Received: 14 January 2022 | Revised: 23 June 2022 | Accepted: 31 July 2022

DOI: 10.1002/ima.22795

RESEARCH ARTICLE

WILEY

Squeeze and multi-context attention for polyp segmentation

Debayan Bhattacharya[1,2] | **Dennis Eggert**[2] | **Christian Betz**[2] | **Alexander Schlaefer**[1]

[1]Institute of Medical Technology and Intelligent Systems, Hamburg University of Technology, Hamburg, Germany

[2]Clinic for Ears, Nose and Throat, University Medical Center, Hamburg, Germany

Correspondence
Debayan Bhattacharya, Institute of Medical Technology and Intelligent Systems, Hamburg University of Technology, Am Schwarzenberg-Campus 1, Hamburg 21073, Germany.
Email: debayan.bhattacharya@tuhh.de

Funding information
Free and Hanseatic City of Hamburg; Hamburg University of Technology; University Hospital Hamburg-Eppendorf

Abstract

Artificial Intelligence-based Computer Aided Diagnostics (AI-CADx) have been proposed to help physicians in reducing misdetection of polyps in colonoscopy examination. The heterogeneity of a polyp's appearance makes detection challenging for physicians and AI-CADx. Towards building better AI-CADx, we propose an attention module called Squeeze and Multi-Context Attention (SMCA) that re-calibrates a feature map by providing channel and spatial attention by taking into consideration highly activated features and context of the features at multiple receptive fields simultaneously. We test the effectiveness of SMCA by incorporating it into the encoder of five popular segmentation models. We use five public datasets and construct intra-dataset and inter-dataset test sets to evaluate the generalizing capability of models with SMCA. Our intra-dataset evaluation shows that U-Net with SMCA and without SMCA has a precision of 0.86 ± 0.01 and 0.76 ± 0.02 respectively on CVC-ClinicDB. Our inter-dataset evaluation reveals that U-Net with SMCA and without SMCA has a precision of 0.62 ± 0.01 and 0.55 ± 0.09 respectively when trained on Kvasir-SEG and tested on CVC-ColonDB. Similar results are observed using other segmentation models and other public datasets. In conclusion, we demonstrate that incorporating SMCA into the segmentation models leads to an increase in generalizing capability of the segmentation models.

KEYWORDS

attention, attention gate, polyp segmentation, squeeze and excite, squeeze and multi-context, U-Net

1 | INTRODUCTION

Colon Cancer can be fatal if not detected early and as such, poses a huge risk to public health. It is the third most common cause of cancer in the US.[1] One of the earliest signs of colon cancer is the emergence of polyps in the colon and rectum. Early detection and removal of polyps can increase the survival rate to 90%.[2] To this end, colonoscopy is performed to detect the presence of colorectal polyps. The problem with manual inspection is that polyps can be misdetected because they have heterogeneous morphological characteristics. Hence, there is an ongoing effort to develop Computer Aided Diagnosis Systems (CADx) that limit the number of misdetections.[3]

Artificial Intelligence (AI) based Polyp Segmentation is a paradigm of AI-CADx where an AI model is

purposed with the task of classifying the pixels that belong to polyps in images. Specifically, deep learning-based AI methods show promising results.[4] It is believed that AI-CADx will reduce the burden of a physician and lead to better patient care. It is also argued that CADx solutions could potentially be an alternative to manual screening. Therefore, it is of paramount importance that the accuracy and precision of deep learning-based AI-CADx are improved.

As argued by Jha et al.,[5] robustness and generalizability are two key aspects that need to be handled if we want CADx systems in clinical practice. The robustness is the ability of the CADx to perform reliably within an accepted error margin for all kinds of colonoscopic images. Generalization is the ability of the CADx to segment polyps reliably and accurately from images belonging to a wide range of image distributions. Solving these two aspects are key to making reliable AI-CADx for polyp segmentation. Figure 1 shows the variations in appearance and morphological features of polyps across different datasets.

Towards learning robust and generalizing features for polyp segmentation, we propose a module called "Squeeze and Multi-Context Attention" (SMCA), an attention module that re-calibrates feature maps based on attention weights computed from the aggregated polyp and context features at multiple receptive fields. In doing so, we leverage the global context and the local context at multiple receptive fields to provide spatial and channel attention. In comparison, Squeeze and Excite (SE)[10] module extracts only global context through global average pooling to provide channel attention. Attention gates (AG)[11] provide spatial attention by calculating attention weights from coarser signals for each feature in a feature map. However, our module combines the channel attention mechanism from SE and the spatial attention mechanism from AG to compute attention weights that provide attention in the channel and spatial dimensions. Additionally, we perform the channel and spatial attention at multiple receptive fields. A point to note is that SMCA is a self-attention module whereas AG is an attention module. We evaluate the effectiveness of our module by incorporating it into multiple deep learning-based segmentation models, namely: U-Net,[12] Attention U-Net,[11] R2U-Net,[13] R2AU-Net[14] and ResUNet++.[15] In ResUNet++, we replace SE module with SMCA module. Towards robustness, we evaluate the five models with and without SMCA on four public datasets. Towards generalization, we construct inter-dataset test sets and evaluate the segmentation models with and without SMCA on them. Finally, we compare the attention maps of the convolution kernels of U-Net with and without our SMCA module using Grad-Cam++[16] to qualitatively illustrate the differences in the feature representation.

In summary, our contributions are as follows:

- We propose an attention module called SMCA that takes global and local context at multiple receptive fields to re-calibrate the feature maps.

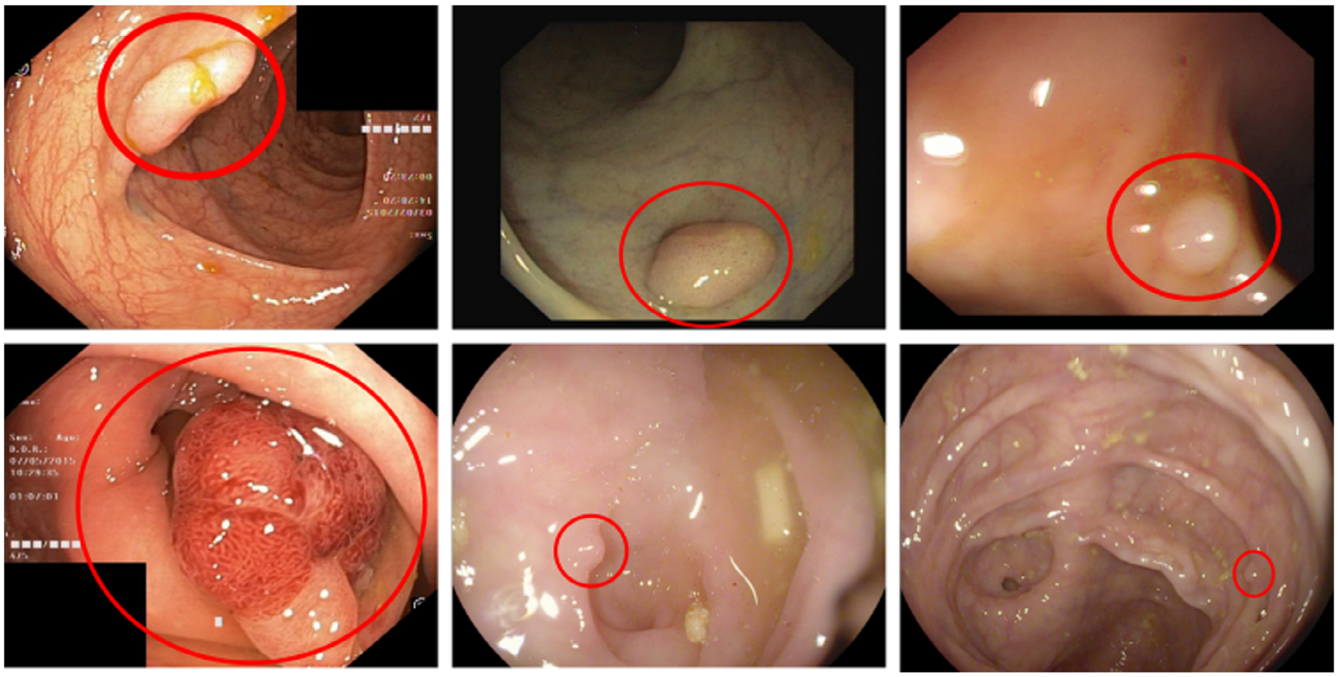

FIGURE 1 Sample images from Kvasir-Seg,[6] ETIS-Larib,[7] CVC-ColonDB[8] and CVC-ClinicDB[9] illustrating the variations in appearance and morphological features of polyps shown with red circles

- We check the performance changes due to SMCA by extensively evaluating five models with and without SMCA through five-fold cross validation on four public datasets that is, Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB and Kvasir-Sessile.
- We check the generalizing ability through extensive inter-dataset evaluation that is, we train our models with and without SMCA on Kvasir-SEG and CVC-Clinic. We then evaluate these models on four public datasets, which the model has not seen before.
- We plot the attention maps of the convolution kernels of a U-Net with and without our SMCA. The comparison of the attention maps at multiple hierarchies highlights the differences in the feature representation of the two models.

2 | RELATED WORK

Previously, hand-crafted features were used to detect and segment polyps. In Reference 8, authors proposed a three-stage process for polyp segmentation. They performed region segmentation followed by region description and finally, region classification. The authors of[17] used shape as a discriminatory feature instead of texture. The reason behind their proposal was that small polyps had predominantly elliptical shapes. In Reference 18, authors proposed a dictionary learning approach by extracting hue histogram features and used support vector machine to classify normal and polyp images. However, the limitation of hand-crafted features is that they do not generalize well to unseen images. Furthermore, the complexity of these proposed solutions greatly limit the applicability in real-world scenarios. The limitations posed by hand-crafted feature extraction have been circumvented by using Convolutional Neural Networks (CNN). CNNs have shown great success in the polyp segmentation task. In the MICCAI polyp segmentation challenge, most of the proposed models were based on CNNs and the winning model was also a CNN.[19] Since U-Net[12] came into existence, it and its variants have been commonly used in medical image segmentation.[20]

From the literature, it can be observed that the modifications proposed by authors have mostly been in convolution operations, attention blocks and feature aggregation blocks. With respect to changes in convolution operations, Alam et al.[21] replaced the encoder of U-Net with ResNet-50 and Sun et al.[22] extracted better features by using dilated convolution. Towards the use of attention blocks, one of the earliest architectures was the Attention U-Net[11] which incorporated attention gates to improve segmentation of abdominal regions from CT images. Rundo et al.[23] introduced SE modules into U-Net to improve prostate zonal segmentation. On similar lines, Jha et al.[15] created a variant of ResUNet[24] for polyp segmentation by introducing SE module and attention gates. In Reference 25, the authors introduced a spatial attention layer to a U-Net for the task of polyp segmentation. In Reference 26, authors introduced an attention module called "Focus Gate" that uses spatial and channel attention to calculate the attention weights. The authors demonstrated that their dual attention-gated U-Net called "Focus Net" outperformed state-of-the-art models. With respect to feature aggregation blocks, Mahmud et al. proposed PolypSeg-Net[27] where sequential depth dilated inception (DDI) blocks were used to aggregate features from different receptive fields.

