Neural Networks 121 (2020) 74–87

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

# MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation

Nabil Ibtehaz [a], M. Sohel Rahman [b,*]

[a] Samsung R&D Institute, Bangladesh
[b] Department of CSE, BUET, ECE Building, West Palasi, Dhaka 1205, Bangladesh

ARTICLE INFO

*Article history:*
Received 9 February 2019
Received in revised form 17 August 2019
Accepted 22 August 2019
Available online 4 September 2019

*Keywords:*
Convolutional neural networks
Medical imaging
Semantic segmentation
U-Net

ABSTRACT

In recent years Deep Learning has brought about a breakthrough in Medical Image Segmentation. In this regard, U-Net has been the most popular architecture in the medical imaging community. Despite outstanding overall performance in segmenting multimodal medical images, through extensive experimentations on some challenging datasets, we demonstrate 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. Following these modifications, we develop a novel architecture, MultiResUNet, as the potential successor to the U-Net architecture. We have tested and compared MultiResUNet with the classical U-Net on a vast repertoire of multimodal medical images. Although only slight improvements in the cases of ideal images are noticed, remarkable gains in performance have been attained for the challenging ones. 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. We have also discussed and highlighted some qualitatively superior aspects of MultiResUNet over classical U-Net that are not really reflected in the quantitative measures.

## 1. Introduction

Since the inception of digital medical imaging equipment, 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 (Schindelin, Rueden, Hiner, & Eliceiri, 2015). The task of a computer-aided medical image analysis tool is twofold: segmentation and diagnosis. In the general Semantic Segmentation problem, the objective is to partition an image into a set of non-overlapping regions, which allows the homogeneous pixels to be clustered together (McGuinness & O'connor, 2010). However, in the context of medical images, the interest often lies in distinguishing some interesting areas of the image only, like the tumor regions (Codella et al., 2018), organs (Yang et al., 2018) etc. This enables the doctors to analyze only the significant parts of the otherwise incomprehensible multimodal medical images (Naik et al., 2008). Furthermore, often the segmented images are used to compute various features that may be leveraged in the diagnosis (Rouhi, Jafari, Kasaei, & Keshavarzian, 2015). Therefore, image segmentation is of utmost importance and has tremendous application in the domain of Biomedical Engineering.

Owing to the profound significance of medical image segmentation and the complexity associated with doing that manually, a vast number of automated medical image segmentation methods have been developed, mostly focusing 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 a huge variety of data (Pham, Xu, & Prince, 2000). Consequently, more adaptive algorithms were developed relying on geometric shape priors with tools of soft-computing (Mesejo, Valsecchi, Marrakchi-Kacem, Cagnoni, & Damas, 2015) and fuzzy systems (Zheng, Jeon, Xu, Wu, & Zhang, 2015). Nevertheless, these methods suffer from human biases and cannot deal with the variances in real-world data.

Recent advancements in deep learning (LeCun, Bengio, & Hinton, 2015) have shown a lot of promises towards solving issues discussed above. In this regard, Convolutional Neural Networks (CNN) (LeCun, Bottou, Bengio, & Haffner, 1998) have been the most ground-breaking addition, which are dominating the field of Computer Vision. CNNs have been responsible for the phenomenal advancements in tasks like object classification (Krizhevsky, Sutskever, & Hinton, 2012), object localization (Sermanet et al., 2013) etc., and the continuous improvements to CNN architectures are bringing about further radical progress (He, Zhang,

* Corresponding author.
*E-mail addresses:* n.ibtehaz@samsung.com (N. Ibtehaz), msrahman@cse.buet.ac.bd (M.S. Rahman).

https://doi.org/10.1016/j.neunet.2019.08.025
0893-6080/© 2019 Elsevier Ltd. All rights reserved.

Ren, & Sun, 2016; Simonyan & Zisserman, 2014; Szegedy, Ioffe, Vanhoucke, & Alemi, 2017; Szegedy et al., 2015). Semantic Segmentation tasks have also been revolutionalized by Convolutional Networks. Since CNNs are more intuitive in performing object classification, Ciresan, Giusti, Gambardella, and Schmidhuber (2012) presented a sliding window based pipeline to perform semantic segmentation using CNN. Long, Shelhamer, and Darrell (2015) proposed a fully convolutional network (FCN) to perform end-to-end image segmentation, which outperformed the existing approaches. Badrinarayanan, Kendall, and Cipolla (2015) 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 up-samples the feature maps to predict the segmentation masks. Chen, Papandreou, Kokkinos, Murphy, and Yuille (2018) presented DeepLab 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 (Litjens & Li et al., 2017). Nevertheless, CNNs have shown great promises in medical image segmentation in recent years (Anwar & Li et al., 2018; Litjens et al., 2017), and most of the credits go to U-Net (Ronneberger, Fischer, & Brox, 2015). The structure of U-Net is quite similar to SegNet, comprising an encoder and a decoder network. Furthermore, the corresponding 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 the extent that it has become the de-facto standard in medical image segmentation (Litjens et al., 2017). U-Net and U-Net like models have been successfully used in segmenting biomedical images of neuronal structures (Ronneberger et al., 2015), liver (Christ & Li et al., 2016), skin lesion (Lin, Michael, Kalra, & Tizhoosh, 2017), colon histology (Sirinukunwattana & Li et al., 2017), kidney (Çiçek, Abdulkadir, Lienkamp, Brox, & Ronneberger, 2016), vascular boundary (Merkow, Marsden, Kriegman, & Tu, 2016), lung nodule (Setio & Li et al., 2017), prostate (Yu, Yang, Chen, Qin, & Heng, 2017), 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 the field of 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 is 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 performance.
- We also experiment with a standard 3D extension of MultiResUNet on a particular 3D MRI dataset, and it outperforms the enhanced 3D U-Net as well.
- We qualitatively 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 (Long et al., 2015) and SegNet (Badrinarayanan et al., 2015), U-Net (Ronneberger et al., 2015) uses a network of convolutional layers entirely 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, 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 down-sampling, the number of filters in the convolutional layers is doubled. Finally, a progression of two 3 × 3 convolution operations connects the *Encoder* to the *Decoder*.

On the other hand, the *Decoder* first up-samples the feature map using a 2 × 2 transposed convolution operation (Zeiler, Krishnan, Taylor, & Fergus, 2010), reducing the feature channels by half. Then a sequence of two 3 × 3 convolution operations is performed again. Similar to the *Encoder*, this succession of up-sampling and two convolution operations is repeated four times, halving the number of filters at 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 (LeCun et al., 2015); 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 transferred 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 (Drozdzal, Vorontsov, Chartrand, Kadoury, & Pal, 2016). The network architecture is illustrated in Fig. 1.

Subsequently, the U-Net architecture was extended through a few modifications to 3D U-Net for volumetric segmentation (Çiçek et al., 2016). 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, along with using batch normalization (Ioffe & Szegedy, 2015), the number of filters was doubled before the pooling layers to avoid bottlenecks (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016). The original U-Net (Ronneberger et al., 2015) did not use batch normalization. However, when experimented with it later, the results revealed, perhaps astonishingly, that batch normalization may even hurt the performance sometimes (Çiçek et al., 2016).

