MultiResUNet : Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation

Nabil Ibtehaz[1] and M. Sohel Rahman[1,*]

[1]Department of CSE, BUET,
ECE Building, West Palasi, Dhaka-1205, Bangladesh
[*]Corresponding author

1017052037@grad.cse.buet.ac.bd , msrahman@cse.buet.ac.bd

February 12, 2019

Abstract

In recent years Deep Learning has brought about a breakthrough in Medical Image Segmentation. U-Net is the most prominent deep network in this regard, which has been the most popular architecture in the medical imaging community. Despite outstanding overall performance in segmenting multimodal medical images, from extensive experimentations on challenging datasets, we found out that the classical U-Net architecture seems to be lacking in certain aspects. Therefore, we propose some modifications to improve upon the already state-of-the-art U-Net model. Hence, following the modifications we develop a novel architecture MultiResUNet as the potential successor to the successful U-Net architecture. We have compared our proposed architecture MultiResUNet with the classical U-Net on a vast repertoire of multimodal medical images. Albeit slight improvements in the cases of ideal images, a remarkable gain in performance has been attained for challenging images. We have evaluated our model on five different datasets, each with their own unique challenges, and have obtained a relative improvement in performance of 10.15%, 5.07%, 2.63%, 1.41%, and 0.62% respectively.

Index terms— Convolutional Neural Networks, Medical Imaging, Semantic Segmentation, U-Net

1 Introduction

Since the introduction of digital medical imaging equipments, significant attention has been drawn towards applying image processing techniques in analyzing medical images. Multidisciplinary researchers have been working diligently for decades to develop automated diagnosis systems, and to this day it is one of the most active research areas [1]. The task of a computer aided medical image analysis tool is twofold: segmentation and diagnosis. In the general Semantic Segmentation problem, the objective is partitioning an image into a set of non-overlapping regions, which allows the homogeneous pixels to be clustered together [2]. However, in the context of medical images the interest often lies in distinguishing only some interesting areas of the image, like the tumor regions [3], organs [4] etc. This enables the doctors to analyze only the significant parts of the otherwise incomprehensible multimodal medical images [5]. Furthermore, often the segmented images are used to compute various features that may be leveraged in the diagnosis [6]. Therefore, image segmentation is of utmost importance and application in the domain of Biomedical Engineering.

Owing to the profound significance of medical image segmentation and the complexity associated with manual segmentation, a vast number of automated medical image segmentation methods have been developed, mostly focused on images of specific modalities. In the early days, simple rule-based approaches were followed; however, those methods failed to maintain robustness when tested on huge variety of data [7]. Consequently, more adaptive algorithms were developed relying on geometric shape priors with tools of soft-computing [8] and fuzzy algorithms [9]. Nevertheless, these methods suffer from human biases and can not deal with the amount of variance in real world data.

Recent advancements in deep learning [10] have shwon a lot of promises towards solving such issues. In this regard, Convolutional Neural Networks (CNN) [11] have been the most ground-breaking addition, which is dominating the field of Computer Vision. CNNs have been responsible for the phenomenal advancements in tasks like object classification [12], object localization [13] etc., and the continuous improvements to CNN architectures are bringing further radical progresses [14, 15, 16, 17]. Semantic Segmentation tasks have also been revolutionalized by Convolutional Networks. Since CNNs are more intuitive in performing object classification, Ciresan et al. [18] presented a sliding window based pipeline to perform semantic segmentation using CNN. Long et al. [19] proposed a fully convolutional network (FCN) to perform end-to-end image segmentation, which surpassed the existing approaches. Badrinarayanan et al. [20] improved upon FCN, by developing a novel architecture namely, SegNet. SegNet consists of a 13 layer deep encoder network that extracts spatial features from the image, and a corresponding 13 layer deep decoder network that upsamples the feature maps to predict the segmentation masks. Chen et al. [21] presented DeepLap and performed semantic segmentation using atrous convolutions.

In spite of initiating a breakthrough in computer vision tasks, a major drawback of the CNN architectures is that they require massive volumes of training data. Unfortunately, in the context of medical images, not only the acquisition of images is expensive and complicated, accurate annotation thereof adds even more to the complexity [22]. Nevertheless, CNNs have shown great promise in medical image segmentation in recent years [22, 23], and most of the credit go to U-Net [24]. The structure of U-Net is quite similar to SegNet, comprising an encoder and a decoder network. Furthermore, the correspond-

ing layers of the encoder and decoder network are connected by skip connections, prior to a pooling and subsequent to a deconvolution operation respectively. U-Net has been showing impressive potential in segmenting medical images, even with a scarce amount of labeled training data, to an extent that it has become the de-facto standard in medical image segmentation [22]. U-Net and U-Net like models have been successfully used in segmenting biomedical images of neuronal structures [24], liver [25], skin lesion [26], colon histology [27], kidney [28], vascular boundary [29], lung nodule [30], prostate [31], etc. and the list goes on.

In this paper, in parallel to appreciating the capabilities of U-Net, the most popular and successful deep learning model for biomedical image segmentation, we diligently scrutinize the network architecture to discover some potential scopes of improvement. We argue and hypothesize that the U-Net architecture may be lacking in certain criteria and based on contemporary advancements in deep computer vision we propose some alterations to it. In the sequel, we develop a novel model called MultiResUNet, an enhanced version of U-Net, that we believe will significantly advance the state of the art in the domain of general multimodal biomedical image segmentation. We put our model to test using a variety of medical images originating from different modalities, and even with 3D medical images. From extensive experimentation with this diverse set of medical images, it was found that MultiResUNet overshadows the classical U-Net model in all the cases even with slightly less number of parameters.

The contributions of this paper can be summarized as follows:

- We analyze the U-Net model architecture in depth, and conjecture some potential opportunities for further enhancements
- Based on the probable scopes for improvement, we propose MultiResUNet, which is an enhanced version of the standard U-Net architecture.
- We experiment with different public medical image datasets of different modalities, and MultiResUNet shows superior accuracy.
- We also experiment with a 3D version of MultiResUNet, and it outperforms the standard 3D U-Net as well.
- Particularly, we examine some very challenging images and observe a significant improvement in using MultiResUNet over U-Net.

2 Overview of the UNet Architecture

Similar to FCN [19] and SegNet [20], U-Net [24] uses a network entirely of convolutional layers to perform the task of semantic segmentation. The network architecture is symmetric, having an *Encoder* that extracts spatial features from the image, and a *Decoder* that constructs the segmentation map from the encoded features. The *Encoder* follows the typical formation of a convolutional network. It involves a sequence of two 3×3 convolution operations, which is followed by a max pooling operation with a pooling size of 2×2 and stride of 2. This sequence is repeated four times, and after each downsampling

the number of filters in the convolutional layers are doubled. Finally, a progression of two 3×3 convolution operations connects the *Encoder* to the *Decoder*.

On the contrary, the *Decoder* first up-samples the feature map using a 2×2 transposed convolution operation [32], reducing the feature channels by half. Then again a sequence of two 3×3 convolution operations is performed. Similar to the *Encoder*, this succession of up-sampling and two convolution operations is repeated four times, halving the number of filters in each stage. Finally, a 1×1 convolution operation is performed to generate the final segmentation map. All convolutional layers in this architecture except for the final one use the *ReLU* (Rectified Linear Unit) activation function [10]; the final convolutional layer uses a *Sigmoid* activation function.

Perhaps, the most ingenious aspect of the U-Net architecture is the introduction of skip connections. In all the four levels, the output of the convolutional layer, prior to the pooling operation of the *Encoder* is transfered to the *Decoder*. These feature maps are then concatenated with the output of the upsampling operation, and the concatenated feature map is propagated to the successive layers. These skip connections allow the network to retrieve the spatial information lost by pooling operations [33]. The network architecture is illustrated in Figure 1.

