Retracted: Deep learning for real-time semantic segmentation: Application in ultrasound imaging [Pattern Recognition Letters 144 (2021) 27–34]

Abdeldjalil Ouahabi, Abdelmalik Taleb-Ahmed

Pattern Recognition Letters
journal homepage: www.elsevier.com

Deep learning for real-time semantic segmentation: Application in ultrasound imaging

Abdeldjalil Ouahabi[a,**], Abdelmalik Taleb-Ahmed[b]

[a] *University of Tours, Polytech Tours, Electrical Engineering Department, Imaging and Brain INSERM U 930, Tours 37200, France.*
[b] *IEMN DOAE UMR CNRS 8520, Univ. Polytechnique Hauts-de-France (UPHF), 59313 Valenciennes, France.*

ABSTRACT

A real-time architecture of medical image semantic segmentation called Fully Convolution dense Dilated Network, is proposed to improve the segmentation efficiency while ensuring high accuracy. Considering low resolution and contrast, interferences of shadows, as well as differences in nodules position and size, accurate ultrasound images segmentation cannot be obtained easily. Therefore, a novel layer that integrates the advantages of dense connectivity, dilated convolutions and factorized filters, is proposed in an attempt to remain efficient while retaining remarkable accuracy. Dense connectivity combines low-level fine segmentation with high-level coarse segmentation to extract more features from ultrasound images. Dilated convolution can expand the receptive field of the filter, and the problem of differences in nodules size and position can be solved with different sizes of filters. This study also introduces factorized filters into the network to further optimize the efficiency of the model. In addition, aiming at the class imbalance problem in medical image semantic segmentation, a loss function optimization method is proposed which further improves the accuracy of the network. A thorough set of experiments based on thyroid dataset show that the proposed model achieves state-of-the-art performance in terms of robustness and efficiency.

1. Introduction

It is well known that ultrasound is less expensive, simpler but still effective, readily available, safer and nonionizing than other modalities (**?**) (**?**) making it an ideal front-line tool for diagnostic imaging. Ultrasonic techniques have been considerably developed for medical imaging. Performance reached by modern devices are quite spectacular. Ongoing research in 3D imaging (**?**) (**?**) (**?**), Doppler ultrasound (**?**), ultrasound with contrast media (**?**), sensor miniaturization (**?**) (**?**), characterization tissue (**?**) (**?**) and ultrasound therapy (heating and microbubble delivery of drugs) (**?**), open new avenues for vital medical applications directly accessible through the skin, such as the uterus, liver and bile ducts, kidneys, spleen, breasts, thyroid, etc.

In this study, we are interested in thyroid ultrasound imaging, a clinical technique widely used for the diagnosis of nodules. However, it remains difficult to detect and recognize the nodules due to low contrast and low signal-to-noise ratio. By noise, we mean anything that can hinder the practitioner in establishing an efficient diagnosis, in particular to distinguish between benign and malignant thyroid nodules. According to some studies (**?**), the current practice of monitoring benign thyroid nodules with an evaluation by ultrasound imaging of their growth to diagnose cancer is not conclusive. Efficient alternative strategies are then necessary after an initially benign USFNA (ultrasound-guided fine needle aspiration biopsies). In order to solve these problems, a computer-aided diagnosis (CAD)-based method has been developed to accurately classify thyroid nodules. Efficient segmentation is an essential prerequisite for the classification of thyroid nodules. However, due to the low resolution and contrast of the ultrasound images, as well as large shadow, automatic segmentation of ultrasound images is a challenging task. In addition, the size and position of nodules in the ultrasound images are different from one another which greatly affect the accuracy of nodule segmentation. The automatic segmentation methods, based on active contour and level set (**?**)

**Corresponding author: Tel.: +33 603 894 463;
e-mail: ouahabi@univ-tours.fr (Abdeldjalil Ouahabi)

(**?**), have been used for segmenting thyroid nodules in ultrasound images, but these methods need to extract Regions of Interest (ROI) first, which has a great impact on the segmentation results. In addition, these methods are time-consuming and can not meet the requirements of real-time. Machine learning approaches have achieved good results in the segmentation or thyroid nodules in ultrasound images. However, these conventional methods require a complex process of extracting handcrafted features from images.

