

# Polyp-SES: Automatic Polyp Segmentation with Self-Enriched Semantic Model

Quang Vinh Nguyen<sup>1</sup>, Thanh Hoang Son Vo<sup>1</sup>, Sae-Ryung Kang<sup>2</sup>, and Soo-Hyung Kim<sup>1</sup>

<sup>1</sup> Chonnam National University, Gwangju, South Korea  
`{vinhbn28, hoangsonvothan, shkim}@jnu.ac.kr`

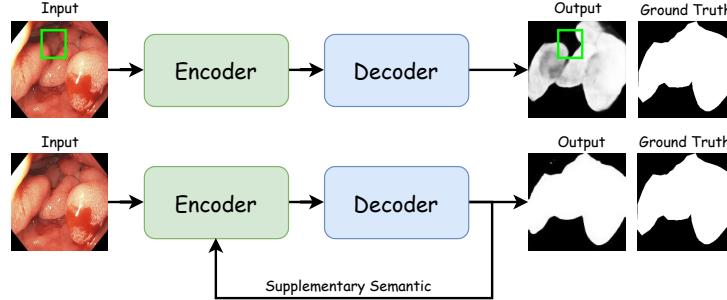
<sup>2</sup> Chonnam National University Hwasun Hospital and Medical School, Hwasun,  
South Korea  
`campanella9@naver.com`

**Abstract.** Automatic polyp segmentation is crucial for effective diagnosis and treatment in colonoscopy images. Traditional methods encounter significant challenges in accurately delineating polyps due to limitations in feature representation and the handling of variability in polyp appearance. Deep learning techniques, including CNN and Transformer-based methods, have been explored to improve polyp segmentation accuracy. However, existing approaches often neglect additional semantics, restricting their ability to acquire adequate contexts of polyps in colonoscopy images. In this paper, we propose an innovative method named “Automatic Polyp Segmentation with Self-Enriched Semantic Model” to address these limitations. First, we extract a sequence of features from an input image and decode high-level features to generate an initial segmentation mask. Using the proposed self-enriched semantic module, we query potential semantics and augment deep features with additional semantics, thereby aiding the model in understanding context more effectively. Extensive experiments show superior segmentation performance of the proposed method against state-of-the-art polyp segmentation baselines across five polyp benchmarks in both superior learning and generalization capabilities.

**Keywords:** Polyp Segmentation · Medical Image Segmentation · Deep Learning

## 1 Introduction

Medical image segmentation is the process of delineating regions of interest (ROI) within medical images, such as X-rays, MRI scans, CT scans, or histological slides, into meaningful and distinct anatomical structures. This process can assist clinicians in quantitative analysis, diagnosis, treatment planning, and monitoring of diseases. Traditional image processing techniques [9, 21, 27, 42] have been widely adopted in the medical image segmentation tasks. These methods primarily rely on handcrafted features potentially limiting their generalizability to complex image structures, noise, or variability in image quality. With



**Fig. 1:** Deep learning-based automatic polyp segmentation methods often include encoder and decoder parts. Contemporary models struggle to identify and categorize challenging features highlighted within green-bordered areas. This region appears relatively blurry and distinct from the surrounding polyp objects leading to confusion between normal tissues and actual polyps, thereby causing segmentation failures. Providing supplementary semantics promotes the model to obtain comprehensive contextual information about polyp objects, leading to a greatly segmentation performance.

the rapid development of deep learning, efficient and reliable segmentation solutions [3, 10, 37, 41] was introduced in the domain of the medical image segmentation.

Polyp segmentation is a critical task in medical imaging aimed at accurately identifying and delineating polyps within endoscopic or colonoscopic images. The challenges of identifying and segmenting polyps in medical images can be summarized for several following primary factors. **Firstly**, the variability in color, shape, size and texture among polyps poses a significant difficulty to precise segmentation. **Secondly**, the presence of noise, artifacts, or overlapping structures in the image can obscure polyps, increasing the complexity of the segmentation task. **Thirdly**, polyps may manifest in diverse positions within the gastrointestinal tract, each with its distinct characteristics, thus contributing to the overall variability and difficulty of the segmentation.

Traditional methods [19, 29, 30, 43] for polyp segmentation have often faced challenges such as sensitivity to image variations, the need for manual parameter adjustment, limited adaptability, and difficulties in handling noise and artifacts. Efforts leveraging deep learning techniques have been proposed to enhance the effectiveness of automatic polyp segmentation. Firstly, CNN-based methods [7, 8, 22–24, 26, 36, 49] exploit the ability of neural networks to automatically learn discriminative features and generate precise segmentation results. Despite the considerable success of CNN-based approaches, the limited receptive field poses challenges in capturing global representations. Transformer-based techniques [5, 14, 34, 39, 48], on the other hand, excel in capturing global dependencies and long-range contextual information more effectively compared to CNNs, resulting in superior performance in the polyp segmentation task. Nonetheless, transformer-based methodologies encounter difficulties in capturing fine-grained

details, which are crucial for accurately detecting and locating polyp objects. In addressing visual comprehension concerns involving semantic segmentation and object recognition, contextual information provides valuable insights that aid in disambiguating similar-looking objects, resolving occlusions, and improving the overall understanding of the visual scene. Notably, the approaches mentioned above primarily relying on poor contextual information neglecting to provide additional semantics. This limitation presents several challenges in comprehending adequate contexts of polyps, which are frequently characterized by noises, ambiguous boundaries, and intricate foregrounds. We have discovered that providing supplementary semantics can assist the model in obtaining comprehensive contextual information about polyp objects, potentially leading to a significant enhancement in segmentation performance as shown in Fig. 1.

