An Embedding-Unleashing Video Polyp Segmentation Framework via Region Linking and Scale Alignment

Zhixue Fang[1], Xinrong Guo[1], Jingyin Lin[1], Huisi Wu[1]*, Jing Qin[2]

[1] College of Computer Science and Software Engineering, Shenzhen University
[2] Centre for Smart Health, The Hong Kong Polytechnic University
hswu@szu.edu.cn

Abstract

Automatic polyp segmentation from colonoscopy videos is a critical task for the development of computer-aided screening and diagnosis systems. However, accurate and real-time video polyp segmentation (VPS) is a very challenging task due to low contrast between background and polyps and frame-to-frame dramatic variations in colonoscopy videos. We propose a novel embedding-unleashing framework consisting of a proposal-generative network (PGN) and an appearance-embedding network (AEN) to comprehensively address these challenges. Our framework, for the first time, models VPS as an appearance-level semantic embedding process to facilitate generate more global information to counteract background disturbances and dramatic variations. Specifically, PGN is a video segmentation network to obtain segmentation mask proposals, while AEN is a network we specially designed to produce appearance-level embedding semantics for PGN, thereby unleashing the capability of PGN in VPS. Our AEN consists of a cross-scale region linking (CRL) module and a cross-wise scale alignment (CSA) module. The former screens reliable background information against background disturbances by constructing linking of region semantics, while the latter performs the scale alignment to resist dramatic variations by modeling the center-perceived motion dependence with a cross-wise manner. We further introduce a parameter-free semantic interaction to embed the semantics of AEN into PGN to obtain the segmentation results. Extensive experiments on CVC-612 and SUN-SEG demonstrate that our approach achieves better performance than other state-of-the-art methods. Codes are available at https://github.com/zhixue-fang/EUVPS.

Introduction

Colorectal cancer (CRC), a gastrointestinal malignancy, is the leading cause of cancer-related deaths worldwide (Center et al. 2009). Fortunately, regular screening for polyps, which are precursors to CRC, is an effective CRC prevention method. Colonoscopy is a commonly used technique to assist doctors in screening and removing polyps. However, the entire diagnostic process largely relies on the doctor's experience, and lack of experience may lead to missed screening of precancerous lesions. Therefore, accurate and real-time automatic polyp segmentation from colonoscopy videos is highly demanded in clinical practice.

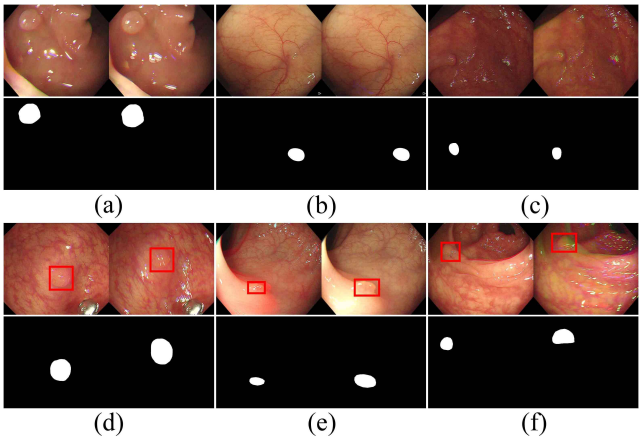

Figure 1: Challenges in VPS, including background disturbances (a)-(c) and dramatic variations (d)-(f). (a)-(c) low contrast between background and polyps. (d) position variation. (e) size variation. (f) shape variation. Note that (a)-(f) are strictly adjacent two frames from SUN-SEG.

Recently, many deep learning methods have achieved remarkable success in *image* polyp segmentation (IPS) (Fan et al. 2020; Wu et al. 2021b,a, 2022; Zhou et al. 2023). However, these methods still struggle to localize the exact location of polyps due to the low contrast between the polyp and the background (Li et al. 2022; Wu et al. 2023), as shown in Figure 1 (a-c). In addition, real-world clinical diagnosis is a dynamic procedure, which presents dramatic variations of polyps in consecutive frames in a video, as shown in Figure 1 (d-f). The existing image-based methods cannot effectively deal with these variations in a real-time manner. To the end, recently, some video polyp segmentation (VPS) methods (Puyal et al. 2020; Ji et al. 2021a, 2022; Li et al. 2022) have been proposed, aiming at capturing both static semantics and frame-to-frame dynamic semantics to improve the segmentation performance.

These VPS methods employ hybrid 2D/3D architectures (Puyal et al. 2020) or normalized self-attention mechanisms (Ji et al. 2021a, 2022) to highlight temporal information be-

*Corresponding author. Email: hswu@szu.edu.cn.

tween frames. By incorporating temporal consistency, these methods surpass traditional IPS approaches in VPS performance. However, prevalent VPS methods model both static and dynamic semantics at pixel-level with dense features, limiting their effectiveness in addressing background disturbances and dramatic variations among frames.

In this paper, we propose a novel embedding-unleashing framework consisting of a proposal-generative network (PGN) and an appearance-embedding network (AEN). To our knowledge, our method for the first time models VPS as an appearance-level semantic embedding process in order to obtain richer semantic information to tackle background disturbances and dramatic variations in colonoscopy videos. Specifically, the PGN, as a video segmentation network, provides mask proposals, while the AEN is harnessed to produce appearance-level embedding semantics for the PGN. Our designed AEN consists of a CRL module and a CSA module. The former screens reliable background semantics against background disturbances in a frame by constructing global linking between region semantics, while the latter achieves the cross-wise scale alignment to resist dramatic variations by modeling the center-perceived motion dependence. Relying on static and dynamic appearance-level embedding semantics generated from AEN, parameter-free semantic interaction embeds the semantics of AEN into PGN to obtain the final segmentation results. Extensive experiments on CVC-612 (Bernal et al. 2015) and SUN-SEG (Ji et al. 2022) demonstrate that our approach achieves better performance than other state-of-the-art methods. Our major contributions are summarized as follows:

- We propose a novel embedding-unleashing framework consisting of a PGN and an AEN to formulate the VPS as an appearance-level semantic embedding process, aiming at fully unleashing the capability of PGN in VPS task.
- Our specially designed AEN (CRL+CSA) not only resists background disturbances through the linking of region but also resists dramatic variations via center-perceived cross-wise scale alignment.
- Our method achieves state-of-the-art performance on two benchmark datasets and outperforms other competitors on three metrics with a real-time inference speed.

