A Survey on Deep Learning for Polyp Segmentation: Techniques, Challenges and Future Trends

Jiaxin Mei, Tao Zhou, Kaiwen Huang, Yizhe Zhang, Yi Zhou, Ye Wu, Huazhu Fu, *Senior Member, IEEE*

Abstract—**Early detection and assessment of polyps play a crucial role in the prevention and treatment of colorectal cancer (CRC). Polyp segmentation provides an effective solution to assist clinicians in accurately locating and segmenting polyp regions. In the past, people often relied on manually extracted lower-level features such as color, texture, and shape, which often had issues capturing global context and lacked robustness to complex scenarios. With the advent of deep learning, more and more outstanding medical image segmentation algorithms based on deep learning networks have emerged, making significant progress in this field. This paper provides a comprehensive review of polyp segmentation algorithms. We first review some traditional algorithms based on manually extracted features and deep segmentation algorithms, then detail benchmark datasets related to the topic. Specifically, we carry out a comprehensive evaluation of recent deep learning models and results based on polyp sizes, considering the pain points of research topics and differences in network structures. Finally, we discuss the challenges of polyp segmentation and future trends in this field. The models, benchmark datasets, and source code links we collected are all published at https://github.com/taozh2017/Awesome-Polyp-Segmentation.**

Index Terms—**Polyp Segmentation, Deep Learning, Comprehensive Evaluation, Medical Imaging.**

I. INTRODUCTION

POLYP segmentation is a crucial task in medical image analysis that aims to automatically identify and segment polyp regions within the colon. Its primary objective is to assist clinical doctors in efficiently and accurately locating and delineating these regions, providing vital support for early diagnosis and treatment of colorectal cancer (CRC) [1]. Polyps exhibit varying sizes and shapes at different stages of development [2], and their precise segmentation poses challenges due to their strong adherence to adjacent organs or mucosa [3]. Despite significant progress made in the domain of polyp segmentation, it still faces several challenges, such as limited annotated data, unclear boundaries, complex foregrounds, and real-time demand [4]–[7].

In the early stage, polyp segmentation primarily relied on manually extracted features [16], [17]. For example, Tajbakhsh *et al.* [16] proposed a method that utilized shape features and surrounding environmental information to automatically detect polyps in colonoscopy videos. Iwahori *et al.* [17] utilized edge and color information to generate a likelihood map, extracted the Directional Gradient Histogram

J. Mei, T. Zhou, K. Huang, Y. Zhang and Y. Wu are with the PCA Lab, and the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China. Y. Zhou is with the School of Computer Science and Engineering, Southeast University, Nanjing, China. H. Fu is with the Institute of High Performance Computing, A*STAR, Singapore. Corresponding author: *Tao Zhou* (e-mail: taozhou.ai@gmail.com).
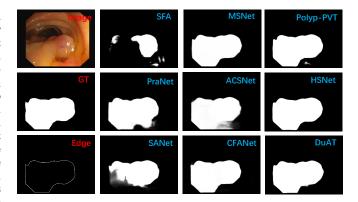

Fig. 1. Comparison of polyp segmentation results on some sample images using six CNN-based models (*i.e.*, SFA [8], PraNet [9], SANet [10], MSNet [11], ACSNet [12], and CFA-Net [7]) and three transformer-based models (*i.e.*, Polyp-PVT [13], HSNet [14], and DuAT [15]).

(DGH) features, and applied a random forest classifier to classify the detected regions as polyp areas or not. However, relying on manually extracted low-level features for segmentation tasks makes it difficult to handle complex scenarios and does not effectively utilize global contextual information [18].

In recent years, deep learning-based polyp segmentation models have made significant advancements, showcasing impressive capabilities in locating and segmenting polyp regions. For instance, Wei *et al.* [10] proposed a Shallow Attention Network, which leverages low-level features to mitigate degradation caused by multiple downsampling. They also proposed an innovative color-swapping method to reduce color dependency by exchanging color statistical data. To address the issue of blurry segmentation edges, Zhao *et al.* [11] proposed a Multi-Scale Subtraction Network, incorporating a subtraction unit to extract differential features between adjacent levels in the encoder. This network assigns different receptive fields to various levels of these units in a pyramid-like fashion, enabling the extraction of rich and diverse multi-scale differential information. To handle variations in sizes, Zhou *et al.* [7] presented a Cross-Level Feature Aggregation Network, employing a dual-stream structure-based segmentation network and a layer-by-layer fusion strategy for effective handling of scale variations and integration of high-level semantic information with low-level features. Rahman *et al.* [31] proposed a Cascade Attention Decoder, which effectively addresses the issue of inconsistent feature sizes by a hierarchical structure of transformers and attention-based convolutional modules to aggregate multi-level features and capture global and local contexts. Considering variations across datasets collected from different devices, Yang
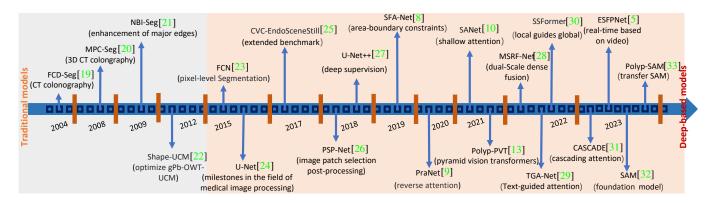
Fig. 2. A brief chronology of polyp segmentation. Methods before 2015 were based on handcrafted features combined with machine learning algorithms. The development of U-Net [24] and FCN [23] since 2015 has greatly propelled the advancement of deep learning techniques in polyp segmentation. More details can be found in Sec. II.

et al. [34] presented a mutual-prototype adaptation model to reduce domain shifts in multi-centers and multi-devices colonoscopy datasets. Furthermore, Jha et al. [4] proposed a Transformer-based Residual Network, which demonstrates strong generalization ability in multi-center external testing. In addition to innovative network architectures, refined models based on general segmentation networks have demonstrated promising results in polyp segmentation. For example, Li et al. [33] fine-tuned the Polyp-SAM model based on the Segment Anything Model (SAM) [32], which performs well in polyp segmentation tasks. To provide a clearer depiction of the progress in polyp segmentation tasks, we show a concise timeline in Fig. 2.

In this paper, we present a comprehensive study and survey of polyp segmentation methods. This survey first reviews existing polyp segmentation methods from different perspectives. Additionally, we conduct a thorough evaluation of various representative polyp segmentation models and analyze their respective advantages. Furthermore, we discuss future challenges and potential research directions.

A. Related Reviews and Surveys

There have been several recent investigations and reviews closely related to the field of polyp segmentation. For instance, Gupta et al. [35] reviewed deep-learning-based methods for efficient colorectal cancer screening, specifically focusing on polyp segmentation. They highlighted the importance of addressing challenges related to data scarcity, transfer learning, and interpretability in future research. Sanchez et al. [36] performed a systematic review of 35 studies since 2015 that utilized deep learning techniques for polyp detection, localization, and segmentation. Furthermore, Xiao et al. [37] classified and reviewed the segmentation methods involving UNet-based Transformers and other model-based Transformers in medical images. Their work summarized Transformer-based segmentation models applied to various anatomical regions such as abdominal organs, heart, brain, and lungs, based on relevant studies in the past two years.

There have been some studies that reviewed the achievements in medical image processing. Qureshi et al. [38] conducted a survey on the latest developments in medical image segmentation techniques, focusing on computational image processing and machine learning methods. They examined the contributions of different architectures to medical image segmentation, discussed the advantages, identified

open challenges, and highlighted potential future directions. Chowdhary et al. [39] investigated various segmentation and feature extraction methods used for preprocessing in medical images. Liu et al. [40] provided a comprehensive review of U-Net architectures in medical image segmentation tasks. They focused on the architecture, expansion mechanism, and application fields of U-Net architectures. Thisanke et al. [41] discussed different Vision Transformer (ViT) architectures applicable to semantic segmentation tasks and analyzed how their evolution has contributed to dense prediction tasks. Their investigation aimed to review and compare the performance of ViT architectures designed for semantic segmentation using benchmark datasets. Bennai et al. [42] conducted a comparative study of recently published multi-agent methods dedicated to medical image segmentation. Their work aimed to provide insights into the performance and effectiveness of these methods in the context of medical image analysis. Overall, these studies contribute to the understanding and advancement of medical image processing and segmentation techniques, highlighting the progress made, existing challenges, and future research directions in this field.

In contrast to previous reviews on polyp segmentation or medical image segmentation, this paper aims to provide a comprehensive review of polyp segmentation methods. It covers both traditional algorithms and deep learning model-based approaches for polyp segmentation. The paper systematically and comprehensively analyzes the strengths, limitations, and outcomes of these methods.

B. Contributions

Our main contributions can be summarized as follows:

- We provide a systematic review of polyp segmentation models from different perspectives. Our study encompasses a thorough examination of both deep learning-based and traditional methods, boundary-aware models, attention-aware models, and feature fusion models. We discuss the strengths associated with each methodological approach, offering valuable insights into their respective applicability and performance characteristics.
- We have conducted an analysis of five publicly available colonoscopy image datasets specifically curated for polyp segmentation. Additionally, we present a comprehensive overview of commonly used evaluation metrics employed in polyp segmentation tasks.

- Moreover, we provide a comprehensive as well as a scale-based evaluation of several representative polyp segmentation models.
- We delve deeply into the numerous challenges encountered in polyp segmentation and offer a comprehensive analysis of each one. Furthermore, we propose potential directions for future research that can help tackle these challenges and advance the field.

C. Organization

The structure of this review is organized as follows. In Sec. II, we review the existing polyp segmentation models from different aspects. In Sec. III, we summarize and provide detailed information and usage of the current publicly available datasets used for polyp segmentation. Then, we conduct a comprehensive assessment of the segmentation performance for polyp sizes and an analysis of the advantages and disadvantages of several representative polyp segmentation models n Sec. IV. After that, in Sec. V, we discuss the challenges and future trends for development in this field. Finally, we conclude the paper in Sec. VI.

II. POLYP SEGMENTATION MODELS

Over the past decade, significant efforts have been made to enhance the performance of automatic polyp segmentation models, improving the detection capability of colonoscopy, and reducing disease risks. Recent research on polyp segmentation has leveraged popular deep-learning methods to achieve noteworthy outcomes. In contrast, earlier approaches relied primarily on manually engineered features for polyp segmentation. A summary of these models can be found in Table I,II. To provide a comprehensive review of these polyp segmentation algorithms, we will introduce them from the following aspects.

(1) **Traditional Models**: They primarily rely on manually designed features, such as color, texture, and shape information, to formulate for algorithm design. (2) **Deep models**: Deep learning models automatically learn deep features, capable of handling more complex structures, and offering stronger expression. (3) **Boundary-aware models**: Edge information is crucial in providing boundary cues to boost the segmentation performance, therefore we will discuss the application of edge information in some existing models. (4) **Attention-aware models**: attention mechanisms have widely been applied in various visual tasks. We conduct a comprehensive review of related works on polyp segmentation to analyze different attention strategies. This analysis provides insights into the potential design of attention modules in future works. (5) **Feature fusion models**: The integration and utilization of multi-level features often contribute significantly to enhancing model performance. Therefore, we investigate the effectiveness of feature fusion strategies in polyp segmentation models.

A. Traditional Models

Early works primarily relied on manually designed features, such as color, texture, and shape, and then employed traditional machine learning techniques for heuristic modeling. For instance, Yao *et al.* [19] proposed an automatic polyp segmentation method that combines knowledge-guided intensity adjustment, fuzzy c-means clustering, and deformable models. Lu *et al.* [20] proposed a three-stage probabilistic binary classification method that integrates low-level and mid-level information to segment polyps in 3D CT colonography. Gross *et al.* [21] studied a segmentation algorithm that enhances primary edges through multiscale filtering. Ganz *et al.* [22] utilized prior knowledge of polyp shapes using NBI Narrow Band Imaging (NBI) to optimize the inherent scale selection problem of gPb-OWT-UCM. This optimization aims to achieve better segmentation results by incorporating the shape information of polyps.