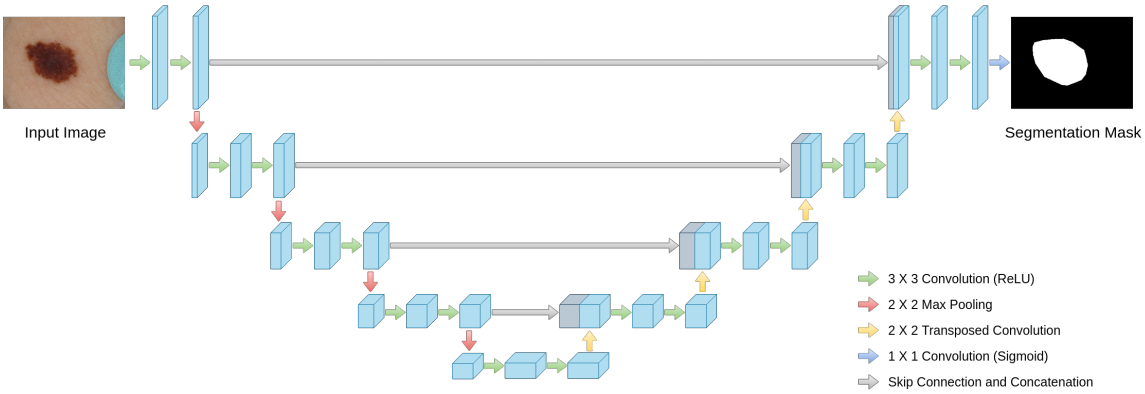


Figure 1: The U-Net Architecture. The model comprises an encoder and a decoder pathway, with skip connections between the corresponding layers.

Subsequently, the U-Net architecture was extended through a few modifications to 3D U-Net for volumetric segmentation [28]. In particular, the two dimensional convolution, max pooling, transposed convolution operations were replaced by their three dimensional counterparts. However, in order to limit the number of parameters, the depth of the network was reduced by one. Moreover, the number of filters were doubled before the pooling layers to avoid bottlenecks [34]. The original U-Net [24] did not use batch normalization [35], however, they were experimented with in the 3D U-Net and astonishingly the results revealed that batch normalization may sometime even hurt the performance [28].

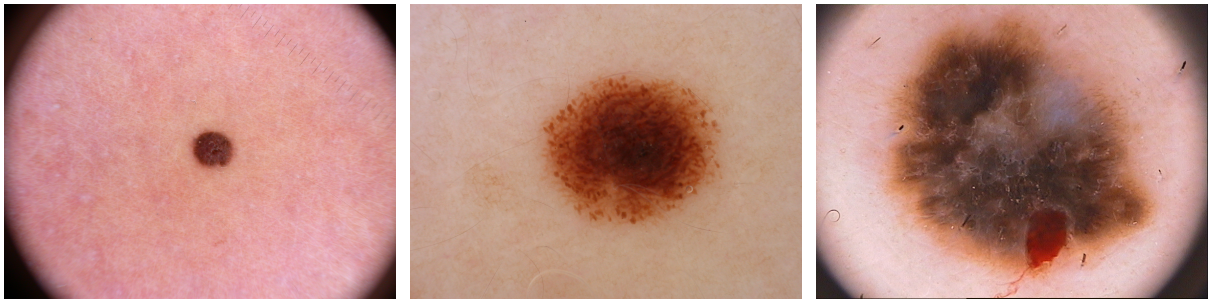
3 Motivations and High Level Considerations

U-Net has been a remarkable and the most popular Deep Network Architecture in medical imaging community, defining the state of the art in medical image segmentation [33].

However, thorough contemplation of the U-Net architecture and drawing some parallels to the recent advancement of deep computer vision leads to some useful observations as described in the following subsections.

3.1 Variation of Scale in Medical Images

In medical image segmentation, we are interested in segmenting cell necluei [36], organs [4], tumors [3] etc. from images originating from various modalities. However, in most cases these objects of interest are of irregular and different scales. For example, in Figure 2 we have demonstrated that the scale of skin lesions can greatly vary in dermoscopy images. These situations frequently occur in different types medical image segmentation tasks.



(a) (b) (c)

Figure 2: Variation of Scale in medical images. Fig. 2a, 2b, 2c are examples of dermoscopy images with small, medium and large size of lesions respectively. The images have been taken from the ISIC-2018 dataset.

Therefore, a network should be robust enough to analyze objects at different scales. Although this issue has been addressed in several deep computer vision works, to the best of our knowledge, this issue is still not addressed properly in the domain of medical image segmentation. Serre et al. [37] employed a sequence of fixed Gabor filters of varying scales to acknowledge the variation of scale in the image. Later on, the revolutionary Inception architecture [15] introduced Inception blocks, that utilizes convolutional layers of varying kernel sizes in parallel to inspect the points of interest in images from different scales. These perceptions obtained at different scales are combined together and passed on deeper into the network.

In the U-Net architecture, after each pooling layer and transposed convolutional layer a sequence of two 3×3 convolutional layers are used. As explained in [34], this series of two 3×3 convolutional operation actually resembles a 5×5 convolutional operation. Therefore, following the approach of Inception network, the simplest way to augment U-Net with multi-resolutional analysis is to incorporate 3×3, and 7×7 convolution operations in parallel to the 5×5 convolution operation, as shown in Figure 3a.

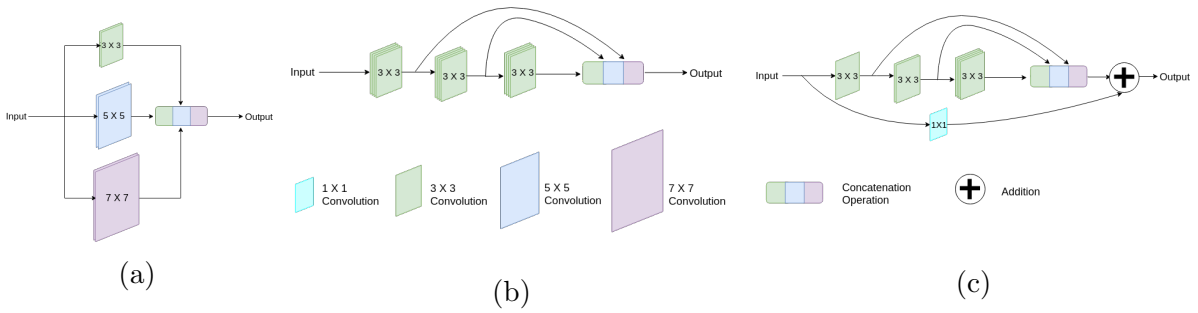
Therefore, replacing the convolutional layers with Inception-like blocks should facilitate the U-Net architecture to reconcile the features learnt from the image at different scales. Another possible option is to use strided convolutions [38], but in our experiments it was overshadowed by U-Net using Inception-like blocks. Despite the gain in performance, the introduction of additional convolutional layers in parallel extravagantly

In U-Net: 2 3x3 convolution = 1 5x5 convolution

81 x 81 $\frac{81-3}{1}+1 = 79$ 77

increases the memory requirement. Therefore, we improvise with the following ideas borrowed from [34]. We factorize the bigger, more demanding 5×5 and 7×7 convolutional layers, using a sequence of smaller and lightweight 3×3 convolutional blocks, as shown in Figure 3b. The outputs of the 2nd and 3rd 3×3 convolutional blocks effectively approximate the 5×5 and 7×7 convolution operations respectively. We hence take the outputs from the three convolutional blocks and concatenate them together to extract the spatial features from different scales. From our experiments, it was seen that the results of this compact block closely resemble that of the memory intensive Inception-like block decsribed earlier. This outcome is in line with the findings of [34], as the adjacent layers of a vision network are expected to be correlated.