Motivated by discussed concerns, our study introduces a new novel approach for automatic polyp segmentation task namely “Automatic Polyp Segmentation with Self-Enriched Semantic Model”. Initially, we employ an encoder to extract a sequence of multi-scale features. Subsequently, we introduce a Local-to-Global Spatial Fusion (LGSF) mechanism to capture both local and global spatial features before decoding them to generate an initial global feature map. Leveraging the proposed Self-Enriched Semantic (SES) module, we augment deep features with additional semantics, thereby aiding the model in understanding context more effectively. Our proposed solution achieves competitive segmentation performance compared to state-of-the-art baselines, showcasing proficiency in learning and generalization capabilities. Notably, it effectively addresses the limitations of prior models when operating in challenging contexts.

## 2 Related Work

### 2.1 Automatic Polyp Segmentation.

Traditional methods [19, 29, 30, 43] primarily rely on low-level features such as geometric characteristics, which often result in missed or inaccurate detections due to similarities with neighboring tissues. Recent advancements in deep learning have revolutionized the polyp segmentation by autonomously learning complex features. Among these innovations, U-Net [26] obtains significant improvements across various medical imaging tasks. By the simplicity and effectiveness design. ACSNet [47] refines the conventional skip connections within the U-Net [26] and selects adaptive features based on a channel attention mechanism. Planet [7] utilizes reverse attention mechanisms to refine boundary details in the global feature map through iterative stages enhancing segmentation predictions. MSNet [49] introduces multi-scale subtraction architecture in order to capture intricate details, eliminate redundancy and complementary information between the multi-scale features. SSformer [34] adopts a systematic feature fusion approach, gradually integrating both local and global contextual information, resulting in precise object delineation and boundary detection, while also capturing fine-grained details and comprehensive scene context. Polyp-PVT [5] presents a similarity aggregation module to extract local pixel and global semantic cues from the polyp area.

## 2.2 Vision Transformer.

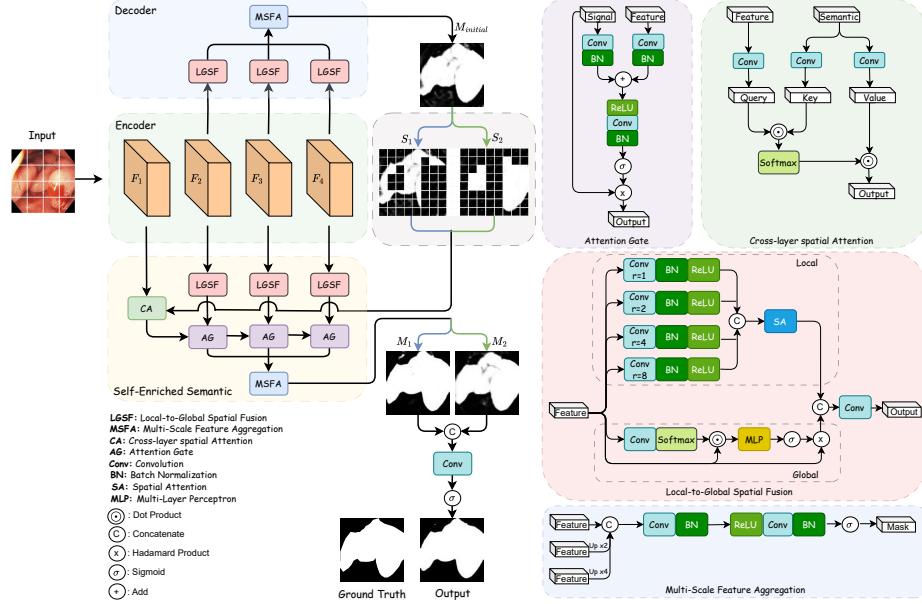
Transformer [31], initially successful in NLP, has garnered prominence for their potential in computer vision tasks. Leveraging the transformer mechanism, ViTs [6] effectively captures global dependencies and long-range spatial relationships, enabling comprehensive predictions based on the entire image context. By employing shifted windows instead of fixed-size patches, Swin [16] captures spatial relationships between adjacent patches, leading to enhanced feature representation and learning capabilities. Pyramid Vision Transformer (PVT) [35] incorporates a pyramid feature extraction mechanism to capture multi-scale information from input images. Through the combination of features from various scales, PVT [35] demonstrates proficiency in capturing both local details and global context, facilitating accurate dense prediction. UniFormer [15] integrates the strengths of convolutional neural networks (CNNs) and vision transformers (ViTs) into a concise transformer format. This innovative design empowers UniFormer [15] to efficiently capture both local redundancies and complex global dependencies, facilitating effective representation learning. MetaFormer [46] has recently demonstrated the commendable performance in computer vision tasks. This study meticulously examines various token mixers, spanning from basic operators like identity mapping or global random mixing to established techniques such as separable convolution and vanilla self-attention.

## 3 Method

As depicted in Fig. 2, our automatic polyp segmentation solution contains three principal components: Encoder, Decoder, and Self-Enriched Semantic (SES). The first is the Encoder, pretrained on ImageNet [4], extracts multi-scale features from an input image. The second contribution is the Decoder which employs Local-to-Global Spatial Fusion (LGSF) to capture both the global and local spatial features to achieve robust feature representation. Subsequently, refined features are aggregated to locate polyp objects and generate an initial global feature map. In the end, the SES component queries potential semantics from the initial global feature map and send them to high-level features to detect and relocate polyp objects accurately.