From the aforementioned works, we observed that using channel and spatial attention blocks and aggregating features from multiple receptive fields were beneficial for segmentation. SMCA module was constructed with these two ideas in mind. Specifically, our SMCA module captures information at multiple receptive fields of a feature map by using average and max pooling of varying kernel sizes. The extracted information from the multiple receptive fields is passed through convolutional blocks to calculate spatial and channel attention weights per receptive field. The channel and spatial attention weights from multiple receptive fields are combined to calculate the final attention weights which are used to re-calibrate the original feature map.

The literature also revealed that majority of the existing works evaluated their models on test sets, which were derived from the same datasets.[28–31] An exception to this trend was the recently published work of Jha et al.[5] They performed inter-dataset evaluation to prove the generalizing capability of their proposed model. However, they performed one-fold cross validation for all their experiments. We take this a step further and perform five-fold cross validation experiments to prove the advantages of incorporating SMCA into models to increase its generalizing ability.

3 | METHODOLOGY

3.1 | Network architecture

In this sub-section, we briefly describe the various models we considered for this study and illustrate through diagrams where we placed the SMCA module in the models.

3.1.1 | U-Net architecture

The architecture of the proposed U-Net with SMCA is shown in Figure 2. It consists of encoder, decoder and

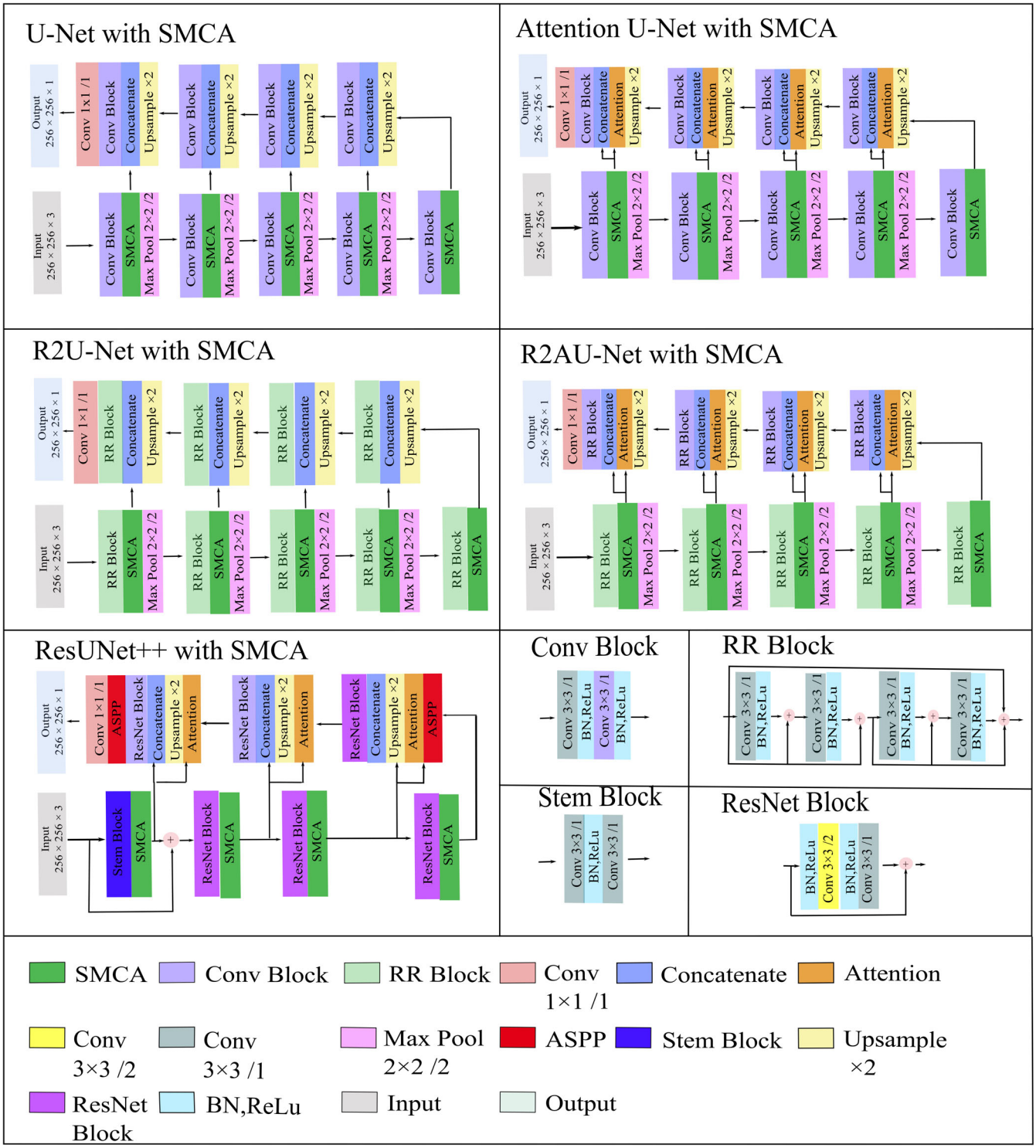

FIGURE 2 The five segmentation models used in our work. The original models do not have SMCA module. In ResUNet++, we replaced SE Layer with SMCA. /1 and /2 represent stride 1 and stride 2. 2 × 2 and 3 × 3 denote kernel sizes. ×2 next to Upsample denotes the scale of upsampling. All Upsampling operations are bilinear interpolation

the SMCA module. The encoder extracts feature through a series of encoding blocks. As information passes down the encoder block, low-level features are converted to high-level features. An encoder block is a series of two convolution operations followed by SMCA and Max Pooling. Before information is passed to the next encoder block, SMCA enhances the features extracted. In the baseline U-Net, the SMCA module is not present. The number of kernels increases in subsequent encoder blocks as follows: 32, 64, 128, 256 and 512. The decoder

block is similar to the encoder with the additional operation being that it concatenates features from the encoder with the upsampled features from the previous decoder block. The decoder kernels decrease in every subsequent decoder block as follows: 256, 128, 64 and 32.

3.1.2 | Attention U-Net architecture

The architecture of the proposed Attention U-Net with SMCA is shown in Figure 2. It consists of encoder, decoder, SMCA module and an additional attention gate.[32] Similar to U-Net, SMCA enhances the features extracted from an encoder block and then the feature is passed to the next block. The SMCA module is absent in the baseline Attention U-Net. The number of kernels increases after every encoder block as follows: 32, 64, 128, 256 and 512. On the other hand, the de-coder kernels decrease as follows: 256, 128, 64 and 32.

3.1.3 | R2U-Net architecture

In this variation of U-Net, Recurrent Residual Convolutional Neural Networks (RRCNN) are introduced. The authors propose the inclusion of these two modules primarily for two reasons. First, the inclusion of residual units helps in training deep architectures as it minimises the occurrence of vanishing and exploding gradients. Secondly, recurrent units ensure better feature representations arising from the accumulation of feature maps. This network achieved state-of-the-art in Skin Lesion Segmentation.[33] The encoder and decoder structures are shown in Figure 2. The number of kernels increases after every encoder block as follows: 64, 128, 256, 512 and 1024. On the decoder side, the kernels reduce with every decoder block as follows: 512, 256, 128 and 64.

3.1.4 | R2AU-Net architecture

In this variation of U-Net, attention gates introduced in Attention U-Nets are used in R2U-Net. Inclusion of attention gate further strengthens the feature representation of the network. The encoder and decoder structures are shown in Figure 2. The number of kernels used in the encoder and decoder blocks is the same as R2U-Net.

3.1.5 | ResUNet++ architecture

ResUNet++ is a segmentation model constructed to improve polyp segmentation performance. This model is built upon ResUNet.[24] This architecture has feature enhancement modules such as Residual Blocks, SE module, Attention Gates and Atrous Spatial Pyramid Pooling (ASPP).[34] The SE layers are included after every residual block in the encoder. Additional skip connections are introduced to propagate information from the encoder blocks to attention gates. The filters in the encoder section increase as follows: 32, 64, 128, 256 and 512. In the decoder section, the filters decrease with each decoder block as follows: 512, 256, 128, 64 and 32. Altogether, ResUNet++ has one stem block (See Figure 2), three encoder blocks and three decoder blocks. The final decoder block has an ASPP layer and a 1×1 convolution for channel reduction.

3.2 | Feature enhancement modules

We consider feature enhancement modules to encompass modules that manipulate feature maps through convolution operations or recalibrate the feature maps by computing attention weights. In this section, we describe the various feature enhancement modules used in the five segmentation models.

3.2.1 | Attention gates

Attention gates were first proposed by Chen et al.[35] Since its introduction, several segmentation models have been used it. In our work, three of the models (Attention U-Net, R2AU-Net, ResU-net++) use attention gates. The reason for using the attention mechanism is that it highlights the relevant information in a feature map while suppressing the irrelevant information. In doing so, the feature representation of the segmentation model is strengthened and therefore, semantic information is preserved as information hows through the network.

3.2.2 | ResNet block

As more layers are added to a network, gradients may either vanish or explode.[36] This can result in the network not converging during training. To alleviate this problem, residual blocks have been introduced. Residual blocks create a short connection from the input that is added to the output. With this simple trick, the gradients how properly during backprogation and vanishing and exploding gradients are prevented. Altogether, residual units are a combination of two convolution layers, Batch Normalization (BN), ReLu and short connection. Residual Blocks have been used in ResUNet++. A diagram of the ResNet Block is shown in the bottom right corner of Figure 2.

3.2.3 | Residual recurrent block (RR Block)

While Residual blocks typically short the input after two consecutive convolution layers, the RR blocks create a short connection between input and output after every convolution layer. A diagram of the Residual Recurrent Block is shown in the bottom right corner of Figure 2. In this diagram, two recurrent blocks are connected sequentially.

3.2.4 | Squeeze and excite

Formally, the SE Layer is described as follows:

$$\mathbf{X} = w_{se} \times \mathbf{x} + F(\mathbf{x}; \phi_1, \phi_2) \tag{1}$$

F (.) is a residual block parameterized by two convolutional layers ϕ_1 and ϕ_2. $\mathbf{x} \in \mathrm{R}^{C \times H \times W}$ and $\mathbf{X} \in \mathrm{R}^{C \times H' \times W'}$ are the input and output feature maps. w_{se}. are the excitation weights and it is computed as follows:

$$w_{se} = \Upsilon(W_2 \times RELU(W_1 \times GAP(\mathbf{x}))) \tag{2}$$

RELU is the relu activation and GAP is the global average pooling operation. Υ denotes the sigmoid activation. SE module re-weights the features across the channel dimension by applying GAP on the individual channels of the feature map $\mathbf{x}$. GAP reduces the feature map to a scalar. The vector produced by GAP is fed to two consecutive linear layers parameterised by W_1 and W_2. The final sigmoid layer is used to compute the 'excitation' weights. The excitation weights are used to reweight the channels of features as shown in Equation (1).