Despite that this modification greatly reduces the memory requirement, it is still quite demanding. This is mostly due to the fact that in a deep network if two convolutional layers are present in a succession, then the number of filters in the first one has a quadratic effect over the memory [15]. Therefore, instead of keeping all the three consecutive convolutional layers of equal number of filters, we gradually increase the filters in those (from 1 to 3), to prevent the memory requirement of the earlier layers from exceedingly propagating to the deeper part of the network. We also add a residual connection because of their efficacy in biomedical image segmentaion [33], and for the introduction of the 1×1 convolutional layers, which may allow us to comprehend some additional spatial information. We call this arrangement a '*MultiRes block*', as shown in Figure 3c.



(a) (b) (c)

Figure 3: Devolping the Proposed *MultiRes* block. We start with a simple Inception like block by using 3×3, 5×5 and 7×7 convolutional filters in parallel and concatenating the generated feature maps (Fig. 3a). This allows us to reconcile spatial features from different context size. Instead of using the 3×3, 5×5 and 7×7 filters in parallel, we factorize the bigger and more expensive 5×5 and 7×7 filters as a succession of 3×3 filters (Fig. 3b) . Fig 3c illustrates the *MultiRes* block, where we have increased the number of filters in the successive three layers gradually and added a residual connection (along with 1×1 filter for conserving dimensions).

3.2 Probable Semantic Gap between the Corresponding Levels of Encoder-Decoder

An ingenious contribution of the U-Net architecture was the introduction of shortcut connections between the corresponding layers before and after the max-pooling and the deconvolution layers respectively. This enables the network to propagate from encoder to decoder, the spatial information that gets lost during the pooling operation.

Despite preserving the dissipated spatial features, a flaw of the skip connections may be speculated as follows. For instance, the first shortcut connection bridges the encoder

before the first pooling with the decoder after the last deconvolution operation. Here, the features coming from the encoder are supposed to be lower level features as they are computed in the earlier layers of the network. On the contrary, the decoder features are supposed to be of much more higher level, since they are computed at the very deep layers of the network. Therefore, they go through more processing. Hence, we observe a possible semantic gap between the two sets of features being merged. We conjecture that the fusion of these two arguably incompatible sets of features could cause some discrepancy throughout the learning and thus adversely affect the prediction procedure. It may be noted that, the amount of discrepancy is likely to decrease gradually as we move towards the succeeding shortcut connections. This can be attributed to the fact that, not only the features from the encoder are going through more processing, but also we are fusing them with decoder features of much juvenile layers.

Therefore, to alleviate the disparity between the encoder-decoder features, we propose to incorporate some convolutional layers along the shortcut connections. Our hypothesis is that these additional non-linear transformations on the features propagating from the encoder stage should account for the further processing done during the by decoder stage therein. Furthermore, instead of using the usual convolutional layers we introduce residual connections to them as they make the learning easier [17] and is proven to be having great potential in medical image analysis [33]. This idea is inspired from the image to image conversion using convolutional neural networks [39], where pooling layers are not favorable for the loss of information. Thus, instead of simply concatenating the feature maps from the encoder stage to the decoder stage, we first pass them through a chain of convolutional layers with residual connections, and then concatenate with the decoder features. We call this proposed shortcut path '*Res path*', illustrated in Fig. 4. Specifically, 3×3 filters are used in the convolutional layers and 1×1 filters accompany the residual connections.

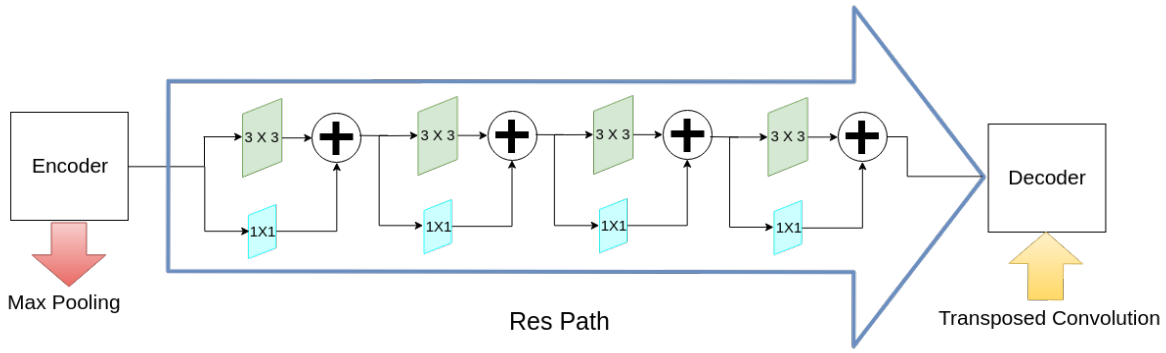


Figure 4: Proposed *Res* path. Instead of combining the the encoder feature maps with the decoder feature in a straight-forward manner, we pass the encoder features through a sequence of convolutional layers. These additional non-linear operations are expected to reduce the semantic gap between encoder and decoder features. Furthermore, residual connections are also introduced as they make the learning easier and are very useful in deep convolutional networks.

4 Proposed Architecture

In the MultiResUNet model, we replace the sequence of two convolutional layers with the proposed *MultiRes* block as introduced in Section 3.2. For each of the *MultiRes* blocks, we assign a parameter W, which controls the number of filters of the convolutional layers inside that block. To maintain a comparable relation between the numbers of parameters in original U-Net and the proposed model, we compute the value of W as follows:

$$W = \alpha \times U \tag{1}$$

Here, U is the number of filters in the corresponding layer of the U-Net and α is a scaler coefficient. Decomposing W to U and α provides a convenient way to both controlling the number of parameters and keeping them comparable to U-Net. We compare our proposed model with an U-Net, having $\#filters = [32, 64, 128, 256, 512]$ along the levels, which are also the values of U in our model. We selected $\alpha = 1.67$ as it keeps the number of parameters in our model slightly below that of the U-Net.

In Section 3.2, we pointed out that it is beneficial to gradually increase the number of filters in the successive convolutional layers inside a *MultiRes* block, instead of keeping them the same. Hence, we assign $\lfloor \frac{W}{6} \rfloor$, $\lfloor \frac{W}{3} \rfloor$ and $\lfloor \frac{W}{2} \rfloor$ filters to the three successive convolutional layers respectively, as this combination achieved the best results in our experiments. Also it can be noted that similar to the U-Net architecture, after each pooling or deconvolution operation the value of W gets doubled.

In addition to introducing the *MultiRes* blocks, we also replace the ordinary shortcut connections with the proposed *Res* paths. Therefore, we apply some convolution operations on the feature maps propagating from the encoder stage to the decoder stage. In Section 3.1, we hypothesized that the intensity of the semantic gap between the encoder and decoder feature maps are likely to decrease as we move towards the inner shortcut paths. Therefore, we also gradually reduce the number of convolutional blocks used along the *Res* paths. In particular, we use $4, 3, 2, 1$ convolutional blocks respectively along the four *Res* paths. Also, in order to account for the number of feature maps in encoder-decoder, we use $32, 64, 128, 256$ filters in the blocks of the four *Res* paths respectively.

All the convolutional layers except for the output layer, used in this network is activated by the *ReLU* (Rectified Linear Unit) activation function [10], and are batch-normalized [35]. Similar to the U-Net model, the output layer is activated by a *Sigmoid* activation function. We present a diagram of the proposed MultiResUNet model in Fig. 5. The architectural details are described in Table 1.