B. Deep Models

However, the aforementioned methods are constrained by the limited information representation capabilities of manual features. They lack generalization capabilities and are not suitable for large-scale deployment. As a result, there is a growing reliance on deep features to handle these limitations. We will review some representative deep learning-based methods in the field of polyp segmentation.

1) CNN-based Methods. Thanks to the development of Convolutional Neural Networks (CNN), especially with the introduction of U-Net [24], many models inspired by this architecture have demonstrated promising results.

- **ACSNet** [12] modifies the skip connections in U-Net into local context extraction modules and adds a global information extraction module. The features are integrated and then adaptively selected based on a channel attention strategy.
- **EU-Net** [48] is an enhanced U-Net framework that enhances semantic information and introduces an adaptive global context module to extract key features. It improves the quality of features at each layer, thereby enhancing the final segmentation performance.
- **MSNet** [11] designs a subtraction unit to generate difference features between adjacent layers and pyramidically equips it with different receptive fields to capture multi-scale information. In addition, it introduces LossNet to supervise the perceptual features at each layer.
- **PEFNet** [81] utilizes an improved U-Net in the merging phase and embeds new location feature information. Thanks to the rich knowledge of positional information and concatenated features, this model achieves higher accuracy and universality in polyp segmentation.

2) Transformer-based Methods. Although CNNs have been successful in various computer vision tasks, they have limitations in capturing long-range dependencies. However, the introduction of Transformer models, originally popular in natural language processing, has revolutionized the field of computer vision. Specifically, Vision Transformers (ViTs) [82] have emerged as a powerful approach for image understanding and have contributed to the development of numerous algorithms that leverage their strengths.

- **MSRAformer** [59]. It adopts a Swin Transformer as the encoder with a pyramid structure to extract features at different stages and utilizes a multi-scale channel attention module to extract multi-scale feature information. It also adds a spatial reverse attention module to supplement edge structure and detail information, thereby exhibiting strong generalization capability and performance.

TABLE I
SUMMARY OF POLYP SEGMENTATION METHODS (PUBLISHED FROM 2019 TO 2021).

#	Year	Method	Pub.	Backbone	Description	Code
1	2019	SFA [8]	MICCAI	light UNet	Boundary-sensitive loss; Selective feature aggregation	link
2	2019	ResUNet++ [43]	ISM	ResUNet	Squeeze and excitation blocks; ASPP; Attention blocks	link
3	2020	PolypSeg [44]	MICCAI	UNet	Improved attention mechanism; Separable convolution	N/A
4	2020	ThresholdNet [1]	TMI	DeepLabv3+	Confidence-guided manifold mixup; Threshold loss	link
5	2020	ACSNet [12]	MICCAI	ResNet34	Adaptively select; Aggregate context features through channel attention	link
6	2020	PraNet [9]	MICCAI	Res2Net	Parallel partial decoders; Reverse attention	link
7	2021	DDANet [45]	PR	ResUNet	Dual-decoder attention; Out-of-training-set testing.	link
8	2021	GMSRF-Net [46]	ICPR	ResNet50	Cross-multi-scale attention; Multi-scale feature selection	link
9	2021	HarDNet-MSEG [47]	Arxiv	HarDNet68	Cascaded partial decoder; Dense aggregation	link
10	2021	EU-Net [48]	CRV	ResNet34	Semantic feature; Adaptive global context module	link
11	2021	FANet [49]	TNNLS	N/A	Feedback attention learning; Iterative refining; Embedded run-length encoding strategy	link
12	2021	Polyp-PVT [13]	CAAI AIR	PVT	Cascaded fusion; Camouflage identification; Similarity aggregation	link
13	2021	UACANet [50]	ACM MM	Res2Net	Parallel axial attention; Uncertainty augmented context attention	link
14	2021	C2FNet [51]	IJCAI	Res2Net-50	Context-aware cross-level fusion; Dual-branch global context	link
15	2021	ResUNet++ + TTA + CRF [52]	JBHI	ResUNet	Conditional random fields; Test-time augmentation	link
16	2021	MPA-DA [34]	JBHI	ResNet-101	Inter-prototype adaptation network; Progressive self-training; Disentangled reconstruction	link
17	2021	TransFuse [53]	MICCAI	ResNet-34 + DeiT-S	Self-attention; Bilinear Hadamard product; Gated skip-connection	link
18	2021	SANet [10]	MICCAI	Res2Net	Color exchange; Shallow attention	link
19	2021	STFT [54]	MICCAI	ResNet-50	Spatial-temporal feature transformation; Deformable convolutions and channel-aware attention	link
20	2021	LOD-Net [55]	MICCAI	ResNet + FPN	Model the probability of each pixel locating in border region; Adaptive thresholding policy	link
21	2021	MSNet [11]	MICCAI	Res2Net-50	Multi-scale subtraction; Training-free network	link
22	2021	CCBANet [56]	MICCAI	ResNet34	Cascading context; Balancing attention	link
23	2021	HRENet [57]	MICCAI	ResNet34	Hard region enhancement; Adaptive feature aggregation; Edge and structure consistency aware loss	link

• **DuAT** [15] is a dual-aggregation transformer network for polyp segmentation. It includes a global-to-local spatial aggregation module for aggregating global and local spatial features and locating multi-scale objects. This method also employs a selective boundary aggregation module to integrate low-level edge features and high-level semantic features.

• **SSFormer** [30] incorporates PVTv2 and Segformer as the encoder while introducing a novel progressive local decoder. This decoder is specifically designed to complement the pyramid Transformer backbone by emphasizing local features and mitigating attention dispersion.

• **ColonFormer** [64] adopts a lightweight architecture based on the transformer as the encoder and uses a hierarchical network structure for learning multi-level features in the decoder. Additionally, the model incorporates a novel skip connection technique that refines polyp boundary information, resulting in precise segmentation outcomes.

• **TransNetR** [4] is a transformer-based residual network, comprising a pre-trained encoder, three decoder blocks, and an upsampling layer, demonstrating excellent real-time processing speed and multi-center generalization capability.

• **Polyp-PVT** [13] comprises a cascaded fusion module that combines high-level semantic and position information, a camouflage recognition module that captures low-level features, and a similarity aggregation module that extends high-level features throughout the entire region. This integrated approach effectively mitigates noise in the features and yields substantial improvements in polyp segmentation performance.

3) Hybrid Methods. Furthermore, numerous models have combined the strengths of both CNN and Transformer, capturing both local context information and long-range dependencies to significantly enhance segmentation performance.

• **TransFuse** [53] combines Transformer and CNN in a parallel manner to capture global dependencies and low-level spatial details. When fusing the multi-level features from the two branches, it incorporates a self-attention mechanism and a multi-modal fusion mechanism. Moreover, spatial attention is used to enhance local details and suppress irrelevant regions while modeling fine-grained interactions between the two branches to suppress noise in low-level features.

• **LAPFormer** [67] employs a hierarchical transformer encoder to extract global features and combines it with a CNN decoder to capture the local appearance of polyps. Additionally, it introduces a progressive feature fusion module to integrate multi-scale features and incorporates a feature refinement module and a feature selection module for feature handling.

• **PPFormer** [68] adopts a shallow CNN encoder and deep Transformer-based encoder to extract features. It then uses prediction maps to guide self-attention for enhanced boundary perception.

• **HSNet** [14] utilizes a dual-branch structure composed of Transformer and CNN networks to capture both long-range dependencies and local appearance details. Additionally, an interaction mechanism is incorporated to facilitate the exchange of semantic information among different network

TABLE II
SUMMARY OF POLYP SEGMENTATION METHODS (PUBLISHED FROM 2022 TO 2023).

#	Year	Method	Pub.	Backbone	Description	Code
24	2022	MSRF-Net [28]	JBHI	N/A	Multi-Scale residual fusion; Dual-scale dense fusion	link
25	2022	TGANet [29]	MICCAI	ResNet50	Text-guided attention; Weights the text-based embeddings	link
26	2022	PolypSeg+ [58]	TCYB	ResNet50	Adaptive scale context module; Lightweight attention mechanism	link
27	2022	MSRAformer [59]	CBM	Swin Transformer	Multiscale spatial reverse attention	link
28	2022	HSNet [14]	CBM	PVTv2	Cross-semantic attention; Hybrid semantic complementary; Multi-scale prediction	link
29	2022	FuzzyNet [60]	NeurIPS	Res2Net/ ConvNext/ PVT	Fuzzy attention; Focus on the blurry pixels	link
30	2022	LDNet [61]	MICCAI	Res2Net	Lesion-aware dynamic kernel; Self-attention	link
31	2022	HarDNet-DFUS [62]	Arxiv	HarDNetV2	Real-time model; Enhanced backbone	link
32	2022	BDG-Net [63]	SPIE MI	EfficientNet-B5	Boundary distribution guided; Boundary distribution generate	link
33	2022	ColonFormer [64]	Access	MiT	Integrate a hierarchical Transformer and a hierarchical pyramid CNN; Residual axial attention	link
34	2022	FCBFormer [65]	MIUA	PVTv2	Improved progressive locality decoder; Fully convolutional branch + Transformer bramch	link
35	2022	DCRNet [66]	ISBI	ResNet-34	Duplex contextual relation network; Cross-image contextual relations	link
36	2022	SSFormer [30]	MICCAI	PVTv2	Aggregate local and global features stepwise	link
37	2022	DuAT [15]	PRCV	PVT	Dual-aggregation transformer; Global-to-local spatial aggregation; Selective boundary aggregation	link
38	2022	LAPFormer [67]	Arxiv	MiT-B1	Hierarchical transformer encoder and CNN decoder; Progressive feature fusion	N/A
39	2022	PPFormer [68]	MICCAI	CvT	Shallow CNN encoder; Deep Transformer-based encoder	N/A
40	2022	BSCA-Net [69]	PR	Res2Net	Bit-plane slicing information; Segmentation squeeze bottleneck union module; Multi-Path Connection Attention	N/A
41	2022	BoxPolyp [70]	MICCAI	Res2Net/ PVT	Box annotations; Fusion filter sampling module	N/A
42	2022	ICBNet [71]	BIBM	PVT	Iterative feedback learning strategy; Context and boundary-aware information	N/A
43	2022	CLD-Net [72]	BIBM	MiT	Small polyp segmentation; Local edge feature extraction	N/A
44	2022	BANet [73]	PRCV	Res2Net-50	Attention-aware localization; Residual pyramid convolution	N/A
45	2022	CaraNet [74]	JMI	Res2Net	Context axial reverse attention	link
46	2023	APCNet [75]	TIM	ResNet50	Attention-guided multi-level aggregation strategy; Complementary information from different layers	N/A
47	2023	RA-DENet [76]	CBM	Res2Net	Improved reverse attention; Distraction elimination	N/A
48	2023	EFB-Seg [77]	Neurocomputing	ConvNet	Boundary Embedding; Semantic offset field learned	N/A
49	2023	PPNet [78]	CBM	P2T	Channel attention; Pyramid feature fusion	N/A
50	2023	Fu-TransHNet [6]	Arxiv	HardNet68	CNN and Transformer; Multi-view learning	N/A
51	2023	DilatedSegNet [79]	MMM	ResNet50	Dilated convolution pooling block; Convolutional attention	link
52	2023	FeDNet [80]	BSPC	PVT	Decouple edge features and main body features	link
53	2023	PEFNet [81]	MMM	EfficientNet V2-L	Positional encoding and information fusion	link
54	2023	Polyp-SAM [33]	Arxiv	ViT	Finetuned SAM model for polyp segmentation [32].	link
55	2023	ESFPNet [5]	SPIE MI	MiT	Efficient stage-wise feature pyramid decoder	link
56	2023	TransNetR [4]	MIDL	ResNet50	Transformer-based residual network; Multi-center out-of-distribution testing	link
57	2023	CASCADE [31]	WACV	PVTv2/ TransUNet	Cascaded attention-based decoder; Multi-stage loss optimization; Feature aggregation	link
58	2023	CFA-Net [7]	PR	Res2Net-50	Cross-level feature aggregated; Boundary aggregated	link

layers, bridging the gap between low-level and high-level features. This design enhances the model's capability to incorporate comprehensive information and improve overall performance.

