FANet: A Feedback Attention Network for Improved Biomedical Image Segmentation

Nikhil Kumar Tomar, Debesh Jha, *Member, IEEE*, Michael A. Riegler, *Member, IEEE*, Håvard D. Johansen, *Member, IEEE*, Dag Johansen, *Member, IEEE*, Jens Rittscher, Pål Halvorsen, *Member, IEEE*, and Sharib Ali, *Member, IEEE*

***Abstract*— The increase of available large clinical and experimental datasets has contributed to a substantial amount of important contributions in the area of biomedical image analysis. Image segmentation, which is crucial for any quantitative analysis, has especially attracted attention. Recent hardware advancement has led to the success of deep learning approaches. However, although deep learning models are being trained on large datasets, existing methods do not use the information from different learning epochs effectively. In this work, we leverage the information of each training epoch to prune the prediction maps of the subsequent epochs. We propose a novel architecture called feedback attention network (FANet) that unifies the previous epoch mask with the feature map of the current training epoch. The previous epoch mask is then used to provide hard attention to the learned feature maps at different convolutional layers. The network also allows rectifying the predictions in an iterative fashion during the test time. We show that our proposed *feedback attention* model provides a substantial improvement on most segmentation metrics tested on seven publicly available biomedical imaging datasets demonstrating the effectiveness of FANet. The source code is available at https://github.com/nikhilroxtomar/FANet.**

***Index Terms*— Cell nuclei, colon polyps, deep learning, feedback attention, lung segmentation, medical image segmentation, retinal vessels, skin lesion.**

Manuscript received March 31, 2021; revised December 19, 2021 and February 11, 2022; accepted March 2, 2022. This work was supported in part by the Research Council of Norway under Grant 263248 and Grant 270053 and in part by the National Institute for Health Research Oxford Biomedical Research Centre. The work of Debesh Jha was supported by the Research Council of Norway (PRIVATON) under Project 263248. The work of Sharib Ali was supported by the National Institute for Health Research Oxford Biomedical Research Centre. *(Corresponding author: Debesh Jha.)*

Nikhil Kumar Tomar is with SimulaMet, 0167 Oslo, Norway.

Debesh Jha and Michael A. Riegler are with SimulaMet, 0167 Oslo, Norway, and also with the Department of Computer Science, UiT The Arctic University of Norway, 9037 Tromsø, Norway (e-mail: debesh@simula.no).

Håvard D. Johansen and Dag Johansen are with the Department of Computer Science, UiT The Arctic University of Norway, 9037 Tromsø, Norway.

Jens Rittscher is with the Department of Engineering Science and the Li Ka Shing Centre for Health Information and Discovery, Big Data Institute, University of Oxford, Oxford OX3 7LF, U.K., and also with the Oxford NIHR Biomedical Research Centre, Oxford University Hospitals Trust, Oxford OX3 9DU, U.K.

Pål Halvorsen is with SimulaMet, 0167 Oslo, Norway, and also with the Department of Computer Science, Oslo Metropolitan University, 0167 Oslo, Norway.

Sharib Ali is with the Department of Engineering Science, University of Oxford, Oxford OX3 7LF, U.K., and also with the Oxford National Institute for Health Research Biomedical Research Centre, Oxford OX4 2PG, U.K. (e-mail: sharib.ali@eng.ox.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3159394.

Digital Object Identifier 10.1109/TNNLS.2022.3159394

I. Introduction

IMAGE segmentation is one of the most studied problems in computer vision, where the main goal is to classify each pixel of an image to a specific class instance. This can either be pixels of any arbitrary objects, such as cars or humans in natural scene data [1], satellite data in remote sensing [2], [3], or pixels of cancerous area or cells in biomedical imaging data [4]. Substantial progress has been made in biomedical imaging due to which various modalities, such as X-ray, computerized tomography (CT), magnetic resonance imaging (MRI), endoscopy imaging, fundus imaging, electron microscopy (EM), and histology imaging exists. While machine learning (ML) methods usually provide improved performance over traditional computer vision methods, most of them require ground truth labels from domain experts, which are often scarce and may not represent enough variability in biomedical imaging data. This can affect ML models resulting in only suboptimal predictions. Furthermore, existing methods for semantic segmentation are based on a single-step prediction process that does not allow them to rectify their own predicted segmentation masks. Thus, these networks are constrained to only one set of learned weights that may not be enough to capture inter- and intra-class differences present in biomedical imaging data. In this work, we introduce an iterative approach that can refine the segmentation masks from previous mask predictions in a few iterative steps. This iteration process enables the network to steer toward the improved feature representation by taking advantage of subsequent attention mechanisms from the previous mask, unlike classically used one-step segmentation methods [1], [5]. Thus, aggregating these results over a few iterations provides improved segmentation masks (see illustration in Fig. 1).

Current developments of convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention modules have improved automated methods in biomedical image analysis. Widely used supervised end-to-end CNNs methods require a large and diverse training dataset to avoid overfitting. A recurrent neural network can be used to preserve the model compactness and can be effectively used for segmentation tasks in resource-constrained settings via an iterative update on internal states of network layers [6]. However, they are known for their memory inefficient memory-bandwidth-bound computation and model complexity [7]. In addition, spatial-visual attention mechanisms used for image captioning for natural scene images [8] and for medical

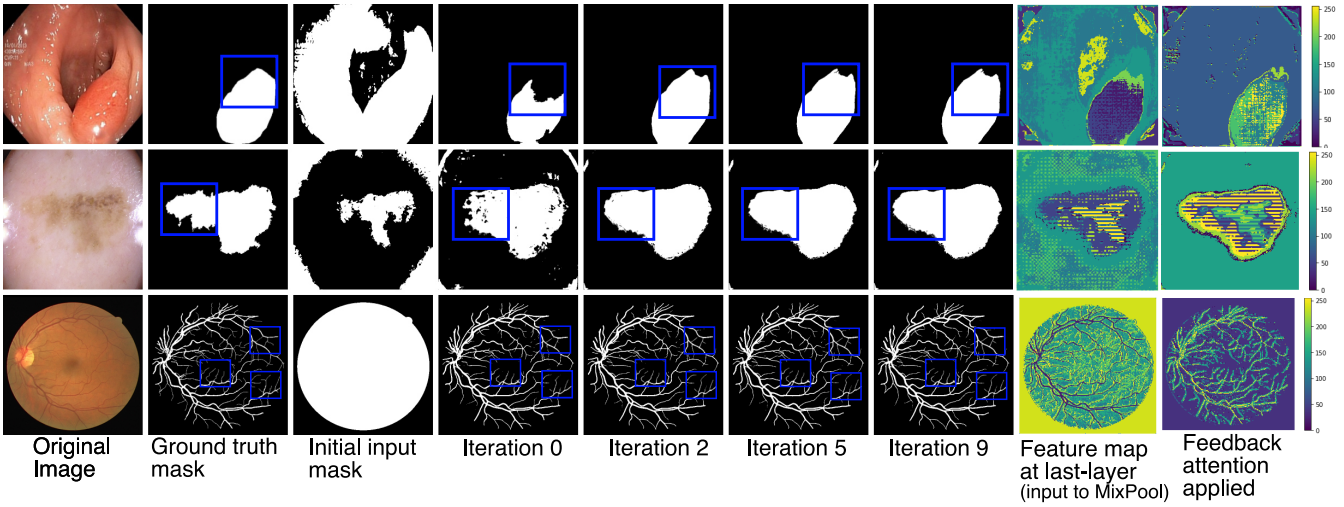

Fig. 1. Semantic segmentation using our FANet architecture. Otsu thresholding is used for generating the initial mask used during 0th iteration. Then, the predictions are iteratively updated with the predicted mask. It can be observed that already at the second iteration, the results converge. The corresponding feature maps before and after feedback attention at the last decoder layer of our FANet are shown, as color images on the right.

image segmentation [9] showed improvements in terms of both model convergence and performance metric. An attention mechanism allows networks to focus on a concrete class instance, thereby penalizing nonspecific regions. Our model is, thus, inspired by the success of both visual attention mechanisms and the recurrent learning paradigm.

A mask-guided contrastive attention model was used by Song *et al.* [10] to deal with the background clutter. Unlike classical training mechanisms and motivated by the work of Song *et al.* [10], we propose to propagate the sample-specific mask output from the previous epoch to the successive epoch in a recursive fashion. Such a feedback mechanism can provide prior information that can help to learn sample variability, thereby enabling to train effectively on diverse datasets. Here, iterative prediction can be used to prune the predicted masks during the inference (see Fig. 1). This allows the network to learn both local and global features that can rectify the mask output from the learned weights. Unlike test-time augmentation (TTA) [11], where different transforms are utilized to mimic sample representations and data diversity, we embed mask rectification during the training process. To our knowledge, feedback attention network (FANet) is the first deep learning model that incorporates the ability to self-rectify its predictions without requiring heavy transformations, ensemble strategies, and prior sample-specific knowledge. FANet uses a single end-to-end trainable network that allows information propagation during both train and test time.

A feedback mechanism during the training is the central to our novel FANet approach for semantic segmentation. The predicted map of each sample from the previous epoch unified with the current state feature map is used to provide attention. FANet uses an attention mechanism to different feature scales in the network, allowing it to capture variability in image samples. In addition, our residual block with squeeze and excitation (SE) layer allows us to improve channel interdependencies, which can be critical to tackling image quality issues. The main contributions of this work can be summarized as follows.

