformer encoders lack sufficient mechanisms for fine boundary identification in limited data datasets, whereas attention-enhanced and compound-scaled convolutional models excel by adaptively focusing on relevant anatomical structures. Finally, soft voting ensembles leverage complementary design strengths to deliver the most generalizable

performance, highlighting the value of architectural diversity in clinical deployment scenarios.

6.2 Study Limitations

Despite these advances, several methodological limitations exist in our analysis that require careful consideration. The evaluation is conducted on a single dataset (PolypGen), and generalization to alternative polyp datasets or imaging modalities remains to be validated. While PolypGen provides valuable multicenter data, the relatively small training dataset size (1,537 images) represents a significant constraint for deep learning model optimization. This limited data availability may have prevented the models from reaching their full potential, particularly for transformer-based architectures that typically require substantially larger datasets to achieve optimal performance.

The training protocol maintains consistency across all models to ensure fair comparison, but individual architectures might benefit from specialized hyperparameter optimization strategies. Our study employed training from scratch rather than utilizing pretrained weights, which could have improved performance across all architectures, especially given the limited training data. The absence of comprehensive hyperparameter optimization for each individual model represents another limitation, as architecture specific tuning might have achieved different performance rankings.

Additional methodological constraints include the lack of external validation on independent datasets from different institutions or imaging systems, which limits the assessment of true generalization capabilities. The evaluation focuses primarily on segmentation accuracy metrics without considering computational efficiency, model complexity, or real time deployment feasibility, which are crucial factors for clinical implementation. Furthermore, the study does not address the interpretability challenges inherent in deep learning models, particularly important in medical applications where understanding model decision-making processes is essential for clinical trust and adoption.

The dataset itself presents certain limitations, including potential annotation inconsistencies between different medical centers and the exclusion of ambiguous cases that might be encountered in real clinical scenarios. The class imbalance between polyp and non-polyp regions, common in medical segmentation tasks, may have influenced model performance and evaluation metrics. Additionally, the study does not account for the variability in imaging conditions, equipment differences, or patient demographics across the six medical centers, factors that could significantly impact model robustness in clinical deployment.

Finally, the ensemble methodology, while demonstrating superior performance, introduces additional complexity and computational overhead that may limit practical implementation in resource constrained clinical environments. The soft voting approach requires running multiple models simultaneously, potentially making real-time applications challenging without specialized hardware infrastructure.

6.3 Future research

Our study highlights several directions for improving polyp segmentation. First, building larger and more varied datasets is crucial. Multicenter collections with thousands of annotated images have already shown better generalization across different endoscopy devices and patient groups. Expanding PolypGen in this way would reduce overfitting and give a more reliable measure of model robustness. Moreover, further research could focus on incorporating temporal information from video sequences, which are available on PolypGen, but were not utilized in this study.

Second, using pretrained models and then finetuning them for polyp segmentation can greatly boost performance, especially when labeled data are scarce. Networks initialized on large natural or medical image datasets converge faster and train more stably. Applying this transfer learning approach to both convolutional and transformer backbones in polyp segmentation is a promising strategy [44], [17].

Third, more work on ensemble methods is needed. Instead of averaging outputs, a separate "weighting" network can learn which model to trust for each image, based on its features. Such learned ensembling has outperformed static combinations in other medical imaging tasks. Dynamic model selection and confidence based weighting could lead to more accurate and reliable segmentation across a wide range of polyp appearances [60], [18], [50].

Fourth, EffiSegNet-B4's top performance showcases the value of testing different backbones and variants. Systematically comparing other EfficientNet scales, adding new convolutional blocks or lightweight attention modules, and even mixing convolution and transformer layers could reveal designs with even higher accuracy and lower computation cost.

Finally, combining these advances: larger, multicenter datasets, pretrained fine□tuning, adaptive ensembling, and backbone exploration offers a roadmap to overcome current limitations. Together, these steps should achieve polyp segmentation models that are more generalizable, stable, and ready for clinical use.