Related Works

Polyp Segmentation

With the development of deep learning, remarkable progress has been made in IPS. FCN-based methods (Brandao et al. 2017; Akbari et al. 2018) and U-Net-based methods (Zhou et al. 2019; Zhang et al. 2020; Wu et al. 2021b) are used to extract accurate semantic information. However, the above methods are limited by fuzzy boundaries. To alleviate the above problem, some methods (Fan et al. 2020; Cheng et al. 2021) based on boundary constraints are used to achieve finer segmentation. Furthermore, some Transformer-based methods (Li et al. 2021; Park and Lee 2022; Ren et al. 2023) also achieve better performance in IPS. However, these image-based methods still suffer from low contrast between background and polyps. Furthermore, these methods ignore the frame-to-frame temporal information and fail to capture the dramatic variations in colonoscopy videos, thus performing poorly in VPS task.

Different from IPS methods, VPS methods need to consider frame-to-frame temporal information. To this end, a hybrid 2/3D CNN (Puyal et al. 2020) is proposed to consider the aggregation of spatial-temporal correlation. PNS-Net (Ji et al. 2021a) proposed a normalized self-attention module to obtain temporal information. Based on PNS-Net, PNS+ (Ji et al. 2022) incorporated a global-to-local learning strategy to balance short-term and long-term dependencies. TC-Net (Xu et al. 2022) are proposed to model temporal correlations based on original video and captured frames. However, modeling static and dynamic semantics from dense features directly using pixel-level manners limits these methods. Different from the above methods, we propose a novel embedding-unleashing framework to address challenges in VPS via the linking of region semantics and the center-perceived cross-wise scale alignment, which for the first time models the VPS as an appearance-level semantic embedding to obtain powerful spatial-temporal semantics.

Embedding-based Semantic Segmentation

The concept of embedding semantics is widely used in various semantic segmentation tasks, which are usually based on multi-network or multi-branch designs. In particular, DED-Net (Galdran, Carneiro, and Ballester 2021) uses two cascaded networks to provide embedding semantics, while AuxNet (Zhang et al. 2022) and SAN (Xu et al. 2023) are proposed to assist another network via providing auxiliary information as embedding semantics. Moreover, multi-branch methods (Zhang, Liu, and Hu 2021; Xu, Xiong, and Bhattacharyya 2023; Su et al. 2023) usually assign different tasks to different branches to obtain embedding semantics with special meaning. However, above embedding-based methods still use pixel-level manners to achieve segmentation from dense features, which is sensitive to background disturbances and dramatic variations in VPS. Therefore, these methods cannot be directly transferred to the VPS task. Different with above methods, we introduce an AEN for the challenges in VPS to model the VPS task as an appearance-level semantic embedding process for the first time. In fact, the appearance-level embedding of AEN can fully unleash the capability of PGN in VPS.

Methods

Overview

The architecture of our proposed method is illustrated in Figure 2, which is a novel embedding-unleashing framework where semantic embedding process is realized between a PGN and an AEN. To resolve background disturbances caused by the low contrast between background and polyps, we propose a CRL module to enhance the static semantics by constructing global linking of region semantics, thereby improving the ability to resist background disturbances. Furthermore, considering that the impact of frame-to-frame dramatic variations of polyps, we propose a CSA module to enhance the dynamic semantics through

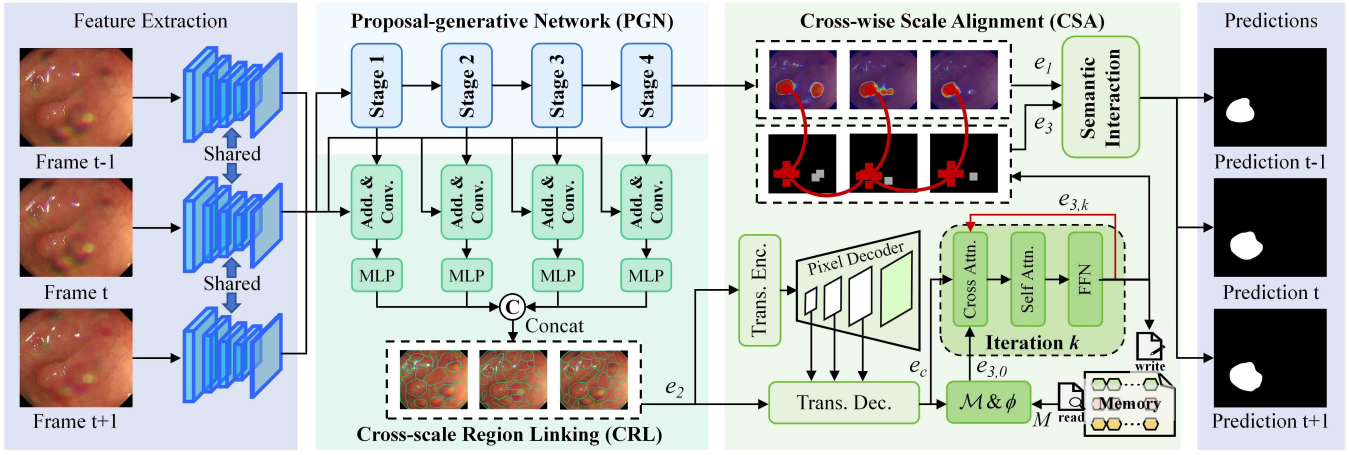

Figure 2: Overview of our proposed embedding-unleashing framework, which consists of a PGN and an AEN (CRL + CSA) for VPS. Our method exploits the mask proposals of PGN and the special embedding semantics of AEN to model VPS as an appearance-level semantic embedding process for the first time. Moreover, our method abandons the previous up-sampling predictor, and the segmentation result is only realized through a parameter-free semantic interaction.