1) *Feedback Attention Learning:* A novel mechanism to utilize the variability present in each training sample. The mask outputs are propagated from one to subsequent epochs to suppress the unwanted feature clutter.
2) *Iterative Refining of Prediction Masks:* Using feedback information helps in refining the predicted masks in training and inference. During testing, we iterate over the input image and keep updating the input mask with the predicted mask for up to ten iterations (empirically set).
3) *Embedded run-length encoding strategy:* Binary mask outputs of each sample are efficiently compressed before being propagated to the next epoch. This provides a memory-efficient mechanism for passing sample-specific masks.
4) *Systematic Evaluation:* Experiments on seven vastly different biomedical datasets suggest that FANet outperforms other state-of-the-art (SOTA) algorithms.
5) *Efficient Training:* FANet achieves near SOTA performance with far fewer training epochs.

II. Related Work

In this section, we summarize relevant advances in medical image segmentation and feedback attention networks. We also highlight recent contributions to iterative refinement methods for image segmentation.

A. Biomedical Image Segmentation

The basis of most modern CNN-based semantic segmentation architectures are either fully convolutional network (FCN) [12] or an encoder–decoder architecture, such as U-Net [5] originally designed for cell segmentation. Various modifications of these networks have been proposed both for semantic segmentation of natural images [13], [14] and biomedical image segmentation [9], [15]–[20]. In general, in the encoder, the image content is encoded using multiple convolutions to capture from low-level to high-level features, whereas in the decoder part of the network the prediction masks are obtained by multiple upsampling mechanisms or deconvolution operations. Methods like PSPNet [13] and DeepLab [1] incorporate convolutional feature maps of varying

resolutions to segment both small and large-sized objects effectively. While PSPNet used a pyramid pooling module, DeepLab used atrous spatial pyramidal pooling (ASPP) for encoding the multiscale contextual information. Both PSPNet- and DeepLab-based architectures have been used widely in the medical imaging community for biomedical image segmentation [21], [22].

B. Feedback Attention Networks

Visual attention has been widely used in computer vision for pose estimation [23], object detection [24], and image segmentation [25], [26]. Chu *et al.* [23] incorporated the multicontext attention method into their end-to-end eight stack hourglass CNN network where each subnetwork of the hourglass generated a multiresolution attention map. Attention mechanisms [27], [28] have also been utilized for posing explicit focus on the target region in medical imaging. Schlemper *et al.* [28] proposed a novel attention gate model that automatically learned to focus on the target structure of the varying shape and sizes by suppressing the irrelevant features and highlighting the silent feature for the specified medical image segmentation task. Attention U-Net [9] used a gated operation in the U-Net architecture to focus on the target abdominal regions of CT datasets. Feedback mechanism for attention using two U-Net architectures with shared weights was used for cell segmentation [29], [30]. The latter used a standard U-Net architecture with the second U-Net incorporating ConvLSTM [31] to store the feature map (input-to-state) from the first U-Net network. However, feedback is only applied to the same epoch with state-to-state transitions. On the contrary, our approach utilizes a feedback mechanism that propagates information flow from the previous epoch to the current epoch in an attention mechanism. We employ the predicted masks from the previous epoch as hard attention to prune the segmentation output.

C. Iterative Refinement for Segmentation

Iterative refinement of the segmentation mask by feeding the input image and the predicted segmentation mask to a modified U-Net architecture was done by Mosinska *et al.* [32]. In this work, we used an iterative refinement pipeline to enhance the quality of the predicted segmentation mask. Similarly, iterative update of latent space and minimization of the structure similarity index measure (SSIM) loss was used to refine the predicted segmentation maps during test time in [33]. Recently, iterative refinement strategies have also been used for pose estimation [34], [35] that used consecutive modules for refinement of the predictions with a loss function for the evaluation of output in each module. These iterative refinement processes show improved predictions and are able to handle domain shifts or object shape variability without requiring very deep networks [33]. However, a major bottleneck in these methods is the requirement of a large number of iterations for model convergence. Unlike these methods, our proposed FANet provides attention to the specific region-of-interest and can prune the predicted segmentation masks in less than ten iterations without requiring any optimization scheme.

III. Method

In this section, we describe the components of the proposed FANet architecture. The overall design along with the proposed feedback attention learning mechanism is shown in Fig. 2.

A. SE-Residual Block

Deeper networks improve the performance of the model significantly, but an increase in depth can cause either vanishing or exploding gradients problem [43]. To deal with this, we take advantage of shortcut connections between layers in the residual learning paradigm. Our squeeze and excite residual (SE-residual) block uses two 3×3 convolutions and an identity mapping, where each convolution layer is followed by a batch normalization (BN) layer and an rectified linear unit (ReLU) nonlinear activation function. The identity mapping is used to connect the input and the output of the convolution layer [Fig. 2(a)].

Similar to the work by Hu *et al.* [44], we add an SE layer in the residual network. The SE layer acts as a content-aware mechanism that reweights each channel accordingly to create robust representations. Hence, it allows the network to become more sensitive to significant features while suppressing irrelevant features. This goal is accomplished in two steps. First, the feature maps are squeezed by using the global average pooling to get a global understanding of each channel. The squeeze operation results in a feature vector of size n, where n refers to the number of channels. In the second step: excitation, this feature vector is fed through a two-layered feed-forward neural network, where the number of features is first reduced and then expanded to the original size n. Now, this n-sized vector represents the weight of the original feature maps, which is used to scale each channel.

B. MixPool Block

The proposed MixPool block shown in Fig. 2(b) is used in multiple layers of our FANet architecture. This block facilitates the flow of samplewise feedback information between consecutive epochs providing hard attention to the learned features from the SE-residual block. The layer provides focus to the relevant features in both contraction path and expansion path layers. The 'hard' attention map consists of the values 0 and 1, i.e., attention to a specific region only unlike soft attention where the probability map is estimated. The advantage of hard attention is that it allows to keep only the important features and ignore irrelevant features. During the elementwise multiplication, the values from the input feature map, if multiplied by 0, becomes 0, leaving the essential features for further operations. Another advantage of such methods is their computational speed, scalability, and ease of interpretation [8], [45]. The input mask used during training is compressed using the run-length encoding technique to save the memory footprint.

As in Fig. 2(b) and (c), fist feature maps from the SE-residual blocks F_l in each layer is passed through a 3×3 convolution followed by a BN and an ReLU

