Published in final edited form as:

IEEE Trans Med Imaging. 2019 January ; 38(1): 134–144. doi:10.1109/TMI.2018.2857800.

Multiple Resolution Residually Connected Feature Streams For Automatic Lung Tumor Segmentation From CT Images

Jue Jiang[1], **Yu-chi Hu**[1], **Chia-Ju Liu**[2], **Darragh Halpenny**[3], **Matthew D. Hellmann**[4], **Joseph O. Deasy**[1], **Gig Mageras**[1], and **Harini Veeraraghavan**[1]

[1]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, USA.

[2]Department of Nuclear Medicine, National Taiwan University Hospital Yunlin Branch, Taiwan.

[3]Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, USA.

[4]Department of Medical Oncology, Memorial Sloan Kettering Cancer Center, New York, USA.

Abstract

Volumetric lung tumor segmentation and accurate longitudinal tracking of tumor volume changes from computed tomography (CT) images are essential for monitoring tumor response to therapy. Hence, we developed two multiple resolution residually connected network (MRRN) formulations called incremental-MRRN and dense-MRRN. Our networks simultaneously combine features across multiple image resolution and feature levels through residual connections to detect and segment lung tumors. We evaluated our method on a total of 1210 non-small cell (NSCLC) lung tumors and nodules from three datasets consisting of 377 tumors from the open-source Cancer Imaging Archive (TCIA), 304 advanced stage NSCLC treated with anti-PD-1 checkpoint immunotherapy from internal institution MSKCC dataset, and 529 lung nodules from the Lung Image Database Consortium (LIDC). The algorithm was trained using the 377 tumors from the TCIA dataset and validated on the MSKCC and tested on LIDC datasets. The segmentation accuracy compared to expert delineations was evaluated by computing the Dice Similarity Coefficient (DSC), Hausdorff distances, sensitivity and precision metrics. Our best performing incremental-MRRN method produced the highest DSC of 0.74±0.13 for TCIA, 0.75±0.12 for MSKCC and 0.68±0.23 for the LIDC datasets. There was no significant difference in the estimations of volumetric tumor changes computed using the incremental-MRRN method compared with expert segmentation. In summary, we have developed a multi-scale CNN approach for volumetrically segmenting lung tumors which enables accurate, automated identification of and serial measurement of tumor volumes in the lung.

Keywords

Deep learning; Segmentation; Longitudinal; Lung cancer; Detection

Correspondence to: Harini Veeraraghavan.

I. Introduction

Lung cancer is the most common cause of cancer death worldwide[1]. In patients with lung cancer treated with systemic therapy, the relative benefits of treatment are routinely determined by measurement of changes in size of tumor lesions, usually using uni-dimensional measurement, such as RECIST v1.1[2]. The applications of automatic tumor segmentation are broad, including measuring treatment response, planning of radiation treatment, and to facilitate extraction of robust features for high-throughput radiomics. Manual delineation of tumor volumes is extremely laborious and prior studies [3] have shown that semi-automatic computer-generated segmentations are more repeatable than manual delineations especially for radiomics analysis [4]. Representative semi-automated tumor segmentation approaches applied to lung cancers include single-click ensemble methods[5] and marker controlled watershed method[6]. However, such methods when applied to lung nodule segmentation [7] produce varying results [8]. Interactive methods [9, 10] that adapt their segmentation to user inputs suffer from inter-rater variability[11].

Reproducible segmentation is essential for longitudinal monitoring of tumor response to therapy. In prior studies, we showed that learning even on a tumor-by-tumor basis can lead to more reproducible tumor segmentations for multiple cancers [11, 12]. Fully automatic convolutional neural network (CNN)[13–15] based approaches such as AlexNet [15], VGG [16], GoogleNet [17] have shown remarkable success in a variety of computer vision and medical image analysis tasks (UNet[18], Vnet[19]). Residual networks (ResNet) [20] achieve fast and stable training irrespective of the network depth [21] and are robust to layer removal at training and inference time [22] due to learning through iterative feature refinement [23].

However, the residual connections alone used in ResNet do not eliminate the issue of poor localization and blurring resulting from successive pooling operations which is undesiderable for segmentation. Therefore, the full resolution residual neural network (FRRN) [24] extended ResNet by passing features at full image resolution to each layer. By concatenating features with lower resolution features, FRRN has demonstrated better segmentation performance compared with six different CNNs when using street images. Our work extends the FRRN by residually combining features computed at multiple image resolutions, whereby, a dense feature representation is computed by simultaneously combining feature maps at multiple image resolutions and feature levels. Such a dense feature representation increases the network capacity and ultimately enables the network to recover the input image spatial resolution better than the existing methods. Our contribution consists of two different multiple resolution residual network (MRRN) called the incremental and dense MRRN. Feature map input in each residual stream is produced through pooling (for dense MRRN) and followed by convolutions with residual connections (for the incremental MRRN). Additionally, the feature maps in each residual stream are refined as they are combined with subsequent layers.

The paper is organized as follows: background and motivation are in Section II, followed by the proposed method in Section III; our experiment design is in Section IV; the results and

discussion are in Section V and Section VI, respectively. Finally, we present conclusions based on our findings.

II. Background and Motivation

Fig.1 depicts the schematic of the various network used in this is work namely, the UNet, the FRRN, and the proposed incremental and dense MRRN methods. UNet (Fig.1 a) employs sequential long skip connections to concatenate features computed at different feature resolutions in the encoding path with those in the decoding path. Long skip connection restricts feature map combination to those produced at the same image resolution and leads to under-utilization of the rich feature representation in the image. ResNet [20] and highway networks [25] reduce feature under-utilization by connecting output from the previous layer to the input of a current layer through an element-wise summation that enables local feature combination with immediate higher resolution. The FRRN [24] (depicted in Fig.1b) extends ResNet by maintaining features maps computed at full image resolution that are combined with lower resolution feature maps in addition to residual concatenation with features of immediately higher resolution.

We propose two multiple resolution residual network (MRRN) for extracting dense feature representations using incremental MRRN (Fig.1 c) and dense MRRN (Fig.1d) that pass feature maps of different levels at varying image resolutions. Supplementary Table I depicts the comparison between the multiple networks.

III. Method

We first define the various components used in our networks. Then, we describe those components in detail.

- *Residual streams* carry feature maps computed at a particular image resolution and provide the residual inputs to residual connection units (RCU). For example, zero residual stream, shown by horizontal black arrow in Fig.1, carries feature maps at full image resolution (R). The N^{th} residual stream ($N > 0$) as used in MRRN is generated by N pooling operations and carries feature maps at $R/2^N$ resolution.
- *CNN Block* (Fig. 2) consists of 3×3 convolutions, batch normalization (BN) and ReLU activation and is used for feature extraction.
- *Residual Unit* or RU (Fig. 2) is a convolutional layer with residual connection placed at the beginning and at the end of the network similar to the FRRN[24].
- *Residual Connection Unit* (RCU) (Fig. 2) is the work horse for the network where the feature maps are computed through a sequence of convolutions using CNN blocks.
- *Residual Connection Unit Block (RCUB)* (Fig. 2) consists of sequentially connected RCUs.

The details of RCU and RCUB are described in III.A and III.B, respectively.

A. Residual Connection Unit

The residual connection unit (RCU) constitutes the filters used in each layer. The structure of a RCU is shown in Fig. 2 (a). RCU has two inputs, namely, residual stream feature map and regular CNN feature map and two outputs, namely, the regular and residual feature maps. The residual stream input consists of feature maps from one of the preceding higher resolution residual streams. The residual feature map input is downsamped through pooling and then concatenated with the regular CNN feature map computed from the previous layer. A RCU is composed of one or more CNN blocks (Fig. 2). The CNN blocks output the processed regular feature map that is passed as input to the subsequent RCU block or the next CNN layer. The residual output is computed through a 1×1 convolution with upsampling of the regular output and passed back to the corresponding residual input stream. In summary, the RCU residually connects feature maps from residual streams and the feature maps at particular resolutions and jointly learns features between different residual streams to refine features computed at lower resolution [23].

B. RCU Block

The RCU block (RCUB) is composed of sequentially connected RCUs. We distinguish between two different RCU blocks, namely, the incremental RCU block and the dense RCU block (Fig.2 (a)) based on their inputs and the number of RCUs used in those blocks.

C. Incremental RCUB

An incremental RCUB consists of serially connected RCUs. Each RCU successively refines feature maps of increasing feature resolution starting from the nearest lower stream up to the 0^{th} stream. Fig. 2(a) depicts an example of incremental RCU block at 2^{nd} residual stream. The dot-solid arrow (see legend in Fig.2) is an abstraction that represents a bidirectional connection between the residual stream and the current layer. RCU blocks are only placed in streams that can have input from at least one residual stream. Therefore, the 0^{th} residual stream does not contain a RCU block. We used RCU×3 in the 1^{st} residual stream similar to the FRRN. A residual stream produced using N (N>1) pooling operations contains (N+1) RCUs in its RCU block. The first RCU in a RCU block is residually connected to the 0^{th} stream to use the full resolution information. The subsequent RCUs are sequentially connected to the residual streams of increasing spatial resolution starting from the immediately preceding residual stream.