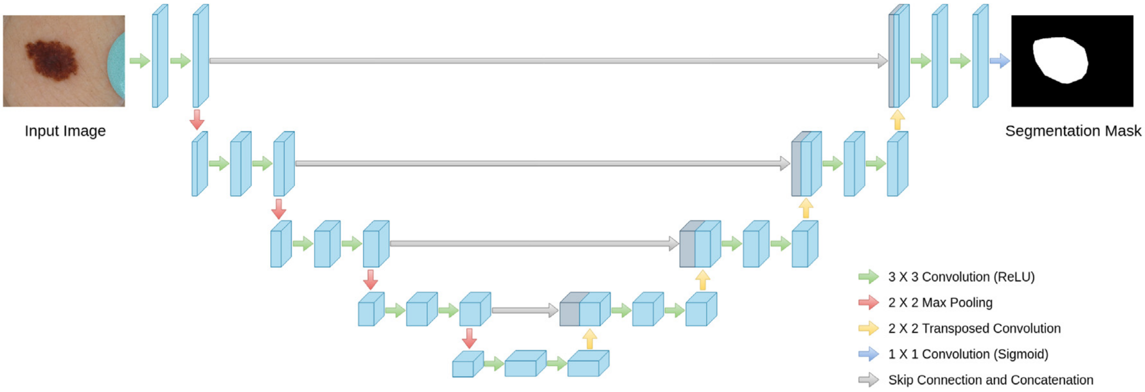

Input Image
Segmentation Mask
3 X 3 Convolution (ReLU)
2 X 2 Max Pooling
2 X 2 Transposed Convolution
1 X 1 Convolution (Sigmoid)
Skip Connection and Concatenation

**Fig. 1.** The classic U-Net architecture. The model comprises an encoder and a decoder pathway, with skip connections between the corresponding layers.

## 3. Motivations and high level considerations

U-Net has been a remarkable and the most popular deep network architecture in the medical imaging community, defining the state of the art in medical image segmentation (Drozdzal et al., 2016). However, through deep contemplation of the U-Net architecture and drawing some parallels to the recent advancement in the field of deep computer vision, we have made some insightful and useful observations (as described in the following subsections); these observations in the sequel have led us to some ideas for improvement.

### *3.1. Variation of scale in medical images*

In medical image segmentation, we are interested in segmenting cell nuclei (Coelho, Shariff, & Murphy, 2009), organs (Yang et al., 2018), tumors (Codella et al., 2018) etc. from images originating from various modalities. However, in most cases, these objects of interest are of irregular and different scales. For example, Fig. 2 demonstrates that the scale of skin lesions can greatly vary in dermoscopy images. These situations frequently occur in different types of medical image segmentation tasks.

Therefore, a network should be robust enough to analyze objects at different scales. Although this issue has somewhat been addressed in several deep computer vision works (Szegedy et al., 2017, 2015, 2016; Zhao, Shi, Qi, Wang, & Jia, 2017), to the best of our knowledge, it is still not addressed comprehensively in the domain of medical image segmentation. Serre, Wolf, Bileschi, Riesenhuber, and Poggio (2007) 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 (Szegedy et al., 2015) introduced Inception blocks that utilized 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, were 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 is used. As explained in Szegedy et al. (2016), this series of two 3 × 3 convolutional operation actually resembles a 5 × 5 convolutional operation. Therefore, following the approach of the Inception network, the simplest way to augment U-Net with a multi-resolutional analysis capability is to incorporate 3 × 3, and 7 × 7 convolution operations in parallel to the 5 × 5 convolution operation, as shown in Fig. 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 (Wang et al., 2018), but in our experiments, it is overshadowed by the former. Despite the gain in performance, the introduction of additional convolutional layers in parallel extravagantly increases the memory requirement. Therefore, we improvise with the following ideas borrowed from Szegedy et al. (2016). 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 Fig. 3b. The outputs of the 2nd and the 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 is seen that the results of this compact block closely resemble that of the memory-intensive Inception-like block described earlier. This outcome is in line with the findings of Szegedy et al. (2016), as the adjacent layers of a vision network are expected to be correlated.

Although 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 (Szegedy et al., 2015). Therefore, instead of keeping all the three consecutive convolutional layers with an 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 segmentation (Drozdzal et al., 2016) as well as to introduce 1 × 1 convolutional layers, which may allow us to comprehend some additional spatial information. We call this arrangement a '*MultiRes block*', as shown in Fig. 3c.

### *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 the max-pooling and after the deconvolution operations. This enables the network to propagate the spatial information that gets lost during the pooling operation from encoder to decoder.

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

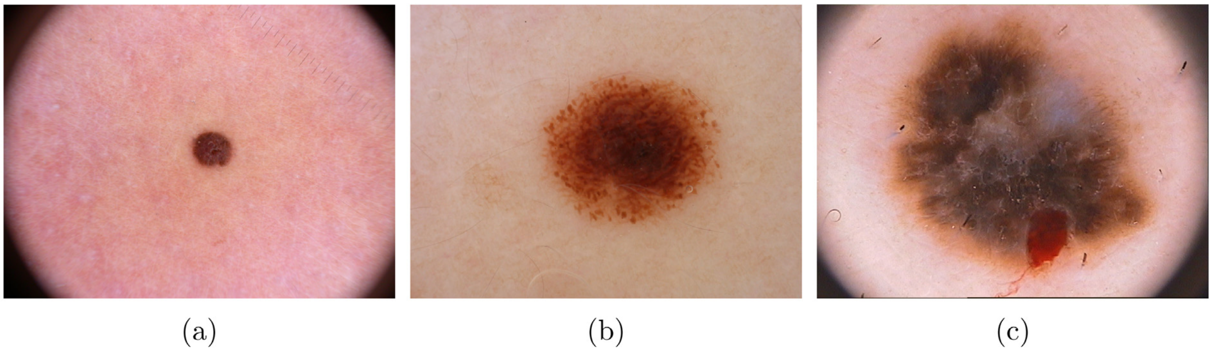


**Fig. 2.** Variation of scale in medical images. Fig. 2a, 2b, 2c are examples of dermoscopy images with small, medium and large size of lesions, respectively. The images have been taken from the ISIC-2018 dataset (Codella et al., 2018).

