Polyp Segmentation with Multi-Scale Guidance and Multi-Level Supervision

Yingjie Wu[1,2], Huiqian Li[3,2], Peiliang Huang[4], Jieru Yao[4*],
[1] AHU-IAI AI Joint Laboratory, Anhui University, Hefei, China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China
[3] Institute of Advanced Technology, University of Science and Technology of China, Hefei, China
[4] School of Automation, Northwestern Polytechnical University, Xi'an, China
Email: ucal666@163.com; huiqian@mail.ustc.edu.cn; peilianghuang2017@gmail.com; jieruyao2018@gmail.com;

***Abstract*—Colorectal endoscopy is an effective method for detecting and treating colorectal polyps before they evolve into colorectal cancer. However, accurately segmenting polyps in endoscopic images is challenging due to their diverse appearance in terms of size, color, texture, and indistinct borders with their surroundings. Existing methods usually first employ a single complex encoder to handle all polyps, ignoring the differences in size and shape; each decoding step only uses output from adjacent layers, ignoring important global information; and they only provide supervision to the final output, which results in inefficient segmentation of challenging samples with obscure visual characteristics. These issues lead to poor performance as data-set diversity increases. In order to address these limitations, we propose a multi-scale attention detector that can handle various polyp types and sizes. The detector employs a semantic express module to capture crucial global information and a multi-level and multi-view supervision mechanism to segment polyps. Experiments are conducted on diverse datasets including complex SUN-SEG and smaller Kvasir and CVC, and the results demonstrate that our proposed model achieves state-of-the-art performance.**

***Index Terms*—Colorectal cancer, polyp segmentation, multi-scale guidance, multi-level supervision, attention mechanism**

I. Introduction

Colorectal cancer (CRC), frequently caused by polyps [2], ranks third in the number of new cancer cases over the world [1], posing a serious threat to human health. Early diagnosis is essential for the treatment of CRC [3] since the survival rate of patients with early-stage CRC can reach more than 95%, but less than 35% [47] as the patients rapidly deteriorate to the end stage due to ignorance. Colonoscopy is such an effective technique for CRC detection and prevention since it provides definitive information about the location, appearance, and status of polyps, enabling physicians to remove these polyps before they develop into CRC. Studies have shown that early detection and removal of polyps can reduce the probability of CRC by 30% [4].

Developing automated tools for the important but highly repetitive task of polyp identification and segmentation has long been a popular area of research. However, it is a challenging task due to the diversity of the polyp itself and its faint border with the surroundings [5]. Early polyp segmentation methods relied on a large number of manually designed and extracted features, usually training classifiers [48] between polyp and surroundings based on combinations of color and texture or trying to analyze the fluctuations of the intestinal wall for polyp detection. But all those methods suffer from low accuracy and speed until the advent of deep learning methods and the increasing number of publicly available data sets.

Although deep learning methods have made great progress in polyp segmentation, they still have many shortcomings making them cannot satisfy accuracy and speed at the same time. PRA [19] designed a complex decoder to roughly locate the polyps first, then accurately outline the contours according to the local features, but it will inevitably cause omissions to detect all polyps with a single encoder. U-Net++ [12] and SFA [20] consider multi-level and multi-scale encoders, but only use the last layer of decoder to calculate the loss function, ignoring the importance of strengthening the supervision of the decoding process to distinguish whether uncertain regions belong to foreground or background. Based on video analysis, methods such as PNS [15] and PNS+ [16] that introduce temporal relationships increase the complexity of the model in order to achieve higher accuracy, resulting in their slow inference speed and cannot be used in actual diagnosis. In addition, the above methods only use the features of adjacent layers in the decoding process, ignoring the supplementary effect of global information on localizing features.

In order to overcome the shortcomings of existing methods, we propose a novel and lightweight multi-scale guidance and multi-level supervision model for polyp segmentation. Our insight comes from the following facts: Firstly, the diversity of polyp morphology determines that the encoders of the model should also be designed with different scales. Secondly, the up-sampling operations are essentially imagination of the lost details and the single supervision imposed on the last decoder layer is insufficient to deal with this problem especial on inconspicuous polyps. It is therefore important to add supervision in time to remind the model which parts belong to the foreground. Furthermore, global features are important in locating objects, providing abstract semantic information about objects, and should not be wasted. We employ a multi-scale convolutional attention encoder to cope with polyps of varying sizes and shapes. A semantic express module is used to pass

This work was supported by the Institute of Artificial Intelligence,Hefei Comprehensive National Science Center Project under Grant(21KT008).

the diluted global semantic information directly to the decoder. Besides, we propose a multi-level and class activate map based multi-view learning strategy to strengthen the supervision of decoding and encoding process. In brief, the contributions of this paper are threefold:

- (1) We propose a novel polyp segmentation model that can deal with morphological differences in polyps such as size and shape, and can better distinguish hard polyps with inconspicuous visual features.
- (2) To achieve these goals, we employ an multi-scale convolutional attention module and a semantic express module to enhance the U-Net. The proposed multi-level and multi-view learning strategy make features more distinguishable.
- (3) Experimental results demonstrate that the proposed approach can outperform existing state-of-the-art methods on SUN-SEG [16], kvasir [36] and CVC [34] especially on hard samples. Moreover, it achieves 30 FPS real-time prediction with a 320x320 input size using one NVIDIA 1080Ti GPU.

