Journal of
Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Recurrent residual U-Net for medical image segmentation

Md Zahangir Alom
Chris Yakopcic
Mahmudul Hasan
Tarek M. Taha
Vijayan K. Asari

SPIE.

Md Zahangir Alom, Chris Yakopcic, Mahmudul Hasan, Tarek M. Taha, Vijayan K. Asari, "Recurrent residual U-Net for medical image segmentation," *J. Med. Imag.* **6**(1), 014006 (2019), doi: 10.1117/1.JMI.6.1.014006.

Recurrent residual U-Net for medical image segmentation

Md Zahangir Alom,[a,*] Chris Yakopcic,[a] Mahmudul Hasan,[b] Tarek M. Taha,[a] and Vijayan K. Asari[a]
[a]University of Dayton, Department of Electrical and Computer Engineering, Dayton, Ohio, United States
[b]Comcast Labs, Washington, DC, United States

Abstract. Deep learning (DL)-based semantic segmentation methods have been providing state-of-the-art performance in the past few years. More specifically, these techniques have been successfully applied in medical image classification, segmentation, and detection tasks. One DL technique, U-Net, has become one of the most popular for these applications. We propose a recurrent U-Net model and a recurrent residual U-Net model, which are named RU-Net and R2U-Net, respectively. The proposed models utilize the power of U-Net, residual networks, and recurrent convolutional neural networks. There are several advantages to using these proposed architectures for segmentation tasks. First, a residual unit helps when training deep architectures. Second, feature accumulation with recurrent residual convolutional layers ensures better feature representation for segmentation tasks. Third, it allows us to design better U-Net architectures with the same number of network parameters with better performance for medical image segmentation. The proposed models are tested on three benchmark datasets, such as blood vessel segmentation in retinal images, skin cancer segmentation, and lung lesion segmentation. The experimental results show superior performance on segmentation tasks compared to equivalent models, including a variant of a fully connected convolutional neural network called SegNet, U-Net, and residual U-Net. © *2019 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.JMI.6.1.014006]

Keywords: medical imaging; semantic segmentation; convolutional neural networks; U-Net; residual U-Net; recurrent U-Net; recurrent residual U-Net.

Paper 18224RR received Oct. 2, 2018; accepted for publication Mar. 5, 2019; published online Mar. 27, 2019.

1 Introduction

Nowadays deep learning (DL) provides state-of-the-art performance for image classification,[1] segmentation,[2] detection and tracking,[3] and captioning.[4] Since 2012, several deep convolutional neural network (DCNN) models have been proposed such as AlexNet,[1] VGG,[5] GoogleNet,[6] Residual Net,[7] DenseNet,[8] and CapsuleNet.[9] A DL-based approach (CNN, in particular) provides a state-of-the-art performance for classification, segmentation, and detection tasks for several recently developed advanced methods, including activation functions, improved regularization techniques, and optimization approaches.[1,10] However, in most cases, models are explored and evaluated using classification tasks on very large-scale datasets such as ImageNet,[1] where the outputs of the classification tasks are labels or probability values. Alternatively, small models with architectural variants are used for semantic image segmentation tasks. For example, a fully convolutional network (FCN) also provides state-of-the-art results for image segmentation tasks in computer vision.[2] Another variant of FCN, SegNet, has also been proposed.[11]

Owing to the great success of deep convolutional neural networks (DCNNs) in the field of computer vision, different variants of this approach are applied in different modalities of medical imaging, including segmentation, classification, detection, registration, and medical information processing. Medical imaging comes from different imaging techniques, such as computer tomography (CT), ultrasound, x-ray, and magnetic resonance imaging (MRI). The goal of computer-aided diagnosis is to obtain a faster and better diagnosis to ensure better treatment of a large number of people at the same time. In addition, efficient automatic processing reduces human error and significantly reduces overall time and cost. Due to the slow process and tedious nature of manual segmentation approaches, there is a significant demand for computer algorithms that can perform segmentation quickly and accurately without human interaction. However, there are some limitations to medical image segmentation, including data scarcity and class imbalance. Most of the time, a large number of labels (e.g., in thousands) are not available for training for several reasons.[12] Labeling the dataset requires an expert in this field, which is expensive, and it requires a lot of effort and time. Sometimes, different data transformation or augmentation techniques (data whitening, rotation, translation, and scaling) are applied for increasing the number of labeled samples available.[13–15] In addition, patch-based approaches are used for solving class imbalance problems. In this work, we have evaluated the proposed approaches on both patch-based and entire image-based approaches. However, to switch from the patch-based approach to the pixel-based approach that works with the entire image, we must be aware of the class imbalance problem. In the case of semantic segmentation, the image backgrounds are assigned a label and the foreground or target regions are assigned with different classes. Therefore, the class imbalance problem is resolved without any trouble. Two advanced techniques, including cross-entropy loss and Dice similarity, have been introduced for efficient training of classification and segmentation tasks in Refs. 14 and 15.

Furthermore, in medical image processing, global localization and context modulation are very often applied for

[*] Address all correspondence to Md Zahangir Alom, E-mail: alomm1@udayton.edu

2329-4302/2019/$25.00 © 2019 SPIE

localization tasks. Each pixel is assigned a class label with the desired boundary that is related to the contour of the target lesion in identification tasks. To define these target lesion boundaries, we must emphasize the related pixels. Landmark detection in medical imaging[16,17] is one such example. There were several traditional machine-learning and image-processing techniques available for medical image segmentation tasks before the DL revolution, including amplitude segmentation based on histogram features,[18] the region-based segmentation method,[19] and the graph-cut approach.[20] However, semantic segmentation approaches that utilize DL have become very popular in recent years in the field of medical image segmentation, lesion detection, and localization.[21] In addition, DL-based approaches are known as universal learning approaches, where a single model can be utilized efficiently in different modalities of medical imaging such as MRI, CT, and x-ray.

According to a recent survey, DL approaches are applied to almost all modalities of medical imaging.[21,22] Furthermore, a large number of papers have been published on segmentation tasks in different modalities of medical imaging.[21,22] A DCNN-based brain tumor segmentation and detection method were proposed in Ref. 23. From an architectural point of view, the CNN model for classification tasks requires an encoding unit and provides class probability as an output. In classification tasks, we have performed convolution operations with activation functions followed by subsampling layers, and this reduces the dimensionality of the feature maps. As the input samples traverse through the layers of the network, the number of feature maps increases but the dimensionality of the feature maps decreases. This is shown in the first part of the model (in green) in Fig. 2. Since the number of feature maps increases in the deeper layers, the number of network parameters also increases. Eventually, the softmax operations are performed at the end of the network to compute the probability of the target classes.

As opposed to classification tasks, the architecture of segmentation tasks requires both convolutional encoding and decoding units. The encoding unit is used to encode input images into a larger number of maps with lower dimensionality. The decoding unit is used to perform upconvolution (transpose convolution, or what is occasionally called deconvolution) operations to produce segmentation maps with the same dimensionality as the original input image. Therefore, the architecture for the segmentation tasks generally requires almost double the number of network parameters when compared to the architecture for the classification tasks. Thus, it is important to design efficient DCNN architectures for segmentation tasks, which can ensure better performance with fewer numbers of network parameters.

This research demonstrates two modified and improved segmentation models: one using recurrent convolution networks and another using recurrent residual convolutional networks. To accomplish our goals, the proposed models are evaluated on different modalities of medical imaging, as shown in Fig. 1. The contributions of this work can be summarized as follows:

- Two new models called recurrent U-Net (RU-Net) and recurrent residual U-Net (R2U-Net) are introduced for medical image segmentation.
- Experiments are conducted on three different modalities of medical imaging, including retinal blood vessel segmentation, skin cancer segmentation, and lung segmentation (LS).
- Performance evaluation of the proposed models is conducted by the patch-based method for retinal blood vessel segmentation tasks and by the end-to-end image-based approach for skin lesion and LS tasks.
- Comparison against recently proposed state-of-the-art methods shows superior performance against equivalent models with the same number of network parameters.
- Empirical evaluation is conducted on the robustness of the proposed R2U-Net model against SegNet[11] and U-Net[13] based on the trade-off between the number of training samples and performance during the training, validation, and testing phases.

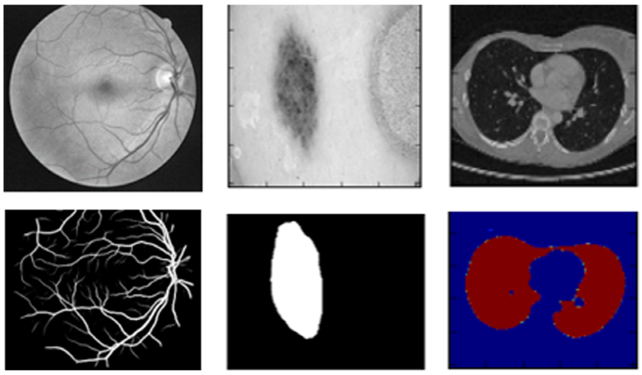


Fig. 1 Medical image segmentation examples displaying RBVS on the left, skin cancer lesion segmentation in the middle, and LS on the right.

The paper is organized as follows: Sec. 2 discusses related work. The architectures of the proposed RU-Net and R2U-Net models are presented in Sec. 3. Section 4 explains experimental setup and performance metrics. The datasets' details and discussion on experimental results are given in Sec. 5. The comparison on experimental results against U-Net and SegNet is given in Sec. 6. The conclusion and future direction are discussed in Sec. 7.

2 Related Works

Semantic segmentation is an active research area where DCNNs are used to classify each pixel in the image individually, which is fueled by different challenging datasets in the fields of computer vision and medical imaging.[23–26] Before the DL revolution, the traditional machine-learning approach mostly relied on hand-engineered features that were used for classifying pixels independently. In the past few years, a lot of models have been proposed that have proved that deeper networks are better for recognition and segmentation tasks.[5] However, training very deep models are difficult due to the vanishing gradient problem, which is resolved by implementing modern activation functions such as rectified linear units (ReLUs) or exponential linear units.[5,6] Another solution to this problem was proposed by He et al.,[27] a deep residual model that overcomes the problem utilizing identity mapping to facilitate the training process.

In addition, CNN-based segmentation methods based on the FCN provide superior performance for natural image segmentation.[2] The performance of FCN has improved with recurrent neural networks, which are fine-tuned on very large datasets.[28] Semantic image segmentation with DeepLab is currently one of the state-of-the-art methods.[29] SegNet consists of two parts: the encoding network, which is a 13-layer VGG-16 network,[5] and the corresponding decoding network that uses pixel-wise classification layers. The main contribution of Ref. 11 is the way in which the decoder upsamples its lower resolution input feature maps. Later, an improved version of SegNet, which is called Bayesian SegNet, was proposed in 2015.[30] Most of these architectures are explored using computer vision applications. However, there are some DL models that have been proposed specifically for the medical image segmentation, as they consider data insufficiency and class imbalance problems.

