GMSRF-Net: An Improved generalizability with Global Multi-Scale Residual Fusion Network for Polyp Segmentation

Abhishek Srivastava*, Sukalpa Chanda†, Debesh Jha‡, Umapada Pal* and Sharib Ali§,¶

*Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India
†Department of Computer Science and Communication, Østfold University College, Halden, Norway
‡ UiT The Arctic University of Norway, Tromsø, Norway
§ Department of Engineering Science, University of Oxford, UK
¶ School of Computing, University of Leeds, Leeds, UK
Email: s.s.ali@leeds.ac.uk

***Abstract*—Colonoscopy is a gold standard procedure but is highly operator-dependent. Efforts have been made to automate the detection and segmentation of polyps, a precancerous precursor, to effectively minimize missed rate. Widely used computer-aided polyp segmentation systems actuated by encoder-decoder have achieved high performance in terms of accuracy. However, polyp segmentation datasets collected from varied centers can follow different imaging protocols leading to difference in data distribution. As a result, most methods suffer from performance drop when trained and tested on different distributions and therefore, require re-training for each specific dataset. We address this generalizability issue by proposing a global multi-scale residual fusion network (GMSRF-Net). Our proposed network maintains high-resolution representations by performing multi-scale fusion operations across all resolution scales through dense connections while preserving low-level information. To further leverage scale information, we design cross multi-scale attention (CMSA) module that uses multi-scale features to identify, keep, and propagate informative features. Additionally, we introduce multi-scale feature selection (MSFS) modules to perform channel-wise attention that gates irrelevant features gathered through global multi-scale fusion within the GMSRF-Net. The repeated fusion operations gated by CMSA and MSFS demonstrate improved generalizability of our network.**

Experiments conducted on two different polyp segmentation datasets show that our proposed GMSRF-Net outperforms the previous top-performing state-of-the-art method by 8.34% and 10.31% on unseen CVC-ClinicDB and on unseen Kvasir-SEG, in terms of dice coefficient. Additionally, when tested on unseen CVC-ColonDB, we surpass the state-of-the-art method by 9.38% and 4.04% in terms of dice coefficient, when source dataset is Kvasir-SEG and CVC-ClinicDB, respectively.

I. Introduction

Colorectal cancer (CRC) has been consistently ranked third in terms of prevalence [1]. The leading cause of CRC is colorectal adenomatous polyps, and thus identification and resection of polyps can reduce the occurrence of CRC. Colonoscopy serves as a gold standard technique for surveillance and treatment. Studies have shown that timely colonoscopy can reduce the chances of CRC by 30% [2]. However, the success of careful identification of malicious polyps and their subsequent resection depends on the ability and experience of clinicians which makes it prone to human error. Such factors eventually lead to a high polyp missed rate [3]. Artificial intelligence (AI) driven methods can be effective and provide precise detection and segmentation of polyps.

With the advent of convolution neural networks (CNNs) research for the polyp segmentation task has been widely conducted to reduce operator-dependence problems in colonoscopy. However, the variations in structures and size of polyps and fluctuation of contrast between polyps and their immediate surrounding make it a challenging task. Whilst methods such as U-Net [4], U-Net++ [5], PraNet [6], UACA-Net [7], MSRF-Net [8] have demonstrated higher metric performances, however, performance of these methods fall considerably when the intervention of imaging protocols are different [9]. This is widely the case in different hospitals performing colonoscopy. The imaging protocols used to acquire colonoscopy images at most times vary over different medical institutions, thus performance compromise of most data driven methods is an major bottleneck as the acquisition techniques often change within the same hospital as well. Also, retraining network for each specific center consumes resources and cannot be used in resource constrained settings such as community hospitals. It is thus important to develop generalizable methods that can be used on unseen datasets without requiring to retrain them. In this work, we have designed a generalizable CNN architecture for polyp segmentation to mitigate these issues and demonstrate the effectiveness of our method on three publicly available datasets.

We can observe various reincarnations of the U-Net developed for polyp segmentation task in [5], [10], [11]. Similarly, PraNet [6] aggregated deep features in their parallel partial decoder to form initial guidance area maps. ColonSegNet [12] used only two encoder and two decoder layers that made their network parameters relatively smaller enabling a faster inference time. UACA-Net [7] used a saliency map for each level in the decoder to calculate foreground, background, and uncertain