II. Related Works

A. Deep Learning based Medical Image Segmentation

The early deep learning medical image segmentation method is usually based on FCN [49], which first performs down-sampling to extract features and up-sample the last few layers of feature maps to the original size as the prediction result. While various attempts have been made in this field, it was not until the introduction of U-Net [6] that significant improvements were observed. Two corresponding branches are consisted in this architecture: the contracting path, which performs down-sampling and provides global information, and the expansion path, which performs up-sampling and enable seize localized information. One key feature that make U-Net particularly superior is skip connections which concatenate features from the contracting path [7]. This approach serves two purposes: firstly, it eliminates redundant noise caused by up-sampling operations in the decoder, secondly, it mitigates the adverse effects of information loss resulting from repeated pooling during the encoding process. Variants such as ResUNet [8] and ResUNet++ [9] have been proposed based on the renowned ResNet [10]. These variants aim to overcome the challenges associated with training very deep neural networks, as increasing the depth of layers can rather lead to recession and even degradation in performance. In these models, a residual connection is applied before the down-sampling or up-sampling manipulation during the homologous branches [11]. The input to the former convolution layer is added to the result of the later convolution layer at each block. The utilization of residual operation assists to solve issues like gradient explosion or disappearance and mitigates the degradation problem, thereby enabling the design of deeper neural networks. U-Net++ [12] is an advanced extension inspired by recently DenseNet [13]. U-Net++ [12] introduces redundant skip connection units between every two corresponding blocks. Each skip connection unit accepts all feature maps from previous units at the identical level, as well as up-sampled feature maps from the direct lower unit [14]. It assumes that both low-level and high-level encoders play crucial roles, and determines the optimal number of skip connections through an adaptive learning process.

But these methods adopt a single-scale filter, ignoring the diversity of polyps, while our encoder adopts convolutions of different sizes to extract features in parallel to fully detect polyps of various shapes and sizes. Furthermore, they only apply supervision at the last decoder layer, which poor in segment harder samples that hold an indistinct border with the surrounding mucosa. We apply supervision at each encoder to enhance the distinguish of uncertain regions.

B. Colonoscopy Polyp Segmentation

Methods before the deep learning era relied on extracting a combination of hand-crafted features, such as color and texture to distinguish a polyp from its surroundings [50]. However, these methods perform poorly due to the low quality of hand-extracted features. After the introduction of CNN, the aforementioned U-Net [6] and its variants such as U-Net++ have become the basic models for polyp segmentation. PNS [15] and PNS+ [16] introduced a self-attention module to strengthen the original architecture. They incorporated multiple adjacent frames of images into the model to explore the temporal and spatial relationships of the images. However, this approach led to a waste of hardware resources, and the cross-attention mechanism was just applied to global features, neglecting the local details [17]. To enhance specific semantic information of features, researchers designed task-driven networks that leverage the inherent characteristics of polyp data sets. For instance, ACS [18] employed an attention-gating mechanism to balance the impact of global classification information and local semantic information on samples with varying complexities. When the samples are complex, the model dynamically increased the weight of local information to achieve improved segmentation results. PRA [19] adopted a method that emulates a doctor's diagnostic process. It firstly predicted the rough area then used a reverse attention mechanism to obtain refined boundary clues for polyp segmentation. SFA [20] decomposed the polyp segmentation task into two steps: first locating the approximate position of the polyp, and then accurately determining the boundary. Additionally, they introduced a new loss function that focuses more on boundary to identically enhance both inner region and boundary detection. However, it is important to note that these approaches rely heavily on their own assumptions rather than established facts.

Existing models either apply the attention mechanism to explore the semantic connection between adjacent frames , adding excessive parameters to the model, or use it only when the encoder interacts with the decoder like ACS. However, we use the attention mechanism throughout the encoding process. In addition, unlike these methods that only use local information for decoding, we emphasize the role of global features in the decoding process.

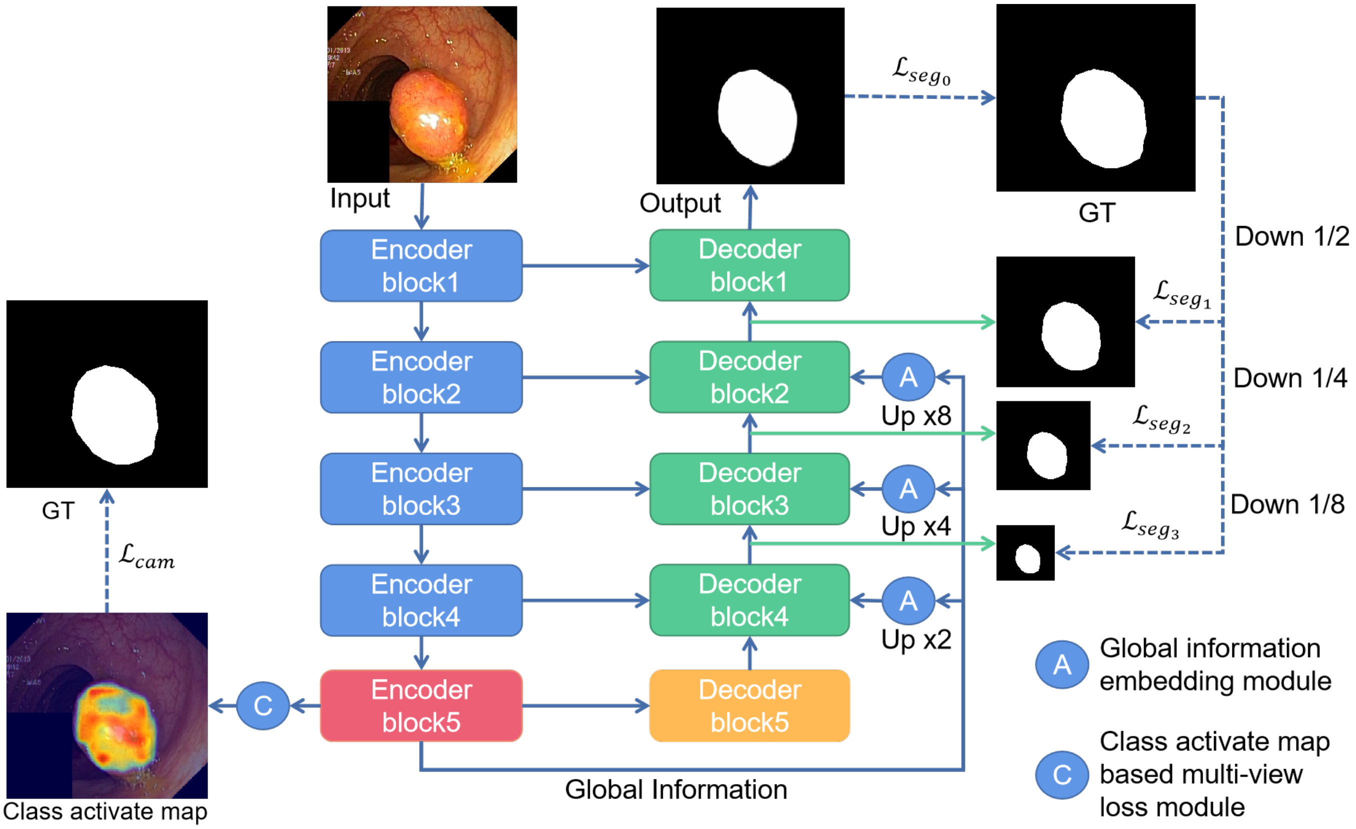

Fig. 1. The overall architecture of our method which consists of a multi-scale convolutional attention based encoder and a decoder with semantic express module which directly preserve and embed global information through the global information embedding module. This module is designed to effectively segment smaller and less visible polyps. All of details will be discussed in section 3.