• **Fu-TransHNet** [6] designs a novel feature fusion module to fully make use of the local and global features obtained from CNN and Transformer networks. The fusion module enables dense fusion of features at the same scale and multiple scales. Furthermore, weights for the CNN and Transformer branches and the fusion module are obtained through multi-view learning, resulting in flexibility to achieve optimal performance.

C. Boundary-aware Models

The utilization of edge information as a guiding strategy was initially prevalent in object detection and has been increasingly applied to polyp segmentation tasks. Given the specific nature of medical image segmentation, precise edge-aware representation provides particular significance. In the following sections, we will review several models that exhibit exceptional capability in perceiving and incorporating edge information.

• **FeDNet** [80] simultaneously optimizes the main body and edges to improve polyp segmentation performance. It explicitly decouples the input features into body features and edge features and then conducts targeted optimization by

introducing a feature decoupling module.

• **BSCA-Net** [69] utilizes bit-plane slicing information to effectively extract boundary information. Additionally, a Segmentation Squeeze Bottleneck Union module is designed to utilize geometric information from different perspectives, and Multi-Path Connection Attention Decoders and Multi-Path Attention Connection Encoders are used to further enhance the network performance for polyp segmentation.

• **BoxPolyp** [70] mitigates overfitting issues by using box annotations, iteratively enhancing the segmentation model to generate fine-grained polyp regions. A Fusion Filter Sampling module is designed to generate pixel-level pseudo labels from less noisy box annotations, significantly improving performance.

• **BDG-Net** [63] is a Boundary Distribution Guided Network. It utilizes a boundary distribution generation module to aggregate high-level features, which are taken as supplementary spatial information and fed to the Boundary Distribution Guided Decoder (BDGD) for guiding polyp segmentation. Additionally, the BDGD adopts a multi-scale feature interaction strategy to cope with size variations.

• **ICBNet** [71] innovatively adopts an iterative feedback learning strategy, supplementing and perfecting encoder features from preliminary segmentation and boundary predictions using context and boundary-aware information. This strategy is iteratively employed to achieve progressive optimization improvements. Furthermore, a dual-branch iterative feedback unit is developed to enhance features under the guidance of segmentation and boundary prediction.

• **CLD-Net** [72] focuses on addressing the issue of feature loss during downsampling and effectively tackles the challenge of small polyp segmentation. It achieves this by incorporating a local edge feature extraction module and a local-global feature fusion module. The model initially extracts a series of edge features using a progressive strategy, subsequently handles noise, and finally integrates the edge features into the global features through an upsampling fusion strategy.

• **BANet** [73] accurately identifies the main location of polyps through an attention-aware localization module. In addition, it mines polyp boundary information through a residual pyramid convolution module and leverages boundary information for constrained polyp region prediction via a boundary-guided refinement module, thereby achieving more accurate segmentation.

• **SFA** [8] improves segmentation performance by constructing a Selective Feature Aggregation Network with region and boundary constraints. It primarily consists of a shared encoder that predicts the polyp region and two mutually constrained decoders that extract edge information. By introducing three upwardly cascaded components between the encoder and decoder and embedding selective kernel modules into the convolutional layers, it achieves selective feature aggregation. Furthermore, a novel boundary-sensitive loss function is innovatively proposed to measure the dependency between regions and boundaries.

• **FCBFormer** [65] fully leverages the strengths of fully convolutional networks (FCNs) and transformers for polyp segmentation.

D. Attention-aware Models

In the field of polyp segmentation, achieving superior performance requires the ability to prioritize relevant information rather than treating all information equally. By incorporating an attention mechanism, this issue can be effectively addressed, allowing the model to focus on the most important features and ultimately improving segmentation performance.

• **CASCADE** [31] takes advantage of the multi-scale features using a hierarchical visual transformer. It includes an attention gate that fuses skip-connection features and a convolutional attention module that suppresses background information to enhance distant and local contexts.

• **APCNet** [75] extracts multi-level features from a pyramid structure, then presents an attention-guided multi-level aggregation strategy, enhancing each layer's context features by leveraging complementary information from different layers.

• **RA-DENet** [76] enhances the representation of different regions through inverse attention, then eliminates noise through distractor removal. It extracts low-level polyp features to obtain edge features, and by connecting these edge features with refined polyp features, resulting in promising polyp segmentation performance.

• **TGANet** [29] incorporates a text attention mechanism, which utilizes features related to the size and number of polyps to adapt to varying polyp sizes and effectively handle scenarios with multiple polyps. By assigning weights to the text-based embeddings through an auxiliary classification task, the network can learn additional feature representations and enhance its overall performance.

• **LDNet** [61] extracts global context features from the input image and then updates them iteratively based on the lesion features predicted by the segmentation. Then, a self-attention module is presented to capture remote context relationships and improve segmentation performance. This model exhibits strong segmentation performance and generalization ability.

• **SANet** [10] eliminates the impact of colors through a color swap operation, then filters out background noise from shallow features based on a shallow attention module. Additionally, it addresses the pixel imbalance issue in small polyps using a probability correction strategy. Thanks to these measures, it performs well on small polyp tasks.

E. Feature Fusion Models

In the domain of semantic segmentation, incorporating multi-scale features is critical for effectively dealing with variations in object sizes. Furthermore, integrating multi-level features through feature fusion and leveraging both high-level and low-level features can greatly enhance segmentation performance. In the context of polyp segmentation tasks, several feature fusion strategies have been employed to achieve this objective.

• **CFA-Net** [7] is a novel cross-level feature aggregation network that adopts a hierarchical strategy to incorporate edge features into the dual-stream segmentation network. Additionally, the model proposes a cross-layer feature fusion module to integrate adjacent features from different levels.

• **EFB-Seg** [77] enhances multi-level feature fusion by introducing a feature fusion module that utilizes the learned

semantic offset field to align multi-level feature maps, thereby addressing the issue of feature misalignment.

- **MSRF-Net** [28]. This model innovatively uses a dual-scale dense fusion block to exchange multi-scale features with different receptive fields. It can preserve resolution, and propagate high-level and low-level features to achieve more accurate segmentation results.

- **DCRNet** [66] captures both intra-image and inter-image context relationships. Within the image, a positional attention module is presented to capture pixel-level context information. Between images, feature enhancement is achieved by embedding context relation matrices, then relationship fusion is implemented through region cross-batch memory.

- **PPNet** [78] uses a channel attention scheme in the pyramid feature fusion module to learn global context features, thereby guiding the information transformation of the decoder branches. Furthermore, it introduces a memory-retained pyramid pooling module in each side branch of the encoder to enhance feature extraction effectiveness.

- **PraNet** [9] aggregates high-level features using parallel partial decoders, uses a reverse attention module to mine boundary clues, and establishes the relationship between region and boundary cues.

- **PolypSeg** [44] aggregates multi-scale context information and focuses on the target area using an improved attention mechanism. It then eliminates background noise from low-level features, enhancing the feature fusion between high-level and low-level features. Furthermore, the model reduces computational cost using depthwise separable convolution.

F. Video Polyp Segmentation

Accurate and real-time polyp image segmentation presents significant challenges due to the dependence on annotation quality and the complexity of deep learning models. To facilitate the deployment of automated segmentation methods in clinical settings, a shift in focus has been observed in some studies toward video-based polyp segmentation methods. By considering temporal information, these approaches aim to overcome the limitations of single-image segmentation and enable more precise and efficient segmentation in real-time scenarios within clinical requirements.

- **ESFPNet** [5] constructs a pre-trained Mixed Transformer (MiT) encoder and an efficient stage-wise feature pyramid decoder, in which the MiT uses overlapping path merging modules and self-attention prediction, ultimately showing efficient performance and potential applicability in related fields.

- **PNS+** [83] extracts long-term spatiotemporal representations using global encoders and local encoders and refines them gradually with normalized self-attention blocks. They introduce a frame-by-frame annotated video polyp segmentation dataset called SUN-SEG, which consists of 158,690 colonoscopy video frames. Extensive experiments have shown that PNS+ exhibits the best performance and real-time inference speed.

- **SSTAN** [84] presents a semi-supervised video polyp segmentation task that only requires sparsely annotated frames for training. It introduces a novel spatio-temporal attention structure, consisting of temporal local context attention modules that refine the current prediction using predicted results
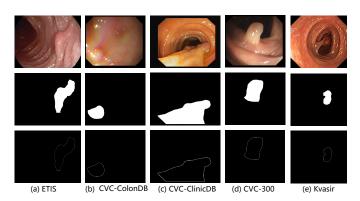

Fig. 3. Examples of images, ground truth maps, and edges in five polyp segmentation datasets, including (a) ETIS-LaribPolypDB [87], (b) CVC-ColonDB [16], (c) CVC-ClinicDB [18], (d) CVC-300 [25], and (e) Kvasir-SEG [88]. In each dataset, the image, ground truth maps, and edges are shown from top to bottom.

from nearby frames, and spatial-temporal attention modules that capture long-range dependencies in both time and space using hybrid transformers.

- **PNS-Net** [85] leverages standard self-attention modules and CNN to efficiently learn representations from polyp videos in a real-time manner without the need for post-processing.

- **NanoNet** [86] has fewer parameters and can be integrated with mobile and embedded devices. It utilizes a pre-trained MobileNetV2 as the encoder. In the architecture between the encoder and decoder, a modified residual block is incorporated to enhance the generalization capability of the decoder.

III. POLYP SEGMENTATION DATASETS

The rapid progress in the field of medical image segmentation has led to the construction of various public benchmark datasets specifically designed for polyp segmentation tasks. These datasets have emerged in recent years and serve as standardized evaluation platforms for assessing the performance of different segmentation models. By utilizing these benchmark datasets, researchers can compare their methods against established baselines, facilitate reproducible research, and foster further advancements in the field of polyp segmentation. Table III summarizes eight popular image-level polyp segmentation datasets, and Fig. 3 shows examples of images (including edge maps, and annotations) from these datasets. Moreover, we provide some details about each dataset below. Please note that some video-level polyp segmentation datasets are summarized in Table III, and some works [16], [18], [83], [89], [90] are based on them.

- **ETIS-LaribPolypDB** [87] collects early colorectal polyp images, consisting of 196 polyp instances of size 966×1225.

- **CVC-ClinicDB** [18] comes from clinical cases at a hospital in Barcelona, Spain, and is generated from 23 different standard white light colonoscopy intervention videos. This dataset contains 612 high-resolution color images of size 576×768, all derived from clinical colonoscopy examinations. Each image comes with a corresponding manual annotation file that clearly delineates the location of the polyp.

- **CVC-ColonDB** [16] is maintained by the Computer Vision Center (CVC) in Barcelona. It includes 380 colonoscopy images of size 500×574 and segmentation masks manually annotated to precisely mark the location of polyps.

- **CVC-300** [25] includes 60 colonoscopy images with a resolution of 500×574.
- **CVC-EndoSceneStill** [25] includes CVC-ClinicDB and CVC-300, thereby containing 912 colonoscopy examination images along with corresponding annotations.
- **Kvasir-SEG** [88] contains $1,000$ images of gastrointestinal polyps along with corresponding segmentation masks and bounding boxes. These were manually annotated by a doctor and validated by a gastroenterology expert.
- **PICCOLO** [91] consists of $3,433$ clinical colonoscopy images from 48 patients, including white light and narrowband imaging images. It also provides annotations including the number and size of polyps detected during colonoscopy. The data is divided into a training set $(2,203)$, a validation set (897), and a test set (333).
- **PolypGen** [92] originates from colonoscopy detection images of over 300 patients from six different centers. It includes both single-frame data and sequence data, containing $3,762$ annotated polyp labels. The delineation of polyp boundaries has been validated by six senior gastroenterology experts. Specifically, the dataset also contains $4,275$ negative samples.