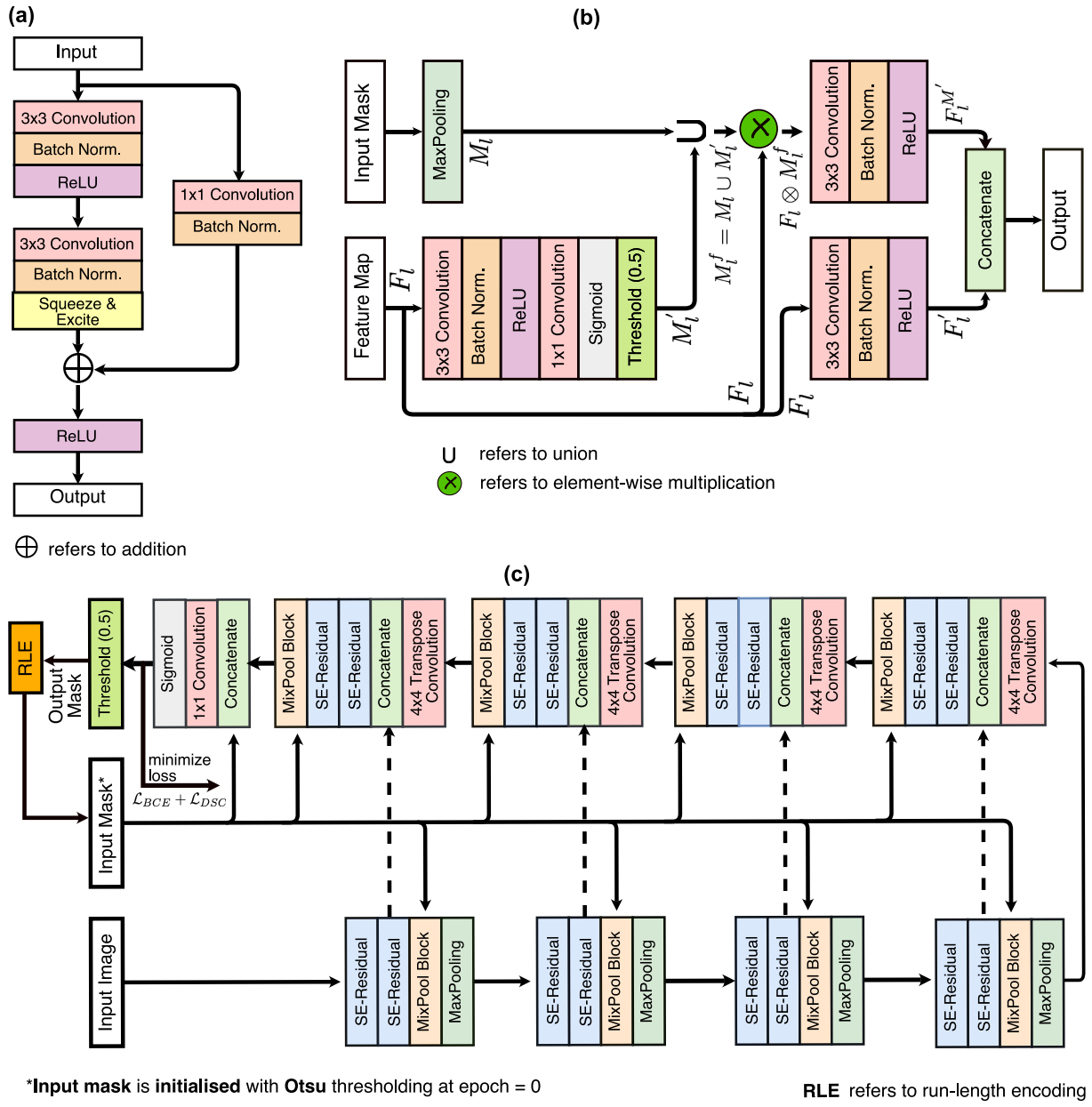

Fig. 2. FANet with SE-residual block and MixPool block. (a) SE-residual block integrated with a squeeze and excite layer uses 1×1 convolution to concatenate the high-resolution feature representation with the encoded feature vector. (b) MixPool block represents the attention mechanism in our network. The input mask is downscaled to the corresponding layer feature map size M_l which is fused with the masked feature map representation M_l' for hard attention of input feature in that layer F_l. Finally, the attenuated feature map $F_l^{M'}$ and the feature maps F_l' are both concatenated. (c) Proposed FANet showing the complete network architecture. Encoder–decoder architecture with skip-connections (in dotted arrows) from SE-residual blocks to preserve high- and intermediate-resolution feature representations and MixPool block connections (with solid arrows) that allow feedbacking the previous mask predictions.

activation function. Then, we apply a 1×1 convolution and a sigmoid activation function $\sigma(\cdot)$ with a threshold of 0.5 to obtain the binary mask M_l' to contribute to the *spatial attention map generation* given by

$$M_l' = \sigma(\text{conv}(F_l)) = \begin{cases} 1, & \text{if } \sigma(\cdot) \geq 0.5 \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Second, we apply appropriate max-pooling on the input mask (from the previous epoch) and resize it to the size of the spatial attention map M_l'. A union operation is then applied between the resized mask and the spatial attention map. This confirms that we obtain the feature from both the feedback and the spatial attention maps to further create a new unified spatial attention map. Next, an elementwise multiplication operation is applied between the unified mask and the original feature map that suppresses the irrelevant features and enhances the important ones. The enhanced and the original feature maps are then followed by a 3×3 convolution, BN, and an ReLU. These operations are used to improve the network's ability to learn nonlinearity in the model prediction.

Finally, we concatenate the output of both activation functions, which constitutes the output of our MixPool block

given by

$$\text{Output}_{\text{MixPool}} = F_l^{'} \frown \left(F_l \otimes \left(M_l \cup M_l^{\prime}\right)\right)^{\prime} \tag{2}$$

where $\frown$ denotes the concatenation operator, $\otimes$ is elementwise multiplication, and $\cup$ represents the union operation.

C. Proposed FANet Architecture

The block diagram of FANet is shown in Fig. 2(c). It uses an encoder–decoder design common to many semantic segmentation architectures. We combine the strength of a residual network enhanced with SE as an SE-residual block and MixPool block that facilitates the attention and propagation of information flow from the current learning paradigm and that of the previous epoch. We implement a recurrent learning mechanism in both encoder and decoder layers that allow us to achieve efficient segmentation. The MixPool block uses the previous segmentation map (as an input mask through RLE encoding), which contains the information from prior training and uses to improve the semantic representation of the feature maps.

We first use the Otsu thresholding [46] to generate an initial input mask for training the proposed architectural model. The variability in the input mask is refined over the training epochs and the model learns over time to prune input or previous epoch masks with learned semantically meaningful features together. To achieve this, we use the novel MixPool block that uses the input mask and applies hard attention over the subsequent input feature maps. The hard attention enables the network to highlight semantically meaningful features for the target region-of-interest in the entire network. The network, thus, not only learns to predict features maps but also strengthens a joint pruning mechanism that is dependent on the input mask. As a result, the devised network is able to rectify the predicted segmentation maps in an iterative fashion, unlike conventional methods which do not have such pruning ability. This provides a strong rationale behind our work that is applicable beyond single-step inference prediction with the capability of refining prediction maps.

The proposed network architecture is an FCNN consisting of four encoder and four decoder blocks. The encoder takes the input image, downsamples it gradually, and encodes it in a compact representation. Then, the decoder takes this compact representation and tries to reconstruct the semantic representation by gradually upsampling it and combining the features from the encoder. Finally, we receive a pixelwise categorization of the input image. Both the encoder and the decoder are built using the SE-residual block, and an additional concatenation of the original resolution feature representation in the encoder is added at each resolution scale. This mechanism minimizes the loss of feature representations during downscaling and upscaling processes.

Each encoder network starts with two SE-residual blocks, which consist of two 3×3 convolutions and a shortcut connection, known as identity mapping, connecting the input and output of the two convolution layers. Each convolution is followed by a BN and a ReLU activation function. The output of the second SE-residual block acts as a skip connection for the corresponding decoder block. After that, it is followed by the MixPool block, which has the previous epoch segmentation mask and provides a hard-attention over the incoming feature maps. This process is repeated for each of the downscaled layers.

Each decoder network starts with a 4×4 transpose convolution that doubles the spatial dimensions of the incoming feature maps. These feature maps are concatenated with feature maps from the corresponding encoder block through skip connections. The skip connections help to propagate the information from the upper layers, which are sometimes lost due to the depth of the network. The skip connections are followed by two SE-residual blocks, which help to eliminate the problem of vanishing gradient. The MixPool block that utilizes the segmentation mask from the previous epoch is then applied to create a hard attention over the learned feature maps. Next, we concatenate the feature maps from the last decoder block and the segmentation mask from the previous epoch. Finally, we apply a 1×1 convolution with the sigmoid activation function. The output of this is used to both minimize the training loss, using a combined binary cross entropy and dice loss, and to generate segmentation masks that are stored as a run-length encoded compression for each sample and propagated during the next epoch. The RLE is updated after each epoch. Similarly, the network learns to adapt the weights in iterative training, this mechanism is also utilized during the test time. As shown in Fig. 1, test results are pruned in a few iterations during the test time. Unlike many methods in the literature [32], [33], we utilize the same network without any complementary loss function optimization.

IV. Experiments

A. Setup

1) Dataset and Evaluation Metrics: To evaluate the proposed architecture, we have selected seven datasets that capture different segmentation tasks in biomedical imaging. The details of each dataset can be found in Table I. The dataset images contain the images of organs and lesions acquired under different imaging protocols. For the retina vessel segmentation task, we use the DRIVE and CHASE-DB1 datasets. These two datasets are aimed at various diseases related to diseases of retina vessels, such as retinopathy, retinal vein occlusion, and retinal artery occlusion. The International Skin Imaging Collaboration (ISIC) 2018 dataset, which is a dermoscopy dataset that is useful in the diagnosis of skin cancer, is the third dataset focused on medical imaging data. This dataset contains a wide variety of skin cancer images of different sizes and shapes, which helps in a better understanding of the disease. We have further included the Kvasir-SEG and CVC-ClinicDB colonoscopy datasets. These datasets contain the image frames extracted from different colonoscopy interventions and are focused on colorectal polyps that are one of the cancer precursors in the colon and rectum. It highly increases the chance of avoiding lethal cancer by early detection. In addition, we have included two datasets acquired from biological imaging aimed at understanding the cellular processes. These include the 2018 Data Science Bowl (DSB)

TABLE I

DETAILS OF THE BIOMEDICAL DATASETS USED IN OUR EXPERIMENTS. "TRAIN," "TRAIN AFTER AUG.," AND "TEST" DENOTE THE NUMBER OF TRAINING SAMPLES, NUMBER OF TRAINING SAMPLES AFTER IMAGE AUGMENTATION, AND NUMBER OF TEST SAMPLES, RESPECTIVELY

Dataset	Images	Size	Train	Train after aug.	Test	Application
Kvasir-SEG [36]	1000	Variable	880	16720	120	Colonoscopy
CVC-ClinicDB [37]	612	384×288	490	14210	61	Colonoscopy
2018 Data Science Bowl [38]	670	256×256	335	10720	134	Nuclie
ISIC 2018 (Lesion Boundary Segmentation) [39], [40]	2596	Variable	1815	39930	259	Dermoscopy
EM dataset [4]	30	512×512	24	384	3	Cell
DRIVE Database [41]	40	584×565	20	640	20	Retina Vessel
CHASE-DB1 [42]	28	584×565	20	640	8	Retina Vessel

and the EM datasets. The 2018 DSB dataset contains images with a large number of variable-shaped nuclei acquired from different cell types, magnification, and imaging modalities. This dataset is designed for automated nuclei segmentation. Similarly, the EM dataset contains the transmission EM images of the neural structures of the Drosophila nerve cord. This dataset is aimed at the automated segmentation of the neural structures. All experiments on these datasets are conducted on the same train, validation, and test splits as provided by the previously published works reported in this article.

To evaluate SOTA deep learning methods and our proposed FANet, we have used standard evaluation metrics that includes dice coefficient (DSC) (a.k.a. F1), mean intersection over union (mIoU), precision, and recall. We have additionally calculated specificity for those datasets where this metric was previously used for benchmarking.