SE module provides channel attention by encoding the global context. Essentially, GAP reduces the features across the channel dimension to a vector of scalars, which represents the encoding of global context. The vector of scalars are passed through a Fully Connected (FC) network with one hidden layer. The hidden layer, which is of lower dimension than the channel dimension, in conjunction with the sigmoid activation function capture non-linear dependencies that exist across the channel dimension of the feature map. Through this process, features which are more important are scaled higher than features which contribute lesser to the segmentation task. The features are scaled along the channel dimension through the global context encoding. However, SE module uses only one receptive field dependent on the height and width of the feature map to provide channel attention. It was observed that the combination of different receptive fields boost semantic segmentation performance suggesting that both local and global context is beneficial for semantic segmentation.[37,38] Therefore, we argue that capturing only the global context to reweight the feature along the channel dimension is insufficient. We propose using global and local level context at multiple receptive fields to re-weight feature maps. To this end, we propose SMCA that we discuss in the next section.

3.2.5 | Squeeze and multi-context attention

We propose a module that uses global and local context for re-weighting the feature maps. SMCA encodes global context using SE module and encodes local context at multiple receptive fields using Average and Max Pooling operations. Average and Max Pooling of various strides and kernel sizes capture the local context at various receptive fields. They are inexpensive as they do not have any learnable parameters. In our experiments, we use strides of 2, 4 and 8 and kernels of size 2, 4 and 8, respectively to capture the local context at increasing receptive fields. The average and max pooling operations are followed by squeeze operation through 1×1 convolutions and convolution operations through 3×3 kernels that capture relavant channel and spatial information. The outputs of the 'Conv Squeeze Block' and 'Conv Normal Block' (See Figure 3) are added. The channel interdependencies are captured by the 'Conv Squeeze Block' and the relevant spatial information is preserved by the 'Conv Normal Block' thus providing channel and spatial attention respectively. Formally, we can define the SMCA module as follows:

$$\mathbf{X} = w_{smca} \times \mathbf{x} + \mathbf{x} \tag{3}$$

where $\mathbf{x}$ is the input feature map, $\mathbf{X}$ is the output feature map and w_{smca} is the multi-context attention weights used to recalibrate the input map. w_{smca} is computed of three spatial and channel attention weights at different receptive fields as shown as follows:

$$w_1 = F(MP(x,2) + AP(x,2)) + F_{sq}(MP(x,2) + AP(x,2), r) \tag{4}$$

$$w_2 = F(MP(x,4) + AP(x,4)) + F_{sq}(MP(x,4) + AP(x,4), r) \tag{5}$$

$$w_3 = F(MP(x,8) + AP(x,8)) + F_{sq}(MP(x,8) + AP(x,8), r) \tag{6}$$

For the sake of brevity, we remove the parameterization notations for the residual block F (.) and the 'squeeze' block F_{sq} (., r). F_{sq} (., r) is a special

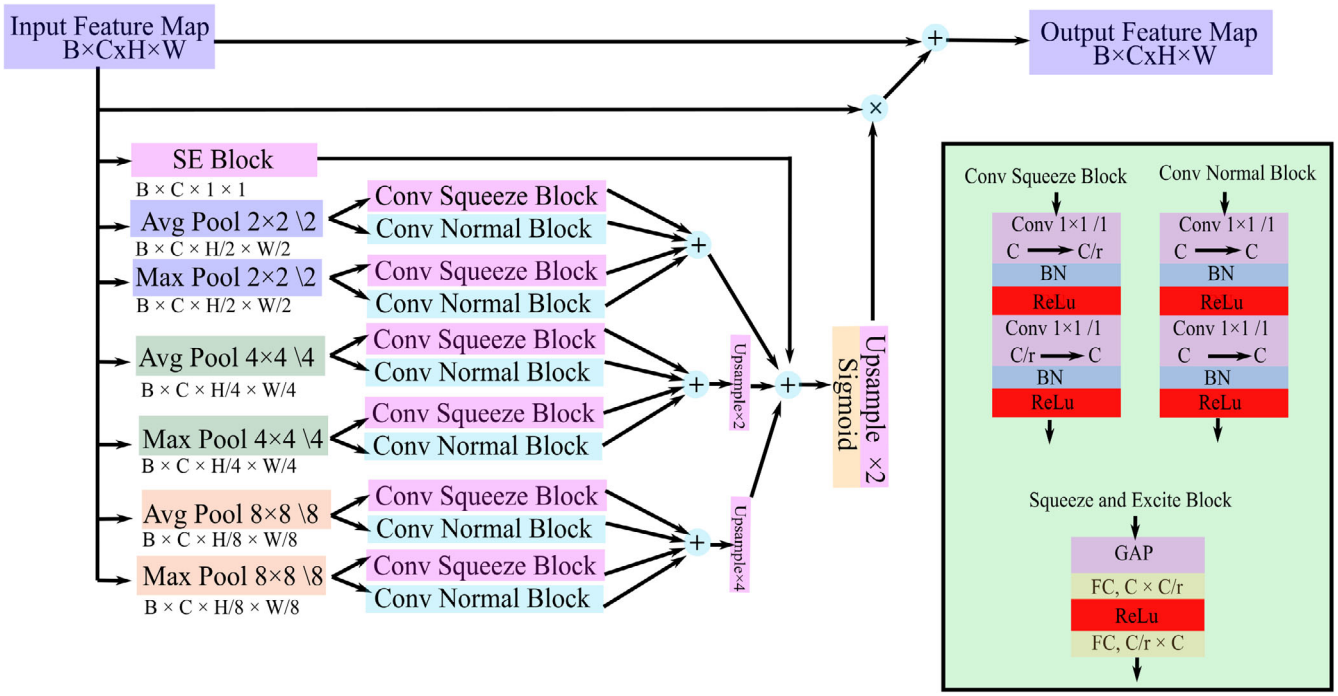

FIGURE 3 Squeeze and Multi-Context Attention module. The resolutions mentioned below the SE module, Avg Pool and Max Pool denote the output resolution after the corresponding operation. Conv Squeeze Block, Conv Normal block and SE module are illustrated inside the green box. BN is abbreviation of Batch Normalization. The input and output nodes (eg. C × C/r) are mentioned in the fully connected (FC) box. r is the channel compression ratio

convolutional block where bottleneck is introduced in the channel dimension by reducing the channel dimension by a factor of r using 1×1 convolution. AP (., n) and MP (., n) represent the Average and Max Pooling operations where n denotes the stride n and kernel size $n \times n$. Finally, the multi-context weights are upsampled bilinearly by the corresponding factor to match the input feature map dimensions. Formally it can be described as follows:

$$w_{smca} = \varphi(\mathrm{Y}(\varphi(w_1, 1) + \varphi(w_2, 2) + \varphi(w_3, 4) + w_{se}), 2) \quad (7)$$

where φ (., k) denotes the upsampling operation by a factor k.

4 | EXPERIMENTS

4.1 | Dataset details

We have used the following datasets for training and evaluating our models (Table 1).

- KVASIR-SEG contains 1000 images annotated by endoscopists from Oslo University. Each image contains atleast one polyp.
- CVC-ColonDB consists of 380 images from 15 colonoscopy videos. Each image shows at least one polyp.
- CVC-ClinicDB consists of 612 images. Each image contains at least one polyp.
- KVASIR-Sessile consists of 196 images of polyps smaller than 10 mm. This dataset is a subset of Kvasir-SEG dataset.
- ETIS-Larib Polyp DB contains 196 images. Each image contains at least one polyp. This dataset is only used as test set for inter-dataset evaluation.

4.2 | Implementation details

For our intra-dataset experiments, from each dataset, 10% of the images were randomly selected to construct the test set. The remaining images in the dataset were split into five portions of equal sizes. Leave-one-fold-out strategy was then used to construct the training and cross-validation sets. We evaluated U-Net, Attention U-Net, R2U-Net, R2AU-Net and ResUNet++ with and without SMCA on Kvasir-SEG, CVC-ColonDB, CVC-ClinicDB and Kvasir-Sessile datasets. In our inter-dataset experiments, we randomly shu ed the Kvasir-SEG and CVC-ClinicDB dataset five times and split it into five pairs of training and cross-validation sets. The training

TABLE 1 Polyp segmentation datasets used in our experiments

Dataset	Train images	Validation images	Test images	Input size
Kvasir-SEG	720	180	100	Variable
CVC-ColonDB	273	69	38	574 × 500
CVC-ClinicDB	440	110	62	384 × 288
Kvasir-Sessile	140	36	20	Variable
ETIS-Larib Polyp DB	-	-	196	1225 × 966

and cross-validation split ratio was 90:10. We trained all the models with and without SMCA on five training sets of Kvasir-SEG and tested them on CVC-ColonDB, CVC-ClinicDB and ETIS-Larib Polyp DB. Similary, we trained the models on five training sets of CVC-ClinicDB and tested them on CVC-ColonDB, Kvasir-SEG and ETIS-Larib Polyp DB. We employed the cross-validation set derived from the training dataset to save the best-performing model. All the architectures were implemented using PyTorch and trained on NVIDIA RTX 3090, 24 GB RAM. All the models were trained for 65 epochs with a learning rate of 1e-4 for the first 50 epochs and 1e-5 for the last 15 epochs. We have used data augmentation such as random rotation, horizontal and vertical hip. Batch size of 8 was used for all the experiments. Each batch of images was scaled to 196 × 196, 256 × 256 and 512 × 512 resolutions for training. All the models were evaluated on 256 × 256 sized images.

4.3 | Loss function

The choice of loss function is particularly crucial in polyp segmentation as we have an imbalance between the number of positive class samples (polyp pixels) and negative class samples (background pixels). If the class imbalance is not considered in the loss function, the model may converge to sub-optimal solution. Additionally, in medical applications, reducing false negative predictions typically takes precedence over reducing false positive prediction. Concretely, segmenting polyp pixels is more important than falsely segmenting non-polyp pixels as polyps. Therefore, there have been several works that tackled the class imbalance problem. Yeung et al.[39] propose a unified asymmetric focal loss that prevents suppression of gradients of classes that occur infrequently. Additionally, Ma et al.[40] perform a thorough analysis of the contribution of 20 loss functions on 4 segmentation tasks. The literature reveals that Tversky loss[41] can weigh the influence of false negative class prediction over false positive class prediction when computing the gradients for model training. Therefore, we use Tversky loss for our experiments. It is an asymmetric similarity measure between predicted segmentation map and ground truth map. It is a generalization of Dice similarity coefficient (DSC) and Jaccard index. The Tversky loss is calculated as the mean of Tversky index (TI). Tversky Index is calculated as follows:

$$TI_i = \frac{TP}{TP + \alpha \times FP + \beta \times FN} \tag{8}$$

where i is the ith pair of predicted and ground truth segmentation map. TP, FN and FP are the true positive, false negative and false positive count. α and β are the weights associated with the false positive and false negative count. $\beta > \alpha$ forces the model to improve the recall more than precision and vice versa. We set $\alpha = 0.4$ and $\beta = 0.6$ based on grid search. We use these values for all our experiments.