7 Conclusions

This work compares six state-of-the-art segmentation architectures on the PolypGen dataset under a uniform training and evaluation protocol. We trained each model from scratch for 300 epochs, recorded validation loss and accuracy curves, and evaluated performance using accuracy, Mean Dice, F_1 score, mean IoU, precision and recall. An ensemble of EffiSegNet-B4-32, Attention U-Net, and SegResNet via soft voting was also constructed to assess the benefits of model combination.

Our experiments showed that EffiSegNet-B4-32 achieved the lowest validation loss and highest accuracy among single models, owing to its compound scaling and SE modules that capture multiscale polyp features effectively. Attention U-Net ranked second by using attention gates to filter out irrelevant background signals, leading to higher recall and boundary precision. SegResNet offered solid intermediate performance through its residual connections, while transformer-based models (Swin-UNETR and UNETR) suffered from training instability and lower Dice scores, due to fixed patch partitioning and limited training data. The ensemble method outperformed all single models, demonstrating that architectural diversity and probability based soft voting achieves the most robust segmentation results.

From these findings, we conclude that convolutional models augmented with selective feature mechanisms, either through SE blocks or attention gates, are the best suited for polyp segmentation in limited data settings. Pure transformer encoders lack sufficient local bias for precise boundary delineation when data are scarce. Model ensembles further enhance robustness by leveraging complementary strengths. These insights inform the design of future polyp-specific networks and practical deployment strategies in clinical endoscopy.

8 References

[1] Abubaker Abdelrahman and Serestina Viriri. “EfficientNet family U-Net models for deep learning semantic segmentation of kidney tumors on CT images”. In: *Frontiers in Computer Science* Volume 5 - 2023 (2023). ISSN: 2624-9898. DOI: 10.3389/fcomp.2023.1235622. URL: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2023.1235622.

[2] Hesham Abuelhasan, Amaal Oshah, and Ahmed Rgibi. “Enhancing Medical Image Segmentation Based on Loss Functions Integration”. In: *Sohag Engineering Journal* 5.1 (2025), pp. 93–100. ISSN: 2735-5888. DOI: 10.21608/sej.2025.357874.1073. eprint: https://sej.journals.ekb.eg/article_419898_1554427bcbf0ded07f87e0e29eee7ae1.pdf. URL: https://sej.journals.ekb.eg/article_419898.html.

[3] Ganesh Agrawal et al. “Global burden and trends of colorectal cancer”. In: *European Journal of Cancer* XX (2025). Accessed in 2025 August, pp. 55–70. URL: https://doi.org/xxx/xxxx.

[4] Ahmed Al Qurri and Mohamed Almekkawy. “Improved UNet with attention for medical image segmentation”. en. In: *Sensors (Basel)* 23.20 (Oct. 2023), p. 8589.

[5] Saleh Alaraimi et al. “Transfer learning networks with skip connections for classification of brain tumors”. In: *International Journal of Imaging Systems and Technology* 31.3 (2021), pp. 1564–1582. DOI: https://doi.org/10.1002/ima.22546. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ima.22546. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ima.22546.

[6] Alaa Ali, Sun Min, Dong Zhou, et al. “PolypGen: A Multi-center Polyp Dataset for Generalization Assessment”. In: *arXiv preprint arXiv:2301.10805* (2023). URL: https://arxiv.org/abs/2301.10805.

[7] Sharib Ali et al. “Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge”. In: *Scientific Reports* 14.1 (Jan. 2024), p. 2032.

[8] Reza Azad et al. *Enhancing Medical Image Segmentation with TransCeption: A Multi-Scale Feature Fusion Approach*. 2023. arXiv: 2301.10847 [cs.CV]. URL: https://arxiv.org/abs/2301.10847.

[9] Reza Azad et al. *Loss Functions in the Era of Semantic Segmentation: A Survey and Outlook*. 2023. arXiv: 2312.05391 [cs.CV]. URL: https://arxiv.org/abs/2312.05391.

[10] Ashwini B et al. “Efficient skip connections-based residual network (ESRNet) for brain tumor classification”. en. In: *Diagnostics (Basel)* 13.20 (Oct. 2023).