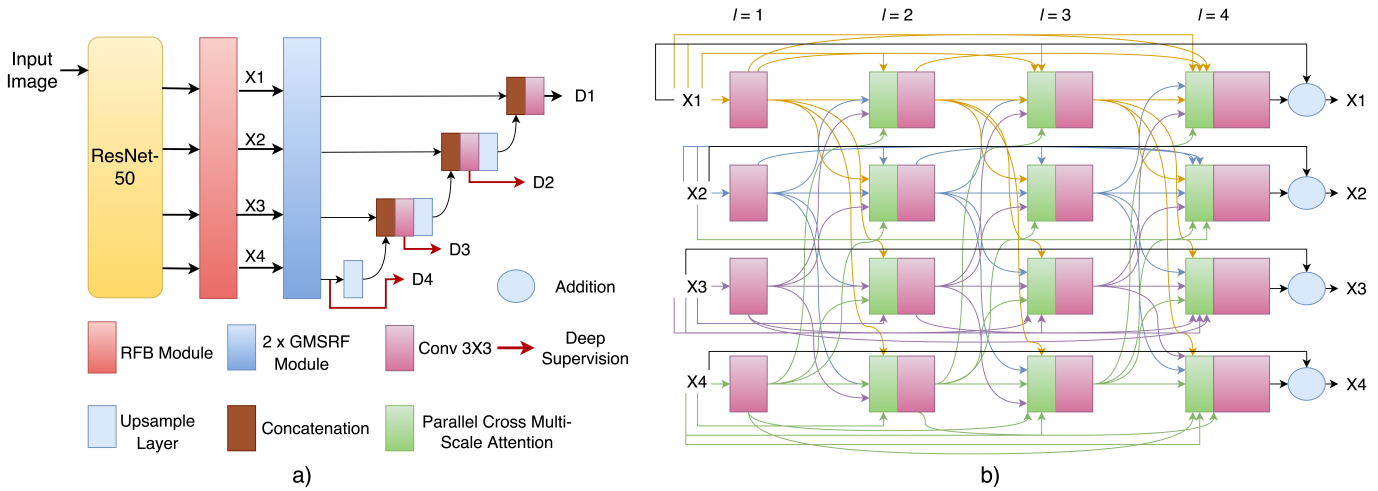

Figure 1: **The proposed global multi-scale residual fusion network (GMSRF-Net) architecture:** (On left) Overview of a complete GMSRF-Net (left). Here, images are encoded using a pre-trained ResNet50 and different scaled features are fused using two GMSRF-modules before being used by the decoder for constructing the mask. (On right) The GMSRF module that indicates the feature fusion at multiple cross-scales (colored lines). Also, feature fusion within the same layers at the output is shown by dark black lines.

area maps. However, a major drawback with encoder-decoder architectures like U-Net is that shallow features from the encoder and deep features from the decoder suffer from semantic gap [13]. Deeplabv3+ [14] introduced atrous spatial pyramid pooling with skip connections to aggregate global multi-scale context. Wang et al. [15] designed a network where spatial precision is not compromised by maintaining high-resolution representations throughout the process. Here, multi-scale fusion is performed by repeated cross-scale fusion of features for all resolution scales. Inspired by deep fusion [16], [17], MSRF-Net [8] increased the number of fusion operations by introducing dual-scale dense fusion blocks, which allowed the preservation of both high- and low-level features for all resolution scales demonstrating the superior generalizability of MSRF-Net on polyp segmentation task. Building upon these concepts we aim to further improve generalizability of polyp segmentation task for different clinical settings. For this, we introduce a global multi-scale residual fusion network "GMSRF-Net". Our main contributions include: 1) a densely connected multi-scale fusion mechanism that fuses features from all resolution scales at once, 2) increase of the frequency of fusion operations while maintaining global multi-scale context, and 3) integration of global multi-scale fusion operations and dense connections to preserve spatially relevant low-level features while generating high-level features capable of capturing global context. We have designed a novel cross multi-scale attention (CMSA) mechanism. These attention maps formed by the aggregation of multi-scale context boost the feature map representations in all resolution scales. Our multi-scale feature selection (MSFS) module, applies channel-wise attention on the features fused from all scales to further amplify the relevant features. Further, we demonstrate the improved generalizability of the proposed approach compared to former state-of-the-art (SOTA) methods.

II. Materials and Method

A. Materials

We have used three publicly available polyp segmentation datasets: Kvasir-SEG [18], CVC-ClinicDB [19], CVC-ColonDB [20]. Kvasir-SEG was acquired in Vestre Viken Health Trust in Norway, CVC-ClinicDB was obtained in Hospital Clinic in Barcelona, and CVC-ColonDB was procured by Computer Vision Center and Computer Science Department, Universitat Autonoma de Barcelona. The dataset acquisition protocols used in the three distinct centres differed in terms of resolution, type of scope, and frame rejection criterion.

To demonstrate the effectiveness of our technique, we perform six experiments with different setups. Two experiments were carried out when the training and testing datasets were the same (i.e., same center data). Additionally, to establish the generalization capacity of our network, we trained and tested our model on different datasets (referred to as "unseen" in this paper). We trained our proposed approach on Kvasir-SEG and tested it on CVC-ClinicDB and CVC-ColonDB. Additionally, we used CVC-ClinicDB as the training dataset and tested our model on Kvasir-SEG and CVC-ColonDB datasets.

