This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3138024, IEEE Journal of Biomedical and Health Informatics

MSRF-Net: A Multi-Scale Residual Fusion Network for Biomedical Image Segmentation

Abhishek Srivastava, Debesh Jha, *Member, IEEE*, Sukalpa Chanda, *Member, IEEE*, Umapada Pal, *Senior Member, IEEE*, Håvard D. Johansen, *Member, IEEE*, Dag Johansen, *Member, IEEE*, Michael A. Riegler, *Member, IEEE*, Sharib Ali, *Member, IEEE*, Pål Halvorsen *Member, IEEE*

***Abstract*— Methods based on convolutional neural networks have improved the performance of biomedical image segmentation. However, most of these methods cannot efficiently segment objects of variable sizes and train on small and biased datasets, which are common for biomedical use cases. While methods exist that incorporate multi-scale fusion approaches to address the challenges arising with variable sizes, they usually use complex models that are more suitable for general semantic segmentation problems. In this paper, we propose a novel architecture called Multi-Scale Residual Fusion Network (MSRF-Net), which is specially designed for medical image segmentation. The proposed MSRF-Net is able to exchange multi-scale features of varying receptive fields using a Dual-Scale Dense Fusion (DSDF) block. Our DSDF block can exchange information rigorously across two different resolution scales, and our MSRF sub-network uses multiple DSDF blocks in sequence to perform multi-scale fusion. This allows the preservation of resolution, improved information flow and propagation of both high- and low-level features to obtain accurate segmentation maps. The proposed MSRF-Net allows to capture object variabilities and provides improved results on different biomedical datasets. Extensive experiments on MSRF-Net demonstrate that the proposed method outperforms the cutting-edge medical image segmentation methods on four publicly available datasets. We achieve the Dice Coefficient (DSC) of 0.9217, 0.9420, and 0.9224, 0.8824 on Kvasir-SEG, CVC-ClinicDB, 2018 Data Science Bowl dataset, and ISIC-2018 skin lesion segmentation challenge dataset respectively. We further conducted generalizability tests and achieved DSC of 0.7921 and 0.7575 on CVC-ClinicDB and Kvasir-SEG, respectively.**

***Index Terms*— Medical image segmentation, MSRF-Net, multi-scale fusion, generalization, colonoscopy**

I. Introduction

MEDICAL image segmentation is an essential task in clinical diagnosis and has been extensively studied by the medical image analysis community [1]–[3]. The semantic segmentation results can help identify regions of interest for lesion assessment, such as polyps in the colon, to inspect if they are cancerous and remove them if necessary. Thus, the segmentation results can help to detect missed lesions, prevent diseases, and improve therapy planning and treatment. The significant challenge in medical imaging is the requirement of a large number of high-quality labeled and annotated datasets. This is a key factor in the development of robust algorithm for automated medical image segmentation task.

The manual pixel-wise annotation of medical image data is very time-consuming, requires collaborations with experienced medical experts, and is costly. During the annotation of the regions in medical images (for example, polyps in still frames), the guidelines and protocols are set based on which expert performs the annotations. However, there might exist discrepancies among the experts, e.g., while considering a particular area in the lesion as cancerous or non-cancerous. Additionally, the lack of standard annotation protocols for various imaging modalities and low image quality can influence annotation quality. Other factors such as the annotator's attentiveness, type of display device, image-annotation software and data misinterpretation due to lightning conditions can also affect the quality of annotations [4]. An alternative solution to manual image segmentation is an automated computer aided segmentation based diagnosis-assisting system that can provide a faster, more accurate, and more reliable solution to transform clinical procedures and improve patient care. Computer aided diagnosis will reduce the expert's burden and also reduce the overall treatment cost. Due to the diverse nature of medical-imaging data, computer aided diagnosis based segmentation models must be robust to variations in imaging modalities [5].

In the past years, Convolutional Neural Networks (CNNs) based approaches have overcome the limitations of traditional segmentation methods [6] in various medical imaging modalities such as X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), endoscopy, wireless capsule endoscopy, dermatoscopy, and in high-throughput imaging like histopathology and electron microscopy. Modern semantic and instance segmentation architectures are usually encoder-decoder based networks [7], [8]. The success of deep encoder-decoder based CNNs is largely due to their skip connections, which allows the propagation of deep, semantically meaningful, and dense feature maps from the encoder network to the decoder sub-networks [9], [10]. However, encoder-decoder based image segmentation architectures have limitations in optimal depth and design of the skip connections [11]. The

A. Srivastava is with Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India
D. Jha is with SimulaMet, Oslo, Norway and UiT The Arctic University of Norway, Tromsø, Norway (corresponding email: debesh@simula.no)
S. Chanda is with Østfold University College, Halden, Norway
U. Pal is with Indian Statistical Institute, Kolkata, India
H. D. Johansen and D. Johansen are with UiT The Arctic University of Norway, Tromsø, Norway
M. A. Riegler is with SimulaMet, Oslo, Norway and UiT The Arctic University of Norway, Tromsø, Norway
S. Ali is with the Department of Engineering Science, University of Oxford, and Oxford NIHR Biomedical Research Centre, Oxford, UK (corresponding email: sharib.ali@eng.ox.ac.uk)
P. Halvorsen is with SimulaMet, Oslo, Norway and Oslo Metropolitan University, Oslo, Norway
S. Ali and P. Halvorsen: Shared senior authorship

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3138024, IEEE Journal of Biomedical and Health Informatics

optimal depth of the architectures can vary from one biomedical application to another. The number of samples in the dataset used in training also contributes to the limitation on the complexity of the network. The design of skip connections are sometimes unnecessarily restrictive, demanding the fusion of the same-scale encoder and decoder feature maps. Moreover, traditional CNN methods do not make use of the hierarchical features.

In this paper, we propose a novel medical image segmentation architecture, called *MSRF-Net*, which aims to overcome the above discussed limitations. Our proposed MSRF-Net maintains high-resolution representation throughout the process, which is conducive to potentially achieving high spatial accuracy. The MSRF-Net utilizes a novel Dual-Scale Dense Fusion (DSDF) block that performs dual scale feature exchange and a sub-network that exchanges multi-scale features using the DSDF block. The DSDF block takes two different scale inputs and employs a residual dense block that exchanges information across different scales after each convolutional layer in their corresponding dense blocks. The densely connected nature of blocks allows relevant high- and low-level features to be preserved for the final segmentation map prediction. The multi-scale information exchange in our network preserves both high- and low-resolution feature representations, thereby producing finer, richer and spatially accurate segmentation maps. The repeated multi-scale fusion helps in enhancing the high-resolution feature representations with the information propagated by low-resolution representations. Further, layers of residual networks allow redundant DSDF blocks to die out, and only the most relevant extracted features contribute to the predicted segmentation maps.

Additionally, we propose adding a complimentary gated shape stream that can leverage the combination of high- and low-level features to compute shape boundaries accurately. We have evaluated the MSRF-Net segmentation model using four publicly available biomedical datasets. The results demonstrate that the proposed MSRF-Net outperforms the State Of The Art (SOTA) segmentation methods on most standard computer vision evaluation metrics.

The main contributions of this work are as following:

1) Our proposed MSRF-Net architecture is based on a DSDF block that comprises of residual dense connections and exchanging information across multiple scales. This allows both high-resolution and low-resolution features to propagate, thereby extracting semantically meaningful features that improve segmentation performance on various biomedical datasets.
2) MSRF-Net computes the multi-scale features and fuses them effectively using a DSDF block. The residual nature of DSDF block improves gradient flow which improves the training efficiency, i.e., reducing the need for large datasets for training.
3) The effectiveness of MSRF-Net is demonstrated on four public datasets: Kvasir-SEG [12], CVC-ClinicDB [13], 2018 Data Science Bowl (DSB) Challenge [2], and ISIC 2018 Challenge [14], [15]. We conduct a generalizability study of the proposed network for which we trained our model on Kvasir-SEG and tested on the CVC-ClinicDB and vice versa. The experimental results and their comparison with established computer vision methods confirmed that our approach is more generalizable.

II. Related work

A. Medical image segmentation

Long et al. [16] proposed a Fully Convolutional Network (FCN) that included only convolutional layers for semantic segmentation. Subsequently, Ronneberger et al. [17] modified the FCN with an encoder-decoder U-Net architecture for segmentation of HeLa cells and neuronal structures of electron microscopic stacks. In the U-Net [17], low- and high-level feature maps are combined through skip connections. The high-level feature maps are processed by deeper layers of the encoder network and propagated through the decoder whereas, the low-level features are propagated from the initial layers of the network. This may cause a semantic gap between the high- and low-level features. Ibtehaz et al. [18] proposed to add convolutional units along the path of skip connections to reduce the semantic gap. Oktay et al. [19] proposed an attention U-Net that used an attention block to alter the feature maps propagated through the skip-connections. Here, the previous decoder block output was used to form a gating mechanism to prune unnecessary spatial features passing from the skip-connections and to keep only the relevant features. In addition, various other extensions of the U-Net have been proposed [10], [11], [20]–[22]. To incorporate global context information for the task of scene parsing, PSPNet [23] generated hierarchical feature maps through a Pyramid Pooling Module (PPM). Similarly, Chen et al. [24] used the Atrous Spatial Pyramid Pooling (ASPP) to aggregate the global features. Later, the same group proposed the DeepLabV3+ [25] architecture that used skip connections between the encoder and decoder. Both of these networks have been widely used by the biomedical imaging community [26]–[28].