D. Dense RCU Block

Unlike the incremental RCUB, every dense RCUB integrates information residually only from the immediate higher spatial resolution feature maps. The additional RCU blocks in a residual stream further refine the feature maps computed at that stream thereby, resulting in dense residual feature maps. Each RCUB has 2 RCUs to refine features from preceding residual stream with the feature maps at current resolution.

E. Incremental MRRN

Incremental MRRN is composed of multiple incremental RCUB and residual feature streams (Fig. 2(b)). The two ends of all residual streams except the 0^{th} and the last stream

contain two RCUB to encode and decode the information. The 0th residual stream does not contain any RCUB while the last stream contains one RCUB. Feature integration from multiple resolutions obviates the need for additional unpooling following RCUBs as used in conventional encoder-decoder like the Unet. The incremental MRRN at a given layer n with multiple residual feature streams k *(=0,1,… N)*, preceding n is defined as:

$$\begin{aligned} z_n^k &= z_{n-1}^k + H(y_{f(n,k)-1}^{p(n-1,k)}, z_{n-1}^k; W_n^k) \\ y_{f(n,k)}^{p(n,k)} &= G(y_{f(n,k)-1}^{p(n-1,k)}, z_{n-1}^k; W_n^k) \quad k = 0, 1, \ldots N, \end{aligned} \tag{1}$$

where, N is the total number of residual feature streams at n; Z_n^k is the n-th residual connection of k^{th} feature stream; $y_{f(n,k)}^{p(n,k)}$ is the output produced by concatenating Z_n^k with p(n,k)-th residual feature stream after f(n,k) RCU connection (see Supplementary Table I (c)). Note that (1) is identical to the FRRN formulation following the first pooling operation when k=0. The number of residual connections to RCUB increase and decrease with addition of pooling and unpooling operations, respectively.

The backpropagated loss l is computed by taking the derivative with respect to the weights $W_{n\text{-}1}$ at layer n as:

$$\begin{aligned} \frac{\partial l}{\partial W_{n-1}^k} &= \frac{\partial l}{\partial y_{f(n-1,k)}^{p(n-1,k)}} \frac{\partial y_{f(n-1,k)}^{p(n-1,k)}}{\partial W_{n-1}^k} + \frac{\partial l}{\partial z_{n-1}^k} \frac{\partial z_{n-1}^k}{\partial W_{n-1}^k} \\ &= \frac{\partial l}{\partial y_{f(n-1,k)}^{p(n-1,k)}} \frac{\partial y_{f(n-1,k)}^{p(n-1,k)}}{\partial W_{n-1}^k} + \frac{\partial z_n^k}{\partial W_n^k} \frac{\partial l}{\partial z_n^k} \frac{\partial z_n^k}{\partial z_{n-1}^k} \\ &= \frac{\partial l}{\partial y_{f(n-1,k)}^{p(n-1,k)}} \frac{\partial y_{f(n-1,k)}^{p(n-1,k)}}{\partial W_{n-1}^k} + \frac{\partial z_n^k}{\partial W_n^k} \left(\frac{\partial l}{\partial z_n^k} + \frac{\partial l}{\partial z_n^k} \frac{\partial H(y_{f(n,k)-1}^{p(n-1,k)}, z_{n-1}^k; W_n^k)}{\partial z_{n-1}^k} \right) \\ &= \frac{\partial l}{\partial y_{f(n-1,k)}^{p(n-1,k)}} \frac{\partial y_{f(n-1,k)}^{p(n-1,k)}}{\partial W_{n-1}^k} + \frac{\partial z_n^k}{\partial W_n^k} \left(\frac{\partial l}{\partial z_n^k} + \frac{\partial l}{\partial z_n^k} \frac{\partial H\left(y_{f(n-1,k)}^{p(n-1,k)}, z_{n-1}^k; W_n^k\right)}{\partial y_{f(n,k)-1}^{p(n-1,k)}} \right. \\ &\left. \cdot \prod_{i=f(n-1,k)}^{f(n,k)-1} \frac{\partial G(y_i)}{\partial y_i} \times \frac{\partial y_{f(n-1,k)}^{p(n-1,k)}}{\partial z_{n-1}^k} \right), \end{aligned} \tag{2}$$

As seen, the gradient $\frac{\partial l}{\partial z_n^k}$ is independent of depth similar to the FRRN formulation.

However, our formulation adds an addition term $\prod_{i=f(n-1,k)}^{f(n,k)-1} \frac{\partial G(y_i)}{\partial y_i}$ to further constrain weight update and regularize learning.

F. Dense MRRN

Dense MRRN uses down-sampled feature maps computed at image resolution as input to the first RCUB in all streams (Fig. 2(c)) to produce dense residual streams. Similar to incremental MRRN, the 0th residual stream does not contain RCUB. All other residual

streams have as many intermediate RCUBs corresponding to the number of following residual streams placed in between the end RCUBs. The feature maps are computed as:

$$z_n^k = z_{n-1}^k + H(z_{n-1}^k, z_{n'}^{k+1}; W_n^k) \qquad (3)$$
$$z_{n'}^{k+1} = z_{n'-1}^{k+1} + H(z_{n'-1}^{k+1}, z_{n''}^{k+2}; W_{n'}^{k+1}),$$

where W_n^k is the parameter of nth RCU output refined using residual stream k (Supplementary Table I (d)). The derivative of the loss l with respect to W_{n-1}^k is calculated as:

$$\begin{aligned}
\frac{\partial l}{\partial W_{n-1}^k} &= \frac{\partial l}{\partial z_{n-1}^k}\frac{\partial z_{n-1}^k}{\partial W_{n-1}^k} + \frac{\partial l}{\partial z_{n'-1}^k}\frac{\partial z_{n'-1}^k}{\partial W_{n-1}^k} \qquad (4)\\
&= \frac{\partial l}{\partial z_{n'-1}^k}\frac{\partial z_{n'-1}^k}{\partial W_{n-1}^k} + \frac{\partial z_{n-1}^k}{\partial W_{n-1}^k}\frac{\partial l}{\partial z_n^k}\frac{\partial z_n^k}{\partial z_{n-1}^k}\\
&= \frac{\partial l}{\partial z_{n'-1}^k}\frac{\partial z_{n'-1}^k}{\partial W_{n-1}^k} + \frac{\partial z_{n-1}^k}{\partial W_{n-1}^k}\left(\frac{\partial l}{\partial z_n^k} + \frac{\partial l}{\partial z_n^k}\frac{\partial H\left(z_{n-1}^k, z_{n'}^{k+1}; W_n^k\right)}{\partial z_{n-1}^k}\right)\\
&= \frac{\partial l}{\partial z_{n'-1}^k}\frac{\partial z_{n'-1}^k}{\partial W_{n-1}^k} + \frac{\partial z_{n-1}^k}{\partial W_{n-1}^k}\left(\frac{\partial l}{\partial z_n^k} + \frac{\partial l}{\partial z_n^k}\frac{\partial H\left(z_{n-1}^k, z_{n'}^{k+1}; W_n^k\right)}{\partial z_{n-1}^k}\frac{\partial z_{n'}^{k+1}}{\partial z_{n-1}^k}\right)\\
&= \frac{\partial l}{\partial z_{n'-1}^k}\frac{\partial z_{n'-1}^k}{\partial W_{n-1}^k} + \frac{\partial z_{n-1}^k}{\partial W_{n-1}^k}\frac{\partial l}{\partial z_n^k}\left(\frac{\partial l}{\partial z_n^k} + \frac{\partial l}{\partial z_n^k}\frac{\partial H(z_{n-1}^k, z_{n'}^{k+1}; W_n^k)}{\partial z_{n-1}^k}\cdot\frac{\partial z_{n'}^{k+1}}{\partial z_{n'-1}^{k+1}}\frac{\partial z_{n'-1}^{k+1}}{\partial z_{n-1}^k}\right)\\
&= \frac{\partial l}{\partial z_{n'-1}^k}\frac{\partial z_{n'-1}^k}{\partial W_{n-1}^k} + \frac{\partial z_{n-1}^k}{\partial W_{n-1}^k}\frac{\partial l}{\partial z_n^k}\left(\frac{\partial l}{\partial z_n^k} + \frac{\partial l}{\partial z_n^k}\frac{\partial H\left(z_{n-1}^k, z_{n'}^{k+1}; W_n^k\right)}{\partial z_{n-1}^k}\cdot(1\right.\\
&\left. + \frac{\partial H\left(z_{n'-1}^{k+1}, z_{n''}^{k+2}; W_{n'}^{k+1}\right)}{\partial z_{n'-1}^k}\right)\frac{\partial z_{n'-1}^{k+1}}{\partial z_{n-1}^k}\Big),
\end{aligned}$$

where, $(1 + \frac{\partial H\left(z_{n'-1}^{k+1}, z_{n''}^{k+2}; W_{n'}^{k+1}\right)}{\partial z_{n'-1}^k})$ results from the presence of intermediate RCUBs.

