Li-SegPNet: Encoder-Decoder Mode Lightweight Segmentation Network for Colorectal Polyps Analysis

Pallabi Sharma, *Student Member, IEEE*, Anmol Gautam, Pallab Maji, Ram Bilas Pachori, *Senior Member, IEEE*, and Bunil Kumar Balabantaray*, Member, IEEE*

Abstract*—*Objective: **One of the fundamental and crucial tasks for the automated diagnosis of colorectal cancer is the segmentation of the acute gastrointestinal lesions, most commonly colorectal polyps. Therefore, in this work, we present a novel lightweight encoder-decoder mode of architecture with the attention mechanism to address this challenging task. *Methods:* The proposed Li-SegPNet architecture harnesses cross-dimensional interaction in feature maps with novel encoder block with modified triplet attention. We have used atrous spatial pyramid pooling to handle the problem of segmenting objects at multiple scales. We also address the semantic gap between the encoder and decoder through a modified skip connection using attention gating. *Results:* We applied our model to colonoscopy still images and trained and validated it on two publicly available datasets, Kvasir-SEG and CVC-ClinicDB. We achieve mean Intersection-Over-Union (mIoU) and dice scores of 0.88, 0.9058 and 0.8969, 0.9372 on Kvasir-SEG and CVC-ClinicDB, respectively. We analyze the generalizability of Li-SegPNet by testing it on two independent previously unseen datasets, Hyper-Kvasir and EndoTect 2020, and establish the model efficiency in cross-dataset evaluation. We employ multi-scale testing to examine the model performance on different sizes of polyps. Li-SegPNet performs best on medium-sized polyps with a mIoU and dice score of 0.9086 and 0.9137, respectively on the Kvasir-SEG dataset and 0.9425, 0.9434 of mIoU and dice score, respectively on CVC-ClinicDB. *Conclusion:* The experimental results convey that we establish a new benchmark on these four datasets for the segmentation of polyps. *Significance:* The proposed model can be used as a new benchmark model for polyps segmentation. Lesser parameters in comparison to other models give the edge in the applicability of the proposed Li-SegPNet model in real-time clinical analysis.**

***Index Terms*—Deep learning, attention, colon cancer, polyps segmentation.**

I. Introduction

IN TERMS of cancer incidence, colorectal cancer (CRC) is ranked third, and its death rate is the second-highest in the worldwide scenario [1]. Colorectal cancer usually begins as polyps, which may be benign, i.e., non-cancerous, at its initial stage. Polyps are tiny clumps of cells. Over time polyps may become cancerous. Cancerous polyps can grow through 5 stages. Based on its growth and location, it belongs to any one of the stages from stage 0 to stage 4. The overall survival rate of colon cancer is 63%, but in conditions, like if it is treated in a localized stage (cancer has not spread outside of the colon or rectum), the survival rate is 91% [2], [3]. However, when it spreads to distant body parts, the survival rate decreases to 14%. Therefore, early detection and removal are necessary to prevent CRC and reduce mortality rates. Medical science suggests regular colonoscopy screenings to detect and remove polyps at their early stage [4]. However, the manual process is tedious, and due to varying human expertise levels, polyps can be missed during colonoscopy depending on their type and size [5].

Underlying challenges in manual polyps detection through colonoscopy provide an opportunity to incorporate an automated system to improve the efficiency of the process. We perform segmentation and classification to develop a decision support system for automatic analysis of CRC. Even though various approaches have been developed by several research groups to deal with the automatic polyps classification or segmentation methods, benchmarking state-of-the-art (SOTA) models is still an open problem to the research community. This is because of the growing variety of computer vision (CV) technologies that have been developed and applied to different polyp datasets. Therefore, in this work, we focus on segmentation of the region of polyps from colonoscopy still images experimented on four publicly available dataset. Previous work on literature segments the polyp region using basic image analysis based on texture features or morphological features [6]. A domain expert needs to analyze the images and identify the operations to perform on the images to extract the necessary information. Owing to the limitations of traditional image processing tasks, researchers used machine learning in medical image analysis [7]. However, these techniques also cannot eliminate the domain expert's overload since the segmentation's efficiency is dependent completely on the features used during the training of the machine learning model.

Manuscript received 15 June 2022; revised 10 September 2022; accepted 13 October 2022. Date of publication 21 October 2022; date of current version 21 March 2023. *(Corresponding author: Pallabi Sharma.)*

Pallabi Sharma is with the Department of Computer Science and Engineering, National Institute of Technology Meghalaya, Shillong 793003, India (e-mail: pallabisharma@nitm.ac.in).

Anmol Gautam and Bunil Kumar Balabantaray are with the Department of Computer Science and Engineering, National Institute of Technology Meghalaya, India.

Pallab Maji is with the NVIDIA Graphics Pvt. Ltd., India.

Ram Bilas Pachori is with the Department of Electrical Engineering, Indian Institute of Technology Indore, India.

Digital Object Identifier 10.1109/TBME.2022.3216269

In recent years, motivated by the success of deep learning in CV tasks, such as hand-written digit recognition and natural image classification, literature shows a trend of developing prototypes for medical image analysis incorporating deep learning [8], [9]. Recently, encoder-decoder based models have proven their usefulness in the segmentation task [10], [11], [12]. Even though there have been many improvements; still, these models are not accurate in all situations. Accurate segmentation of the polyps region is still an open research problem, and the medical science community demands a more accurate model to use in real clinical applications. Therefore, in this work, we propose an encoder-decoder-based model that outperforms the SOTA models for polyps segmentation. In this model, we use the modified triplet attention (MTA) module to inherit cross-dimensional interaction and improve the skip connection to transfer refined spatial information from encoder to decoder. It also aids in the harnessing of multi-scale information during the encoding stage. We use features with different spatial resolutions from a pre-trained ResNet-50 [13] to feed each block of the encoder to boost the encoder performance and establish a new benchmark in the direction of developing an automatic system for CRC analysis. The summary of the contributions is as follows:

- We propose Li-SegPNet, a novel encoder-decoder style model for polyps segmentation. Li-SegPNet significantly improves the segmentation results for the colorectal polyps with reasonable parameters. The novelty of the proposed model lies primarily in the unique architecture of the encoder and decoder block, along with the special skip connection. The use of the MTA module with RCB block helps preserve channel interdependencies and capture effective discriminative features to recognize polyps on the different scale while keeping the model lightweight in terms of parameters.
- The proposed MTA module used in the encoder and skip connection facilitates cross-dimensional interaction. MTA performs separate pooling for width, height, and channel as a separate branch. The output of each attention block of the MTA is multiplied element-wise to avoid internal covariate shift. This is the first work that uses an attention mechanism based on cross-dimensional interaction in polyps segmentation.
- Furthermore, Li-SegPNet introduces an attention-skip connection, a novel skip connection from the encoder to the decoder, to reduce the semantic gap between the encoder and decoder. The skip connection uses an attention gate, takes the incoming skip connection from the encoder as input along with the upsampled output from the lower-level decoder, and processes it through the MTA module. This helps the proposed Li-SegPNet to overcome the challenge of spatial information loss due to max-pooling and achieves impressive results.
- Our model adopts a transfer learning strategy to solve the problem of lack of large datasets. We use weights of the first few layers of ResNet-50 [13] pre-trained on the ImageNet dataset to feed each encoder block. Pre-trained weights are used as a weight initialization scheme for the proposed Li-SegPNet model. The weights taken from different blocks of ResNet-50 harness different spatial resolutions and boost the performance of each block of the encoder. This improves the overall performance of the proposed Li-SegPNet.
- A comprehensive comparison of the SOTA CV baseline methods is presented. We train and validate our model on two publicly available datasets, Kvasir-SEG [14] and CVC-ClinicDB [15] to evaluate our proposed Li-SegPNet. We establish the generalizability of Li-SegPNet by testing it on Hyper-Kvasir [16] and EndoTect 2020 [17] datasets. The experimental results demonstrate that we establish a new benchmark on these four datasets for the segmentation of polyps.