Hu et al. [29] proposed SE-Net, which pioneered channel-wise attention. The Squeeze and Excitation (S&E) block was able to model interdependencies between the channels and derive a global information map that helps in emphasizing relevant features and suppressing irrelevant features. FED-Net [30] incorporated these S&E blocks in their modified U-Net architecture. Kaul et al. [31] incorporated both types of attention, i.e., spatial and channel-wise attention, in their proposed FocusNet. Jha et al. [20] modified ResUNet [32] adding ASPP, S&E block [29] and attention mechanisms to boost the performance of the network further. Taikkawa et al. [33] proposed Gated-SCNN, which pioneered the idea of gated shape stream to generate finer segmentation maps leveraging the shape and boundaries of the target object. The shape stream was recently also employed by Sun at al. [22] to capture the shape and boundaries of the target segmentation map for medical segmentation problems. Fan et al. [34] devised a parallel partial decoder (PraNet) that aggregated high-level features to generate a guidance map that estimates a rough location of the region of interest. The guidance map used in PraNet was then used with a reverse attention module to extract finer boundaries from the low-level features. Kim et al. [35] modified the U-Net architecture

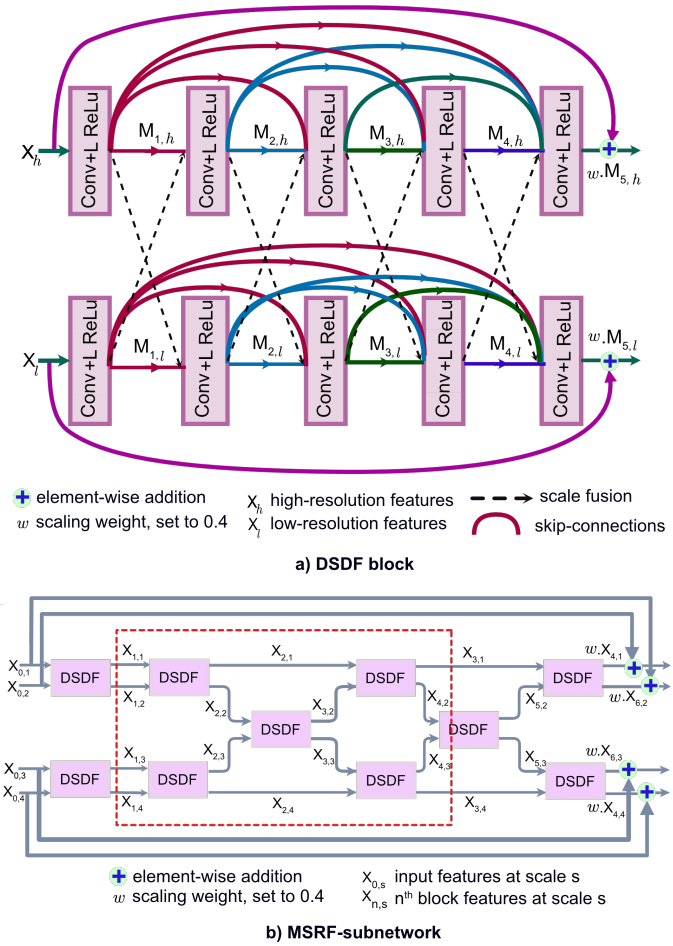

Fig. 1. Components of our MSRF-Net, a) Proposed Dual-Scale Dense Fusion (DSDF) block and b) Multi-Scale Residual Fusion (MSRF). Dotted rectangle block in (b) represents multi-scale feature exchange in MSRF-Net

and added additional encoder and decoder modules. Saliency maps computed by a prediction module in the UACA-Net is used to compute foreground, background and uncertain area maps for each representation. The relationship between each representation is computed and used by the next prediction module. A detailed summary of advances of deep-learning based methodologies in medical image segmentation can be found in [36]–[38].

B. Residual dense blocks

Dense connections are a unique approach of improving information flow and keeping a collection of diversified features. The architectures based on dense connections are characterized by each layer receiving inputs from all previous layers. Various medical image segmentation methods [11], [39]–[42] leverage the diversified features captured by such dense connections to improve segmentation performance. Guan et al. [39] modified the U-Net architecture by substituting standard encoder-decoder units with densely connected convolutional units. Zhou et al. [11] conceived an architecture where the encoder and decoder are connected through dense and nested skip pathways for efficient feature fusion between the feature maps of encoder and decoder. Zhang et al. [40] proposed Residual Dense Blocks (RDB) to extract local features via densely connected convolutional layers. Additionally, their architecture allowed them to connect the previous RDB block to all the current RDB blocks and a final global fusion through 1×1 convolutions for maintaining global hierarchical feature extraction. In ResUNet [41] and Residual Dense U-Net (RD-U-Net) [42], the RDBs are included in a standard U-Net based architecture to make use of hierarchical features. Dolz el al. [43] proposed HyperDense-Net, which introduced a two-stream CNN designed to process each modality in a separate stream for multi-modal image segmentation. The dense connections were used across layers of the same path and also between layers of a different path, therefore, increasing the capacity of the network to learn more complex combination between different modality.

C. Multi-scale fusion

Maintaining a high-resolution representation of the image is important for segmentation architectures to precisely capture the spatial information and give accurate segmentation maps [44]. Rather than recovering such representations from low-level representations, multi-scale fusion can help exchange high- and low-resolution features throughout the segmentation process. Wang et al. [44] demonstrated that such exchange of features improves the flow of high-resolution features and can potentially lead to a more spatially accurate segmentation map. They achieved this by processing all the resolution streams in parallel, keeping the resolution representation for each resolution, and performing the feature fusion across all resolution scales.

The previous works by Ronneberger et al. [17] and Badrinarayanan et al. [45] used skip-connections to concatenate high-resolution feature representations at each level with the upscaled features in the decoder to preserve both high- and low-resolution feature representations. Zhao et al. [23] used pyramid pooling to perform multi-resolution fusion while Chen et al. [24] used ASPP and multiple Atrous convolutions with different sampling rates. Similarly, Yang et al. [46] used densely connected atrous convolutional layers in their DenseASPP network to gather multi-scale features with a large range of receptive fields. Lin et al. [47] proposed ZigZagNet, which fused multi-resolution features by exchanging information in a zig-zag fashion between the encoder-decoder architecture. Wang et al. [48] proposed Deeply-Fused Nets that applies fusion of intermediate resolutions allowing varying receptive fields with different sizes. Additionally, the authors used the same-sized receptive field derived from two other base networks to capture different characteristics in the extracted features. Deep fusion was further studied in [44], [49], [50].

D. Our approach

To address the challenges of the existing approaches, we introduce a DSDF block that takes two different scale features as input. While propagating information flow in the same resolution, the DSDF block also performs a cross resolution fusion. This establishes a dual-scale fusion of features that inherit both high- and low-resolution feature representations. An

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3138024, IEEE Journal of Biomedical and Health Informatics

encoder network is used to feed the feature representations to the MSRF sub-network that consists of multiple DSDF blocks, thereby performing multi-scale feature exchange. Later, decoder layers with skip-connections from our sub-network and a triple attention mechanism are used to process our fused feature maps together with the shape stream. It is to be noted that the fusion strategy is interchangeable, i.e., low-to-high resolution and vice-versa.

III. The MSRF-Net architecture

Figure 2(a) represents the MSRF-Net that consists of an encoder block, the MSRF sub-network, a shape stream block, and a decoder block. The encoder block consists of squeeze and excitation modules, and the MSRF sub-network is used to process low-level feature maps extracted at each resolution scale of the encoder. The MSRF sub-network incorporates several DSDF blocks. A gated shape stream is applied after the MSRF sub-network, and decoders consisting of triple attention blocks are used in the proposed architecture. A triple attention block has the advantage of using spatial and channel-wise attention along with spatially gated attention, where irrelevant features from MSRF sub-network are pruned. Below, we briefly describe each component of our MSRF-Net.

A. Encoder

The encoder blocks (E1–E4) in Figure 2(a) are comprised of two consecutive convolutions followed by a squeeze and excitation module. The S&E block in the network increases the network's representative power by computing the interdependencies between channels. During the squeezing step, global average pooling is used to aggregate feature maps across the channel's spatial dimensions. In the excitation step, a collection of per-channel weights are produced to capture channel-wise dependencies [29]. At each encoder stage, max pooling with the stride of two is used for downscaling the resolution, and drop out is utilized for the model regularization.

B. The DSDF block and MSRF sub-network

Maintaining the resolution throughout the feature encoding process can help the target images become more semantically richer and spatially accurate. The DSDF block helps to exchange information between scales, preserve low-level features, and improves information flow while maintaining resolution. The block has two parallel streams for two different resolution scales (Figure 1(a)). If we let a 3×3 convolution followed by a LeakyRelu activation be represented by the operation $\mathrm{CLR}(\cdot)$, then each stream has a densely connected residual block with five CLR operations in series. The output feature map $M_{d,h}$ of the d-th CLR operation is computed from the high-resolution input X_h as follows:

$$M_{d,h} = \mathrm{CLR}(M_{d-1,h} \oplus M_{d-1,l} \oplus M_{d-2,h} \oplus \cdots \oplus M_{0,h}) \quad (1)$$

Here, $\oplus$ is the concatenation operation, and h represents CLR operation is on the higher resolution stream of the DSDF block. $M_{d-1,l}$ is processed by a transposed convolutional layer with a 3×3 kernel size and stride of 2 before being concatenated. Similarly, for lower resolution stream the output of the d-th CLR operation is denoted by $M_{d,l}$ and represented as:

$$M_{d,l} = \mathrm{CLR}(M_{d-1,l} \oplus M_{d-1,h} \oplus M_{d-2,l} \oplus \cdots \oplus M_{0,l}) \quad (2)$$

Here, $M_{d-1,h}$ is processed by a convolutional layer with kernel size of 3×3 and stride of 2 before being concatenated. In Equation 1 and Equation 2, d ranges from $1 \leq d \leq 5$. Initially, X_h (or $M_{0,h}$) and X_l (or $M_{0,l}$) are the higher and lower resolution stream input, respectively. The output of each CLR has k output channels denoting the growth factor, which regulates the amount of new features the layer can extract and propagate further in the network. Since the growth factor varies for each scale, we only use two scales at once in the DSDF to reduce the model's computational complexity for making the training feasible. Further, local residual learning is used to improve information flow, and residual scaling is used to prevent instability [51], [52]. Scaling factor $0 \leq w \leq 1$ can be used for residual scaling. The final output of the DSDF block can be written as (see Figure 1(a)):

$$X_r = w \times M_{5,r} + X_r, \quad (3)$$

where $r \in [h, l]$ is the resolution with h indicating high-resolution representation and l for low resolution representation.

Next, we present an MSRF sub-network that comprises of several DSDF blocks to achieve a global multi-scale context using the dual-scale fusion mechanism. As shown in [40], our approach has a contiguous memory mechanism that allows retaining multi-scale feature representations since the inputs of each DSDF is passed to each subsequent DSDF blocks in the same resolution stream.