In this paper, from the point of view of an efficient network structure, a high-accuracy real-time semantic segmentation network is designed to efficiently and precisely segment nodules from thyroid images. The proposed network is called: Fully Convolutional dense Dilated Net (FCdDN). In order to retain more detailed information in low resolution and contrast images, this network uses a structure like U-Net (**?**) to input the feature map generated in the process from down-sampling phase to the up-sampling phase by using skip connections. The proposed network integrates the advantages of dense connectivity, dilated convolution and factorized filters to design a novel layer as the basic structure of the network. The framework of the network is shown in Fig. 1. To further improve our network, we propose an optimization method based on cross- entropy loss function.

The rest of our study is structured as follows. Section 2 gives a brief related works, and Section 3 provides a detailed description of the proposed architecture. Experimental evaluation as well as the comparison with the existing works are described in Section 4. Finally, the study is concluded in Section 5.

2. Related works

Computer vision tasks have achieved remarkable success thanks to deep learning (**?**), especially the development of convolutional neural networks (CNNs) in medical image segmentation (**?**) (**?**). Segmentation of thyroid nodules from ultrasound images is one of them (**?**), which has achieved satisfactory results. The models under the guidance of deep learning should not only be high in accuracy but also high in real-time performance. To balance high accuracy and computing resources, most existing methods focus on network pruning (**?**), low-bit quantization (**?**) and the design of an efficient network structure. However, both network pruning and low-bit quantization have to be processed on the trained models, thus inevitably affecting the accuracy of the models. In contrast, the design of an efficient network structure can reduce the required computational resources without loss of accuracy.

Recently, a new method based on deep learning called multi-output (or multi-prong) convolutional neural network (MPCNN) algorithm with dilated convolutional layers is proposed in segmentation of thyroid nodules from clinical ultrasound B-mode scans (**?**). This promising method was compared to our approach (see Table 4) and can be used for detection, segmentation, size estimation, volume estimation, and generating thyroid maps for thyroid nodules. It should be noted that encoder-decoder architectures have proven to be very effective in semantic segmentation (**?**), **?**, (**?**), (**?**), (**?**) (**?**). Some architectures, e.g. (**?**), use for semantic segmentation an extension of DenseNets (**?**) in a symmetrical way: 2 dense blocks for the encoder and 2 others on the decoder side, and a dense block in a center. Our choice (see Fig. 1) is motivated by the following considerations: A configuration based on a dense dilated alleviates the problem of the leakage gradient, strengthens feature propagation, encourages feature reuse and considerably reduces the number of parameters. Moreover, drawing inspiration from recent works on convolutional networks, we tested several encoder-decoder architectures for semantic segmentation of thyroid nodule images. The best accuracy-computational resources compromise is the one presented in Table 2.

3. Proposed architecture

The aim of this paper is to establish an efficient semantic segmentation model. We propose a novel layer that integrates the advantages of dense connectivity (**?**), dilated convolutions (**?**) and factorized filters (**?**) in an attempt to remain efficient while retain remarkable accuracy. This novel layer is the core of our architecture. In order to further improve the accuracy of the network, a loss function optimization method is also proposed.

3.1. 1D Dilated layer

Dense connectivity can make more effective use of features by enhancing feature delivery. Besides, using dense connectivity allows network to maintain high accuracy with very few parameters. Dense connectivity connects each layer of the network directly to its front layer. Therefore, layer i can directly use feature maps of all previous layers:

$$x_i = H_i([x_0, x_1, \cdots, x_{i-1}]) \tag{1}$$

where $[x_0, x_1, \cdots, x_{i-1}]$ represents the concatenation of feature mapping from layer 0 to layer $i-1$. H_i is composed succesively of BN, ReLU, convolution and dropout: BN stands for batch normalization, and ReLU for rectified linear unit.

The layer i outputs k feature maps, where k, the growth rate parameter, is generally set to a smaller value (e.g. $k = 16$). An example of a standard layer with dense connectivity is shown in Fig. 2(a), in which the size of convolution kernel is 3×3, and the input is represented by c.