The organization of the paper is as follows: Section II discusses the related works followed by a detail discussion on the proposed model and its underlying methodology in Section III. Sections IV presents the datasets used in this work and the experimental setup, along with the details of the evaluation criteria of the proposed model. Section V reports the experimental results and the discussion on ablation studies of the proposed method. Section VI includes the conclusion.

II. Related Work

Automating the process of polyp segmentation is an open research area. Different architectures have been proposed in this pursuit, and among them, the encoder-decoder-style deep neural networks are widely used. Fully convolutional neural networks (FCNN) is the first encoder-decoder style of architecture proposed by Long et al. [18]. Unet [10], is proposed by Ronnerberger et al., which is a similar encoder-decoder style architecture with a skip connection developed particularly intending to solve medical image segmentation problems.

To improve the feature extraction, Hu et al. [19] introduced the squeeze and excitation module to produce a channel-wise attention map by using global average pooling and multi-layer perceptron. Woo et al. [20] introduced convolutional block attention module (CBAM), which generated spatial and channel attention maps together. Gain in performance using an attention module motivates us to use attention in medical image segmentation as well. To this end, Tomar et al. [21] proposed a new dual decoder attention network (DDANet) as an FCNN. On the Kvasir-SEG [14] data set, the model achieves a dice score of 0.8576 and mIoU of 0.78. Safarov et al. [22] and Gautam et al. [23] also used attention modules combined with dense blocks, residual blocks, and atrous or dilated convolutions to capture the long-range relationships and to preserve the spatial information from the input data. Atrous convolution has allowed the network to aggregate multi-scale contextual information without losing the resolution by increasing the receptive field area.

Mahmud et al. [24] introduced a depth dilated inception (DDI) module to enclose the different receptive areas to obtain the most generalized features. Their proposed PolypSegNet achieved 0.8872 dice score and 0.8256 Intersection-Over-Union (IoU) on Kvasir-SEG dataset and 0.9152 dice score and 0.8462 IoU on Clinic-DB dataset. Nevertheless, there are improvements in the segmentation task these models requires high computational power due to large number of trainable parameters.

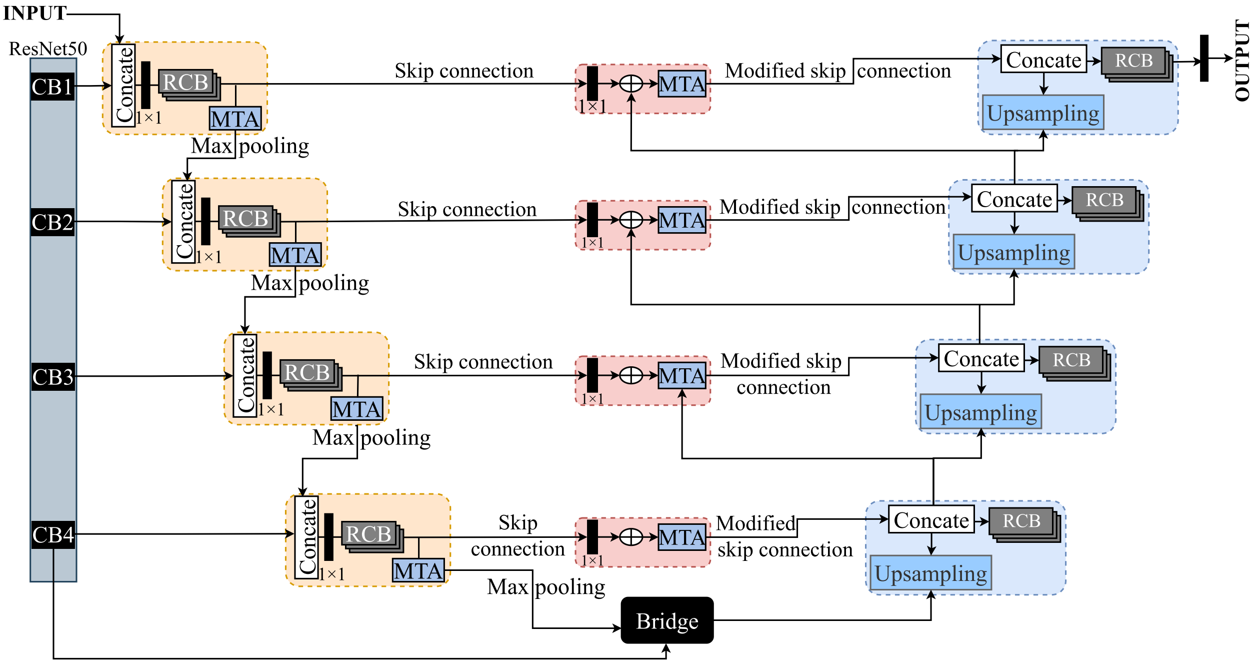


Fig. 1. The proposed architecture of encoder-decoder mode lightweight segmentation network (Li-SegPNet) for colorectal polyps analysis.

However, lightweight models are a crucial requirement for deploying a deep-learning model in hardware devices for clinical use. Considering this requirement, Jha et al. [25] used MobileNetV2 [26] as encoder for feature extraction in their proposed NanoNet architecture. The decoder is a modified version of the residual block [13]. Again, Jha et al. [11] proposed another lightweight architecture, ColoSegNet, that uses residual blocks along with squeeze and excitation networks [19]. ColoSegNet achieved an IoU of 0.81.

Literature reveals that there is an opportunity to contribute to the advancement of polyp segmentation, and a more efficient network will give an edge to the clinical applicability of the automatic model in CRC analysis.

III. Methodology

This section describes our proposed Li-SegPNet model in detail and its underlying concept in terms of polyps segmentation. We first introduce the target problem, followed by a detailed explanation of each module of the proposed Li-SegPNet architecture.

A. Problem Formulation

Let $X \in \mathbb{R}^{i \times j \times k}$ represent the input image space and $Y \in \mathbb{R}^{i \times j \times 1}$ denote the output segmented binary mask. For segmentation, each pixel corresponding to polyps has been assigned a label 1, and background pixels are labeled as 0. Let $G \in \mathbb{R}^{i \times j \times 1}$ be the ground truth mask where each pixel $X \in [0, 1]$. Given the pair of (X, G), the objective is to learn the mapping $F : X \rightarrow G$ given G as a label mask during training. The learned mapping is used to generate the segmented mask Y for a given input image X, i.e., $Y = F(X)$ in the inference stage.