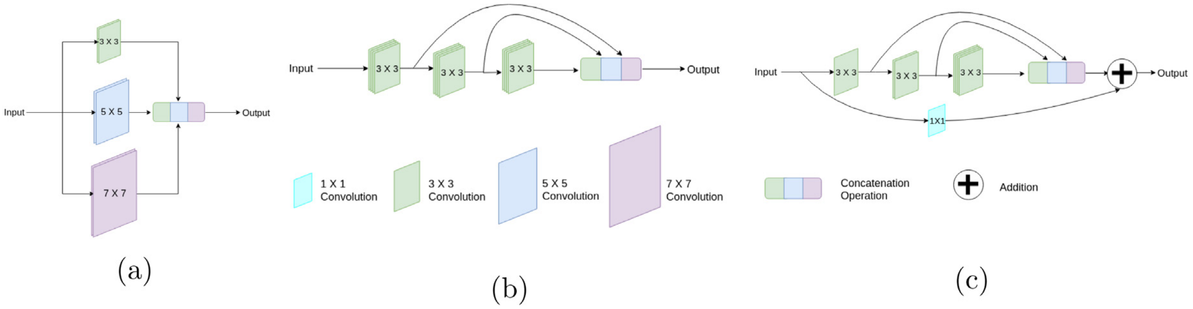


**Fig. 3.** Developing the proposed *MultiRes* block. We start with a simple Inception-like block by using 3 × 3, 5 × 5 and 7 × 7 convolutional filters in parallel and concatenating the generated feature maps (Fig. 3a). This allows us to reconcile spatial features from different context size. Subsequently, 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 filters for conserving dimensions).

the very deep layers of the network, thereby, going 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 thereby adversely affecting 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 in later stages, 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. We hypothesize that these additional non-linear transformations on the features propagating from the encoder stage should account for or somewhat balance the possible semantic gap (alluded to above) introduced by the higher degree of processing by the deeper decoder stages. Furthermore, instead of using the usual convolutional layers, we introduce residual connections to them as they make the learning easier (Szegedy et al., 2017) and are proven to have great potential in medical image analysis (Drozdzal et al., 2016). This idea is inspired from the image to image conversion using convolutional neural networks (Mao, Shen, & Yang, 2016), where pooling layers are not favorable for the loss of information. Thus, instead of simply concatenating the feature maps from the encoder stages to the decoder stages, we first pass them through a chain of convolutional layers with residual connections and then concatenate them with the decoder features. We call this proposed shortcut path '*Res path*', illustrated in Fig. 4. More specifically, 3 × 3 filters are used in the convolutional layers and 1 × 1 filters accompany the residual connections.

## 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 $W$ that controls the number of filters of the convolutional layers inside that block. To maintain a comparable relationship between the numbers of parameters in the 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 U-Net and $\alpha$ is a scalar coefficient. This provides us with a convenient way to both control the number of parameters and keep them comparable to that of 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 set $\alpha = 1.67$ as it keeps the number of parameters in our model slightly below that of U-Net.

In Section 3.2, we have 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 has 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 or halved respectively.

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

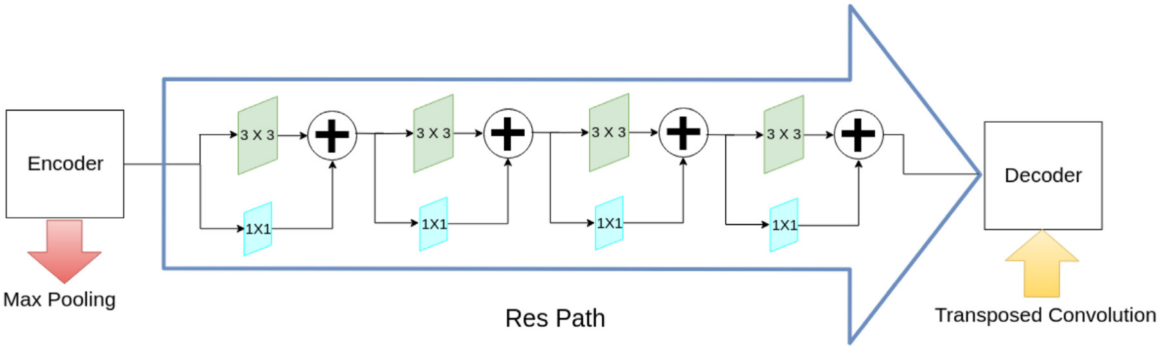


**Fig. 4.** Proposed *Res* path. Instead of combining the encoder feature maps with the decoder feature in a straight-forward manner, we pass the encoder features through a sequence of convolutional layers. These additional non-linear operations are expected to reduce the semantic gap between encoder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

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 in this network, except for the output layer, are activated by the *ReLU* (Rectified Linear Unit) activation function (LeCun et al., 2015), and are batch-normalized (Ioffe & Szegedy, 2015). 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 is challenging as compared to the traditional computer vision datasets. Expensive imaging equipment, sophisticated image acquisition pipelines, necessity of expert annotation, issues of privacy — all add to the complexity of developing medical imaging datasets (Litjens et al., 2017). As a result, only a few public medical imaging benchmark datasets exist, containing only a handful of images each. In order to assess the efficacy of the proposed architecture, we have tested and evaluated it on a variety of image modalities. More specifically, we have selected datasets that are as heterogeneous as possible from each other. Also, each of these datasets poses a unique challenge of its own (more details are given in Sections 7 and 8). The datasets used in the experiments are briefly described below (also see Table 2 for an overview).

### 5.1. Fluorescence microscopy images

We have used the fluorescence microscopy image dataset developed by Murphy Lab (Coelho et al., 2009). 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 are segmented manually by experts. The nuclei are irregular in terms of brightness and the images often contain noticeable debris, making this a challenging dataset of bright-field microscopy images. The original resolution of the images range from 1344 × 1024 to 1349 × 1030; they have been resized to 256 × 256 for computational constraints.

### 5.2. Electron microscopy images

To observe the effectiveness of the architecture with electron microscopy images, we have used the dataset of the ISBI-2012: 2D EM segmentation challenge (Arganda-Carreras et al., 2015; Cardona et al., 2010). This dataset contains only 30 images from a serial section Transmission Electron Microscopy (ssTEM) of the Drosophila first instar larva ventral nerve cord (Cardona et al., 2010). 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 images

We have 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 (Codella et al., 2018) and the HAM10000 dataset (Tschandl, Rosendahl, & Kittler, 2018). The compiled dataset contains a total of 2594 images of different types of skin lesions with expert annotation. The images are of various resolutions, but they have all been resized to 256 × 192, maintaining the average aspect ratio.

### 5.4. Endoscopy images

We have used the CVC-ClinicDB (Bernal et al., 2015), 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 have been 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 images

All the datasets described above contain 2D medical images. In order to evaluate our proposed architecture with 3D medical images, we have used the magnetic resonance images (MRI) from the BraTS17 competition database (Bakas et al., 2017; Menze et al., 2015). 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 × 240 × 155 but have been resized to 80 × 80 × 48 for 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.

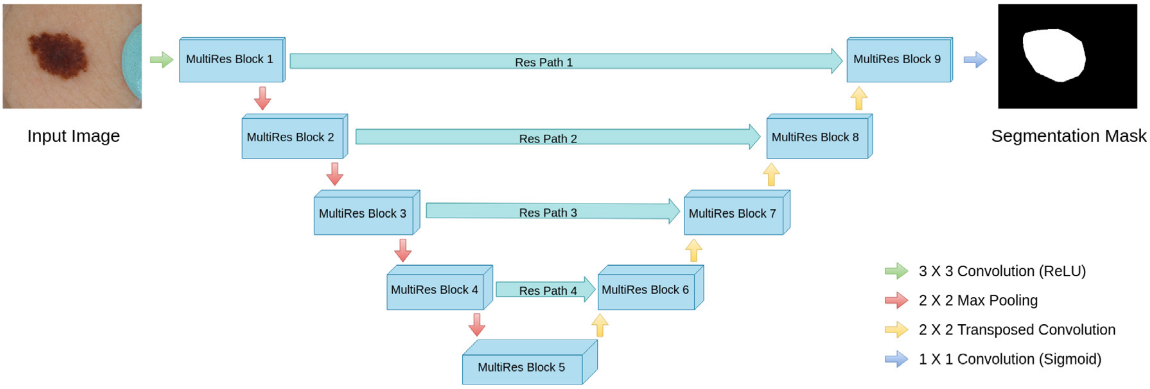


**Fig. 5.** Proposed MultiResUNet architecture. We replace the sequences of two convolutional layers in the U-Net architecture with the proposed *MultiRes* blocks. Furthermore, instead of using plain shortcut (skip) connections, we use the proposed *Res* paths.