the center-perceived cross-wise scale alignment, thus modeling robust temporal consistency. In this way, our AEN (CRL + CSA) can fully unleash the capability of PGN in VPS by appearance-level embedding semantics. With this embedding-unleashing framework, our proposed method models the VPS as an appearance-level semantic embedding process to learn powerful spatial-temporal information, and achieves promising segmentation accuracy.

Embedding-unleashing Perspective

Before introducing our technical details, we discuss the embedding-unleashing perspective to help readers better understand our method. As discussed above, video polyp semantics include background information (static semantics) and polyp motion information (dynamic semantics), both of which play a crucial role in the performance of VPS. Some early attempts (Puyal et al. 2020; Ji et al. 2021a, 2022) to learn static and dynamic semantics via pixel-level manners from dense features. However, the pixel-level criterion is limited by background disturbances and dramatic variations in VPS due to poor appearance perception.

Modeling the VPS as an appearance-level semantic embedding process improves appearance perception due to both the region linking and the center-perceived scale alignment force the model to understand the VPS task in an appearance view. To reach this goal, we introduce a PGN and an AEN to compose our novel design, called embedding-unleashing framework. In this way, the VPS task can be described as three steps: (1) PGN provides mask proposals; (2) AEN utilizes the features provided by PGN and backbone to generate appearance-level embedding information; (3) parameter-free interaction utilizes mask proposals and embedding information to obtain segmentation results. Different from the previous methods, with the perspective of embedding-unleashing, our method fully leverages the appearance-level embedding semantics of AEN to unleash the capability of PGN in VPS, thus modeling powerful spatial-temporal information.

Specifically, our embedding-unleashing design can be described as three sets of learnable embedding semantics (e_1, e_2 and e_3) and a parameter-free semantic interaction θ. Hence, our segmentation can be expressed as follows:

$$P = \theta(e_1, e_2, e_3), \tag{1}$$

where P represents segmentation results. Currently, the definition of e_1 can be given:

$$e_1 = \tau(F) \in \mathbb{R}^{L \times q \times H \times W}, \tag{2}$$

where τ denotes PGN, and F represents features obtained by backbone, L is the clip length, q is the number of channels, and $H \times W$ is the resolution of the feature map.

In training stage, we define $e_i = f_i(\pi_i)$, where f_i and π_i represent the generation process of e_i and the parameters used respectively, and $i \in \{1, 2, 3\}$. In this way, the learning objective of our proposed framework is defined as:

$$\min_{\mathcal{P}} \mathcal{L}_1 + \mathcal{L}_2 + \cdots + \mathcal{L}_n, \tag{3}$$

where $\mathcal{L}_j$ $(1 \leq j \leq n)$ represents loss function used, and $\mathcal{P}$ is the set of $\{\pi_i\}_{i=1}^{3}$.

Cross-scale Region Linking

The quality of background information can be muddied by background disturbances present in polyp frames. To mitigate these disturbances, we propose a CRL module to model reliable background information. Some methods use attention mechanism to establish pixel-to-pixel global connections to model background information (Ji et al. 2021a, 2022; Fan et al. 2020). However, modeling background information under pixel-to-pixel manner is limited by the low contrast between background and polyps. Considering the great pixel similarity of background and polyp at the static semantic level, different from previous methods, we establish region linking semantics with a cross-scale manner, and

use this global linking of region to require the model to view the feature map at the appearance-level instead of pixel-level, thereby resisting background disturbances.

Our CRL module takes as input the stage features $F^{'} = \{F_1^{'}, F_2^{'}, F_3^{'}, F_4^{'}\}$ output by PGN and the original features $F = \{F_1, F_2, F_3, F_4\}$ output by backbone. For brevity of description, we only illustrate $F_i^{'} \in \mathbb{R}^{L\times C\times H\times W}$ and $F_i \in \mathbb{R}^{L\times C\times H\times W}$ because of similar operations on other features, where C is the channel number of stage features. To augment $F_i^{'}$ at the static semantic level, we first add F_i as the augmentation source to $F_i^{'}$ with element-wise manner. Then, we use the 1×1 convolutional layer to expand the information in the class dimension, and two 3×3 convolutional layers to share spatial information for a group of pixels to model region-sharing background information. To remove the disturbance caused by the low contrast between the background and the polyp, we can model the sharing-to-sharing global linking between regions to require the model to view the feature map at appearance level. Specifically, we leverage a simple MLP project to pull this inter-region sharing into the same embedding space. In this way, a spatial point in the embedding space can represent the sharing of multiple regions, and of course the entire embedding vector can establish the linking of all regions.