IV. MODEL EVALUATION AND ANALYSIS

A. Evaluation Metrics

We briefly review several popular metrics for polyp segmentation task, *i.e.*, Intersection over Union (IoU), precision-recall (PR), specificity, Dice coefficient (Dice), F-measure (F_β) [94], mean absolute error (MAE) [95], structural measure (S_α) [96], and enhanced-alignment measure (E_ϕ) [97].

First, let's introduce some parameters. *TP* represents True Positives, which indicates the number of positive samples that the model correctly predicts as positive. *FP* represents False Positives, which represents the number of negative samples that the model incorrectly predicts as positive. *FN* represents False Negatives, which represents the number of positive samples that the model incorrectly predicts as negative.

- **PR**. Precision represents the proportion of correctly predicted positive samples by the model. Recall (also known as Sensitivity) represents the proportion of true positive samples successfully detected by the model out of all actual positive samples. PR (Precision-Recall) curve can be plotted by taking recall as the x-axis and precision as the y-axis.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \ \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (1)$$

- **IoU**. Intersection over Union, also known as the Jaccard coefficient, is used to measure the overlap between the predicted result and the ground truth target. It is commonly used in handling imbalanced datasets.

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \quad (2)$$

- **F-measure**(F_β). The F-measure, also known as the F1-score, combines precision and recall by considering their harmonic mean. It allows adjusting the weights of precision and recall to address specific problems and is effective in handling imbalanced datasets.

$$F_\beta = (1 + \beta^2) \frac{(\text{Precision} \cdot \text{Recall})}{(\beta^2 \cdot \text{Precision}) + \text{Recall}}. \quad (3)$$

- **Dice**. Dice coefficient is a commonly used evaluation metric for measuring the similarity between predicted segmentation results and the ground truth segmentation targets. It provides a score ranging from 0 to 1, where 1 indicates a perfect match and 0 indicates no match at all.

$$\text{Dice} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}. \quad (4)$$

- **S-measure** (S_α). Structure-measure [96] evaluates the structural similarity of segmentation results.

$$S_\alpha = \alpha \cdot S_O + (1 - \alpha) \cdot S_R. \quad (5)$$

In this context, S_O represents the Object Similarity, which measures the overlap between the segmentation result and the ground truth target. S_R represents the Region Similarity, which assesses the structural similarity between the segmentation result and the ground truth target. The parameter α is used to balance the weights between these two measures.

- **E-measure** (E_ϕ). Enhanced-measure [97] is used to evaluate region coverage.

$$E_\phi = \frac{(1 + \phi^2) \cdot S_O \cdot S_R}{\phi^2 \cdot S_O + S_R}, \quad (6)$$

where, the parameter ϕ is a balancing parameter that adjusts the weight ratio between accuracy and region coverage.

- **MAE**. Mean Absolute Error (MAE) is commonly used for evaluating regression tasks.

$$\text{MAE} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |S_{i,j} - G_{i,j}|, \quad (7)$$

where W and H denote the width and height of the map, S represents the predicted segmentation map, and G represents the ground truth segmentation map.

- **Specificity**. Specificity measures the model's ability to recognize negative samples. Specifically, Specificity can be calculated using the following formula:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \quad (8)$$

The value of Specificity ranges between 0 and 1, with a higher value indicating a stronger ability of the model to recognize negative samples.

B. Performance Comparison and Analysis

1) Overall Evaluation: To quantify the performance of different models, we conducted a comprehensive evaluation of 24 representative polyp segmentation models, including 1) eighteen CNN-based models: UNet [24], UNet++ [27], SFA [8], PraNet [9], ACSNet [12], MSEG [47], EU-Net [48], SANet [10], MSNet [11], UACANet-S [50], UACANet-L [50], C2FNet [51], DCRNet [66], BDG-Net [63], CaraNet [74], EFA-Net [98], CFA-Net [7], and M2SNet [99], and 2) six Transformer-based methods: Polyp-PVT [13], HSNet [14], DuAT [15], ESFPNet [5], FeDNet [80], and SAM-based poly segmentation model (with three different backbones, *i.e.*, SAM-B, SAM-H, and SAM-L) [100]. Firstly, we individually evaluated the performance of the aforementioned 24 models on Dice, IoU, S_α, F_β, E_ϕ, and MAE metrics across five public datasets (ETIS-LaribPolypDB [87], CVC-ColonDB [16], CVC-ClinicDB [18], CVC-300 [25], and

TABLE III
ACCORDING TO THE YEAR (YEAR), PUBLICATION JOURNAL (PUB.), DATASET SIZE (SIZE), NUMBER OF OBJECTS IN THE IMAGES (OBJ.), AND RESOLUTION (RESOLUTION), STATISTICS ON THE DATASETS ARE CONDUCTED. MORE DETAILED INFORMATION ABOUT EACH DATASET CAN BE FOUND IN SEC. III. THESE DATASETS CAN BE DOWNLOADED FROM OUR WEBSITE: HTTPS://GITHUB.COM/TAOZH2017/AWESOME-POLYP-SEGMENTATION.

	#	Dataset	Year	Pub.	Size	Obj.	Resolution
Image-level	1	**ETIS-LaribPolypDB** [87]	2014	IJCARS	196	Multiple	1225×966
	2	**CVC-ColonDB** [16]	2015	TMI	380	One	574×500
	3	**CVC-ClinicDB** [18]	2015	CMIG	612	Multiple	768×576
	4	**CVC-300** [25]	2017	JHE	60	One	574×500
	5	**CVC-EndoSceneStill** [25]	2017	JHE	912	Multiple	$[574 \sim 768] \times [500 \sim 576]$
	6	**Kvasir-SEG** [88]	2020	MMM	1,000	Multiple	$[487 \sim 1072] \times [332 \sim 1920]$
	7	**PICCOLO** [91]	2020	AS	3,433	Multiple	$[854 \sim 1920] \times [480 \sim 1080]$
	8	**PolypGen** [92]	2021	SD	8,037	Multiple	$[384 \sim 1920] \times [288 \sim 1080]$
Video-level	9	**ASU-Mayo Clinic** [16]	2016	TMI	36,458	One	688×550
	10	**CVC-ClinicVideoDB** [93]	2017	GIANA	11,954	Multiple	*N/A*
	11	**LDPolypVideo** [89]	2021	MICCAI	40,266	Multiple	560×480
	12	**SUN-SEG** [83]	2022	MIR	158,690	One	$[1158 \sim 1240] \times [1008 \sim 1080]$

TABLE IV
BENCHMARK RESULTS OF 24 REPRESENTATIVE POLYP SEGMENTATION MODELS (18 CNN-BASED AND 6 TRANSFORMER-BASED MODELS) ON FIVE COMMONLY USED DATASETS IN TERMS OF Dice, IoU, AND S_α. THE THREE BEST RESULTS ARE SHOWN IN RED, BLUE, AND GREEN FONTS.

Method	Pub.	ETIS-Larib			CVC-ColonDB			CVC-ClinicDB			CVC-300			Kvasir		
		Dice	IoU	S_α	Dice	IoU	S_α	Dice	IoU	S_α	Dice	IoU	S_α	Dice	IoU	S_α
UNet [24]	MICCAI 2015	.398	.335	.684	.504	.436	.710	.823	.755	.889	.710	.627	.843	.818	.746	.858
UNet++ [27]	MICCAI 2018	.401	.344	.683	.482	.408	.692	.794	.729	.873	.707	.624	.839	.821	.743	.862
SFA [8]	MICCAI 2018	.297	.217	.557	.456	.337	.628	.700	.607	.793	.467	.329	.640	.723	.611	.782
PraNet [9]	MICCAI 2020	.628	.567	.794	.712	.640	.820	.899	.849	.936	.871	.797	.925	.898	.840	.915
ACSNet [12]	MICCAI 2020	.578	.509	.754	.716	.649	.829	.882	.826	.927	.863	.787	.923	.898	.838	.920
MSEG [47]	ArXiv 2021	.700	.630	.828	.735	.666	.834	.909	.864	.938	.874	.804	.924	.897	.839	.912
EU-Net [48]	CRV 2021	.687	.609	.793	.756	.681	.831	.902	.846	.936	.837	.765	.904	.908	.854	.917
SANet [10]	MICCAI 2021	.750	.654	.849	.753	.670	.837	.916	.859	.939	.888	.815	.928	.904	.847	.915
MSNet [11]	MICCAI 2021	.723	.652	.845	.751	.671	.838	.918	.869	.946	.865	.799	.926	.905	.849	.923
UACANet-S [50]	ACM MM 2021	.694	.615	.815	.783	.704	.847	.916	.870	.939	.902	.837	.934	.905	.852	.914
UACANet-L [50]	ACM MM 2021	.766	.689	.859	.751	.678	.835	.926	.880	.942	.910	.849	.938	.912	.859	.917
C2FNet [51]	IJCAI 2021	.699	.624	.827	.724	.650	.826	.919	.872	.941	.874	.801	.927	.886	.831	.905
DCRNet [66]	ISBI 2022	.556	.496	.736	.704	.631	.821	.896	.844	.933	.856	.788	.921	.886	.825	.911
BDG-Net [63]	SPIE MI 2022	.752	.681	.860	.797	.723	.870	.905	.857	.936	.902	.837	.940	.915	.863	.920
CaraNet [74]	SPIE MI 2022	.747	.672	.868	.773	.689	.853	.936	.887	.954	.900	.838	.940	.916	.865	.929
EFA-Net [98]	Arxiv 2023	.749	.670	.858	.774	.696	.855	.919	.871	.943	.894	.830	.941	.914	.861	.929
CFANet [7]	PR 2023	.732	.655	.845	.743	.665	.835	.932	.883	.950	.893	.827	.938	.915	.861	.924
M2SNet [99]	Arxiv 2023	.723	.652	.845	.751	.671	.838	.918	.869	.946	.865	.799	.926	.905	.849	.923
HSNet [14]	CBM 2022	.808	.734	.882	.810	.735	.868	.948	.905	.953	.903	.839	.937	.926	.877	.927
DuAT [15]	Arxiv 2022	.822	.746	.889	.819	.737	.873	.948	.906	.956	.901	.840	.940	.924	.876	.929
Polyp-PVT [13]	AIR 2023	.787	.706	.871	.808	.727	.865	.937	.889	.949	.900	.833	.935	.917	.864	.925
ESFPNet [5]	MI 2023	.823	.748	.891	.811	.730	.864	.928	.883	.943	.902	.836	.934	.917	.866	.923
FeDNet [80]	BSPC 2023	.810	.733	.892	.823	.744	.878	.930	.885	.949	.911	.848	.946	.924	.876	.933
SAM-B [100]	Arxiv 2023	.406	.370	.672	.215	.188	.553	.268	.231	.572	.371	.339	.650	.515	.459	.682
SAM-H [100]	Arxiv 2023	.517	.477	.730	.441	.396	.676	.547	.500	.738	.651	.606	.812	.778	.707	.829
SAM-L [100]	Arxiv 2023	.551	.507	.751	.468	.422	.690	.578	.526	.744	.726	.676	.849	.782	.710	.832

Kvasir-SEG [88]). The evaluation results are presented in Tab. IV and Tab. V. Secondly, we also report the mean values of Dice and MAE across the five datasets for each model in Fig. 5. It is worth noting that better models are shown in the upper left corner (*i.e.*, with a larger mDice and smaller MAE). From the results shown in Fig. 5, we have the following observations:

- **CNN vs. Transformer**. Compared with CNN-based models, Transformer-based methods obtain significantly better performance. Due to its feature extraction network structure based on the self-attention mechanism, the Transformer can effectively capture global context information.
- **Comparison of Deep Models**. Among the deep learning-

TABLE V
BENCHMARK RESULTS OF 24 REPRESENTATIVE POLYP SEGMENTATION MODELS (18 CNN-BASED AND 6 TRANSFORMER-BASED MODELS) ON FIVE COMMONLY USED DATASETS IN TERMS OF F_β, E_ϕ, AND MAE. THE THREE BEST RESULTS ARE SHOWN IN RED, BLUE, AND GREEN FONTS.