Finally, the Tversky loss over the mini-batch of size B can be defined as follows:

$$L_i = 1 - TI_i \tag{9}$$

$$L = \frac{1}{B}\overset{B}{_{i=1}} L_i \tag{10}$$

4.4 | Evaluation metrics

The models are evaluated using DSC, mean intersection over union (mIoU), precision and recall. The metrics are computed as follows:

$$DSC = \frac{2 \times TP}{2 \times TP + FP + TN} \tag{11}$$

$$IoU = \frac{TP}{TP + FP + TN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

where TP or True Positive is the total number of pixels in the predicted segmentation mask classified as polyp pixels and are actually polyp pixels, FP or False Positive is the total number of pixels classified as back-ground pixels but are polyp pixels and TN or True Negative is the total number of pixels, which belong to the background class and are predicted as background pixels.

5 | RESULTS

In this section, we report the findings of our intra-dataset and inter-dataset experiments. First, we report the segmentation metrics of each model separately. To this end, we report each model with and without SMCA and report the performance differences. Next, we report the results of our inter-dataset experiments by taking all the models together.

The qualitative comparison of our intra-dataset experiments is shown in Figure 4. The qualitative comparison of our inter-dataset experiments with training sets Kvasir-SEG and CVC-ClinicDB are shown in Figures 5 and 6, respectively.

5.1 | Evaluation of U-Net

The results of our intra-dataset experiments on U-Net are presented in Table 2. We observe that SMCA improves all the metrics for Kvasir-SEG, CVC-Clinic and Kvasir-Sessile. Notably, the DSC improves by 5.1%, mIoU by 8.8%, precision by 7.8% and recall by 2.3% for CVC-ClinicDB. SMCA brings improvement to Kvasir-Sessile dataset which contains images of polyps (less than 10 mm) that are hard to segment. Specifically, the DSC improves by 2.2%, mIoU by 3%, precision by 9.5% and recall by 9%.

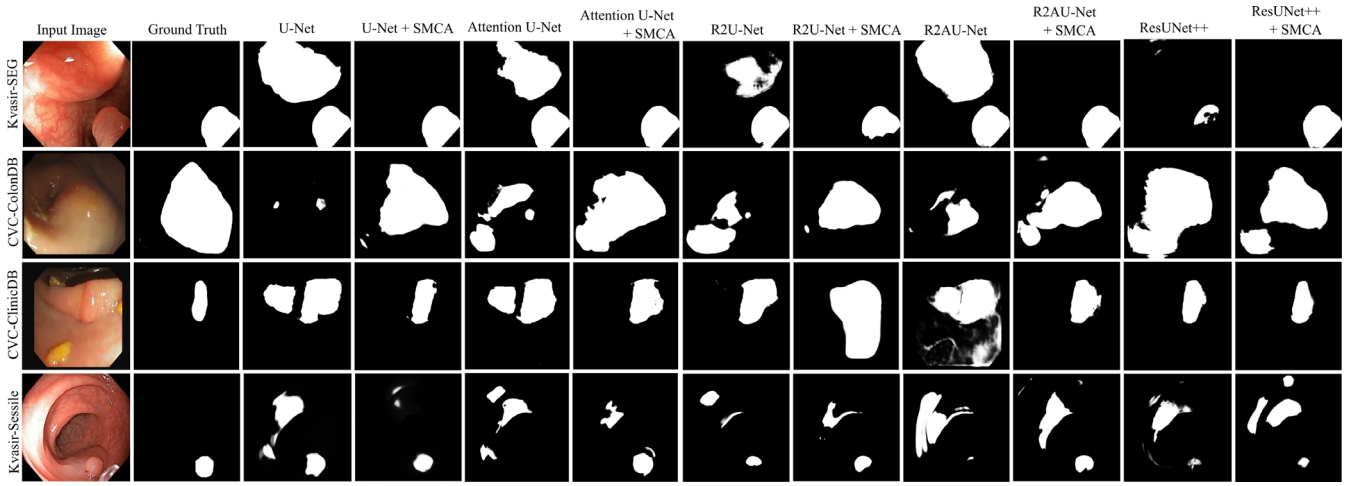

FIGURE 4 Qualitative comparison of models U-Net, Attention U-Net, R2U-Net, R2AU-Net and ResUNet++ with and without SMCA on intra-dataset evaluation. The figure contains sampled images from Kvasir-SEG, CVC-ColonDB, CVC-ClinicDB and Kvasir-Sessile. The figures have problematic artifacts such as high rejection, poor illumination, gastrointestinal remnants and motion blur. Despite the performance degrading artifacts, models with SMCA are still able to predict segmentation maps that are similar to the ground truth masks

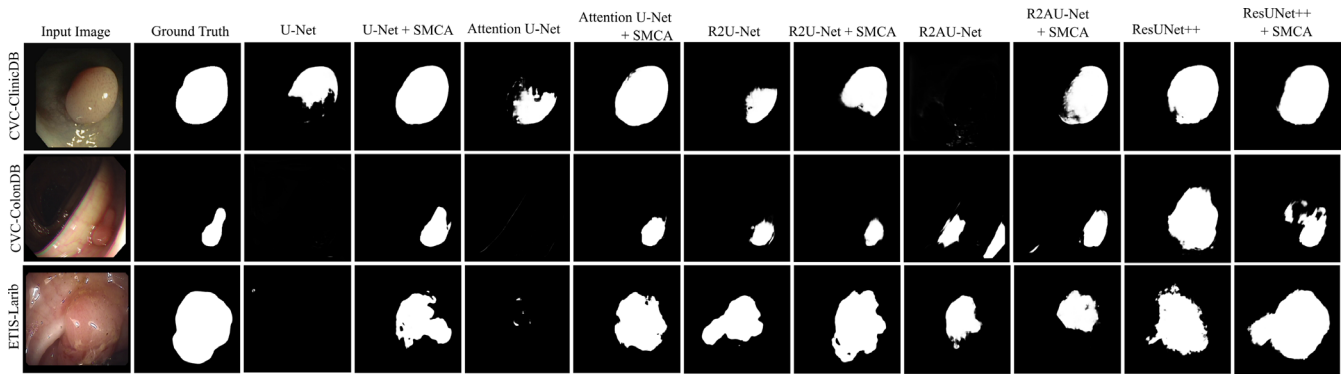

FIGURE 5 Qualitative comparison of models U-Net, Attention U-Net, R2U-Net, R2AU-Net and ResUNet++ with and without SMCA trained on Kvasir-SEG and evaluated on CVC-ClinicDB, CVC-ColonDB and ETIS-Larib Polyp DB. It can be observed that SMCA learns features, which are less sensitive to non-polyp pixels and in general learn more robust features. This helps the model generalize better to new datasets

5.2 | Evaluation of attention U-Net

The results of our intra-dataset experiments on Attention U-Nets are presented in Table 3. We report that SMCA shows improvement of all metrics for all the four datasets. The largest improvement is shown on CVC-ColonDB with increase of 65% for DSC, 100% for mIoU, 86.4% for precision and 5% for recall. Similar to U-Nets, we observe that SMCA improves the performance on Kvasir-Sessile dataset. Another observation is that the Attention U-Net (with and without SMCA) performs worse relative to U-Net on Kvasir-SEG, CVC-ColonDB, CVC-ClinicDB and Kvasir-Sessile.

5.3 | Evaluation of R2U-Net

Table 4 shows the results of our intra-dataset evaluation on R2U-Net. The results show a general improvement trend for all four datasets. The model trained on Kvasir-SEG and CVC-ColonDB show notable improvements due to SMCA. We find that the DSC, mIoU, precision and recall improve by 23%, 35.8%, 30.7% and 6.4% on Kvasir-SEG. The performance improvements on Kvasir-Sessile are negligible in comparison to the other three datasets.

5.4 | Evaluation of R2AU-Net

Table 5 shows the results of our intra-dataset evaluation of R2AU-Net. We report that SMCA shows a general improvement in segmentation metrics on all the datasets. Similar to R2U-Net, the model trained on Kvasir-SEG and CVC-ColonDB shows notable improvements due to SMCA. We find that the DSC, mIoU, precision and recall improve by 17.3%, 25.8%, 27.9% and 6.4% on Kvasir-SEG. We also observe that the recall of R2AU-Net with SMCA is almost at par with R2AU-Net without SMCA. Furthermore, the performance improvements on Kvasir-Sessile are negligible in comparison to the other three datasets.

5.5 | Inter-dataset evaluation

In this section, we report the results of our inter-dataset experiments. The purpose of the inter-dataset evaluation is

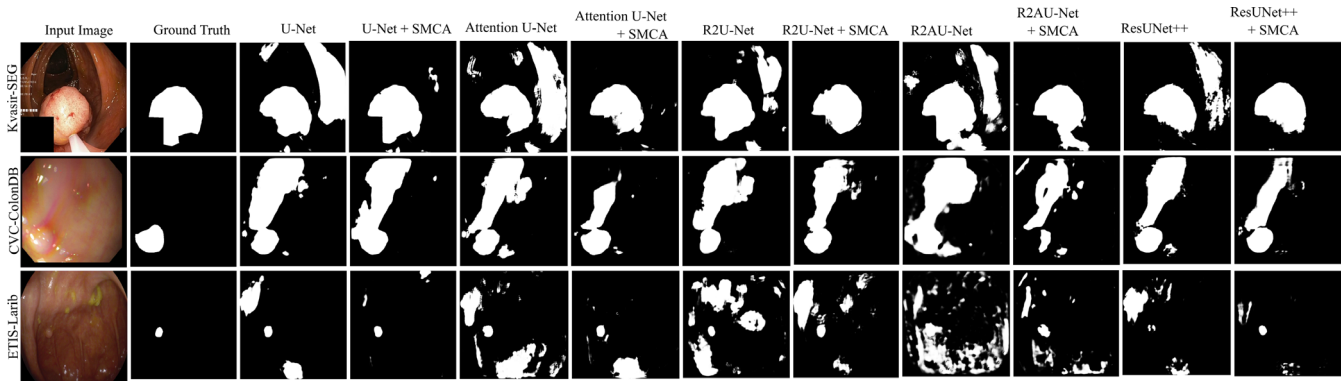

FIGURE 6 Qualitative comparison of models U-Net, Attention U-Net, R2U-Net, R2AU-Net and ResUNet++ with and without SMCA trained on CVC-ClinicDB and evaluated on Kvasir-SEG, CVC-ColonDB and ETIS-Larib Polyp DB. In general, the segmentation maps of models with SMCA are more similar to the ground truth mask

TABLE 2 Evaluation of U-Net with and without SMCA on four public datasets

Dataset	SMCA	DSC	mIoU	Precision	Recall
Kvasir-SEG		0.83 ± 0.02	0.75 ± 0.02	0.83 ± 0.01	0.88 ± 0.02
Kvasir-SEG	✓	**0.85 ± 0.01**	**0.78 ± 0.01**	**0.86 ± 0.01**	**0.89 ± 0.01**
CVC-ColonDB		**0.75 ± 0.01**	**0.65 ± 0.01**	**0.78 ± 0.02**	**0.78 ± 0.01**
CVC-ColonDB	✓	0.72 ± 0.04	0.62 ± 0.06	0.72 ± 0.08	0.78 ± 0.01
CVC-ClinicDB		0.78 ± 0.02	0.68 ± 0.02	0.76 ± 0.02	0.86 ± 0.02
CVC-ClinicDB	✓	**0.82 ± 0.02**	**0.74 ± 0.02**	**0.82 ± 0.02**	**0.88 ± 0.01**
Kvasir-Sessile		0.31 ± 0.03	0.20 ± 0.02	0.42 ± 0.04	0.48 ± 0.04
Kvasir-Sessile	✓	**0.38 ± 0.04**	**0.26 ± 0.03**	**0.46 ± 0.04**	**0.57 ± 0.08**

Note: Bold value represent the highest value achieved.