Reviewing features under this appearance-level perspective, the low-contrast disturbance between background and polyps can be effectively reduced. In this way, we can get and define our second set of embedding semantics called the region-linking semantics, which are described as follows:

$$e_2^i = \{e_2^i(s) \in \mathbb{R}^{L\times c\times D}\}_{s=1}^{S}, \tag{4}$$

where S is the number of region linking semantics, D denotes the size of embedding space, and c is the number of classes. Finally, we concatenate e_2^i into e_2 as follows:

$$e_2 = Concat[e_2^i] \in \mathbb{R}^{L\times c\times q\times D}, \tag{5}$$

where $i \in \{1, 2, 3, 4\}$, and $q = 4S$.

To sum up, e_2 contains the background information, which is used to refine the static semantic. It resists background disturbances by region linking of visual features at the appearance level and learns to screen reliable background semantics by forcing the model to focus on appearance semantics. Furthermore, as a cross-scale connection, e_2 has the ability to achieve background refinement.

Cross-wise Scale Alignment

In our task, continuous polyp frame information has two import implications. Firstly, the information across frames enhances perception and tracking, and reduces temporal inconsistency. Secondly, the information allows us to accurately define the motion state of polyps, which is beneficial to lock more reliable semantics. To comprehensively exploit the information between consecutive frames, previous attention-based methods (Ji et al. 2021a, 2022) which model the global connection by viewing the pixel in feature map of each frame as visual token. However, due to the pixel-level similarity of the same polyp between consecutive frames, directly modeling dependencies from dense features creates redundancy. Furthermore, dramatic variations resulting in long spatial distances between associated tokens require those methods to model dense features into long sequences, which may limit the performance of attention mechanisms as demonstrated in (Chu et al. 2021; Heo et al. 2021; Liu et al. 2021). Compared with above methods, we propose a CSA module which leverages a set of learnable polyp-center semantics for scale alignment to model center-perceived motion dependencies. By this way, our method is able to more stably capture the motion state of polyps because the center-perceived dependence is more friendly to dramatic variations than the pixel-perceived dependence.

The fact that videos can be represented by a set of vectors represented by object centers rather than pixel-level information as demonstrated in (Heo et al. 2022, 2023), which encourages us to directly align spatial points on the embedding space to map the motion state of polyps. Therefore, we directly exploit e_2 to establish temporal consistency.

As shown in Figure 2, taking e_2 as input, we first model dynamic information as center-perceived polyp motion information. We also use the Transformer (Vaswani et al. 2017) layers like (Heo et al. 2022, 2023) to center polyp motion information. In addition, we introduce a pixel-decoder (Cheng et al. 2022) to generate learnable per-position embedding bias for the embedding space to cope with the dramatic variations in VPS. In this way, centralized motion information is anti-redundant, and friendly to dramatic variations. Finally, we express the process of polyp motion information centralization as follows:

$$e_c = \mathcal{D}_2(e_2, \mathcal{D}_1(\mathcal{E}(e_2))), \tag{6}$$

where $\mathcal{E}$ represents Transformer encoder, $\mathcal{D}_1$ denotes pixel-decoder, and $\mathcal{D}_2$ is Transformer decoder, respectively.

Given e_c rich in polyp-center motion information, we further introduce the joint attention and memory mechanism to model center-to-center alignment instead of pixel-to-pixel. Specifically, to obtain long-term motion perception, we first use a series of matrix operations to implant clip-level memory information $M \in \mathbb{R}^{m\times L\times c\times q\times D}$ into e_c as follows:

$$e_c^{'} = \phi(\mathcal{M}(e_c, M)) \in \mathbb{R}^{L\times c\times q\times D}, \tag{7}$$

where m is the memory length, $\mathcal{M}$ is a combination of dot product and dimension summation, and ϕ is the Softmax function used to normalize the memory information. Then we use cross-attention to establish clip-wise center alignment between memory information and current information (Ke et al. 2021; Han et al. 2022; Heo et al. 2022, 2023). In this way, clip-wise center alignment enables our model to perceive the motion state of the entire video, thereby resisting dramatic variations in VPS. In addition, considering the overall perception of the video as the only motion information reference leads to many uncertainties due to the harsh motion state will cause the current clip to be irrelevant to the memory information soon. Therefore, we also introduce a simple but effective self-attention mechanism (Sun et al. 2022; Su et al. 2023) to model frame-wise center alignment. With this frame-wise center alignment, our model is able to mine effective temporal information from adjacent frames

rather than the entire video, thus also being robust against dramatic variations in VPS.

To summarize, two different alignments (clip-wise and frame-wise) enable our model to obtain robust temporal consistency. For long-term temporal consistency, the clip-wise center alignment represents the perception of the entire video, and for short-term temporal consistency, the frame-wise center alignment enables the segmentation results to be smooth across consecutive frames. So far, we can express the alignment process of the polyp-center, while giving the definition of $e_3 = e_{3,m} \in \mathbb{R}^{L\times c\times q\times D}$ via m iterations:

$$e_{3,k} = \mathcal{F}_{ffn}(\mathcal{F}_{sa}(\mathcal{F}_{ca}(e_{3,k-1}, e_c))), e_{3,0} = e_c^{'}, \quad (8)$$

where $\mathcal{F}_{ca}$ is cross-attention, $\mathcal{F}_{sa}$ is self-attention, $\mathcal{F}_{ffn}$ is feed-forward network, and $k \in \{1, 2, \cdots, m\}$. Note that because CRL and CSA are decoupled, e_3 not only has reliable background information, but also has temporal information that is resistant to dramatic variations in VPS.

Parameter-free Semantic Interaction

Finally, we introduce the parameter-free semantic interaction process θ. After obtaining e_3 enriched with background refinement and accurate polyp motion state, we employ a straightforward yet effective dot product and dimension-wise addition to embed the rich semantics from e_3 into e_1, thereby completing our final semantic interaction. In this way, the VPS performance of PGN can be fully unleashed by the special embedding semantics of AEN. The entire semantic interaction process can be described as follows:

$$P = \varphi(\sum_{d=1}^{D} e_1 \cdot e_3) \in \mathbb{R}^{L\times c\times H\times W}, \quad (9)$$

where $\cdot$ represents the dot product, and φ is the Sigmoid function. Different with previous methods, we do not require an additional up-sampling predictor.