**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 |

**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 × 240 × 155 | 80 × 80 × 48 |

### *5.6. Challenges in the datasets*

As has been mentioned above, each of the datasets used in the experiments poses unique challenges of its own. In the Fluorescence Microscopy image dataset, there exist some images with bright objects that are apparently almost indistinguishable from the actual nuclei. These objects act as outliers in images with well-defined contrast between the foreground and the background. ISBI-2012 Electron Microscopy dataset presents a different type of challenge. In this dataset, the region being segmented (i.e., region of interest) covers the majority of the image; thus a tendency can be observed to over-segment the images. The Dermoscopy dataset of ISIC-2018 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, that are present both in the background and in the foreground, make pattern recognition more difficult. In the Endoscopy dataset, the boundaries between the polyps and the background are so vague that often it becomes difficult to distinguish even for a trained operator (Bernal et al., 2017). In addition, the polyps are diverse in terms of shape, size, structure, orientation etc. BRATS17 MRI dataset, on the other hand, is a dataset containing 3D images. Therefore, it brings forth the challenges of segmenting 3D volumes. Moreover, the actual tumors therein are pretty tiny compared to the whole volume; on average the tumors cover about only 1% of the entire brain scan.

## 6. Experiments

We have used Python, more specifically Python3 programming language (Van Rossum et al., 2007) to conduct the experiments. The network models have been implemented using Keras (Chollet et al., 2015) with Tensorflow backend (Abadi et al., 2016). Our model implementation is available in the following github repository:

https://github.com/nibtehaz/MultiResUNet

The experiments have been 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 have compared its performance with the U-Net architecture as the baseline. To keep the number of parameters comparable to our proposed MultiResUNet, we have implemented the original U-Net (Badrinarayanan et al., 2015) 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 have used the 3D counterpart of the U-Net as described in the original paper (Çiçek et al., 2016). 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 is presented in Table 3. Although not utterly significant, in both the cases, MultiResUNet requires a slightly lesser number of parameters.

### 6.2. Pre-processing/post-processing

The objective of our experiments is to investigate the performance of the proposed MultiResUNet architecture as compared to the original U-Net, as a general model. Therefore, no domain-specific pre-processing has been applied. The only pre-processing we have applied is that the input images have been resized to fit into the GPU memory and the pixel values were divided by 255 to bring them to the $[0\ldots1]$ range. Similarly, no application-specific post-processing has been performed. Since the final layer is activated by a Sigmoid function, it produces outputs in the range $[0\ldots1]$. Therefore, we have 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 be $Y$, and the segmentation mask predicted by the model be $\hat{Y}$. For a pixel $px$, the network predicts $\hat{y}_{px}$, whereas, the ground truth value is $y_{px}$. The binary cross-entropy loss for that image is defined as:

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

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

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

We have minimized the binary cross-entropy loss and hence have trained the model using the Adam optimizer (Kingma & Ba, 2014). 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 (Duchi, Hazan, & Singer, 2011) and RMSProp (Tieleman & Hinton, 2012); therefore Adam has been often used as the default choice, in benchmarking deep learning models (Ruder, 2016). Adam has a number of parameters including $\beta_1$ and $\beta_2$, which control the decay of first and second moments respectively. In this work, we have used Adam with the parameters mentioned in the original paper. The models have been trained for 150 epochs using Adam optimizer. The reason for selecting 150 as the number of epochs is due to the fact that after 150 epochs no further improvement is noticed in either of the networks.

### 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 a false sense of superiority, inflated by the perfection of detecting the background. Hence, the Jaccard Index has been widely used to evaluate and benchmark image segmentation and object localization algorithms (McGuinness & O'connor, 2010). Jaccard Index for two sets $A$ and $B$ are defined as the ratio of the intersection and union of the two sets:

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

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 $D_1, D_2, \ldots, D_k$ of equal or near-equal size (Kohavi et al., 1995). 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 have performed 5-Fold Cross Validation tests on each of the different datasets.

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

## 7. Results

### 7.1. MultiResUNet consistently outperforms U-Net

As described in Sections 5 and 6, to evaluate the performance of the proposed architecture, we have conducted experiments

**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 |

**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 |

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 are recorded and are combined 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 MultiResUNet outperforms the base 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, these improvements are of great significance. For Fluorescence Microscopy images as well, our model achieves a 2.6326% relative improvement over U-Net, and despite having a slightly lesser number of parameters, it still achieves a relative improvement of 1.4104% for MRI images. Only for Electron Microscopy images, U-Net seems to be on par with MultiResUNet, 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 Fig. 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 convergence much faster. This can be attributed to the synergy between residual connections and batch normalization (Drozdzal et al., 2016). Moreover, apart from Fig. 6d (i.e., for the Electron Microscopy dataset), in all other cases the MultiResUNet model consistently outperforms 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 fluctuations, the standard deviation of the performance of the MultiResUNet is much smaller (please refer to the Supplementary Material 3 for a more precise idea); 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 (Fig. 7c); our proposed model, however, performs better albeit only slightly (Fig. 7d). 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 a lack of clear boundaries. On such cases, the U-Net model either under-segmented (Fig. 8a) or over-segmented (Fig. 8b) the polyps. Our proposed MultiResUNet, on the other hand, performed considerably better in both cases. However, there are some images where both the models faced complications, but even in those cases, MultiResUNet's performance was superior (Fig. 8c). Dermoscopic images have comparatively clearer defined boundaries; still, in those cases, MultiResUNet delineates the boundaries better (Fig. 8d). The same phenomenon 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 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 appears similar to the region of interest. These two