TABLE 3 Evaluation of attention U-Net with and without SMCA on four public datasets

Dataset	SMCA	DSC	mIoU	Precision	Recall
Kvasir-SEG		0.82 ± 0.01	0.74 ± 0.01	0.82 ± 0.01	0.88 ± 0.002
Kvasir-SEG	✓	**0.84 ± 0.01**	**0.76 ± 0.01**	**0.85 ± 0.01**	**0.88 ± 0.001**
CVC-ColonDB		0.41 ± 0.09	0.28 ± 0.07	0.37 ± 0.07	0.69 ± 0.07
CVC-ColonDB	✓	**0.68 ± 0.07**	**0.56 ± 0.08**	**0.69 ± 0.08**	**0.73 ± 0.04**
CVC-ClinicDB		0.72 ± 0.03	0.61 ± 0.03	0.70 ± 0.03	0.83 ± 0.02
CVC-ClinicDB	✓	**0.77 ± 0.02**	**0.68 ± 0.03**	**0.77 ± 0.03**	**0.84 ± 0.01**
Kvasir-Sessile		0.29 ± 0.04	0.18 ± 0.03	0.34 ± 0.04	0.51 ± 0.06
Kvasir-Sessile	✓	**0.35 ± 0.08**	**0.24 ± 0.08**	**0.43 ± 0.1**	**0.51 ± 0.03**

Note: Bold value represent the highest value achieved.

TABLE 4 Evaluation of R2U-Net with and without SMCA on four public datasets

Dataset	SMCA	DSC	mIoU	Precision	Recall
Kvasir-SEG		0.65 ± 0.11	0.53 ± 0.12	0.65 ± 0.14	0.77 ± 0.04
Kvasir-SEG	✓	**0.80 ± 0.01**	**0.72 ± 0.01**	**0.85 ± 0.03**	**0.82 ± 0.01**
CVC-ColonDB		0.30 ± 0.69	0.20 ± 0.05	0.26 ± 0.06	0.68 ± 0.06
CVC-ColonDB	✓	**0.57 ± 0.09**	**0.44 ± 0.09**	**0.53 ± 0.11**	**0.71 ± 0.04**
CVC-ClinicDB		**0.65 ± 0.06**	0.53 ± 0.07	**0.65 ± 0.09**	**0.77 ± 0.02**
CVC-ClinicDB	✓	0.65 ± 0.14	**0.54 ± 0.15**	0.64 ± 0.15	0.77 ± 0.08
Kvasir-Sessile		**0.27 ± 0.05**	0.17 ± 0.05	0.39 ± 0.07	**0.40 ± 0.06**
Kvasir-Sessile	✓	0.26 ± 0.07	**0.18 ± 0.04**	**0.46 ± 0.04**	0.34 ± 0.12

Note: Bold value represent the highest value achieved.

TABLE 5 Evaluation of R2AU-Net with and without SMCA on four public datasets

Dataset	SMCA	DSC	mIoU	Precision	Recall
Kvasir-SEG		0.69 ± 0.07	0.58 ± 0.08	0.68 ± 0.11	**0.83 ± 0.03**
Kvasir-SEG	✓	**0.81 ± 0.02**	**0.73 ± 0.03**	**0.87 ± 0.03**	0.81 ± 0.05
CVC-ColonDB		0.36 ± 0.11	0.25 ± 0.08	0.31 ± 0.11	0.73 ± 0.01
CVC-ColonDB	✓	**0.54 ± 0.11**	**0.41 ± 0.11**	**0.48 ± 0.12**	**0.75 ± 0.04**
CVC-ClinicDB		0.61 ± 0.03	0.48 ± 0.04	0.56 ± 0.04	**0.82 ± 0.01**
CVC-ClinicDB	✓	**0.69 ± 0.06**	**0.58 ± 0.07**	**0.67 ± 0.08**	0.81 ± 0.03
Kvasir-Sessile		0.23 ± 0.05	0.15 ± 0.03	0.37 ± 0.03	**0.34 ± 0.08**
Kvasir-Sessile	✓	**0.24 ± 0.09**	**0.16 ± 0.06**	**0.42 ± 0.05**	0.33 ± 0.15

Note: Bold value represent the highest value achieved.

to further test the generalizability of models with our SMCA module when the test set is not derived from the same dataset. We use images of Kvasir-SEG and CVC-ClinicDB to construct our training sets similar to Jha et al.[5] The images in all these datasets are recorded with different imaging apparatus, have imaging artifacts such as illumination changes, motion blurring, gastrointestinal artifacts, and so forth. Furthermore, the shape and appearance of the polyps vary from dataset to dataset. Therefore, it is expected that there will be a drop in segmentation performance.

5.6 | Evaluation of ResUNet++

Table 6 presents the findings of our intra-dataset evaluation on ResUNet++. We note an overall performance improvement on all the datasets with notable improvements shown on Kvasir-SEG dataset. The DSC, mIoU, precision and recall improve by 8.2%, 11.3%, 2.7% and 8.8%. The performance improvements on Kvasir-Sessile are not significant in comparison to the improvements on the other three datasets.

Dataset	SMCA	DSC	mIoU	Precision	Recall
Kvasir-SEG		0.72 ± 0.02	0.62 ± 0.02	0.75 ± 0.03	0.79 ± 0.003
Kvasir-SEG	✓	**0.78 ± 0.01**	**0.69 ± 0.02**	**0.77 ± 0.02**	**0.86 ± 0.02**
CVC-ColonDB		0.46 ± 0.03	0.33 ± 0.03	0.49 ± 0.03	0.58 ± 0.1
CVC-ColonDB	✓	**0.52 ± 0.07**	**0.39 ± 0.07**	**0.50 ± 0.11**	**0.67 ± 0.04**
CVC-ClinicDB		0.69 ± 0.05	0.58 ± 0.06	0.68 ± 0.05	0.79 ± 0.03
CVC-ClinicDB	✓	**0.73 ± 0.02**	**0.62 ± 0.02**	**0.70 ± 0.03**	**0.82 ± 0.01**
Kvasir-Sessile		0.30 ± 0.03	0.20 ± 0.03	0.38 ± 0.05	0.49 ± 0.04
Kvasir-Sessile	✓	**0.31 ± 0.02**	**0.21 ± 0.02**	**0.39 ± 0.02**	**0.50 ± 0.02**

Note: Bold value represent the highest value achieved.

TABLE 6 Evaluation of ResUNet++ with and without SMCA on four public datasets

5.6.1 | Kvasir-SEG as training set

Table 7 shows the results of our experiments using Kvasir-SEG as training set. Despite the changes in image distribution across the datasets, it can be seen that SMCA improves the segmentation performance of most of the models. For example, U-Net with SMCA trained on Kvasir-SEG shows 29.7%, 30.7%, 12.7%, 27.7% improvements on DSC, mIoU, precision and recall respectively when tested on CVC-ColonDB. Similarly, ResUNet++ with SMCA trained on Kvasir-SEG shows 25.5%, 29%, 27.9% and 9.6% on DSC, mIoU, precision and recall over the baseline ResUNet++ when tested on CVC-ColonDB.

5.6.2 | CVC-ClinicDB as training set

Table 8 shows the inter-dataset evaluation of the five models with and without SMCA trained on CVC-ClinicDB. Altogether, we report improvements in most of our models. For example, U-Net with SMCA shows 35%, 50%, 41% and 3% increase in DSC, mIoU, precision and recall when tested on CVC-ColonDB compared to baseline U-Net. An observation that can be drawn is that the inter-dataset performance of models trained on Kvasir-SEG is better than models trained on CVC-ClinicDB. We believe this is the case because the images of Kvasir-SEG have higher contrast than CVC-Clinic and also, the polyps in Kvasir-SEG are more diverse in size, shape, color and appearance. We conjecture that these attributes of the training set play a role.

5.7 | Choice of channel compression ratio

Choosing the correct channel compression ratio is important as it is mainly responsible for re-weighting the information across the channel dimension. Therefore, we performed experiments to find the ideal channel compression ratio r for our SMCA module. We chose U-Net as our baseline architecture and used Kvasir-SEG dataset to perform a five-fold cross validation experiment. Observing the results in Table 9, we chose the channel compression ratio 2 for all our intra-dataset and inter-dataset experiments.

6 | DISCUSSION

6.1 | Summary of results

Looking at the quantitative results of intra-dataset experiments (See Tables 2–6), we can draw the following observations: (i) SMCA improves the performance when incorporated into five popular segmentation models; (ii) SMCA has a greater impact on larger models than on smaller models (see Tables 4 and 5 vs. Table 2); (iii) On an average, all the models perform the best on Kvasir-SEG, followed by CVC-ClinicDB, CVC-ColonDB and Kvasir-Sessile; (iv) SMCA when incorporated into ResUNet++ performs better than the baseline. Our results indicate that the SMCA is a better attention module compared to SE module. When observing the results of the inter-dataset experiment, we can draw the following observations: (i) Models with SMCA perform better than models without SMCA; (ii) Models generalize better when trained on Kvasir-SEG than on CVC-Clinic; (iii) Models with fewer trainable parameters perform better than models with more parameters.