III. Methods

A. The Overall Framework

The architecture of our enhanced U-Net model is shown in Fig.1. We adopt a multi-scale attention-based encoder which has four sequential blocks that work like an FPN [21] to cope with polyps of varying size and shape. The decoder branch also has four blocks and each of them receives three feature maps from the semantic express module(SEM), corresponding encoder, and lower-level decoder. The later two equal-sized feature maps will be channel-wise concatenated while the feature from SEM will be processed by an embedding module. The SEM captures and preserves the global context information and densely concatenates each layer in the decoder path. The embedding module bridge the semantic gap caused by rough global features. Besides, our model outputs all the decoding results and compare them with resized gt to construct multi-level losses to better distinguish fuzzy areas. Finaly, we use a class activate map based multi-view loss to make the features more focused on areas where polyps are likely present.

B. Multi-scale Convolutional Attention Module

The encoder follows a pyramid structure similar to FPN [21] to down-sample and extracts image features while preserving the spatial structure of the image. Inspired by VIT [22], [23], we design a multi-scale convolutional attention(MSA) module as the backbone of the encoder, which combines the advantages of the two basic architectures in the visual field: the convolutional structure focuses on the local features of the image and can preserve the spatial features of the image, while the attention mechanism can make up for the neglect of the global features of the convolutional architecture [24]. The Depth-wise convolution proposed by mobile-net [25] is an efficient and lightweight new form of CNN. It only uses one convolution kernel for down-sampling processing for each channel of the input feature map, and then uses 1x1 point-wise convolution to adjust the output channel to maximize the receptive field without introducing too many parameters. As shown in Fig.2, each multi-scale and attention block consists of three components: the first is a 5x5 depth-wise convolution to achieve down-sampling, followed by a batch normalization layer [26] since batch normalization gains more for the segmentation performance than layer normalization. To deal with polyps of different sizes, the results of down-sampling will be processed in parallel by three depth-wise convolutions with sizes to extract features at different scales. After the channel is adjusted by 1x1 point-wise convolution, all the feature maps mentioned above and the unprocessed original map will be concatenated through the channel, and

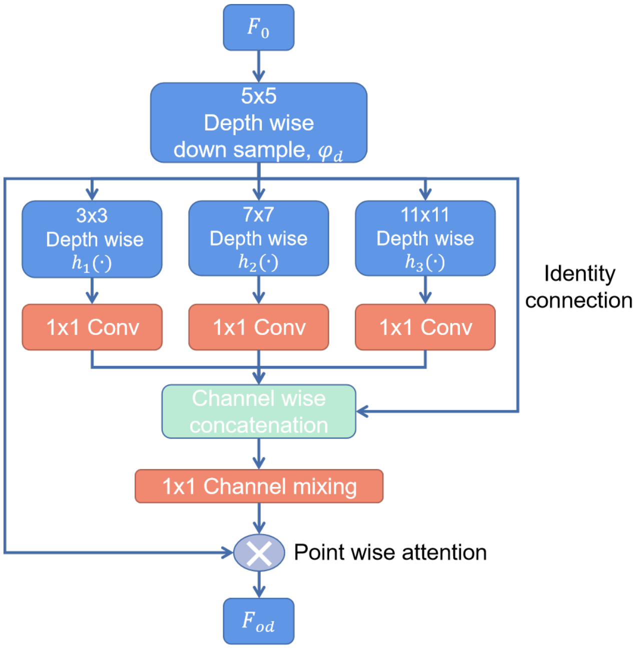

Fig. 2. Input features down-sampled in MSA will pass through a multi-scale detector. The detection result is used as the attention weight of the feature map after downsampling

then adjusted by 1x1 convolution as the attention weight, and these weights are compared with the previous down-sampling results element by element. The product becomes the final output of the encoder. Mathematically, the encoder proceeds as follows:

$$\mathbf{F_1} = \varphi_d\left(\mathbf{F_0}\right), \tag{1}$$

$$\mathbf{F_{oe}} = \left(\mathbf{F_1} + \sum_{i=1}^{3} h_i\left(\mathbf{F_1}\right)\right) \otimes \mathbf{F_1}. \tag{2}$$

Among them, F_1 represents the result of down-sampling, $\varphi_d\left(\cdot\right)$ indicates processing F_1 with a 5x5 depth wise convolution and performing down-sampling. $h_i\left(\cdot\right)$ are i-th multi-scale depth wise convolution without down-sampleing. $\sum_{i=1}^{3}$ means summing the results of three multi-scale detectors. $\otimes$ represents the attention mechanism, F_{oe} is the output of the encoder of this layer.

C. Semantic Express Module

As shown in Fig.3, a decoder block consists of a series of convolutional blocks stacked together and progressively up-sampled back to the original size using bilinear interpolation. It has been pointed out in [27]that global and aggressive semantic information is instrumental for discovering the specific locations of objects. Meanwhile, lower or mid-level features are essential to bringing the deep extracted features from a coarse and condensed state to a fine level [28]. But one of the important problems with the U-shaped architecture to be resolved is that higher-level features are gradually diluted as they are transferred to the lower layers, and the model tends to focus on local features at the expense of capturing the content of the image as a whole. Regarding the lack of high-level semantic information for fine-level feature maps in the bottom-up decoder, we adopt a semantic express module, it consists of a sub-module that holds high-level semantic information and a series of information flows (IFs). This module copies the output of the last encoder and the IFs are independent of the decoder, using bilinear interpolation up-sampling to pass the high-level semantic information directly to each decoder and participate in the decoding operation together. In order to make the high-magnification up-sampling process smoother, we use a global information embedding sub-module in Figure 3, which first performs up-sampling on the corresponding scale of the global features, and then performs the following processing on the aforementioned results in parallel in multiple sub-spaces. It first samples and then up-samples the features at the same ratio to fully cope with the different sizes of objects in the global features, making their features fully fused. In this way, we explicitly increase the weight of the global semantic information in each part of the bottom-up path to ensure that the positional information is not diluted when building the U-shape network.