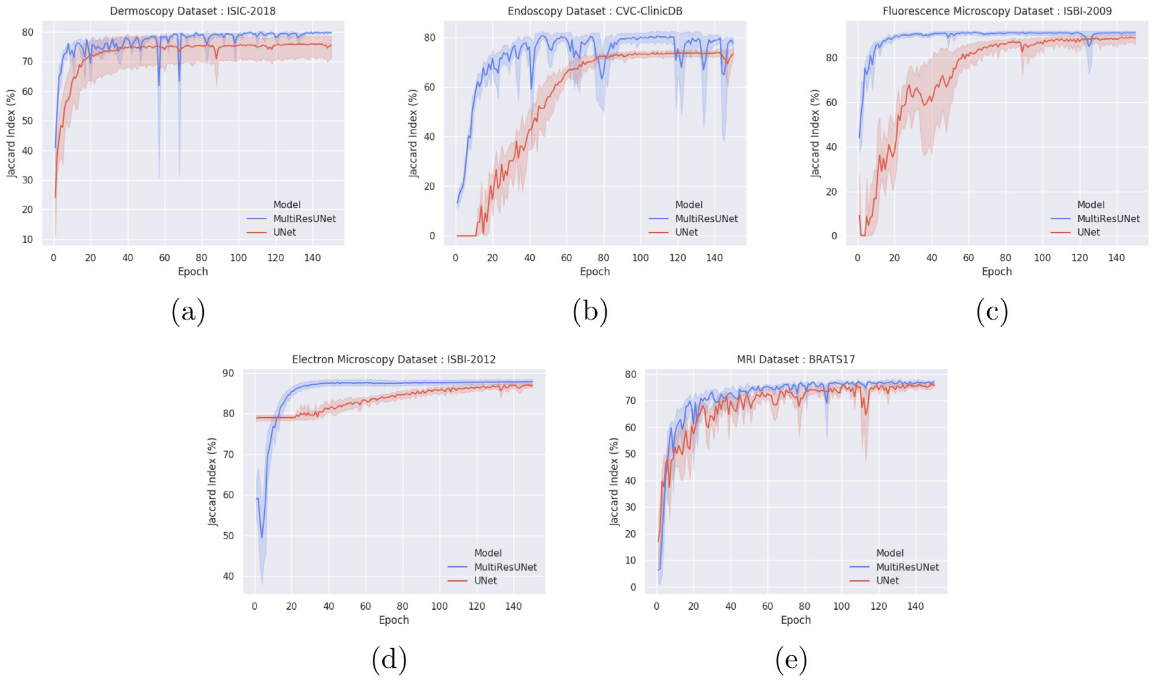


**Fig. 6.** Progress of the validation performance with the number of epochs. We record the value of Jaccard Index on validation data after each epoch. It can be observed that not only MultiResUNet outperforms the U-Net model, but also the standard deviation of MultiResUNet is much smaller (please refer to the Supplementary Material 3 for a more precise idea).

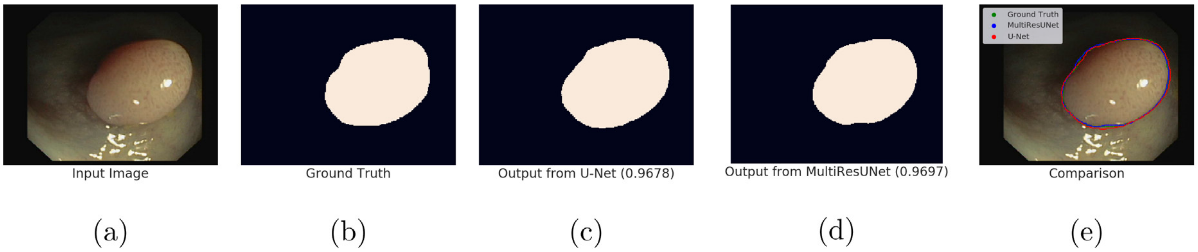


**Fig. 7.** Segmenting a polyp with clearly visible boundary (7a). U-Net manages to segment the polyp with a high level of performance (J.I. = 0.9678) (7c). MultiResUNet performs only slightly better (J.I. = 0.9697) (7d). Both the models seem to segment the polyp close to the ground truth (7b, 7e).

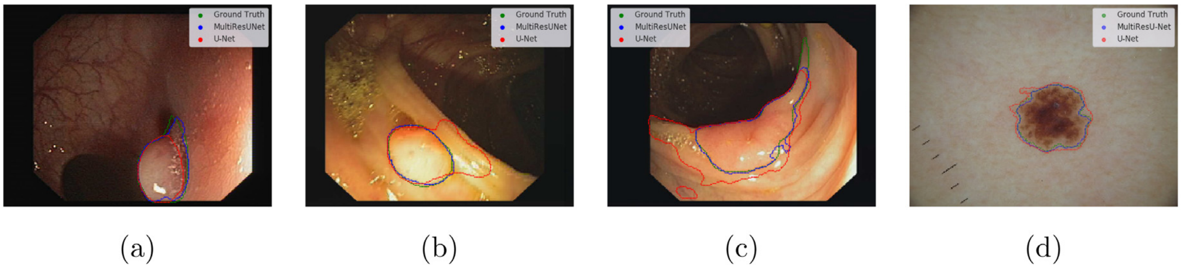
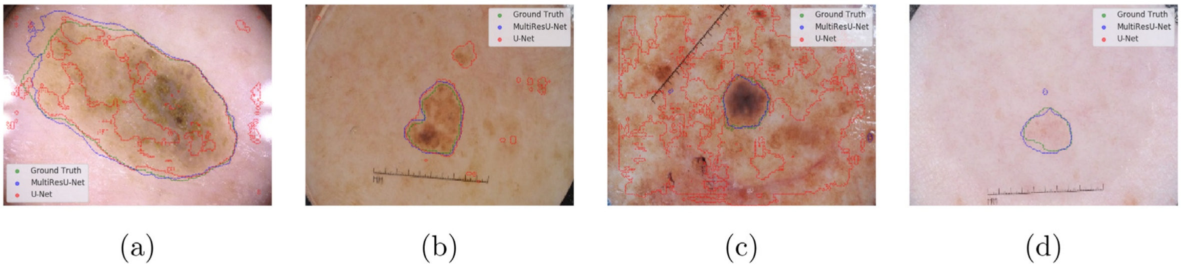


**Fig. 8.** Segmenting images with vague boundaries. This issue is more prominent for Colon Endoscopy images. U-Net seems to either under-segment (8a), or over-segment (8b) the polyps. MultiResUNet manages to segment polyps of such situation much better (8a, 8b). However, some images are too problematic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

cases lead to loss of information and false classifications respectively. Fortunately, the Dermoscopy image dataset we have used contains images with such confusing cases, allowing us to analyze and compare the behavior and performance of the two models thereon.

In spite of segmenting the images having a 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 tends to vary a bit, U-Net was unable to segment the foreground as a continuous region. It rather predicted a set of scattered regions (Fig. 9a), confusing the foreground as background thereby causing 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 rougher the background becomes, the more false predictions are made (Fig. 9c). Furthermore, in some dreadfully adverse situations, where due to irregularity,



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

**Fig. 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 sometimes classifies them as the foreground (9b), the more irregular the background, the more false predictions are made (9c). At the other extreme, for images where difference between foreground and background is too subtle, U-Net misses the foreground completely (9d). Though the segmentations produced by MultiResUNet in these challenging cases are not perfect, they have been consistently better than that of the U-Net.

the difference between background and foreground is too subtle, the U-Net model failed to make any prediction at all (Fig. 9d). Although in such challenging cases the segmentation of MultiResUNet is not perfect, it performs far superior (5.065% better to be specific) 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 *ResPaths*, 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 shown in Fig. 10.

It can be observed that the input image (Fig. 10a) is infected with some small particles that are not actual cell nuclei (Fig. 10b). 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 (Fig. 10c). On the other hand, our proposed MultiResUNet has been able to reject those outliers (Fig. 10d). Since the outliers are pretty tiny, false predictions made by the U-Net model do 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 alike outliers are present in other datasets as well, and MultiResUNet has been able to segment the images reliably without making false predictions. The additional convolutional operations along the proposed *Res* path are likely to contribute towards this success. Since in the classical U-Net model the lower level features from the encoder network are utilized in making the final prediction, visually similar outliers can outsmart the network. This somewhat rare quality of MultiResUNet, however, is not properly reflected in a quantitative manner in the evaluation metric for the reasons mentioned above.

