Fine-Grained Segmentation Using Hierarchical Dilated Neural Networks

Sihang Zhou[1,2], Dong Nie[2], Ehsan Adeli[3], Yaozong Gao[4], Li Wang[2], Jianping Yin[5], and Dinggang Shen[2(✉)]

[1] College of Computer, National University of Defense Technology, Changsha 410073, Hunan, China
[2] Department of Radiology and BRIC, UNC at Chapel Hill, Chapel Hill, NC, USA
dgshen@med.unc.edu
[3] Stanford University, Stanford, CA 94305, USA
[4] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
[5] Dongguan University of Technology, Dongguan 523808, Guangdong, China

Abstract. Image segmentation is a crucial step in many computer-aided medical image analysis tasks, e.g., automated radiation therapy. However, *low tissue-contrast* and large amounts of *artifacts* in medical images, i.e., CT or MR images, corrupt the true boundaries of the target tissues and adversely influence the precision of boundary localization in segmentation. To precisely locate blurry and missing boundaries, human observers often use high-resolution context information from neighboring regions. To extract such information and achieve fine-grained segmentation (high accuracy on the boundary regions and small-scale targets), we propose a novel hierarchical dilated network. In the hierarchy, to maintain precise location information, we adopt dilated residual convolutional blocks as basic building blocks to reduce the dependency of the network on downsampling for receptive field enlargement and semantic information extraction. Then, by concatenating the intermediate feature maps of the serially-connected dilated residual convolutional blocks, the resultant hierarchical dilated module (HD-module) can encourage more smooth information flow and better utilization of both high-level semantic information and low-level textural information. Finally, we integrate several HD-modules in different resolutions in a parallel connection fashion to finely collect information from multiple (more than 12) scales for the network. The integration is defined by a novel late fusion module proposed in this paper. Experimental results on pelvic organ CT image segmentation demonstrate the superior performance of our proposed algorithm to the state-of-the-art deep learning segmentation algorithms, especially in localizing the organ boundaries.

1 Introduction

Image segmentation is an essential component in computer-aided diagnosis and therapy systems, for example, dose planning for imaging-guided radiation

D. Shen—This work was supported in part by the National Key R&D Program of China 2018YFB1003203 and NIH grant CA206100.

© Springer Nature Switzerland AG 2018
A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11073, pp. 488–496, 2018.
https://doi.org/10.1007/978-3-030-00937-3_56

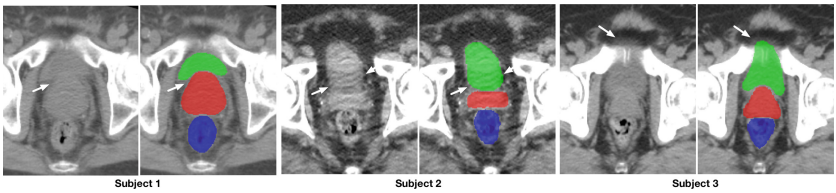

Fig. 1. Illustration of the blurry and vanishing boundaries in pelvic CT images. The green, red and blue masks indicate segmentation ground-truth of bladder, prostate, and rectum, respectively.

therapy (IGRT) and quantitative analysis for disease diagnosis. To obtain reliable segmentation for these applications, not only a robust detection of global object contours is required, a fine localization of tissue boundaries and small-scale structures is also fundamental. Nevertheless, the defection of image quality due to acquisition and process operations of medical images poses challenges to researchers in designing dependable segmentation algorithms.

Take the pelvic CT image as an example. The low soft-tissue-contrast makes the boundaries of target organs vague and hard to detect. This makes the nearby organs visually merged as a whole (see Fig. 1). In addition, different kinds of artifacts, e.g., metal, motion, and wind-mild artifacts, corrupt the real boundaries of organs and, more seriously, split the holistic organs into isolated parts with various sizes and shapes by generating fake boundaries (see Subject 2 in Fig. 1).