B. Method

In this section, we present the architecture of our GMSRF-Net (see Fig. 1). GMSRF-Net uses global multi-scale feature fusion mechanism which incorporates attention across multi-scales and subsequent multi-scale feature selection module for accurate and generalizable segmentation of polyps. The encoder, two GMSRF modules, and decoder are detailed in the following subsections. Unlike MSRF-Net [8], our GMSRF-Net is capable of fusing features across all scales, increasing the

available pathways for a feature set to propagate before being used for forming the final segmentation map.

1) Encoder block: The colonoscopy images are first processed by ResNet50 [21] backbone pre-trained on ImageNet. The number of feature maps for all scales is reduced by Receptive Field Blocks (RFBs) [22] to reduce the computational cost incurred by the following global multi-scale residual fusion (GMSRF) module and the decoder network (see Fig. 1(a)). Here, the features generated by the RFB module be denoted as X_i where $i \in \{1, 2, 3, 4\}$ denote scales. The GMSRF module, the parallel cross multi-scale attention, and the multi-scale feature selection module entailed within it are described in the following subsection.

2) Global Multi-Scale Residual Fusion block: Let $[X_1, X_2, X_3, X_4]$ denote features of distinct spatial resolutions (see Fig. 1(b)). In the initial layer $l = 1$, where l represents the layer number in GMSRF module, each set of feature maps undergoes a convolution operation with output number of feature maps set as k. Here, k is the growth factor [23] and controls the amount of features generated by each convolution operation in the entire densely connected multi-scale fusion mechanism. We use two GMSRF modules consecutively in our network (see Fig. 1(a)).

a) Cross multi-scale attention maps (CMSA): are calculated for each scale concurrently. Eq. 1 represents how the $l'th$ CMSA is calculated for the $i'th$ scale. $\{X_w, X_y, X_z\} \neq X_i$ are first transformed to the spatial resolution size of the $i'th$ scale by suitable convolution or transposed convolution operations (see Fig. 2). They are concatenated and then processed by a 3×3 convolution operation, to effectively fuse the features of selected scales.

$$X_{att,\hat{i},l} = Conv_{1\times1}(Conv_{3\times3}(X_{w,l-1} \oplus X_{y,l-1} \oplus X_{z,l-1})), \{w, y, z\} \neq i, \{i, w, y, z\} \in \{1, 2, 3, 4\} \quad (1)$$

Here, $X_{att,\hat{i},l}$ represents the CMSA map for the $i'th$ scale of the $l'th$ layer and $\oplus$ represents concatenation operation. Attention maps are then generated to identify spatial locations based on the fused multi-scale features of parallel resolution streams. The information conveyed from low-resolution streams helps to boost the feature maps in the high-resolution stream and vice versa. The subsequent combination with the CMSA module allows the selection of features that are relevant towards identifying the region-of-interest.

b) Global multi-scale residual fusion (GMSRF): is performed as described in Eq. (2). The $l'th$ convolutional layer in the $i'th$ resolution stream receives concatenated feature maps from $l-1'th$ convolutional layer from all resolution scales and previous convolutional layers for the same resolution stream (see Fig. 1(b)). This global multi-scale fusion with densely connected blocks increases the number of paths through which feature maps can propagate and undergoes varying operations before contributing to the final segmentation map prediction.

$$X_{i,l} = Conv_{3\times3}(X_{i,0} \cdots X_{i,l-1} \oplus X_{w,l-1} \oplus X_{y,l-1} \oplus X_{z,l-1}), \{w, y, z\} \neq i, \{i, w, y, z\} \in \{1, 2, 3, 4\} \quad (2)$$

The feature maps can capture the global multi-scale context at each layer of the densely connected mechanism. Eq. (3) describes how CMSA maps are used to identify and propagate relevant features of the $i'th$ scale stream forward.

$$X_{i,l} = X_{i,l} \otimes X_{att,\hat{i},l} \quad (3)$$

c) Multi-scale feature selection (MSFS): module, is the next step where channel-wise attention is applied on the fused features using squeeze and excitation (S&E) block [24] (refer to Fig. 2). This enables the amplification of salient channels transmitted by various scale streams. The suppression of irrelevant channels by this module is also conducive to a higher level of accuracy. Residual connection from the input of the GMSRF module is added to improve gradient flow. For simplification purposes, we use the $i'th$ scale while describing this mechanism.

3) Decoder: To fully establish the contribution of our GMSRF-Net, we choose to use a vanilla decoder (see Fig. 1(a)). X_i is the output of the GMSRF module for the $i'th$ scale. Each decoder block upscales the output from the previous decoder block and concatenates the resultant feature maps from the same scale output of the GMSRF module (see Equation 4).

$$D_i = Conv_{3\times3}(TransConv(D_{i-1}) \oplus X_i) \quad (4)$$

Here, TransConv is strided transposed convolutional layer and initially $D_4 = X_4$. The output of all decoder blocks, i.e D_4, D_3, D_2, are upscaled to the size of ground truth maps for improved gradient flow and regularization.