### 7.6. Note on segmenting the majority class

The Electron Microscopy dataset is quite interesting and unorthodox as in this dataset the region of interest under consideration actually comprises the major portion of an image. This is a rare incident since usually, the region of interest consists of a small portion of the image. This brings out 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. 11a. Here, it can be observed quite astonishingly that the majority of the image is actually foreground (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. 11c). MultiResUNet, on the other hand, has managed to segment the regions with properly defined separations among them (Fig. 11d). 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 that the two segmentations (i.e., the results of MultiResUNet and U-Net) are very different from each other, the respective Jaccard Index values 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 Jaccard Index falls short in adequately representing the accuracy while segmenting the majority class. Despite predicting much inferior segmentations, for this reason, the Jaccard Index of U-Net is 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.

### 7.7. Ablation study

An ablation study has been conducted to investigate the individual contributions of the *MultiRes* blocks and the *Res* paths. The experiments were performed from two ends; in the first case we just included the *Res* paths in a plain U-Net model, and in the second, we replaced the pair of convolutional blocks with *MultiRes* blocks. Hence, a comparison has been made among U-Net, U-Net with *Res* paths, U-Net with *MultiRes* blocks and the MultiResUNet. We have selected the CVC-ClinicDB dataset for this ablation study as it is the most challenging dataset, used in our experiments. The results of the 5-fold cross-validation test are presented in Table 5. It can be observed from the table that inclusion of *Res* paths improves performance over the classical U-Net. Introducing *MultiRes* blocks (alone, without *Res* path) is even more effective as is evident from the table. However, when both *Res* paths and *MultiRes* blocks are used (i.e., the proposed *MultiResUNet*), the synergy between these two components yields the best results.

On the other hand, from empirical observation, it was observed that after introducing the *MultiRes* blocks the model was more successful in fine detection of the edges and the distinctive

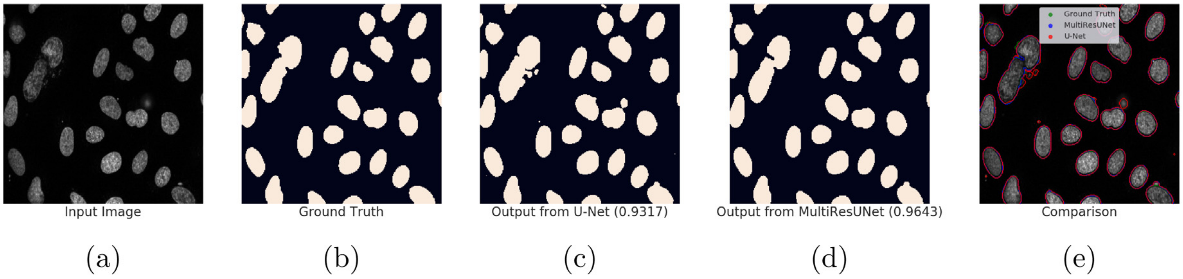


**Fig. 10.** Segmenting images containing outliers. In the fluorescence microscopy images, the exist some bright particles, visually very similar to the cell nuclei under analysis (10a). Although MultiResUNet can identify and reject those outliers (10d), U-Net seems to have mis-classified them (10c). This becomes more apparent from the comparison presented in (10e).

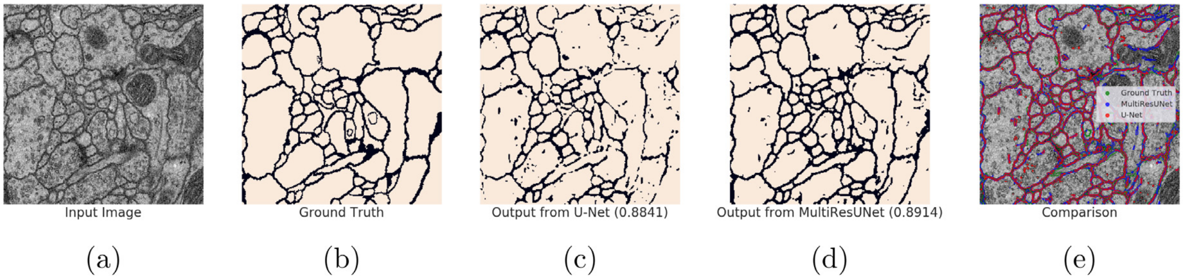


**Fig. 11.** Segmenting the majority class. Here we can observe that the region of interest comprises most part of the image (11b). Despite the values of Jaccard Index for both U-Net (11c) and MultiResUNet (11d) are quite similar, visually the segmentation masks are very different (11e). It can be seen that the 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 Jaccard Index are inflated.

**Table 5**
Ablation study investigating the individual contributions of *MultiRes* blocks and *Res* paths. The results are obtained from CVC-ClinicDB through a 5-fold cross-validation.

| Model | Fold = 1 | Fold = 2 | Fold = 3 | Fold = 4 | Fold = 5 | Average |
|---|---|---|---|---|---|---|
| U-Net | 73.71 | 72.08 | 74.89 | 75.58 | 76.23 | 74.5 |
| Only ResPath | 75.85 | 73.99 | 77.45 | 74.31 | 77.67 | 75.85 |
| Only ResBlock | 82.31 | 78.25 | 83.14 | 81.06 | 83.84 | 81.72 |
| MultiResUnet | 81.88 | 79.89 | 83.03 | 81 | 84.49 | 82.06 |

patterns or textures of the objects. Therefore, we hypothesize that the inclusion of filters of different scales are allowing the model to distinguish the boundaries better. On the contrary, it was sometimes seen that models with only *MultiRes* blocks resulted in some discontinuity within the segmented regions. Adding *Res* paths improved the segmentation performance of such images. Therefore, we conjecture that *Res* paths, by alleviating the semantic gap between the encoder and decoder networks, actually make the continuous homogeneous regions more homogeneous. This is based on the intuition that certain signatures of the feature maps propagated from the encoder network are likely to have been lost during the pooling operation, but the corresponding feature maps in decoder network may have generated those signatures. Therefore, some additional convolutional operation along the concatenation path aids in the proper fusion of these two sets of feature maps, preserving the homogeneity throughout. Some experimental results are presented in Fig. 12. However, it may also be noted that occasionally models with only *Res* path produce segmentations with discontinuity and sometimes models involving *MultiRes* blocks miss some vague edges.

**Table 6**
Impact of data augmentation on both the models. The results are obtained from CVC-ClinicDB through 5-fold cross-validation tests.

| Model | Fold = 1 | Fold = 2 | Fold = 3 | Fold = 4 | Fold = 5 | Average |
|---|---|---|---|---|---|---|
| Without data augmentation | | | | | | |
| U-Net | 73.71 | 72.08 | 74.89 | 75.58 | 76.23 | 74.498 |
| MultiResUNet | 81.88 | 79.89 | 83.03 | 81.00 | 84.49 | 82.057 |
| With data augmentation | | | | | | |
| U-Net | 80.47 | 78.45 | 80.97 | 77.02 | 79.29 | 79.240 |
| MultiResUNet | 85.06 | 83.18 | 87.49 | 83.35 | 85.78 | 84.972 |

### *7.8. Note on data augmentation*

It is a well-known fact that data augmentation significantly improves the performance of Convolutional Neural Networks. However, the results presented so far are obtained without using any data augmentation. Since we have opted to evaluate the general behavior of the two models (i.e., MultiResUNet and classical U-Net), we have been inclined towards testing the models without any data augmentation. The primary reasoning behind this is that the lack of data augmentation will make the training task difficult for both the models and act as an additional adversity.