Numerous methods have been proposed in the literature to solve the problem of blurry image segmentation. Among the recently proposed algorithms, deep learning methods that are equipped with end-to-end learning mechanisms and representative features have become indispensable components and helped the corresponding algorithms to achieve state-of-the-art performances in many applications. For example, in [9], Oktay *et al.* integrated shape priors into a convolutional network through a novel regularization model to constrain the network of making appropriate estimation in the corrupted areas. In [6], Chen *et al.* introduced a multi-task network structure to simultaneously conduct image segmentation and boundary delineation to achieve better boundary localization performance. A large improvement has been made by the recently proposed algorithms. In the mainstream deep learning-based segmentation methods, to achieve good segmentation accuracy, high-resolution location information (provided by skip connections) is integrated with robust semantic information (extracted by downsampling and convolutions) to allow the network making local estimation with global guidance. However, both these kinds of information cannot help accurately locate the blurry boundaries contaminated by noise and surrounded by fake boundaries, thus posing the corresponding algorithms under potential failure in fine-grained medical image segmentation.

In this paper, to better detect the blurry boundary and tiny semantic structures, we propose a novel hierarchical dilated network. The main idea of our design is to first extract high-resolution context information, which is accurate

for localization and abundant in semantics. Then, based on the obtained high-resolution information, we endow our network the ability to infer the precise location of boundaries at blurry areas by collecting tiny but important clues and through observing the surrounding contour tendency in high resolution. To implement this idea, in the designed network, dilation is adopted to replace downsampling for receptive field enlargement to maintain precise location information. Also, by absorbing both the strength of DenseNet (the feature propagation and reuse mechanism) [3] and ResNet (the iterative feature refinement mechanism) [1], we concatenate the intermediate feature maps of several serially-connected dilated residual convolutional blocks and propose our hierarchical dilated module (HD-module). Then, different from the structures of ResNet and DenseNet, which link the dense blocks and residual blocks in a serial manner, we use parallel connections to integrate several deeply supervised HD-modules in different resolutions and construct our proposed hierarchical dilated neural network (HD-Net). After that, a late fusion module is introduced to further merge intermediate results from different HD-modules. In summary, the advantages of the proposed method are three-fold: (1) It can provide a better balance between *what* and *where* by providing high-resolution semantic information, thus helping improve the accuracy on blurry image segmentation; (2) It can endow sufficient context information to tiny structures and achieve better segmentation results on targets with small sizes; (3) It achieves smoother information flow and more elaborate utilization of multi-level (semantic and textural) and multi-scale information. Extensive experiments indicate superior performance of our method to the state-of-the-art deep learning medical image segmentation algorithms.

2 Method

In this section, we introduce our proposed hierarchical dilated neural network (HD-Net) for fine-grained medical image segmentation.

2.1 Hierarchical Dilated Network

Hierarchical Dilated Module (HD-Module). In order to extract high-resolution context information and protect the tiny semantic structure, we select dilated residual blocks as basic building blocks for our network. These blocks can arbitrarily enlarge the receptive field and efficiently extract context information without any compromise on the location precision. Also, the dilation operations eliminate the dependency on downsampling of the networks, thus allowing the tiny but important structures within images to be finely protected for more accurate segmentation. Our proposed hierarchical dilated module is constructed by concatenating the intermediate feature maps of several serially-connected dilated residual convolutional blocks (see Fig. 2). In the designed module, because of the combination of dense connections (concatenation) and residual connections, more smooth information flow is encouraged, and also, more comprehensive multi-level (textural and semantic) and multi-scale information is finely preserved.

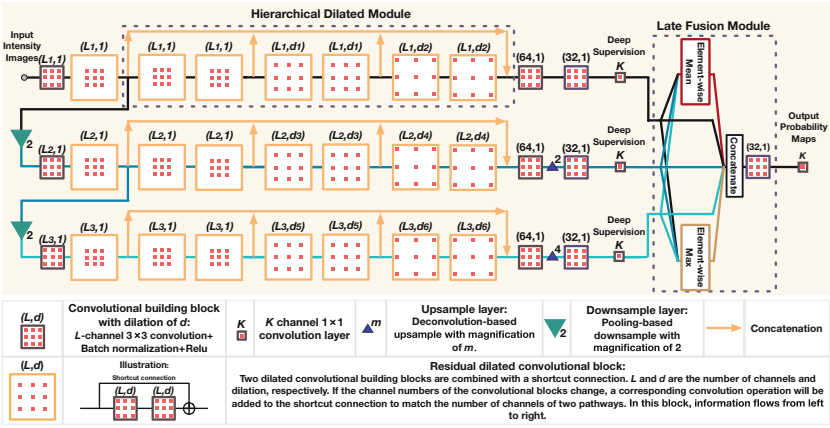

Fig. 2. Proposed hierarchical dilated network (HD-Net).