4) Loss Function: We use a dual loss function $\mathcal{L}_{\text{DUAL}} = \lambda_1 \mathcal{L}_{\text{BCE}} + \lambda_2 \mathcal{L}_{\text{IoU}}$, where $\mathcal{L}_{\text{DUAL}}$ is a combination of weighted intersection over union loss ($\mathcal{L}_{\text{IoU}}$) and binary cross entropy ($\mathcal{L}_{\text{BCE}}$). Each of the loss components are equally weighted i.e. $\lambda_1 = \lambda_2 = 1$. For all supervise segmentation maps generated by all decoder levels, the total loss function is given by: $\mathcal{L}_{\text{GMSRF}} = \sum_{i=1}^{i=4} \mathcal{L}_{\text{DUAL}}(D_i)$, where i is the number of decoder layers.

III. Experiments

A. Experimental setup

We evaluate our proposed GMSRF-Net on Kvasir-SEG, CVC-ClinicDB, and CVC-ColonDB. All images are resized to 256×256 as a pre-processing step. We reserve 80% data for training, 10% for validation, and 10% for testing. The entire CVC-ColonDB dataset is used for testing. The training set is augmented using techniques like random flipping, cropping, color jittering etc. The growth factor k used in the GMSRF module is set to 0.4. We train the network for 50 epochs using Adam optimizer with initial learning rate of $1e-4$ and batch size of 8. ResNet-50 pre-trained on ImageNet is used as as backbone, where we do not freeze the weights. We use the author released source code for all baselines. Each method used for comparison follow the same training, testing and validation split. All experiments were performed on an NVIDIA DGX-2 machine using NVIDIA V100.

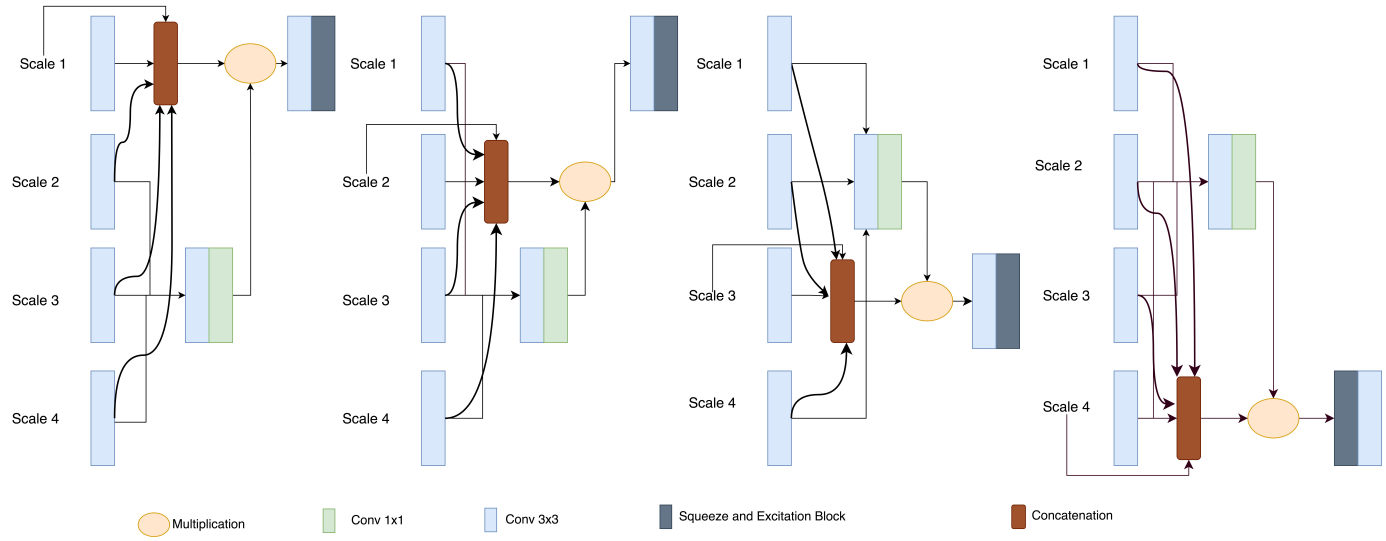

Figure 2: **Computation of (CMSA) and (MSFS) for all scales:** Illustration of how the feature fusion across multiple scales is carried out in our CMSA and MSFS blocks. CMSA maps are computed for each scale in parallel for a particular layer, The bold arrows elucidate the selection and flow of different scale features. These features are then used for the generation of CMSA maps that are then multiplied with the fused multi-scale features. Finally, the MSFS module is used for further gating irrelevant channels.

Table I: Result comparison on seen and unseen dataset. Here we have used Kvasir-SEG as seen (source) dataset while CVC-ClinicDB and CVC-ColonDB have been used as unseen datasets to assess generalizability of our proposed approach.