5 Datasets

Curation of medical imaging datasets are challenging compared to the traditional computer vision datasets. Expensive imaging equipments, sophisticated image acquisition pipelines, necessity of expert annotation, issues of privacy-all adds to the complexity of developing medical imaging datasets [22]. As a result, only a few public medical imaging benchmark datasets exist, and they only contain a handful of images each. In order to assess the efficacy of the proposed architecture, we tried to evaluate it on a variety of image modalities. More specifically, we selected datasets that are as heterogeneous as possible

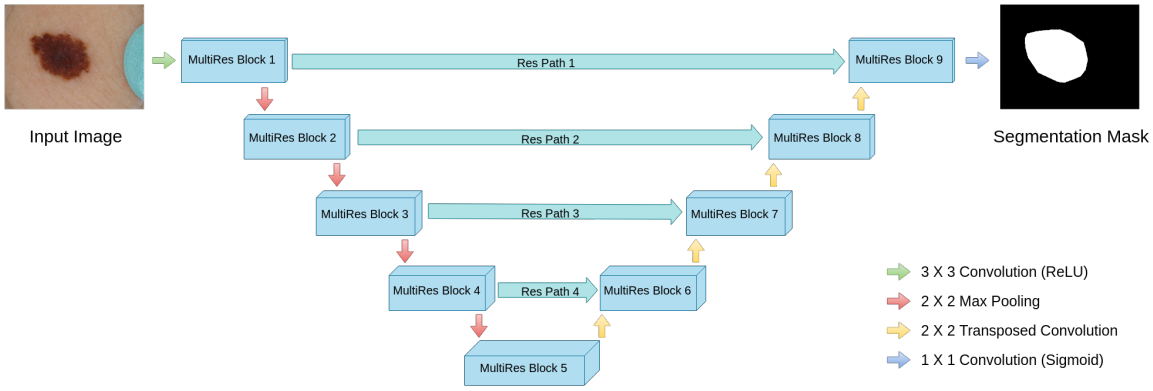


Figure 5: Proposed MultiResUNet architecture. We replace the sequences of two convolutional layers in the U-Net architectures with the proposed *MultiRes* block. Furthermore, instead of using plain shortcut connections, we use the proposed *Res* paths.

from each other. Also, each of these datasets poses a unique challenge of their own (more details are given in Section 7 and Section 8). The datasets used in the experiments are briefly described below (also see Table 2 for an overview).

5.1 Fluorescence Microscopy Image

We used the fluorescence microscopy image dataset developed by Murphy Lab [36]. This dataset contains 97 fluorescence microscopy images and a total of 4009 cells are contained in these images. Half of the cells are U2OS cells and the other half comprises NIH3T3 cells. The nuclei were segmented manually by experts. The nuclei are irregular in terms of brigtness and the imags often contain noticeable debris, making this a challenging dataset of bright-field microscopy images. The original resolution of the images range from 1349×1030 to 1344×1024; they have been resized to 256×256 for computational constraints.

5.2 Electron Microscopy Image

To observe the effectiveness of the architecture with electron microscopy images, we used the dataset of the ISBI-2012 : 2D EM segmentation challenge [40, 41]. This dataset contains only 30 images from a serial section Transmission Electron Microscopy (ssTEM) of the Drosophila first instar larva ventral nerve cord [41]. The images face slight alignment errors, and are corrupted with noises. The resolution of the images is 512×512, but they have been resized to 256×256 due to computational limitations.

5.3 Dermoscopy Image

We acquired the dermoscopy images from the ISIC-2018 : Lesion Boundary Segmentation challenge dataset. The data for this challenge were extracted from the ISIC-2017 dataset [3] and the HAM10000 dataset [42]. The compiled dataset contains a total of 2594 images of different types of skin lesions with expert annotation. The images are of various

Table 1: MultiResUNet Architecture Details

MultiResUNet					
Block	Layer(filter size)	#filters	Path	Layer(filter size)	#filters
MultiRes Block 1	Conv2D(3,3)	8	Res Path 1	Conv2D(3,3)	32
	Conv2D(3,3)	17		Conv2D(1,1)	32
MultiRes Block 9	Conv2D(3,3)	26		Conv2D(3,3)	32
	Conv2D(1,1)	51		Conv2D(1,1)	32
MultiRes Block 2	Conv2D(3,3)	17		Conv2D(3,3)	32
	Conv2D(3,3)	35		Conv2D(1,1)	32
MultiRes Block 8	Conv2D(3,3)	53		Conv2D(3,3)	32
	Conv2D(1,1)	105		Conv2D(1,1)	32
MultiRes Block 3	Conv2D(3,3)	35	Res Path 2	Conv2D(3,3)	64
	Conv2D(3,3)	71		Conv2D(1,1)	64
MultiRes Block 7	Conv2D(3,3)	106		Conv2D(3,3)	64
	Conv2D(1,1)	212		Conv2D(1,1)	64
MultiRes Block 4	Conv2D(3,3)	71		Conv2D(3,3)	64
	Conv2D(3,3)	142		Conv2D(1,1)	64
MultiRes Block 6	Conv2D(3,3)	213	ResPath 3	Conv2D(3,3)	128
	Conv2D(1,1)	426		Conv2D(1,1)	128
MultiRes Block 5	Conv2D(3,3)	142		Conv2D(3,3)	128
	Conv2D(3,3)	284		Conv2D(1,1)	128
	Conv2D(3,3)	427	Res Path 4	Conv2D(3,3)	256
	Conv2D(1,1)	853		Conv2D(1,1)	256

resolutions, but they have all been resized to 256×192, maintaining the average aspect ratio, for computational purposes.

5.4 Endoscopy Image

We used the CVC-ClinicDB [43], a colonoscopy image database for our experiments with endoscopy images. The images of this dataset were extracted from frames of 29 colonoscopy video sequences. Only the images with polyps were considered, resulting in a total of 612 images. The images are originally of resolution 384×288, but have been resized to 256×192, maintaining the aspect ratio.

5.5 Magnetic Resonance Image

All the datasets described previously contains 2D medical images. In order to evaluate our proposed architecture with 3D medical images, we used the magnetic resonance images (MRI) from the BraTS17 competition database [44, 45]. This dataset contains 210 glioblastoma (HGG) and 75 lower grade glioma (LGG) multimodal MRI scans. These multimodal scans include native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2) and T2 Fluid Attenuated Inversion Recovery (FLAIR) volumes, which were acquired following different clinical protocols and various scanners from 19 institutions. The images are of dimensions $240 \times 240 \times 155$ but have been resized to $80 \times 80 \times 48$ for

Table 2: Overview of the Datasets.

Modality	Dataset	No. of images	Original Resolution	Input Resolution
Fluorescence Microscopy	Murphy Lab	97	Variable	256×256
Electron Microscopy	ISBI-2012	30	512×512	256×256
Dermoscopy	ISIC-2018	2594	Variable	256×192
Endoscopy	CVC-ClinicDB	612	384×288	256×192
MRI	BraTS17	210 HGG + 75 LGG	$240 \times 240 \times 155$	$80 \times 80 \times 48$

computational ease. All the four modalities, namely, T1, T1Gd, T2 and FLAIR are used as four different channels in evaluating the 3D variant of our model.

6 Experiments

We used Python, more specifically Python3 programming language to conduct the experiments [46]. The network models were implemented using Keras [47] with Tensorflow backend [48]. Our model implementation is available in the following github repository:

`https://github.com/nibtehaz/MultiResUNet`

The experimets were conducted in a desktop computer with intel core i7-7700 processor (3.6 GHz, 8 MB cache) CPU, 16 GB RAM, and NVIDIA TITAN Xp (12 GB, 1582 MHz) GPU.

6.1 Baseline Model

Since the proposed architecture, MultiResUNet, is targeted towards improving the state of the art U-Net architecture for medical image segmentation, we compared its performance with the U-Net architecture as the baseline. To keep the number of parameters comparable to our proposed MultiResUNet, we implemented the original U-Net [20] having five layer deep encoder and decoder, with filter numbers of 32, 64, 128, 256, 512.