2) Implementation Details: All the training is performed on a Volta 100 GPU and an NVIDIA DGX-2 system using the PyTorch 1.6 framework. For test inference, we have used an NVIDIA GTX 1050 Ti GPU for our method and all SOTA methods used in the article as this hardware is widely available. Our model is trained for 100 epochs *(empirically set)* using an Adam optimizer with a learning rate of $1e^{-4}$ for all the experiments except for the digital retinal images for vessel extraction (DRIVE) and the CHASE-DB1 dataset, where the learning rate was adjusted to $1e^{-3}$ due to the small size of the training dataset. Datasets were chosen such that the efficiency of our model could be compared with the SOTA methods. A combination of binary cross entropy and dice loss has been used as the loss function. ReduceLROnPlateau callback was used to monitor the learning rate and adjust it to obtain optimal training performance. All the images used in the study were resized to 512×512 except for the 2018 DSB and the CVC-ClinicDB dataset, where images were resized to 256×256. Data augmentation, such as random crop, flipping, rotation, elastic transformation, grid distortion, optical distortion, grayscale conversion, random brightness, contrast, channel, and course dropout were used.

3) Ablation Study: In order to evaluate the strength of our proposed FANet architecture, we perform a thorough ablation study. For this, we have used all seven datasets and evaluated several metrics for baseline (FANet without MixPool), baseline with MixPool, and the combination of baseline, MixPool, and feedback (proposed).

TABLE II

RESULTS ON THE KVASIR-SEG [36]

Method	Backbone	F1	mIoU	Recall	Prec.
U-Net [5]	-	0.5969	0.4713	0.6171	0.6722
ResUNet [47]	-	0.6902	0.5721	0.7248	0.7454
ResUNet++ [18]	-	0.7143	0.6126	0.7419	0.7836
FCN8 [12]	VGG 16	0.8310	0.7365	0.8346	0.8817
HRNet [14]	-	0.8446	0.7592	0.8588	0.8778
DoubleU-Net [19]	VGG 19	0.8129	0.7332	0.8402	0.8611
PSPNet [13]	ResNet50	0.8406	0.7444	0.8357	0.8901
DeepLabv3+ [48]	MobileNet	0.8425	0.7575	0.8377	0.9014
DeepLabv3+ [48]	ResNet50	0.8572	0.7759	0.8616	0.8907
DeepLabv3+ [48]	ResNet101	0.8643	0.7862	0.8592	**0.9064**
U-Net [5]	VGG19	0.7535	0.6571	0.7364	0.8565
U-Net++ [49]	-	0.8002	0.7000	0.8716	0.7992
Attention U-Net [9]	-	0.7944	0.6959	0.8383	0.8287
FANet	-	**0.8803**	**0.8153**	**0.9058**	0.9005

B. Results

In the following, we present quantitative results on seven different biomedical imaging datasets and compare them with corresponding SOTA methods.

1) Results on Kvasir-SEG: Kvasir-SEG [36] is a publicly available polyp segmentation dataset acquired from clinical colonoscopy procedures. This dataset has been widely used for algorithm benchmarking. We have trained our model and compared it with recent SOTA methods on Kvasir-SEG. A comparison with widely accepted segmentation methods with different backbones (see Table II) shows that our approach is improved performance compared with the SOTA methods (on the same train-test split). Our FANet outperforms all the SOTA methods on almost all metrics. While outperforming most U-Net and its variants, it can be observed that FANet achieved an F1 score of 0.8803, which is 1.6% and 3.57% better than the most accurate DeepLabv3+ with ResNet101 backbone and the recent HRNet.

2) Results on CVC-ClinicDB Dataset: CVC-ClinicDB is another commonly used dataset for colonoscopy image analysis. FANet architecture outperforms all the SOTA methods on this dataset by a large margin with F1 of 0.9355, mIoU of 0.8937, recall of 0.9339, and precision of 0.9401 (see Table III). FANet achieves the best tradeoff between recall and precision compared to the ResUNet-based architectures [18], [47]. The strength of the FANet can be observed by the large improvement of 23.17% in the recall and 5.24% in the precision over the SOTA ResUNet++ [18]. The recall

TABLE III
RESULTS ON THE CVC-CLINICDB [37]

Method	F1	mIoU	Recall	Precision
U-Net (MICCAI'15) [5]	0.8230	0.7550	-	-
ResUNet-mod [47]	0.7788	0.4545	0.6683	0.8877
ResUNet++ [18]	0.7955	0.7962	0.7022	0.8785
SFA (MICCAI'19) [50]	0.7000	0.6070	-	-
PraNet [17]	0.8990	0.8490	-	-
U-Net++ [49]	0.9377	0.8890	0.9405	0.9432
Attention U-Net [9]	0.9325	0.8856	0.9276	0.9546
FANet	**0.9355**	**0.8937**	**0.9339**	**0.9401**

TABLE IV
RESULTS ON THE 2018 DSB [38]

Method	Backbone	F1	mIoU	Recall	Prec.
U-Net [5]	ResNet101	0.7573	0.9103	-	-
DoubleU-Net [19]	VGG19	0.7683	0.8407	0.6407	**0.9596**
U-Net++ [49]	-	0.9117	0.8477	0.9203	0.9107
Attention U-Net [9]	-	0.9179	0.8570	0.9183	0.9235
FANet	None	**0.9176**	0.8569	**0.9222**	0.9194

TABLE V
RESULTS ON THE ISIC 2018 (SKIN CANCER SEGMENTATION) [39], [40]

Method	F1	mIoU	Recall	Spec.	Prec.
U-Net [5]	0.6740	0.5490	0.7080	0.9640	-
R2U-Net [52]	0.6790	0.5810	0.7920	0.9280	-
Attention R2U-Net [52]	0.6910	0.5920	0.7260	0.9710	-
BCDU-Net (d=1) [53]	0.8470	-	0.7830	0.9800	-
BCDU-Net (d=3) [53]	0.8510	-	0.7850	**0.9820**	-
U-Net++ [49]	0.8088	0.7319	0.8450	0.9110	0.8648
Attention U-Net [9]	0.8205	0.7346	0.8516	0.9135	0.8645
FANet	**0.8731**	**0.8023**	**0.8650**	0.9611	**0.9235**

suggests that our method is more clinically preferable than the SOTA. A higher recall is desired in the systems used for clinical diagnosis [51].

3) Results on 2018 Data Science Bowl: Cell nuclei segmentation in microscopy imaging is a common task in the biological image analysis [38]. We used the publicly available 2018 DSB challenge dataset and compared our results with the SOTA methods. Table IV shows that FANet produces an F1 of 0.9176, mIoU of 0.8569, and recall of 0.9222 with an improvement of 2.02% in F1 with respect to SOTA UNet++ [16] and 28.15% improvement in recall compared with the best performing DoubleU-Net [19]. In general, FANet achieves the best tradeoff between precision and recall compared with the SOTA methods resulting in the highest F1 score (0.9176). The qualitative results with 2018 DSB also show that the predicted FANet produces high-quality segmentation masks for cell nuclei with respect to the ground truth (see Fig. 3).

4) Results on ISIC 2018 Dataset: Skin cancer is one of the most commonly diagnosed cancers in the U.S. Early detection of melanoma can improve the five-year survival rate and help prevent it in 99% of the cases [54]. Table V shows the results on the publicly available ISIC 2018 dataset. FANet outperformed all the methods on almost all evaluation metrics

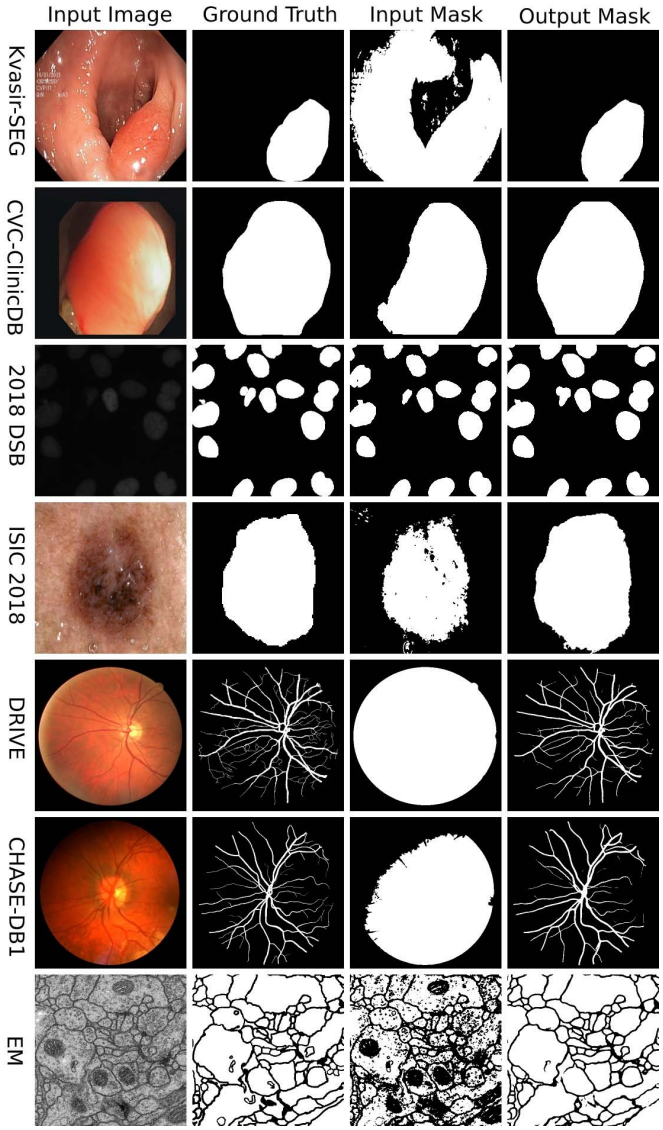

Fig. 3. Qualitative results of FANet on seven biomedical image segmentation datasets. The initial "input mask" is generated using *Otsu thresholding*. The "output mask" is the predicted segmentation mask from the FANet model.