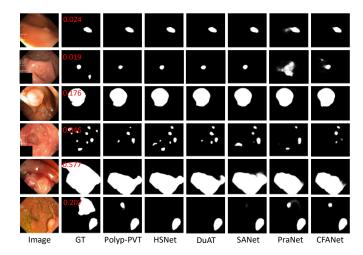| Method	Pub.	ETIS-Larib			CVC-ColonDB			CVC-ClinicDB			CVC-300			Kvasir		
		F_β	E_ϕ	MAE	F_β	E_ϕ	MAE	F_β	E_ϕ	MAE	F_β	E_ϕ	MAE	F_β	E_ϕ	MAE
UNet [24]	MICCAI 2015	.366	.643	.036	.491	.692	.059	.811	.913	.019	.684	.848	.022	.794	.881	.055
UNet++ [27]	MICCAI 2018	.390	.629	.035	.467	.680	.061	.785	.891	.022	.687	.834	.018	.808	.886	.048
SFA [8]	MICCAI 2018	.231	.531	.109	.366	.661	.094	.647	.840	.042	.341	.644	.065	.670	.834	.075
PraNet [9]	MICCAI 2020	.600	.808	.031	.699	.847	.043	.896	.963	.009	.843	.950	.010	.885	.944	.030
ACSNet [12]	MICCAI 2020	.530	.737	.059	.697	.839	.039	.873	.947	.011	.825	.939	.013	.882	.941	.032
MSEG [47]	ArXiv 2021	.671	.855	.015	.724	.859	.038	.907	.961	.007	.852	.948	.009	.885	.942	.028
EU-Net [48]	CRV 2021	.636	.807	.067	.730	.863	.045	.891	.959	.011	.805	.918	.015	.893	.951	.028
SANet [10]	MICCAI 2021	.685	.881	.015	.726	.869	.043	.909	.971	.012	.859	.962	.008	.892	.949	.028
MSNet [11]	MICCAI 2021	.677	.875	.020	.736	.872	.041	.913	.973	.008	.848	.945	.010	.892	.947	.028
UACANet-S [50]	ACM MM 2021	.650	.848	.023	.772	.894	.034	.917	.965	.008	.886	.974	.006	.897	.948	.026
UACANet-L [50]	ACM MM 2021	.740	.903	.012	.746	.875	.039	.928	.974	.006	.901	.977	.005	.902	.955	.025
C2FNet [51]	IJCAI 2021	.668	.860	.022	.705	.854	.044	.906	.969	.009	.844	.949	.009	.869	.929	.036
DCRNet [66]	ISBI 2022	.506	.742	.096	.684	.840	.052	.890	.964	.010	.830	.943	.010	.868	.933	.035
BDG-Net [63]	SPIE MI 2022	.719	.901	.014	.781	.901	.028	.898	.959	.008	.883	.969	.005	.906	.959	.025
CaraNet [74]	SPIE MI 2022	.709	.875	.017	.729	.880	.042	.931	.985	.007	.887	.977	.007	.909	.962	.023
EFA-Net [98]	Arxiv 2023	.698	.872	.018	.753	.884	.036	.916	.972	.009	.878	.961	.009	.906	.955	.024
CFANet [7]	PR 2023	.693	.881	.014	.728	.869	.039	.924	.981	.007	.875	.962	.008	.903	.956	.023
M2SNet [99]	Arxiv 2023	.677	.875	.020	.736	.872	.041	.913	.973	.008	.848	.945	.010	.892	.947	.028
HSNet [14]	CBM 2022	.777	.904	.021	.796	.912	.032	.951	.990	.006	.887	.970	.007	.918	.961	.023
DuAT [15]	Arxiv 2022	.789	.917	.013	.805	.922	.026	.950	.990	.006	.890	.965	.005	.916	.960	.023
Polyp-PVT [13]	AIR 2023	.750	.906	.013	.795	.913	.031	.936	.985	.006	.884	.973	.007	.911	.956	.023
ESFPNet [5]	MI 2023	.786	.930	.012	.798	.908	.030	.930	.976	.007	.882	.970	.006	.913	.957	.024
FeDNet [80]	BSPC 2023	.773	.931	.016	.809	.918	.029	.928	.978	.007	.897	.976	.006	.918	.963	.021
SAM-B [100]	Arxiv 2023	.404	.574	.035	.210	.412	.077	.259	.431	.092	.374	.563	.058	.509	.624	.104
SAM-H [100]	Arxiv 2023	.513	.658	.029	.434	.585	.056	.546	.676	.040	.653	.765	.020	.769	.828	.062
SAM-L [100]	Arxiv 2023	.544	.686	.030	.463	.607	.054	.563	.683	.057	.729	.824	.020	.773	.834	.061

Fig. 4. Some images of polyps in large, medium, and small sizes, along with the segmentation maps of six typical models, including three CNN-based models: PraNet [9], SANet [10], and CFA-Net [7], and three transformer-based models: Polyp-PVT [13], HSNet [14], and DuAT [15]. The numbers on the GT map represent the proportion of polyp pixels to the total number of pixels in the image. The images are sourced from the Kvasir-SEG [88] dataset.

based models, DuAT [15], FeDNet [80], ESFPNet [5], Polyp-PVT [13], HSNet [14], and BDG-Net [63] obtain much better performance.

Moreover, Fig. 6 shows the PR and F-measure curves for the 23 representative polyp segmentation models on five datasets (ETIS-LaribPolypDB [87], CVC-ColonDB [16], CVC-ClinicDB [18], CVC-300 [25], and Kvasir-SEG [88]).

To provide a deeper understanding of the better-performing models, we will discuss the main characteristics of the following six models in the sections below.

• DuAT [15]. Dual-Aggregation Transformer Network uses a transformer based on a pyramid structure as an encoder, and the decoder adopts a dual-stream design, building Local Spatial Attention modules and Global Spatial Attention modules to enhance the segmentation performance of polyps of different sizes.

• FeDNet [80] decouples the input features into high-frequency edge features and low-frequency main body features through a feature decoupling operator, and then predicts by fusing the optimized features through a feature fusion operator.

• ESFPNet [5] utilizes a pre-trained Mixed Transformer encoder and an efficient Stage-wise Feature Pyramid decoder structure, which fuses features from deep to shallow layers and also performs linear fusion of features from global to local, and connects them with intermediate aggregated features to obtain the final segmentation result.

• Polyp-PVT [13] introduces a Pyramid Vision Transformer encoder to extract multi-scale features with long-range dependencies, utilizes high-level features for side output supervision, incorporates attention mechanisms to enhance low-level features and eliminate noise, and finally fuses multi-
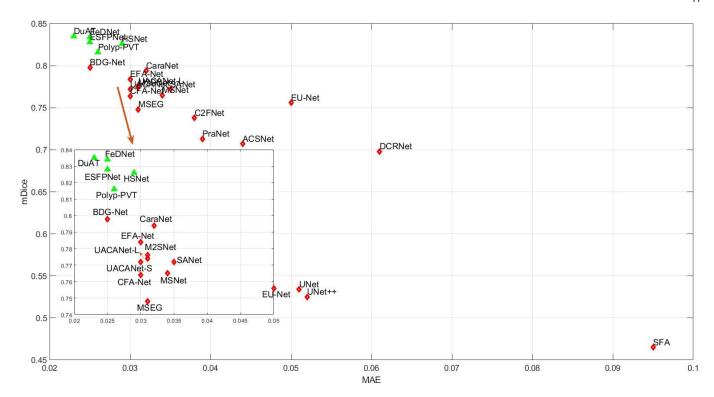
Fig. 5. A comprehensive evaluation is conducted on 23 representative deep-learning models, including UNet [24], UNet++ [27], SFA [8], PraNet [9], ACSNet [12], MSEG [47], EU-Net [48], SANet [10], MSNet [11], UACANet-S [50], UACANet-L [50], C2FNet [51], DCRNet [66], BDG-Net [63], CaraNet [74], EFA-Net [98], CFA-Net [7], M2SNet [99], Polyp-PVT [13], HSNet [14], DuAT [15], ESFPNet [5], and FeDNet [80]. We report the average Dice and MAE values for each model across five datasets (*i.e.*., ETIS-LaribPolypDB [87], CVC-ColonDB [16], CVC-ClinicDB [18], CVC-300 [25], and Kvasir-SEG [88]). Please note that the models represented in the top left corner are better (*i.e.*., they have larger Dice scores and smaller MAE values). In this context, the green triangles represent Transformer-based models, while the red diamonds signify CNN-based models.

TABLE VI

PERFORMANCE STUDIES BASED ON POLYP SIZE. COMPARISON RESULTS FOR 24 REPRESENTATIVE POLYP SEGMENTATION MODELS (18 CNN-BASED MODELS AND 6 TRANSFORMER-BASED MODELS, IN THIS CONTEXT, SAM [100] REFERS TO SAM-L) ARE PROVIDED IN TERMS OF MAE, MDICE, AND S_α. THE THREE BEST RESULTS ARE SHOWN IN RED, BLUE, AND GREEN FONTS.

	Scale	UNet [24]	UNet++ [27]	SFA [8]	PraNet [9]	ACSNet [12]	MSEG [47]	EU-Net [48]	SANet [10]	MSNet [11]	UACANet-S [50]	UACANet-L [50]	C2FNet [51]	DCRNet [66]	BDG-Net [63]	CaraNet [74]	EFA-Net [98]	CFA-Net [7]	M2SNet [99]	Polyp-PVT [13]	HSNet [14]	DuAT [15]	ESFPNet [5]	FeDNet [80]	SAM1 [100]
							CNN-based models													Transformer-based models					
MAE	Small	.026	.015	.109	.037	.034	.014	.041	.014	.019	.020	.012	.026	.068	.011	.022	.017	.011	.021	.017	.023	.012	.014	.012	.020
	Medium	.040	.038	.062	.021	.030	.016	.035	.018	.019	.016	.017	.019	.035	.013	.021	.019	.017	.016	.014	.013	.013	.012	.016	.038
	Large	.173	.215	.201	.120	.135	.138	.140	.164	.139	.120	.142	.150	.152	.114	.106	.112	.138	.122	.101	.112	.094	.107	.099	.199
	Overall	.051	.052	.095	.039	.044	.031	.050	.035	.034	.030	.031	.038	.061	.025	.029	.030	.030	.031	.026	.029	.023	.025	.025	.051
mDice	Small	.306	.373	.158	.499	.550	.603	.620	.680	.642	.624	.687	.569	.541	.674	.667	.669	.650	.634	.705	.734	.743	.727	.766	.501
	Medium	.666	.637	.632	.837	.804	.856	.852	.866	.860	.875	.857	.859	.806	.887	.870	.860	.855	.872	.890	.898	.896	.900	.884	.563
	Large	.590	.462	.589	.763	.717	.677	.711	.617	.692	.734	.672	.682	.659	.753	.817	.775	.683	.736	.803	.763	.823	.794	.804	.721
	Overall	.534	.525	.465	.713	.707	.748	.756	.772	.765	.772	.776	.738	.698	.798	.794	.784	.764	.774	.816	.826	.835	.828	.834	.561
S_α	Small	.641	.679	.493	.732	.756	.786	.778	.826	.804	.780	.824	.769	.747	.824	.827	.825	.809	.807	.833	.849	.854	.842	.871	.737
	Medium	.800	.784	.736	.893	.876	.904	.885	.902	.907	.907	.897	.903	.878	.922	.912	.905	.902	.909	.916	.922	.921	.922	.918	.747
	Large	.684	.590	.639	.791	.772	.743	.745	.698	.751	.768	.735	.741	.731	.790	.824	.801	.745	.768	.806	.782	.823	.796	.813	.696
	Overall	.731	.724	.642	.826	.822	.844	.831	.851	.853	.846	.852	.837	.815	.873	.872	.865	.851	.857	.874	.880	.886	.879	.889	.737

level features.