Also, as the baseline for 3D image segmentation we used the 3D counterpart of the U-Net as described in the original paper [28]. The 3D version of the MultiResUNet is constructed simply by substituting the 2D convolutional layers, pooling layers and transposed convolution layers, with their 3D variants respectively, without any further alterations.

The number of parameters of the models are presented in Table 3. In both the cases the proposed networks require slightly lesser number of parameters.

Table 3: Models used in our experiments

2D		3D	
Model	Parameters	Model	Parameters
U-Net (baseline)	7,759,521	3D U-Net (baseline)	19,078,593
MultiResUNet (proposed)	7,262,750	MultiResUNet 3D (proposed)	18,657,689

6.2 Pre-processing / Post-processing

The objective of the experiments is to investigate the superiority of the proposed MultiResUNet architecture over the original U-Net as a general model. Therefore, no domain specific pre-processing was performed. The only pre-processing the input images went through was that they were resized to fit into the GPU memory and the pixel values were divided by 255 to bring them to the $[0 \dots 1]$ range. Similarly, no application specific post-processing was performed. Since, the final layer is activated by a Sigmoid function, it produces outputs in the range $[0 \dots 1]$. Therefore, we applied a threshold of 0.5 to obtain the segmentation map of the input images.

6.3 Training Methodology

The task of semantic segmentation is to predict the individual pixels whether they represent a point of interest, or are merely a part of the background. Therefore, this problem ultimately reduces to a pixel-wise binary classification problem. Hence, as the loss function of the network we simply took the binary cross-entropy function and minimized it.

Let, for an image X, the ground truth segmentation mask is Y, and the segmentation mask predicted by the model is $\hat{Y}$. For a pixel px, the network predicts $\hat{y}_{px}$, whereas, the ground truth value is y_{px}. The binay cross-entropy loss for that image is defined as:

$$\textit{Cross Entropy}(X, Y, \hat{Y}) = \sum_{px \in X} -(y_{px} \log(\hat{y}_{px}) + (1 - y_{px}) \log(1 - \hat{y}_{px})) \tag{2}$$

For a batch containing n images the loss function J becomes,

$$J = \frac{1}{n} \sum_{i=1}^{n} \textit{Cross Entropy}(X_i, Y_i, \hat{Y}_i) \tag{3}$$

We minimized the binary cross-entropy loss, hence trained the model using the Adam optimizer [49]. Adam adaptively computes different learning rates for different parameters from estimates of first and second moments of the gradients. This idea, in fact combines the advantages of both AdaGrad [50] and RMSProp [51]; therefore Adam has been often used in benchmarking deep learning models as the default choice [52]. Adam has a number of parameters including β_1 and β_2 which control the decay of first and second moment respectively. However, in this work we used Adam with the parameters mentioned in the original paper. The models were trained for 150 epochs using Adam optimizer. The reason of selecting 150 as the number of epochs is due to the fact that after 150 epochs neither of the models were showing any further improvements.

6.4 Evaluation Metric

In semantic segmentation, usually the points of interest comprise a small segment of the entire image. Therefore, metrics like precision, recall are inadequate and often lead to false sense of superiority, inflated by the perfection of detecting the background. Hence, Jaccard Index has been widely used to evaluate and benchmark image segmentation and

object localization algorithms [2]. Jaccard Index for two sets A and B are defined as the ratio of the intersection and union of the two sets:

$$Jaccard\ Index = \frac{Intersection}{Union} = \frac{A \cap B}{A \cup B} \tag{4}$$

In our case, the set A represents the ground truth binary segmentation mask Y, and set B corresponds to the predicted binary segmentation mask $\hat{Y}$. Therefore, by taking the Jaccard Index as the metric, we not only emphasize on precise segmentation, but also penalize under-segmentation and over-segmentation.

6.5 k-Fold Cross Validation

Cross-Validation tests estimate the general effectiveness of an algorithm on an independent dataset, ensuring a balance between bias and variance. In a k-Fold cross-validation test, the dataset D is randomly split into k mutually exclusive subsets $D_1, D_2, \cdots, D_k$ of equal or near equal size [53]. The algorithm is run k times subsequently, each time taking one of the k splits as the validation set and the rest as the training set. In order to evaluate the segmentation accuracy of both the baseline U-Net and proposed MultiResUNet architecture, we performed 5-Fold Cross Validation tests on each of the different datasets.

Since, this is a deep learning pipeline, the best performing result on the validation set achieved through the total number of epochs (150 in our case) performed is recorded in each run. Finally, combining the results of all the k runs gives an overall estimation of the performance of the algorithm.

7 Results

7.1 MultiResUNet Consistently Outperforms U-Net

As described in Section 5 and Section 6 to evaluate the performance of the proposed architecture, we performed experiments with diversified classes of medical images, each with a unique challenge of its own. In particular, we have performed 5-fold cross validation and observed the performance of our proposed MultiResUNet and the baseline, U-Net. In each run the best results obtained on the validation set through the 150 epochs performed was noted and they were summarised from the 5 runs to obtain the final result.

The results of the 5-Fold Cross Validation for both the proposed MultiResUNet model and baseline U-Net model on the different datasets are presented in Table 4. It should be noted that for better readability the fractional Jaccard Index values have been converted to percentage ratios (%).

From the table, It can be observed that our proposed model outperforms the U-Net architecture in segmenting all different types of medical images. Most notably, remarkable improvements are observed for Dermoscopy and Endoscopy images. These images tend to be a bit less uniform and often they appear confusing even to a trained eye (more details are discussed in a later section). Therefore, this improvement is of great significance. For Fluorescence Microscopy images our model also achieve a 2.6326% relative improvement over U-Net, and despite having a lesser number of parameters, still it achieves a relative

Table 4: Results of 5-fold cross validation. Here, we present the best obtained results in the five folds, of both U-Net and MultiResUNet, for all the datasets used. We also mention the relative improvement of MultiResUNet over U-Net. It should be noted that, for better readability the fractional values of Jaccard Index have been converted to percentage ratios (%).

Modality	MultiResUNet (%)	U-Net (%)	Relative Improvement (%)
Dermoscopy	80.2988 ± 0.3717	76.4277 ± 4.5183	5.065
Endoscopy	82.0574 ± 1.5953	74.4984 ± 1.4704	10.1465
Fluorescence Microscopy	91.6537 ± 0.9563	89.3027 ± 2.1950	2.6326
Electron Microscopy	87.9477 ± 0.7741	87.4092 ± 0.7071	0.6161
MRI	78.1936 ± 0.7868	77.1061 ± 0.7768	1.4104

improvement of 1.4104% for MRI images. Only for Electron Microscopy images U-Net seems to be on par with our proposed model, yet in that case the latter obtains slightly better results (relative improvement of 0.6161%).