G. Implementation details

We trained the networks using the Tensorflow [26] library on Nvidia GTX 1080Ti with 12 GB memory processor. A kernel size of 3×3 with 32 features was used to produce features following the 0^{th} residual stream. The number of features was increased at a rate of 2^{k+5} with each additional pooling ultimately resulting in 28,620,001 parameters for incremental MRRN and 25,542,241 parameters for dense MRRN. We used a batch size of 16 due to GPU memory constraints and employed the ADAM algorithm [27] with the initial learning rate of $1e^{-4}$. We employed early stopping with a maximum epoch number of 100 to prevent overfitting. We tried to reduce the impact of data imbalance during training by using the Dice loss [19], which is defined as:

$$L_{DSC} = 1 - \frac{2\sum_i p_i g_i}{\sum_i p_i + \sum_i p_i}, \quad (5)$$

where $p_i \in [0,1]$ represents the i^{th} output of the last layer and $g_i \in [0,1]$ is the ground truth label. Additionally, we implemented real-time data augmentation during training by using image flipping, shifting, elastic deformation, with Gaussian noise.

The final volumetric segmentation was obtained by combining the segmentations from the individual 2D slices using connected components extraction. No additional postprocessing of the segmentations was performed.

IV. Experimental setup

A. Datasets

We used three datasets for the analysis, namely, the open-source The Cancer Imaging Archive (TCIA) dataset [28] of 377 NSCLC scanned at the MAASTRO clinic, internal institution (MSKCC) dataset of 304 tumors from 50 NSCLC patients treated with PD-1 inhibitor immunotherapy using pembrolizumab, and the Lung Image Database Consortium (LIDC)[29] dataset consisting of 2669 nodules from 1018 patients from seven different institutions.

The tumors in the TCIA and MSKCC were confirmed advanced stage malignant tumors while those from the LIDC were lung nodules (>3mm and confirmed by atleast 1 radiologist) with varying degrees of malignancy. The MSKCC dataset also included longitudinal CT scans imaged before and every 9 week during therapy upto a maximum of 17 weeks after treatment initiation. The training (TCIA) and validation set (MSKCC) consisted of contrast enhanced CT scans while the testing set from the LIDC included both regular dose and low-dose CT images. Tumor contours for the TCIA were verified and edited when necessary by radiologist (L.C.J). All ground truth tumor contours in the MSKCC dataset were confirmed by a chest radiologist (DH) with several years of experience reviewing chest CT images. The LIDC datasets were manually delineated by four radiologists who also assigned a malignancy score between 1–5 with 1 being low and 5 being high malignancy. Out of 2669, 928 were confirmed and delineated by all four radiologists and were at least 3mm in diameter. As used in prior works [30, 31], we only analyzed 529 nodules of the 928 that were assigned an average malignancy score of > 3. This resulted in a total of 1210 analyzed tumors from all the three datasets.

There was a wide variation in the size and distribution of tumors between the datasets. The tumors in the TCIA dataset ranged in size from 1.88cc to 1033cc, MSKCC from 2.96cc to 413.8cc, and the LIDC from 0.031cc to 19.18cc (all calculated according to radioligist delineation). The distribution of tumor sizes discretized by size are shown in Table I. Finally, predominant lesions in the TCIA and MSKCC dataset were located in the mediastinum or attached to chest wall while the tumors in LIDC were the most frequent in the lung parenchyma (Table II).

B. Training and testing experiments

We used the TCIA dataset as the training cohort as it had the largest tumors and contained many difficult to detect tumors such as those attached to the mediastinum and the chest wall. The MSKCC dataset was used as validation set for selecting the best model. The LIDC dataset was used as an independent test set. Additionally, we performed K=5 fold cross-validation on the TCIA dataset with part of the data withheld to evaluate and report accuracy on the data not used for training. Training was performed using patches of size 160×160 or region of interest (ROI) centered around the tumor. All ROIs were resized to the size of 160×160 resulting in a total number of 57793 training image samples followed by data augmentation as described previously.

The absence of any fully connected layer enables evaluation of images of sizes other than the 160×160 used in training. The MSKCC and LIDC datasets consisted of 256×256 and 160×160 images, respectively. Detections from individual slices were then stacked into full volumetric segmentation through connected components extraction.

C. Compared methods

1. Random forest with fully connected conditional random field (RF+fCRF): We implemented a random forest (RF) classifier with fully connected conditional random field (CRF) approach for the lung tumor segmentation as described in [32]. Similar to [33] we used maximum CT intensity, minimum CT intensity, mean minimum CT intensity, mean gradient, std gradient, entropy and Gabor edge features at angles ($\theta=0, \pi/4, \pi/2, 3\pi/4$) at bandwidth of ($\gamma$ = 0.1, 0.2, 0.3, 0.4). The RF classifier was trained with (k=5) fold cross-validation using 50 trees and maximum depth of 30. Voxel-wise RF classifications were refined by a fully connected CRF as in [32].
2. Unet: Unet is a commonly used neural network for medical image segmentation and has been applied on a variety of medical image segmentation tasks including in liver[34], and breast[35]. We did not use the fully connected layer as propsed in [18] to enable application of the network for segmenting images of varying sizes.
3. SegNet: We used the SegNet [36] that combines an encoder and decoder structure. The encoder network consists of 13 convolutional layers as the first 13 convolutional layers in the VGG16 network and a corresponding number of decoder layers.
4. FRRN: FRRN[24] can be considered as a special case of our proposed methods where the FRRN uses only the 0^{th} residual stream (carrying full resolution feature).

D. Evaluation Metrics

We quantitatively evaluated the segmentation accuracy using Dice overlap coefficient (DSC), sensitivity, precision, and Hausdorff distance. The DSC calculates the overlap between the segmentation results and ground truth, and is defined as:

$$DSC = \frac{2TP}{FP + 2TP + FN}, \quad (6)$$

where, TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives. Sensitivity and precision metrics were computed as,

$$Sensitivity = \frac{TP}{TP + FN}, \quad (7)$$

$$\text{Pr}ecision = \frac{TP}{TP + FP}, \quad (8)$$

The Hausdorff distance is defined as:

$$Haus(P, T) = \max\{\sup_{p \in S_P} \inf_{t \in S_T} d(p, t), \sup_{t \in S_T} \inf_{p \in S_P} d(t, p)\}, \quad (9)$$

where, P and T are ground truth and segmented volumes, and p,t are points on P and T, respectively. S_p and S_t correspond to the surface of P and T, respectively. We used Hausdorff Distance (95%) as suggested in [37] to remove undue influence of outliers.

E. Tumor detection rate

We evaluated the tumor detection rate of the analyzed methods by using segmentation overlap threshold determined using the DSC metric. Tumors were considered detected when the overlap between the segmentation and ground truth was at least τ, computed as

$$\tau = \frac{TP}{N}, \quad (10)$$

where, TP is the true positives and N is the total number of ground truth tumor pixels. Tumor detection rate is then

$$Rate = \frac{D_{tumor}}{N_{tumor}}, \quad (11)$$

where, D_{tumor} is the detected tumor number and N_{tumor} is the total tumor numbers.

F. Longitudinal tracking of tumor changes and relation to outcomes

Thirty six patients in the MSKCC dataset were imaged at multiple times (≥ 3) starting from baseline and during treatment (every 9 weeks) with immunotherapy. Only the best

performing segmentation method was used for this analysis. The relative volume with respect to the baseline volume (v_0) was computed as:

$$v_i' = \frac{v_i}{v_0}, \quad (12)$$

where, v_i is the volume at the time $t = i$. Next, we computed the average slope of the relative volumes from baseline to the last available imaged time (t ranged from 3 to 13) for both the expert and the algorithm segmentations. The two trends were compared statistically using paired Student's T-test.

G. Preliminary tests to study the influence of residual feature streams on segmentation performance

We evaluated the impact of features from the residual feature streams using weight sparsity with respect to layer depths for FRRN, incremental MRRN and dense MRRN. We only chose these three methods as these are the only ones that used residual streams. The weight sparsity was defined as:

$$S_i = \frac{N_i(|w| < T)}{N_i}, \quad (13)$$

where $N_i(|w<T|)$ is the number of weights whose absolute value is less than a threshold T and N_i is the total number of weight in layer of depth i. Larger sparsification at a certain layer indicates that fewer features from that layer are used which in turn means that those features have lower impact on the performance of the network compared to another layer with smaller sparsification.