One of the first and most popular approaches for semantic medical image segmentation is the U-Net.[13] According to the U-Net architecture, the network consists of two main parts: the convolutional encoding and decoding units. The basic convolution operations are performed followed by ReLU activation in both parts of the network. For downsampling in the encoding unit, 2×2 max-pooling operations are performed. In the decoding phase, the convolution transpose (representing upconvolution or deconvolution) operations are performed to upsample the feature maps. The very first version of U-Net had been used for cropping and copying feature maps from the encoding unit to the decoding unit. The U-Net model provides several advantages for segmentation tasks: first, this model allows the use of global location and context at the same time. Second, it works with very few training samples and provides better performance for segmentation tasks.[13] Third, an end-to-end pipeline processes the entire image in the forward pass and directly produces segmentation maps. This ensures that U-Net preserves the full context of the input images, which is a major advantage when compared to patch-based segmentation approaches.[13,15]

However, U-Net is not only limited to applications in the domain of medical imaging, but nowadays this model is also applied for computer vision tasks.[31,32] Meanwhile, different variants of U-Net models have been proposed, including a very simple variant of U-Net for CNN-based segmentation of medical imaging data.[33] In this model, two modifications are made to the original design of U-Net: first, a combination of multiple segmentation maps and forward feature maps are summed (element-wise) from one part of the network to the other. The feature maps are taken from different layers of the encoding and decoding units, and finally, summation (element-wise) is performed outside of the encoding and decoding units. The authors report promising performance improvement during training with better convergence compared to U-Net, but no benefit has been observed when using a summation of features during the testing phase.[33] However, this concept proved that feature summation impacts the performance of a network. The importance of skipped connections for biomedical image segmentation tasks has been empirically evaluated with U-Net and residual networks.[34] The deep contour-aware network had been proposed in 2016, which can extract multilevel contextual features using a hierarchical architecture for accurate gland segmentation of histology images, and it shows very good performance for segmentation.[35] Furthermore, Nabla-Net, a deep dig-like convolutional architecture, had been proposed for segmentation in 2017.[36]

Other DL approaches have been proposed based on U-Net for three-dimensional (3-D) medical image segmentation tasks as well. The 3-D U-Net architecture for volumetric segmentation learned from sparsely annotated volumetric images.[14] A powerful end-to-end 3-D medical image segmentation system based on volumetric images called V-Net has been proposed, which consists of an FCN with residual connections.[15] This paper also introduces a Dice loss layer.[15] Furthermore, a 3-D deeply supervised approach for automated segmentation of volumetric medical images was presented in Ref. 37. HighRes3DNet was proposed using residual networks for 3-D segmentation tasks in 2016.[38] In 2017, a CNN-based brain tumor segmentation approach was proposed using a 3-D CNN model with a fully connected conditional random field.[39] Pancreas segmentation was proposed in Ref. 40, and VoxResNet was proposed in 2016 where a deep voxel-wise residual network was used for brain segmentation. This architecture utilized residual networks and summation of feature maps from different layers.[41]

Alternatively, we have proposed two models for semantic segmentation based on the architecture of U-Net in this paper. The proposed recurrent CNN (RCNN) model based on U-Net is named RU-Net, which is shown in Fig. 2. In addition, we have proposed a residual RCNN (RRCNN)-based U-Net model, which is called R2U-Net. Section 3 provides the architectural details of both these models.

3 RU-Net and R2U-Net Architectures

3.1 *RU-Net and R2U-Net Model Details*

Inspired by the deep residual model,[7] the RCNN,[42] and the U-Net[13] model, we propose two models for segmentation tasks

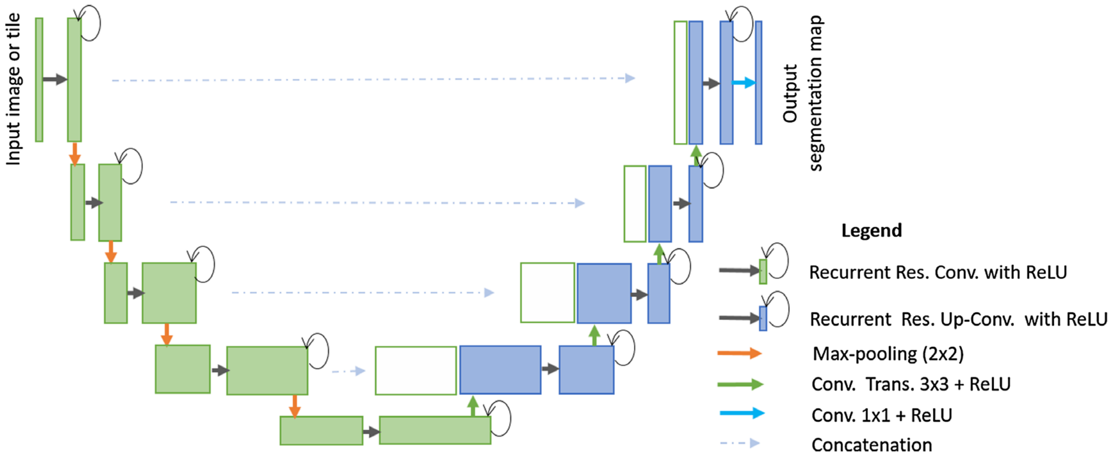


Fig. 2 The RU-Net architecture with convolutional encoding and decoding units using RCLs, which is based on a U-Net architecture. The residual units are used with the RCL for R2U-Net architectures.

that are named RU-Net and R2U-Net. These two approaches utilize the strengths of all three recently developed DL models. The RCNN and its variants have already shown superior performance on object recognition tasks using different benchmarks.[43,44] The recurrent residual convolutional operations can be demonstrated mathematically, according to the improved residual networks in Ref. 44. The operations of the recurrent convolutional layers (RCLs) are performed with respect to the discrete time steps that are expressed according to the RCNN.[42] Let us consider the x_l input sample in the l'th layer of the RRCNN block and a center pixel of a patch located at (i, j) in an input sample on the k'th feature map in the RCL. In addition, let us assume that the output of the network $O^l_{ijk}(t)$ is at the time step t. The output can be expressed as follows:

$$O^l_{ijk}(t) = (w^f_k)^T * x^{f(i,j)}_l(t) + (w^r_k)^T * x^{r(i,j)}_l(t-1) + b_k. \quad (1)$$

Here, $x^{f(i,j)}_l(t)$ and $x^{r(i,j)}_l(t-1)$ are the inputs to the standard convolutional layers and the l'th RCL, respectively. The w^f_k and w^r_k values are the weights of the standard convolutional layer and the RCL of the k'th feature map, respectively, and b_k is the bias. The outputs of the RCL are fed to the standard ReLU activation function f and are expressed as

$$\mathcal{F}(x_l, w_l) = f[O^l_{ijk}(t)] = \max[0, O^l_{ijk}(t)], \quad (2)$$

where $\mathcal{F}(x_l, w_l)$ represents the outputs from of l'th layer of the RCNN unit. The output of $\mathcal{F}(x_l, w_l)$ is used for downsampling and upsampling layers in the convolutional encoding and decoding units of the RU-Net model, respectively. In the case of R2U-Net, the final outputs of the RCNN unit are passed through the residual unit, as shown in Fig. 3(d). Let us consider the output of the RRCNN block to be x_{l+1}, then it can be calculated as follows:

$$x_{l+1} = x_l + \mathcal{F}(x_l, w_l). \quad (3)$$

Here, x_l represents the input samples of the RRCNN block. The x_{l+1} sample is the input for the immediately succeeding subsampling or upsampling layers in the encoding and decoding convolutional units of the R2U-Net model. However, the number of feature maps and the dimensions of the feature maps for the residual units are the same as in the RRCNN block, which is shown in Fig. 3(d).

The proposed DL models are the building blocks of the stacked convolutional units, which are shown in Figs. 3(b) and 3(d). Four different architectures are evaluated in this work. First, the U-Net with forward convolution layers and feature concatenation is applied as an alternative to the crop-and-copy method found in the primary version of U-Net.[13] The basic convolutional unit of this model is shown in Fig. 3(a). Second, the U-Net model with forward convolutional layers with residual connectivity is used, which is often called a residual U-Net (or a ResU-Net) and is shown in Fig. 3(c).[15,31] The third architecture is the U-Net model with forward RCLs, as shown in Fig. 3(b), which is named RU-Net. Finally, the last architecture is the U-Net model with recurrent convolution layers with residual connectivity, as shown in Fig. 3(d), which is named R2U-Net. The pictorial representation of the unfolded RCL layers with respect to time step is shown in Fig. 4. Here, $t = 2$ (0 to 2), refers to the recurrent convolutional operation

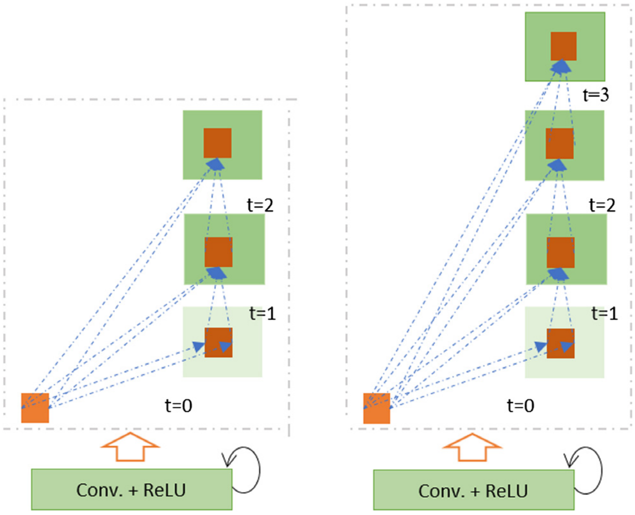


Fig. 4 The lower part of units represents RCUs and upper parts are for unfolded RCUs for $t = 2$ (left) and $t = 3$ (right). For $t = 2$, we have used one forward convolutional layer followed by two RCLs; on the other hand, for $t = 3$, one forward convolutional layer is used followed by three RCLs. The orange and blue arrows represent the equivalent representation of folded and unfolded RCUs and the convolutional operation with respect to different time steps, respectively. The orange and green rectangles indicate the kernels and the feature maps for the respective layers.