7.2 MultiResUNet can Obtain Better Results in Less Number of Epochs

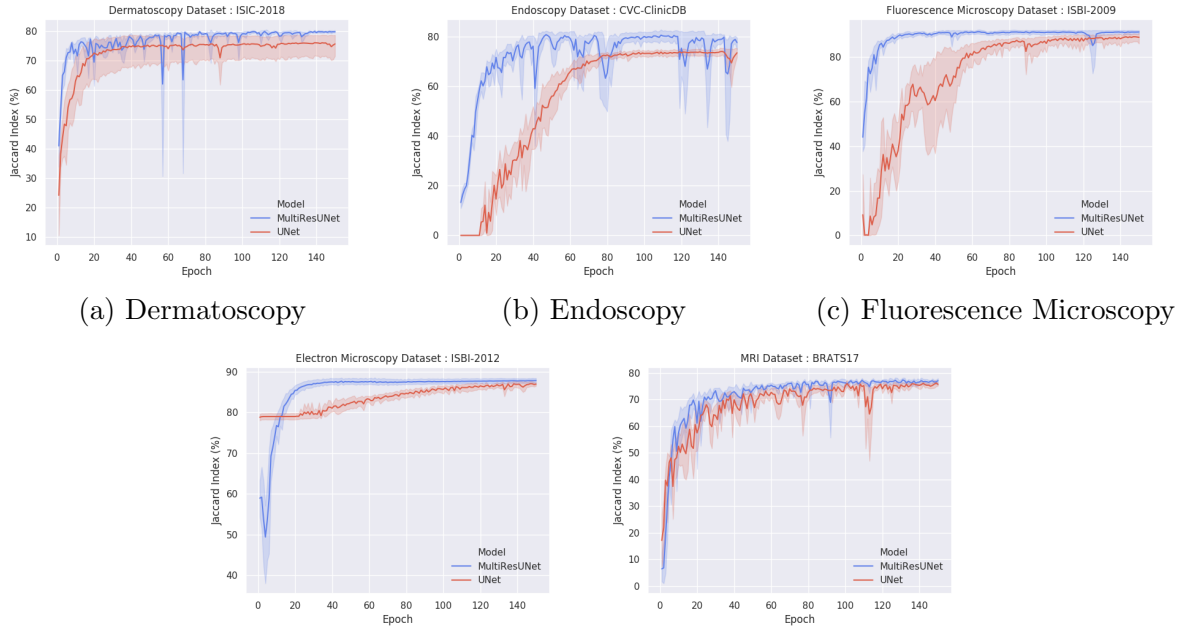
In addition to analyzing the best performing models from each run, we also monitored how the model performance progressed with epochs. In Figure 6, the performance on the validation data on each epoch is shown, for all the datasets. We have presented the band of Jaccard Index values at a certain epoch in the 5-fold cross validation. It can be noted that for all the cases our proposed model attains convergance much faster. This can be attributed to the synergy between residual connections and batch normalization [33]. Moreover, apart from Fig. 6d in all other cases the MultiResUNet model consistently outperformed the classical U-Net model. In spite of lagging behind the U-Net at the beginning for the electron microscopy images (Fig. 6d), eventually the MultiResUNet model converges at a better accuracy than U-Net. Another remarkable observation from the experiments is that except for some minor flactuations, the standard deviation of the performance of the MultiResUNet is much smaller; this indicates the reliability and the robustness of the proposed model.

These results, therefore, suggest that using the proposed MultiResUNet architecture, we are likely to obtain superior results in less number of training epochs as compared to the classical U-Net architecture.

7.3 MultiResUNet Delineates Faint Boundaries Better

Being the current state of the art model for medical image segmentation, U-Net has demonstrated quite satisfactory results in our experiments. For instance, in Fig. 7, for a polyp with clearly distinguishable boundary the U-Net model manages to segment it with a high value of Jaccard Index; our proposed model however performs better albeit only slightly.

But as we study more and more challenging images, especially with not so much conspicuous boundaries, U-Net seems to be struggling a bit (Fig. 8). The colon polyp images often suffer from the lack of clear boundaries. On such cases, the U-Net model



(a) Dermatoscopy (b) Endoscopy (c) Fluorescence Microscopy

(d) Electron Microscopy (e) MRI

Figure 6: Progress of the validation performance with the number of epochs. We record the value of Jacard Index on validation data after each epoch. It can be observed that not only MultiResUNet outperforms the U-Net model, but also the standard deviation of MultiResUNet is much smaller.

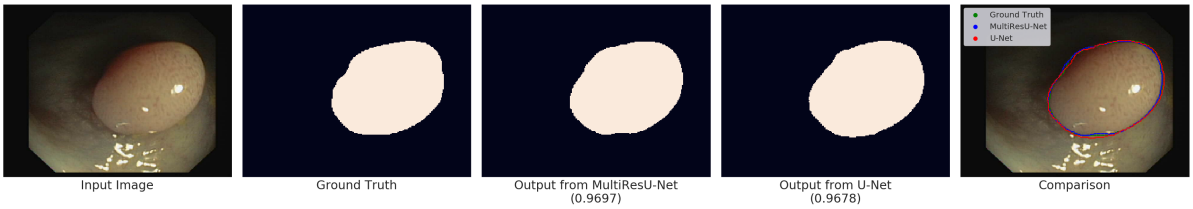
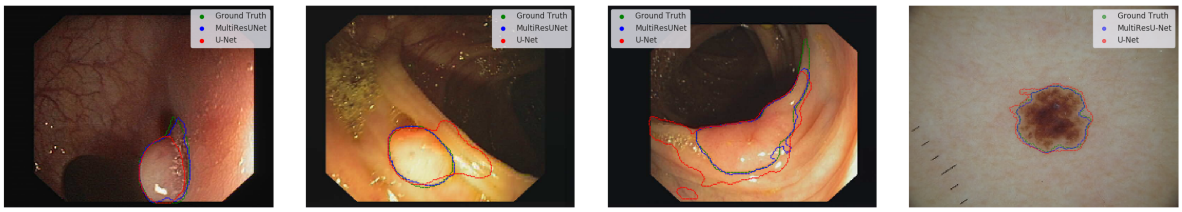


Figure 7: Segmenting a polyp with clearly visiable boundary. U-Net manages to segment the polyp with a high level of performance (J.I. = 0.9678). MultiResUNet segments just slightly better (J.I. = 0.9697)

either under-segmented (Fig. 8a) or over-segmented (Fig. 8b) the polyps. Our proposed MultiResUNet, on the other hand, performed considerably better in both the cases. However, there are some images where both the models faced complications, but in those cases MultiResUNet's performance was superior(Fig. 8c). Dermoscopic images have comparatively clearer defined boundaries; still in those cases MultiResUNet delineats the boundaries better (Fig. 8d). Same was observed for other types of images. We hypothesize that the use of multiple filter sizes allows MultiResUNet to perform better pixel perfect segmentation.

7.4 MultiResUNet is More Immune to Perturbations

The core concept of semantic segmentation is to cluster the homologous regions of an image together. However, often in real world medical images the homologous regions get deviated due to various types of noises, artifacts and irregularities in general. This



(a) (b) (c) (d)

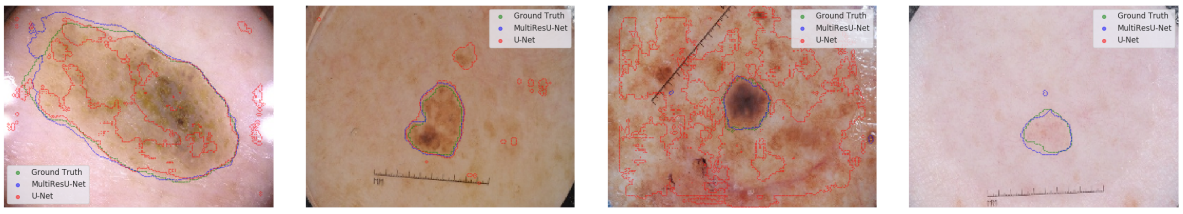
Figure 8: Segmenting images with vague boundaries. This issue is more prominent for Colon Endoscopy images. U-Net seems to either under-segment (8a), or over-segment (8b) the polyps. MultiResUNet manages to segment polyps of such situation much better. However, some images are too problematic even for MultiResUNet, but in those cases as well it performs beter than U-Net (8c). Even in dermoscopy images, where there exists a clear boundary, U-Net sometimes produced some irregularities along the boundaries, but MultiResUNet was much more robust (8d).

makes it challenging to distinguish between the region of interest and background in medical images. As a result, instead of obtaining a continuous segmented region, we are often left with a collection of fractured segmented regions. At the other extreme, due to textures and perturbations the plain background sometimes appear similar to the region of interest. These two cases lead to loss of information and false classifications respectively. Fortunately, the Dermatoscopy image dataset we have used contains images with such confusing cases, allowing us to analyze and compare the behaviour and performance of the two models.