(F1, mIoU, and recall). FANet achieved 0.8731 on F1 and recall of 0.8650 with an improvement of 2.21% and 8.00%, respectively, over the most accurate SOTA BCDU-Net ($d = 3$) method. A competitive specificity and precision were also recorded. From the qualitative results in Fig. 3, we can see that the input mask produced by Otsu thresholding shows under segmentation, which is improved significantly using FANet. The masks produced by FANet have smooth boundaries.

5) Results on DRIVE Dataset: The automated segmentation of vessels in fundus images can assist in the diagnosis and treatment of diabetic retinopathy. The quantitative result on the publicly available DRIVE dataset is presented in Table VI. We can observe that the proposed FANet achieves an F1 score of 0.8183, mIoU of 0.6927, recall of 0.8215, and precision of 0.8189. The proposed method achieves an improvement of 4.24% in the recall over SOTA IterNet [57]. Although the F1 of the IterNet is 0.35% higher than FANet, the recall is relatively lower, and other metrics, such as mIoU and

TABLE VI
RESULTS ON THE DRIVE DATASET [41]

Method	F1	mIoU	Recall	Spec.	Prec.
U-Net [5]	0.8174	-	0.7822	0.9808	-
Residual U-Net [52]	0.8149	-	0.7726	0.9820	-
Recurrent U-Net [52]	0.8155	-	0.7751	0.9816	-
R2U-Net [52]	0.8171	-	0.7792	0.9813	-
DenseBlock-UNet [55]	0.8146	-	0.7928	0.9776	-
DUNet [56]	0.8190	-	0.7863	0.9805	-
IterNet [57]	**0.8218**	-	0.7791	0.9831	-
IterNet(Patched) [57]	0.8205	-	0.7235	**0.9838**	-
U-Net++ [49]	0.7960	0.6615	0.7903	0.9818	0.8070
Attention U-Net [9]	0.7984	0.6648	0.7877	0.9827	0.8146
FANet	0.8183	**0.6927**	**0.8215**	0.9826	**0.8189**

TABLE VII
RESULTS ON THE CHASE-DB1 DATASET [42]

Method	F1	mIoU	Recall	Spec.	Prec.
U-Net [5]	0.7993	-	0.7840	0.9880	-
DenseBlock-UNet [55]	0.8005	-	0.8177	0.9848	-
DUNet [56]	0.8000	-	0.7858	0.9880	-
IterNet [57]	0.8072	-	0.7969	**0.9881**	-
U-Net++ [49]	0.7954	0.6606	0.8114	0.9847	0.7818
Attention U-Net [9]	0.7941	0.6589	0.8049	0.9852	0.7852
FANet	**0.8108**	**0.6820**	**0.8544**	0.9830	**0.7722**

precision, are not presented. For our proposed FANet, the precision of 0.8189 is well balanced with the obtained recall. The higher recall produced by FANet shows that our method is more clinically relevant. The quality of the segmentation masks in Fig. 3 demonstrates the efficiency of FANet.

6) Results on CHASE-DB1 Dataset: CHASE-DB1 is the second retinal image segmentation dataset used to evaluate our method. For this dataset, there is no official training and test split. We have used 20 images to train our model and 8 images to test as reported in the work of Li *et al.* [57]. From Table VII, we can observe that our method achieved the highest F1 of 0.8108, mIoU of 0.6820, and the highest recall of 0.8544. FANet achieved an improvement of 3.67% in the recall compared with the SOTA DenseBlock-UNet.

7) Results on EM Dataset: The EM dataset aims to develop an automatic ML algorithm for the segmentation of the neural structures so that difficulties due to manual labeling can be resolved. Table VIII shows the quantitative results on the EM dataset. The proposed FANet also obtains F1 of 0.9547, mIoU of 0.9134, and a recall of 0.9568. The presented results demonstrate that FANet produces SOTA results, surpassing other recent methods in terms of mIoU metric that was used by other methods for comparison.

C. Qualitative Results

The qualitative results on all seven datasets are presented in Fig. 3. It can be observed that for colonoscopy datasets (Kvasir-SEG and CVC-ClinicDB), even though the initial input mask covers the entirety of the image, our model is able to prune and provide accurate masks. The same can be observed for the two retina vessel segmentation datasets, DRIVE and CHASE-DB1. It can be observed that our model

TABLE VIII
RESULTS ON THE EM DATASET [4]

Method	F1	mIoU	Recall	Specificity	Prec.
U-Net [5]	-	0.8830	-	-	-
Wide U-Net [15]	-	0.8837	-	-	-
U-Net++ [49]	0.9495	0.9038	0.9520	0.7875	0.9474
Attention U-Net [9]	0.9492	0.9033	0.9502	0.7912	0.9484
FANet	**0.9547**	**0.9134**	**0.9568**	**0.8096**	**0.9529**

is able to segment the challenging retinal vessels, including small retinal vessel bifurcations, and it well resembles the ground truth mask. For 2018 DSB, ISIC-2018, and EM cell data, again, the input masks are finely rectified, achieving close to ground truth results by the proposed FANet model.

D. Ablation Study

In this section, we ablate our model architecture and present extensive experimental results related to the effectiveness of the proposed FANet. To evaluate the contribution of the MixPool block and the feedback, we created the following configurations.

1) *Baseline (B1):* It refers to the FANet without the MixPool block, which means "no feedback mechanism" or "iterative pruning." We require the MixPool block to provide feedback as it unifies the attention from the network feature map and input mask [refer to Fig. 2(b)].
2) *Baseline + MixPool (B2):* We integrate the MixPool block in all the encoder blocks and decoder blocks. During the inference, we directly apply the trained model weights with the Ostu thresholding (initial input mask) only once, i.e., no iterative pruning is used.
3) *Baseline + MixPool(E1, D4) + Feedback (B3):* Here, we integrate the MixPool block in the first encoder block and the last decoder block. Feedback (iterative pruning) is used during the inference.
4) *Baseline + MixPool + Feedback (B4):* This is the final FANet architecture, with MixPool block in all encoder and decoder blocks and the feedback (iterative pruning) mechanism is used during the inference.

Table IX presents the ablation results on these four configurations performed on all seven datasets. In the following, we provide detailed analyses of the use of different model architectural settings and validate them with the above-described four network configurations (**B1–B4**):

1) Effectiveness of MixPool Block: The MixPool block is an essential part of the proposed FANet architecture. It uses the previously predicted mask as the attention to improve the semantically meaningful features and allows higher level abstractions. The effectiveness of the MixPool block can be evaluated by comparing the network configurations B1 and B4.

From the experiments in Table IX, we can conclude that the B4 outperforms the B1 on all the datasets. On the F1 metric, B4 shows an improvement of 2.87% on the Kvasir-SEG dataset, 1.89% improvement on the CVC-ClinicDB, 0.55% improvement on the 2018 DSB dataset, 0.84% improvement on the ISIC 2018 dataset, 0.11% improvement on the DRIVE dataset, 2.92% improvement on the CHASE-DB1 dataset, and

TABLE IX

DETAILED ABLATION STUDY OF THE FANET ARCHITECTURE. FLOP IS CALCULATED IN TERMS OF GMAC. "REC" STANDS FOR RECALL, "PREC" STANDS FOR PRECISION, "SPEC" STANDS FOR SPECIFICITY, "ACC" STANDS FOR ACCURACY, AND "PARAM" STANDS FOR TOTAL NUMBER PARAMETERS. B1–B4 DENOTE DIFFERENT NETWORK CONFIGURATIONS