Loss Functions

For the supervised learning of the model, we adopt the binary cross-entropy loss $\mathcal{L}_{bce}$ like (Ji et al. 2022) to guide the convergence process of the model. Furthermore, the Dice loss $\mathcal{L}_{dice}$ (Milletari, Navab, and Ahmadi 2016) becomes another part of our loss, given the irregularity in the distribution of polyps across sequences. In summary, we can formulate the total loss with two components as follows:

$$\mathcal{L}_{total} = \frac{1}{L}\sum_{i=1}^{L} \alpha \cdot \mathcal{L}_{bce}(P_i, G_i) + \beta \cdot \mathcal{L}_{dice}(P_i, G_i), \quad (10)$$

where P_i represents the prediction, and G_i represents the corresponding ground truth (GT).

Experiments

Datasets and Evaluation Metrics

We evaluated our method on two public datasets, including CVC-612 (Bernal et al. 2015) and SUN-SEG (Ji et al. 2022).

- **CVC-612** contains 612 frames from 31 colonoscopy sequences with a resolution of 384×288. However, it is not strictly a video dataset because most frames from the same sequence are not really adjacent.
- **SUN-SEG** is the latest large-scale dataset for VPS, which contains $158,690$ frames from $1,013$ sequences. In fact, the data used is $49,136$ frames from 285 sequences, including a training dataset with $19,544$ frames from 112 sequences and two test datasets (SUN-SEG-Easy with $17,070$ frames from 119 sequences and SUN-SEG-Hard with $12,522$ frames from 54 sequences).

In our experiments, we introduced three widely used metrics to evaluate our method, including structure measure (S_α, $\alpha = 0.5$), maximum intersection over union (maxIoU), and maximum dice coefficient (maxDice) like (Ji et al. 2021a).

Implementation Details

Our proposed method is designed on top of $detectron2$ (Wu et al. 2019) with a single NVDIA GeForce RTX 3090TI GPU. Res2Net-50 (R2-50) (Gao et al. 2019) and HRNet-W48 (H-W48) (Sun et al. 2019; Wang et al. 2020) are used as our backbones, which are all pre-trained on ImageNet. The input is an arbitrary number of post-processed clips of length 3 that compose a complete colonoscopy video. All frames are unified to a resolution of 320×448. The AdamW and poly learning rate schedule are used for optimizing parameters, with an initial learning rate $= 0.0001$ and a weight decay $= 0.1$. In training stage, we set the number of iterations $30K$ and $3K$ for SUN-SEG and CVC-612, respectively. We set $m = 3$ for Equation 8. The total loss $\mathcal{L}_{total}$ is balanced with $\alpha = 5$ and $\beta = 2$. For SUN-SEG, we separate 20% from the training set as the validation set. For CVC-612, we split the training set, validation set, and test set with a ratio of $6:2:2$.

Ablation Studies

All ablation experiments were designed on the SUN-SEG. We apply CFFM (Sun et al. 2022) and Res2Net-50 as the PGN and backbone, respectively, while training the model using the parameters in the implementation details.

Stability of AEN. To verify the stability of AEN, we composed different embedding-unleashing designs by replacing PGN. As shown in Table 1, each of the embedding-unleashing frameworks achieves improvements in all metrics compared to segmentation using PGN alone, showing that our AEN can indeed provide useful semantic information for PGN. We also verify the effectiveness of AEN through the visualization of t-SNE. As shown in Figure 3, our AEN is able to achieve a clear division between lesion regions and the opposite.

Impact of CRL module. To verify the importance of CRL in the AEN, we explore the performance by removing the CRL module. In fact, we replace the CRL with a linear layer to ensure that features can be processed by CSA module. As shown in Table 2 and Figure 4, the improvement in metrics and the optimization in visualization prove that the global linking of region semantics is effective.

Method	Class	Backbone	SUN-SEG-Easy			SUN-SEG-Hard			CVC-612		
			S_α	maxDice	maxIoU	S_α	maxDice	maxIoU	S_α	maxDice	maxIoU
PIDNet	NIS	-	0.798	0.710	0.642	0.776	0.703	0.626	0.883	0.861	0.788
FSNet	NVS	R2-50	0.781	0.729	0.646	0.768	0.722	0.633	0.878	0.851	0.777
GenVIS	NVS	R2-50	0.812	0.768	0.685	0.798	0.743	0.649	0.893	0.751	0.783
CFFM	NVS	R2-50	0.817	0.772	0.692	0.810	0.747	0.657	0.898	0.857	0.792
PraNet	IPS	R2-50	0.778	0.683	0.605	0.752	0.654	0.562	0.882	0.848	0.778
META-UNet	IPS	R2-50	0.803	0.763	0.678	0.796	0.741	0.645	0.905	0.864	0.793
PNSNet	VPS	R2-50	0.793	0.711	0.638	0.781	0.703	0.624	0.892	0.855	0.786
PNS+	VPS	R2-50	0.823	0.774	0.698	0.812	0.753	0.663	0.903	0.863	0.794
AEN + GenVIS	VPS	R2-50	**0.834**	**0.792**	**0.711**	**0.821**	**0.773**	**0.674**	**0.905**	**0.866**	**0.795**
AEN + PNS+	VPS	R2-50	**0.836**	**0.794**	**0.714**	**0.820**	**0.772**	**0.676**	**0.912**	**0.872**	**0.802**
AEN + CFFM	VPS	R2-50	**0.847**	**0.810**	**0.728**	**0.823**	**0.786**	**0.686**	**0.916**	**0.878**	**0.805**
AEN + CFFM	VPS	H-W48	**0.869**	**0.830**	**0.746**	**0.832**	**0.804**	**0.694**	**0.924**	**0.882**	**0.812**

Table 1: Quantitative comparison with different state-of-the-art methods on SUN-SEG and CVC-612 test sets.