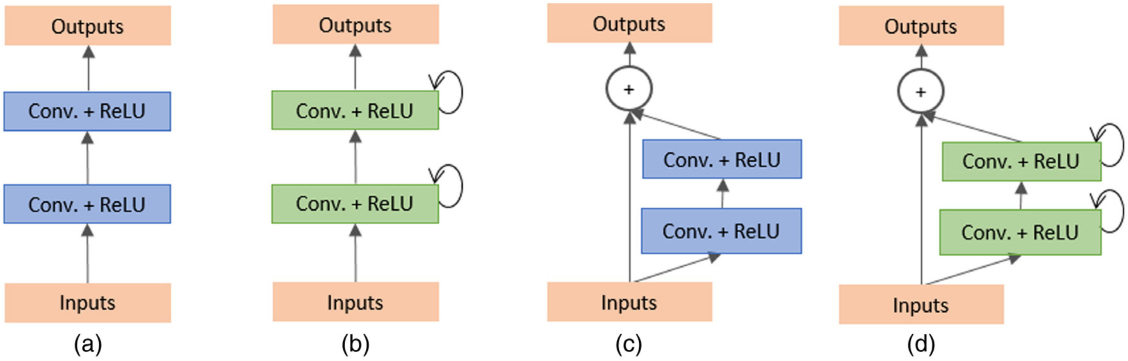


Fig. 3 Different variants of the convolutional and recurrent convolutional units (RCUs) including (a) the forward convolutional unit, (b) the recurrent convolutional block, (c) the residual convolutional unit, and (d) the recurrent residual convolutional unit.

that includes one single convolution layer followed by two subsequential RCLs.

In this implementation, we have applied concatenation to the feature maps from the encoding unit to the decoding unit for the RU-Net and R2U-Net models. The differences between the proposed models with respect to the U-Net model are threefold. This architecture consists of convolutional encoding and decoding units that are the same as those used in the U-Net model. However, the RCLs (and RCLs with residual units) are used instead of regular forward convolutional layers in both the encoding and decoding units. The residual unit with RCLs helps to develop a more efficient deeper model. Second, the efficient feature accumulation method is included in the RCL units of both the proposed models. The effectiveness of feature accumulation from one part of the network to the other is shown in the CNN-based segmentation approach for medical imaging. In this model, the element-wise feature summation is performed outside the U-Net model.[33] The U-Net model only shows the benefit during the training process in the form of better convergence. However, our proposed models show benefits for both training and testing phases due to the feature accumulation inside the model. The feature accumulation with respect to different time steps ensures better and stronger feature representation. Thus, it helps in extracting very-low-level features that are essential for segmentation tasks for different modalities of medical imaging (e.g., blood vessel segmentation). Third, we have removed the cropping and copying unit from the basic U-Net model and use only concatenation operations. Therefore, with all the above-mentioned changes, the proposed models are much better compared to equivalent SegNet, U-Net, and ResU-Net models, which ensure better performance with the same or fewer number of network parameters.

There are several advantages of using the proposed architectures when compared to U-Net. The first is the efficiency in terms of the number of network parameters. The proposed RU-Net and R2U-Net architectures are designed to have the same number of network parameters, when compared to U-Net and ResU-Net, and the RU-Net and R2U-Net models show better performance on segmentation tasks. The recurrent and residual operations do not increase the number of network parameters. However, they do have a significant impact on training and testing performance, which is shown through an empirical evaluation with a set of experiments in the following sections.[44] This approach is also generalizable, as it can easily be applied to DL models based on SegNet,[11] 3D-U-Net,[14] and V-Net[15] with improved performance for segmentation tasks.

3.2 *Model Architecture and Parameters*

Table 1 Architectural details, the numbers of feature maps in the convolutional blocks, and the number of network parameters for RBVS, SLS, and LS.

Dataset	t	Network architectures	Number of parameters (in millions)
RBVS + LS	2	1 → 16(3) → 32(3) → 64(3) → 128(3) → 64(3) → 32(3) → 16(3) → 1	0.841
LS	3	1 → 16(4) → 32(4) → 64(4) → 128(4) → 64(4) → 32(4) → 16(4) → 1	1.037
SLS + RBVS	2	1 → 32(3) → 64(3) → 128(3) → 256(3) → 512(3) → 256(3) → 128(3) → 64(3) → 32(3) → 1	13.34

We have conducted experiments using several different models, including SegNet,[11] U-Net,[13] ResU-Net,[31] RU-Net, and R2U-Net. These models are evaluated with different numbers of convolutional layers in the convolutional blocks, and the numbers of layers are determined with respect to time step t. The network architectures along with the corresponding numbers of feature maps in different convolutional blocks are shown in Table 1. From the table, it can be clearly seen in rows 2 and 4 that the numbers of feature maps in the convolutional blocks remain the same; however, as a convolutional layer is added in the convolutional block when $t = 3$, the number of network parameters increases. Feature fusion is performed with an element-wise addition operation in different residual, recurrent, and recurrent residual units. In the encoding unit of the network, each convolutional block consists of two or three RCLs, where 3 × 3 convolutional kernels are applied, proceeded by ReLU activation layers, followed by a batch normalization layer. For down-sampling, a 2 × 2 max-pooling layer followed by a 1 × 1 convolutional layer is used between the convolutional blocks. In the decoding unit, each block consists of a convolutional transpose layer followed by two convolutional layers and a concatenation layer. We have empirically evaluated different fusion techniques, including addition, concatenation, and addition and concatenation between encoding and decoding units. The concatenation operations perform better compared to the other two methods. Therefore, the concatenation operations are used between the features in the encoding and decoding units in the network. The features are then mapped to a single output feature map, where 1 × 1 convolutional kernels are used with a sigmoid activation function. Finally, the segmentation region is generated with a threshold (T), which is empirically set at 0.5 in our experiment.

The architecture shown in the fourth row in Table 1 is used for retina blood vessel segmentation on the DRIVE dataset, as well as skin cancer segmentation. We have also implemented the SegNet model[11] with similar architecture and a similar number of feature maps for impartial comparison in the cases of skin cancer lesions and LS. The architecture we used can be written as 1 → 32(3) → 64(3) → 128(3) → 256(3) → 512(3) → 256(3) → 128(3) → 64(3) → 32(3) → 1 in the SegNet model for skin cancer lesion segmentation, where each convolutional block contains three convolutional layers and a batch normalization layer that requires 14.94M network parameters. For LS, the architecture can be written as 1 → 32(3) → 64(3) → 128(3) → 256(3) → 128(3) → 64(3) → 32(3) → 1 for the SegNet model (three convolutional layers and a batch normalization layer are used in each block), which requires 1.7M network parameters.

4 Experimental Setup and Evaluation Metrics

4.1 *Experimental Setup*

To demonstrate the performance of the RU-Net and R2U-Net models, we have tested them on three different medical imaging datasets. These include blood vessel segmentation from retina images (DRIVE, STARE, and CHASE_DB1, as shown in Fig. 5), skin cancer lesion segmentation, and LS from

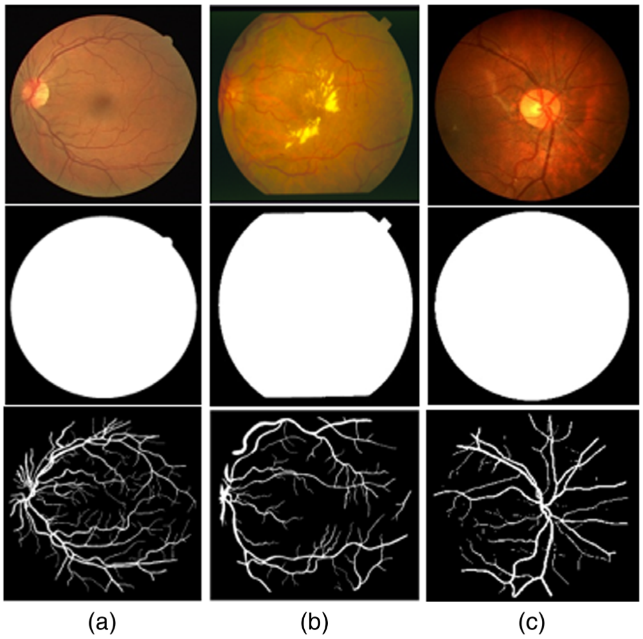


Fig. 5 Example images from training datasets where (a) is taken from the DRIVE dataset, (b) is taken from the STARE dataset, and (c) is taken from the CHASE-DB1 dataset. The first row shows the original images, the second row shows the FOVs, and third row shows the target outputs.

two-dimensional (2-D) images. For this implementation, the Keras and TensorFlow frameworks are used on a single graphics processing units machine with 56 G of RAM and an NVIDIA GEFORCE GTX-980 Ti with 6 GB of memory.

4.2 *Evaluation Metrics*

For quantitative analysis of the experimental results, several performance metrics are considered, including accuracy (AC), sensitivity (SE), specificity (SP), F1-score, Dice coefficient (DC), and Jaccard index (JA). To do this, we also use the variables true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The overall AC is calculated using Eq. (4), and SE and SP are calculated using Eq. (5).

$$\mathrm{AC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \tag{4}$$

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \qquad \mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}. \tag{5}$$

Furthermore, DC and JA are calculated using the following equation:

$$\mathrm{DC} = \frac{2.\mathrm{TP}}{2.\mathrm{TP}+\mathrm{FN}+\mathrm{FP}} \qquad \mathrm{JA} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}}. \tag{6}$$

In addition, we have also conducted an experiment to determine the Dice index (DI) loss function according to Ref. 45, and the Jaccard similarity score (JS) is represented using Eq. (7), as in Ref. 46. Here, GT refers to the ground truth and SR refers to the segmentation result.

$$\mathrm{DI}(\mathrm{GT},\mathrm{SR}) = 2\frac{|\mathrm{GT}\cap\mathrm{SR}|}{|\mathrm{GT}|+|\mathrm{SR}|} \qquad \mathrm{JS}(\mathrm{GT},\mathrm{SR}) == \frac{|\mathrm{GT}\cap\mathrm{SR}|}{|\mathrm{GT}\cup\mathrm{SR}|}. \tag{7}$$

The F1-score is calculated according to the following equation:

$$\mathrm{F1}-\mathrm{score} = 2\times\frac{\mathrm{precision}\times\mathrm{recall}}{\mathrm{precision}+\mathrm{recall}}, \tag{8}$$

where the precision and recall are expressed as

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}; \qquad \mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{9}$$

The area under the curve (AUC) and the receiver operating characteristics (ROC) curve are common evaluation measures for medical image segmentation tasks. In this experiment, we had utilized both analytical methods to evaluate the performance of the proposed approaches and had compared our results to the existing state-of-the-art techniques.