B. Li-SegPNet Architecture

The proposed Li-SegPNet architecture is composed of a novel encoder-decoder block. Fig. 1. shows the overall architecture of the Li-SegPNet. Each encoder is made up of a residual block, MTA module, and max pooling block of 2×2 window and stride of 2. We take the output from pre-trained ResNet-50 at four different resolutions. The reason behind choosing ResNet-50 to transfer-learned weights for the proposed Li-SegPNet is its superior performance on medical image analysis [27], [28], [29], [30], [31]. Each output from one to fourth convolutional block of ResNet-50 is given to the respective encoder block as input along with input from the previous encoder. The base encoder receives input from ResNet-50 only along with the original input image. As shown in Fig. 1, each encoder except the base encoder takes two different inputs, one from the previous encoder and the other from the respective layer from ResNet-50, and performs concatenation. Then, 1×1 reduction is taken on concatenated features which are fed to the residual convolution block (RCB). The output from RCB is given to the MTA module to capture the multi-scale information, and then the max-pooling operation is performed to reduce the spatial dimension. The reduced feature map generated from the attention maps becomes the input to the next encoder.

In the encoder-decoder style of a network, skip connection is used to overcome the challenge of spatial information loss due to max-pooling. However, skip connection introduces a semantic gap between the encoder and decoder. Therefore, we have proposed a modified skip connection, i.e., attention-skip, that uses an attention gate to enhance the skip connections. The attention gate takes the incoming skip connection, i.e., the output from the RCB of the encoder block and the output from the lower-level decoder as input. Attention gate concatenates them and performs a 1 × 1 convolution followed by Rectified Linear Unit (ReLU) activation that helps in combining spatial information and low-level features. Therefore, the output feature map, which is generated from the MTA block of the attention gate, gives richer features to the decoder containing spatial and high-level information.

In each decoder block, the features generated from the attention-skip connection are concatenated with up-sampled features and go through the RCB block. The decoder performs up-sampling using transposed convolution. In the end, to generate a final segmented map, we perform a convolution operation with sigmoid activation to generate the segmentation mask.

Furthermore, the adoption of the transfer learning strategy in the proposed Li-SegPNet model solves the problem of the lack of large datasets. We use weights of the first few layers of ResNet-50 [13] pre-trained on the ImageNet dataset to feed each block of the encoder. Pre-trained weights are used as a weight initialization scheme for the proposed Li-SegPNet model. Weights are taken from different blocks of ResNet-50 that harness different spatial resolutions and boost the performance of each block of the encoder. This improves the overall performance of the Li-SegPNet.

In brief, the proposed novel architecture of Li-SegPNet overcomes the challenge of performance variance with respect to the object size using the unique encoder, decoder block, and attention-skip. The Novel MTA module, with its cross-dimensional interaction property, helps to improve the performance compared to other SOTA models keeping reasonable parameters. Moreover, transfer learned weight initialization helps overcome the problem of the lack of a large dataset.

C. Residual Convolutional Block (RCB)

The residual blocks allow us to build a deeper network and help preserve channel inter-dependencies. Without the residual unit, there is an issue of degradation and vanishing gradient. He et al. [13] proposed the residual unit as a solution to address this problem. Each residual unit is given by this general form specific to our implementation,

$$y' = H(x', W') + f_{1\times1}(x') \tag{1}$$

Here, x' is the input tensor, and y' is the output of the RCB. The residual connection goes through 1 × 1 convolution. These residual units form the backbone of the encoders. They are made up of two 3 × 3 convolutional layers followed by batch normalization (BN). The ReLU activation is used to introduce non-linearity.

D. Modified Triplet Attention (MTA)

Triplet attention (TA) is a computationally inexpensive mechanism to enhance the performance of downstream tasks. It exploits cross-dimensional interaction in the input feature map. Misra et al. [32] have proposed a TA module to provide an almost parameter-free attention mechanism to generate both channel and spatial attention maps. They have propounded the idea of cross-dimensional interaction to capture the interaction between spatial and channel dimensions of the input. Considering the advantages of attention, we propose a modified triplet attention (MTA) module composed of a TA module that uses a unique feature aggregation step that helps in improving the overall performance. The proposed MTA module is shown in Fig. 2.

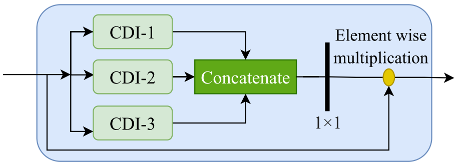


Fig. 2. Modified triplet attention (MTA) Module.

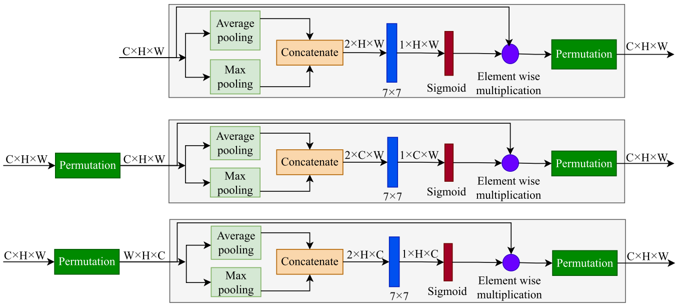


Fig. 3. Architecture of the cross dimensional interaction (CDI) block. C, H, W denotes channel, height, and width respectively.

The proposed MTA module has three branches to compute interaction between channel and spatial dimensions. The input flows through three different paths; each path computes a particular cross-dimensional interaction (CDI) as shown in Fig. 3. In the CDI-1, max pooling and average pooling are performed on the channel dimension of the input (X) and then concatenated to produce a tensor of dimensions (2 × H × W). This is passed through 7 × 7 convolution and BN. The attention weights (O_1) are produced in the dimension (1 × H × W) by passing through sigmoid activation. Final output of the path (${O_1}'$) is generated by element-wise multiplying X and O_1.

$$O_1 = \sigma\left(f_{BN}\left(f^{7\times7}\left(f_{\text{concat}}\left(\sqsubset F^S_{\text{avg}}, F^S_{\text{max}} \sqsupset\right)\right)\right)\right) \tag{2}$$

$${O_1}' = X \odot O_1 \tag{3}$$

Where, ${O_1}'$ is the output from CDI-1. In CDI-2 and CDI-3, X is rotated to get X' and X'' respectively, where X' is $(H \times C \times W)$ and X'' is $(W \times H \times C)$. Final output generated from these two paths can be given by following equations.

$$O_2 = \sigma\left(f_{BN}\left(f^{7\times7}\left(f_{\text{concat}}\left(\sqsubset F'_{\text{avg}}, F'_{\text{max}} \sqsupset\right)\right)\right)\right) \tag{4}$$

$${O_2}' = X' \odot O_2 \tag{5}$$

Where, ${O_2}'$ is the output from CDI-2. Where, F' is calculated by rotating the tensor in the following way

$F \in \mathbb{R}^{C\times H\times W}$ then, $F' \in \mathbb{R}^{H\times C\times W}$ where, $F' = f_{\text{perm}}(F)$ Similarly,

$$O_3 = \sigma\left(f_{BN}\left(f^{7\times7}\left(f_{\text{concat}}\left(\sqsubset F_{\text{avg}},'' F''_{\text{max}} \sqsupset\right)\right)\right)\right) \tag{6}$$

$${O_3}' = X' \odot O_3 \tag{7}$$

Where, ${O_3}'$ is the output from CDI-3. $F'' \in \mathbb{R}^{W\times H\times C}$, where $F'' = f_{\text{perm}}(F)$ In CDI-2 channel interacts with the width dimension, and in CDI-3, the channel interacts with the height dimension. In the MTA module, the output of each path is concatenated, and 1×1 convolution is taken, followed by sigmoid activation. Finally, the output is multiplied element-wise with the input to produce the final attention map, which goes through BN. The following equation gives the final output of the MTA module.

$$X' = X \odot \left(f_{BN}\left(f_{1\times 1}\left(\sqsubset O_1'; O_2'; O_3' \sqsupset\right)\right)\right) \tag{8}$$

$f_{perm}(.)$ represents permutation of the tensor along an axis. $f_{BN}(.)$ represents BN, $f_{k\times k}(.)$ is $k \times k$ convolution and $f_{concat}(.)$ is concatenation operation on the input tensors along respective axis. σ represents sigmoid activation function. We have placed the BN layer to avoid internal covariate shift due to three paths in the triplets.