V. Results

A. Impact of training iterations on error

Fig. 3 shows the changes in the error for the training (TCIA) and the validation (MSKCC) data using the various networks. As shown, both incremental and dense MRRN lead to the largest decrease in the DSC loss for the validation dataset. On the other hand, the Unet and the SegNet are among the methods achieveing the largest validation error despite achieving the lowest training error earliest compared to the other networks. These results indicate that the same two methods potentially overfit compared to the proposed MRRN methods. The star (see Fig. 3) corresponds to the lowest validation error achieved during training. The models with lowest validation error of different methods were retained for testing on new datasets.

B. Incremental MRRN detected the largest number of tumors

Fig. 4 shows the detection rate and sensitivity computed by varying the segmentation overlap thresholds ranging from $\tau = 0.1$ to 0.9 with increment of 0.1 when using the FRRN

method to benchmark performance. The average DSC increases and the standard deviation decreases as the threshold for tumors to be considered detected increases. Correspondingly the number of detected tumors or the detection rate decreases. We note that the detection threshold does not influence the false positive rate but only the detected number of tumors or the true positive rate. We chose the threshold of 0.5 as used in prior work [38]. Table III shows the detection rate for the analyzed datasets and methods using a detection threshold of 0.5. As shown, incremental MRRN achieved the highest detection rate while dense MRRN outperformed the previously proposed methods.

We computed segmentation accuracies from tumors that were detected by atleast one of the methods to assess clinical usability of the various methods. This resulted in 354, 299 and 455 tumors detected for TCIA, MSKCC and LIDC for segmentation analysis.

C. Incremental and dense MRRN produced the most accurate segmentations.

Fig. 5 shows segmentation performance using the various methods for the TCIA, the MSKCC and the LIDC datasets for tumors attached to mediastinum in Fig. 5 (i,iii), to the chestwall in Fig. 5 (ii, iv, v) and for tumors enclosed within lung parenchyma in Fig. 5 (vi). The blue contour corresponds to expert delineation, while red contour corresponds to the algorithm segmentation results. The non-CNN based RF+CRF method yielded the least accurate segmentation in all three datasets across all analyzed performance metrics (see Fig. 5, Table IV). As shown in Fig. 5, Unet and SegNet produced worse segmentations including under and over-segmentation compared to the proposed methods. FRRN improved over both Unet and SegNet but still produced over-segmentation as shown in the case in Fig. 5 (ii, iii). On the other hand, both incremental MRRN and Dense MRRN produced close to expert segmentation in the presented cases.

Table IV shows the overall segmentation performance of all analyzed methods and the datasets using multiple metrics. Significant test results comparing the proposed MRRN methods with other methods are also shown. Both incremental and dense MRRNs achieved significantly higher DSC than the RF+CRF and Unet for all analyzed datasets, and significantly higher DSC compared to SegNet using internal and external validation datasets. Furthermore, both incremental and dense MRRNs achieved significantly lower Hasudorff distance (HD95) compared to all of the compared methods in all datasets. The sensitivity and precision metrics using either incremental or dense MRRN were slightly improved than the compared methods.

Tumor size: We evaluated the effect of tumor size (Fig. 6) and the location of the tumor (Fig. 7) on the performance using the various methods using DSC and HD95 metrics. The MRRN methods achieved lower HD95 values independent of the tumor size in all the datasets. Both incremental and dense MRRNs outperformed Unet and SegNet independent of the tumor size using DSC and HD95 metrics. The incremental MRRN produced slightly higher DSC compared with the FRRN in the TCIA and MSKCC datasets independent of tumor size and a markedly higher DSC for very small tumors in the LIDC dataset. Nevertheless, very small tumors are challenging to segment for all methods. The highest DSC accuracy for very small tumors on average for the incremental MRRN was (TCIA:

0.70 ± 0.06, MSKCC: 0.74 ± 0.11, LIDC: 0.68 ± 0.23) compared with the lowest accuracy achieved by the RF+fCRF (TCIA: 0.29 ± 0.20, MSKCC: 0.40 ± 0.22, LIDC: 0.35 ± 0.27).

Tumor location: As shown in Fig. 7, the proposed incremental and dense MRRNs and the original FRRN outperformed RF+CRF, Unet and Segnet on all three datasets using both DSC and HD95 metrics independent of tumor location. The overall DSC accuracy achieved by the best incremental MRRN for most difficult tumors, namely, those attached to the mediastinum was (MSKCC: 0.72±0.15, TCIA: 0.72±0.17, LIDC: 0.79 ± 0.10) when compared with the Unet (TCIA: 0.59± 0.15, MSKCC: 0.43±0.20, 0.58 ± 0.17) method.

D. All analyzed methods produced consistent accuracies when evaluated against multiple radiologists in LIDC

All methods produced consistent segmentation independent of radiologists (Fig.8). The segmentation concordance using the various performance metrics across the four radioligists as the reference was: DSC 0.76±0.12 and HD95: 1.75±1.37 mm. Example segmentations produced by incremental MRRN from 8 randomly chosen cases together with the radiologists segmentations are shown in Fig.9. Two cases in Fig.9 (b) and (d) are missing delineation from radiologist R2.

E. Incremental MRRN produced highly similar tumor volumes as radiologist for longitudinal monitoring of tumor response to immunotherapy

The best segmentation method, namely, incremental MRRN was used in the analysis which resulted in highly similar volumes changes (incremental MRRN 0.13± 0.12 and expert 0.09 ± 0.08) (Fig. 10). There was no significant difference trend in tumor volume changes between expert and algorithmic semgentations ($p = 0.7$). The average trend difference between the incremental MRRN and the expert delineation was 0.05 ± 0.05.

Fig. 10(a) shows two example patients, one with acquired resistance to treatment (no longer responding to immunotherapy) shown in red and the second with durable response (showing long term benefit from treatment) in blue. Solid lines correspond to the incremental MRRN segmentations whereas the dashed lines correspond to the expert delineation. Although the algorithm and the expert segmentation are not in perfect agreement (e.g. the algorithm results in over-segmentation since the tumor is difficult to differentiate from adjacent mediastinal pleura at the later time points in the patient with acquired resistance), the trends in volumetric changes are in agreement. The images and the segmentations are visualized for the same two patients in Fig. 10 (b) for acquired resistance and in Fig. 10 (c) for durable response for segmentations computed from baseline and the first and second on-treatment scans separated by 9 weeks each.

F. Deep features from incremental MRRN are preferentially used more than deep features in the FRRN for segmentation.

Using a weight sparsity threshold of T=0.01, incremental MRRN shows a consistently decreasing feature sparsity as network depth increases in Fig. 11. FRRN on the other hand shows no such trend with all layers getting more or less similar sparsity across network depth.

VI. Discussion

We developed, implemented and tested two different multiple resolution residual deep neural network that simultaneously combine features at different resolutions for segmenting lung tumors. We evaluated our approach on segmenting the largest number of lung tumors from three different datasets consisting of an internal, and two external datasets from multiple institutions. Our approach showed significantly better performance compared with some of the existing deep neural network formulations typically used for medical image analysis and segmentation. We also benchmarked the performance of the CNN-based methods with a shallow learning-based RF+CRF method. Our results confirm prior studies that showed the improved performance of deep CNN methods over shallow learning methods. Our methods also achieved the best segmentation performance independent of tumor size and location. Furthermore, our best performing network, namely, the incremental-MRRN showed good concordance with expert delineation in longitudinally tracking tumor volumes in patients with advanced stage cancers and treated with immunotherapy. It is notable that tumors treated with immunotherapy undergo changes both in volume and appearance on CT images. Our approach shows that detection and tracking of such tumors is feasible. To our knowledge, this is the first study that has employed deep learning-based auto-segmentation for analyzing continuously changing tumors treated with immunotherapy.

Our results confirm the previous perspective that combining features from all levels[38] is useful for segmentation. Specifically, our method significantly outperformed traditional approaches like Unet that incrementally concatenate features of different resolution as the images and features are passed from layer to layer. Earlier works that employed multi-resolution features include hypercolumn approach [38], RefineNet [39], and FRRN [24]. Our approach improves on the FRRN method and is somewhat similar to the RefineNet method where features from multiple resolutions are merged through a smooth incremental update such as in the incremental MRRN. However, both incremental and dense MRRNs add connections that enable passing features from all resolutions. Our approach does not add more computational effort as in the hypercolumn method and can be applied to generate segmentations from reasonably sized images; hypercolumn method was restricted to 50×50 images. We used 160×160 images throughout the analysis. However, even larger sized images can be used.