Factorized filters are an effective method to reduce the parameters without yielding the performance of the model under the condition that the sizes of the feature images remain the same. We propose to redesign the standard layer by using factorized filters to further improve the network efficiency (Fig. 2(b)). The study in (**?**) shows that any convolution kernel of $h \times w$ can be decomposed into two consecutive 1D convolution kernels of $h \times 1$ and $1 \times w$. In this study, the 1D convolution with $h = w = 3$ is used to redesign the layer in a better way. It can be seen that the number of parameters before and after using factorized filter are 9 and 6 respectively, resulting in one third decrease. Let H_i^1 and H_i^2 be the same function as H_i, except that they transform the 3×3 convolution into the 3×1 and 1×3 convolution respectively. If H_i in equation (1) is represented by

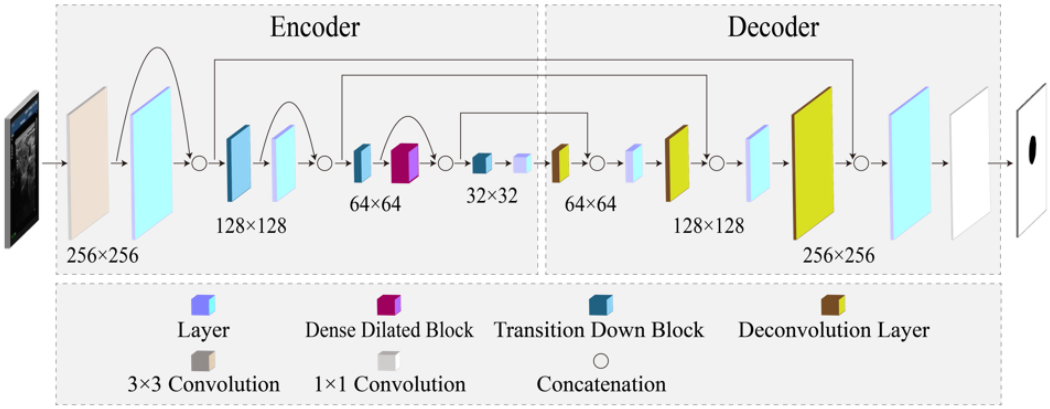

Fig. 1: The framework of Fully Convolutional dense Dilated Net (FCdDN). The volume in the figure corresponds to the layer in Table II. Details of layer, transition down Block and Dilated Block is given in Section 3.

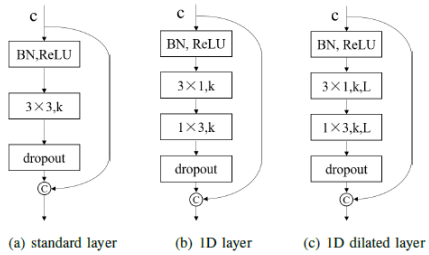

Fig. 2: Dense layer architecture: standard dense layer (a), dense layer of using factorized filters (b), dense layer of using factorized filters and dilated convolution (c). Each layer is composed successively of BN, ReLU, convolution and dropout. c stands for the input of layer, k for the output dimension of layer, L for the dilation rate of dilated convolution in layer and ©for the concatenation.

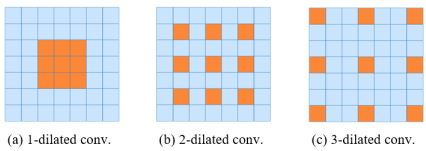

Fig. 3: Dilated convolution: the receptive field size of 1-dilated convolution is 3, the receptive field size of 2-dilated convolution is 5, the receptive field size of 3-dilated convolution is 7.

H_i^1 and H_i^2, then equation (1) can be expressed as follows:

$$x_i = H_i^2(H_i^1(x_0, x_1, \cdots, x_{i-1})) \tag{2}$$

The down-sampling in the semantic segmentation process can give the convolution kernel a larger receptive field, which enables the convolution kernel to obtain more context information. However, down-sampling can result in the loss of spatial information, such as accurate edge shape. Therefore, the aim is to limit the number of down-sampling as much as possible. For the reduced receptive field that changes according to the down-sampling, the dilated convolution is applied to obtain a larger receptive field without losing the spatial information of the image. And the number of parameters of convolution kernels remains unchanged. In dilated convolution, a small size kernel with $k \times k$ filter is enlarged to $k + (k-1)(L-1)$ with dilated rate L. Thus, it allows flexible aggregation of the multi-scale contextual information while keeping the same resolution. Examples can be found in Fig. 3 where standard convolution gets 3×3 receptive field and two dilated convolutions deliver 5×5 and 7×7 receptive fields respectively.