E. Loss Function

In this work, we address the problem of class imbalance through a region-based loss function. In biomedical image segmentation, a class imbalance arises when the number of pixels representing area of interest is significantly lesser compared to the background regions. Therefore, instead of using the distribution-based loss function, we have opted for region-based loss functions. We use two region-based loss functions, Dice loss [33] and focal tversky loss [34] for our experimentation. Using dice loss, we have optimized the parameters to get the best dice score and IoU. Dice score can be seen in general form by this equation 9 where ground truth labels, G_{ic} belongs to 0, 1, where 0 and 1 represent two classes in binary segmentation. Predicted labels, P_{ic} belongs to [0, 1]. Let the total number of pixels be K.

$$\text{Dice score} = \frac{\sum_{i=1}^{K} P_{ic} \cdot G_{ic}}{\sum_{i=1}^{K} P_{ic} + G_{ic} + \epsilon} \tag{9}$$

Here epsilon is the smoothening factor to prevent division by zero. Dice loss penalizes when region overlap is poor.

$$L_{\text{Dice}} = \sum_{c} 1 - \text{Dice score} \tag{10}$$

Dice loss gives same weight to false positives and false negatives leading to high precision and low recall. We strictly want to minimize this difference. Also, false negatives must be given more attention than false positives. Tversky similarity index is more general form of dice score and can be represented as,

$$TL_c = \frac{\sum_{i=1}^{K} P_{ic} \cdot G_{ic} + \epsilon}{\sum_{i=1}^{K} P_{ic} \cdot G_{ic} + \alpha \sum_{i=1}^{K} P_{i\bar{c}} \cdot G_{ic} + \beta \sum_{i=1}^{K} P_{ic} \cdot G_{ic}} \tag{11}$$

where, c and $\bar{c}$ are class labels. Focal tversky loss is given by the following equation,

$$L_{TF} = \sum_{C} (1 - TL_C)^{\frac{1}{\gamma}} \tag{12}$$

The advantage of focal tversky loss [34] is that it can help in learning examples with small regions of interest, and hence it is suitable for our problem statement of polyp segmentation.

In addition, an analysis of the targeted polyps segmentation task using Li-SegPNet with a combination of one of the best performing loss functions proposed by Hatamizadeh et al. [35] on medical image classification task is included in this work to confirm the choice of best suitable loss function. They proposed a loss function that is a combination of soft-dice and cross-entropy loss (CSDCE) and can be represent by the following equation,

$$L_{CSDCE} = 1 - \frac{2}{C} \sum_{c=1}^{C} \frac{\sum_{i=1}^{K} G_{i,c} Y_{i,c}}{\sum_{i=1}^{C} G_{i,c}^2 + \sum_{i=1}^{C} Y_{i,c}^2} - \frac{1}{K} \sum_{i=1}^{C} \sum_{j=1}^{K} G_{i,c} \log Y_{i,c} \tag{13}$$

We have extensively analyzed the proposed model with all three loss function.

IV. Experimental Setup

A. Dataset

We use two datasets, Kvasir-SEG, and CVC-ClinicDB, to evaluate the proposed method for the segmentation tasks of polyps. To check the robustness and generalizability, we perform extensive testing on another two datasets, EndoTect 2020 and Hyper-Kvasir [36]. These two datasets are used only for testing, and images of these two datasets are not seen by the model during training. The details of all four datasets are summarized below.

- Kvasir-SEG: This dataset contains 1000 polyp images acquired by the high-resolution electromagnetic imaging system, i.e., ScopeGuide, Olympus Europe, their corresponding masks, and bounding box information. The images and their corresponding ground truths are used for the segmentation. The resolution of the images in this dataset ranges from 332 × 487 to 1920 × 1072 pixels. The dataset can be downloaded at https://datasets.simula.no/Kvasir-SEG/.
- CVC-ClinicDB: This is an open-access dataset of 612 images with a resolution of 384 × 288 from 31 colonoscopy sequences. The dataset can be downloaded from https://polyp.grand-challenge.org/site/Polyp/CVCClinicDB/
- EndoTect 2020: This dataset contains 1000 original images of polyps along with its corresponding segmentation mask. The images are in JPEG-compressed format. The dataset can be downloaded from https://endotect.com/
- Hyper-Kvasir: This dataset provides the original image and a segmentation mask of 1,000 images from the

polyp class. The dataset can be downloaded from https://datasets.simula.no/hyper-kvasir/

For all the data sets, the given mask is in the form of a binary image. The region of interest, i.e., the polyps pixels, is represented by the foreground (white), while the background (in black) does not contain polyp pixels.

B. Data-Augmentation

Generating annotated data instances for a biomedical task is an endeavor, and annotating new images is time-consuming and costly. For quality annotations, expensive medical expertise is needed. Some of the previous work used privately generated datasets. However, due to privacy and ethical issues, it is more difficult to share those medical data than natural images. Therefore, the availability of publicly accessible medical image datasets is limited. In our case, we have a maximum of 1000 annotated images of polyps in each dataset. Convolutional neural network (CNN) based models are data-hungry in nature, and it is generally accepted that an increase in the size of data improves performance. Therefore, to compensate for the data limitation, we use data augmentation using Albumentations [37] library. We perform offline data augmentations on the training set. We perform a total of 17 augmentations, center crop, random rotate, grid distortion, horizontal flip, vertical flip, grayscale conversion,coarse dropout, RGB Shift, etc., on each image from the training set.

C. Evaluation Metrics

The performance of each model was evaluated based on different metrics. In this work, the positive class label is one means pixel belongs to the polyp class, and the negative class label is zero, i.e., background class. True positive (TP) is an outcome where the model correctly labels the pixel to the positive class. True negative (TN) is an outcome where the model correctly labels the pixel to negative class. False positive (FP) means the pixel of a negative class is predicted as the positive class. False negative (FN) means the pixel of a positive class is predicted as a negative class.

Dice score is the measure of similarity between the predicted segmented mask and labeled mask. Equation 9 is the continuous version of the dice score, which helps to make it differential and hence use it for loss calculation and back-propagation. The Dice Score given in equation 14 is a discrete form of the same.

$$\text{Dice score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{14}$$

For segmentation tasks, IoU is commonly used. IoU is measured by the ratio of overlapped area to the union of the area. IoU is also known as Jaccard Index.

$$\text{IoU} = \frac{\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{15}$$

mIoU is also used to measure the performance of the segmentation task. It averages the IoU for each class.

Furthermore, from the literature it is seen that precision and recall are two measures based on pixel accuracy to evaluate the work performance of CNN model. Precision is the measure of correct positive predictions given total positive predictions. A recall is the measure of correct positive predictions made by the model out of all positive samples. Precision and recall can be calculated as follows,

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{17}$$

Furthermore, to indicate the correctness of the prediction in pixel level, receiver operating characteristic (ROC) curve is used and the area under the ROC (AUC-ROC) plot is used to compare the performance of the models. ROC curves are plotted considering the ratio of true-positive-rate (TPR) to false-positive-rate (FPR), where TPR is exactly same as recall, and FPR can be calculated as,

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{18}$$

V. Results and Discussion

This section presents the results and analysis of the proposed Li-SegPNet on polyps segmentation followed by a comparative analysis of Li-SegPNet with SOTA segmentation model. The section further discusses the performance of Li-SegPNet based on the size of polyps.