Algorithm 1 MSRF sub-network

Information exchange across all scales in MSRF Sub-network
N is no. of DSDF layers ($N = 6$ in Figure 1(b))
$H \leftarrow X_{\hat{h},1}, X_{\hat{h},3}, X_{\hat{h}+1,1}, X_{\hat{h}+1,3}, \ldots$ (High-res. input)
$L \leftarrow X_{\hat{l},2}, X_{\hat{l},4}, X_{\hat{l}+1,2}, X_{\hat{l}+1,4}, \ldots$ (Low-res. input)
$p \in \{1, 3\}$ and $q \in \{2, 4\}$ are scale pairs
$X_{\hat{h}+1,p}, X_{\hat{l}+1,q} = \mathrm{DSDF}(X_{\hat{h},p}, X_{\hat{l},q})$
Update: $\tilde{X}_{\hat{h},p} = X_{\hat{h}+1,p}$, $\tilde{X}_{\hat{l},q} = X_{\hat{l}+1,q}$
for $2 \leq \mathrm{L} \leq N - 3$ **do**
    $X_{\hat{h}+1,p}, X_{\hat{l}+1,q} = \mathrm{DSDF}(X_{\hat{h},p}, X_{\hat{l},q})$
    Update: $\tilde{X}_{\hat{h},p}, \tilde{X}_{\hat{l},q} = X_{\hat{h}+1,p}, \tilde{X}_{\hat{l}+1,q}$
    $X_{\hat{l}+1,2}, \tilde{X}_{\hat{h}+1,3} = \mathrm{DSDF}(X_{\hat{l},2}, X_{\hat{h},3})$
    Update: $\tilde{X}_{\hat{l},2}, \tilde{X}_{\hat{h},3} = X_{\hat{l}+1,2}, \tilde{X}_{\hat{h}+1,3}$
end for
$X_{\hat{h}+1,p}, X_{\hat{l}+1,q} = \mathrm{DSDF}(X_{\hat{h},p}, X_{\hat{l},q})$
Update: $\tilde{X}_{\hat{h},p} = w.X_{\hat{h}+1,p} + X_{0,p}$, $\tilde{X}_{\hat{l},q} = w.X_{\hat{l}+1,q} + X_{0,q}$

In Algorithm 1, we define inputs in the MSRF sub-network as the process of demarcating all the resolution scale pairs and feeding them in their respective DSDF blocks. For this, we start with the first layer with each layer consisting of four resolution scales with H and L representing a high-resolution and low-resolution set of features, respectively, and

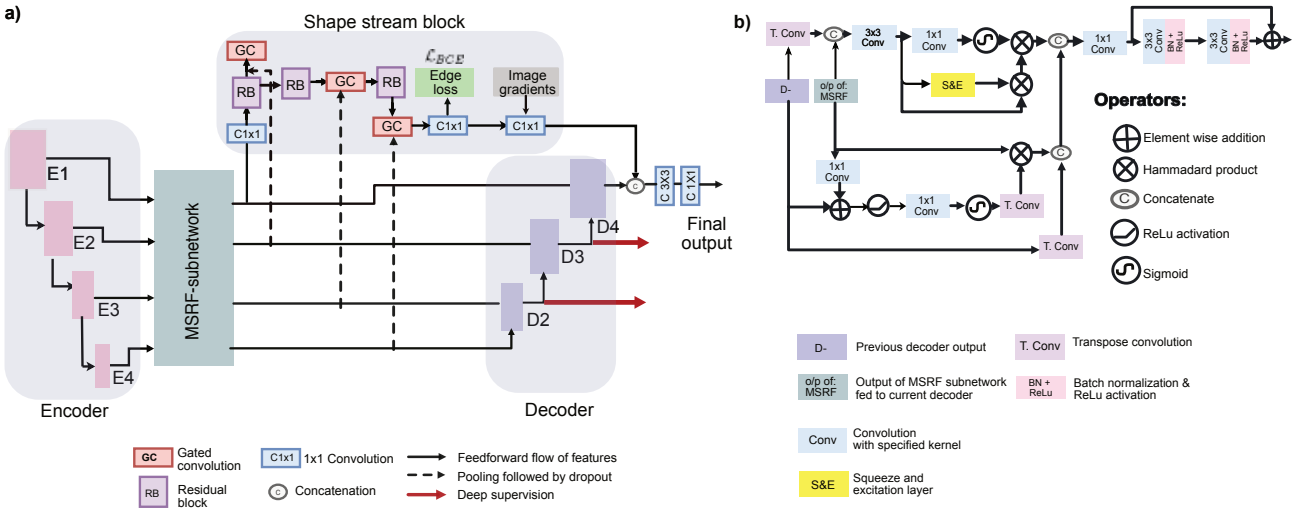

Fig. 2. **The proposed MSRF-Net architectures.** a) Overall block diagram of our network and b) overview of our decoder network.

each respective block is denoted by $\hat{h}$ and $\hat{l}$. The $\mathrm{DSDF}(\cdot)$ function performs feature fusion across scales in the DSDF block, where $\left(X_{\hat{h},p}, X_{\hat{l},q}\right)$ is jointly computed from the p and q scale pairs. Moreover, $\tilde{X}$ represents the feature exchange in the center DSDF. Already after the fourth layer of the MSRF sub-network, we effectively exchange features across all scales and attain global multi-scale fusion (refer to the red rectangular block in Figure 1(b)). We can observe that $X_{0,r}, \forall r \in \{1,2,3,4\}$ is able to transmit its features to all the parallel resolution representations through multiple DSDF blocks. Using this method, we exchange features globally in a more effective way, even when the number of resolution scales is greater than 4. Similar to the DSDF block, the output of the last layer of the sub-network is again scaled by w and added to the original input of the MSRF sub-network.

C. Shape stream

We have incorporated the gated shape stream [33] in MSRF-Net for the shape prediction (see the shape stream block in Figure 2(a)). The DSDF blocks can extract relevant high-level feature representations that include important information about shape and boundaries and can be used in the shape stream. Similar to [22], we define S_l as the shape stream feature maps where l is the number of layers and X is the output of the MSRF-sub-network. Bilinear interpolation is used so that X can match spatial dimensions of S_l, attention map α_l at the gated convolution is computed as:

$$\alpha_l = \sigma\left(\mathrm{Conv}_{1\times1}\left(S_l \oplus X\right)\right) \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid activation function. Finally, S_{l+1} is computed as $S_{l+1} = \mathrm{RB}(S_l \times \alpha)$, where RB represents residual block with two CLR operations followed by a skip-connection. The output of the shape stream is concatenated with the image gradients of the input image and merged with the original segmentation stream before the last CLR operation. This is done to increase the spatial accuracy of the segmentation map.

D. Decoder

The decoder block (D2–D4) has skip-connections from the MSRF sub-network and the previous decoder output (say D^-) except for D2, where the previous layer connection is the MSRF sub-network output of the E4 (Figure 2(a)). In the decoder block (Figure 2(b)), we use two attention mechanisms. The first attention mechanism applies channel and spatial attention, whereas the second attention uses a gating mechanism. We have used a S&E block for the calculation of channel-wise scale coefficients denoted by $X_{\alpha_{se}}$. Spatial attention is also calculated at the same top stream where the input channels C are reduced to 1 using 1×1 convolution. The sigmoid activation function $\sigma(\cdot)$ is used to scale the values between 0 and 1 to produce an activation map, which is stacked C times to give X_{α_s}. The output of the spatial and channel attention can be represented as:

$$D_{sc} = (X_{\alpha_s} + 1) \otimes X_{\alpha_{se}} \tag{5}$$

where $\otimes$ denotes the Hadamard product, and X_{α_s} is increased by a magnitude of 1 to amplify relevant features determined by the activation map. We also use the attention gated mechanism [19]. Let the features coming from MSRF-Net be X, and the output from the previous decoder block be D^-, then the attention coefficients can be calculated as:

$$D_{AG} = \Omega\left(\sigma\left(\Psi(\theta(X) + \phi(D^-))\right)\right) \tag{6}$$

where $\theta(\cdot)$ is the convolution operation with stride 2, kernel size 1, and G channel outputs; $\phi(\cdot)$ is a convolution operation with stride 1 and kernel size 1×1 applied to D^- giving the same G channels; and $\Psi(\cdot)$ is convolution function with 1×1 kernel size applied to a combined features from $\theta(\cdot)$ and $\phi(\cdot)$ making output channel equal to 1. Finally, $\sigma(\cdot)$ is applied to obtain the activation map on which transpose convolution operation $\Omega(\cdot)$ is applied. D_{AG} captures the contextual information and identifies the target regions and structures of the image. $\tilde{D}_{AG} = D_{AG} \otimes X$ allows the irrelevant features to be pruned and relevant target structure and regions to be propagated further. $\tilde{D}_{AG}$ is updated as:

$$\tilde{D}_{AG} = \tilde{D}_{AG} \oplus \Omega(D^-) \tag{7}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3138024, IEEE Journal of Biomedical and Health Informatics

Now, the final output of the *triple attention decoder block* (i.e., the combination of channel, spatial and gated spatial attention) is $D_\alpha = D_{sc} \oplus \tilde{D}_{AG}$, which is then followed by two CLR operations.

E. Loss computation

We have used binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$ as defined in Equation 8 where y is the ground truth value and $\hat{y}$ is the predicted value. We have also used dice loss $\mathcal{L}_{\text{DCS}}$, which is defined in Equation 9.

$$\mathcal{L}_{\text{BCE}} = (y-1)\log(1-\hat{y}) - y\log\hat{y} \tag{8}$$

$$\mathcal{L}_{\text{DCS}} = 1 - \frac{2y\hat{y}+1}{y+\hat{y}+1} \tag{9}$$

The sum of the two loss functions, $\mathcal{L}_{\text{comb}} = \lambda_1\mathcal{L}_{\text{BCE}} + \lambda_2\mathcal{L}_{\text{DCS}}$, is used for gradient minimization between the predicted maps and the labels, while only $\mathcal{L}_{\text{BCE}}$ has been used for shape stream. Here, we set the values of λ_1 and λ_2 to 1. For the latter loss, predicted edge maps and ground truth maps are used during computation. Deep supervision is also used to improve the flow of the gradients and regularization [53]. Thus, our final loss function can be represented as:

$$\mathcal{L}_{\text{MSRF}} = \alpha\mathcal{L}_{\text{comb}} + \beta_1\mathcal{L}_{\text{comb}}^{DS^0} + \beta_2\mathcal{L}_{\text{comb}}^{DS^1} + \gamma\mathcal{L}_{\text{BCE}}^{SS} \tag{10}$$

where $\mathcal{L}_{\text{comb}}^{DS^0}$ and $\mathcal{L}_{\text{comb}}^{DS^1}$ representing the two deep supervision outputs losses (see Figure 2(a)) and $\mathcal{L}^{SS}$ is the loss computed for the shape stream. Here, we set the values of $\alpha = 1$, $\beta_1 = 1$, $\beta_2 = 1$ and $\gamma = 1$ for our experiments.