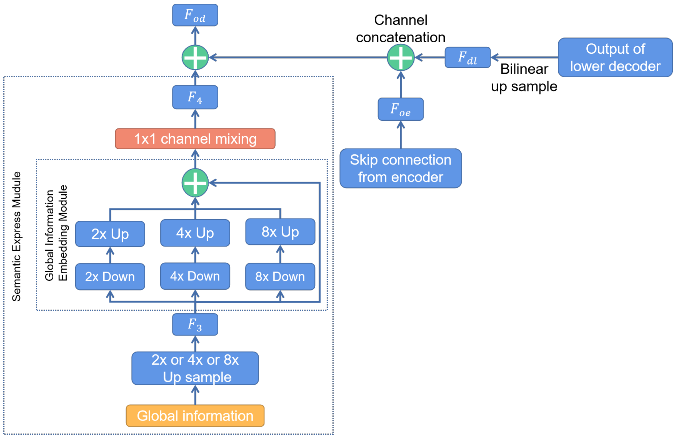

Fig. 3. Each decoder block accepts upsampled lower decoder outputs, skip connection from peer encoders, and the global information passed from the semantic express module.

$$\mathbf{F_4} = \mathbf{F_3} + \sum_{i=1}^{3}\left(g_i\left(\sum_{j=1}^{3}\phi_j\left(\mathbf{F_3}\right)\right)\right), \tag{3}$$

$$\mathbf{F_{od}} = \psi_{3\times3}\left(f_c\left(\mathbf{F_{dl}}, \mathbf{F_{oe}}\right) + \mathbf{F_4}\right). \tag{4}$$

Among them F_3 represents the output of SEM, F_4 is the result of FAM, $\phi_j\left(\cdot\right)$ and $g_i\left(\cdot\right)$ represents down-sampling or up-sampling using depth-wise convolutions, $\sum_{j=1}^{3}$ and $\sum_{i=1}^{3}$ represent the summation of down-sampling or up-sampling results respectively. $f_c\left(\cdot\right)$ refers to the function connecting two feature maps in the channel direction, $\psi_{3\times3}$ is a is a convolutional layer with a kernel size of $\{3\times3\}$to eliminate the semantic gap, F_{od} is the result of this decoder.

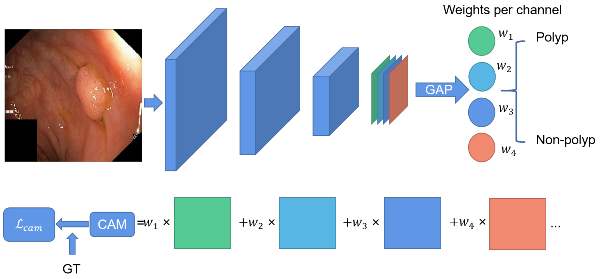

Fig. 4. The CAM loss is used to make the encoder more focused on the foreground of the image since the CAM obtained by weighting and summing the results of the encoder has certain targeting capabilities.

D. Multi-level and Multi-view Training Strategy

Segmentation networks tend to apply supervision only at the last layer of decoders, measuring the difference between decoding results and gt to guide the training procession. As a result, traditional U-Net has two shortcomings: firstly, U-like networks have complex up-sampling operations, which in a way ”imagine” the details of the image without specific supervision, and the final segmentation supervision alone is not sufficient to cope with obscure targets. Secondly, the process of connecting the feature map from the encoder block to the corresponding decoder, and then processing it through certain convolutional layers to eliminate the blending effect, also lacks a clear objective. In this method, multi-level supervision is designed to solve the above problem by up-sampling the output of each decoder using bilinear interpolation and comparing it with gt or by directly resizing gt to the appropriate scale and comparing it with them as part of the final loss. In terms of the specific calculation, we follow the adaptive pixel intensity loss (API) proposed by TRACER [29], which introduces the concept of pixel intensity, assigning a weight to each pixel point to re-balance the various components of the ground truth, based on the idea that, compared to pixels that highlight the background of the image and the center of the target edges and the pixels adjacent to the edges need more attention. It, therefore, gives more weight to the edges, forcing the model to better segment the edge region and obtain a finer mask.

Some studies [30] pointed out that the global average pooling layer(GAP) proposed in [31] enables the convolutional neural network to have remarkable localization ability despite being trained on image-level labels. Further experiments have shown that both GoogLeNet and VGG have some object localization capability after using GAP to process the output of the last convolutional layer, in other words, this method is a weakly supervised target detection solution because it allows the model to focus more on target-related areas of the image. As shown in Fig.4, for each specific class, GAP generates a set of weights corresponding to the last convolutional layer of the network, which is multiplied by the corresponding feature map and combined in the channel direction to form a class activates map [32](CAM) corresponding to the class. The visualization results suggest that the CAM assigns greater weights to pixels that may represent semantic information about the image. Inspired by these detection insights, We add a binary classification task to the last layer of the encoder, adding the classification losses to the final loss function and using the classification weights formed by the GAP to construct a class activation map (CAM) for the class of polyps. The CAM is compared with gt to calculate an additional segmentation loss, which is added to the final loss function, noting that the classification loss is valid for all samples during training, while the segmentation loss is only valid for positive samples. This process can be formalized as follows:

$$\mathbf{F^k} = \frac{1}{x+y}\sum_{x,y\in\Omega} f_k\left(x,y\right), \tag{5}$$

$$s_c = \sum_{k=1}^{K} w_k^c \mathbf{F^k} \frac{1}{x+y}\sum_{x,y\in\Omega}\sum_{k=1}^{K} w_k^c f_k\left(x,y\right), \tag{6}$$