Method	**Source data "Kvasir-SEG"**				**Unseen dataset "CVC-ClinicDB"**				**Unseen dataset "CVC-ColonDB"**			
	DSC	**mIoU**	**Recall**	**Precision**	**DSC**	**mIoU**	**Recall**	**Precision**	**DSC**	**mIoU**	**Recall**	**Precision**
U-Net [4]	0.8629	0.8176	0.9094	0.8901	0.7172	0.6133	0.7255	0.7986	0.5106	0.3848	0.4497	**0.7273**
U-Net++ [5]	0.7475	0.6313	0.6865	0.8871	0.4265	0.3345	0.3939	0.6894	0.3126	0.2532	0.3053	0.5973
Deeplabv3+ (Xception) [14]	0.8965	0.8575	0.8984	0.9496	0.6509	0.5385	0.6251	0.7947	0.5197	0.4296	0.5047	0.7429
Deeplabv3+ (Mobilenet) [14]	0.8656	0.8186	0.8808	0.9205	0.6303	0.4825	0.5957	0.7173	0.4318	0.3503	0.4756	0.5708
HRNetV2-W18 -Smallv2 [25]	0.8179	0.7470	0.8016	0.8696	0.6428	0.5513	0.6811	0.7253	0.3597	0.2925	0.4382	0.4099
HRNetV2-W48 [25]	0.8896	0.8262	0.8973	0.9056	0.7901	0.6953	0.8796	0.7694	0.5180	0.4462	0.6159	0.5393
ColonSegNet [12]	0.8203	0.7435	0.8124	0.8832	0.6895	0.5813	0.7862	0.7177	0.3936	0.3005	0.4597	0.4884
PraNet [6]	0.9078	0.8561	0.9034	0.9352	0.7225	0.6328	0.7531	0.7888	0.4859	0.4220	0.5059	0.5380
UACANet-S [7]	0.8800	0.8250	0.8701	0.9229	0.5683	0.4907	0.5792	0.7095	0.2890	0.2400	0.2869	0.4156
UACANet-L [7]	0.9014	0.8555	0.8897	0.9381	0.5589	0.4849	0.5800	0.6775	0.2973	0.2545	0.2923	0.4166
MSRF-Net [8]	0.9217	**0.8914**	0.9198	**0.9666**	0.7921	0.6498	0.9001	0.7000	0.5391	0.4017	**0.8372**	0.4357
GMSRF-Net	**0.9263**	0.8843	**0.9402**	0.9310	**0.8755**	**0.8091**	**0.9106**	**0.8588**	**0.6329**	**0.5611**	0.6895	0.6513

B. Evaluation metrics

For the evaluation of our model, we have chosen Sørensen–dice coefficient (DSC), mean intersection over union (mIoU), precision, and recall.

$$\text{DSC} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{5}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{6}$$

$$\text{Recall } (Rec.) = \frac{TP}{TP + FN} \tag{7}$$

$$\text{Precision } (Prec.) = \frac{TP}{TP + FP} \tag{8}$$

In the equations (5)-(8), TP, FP, TN, FN represents true positives, false positives, true negatives, and false negatives, respectively, for the classification outputs.

C. Results and discussion

We provide quantitative results to demonstrate competitiveness of our approach and improved generalizability against SOTA methods. For this algorithms are evaluated on both test split from source data and unseen test dataset (unseen, not used during training). Additionally, we provide visual comparisons and network ablation to demonstrate the effectiveness of our approach.

1) Quantitative results and generalizability assessment: From Table I, it can be observed that our GMSRF-Net is competitive to MSRF-Net on the Kvasir-SEG for the same source

Table II: Result comparison on seen and unseen dataset. Here, we have used CVC-ClinicDB as seen (source) dataset while Kvasir-SEG and CVC-ColonDB has been used as unseen datasets to assess generalizability of our proposed approach.

Method	Source data "CVC-ClinicDB"				Unseen dataset "Kvasir-SEG"				Unseen dataset "CVC-ColonDB"			
	DSC	mIoU	Recall	Precision	DSC	mIoU	Recall	Precision	DSC	mIoU	Recall	Precision
U-Net [4]	0.9145	0.8654	0.9178	0.9381	0.6222	0.4588	0.5129	0.8133	0.5334	0.3745	0.5685	0.5232
U-Net++ [5]	0.8453	0.7559	0.8917	0.8323	0.5926	0.4564	0.7352	0.5462	0.3702	0.2372	0.4465	0.3360
Deeplabv3+ (Xception) [14]	0.8897	0.8706	0.9251	0.9366	0.6746	0.5327	0.7757	0.6296	0.4834	0.3657	0.5021	0.5739
Deeplabv3+ (Mobilenet) [14]	0.8985	0.8588	0.9160	0.9287	0.6474	0.5098	0.6632	0.6878	0.5070	0.3749	0.5305	0.5612
HRNetV2-W18 -Smallv2 [25]	0.9073	0.8457	0.9137	0.9191	0.7012	0.6009	0.7184	0.7666	0.5749	0.4937	0.6010	0.6237
HRNetV2-W48 [25]	0.9244	0.8747	0.9234	0.9296	0.7404	0.6233	0.7293	0.8511	0.6294	0.5571	0.6620	0.6715
ColonSegNet [12]	0.9132	0.8600	0.9072	0.9292	0.6324	0.5183	0.6112	0.7897	0.4797	0.3822	0.5356	0.5616
PraNet [6]	0.9072	0.8575	0.9227	0.9134	0.7293	0.6262	0.8007	0.7623	0.5875	0.5186	0.6451	0.6334
UACANet-S [7]	0.9190	0.8700	0.9285	0.9201	0.6945	0.5894	0.7692	0.7377	0.5491	0.4669	0.6229	0.5880
UACANet-L [7]	0.9098	0.8649	0.9174	0.9114	0.7312	0.6383	0.7417	0.8314	0.5448	0.4803	0.5591	0.6121
MSRF-Net [8]	**0.9420**	**0.9043**	**0.9567**	**0.9427**	0.7575	0.6337	0.7197	0.8414	0.6308	0.4310	0.5228	**0.7106**
GMSRF-Net	0.9326	0.8882	0.9376	0.9307	**0.8606**	**0.7877**	**0.8641**	**0.9056**	**0.6712**	**0.6018**	**0.7121**	0.6849