It is interesting to recall the definition of 2-D dilated convolution:

$$y(m,n) = \sum_{i=1}^{M}\sum_{j=1}^{N} x(m + L \times i, n + L \times j)w(i,j) \tag{3}$$

$y(m,n)$ is the output of dilated convolution from input $x(m,n)$ and a filter $w(i,j)$ with the length and the width of M and N respectively. Note that when the dilation rate L is 1, dilated convolutions are the same as standard convolutions.

The idea of dilated filters was developed in the "algorithm trous" for efficient multi-resolution analysis based on wavelets (**?**). Significant improvements in the accuracy of segmentation tasks have been achieved by dilated convolutional layers which makes them an interesting alternative to conventional pooling layers. Although pooling layers (e.g., max pooling) are widely used for maintaining invariance and controlling over fitting, they also dramatically reduce the spatial resolution meaning the spatial information of feature map is lost. Hence the advantage of a compromise between dilated convolution and max pooling. This is where our approach fits in. Dilation is also equivalent to upsample convolutional filters by inserting zeros between weights, as illustrated in Fig. 3. It enlarges the receptive field, but does not require training extra parameters. Dilated convolutions can be used in cascade to build multi-layer networks as illustrated in Fig. 4.

3.2. Dense Dilated Block

Inspired by Dense Block in FC-DenseNet (**?**), we design the dense dilated block by combining dense connectivity, dilated convolution and factorized filters. The structure of the dense dilated block is shown in Fig. 4, where the dilation rate of dilated convolution is $L = 2^N$, N being the order of convolution layers.

Convolutions in the dense dilated block have different dilation rates, which can achieve multi-scale information for fusion.

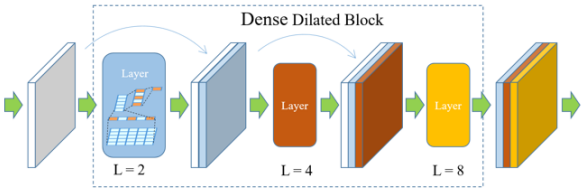

Fig. 4: Dense Dilated Block with three convolutions. L represents the dilation rate of dilated convolution in layer. Layer outputs the feature maps with the same color. The superposition of different color feature maps represents concatenation. The final feature map is the concatenation of feature maps of the three layers.

Such fusion can solve the problem of large differences in size and location of nodules, thus improving the segmentation accuracy of thyroid nodules.

3.3. Network architecture

The proposed network architecture is shown in Fig. 1, which contains operations of each layer from RGB images to pixel classification, the size of each volume representing the size of output. The architecture of the layer is shown in Table 1. There are two kinds of architecture in this table, the asymmetrical architecture on the left and the standard architecture on the right, both composed in turn of BN, ReLU, convolution and dropout. The difference between the left and the right is that the asymmetric architecture used a 3×1 convolution and a 1×3 convolution instead of a 3×3 convolution (see expressions (1) and (2)).

Table 1: Layer architecture

Layer	
Batch Normalization	
ReLU	
3×1 Convolution 1×3 Convolution	3×3 Convolution
Dropout = 0.2	

The down-sampling block consists of BN, ReLU, a convolution of 1×1, dropout = 0.2 and a max pooling layer size of 2×2 in turn. The deconvolution layer is a transpose convolution in the size of 3×3, with stride 2. The description of the architecture is detailed in Table 2, in which the layers 1 to 8 constitute an encoder, and the layers 9 to 15 constitute a decoder. The network comprises 8 convolution layers, one dense dilated block, three transition down blocks, and three deconvolution layers. Table 3 shows the architecture of Transition Down Block. On the first layer is a standard 3×3 convolution with stride 1. On the last layer is a standard 1×1 convolution with stride 1. The other layers all use factorized filters and dilated convolution, the dilation rate being fixed at 2. The sixth layer is dense Dilated Block, containing 4 factorized layers. According to Section 3.2, the dilation rates of convolutions are 2, 4, 8, and 16. In this paper, the growth rate in the network is set as $k = 16$.