$$p_c = \frac{e^{s_c}}{e^{s_1}+e^{s_2}+\cdots e^{s_K}}, \tag{7}$$

$$\mathbf{F_{re}} = w_k^c f_k\left(x,y\right). \tag{8}$$

Here x,y is the spatial location of a pixel point. K indicates all categories. Ω denotes the entire feature map points. F^k denotes the global average pooling result for the K-th channel,$f_k\left(x,y\right)$ is the activation of the corresponding spatial position. w_k^c denotes the importance of F^k for the model to judge the target as class c. The probability of the target being class C will be given by the p_c. e^{s_i} denotes the probabilistic activation value of class c. $\sum_{x,y\in\Omega}$ and $\sum_{k=1}^{K}$ means summation in spatial and channel direction. F_{re} indicates the final result of the encoder. The final loss function consists of three components, a multi-level segmentation loss with pixel intensity, a CAM loss for positive samples and a classification loss for all samples.

$$\mathcal{L}_{seg} = \sum_{i=1}^{4}\left(\mathcal{L}_{api}\left(\phi_{up}\left(\mathbf{F_{d_i}}, s_i\right)\right)\right), \tag{9}$$

$$\mathcal{L}_{cam} = -\sum_{m,n\in\Omega}\left(\mathcal{Y}_{m,n}\log\mathcal{P}_{m,n} + \left(1-\mathcal{Y}_{m,n}\right)\log\left(1-\mathcal{P}_{m,n}\right)\right), \tag{10}$$

$$\mathcal{L}_{cls} = -\left(y_i\log p_i + \left(1-y_i\log\left(1-p_i\right)\right)\right). \tag{11}$$

Here,F_{d_i} is i-th decoding result, $\sum_i$ means making use of all decoders. $\phi_{up}\left(\cdot\right)$ means to up-sample F_{d_i} by a factor of s_i using bilinear interpolation $\mathcal{L}_{api}$ is the adaptive pixel loss, $\mathcal{Y}_{m,n}$ denotes the pixel level label of location (m,n). $\mathcal{P}_{m,n}$ denotes the probability the pixel is positive sample. $\sum_{m,n}$ means adding all pixel loss in the image. y_i indicates whether the sample contains polyps, p_i denotes the probability that the model predicts the image as a positive sample. These segment loss $\mathcal{L}_{seg}$, $\mathcal{L}_{cam}$ and the categorical loss $\mathcal{L}_{cls}$ are added together to form the complete loss function.

IV. Experiment Design

A. Experimental Settings

We evaluated our method on two types of publicly available polyp segmentation data sets, the first being the earlier and smaller ETIS [33], CVC-ClinicDB/CVC-612 [34], CVC-ColonDB [35], Kvasir [36], and the second being the more recently proposed larger SUN-SEG [16] data set, which contains 19544 training images and tens of thousands of test images distributed across multiple settings. We compare the model in this paper with several state-of-the-art medical image segmentation methods: U-Net [6],U-Net++ [12] ,ResUNet++ [9],SFA [20],PRA [19],ACS [18]. Two video segmentation methods are also compared: PNS [15] PNS+ [16], where the results for PRA, ACS, PNS+ are obtained from publicly available code, using default settings, and the rest are quoted from experimental results of paper PRA and PNS+.

Our implementation is based on PyTorch and all training and testing can be done on a single 1080TI, with the input image set to 320x320 and processed by certain data enhancements and augmentations. Before starting training on the target data set the encoder was pre-trained on imagenet to migrate the knowledge of the natural images and obtain good initial weights. we set the initial learning rate to 5×10^{-5} and employed a weight decay of 1×10^{-4}. We choose the AdamW optimizer to optimize the loss function and train the model. Batch size and gradient clipping are set to 16 and 2 respectively. The training epoch is set to 50 since we find that this ensures adequate training and saturation of the model for all data sets. For the first type of data set, we follow the PRA correlation settings: i.e. the images from the Kvasir and CVC-ClinicDB data sets are randomly split into two parts, 80% of which are used for training and 20% for testing. Our training-related parameters were set as follows table. In terms of evaluation metrics, we mainly use two widely used evaluation metrics in the field of semantic segmentation, Dice and IoU for quantitative evaluation. To gain a deeper understanding of the model performance and to compare more comprehensively with existing models, we inherited the evaluation system from PRA and introduced other metrics that are widely used in the field of target detection [37] including F_{β}^{α}, E_{φ}^{mn} for evaluating pixel-level similarity, Sen and S_{α} for evaluating global-level similarity. We used the open source evaluation toolkit, which can be found at the official website of VPS in GitHub [38].

B. Comparison to the State-of-the-art Methods

For the first type of data set, we set up separate experiments to test the learning and generalization capabilities of the model, following the PRA setup. The remaining 20% of the data from Kvasir and CVC-612, which had been present in the training set, were used to verify the model's ability to fit the training data. We tested the model's generalization ability on the other two data sets. These data sets that do not appear in the training set have their own challenging circumstances and properties. CVC-ColonDB is a small scale data set containing 380 images from 15 real video sequences. ETIS is an earlier released small scale data set containing 196 polyp images. All images were used as data for our test set. The results of our experiments are shown in Table.Iand 3, which show that our model achieves SOTA in both its ability to fit on the seen data set and its ability to generalize on the unseen data set. Besides, we get the better performance on hard test set. For the second type of data set, we only used positive samples from SUN-SEG, and negative samples will be used in subsequent ablation experiments. Following the data set partitioning provided by SUN-SEG, we trained with all training sets and tested only on easy/unseen and hard/unseen, as data sets that have not appeared in the training set are more responsive to the generalization ability of the model. The results are shown in Table.4, where our model outperforms all image-based and video-based methods. In Fig.5, we give the results of the model's polyp segmentation on the CVC-clinicDB validation set. Our model can precisely locate and segment polyp tissue in many challenging scenarios, e.g., different sizes, shapes, textures, etc.