However, we have conducted a limited study with the models employing data augmentation. Again, we have used the CVC-ClinicDB dataset as it has been the most challenging one. We have randomly flipped, rotated or done both and have increased the training data up to three times thereby. Then we have evaluated the performance of the baseline U-Net and the proposed MultiResUNet model. The results with and without data augmentation are presented in Table 6.

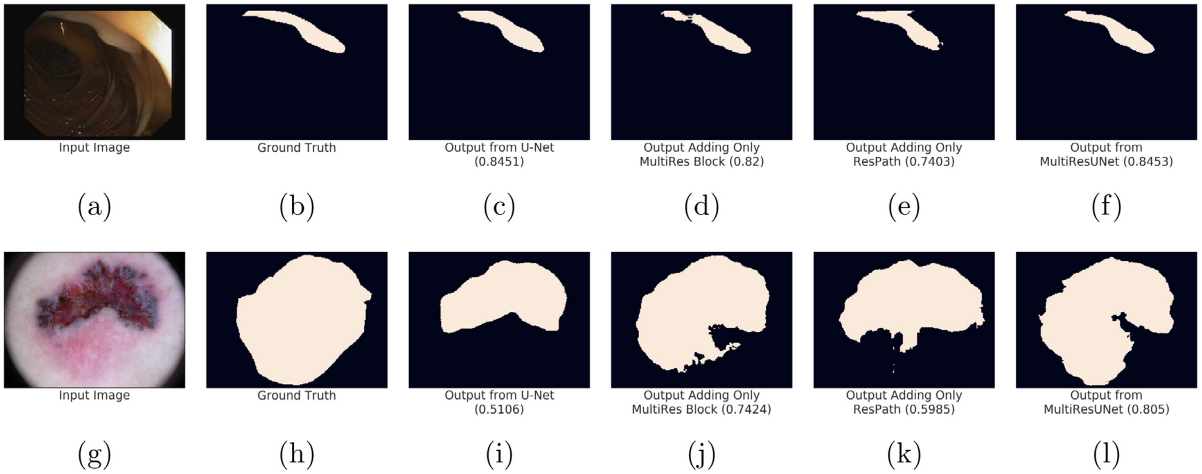

Input Image
Ground Truth
Output from U-Net (0.8451)
Output Adding Only MultiRes Block (0.82)
Output Adding Only ResPath (0.7403)
Output from MultiResUNet (0.8453)
(a)
(b)
(c)
(d)
(e)
(f)
Input Image
Ground Truth
Output from U-Net (0.5106)
Output Adding Only MultiRes Block (0.7424)
Output Adding Only ResPath (0.5985)
Output from MultiResUNet (0.805)
(g)
(h)
(i)
(j)
(k)
(l)

**Fig. 12.** Some empirical results investigating the individual contributions of the *MultiRes* blocks and *Res* paths. 12a and 12g present two example images from CVC-ClinicDB and ISIC-2018 respectively, with 12b and 12h demonstrating the ground truth segmentations. It can be observed that adding either *MultiRes* block (12d, 12j) or *Res* path (12e, 12k), achieves improvement over the standard U-Net (12c, 12i). Further observation reveals that the segmentation obtained from models with only *MultiRes* blocks, though improves the overall boundary tracking, contains discontinuities therein (12d, 12j). On the other hand, models with only *Res* paths alleviate internal discontinuities but perform poorly in boundary tracking (12e, 12k). However, in *MultiResUNet* the synergy between the two components settles a balance of these two kinds of behavior, and obtains superior outcomes (12f, 12l).

From the experimental results, it can be observed that both the models demonstrate improved performance with data augmentation and MultiResUNet does perform better than classical U-Net. However, for the baseline U-Net model, the improvement (4.74%) appears to be a bit higher compared to that of MultiResUNet (2.91%). It is quite reasonable, as the baseline U-Net was lagging far behind the MultiResUNet in this dataset with an accuracy of 74.498%; the data augmentation paved the way for the U-Net model to recognize some patterns that were comparatively easily learnt by the MultiResUNet model (82.057% previously) without data augmentation. Also, it is intuitive that the nearer the predicted segmentation is to the perfect segmentation, the harder it is to improve it further; this reasoning can be attributed to the lower improvement of MultiResUNet after data augmentation as compared to U-Net.

## 8. Conclusion

In this work, we started by analyzing the U-Net architecture diligently, with the hope of finding potential rooms for improvement. We noticed 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 have 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-resolutional analysis, we have proposed *MultiRes* blocks. We took inspirations from Inception blocks and formulated a compact analogous structure, that is comparatively lightweight and demands less memory. Incorporating these modifications, we have 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 foreground, 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 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 MRI dataset, on the other hand, contains multimodal 3D images, which is a different problem altogether.

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% has been observed in using MultiResUNet over U-Net (Table 4). Not only the segmentations generated by MultiResUNet attain a higher score in the evaluation metric, but they are also visually more similar to the ground truth. Furthermore, on the very challenging images, U-Net tends to over-segment, under-segment, make false predictions and even miss the objects completely. On the contrary, in the experiments, MultiResUNet has appeared to be more reliable and robust. MultiResUNet has managed to detect even the most subtle boundaries, has been resilient in segmenting images with a lot of perturbations, and has been rejectable to the outliers. Even in segmenting the majority class, where the U-Net tends to over-segment, MultiResUNet manages to capture the fine details. Furthermore, the straightforward 3D adaptation of MultiResUNet has performed better than the 3D U-Net, which is not just a straightforward 3D implementation of the U-Net, in fact, is an enhanced and improved version. It should be noted that the segmentations generated by the proposed MultiResUNet are not perfect, but in most of the cases, it outperforms the classical U-Net by a moderate margin.

Therefore, we believe that MultiResUNet architecture can be a potential successor to the classical U-Net architecture as the state of the art. Though in this work we have kept our analysis limited to a boilerplate configuration, from additional experiments it has been observed that use of advanced loss functions, e.g., dice loss function (Milletari, Navab, & Ahmadi, 2016) significantly improves the segmentation accuracy; data augmentation benefits the model as well. 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, as more public datasets of medical images of different modalities are curated, we would like to evaluate our model performance on those datasets as well. Furthermore, we are interested in experimenting by applying several domain and application specific pre-processing and post-processing schemes to our model for specific problems. 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.

## Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the Titan XP GPU for this research.

This research work is part of the research project titled 'Machine Intelligence Aided Diagnosis of Cervical Cancer' under the innovation fund of ICT Division, Government of the Peoples' Republic of Bangladesh.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.neunet.2019.08.025.

## References

Abadi, Martín, Barham, Paul, Chen, Jianmin, Chen, Zhifeng, Davis, Andy, Dean, Jeffrey, et al. (2016). Tensorflow: a system for large-scale machine learning. In *OSDI, Vol. 16* (pp. 265–283).