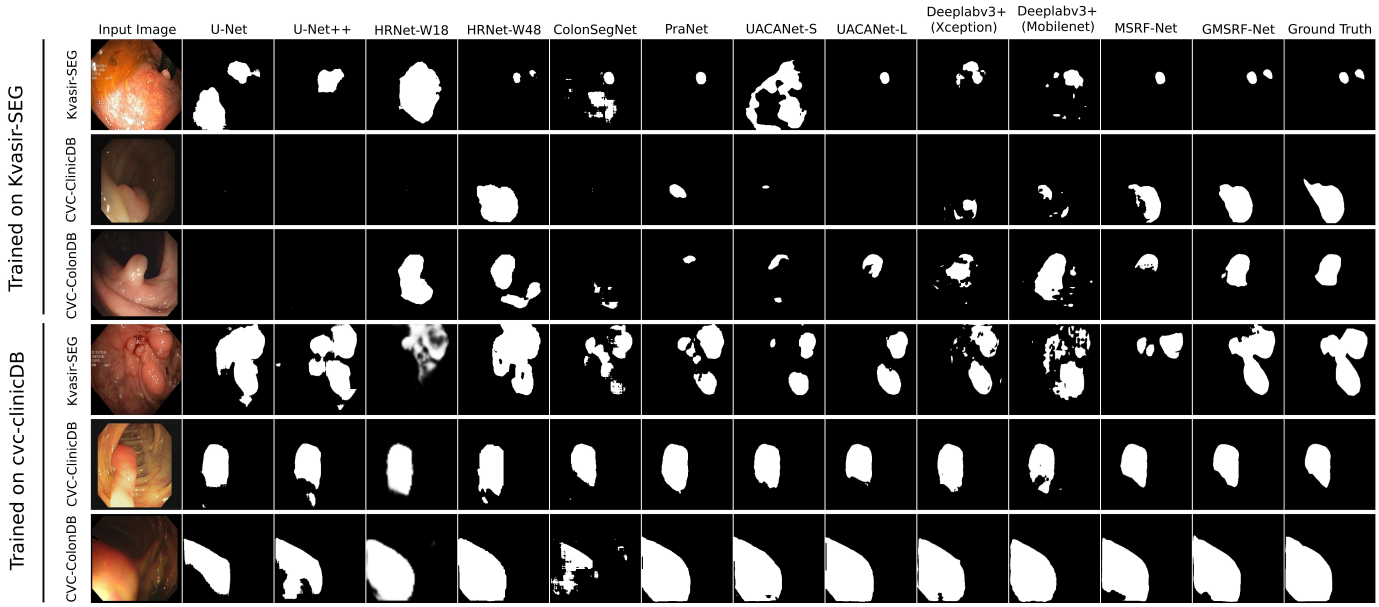

Figure 3: Qualitative comparison of GMSRF-Net with other state-of-the-art methods

data while outperforming on unseen data (CVC-ClinicDB). An increase of 8.34%, 15.93%, 1.05%, 15.88% in dice coefficient (DSC), mIoU, recall and precision, respectively, can be seen when compared with the best performing SOTA method MSRF-Net. Moreover, when tested on CVC-ColonDB as an unseen dataset, GMSRF-Net achieves 9.31% increase DSC as compared to MSRF-Net, and an 11.49% increase in mIoU as compared to HRNetV2-W48. A similar trend can be noted in Table II where our proposed method outperformed SOTA methods on unseen Kvasir-SEG by large margins on all metrics, i.e., an improvement of 10.31%, 15.40%, 14.44% and 6.42% on DSC, mIoU, recall and precision, respectively. Furthermore, when tested on unseen CVC-ColonDB, we observed an improvement of 4.04%, 4.47%, 5.01% on DSC, mIoU, recall and precision, respectively, over SOTA methods. Moreover, we can observe that GMSRF-Net achieves a DSC of 0.8606 when trained on CVC-ClinicDB and tested on Kvasir-SEG (see Table II). However, some networks such as U-Net++, ColonSegNet, and HRNetV2-W18-Smallv2 reports relatively lower performance even when they are trained and tested on Kvasir-SEG (see Table I).