TABLE I
Experimental results on kvasir and CVC-612 for testing the learning ability of the model

Dataset	Method	Publish	Dice	IoU	F_{β}^{w}	S_{α}	E_{φ}^{mn}
Kvasir	U-Net[6]	MICCAI2015	0.818	0.746	0.794	0.858	0.893
	U-Net++[12]	TMI2019	0.821	0.743	0.808	0.862	0.910
	ResUNet-Mod[39]	GRSL2018	0.791	N/A	N/A	N/A	N/A
	ResUNet++[9]	ISM2019	0.813	0.793	N/A	N/A	N/A
	SFA[20]	MICCAI2019	0.723	0.611	0.670	0.782	0.849
	PraNet[19]	MICCAI2020	0.898	0.840	0.885	0.915	0.948
	ours		**0.912**	**0.851**	**0.887**	**0.921**	**0.953**
CVC-612	U-Net[6]	MICCAI2015	0.823	0.755	0.811	0.889	0.954
	U-Net++[12]	TMI2019	0.794	0.729	0.785	0.873	0.931
	ResUNet-Mod[39]	GRSL2018	0.779	N/A	N/A	N/A	N/A
	ResUNet++[9]	ISM2019	0.796	0.796	N/A	N/A	N/A
	SFA[20]	MICCAI2019	0.700	0.607	0.647	0.793	0.885
	PNS[15]	MICCAI2021	0.860	0.795	N/A	0.903	0.903
	ours		**0.892**	**0.836**	**0.876**	**0.928**	**0.948**

TABLE II
Results from CVC-ColonDB and ETIS for testing the generalisation ability of the model

Dataset	Method	Publish	Dice	IoU	F_{β}^{w}	S_{α}	E_{φ}^{mn}
CVC-ColonDB	U-Net[6]	MICCAI2015	0.512	0.444	0.498	0.712	0.776
	U-Net++[12]	TMI2019	0.483	0.41	0.467	0.691	0.760
	SFA[20]	MICCAI2019	0.469	0.347	0.379	0.634	0.765
	PraNet[19]	MICCAI2020	0.709	0.640	0.696	0.819	0.869
	ours		**0.747**	**0.662**	**0.712**	**0.834**	**0.863**
ETIS	U-Net[6]	MICCAI2015	0.398	0.335	0.366	0.684	0.740
	U-Net++[12]	TMI2019	0.401	0.344	0.390	0.683	0.776
	SFA[20]	MICCAI2019	0.297	0.217	0.231	0.557	0.633
	PraNet[19]	MICCAI2020	0.628	0.567	0.600	0.794	0.841
	ours		**0.746**	**0.658**	**0.684**	**0.844**	**0.869**

C. Ablation Study

In this section, we conduct thorough testing on the SUN-SEG dataset to comprehensively evaluate each component of our model. Our objective is to gain a deeper understanding of our model's capabilities and showcase the effectiveness of our design. We begin by assessing the efficacy of the Semantic

TABLE III
EXPERIMENTAL RESULTS ON SUN-SEG, TESTED ONLY ON THE UNSEEN SUBGROUP IN ORDER TO MAKE THE RESULTS MORE CONVINCING

		SUN-SEG-Easy (Unseen)						SUN-SEG-Hard (Unseen)					
	Model	S_α	E_φ^{mn}	F_β^w	F_β^{mn}	Dice	Sen	S_α	E_φ^{mn}	F_β^w	F_β^{mn}	Dice	Sen
IMAGE	UNet	0.669	0.677	0.459	0.528	0.530	0.420	0.670	0.679	0.457	0.527	0.542	0.429
	UNet++[6]	0.684	0.687	0.491	0.553	0.559	0.457	0.685	0.697	0.480	0.544	0.554	0.467
	ACSNet[18]	0.782	0.779	0.642	0.688	0.713	0.601	0.783	0.787	0.636	0.684	0.708	0.618
	PraNet[19]	0.733	0.753	0.572	0.632	0.621	0.524	0.717	0.735	0.544	0.607	0.598	0.512
	SANet[40]	0.720	0.745	0.566	0.634	0.649	0.521	0.706	0.743	0.526	0.580	0.598	0.505
VIDEO	COSNet[41]	0.654	0.600	0.431	0.496	0.596	0.359	0.670	0.627	0.443	0.506	0.606	0.380
	MAT[42]	0.77	0.737	0.575	0.641	0.710	0.542	0.785	0.755	0.578	0.645	0.712	0.579
	PCSA[43]	0.680	0.660	0.451	0.519	0.592	0.398	0.682	0.660	0.442	0.510	0.584	0.415
	2/3D[44]	0.786	0.777	0.652	0.708	0.722	0.603	0.786	0.775	0.634	0.688	0.706	0.607
	AMD[45]	0.474	0.533	0.133	0.146	0.266	0.222	0.472	0.527	0.128	0.141	0.252	0.213
	DCF[46]	0.523	0.514	0.270	0.312	0.325	0.340	0.514	0.522	0.263	0.303	0.317	0.364
	FSNet[47]	0.725	0.695	0.551	0.630	0.702	0.493	0.724	0.694	0.541	0.611	0.699	0.491
	PNSNet[15]	0.767	0.744	0.616	0.664	0.676	0.574	0.767	0.755	0.609	0.656	0.675	0.579
	PNS+[16]	0.806	0.798	0.676	0.730	0.756	0.630	0.797	0.793	0.653	0.709	0.737	0.623
	Ours	**0.842**	**0.869**	**0.754**	**0.809**	**0.772**	**0.738**	**0.852**	**0.892**	**0.749**	**0.807**	**0.786**	**0.763**
	Ours w/ cam	**0.856**	**0.895**	**0.777**	**0.822**	**0.795**	**0.760**	**0.863**	**0.909**	**0.774**	**0.812**	**0.801**	**0.795**