CRL	CSA	SUN-SEG-Easy		SUN-SEG-Hard	
		maxDice	maxIoU	maxDice	maxIoU
		0.772	0.692	0.747	0.657
✓		0.786	0.704	0.766	0.671
	✓	0.793	0.716	0.774	0.682
✓	✓	**0.810**	**0.728**	**0.786**	**0.686**

Table 2: Statistical comparison of our ablation studies over different components on SUN-SEG test set.

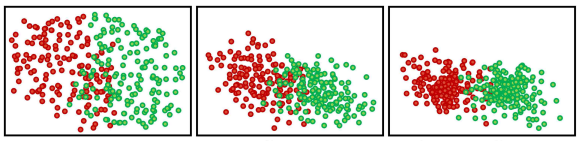

Figure 3: t-SNE visualization of features. Green represents lesion regions, while red represents the opposite.

Ablation of CSA module. To demonstrate the effectiveness of CSA module, we also explore performance changes by means of deletion. As shown in Table 2, the improvement on all metrics indicates that our CSA module can provide powerful and stable temporal information. Figure 4 also visually demonstrates that our CSA module can optimize mask proposals to provide more accurate segmentation results. Moreover, we also consider the impact of the two alignments (clip-wise and frame-wise) on performance. As shown in Table 3, both manner lead to performance gains, suggesting that the center-perceived motion dependence can provide more reliable motion semantics by modeling clip-wise and frame-wise scale alignments.

Comparison with State-of-the-art Methods

We compared our method with eight state-of-the-art competitors over the datasets SUN-SEG and CVC-612, including PraNet (Fan et al. 2020), FSNet (Ji et al. 2021b), PNSNet (Ji et al. 2021a), PNS+ (Ji et al. 2022), CFFM (Sun et al. 2022), META-UNet (Wu, Zhao, and Wang 2023), GenVIS (Heo et al. 2023), and PIDNet (Xu, Xiong, and Bhattacharyya 2023). The above methods can be divided into four categories: (1) natural image segmentation (NIS), (2) natural video segmentation (NVS), (3) IPS, and (4) VPS. All controllable training parameters are set to the same value.

The comparison results between our method and above state-of-the-art methods on the SUN-SEG and CVC-612 are shown in Table 1. All R2-50-based embedding-unleashing designs outperform other state-of-the-art methods in all metrics, illustrating the robustness of our method. Moreover, the heavyweight H-W48 also brings reasonable gains.

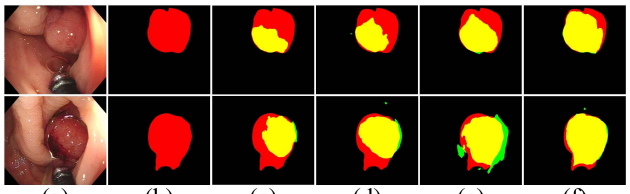

Figure 4: Visualization of module ablation on SUN-SEG-Hard test set. (a) Frame. (b) GT. (c) CFFM. (d) Ours only w/ CRL. (e) Ours only w/ CSA. (f) Ours w/ (CRL + CSA). Red, green and yellow represent the GT, prediction and their overlapping regions, respectively.

Alignment	SUN-SEG-Easy		SUN-SEG-Hard	
	maxDice	maxIoU	maxDice	maxIoU
w/o alignment	0.790	0.702	0.771	0.672
only frame-wise	0.798	0.713	0.778	0.679
only clip-wise	0.803	0.716	0.776	0.676
cross-wise	**0.810**	**0.728**	**0.786**	**0.686**

Table 3: Ablation studies of different scale alignments on SUN-SEG test set. We perform different alignments by removing the attention layer with specific capabilities.

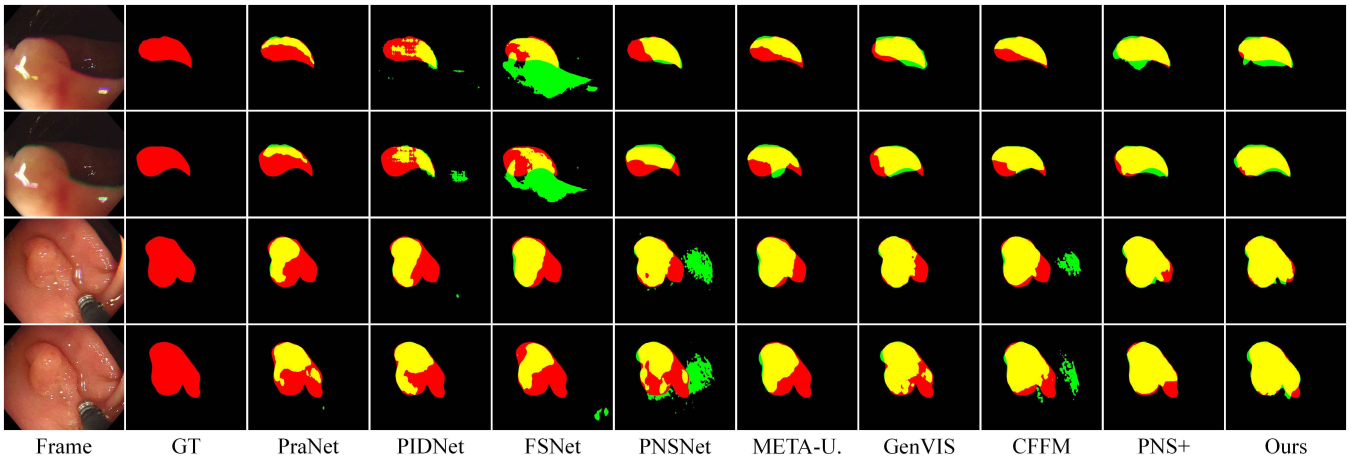

Figure 5: Visual comparison with different state-of-the-art methods on the SUN-SEG-Easy and SUN-SEG-Hard test sets. Red, green and yellow represent the GT, prediction and their overlapping regions, respectively. Ours is AEN+CFFM (Res2Net-50).