### 3.1 Encoder Backbone

In computer vision tasks, the encoder plays a crucial role in capturing spatial information and contextual cues from input images. Transformer-based encoding methods [5, 14, 48] offer the ability to capture long-range dependence information across different areas within the input image. Metaformer [46] has recently introduced new insights into designing transformer architecture and has shown significant performance improvements in various computer vision tasks. Motivated by these findings, our study adopts a vision metaformer encoder known as Caformer [46] as a reliable and competitive backbone for feature extraction.



**Fig. 2:** Overview of our architecture. The proposed method consists of an Encoder (Section 3.1), a Decoder (Section 3.2) and a Self-Enriched Semantic (Section 3.3) module. The Encoder extracts a sequence of multi-scale features from an input image. The Decoder aggregates high-level features to generate an initial segmentation mask. The Self-Enriched Semantic provides supplementary semantics to high-level features to re-locate polyp objects.

Given an input image  $I \in \mathbb{R}^{H \times W \times 3}$ , the encoder extracts four levels of features denoted as  $\{F_i | i \in (1, 2, 3, 4)\}$ . Among these feature maps,  $F_1$  provides detailed appearance information, while  $F_2, F_3$ , and  $F_4$  offer high-level features.

$$F_i = \varphi_{\text{Caformer}}(I) \in \mathbb{R}^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i} \quad (1)$$

Where  $H, W, C$  represent the height, width spatial, channel dimensions, respectively. In practice, we set  $H$  and  $W$  to 352,  $C_i \in (64, 128, 320, 512)$ .

### 3.2 Global Feature Map Aggregation

The encoder features represent crucial and distinctive information essential for detecting polyp objects. Local features capture intricate details and boundaries, whereas global features provide contextual insights and spatial relationships among various structures. To effectively capture both global and local spatial information, we propose Local-to-Global Spatial Fusion (LGSF), as illustrated in Fig. 2.

The local stage conducts four parallel dilated convolutions [45] with a dilation rate of  $\{1, 2, 4, 8\}$  to extract local features at various spatial scales. Each

dilated convolution is followed by batch normalization (BN) and a rectified linear unit (ReLU). Resultant features from four dilated convolutions are aggregated to obtain the local feature representation. The resulting feature representation is then processed by a spatial attention mechanism (SA) [38] to suppress the irrelevant regions. The detail of the process is provided below:

$$\mathbf{F}_{local} = SA(Concatenate(C_{3 \times 3}^{r=1}(\mathbf{F}), C_{3 \times 3}^{r=2}(\mathbf{F}), C_{3 \times 3}^{r=4}(\mathbf{F}), C_{3 \times 3}^{r=8}(\mathbf{F}))) \quad (2)$$

The global stage incorporates non-local operation [2] to explore the long-range relationships between each pixel in the spatial space. This stage applies a convolution layer to obtain a feature representation. The resulting feature representation is transposed and followed by a Softmax function, and a Hadamard operation on the input feature  $F$  to create a pixel relationship context. This context is then passed through an MLP layer to enhance the relationship representation. In the end, the resulting representation is followed by a sigmoid ( $\sigma$ ) function and another Hadamard operation on the input feature  $F$  to attain the global feature representation. The global stage can be delineated as follows:

$$\mathbf{F}_{global} = \mathbf{F} \times \sigma(MLP(\mathbf{F} \odot Softmax((C_{1 \times 1}(\mathbf{F}))^T))) \quad (3)$$

The final feature representations are obtained by combining local and global information, followed by a convolution layer. Multi-Scale Feature Aggregation (MSFA) module is used to synthesize multi-scale feature representations. This module fuses refined high-level features through a process as depicted in Fig. 2. The first two features,  $F_3$  and  $F_4$  undergo bilinear upsampling to match the spatial dimensions of all three features before they are concatenated. To enhance the capture of non-linear features, we further employ a series of convolutional layers, BN and ReLU. Finally, a sigmoid ( $\sigma$ ) function is conducted to produce the output  $M_{initial}$ , which serves as the initial global feature map.

### 3.3 Self-Enriched Semantic

The shallow layer features are closer to the input more than deep layer features, they preserve more of the original image's details and structure. Therefore, we leverage the low-level features  $F_1$  to query implicit semantics from the initial global feature map, thereby providing supplementary semantics to the deep features. The semantic-enriched deep features are then decoded to yield two semantic-enriched segmentation,  $M_1$  and  $M_2$ . The detailed structure of the Self-Enriched Semantic (SES) module is displayed in Fig. 2. The process can be formulated as following:

$$\mathbf{M}_i = MSFA(\mathbf{F}_2^{rich}, \mathbf{F}_3^{rich}, \mathbf{F}_4^{rich}) \quad (4)$$

$$\mathbf{M} = \sigma(C_{1 \times 1}(Concatenate(\mathbf{M}_1, \mathbf{M}_2))) \quad (5)$$

Firstly, we consider the distribution of pixel values in patch-level images to represent the initial global feature map  $M_{initial}$  to two distinct types of semantic

areas including  $S_1$  and  $S_2$ . Considering patch-level images where exist polyp objects,  $S_1$  includes patches with pixel values varying from 0 and 1, often indicating noise or ambiguous boundaries that have not been sufficiently explored, whereas  $S_2$  consists of patches with pixel values closer to 1, representing solid structures of polyp objects. This categorization helps us differentiate between areas of interest and those that may introduce variability or noise. We then employ the  $F_1$  as query and the  $S_1$  acts as key-value pairs, applying Cross-layer spatial Attention (CA) [31] to ascertain the relevance of  $F_1$  and  $S_1$ . The resultant feature is then sent to high-level features through Attention Gate (AG) units progressively. In the end, we fuse semantic-enriched high-level features using Multi-Scale Fusion Aggregation (MSFA) to achieve a semantic-enriched global feature map,  $M_1$ . By Implementing the same operation to  $F_1$  and  $S_2$ , we also obtain  $M_2$ . Finally, we concatenate  $M_1$  and  $M_2$  before passing them through a convolution layer followed by a sigmoid ( $\sigma$ ) function to predict the final global feature map,  $M$ .