In spite of segmenting the images with near consistent background and approximately undeviating foreground with almost perfection, the baseline U-Net model seems to struggle quite a bit in the presence of perturbations in images (Fig. 9). In images where the foreground object tend to vary a bit therein, U-Net, was unable to segment the forground as a continuous region. It rather predicted a set of scattered regions (Fig. 9a), confusing the forground as background and thus caused the loss of some valuable information. On the other hand, for images where the background is not uniform, the U-Net model seems to make some false predictions (Fig. 9b). The more rough the background becomes, the more false predictions are made (Fig. 9c). Furthermore, in some dreadfully adverse situations, where due to irregularity the difference between background and forground are too subtle, the U-Net model failed to make any predictions at all (Fig. 9d). Although in such challenging cases the segmentation of MultiResUNet is not perfect, it performs far superior than the classical U-Net model as shown in Fig. 9. It is worth noting here that in the initial stages of our experiments, prior to using the *ResPath*s, our proposed model was also being affected by such perturbations. Therefore, we conjecture that applying additional non-linear operations on the encoder feature maps makes it robust against perturbations.

7.5 MultiResUNet is More Reliable Against Outliers

Often in medical images some outliers are present, which in spite of being visually quite similar, are different from what we are interested in segmenting. Particularly, in the Fluorescence Microscopy image dataset, there exist some images with bright objects, that are apparently almost indistinguishable from the actual nuclei. Such an example is



(a) (b) (c) (d)

Figure 9: Segmenting images with irregularities. For images where the foreground is not consistent throughout, instead of segmenting it as a continuous region, U-Net seems to have predicted a set of small regions (9a). For images with rough backgrounds U-Net classified them as the foreground (9b), the more irregular the background, the more false predictions were made (9c). To the other extreme, for images where difference between foreground and background is too subtle, U-Net missed the foreground completely (9d). Though the segmentations produced by MultiResUNet in these challenging cases are not perfect, they were consistently better than that of the U-Net.

shown in Fig. 10.

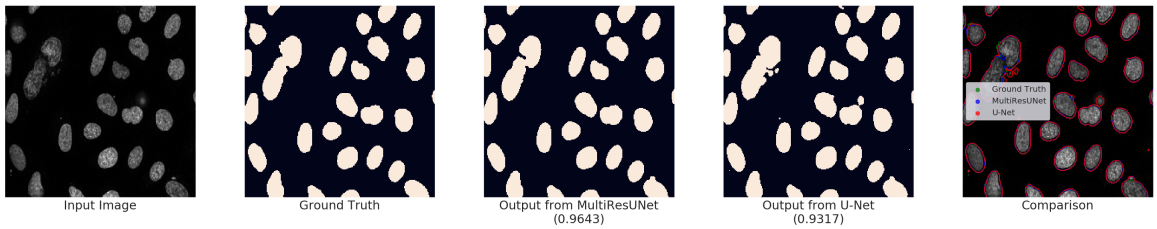


Figure 10: Segmenting images containing outliers. In the fluorescence microscopy images, the exist some bright particles, visually very similar to the cell nuclei under analysis. Although MultiResUNet can identify and reject those outliers, U-Net seems to have missclassified them.

It can be observed that the input image is infected with some small particles that are not actual cell nuclei. However, if we study the segmentation mask generated from U-Net, it turns out that U-Net has mistakenly predicted those outlier particles to be cell nuclei. On the other hand, our proposed MultiResUNet seems to be able to reject those outliers. Since the outliers are pretty tiny, false predictions made by the U-Net model does not hurt the value of Jaccard Index that much (0.9317 instead of 0.9643, when outliers are filtered out). Nevertheless, being able to segregate these outliers are of substantial significance. It can be noted that, similar types of visually similar outliers were present in other datasets as well, and MultiResUNet was able to segment the images reliably without making false predictions.

7.6 Note on Segmenting the Majority Class

The Electron Microscopy dataset we used in our experiments is quite interesting and unorthodox as in this dataset the region of interest under consideration actually comprises the majority of the images. This is a rare incident since usually the region of interest consists of a small portion of the image. This brings a different type of challenge as in such a case the models tend to over-segment the images unnecessarily to minimize the losses during training. A relevant example is presented in Fig. 11.

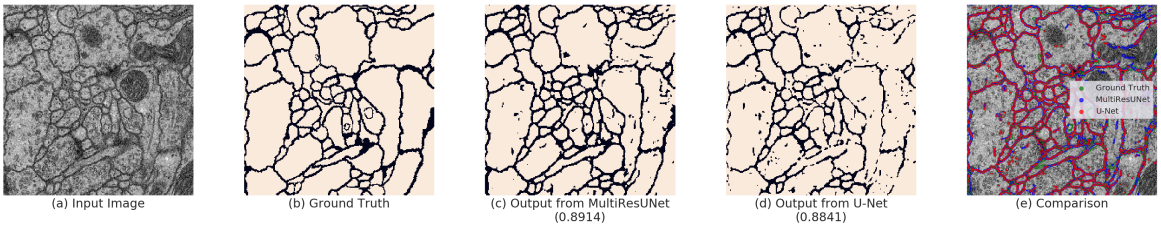


Figure 11: Segmenting the Majority Class. Here we can observe that the region of interest comprises most part of the image. Despite the values of Jacard Index for both U-Net and MultiResUNet are quite similar, visually the segmentation mask are very different. It can be seen that segmentation mask generated from MultiResUNet captures most of the fine separating lines, but U-Net tends to miss them. Moreover, there are some clusters of background pixels which although are missed by U-Net, have been roughly identified by MultiResUNet. Since the class being segmented is the majority class, the values of Jacard Index are inflated.

Here, it can be observed quite astonishingly that the majority of the image is actually forground (Fig. 11b), with some narrow separations among them by the background, i.e., membranes in this context. If we analyze the segmentation predicted by U-Net, it appears that those fine lines of separation have often been missed (Fig. 11d). MultiResUNet, on the other hand, have managed to segment the regions with properly defined separations among them (Fig. 11c). Also, it can be observed that there are some small clusters of background pixels, which have been captured with some success in the segmentation mask predicted by MultiResUNet, but are almost non-existent in the segmentation performed by U-Net. Furthermore, the result generated by MultiResUNet seems to be more immune to the noises present in the image.

Despite the two segmentations (i.e. the results of MultiResUNet and U-Net) are very different from each other, the values of the respective Jaccard Index are quite alike (0.8914 and 0.8841 as shown in Fig. 11). This is due to the fact that, the metric Jaccard Index has been inflated with the results of segmenting the majority class of the image. Therefore, the value of Jaccard Index is not that much of a proper representative of accuracy while segmenting the majority class. Despite predicting much inferior segmentations, for this reason the Jaccard Index of U-Net are very close to that of MultiResUNet. Thus, among all the different datasets, the improvement in terms of metric has been underwhelming in this dataset but the predicted segmentations are more accurate visually.

8 Conclusion

In this work, we started by analyzing the U-Net architecture diligently, with the hope of finding potential rooms for improvement. We anticipated some discrepancy between the features passed from the encoder network and the features propagating through the decoder network. To reconcile these two incompatible sets of features, we proposed *Res* paths, that introduce some additional processing to make the two feature maps more homogeneous. Furthermore, to augment U-Net with the ability of multi-resolutiona analysis, we proposed *MultiRes* blocks. We took inspirations from Inception blocks and formulated a compact analogous structure, that was lightweight and demanded less memory. Incorporating these modifications, we developed a novel architecture, MultiResUNet.