We develop our model using Keras framework [38] with TensorFlow [39] as back end and Quadro RTX 6000 for training the models. During training, to finalize the optimized hyper-parameter, we use hold-out as a cross-validation strategy. Data splitting is performed in the ratio of 8:1:1 for training, validation, and testing purposes. The images are resized to dimensions 256 × 256. We optimize our model using Adam [40] with a learning rate of 1e-4. To handle the memory requirements, we use a batch size of 16. We train our models for 50 epochs. EarlyStopping and ReduceLROnPlateau are used to handle the overfitting issue. Again, to check the generalizability of the proposed model, we chose cross-dataset evaluation over cross-validation because of the computational cost of cross-validation. As we know, cross-validation gives the idea of how the model will generalize to an unknown dataset. Therefore, rather than training and testing on a single dataset, we opted for across datasets for better generalizability. We trained and validated our model on, Kvasir-SEG and CVC-ClinicDB and tested it on Hyper-Kvasir and EndoTect 2020 datasets.

The loss-curves acquired during training of the Li-SegPNet model on the Kvasir-SEG dataset and CVC-clinicDB are shown in Fig. 4. Convergence of the model is clearly visible on the graph in under 50 epochs. The predicted mask for sample test images are shown in Fig. 5. The input image, ground truth, and prediction masks are presented in the first, second, and third columns, respectively.

A. Effect of Each Component of Li-SegPNet

To finalize the hyper-parameter used in the proposed Li-SegPNet, we perform an empirical analysis on both Kvasir-SEG and CVC-ClinicDB datasets.

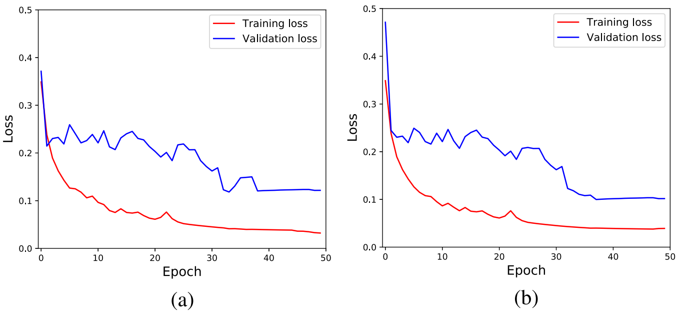


Fig. 4. Loss convergence plot obtained by Li-SegPNet on (a) Kvasir-SEG dataset, (b) CVC-ClinicDB dataset.

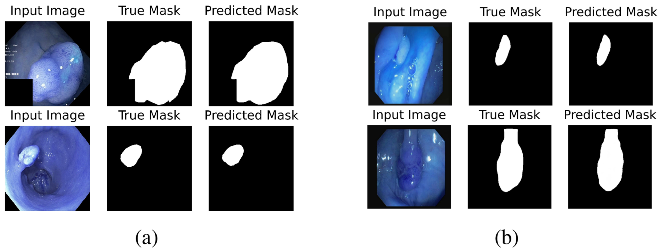


Fig. 5. Predicted mask obtained using Li-SegPNet on (a) Kvasir-SEG dataset, (b) CVC-ClinicDB dataset.

TABLE I
PERFORMANCE COMPARISON OF LI-SEGPNET ON TEST SET OF KVASIR-SEG AND CVC-CLINICDB DATASETS USING DIFFERENT LOSS FUNCTION

Dataset	Loss-function	Dice score	IoU	mIoU	Precision	Recall
Kvasir-SEG	Dice loss	**0.9058**	**0.8283**	**0.8800**	**0.9424**	**0.8509**
	Focal tversky loss	0.8970	0.8137	0.8758	0.9315	0.8471
	CSDCE	0.8765	0.7825	0.4235	0.9411	0.8173
CVC-ClinicDB	Dice loss	**0.9372**	**0.8820**	**0.9176**	**0.9642**	**0.9055**
	Focal tversky loss	0.9250	0.8607	0.8969	0.9594	0.8726
	CSDCE	0.9065	0.8125	0.5051	0.9512	0.8927

*Bold fonts indicate the best performance.

1) Effect of Different Loss Function on Li-SegPNet: We study the performance of Li-SegPNet using three different loss functions. The results shown in Table I present the quantitative performance of Li-SegPNet using dice loss, focal-tversky loss, and CSDCE loss function, and the results indicate that the dice loss performs more efficiently compared to the other two loss functions on both the datasets.

In the Kvasir-SEG dataset, we have achieved a mIoU of 0.8800, a dice score of 0.9058 with dice loss which is higher than that of focal tversky loss (0.8758, 0.8970) and CSDCE (0.4235, 0.8765). In the CVC-ClinicDB dataset as well, dice loss has outperformed focal tversky loss and CSDCE loss function in our analysis. As a result, we have opted to use dice loss to train Li-SegPNet for all further experiments.

2) Effect of Pre-Trained CNN: To boost the encoder performance, Li-SegPNet uses transfer learned weight initialization scheme. It uses ResNet-50 pre-trained on ImageNet dataset. We see from Table II that uses of pre-trained network aid in performance improvement. The reason behind choosing ResNet-50 as a pre-trained network is its superior performance as a feature extractor among all other CNN model in medical image analysis task [28], [29], [30], [31], [41], [42]. To confirm the choice of pre-trained model, we perform a study on the impact of additional three popular CNN, i.e., EfficientNet, MobileNetV1 and MobileNetV2, along with the ResNet-50. The results indicate that the choice of pre-trained model does not hamper to a great extent. However, the effect of a pre-trained model on encoder performance is quite visible when compared to its performance without the use of a pre-trained model.

TABLE II
PERFORMANCE ANALYSIS OF LI-SEGPNET USING DIFFERENT PRE-TRAINED CNN

Dataset	Pre-trained CNN	Dice score	IoU	mIoU	Precision	Recall
Kvasir-SEG	ResNet-50	**0.9058**	**0.8283**	**0.8800**	**0.9424**	**0.8509**
	EfficientNet	0.9007	0.8237	0.8798	0.9415	0.8571
	MobileNetV1	0.9007	0.8201	0.8716	0.9395	0.8471
	MobileNetV2	0.9053	0.8224	0.8724	0.9405	0.8487
	No-pretrained CNN	0.8407	0.7290	0.8201	0.8744	0.8110
CVC-ClinicDB	ResNet-50	**0.9250**	**0.8607**	**0.8969**	**0.9642**	**0.8726**
	EfficientNet	0.9200	0.8572	0.8896	0.9542	0.8672
	MobileNetV1	0.9165	0.8472	0.8689	0.9413	0.8497
	MobileNetV2	0.9193	0.8502	0.8714	0.9496	0.8519
	No-pretrained CNN	0.8904	0.8227	0.8502	0.9133	0.8276

*Bold fonts indicate the best performance.