IV. Experimental Setup

A. Dataset

To evaluate the effectiveness of the MSRF-Net, we have used four publicly available biomedical imaging datasets; Kvasir-SEG [12], CVC-ClinicDB [13], 2018 Data Science Bowl [2], and ISIC-2018 Challenge [14], [15]. The details about the datasets, number of training and testing samples used, and their availability is presented in Table I. All of these datasets consist of the images and their corresponding ground truth masks. An example of each dataset can be found in Figure 3. The chosen datasets are commonly used in biomedical image segmentation. The main reason for choosing diverse imaging modalities datasets is to evaluate the performance and robustness of the proposed method.

B. Evaluation metrics

Standard computer vision metrics for medical image segmentation such as Dice Coefficient (DSC), mean Intersection over Union (mIoU), recall, precision, and Frames Per Second (FPS) have been used for the evaluation of our experimental results. The standard deviations for DSC, mIoU, r and p are also provided. Additionally, we conduct a paired t-test between the DSC achieved by our proposed MSRF-Net and the DSC attained by other SOTA methods. The p-values of the paired t-tests are also reported.

TABLE I
THE MEDICAL DATASETS USED IN OUR EXPERIMENTS.

Dataset	Images	Input size	Train	Valid	Test
Kvasir-SEG [12]	1000	Variable	800	100	100
CVC-ClinicDB [13]	612	384×288	490	61	61
2018 Data Science Bowl [2]	670	256×256	536	67	67
ISIC-2018 Challenge [14], [15]	2596	384×512	2078	259	259

C. Implementation details

We have implemented the proposed architecture using the Keras framework [54] with TensorFlow [55] as backend. All experiments are conducted on an NVIDIA DGX-2 machine that uses NVIDIA V100 Tensor Core GPUs. The Adam optimizer was used with a learning rate of 0.0001, and a dropout regularization with $p = 0.2$ was used. The scaling factor for our DSDF and MSRF sub-network was set to 0.4 ($w = 0.4$). The growth factor k is set to 16, 32, and 64 for resolution scale pairs in the DSDF. For Kvasir-SEG and 2018 DSB, the images are resized to 256×256. ISIC-2018 images are resized to 384×512, and images from CVC-ClinicDB are resized to 384×288 resolution. We have used the batch size of 16 for Kvasir-SEG and 2018 DSB, eight for CVC-ClinicDB, and four for the ISIC-2018 Challenge dataset. We have empirically set the number of epochs for all datasets to 200 epochs. We have used 80% of the dataset for training, 10% for validation, and the remaining 10% for testing. Data augmentation techniques such as random cropping, random rotation, horizontal flipping, vertical flipping, and grid distortion were applied. It is to be noted that we have used open-source code provided by the respective authors for all the baseline comparisons. The proposed model is available at https://github.com/NoviceMAn-prog/MSRF-Net.

V. Results

A. SOTA method comparisons

In this section, we present the comparison of our MSRF-Net with other SOTA methods.

1) Comparison on Kvasir-SEG

Early detection of polyps, before they potentially change into colorectal cancer, can improve the survival rate [58]. Therefore, we have selected two popular colonoscopy datasets in our experiment. The first colonoscopy dataset is Kvasir-SEG. We report the quantitative evaluation of MSRF-Net in Table II and qualitative results in Figure 3. From the quantitative results, we can observe that our method outperforms all the other SOTA methods on all metrics. It achieves 1.39% improvement on DSC as compared to PraNet [34], 3.39% improvement on mIoU as compared Deeplabv3+ with Xception backbone [25]. Our method also achieves an improvement of 1.70% on precision and 1.04% on recall as compared to Deeplabv3+ with Xception backbone and U-Net [17], respectively. The network's ability to segment polyps can be observed from the ground truth comparison with the predicted mask. (Figure 3).

2) Comparison on CVC-ClinicDB

CVC-ClinicDB is the second colonoscopy dataset used in our experiment. The quantitative results from Table III show

TABLE II

RESULT COMPARISON ON THE KVASIR-SEG DATASET. WE HAVE NOT COMPUTED PAIRED T-TEST (P VALUES) FOR THE SAME NETWORK (MSRF-NET).

Method	DSC	mIoU	Recall	Precision	P-values	Parameters	FPS
U-Net [17]	0.8629 ± 0.2334	0.8176 ± 0.2465	0.9094 ± 0.2216	0.8901 ± 0.2313	1.559e-02	7.11M	41.04
U-Net++ [11]	0.7475 ± 0.2664	0.6313 ± 0.2788	0.6865 ± 0.2888	0.8871 ± 0.2689	4.363e-10	9.04M	30.67
ResUNet++ [20]	0.8189 ± 0.2652	0.7918 ± 0.2819	0.8372 ± 0.2751	0.9255 ± 0.2545	6.518e-06	4.07M	15.92
Deeplabv3+(Xception) [25]	0.8965 ± 0.2072	0.8575 ± 0.2290	0.8984 ± 0.2099	0.9496 ± 0.1801	3.660e-01	41.25M	49.11
Deeplabv3+ (Mobilenet) [25]	0.8656 ± 0.2032	0.8186 ± 0.2222	0.8808 ± 0.2105	0.9205 ± 0.1980	1.409e-04	2.14M	118.50
DoubleUNet [10]	0.8699 ± 0.1585	0.8166 ± 0.1933	0.9039 ± 0.1810	0.8942 ± 0.1586	8.109e-05	29.29M	7.46
HRNetV2-W18-Smallv2 [44]	0.8179 ± 0.2067	0.7470 ± 0.2622	0.8016 ± 0.2681	0.8696 ± 0.2494	4.844e-06	26.20M	52.68
HRNetV2-W48 [44]	0.8896 ± 0.1200	0.8262 ± 0.1856	0.8973 ± 0.1719	0.9056 ± 0.1492	5.935e-04	65.84M	29.79
ColonSegNet [5]	0.8203 ± 0.2295	0.7435 ± 0.2539	0.8124 ± 0.2494	0.8832 ± 0.1985	8.692e-07	5.01M	24.26
DDANet [56]	0.8915 ± 0.1880	0.8393 ± 0.2126	0.8927 ± 0.2093	0.9213 ± 0.1604	2.558e-02	6.83M	7.76
ResUNet++ + CRF$^{\diamond}$ [57]	0.7965 ± 0.2707	0.8250 ±0.2832	0.8119 ± 0.2803	0.8045 ± 0.2638	1.721e-06	4.02M	15.12
PraNet [34]	0.9078 ± 0.1543	0.8561 ± 0.1823	0.9034 ± 0.1719	0.9352 ± 0.1358	3.693e-01	32.54M	48.25
UACANet-S [35]	0.8800 ± 0.2042	0.8250 ± 0.2215	0.8701 ± 0.2136	0.9229 ± 0.1758	3.871e-04	26.90M	32.58
UACANet-L [35]	0.9014 ± 0.1878	0.8555 ± 0.2098	0.8897 ± 0.2148	0.9381 ± 0.1458	4.878e-01	69.15M	28.40
MSRF-Net (Ours)	**0.9217** ± 0.1685	**0.8914** ± 0.1938	**0.9198** ± 0.1919	**0.9666** ± 0.1379	-	18.38M	14.38

TABLE III

RESULT COMPARISON ON THE CVC-CLINICDB.

Method	DSC	mIoU	Recall	Precision	P-values	Parameters	FPS
U-Net [17]	0.9145 ± 0.1390	0.8654 ± 0.1514	0.9178 ± 0.1309	0.9381 ± 0.1594	1.000e+00*	7.11M	22.84
U-Net++ [11]	0.8453 ± 0.1516	0.7559 ± 0.1188	0.8917 ± 0.2594	0.8323 ± 0.2713	6.400e-04	9.04M	17.60
ResUNet++$^{\diamond}$ [20]	0.9075 ± 0.1455	0.8587 ± 0.1616	0.9156 ± 0.1372	0.9325 ± 0.1593	1.000e+00*	4.07M	15.71
Deeplabv3+ (Xception) [25]	0.8897 ± 0.1895	0.8706 ± 0.2036	0.9251 ± 0.1965	0.9366 ± 0.1621	6.723e-01	41.25M	29.08
Deeplabv3+ (Mobilenet) [25]	0.8985 ± 0.1385	0.8588 ± 0.1544	0.9160 ± 0.1562	0.9287 ± 0.1378	2.838e-01	2.14M	55.68
DoubleU-Net [10]	0.9272 ± 0.1761	0.8889 ± 0.1896	0.9395 ± 0.1800	**0.9592** ± 0.1609	1.000e+00*	29.29M	7.46
HRNetV2-W18-Smallv2 [44]	0.9073 ± 0.1158	0.8457 ± 0.1477	0.9137 ± 0.1293	0.9191 ± 0.1060	1.000e+00*	26.20M	57.47
HRNetV2-W48 [44]	0.9244 ± 0.1280	0.8747 ± 0.1332	0.9234 ± 0.1356	0.9296 ± 0.1318	1.000e+00*	65.84M	29.76
ColonSegNet [5]	0.9132 ± 0.1416	0.8600 ± 0.1585	0.9072 ± 0.1564	0.9292 ± 0.1393	1.000e+00*	5.01M	7.98
DDANet [56]	0.9233 ± 0.1339	0.8747 ± 0.1438	0.9271 ± 0.1285	0.9259 ± 0.1495	1.000e+00*	6.83M	58.15
ResUNet++ + CRF$^{\diamond}$ [57]	0.8815 ± 0.1507	0.8899 ± 0.1685	0.8970 ± 0.1581	0.8674 ± 0.1605	1.102e-01	4.02M	8.55
PraNet [34]	0.9072 ± 0.1695	0.8575 ± 0.1823	0.9227 ± 0.1657	0.9134 ± 0.1614	1.000e+00*	32.54M	47.92
UACANet-S [35]	0.9190 ± 0.1426	0.8700 ± 0.1556	0.9285 ± 0.1356	0.9201 ± 0.1574	1.000e+00*	26.90M	31.79
UACANet-L [35]	0.9098 ± 0.1829	0.8649 ± 0.1872	0.9174 ± 0.1772	0.9114 ± 0.1950	1.000e+00*	69.15M	32.81
MSRF-Net (Ours)	**0.9420** ± 0.0804	**0.9043** ± 0.1009	**0.9567** ± 0.0620	0.9427 ± 0.0994	-	18.38M	12.50