• HSNet [14] is also based on a PVT encoder and suppresses noise information in low-level features by modeling the semantic spatial relationships and channel dependencies of lower-layer features. It bridges feature disparities through a semantic interaction mechanism and captures long-range dependencies and local appearance details via a dual-branch structure.

• BDG-Net [63] aggregates high-level features to generate a boundary distribution map, which is fed into a boundary distribution-guided decoder and employs a multi-scale feature

interaction strategy to enhance segmentation precision.

2) Scale-based Evaluation: To investigate the influence of scale variations, we carry out evaluations on several representative polyp segmentation models. To achieve this evaluation, we compute the ratio (r) of the size of the polyp body area in a given ground truth image, which is used to characterize the size of the polyp. For this purpose, three types of polyp scales are defined: 1) when r is less than 0.025, the polyp is considered "small"; 2) when r is more than 0.2, it is considered "large"; 3) when the ratio is within the range $[0.025, 0.2]$, we call it "medium". In addition,
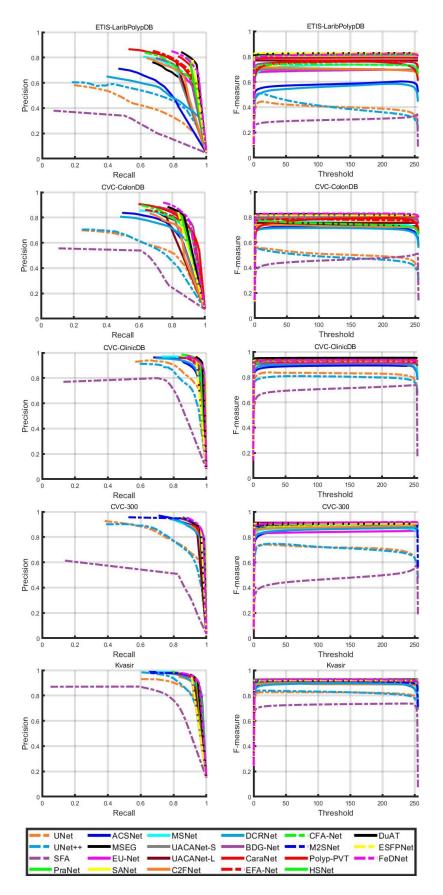
Fig. 6. The PR curves and F-measures under different thresholds for 23 deep polyp segmentation models on five datasets in ETIS-LaribPolypDB [87], CVC-ColonDB [16], CVC-ClinicDB [18], CVC-300 [25], and Kvasir-SEG [88].

we mix and classify the five colonoscopy datasets, deriving a newly constructed dataset containing 212, 334, and 77 images of "small", "medium", and "large" types, respectively. Fig. 7 presents the comparison results associated with scale variations. Segmentation performance is represented by three metrics (*i.e.*, IoU, F_β, and E_ϕ).

Some sample images with different scales of polyps are shown in Fig. 4. Visual comparison results of the scale variation evaluation are shown in Tab. VI. According to the results, we can draw the following conclusions: (1) In terms of overall performance, DuAT [15] and FeDNet [80] perform better across all types, and (2) Vertical analysis reveals that most models achieve better performance in segmenting "medium" polyps while they exhibit relatively lower performance in segmenting other types.

V. CHALLENGES AND FUTURE TRENDS

A. Effective Network Structure

Edge Feature Extraction. The diverse shapes of polyps and the complex background of endoscopic images, combined with blurred edges and adherences of polyps often pose challenges for accurate segmentation. To further improve accuracy and robustness, and to leverage richer and more complex image features, some models introduce edge feature extraction modules ([8], [65], [71], [73]) to utilize low-level features to assist in polyp segmentation. Most polyp segmentation models capture the edge information by utilizing low-level features extracted by backbone networks, which are then aggregated with high-level features to accomplish segmentation. However, after operations like convolution, the extracted image features, including edge and body features, become intertwined, making decoupling extremely challenging. Additionally, as the tasks of feature extraction and segmentation are not orthogonal, incorrect edge estimation can lead to error propagation. Therefore, an optional choice is to decouple the original images using traditional image processing methods. There has already been some work in the fields of object detection and other semantic segmentation domains. Shan *et al.* [101] utilized Fourier Transform to derive high and low-frequency components from images. These components are then input into two parallel branches to obtain edge and body features, which are subsequently merged to perform semantic segmentation. Cong *et al.* [102] designed a full-frequency perception module based on Octave Convolution. This module can automatically learn low and high-frequency features for coarse positioning, providing auxiliary information for segmentation. FeDNet [80] employs Laplacian pyramid decomposition to decouple the input features into high-frequency edge features and low-frequency body features. Following this, these two types of features undergo deep supervised optimization.

Dual-stream Structure. CNNs exhibit a powerful capability to extract local features, such as texture and shape. However, CNNs are relatively weaker in capturing long-distance global dependencies. In this context, the Transformer can serve as a supplement. Due to its self-attention mechanism, it possesses an excellent ability to capture global information. Therefore, introducing a dual-stream structure in existing polyp segmentation models can integrate local information captured by CNNs and global information captured by the Transformer, enhancing the performance of polyp segmentation. In addition, this structure may improve the model's capabilities to handle complex, uneven backgrounds and noise interference, thereby increasing the robustness of the model [6], [14], [53], [67], [68]. It is important to note that how to effectively merge the outputs from the two networks and balance their weights remains a problem. However, the tremendous potential and possibilities undoubtedly justify further exploration and research.

B. Different Supervision Strategies

Existing polyp segmentation models typically use fully-supervised strategies to learn features and produce segmentation results. However, annotating colonoscopy data is time-consuming and labor-intensive, especially for video data. To alleviate this issue, attention has been shifted towards weakly-supervised and semi-supervised learning, applying them to polyp segmentation tasks [15], [103]. Actually, semi-supervised or weakly supervised methods have been widely applied in the field of medical image segmentation. For example, [104] proposed a novel data augmentation approach for medical image segmentation that does not lose essential semantic information of the key objects. [105] redefined the traditional per-pixel segmentation task as a contour regression problem and modeled the position uncertainty. [106] introduced a semi-supervised medical image segmentation technique that first trains the segmentation model on a small number of unlabeled images, generates initial labels for them, and also introduces a consistency-based pseudo-label enhancement scheme to improve the quality of the model's predictions. Therefore, in the future, semi/weakly supervised methods can be used for image-level labeling and pseudo-annotation to improve the accuracy of polyp segmentation.

C. Clinical Requirements

Dataset Collection. The shape, texture, and color of polyps can vary depending on the time and stage of the disease. Colonoscopy data from multiple centers also often exhibit different morphologies. Existing polyp segmentation datasets mostly consist of images containing a single polyp, and there are not many specialized datasets with a large number of images. Only 23 out of the 623 images in the aforementioned five datasets (ETIS-LaribPolypDB [87], CVC-ColonDB [16], CVC-ClinicDB [18], CVC-300 [25], and Kvasir-SEG [88]) contain multiple polyps, as such, models trained on these datasets often perform poorly on tasks involving multiple polyps. Although there are multiple public datasets available for polyp segmentation, their scale is quite limited. For instance, the largest dataset we present is PolypGen [92], which contains 3, 762 images. Most of the data in these datasets comes from colonoscopy images and video slices. Therefore, it is necessary to develop a new large-scale polyp segmentation dataset to serve as a baseline for future research. In addition to this, collecting datasets for complex specialized scenarios is also a potential direction. For example, constructing multi-center datasets, multi-target datasets, and specialized datasets for small or large polyps could enhance the model's performance in different scenarios.
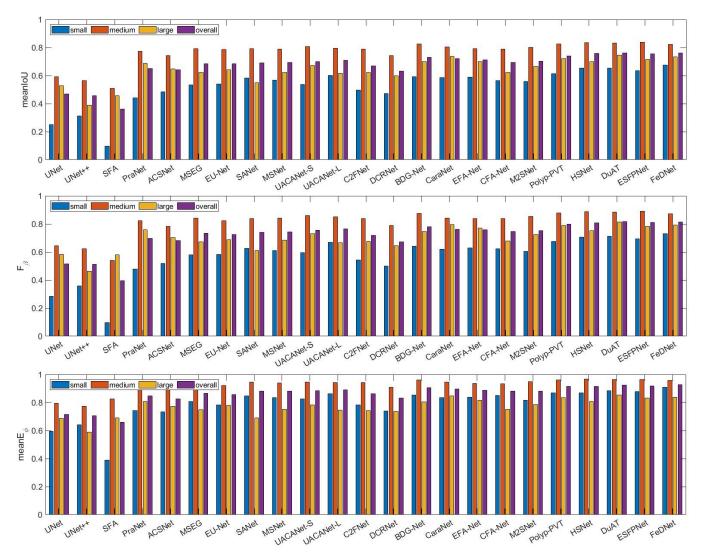
Fig. 7. Performance study based on the size of polyps (*i.e.*, small vs. medium vs. large). The comparison results on 23 representative polyp segmentation models (*i.e.*,) are given in terms of mIoU (top), F_β (medium), and E_ϕ (bottom).

Cross-domain Segmentation. Deep learning-based polyp segmentation methods have achieved promising performance, but they often suffer from performance degradation when applied to unseen target domain datasets collected from different imaging devices. Thus, it is still challenging to apply existing polyp segmentation methods to unseen datasets. More importantly, manual annotation for new target datasets is tedious and labor-intensive, leveraging knowledge learned from the labeled source domains to boost the performance in the unlabeled target domain is highly demanded in clinical practice. To achieve this, Yang *et al.* [34] proposed a mutual-prototype adaptation network for cross-domain polyp segmentation, which significantly reduces the gap between the two domains to improve the segmentation performance on the target domain datasets. Thus, this direction deserves further exploration to develop more cross-domain segmentation models in this task.

Real-Time Polyp Segmentation. It is worth noting that real-time segmentation is important for this task, as it is anticipated that the segmented results can be immediately presented to the doctor during the colonoscopy procedures for further decisions and treatments. However, the current deep learning-based models often require huge computation complexity,

making them challenging to apply for real-time segmentation. Several real-time polyp segmentation models have been developed [44], [86], [107]–[112]. Indeed, the development of efficient lightweight networks for polyp segmentation without sacrificing performance is of utmost importance. However, it poses a significant challenge due to the inherent trade-off between model complexity and efficiency. Efficient lightweight networks can enable real-time segmentation, reduce computational costs, and facilitate deployment in resource-constrained clinical settings. Thus, overcoming this challenge involves finding innovative solutions that strike a balance between model complexity and performance, ultimately enhancing the practical applicability of polyp segmentation algorithms in clinical practice.

D. Ethical Issues

The specificity of medical issues often raises privacy concerns when using real patient examination data from hospitals. Moreover, there are inherent differences between data obtained from different centers. Models trained on data from a single center tend to perform worse when applied to unseen data acquired from different scanners or other centers. Thus, it becomes crucial to leverage the knowledge gained

from labeled source domains to enhance the performance in unlabeled target domains. The goal is to mitigate the domain shift observed in colonoscopy images sourced from multiple centers and devices. Federated learning emerges as a promising approach in this context, enabling multiple centers to collaboratively learn a shared prediction model while ensuring privacy protection. For instance, Liu *et al.* [113] presented a novel method of incidental learning in continuous frequency space, enabling diverse endpoints to utilize multi-source data distributions while addressing challenging constraints associated with data dispersion.

VI. Conclusion

In this paper, to the best of our knowledge, we provide the first comprehensive review of the development and evaluation in the field of polyp segmentation. We categorize models into traditional and deep ones initially, then focus on reviewing existing deep models from various perspectives, followed by a summary of popular polyp segmentation datasets and providing detailed information for each dataset. Following that, we conduct a comprehensive assessment as well as an evaluation based on polyp sizes for 24 representative deep learning-based polyp segmentation models. Specifically, we perform a size-based performance analysis by constructing a new dataset for 24 representative polyp segmentation models. Furthermore, we discuss some challenges and highlight open directions for future research. Although significant progress has been made in the field of polyp segmentation over the past few decades, there is still ample room for improvement. We hope this survey will spark more interest and understanding in this field. To promote future research for polyp segmentation, we will continue to collect newly released polyp segmentation models at: https://github.com/taozh2017/Awesome-Polyp-Segmentation.