## 4 Experiments

### 4.1 Dataset and Evaluation Metrics

Following recent cutting-edge solutions for the polyp segmentation task, we employ five widely-used benchmark datasets to assess the efficacy of our proposed model. These datasets include Kvasir [13], ClinicDB [1], ColonDB [30], ETIS [28], and EndoScene [32]. Table 1 provides a comprehensive overview of each dataset, including their specific usage details and objectives.

We employ various standard metrics to assess and compare the performance of polyp segmentation algorithms. The Dice score quantifies the spatial agreement between the predicted segmentation mask and the ground truth mask, whereas the IoU score computes the ratio of their overlapping area to the combined area. Both scores range between 0 and 1, with higher values indicating better segmentation performance. The MAE computes the average absolute difference between individual pixels in the predicted and ground truth masks. These evaluation metrics offer a comprehensive assessment of the segmentation performance, considering both spatial alignment and pixel-level precision.

### 4.2 Implementation Details

We utilize the power of RTX 3090 GPU to accelerate both the training and inference stages of our model. Throughout the training process, we monitor various metrics including loss function, mDice, mIoU, and MAE scores to assess the performance and guide the training process. The total duration of training amounts to approximately 2 hours to achieve optimal performance. Detailed training parameters are provided in Table 2.

### 4.3 Comparisons with State-of-the-art Methods

This section conducts a comprehensive evaluation focusing on two critical aspects: Learning ability, which verifies the segmentation performance on the

**Table 1:** Specific usage details and objectives of Kvasir, ClinicDB, ColonDB, ETIS and EndoScene Datasets.

Dataset	Image	Size	Train	Test	Objective
Kvasir	1000	Variable	900	100	Learning
ClinicDB	612	384 × 288	550	62	Learning
ColonDB	380	574 × 500	-	380	Generalization
ETIS	196	1225 × 966	-	196	Generalization
EndoScene	60	574 × 500	-	60	Generalization

**Table 2:** Parameters of training configuration.

Image Size	Batch Size	Epoch	Loss Function
352 × 352	16	200	wBCE [40] + wDice [20]
Optimizer	Learning Rate	Decay Rate	Weight Decay
AdamW [17]	1e-4	1e-1	1e-4

seen dataset, and generalization ability, which evaluates the capacity of the model to generalize effectively to unseen data. A total of sixteen state-of-the-art models from the domain of the polyp segmentation, including U-Net [26], UNet++ [50], PraNet [7], SFA [8], MSEG [12], ACSNet [47], DCRNet [44], EU-Net [25] and SANet [36], alongside newer models such as Polyp-PVT [5], ADSNet [22], CaraNet [18], TransUnet [3], Transfuse [48], UCTransNet [33], SSFormer [34], are collected for comparative analysis. The performance of these models is meticulously evaluated on five benchmark datasets using mDice, mIoU, and Mean Absolute Error (MAE) scores. In order to ensure fairness and reproducibility in our comparative analysis, we meticulously maintained consistency across training, validation, and testing datasets for all assessed models. Following the methodology outlined in PraNet [7], we adopt an identical dataset configuration as illustrated in Table 1, comprising 900 and 548 images sourced from the Kvasir and ClinicDB datasets as the training set, with the remaining 64 and 100 images allocated as the respective test set to evaluate the learning ability. Additionally, we utilize the ColonDB, ETIS, and EndoScene datasets, which were not included in the training phase, to assess generalization ability.

**Learning ability.** In the learning ability experiment, the domain of the test and train set is similar. Table 3 presents the results of different cutting-edge models on the Kvasir and ClinicDB datasets. Our method demonstrates outstanding performance compared to recently published models on both datasets, as evidenced by the mDice, mIoU, and MAE scores. Specifically, our method obtains a mDice score of 0.924, a mIoU score of 0.875 on the Kvasir dataset, outperforming the second-best model ADSNet [22]. On the ClinicDB dataset, our model achieves a

**Table 3:** Learning ability of diverse polyp segmentation baselines on Kvasir & ClinicDB datasets across mDice, mIoU, MAE scores.  $\uparrow$  denotes higher the better and  $\downarrow$  denotes lower the better. **Bold** denotes the best score among the models, and underline denotes the second best.

Methods	Kvasir			ClinicDB		
	mDice $\uparrow$	mIoU $\uparrow$	MAE $\downarrow$	mDice $\uparrow$	mIoU $\uparrow$	MAE $\downarrow$
U-Net [26]	0.818	0.746	0.055	0.823	0.755	0.019
UNet++ [50]	0.821	0.743	0.048	0.794	0.729	0.022
SFA [8]	0.723	0.611	0.075	0.700	0.607	0.042
MSEG [12]	0.897	0.839	0.028	0.909	0.864	<u>0.007</u>
DCRNet [44]	0.886	0.825	0.035	0.896	0.844	0.010
ACSNet [47]	0.898	0.838	0.032	0.882	0.826	0.011
PraNet [7]	0.898	0.840	0.030	0.899	0.849	0.009
EU-Net [25]	0.908	0.854	0.028	0.902	0.846	0.011
SANet [36]	0.904	0.847	0.028	0.916	0.859	0.012
Polyp-PVT [5]	0.917	0.864	<u>0.023</u>	0.937	0.889	<b>0.006</b>
ADSNet [22]	0.920	<u>0.871</u>	<b>0.020</b>	0.938	0.890	<b>0.006</b>
CaraNet [18]	0.918	0.865	<u>0.023</u>	0.936	0.887	<u>0.007</u>
TransUnet [3]	0.913	0.857	0.028	0.935	0.887	0.008
TransFuse [48]	0.920	0.870	<u>0.023</u>	<u>0.942</u>	<u>0.897</u>	<u>0.007</u>
UCTransNet [33]	0.918	0.860	<u>0.023</u>	0.933	0.860	0.008
<b>Polyp-SES</b>	<b>0.924</b>	<b>0.875</b>	<b>0.020</b>	<b>0.945</b>	<b>0.902</b>	<b>0.006</b>

mDice score and mIoU of 0.945 and 0.902, respectively, showcasing an improvement compared to TransFuse [48]. These results underscore the robustness and effectiveness of the proposed method in terms of learning ability.