** large improvements were still observed for some samples with consistent Dice scores (also see standard deviation) for MSRF-Net*

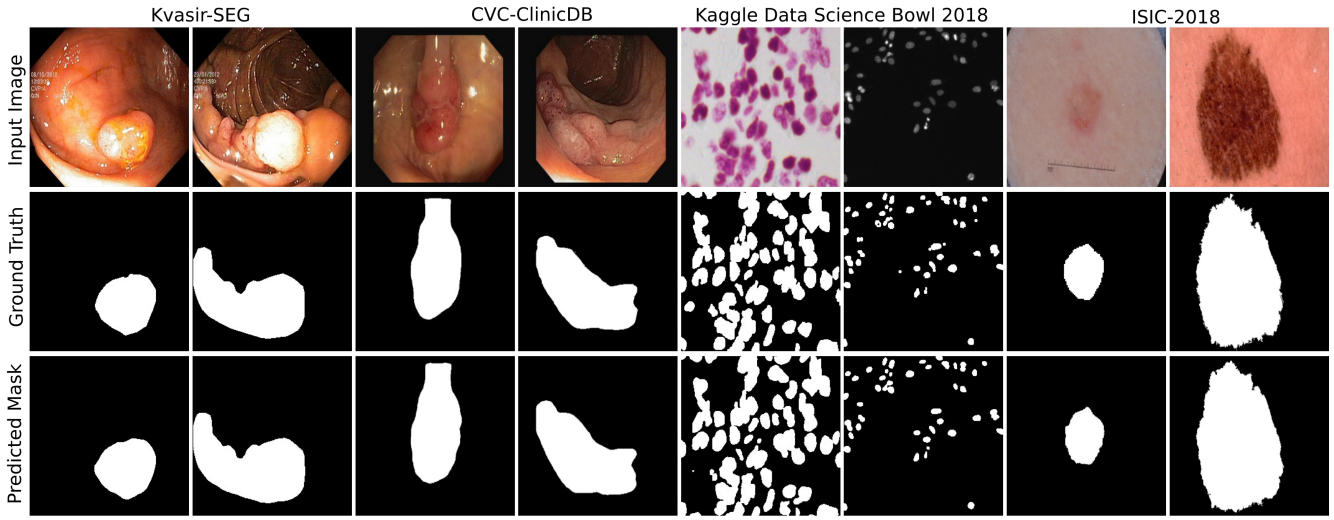

Fig. 3. The Figure shows qualitative results of the MSRF-Net on four biomedical imaging datasets.

that our approach surpasses other SOTA methods and achieves a DSC of 0.9420 ± 0.0804, which is 1.76% improvement in DSC over the best-performing HRNetV2-W48 [44]. We report a mIoU of 0.9043 ± 0.1009 and a recall of 0.9567 ± 0.0620, which is 1.44% improvement in mIoU and 2.82% improvement in recall over SOTA combination of ResUNet++ and conditional random field [57] and UACANet-S [35], respectively. Additionally, MSRF-Net achieves a precision of 0.9427, which is competitive with the best performing DoubleUNet [10]. Our method produces prediction masks

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3138024, IEEE Journal of Biomedical and Health Informatics

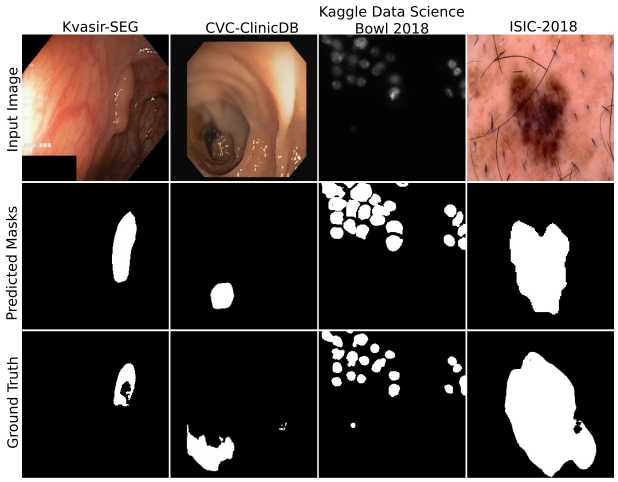

Fig. 4. Qualitative results of the MSRF-Net for sub-optimal cases.

with nearly the same boundaries and shape of the polyp as compared to the ground truth masks (Figure 3).

3) Comparison on 2018 Data Science Bowl

Finding nuclei in a cell from a large variety of microscopy images is a challenging problem. We experiment with the 2018 Data Science Bowl challenge dataset. Table IV shows the comparison of the result of the proposed MSRF-Net with some of the presented approaches. MSRF-Net obtains a DSC of 0.9224 ± 0.0538, mIoU of 0.8534 ± 0.0870, recall of 0.9402 ± 0.0734 and precision of 0.9022 ± 0.0601 which outperforms the best performing ColonSegNet [5] in most metrics (see Table IV). From the qualitative results in Figure 3, we observe that the predicted masks are visually similar to the ground truth masks.

4) Comparison on ISIC-2018 Skin Lesion Segmentation challenge

An automatic diagnosis tool for skin lesions can help in accurate melanoma detection, which is also a commonly occurring cancer and can save life up to 99% [59] of cases. The quantitative results for the ISIC-2018 challenge are shown in Table V. Our method achieved a DSC of 0.8824 ± 0.1602, mIoU of 0.8373 ± 0.1818, recall of 0.8893 ± 0.1889, and precision of 0.9348 ± 0.1488. We can observe an improvement of 0.43% and 1.37% over Deeplabv3+ with the Mobilenet backbone [24] in DSC and mIoU, respectively. We also observe a 0.63% improvement in recall over Deeplabv3+(MobileNet) [24]. Our results are comparable with DoubleU-Net [10] which reports the highest DSC of 0.8938. A higher recall shows that our method is more medically relevant, which is considered as the major strength of our architecture [60]. From Figure 3, we can observe that our method can segment skin lesions of varying sizes accurately.

B. Generalization study

To ensure generalizability, we have trained our model and other SOTA methods on one dataset and then experimented on a new unseen dataset which comes from a different institution, consisting of different cohort populations and acquired using different imaging protocols. To this end, we have used the Kvasir-SEG collected in Vestre Viken Health Trust in Norway for training and tested our trained model on the CVC-ClinicDB, which was captured in Hospital Clinic in Barcelona, Spain. Similarly, we conducted this study on an opposite set-up as well, i.e., training on CVC-ClinicDB and testing on Kvasir-SEG.

1) Generalizability results on CVC-ClinicDB

Table VI shows the generalizability results of the MSRF-Net model trained on Kvasir-SEG and tested on CVC-ClinicDB. Despite using two datasets acquired using two different imaging protocols, MSRF-Net obtained an acceptable DSC of 0.7921 ± 0.2564, mIoU of 0.6498 ± 0.2729, recall of 0.9001 ± 0.2980, and precision of 0.7000 ± 0.1572. We observe that our MSRF-Net performs better than other SOTA methods in terms of DSC. HRNetV2-W48 [44] obtained a competitive DSC of 0.7901.

2) Generalizability results on Kvasir-SEG

Similarly, we present the results of the models trained on CVC-ClinicDB and tested on Kvasir-SEG in Table VII. We report that our model achieves a DSC of 0.7575 ± 0.2643, mIoU of 0.6337 ± 0.2815, recall of 0.7197 ± 0.2775 and precision of 0.8414 ± 0.2731, which outperforms other SOTA methods in DSC and mIoU. The second best performing method is PraNet [34] with DSC of 0.7293 ± 0.3004, and mIoU of 0.6262 ± 0.3128. Our method outperforms PraNet [34] by 2.82% in DSC and 0.75% in mIoU but PraNet [34] records the highest recall of 0.8007.

C. Ablation study

We have conducted an extensive ablation study on the Kvasir-SEG. For this, we ablated the impact of the MSRF sub-network, scaling mechanism used in the network, the effect of the number of DSDF blocks used, the impact of the MSRF sub-network on shape prediction in the shape stream in Section V-C, the effect observed when shape stream and deep supervision is removed from the architecture of MSRF-Net.

Table VIII shows the quantitative results of our ablation study. Initially, we removed the MSRF sub-network which resulted in the least DSC of 0.8771. The addition of a subset of the original MSRF sub-network (The red dotted region in Figure 1(b)) raises the DSC to 0.8986. Further, we removed the DSDF with second and third scale inputs (also see Figure 1(b), where middle DSDF blocks represent them, i.e., layer three and layer five are removed) from the original MSRF sub-network to achieve a DSC of 0.9013. To further investigate the contribution of our MSRF sub-network, we remove the shape stream to achieve a DSC of 0.9194 which is comparable to highest 0.9217 DSC reported by the original MSRF-Net configuration. We disable the triple attention mechanism in the decoder block to get a DSC 0.9067 ± 0.1834. Our ablation on further removing deep supervision resulted in a lower DSC of 0.8988. We also report the effect of using a combination of dice loss and binary cross entropy loss in $\mathcal{L}_{\text{comb}}$ used in Equation 10 to supervise MSRF-Net during training. First, we set $\mathcal{L}_{\text{comb}} = \mathcal{L}_{\text{BCE}}$ and secure a DSC of 0.9059 which was followed by setting $\mathcal{L}_{\text{comb}} = \mathcal{L}_{\text{DCS}}$ which scored a DSC of 0.8861. A similar trend was observed for other metrics.