Method	SUN-SEG-Easy			
	maxDice	Param. (M)	GFLOPs	FPS
CFFM	0.772	26.49	82.98	45
AEN + CFFM	0.810	33.81	93.48	39
GenVIS	0.768	38.99	95.85	34
AEN + GenVIS	0.792	45.36	107.56	27
PNS+	0.774	9.79	53.24	65
AEN + PNS+	0.794	18.71	66.45	56

Table 4: Performance-efficiency comparison with the state-of-the-art methods on SUN-SEG-Easy test set with 320 × 448 resolution. Res2Net-50 is selected as the backbone.

The focus of this paper is to improve accuracy. To comprehensively analyze the strengths of our method, we also conduct some analysis between accuracy and efficiency. In Table 4, all embedding-unleashing designs achieve considerable performance gains with little overhead. In addition, we also compare the efficiency and performance with other state-of-the-art methods. As shown in Figure 6, our embedding-unleashing framework achieves a performance-efficiency trade-off with small overhead. In fact, since AEN reuses the features of backbone and PGN in a non-dense-feature manner, the efficiency of the embedding-unleashing framework is mainly determined by PGN.

We also perform qualitative comparison with state-of-the-art methods. Figure 5 shows that our method performs better segmentation in coping with background disturbances and dramatic variations. It verifies that our method can obtain more reliable semantic information for VPS.

Discussions and Limitations

Although we only compose three embedding-unleashing designs in our experiments, we believe our AEN has the potential to form more excellent VPS methods with other video segmentation networks. Moreover, our method still has some limitations. As shown in Figure 7, facula interference (a-b), small polyps with extremely low-contrast (c-d), and dramatic shape (e-f) may limit our method.

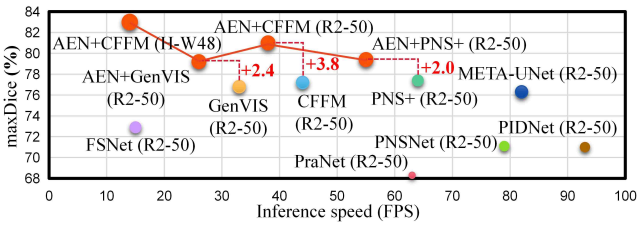

Figure 6: Performance-efficiency comparison with other state-of-the-art methods on SUN-SEG-Easy test set.

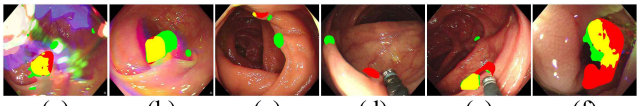

Figure 7: Failure cases. Red, green and yellow represent the GT, prediction and their overlapping regions, respectively.

Conclusion

In this paper, we propose a novel embedding-unleashing framework consisting of a PGN and an AEN, which for the first time models the VPS task as an appearance-level semantic embedding process to improve segmentation performance. PGN serves as a video segmentation network to provide mask proposals. The AEN (CRL + CSA) we designed obtains appearance-level embedding semantics to address the challenges in VPS through region linking and center-perceived cross-wise scale alignment. Finally, segmentation results are obtained by a parameter-free semantic interaction between mask proposals of PGN and embedding semantics of AEN, thus unleashing the capability of PGN in VPS. Our method achieves state-of-the-art results on both the CVC-612 and SUN-SEG test sets with a real-time inference speed.

Acknowledgments

This work was supported partly by National Natural Science Foundation of China (Nos. 62273241 and 61973221), Natural Science Foundation of Guangdong Province, China (No. 2019A1515011165), and the General Research Fund of Hong Kong Research Grants Council (Project no. 15218521).

References

Akbari, M.; Mohrekesh, M.; Nasr-Esfahani, E.; Soroushmehr, S. R.; Karimi, N.; Samavi, S.; and Najarian, K. 2018. Polyp segmentation in colonoscopy images using fully convolutional network. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 69–72. IEEE.

Bernal, J.; Sánchez, F. J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; and Vilariño, F. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43: 99–111.

Brandao, P.; Mazomenos, E.; Ciuti, G.; Caliò, R.; Bianchi, F.; Menciassi, A.; Dario, P.; Koulaouzidis, A.; Arezzo, A.; and Stoyanov, D. 2017. Fully convolutional neural networks for polyp segmentation in colonoscopy. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, 101–107. SPIE.

Center, M. M.; Jemal, A.; Smith, R. A.; and Ward, E. 2009. Worldwide variations in colorectal cancer. *CA: a cancer journal for clinicians*, 59(6): 366–378.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1290–1299.

Cheng, M.; Kong, Z.; Song, G.; Tian, Y.; Liang, Y.; and Chen, J. 2021. Learnable oriented-derivative network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, 720–730. Springer.

Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; and Shen, C. 2021. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34: 9355–9366.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020. Pranet: Parallel reverse attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*, 263–273. Springer.