[11] Yutong Cai and Yong Wang. *MA-Unet: An improved version of Unet based on multi-scale and attention mechanism for medical image segmentation*. 2020. arXiv: 2012.10952 [eess.IV]. URL: https://arxiv.org/abs/2012.10952.

[12] G. Jignesh Chowdary and Zhaozheng Yin. “ Med-Former: A Transformer based Architecture for Medical Image Classification ”. In: *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Vol. LNCS 15011. Springer Nature Switzerland, Oct. 2024.

[13] Peixin Dai, Jingsi Zhang, and Zhitao Shu. *Residual Connection Networks in Medical Image Processing: Exploration of ResUnet++ Model Driven by Human Computer Interaction*. 2024. arXiv: 2412.20709 [eess.IV]. URL: https://arxiv.org/abs/2412.20709.

[14] B Dhiyanesh et al. “EnsembleEdgeFusion: advancing semantic segmentation in microvascular decompression imaging with innovative ensemble techniques”. In: *Scientific Reports* 15.1 (May 2025), p. 17892.

[15] Razvan-Gabriel Dumitru, Darius Peteleaza, and Catalin Craciun. “Using DUCK-Net for polyp image segmentation”. In: *Scientific Reports* 13.1 (June 2023). ISSN: 2045-2322. DOI: 10.1038/s41598-023-36940-5. URL: http://dx.doi.org/10.1038/s41598-023-36940-5.

[16] Kerr Fitzgerald and Bogdan Matuszewski. *FCB-SwinV2 Transformer for Polyp Segmentation*. 2023. arXiv: 2302.01027 [cs.CV]. URL: https://arxiv.org/abs/2302.01027.

[17] Yuxiao Gao et al. “Medical image segmentation: A comprehensive review of deep learning-based methods”. en. In: *Tomography* 11.5 (Apr. 2025), p. 52.

[18] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, and Andreea-Iuliana Miron. *Diversity-Promoting Ensemble for Medical Image Segmentation*. 2022. arXiv: 2210.12388 [eess.IV]. URL: https://arxiv.org/abs/2210.12388.

[19] Hao Guan and Mingxia Liu. “Domain adaptation for medical image analysis: A survey”. en. In: *IEEE Trans. Biomed. Eng.* 69.3 (Mar. 2022), pp. 1173–1185.

[20] Natalie Halvorsen et al. “Benefits, burden, and harms of computer aided polyp detection with artificial intelligence in colorectal cancer screening: microsimulation modelling study”. en. In: *BMJ Med.* 4.1 (Jan. 2025), e001446.

[21] Ali Hatamizadeh et al. *Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images*. 2022. arXiv: 2201.01266 [eess.IV]. URL: https://arxiv.org/abs/2201.01266.

[22] Ali Hatamizadeh et al. “UNETR: Transformers for 3D Medical Image Segmentation”. In: *arXiv preprint arXiv:2103.10504* (2021). URL: https://arxiv.org/abs/2103.10504.

[23] Lina Huang et al. *Segmenting Medical Images: From UNet to Res-UNet and nnUNet*. 2024. arXiv: 2407.04353 [eess.IV]. URL: https://arxiv.org/abs/2407.04353.

[24] D. Jha et al. “ResUNet++: An advanced architecture for medical image segmentation”. In: *IEEE Access* 7 (2019), pp. 118530–118540. DOI: 10.1109/ACCESS.2019.2933937.

[25] Debesh Jha et al. *Kvasir-SEG: A Segmented Polyp Dataset*. 2019. arXiv: 1911.07069 [eess.IV]. URL: https://arxiv.org/abs/1911.07069.

[26] Debesh Jha et al. “Real-time polyp detection, localization and segmentation in colonoscopy using deep learning”. en. In: *IEEE Access* 9 (Mar. 2021), pp. 40496–40510.