Although our approaches demonstrated improved performance compared to existing methods, there is still room for improvement including for difficult to detect tumors such as those attached to the mediastinum. Another limitation is that we used slice-wise segmentation instead of 3D convolutions and employed a ROI-based training framework wherein, multiple ROIs containing different extent of tumor were generated from the same image to increase the training size. Nevertheless, this is one of the first studies that have developed and tested a deep learning network on large number of lung cancers using multi-institutional datasets and applied the method for longitudinal segmentation and tracking of tumors that change in size and appearance due to treatment with immunotherapy. In summary, we presented a multiple resolution residual network-based deep convolutional neural network approach for generating automatic segmentation of lung tumors.

VII. Conclusion

In this paper, we proposed two neural networks to segment lung tumors from CT images by adding multiple residual streams of varying resolutions. Our results clearly demonstrate the improvement in segmentation accuracy across multiple datasets. Our approach is applicable to longitudinal tracking of tumor volumes for cancers subjected to treatment with immunotherapy, which alters both the size and appearance of tumors on CT. Given the success for lung tumors, the method is promising for other sites as well. Both of our proposed MRRN outperform existing methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by the Breast Cancer Research Foundation (BRCF), and partially by the MSK Cancer Center Support Grant/Core Grant (P30 CA008748).

VIII. REFERENCE

[1]. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, et al., “Cancer statistics, 2008,” CA: a cancer journal for clinicians, vol. 58, pp. 71–96, 2008. [PubMed: 18287387]

[2]. Eisenhauer E, Therasse P, Bogaerts J, Schwartz L, Sargent D, Ford R, et al., “New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1),” European journal of cancer, vol. 45, pp. 228–247, 2009. [PubMed: 19097774]

[3]. Velazquez ER, Parmar C, Jermoumi M, Mak RH, Van Baardwijk A, Fennessy FM, et al., “Volumetric CT-based segmentation of NSCLC using 3D-Slicer,” Scientific reports, vol. 3, pp. 3529, 2013. [PubMed: 24346241]

[4]. Parmar C, Velazquez ER, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al., “Robust radiomics feature quantification using semiautomatic volumetric segmentation,” PloS one, vol. 9, pp. e102107, 2014. [PubMed: 25025374]

[5]. Gu Y, Kumar V, Hall LO, Goldgof DB, Li C-Y, Korn R, et al., “Automated delineation of lung tumors from CT images using a single click ensemble segmentation approach,” Pattern Recognition, vol. 46, pp. 692–702, 2013. [PubMed: 23459617]

[6]. Tan Y, Schwartz LH, and Zhao B, “Segmentation of lung lesions on CT scans using watershed, active contours, and Markov random field,” Medical physics, pp.043502, vol. 40, 2013. [PubMed: 23556926]

[7]. Kalpathy-Cramer J, Zhao B, Goldgof D, Gu Y, Wang X, Yang H, et al., “A comparison of lung nodule segmentation algorithms: methods and results from a multi-institutional study,” Journal of digital imaging, vol. 29, pp. 476–487, 2016. [PubMed: 26847203]

[8]. Balagurunathan Y, Beers A, Kalpathy-Cramer J, McNitt-Gray M, Hadjiiski L, Zhao B, et al., “Semi‐automated pulmonary nodule interval segmentation using the NLST data,” Medical physics, vol. 45, pp. 1093–1107, 2018. [PubMed: 29363773]

[9]. Grove O, Berglund AE, Schabath MB, Aerts HJ, Dekker A, Wang H, et al., “Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma,” PloS one, vol. 10, p. e0118261, 2015.

[10]. Egger J, Kapur T, Fedorov A, Pieper S, Miller JV, Veeraraghavan H, et al., “GBM volumetry using the 3D Slicer medical image computing platform,” Scientific reports, vol. 3, 2013.

[11]. Veeraraghavan H, Dashevsky BZ, Onishi N, Sadinski M, Morris E, Deasy JO, et al., “Appearance Constrained Semi-Automatic Segmentation from DCE-MRI is Reproducible and Feasible for

Breast Cancer Radiomics: A Feasibility Study," Scientific reports, vol. 8, p. 4838, 2018. [PubMed: 29556054]

[12]. Veeraraghavan H and Miller JV, "Active learning guided interactions for consistent image segmentation with reduced user interactions," in Biomedical Imaging: From Nano to Macro, IEEE International Symposium on, pp. 1645–1648, 2011.

[13]. LeCun Y, Bottou L, Bengio Y, and Haffner P, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, pp. 2278–2324, 1998.

[14]. LeCun Y, Bengio Y, and Hinton G, "Deep learning," Nature, vol. 521, pp. 436–444, 2015. [PubMed: 26017442]

[15]. Krizhevsky A, Sutskever I, and Hinton GE, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, pp. 1097–1105, 2012.

[16]. Simonyan K and Zisserman A, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[17]. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al., "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9, 2015.

[18]. Ronneberger O, Fischer P, and Brox T, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241, 2015.

[19]. Milletari F, Navab N, and Ahmadi S-A, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in 3D Vision (3DV), Fourth International Conference on, pp. 565–571, 2016.

[20]. He K, Zhang X, Ren S, and Sun J, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

[21]. He K, Zhang X, Ren S, and Sun J, "Identity mappings in deep residual networks," in European Conference on Computer Vision, pp. 630–645, 2016.

[22]. Veit A, Wilber M, and Belongie S, "Residual networks are exponential ensembles of relatively shallow networks," arXiv preprint arXiv:1605.06431, vol. 1, 2016.

[23]. Greff K, Srivastava RK, and Schmidhuber J, "Highway and residual networks learn unrolled iterative estimation," arXiv preprint arXiv:1612.07771, 2016.

[24]. Pohlen T, Hermans A, Mathias M, and Leibe B, "Full-Resolution Residual Networks for Semantic Segmentation in Street Scenes," arXiv preprint arXiv:1611.08323, 2016.

[25]. Srivastava RK, Greff K, and Schmidhuber J, "Highway networks," arXiv preprint arXiv: 1505.00387, 2015.

[26]. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," arXiv preprint arXiv:1603.04467, 2016.

[27]. Kingma D and Ba J, "Adam: A method for stochastic optimization," arXiv preprint arXiv: 1412.6980, 2014.

[28]. Aerts H, Rios Velazquez E, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al., "Data from NSCLC-radiomics," Cancer Imaging Archive, 2015.

[29]. Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al., "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," Medical physics, vol. 38, pp. 915–931, 2011. [PubMed: 21452728]

[30]. Shen W, Zhou M, Yang F, Yang C, and Tian J, "Multi-scale convolutional neural networks for lung nodule classification," in International Conference on Information Processing in Medical Imaging, pp. 588–599, 2015.

[31]. Han F, Zhang G, Wang H, Song B, Lu H, Zhao D, et al., "A texture feature analysis for diagnosis of pulmonary nodules using LIDC-IDRI database," in Medical Imaging Physics and Engineering (ICMIPE), IEEE International Conference on, pp. 14–18, 2013.

[32]. Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," Medical image analysis, vol. 36, pp. 61–78, 2017. [PubMed: 27865153]

[33]. Hardie RC, Rogers SK, Wilson T, and Rogers A, “Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs,” Medical Image Analysis, vol. 12, pp. 240–258, 2008. [PubMed: 18178123]

[34]. Christ PF, Elshaer MEA, Ettlinger F, Tatavarty S, Bickel M, Bilic P, et al., “Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields,” in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 415–423, 2016.

[35]. Dalmış MU, Gubern-Mérida A, Vreemann S, Karssemeijer N, Mann R, and Platel B, “A computer-aided diagnosis system for breast DCE-MRI at high spatiotemporal resolution,” Medical physics, vol. 43, pp. 84–94, 2016. [PubMed: 26745902]

[36]. Badrinarayanan V, Kendall A, and Cipolla R, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” arXiv preprint arXiv:1511.00561, 2015.

[37]. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al., “The multimodal brain tumor image segmentation benchmark (BRATS),” IEEE transactions on medical imaging, vol. 34, pp. 1993–2024, 2015. [PubMed: 25494501]

[38]. Hariharan B, Arbeláez P, Girshick R, and Malik J, “Hypercolumns for object segmentation and fine-grained localization,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 447–456, 2015.

[39]. Lin G, Milan A, Shen C, and Reid I, “Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation,” arXiv preprint arXiv:1611.06612, 2016.

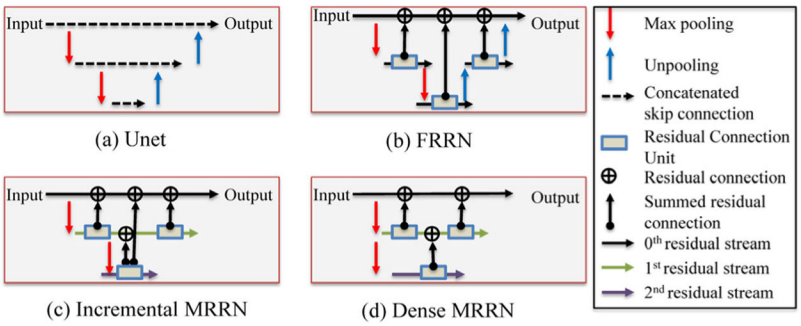

Fig. 1.