References

[1] X. Guo, C. Yang, Y. Liu, and Y. Yuan, "Learn to threshold: Threshold-net with confidence-guided manifold mixup for polyp segmentation," *IEEE TMI*, vol. 40, no. 4, pp. 1134–1146, 2020.

[2] X. Yang, Q. Wei, C. Zhang, K. Zhou, L. Kong, and W. Jiang, "Colon polyp detection and segmentation based on improved mrcnn," *IEEE TIM*, vol. 70, pp. 1–10, 2020.

[3] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE TPAMI*, vol. 44, no. 10, pp. 6024–6042, 2021.

[4] D. Jha, N. K. Tomar, V. Sharma, and U. Bagci, "Transnetr: Transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing," *arXiv preprint arXiv:2303.07428*, 2023.

[5] Q. Chang, D. Ahmad, J. Toth, R. Bascom, and W. E. Higgins, "Esfpnet: efficient deep learning architecture for real-time lesion segmentation in autofluorescence bronchoscopic video," in *Medical Imaging: Medical Imaging: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 12468. SPIE, 2023, p. 1246803.

[6] Y. Wang, Z. Deng, Q. Lou, S. Hu, K.-s. Choi, and S. Wang, "Cooperation learning enhanced colonic polyp segmentation based on transformer-cnn fusion," *arXiv preprint arXiv:2301.06892*, 2023.

[7] T. Zhou, Y. Zhou, K. He, C. Gong, J. Yang, H. Fu, and D. Shen, "Cross-level feature aggregation network for polyp segmentation," *PR*, vol. 140, p. 109555, 2023.

[8] Y. Fang, C. Chen, Y. Yuan, and K.-y. Tong, "Selective feature aggregation network with area-boundary constraints for polyp segmentation," in *MICCAI*. Springer, 2019, pp. 302–310.

[9] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *MICCAI*. Springer, 2020, pp. 263–273.

[10] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui, "Shallow attention network for polyp segmentation," in *MICCAI*. Springer, 2021, pp. 699–708.

[11] X. Zhao, L. Zhang, and H. Lu, "Automatic polyp segmentation via multi-scale subtraction network," in *MICCAI*. Springer, 2021, pp. 120–130.

[12] R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, and Y. Yu, "Adaptive context selection for polyp segmentation," in *MICCAI*. Springer, 2020, pp. 253–262.

[13] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-pvt: Polyp segmentation with pyramid vision transformers," *CAAI AIR*, 2023.

[14] W. Zhang, C. Fu, Y. Zheng, F. Zhang, Y. Zhao, and C.-W. Sham, "Hsnet: A hybrid semantic network for polyp segmentation," *Computers in biology and medicine*, vol. 150, p. 106173, 2022.

[15] F. Tang, Z. Xu, Q. Huang, J. Wang, X. Hou, J. Su, and J. Liu, "Duat: Dual-aggregation transformer network for medical image segmentation," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 2023, pp. 343–356.

[16] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE TMI*, vol. 35, no. 2, pp. 630–644, 2015.

[17] Y. Iwahori, H. Hagi, H. Usami, R. J. Woodham, A. Wang, M. K. Bhuyan, and K. Kasugai, "Automatic polyp detection from endoscope image using likelihood map based on edge information," in *ICPRAM*, vol. 2. SciTePress, 2017, pp. 402–409.

[18] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *CMIG*, vol. 43, pp. 99–111, 2015.

[19] J. Yao, M. Miller, M. Franaszek, and R. M. Summers, "Colonic polyp segmentation in ct colonography-based on fuzzy clustering and deformable models," *IEEE TMI*, vol. 23, no. 11, pp. 1344–1352, 2004.

[20] L. Lu, A. Barbu, M. Wolf, J. Liang, M. Salganicoff, and D. Comaniciu, "Accurate polyp segmentation for 3d ct colongraphy using multi-staged probabilistic binary learning and compositional model," in *IEEE CVPR*, 2008, pp. 1–8.

[21] S. Gross, M. Kennel, T. Stehle, J. Wulff, J. Tischendorf, C. Trautwein, and T. Aach, "Polyp segmentation in nbi colonoscopy," in *Bildverarbeitung für die Medizin*. Springer, 2009, pp. 252–256.

[22] M. Ganz, X. Yang, and G. Slabaugh, "Automatic segmentation of polyps in colonoscopic narrow-band imaging data," *IEEE TBE*, vol. 59, no. 8, pp. 2144–2151, 2012.

[23] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE CVPR*, 2015, pp. 3431–3440.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[25] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, A. Courville *et al.*, "A benchmark for endoluminal scene segmentation of colonoscopy images," *JHE*, vol. 2017, 2017.

[26] M. Akbari, M. Mohrekesh, E. Nasr-Esfahani, S. R. Soroushmehr, N. Karimi, S. Samavi, and K. Najarian, "Polyp segmentation in colonoscopy images using fully convolutional network," in *Proceedings of the IEEE International Conference on Medicine and Biology Society*, 2018, pp. 69–72.

[27] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.

[28] A. Srivastava, D. Jha, S. Chanda, U. Pal, H. D. Johansen, D. Johansen, M. A. Riegler, S. Ali, and P. Halvorsen, "Msrf-net: A multi-scale residual fusion network for biomedical image segmentation," *IEEE JBHI*, vol. 26, no. 5, pp. 2252–2263, 2021.

[29] N. K. Tomar, D. Jha, U. Bagci, and S. Ali, "Tganet: Text-guided attention for improved polyp segmentation," in *MICCAI*. Springer, 2022, pp. 151–160.

[30] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, and S. Song, "Stepwise feature fusion: Local guides global," in *MICCAI*. Springer, 2022, pp. 110–120.

[31] M. M. Rahman and R. Marculescu, "Medical image segmentation via cascaded attention decoding," in *IEEE/CVF WACVW*, 2023, pp. 6222–6231.

[32] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[33] Y. Li, M. Hu, and X. Yang, "Polyp-sam: Transfer sam for polyp segmentation," *arXiv preprint arXiv:2305.00293*, 2023.

[34] C. Yang, X. Guo, M. Zhu, B. Ibragimov, and Y. Yuan, "Mutual-prototype adaptation for cross-domain polyp segmentation," *IEEE JBHI*, vol. 25, no. 10, pp. 3886–3897, 2021.

[35] S. Gupta, G. Sikka, and A. Malik, "A review on deep learning-based polyp segmentation for efficient colorectal cancer screening," in *IEEE ICSCCC*, 2023, pp. 501–506.

[36] L. F. Sanchez-Peralta, L. Bote-Curiel, A. Picon, F. M. Sanchez-Margallo, and J. B. Pagador, "Deep learning to find colorectal polyps in colonoscopy: A systematic literature review," *Artificial Intelligence in Medicine*, vol. 108, p. 101923, 2020.

[37] H. Xiao, L. Li, Q. Liu, X. Zhu, and Q. Zhang, "Transformers in medical image segmentation: A review," *BSPC*, vol. 84, p. 104791, 2023.

[38] I. Qureshi, J. Yan, Q. Abbas, K. Shaheed, A. B. Riaz, A. Wahid, M. W. J. Khan, and P. Szczuko, "Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends," *Information Fusion*, 2022.

[39] C. L. Chowdhary and D. P. Acharjya, "Segmentation and feature extraction in medical imaging: a systematic review," *Procedia Computer Science*, vol. 167, pp. 26–36, 2020.

[40] L. Liu, J. Cheng, Q. Quan, F.-X. Wu, Y.-P. Wang, and J. Wang, "A survey on u-shaped networks in medical image segmentations," *Neurocomputing*, vol. 409, pp. 244–258, 2020.

[41] H. Thisanke, C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, and D. Herath, "Semantic segmentation using vision transformers: A survey," *EAAI*, vol. 126, p. 106669, 2023.

[42] M. T. Bennai, Z. Guessoum, S. Mazouzi, S. Cormier, and M. Mezghiche, "Multi-agent medical image segmentation: A survey," *CMPB*, p. 107444, 2023.

[43] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," in *IEEE ISM*. IEEE, 2019, pp. 225–2255.

[44] J. Zhong, W. Wang, H. Wu, Z. Wen, and J. Qin, "Polypseg: An efficient context-aware network for polyp segmentation from colonoscopy videos," in *MICCAI*. Springer, 2020, pp. 285–294.

[45] N. K. Tomar, D. Jha, S. Ali, H. D. Johansen, D. Johansen, M. A. Riegler, and P. Halvorsen, "Ddanet: Dual decoder attention network for automatic polyp segmentation," in *Pattern Recognition. ICPR International Workshops and Challenges*. Springer, 2021, pp. 307–314.

[46] A. Srivastava, S. Chanda, D. Jha, U. Pal, and S. Ali, "Gmsrf-net: An improved generalizability with global multi-scale residual fusion network for polyp segmentation," in *IEEE ICPR*, 2022, pp. 4321–4327.

[47] C.-H. Huang, H.-Y. Wu, and Y.-L. Lin, "Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps," *arXiv preprint arXiv:2101.07172*, 2021.

[48] K. Patel, A. M. Bur, and G. Wang, "Enhanced u-net: A feature enhancement network for polyp segmentation," in *IEEE CRV*, 2021, pp. 181–188.

[49] N. K. Tomar, D. Jha, M. A. Riegler, H. D. Johansen, D. Johansen, J. Rittscher, P. Halvorsen, and S. Ali, "Fanet: A feedback attention network for improved biomedical image segmentation," *IEEE TNNLS*, 2022.

[50] T. Kim, H. Lee, and D. Kim, "Uacanet: Uncertainty augmented context attention for polyp segmentation," in *ACM MM*, 2021, pp. 2167–2175.

[51] Y. Sun, G. Chen, T. Zhou, Y. Zhang, and N. Liu, "Context-aware cross-level fusion network for camouflaged object detection," *arXiv preprint arXiv:2105.12555*, 2021.

[52] D. Jha, P. H. Smedsrud, D. Johansen, T. de Lange, H. D. Johansen, P. Halvorsen, and M. A. Riegler, "A comprehensive study on colorectal polyp segmentation with resunet++, conditional random field and test-time augmentation," *IEEE JBHI*, vol. 25, no. 6, pp. 2029–2040, 2021.

[53] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *MICCAI*. Springer, 2021, pp. 14–24.

[54] L. Wu, Z. Hu, Y. Ji, P. Luo, and S. Zhang, "Multi-frame collaboration for effective endoscopic video polyp detection via spatial-temporal feature transformation," in *MICCAI*. Springer, 2021, pp. 302–312.

[55] M. Cheng, Z. Kong, G. Song, Y. Tian, Y. Liang, and J. Chen, "Learnable oriented-derivative network for polyp segmentation," in *MICCAI*. Springer, 2021, pp. 720–730.

[56] T.-C. Nguyen, T.-P. Nguyen, G.-H. Diep, A.-H. Tran-Dinh, T. V. Nguyen, and M.-T. Tran, "Ccbanet: cascading context and balancing attention for polyp segmentation," in *MICCAI*. Springer, 2021, pp. 633–643.

[57] Y. Shen, X. Jia, and M. Q.-H. Meng, "Hrenet: A hard region enhancement network for polyp segmentation," in *MICCAI*. Springer, 2021, pp. 559–568.

[58] H. Wu, Z. Zhao, J. Zhong, W. Wang, Z. Wen, and J. Qin, "Polypseg+: A lightweight context-aware network for real-time polyp segmentation," *IEEE TCYB*, vol. 53, no. 4, pp. 2610–2621, 2022.