5 Experimental Results

5.1 *Blood Vessel Segmentation*

We have experimented on three different popular datasets for retinal blood vessel segmentation, including DRIVE,[47] STARE,[48] and CHASE_DB1.[49]

5.1.1 *Databases details*

The DRIVE dataset consists of 40 color retina images, of which 20 samples are used for training and the remaining 20 samples are used for testing. The size of each original image is 565 × 584 pixels.[47] To develop a square dataset, the images are cropped to only contain the data from columns 9 to 574, which then makes each image size 565 × 565 pixels. In this implementation, we consider 190,000 randomly selected patches from 20 of the images in the DRIVE dataset, where 171,000 patches are used for training, and the remaining 19,000 patches are used for validation. The size of each patch is 48 × 48 for all the three datasets, as shown in Fig. 6. The second dataset, STARE, contains 20 color images, and each image has a size of 700 × 605 pixels.[48,50] Owing to the small number of samples

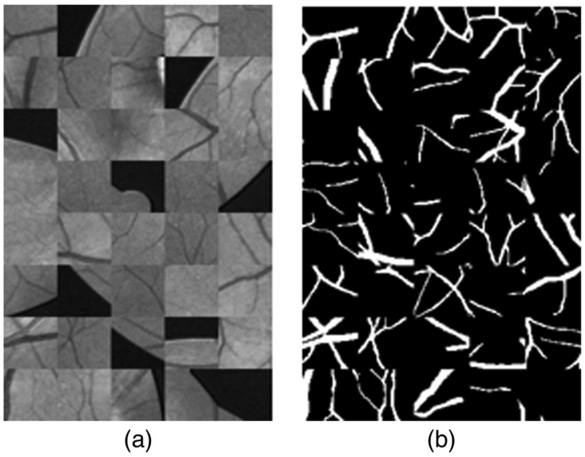


Fig. 6 Example patches are shown in (a) and the corresponding outputs of the patches are shown in (b).

in the STARE dataset, two approaches are often applied for training and testing when using this dataset. First, training is sometimes performed with randomly selected samples from all 20 images.[51]

Another approach is the "leave-one-out" method, where in each trial one image is selected for testing, and training is conducted on the remaining 19 samples.[49,52] Therefore, there is no overlap between the training and testing samples. In this implementation, we used the "leave-one-out" approach for the STARE dataset. The CHASE_DB1 dataset contains 28 color retina images, and the size of each image is 999 × 960 pixels.[49] The images in this dataset are collected from both the left and right eyes of 14 school children. The dataset is divided into two sets where samples are selected randomly. A 20-sample set is used for training and the remaining 8 samples are used for testing.

As the dimensionality of the input data in the STARE and CHASE_DB1 datasets is larger than that of the DRIVE dataset, we considered 250,000 patches in total from 20 images for both STARE and CHASE_DB1 datasets. In this case, 225,000 patches are used for training and the remaining 25,000 patches are used for validation. As the binary field of view (FOV) (which is shown in the second row of Fig. 5) is not available for the STARE and CHASE_DB1 datasets, we generated FOV masks using a similar technique to the one described in Ref. 52. One advantage of the patch-based approach is that the patches give the network access to local information about the pixels, which has an impact on the overall prediction. Furthermore, it ensures that the classes of the input data are balanced. The input patches are randomly sampled over an entire image, which also includes the outside region of the FOV.

5.1.2 *Experimental results*

Owing to the data scarcity of retinal blood vessel segmentation datasets, the patch-based approach is used during training and testing phases. We used a random initialization method and a stochastic gradient descent optimization approach, with categorical cross-entropy loss, a batch size 32, and 150 epochs in this implementation.

Results of DRIVE dataset. Figure 7 shows the training and validation AC when using the DRIVE dataset. The proposed R2U-Net and RU-Net models provide better performance during both the training and the validation phases, when compared to the U-Net and ResU-Net models. Quantitative results are achieved with the four different models using the DRIVE dataset, and the results are shown in Table 2. The overall AC and AUC are considered when comparing the performance of the proposed methods in most cases. The results we have achieved with the proposed models with 0.841M network parameters (Table 1, second row) are higher than those obtained when using the state-of-the-art approaches in most cases. However, to compare with the most recently proposed method,[57] a deeper R2U-Net is evaluated with 13.34M network parameters (Table 1, fourth row) that showed the highest AC (0.9613) and a better AUC of 0.979. Most importantly, we can observe that the proposed RU-Net and R2U-Net models provide better performance in terms of AC and AUC, compared to the U-Net and RU-Net models. The precise segmentation results achieved with the proposed R2U-Net model are shown in Fig. 8(a).

Results of STARE dataset. The quantitative results when using the STARE dataset, along with a comparison to the existing methods, are shown in Table 2. In 2016, a cross-modality learning approach was proposed by Li et al.[56] and had reported AC of ~0.9628 for STARE dataset, which had been previously the highest recorded result. Recently, Zhao et al.[57] proposed a method with a weighted symmetry filter and showed an AC of 0.9570. In this work, we have used the "leave-one-out" method and have reported the average results of five different trials. We have achieved an AC of 0.9712 with the R2U-Net model for the STARE dataset, which is 0.84% and 1.42% better than the results obtained when using the methods proposed by Li et al. and Zhao et al., respectively. In addition, the RU-Net and R2U-Net models outperform the U-Net and ResU-Net models in this experiment. The R2U-Net model shows 0.22% and 0.12% better AC compared to U-Net and ResU-Net, respectively. The qualitative results of R2U-Net when using the STARE dataset are shown in Fig. 8(b).

Results of CHASE_DB1 dataset. The results of the quantitative analysis are given in Table 1. From the table, it can be seen that the RU-Net and R2U-Net models provide better performance than the U-Net and ResU-Net models when applying

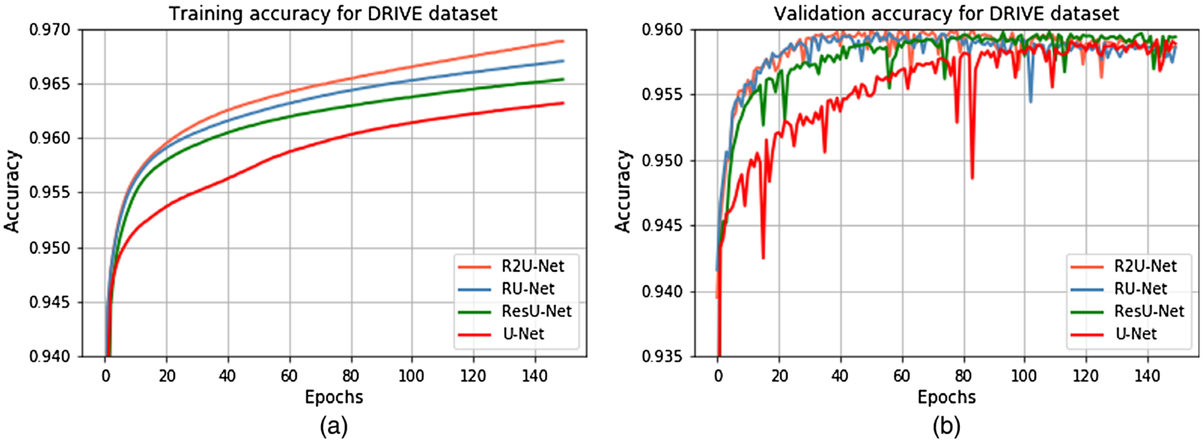


Fig. 7 Training and validation AC of the proposed RU-Net and R2U-Net models compared to the ResU-Net and U-Net models for blood vessel segmentation task. (a) Training AC and (b) validation.

Table 2 Experimental results of the proposed approaches for RBVS and their comparison with other traditional and DL-based approaches.

Dataset	Methods	Year	SE	SP	AC	AUC
DRIVE	Cheng et al.[51]	2014	o.7252	0.9798	0.9474	0.9648
	Azzopardi et al.[53]	2015	0.7655	0.9704	0.9442	0.9614
	Roychowdhury et al.[54]	2016	0.7250	0.9830	0.9520	0.9620
	Liskowski and Krawiec[55]	2016	0.7763	0.9768	0.9495	0.9720
	Li et al.[56]	2016	0.7569	0.9816	0.9527	0.9738
	Zhao et al.[57]	2018	0.7740	0.9790	0.9580	0.9750
	U-Net (1.07M)	2018	0.7537	0.9820	0.9531	0.9755
	ResU-Net (1.07M)	2018	0.7726	0.9820	0.9553	0.9779
	RU-Net (1.07M)	2018	0.7751	0.9816	0.9556	0.9782
	R2U-Net (1.07M)	2018	0.7792	0.9813	0.9556	0.9784
	R2U-Net (13.34M)	2018	0.7661	0.9807	**0.9613**	0.9793
STARE	Marín et al.[58]	2011	0.6940	0.9770	0.9520	0.9820
	Fraz et al.[59]	2012	0.7548	0.9763	0.9534	0.9768
	Roychowdhury et al.[54]	2016	0.7720	0.9730	0.9510	0.9690
	Liskowski and Krawiec[55]	2016	0.7867	0.9754	0.9566	0.9785
	Li et al.[56]	2016	0.7726	0.9844	0.9628	0.9879
	Zhao et al.[57]	2018	0.7880	0.9760	0.9570	0.9590
	U-Net (1.07M)	2018	0.8270	0.9842	0.9690	0.9898
	ResU-Net (1.07M)	2018	0.8203	0.9856	0.9700	0.9904
	RU-Net (1.07M)	2018	0.8108	0.9871	0.9706	0.9909
	R2U-Net (1.07M)	2018	0.8298	0.9862	**0.9712**	0.9914
CHASE_DB1	Fraz et al.[59]	2012	0.7224	0.9711	0.9469	0.9712
	Fraz et al.[60]	2014	—	—	0.9524	0.9760
	Azzopardi et al.[53]	2015	0.7655	0.9704	0.9442	0.9614
	Roychowdhury et al.[54]	2016	0.7201	0.9824	0.9530	0.9532
	Azzopardi et al.[53]	2016	0.7507	0.9793	0.9581	0.9793
	U-Net (1.07M)	2018	0.8288	0.9701	0.9578	0.9772
	ResU-Net(1.07M)	2018	0.7726	0.9820	0.9553	0.9779
	RU-Net (1.07M)	2018	0.7459	0.9836	0.9622	0.9803
	R2U-Net (1.07M)	2018	0.7756	0.9820	**0.9634**	**0.9815**

Note: Bold values indicate the highest testing accuracy for the task.

the CHASE-DB1 dataset. In addition, the proposed methods are compared against the recently proposed approaches for blood vessel segmentation using the CHASE_DB1 dataset. Li et al.[56] proposed an approach with cross-modality learning and achieved an AC of 0.9581. However, we have achieved an AC of ∼0.9634 with the R2U-Net model, which is about 0.53% improvement, compared to the result in Ref. 56. The precise segmentation results with the proposed R2U-Net model on the CHASE_DB1 dataset are shown in Fig. 8(c).