6.2 | Discussion on intra-dataset evaluation

From the intra-dataset experiments, we conclude that models with SMCA show improvements in segmentation metrics. This demonstrates that our module is versatile and can act as a plug-in module to various deep learning architectures. We see that lightweight models such as the

TABLE 7 Inter dataset evaluation of models trained on Kvasir-SEG

		Baseline				Ours			
Test set	**Model**	**DSC**	**mIoU**	**Precision**	**Recall**	**DSC**	**mIoU**	**Precision**	**Recall**
CVC-ColonDB	UNet	0.47 ±0.09	0.39 ± 0.08	0.55 ± 0.09	0.54 ± 0.07	**0.61 ± 0.02**	**0.51 ± 0.01**	**0.62 ± 0.01**	**0.69 ± 0.04**
	Attention UNet	0.55 ± 0.03	0.46 ± 0.03	0.61 ± 0.02	0.61 ± 0.03	**0.59 ± 0.02**	**0.51 ± 0.02**	**0.61 ± 0.01**	**0.67 ± 0.04**
	R2UNet	0.33 ± 0.09	0.27 ± 0.08	0.42 ± 0.08	0.40 ± 0.13	**0.49 ± 0.09**	**0.41 ± 0.08**	**0.58 ± 0.05**	**0.52 ± 0.09**
	R2AUNet	0.39 ± 0.03	0.31 ± 0.03	0.46 ± 0.07	0.48 ± 0.03	**0.42 ± 0.06**	**0.35 ± 0.05**	**0.53 ± 0.04**	**0.45 ± 0.08**
	ResUNet++	0.39 ± 0.04	0.31 ± 0.04	0.43 ± 0.05	0.52 ± 0.04	**0.49 ± 0.03**	**0.40 ± 0.03**	**0.55 ± 0.06**	**0.57 ± 0.02**
CVC-ClinicDB	UNet	0.62 ± 0.08	0.53 ± 0.08	0.71 ± 0.07	0.67 ± 0.07	**0.75 ± 0.02**	**0.65 ± 0.02**	**0.78 ± 0.03**	**0.80 ± 0.03**
	Attention UNet	0.69 ± 0.02	0.60 ± 0.02	0.76 ± 0.01	0.73 ± 0.03	**0.72 ± 0.02**	**0.63 ± 0.02**	**0.76 ± 0.01**	**0.77 ± 0.03**
	R2UNet	0.46 ± 0.11	0.38 ± 0.10	0.56 ± 0.13	0.54 ± 0.12	**0.67 ± 0.05**	**0.58 ± 0.05**	**0.76 ± 0.03**	**0.69 ± 0.07**
	R2AUNet	0.54 ± 0.02	0.45 ± 0.02	0.65 ± 0.04	0.61 ± 0.03	**0.64 ± 0.04**	**0.55 ± 0.04**	**0.75 ± 0.02**	**0.66 ± 0.06**
	ResUNet++	0.57 ± 0.03	0.47 ± 0.03	0.62 ± 0.04	0.70 ± 0.02	**0.61 ± 0.02**	**0.52 ± 0.03**	**0.66 ± 0.05**	**0.72 ± 0.05**
ETIS-LaribPolpyDB	UNet	0.38 ± 0.08	0.31 ± 0.08	0.40 ± 0.09	0.50 ± 0.09	**0.47 ± 0.03**	**0.40 ± 0.02**	**0.45 ± 0.03**	**0.60 ± 0.04**
	Attention UNet	0.41 ± 0.01	0.35 ± 0.01	0.42 ± 0.02	0.50 ± 0.03	**0.46 ± 0.07**	**0.39 ± 0.006**	**0.45 ± 0.01**	**0.59 ± 0.03**
	R2UNet	0.27 ± 0.07	0.21 ± 0.06	0.31 ± 0.07	0.40 ± 0.01	**0.39 ± 0.04**	**0.33 ± 0.04**	**0.41 ± 0.05**	**0.56 ± 0.05**
	R2AUNet	0.30 ± 0.03	0.24 ± 0.03	0.33 ± 0.06	0.43 ± 0.09	**0.35 ± 0.05**	**0.29 ± 0.05**	**0.38 ± 0.06**	**0.43 ± 0.05**
	ResUNet++	0.27 ± 0.01	0.22 ± 0.01	0.28 ± 0.01	0.40 ± 0.05	**0.36 ± 0.03**	**0.30 ± 0.02**	**0.37 ± 0.03**	**0.50 ± 0.01**

Note: Bold value represent the highest value achieved.

TABLE 8 Inter Dataset Evaluation of models trained on CVC-ClinicDB

		Baseline				Ours			
Test Set	**Model**	**DSC**	**mIoU**	**Precision**	**Recall**	**DSC**	**mIoU**	**Precision**	**Recall**
CVC-ColonDB	UNet	0.40 ± 0.08	0.30 ± 0.08	0.39 ± 0.09	0.66 ± 0.04	**0.54 ± 0.03**	**0.45 ± 0.03**	**0.55 ± 0.04**	**0.68 ± 0.02**
	Attention UNet	0.42 ± 0.03	0.32 ± 0.03	0.41 ± 0.02	**0.67 ± 0.03**	**0.48 ± 0.04**	**0.38 ± 0.04**	**0.49 ± 0.04**	0.62 ± 0.05
	R2UNet	**0.38 ± 0.04**	**0.29 ± 0.04**	**0.39 ± 0.06**	**0.61 ± 0.03**	0.35 ± 0.08	0.26 ± 0.08	0.35 ± 0.08	0.55 ± 0.09
	R2AUNet	0.32 ± 0.02	0.22 ± 0.01	0.28 ± 0.01	**0.68 ± 0.02**	**0.37 ± 0.07**	**0.28 ± 0.07**	**0.38 ± 0.08**	0.54 ± 0.04
	ResUNet++	0.27 ± 0.03	0.19 ± 0.02	0.30 ± 0.03	0.44 ± 0.04	**0.33 ± 0.02**	**0.24 ± 0.02**	**0.35 ± 0.03**	**0.52 ± 0.08**
Kvasir-SEG	UNet	0.43 ± 0.04	0.31 ± 0.03	0.36 ± 0.05	**0.85 ± 0.05**	**0.58 ± 0.05**	**0.47 ± 0.05**	**0.57 ± 0.08**	0.78 ± 0.06
	Attention UNet	0.41 ± 0.02	0.29 ± 0.01	0.32 ± 0.02	**0.88 ± 0.03**	**0.57 ± 0.03**	**0.45 ± 0.03**	**0.61 ± 0.03**	0.67 ± 0.03
	R2UNet	0.45 ± 0.03	0.32 ± 0.03	0.39 ± 0.05	**0.78 ± 0.05**	**0.48 ± 0.04**	**0.36 ± 0.05**	**0.46 ± 0.04**	0.71 ± 0.09
	R2AUNet	0.40 ± 0.01	0.27 ± 0.01	0.31 ± 0.02	**0.87 ± 0.02**	**0.47 ± 0.03**	**0.35 ± 0.04**	**0.49 ± 0.09**	0.65 ± 0.05
	ResUNet++	**0.46 ± 0.02**	**0.34 ± 0.02**	**0.44 ± 0.05**	0.71 ± 0.05	0.45 ± 0.03	0.33 ± 0.02	0.40 ± 0.04	**0.77 ± 0.03**
ETIS-LaribPolpyDB	UNet	0.15 ± 0.03	0.09 ± 0.02	0.11 ± 0.02	**0.76 ± 0.12**	**0.21 ± 0.03**	**0.15 ± 0.03**	**0.20 ± 0.06**	0.68 ± 0.24
	Attention UNet	0.13 ± 0.01	0.08 ± 0.01	0.09 ± 0.01	**0.83 ± 0.10**	**0.20 ± 0.01**	**0.14 ± 0.01**	**0.18 ± 0.01**	0.64 ± 0.12
	R2UNet	0.15 ± 0.02	0.10 ± 0.01	0.11 ± 0.02	**0.64 ± 0.14**	**0.17 ± 0.01**	**0.12 ± 0.01**	**0.14 ± 0.01**	0.57 ± 0.16
	R2AUNet	0.12 ± 0.01	0.07 ± 0.01	0.08 ± 0.01	**0.82 ± 0.05**	**0.16 ± 0.03**	**0.11 ± 0.02**	**0.15 ± 0.03**	0.51 ± 0.16
	ResUNet++	0.15 ± 0.02	0.10 ± 0.02	0.12 ± 0.03	**0.73 ± 0.12**	**0.16 ± 0.02**	**0.11 ± 0.01**	**0.13 ± 0.03**	0.67 ± 0.12

Note: Bold value represent the highest value achieved.

TABLE 9 Choice of channel compression ratio

Channel compression ratio r	DSC	mIoU	Precision	Recall
2	0.856 ± 0.01	0.78 ± 0.01	0.86 ± 0.01	0.897 ± 0.01
4	0.81 ± 0.03	0.73 ± 0.04	0.81 ± 0.04	0.87 ± 0.03
8	0.851 ± 0.01	0.77 ± 0.01	0.85 ± 0.02	0.893 ± 0.02

Note: Bold value represent the highest value achieved.

U-Net perform better in all the datasets compared to models with more parameters (ResUNet++, Attention U-Net, R2U-Net and R2AU-Net). We believe this to be the case because our training dataset is small due to the five-fold cross validation experiments. As such, chances of overfitting larger and deeper models are higher than shallow models[42] when training on small datasets. The authors of ResUNet++[15] use augmentation schemes such as center crop, random crop, horizontal hip, vertical hip, scale augmentation, random rotation, cutout, brightness augmentation, and so forth. In our case, we use only random vertical and horizontal hip. Thus, we argue that using more augmentation methods will improve the performance of the larger models. Additionally, we observe that the boost in metrics due to SMCA to larger models is greater than the boost in metrics on U-Nets (See Table 4 vs. Table 2). We conjecture that the SMCA is able to counter the overfitting tendency by introducing a regularizing effect. The regularising effect is more prominent in larger models and thereof, the improvement in segmentation performance due to SMCA is more in larger models than smaller models.

6.3 | Discussion on inter-dataset evaluation

The inter-dataset evaluation of models is an important and necessary technique to test generalizing capabilities of the models. Our work builds on the cross-dataset experiments of Jha et al.[5] We believe that inter-dataset evaluation of models is important if we want to realise AI-CADx in clinical settings. Deep learning models perform poorly when the test set and training set have diverging image distributions. We think training and test set image distribution divergence will be a common problem faced in the polyp segmentation domain primarily because the images are recorded under different conditions (e.g., with different recording devices, different light scources, etc.). Furthermore, as mentioned previously, the polyps appear in various sizes, shapes and appearances. Additionally, the experience of the physician who is doing the colonoscopy will also affect the quality of the images. As such, performing inter-dataset evaluation should become a standard criteria to demon. strate the generalizing capabilities of AI-CADx for polyp segmentation. Our work is a step forward in this direction. In our work, we perform inter-dataset evaluation to demonstrate the improvements in generalizing capability of baseline models due to SMCA module.