Method	mIoU	F1	Rec	Prec	Spec	Acc	F2	Param	Flops	FPS	Image Size
Dataset: Kvasir-SEG											
Baseline (FANet without MixPool, **B1**)	0.7732	0.8516	0.8835	0.8710	0.9783	0.9563	0.8614	5.76M	70.38	104.20	512×512
Baseline + MixPool (**B2**)	0.6378	0.7302	0.6982	**0.9098**	**0.9815**	0.9412	0.7039	7.72M	94.75	66.75	512×512
Baseline + MixPool(E1, D4) + feedback (**B3**)	0.7688	0.8460	0.9047	0.8479	0.9576	0.9474	0.8699	5.78M	76.53	101.10	512×512
Baseline + MixPool + feedback (**B4**)	**0.8153**	**0.8803**	**0.9058**	0.9005	0.9794	**0.9667**	**0.8872**	7.72M	94.75	68.18	512×512
Dataset: CVC-ClinicDB											
Baseline (FANet without MixPool, **B1**)	0.8619	0.9166	0.9310	0.9247	0.9934	0.9877	0.9194	5.76M	70.38	103.46	256×256
Baseline + MixPool (**B2**)	0.8541	0.9108	0.9026	0.9296	0.9943	0.9864	0.9048	7.72M	94.75	67.490	256×256
Baseline + MixPool(E1, D4) + feedback (**B3**)	0.8729	0.9162	0.9052	**0.9462**	0.9941	0.9889	0.9093	5.78M	76.53	99.03	256×256
Baseline + MixPool + feedback (**B4**)	**0.8937**	**0.9355**	**0.9339**	0.9401	**0.9948**	**0.9916**	**0.9342**	7.72M	94.75	67.910	256×256
Dataset: 2018 Data Science Bowl											
Baseline (FANet without MixPool, **B1**)	0.8495	0.9121	0.9047	0.9283	0.9871	0.9800	0.9068	5.76M	70.38	114.82	256×256
Baseline + MixPool (**B2**)	0.8158	0.8893	0.8665	**0.9289**	**0.9887**	0.9751	0.8733	7.72M	94.75	69.64	256×256
Baseline + MixPool(E1, D4) + feedback (**B3**)	0.8552	0.9165	0.9189	0.9199	0.9863	**0.9802**	0.9173	5.78M	76.53	100.27	256×256
Baseline + MixPool + feedback (**B4**)	**0.8569**	**0.9176**	**0.9222**	0.9194	0.9860	0.9800	**0.9195**	7.72M	94.75	69.22	256×256
Dataset: ISIC 2018											
Baseline (FANet without MixPool, **B1**)	0.7908	0.8647	**0.9033**	0.8780	0.9151	0.9151	**0.8778**	5.76M	70.38	111.95	512×512
Baseline + MixPool (**B2**)	0.7486	0.8303	0.8049	0.9214	**0.9617**	0.9211	0.8081	7.72M	94.75	65.91	512×512
Baseline + MixPool(E1, D4) + feedback (**B3**)	**0.8078**	**0.8780**	0.8746	**0.9252**	0.9614	**0.9374**	0.8719	5.78M	76.53	99.06	512×512
Baseline + MixPool + feedback (**B4**)	0.8023	0.8731	0.8650	0.9235	0.9611	0.9351	0.8630	7.72M	94.75	71.02	512×512
Dataset: DRIVE											
Baseline (FANet without MixPool, **B1**)	0.6912	0.8172	0.8048	**0.8339**	**0.9846**	**0.9687**	0.8093	5.76M	70.38	103.68	512×512
Baseline + MixPool (**B2**)	0.6895	0.8161	**0.8219**	0.8145	0.9820	0.9678	0.8190	7.72M	94.75	68.47	512×512
Baseline + MixPool(E1, D4) + feedback (**B3**)	**0.6928**	**0.8183**	0.8124	0.8280	0.9839	**0.9687**	0.8142	5.78M	76.53	95.30	512×512
Baseline + MixPool + feedback (**B4**)	0.6927	**0.8183**	0.8215	0.8189	0.9826	0.9683	**0.8197**	7.72M	94.75	70.66	512×512
Dataset: CHASE-DB1											
Baseline (FANet without MixPool, **B1**)	0.6419	0.7816	0.7876	0.7768	0.9848	0.9723	0.7850	5.76M	70.38	95.77	512×512
Baseline + MixPool (**B2**)	0.5419	0.7009	0.8116	0.6209	0.9664	0.9565	0.7625	7.72M	94.75	65.03	512×512
Baseline + MixPool(E1, D4) + feedback (**B3**)	**0.6877**	**0.8147**	0.8372	**0.7948**	**0.9855**	**0.9760**	0.8279	5.78M	76.53	99.00	512×512
Baseline + MixPool + feedback (**B4**)	0.6820	0.8108	**0.8544**	0.7722	0.9830	0.9749	**0.8363**	7.72M	94.75	71.67	512×512
Dataset: EM											
Baseline (FANet without MixPool, **B1**)	0.9128	0.9544	**0.9597**	0.9495	0.7946	0.9263	**0.9575**	5.76M	70.38	79.59	512×512
Baseline + MixPool (**B2**)	0.9121	0.9540	0.9596	0.9488	0.7918	0.9257	0.9573	7.72M	94.75	59.15	512×512
Baseline + MixPool(E1, D4) + feedback (**B3**)	0.9042	0.9497	0.9404	**0.9594**	**0.8378**	0.9198	0.9441	5.78M	76.53	90.62	512×512
Baseline + MixPool + feedback (**B4**)	**0.9134**	**0.9547**	0.9568	0.9529	0.8096	**0.9271**	0.9559	7.72M	94.75	70.70	512×512

a 0.03% improvement on the EM dataset. These performance gains are significant, and thus, demonstrate the effectiveness of the use of MixPool block in the proposed FANet.

2) Optimum Position of MixPool Block in FANet Architecture: The positioning of the MixPool is an important factor determining the performance of the model. In the FANet (B4), we integrate the MixPool block in all the encoder blocks and the decoder blocks. In B3, we integrate the MixPool block in the first encoder block and the last decoder block only. To evaluate the effectiveness of the integrating MixPool block, we compare B3 with B4 in Table IX. It can be observed that out of the seven datasets, on three datasets, i.e., Kvasir-SEG, CVC-ClinicDB, and 2018 DSB, a significant improvement in B4 is observed

TABLE X
ALGORITHM COMPLEXITY OF OUR PROPOSED FANET AND OTHER SOTA METHODS

Method	Params. (million)	Flops (GMac)	Inf. Time (in ms.)	Image size (pixels)
U-Net [5]	31.04	219.01	3.14	512×512
ResU-Net [47]	8.22	181.68	2.93	512×512
U-Net++ [49]	9.16	138.6	4.07	512×512
Attention U-Net [9]	34.88	266.54	4.47	512×512
FANet	7.72	94.75	8.25	512×512

Params.: parameters; Inf.: inference; ms: milliseconds

as compared with the B3. On the F1 metric, we can observe that B4 achieves an improvement of 3.43% on Kvasir-SEG, 1.93% on CVC-ClinicDB, 0.11% on the 2018 DSB, and 0.5% on the EM dataset.

3) Significance of Feedback During Evaluation: The proposed architecture uses the feedback information (input mask) while training. This feedback mechanism is also used during the evaluation for iterative pruning of the predicted mask. To evaluate the effectiveness of the feedback mechanism, we compare the B2 (FANet without feedback) with the B4 (FANet with feedback) in Table IX. On all the datasets, we used feedback during inference and compared its performance with the model without feedback. We can observe that the majority of performance gains in mIoU and F1. For Kvasir-SEG, B4 shows a 17.75% improvement in the mIoU, 15.01% improvement in the F1, and a 20.76% improvement in the recall. Likewise, on the CVC-ClinicDB, we can see that B4 has 3.96% improvement in mIoU and 2.27% improvement in the F1.

E. Algorithm Efficiency

We have analyzed the algorithm efficiency in terms of the number of parameters, flops, and inference time for SOTA methods and FANet (see Table X). During the architectural design, we limit the number of trainable parameters in order to minimize the computational cost of our model. The proposed FANet has only 7.72 million parameters and 94.75 GMac flops, i.e., FANet has the least number of parameters and flops as compared with other deeper architectures. However, our inference time is higher than the other baseline networks, which is due to the introduction of a novel MixPool block in the FANet that incorporates additional operations, such as elementwise multiplication from the readout of the RLE encoded mask that resulted in larger computational time. However, in terms of FPS per iteration this is still above 60 (see Table IX). In the FANet, the MixPool block facilitates attention and propagation of information flow from the current learning paradigm and that of the previous epoch, which helps to achieve a performance boost (refer Table IX). To verify the efficiency of the MixPool block, we have compared our network with and without the MixPool block in Table IX. It is also evident that removing the MixPool block reduces the overall performance in all datasets.

F. Extended Ablation Study

We have performed an extended ablation study to demonstrate the architectural effectiveness of the proposed FANet.

TABLE XI
EXTENDED ABLATION STUDY DEMONSTRATING EFFECTIVENESS OF OUR PROPOSED FANET ARCHITECTURE ON THE KVASIR-SEG DATASET. HERE, WE ABLATE OUR NETWORK USING DIFFERENT CONFIGURATION THAT INCLUDES: 1) REMOVING F_l OF MIXPOOL BLOCK IN FANET [FIG. 2(b)]; 2) REMOVING AND ADDING SE-RESIDUAL NETWORKS IN FANET [FIG. 2(c)]; AND 3) SERIES CONCATENATION OF FANET IN CONTRAST TO ITERATIVE MECHANISM

Method	F1	mIoU	Recall	Prec.
MixPool block (w/o F_l)	0.6527	0.5523	0.7175	0.7255
One SE block only	0.7641	0.6810	**0.9531**	0.7133
Three SE blocks	0.8570	0.7908	0.8974	0.8768
Four SE blocks	0.8158	0.7417	0.8528	0.8605
Four FANet (in series)	0.8682	0.8004	0.8994	0.8913
Two SE in FANet (Ours)	**0.8803**	**0.8153**	0.9058	**0.9005**

Here, we begin with the experimental verification of the MixPool block by removing its certain components. From Table XI, we can observe a performance drop, F1 drops by 22.76% and mIoU drops by 26.3% when feature map F_l is not used during feature concatenation. In order to justify the use of two SE-residual blocks in our proposed FANet, we conducted studies by removing and adding SE-residual blocks from the FANet. We first began by removing an SE-residual block for which we observed an 11.62% drop in F1 and 13.43% drop in mIoU.

Furthermore, we modified the FANet architecture by adding three SE-residual blocks and we observed again a decrease in the performance. For this case, F1 drops by 2.33% and mIoU by 2.45%. Next, we added one more SE-residual block and a severe performance drop can be observed. The F1 dropped by 6.45% and the mIoU drops by 7.36%. In our proposed architecture, we used iterative pruning. However, we experimented with an alternative strategy by concatenating the four FANet together in a series. From this experiment, we observed a drop of F1 by 1.21% and 1.49% drop in mIoU with nearly four times increase in the number of trainable parameters.