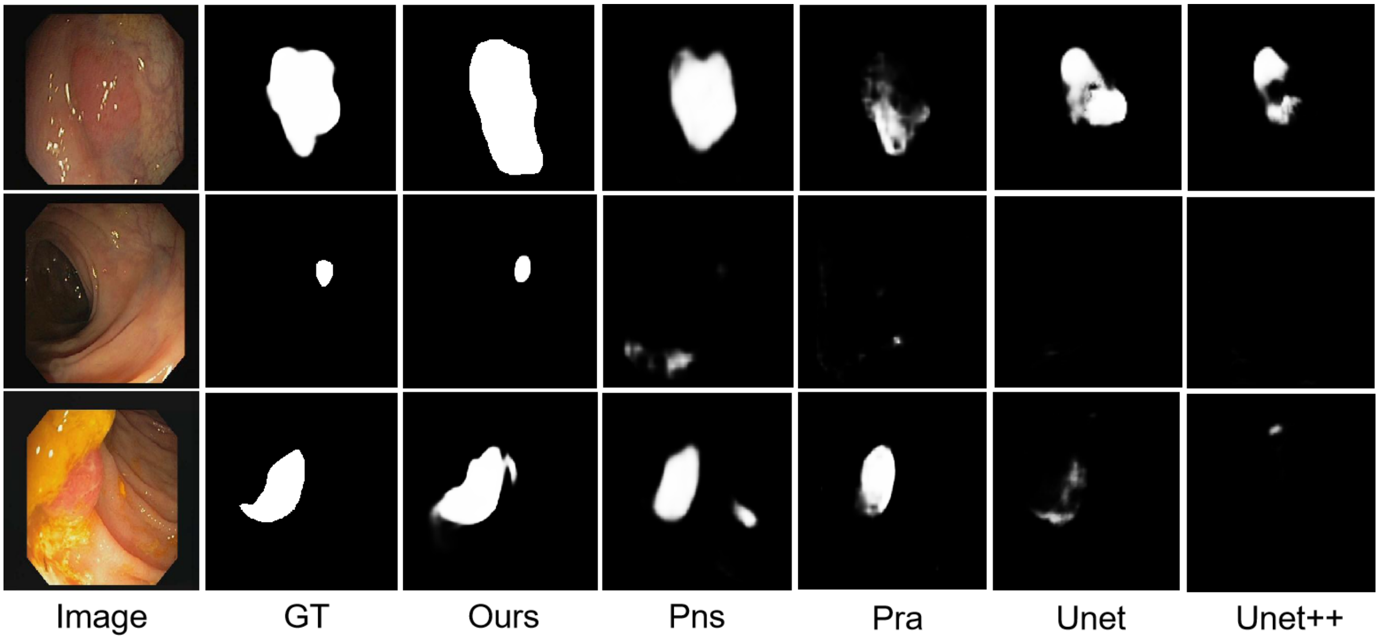

Fig. 5. Visualization of our method on the CVC-612, with each row representing one sample. The meaning of each column has been indicated in the figure.

TABLE IV
THE ABLATION STUDY OF SEMANTIC EXPRESS MODULE ON SUN-SEG EASY/UNSEEN

D4	D3	D2	S_α	E_φ^{mn}	F_β^w	F_β^{mn}	Dice	Sen	IoU
×	×	×	0.808	0.820	0.704	0.757	0.715	0.669	0.634
√	×	×	0.815	0.831	0.719	0.772	0.729	0.678	0.648
√	√	×	0.822	0.848	0.730	0.786	0.745	0.697	0.657
√	√	√	**0.842**	**0.869**	**0.754**	**0.809**	**0.772**	**0.738**	**0.688**

TABLE V
THE ABLATION STUDY OF VARIOUS SUPERVISION ON SUN-SEG HARD/UNSEEN

supervision	S_α	E_φ^{mn}	F_β^w	F_β^{mn}	Dice	Sen	IoU
single level	0.842	0.877	0.747	0.791	0.770	0.753	0.685
multi level	0.852	0.892	0.749	0.807	0.786	0.763	0.702
multi level w/i CAM	**0.863**	**0.909**	**0.774**	**0.812**	**0.801**	**0.795**	**0.724**

Express Module (SEM). To accomplish this, we incrementally introduce global information flows to decoders D2-D4 and evaluate their performance on easy/unseen test set. These models are trained using identical settings on the SUN-SEG training set. The results, presented in Table.5, demonstrate that the model's performance consistently improves as global information is incorporated into more decoder layers. Next, we evaluate the performance of the model on both easy and hard test sets after incorporating the CAM LOSS. Table.4 displays the results, illustrating that the addition of CAM LOSS yields improvements across all test sets. Finally, we examine the impact of various supervision methods on the model's performance on the hard/unseen data set which can be seen in Table.6 We compare the outcomes of applying supervision solely on the output of the last decoder layer (single supervision) with multi-level supervision, which involves utilizing all decoder outputs. Additionally, we consider the most comprehensive approach, multi-level supervision combined with CAM LOSS. Our findings indicate that both multi-angle and multi-level supervision effectively enhance the model's performance. Overall, our experiments demonstrate the effectiveness of the SEM, CAM loss, and various supervision methods in improving the performance of our model on different test sets.

V. Conclusion

We propose a enhanced U-Net network architecture for automatic polyp segmentation in colonoscopy images. The proposed method includes an encoder with a multi-scale convolutional attention mechanism and a decoder with a semantic express module. Comprehensive experiments on the widely used benchmark datasets demonstrate that the proposed approach achieves state-of-the-art performance under several different experimental settings, especially demonstrating remarkable accuracy (mdice>0.9) on the kvasir-seg dataset. Moreover, our model exhibits exceptional inference speed, achieving over 30 frames per second (FPS) on one NVIDIA 1080Ti GPU when processing 320×320 image inputs. In comparison to the leading PRA and PNS+ models, our approach shows superior learning ability, generalization capacity, and real-time segmentation efficiency.

References

[1] Silva,J.,Histace,A.,Romain,O.,Dray, X., Granado,B.:Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International Journal of Computer Assisted Radiology and Surgery 9(2), 283–293

[2] High-grade dysplasia and invasive carcinoma in colorectal adenomas: A multivariate analysis of the impact of adenoma and patient characteristics[J].European Journal of Gastroenterology and Hepatology,2002,14(2):183–188.