TABLE III
RESULTS OF ABLATION STUDY WITH DIFFERENT ARCHITECTURAL CONFIGURATION (TESTED ON KVASIR-SEG DATASET)

Architectural choice		Dice score	IoU	mIoU	Precision	Recall
MTA at encoder	MTA at skip connection					
✗	✗	0.7673	0.6232	0.7695	0.8143	0.7210
✗	✓	0.8407	0.7290	0.8202	0.8744	0.8110
✓	✗	0.8532	0.7451	0.8541	0.8914	0.8009
✓	✓	**0.9058**	**0.8283**	**0.8800**	**0.9424**	**0.8509**

*Bold fonts indicate the best performance.

3) Effect of MTA Module on Performance of Li-SegPNet: To further understand the insights of the architecture of the proposed Li-SegPNet, we investigated the impact of the proposed MTA module by ablation study and reported the results on Table III. As described in section III-B, the MTA module is used in both the encoder block and the modified skip connection in the proposed Li-SegPNet. Therefore, first we removed MTA from the skip connection while keeping the rest of the architecture unchanged, we achieved the dice score of 0.8532 and IoU of 0.7451. Next we experimented Li-SegPNet architecture by removing the MTA module from every encoder block, and achieved dice score and IoU is 0.8407 and 0.7290, respectively. Further, we experimented by removing the MTA from the encoder as well as from the skip-connection, that leads to a further decrease of dice score to 0.7673 and IoU to 0.6232. The MTA's capacity to extract cross-dimensional characteristics aids in enhancing the segmentation's performance, as we see from the ablation study. With the addition of MTA, the problem of spatial information loss in the skip connection is also resolved, and the suggested Li-SegPNet shows significant performance gains over the standard skip connection.

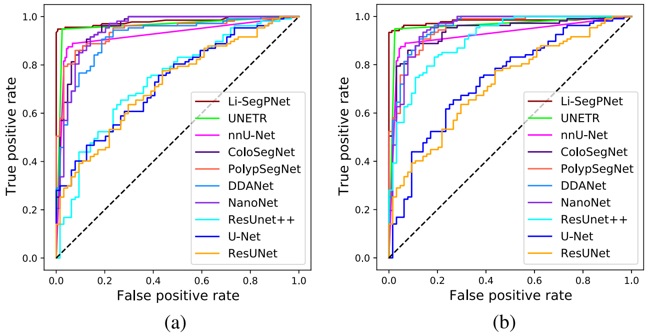
B. Performance Comparison of Li-SegPNet With Similar Works

Performance of Li-SegPNet is compared with nine SOTA deep learning based models for polyps segmentation namely, U-Net [10], ResUNet [43], ResUNet++ [43], NanoNet [25], ColoSegNet [11], DDANet [21], UNETR [35], PolypSegNet [24], and nnU-Net [44]. For the fair comparison, all the

TABLE IV
QUANTITATIVE COMPARISON OF THE PROPOSED MODEL WITH SOTA CNN BASED WORKS FOR POLYPS SEGMENTATION

Model	Kvasir-SEG					CVC-ClinicDB				
	Dice score	IoU	mIoU	Precision	Recall	Dice score	IoU	mIoU	Precision	Recall
ResUNet [43]	0.7484	0.6012	0.7426	0.8048	0.6993	0.4510	0.5643	0.4570	0.5614	0.5775
U-Net [10]	0.7673	0.6232	0.7695	0.8143	0.7210	0.8780	0.7673	0.7880	0.9330	0.7870
ResUNet++ [43]	0.8133	0.5859	0.7927	0.877	0.7064	0.8260	0.7134	0.7311	0.8511	0.8003
NanoNet [25]	0.8227	0.6988	0.7282	0.8367	0.8588	0.7880	0.7800	0.7180	0.8220	0.6870
ColoSegNet [11]	0.8209	0.8100	0.6980	0.8435	0.8496	0.8862	0.8531	0.8248	0.9017	0.8828
DDANet [21]	0.8576	0.7548	0.7800	0.8643	0.8880	0.9075	0.7936	0.8587	0.9325	0.8456
UNETR [35]	0.8810	0.7880	0.7650	0.9161	0.9166	0.9090	0.8370	0.7860	0.9424	0.8515
PolypSegNet [24]	0.8872	0.8256	0.8564	0.9168	**0.9254**	0.9152	0.8462	0.7946	0.9621	**0.9113**
nnU-Net [44]	0.8977	0.8178	0.8106	0.9085	0.8271	0.9060	0.8532	0.7912	0.9587	0.8325
Li-SegPNet (Proposed)	**0.9058**	**0.8283**	**0.8800**	**0.9424**	0.8509	**0.9250**	**0.8607**	**0.8969**	**0.9642**	0.8726

*Bold fonts indicate the best performance.



(a) (b)

Fig. 6. AUC-ROC curves on (a) Kvasir-SEG dataset, and (b) CVC-ClinicDB dataset.

models are experimented on the same dataset using hold-out as a cross validation strategy with the ratio of 8:1:1 for training, validation, and testing. The only parameters necessary to define in its instantiation of UNETR are feature size and the position embedding layer type. We kept the initial feature size similar to original U-Net model i.e., 16 and position embedding layer type as convolution for fair comparison. For nnU-Net as well, we kept the backbone similar to its original paper by [44]. The remaining hyper-parameter and the architecture kept unchanged for all the models. The results on the test set of both the datasets are shown in Table IV and the results indicate that we have achieved SOTA score on polyps segmentation with our proposed Li-SegPNet model in terms of performance. Li-SegPNet achieved mIoU of 0.8800, and dice score of 0.9058, precision and recall are 0.9424 and 0.8509, respectively on the test set of Kvasir-SEG dataset. In the CVC-ClinicDB dataset, we have achieved a mIoU of 0.8969, dice score of 0.9250, precision and recall of 0.9642 and 0.8726 respectively. Furthermore, AUC-ROC is shown in Fig. 6. From the AUC-ROC curves, it is comprehensible that the proposed model can accurately differentiate the polyps pixels and the background pixels than the others. It can be observed that the Li-SegPNet model achieves higher performance than the SOTA models in spite of being significantly lighter.

1) Comparative Analysis of Li-SegPNet in Terms of Model Parameters: We compare the parameters of the proposed Li-SegPNet model with some of the top performing CNN based models for polyps segmentation. The Li-SegPNet has a lesser number of parameters as compared to vanilla U-Net, ResUNet++, UNETR and nnU-Net, and the number of parameters of each of the models is included in the Table V. As discussed earlier, lightweight models are a crucial requirement for deploying a deep-learning model in hardware devices for clinical use. Thus, comparatively lesser parameter is one of the notable contribution of the this work.

TABLE V
COMPARISON IN TERMS OF PARAMETERS

Model	Number of parameters
UNETR	92,580,000
nnU-Net	62,000,000
U-Net [10]	31,040,000
ResUNet++ [43]	16,242,036
Li-SegPNet (proposed)	11,106,697

C. Analysis of Generalization and Robustness of Li-SegPNet

To test the generalization capacity, we opted for across dataset validation. We trained our models on Kvasir-SEG and tested it on EndoTect 2020 and Hyper-Kvasir datasets. A descriptive statistical analysis are presented in Fig. 7 and the qualitative results of the predicted mask are shown on Fig. 8. There are some hazy edges in a few of the predicted mask for test images, but overall performance is quite impressive. The proposed LiSegPNet model has reached the dice score of 0.9652 on the Hyper-Kvasir dataset and 0.8620 on the EndoTect 2020 data set.