Schematic structure of Unet, FRRN and proposed incremental MRRN and dense MRRN.

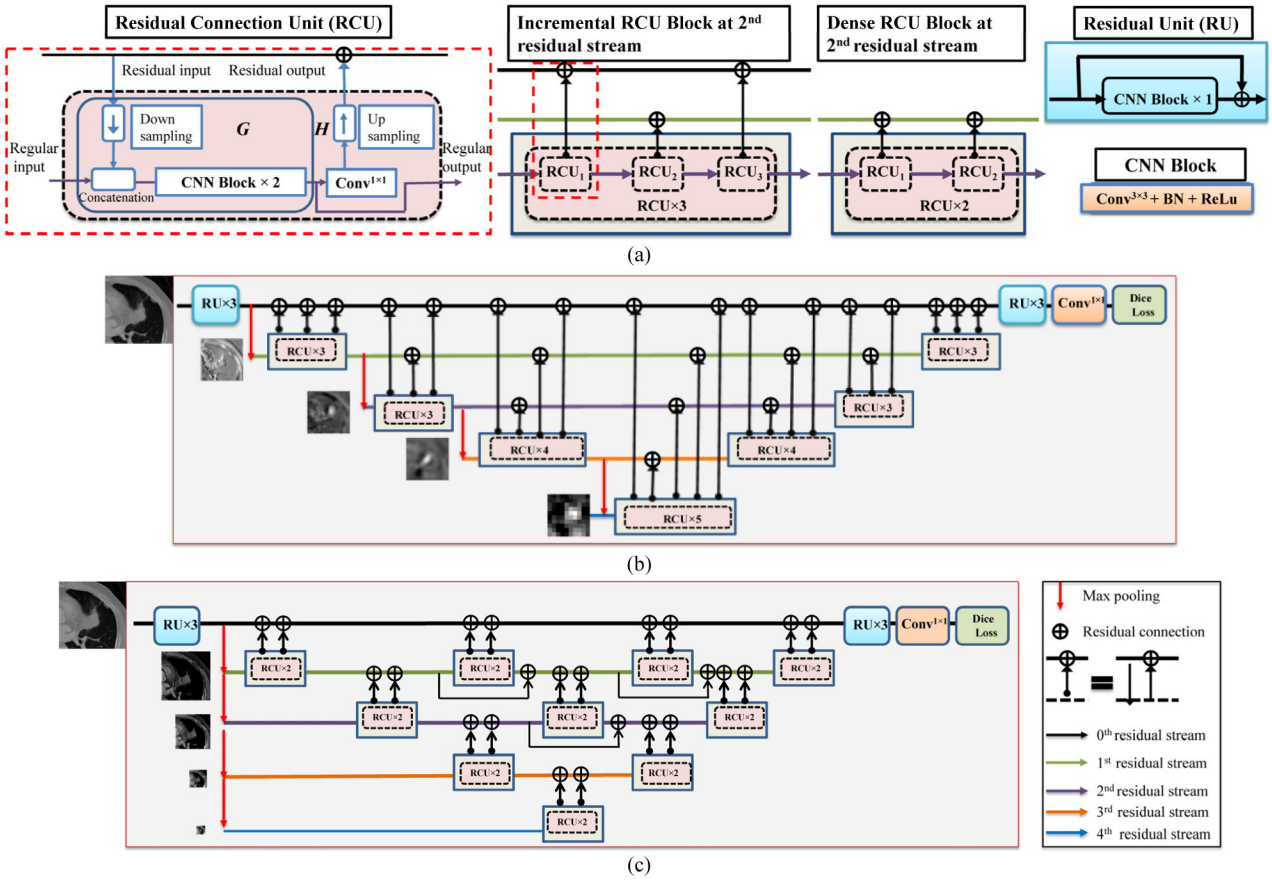

Fig. 2.

Incremental and dense MRRN showing (a) the configuration of each component; (b) incremental MRRN; (c) dense MRRN

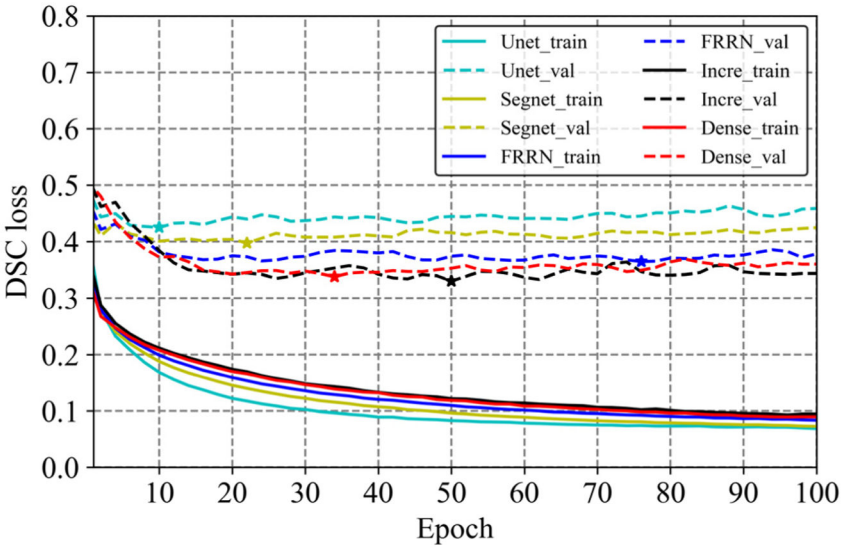

Fig. 3.

Changes in training and validation errors for the analyzed networks with training time.

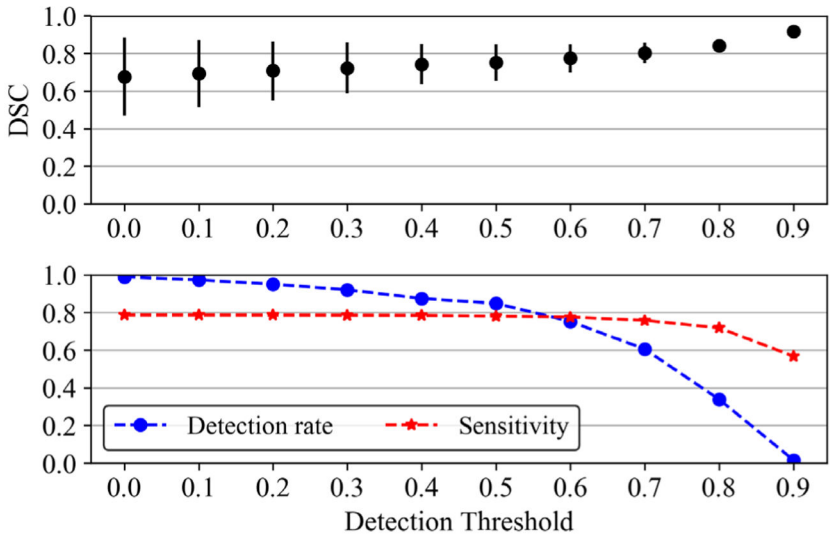

Fig. 4.

Tumor detection rates with varying detection thresholds for FRRN.

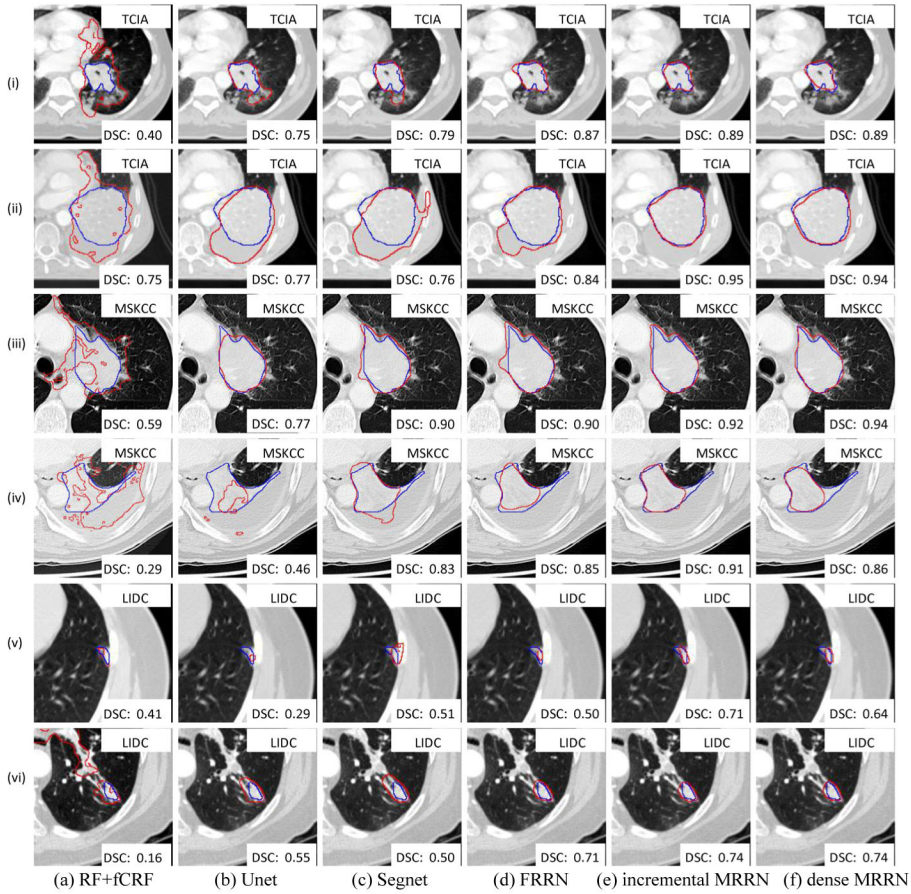

Fig. 5.