[3] David A Rosman, Judith Bamporiki, Rebecca Stein-Wexler, and Robert D Harris.2019.Developing diagnostic radiology training in low resource countries. Current Radiology Reports

[4] Haggar, F.A., Boushey, R.P.: Colorectal cancer epidemiology: incidence, mortality,survival, and risk factors. Clinics in colon and rectal surgery 22(04), 191–197

[5] Jia, X., Xing, X., Yuan, Y., Xing, L., Meng, M.Q.H.: Wireless capsule endoscopy:A new tool for cancer screening in the colon with deep-learning-based polyp recognition. Proceedings of the IEEE 108(1)

[6] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015.

[7] Siddique N, Paheding S, Elkin C P, et al. U-net and its variants for medical image segmentation: A review of theory and applications[J]. Ieee Access, 2021, 9: 82031-82057.

[8] Diakogiannis F I, Waldner F, Caccetta P, et al. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 162: 94-114.

[9] Jha D, Smedsrud P H, Riegler M A, et al. Resunet++: An advanced architecture for medical image segmentation[C]//2019 IEEE International Symposium on Multimedia (ISM). IEEE, 2019: 225-2255.

[10] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[11] Minaee S, Boykov Y, Porikli F, et al. Image segmentation using deep learning: A survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(7): 3523-3542.

[12] Zhou Z, Siddiquee M M R, Tajbakhsh N, et al. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation[J]. IEEE transactions on medical imaging, 2019, 39(6): 1856-1867.

[13] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[14] Jha D, Riegler M A, Johansen D, et al. Doubleu-net: A deep convolutional neural network for medical image segmentation[C]//2020 IEEE 33rd International symposium on computer-based medical systems (CBMS). IEEE, 2020: 558-564.

[15] 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 International Conference on Medical Image Computing and Computer Assisted Intervention. Strasbourg, France: Springer, 2021, pp. 142–152

[16] Ji G P, Xiao G, Chou Y C, et al. Video polyp segmentation: A deep learning perspective[J]. Machine Intelligence Research, 2022: 1-19.

[17] Khan S, Naseer M, Hayat M, et al. Transformers in vision: A survey[J]. ACM computing surveys (CSUR), 2022, 54(10s): 1-41.

[18] R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, and Y. Yu, "Adaptive context selection for polyp segmentation," in International Conference on Medical Image Computing and Computer Assisted Intervention. Lima, Peru: Springer, 2020, pp. 253–262

[19] 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 International Conference on Medical Image Computing and Computer Assisted Intervention. Lima, Peru: Springer, 2020,pp. 263–273

[20] Fang, Y., Chen, C., Yuan, Y., Tong, K.y.: Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: MICCAI. pp. 302–310.Springer (2019)

[21] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[22] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[23] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[24] Guo M H, Lu C Z, Hou Q, et al. Segnext: Rethinking convolutional attention design for semantic segmentation[J]. arXiv preprint arXiv:2209.08575, 2022.

[25] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[26] Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Int. Conf. Mach. Learn. pp. 448–456. PMLR (2015)

[27] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji,Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. IEEE TPAMI,41(4):815–828, 2019.

[28] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[29] Lee M S, Shin W S, Han S W. TRACER: Extreme Attention Guided Salient Object Tracing Network (Student Abstract)[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(11): 12993-12994.

[30] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba.Object detectors emerge in deep scene cnns. International Conference on Learning Representations, 2015.

[31] M. Lin, Q. Chen, and S. Yan. Network in network. International Conference on Learning Representations, 2014.

[32] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2921-2929.

[33] Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International Journal of Computer Assisted Radiology and Surgery 9(2), 283–293 (2014)

[34] Bernal, J., S´anchez, F.J., Fern´andez-Esparrach, D., Rodrguez, C., Vilarino, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs saliency maps from physicians. CMIG 43, 99–111 (2015)

[35] Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE TMI 35(2), 630–644 (2015)

[36] Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D.,Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MMM. pp. 451–462. Springer (2020)

[37] Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: IEEE CVPR (2020)

[38] https://github.com/GewelsJI/VPS/tree/main/eval

[39] Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters 15(5), 749–753 (2018)

[40] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui, “Shallow attention network for polyp segmentation,” in International Conference on Medical Image Computing and Computer Assisted Intervention. Strasbourg, France: Springer, 2021, pp. 699–70

[41] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, “See more, know more: Unsupervised video object segmentation with co-attention siamese networks,” in Conference on computer vision and pattern recognition.

[42] T. Zhou, J. Li, S. Wang, R. Tao, and J. Shen, “Matnet: Motion-attentive transition network for zero-shot video object segmentation,” Transactions on image processing, vol. 29, pp. 8326–8338, 2020,

[43] Y. Gu, L. Wang, Z. Wang, Y. Liu, M.- M. Cheng, and S.-P. Lu, “Pyramid constrained self-attention network for fast video salient object detection,” in AAAI Conference on Artificial Intelligence, vol. 34. New York, New York

[44] J. G.-B. Puyal, K. K. Bhatia, P. Brandao, O. F. Ahmad, D. Toth, R. Kader,L. Lovat, P. Mountney, and D. Stoyanov,“Endoscopic polyp segmentation using a hybrid 2d/3d cnn,” in International Conference on Medical Image Computing and Computer Assisted Intervention.

[45] R. Liu, Z. Wu, S. Yu, and S. Lin, “The emergence of objectness: Learning zero-shot segmentation from videos,” in Advances in neural information processing systems. [Online]:Curran Associates, Inc., 2021.

[46] M. Zhang, J. Liu, Y. Wang, Y. Piao, S. Yao, W. Ji, J. Li, H. Lu, and Z. Luo, “Dynamic context-sensitive filtering network for video salient object detection,” in International conference on computer vision.IEEE, 2021, pp. 1553–1563,

[47] Bernal, J., S´anchez, J., Vilarino, F.: Towards automatic polyp detection with a polyp appearance model. PR 45(9), 3166–3182 (2012)

[48] Mamonov, A.V., Figueiredo, I.N., Figueiredo, P.N., Tsai, Y.H.R.: Automated polyp detection in colon capsule endoscopy. IEEE TMI 33(7), 1488–1502 (2014)

[49] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[50] Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Integrating online and offline threedimensional deep learning for automated polyp detection in colonoscopy videos.IEEE JBHI 21(1), 65–75