V. DISCUSSION

While deep learning semantic segmentation has been widely implemented, to the best of our knowledge, only direct inference strategies have been published to date. In this work, we utilize a segmentation map pruning mechanism that demonstrates a clear advantage over the current state-of-the-art models due to its ability to self-rectify the predicted mask during the evaluation process (see Tables II–VIII). The process of self-rectification or iterative pruning helps to improve the performance of the proposed FANet architecture. This improvement is due to the feedback provided by the input mask in the MixPool block which is further validated from our two ablation studies (Tables IX and XI). Furthermore, a joint configuration together with mask and the feature embeddings allow learning to achieve better feature representation of target regions and learning to adjust weights dependent on the input mask. This establishes an effective pruning mechanism of the network enabling input mask to be steered in the direction of the relevant learned features of the network. In addition, it can

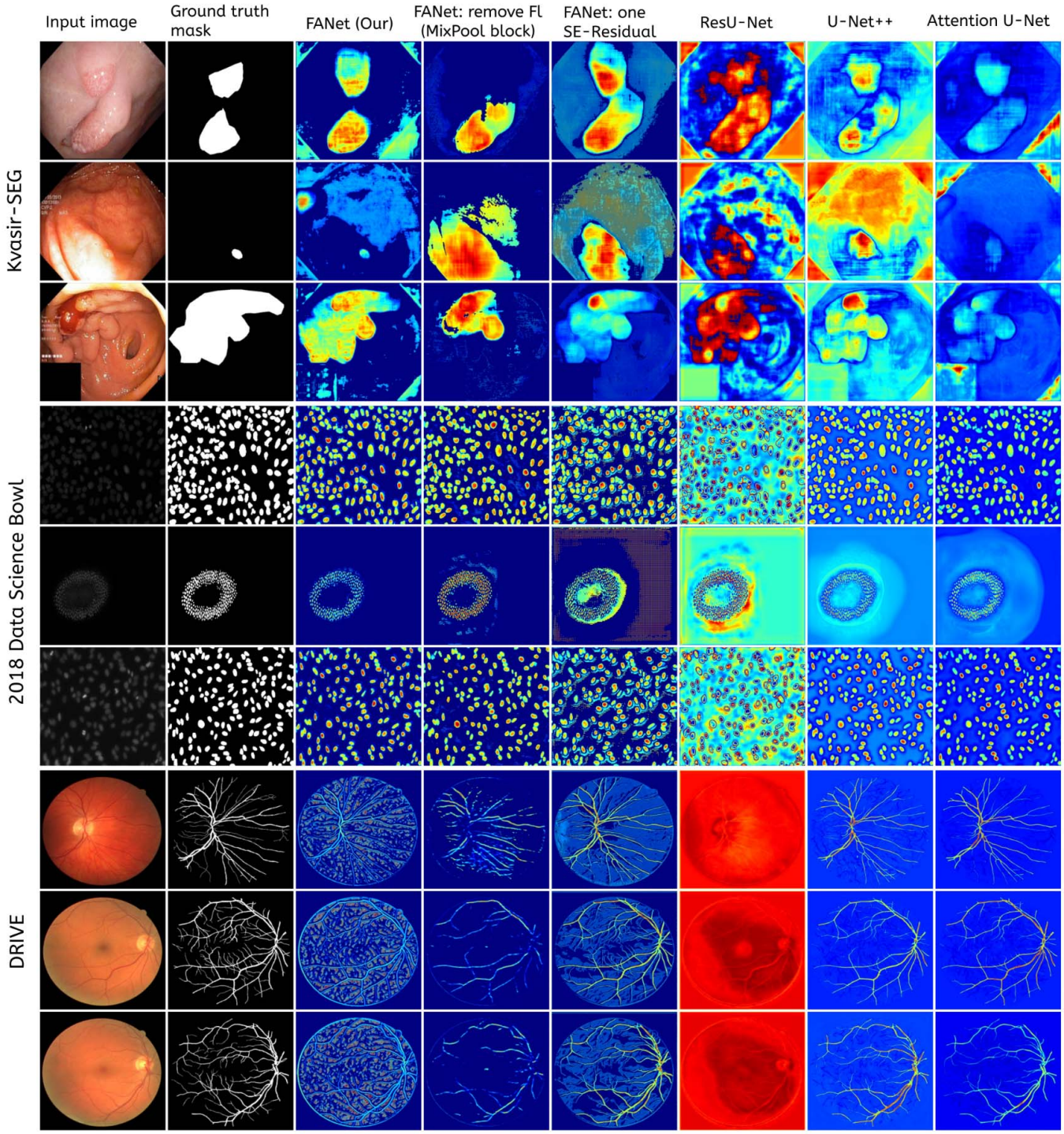

Fig. 4. Comparison of the intermediate feature map of the different networks on the Kvasir-SEG, 2018 DSB, and DRIVE datasets. For each dataset, we have included three diverse images. The provided heat maps demonstrate the impact of the weights for different networks. Here, red and yellow regions in the heat map refer to the most important features, and the blue region refers to the region of less importance. From the heat map, it can be observed that FANet has a better feature representation than other baseline networks for most of the datasets. F_l represents the input feature map in the MixPool block (refer Fig. 2).

capture the variability in datasets (e.g., shape distributions, surface morphology, and so on), allowing the network to rectify the predicted/input masks.

Table IX shows the complete ablation study of the MixPool block in the FANet architecture. In this ablation, we provide experimental results with (proposed network, B4) and without the MixPool block (B1). Here, B1 refers to the "no feedback mechanism" as no MixPool block is applied. In the proposed FANet, we require the MixPool block to provide feedback through a unified attention mechanism taking into account the network feature map and the input mask from the previous epoch [refer to Fig. 2(b)]. However, for the MixPool block without feedback (i.e., B2), we provide the attention from the generated feature map and the input mask but we do not perform the iterative pruning during the evaluation. Thus, even though B2 and B4 networks have the same number of

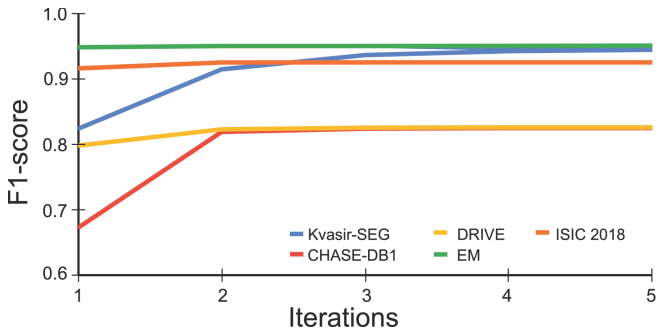

Fig. 5. Iterative pruning on different dataset images.

parameters (7.72 million parameters), the removal of the feedback mechanism affects the algorithm performance (see Table IX). SE-residual blocks that serve as a self-attention mechanism on feature channels by performing global average pooling followed by multilayer perceptron that allows to explicitly model the interdependencies between feature channels. Furthermore, in our network, we introduce spatial attention mechanisms. Multiple SE-residual blocks allow learning complex nonlinear feature interdependencies (also see Table XI for different combinations of SE-residual blocks). Furthermore, other ablation experiments, such as series concatenation of FANet, and the removal of F_l layer in MixPool block in Table XI showed that the proposed FANet achieves the highest performance. This justifies the importance of different components integrated into the proposed FANet architecture. Furthermore, qualitative results in Fig. 4 demonstrate the effectiveness of our network over different configurations, for example, removing of F_l layer in the MixPool block and using only one SE-residual block. In addition, it can be observed that FANet has more apparent segmentation maps, which are easily distinguishable regions from the background, than the SOTA methods.

With the introduction of the iterative pruning in our FANet architecture, we introduce a new hyperparameter, i.e., the number of iterations during the evaluation. The optimal number of iterations is 10, which was empirically established across datasets. The computed number of iterations is the same for all datasets. For this, we have plotted a graph (Fig. 5) showing the iterative pruning on different dataset images. From the graph, it is observed that there is a significant improvement from iteration 1 to 5. However, from 5 to 10 iterations, there is a minor to negligible improvement. Thus, we considered the highest of 10 iterations during the evaluation. The iterative pruning over the input image increases the inference time. However, this process allows us to refine the predicted segmentation masks, unlike most current methods. For obtaining a better tradeoff between efficiency and accuracy, we advise using a lesser number of iterations. We plot the F1 score for different dataset images for five iterations during evaluation. Fig. 5 shows that our proposed FANet benefits with just two iterations. In addition, we have used the NVIDIA GTX 1050 Ti (released in 2016) for inference, and thus using a more recent GPU with higher performance can provide better inference time.

VI. Conclusion

With the FANet architecture, we proposed a novel approach for biomedical image segmentation that can self-rectify the predicted masks. By introducing a feedback mechanism, we achieved an improvement on seven publicly available biomedical datasets when compared with existing SOTA methods. Our approach requires far fewer epochs for training and is well-suited to diverse biomedical imaging datasets. The feedback mechanism integrated into the FANet design effectively acts as hard attention that is used with the existing feature maps to boost the strength of feature representations. The experimental results demonstrate that the proposed architecture achieves accurate and consistent segmentation results across several biomedical imaging datasets despite its simple and straightforward network architecture. The ablation study also reveals that FANet requires less training time to achieve near SOTA performance. In the future, we will use a contrastive learning approach to improve the performance of FANet further and test it on additional multimodal biomedical images.