D. Performance Analysis of Li-SegPNet Based on the Polyps Size

In order to test the proposed model's performance on different sizes of polyps, we have carried out the multi-scale analysis of polyps on both Kvasir-SEG and CVC-clinicDB datasets. Validation data and test data were combined and divided based on the size of polyps. Pixel count representing polyps in the ground truth masks is taken as the dividing criteria. In the Kvasir-SEG data set, the number of small, medium, and large-sized polyp images are 50, 74, and 76, respectively. The results of the scale-wise analysis are given in Table VI. The model performed best on medium-sized polyps with a mIoU of 0.9086, dice score of 0.9137, precision, and recall of 0.9111 and 0.9193, respectively.

On CVC-ClinicDB, we have done a similar analysis. Out of 122 test and validation images, 30 belongs to small-sized

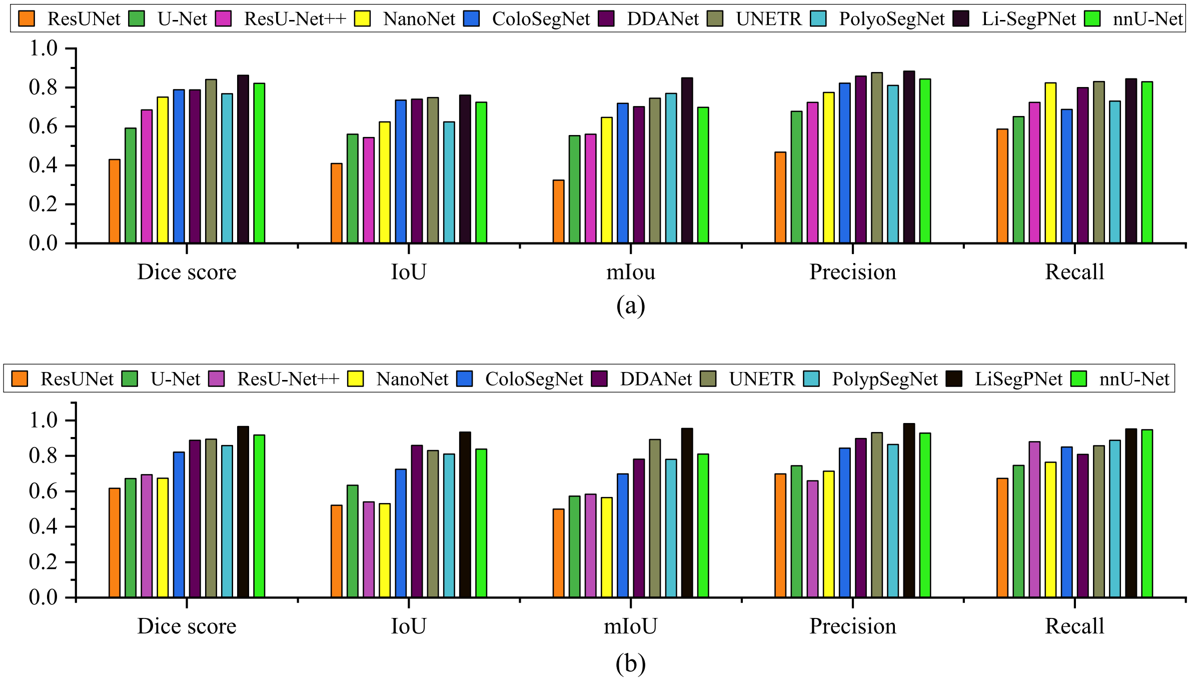


Fig. 7. A descriptive statistical analysis of Li-SegPNet's performance on generalizibilty. Comparison of Li-SegPNet with nine other SOTA models are shown where across dataset validation is adopted. All the models are trained on Kvasir-SEG dataset and tested on (a)EndoTect 2020 dataset, and (b) Hyper-Kvasir dataset.

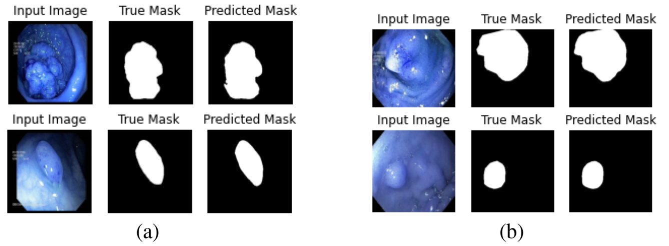


Fig. 8. Sample predicted mask generated using Li-SegPNet that is trained on Kvasir-SEG and tested on (a) EndoTect 2020 dataset, (b) Hyper-Kvasir dataset.

TABLE VI
EVALUATION RESULTS OF LI-SEGPNET BASED ON THE SIZE OF POLYPS

Dataset	Polyp size	Dice	IoU	mIoU	Precision	Recall
Kvasir-SEG	small	0.8377	0.7245	0.8731	0.8573	0.8848
	medium	**0.9137**	**0.8413**	**0.9086**	**0.9111**	**0.9193**
	large	0.8760	0.7817	0.8338	0.9503	0.8099
CVC-ClinicDB	small	0.8944	0.8105	0.8921	0.9312	0.8570
	medium	**0.9434**	**0.8929**	**0.9425**	**0.9396**	**0.9496**
	large	0.9230	0.8581	0.8835	0.9648	0.8860

*Bold fonts indicate the best performance.

polyps, 44 are medium-sized polyps, and 48 are large-sized. In multi-scale analysis, we found our model to perform best on medium-sized polyps with a mIoU of 0.9425 and dice score of 0.9434. The qualitative results is given in Fig. 9. From these analysis it is comprehensible that the use of ResNet-50 has helped to boost the performance of feature extraction, and MTA aided the process of incorporating multi-scale features in feature extraction immensely. The obtained results convey that we efficiently addressed the class imbalance and data insufficiency by our proposed Li-SegPNet using offline augmentation and dice loss.

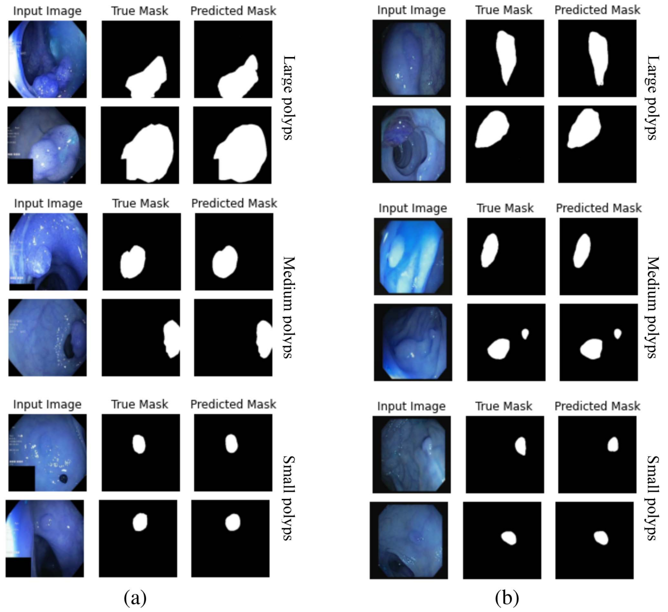


Fig. 9. Qualitative results obtained using Li-SegPNet on three different size polyps. (a) Kvasir-SEG dataset. (b) CVC-ClinicDB dataset.