[27] Wang Jiangtao, Nur Intan Raihana Ruhaiyem, and Fu Panpan. *A Comprehensive Review of U-Net and Its Variants: Advances and Applications in Medical Image Segmentation*. 2025. arXiv: 2502.06895 [eess.IV]. URL: https://arxiv.org/abs/2502.06895.

[28] Garrett G R J Johnson et al. “Colorectal polyp classification and management of complex polyps for surgeon endoscopists”. en. In: *Can. J. Surg.* 66.5 (Sept. 2023), E491–E498.

[29] Kunal Kawadkar. *Comparative Analysis of Vision Transformers and Convolutional Neural Networks for Medical Image Classification*. 2025. arXiv: 2507.21156 [eess.IV]. URL: https://arxiv.org/abs/2507.21156.

[30] Oz Kilim et al. “Physical imaging parameter variation drives domain shift”. In: *Scientific Reports* 12.1 (Dec. 2022), p. 21302.

[31] Ji Woong Kim, Aisha Urooj Khan, and Imon Banerjee. “Systematic review of hybrid vision transformer architectures for radiological image analysis”. en. In: *J. Imaging Inform. Med.* (Jan. 2025).

[32] Nam Hee Kim et al. “Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies”. en. In: *Intest. Res.* 15.3 (July 2017), pp. 411–418.

[33] Yong Bae Kim. “Smartphone-based polyp detection: a first step towards an open-source AI framework”. In: *Journal of Medical Artificial Intelligence* 8.0 (2025). ISSN: 2617-2496. URL: https://jmai.amegroups.org/article/view/9824.

[34] Shin-Ei Kudo et al. “Artificial intelligence and computer-aided diagnosis for colonoscopy: where do we stand now?” en. In: *Transl. Gastroenterol. Hepatol.* 6 (Oct. 2021), p. 64.

[35] Jun Li et al. “Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives”. en. In: *Med. Image Anal.* 85.102762 (Apr. 2023), p. 102762.

[36] Yabei Li et al. “Mechanisms and Applications of Attention in Medical Image Segmentation: A Review: Subtitle Is Not Required, Please Write It Here If Your Article Has One”. In: *Academic Journal of Science and Technology* 5.3 (May 2023), pp. 237–243. DOI: 10.54097/ajst.v5i3.8021. URL: https://drpress.org/ojs/index.php/ajst/article/view/8021.

[37] P Lijin et al. “PolySegNet: improving polyp segmentation through swin transformer and vision transformer fusion”. en. In: *Biomed. Eng. Lett.* 14.6 (Nov. 2024), pp. 1421–1431.

[38] Akis Linardos et al. “Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease”. en. In: *Sci. Rep.* 12.1 (Mar. 2022), p. 3551.

[39] Takahisa Matsuda, Ai Fujimoto, and Yoshinori Igarashi. “Colorectal cancer: Epidemiology, risk factors, and public health strategies”. en. In: *Digestion* 106.2 (Feb. 2025), pp. 91–99.

[40] Eileen Morgan et al. “Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN”. en. In: *Gut* 72.2 (Feb. 2023), pp. 338–344.

[41] Aminu Musa, Rajesh Prasad, and Monica Hernandez. “Addressing cross-population domain shift in chest X-ray classification through supervised adversarial domain adaptation”. In: *Scientific Reports* 15.1 (Apr. 2025), p. 11383.

[42] Andriy Myronenko. “3D MRI brain tumor segmentation using autoencoder regularization”. In: *CoRR* abs/1810.11654 (2018). arXiv: 1810.11654. URL: http://arxiv.org/abs/1810.11654.

[43] Loris Nanni et al. “Ensembles of convolutional neural networks and transformers for polyp segmentation”. en. In: *Sensors (Basel)* 23.10 (May 2023).

[44] Quang Vinh Nguyen et al. *Polyp-SES: Automatic Polyp Segmentation with Self-Enriched Semantic Model*. 2024. arXiv: 2410.01210 [cs.CV]. URL: https://arxiv.org/abs/2410.01210.

[45] Ozan Oktay et al. “Attention U-Net: Learning Where to Look for the Pancreas”. In: (2018). URL: https://arxiv.org/abs/1804.03999.