VI. Discussion

Multi-scale fusion methodologies have been studied previously, however, there are some disadvantages. For example,

TABLE IV
RESULTS ON THE 2018 DATA SCIENCE BOWL

Method	DSC	mIoU	Recall	Precision	P-values	Parameters	FPS
U-Net [17]	0.9080 ± 0.0638	0.8314 ± 0.1019	0.9029 ± 0.0981	0.9130 ± 0.0719	1.308e-03	7.11M	25.02
U-Net++ [11]	0.7705 ± 0.3010	0.5265 ± 0.3078	0.7159 ± 0.3171	0.6657 ± 0.2745	4.832e-04	9.04M	21.86
ResUNet++ [20]	0.9098 ± 0.0797	0.8370 ± 0.1154	0.9169 ± 0.0947	0.9057 ± 0.0853	2.420e-01	4.07M	19.81
Deeplabv3+ (Xception) [25]	0.8857 ± 0.1674	0.8367 ± 0.1702	0.9141 ± 0.1751	0.9081 ± 0.1689	4.641e-01	41.25M	16.20
Deeplabv3+ (Mobilenet) [25]	0.8239 ± 0.1613	0.7402 ± 0.1618	0.8896 ± 0.1720	0.8151 ± 0.1657	1.044e-06	2.14M	23.70
DoubleUNet [10]	0.9109 ± 0.0876	0.8429 ± 0.1109	0.9278 ± 0.0962	0.9020 ± 0.0997	8.929e-02	29.29M	7.47
HRNetV2-W18-Smallv2 [44]	0.8495 ± 0.3267	0.7585 ± 0.1521	0.8640 ± 0.1659	0.8398 ± 0.1602	4.978e-04	26.20M	58.03
HRNetV2-W48 [44]	0.8488 ± 0.1470	0.7588 ± 0.1499	0.8359 ± 0.1618	0.8913 ± 0.0550	8.460e-05	65.84M	29.41
ColonSegNet [5]	0.9197 ± 0.0605	0.8466 ± 0.0953	0.9153 ± 0.0917	**0.9312** ± 0.0532	6.433e-01	5.01M	16.56
DDANet [56]	0.9182 ± 0.0684	0.8452 ± 0.1037	0.9139 ± 0.0964	0.9289 ± 0.0575	5.922e-01	6.83M	19.02
ResUNet++ + CRF [57]	0.7806 ± 0.2223	0.7322 ± 0.2386	0.7534 ± 0.2558	0.6308 ± 0.1752	6.971e-06	4.02M	72.78
PraNet [34]	0.8751 ± 0.0871	0.7868 ± 0.1169	0.9182 ± 0.0736	0.8438 ± 0.1138	1.553e-07	32.54M	11.88
UACANet-S [35]	0.8687 ± 0.0913	0.7774 ± 0.1208	0.9092 ± 0.0960	0.8385 ± 0.1115	2.940e-08	26.90M	28.08
UACANet-L [35]	0.8688 ± 0.0999	0.7791 ± 0.1283	0.9061 ± 0.1057	0.8414 ± 0.1141	8.975e-07	69.15M	32.26
MSRF-Net (Ours)	**0.9224** ± 0.0538	**0.8534** ± 0.0870	**0.9402** ± 0.0734	0.9022 ± 0.0601	-	18.38M	6.84

TABLE V
RESULTS ON THE ISIC-2018 SKIN LESION SEGMENTATION CHALLENGE

Method	DSC	mIoU	Recall	Precision	P-values	Parameters	FPS
U-Net [17]	0.8554 ± 0.1848	0.7847 ± 0.2094	0.8204 ± 0.2186	**0.9474** ± 0.1296	1.315e-02	7.11M	79.43
U-Net++ [11]	0.8094 ± 0.2261	0.7288 ± 0.2452	0.7866 ± 0.2369	0.9084 ± 0.2222	6.336e-08	9.04M	60.44
ResUNet++ [20]	0.8557 ± 0.2014	0.8135 ± 0.2210	0.8801 ± 0.2320	0.8676 ± 0.1562	1.813e-02	4.07M	40.93
Deeplabv3+ (Xception) [25]	0.8772 ± 0.1465	0.8128 ± 0.1806	0.8681 ± 0.1792	0.9272 ± 0.13602	4.314e-02	41.25M	43.53
Deeplabv3+ (Mobilenet) [25]	0.8781 ± 0.1371	0.8236 ± 0.1711	0.8830 ± 0.1725	0.9244 ± 0.1317	9.503e-02	2.14M	61.10
HRNetV2-W18-Smallv2 [44]	0.8561 ± 0.3696	0.7821 ± 0.2091	0.8556 ± 0.2029	0.8974 ± 0.1862	8.780e-03	26.20M	57.47
HRNetV2-W48 [44]	0.8667 ± 0.2453	0.8109 ± 0.2630	0.8584 ± 0.2936	0.9155 ± 0.2755	4.270e-02	65.84M	28.54
DoubleU-Net [10]	**0.8938** ± 0.1362	0.8212 ± 0.1659	0.8780 ± 0.1573	0.9459 ± 0.1353	1.000e+00	29.29M	7.46
ResUNet++ + CRF [57]	0.8688 ± 0.1719	0.8209 ± 0.1971	0.8826 ± 0.2063	0.8736 ± 0.1540	1.813e-02	4.02M	79.11
MSRF-Net (Ours)	0.8824 ± 0.1602	**0.8373** ± 0.1818	**0.8893** ± 0.1889	0.9348 ± 0.1488	-	18.38M	16.10

TABLE VI
GENERALIZABILITY RESULTS OF THE MODELS TRAINED ON KVASIR-SEG AND TESTED ON CVC-CLINICDB

Method	DSC	mIoU	Recall	Precision	Parameters	FPS
U-Net [17]	0.7172 ± 0.2911	0.6133 ± 0.2870	0.7255 ± 0.3246	**0.7986** ± 0.2775	7.11M	24.15
U-Net++ [11]	0.4265 ± 0.3922	0.3345 ± 0.3518	0.3939 ± 0.4480	0.6894 ± 0.4111	9.04M	21.18
ResUNet++ [20]	0.5560 ± 0.3436	0.4542 ± 0.3174	0.5795 ± 0.3896	0.6775 ± 0.3579	4.07M	8.84
Deeplabv3+ (Xception) [25]	0.6509 ± 0.3172	0.5385 ± 0.3174	0.6251 ± 0.3621	0.7947 ± 0.3175	41.25M	27.24
Deeplabv3+ (Mobilenet) [25]	0.6303 ± 0.2740	0.4825 ± 0.2716	0.5957 ± 0.3391	0.7173 ± 0.2730	2.10M	73.64
HRNetV2-W18-Smallv2 [44]	0.6428 ± 0.3003	0.5513 ± 0.3213	0.6811 ± 0.3753	0.7253 ± 0.3191	26.20M	57.32
HRNetV2-W48 [44]	0.7901 ± 0.2280	**0.6953** ± 0.2455	0.8796 ± 0.1746	0.7694 ± 0.2642	65.84M	29.24
ColonSegNet [5]	0.6895 ± 0.2716	0.5813 ± 0.2754	0.7862 ± 0.2965	0.7177 ± 0.2897	5.01M	17.01
ResUNet++ + CRF [57]	0.6502 ± 0.3381	0.7417 ± 0.3147	0.7047 ± 0.3631	0.6277 ± 0.3392	4.02M	80.31
PraNet [34]	0.7225 ± 0.2931	0.6328 ± 0.3028	0.7531 ± 0.3390	0.7888 ± 0.2953	32.54M	30.91
UACANet-S [35]	0.5683 ± 0.3799	0.4907 ± 0.3631	0.5792 ± 0.4101	0.7095 ± 0.3833	26.90M	32.61
UACANet-L [35]	0.5589 ± 0.3899	0.4849 ± 0.3689	0.5800 ± 0.4276	0.6775 ± 0.3906	69.15M	32.20
MSRF-Net (Ours)	**0.7921** ± 0.2564	0.6498 ± 0.2729	**0.9001** ± 0.2980	0.7000 ± 0.1572	18.38M	10.94

TABLE VII
GENERALIZABILITY RESULTS OF THE MODELS TRAINED ON CVC-CLINICDB AND TESTED ON KVASIR-SEG

Method	DSC	mIoU	Recall	Precision	Parameters	FPS
U-Net [17]	0.6222 ± 0.2595	0.4588 ± 0.2609	0.5129 ± 0.1766	0.8133 ± 0.3059	7.11M	35.18
U-Net++ [11]	0.5926 ± 0.2363	0.4564 ± 0.2321	0.7352 ± 0.2368	0.5462 ± 0.2902	9.04M	16.85
ResUNet++ [20]	0.5147 ± 0.3138	0.4082 ± 0.2940	0.7181 ± 0.3331	0.4860 ± 0.3530	4.07M	25.86
HRNetV2-W18-Smallv2 [44]	0.7012 ± 0.0680	0.6009 ± 0.2889	0.7184 ± 0.3021	0.7666 ± 0.2863	26.20M	57.89
HRNetV2-W48 [44]	0.7404 ± 0.1489	0.6233 ± 0.2834	0.7293 ± 0.2830	**0.8511** ± 0.2563	65.84M	29.80
Deeplabv3+ (Xception) [25]	0.6746 ± 0.2746	0.5327 ± 0.2788	0.7757 ± 0.2967	0.6296 ± 0.2699	41.25M	39.62
Deeplabv3+ (Mobilenet) [25]	0.6474 ± 0.2634	0.5098 ± 0.2653	0.6632 ± 0.2611	0.6878 ± 0.2916	2.10M	78.63
ColonSegNet [5]	0.6324 ± 0.2772	0.5183 ± 0.2830	0.6112 ± 0.3058	0.7897 ± 0.2731	5.01M	7.92
ResUNet++ + CRF [57]	0.4200 ± 0.3150	0.6096 ± 0.2770	0.3782 ± 0.3132	0.6711 ± 0.4004	4.02M	46.98
PraNet [34]	0.7293 ± 0.3004	0.6262 ± 0.3128	**0.8007** ± 0.2675	0.7623 ± 0.3243	32.54M	49.61
UACANet-S [35]	0.6945 ± 0.2634	0.5894 ± 0.2911	0.7692 ± 0.2424	0.7377 ± 0.3119	26.90M	32.89
UACANet-L [35]	0.7312 ± 0.2724	0.6383 ± 0.2961	0.7417 ± 0.2803	0.8314 ± 0.2627	69.15M	32.73
MSRF-Net (Ours)	**0.7575** ± 0.2643	**0.6337** ± 0.2815	0.7197 ± 0.2775	0.8414 ± 0.2731	18.38M	16.24