Galdran, A.; Carneiro, G.; and Ballester, M. A. G. 2021. Double encoder-decoder networks for gastrointestinal polyp segmentation. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I*, 293–307. Springer.

Gao, S.-H.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; and Torr, P. 2019. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2): 652–662.

Han, S. H.; Hwang, S.; Oh, S. W.; Park, Y.; Kim, H.; Kim, M.-J.; and Kim, S. J. 2022. Visolo: Grid-based space-time aggregation for efficient online video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2896–2905.

Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; and Oh, S. J. 2021. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11936–11945.

Heo, M.; Hwang, S.; Hyun, J.; Kim, H.; Oh, S. W.; Lee, J.-Y.; and Kim, S. J. 2023. A generalized framework for video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14623–14632.

Heo, M.; Hwang, S.; Oh, S. W.; Lee, J.-Y.; and Kim, S. J. 2022. VITA: Video Instance Segmentation via Object Token Association. In *Advances in Neural Information Processing Systems*.

Ji, G.-P.; Chou, Y.-C.; Fan, D.-P.; Chen, G.; Fu, H.; Jha, D.; and Shao, L. 2021a. Progressively normalized self-attention network for video polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, 142–152. Springer.

Ji, G.-P.; Fu, K.; Wu, Z.; Fan, D.-P.; Shen, J.; and Shao, L. 2021b. Full-duplex strategy for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4922–4933.

Ji, G.-P.; Xiao, G.; Chou, Y.-C.; Fan, D.-P.; Zhao, K.; Chen, G.; and Van Gool, L. 2022. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, 1–19.

Ke, L.; Li, X.; Danelljan, M.; Tai, Y.-W.; Tang, C.-K.; and Yu, F. 2021. Prototypical cross-attention networks for multiple object tracking and segmentation. *Advances in Neural Information Processing Systems*, 34: 1192–1203.

Li, S.; Sui, X.; Luo, X.; Xu, X.; Liu, Y.; and Goh, R. 2021. Medical Image Segmentation using Squeeze-and-Expansion Transformers. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 807–815. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Li, X.; Xu, J.; Zhang, Y.; Feng, R.; Zhao, R.-W.; Zhang, T.; Lu, X.; and Gao, S. 2022. TCCNet: Temporally Consistent Context-Free Network for Semi-supervised Video Polyp Segmentation. *IJCAI-22*, 1109–1115.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, 565–571. Ieee.

Park, K.-B.; and Lee, J. Y. 2022. SwinE-Net: hybrid deep learning approach to novel polyp segmentation using convolutional neural network and Swin Transformer. *Journal of Computational Design and Engineering*, 9(2): 616–632.

Puyal, J. G.-B.; Bhatia, K. K.; Brandao, P.; Ahmad, O. F.; Toth, D.; Kader, R.; Lovat, L.; Mountney, P.; and Stoyanov, D. 2020. Endoscopic polyp segmentation using a hybrid 2D/3D CNN. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*, 295–305. Springer.

Ren, G.; Lazarou, M.; Yuan, J.; and Stathaki, T. 2023. Towards Automated Polyp Segmentation Using Weakly-and Semi-Supervised Learning and Deformable Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4354–4363.

Su, J.; Yin, R.; Zhang, S.; and Luo, J. 2023. Motion-state Alignment for Video Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3570–3579.

Sun, G.; Liu, Y.; Ding, H.; Probst, T.; and Van Gool, L. 2022. Coarse-to-fine feature mining for video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3126–3137.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5693–5703.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10): 3349–3364.

Wu, H.; Chen, G.; Wen, Z.; and Qin, J. 2021a. Collaborative and adversarial learning of focused and dispersive representations for semi-supervised polyp segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3489–3498.

Wu, H.; Xie, W.; Lin, J.; and Guo, X. 2023. ACL-Net: Semi-supervised Polyp Segmentation via Affinity Contrastive Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2812–2820.

Wu, H.; Zhao, Z.; and Wang, Z. 2023. META-Unet: Multi-Scale Efficient Transformer Attention Unet for Fast and High-Accuracy Polyp Segmentation. *IEEE Transactions on Automation Science and Engineering*.

Wu, H.; Zhao, Z.; Zhong, J.; Wang, W.; Wen, Z.; and Qin, J. 2022. Polypseg+: A lightweight context-aware network for real-time polyp segmentation. *IEEE Transactions on Cybernetics*.

Wu, H.; Zhong, J.; Wang, W.; Wen, Z.; and Qin, J. 2021b. Precise yet efficient semantic calibration and refinement in convnets for real-time polyp segmentation from colonoscopy videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2916–2924.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2. Accessed: 2023-06-28.

Xu, J.; Xiong, Z.; and Bhattacharyya, S. P. 2023. PIDNet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19529–19539.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2945–2954.

Xu, Z.; Qiu, D.; Lin, S.; Zhang, X.; Shi, S.; Zhu, S.; Zhang, F.; and Wan, X. 2022. Temporal Correlation Network for Video Polyp Segmentation. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1317–1322. IEEE.

Zhang, R.; Li, G.; Li, Z.; Cui, S.; Qian, D.; and Yu, Y. 2020. Adaptive context selection for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*, 253–262. Springer.

Zhang, Y.; Borse, S.; Cai, H.; and Porikli, F. 2022. Auxadapt: Stable and efficient test-time adaptation for temporally consistent video semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2339–2348.

Zhang, Y.; Liu, H.; and Hu, Q. 2021. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, 14–24. Springer.

Zhou, T.; Zhou, Y.; He, K.; Gong, C.; Yang, J.; Fu, H.; and Shen, D. 2023. Cross-level Feature Aggregation Network for Polyp Segmentation. *Pattern Recognition*, 140: 109555.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2019. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6): 1856–1867.