[46] Susan Y Quan et al. “Clinical evaluation of a real-time artificial intelligence-based polyp detection system: a US multi-center pilot study”. en. In: *Sci. Rep.* 12.1 (Apr. 2022), p. 6598.

[47] Douglas K Rex. “Colonoscopy remains an important option for primary screening for colorectal cancer”. en. In: *Dig. Dis. Sci.* 70.5 (May 2025), pp. 1595–1605.

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional networks for biomedical image segmentation”. In: (2015), pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28. URL: https://arxiv.org/abs/1505.04597.

[49] Ming-Hung Shen et al. “Deep learning empowers endoscopic detection and polyps classification: A multiple-hospital study”. en. In: *Diagnostics (Basel)* 13.8 (Apr. 2023).

[50] Syed Waseem Abbas Sherazi, Jang-Whan Bae, and Jong Yun Lee. “A soft voting ensemble classifier for early prediction and diagnosis of occurrences of major adverse cardiovascular events for STEMI and NSTEMI during 2-year follow-up in patients with acute coronary syndrome”. en. In: *PLoS One* 16.6 (June 2021), e0249338.

[51] J. Silva et al. “Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer”. In: *International Journal of Computer Assisted Radiology and Surgery* 9.2 (Mar. 2014). Epub 2013 Sep 15, pp. 283–293. DOI: 10.1007/s11548-013-0926-3.

[52] Marco Spadaccini et al. “AI and polyp detection during colonoscopy”. en. In: *Cancers (Basel)* 17.5 (Feb. 2025).

[53] Swetha Kumari T and Vasuki R. “Optimized Computer Vision Model for Accurate Polyp Detection in Endoscopic Procedures”. In: *International Research Journal of Multidisciplinary Technovation* 7.3 (May 2025), pp. 134–147. DOI: 10.54392/irjmt25312. URL: https://journals.asianresassoc.org/index.php/irjmt/article/view/3542.

[54] Satoshi Takahashi et al. “Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review”. en. In: *J. Med. Syst.* 48.1 (Sept. 2024), p. 84.

[55] Mingxing Tan and Quoc V. Le. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. 2020. arXiv: 1905.11946 [cs.LG]. URL: https://arxiv.org/abs/1905.11946.

[56] Yucheng Tang et al. *Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis*. 2022. arXiv: 2111.14791 [cs.CV]. URL: https://arxiv.org/abs/2111.14791.

[57] Ioannis Vezakis et al. “EffiSegNet: Efficient Segmentation Network for Gastrointestinal Polyp Segmentation”. In: *Journal of Medical Imaging* (2024). DOI: 10.1117/1.JMI.XX.XXXXXX.

[58] Frauke Wilm et al. *Rethinking U-net Skip Connections for Biomedical Image Segmentation*. 2024. arXiv: 2402.08276 [eess.IV]. URL: https://arxiv.org/abs/2402.08276.

[59] Leyi Xiao et al. “Enhanced medical image segmentation using U-Net with residual connections and dual attention mechanism”. In: *Engineering Applications of Artificial Intelligence* 153 (2025), p. 110794. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2025.110794. URL: https://www.sciencedirect.com/science/article/pii/S0952197625007948.

[60] Cun Xu et al. “Dual ensemble system for polyp segmentation with submodels adaptive selection ensemble”. In: *Scientific Reports* 14.1 (Mar. 2024), p. 6152.

[61] Wanni Xu, You-Lei Fu, and Dongmei Zhu. “ResNet and its application to medical image processing: Research progress and challenges”. en. In: *Comput. Methods Programs Biomed.* 240.107660 (Oct. 2023), p. 107660.

[62] He Xue et al. “A lighter hybrid feature fusion framework for polyp segmentation”. In: *Scientific Reports* 14.1 (Oct. 2024), p. 23179.

[63] Wenjian Yao et al. “From CNN to transformer: A review of medical image segmentation models”. en. In: *J. Imaging Inform. Med.* 37.4 (Aug. 2024), pp. 1529–1547.