Example segmentation results of different methods from the three datasets. The blue contours correspond to expert delineation and red is the algorithm generated segmentation.

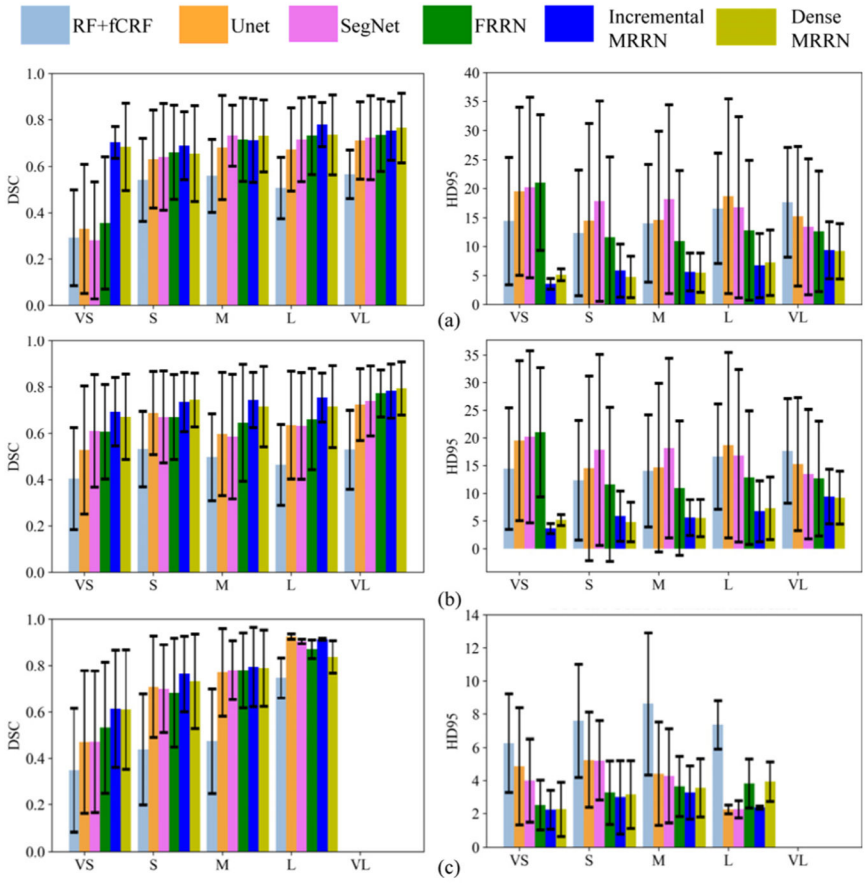

Fig. 6.

Segmentation accuracy of analyzed methods using (a) TCIA, (b) MSKCC, and (c) LIDC with varying tumor sizes is shown for DSC and HD95 metrics. VS: very small tumor; S: small tumor; M: middle size tumor; L: large tumor; VL: very large tumor.

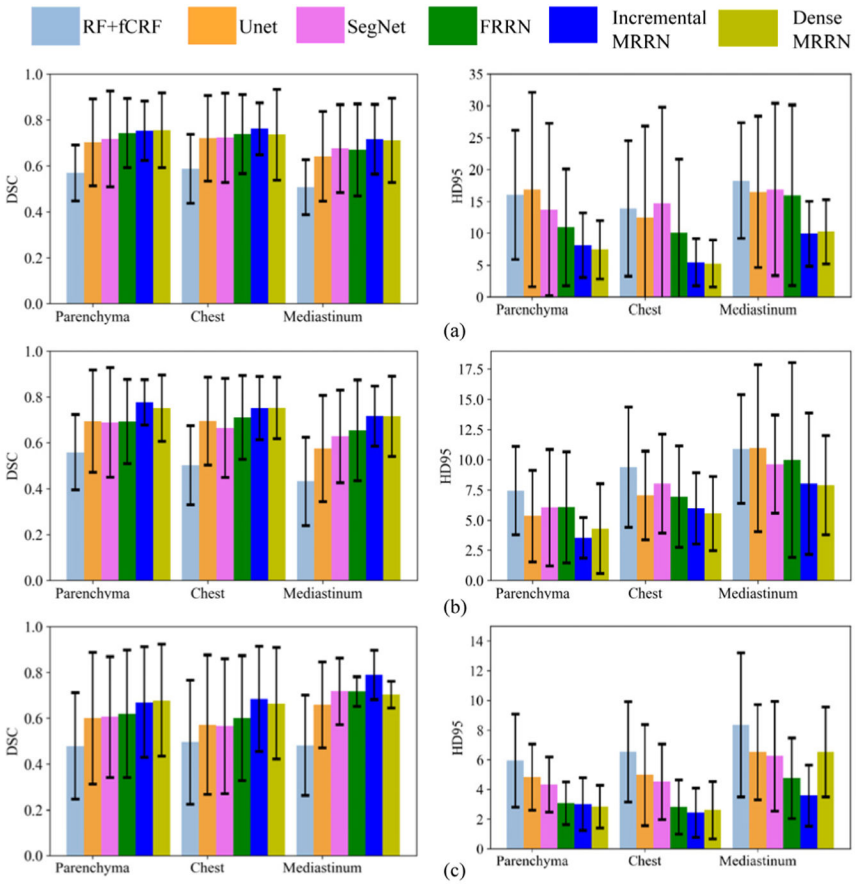

Fig. 7.

Segmentation accuracy using DSC and HD95 metrics of different methods for (a) TCIA, (b) MSKCC and (c) LIDC by tumor locations

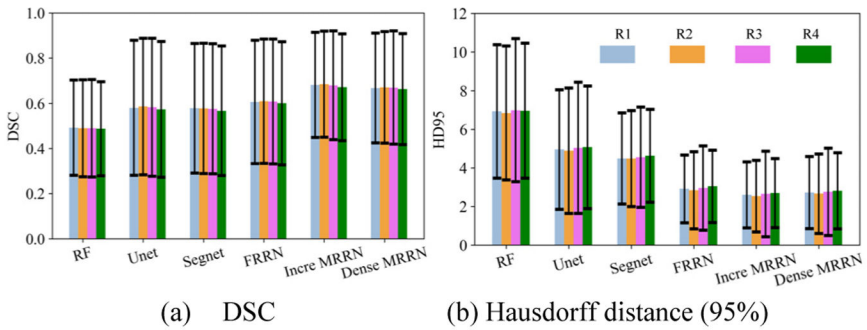

Fig. 8.
Segmentation results of different methods compared to the four radiologists.