2) Qualitative results: Figure 3 (top) illustrates the qualitative superiority achieved by the GMSRF-Net over other SOTA methods when trained on Kvasir-SEG and tested on Kvasir-SEG (same as training), CVC-ClinicDB (unseen) and CVC-ColonDB (unseen). It can be observed that when tested on Kvasir-SEG, GMSRF-Net achieved improvement over the most accurate SOTA MSRF-Net. When tested on unseen datasets, GMSRF-Net is capable of accurately segmenting polyps, whereas multiple methods failed to even locate polyp in some samples. Figure 3 (bottom) also demonstrates the qualitative comparison of GMSRF-Net with other baselines

Table III: Ablation study of GMSRF-Net using Kvasir-SEG as the source dataset while CVC-ClinicDB and CVC-ColonDB has been used as unseen datasets to assess generalizability of the ablated networks.

Ablation design	Source data "Kvasir-SEG"				Unseen dataset "CVC-ClinicDB"				Unseen dataset "CVC-ColonDB"			
	DSC	**mIoU**	**Recall**	**Precision**	**DSC**	**mIoU**	**Recall**	**Precision**	**DSC**	**mIoU**	**Recall**	**Precision**
1 x GMSRF module	0.9218	0.8769	0.9334	0.9307	0.8322	0.7623	0.8668	0.8371	0.6258	0.5520	0.6545	**0.6673**
3 x GMSRF module	0.9162	0.8696	0.9353	0.9199	0.8650	0.7944	0.9103	0.8419	0.5859	0.5226	0.6211	0.5916
w/o CMSA in Scale 1	0.9119	0.8638	0.9217	0.9280	0.8540	0.7750	0.8893	0.8607	0.5978	0.5331	0.6558	0.6139
w/o CMSA in Scale 2	0.9242	0.8794	0.9410	0.9233	0.8735	0.8036	0.9041	**0.8597**	0.5745	0.5101	0.6002	0.6135
w/o CMSA in Scale 3	0.9188	0.8716	0.9339	0.9219	0.8464	0.7773	0.8720	0.8565	0.5922	0.5232	0.6126	0.6310
w/o CMSA in Scale 4	0.9058	0.8514	0.9106	0.9246	0.8333	0.7578	0.8538	0.8633	0.6005	0.5259	0.6344	0.6396
w/o CMSA	0.8993	0.8424	0.9319	0.8897	0.8362	0.7579	0.8981	0.8229	0.5328	0.4627	0.5705	0.5776
w/o MSFS	0.9227	0.8743	0.9426	0.9266	0.8496	0.7719	0.8900	0.8414	0.6204	0.5414	0.6707	0.6424
w/o Deep Supervision	0.9237	0.8803	0.9323	0.9329	0.8700	0.7987	0.8958	0.8603	0.5656	0.5044	0.6130	0.5755
GMSRF-Net	**0.9263**	**0.8843**	**0.9402**	**0.9310**	**0.8755**	**0.8091**	**0.9106**	0.8588	**0.6329**	**0.5611**	**0.6895**	0.6513

where all networks are trained on CVC-ClinicDB (seen) and tested on Kvasir-SEG (unseen), CVC-ClinicDB (seen), and CVC-ColonDB (unseen). Under this scenario, we can observe that our GMSRF-Net again generates predicted masks more visually similar to the ground truth than other SOTA methods for most samples. Our experiments demonstrate that the multi-scale fusion technique that combines features from all resolution scales yields better generalization performances (also see Table I-II). The reason behind this improved generalization ability of GMSRF-Net can be due to use of global multi-scale residual fusion in our network that increases the number of fusion operations together with attention modules (CMSA and MSFS).

3) Ablation study: We perform an ablation study (see Table III) to demonstrate the significance of each component used in our GMSRF-Net. From Table III we can observe the impact of increasing/decreasing the number of GMSRF modules used in the GMSRF-Net. When we reduce the number of GMSRF modules to 1x, we observe a 0.45%, 4.43%, and 0.71% decrease in DSC in Kvasir-SEG, CVC-ClinicDB and CVC-ColonDB, respectively. Whereas, when we increase the number of GMSRF modules to 3x, we observe a 1.01%, 1.05%, and 4.7% decrease in DSC in Kvasir-SEG, CVC-ClinicDB and CVC-ColonDB, respectively. Hence, the experiments mentioned above validate the choice to use 2x GMSRF module in the GMSRF-Net. To determine the contribution of the CMSA module in enhancing the performance gain by progressively gating irrelevant features and amplifying informative features, we ablate the mechanism by removing the multi-scale mechanism in each scale and then by disabling it entirely. In Table III, we can observe that when multi-scale attention is removed in scale 1, scale 2, scale 3, scale 4 we observe a 1.44%, 0.21%, 0.75%, 2.05% decrease in DSC respectively when tested on Kvasir-SEG. A similar trend is observed when the ablated networks are tested on unseen CVC-ClinicDB and CVC-ColonDB. When CMSA is ablated across all scales, we observe a significant drop of 2.70%, 3.93%, 10.01% when tested on Kvasir-SEG, CVC-ClinicDB, and CVC-ColonDB respectively (see Table III). The boost in performance achieved by the MSFS mechanism can be noticed in Table III, where GMSRF-Net without MSFS mechanism suffers from a drop of 0.36%, 2.59%, 1.25% in DSC when tested on Kvasir-SEG, CVC-ClinicDB and CVC-ColonDB.