VI. CONCLUSION

To address the challenges in the polyps segmentation task, in this work, we have proposed a novel deep learning architecture, "Li-SegPNet," that uses a unique encoder-decoder architecture with the MTA mechanism to harness cross-dimensional interaction in the input feature map. To boost the performance of the encoder in extracting polyps features, we used a pre-trained ResNet-50 model. Experimental results show that the proposed model is more efficient than other SOTA models and can be easily used in clinical settings due to its lightweight nature in terms of parameters. The comprehensive analysis of the proposed Li-SegPNet and SOTA models on four independent public datasets has verified the effectiveness and generalizability of our proposed model. Additionally, from the visualization result, it is evident that, even though the dataset contains different sizes of polyps, our model performs efficiently for all sizes of lesions. Future research should focus on reducing the number of trainable parameters while ensuring better performance. We can extend the applicability of our proposed Li-SegPNet for the segmentation of acute lesions from other medical image modalities.

Acknowledgment

The authors would like to thank NVIDIA Higher Education and Research Team for supporting the work. We utilized the resource supported by NVIDIA Academic Hardware Grant Program.

References

[1] X. Yang, X. Ye, and G. Slabaugh, “Multilabel region classification and semantic linking for colon segmentation in CT colonography,” *IEEE Trans. Biomed. Eng.*, vol. 62, no. 3, pp. 948–959, Mar. 2015.

[2] Y. Jiang et al., “Global pattern and trends of colorectal cancer survival: A systematic review of population-based registration data,” *Cancer Biol. Med.*, vol. 19, no. 2, 2022, Art. no. 175.

[3] M. Gao et al., “Comprehensive analyses of correlation and survival reveal informative lncRNA prognostic signatures in colon cancer,” *World J. Surg. Oncol.*, vol. 19, no. 1, pp. 1–15, 2021.

[4] B. Levin et al., “Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: A joint guideline from the American Cancer Society, the US multi-society task force on colorectal cancer, and the American College of Radiology,” *Gastroenterology*, vol. 134, no. 5, pp. 1570–1595, 2008.

[5] Y. Ren et al., “High-performance CAD-CTC scheme using shape index, multiscale enhancement filters, and radiomic features,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1924–1934, Aug. 2017.

[6] K. P. Constantinou et al., “Medical image analysis using AM-FM models and methods,” *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 270–289, 2021.

[7] J. L. Bruse et al., “Detecting clinically meaningful shape clusters in medical image data: Metrics analysis for hierarchical clustering applied to healthy and pathological aortic arches,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 10, pp. 2373–2383, Oct. 2017.

[8] Z. Yu et al., “Melanoma recognition in dermoscopy images via aggregated deep convolutional features,” *IEEE Trans. Biomed. Eng.*, vol. 66, no. 4, pp. 1006–1016, Apr. 2019.

[9] S. Trajanovski et al., “Tongue tumor detection in hyperspectral images using deep learning semantic segmentation,” *IEEE Trans. Biomed. Eng.*, vol. 68, no. 4, pp. 1330–1340, Apr. 2021.

[10] O. Ronneberger et al., “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[11] D. Jha et al., “Real-time polyp detection, localization and segmentation in colonoscopy using deep learning,” *IEEE Access*, vol. 9, pp. 40496–40510, 2021.

[12] Z. Zhang, Q. Liu, and Y. Wang, “Road extraction by deep residual U-Net,” *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[13] K. He et al., “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[14] D. Jha et al., “Kvasir-SEG: A segmented polyp dataset,” in *Proc. Int. Conf. Multimedia Model.*, 2020, pp. 451–462.

[15] J. Bernal et al., “WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians,” *Computerized Med. Imag. Graph.*, vol. 43, pp. 99–111, 2015.

[16] H. Borgli et al., “Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy,” *Sci. Data*, vol. 7, no. 1, pp. 1–14, 2020.

[17] S. A. Hicks et al., “The EndoTect 2020 challenge: Evaluation and comparison of classification, segmentation and inference time for endoscopy,” in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 263–274.

[18] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3431–3440.

[19] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[20] S. Woo et al., “CBAM: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[21] N. K. Tomar et al., “DDANet: Dual decoder attention network for automatic polyp segmentation,” in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 307–314.

[22] S. Safarov and T. K. Whangbo, “A-DenseUNet: Adaptive densely connected U-Net for polyp segmentation in colonoscopy images with atrous convolution,” *Sensors*, vol. 21, no. 4, 2021, Art. no. 1441.

[23] A. Gautam et al., “SAU-NET: Scale aware polyp segmentation using encoder-decoder network,” in *Proc. IEEE Region 10 Symp.*, 2022, pp. 1–5.

[24] T. Mahmud et al., “PolypSegNet: A modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images,” *Comput. Biol. Med.*, vol. 128, 2021, Art. no. 104119.

[25] D. Jha et al., “NanoNet: Real-time polyp segmentation in video capsule endoscopy and colonoscopy,” in *Proc. IEEE 34th Int. Symp. Comput.-Based Med. Syst.*, 2021, pp. 37–43.

[26] M. Sandler, “MobileNetV2: Inverted residuals and linear bottlenecks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[27] N. Zakaria, “Three ResNet deep learning architectures applied in pulmonary pathologies classification,” in *Proc. Int. Conf. Artif. Intell. Cyber Secur. Syst. Privacy*, 2021, pp. 1–8.

[28] M.-J. Tsai and Y.-H. Tao, “Deep learning techniques for the classification of colorectal cancer tissue,” *Electronics*, vol. 10, no. 14, p. 1662, 2021.

[29] Y. Komeda et al., “Artificial intelligence-based endoscopic diagnosis of colorectal polyps using residual networks,” *Plos One*, vol. 16, no. 6, 2021, Art. no. e0253585.

[30] S. H. Kassani and P. H. Kassani, “A comparative study of deep learning architectures on melanoma detection,” *Tissue Cell*, vol. 58, pp. 76–83, 2019.

[31] A. B. Hamida et al., “Deep learning for colon cancer histopathological images analysis,” *Comput. Biol. Med.*, vol. 136, 2021, Art. no. 104730.

[32] D. Misra, “Rotate to attend: Convolutional triplet attention module,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3139–3148.

[33] C. H. Sudre et al., “Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Berlin, Germany: Springer, 2017, pp. 240–248.

[34] N. Abraham and N. M. Khan, “A novel focal tversky loss function with improved attention U-Net for lesion segmentation,” in *Proc. IEEE 16th Int. Symp. Biomed. Imag.*, 2019, pp. 683–687.

[35] A. Hatamizadeh et al., “UNETR: Transformers for 3D medical image segmentation,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 1748–1758.

[36] H. Borgli et al., “HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy,” *Sci. Data*, vol. 7, no. 1, 2020, Art. no. 283.

[37] A. Buslaev et al., “Albumentations: Fast and flexible image augmentations,” *Information*, vol. 11, no. 2, 2020, Art. no. 125.

[38] F. Chollet et al., “Keras,” 2015. [Online]. Available: https://github.com/fchollet/keras

[39] M. Abadi et al., “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015. [Online]. Available: tensorflow.org

[40] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Representations*, 2015.

[41] P. Sharma et al., “Two stage classification with CNN for colorectal cancer detection,” *Oncologie*, vol. 22, no. 3, pp. 129–145, 2020.

[42] P. Sharma et al., “An ensemble-based deep convolutional neural network for computer-aided polyps identification from colonoscopy,” *Front. Genet.*, vol. 13, 2022, Art. no. 844391.

[43] D. Jha et al., “ResUNet++ : An advanced architecture for medical image segmentation,” in *Proc. IEEE Int. Symp. Multimedia*, 2019, pp. 225–2255.

[44] F. Isensee et al., “nnU-Net: Self-adapting framework for U-Net-based medical image segmentation,” *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021.