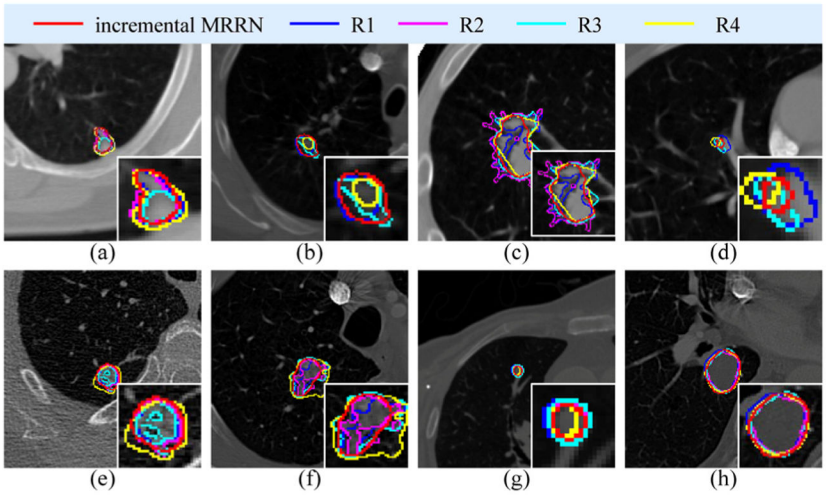

Fig. 9.
Segmentation results of incremental MRRN comparison to the four radiologists

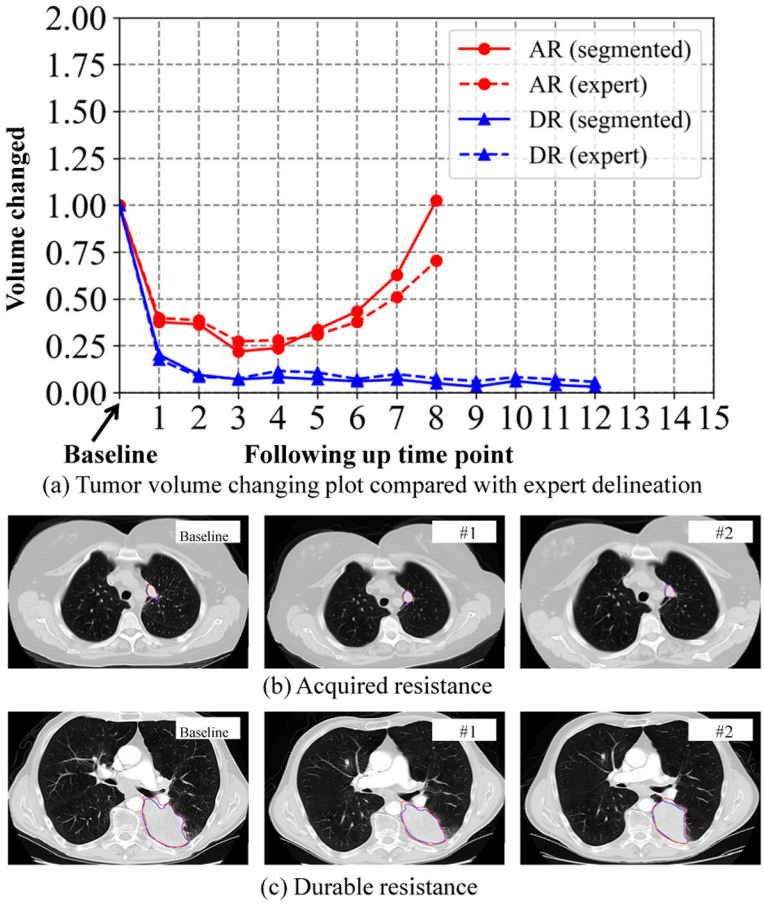

Fig. 10.

Tumor response with respect measured times. The red solid line and red dot line show the acquired resistance (AR) tumor volume changing calculated by incremental MRRN and delineated by expert, respectively. The blue solid line and blue dot line show the durable resistance (DR) tumor volume changing calculated by incremental MRRN and delineated by expert, respectively. (b) and (c) show the segmentation performance at measurement point: baseline, #1 and #2 of the AR and DR in figure 10 (a), respectively. The blue contours show the delineation from expert while the red contour show the segmentation results.

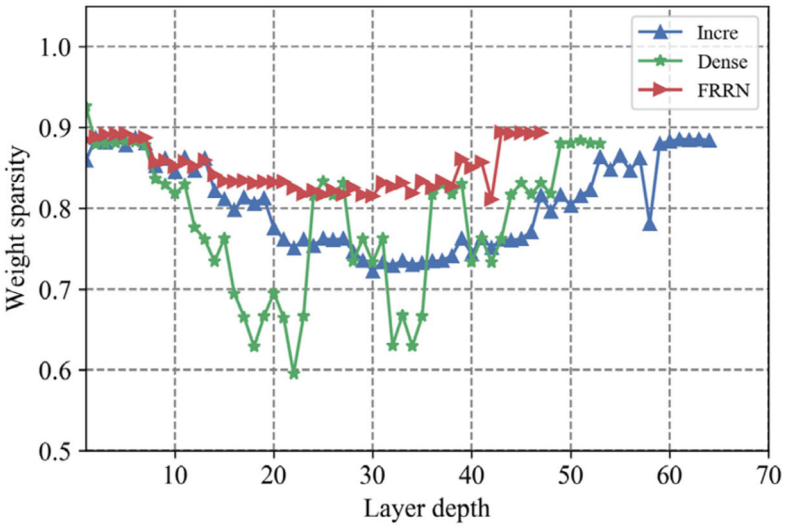

Fig. 11.
Weights sparsity with increasing layer depth.

TABLE I

Tumor sizes in three different datasets (TCIA, MSKCC and LIDC)

Group	Tumor size range(cm^3)	TCIA	MSKCC	LIDC
Very small (VS)	[0,1)	8	38	324
Small (S)	[1,5)	57	75	164
Medium (M)	[5,10)	41	56	38
Large (L)	[10,20)	44	51	3
Very large (VL)	>20	227	84	0

TABLE II

Tumor location in three different datasets (TCIA, MSKCC and LIDC)

Tumor Location	TCIA	MSKCC	LIDC
Within lung parenchyma	103	89	387
Abutting to mediastinum	157	104	6
Attached to chest wall	117	111	136

TABLE III

detection rate under overlap threshold of 0.5 in three different datasets (TCIA, MSKCC and LIDC))

	TCIA	MSKCC	LIDC
Unet	0.80	0.78	0.61
SegNet	0.83	0.80	0.67
FRRN	0.84	0.86	0.64
Incre MRRN	0.89	0.91	0.72
Dense MRRN	0.88	0.93	0.71

TABLE IV

Segmentation accuracies of analyzed methods using multiple metrics, Dice score coefficient (DSC), Hausdorff distance (95%) (HD95), sensitivity and precision, presented as mean and standard deviation (SD). The significant differences between incremental (incre), dense MRRN VS Unet, Segnet, and FRRN are shown as ns ($p \geq 0.05$), Δ($p < 0.05$), □($p < 0.01$) and *($p < 0.001$). The best performing method is indicated using bold font. LIDC dataset is evaluated using the reader 1 delineation.

Data	**Analysis**	**DSC**						**HD95**					
		RF+fCRF	**Unet**	**Segnet**	**FRRN**	**In ere**	**Dense**	**RF+fCRF**	**unet**	**Segnet**	**FRRN**	**Incre**	**Dense**
NSCLC (cross validation)	Mean	0.55	0.68	0.70	0.71	**0.74**$^{*,*,*,\Delta}$	0.73$^{*,*,\Delta,ns}$	16.30	15.51	15.24	12.66	**7.94***,*,*,*	8.10*,*,*,*
	SD	0.13	0.19	0.20	0.18	0.13	0.18	10.29	11.74	14.03	12.23	4.96	5.06
MSKCC (internal validation)	Mean	0.5	0.65	0.66	0.70	**0.75***,*,*,*	0.74*,*,*,*	9.24	7.87	7.92	7.72	**5.85***,*,*,*	5.94*,*,*,*
	SD	0.15	0.22	0.22	0.20	0.12	0.15	4.53	7.92	4.60	6.21	4.40	3.99
LIDC (external validation)	Mean	0.49	0.58	0.57	0.60	**0.68***,*,*,*	0.67*,*,*,*	6.93	4.95	4.48	2.91	**2.60***,***,*	2.72^{*,*,*,ns}
	SD	0.26	0.30	0.29	0.27	0.23	0.23	3.40	3.10	2.36	1.75	1.71	1.87
Data	Analysis	Sensitivity						Precision					
		RF+fCRF	unet	Segnet	FRRN	Incre	Dense	RF+fCRF	unet	Segnet	FRRN	Incre	Dense
NSCLC (cross validation)	Mean	0.42	0.73	0.73	0.75	**0.80***,*,*,*	0.79*,*,*,*	0.59	0.71	0.72	0.73	**0.73**$^{*,ns,\Delta,ns}$	0.73^{*,ns,ns,ns}
	SD	0.15	0.19	0.18	0.19	0.16	0.16	0.23	0.21	0.21	0.19	0.19	2.20
MSKCC (internal validation)	Mean	0.44	0.75	0.72	0.69	**0.82***,*,*,*	0.80*,*,*,*	0.59	0.66	0.69	0.71	**0.72**$^{ns,ns,\Delta,*}$	0.72$^{*,ns,\Delta,*}$
	SD	0.16	0.24	0.27	0.26	0.18	0.20	0.27	0.18	0.18	0.19	0.14	0.15
LIDC (external validation)	Mean	0.66	0.80	0.77	0.76	**0.85***,*,*,*	0.82^{*,ns,*,ns}	0.66	0.64	0.60	0.64	0.67^{ns,ns,*,ns}	**0.70**$^{ns,*,*,\Delta}$
	SD	0.23	0.15	0.20	0.23	0.13	0.15	0.25	0.26	0.26	0.27	0.22	0.21