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

Arganda-Carreras, Ignacio, Turaga, Srinivas C, Berger, Daniel R, Cireşan, Dan, Giusti, Alessandro, Gambardella, Luca M, et al. (2015). Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in Neuroanatomy*, *9*, 142.

Badrinarayanan, Vijay, Kendall, Alex, & Cipolla, Roberto Segnet: A deep convolutional encoder-decoder architecture for image segmentation, arXiv preprint arXiv:1511.00561.

Bakas, Spyridon, Akbari, Hamed, Sotiras, Aristeidis, Bilello, Michel, Rozycki, Martin, Kirby, Justin S, et al. (2017). Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data*, *4*, 170117.

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

Bernal, Jorge, Tajkbaksh, Nima, Sánchez, Francisco Javier, Matuszewski, Bogdan J, Chen, Hao, Yu, Lequan, et al. (2017). Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. *IEEE Transactions on Medical Imaging*, *36*(6), 1231–1249.

Cardona, Albert, Saalfeld, Stephan, Preibisch, Stephan, Schmid, Benjamin, Cheng, Anchi, Pulokas, Jim, et al. (2010). An integrated micro-and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy. *PLoS Biology*, *8*(10), e1000502.

Chen, Liang-Chieh, Papandreou, George, Kokkinos, Iasonas, Murphy, Kevin, & Yuille, Alan L (2018). 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.

Chollet, François, et al. Keras.

Christ, Patrick Ferdinand, Elshaer, Mohamed Ezzeldin A, Ettlinger, Florian, Tatavarty, Sunil, Bickel, Marc, Bilic, Patrick, et al. (2016). 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* (pp. 415–423). Springer.

Çiçek, Özgün, Abdulkadir, Ahmed, Lienkamp, Soeren S, Brox, Thomas, & Ronneberger, Olaf (2016). 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention* (pp. 424–432). Springer.

Ciresan, Dan, Giusti, Alessandro, Gambardella, Luca M, & Schmidhuber, Jürgen (2012). Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems* (pp. 2843–2851).

Codella, Noel CF, Gutman, David, Celebi, M Emre, Helba, Brian, Marchetti, Michael A, Dusza, Stephen W, et al. (2018). 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* (pp. 168–172). IEEE.

Coelho, Luís Pedro, Shariff, Aabid, & Murphy, Robert F. (2009). 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* (pp. 518–521). IEEE.

Drozdzal, Michal, Vorontsov, Eugene, Chartrand, Gabriel, Kadoury, Samuel, & Pal, Chris (2016). The importance of skip connections in biomedical image segmentation. In *Deep learning and data labeling for medical applications* (pp. 179–187). Springer.

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

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

Ioffe, Sergey, & Szegedy, Christian Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167.

Kingma, Diederik P., & Ba, Jimmy Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

Kohavi, Ron, et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai, Vol. 14* (pp. 1137–1145). Montreal, Canada.

Krizhevsky, Alex, Sutskever, Ilya, & Hinton, Geoffrey E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

LeCun, Yann, Bengio, Yoshua, & Hinton, Geoffrey (2015). Deep learning. *Nature*, *521*(7553), 436.

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

Lin, Bill S, Michael, Kevin, Kalra, Shivam, & Tizhoosh, Hamid R (2017). Skin lesion segmentation: U-nets versus clustering. In *2017 IEEE symposium series on computational intelligence (SSCI)* (pp. 1–7). Springer.

Litjens, Geert, Kooi, Thijs, Bejnordi, Babak Ehteshami, Setio, Arnaud Arindra Adiyoso, Ciompi, Francesco, Ghafoorian, Mohsen, et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88.

Long, Jonathan, Shelhamer, Evan, & Darrell, Trevor (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Mao, Xiao-Jiao, Shen, Chunhua, & Yang, Yu-Bin Image restoration using convolutional auto-encoders with symmetric skip connections, arXiv preprint arXiv:1606.08921.

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

Menze, Bjoern H, Jakab, Andras, Bauer, Stefan, Kalpathy-Cramer, Jayashree, Farahani, Keyvan, Kirby, Justin, et al. (2015). The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, *34*(10), 1993.

Merkow, Jameson, Marsden, Alison, Kriegman, David, & Tu, Zhuowen (2016). Dense volume-to-volume vascular boundary detection. In *International conference on medical image computing and computer-assisted intervention* (pp. 371–379). Springer.

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

Milletari, Fausto, Navab, Nassir, & Ahmadi, Seyed-Ahmad (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3d vision (3DV)* (pp. 565–571). IEEE.

Naik, Shivang, Doyle, Scott, Agner, Shannon, Madabhushi, Anant, Feldman, Michael, & Tomaszewski, John (2008). 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* (pp. 284–287). IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research (JMLR)*, *12*, 2825–2830.

Pham, Dzung L., Xu, Chenyang, & Prince, Jerry L. (2000). Current methods in medical image segmentation. *Annual Review of Biomedical Engineering*, *2*(1), 315–337.

Ronneberger, Olaf, Fischer, Philipp, & Brox, Thomas (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

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

Ruder, Sebastian An overview of gradient descent optimization algorithms, arXiv preprint arXiv:1609.04747.

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

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

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

Setio, Arnaud Arindra Adiyoso, Traverso, Alberto, De Bel, Thomas, Berens, Moira SN, van den Bogaard, Cas, Cerello, Piergiorgio, et al. (2017). 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.

Simonyan, Karen, & Zisserman, Andrew Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

Sirinukunwattana, Korsuk, Pluim, Josien PW, Chen, Hao, Qi, Xiaojuan, Heng, Pheng-Ann, Guo, Yun Bo, et al. (2017). Gland segmentation in colon histology images: The glas challenge contest. *Medical Image Analysis*, *35*, 489–502.

Szegedy, Christian, Ioffe, Sergey, Vanhoucke, Vincent, & Alemi, Alexander A (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI, Vol. 4* (p. 12).

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, et al. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

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

Tieleman, Tijmen, & Hinton, Geoffrey (2012). 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.

Tschandl, Philipp, Rosendahl, Cliff, & Kittler, Harald The ham10000 datase: A large collection of multi-source dermatoscopic images of common pigmented skin lesions, arXiv preprint arXiv:1803.10417.

Van Rossum, Guido, et al. (2007). Python programming language. In *USENIX annual technical conference, Vol. 41* (p. 36).

Wang, Panqu, Chen, Pengfei, Yuan, Ye, Liu, Ding, Huang, Zehua, Hou, Xiaodi, et al. (2018). Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)* (pp. 1451–1460). IEEE.

Yang, Jinzhong, Veeraraghavan, Harini, Armato III, Samuel G, Farahani, Keyvan, Kirby, Justin S, Kalpathy-Kramer, Jayashree, et al. (2018). Autosegmentation for thoracic radiation treatment planning: A grand challenge at aapm 2017. *Medical physics*.

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

Zeiler, M. D., Krishnan, D., Taylor, G. W., & Fergus, R. (2010). Deconvolutional networks. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 2528–2535).

Zhao, Hengshuang, Shi, Jianping, Qi, Xiaojuan, Wang, Xiaogang, & Jia, Jiaya (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881–2890).

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