IV. Conclusion

In this paper, we have proposed a global multi-scale feature fusion technique that incorporates together with attention and gating mechanisms (i.e., CMSA and MSFS modules) that allow reliable and robust global feature aggregation. Feature profiling and pruning at each step makes the network capable of addressing variability in samples in varied datasets. Our proposed network maintains high resolution representations and enriches high-resolution features by fusion with low-resolution feature streams and vice versa. The proposed technique achieved significant performance gain on segmentation tasks where the training and testing datasets are from different distributions. The generalization performance of our GMSRF-Net is thus an important step towards improving the generalizability of supervised learning methods. In future, we will extend our work towards quantifying the generalizability of the proposed model on other biomedical imaging datasets.

References

[1] N. Howlader, A. Noone, M. Krapcho, D. Miller, A. Brest, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, D. Lewis *et al.*, "SEER Cancer Statistics Review, 1975–2018, National Cancer Institute. Bethesda, MD," 2018.

[2] F. A. Haggar and R. P. Boushey, "Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors," *Clinics in colon and rectal surgery*, vol. 22, no. 04, pp. 191–197, 2009.

[3] J. G.-B. Puyal, K. K. Bhatia, P. Brandao, O. F. Ahmad, D. Toth, R. Kader, L. Lovat, P. Mountney, and D. Stoyanov, "Endoscopic Polyp Segmentation Using a Hybrid 2D/3D CNN," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 295–305.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *Proc. of Internat. Confer. on Med. Ima. Compu. Comput.-Assis. Interven.*, 2015, pp. 234–241.

[5] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, 2019.

[6] D.-P. Fan *et al.*, "PraNet: parallel reverse attention network for polyp segmentation," in *Proc. of Internat. Confer. on Med. Ima. Compu. Comput.-Assis. Interven.*, 2020, pp. 263–273.

[7] T. Kim, H. Lee, and D. Kim, "UACANet: Uncertainty Augmented Context Attention for Polyp Segmentation," *arXiv preprint arXiv:2107.02368*, 2021.

[8] A. Srivastava, D. Jha, S. Chanda, U. Pal, H. D. Johansen, D. Johansen, M. A. Riegler, S. Ali, and P. Halvorsen, "MSRF-Net: A Multi-Scale Residual Fusion Network for Biomedical Image Segmentation," *arXiv preprint arXiv:2105.07451*, 2021.

[9] S. Ali *et al.*, "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy," *Med. Imag. Anal.*, p. 102002, 2021.

[10] D. Jha *et al.*, "ResUNet++: An advanced architecture for medical image segmentation," in *Proc. of Internat. Sympos. Multime.*, 2019, pp. 225–230.

[11] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep learn. med. ima. anal. multimo. learn. clini. deci. sup.*, 2018, pp. 3–11.

[12] D. Jha *et al.*, "Real-Time Polyp Detection, Localisation and Segmentation in Colonoscopy Using Deep Learning," *IEEE Acc.*, 2021.

[13] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neur. Networ.*, vol. 121, pp. 74–87, 2020.

[14] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. of the Europ. conf. comput. vis.*, 2018, pp. 801–818.

[15] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE trans. patt. analy. mach.*, 2020.

[16] K. Sun, M. Li, D. Liu, and J. Wang, "Igcv3: Interleaved low-rank group convolutions for efficient deep neural networks," *arXiv preprint arXiv:1806.00178*, 2018.

[17] G. Xie, J. Wang, T. Zhang, J. Lai, R. Hong, and G.-J. Qi, "Interleaved structured sparse convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8847–8856.

[18] D. Jha *et al.*, "Kvasir-SEG: A Segmented Polyp Dataset," in *Proc. of Internat. Conf. Multimed. Model.*, 2020, pp. 451–462.

[19] J. Bernal *et al.*, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computer. Medi. Imag. Graph.*, vol. 43, pp. 99–111, 2015.

[20] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2015.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[22] S. Liu, D. Huang *et al.*, "Receptive field block net for accurate and fast object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 385–400.

[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. of Comput. Vis. and Patt. Recogn.*, 2018, pp. 7132–7141.

[25] J. Wang and other, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. on Patt. Analy. Mach. Intelli.*, p. 1–1, 2020.