Acknowledgment

The computations in this article are performed on equipment provided by the Experimental Infrastructure for Exploration of Exascale Computing (eX3). The views expressed are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research, or the Department of Health.

References

[1] 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. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[2] X. Sun, B. Wang, Z. Wang, H. Li, H. Li, and K. Fu, "Research progress on few-shot learning for remote sensing image interpretation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2387–2402, 2021.

[3] Q. He, X. Sun, Z. Yan, and K. Fu, "DABNet: Deformable contextual and boundary-weighted network for cloud detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[4] A. Cardona *et al.*, "An integrated micro- and macroarchitectural analysis of the *Drosophila* brain by computer-assisted serial section electron microscopy," *PLoS Biol.*, vol. 8, no. 10, Oct. 2010, Art. no. e1000502.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[6] W. Wang, K. Yu, J. Hugonot, P. Fua, and M. Salzmann, "Recurrent U-Net for resource-constrained segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2142–2151.

[7] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *CoRR*, vol. abs/1803.01271, pp. 1–15, Apr. 2018.

[8] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015, pp. 2048–2057.

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

[10] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1179–1188.

[11] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, Apr. 2019.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[13] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[14] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[15] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. Int. Workshop Deep Learn. Med. Image Anal.*, 2018, pp. 3–11.

[16] Z. Zhou *et al.*, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Dec. 2019.

[17] D.-P. Fan *et al.*, "PraNet: Parallel reverse attention network for polyp segmentation," in *Proc. MICCAI*, 2020, pp. 263–273.

[18] D. Jha *et al.*, "ResUNet++: An advanced architecture for medical image segmentation," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2019, pp. 225–230.

[19] 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. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 558–564.

[20] R. Wang, S. Chen, C. Ji, J. Fan, and Y. Li, "Boundary-aware context neural network for medical image segmentation," 2020, *arXiv:2005.00966*.

[21] T. Hassan, M. U. Akram, and N. Werghi, "Evaluation of deep segmentation models for the extraction of retinal lesions from multi-modal retinal images," 2020, *arXiv:2006.02662*.

[22] X. Sun, P. Zhang, D. Wang, Y. Cao, and B. Liu, "Colorectal polyp segmentation by U-Net with dilation convolution," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 851–858.

[23] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5669–5678.

[24] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5659–5667.

[25] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.

[26] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10502–10511.

[27] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift Für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[28] J. Schlemper *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, Apr. 2019.

[29] H. Tsuda, E. Shibuya, and K. Hotta, "Feedback attention for cell image segmentation," in *Proc. ECCVW*, 2020, pp. 365–379.

[30] E. Shibuya and K. Hotta, "Feedback U-Net for cell image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 974–975.

[31] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 802–810.

[32] A. Mosinska, P. Marquez-Neila, M. Kozinski, and P. Fua, "Beyond the pixel-wise loss for topology-aware delineation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3136–3145.

[33] P. Pandey, A. K. Tyagi, S. Ambekar, and A. P. Prathosh, "Unsupervised domain adaptation for semantic segmentation of NIR images through generative latent search," in *Proc. ECCV*, 2020, pp. 413–429.

[34] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. ECCV*, 2016, pp. 483–499.

[35] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.

[36] D. Jha *et al.*, "Kvasir-SEG: A segmented polyp dataset," in *Proc. Int. Conf. Multimedia Modeling*, 2020, pp. 451–462.

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

[38] J. C. Caicedo *et al.*, "Nucleus segmentation across imaging experiments: The 2018 data science bowl," *Nature Methods*, vol. 16, no. 12, pp. 1247–1253, 2019.

[39] 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. ISBI*, Apr. 2018, pp. 168–172.

[40] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, Dec. 2018, Art. no. 180161.

[41] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004.

[42] C. G. Owen *et al.*, "Measuring retinal vessel tortuosity in 10-year-old children: Validation of the computer-assisted image analysis of the retina (CAIAR) program," *Investigative Ophthalmol. Vis. Sci.*, vol. 50, no. 5, pp. 2004–2010, May 2009.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[44] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[45] M. Malinowski, C. Doersch, A. Santoro, and P. Battaglia, "Learning visual question answering by bootstrapping hard attention," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–20.

[46] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

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

[48] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[49] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Granada, Spain: Springer, 2018, pp. 3–11.

[50] Y. Fang, C. Chen, Y. Yuan, and K.-Y. Tong, "Selective feature aggregation network with area-boundary constraints for polyp segmentation," in *Proc. MICCAI*, 2019, pp. 302–310.

[51] C. Gilvary, N. Madhukar, J. Elkhader, and O. Elemento, "The missing pieces of artificial intelligence in medicine," *Trends Pharmacolog. Sci.*, vol. 40, no. 8, pp. 555–564, Aug. 2019.

[52] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation," 2018, *arXiv:1802.06955*.

[53] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-directional ConvLSTM U-Net with Densley connected convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–10.

[54] *Cancer Facts & Figures 2018*, American Cancer Society, Atlanta, GA, USA, 2018.

[55] Z. Yao *et al.*, "Eye3DVas: Three-dimensional reconstruction of retinal vascular structures by integrating fundus image features," in *Frontiers in Optics*. Washington, DC, USA: Optical Society of America, 2020.

[56] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "DUNet: A deformable network for retinal vessel segmentation," *Knowl.-Based Syst.*, vol. 178, pp. 149–162, Aug. 2019.

[57] L. Li, M. Verma, Y. Nakashima, H. Nagahara, and R. Kawasaki, "IterNet: Retinal image segmentation utilizing structural redundancy in vessel networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3656–3665.

Nikhil Kumar Tomar is currently pursuing the master's degree in computer application from Indira Gandhi Open University, New Delhi, India.

He is currently a part-time Research Intern with NAAMII, Kathmandu, Nepal. He was with at SimulaMet, Oslo, Norway, where he has worked on different deep learning-based biomedical image analysis problems in close collaboration with researchers. His research interests include computer vision, artificial intelligence, parallel processing, and medical image segmentation.

Debesh Jha (Member, IEEE) received the Ph.D. degree in computer science from the UiT The Arctic University of Norway, Tromsø, Norway, in 2022.

He was a Researcher with the UiT The Arctic University of Norway. He joined the Department of Radiology, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA. His research interests include medical image analysis, machine learning, deep learning, and applications in biomedical and clinical imaging.

Michael A. Riegler (Member, IEEE) received the Ph.D. degree from the University of Oslo, Oslo, Norway, in 2017.

He is currently the Chief Research Scientist with SimulaMet Oslo, and an Associate Professor with UiT The Arctic University of Norway, Tromsø. His research interests include machine learning, video analysis and understanding, image processing, image retrieval, crowd sourcing, social computing, and applications of artificial intelligence in medicine.

Dr. Riegler is a member of the Young Academy of Norway.

Håvard D. Johansen (Member, IEEE) received the Ph.D. degree in computer science from the University of Tromsø, Tromsø, Norway, in 2007, with an emphasis on secure group membership protocols, decentralized access control, and multimedia processing.

He is currently as an Associate Professor with the Department of Computer Science, UiT The Arctic University of Norway, Tromsø. His current research interest includes information-flow control methods for enforcing security, privacy, and regulatory compliance in data-driven AI systems.

Dag Johansen (Member, IEEE) is currently a Full Professor with the Department of Computer Science, UiT The Arctic University of Norway, Tromsø, Norway, where he is addressing fundamental distributed system problems fostering a principled approach to the design and development of robust and efficient distributed systems. He is exploring interdisciplinary research problems in disparate domains, such as clinical medicine, elite athlete performance development, and sustainable commercial fisheries.

Dr. Johansen is an Elected Member of the Norwegian Academy of Technological Sciences.

Jens Rittscher was a Senior Research Scientist and a Manager with GE Global Research, Niskayuna, NY, USA. He is currently a Professor of engineering science with the University of Oxford, Oxford, U.K., with his appointment held jointly between the Institute of Biomedical Engineering, Oxford, and the Nuffield Department of Medicine, Oxford. He is also the Group Leader with the Big Data Institute, Oxford, and an Adjunct Member with the Ludwig Institute for Cancer Research, Oxford, and the Wellcome Centre for Human Genetics, Oxford. His research interests include enabling biomedical imaging through the development of new algorithms and novel computational platforms, with a current focus to improve mechanistic understanding of cancer and patient care through quantitative analysis of image data.

Dr. Rittscher serves on the Executive Committee of *Medical Image Analysis* and the Editorial Board of *Biological Imaging*.

Pål Halvorsen (Member, IEEE) is currently the Chief Research Scientist with SimulaMet, Oslo, Norway, a Full Professor with the Department of Computer Science, Oslo Metropolitan University, Oslo, and an Adjunct Professor with the Department of Informatics, University of Oslo, Oslo. His research interests include distributed multimedia systems, including operating systems, processing, storage, retrieval, communication, distribution, and content analysis.

Sharib Ali (Member, IEEE) received the Ph.D. degree from the University of Lorraine, Nancy, France, in 2016.

He was a Post-Doctoral Researcher with the Biomedical Computer Vision Group and German Cancer Research Center (DKFZ), Heidelberg University, Heidelberg, Germany. He is currently with the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, U.K. He is a Founding Member of NAAMII, Kathmandu, Nepal, and leads the Computational Endoscopy and Surgery Group. His current research interests include developing machine learning techniques for detecting and diagnosing cancer, early cancer precursors and facilitating therapeutic and surgical interventions, and solving a wide range of biomedical problems through computer vision and medical image analysis.