The ROC curve for the highest AUCs of the R2U-Net (with 1.07M network parameters) model on each of the three retina blood vessel segmentation (RBVS) datasets is shown in Fig. 9.

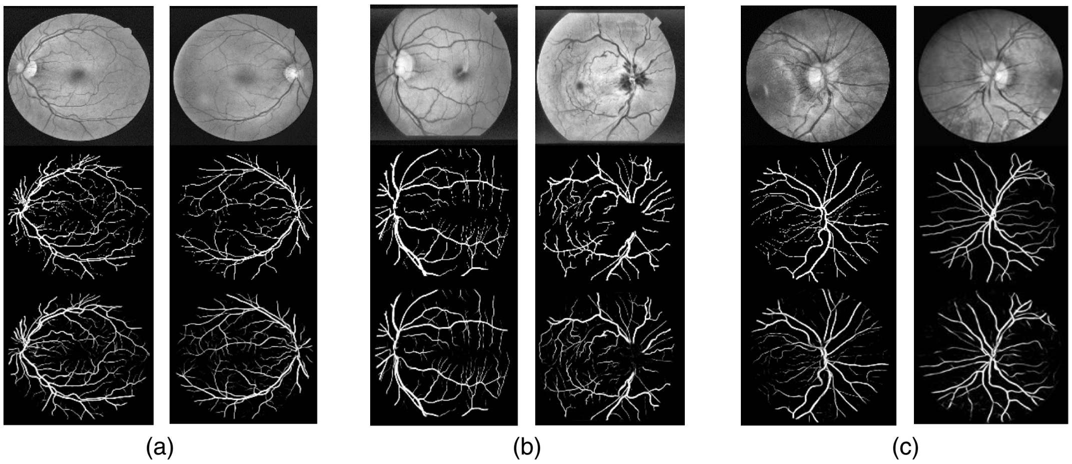


Fig. 8 Experimental outputs for three different datasets for RBVS using R2U-Net. The first row shows input images in grayscale, the second row shows the ground truth, and the third row shows the experimental outputs. The images correspond to the (a) DRIVE, (b) STARE, and (c) CHASE_DB1 datasets.

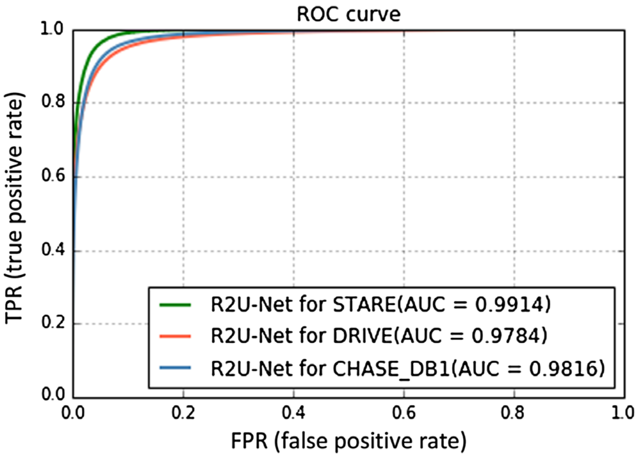


Fig. 9 AUC for RBVS for the best performance achieved with R2U-Net on three different datasets.

5.2 *Skin Cancer Segmentation*

5.2.1 *Database*

This dataset is taken from the Kaggle competition on skin lesion segmentation (SLS) that occurred in 2016.[61] This dataset contains 900 images, along with associated ground-truth samples for training. Another set of 379 images is provided for testing. The original size of each sample is 700 × 900, which is rescaled to 128 × 128 for this implementation. The training samples include the original images, as well as corresponding target binary images containing cancerous or noncancerous lesions. The target pixels are set to a value of either 255 or 0, denoting pixels inside or outside the target lesion, respectively.

5.2.2 *Experimental results*

In this implementation, this dataset was preprocessed with mean subtraction and was normalized according to the standard deviation. We used the ADAM optimization technique with a learning rate of 2×10^{-4} and binary cross-entropy loss. In addition, we also calculated the means squared error during the training and validation phase. In this case, 10% of the samples were used for validation during training with a batch size of 32 and 150 epochs. The training AC of the proposed R2U-Net and RU-Net models was compared with that of the ResU-Net and U-Net models for an end-to-end image-based segmentation approach. The training and the validation AC for all four models are shown in Fig. 10. In both cases, the proposed RU-Net and R2U-Net models showed better performance when compared with the equivalent U-Net and ResU-Net models. This clearly demonstrated the robustness of the learning phase of the proposed models for end-to-end image-based segmentation tasks.

The quantitative results of this experiment are compared against the existing methods, as shown in Table 3. We have evaluated the proposed RU-Net and R2U-Net models with respect to the time step $t = 2$ in the RCL unit. The time step value $t = 2$ means that the RCL unit consists of one forward convolution followed by two RCLs. We compared the proposed approaches against the recently published results using performance metrics, including SE, SP, AC, AUC, and DC. The proposed R2U-Net model provides a testing AC of 0.9472 with a higher AUC, which is 0.9430. Furthermore, the JA and DC are calculated for all models, and the R2U-Net model provides the values 0.9278 for JA and 0.9627 for the DC for SLS. Although we are in the third position in terms of AC compared to ISIC-2016[61] (highest) and FCRN-50[63] (second highest), the proposed R2U-Net models show better performance in term of the DC and JA. These results were achieved with an R2U-Net model with 34 layers that contained ∼13.34M network parameters. The architectural detail is shown in Table 1. However, the AC of the proposed RU-Net and R2U-Net models is still higher when compared to the FCRN-38 networks.[63] In addition, the work presented in Refs. 61 and 63 is evaluated with the VGG-16 and Inception-V3 models for skin cancer lesion segmentation. These models contain ∼138M and 23M network parameters, respectively. Furthermore, the RU-Net and R2U-Net models show higher AC and AUC, compared to the VGG-16[63] and GoolgeNet models.[63] In most cases, the RU-Net and R2U-Net models show better performance against equivalent SegNet,[11] U-Net,[13] and ResU-Net[31] models for SLS.

Some qualitative outputs of the SegNet, U-Net, and R2U-Net model for skin cancer lesion segmentation are shown for visual

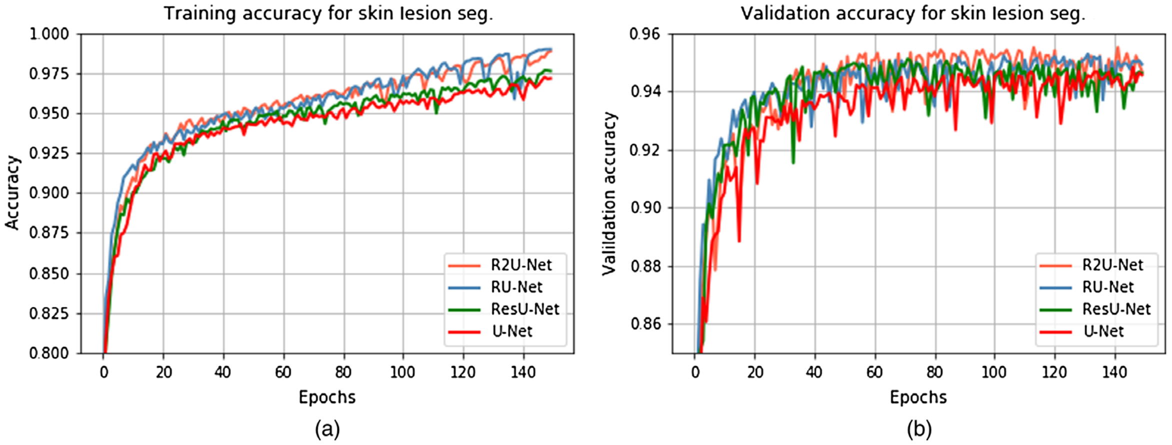


Fig. 10 Training and validation AC of R2U-Net, RU-Net, ResU-Net, and U-Net for SLS. (a) Training AC and (b) validation AC.

Table 3 Experimental results of the proposed approaches for skin cancer lesion segmentation and their comparison with other traditional and DL-based approaches.

Methods	Year	SE	SP	AC	AUC	DC	JA
ISIC-2016[61]	2016	0.910	0.965	**0.953**	—	—	08430
Conv. classifier VGG-16[62]	2017	0.533	—	0.613	0.6420	—	—
Conv. classifier Inception-v3[62]	2017	0.760	—	0.693	0.7390	—	—
VGG-16[63]	2017	0.796	0.945	0.903	—	0.794	0.707
GoogleNet[63]	2017	0.901	0.916	0.916	—	0.848	0.776
FCRN-38[63]	2017	0.882	0.932	0.929	—	0.856	0.785
FCRN-50[63]	2017	0.911	0.957	**0.949**	—	0.897	0.829
FCRN-101[63]	2017	0.903	0.903	0.937	—	0.872	0.803
SegNet[11]	2018	0.9395	0.9222	0.9263	0.9308	0.9502	0.9052
U-Net (16.67M)	2018	0.9457	0.9307	0.9343	0.9324	0.9554	0.9148
ResU-Net (16.67M)	2018	0.9287	0.9479	0.9432	0.9378	0.9608	0.9245
RecU-Net (16.67M)	2018	0.9477	0.9443	0.9458	0.9383	0.9624	0.9273
R2U-Net (16.67M)	2018	**0.9224**	0.9545	**0.9472**	**0.9430**	**0.9627**	**0.9278**

Note: The results of VGG-16 and GoogleNet are taken from Ref. 63. Bold values indicate the highest testing accuracy for the task.

comparison in Fig. 11. In most cases, the target lesions are segmented accurately with a similar shape in ground truth.

However, if we closely observe the outputs in the first, second, and fourth rows of images in Fig. 11, it can be clearly distinguished that the proposed R2U-Net model provides a very similar output shape to the ground truth when compared to the outputs of the SegNet and U-Net models. If we observe the third row of images in Fig. 11, it can be clearly seen that the input image contains three lesions. One is a target lesion, and the other brighter lesions are not targets. The R2U-Net model segments the desired part of the image more accurately when compared to the SegNet and U-Net models. Finally, the fifth row clearly demonstrates that the R2U-Net model provides a very similar shape to the ground truth, which is a much better representation than those obtained from the SegNet and U-Net models. Thus, it can be stated that the R2U-Net model is more capable and robust for skin cancer lesion segmentation.