3.4. Loss function optimization

In object segmentation, as is the case for the segmentation of thyroid nodules, a common issue is class imbalance, so the choice of the loss function is important to overcome this problem. As we have an unbalanced data, the task is very challenging. In this experiment, we have focused on semantic segmentation of thyroid nodules, therefore the number of classes at pixel level is restricted to 2, and to solve the problem of class imbalance in binary thyroid image segmentation, we optimize the cross-entropy loss function. After the soft-max layer, each pixel will get a probability value, and the cross-entropy loss is used to measure the difference between the predicted result p and the ground truth g as defined in (4):

$$L_{CE}(g,p) = -\frac{1}{mn}\sum_{x=1}^{m}\sum_{y=1}^{n}\sum_{c=1}^{2} g_{x,y}\log p_{x,y} \tag{4}$$

where m and n are the length and width of probability distribution, x and y are the horizontal and vertical coordinates of pixels, c is the real class, $g_{x,y}$ and $p_{x,y}$ are real value and probabilities at pixel (x, y) respectively. Since the cross-entropy loss evaluates the class predictions for each vector pixel individually and then averages all the pixels, this can be a problem if the different classes have an unbalanced representation in the image, as training can be dominated by the most common class. This means that the misclassified pixels are also counted when calculating the loss. Several strategies are then possible to optimize this loss function (**?**) (**?**) (**?**). In order to limit the transfer of these misclassified pixels into the network, we propose a mapping function to remap the probability value of these pixels. Mapping the probability value to a smaller value can get a greater loss. It is therefore necessary to choose this probability adequately, for example by introducing a function of the sigmoid type defined as $f_1 = \frac{1}{1+\exp(\frac{\theta}{2}-p_{x,y}\times\theta)}$, where θ is a variable parameter. This function is justified by the fact that the parameter θ confers a degree of freedom allowing a correct classification. The value interval of f_1 mapping is not [0,1], but [$\frac{1}{1+\exp(\frac{\theta}{2})}$, $\frac{1}{1+\exp(\frac{-\theta}{2})}$] and when θ is larger, the mapping interval is closer to [0,1]. However, when θ is too large, it will also have a bad effect on the correctly classified pixels. Therefore, it is necessary to choose an appropriate value for θ. When we use f_1, the probability values of misclassified pixels can be mapped to smaller values, and the probability values of correctly classified pixels can be mapped to larger values. At this point, the network can transfer the attention to the misclassified pixels to a greater extent, thus improving the performance of the network. However, at the later stage of training, when most of the pixels are classified correctly, the network attention will continue to transfer to the incorrectly classified pixels, which will cause the originally correctly classified pixels to be classified incorrectly. In such situation, the accuracy obtained will not always be improved. To meet our expectation, we choose a simple quadratic function as $f_2 = p_{x,y}^2$. As shown in Fig. 5 , the average value between f_1 and f_2 rated f_3 meets our requirements and expectations. Using our optimization method to remap the probability value, we redefine the loss function as follows:

$$L_{CE}(g,p) = -\frac{1}{mn}\sum_{x=1}^{m}\sum_{y=1}^{n}\sum_{c=1}^{2} g_{x,y}\log\frac{f_1(p_{x,y})+f_2(p_{x,y})}{2} \tag{5}$$

This loss function is used to optimize our network and further improve its segmentation performance. Our optimization

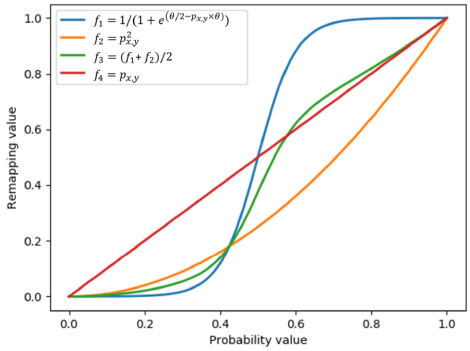

Fig. 5: Probability values mapped to different values through different mapping functions.

method can solve the class imbalance problem because it gives more weight to misclassified pixels, which have no relation to the target size.

4. Experimental evaluation

In this paper, all experiments are conducted with TensorFlow framework. All models are trained with a batch size of 12, and weight decay of 2×10^{-4}. The input image size of all networks is 256×256. In accordance with (**?**) and for fair comparison, we use the optimal parameters for CE-Net where the input image size is 448×448 and Dice loss is used instead of cross entropy loss for all other networks. The initial learning rate is 1×10^{-3} that we divide by a factor of 10 for a certain number of rounds, coming to a total of 100 rounds of training, with the final learning rate being 1×10^{-5}. Adam optimization (**?**) of stochastic gradient descent is used for training. The thyroid dataset used in this experiment includes 3794 ultrasound images. All these ultrasound images come from Piti Salpłtrire Hospital, Sorbonne University, Paris, and were scanned and labeled by the team of doctors in the Thyroid and Endocrine Tumor Unit. To train and test the model, with the help of Ultrasound doctors and radiologists, 3794 ultrasound images were labeled with nodule edge, and the artificial label was segmented, finally developing the dataset of this experiment. The images were randomly divided into training set and test set, of which the training set contains 2530 images and the test set contains 1264 images. All accuracy results are reported using the Intersection-over-Union IoU metric, also known as Jaccard index, true positive fraction TPF and false positive fraction FPF:

$$IoU = \frac{area(A \cap B)}{area(A \cup B)} \tag{6}$$

$$TPF = \frac{area(A \cap B)}{area(A)} \tag{7}$$

$$FPF = \frac{area(B) - area(A \cap B)}{area(C) - area(A)} \tag{8}$$

where A represents the nodule area in ground truth, B represents the nodule area in the predicted results of the model, and C represents ground truth.

4.1. Quantitative results, ablation experiments and analysis

As shown in Table 4, the proposed FCdDN is compared to recent methods (**?**), (**?**), (**?**), (**?**), (**?**), (**?**) and (**?**) in terms of IoU, TPF, FPF, forward pass time and the number of model parameters.

Table 2: Detail architecture of FCdDN. Out-S: output size. Out-F: number of feature maps at layers output.

Layer	Out-S	Type	Out-F
1	256×256	Convolution (3×3)	48
2	256×256	Layer (1D dilated)	16
3	128×128	Transition Down Block	64
4	128×128	Layer (1D dilated)	32
5	64×64	Transition Down Block	96
6	64×64	Dense block(1D dilated)	64
7	32×32	Transition Down Block	160
8	32×32	Layer (1D dilated)	16
9	64×64	Deconvolution layer	176
10	64×64	Layer (1D dilated)	16
11	128×128	Deconvolution layer	112
12	128×128	Layer (1D dilated)	16
13	256×256	Deconvolution layer	80
14	256×256	Layer (1D dilated)	16
15	256×256	Convolution (1×1)	2

Table 3: Transition Down Block Architecture

Transition Down Block
Batch Normalization
ReLU
1×1 Convolution
Dropout = 0.2
2×2 Max Poling

All models are not pre-trained. IoU, TPF, and FPF are obtained by calculating the mean value of all test images. Time means the forward pass time on a single NVIDIA TITAN Xp GPU in milliseconds. Model size was the required space of the model on disk in MB. Parm means the number of parameters. Comprehensive results show that, in light of almost all of the evaluation criteria, the proposed method outperforms the original CE-Net method as well as other state-of-the-art methods for semantic segmentation of thyroid nodules. Indeed, our network can run in real time on a single GPU. It reached 81.70% IoU, 90.50% TPF and 0.25% FPF in the thyroid test dataset. Only 7.8 ms was needed to process each image on a single GPU, making our network one of the fastest networks available. These interesting results are obtained by using a loss function with θ =20, referred to in Table 4 as $Loss_{rmp}$, which consists in optimizing our network by remapping the value of the probabilities (or weights) of the misclassified pixels (**?**), (**?**). In addition, liver tumor segmentation by an improvement of U-Net called Modified U-Net (**?**) shows a dice of 89.72% corresponding to an IoU of81.35%, therefore of the same order as our approach. However, the authors of Modified U-Net use a modality (CT

scan) where the the image is generally of better quality. Compared with state-of-the-art networks, our approach has a similar accuracy, but with a significantly lower number of parameters and a shorter direct transit time. Actually, we should not only consider the number of theoretical parameters but also pay more attention to the actual space required by the model on disk. It requires just 5.9 MB of disk space, enough for designing auxiliary diagnostic tools in mobile terminals and embedded devices.

Since the proposed network integrates the advantages of dense connectivity, dilated convolution and factorized filters to design a basic structure of the network, it is interesting to investigate the effect of each component on the performance of network, and to show the necessity of all three changes. Ablation results are illustrated in Table 6. By analyzing Table 6, we find that the integration of the three components contributes to increase the performance of the proposed network.