**Generalization ability.** We conduct a thorough evaluation of the polyp segmentation baselines to assess their generalization performance on unseen datasets, as shown in Table 4. It can be observed that our method demonstrates competitive performance across all three datasets compared to other techniques. Specifically, our model is higher than the second-best ADSNet [22] on the ColonDB dataset in term of mDice score and mIoU score. On ETIS dataset, although Transfuse [48] exhibits notable performance with a mIoU score of 0.826, its corresponding mDice score is lower at 0.737. In contrast, our results achieve a mDice score of 0.805, outperforming all other models, alongside a mIoU score of 0.756. These findings highlight the stable performance of our proposed approach, which excels in both mDice and mIoU scores where other methods may have limitations. Additionally, our model demonstrates remarkable improvement on the EndoScene dataset, with mDice score, mIoU score, and MAE score of 0.911, 0.847, and 0.005, respectively. These results underscore the superior generalization capability of our proposed method.

**Qualitative results.** We present qualitative results comparing our model with other polyp segmentation baselines across five datasets, depicted in Fig. 3 and Fig. 4. The segmentation results of the compared methods are sourced from the publicly available Polyp-PVT [5]. We can observe that our model produces clear and precise segmentation outcomes across a variety of polyp structures. Furthermore, it effectively identifies and segments polyp objects under different

**Table 4:** Generalization ability of diverse polyp segmentation baselines on ColonDB, ETIS & EndoScene datasets across mDice, mIoU, MAE scores.  $\uparrow$  denotes higher the better and  $\downarrow$  denotes lower the better. **Bold** denotes the best score among the models, and underline denotes the second best.

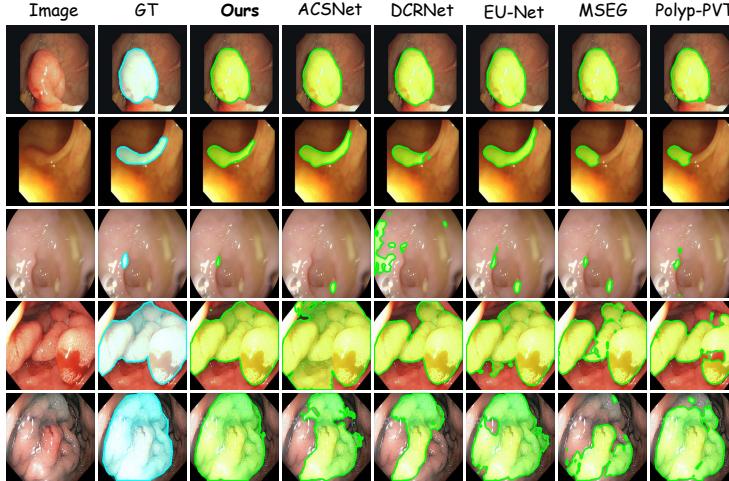
Methods	ColonDB			ETIS			EndoScene		
	mDice $\uparrow$	mIoU $\uparrow$	MAE $\downarrow$	mDice $\uparrow$	mIoU $\uparrow$	MAE $\downarrow$	mDice $\uparrow$	mIoU $\uparrow$	MAE $\downarrow$
U-Net [26]	0.512	0.444	0.061	0.398	0.335	0.036	0.710	0.627	0.022
UNet++ [50]	0.483	0.410	0.064	0.401	0.344	0.035	0.707	0.624	0.018
SFA [8]	0.469	0.347	0.094	0.297	0.217	0.109	0.467	0.329	0.065
MSEG [12]	0.735	0.666	0.038	0.700	0.630	0.015	0.874	0.804	0.009
DCRNet [44]	0.704	0.631	0.052	0.556	0.496	0.096	0.856	0.788	0.010
ACSNet [47]	0.716	0.649	0.039	0.578	0.509	0.059	0.863	0.787	0.013
PraNet [7]	0.712	0.640	0.043	0.628	0.567	0.031	0.871	0.797	0.010
EU-Net [25]	0.756	0.681	0.045	0.687	0.609	0.067	0.837	0.765	0.015
SANet [36]	0.753	0.670	0.043	0.750	0.654	0.015	0.888	0.815	0.008
Polyp-PVT [5]	0.808	0.727	0.031	0.787	0.706	0.013	0.900	0.833	<u>0.007</u>
ADSNet [22]	0.815	0.730	<u>0.029</u>	0.798	0.715	<u>0.012</u>	0.890	0.819	0.010
CaraNet [18]	0.773	0.689	0.042	0.747	0.672	0.017	<u>0.903</u>	<u>0.838</u>	<u>0.007</u>
TransUnet [3]	0.781	0.699	0.036	0.731	<u>0.824</u>	0.021	0.893	0.660	0.009
TransFuse [48]	0.781	0.706	0.035	0.737	<b>0.826</b>	0.020	0.894	0.654	0.009
SSFormer [34]	0.772	0.697	0.036	0.767	0.698	0.016	0.887	0.821	<u>0.007</u>
<b>Polyp-SES</b>	<b>0.817</b>	<b>0.741</b>	<b>0.026</b>	<b>0.805</b>	0.756	<b>0.011</b>	<b>0.911</b>	<b>0.847</b>	<b>0.005</b>

variations in image quality, minimizing artifacts and extraneous regions while maintaining exceptional segmentation accuracy. These findings underscore the efficiency and accuracy of our proposed segmentation algorithm, even in challenging spatial scenarios where previous methods have struggled.