Hierarchical Dilated Network (HD-Net). To comprehensively exploit the diverse high-resolution semantic information from different scales, we further integrate several HD-modules and propose hierarchical dilated network (HD-Net). As we can see at the bottom of Fig. 2, convolution and downsampling operations tightly integrate three HD-modules from different resolutions into the network. Then, after upsampling and deep supervision operations [6], the intermediate probability maps of the three modules are further combined to generate the final output. The numbers of channels L_1, L_2, and L_3 of the three modules are 32, 48 and 72, respectively. The dilation factors are set as $d_1 = 3, d_2 = 5$ for high-resolution, $d_3 = 2, d_4 = 4$ for medium-resolution, and $d_5 = 2, d_6 = 2$ for low-resolution module. In this setting, when generating the output probability maps, multi-scale information from 12 receptive fields with sizes ranging from 7 to 206 is directly visible to the final convolutional layers, making the segmentation result precise and robust.

Late Fusion Module. Element-wise max or average [6] operations are two common fusion strategies in deep learning research. However, these methods treat all the results equally. Therefore, to better fuse the intermediate deeply supervised results from different sub-networks, we propose a late fusion module that weighs the outputs according to their quality and how they convey complementary information compared to other outputs. Specifically, we first generate the element-wise max and average of original outputs as intermediate results, and then automatically merge all the results through convolution. In this way, the enhanced intermediate results are automatically fused with more appropriate weights, to form an end-to-end model.

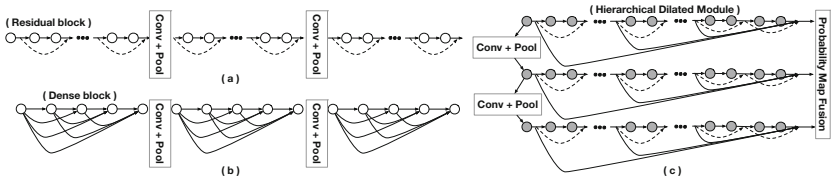

Fig. 3. Sketches of network structures of (a) ResNet, (b) DenseNet and (c) the proposed HD-Net. In the figure, unbroken arcs indicate concatenation, dotted arcs indicate element-wise plus, and straight lines indicate ordinal connections. Solid and hollow circles indicate convolution with and without dilation.

2.2 Comparison with ResNet and DenseNet

As discussed earlier, the proposed HD-Net borrows the advantages of both residual neural networks and dense networks. In this sub-section, we briefly compare the differences between these networks (See Fig. 3 for intuitive comparison).

Intra-block Connections. Residual blocks are constructed in a parallel manner by linking several convolutional layers with identity mapping, while dense blocks are constructed in a serial-parallel manner by densely linking all the preceding layers with the later layers. However, as pointed out by the latest research, although both networks perform great in many applications, the effective paths in residual networks are proved to be relatively shallow [2], which means the information interaction between lower layers and higher layers is not smooth enough. Also, compared to DenseNet, Chen *et al.* [8] argued that too frequent connections from the preceding layers may cause redundancy within the network. To solve the problem of information redundancy of DenseNet, in our network the dilated residual convolutions are selected as basic building blocks. In this building block, dilation can help speed up the process of iterative representation refinement within residual blocks [4], thus making the features extracted by two consecutive dilated residual convolution blocks be more diverse. Moreover, to solve the problem of lacking long-term connections within ResNet, we introduce dense connections into the serially connected dilated residual blocks and encourage a smoother information flow throughout the network.

Inter-block Connections. As far as inter-block connections are concerned, both ResNet and DenseNet use serial connection manners. As can be imagined, this kind of connection may suffer from a risk of blocking the low-layer textural information to be visible to the final segmentation result. Consequently, in our designed network, we adopt a parallel connection between HD-modules to achieve more direct utilization of multi-level information.

The Usage of Downsampling. ResNet and DenseNet mainly use downsampling operations to enlarge receptive field and to extract semantic information.