From Tables 7 and 8 we see that SMCA improves the generalizing capabilities of five segmentation models. Our results indicate that the feature recalibration of SMCA is beneficial towards learning robust features for polyp segmentation. We believe that SMCA learns robust features for multiple reasons. First, the use of max and average pooling with different kernel sizes allow the model to extract highly activated features and average activation of features simultaneously over varying receptive fields. The highly activated features are primarily due to polyp pixels and relevant background information necessary for polyp segmentation. Thus, the max pooling passes the highest activated polyp and background features and the average pooling passes an average of highly activated and lowly activated features which can be considered to be the context information of the polyp. Max and average pool with large kernels help to pass large context information around the polyps. Similarly, max and average pool with small kernels help to pass small context information around the polyp. We can draw an analogy for the working of SMCA to a physician inspecting the area around a suspected polyp lesion to demarcate a polyp mass from a non-polyp surrounding tissue. The use of receptive fields of different sizes in SMCA is analogous to a physician inspecting a large and small area around a suspected polyp lesion. A large area of inspection provides more context of a polyp's position in relation to the colorectal surface whereas a small area of inspection provides the necessary detail to distinguish a polyp lesion from the non-polyp colorectal surface. Second, the resulting feature maps from the multiple average and max pooling operations are passed through "Conv Squeeze Block" and "Conv Normal Block". These blocks are used to provide channel and spatial attention, respectively. Third, the attention weights of the "Conv Squeeze Block" and "Conv Normal Block" computed from multiple receptive fields are added together to compute the final attention weights. This, in effect, allows small and large-sized polyp features and its corresponding contexts to contribute towards recalibration of the original feature map.

We also observe that models trained on Kvasir-SEG show better inter-dataset performance than models

trained on CVC-ClinicDB. We conjecture that the higher contrast and larger variation (in terms of size, shape and appearance) of polyp images on Kvasir-SEG compared to the CVC-ClinicDB enabled the model to learn better generalizing features. Despite the differences in the training dataset, baseline models with SMCA show improvement in segmentation metrics suggesting the robust representations learned due to SMCA.

6.4 | Visualizing the effectiveness of our SMCA module

The ultimate objective of proposing AI-CADx in polyp segmentation is to improve clinical decision-making by using human intelligence capabilities in conjunction with AI-CADx. Further, with the latest advancements in human-in-the-loop annotation tools, generating annotated datasets have become easier.[43] Therefore, the combined capabilities of AI and human intelligence may lead to efficient AI workhows and clinical workhows. While generating annotated datasets have become more efficient, AI models are still considered a black box.[44] As argued by Rundo et al.,[45] one of the many challenges in installing AI-CADx in clinical practice is the lack of interpretability and explainability. Therefore it is of utmost importance to develop AI-CADx that are interpretable. Then these systems can gather trust amongst physicians and patients alike. Interpretability is mostly ignored in many works dealing with polyp segmentation. We believe that the risks posed by AI-CADx deployed in healthcare industry are far greater than in other industries. The risk of a model making a false prediction can have life-threatening consequences. Therefore, it is of utmost importance to understand the decision-making process of machine learning models. This will help in understanding the pitfalls of deep learning models and help in finding techniques to redress them. We believe this will enable research in design and development of network architectures that are more reliable and have better-generalizing capabilities. Our work is a step forward in this direction.

Visualising the feature representation of a model with and without SMCA can offer better insight than simply reporting the segmentation metrics. To this end, we use M3D-CAM's[46] implementation of Grad Cam++ to visualise the gradient weighted attention maps of the two convolution kernels at each "Conv Block" (See Figure 2) in the encoder and decoder of U-Net. In Figure 7, we present the side-by-side comparison of the attention maps from the "Conv Blocks" of baseline U-Net and U-Net with SMCA. The qualitative comparison of attention maps shows the difference in learned representation of both the networks. We visualise the attention maps of the convolution kernels in the "Conv Block" because each "Conv Block" from the second layer onwards receives the re-calibrated feature maps of the SMCA. Furthermore, "Conv Block" is present in both the baseline U-Net and U-Net with SMCA. Therefore, it serves as a good entity to make a fair comparison.

One of the many challenges in segmentation is effectively retaining important semantic information along with high-level concepts as information propagates through the network. Cascade of max pool operations in CNNs result in learning of high-level concepts at the expense of granular information such as edge and color being lost. However, preserving the important low-level features alongside the high-level concepts can improve the precision and accuracy of segmentation maps.[47] Our qualitative analysis indicates that SMCA enables the CNN to preserve low-level features relevant for semantic segmentation in the deeper layers. The re-calibration of the encoder features using the max and average pooling operations of varying kernels help in the extraction of relevant polyp and context features at multiple scales. Additionally, computing spatial and channel attention weights from these extracted features helps in preserving important low-level semantic features while allowing the formation of high-level concepts. Our results indicate that this allows the models with SMCA to make more precise and accurate segmentation maps. Looking at the activation map of "Conv Block" at Conv Layer 3 of Figure 7 for U-Net with SMCA, we observe that the convolution kernels that receive re-calibrated feature maps activate low-level semantic concepts occurring throughout the image. This indicates that SMCA re-calibrates the feature map to preserve important low-level semantic concepts as information propagates through the network. In comparison, the attention maps of the baseline U-Net in Conv Layer 3 shows limited activation implying loss of relevant semantic information in the deeper layers. Similarly, activation maps at Conv Layer 4 of U-Net with SMCA show more activity than activation maps of baseline U-Net. Observing the activation maps at Conv Layer 5, we see that both U-Net and U-Net with SMCA learn high-level concepts. However, U-Net with SMCA does it while preserving the low-level semantic information in the preceeding layers. On the decoder side (See Conv Layer 6, Conv Layer 7, Conv Layer 8, Conv Layer 9), we see that the activation maps are more prominent for U-Net with SMCA and they start resembling the final segmentation map from Conv Layer 7 onwards. We conjecture that the closer resemblance of the activation maps to the predicted segmentation map for U-Net with SMCA is because the decoder is able to use the preserved low-level semantic features passed to it through skip

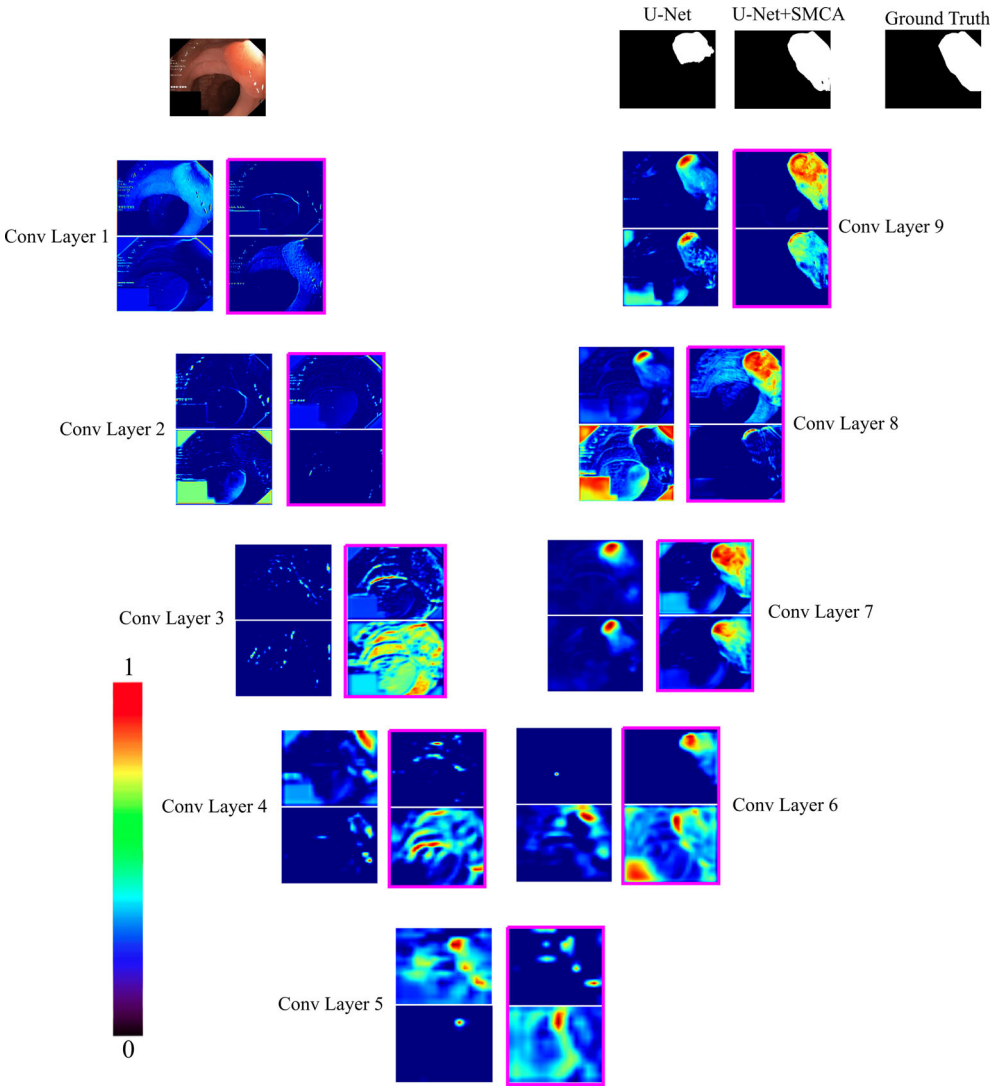

FIGURE 7 Visualization of the attention maps at each convolutional kernel in the bottlenecks. There are two attention maps at each layer because there are two convolution operations at each "Conv Block" (See Figure 2). The attention maps with the pink borders are from the U-Net with SMCA, the attention maps without borders are from U-Net without SMCA

connections. This, in effect, leads to the prediction of more accurate segmentation maps.

6.5 | Limitations

The models do not generalize well to unseen image distributions. All models perform better when test sets and the training set are from the same dataset. Although our module redresses this problem to an extent, there is more progress to be made in generalization. Self-supervision is an emerging area of research that makes models generalize better to unseen distributions.[48] We think there are significant advantages of this learning paradigm which can expedite the implementation of AI-CADx in clinical settings. Furthermore, our work is a retrospective study which is very different from prospective clinical application. The images in the datasets are selected by expert gastroenterologist. Prospective clinical use-case would involve testing the models on colonoscopy videos. Furthermore, our training set consists of polyps being present in every image. Our models are not trained to consider endoscopic images having any polyps.

7 | CONCLUSION

In this paper, we present a novel module called SMCA. We incorporated SMCA to five segmentation models: U-Net, Attention U-Net, R2U-Net, R2AU-Net and ResUNet++. We extensively evaluated the performance of the mentioned models with and without SMCA on four public polyp segmentation datasets (Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, Kvasir-Sessile). We report that models with SMCA perform better than baseline models. To further test the generalizing ability, we perform rigorous inter-dataset experiments. In the first inter-dataset experiment, we train all the models with and without SMCA on Kvasir-SEG and test it on CVC-ColonDB, CVC-ClinicDB and ETIS-Larib Polyp DB. In

the second experiment, we train all the models on CVC-ClinicDB and test them on CVC-ColonDB, Kvasir-SEG and ETIS-Larib Polyp DB. Finally, to better understand the impact of SMCA on features learned by models, we render the attention maps from the convolution kernels of U-Net with and without SMCA using Grad-CAM++.

The qualitative comparison further illustrates that models with SMCA learn features that preserve important semantic cues throughout the depth of the network. This partially suggests why the models with SMCA predict more accurate segmentation maps.