#### 4.4 Ablation Study

In the ablation study section, we conduct experiments to validate the necessity and effectiveness of each proposed module in the overall architecture individually. Our standard polyp segmentation architecture includes an Encoder, Decoder and Self-Enriched Semantic (SES). The ablation studies are conducted on all five polyp datasets, evaluating based on mDice and mIoU scores.

**Effectiveness of Encoder Backbone** In the first ablation study, we assess the effectiveness of different encoder backbones. We use the proposed standard architecture as the baseline and replace diverse encoder backbones, consisting of ResNet50 [11] (CNN), PVT [35] (Transformer), and Caformer [46] (Metaformer). All variants are trained under the same configuration, and the results are summarized in Table 5. It is evident that the standard baseline, with Caformer as the encoder backbone, achieves superior performance with higher mDice and mIoU scores across all five datasets compared to CNN-based or conventional transformer encoder backbones. This demonstrates the effectiveness of exploiting the vision metaformer as encoder backbone in extracting robust features and enhancing polyp segmentation performance.

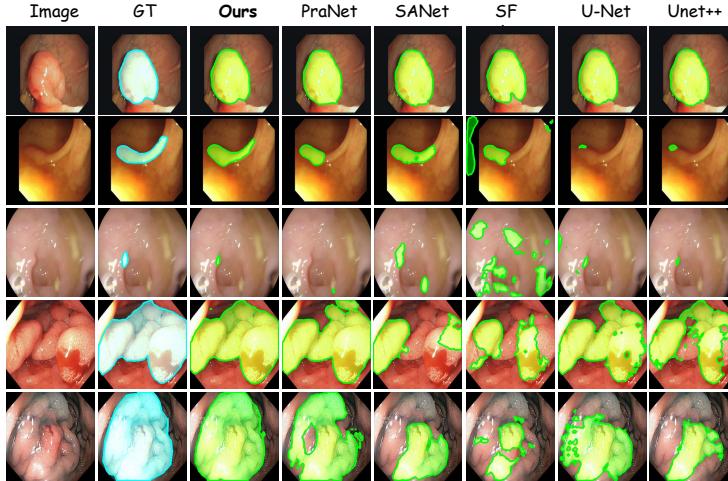


**Fig. 3:** Qualitative results with the current polyp segmentation baselines. Green indicates a predicted mask. It can be found, our proposed model can precisely recognize and segment polyp objects even under the variability in polyp appearance attached to noises, ambiguous boundaries, and intricate foregrounds.

**Table 5:** Ablation study of various encoder backbones over five benchmarks.  $\uparrow$  denotes higher the better. **Bold** denotes the best score.

Encoder	Type	Kvasir		ClinicDB		ColonDB		ETIS		EndoScene	
		mDice $\uparrow$	mIoU $\uparrow$								
ResNet50 [11]	CNN	0.909	0.852	0.932	0.880	0.797	0.722	0.804	0.727	0.895	0.827
PVT [35]	Transformer	0.919	0.870	0.933	0.884	0.804	0.726	0.779	0.695	0.892	0.826
Caformer [46]	Metaformer	<b>0.924</b>	<b>0.875</b>	<b>0.945</b>	<b>0.902</b>	<b>0.817</b>	<b>0.741</b>	<b>0.805</b>	<b>0.756</b>	<b>0.911</b>	<b>0.847</b>

**Effectiveness of Local-to-Global Spatial Fusion** To assess the impact of local and global feature aggregation, we remove the LGSF units from the decoder in the standard architecture, and replace them with  $3 \times 3$  convolution layers. Results presented in Table 6 demonstrate a significant decrease in both mDice and mIoU scores compared to the standard baseline with LGSF units. Furthermore, visualizations of segmentation predictions in Fig. 5 reveal that the absence of LGSF introduces considerable noise. These qualitative and quantitative results prove that LGSF can help model to distinguish polyp tissues and contribute greatly to the polyp segmentation performance. In order to further explore the contribution of the LGSF, we showcase high-level features before and after refinement by the LGSF units in Fig. 6. As can be observed, the LGSF eliminate redundant information from other regions and yield informative characteristics of level-specific features, aiding the model in precisely locating polyp objects and enhancing segmentation performance.

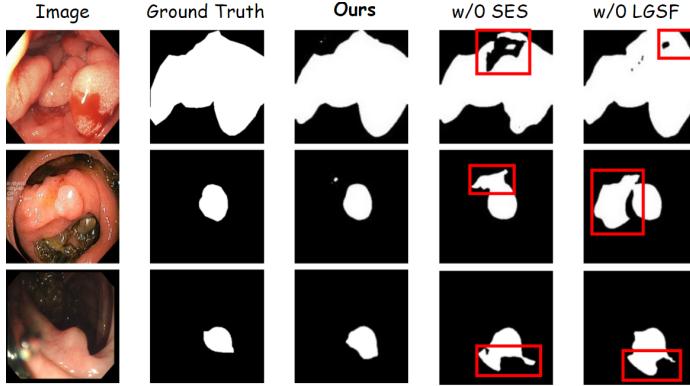


**Fig. 4:** Qualitative results with the current polyp segmentation baselines. Green indicates a predicted mask. It can be found, our proposed model can precisely recognize and segment polyp objects even under the variability in polyp appearance attached to noises, ambiguous boundaries, and intricate foregrounds.