[59] C. Wu, C. Long, S. Li, J. Yang, F. Jiang, and R. Zhou, "Msraformer: Multiscale spatial reverse attention network for polyp segmentation," *CBM*, vol. 151, p. 106274, 2022.

[60] K. B. Patel, F. Li, and G. Wang, "Fuzzynet: A fuzzy attention module for polyp segmentation," in *NeurIPS'22 Workshop on All Things Attention: Bridging Different Perspectives on Attention*, 2022.

[61] R. Zhang, P. Lai, X. Wan, D.-J. Fan, F. Gao, X.-J. Wu, and G. Li, "Lesion-aware dynamic kernel for polyp segmentation," in *MICCAI*. Springer, 2022, pp. 99–109.

[62] T.-Y. Liao, C.-H. Yang, Y.-W. Lo, K.-Y. Lai, P.-H. Shen, and Y.-L. Lin, "Hardnet-dfus: An enhanced harmonically-connected network for diabetic foot ulcer image segmentation and colonoscopy polyp segmentation," *arXiv preprint arXiv:2209.07313*, 2022.

[63] Z. Qiu, Z. Wang, M. Zhang, Z. Xu, J. Fan, and L. Xu, "Bdg-net: boundary distribution guided network for accurate polyp segmentation," in *Medical Imaging: Image Processing*, vol. 12032. SPIE, 2022, pp. 792–799.

[64] N. T. Duc, N. T. Oanh, N. T. Thuy, T. M. Triet, and V. S. Dinh, "Colonformer: An efficient transformer based method for colon polyp segmentation," *IEEE Access*, vol. 10, pp. 80 575–80 586, 2022.

[65] E. Sanderson and B. J. Matuszewski, "Fcn-transformer feature fusion for polyp segmentation," in *MIUA*. Springer, 2022, pp. 892–907.

[66] Z. Yin, K. Liang, Z. Ma, and J. Guo, "Duplex contextual relation network for polyp segmentation," in *IEEE ISBI*. IEEE, 2022, pp. 1–5.

[67] M. Nguyen, T. T. Bui, Q. Van Nguyen, T. T. Nguyen, and T. Van Pham, "Lapformer: A light and accurate polyp segmentation transformer," *arXiv preprint arXiv:2210.04393*, 2022.

[68] L. Cai, M. Wu, L. Chen, W. Bai, M. Yang, S. Lyu, and Q. Zhao, "Using guided self-attention with local information for polyp segmentation," in *MICCAI*. Springer, 2022, pp. 629–638.

[69] Y. Lin, J. Wu, G. Xiao, J. Guo, G. Chen, and J. Ma, "Bsca-net: Bit slicing context attention network for polyp segmentation," *PR*, vol. 132, p. 108917, 2022.

[70] J. Wei, Y. Hu, G. Li, S. Cui, S. Kevin Zhou, and Z. Li, "Boxpolyp: Boost generalized polyp segmentation using extra coarse bounding box annotations," in *MICCAI*. Springer, 2022, pp. 67–77.

[71] Y. Xiao, Z. Chen, L. Wan, L. Yu, and L. Zhu, "Icbnet: Iterative context-boundary feedback network for polyp segmentation," in *IEEE BIBM*, 2022, pp. 1297–1304.

[72] R. Chen, X. Wang, B. Jin, J. Tu, F. Zhu, and Y. Li, "Cld-net: Complement local detail for medical small-object segmentation," in *IEEE BIBM*, 2022, pp. 942–947.

[73] L. Lu, X. Zhou, S. Chen, Z. Chen, J. Yu, H. Tang, and X. Hu, "Boundary-aware polyp segmentation network," in *PRCV*. Springer, 2022, pp. 66–77.

[74] A. Lou, S. Guan, H. Ko, and M. H. Loew, "Caranet: Context axial reverse attention network for segmentation of small medical objects," in *Medical Imaging 2022: Image Processing*, vol. 12032. SPIE, 2022, pp. 81–92.

[75] G. Yue, S. Li, R. Cong, T. Zhou, B. Lei, and T. Wang, "Attention-guided pyramid context network for polyp segmentation in colonoscopy images," *IEEE TIM*, vol. 72, pp. 1–13, 2023.

[76] K. Wang, L. Liu, X. Fu, L. Liu, and W. Peng, "Ra-denet: Reverse attention and distractions elimination network for polyp segmentation," *Computers in Biology and Medicine*, vol. 155, p. 106704, 2023.

[77] Y. Su, J. Cheng, C. Zhong, C. Jiang, J. Ye, and J. He, "Accurate polyp segmentation through enhancing feature fusion and boosting boundary performance," *Neurocomputing*, vol. 545, p. 126233, 2023.

[78] K. Hu, W. Chen, Y. Sun, X. Hu, Q. Zhou, and Z. Zheng, "Ppnet: Pyramid pooling based network for polyp segmentation," *Computers in Biology and Medicine*, vol. 160, p. 107028, 2023.

[79] N. K. Tomar, D. Jha, and U. Bagci, "Dilatedsegnet: A deep dilated segmentation network for polyp segmentation," in *MMM*. Springer, 2023, pp. 334–344.

[80] Y. Su, J. Cheng, C. Zhong, Y. Zhang, J. Ye, J. He, and J. Liu, "Fednet: Feature decoupled network for polyp segmentation from endoscopy images," *BSPC*, vol. 83, p. 104699, 2023.

[81] T.-H. Nguyen-Mau, Q.-H. Trinh, N.-T. Bui, P.-T. V. Thi, M.-V. Nguyen, X.-N. Cao, M.-T. Tran, and H.-D. Nguyen, "Pefnet: Positional embedding feature for polyp segmentation," in *MMM*. Springer, 2023, pp. 240–251.

[82] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[83] G.-P. Ji, G. Xiao, Y.-C. Chou, D.-P. Fan, K. Zhao, G. Chen, and L. Van Gool, "Video polyp segmentation: A deep learning perspective," *MIR*, vol. 19, no. 6, pp. 531–549, 2022.

[84] X. Zhao, Z. Wu, S. Tan, D.-J. Fan, Z. Li, X. Wan, and G. Li, "Semi-supervised spatial temporal attention network for video polyp segmentation," in *MICCAI*. Springer, 2022, pp. 456–466.

[85] G.-P. Ji, Y.-C. Chou, D.-P. Fan, G. Chen, H. Fu, D. Jha, and L. Shao, "Progressively normalized self-attention network for video polyp segmentation," in *MICCAI*. Springer, 2021, pp. 142–152.

[86] D. Jha, N. K. Tomar, S. Ali, M. A. Riegler, H. D. Johansen, D. Johansen, T. de Lange, and P. Halvorsen, "Nanonet: Real-time polyp segmentation in video capsule endoscopy and colonoscopy," in *IEEE CBMS*. IEEE, 2021, pp. 37–43.

[87] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *IJCARS*, vol. 9, pp. 283–293, 2014.

[88] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *MMM*. Springer, 2020, pp. 451–462.

[89] Y. Ma, X. Chen, K. Cheng, Y. Li, and B. Sun, "Ldpolypvideo benchmark: a large-scale colonoscopy video dataset of diverse polyps," in *MICCAI 2021*. Springer, pp. 387–396.

[90] A. Wang, M. Xu, Y. Zhang, M. Islam, and H. Ren, "S2me: Spatial-spectral mutual teaching and ensemble learning for scribble-supervised polyp segmentation," *arXiv preprint arXiv:2306.00451*, 2023.

[91] L. F. Sánchez-Peralta, J. B. Pagador, A. Picón, Á. J. Calderón, F. Polo, N. Andraka, R. Bilbao, B. Glover, C. L. Saratxaga, and F. M. Sánchez-Margallo, "Piccolo white-light and narrow-band imaging colonoscopic dataset: A performance comparative of models and datasets," *Applied Sciences*, vol. 10, no. 23, p. 8501, 2020.

[92] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, M. A. Riegler, K. V. Anonsen *et al.*, "A multi-centre polyp detection and segmentation dataset for generalisability assessment," *Scientific Data*, vol. 10, no. 1, p. 75, 2023.

[93] "Gastrointestinal image analysis (giana) challenge," Online, available: https://endovissub2017-giana.grand-challenge.org/.

[94] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE TIP*, vol. 24, no. 12, pp. 5706–5722, 2015.

[95] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *IEEE CVPR*, 2012, pp. 733–740.

[96] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *IEEE ICCV*, 2017, pp. 4548–4557.

[97] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *IJCAI*, 2018, pp. 698–704.

[98] T. Zhou, Y. Zhang, G. Chen, Y. Zhou, Y. Wu, and D.-P. Fan, "Edge-aware feature aggregation network for polyp segmentation," *arXiv preprint arXiv:2309.10523*, 2023.

[99] X. Zhao, H. Jia, Y. Pang, L. Lv, F. Tian, L. Zhang, W. Sun, and H. Lu, "M2snet: Multi-scale in multi-scale subtraction network for medical image segmentation," *arXiv preprint arXiv:2303.10894*, 2023.

[100] T. Zhou, Y. Zhang, Y. Zhou, Y. Wu, and C. Gong, "Can sam segment polyps?" *arXiv preprint arXiv:2304.07583*, 2023.

[101] L. Shan, X. Li, and W. Wang, "Decouple the high-frequency and low-frequency information of images for semantic segmentation," in *IEEE ICASSP*, 2021, pp. 1805–1809.

[102] R. Cong, M. Sun, S. Zhang, X. Zhou, W. Zhang, and Y. Zhao, "Frequency perception network for camouflaged object detection," in *ACM MM*, 2023, pp. 1179–1189.

[103] H. Wu, G. Chen, Z. Wen, and J. Qin, "Collaborative and adversarial learning of focused and dispersive representations for semi-supervised polyp segmentation," in *IEEE ICCV*, October 2021, pp. 3489–3498.

[104] H. Cho, Y. Han, and W. H. Kim, "Anti-adversarial consistency regularization for data augmentation: Applications to robust medical image segmentation," in *MICCAI*. Springer, 2023, pp. 555–566.

[105] T. Judge, O. Bernard, W.-J. Cho Kim, A. Gomez, A. Chartsias, and P.-M. Jodoin, "Asymmetric contour uncertainty estimation for medical image segmentation," in *MICCAI*. Springer, 2023, pp. 210–220.

[106] Q. Wei, L. Yu, X. Li, W. Shao, C. Xie, L. Xing, and Y. Zhou, "Consistency-guided meta-learning for bootstrapping semi-supervised medical image segmentation," in *MICCAI*. Springer, 2023, pp. 183–193.

[107] H. Wu, Z. Zhao, J. Zhong, W. Wang, Z. Wen, and J. Qin, "Polypseg+: A lightweight context-aware network for real-time polyp segmentation," *IEEE TCYB*, vol. 53, no. 4, pp. 2610–2621, 2022.

[108] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, and P. Halvorsen, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *IEEE Access*, vol. 9, pp. 40 496–40 510, 2021.

[109] I. Wichakam, T. Panboonyuen, C. Udomcharoenchaikit, and P. Vateekul, "Real-time polyps segmentation for colonoscopy video frames using compressed fully convolutional network," in *MMM*. Springer, 2018, pp. 393–404.

[110] N. K. Tomar, A. Shergill, B. Rieders, U. Bagci, and D. Jha, "Transresunet: Transformer based ResU-Net for real-time colonoscopy polyp segmentation," *arXiv preprint arXiv:2206.08985*, 2022.

[111] H. Wu, J. Zhong, W. Wang, Z. Wen, and J. Qin, "Precise yet efficient semantic calibration and refinement in convnets for real-time polyp segmentation from colonoscopy videos," in *AAAI*, vol. 35, no. 4, 2021, pp. 2916–2924.

[112] R. Feng, B. Lei, W. Wang, T. Chen, J. Chen, D. Z. Chen, and J. Wu, "Ssn: A stair-shape network for real-time polyp segmentation in colonoscopy images," in *IEEE ISBI*, 2020, pp. 225–229.

[113] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng, "Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *IEEE CVPR*, 2021, pp. 1013–1023.