5.3 *Lung Segmentation*

5.3.1 *Database*

The Lung Nodule Analysis (LUNA)-16 competition at the Kaggle Data Science Bowl, in 2017, was held to find lung lesions in 2-D and 3-D CT images. This dataset consisted of 267 2-D samples, each containing a sample photograph, and

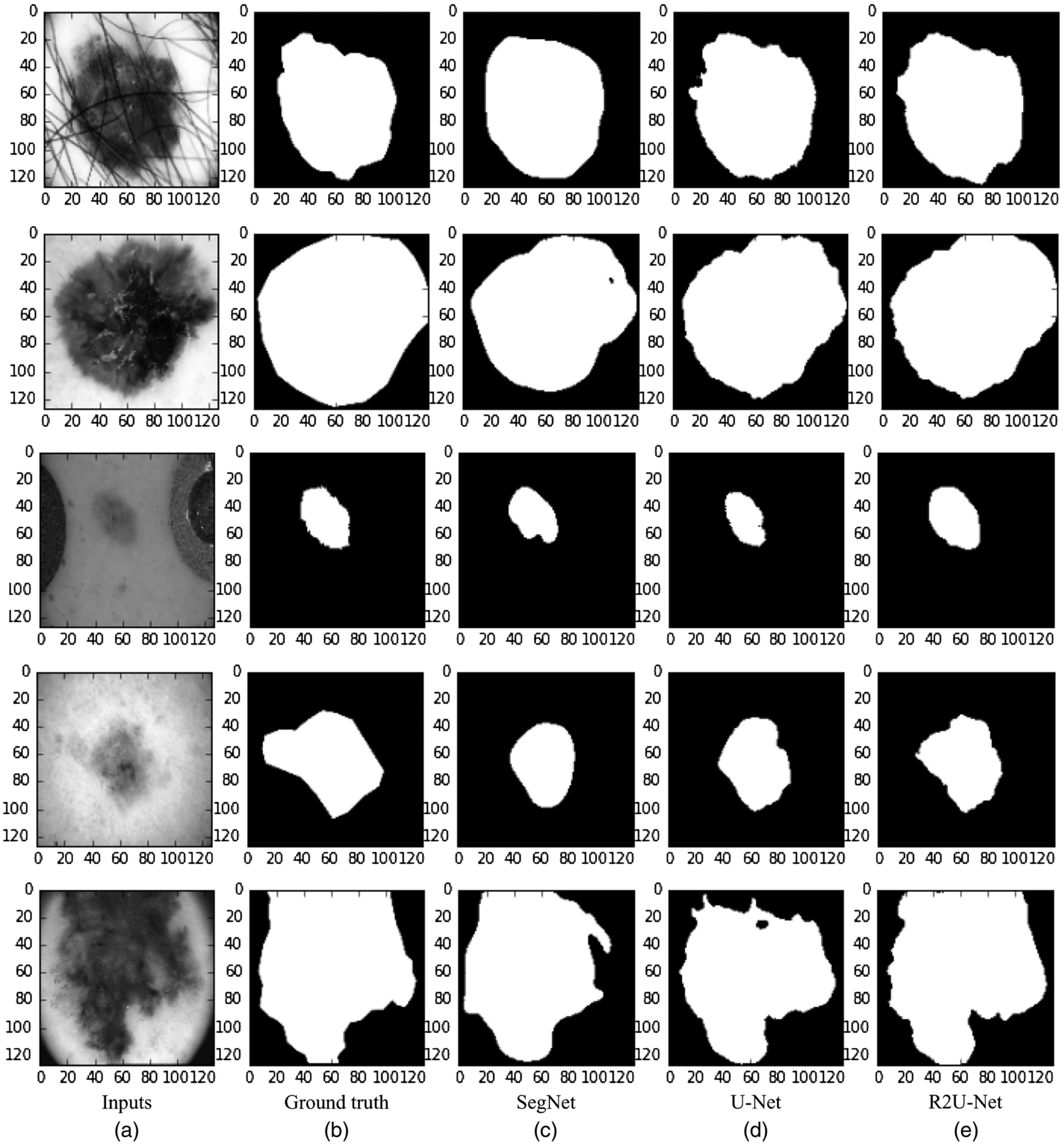


Fig. 11 Illustration of qualitative assessment of the proposed R2U-Net for the skin cancer segmentation task. (a) The input sample, (b) ground truth, (c) the outputs from the SegNet[11] model, (d) the outputs from the U-Net[12] model, and (e) the results of the proposed R2U-Net model.

label image displaying correct LS.[64] For this study, 80% of the images were used for training, and the remaining 20% were used for testing. The original image size was 512 × 512; however, we resized the images to 256 × 256 pixels in this implementation.

5.3.2 *Experimental results*

LS is very important for analyzing lung-related diseases, and it can be applied to lung cancer segmentation and lung pattern classification for identifying other problems. In this experiment, the ADAM optimizer is used with a learning rate of 2×10^{-4}.

We have used DI loss function, according to Eq. (7). In this case, 10% of the samples are used for validation with a batch size of 16 for 150 epochs. Table 4 shows a summary of how well the proposed models performed against the equivalent SegNet,[11] U-Net, and ResU-Net models. In terms of AC, the proposed R2U-Net model has showed 0.26% and 0.55% better testing AC compared to the equivalent SegNet[11] and U-Net[13] models, respectively. In addition, the R2U-Net model provided 0.18% better AC against the ResU-Net model with the same number of network parameters. The qualitative results are shown in Fig. 12, where the first column shows the input samples, the second column represents ground truth, and the third, fourth, and fifth columns show the outputs of the SegNet,[11] U-Net,[13] and R2U-Net models, respectively. It can be visualized that the R2U-Net shows better segmentation results with internal details that are very similar to those displayed in the ground data. If we observe the input, the ground truth, and the output of the

Table 4 The experimental results of the proposed RU-Net and R2U-Net approaches for lung segmentation and their comparison with the SegNet, U-Net, and ResU-Net models for $t = 2$ and $t = 3$.

Methods	Year	SE	SP	JSC	F1-Score	AC	AUC	DI
SegNet (1.02M)[11]	2018	0.9766	0.9791	0.9784	0.9575	0.9784	0.9778	0.9652
SegNet (1.752M)[11]	2018	0.9757	0.9931	0.9887	0.9777	0.9887	0.9844	0.9754
U-Net ($t = 2$)	2018	0.8645	0.9929	0.9635	0.9156	0.9635	0.9287	0.9780
ResU-Net ($t = 2$)	2018	0.9781	0.9975	0.9781	0.9522	0.9781	0.9568	0.9792
RU-Net ($t = 2$)	2018	0.9747	0.9962	0.9911	0.9811	0.9911	0.9855	0.9831
R2U-Net ($t = 2$)	2018	0.9861	0.9940	0.9922	0.9830	0.9922	0.9901	0.9857
U-Net ($t = 2$)	2018	0.9816	0.9945	0.9916	0.9822	0.9916	0.9881	0.9801
ResU-Net ($t = 2$)	2018	0.9838	0.9951	0.9926	0.9833	0.9926	0.9895	0.9825
RU-Net ($t = 2$)	2018	0.9875	0.9959	0.9942	0.9872	0.9942	0.9918	0.9863
R2U-Net ($t = 2$)	2018	0.9912	0.9952	0.9943	0.9879	**0.9944**	**0.9933**	**0.9880**

Note: Bold values indicate the highest testing accuracy for the task.

different approaches in the first and second rows, the outputs of the proposed approaches show better segmentation with more accurate internal details. In the third row, the R2U-Net model clearly defines the inside hole in the left lung, whereas the SegNet[11] and U-Net[13] models do not capture this detail. The last row of images in Fig. 12 shows that the SegNet[11] and U-Net models provide outputs that incorrectly capture parts of the image that are outside of the lesion. On the contrary, the R2U-Net model provides a much more accurate segmentation result. Many models struggle to define the class boundary properly during segmentation tasks.[65] The outputs in Fig. 12 are provided as heatmaps, which show the sharpness of the segmentation borders. These outputs show that the ground truth tends to have a sharper boundary when compared to the model outputs. The ROC with AUCs is shown in Fig. 13. The highest AUC is achieved in the proposed R2U-Net model.

In this implementation, we evaluated both proposed models for patch-based modeling of retinal blood vessel segmentation and end-to-end image-based methods for skin and lung lesion segmentation. In both cases, the proposed models outperformed the existing state-of-the-art methods, including SegNet,[11] U-Net,[13] ResU-Net,[31] and FCRN-38,[63] in terms of AUC and AC on all three datasets. Thus, the quantitative and qualitative results clearly demonstrated the effectiveness of the proposed approach for segmentation tasks.

6 Discussions

6.1 *Trade-off between the Number of Training Samples versus Accuracy*

To further investigate the performance of the proposed R2U-Net model, the trade-off between the number of training samples versus the performance was investigated for the LS dataset. We considered the U-Net and R2U-Net models with $t = 3$, and these models contained 1.07M network parameters. In the case of SegNet,[11] we considered a similar architecture that was proposed in Ref. 11 with 1.7M network parameters. At the beginning of the experiment, the entire dataset was divided into two sets, where 80% of the samples were used for training and validation, and the remaining 20% of the samples were used for testing during each trail. During this experiment, we used different split ratios of [0.9, 0.7, 0.5, 0.3, and 0.1] where the number of training samples was increased, and the number of validation samples was decreased for each successive trail. For example, a split ratio of 0.9 meant that only 10% of the samples were used for training and the remaining 90% of the samples were used for validation. Likewise, a split ratio of 0.7 meant that only 30% of the samples were used for training and the remaining 70% of the samples were used for validation. Figures 14(a) and 14(b) show the training and validation DI coefficient errors (1-DI) with respect to the number of training and validation samples. In each trial, we considered 150 epochs, and the errors presented were the average training and validation errors of the last 20 epochs.

These figures show that the proposed R2U-Net model shows the lowest training and validation error for all of the tested split ratios, except for the result where the split ratio is equal to 0.5 for the validation case.

In this case, the error of the R2U-Net model is only slightly greater than that of the U-Net model. These results clearly demonstrate that the R2U-Net model is a more capable tool when used for extracting, representing, and learning features during the training phase, which ultimately helps in ensuring a better performance. In each trial, we have tested the models with the remaining 20% of the samples, and the testing errors are shown in Fig. 15. The R2U-Net model shows the lowest error for almost all trials relative to the error obtained from the SegNet[11] and U-Net[13] models.