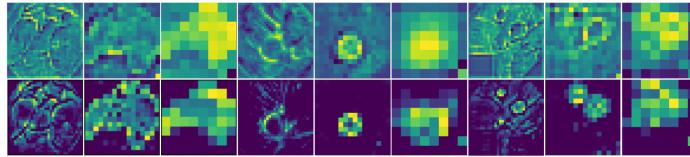
**Table 6:** Ablation study of LGSF and SES over five benchmarks.  $\uparrow$  denotes higher the better. **Bold** denotes the best score

Method	Kvasir		ClinicDB		ColonDB		ETIS		EndoScene	
	mDice $\uparrow$	mIoU $\uparrow$								
w/o SES, LGSF	0.900	0.850	0.909	0.862	0.775	0.699	0.691	0.615	0.891	0.819
w/o SES	0.905	0.853	0.923	0.874	0.784	0.708	0.729	0.654	0.888	0.811
w/o LGSF	0.918	0.869	0.912	0.868	0.781	0.694	0.786	0.702	0.888	0.824
<b>Ours</b>	<b>0.924</b>	<b>0.875</b>	<b>0.945</b>	<b>0.902</b>	<b>0.817</b>	<b>0.741</b>	<b>0.805</b>	<b>0.756</b>	<b>0.911</b>	<b>0.847</b>

**Effectiveness of Self-Enriched Semantic** This ablation study validates the effectiveness of the proposed SES module on the overall architecture. By excluding the SES module from the baseline, we revert to a conventional encoder-decoder structure. The performance presented in Table 6 reveals that the conventional encoder-decoder architecture without SES leads to a deterioration in performance, with lower on mDice score and mIoU score compared our standard model. In Fig. 5, it is apparent that the absence of the SES results in more detailed errors or missed semantic areas. This proves that the SES module facilitates the model to explore potential semantics to give the better global feature map with the comprehensive context. We further investigate the contribution of the SES by visualizing the two semantic-enriched segmentation masks containing  $M_1$  and  $M_2$  in Fig. 7. Notably,  $M_1$  demonstrates the ability to explore potential semantic areas referring to regions denoted as red-bordered boxes where were



**Fig. 5:** Visualization of the ablation study results. As can be seen, removing Self-Enrich Semantic (SES) leads to segmentation failures in challenging semantic areas, whereas the removal of Local-to-Global Spatial Fusion (LGSF) causes incorrectly segmentation results denoted as red-bordered boxes.

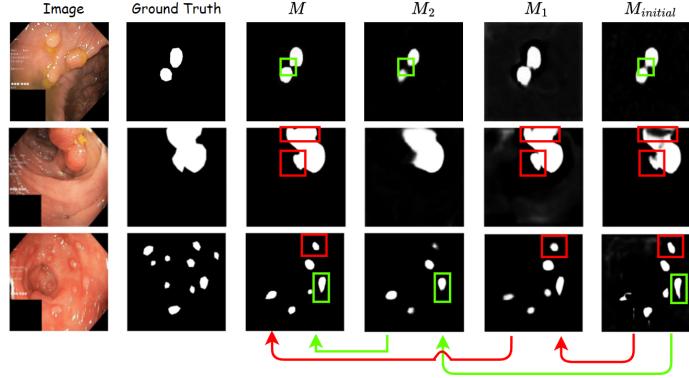


**Fig. 6:** Visualization of the high-level feature maps before and after refining by the Local-to-Global Spatial Fusion (LGSF). The first row is high-level features, the second row is features which are refined by LGSF. As can be observed, the LGSF captures helpful characteristics of level-specific features and removes redundant information from other regions.

not previously captured by  $M_{initial}$ . Meanwhile,  $M_2$  concerns the solid structural components of polyp objects in green-bordered boxes where already captured by  $M_{initial}$ . Taking advantage of  $M_1$  and  $M_2$ , we attain a final global feature map with comprehensive semantics, thereby improving polyp segmentation performance.

## 5 Conclusion

In this paper, we introduce “Automatic Polyp Segmentation with Self-Enriched Semantic Model”, an innovative approach aimed at addressing the limitations of contemporary methods in capturing comprehensive contexts. By leveraging a vision metaformer Encoder, a Decoder, and a Self-Enriched Semantic module, our method effectively enriches deep features with supplementary semantics, improving the model’s understanding of challenging contexts. Through quantitative and qualitative experiments, we demonstrate its effectiveness and superiority



**Fig. 7:** Visualization of  $M_{initial}$ ,  $M_1$ ,  $M_2$  and  $M$  predictions. It is found that the final global feature map  $M$  are constructed and contributed by two semantic-enriched segmentation masks  $M_1$ , and  $M_2$ . For example, in the **first row**,  $M_2$  separates successfully two close polyp objects. In the **second row**,  $M_1$  explores missing semantic areas to reconstruct the feature map. In the **third row**,  $M_1$  and  $M_2$  relocate tiny polyp objects more precisely.

over state-of-the-art models across five polyp benchmarks, evaluated on mDice, mIoU, and MAE metrics, showcasing its proficiency in both learning and generalization abilities. Additionally, we conducted through studies to understand the underlying reasons for its effectiveness, offering valuable insights that can guide future research in medical image segmentation-related tasks, particularly those focused on automatic polyp segmentation.

## Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government(MSIT). This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2024-00437718) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). This study was supported by a grant (HCRI 23038) Chonnam National University Hwasun Hospital Institute for Biomedical Science. The corresponding author is Soo-Hyung Kim.

## References

1. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilarín, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *CMIG* **43**, 99–111 (2015)

2. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: ICCVW. pp. 0–0 (2019)
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
5. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932 (2021)
6. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: MICCAI. pp. 263–273 (2020)
8. 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 (2019)
9. Feng, Y., Zhao, H., Li, X., Zhang, X., Li, H.: A multi-scale 3d otsu thresholding algorithm for medical image segmentation. DSP **60**, 186–199 (2017)
10. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: WACV. pp. 574–584 (2022)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Huang, C.H., Wu, H.Y., Lin, Y.L.: Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. arXiv preprint arXiv:2101.07172 (2021)
13. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvadir-seg: A segmented polyp dataset. In: MMM. pp. 451–462 (2020)
14. Jha, D., Tomar, N.K., Sharma, V., Bagci, U.: Transnert: Transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing. arXiv preprint arXiv:2303.07428 (2023)
15. Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unifying convolution and self-attention for visual recognition. IEEE TPAMI (2023)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
18. Lou, A., Guan, S., Loew, M.: Caranet: context axial reverse attention network for segmentation of small medical objects. Journal of Medical Imaging **10**(1), 014005–014005 (2023)
19. Mamonov, A.V., Figueiredo, I.N., Figueiredo, P.N., Tsai, Y.H.R.: Automated polyp detection in colon capsule endoscopy. IEEE transactions on medical imaging **33**(7), 1488–1502 (2014)
20. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: IC3DV. pp. 565–571. Ieee (2016)
21. Mubarak, D.M.N., Sathik, M.M., Beevi, S.Z., Revathy, K.: A hybrid region growing algorithm for medical image segmentation. IJCSIT pp. 61–70 (2012)

22. Nguyen, Q.V., Huynh, V.T., Kim, S.H.: Adaptation of distinct semantics for uncertain areas in polyp segmentation. In: BMVC (2023)
23. Nguyen, Q.V., Tran, T.T., Pham, V.T.: Gca-net: Geometrical constraints-based advanced network for polyp segmentation. In: 2022 9th NAFOSTED Conference on Information and Computer Science (NICS). pp. 241–246. IEEE (2022)
24. Nguyen, Q.V., Tran, T.T., et al.: Fcmd-net: A full-connection multi-decoder network for polyp segmentation. In: 2022 6th International Conference on Green Technology and Sustainable Development (GTSD). pp. 1070–1075. IEEE (2022)
25. Patel, K., Bur, A.M., Wang, G.: Enhanced u-net: A feature enhancement network for polyp segmentation. In: 2021 18th conference on robots and vision (CRV). pp. 181–188. IEEE (2021)
26. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
27. Shen, T., Li, H., Huang, X.: Active volume models for medical image segmentation. *IEEE TMI* **30**(3), 774–791 (2010)
28. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *IJCARS* **9**, 283–293 (2014)
29. Tajbakhsh, N., Gurudu, S., Liang, J.: Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. *IEEE ISBI* **2015**, 79–83 (2015)
30. 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)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *NeurIPS* **30** (2017)
32. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A., et al.: A benchmark for endoluminal scene segmentation of colonoscopy images. *JHE* **2017** (2017)
33. Wang, H., Cao, P., Wang, J., Zaiane, O.R.: Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: AAAI. vol. 36, pp. 2441–2449 (2022)
34. Wang, J., Huang, Q., Tang, F., Meng, J., Su, J., Song, S.: Stepwise feature fusion: Local guides global. In: MICCAI. pp. 110–120 (2022)
35. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: ICCV. pp. 568–578 (2021)
36. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: MICCAI. pp. 699–708 (2021)
37. Wenxuan, W., Chen, C., Meng, D., Hong, Y., Sen, Z., Jiangyun, L.: Transbts: Multimodal brain tumor segmentation using transformer. In: MICCAI. pp. 109–119 (2021)
38. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
39. Wu, C., Long, C., Li, S., Yang, J., Jiang, F., Zhou, R.: Msraformer: Multiscale spatial reverse attention network for polyp segmentation. *CBM* **151**, 106274 (2022)
40. Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV. pp. 1395–1403 (2015)
41. Xie, Y., Zhang, J., Shen, C., Xia, Y.: Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In: MICCAI. pp. 171–180 (2021)
42. Xu, C., Pham, D.L., Prince, J.L.: Image segmentation using deformable models. *Handbook of medical imaging* **2**(20), 0 (2000)

43. Yao, J., Miller, M., Franaszek, M., Summers, R.M.: Colonic polyp segmentation in ct colonography-based on fuzzy clustering and deformable models. *IEEE TMI* **23**(11), 1344–1352 (2004)
44. Yin, Z., Liang, K., Ma, Z., Guo, J.: Duplex contextual relation network for polyp segmentation. In: *IEEE ISBI*. pp. 1–5 (2022)
45. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
46. Yu, W., Si, C., Zhou, P., Luo, M., Zhou, Y., Feng, J., Yan, S., Wang, X.: Metaformer baselines for vision. *IEEE TPAMI* **46**(2), 896–912 (2024)
47. Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., Yu, Y.: Adaptive context selection for polyp segmentation. In: *MICCAI*. pp. 253–262 (2020)
48. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: *MICCAI*. pp. 14–24 (2021)
49. Zhao, X., Zhang, L., Lu, H.: Automatic polyp segmentation via multi-scale subtraction network. In: *MICCAI*. pp. 120–130 (2021)
50. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE TMI* **39**(6), 1856–1867 (2019)