In medical routine, the operating environment of the model is probably not as high as the configuration used in our experiment. It would then be interesting to evaluate our model on a simple laptop. Thus, for a configuration with an Intel Core processor from the i5 to i9 family, at 3.5 GHz (up to 4.6 GHz), the test results are presented in Table 5, where the forward pass time (or direct transit time) in seconds (the average test result for 100 images) is the element of comparison of the speed of execution. As shown in Table 5, the direct transit time of FCdDN on CPU was, as expected, much lower than that of high accuracy networks. Our networks forward pass time on standard CPU was only around 0.6 s, the second fastest of all compared networks. In addition, in practical applications, more than 16 images can be input into the network at the same time, which can further improve the actual running speed of the network.

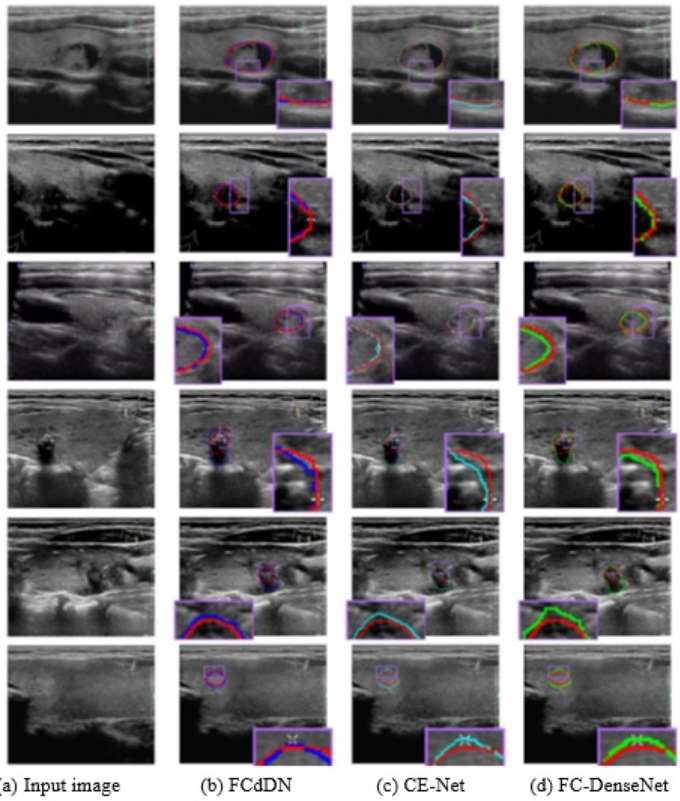

Fig. 6: Qualitative results of the segmentation produced by FCdDN(b), CE-Net(c), FC-DenseNet(d) compared to the ground truth. Ground truth is shown in red, FCdDN predicted results are shown in blue, CE-Net predicted results are show in cyan, FC-DenseNet predicted results are shown in green. A targeted zoom is performed to better visualize the results of the segmentation.

4.2. Qualitative results and analysis

Fig. 6 shows some qualitative segmentation results and compared with ground truth. Curves in different colors were used to mark the original images instead of mask to make a clearer comparison. As can be seen in Fig. 6, the segmentation results of FCdDN are basically similar to those of FC-DenseNet and CE-Net.

The characterization of small targets in ultrasound imaging can pose common and delicate issues. Although the added value of our model lies in its ability to completely separate certain tiny thyroid nodules: In certain medical situations, the effectiveness of our model may be limited in segmentation as is the case with the results of the fourth and fifth lines in Fig. 6 caused by greater calcification of the thyroid nodules. These calcified shadows emerge under the thyroid nodules on ultrasound and the border between these shadows and the thyroid nodules is blurred. The blurry border leads the model to confuse these shadows with the nodules, which makes segmentation less efficient. However when the contrast between the nodule area and the background area is very low, our model can also segment the nodule well, as in Fig. 6 line 6, indicating that the model has learned the essential features of the nodule.

It remains to show that the proposed model is still efficient by segmenting ultrasound images of thyroids from a medical structure and a protocol completely different from those used initially in this study: The digital database of Thyroid Ultrasound Images (`http://cimalab.intec.co/?lang=en&mod=project&id=31`).

This database contains 99 cases and 134 images. Each case is presented as a XML file with the expert's annotation and patient's information. These images have not been pre-processed to remove the noise [(**?**), (**?**)] characterized by black areas, nor to possibly smooth out the gridding artifacts caused by dilated convolutions (**?**). The IoU, TPF and FPF are 79.5%, 88.5% and 0.13% respectively. The qualitative test results are shown in Fig. 7. From the results, we can conclude that our model can segment thyroid ultrasound images of different types without any fine-tunings, which proves that our model is robust.