6.2 *Network Parameters Versus Accuracy*

In our experiments, the U-Net, ResU-Net, RU-Net, and R2U-Net models were utilized with the following architecture: 1 → 16 → 32 → 64 → 128 → 64 → 32 → 16 → 1 for retinal blood vessel segmentation and LS. In the case of the retinal blood vessel segmentation, we used a time step of $t = 2$. This same architecture was tested for lung lesion segmentation for both $t = 2$ and $t = 3$. Even though the number of network

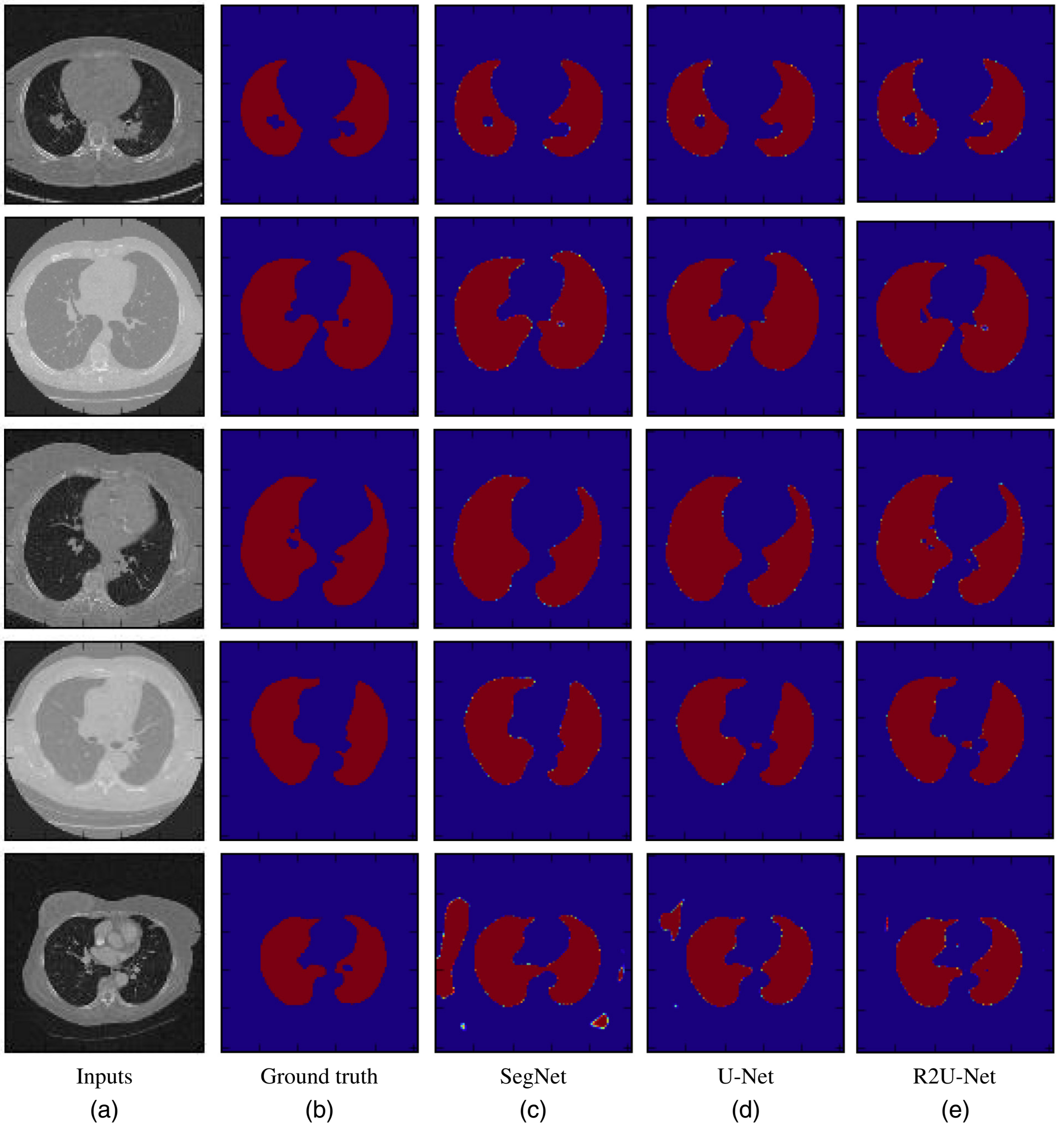


Fig. 12 The experimental results for LS, where (a) shows the inputs, (b) shows the ground truth, (c) shows the outputs of SegNet,[10] (d) shows the outputs of U-Net,[12] and (e) shows the outputs of R2U-Net.

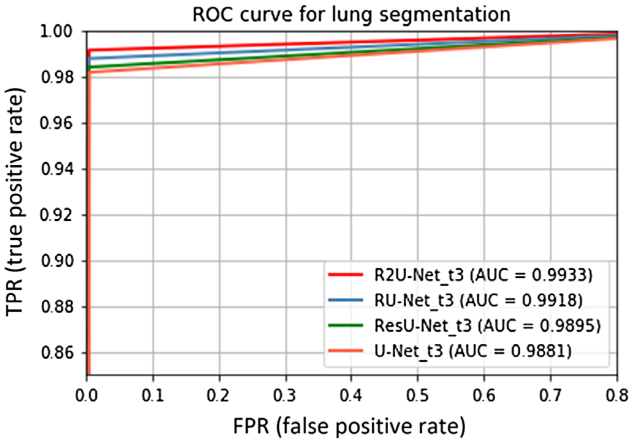
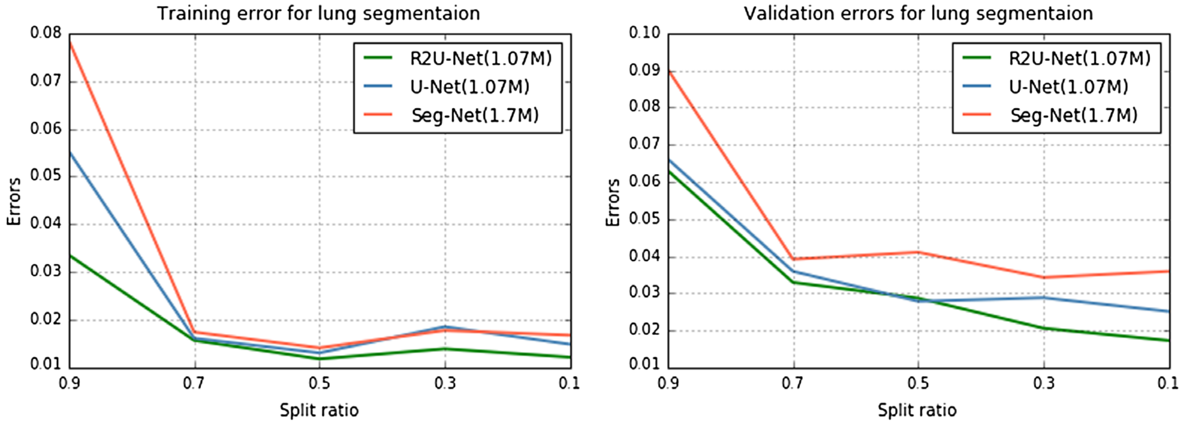


Fig. 13 ROC curve for LS for four different models, where $t = 3$.

parameters slightly increased with respect to the time step in the recurrent convolution layer, improved performance was still observed, as seen in the last rows of Table 4. Furthermore, we implemented an equivalent SegNet[11] model that required 1.73M and 14.94M network parameters, respectively. For skin cancer lesion and LS, the proposed models showed better performance against SegNet[11] when using both 1.07M and 13.34M network parameters, which were around 0.7M and 2.66M less when compared to SegNet.[11] Thus, it can be stated that our model provided better performance with the same or fewer number of network parameters, compared to the SegNet, U-Net, and ResU-Net model. Thus, our proposed model possessed significant advantages in terms of memory and processing time.

6.3 *Computational Times*

The computational time for training per epoch and for segment per sample in the testing phase is shown in Table 5, for all three



(a) (b)

Fig. 14 The performance of the three different models (SegNet, U-Net, and R2U-Net) for different numbers of training and validation samples, where (a) the training DI coefficient errors (1-DI) and (b) validation DI coefficient errors for five different trials are displayed.

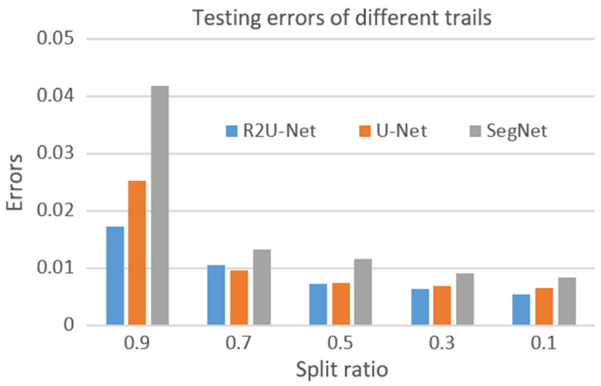


Fig. 15 Testing errors of the R2U-Net, SegNet, and U-Net models for different split ratios for the LS application.

Table 5 Computational time for training and testing phases.

Dataset		Training time (s/epoch)	Testing time (s/sample)
Blood vessel segmentation	DRIVE	209	2.84
	STARE	217	6.42
	CHASE_DB1	283	8.66
Skin cancer segmentation		23	0.32
Lung segmentation		14	1.15

applications. For blood vessel segmentation, the model takes around 209, 217, and 283 s/epoch for the DRIVE, STARE, and CHASE_DB datasets, respectively. The training time for skin cancer and LS tasks are 23 and 14 s, respectively. On the other hand, the processing times during the testing phase for the DRIVE, STARE, and CHASE_DB datasets are 2.84, 6.42, and 8.66 s/sample, respectively. According to Ref. 56, it would take around 90 s on average to segment an entire image (which is equivalent to a few thousand image patches). Alternatively, the proposed R2U-Net approach takes around 6 s/sample, which is an acceptable rate in a clinical use scenario. In addition, when executing skin cancer segmentation and LS, entire images could be segmented in 0.32 and 1.145 s, respectively.

7 Conclusions and Future Works

In this paper, we proposed an extension of the U-Net architecture using RCNNs and recurrent residual CNNs. The proposed models have been called "RU-Net" and "R2U-Net," respectively. These models were evaluated using three different applications in the field of medical imaging, including retinal blood vessel segmentation, skin cancer lesion segmentation, and LS. The experimental results demonstrated that the proposed RU-Net and R2U-Net models showed better performance in most of the cases for segmentation tasks with the same number of network parameters when compared to the existing methods, including the SegNet, U-Net, and ResU-Net models, on all three datasets. The quantitative and qualitative results, as well as the trade-off between the number of training samples versus performance, showed that the proposed RU-Net and R2U-Net models were more capable of learning during training, which ultimately showed better testing performance. In the future, we would like to extend this model to a 3-D architecture to carry out a 3-D medical imaging analysis, including 3-D LS, brain tumor segmentation, and detection from 3-D MRI images.