U-Net [17] uses skip-connections for feature fusion, but the resulting combination of features suffer from semantic gap since it combines low level features of the encoder and high level features of the decoder. Similarly, U-Net++ [11] performs low to high feature fusion to overcome this problem, but high to low feature fusion remains lacking. Pyramid features are fused in Deeplabv3+ [25] while without maintaining the high resolution representations. Similar to our approach, HR-Net builds upon the multi-scale feature fusion process by adding repeated feature fusion while keeping high resolution representation, however, their fusion modules consist of a larger number of trainable parameters and informative low level features are also lost during the segmentation process [61]. The disadvantages stated above can result in Deeplabv3+ [25] and HRNetV2 [62]

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3138024, IEEE Journal of Biomedical and Health Informatics

TABLE VIII
ABLATION STUDY OF MSRF-NET ON THE KVASIR-SEG

Experiment description	DSC	mIoU	Recall	Precision	Parameters	FPS
MSRF-Net (ours)	**0.9217** ± 0.1685	**0.8914** ± 0.1938	0.9198 ± 0.1919	**0.9666** ± 0.1379	18.38M	16.24
Without sub-network	0.8771 ± 0.2062	0.8103 ± 0.2400	0.8911 ± 0.1742	0.8993 ± 0.2172	**2.66M**	**32.64**
Sub-network without scaling	0.9137 ± 0.1772	0.8898 ± 0.2013	0.9625 ± 0.1932	0.9218 ± 0.1484	18.38M	8.92
Sub-network without DSDF (across 2,3 scale)	0.9013 ± 0.2111	0.8782 ± 0.2329	0.9460 ± 0.2098	0.9246 ± 0.2068	17.24M	17.28
Subset of the sub-network	0.8986 ± 0.1997	0.8570 ± 0.1262	0.9228 ± 0.2879	0.9232 ± 0.2666	8.67M	13.10
Without deep supervision	0.8988 ± 0.2060	0.8449 ± 0.2313	0.9053 ± 0.1795	0.9267 ± 0.2152	18.38M	16.58
Without decoder block	0.9067 ± 0.1834	0.8691 ± 0.2080	0.9143 ± 0.1875	0.9461 ± 0.1807	17.41M	16.90
Without shape stream	0.9194 ± 0.1779	0.8907 ± 0.1984	**0.9700** ± 0.1976	0.9159 ± 0.1348	18.37M	17.19
MSRF-Net with $\mathcal{L}_{\text{comb}} = \mathcal{L}_{\text{DCS}}$	0.8861 ± 0.2192	0.8446 ± 0.2389	0.9139 ± 0.2078	0.9176 ± 0.2003	18.38M	16.18
MSRF-Net with $\mathcal{L}_{\text{comb}} = \mathcal{L}_{\text{BCE}}$	0.9059 ± 0.1859	0.8677 ± 0.2116	0.9446 ± 0.1938	0.9143 ± 0.17000	18.38M	16.11

to perform considerably worse on the 2018 Data Science Bowl challenge where finer segmentation maps were required for a high DSC score. The results on Kavisir-SEG, CVC-ClinicDB, and ISIC-2018 also show similar performance gaps between proposed and other multi-scale fusion methods (see Table I - IV).

The proposed MSRF-Net uses DSDF blocks (arranged as described in Algorithm 1) to attain global multi-scale fusion while increasing the frequency of multi-scale fusion operations and reporting a lower computational complexity as compared to HRNetV2 [62]. The DSDF blocks itself allow effective feature fusion between high- and low resolution scales by continuous feature exchange across different scales. Additionally, its residual structure permits the relevant high- and low-level features to be deftly propagated enabling the proposed MSRF-Net to effectively capture the variability in size, shape and structure of the region of interest. We can observe that the residual densely connected nature of the DSDF blocks and its subsequent arrangement allows our proposed MSRF-Net to achieve highest DSC of 0.9217 and mIoU of 0.8914, on the Kvasir-SEG (see Table II). Similarly, we report the highest values for DSC, mIoU and recall of 0.9420, 0.9043 and 0.9567, respectively, on CVC-ClinicDB (see Table III). The ability of our MSRF-Net to recognize smaller and finer cell structures in 2018 Data Science Bowl is evident in Table IV, where we report the best DSC of 0.9224. Additionally, we report best mIoU and recall on the ISIC-2018 skin lesion dataset. Our result is competitive to DoubleUNet in terms of DSC. We present training loss of MSRF-Net (Kvasir-SEG) with respect to the number of epochs elapsed in Figure 6(a). We can see that the model starts converging from epoch number 75 steadily.

In practical clinical environments, the performance of deep learning based segmentation methods decreases due to differences in the imaging protocols and patient variability. The models which are able to generalize across multi-center dataset are more desirable in a clinical setting [63]. MSRF-Net achieves the highest DSC of 0.7921 when trained on Kvasir-SEG and tested on CVC-ClinicDB (see Table VI). Similarly, MSRF-Net achieves highest DSC of 0.7575 and mIoU of and 0.6337, when trained on CVC-ClinicDB and tested on Kvasir-SEG (see Table VII). HRNetV2-W48 [62] was competitive to our method. The above results suggest that our proposed MSRFNet is more generalizable. This can be evidently due to our multi-scale fusion that exploits the feature at different scales, preserving some class representative features.

We performed an ablation study (see Table VIII) to demonstrate that the combination of relevant high- and low-level multi-scale features obtained by the MSRF sub-network is instrumental in recognizing the shape or boundaries of the target object that can boost the segmentation performance. To verify the contribution of the MSRF sub-network, we disable the entire MSRF sub-network from the full network while keeping each component of the network intact and train the model. Table VIII shows that when the MSRF sub-network is removed from the proposed MSRF-Net, the DSC drops by 4.46%. This performance degradation illustrates that each MSRF sub-network contributes to the network. The combination of high- and low-level resolution feature representations of varying receptive fields extracted from the MSRF sub-network contribute significantly towards improving the model's performance. We also ablated if multi-scale fusion was suitable for the entire network. Sub-Network without DSDF refers to the removal of DSDF with 2nd and 3rd scale inputs (also see Figure 1(b), where middle DSDF blocks represent them, i.e., layer three and layer five are removed). Table VIII shows the result when global multi-scale fusion is absent from the network. As a result, we observe a 2.04% performance drop in DSC. Therefore, it is noticeable that the multi-scale fusion used in the MSRF sub-network improves performance. To study the impact of the number of DSDF blocks on the segmentation performance, we reduced the number of DSDF layers from six (ours) to three, i.e., only red rectangular block in Figure 1(b) is used. Even though this enables us to exchange global multi-scale feature representations, our results in Table VIII show that reducing the number of DSDF blocks decreases the DSC by 2.31% .

The sub-network without scaling in Table VIII demonstrates the influence of scaling factor w in the network (see Equation 3). For this experiment, we did not scale the output of DSDF by a constant while adding to the block's input. Drop of 0.80% in DSC was observed when the features were not scaled. Furthermore, our empirical experiments (see Figure 6(b) using different scaling values of w introduced in Equation 3) demonstrate our optimal choice of w to be 0.4.

We design a variant model where, the MSRF sub-network is placed after the shape-stream in the MSRF-Net. Here, we keep the number of parameters same for both the models (i.e., MSRF-Net and variant model) to analyze the impact of MSRF sub-network on the shape stream. The qualitative results (see Figure 5) show that the MSRF-Net can define more precise and more spatially accurate boundaries than the variant model. The

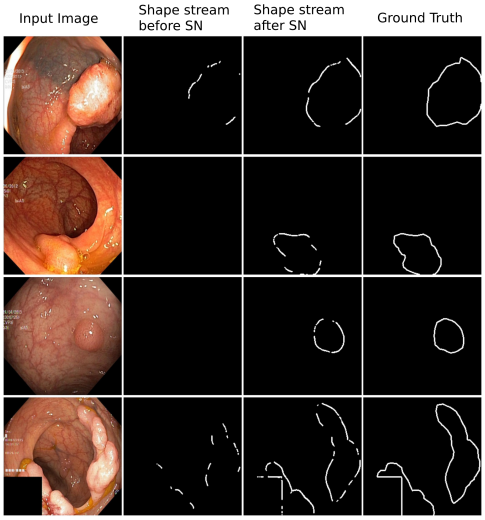

Fig. 5. Qualitative results showing polyp contours when shape stream is used before MSRF sub-network and after the MSRF sub-network in the MSRF-Net

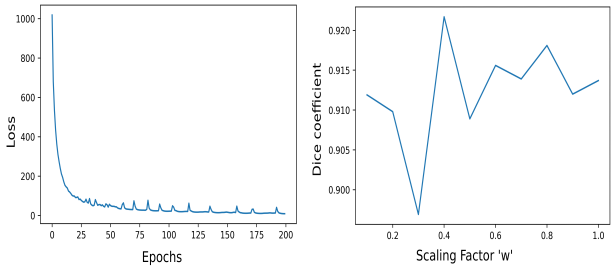

Fig. 6. a) Plot of training loss versus number of epochs and b) Variation in DSC with respect to hyper parameter $\boldsymbol{w}$.

variant model fails to recognize the boundaries of the target structure as it is deprived of the multi-scale features extracted by the MSRF sub-network. This validates our choice of putting the MSRF sub-network before the shape stream block. Only a minor drop of 0.23% in DSC is seen when no shape stream is applied and it still outperforms most SOTA methods.