Among the handful publicly available biomedical image datasets, we selected the ones

that were drastically different from each other. Additionally each of these datasets poses a separate challenge of its own. The Murphy Lab Fluorescence Microscopy dataset is possibly the simplest dataset for performing segmentation, having an acute difference in contrast between the forground, i.e., the cell nuclei and the background, but contains some outliers. The CVC-ClinicDB dataset contains colon endoscopy images where the boundaries between the polyps and the background are so vague that often it becomes difficult to distinguish even for a trained operator. In addition, the polyps are diverse in terms of shape, size, structure, orientation etc., making this dataset indeed a challenging one. On the other hand, the dermoscopy dataset of ISIC-2018 competition contains images of poor contrast to the extent that sometimes the skin lesions seem identical to the background and vice versa. Moreover, various types of textures present in both the background and the foreground make pattern recognition quite difficult. ISBI-2012 electron microscopy dataset presents a different type of challenge. In this dataset the region being segmented covers the majority of the image; thus a tendency is observed to over-segment the images. The BraTS17 MRI dataset, on the other hand contains multimodal 3D images, which is a different problem alltogether.

For perfect or near perfect images U-Net manages to perform segmentation with remarkable accuracy. Our proposed architecture performs only slightly better than U-Net in those cases. However, for intricate images suffering from noises, perturbations, lack of clear boundaries etc., the gain in performance by MultiResUNet dramatically increases. More specifically, for the five datasets a relative improvement in performance of 10.15%, 5.07%, 2.63%, 1.41%, and 0.62% were observed in using MultiResUNet over U-Net (Table 4). Not only the segmentations generated by MultiResUNet attain higher score in the evaluation metric, they are also visually more similar to the ground truth. Furthermore, on the very challenging images U-Net tended to over-segment, under-segment, make false predictions and even miss the objects completely. On the contrary, in the experiments MultiResUNet appeared to be more reliable and robust. MultiResUNet managed to detect even the most subtle boundaries, was resilient in segmenting images with a lot of perturbations, and was rejectable to the outliers. Even in segmenting the majority class, where the U-Net tended to over-segment, MultiResUNet managed to capture the fine details. Furthermore, the 3D adaptation of MultiResUNet performed better than 3D U-Net, which is not just a straightforward 3D implementation of the U-Net, but an enhanced and improved version. It should be noted that, the segmentations generated by the proposed MultiResUNet were not perfect, but in most of the cases it outperformed the classical U-Net by a large margin.

Therefore, we believe our proposed MultiResUNet architecture can be the potential successor to the classical U-Net architecture. The future direction of this research has several branches. In this work, we have been motivated to keep the number of parameters of our model comparable to that of the U-Net model. However, in future we wish to conduct experiments to determine the best set of hyperparameters for the model more exhaustively. Moreover, we would like to evaluate our model performance on medical images originating from other modalities as well. Furthermore, we are interested in experimenting by applying several domain and application specific pre-processing and post-processing schemes to our model. We believe fusing our model to a domain specific expert knowledge based pipeline, and coupling it with proper post-processing stages will improve our model performance further, and allow us to develop better segmentation

methods for diversified applications.

Acknowledgements

The Titan Xp GPU used for this research was the generous donation of NVIDIA Corporation.

Supplementary information

Supplementary Material 1: contains the links to the weights and parameters of the best performing models in each fold for all the datasets.
Supplementary Material 2: provides the random data splits of the datasets obtained using standard methods of Scikit-learn [54].
Supplementary Material 3: presents the detailed experimental results.

References

[1] Johannes Schindelin, Curtis T Rueden, Mark C Hiner, and Kevin W Eliceiri. The imagej ecosystem: an open platform for biomedical image analysis. *Molecular reproduction and development*, 82(7-8):518–529, 2015.

[2] Kevin McGuinness and Noel E O'connor. A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, 43(2):434–444, 2010.

[3] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, 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 *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 168–172. IEEE, 2018.

[4] Jinzhong Yang, Harini Veeraraghavan, Samuel G Armato III, Keyvan Farahani, Justin S Kirby, Jayashree Kalpathy-Kramer, Wouter van Elmpt, Andre Dekker, Xiao Han, Xue Feng, et al. Autosegmentation for thoracic radiation treatment planning: A grand challenge at aapm 2017. *Medical physics*, 2018.

[5] Shivang Naik, Scott Doyle, Shannon Agner, Anant Madabhushi, Michael Feldman, and John Tomaszewski. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pages 284–287. IEEE, 2008.

[6] Rahimeh Rouhi, Mehdi Jafari, Shohreh Kasaei, and Peiman Keshavarzian. Benign and malignant breast tumors classification based on region growing and cnn segmentation. *Expert Systems with Applications*, 42(3):990–1002, 2015.

[7] Dzung L Pham, Chenyang Xu, and Jerry L Prince. Current methods in medical image segmentation. *Annual review of biomedical engineering*, 2(1):315–337, 2000.

[8] Pablo Mesejo, Andrea Valsecchi, Linda Marrakchi-Kacem, Stefano Cagnoni, and Sergio Damas. Biomedical image segmentation using geometric deformable models and metaheuristics. *Computerized Medical Imaging and Graphics*, 43:167–178, 2015.

[9] Yuhui Zheng, Byeungwoo Jeon, Danhua Xu, QM Wu, and Hui Zhang. Image segmentation by generalized hierarchical fuzzy c-means algorithm. *Journal of Intelligent & Fuzzy Systems*, 28(2):961–973, 2015.

[10] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[11] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[13] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.

[18] Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851, 2012.

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[20] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[21] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[22] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[23] Syed Muhammad Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, and Muhammad Khurram Khan. Medical image analysis using convolutional neural networks: a review. *Journal of medical systems*, 42(11):226, 2018.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[25] Patrick Ferdinand Christ, Mohamed Ezzeldin A Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D'Anastasi, et al. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 415–423. Springer, 2016.

[26] Bill S Lin, Kevin Michael, Shivam Kalra, and Hamid R Tizhoosh. Skin lesion segmentation: U-nets versus clustering. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7. IEEE, 2017.

[27] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

[28] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer, 2016.

[29] Jameson Merkow, Alison Marsden, David Kriegman, and Zhuowen Tu. Dense volume-to-volume vascular boundary detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 371–379. Springer, 2016.

[30] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas van den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017.

[31] Lequan Yu, Xin Yang, Hao Chen, Jing Qin, and Pheng-Ann Heng. Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images. In *AAAI*, pages 66–72, 2017.

[32] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2528–2535, June 2010.

[33] Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pages 179–187. Springer, 2016.

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[35] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[36] Luís Pedro Coelho, Aabid Shariff, and Robert F Murphy. Nuclear segmentation in microscope cell images: a hand-segmented dataset and comparison of algorithms. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pages 518–521. IEEE, 2009.

[37] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3):411–426, 2007.

[38] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1451–1460. IEEE, 2018.

[39] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using convolutional auto-encoders with symmetric skip connections. *arXiv preprint arXiv:1606.08921*, 2016.

[40] Ignacio Arganda-Carreras, Srinivas C Turaga, Daniel R Berger, Dan Cireşan, Alessandro Giusti, Luca M Gambardella, Jürgen Schmidhuber, Dmitry Laptev, Sarvesh Dwivedi, Joachim M Buhmann, et al. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in neuroanatomy*, 9:142, 2015.

[41] Albert Cardona, Stephan Saalfeld, Stephan Preibisch, Benjamin Schmid, Anchi Cheng, Jim Pulokas, Pavel Tomancak, and Volker Hartenstein. An integrated micro- and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy. *PLoS biology*, 8(10):e1000502, 2010.

[42] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. *arXiv preprint arXiv:1803.10417*, 2018.

[43] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

[44] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993, 2015.

[45] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4:170117, 2017.

[46] Guido Van Rossum et al. Python programming language. In *USENIX Annual Technical Conference*, volume 41, page 36, 2007.

[47] François Chollet et al. Keras, 2015.

[48] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[49] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[50] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[51] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[52] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[53] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.