5. Conclusion and perspectives

Segmentation of thyroid nodules from ultrasound images is a key tool for diagnosis. However, it is a challenging task due to surrounded similar structures such as lymph nodes, low resolution, low contrast and low signal-to-noise ratio of ultrasound images. In this study, we propose a semantic segmentation network able to run on computer-aided diagnosis equipment in real time with high accuracy and efficiency. This new network based on an integration of dense connectivity, dilated convolution and factorized filters, and optimization of the loss function

Table 4: Test results and comparison. Time means the forward pass time. Model size is the required space of the model on disk. Parm means the number of parameters.

Network	IoU%	TPF%	FPF%	Time	Model size	Param
ENet (**?**)	56.90	74.52	1.28	**5.5ms**	10.9MB	0.36M
U-Net (**?**)	78.16	86.82	0.25	14.5ms	415.3MB	34.50M
ERFNet (**?**)	78.19	89.85	0.35	8.9ms	29.4MB	2.13M
SegNet (**?**)	77.30	85.67	0.23	15.10ms	424.0MB	35.40M
DeepLabv (**?**)	79.75	89.97	0.30	18.7ms	488.6MB	40.35M
FC-DenseNets (**?**)	79.07	**90.50**	0.30	48.0ms	148.7MB	11.12M
CE-Net+Dice Loss (**?**)	70.33	90.36	0.45	19.1ms	488.0MB	40.34M
CE-Net+Cross Entropy (**?**)	81.50	89.52	0.22	19.1ms	488.0MB	40.34M
MPCNN (**?**)	78.80	82.13	0.01	99.1ms	510.0MB	41.00M
FCdDN (ours)	80.50	89.60	0.25	7.8ms	**5.9MB**	0.20M
FCdDN+$Loss_{rmp}$ (ours)	**81.70**	**90.50**	0.25	7.8ms	**5.9MB**	0.20M

Table 5: Ablation results

Network	IoU	TPF	FPF	Time (ms)	Model size	Parameters
No dense connectivity	78.06	85.73	0.21	6.73	3.0 MB	29.3 k
No factorized filters	79.77	89.23	0.28	6.58	4.6 MB	187.8 k
No dilated convolution	74.24	87.02	0.43	5.42	3.8 MB	198.8 k
FCdDN	81.50	90.50	0.25	7.8	5.9 MB	198.8 k

Table 6: Test results on standard CPU

Network	Time
ENet (**?**)	0.21s
U-Net (**?**)	1.52s
ERFNet (**?**)	0.83s
DeepLab (**?**)	1.20s
FC-DenseNets (**?**)	6.73s
CE-Net (**?**)	1.17s
FCdDN (ours)	0.63s

to solve the class imbalance problem at the pixel level. The proposed network achieved segmentation accuracy similar to that of state-of-the-art networks on the thyroid dataset, but with a forward pass time on a single NVIDIA TITAN Xp GPU less than half. From the results of tests on a standard CPU, we can see that the model proposed in this study has achieved an excellent compromise between segmentation accuracy and speed, which makes it suitable for routine medical equipment that has need both robustness and efficiency. In ongoing work, our research team is focusing on fine edge segmentation and shaded areas processing in ultrasound images of thyroid nodules. Thus, the combination of reliable computer-aided diagnosis and clinical prediction results in high-precision performance that benefits the patient and even saves time and money, with equal quality of care: this is our ultimate goal. Tomorrow, the object of ultrasound imaging will no longer be simply visual analysis, but prediction and decision by combining the images and information provided by deep learning.

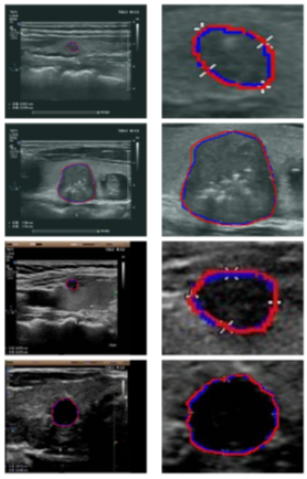

Fig. 7: Qualitative segmentation on database of Thyroid Ultrasound Images. Ground truth is shown in red, FCdDN predicted results are show in blue. First column: Segmentation results on the original image. Second column: Segmentation results on ROI, just for clearer presentation.