Disclosures

No conflicts of interest, financial or otherwise, are declared by authors.

Acknowledgments

We would like to express our very great appreciation to the database providers for providing retinal blood vessel, skin cancer, and lung segmentation databases. This work was supported by the National Science Foundation (NSF) under Award Nos. 1718633 and 1309708.

References

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst.* (2012).

2. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3431–3440 (2015).
3. N. Wang et al., "Transferring rich feature hierarchies for robust visual tracking," arXiv:1501.04587 (2015).
4. J. Mao et al., "Deep captioning with multimodal recurrent neural networks (m-RNN)," arXiv:1412.6632 (2014).
5. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556 (2014).
6. C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.* (2015).
7. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.* (2016).
8. G. Huang et al., "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, 4700–4708 (2017).
9. S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Adv. Neural Inf. Process. Syst.* (2017).
10. M. Z. Alom et al., "The history began from AlexNet: a comprehensive survey on deep learning approaches," arXiv:1803.01164 (2018).
11. V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(12), 2481–2495 (2017).
12. D. Ciresan et al., "Deep neural networks segment neuronal membranes in electron microscopy images," in *Adv. Neural Inf. Process. Syst.* (2012).
13. O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
14. Ö. Çiçek et al., "3D U-Net: learning dense volumetric segmentation from sparse annotation," *Lect. Notes Comput. Sci.* **9901**, 424–432 (2016).
15. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *Fourth Int. Conf. 3D Vision (3DV)*, IEEE (2016).
16. D. Yang et al., "Automated anatomical landmark detection on distal femur surface using convolutional neural network," in *IEEE 12th Int. Symp. Biomed. Imaging*, IEEE (2015).
17. Y. Cai et al., "Multi-modal vertebrae recognition using transformed deep convolution network," *Comput. Med. Imaging Graphics* **51**, 11–19 (2016).
18. N. Ramesh, J.-H. Yoo, and I. K. Sethi, "Thresholding based on histogram approximation," *IEE Proc. Vision, Image Signal Process.* **142**(5), 271–279 (1995).
19. N. Sharma and A. K. Ray, "Computer aided segmentation of medical images based on hybridized approach of edge and region based techniques," in *Proc. Int. Conf. Math. Biol.*, Mathematical Biology Recent Trends, Anamaya Publishers (2006).
20. Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in ND images," in *Proc. Eighth IEEE Int. Conf. Computer Vision*, IEEE, Vol. 1 (2001).
21. G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.* **42**, 60–88 (2017).
22. H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique," *IEEE Trans. Med. Imaging* **35**(5), 1153–1159 (2016).
23. M. Havaei et al., "Brain tumor segmentation with deep neural networks," *Med. Image Anal.* **35**, 18–31 (2017).
24. G. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: a high-definition ground truth database," *Pattern Recognit. Lett.* **30**(2), 88–97 (2009).
25. S. Song, S. P. Lichtenberg, and J. Xiao, "Sun RGB-D: a RGB-D scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 567–576 (2015).
26. M. Kistler et al., "The virtual skeleton database: an open access repository for biomedical research and collaboration," *J. Med. Internet Res.* **15**(11), e245 (2013).
27. K. He et al., "Identity mappings in deep residual networks," *Lect. Notes Comput. Sci.* **9908**, 630–645 (2016).
28. S. Zheng et al., "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 1529–1537 (2015).
29. L.-C. Chen et al., "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Int. Conf. Learn. Represent.* (2015).
30. A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," arXiv:1511.02680 (2015).
31. Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.* **15**(5), 749–753 (2018).
32. R. Li et al., "DeepUNet: a deep fully convolutional network for pixel-level sea-land segmentation," *IEEE J. Selected Topics in Appl. Earth Observations and Remote Sens.* **99**, 1–9 (2018).
33. B. Kayalibay, G. Jensen, and P. van der Smagt, "CNN-based segmentation of medical imaging data," arXiv:1701.03056 (2017).
34. M. Drozdzal et al., "The importance of skip connections in biomedical image segmentation," in *Int. Workshop Large-Scale Annotation of Biomed. Data and Expert Label Synth.*, Springer International Publishing (2016).
35. H. Chen et al., "DCAN: deep contour-aware networks for accurate gland segmentation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.* (2016).
36. R. McKinley et al., "Nabla-net: a deep dag-like convolutional architecture for biomedical image segmentation," *Lect. Notes Comput. Sci.* **10154**, 119–128 (2016).
37. Q. Dou et al., "3D deeply supervised network for automated segmentation of volumetric medical images," *Med. Image Anal.* **41**, 40–54 (2017).
38. W. Li et al., "On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task," *Lect. Notes Comput. Sci.* **10265**, 348–360 (2017).
39. K. Kamnitsas et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.* **36**, 61–78 (2017).
40. H. R. Roth et al., "Deeporgan: multi-level deep convolutional networks for automated pancreas segmentation," *Lect. Notes Comput. Sci.* **9349**, 556–564 (2015).
41. H. Chen et al., "Voxresnet: deep voxelwise residual networks for volumetric brain segmentation," arXiv:1608.05895 (2016).
42. M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.* (2015).
43. M. Z. Alom et al., "Inception recurrent convolutional neural network for object recognition," arXiv:1704.07709 (2017).
44. M. Z. Alom et al., "Improved inception-residual convolutional neural network for object recognition," *Neural Comput. Appl.* , 1–15 (2017).
45. L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology* **26**(3), 297–302 (1945).
46. P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytol.* **11**(2), 37–50 (1912).
47. J. Staal et al., "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imaging* **23**(4), 501–509 (2004).
48. A. D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Trans. Med. Imaging* **19**(3), 203–210 (2000).
49. M. M. Fraz et al., "Blood vessel segmentation methodologies in retinal images—a survey," *Comput. Methods Programs Biomed.* **108**(1), 407–433 (2012).
50. Y. Zhao et al., "Automated vessel segmentation using infinite perimeter active contour model with hybrid region information with application to retinal images," *IEEE Trans. Med. Imaging* **34**(9), 1797–1807 (2015).
51. E. Cheng et al., "Discriminative vessel segmentation in retinal images by fusing context-aware hybrid features," *Mach. Vision Appl.* **25**(7), 1779–1792 (2014).
52. J. V. B. Soares et al., "Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification," *IEEE Trans. Med. Imaging* **25**(9), 1214–1222 (2006).
53. G. Azzopardi et al., "Trainable COSFIRE filters for vessel delineation with application to retinal images," *Med. Image Anal.* **19**(1), 46–57 (2015).
54. S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi, "Blood vessel segmentation of fundus images by major vessel extraction and subimage classification," *IEEE J. Biomed. Health Inf.* **19**(3), 1118–1128 (2015).
55. P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE Trans. Med. Imaging* **35**(11), 2369–2380 (2016).

56. Q. Li et al., "A cross-modality learning approach for vessel segmentation in retinal images," *IEEE Trans. Med. Imaging* **35**(1), 109–118 (2016).
57. Y. Zhao et al., "Automatic 2-D/3-D vessel enhancement in multiple modality images using a weighted symmetry filter," *IEEE Trans. Med. Imaging* **37**(2), 438–450 (2018).
58. D. Marín et al., "A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features," *IEEE Trans. Med. Imaging* **30**(1), 146–158 (2011).
59. M. M. Fraz et al., "An ensemble classification-based approach applied to retinal blood vessel segmentation," *IEEE Trans. Biomed. Eng.* **59**(9), 2538–2548 (2012).
60. M. M. Fraz et al., "Delineation of blood vessels in pediatric retinal images using decision trees-based ensemble classification," *Int. J. Comput. Assisted Radiol. Surg.* **9**(5), 795–811 (2014).
61. D. Gutman et al., "Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC)," arXiv:1605.01397 (2016).
62. J. Burdick et al., "Rethinking skin lesion segmentation in a convolutional classifier," *J. Digital Imaging* **31**(4), 435–440 (2018).
63. L. Yu et al., "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imaging* **36**(4), 994–1004 (2017).
64. Lung Segmentation dataset, https://www.kaggle.com/kmader/finding-lungs-in-ct-data/data (05 December 2017).
65. R. C. Hsu et al., "Contour extraction in medical images using initial boundary pixel selection and segmental contour following," *Multidimension. Syst. Signal Process.* **23**(4), 469–498 (2012).

Md Zahangir Alom is a research engineer at the University of Dayton, Ohio, USA. He has received his BS and MS degrees in computer engineering from the University of Rajshahi, Bangladesh, and Chonbuk National University, South Korea, in 2008 and 2012, respectively. He received his PhD in electrical and computer engineering from the University of Dayton in 2018. His research interests include machine learning, deep learning, medical imaging, and computational pathology. He is a student member of IEEE, member of the International Neural Network Society (INNS), and member of the Digital Pathology Association, USA.

Chris Yakopcic is on the research faculty at the University of Dayton. He has received his BS, MS, and PhD degrees in electrical engineering from the University of Dayton in 2009, 2011, and 2014, respectively. His research includes memristor modeling and circuit design and implementing neural and AI algorithms on low-power hardware. In 2013, he received the IEEE/INNS International Joint Conference on Neural Networks best paper award for a paper on memristor device modeling.

Mahmudul Hasan is a lead researcher in media analytics and content discovery in Comcast Applied AI Research Lab in Washington DC. He has graduated from the University of California, Riverside, with a PhD in computer science in August 2016. Previously, he had received his bachelor's and master's degree in computer science and engineering from Bangladesh University of Engineering and Technology in 2009 and 2011, respectively. His research interest lies in computer vision, machine learning, deep learning, and natural language processing. He has served as a reviewer for many international journals and conferences.

Tarek M. Taha has received his BS degree from DePauw University, Greencastle, Indiana, in 1996 and his BSEE, MSEE, and PhD degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1996, 1998, and 2002, respectively. He is a professor of electrical and computer engineering at the University of Dayton. His research interests include cognitive computing architectures, high performance computing, and architectural performance modeling. He had received the NSF CAREER Award in 2007.

Vijayan K. Asari is a professor in electrical and computer engineering and endowed chair in wide area surveillance at the University of Dayton. He is also the director of the UD Vision Lab. He has received his MTech and PhD degrees in electrical engineering from the Indian Institute of Technology, Madras. He has received several teaching, research, advising, and technical leadership awards.