In summary, SMCA recalibrates the feature maps through simultaneous spatial and channel attention. The spatial and channel attention weights are computed through the extraction of relevant edge and context features at multiple scales. Our results suggest that models with SMCA can segment large and small polyps better than their baseline counterparts. Additionally, we report that SMCA-based models generalize better. This is demonstrated through our extensive intra-dataset and inter-dataset experiments. We think that SMCA will improve the segmentation performance for tasks where the objects to segment appear in different sizes such as brain tumor segmentation[49] where tumors appear in multiple sizes. In non-medical application, SMCA may be beneficial in segmenting regions of interest in urban scenes such as cars, traffic lights and pedestrians[50] which too appear in different sizes. As future work, we want to incorporate SMCA into the decoder network and analyse the changes in the performance. Additionally, we want to analyse the performance changes by pre-training the SMCA-based model through self-supervision.

ACKNOWLEDGMENT

The authors have no conflicts of interests to report. This work has not been submitted for publication anywhere else. This work is funded partially by Hamburg University of Technology (TUHH) and University Hospital Hamburg-Eppendorf (UKE). The authors also acknowledge the partial funding by the Free and Hanseatic City of Hamburg (Interdisciplinary Graduate School "Innovative Technologies in Cancer Diagnostics and Therapy"). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors also acknowledge support for the Open Access fees by TUHH in the funding programme Open Access Publishing. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

This work is funded by the Free and Hanseatic City of Hamburg (Interdisciplinary Graduate School "Innovative Technologies in Cancer Diagnostics and Therapy"). The funder had no role in study design, data collection and analysis, decision to publish, or preparation.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available. We have used five public polyp segmentation datasets in this study and these datasets can be accessed from their corresponding references.

ORCID

Debayan Bhattacharya https://orcid.org/0000-0001-8552-2227

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics. *CA Cancer J Clin*. 2020;70(1):7-30. doi:10.3322/caac.21590
2. Vázquez D, Bernal J, Sánchez FJ, Fernández-Esparrach G, López AM, Romero A, et al. A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images. https://arxiv.org/pdf/1612.00799
3. Tavanapong W, Oh J, Riegler M, Khaleel MI, Mitta B, De Groen PC. Artificial intelligence for colonoscopy: past, present, and future. *IEEE J Biomed Health Inform*. 2022;26(8): 3950-3965.
4. Sánchez-Peralta LF, Bote-Curiel L, Picón A, Sánchez-Margallo FM, Pagador JB. Deep learning to find colorectal polyps in colonoscopy: A systematic literature review. *Artif Intell Med*. 2020;108:101923. https://www.sciencedirect.com/science/article/pii/S0933365719307493
5. Jha D, Smedsrud PH, Johansen D, et al. A comprehensive study on colorectal polyp segmentation with resUNet++, conditional random field and test-time augmentation. *IEEE J Biomed Health Inform*. 2021;25(6):2029-2040.
6. Jha D, Smedsrud PH, Riegler MA, et al. Kvasir-SEG: a Segmented polyp dataset. https://arxiv.org/pdf/1911.07069
7. Silva J, Histace A, Romain O, Dray X, Granado B. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *Int J Comput Assist Radiol Surg*. 2014;9(2):283–293. doi:10.1007/s11548-013-0926-3
8. Jorge Bernal, F J Sánchez, F Vilariño. Towards automatic polyp detection with a polyp appearance model. Undefined 2012; https://www.semanticscholar.org/paper/Towards-automatic-polyp-detection-with-a-polyp-Bernal-S%C3%A1nchez/9d32259f16ac76089c39e84063296f697a76460f
9. Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Rodríguez C, Vilariño F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput Med Imaging Graph*. 2015;43:99–111. doi:10.1016/j.compmedimag.2015.02.007
10. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze and Excitation Networks. https://arxiv.org/pdf/170901507
11. Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, et al. Attention U-Net: Learning Where to Look for the Pancreas. https://arxiv.org/pdf/1804.03999
12. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for Biomedical image segmentation. https://arxiv.org/pdf/1505.04597

13. Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK, Recurrent residual convolutional neural network based on U-net (R2U-net) for medical image segmentation. https://arxiv.org/pdf/1802.06955.
14. Zuo Q, Chen S, Wang Z. R2AU-net: attention recurrent residual convolutional neural network for Multimodal medical image segmentation. *Security Commun Net*. 2021;2021:1-10. https://www.hindawi.com/journals/scn/2021/6625688/
15. Jha D, Smedsrud PH, Riegler MA, Johansen D, de Lange T, Halvorsen P, et al. ResUNet++: an advanced architecture for medical image segmentation. https://arxiv.org/pdf/1911.07067.
16. Chattopadhyay A, Sarkar A, Howlader P. Balasubramanian VN, grad-CAM++: improved visual explanations for deep convolutional networks. https://arxiv.org/pdf/1710.11063
17. Hwang S, Oh J, Tavanapong W, Wong J, de Groen PC. Polyp detection in colonoscopy video using elliptical shape feature. 2007 IEEE International Conference on Image Processing; 2007:II-465–II-468.
18. Shin Y, Balasingham I. Automatic polyp frame screening using patch based combined feature and dictionary learning. *Comput Med Imaging Graph*. 2018;69:33-42. https://www.sciencedirect.com/science/article/pii/S0895611118300922
19. Bernal J, Tajkbaksh N, Sanchez FJ, et al. Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge. *IEEE Trans Med Imaging*. 2017;36(6):1231-1249. https://pubmed.ncbi.nlm.nih.gov/28182555/
20. Liu L, Cheng J, Quan Q, Wu FX, Wang YP, Wang J. A survey on U-shaped networks in medical image segmentations. *Neurocomputing*. 2020;409:244-258. https://www.sciencedirect.com/science/article/pii/S0925231220309218
21. Alam S, Tomar NK, Thakur A, Jha D, Rauniyar A. Automatic Polyp Segmentation using U-Net-ResNet50. https://arxiv.org/pdf/2012.15247
22. Sun X, Zhang P, Wang D, Cao Y, Liu B. Colorectal polyp segmentation by U-net with dilation Convolution. 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA); 2019:851–858.
23. Rundo L, Han C, Nagano Y, et al. USE-net: incorporating squeeze-and-excitation blocks into U-net for prostate zonal segmentation of multi-institutional MRI datasets. *Neurocomputing*. 2019;365:31-43. https://www.sciencedirect.com/science/article/pii/S0925231219309245
24. Zhang Z, Liu Q, Wang Y. Road extraction by deep residual U-net. https://arxiv.org/pdf/1711.10684
25. Sushma B, Raghavendra CK, Prashanth J. CNN based U-net with modified skip connections for colon polyp segmentation. 2021 5th International Conference on Computing Methodologies and Communication (ICCMC); 2021:1762–1766.
26. Yeung M, Sala E, Schönlieb CB, Rundo L. Focus U-net: a novel dual attention-gated CNN for polyp segmentation during colonoscopy. *Comput Biol Med*. 2021;137:104815.
27. Mahmud T, Paul B, Fattah SA. PolypSegNet: A modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images. *Comput Biol Med*. 2020;128:104119.
28. Zhong J, Wang W, Wu H, Wen Z, Qin J. PolypSeg: an efficient context-aware network for polyp segmentation from colonoscopy videos. In: Martel AL, Abol-Maesumi P, Stoyanov D, et al., eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Springer International Publishing; 2020:285-294.
29. Galdran A, Carneiro G, Ballester MAG. Double encoder-decoder networks for gastrointestinal polyp segmentation. https://arxiv.org/pdf/2110.01939
30. Li Q, Yang G, Chen Z, et al. Colorectal polyp segmentation using a fully convolutional neural network. 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI); 2017:1–5.
31. Fan DP, Ji GP, Zhou T, et al. PraNet: parallel reverse attention network for polyp segmentation. https://arxiv.org/pdf/2006.11392
32. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Proces Syst*. 2017;30.
33. Gutman D, Codella NCF, Celebi E, et al. Skin lesion analysis toward melanoma detection: a challenge at the International symposium on Biomedical imaging (ISBI); 2016, Hosted by the International Skin Imaging Collaboration (ISIC). https://arxiv.org/pdf/1605.01397
34. Chen LC, Papandreou G, Schroff F, Adam H. Rethinking Atrous convolution for semantic image Segmentation. https://arxiv.org/pdf/1706.05587
35. Chen LC, Yang Y, Wang J, Xu W, Yuille AL. Attention to scale: scale-aware semantic image segmentation. https://arxiv.org/pdf/1511.03339
36. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. https://arxiv.org/pdf/1512.03385
37. Zhang H, Dana K, Shi J, et al. Context encoding for semantic segmentation. https://arxiv.org/pdf/1803.08904
38. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid Scene Parsing Network. https://arxiv.org/pdf/1612.01105
39. Yeung M, Sala E, Schönlieb CB, Rundo L. Unified Focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Comput Med Imaging Graph*. 2022;95:102026. https://www.sciencedirect.com/science/article/pii/S0895611121001750
40. Ma J, Chen J, Ng M, et al. Loss odyssey in medical image segmentation. *Med Image Anal*. 2021;71:102035. https://www.sciencedirect.com/science/article/pii/S1361841521000815
41. Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. https://arxiv.org/pdf/1706.05721
42. Bejani MM, Ghatee M. A systematic review on overfitting control in shallow and deep neural networks. *Artif Intell Rev*. 2021;54(8):6391–6438. doi:10.1007/s10462-021-09975-1
43. Lutnick B, Ginley B, Govind D, et al. An integrated iterative annotation technique for easing neural network training in medical image analysis. *Nat Mach Intell*. 2019;1(2):112-119. doi:10.1038/s42256-019-0018-3
44. Quinn TP, Jacobs S, Senadeera M, Le V, Coghlan S. The three ghosts of medical AI: can the black-box present deliver? *Artif Intell Med*. 2022;124:102158. https://www.sciencedirect.com/science/article/pii/S0933365721001512
45. Rundo L, Pirrone R, Vitabile S, Sala E, Gambino O. Recent advances of HCI in decision-making tasks for optimized clinical workhows and precision medicine. *J Biomed Inform*. 2020;108:103479. https://www.sciencedirect.com/science/article/pii/S1532046420301076

46. Gotkowski K, Gonzalez C, Bucher A, Mukhopadhyay A. M3d-CAM: a PyTorch library to generate 3D data attention maps for medical deep learning. https://arxiv.org/pdf/2007.00453
47. Wang J, Sun K, Cheng T, et al. Deep high-resolution representation learning for visual recognition. https://arxiv.org/pdf/1908.07919
48. Peng J, Wang Y. Medical image segmentation with limited supervision: a review of deep network models. https://arxiv.org/pdf/2103.00429
49. Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging*. 2014;34(10):1993-2024.
50. Cordts M, Omran M, Ramos S, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3213–3223. doi:10.1109/CVPR.2016.350

How to cite this article: Bhattacharya D, Eggert D, Betz C, Schlaefer A. Squeeze and multi-context attention for polyp segmentation. *Int J Imaging Syst Technol*. 2022;1-20. doi:10.1002/ima.22795