We also investigated the impact of our triple attention block by disabling the mechanism prior to training the MSRF-Net. We disable deep supervision in another experiment, while training MSRF-Net. Both of these experiments showed performance drop compared to our proposed MSRF-Net (1.50% drop in former and 2.29% drop in latter on DSC metric). We also evaluate the impact of the combination of $\mathcal{L}_{\text{BCE}}$ and $\mathcal{L}_{\text{DCS}}$ used in $\mathcal{L}_{\text{comb}}$ (see Section III-E). For this, we trained the MSRF-Net with $\mathcal{L}_{\text{comb}} = \mathcal{L}_{\text{DCS}}$ and then with $\mathcal{L}_{\text{comb}} = \mathcal{L}_{\text{BCE}}$. When $\mathcal{L}_{\text{comb}} = \mathcal{L}_{\text{DCS}} + \mathcal{L}_{\text{BCE}}$, we obtained an increase of 3.56% in DSC, 4.68% in mIoU, 0.59% in recall and 4.90% in precision as compared to the $\mathcal{L}_{\text{comb}} = \mathcal{L}_{\text{DCS}}$ setting. Similar trend was observed when $\mathcal{L}_{\text{comb}}$ was equal to $\mathcal{L}_{\text{BCE}}$ (see Table VIII).

MSRF-Net clearly shows the strength of fusing low- and high-resolution features through DSDF blocks and MSRF sub-network. Alongside, complementary inclusion of scaling factor, deep supervision in the encoder block and triple attention in the decoder block showed further improvements. In Figure 4, we show the qualitative results for the sub-optimal cases. The qualitative results show poor performance for oblique samples in polyp datasets. Similarly, the model also failed for extremely low contrast images with 2018 DSB and scattered similar patches in ISIC 2018.

VII. Conclusion

In this paper, we proposed the MSRF-Net architecture for medical image segmentation that takes advantage of multi-scale resolution features passed through a sequence of DSDF blocks. Such densely connected residual blocks with dual-scale feature exchange enable efficient feature extraction with varying receptive fields. Additionally, we have also shown that the features from DSDF blocks are better suited to capture a target object's entire shape boundaries, even for objects with variable sizes. Our experiments revealed that MSRF-Net outperformed several SOTA methods on four independent biomedical datasets. Our investigation using cross-datasets testing to evaluate the generalizability of the MSRF-Net confirmed that our model can produce competitive results in such scenarios. We also identified some challenges of the proposed method, such as that the model fails when extremely low contrast images are part of the data. For future work, we plan to investigate the identified challenges further and adjust the design of the network to address the challenging cases.

Acknowledgment

D. Jha is funded by the PRIVATON project (#263248) which is funded by Research Council of Norway (RCN). S. Ali is supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC).The computations in this paper were performed on equipment provided by the Experimental Infrastructure for Exploration of Exascale Computing (eX³), which is financially supported by the Research Council of Norway under contract 270053.

References

[1] M. E. Celebi, N. Codella, and A. Halpern, "Dermoscopy image analysis: overview and future directions," *IEEE J. Biomed. Health Inform*, vol. 23, no. 2, pp. 474–478, 2019.

[2] J. C. Caicedo *et al.*, "Nucleus segmentation across imaging experiments: the 2018 data science bowl," *Nat. Meth.*, vol. 16, no. 12, pp. 1247–1253, 2019.

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

[4] M. Lux and M. Riegler, "Annotation of endoscopic videos on mobile devices: a bottom-up approach," in *Proceedings of the 4th ACM Multimedia Systems Conference*, 2013, pp. 141–145.

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

[6] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Imag. Anal.*, vol. 42, pp. 60–88, 2017.

[7] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Ann. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, 2017.

[8] T. Roß *et al.*, "Comparative validation of multi-instance instrument segmentation in endoscopy: results of the robust-mis 2019 challenge," *Med. Imag. Anal.*, p. 101920, 2020.

[9] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Dee. learn. da. label. medi. applicat.*, 2016, pp. 179–187.

[10] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation," in *Proc. of Internat. Sympo. Comp.-Bas. Med. Syst.*, 2020.

[11] 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.

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

[13] 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.

[14] N. C. Codella *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *Proc. of Internat. Sympo. on Biomed. Imag.*, 2018, pp. 168–172.

[15] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scienti. Da.*, vol. 5, p. 180161, 2018.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of Comput. Vis. and Patt. Recogn.*, 2015, pp. 3431–3440.

[17] 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.

[18] 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.

[19] O. Oktay *et al.*, "Attention U-Net: learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

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

[21] 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.

[22] J. Sun, F. Darbehani, M. Zaidi, and B. Wang, "Saunet: shape attentive u-net for interpretable medical image segmentation," in *Proc. of Internat. Confer. on Med. Ima. Compu. Comput.-Assis. Interven.*, 2020, pp. 797–806.

[23] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. of Comput. Vis. and Patt. Recogn.*, 2017, pp. 2881–2890.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. on Patt. Analy. and Mach. Intelli.*, vol. 40, no. 4, pp. 834–848, 2017.

[25] 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.

[26] T. Hassan, M. Usman Akram, and N. Werghi, "Evaluation of Deep Segmentation Models for the Extraction of Retinal Lesions from Multi-modal Retinal Images," *arXiv e-prints*, 2020.

[27] R. M. Rad, P. Saeedi, J. Au, and J. Havelock, "Cell-Net: embryonic cell counting and centroid localization via residual incremental atrous pyramid and progressive upsampling convolution," *IEEE Acc.*, vol. 7, pp. 81 945–81 955, 2019.

[28] Y. Guo, J. Bernal, and B. J Matuszewski, "Polyp segmentation with fully convolutional deep neural networks — extended evaluation study," *Jour. of Imag.*, vol. 6, no. 7, p. 69, 2020.

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

[30] X. Chen, R. Zhang, and P. Yan, "Feature fusion encoder decoder network for automatic liver lesion segmentation," in *Proc. of internat. sympos. biomed. imag.*, 2019, pp. 430–433.

[31] C. Kaul, S. Manandhar, and N. Pears, "Focusnet: An attention-based fully convolutional network for medical image segmentation," in *Proc. of Internat. Sympo. on Biomed. Imag.*, 2019, pp. 455–458.

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

[33] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape CNNs for semantic segmentation," in *Proc. of Comput. Vis. and Patt. Recogn.*, 2019, pp. 5229–5238.

[34] 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.

[35] T. Kim, H. Lee, and D. Kim, "Uacanet: Uncertainty augmented context attention for polyp semgnetaion," *arXiv preprint arXiv:2107.02368*, 2021.

[36] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Jour. of digi. imag.*, vol. 32, no. 4, pp. 582–596, 2019.

[37] D. Sarvamangala and R. V. Kulkarni, "Convolutional neural networks in medical image understanding: a survey," *Evolutionary intelligence*, pp. 1–22, 2021.

[38] L. Liu *et al.*, "A survey on u-shaped networks in medical image segmentations," *Neurocomputing*, vol. 409, pp. 244–258, 2020.

[39] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis, "Fully dense unet for 2-d sparse photoacoustic tomography artifact removal," *IEEE J. Biomed. Health Inform*, vol. 24, no. 2, pp. 568–576, 2019.

[40] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. of Comput. Vis. and Patt. Recogn.*, 2018.

[41] X. o. Yang, "Road Detection via Deep Residual Dense U-Net," in *Pro. of Internat. Joi. Conf. on Neu. Netwo.*, 2019, pp. 1–7.

[42] P. L. K. Ding, Z. Li, Y. Zhou, and B. Li, "Deep residual dense U-Net for resolution enhancement in accelerated MRI acquisition," in *Proc. of Medi. Imag. 2019: Ima. Proce.*, vol. 10949, 2019, p. 109490F.

[43] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 5, pp. 1116–1126, 2018.

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

[45] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. on Patt. Analy. and Mach. Intelli.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[46] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *Proc. of Comput. Vis. and Patt. Recogn.*, 2018, pp. 3684–3692.

[47] D. Lin *et al.*, "Zigzagnet: Fusing top-down and bottom-up context for object segmentation," in *Proc. of Comput. Vis. and Patt. Recogn.*, 2019, pp. 7490–7499.

[48] J. Wang, Z. Wei, T. Zhang, and W. Zeng, "Deeply-fused nets," *arXiv preprint arXiv:1605.07716*, 2016.

[49] T. Zhang, G.-J. Qi, B. Xiao, and J. Wang, "Interleaved group convolutions," in *Proc. of Internat. Conf. Compu. Vis.*, 2017, pp. 4373–4382.

[50] 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.

[51] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. of Comput. Vis. and Patt. Recogn. Worksh.*, 2017, pp. 136–144.

[52] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. of AAAI Conf. Artifi. Intelli.*, vol. 31, no. 1, 2017.

[53] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artifi. intelli. stat.*, 2015, pp. 562–570.

[54] F. Chollet *et al.*, "Keras," 2015.

[55] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} sympo. operat. syst. desi. implement. ({OSDI} 16)*, 2016, pp. 265–283.

[56] N. K. Tomar *et al.*, "DDANet: Dual Decoder Attention Network for Automatic Polyp Segmentation," in *Proc. of the ICPR 2020 Worksh. and Chall.*, 2020.

[57] D. Jha and Others, "A Comprehensive Study on Colorectal Polyp Segmentation with ResUNet++, Conditional Random Field and Test-Time Augmentation," *IEEE J. Biomed. Health Inform*, 2021.

[58] B. Levin *et al.*, "Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the american cancer society, the us multi-society task force on colorectal cancer, and the american college of radiology," *Gastroenterology*, vol. 134, no. 5, pp. 1570–1595, 2008.

[59] A. C. Society, "Cancer facts & figures 2018," 2018.

[60] C. Gilvary, N. Madhukar, J. Elkhader, and O. Elemento, "The missing pieces of artificial intelligence in medicine," *Tren. pharmacolo. sci.*, vol. 40, no. 8, pp. 555–564, 2019.

[61] Z. Xu, W. Zhang, T. Zhang, and J. Li, "Hrcnet: high-resolution context extraction network for semantic segmentation of remote sensing images," *Remote Sensing*, vol. 13, no. 1, p. 71, 2021.

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

[63] S. Ali *et al.*, "PolypGen: a multi-center polyp detection and segmentation dataset for generalisability assessment," *arXiv preprint arXiv:2106.04463*, 2021.