But in our proposed network, dilation becomes the main source of receptive field enlargement. Downsampling is mainly utilized for improving the information diversity and robustness of the proposed network. This setting also makes the design of parallel connections between modules to be more reasonable.

In summary, thanks to the dilation operations and the hierarchical structure, the high-resolution semantic information in different scales is fully exploited. Hence, HD-Net tends to provide a more detailed segmentation result, making it potentially more suitable for fine-grained medical image segmentation.

3 Experiments

Dataset and Implementation Details. To test the effectiveness of the proposed HD-Net, we adopt a pelvic CT image dataset with 339 scans for evaluation. The contours of the three main pelvic organs, i.e., prostate, bladder, and rectum have been delineated by experienced physicians and serve as ground-truth for segmentation. The dataset is randomly divided into training, validation and testing sets with 180, 59 and 100 samples, respectively. The patch size for all the compared networks is $144 \times 208 \times 5$. The implementations of all the compared algorithms are based on Caffe platform. To make a fair comparison, we use Xavier method to initialize parameters, and employ the Adam optimization method with fixed hyper-parameters for all the compared methods. Among the parameters, the learning rate (lr) is set to 0.001, and the decay rate hyper-parameters β_1 and β_2 are set to 0.9 and 0.999, respectively. The batch size of all compared methods is 10. The models are trained for at least 200,000 iterations until we observe a plateau or over-fitting tendency according to validation losses.

Evaluating the Effectiveness of Dilation and Hierarchical Structure in HD-Net. To conduct such an evaluation, we construct three networks for comparison. The first one is the HD-Net introduced in Sect. 2. The second one is an HD-Net without dilation (denoted by H-Net). The third one is constructed by the HD-module but without the hierarchical structure, i.e., with only one pathway (referred to as D-Net). The corresponding Dice similarity coefficient (DSC) and average surface distance (ASD) of these methods are listed in Table 1. Through the results, we can find that the introduction of dilation can contribute an improvement of approximately 1.3% on Dice ratio and 0.16 mm on ASD, while the introduction of hierarchical structure can contribute an improvement of approximately 2.3% on Dice ratio and 0.34 mm on ASD. It verifies the effectiveness of dilation and hierarchical structure in HD-Net.

Evaluating the Effectiveness of Late Fusion Module. From the reported DSC and ASD in Table 2, we can see that, with the help of the late fusion module, the network performance improves compared with the networks using average fusion (Avg-Fuse), max fusion (Max-Fuse), and simple convolution (Conv-Fuse).

Table 1. Evaluation of dilation and hierarchical structure in HD-Net.

Networks	Prostate	Bladder	Rectum	Prostate	Bladder	Rectum
	DSC (%)			ASD (mm)		
H-Net	86.1 ± 4.9	91.6 ± 8.7	85.5 ± 5.5	1.57 ± 0.78	1.58 ± 2.34	1.39 ± 0.50
D-Net	85.3 ± 4.7	91.5 ± 7.6	84.1 ± 5.4	1.62 ± 0.53	1.75 ± 2.53	1.70 ± 0.69
HD-Net	$\mathbf{87.7 \pm 3.7}$	$\mathbf{93.4 \pm 5.5}$	$\mathbf{86.5 \pm 5.2}$	$\mathbf{1.39 \pm 0.36}$	$\mathbf{1.34 \pm 1.75}$	$\mathbf{1.32 \pm 0.50}$

Table 2. Evaluation of the effectiveness of the proposed late fusion module.

Networks	Prostate	Bladder	Rectum	Prostate	Bladder	Rectum
	DSC (%)			ASD (mm)		
Avg-Fuse	87.0 ± 3.9	93.0 ± 6.2	85.7 ± 5.3	1.50 ± 0.44	1.43 ± 2.04	1.42 ± 0.48
Max-Fuse	87.2 ± 3.9	93.2 ± 5.4	86.1 ± 5.3	1.43 ± 0.37	$\mathbf{1.21 \pm 0.92}$	1.47 ± 0.67
Conv-Fuse	87.3 ± 3.9	93.1 ± 5.4	85.9 ± 5.5	1.45 ± 0.42	1.47 ± 2.41	1.48 ± 0.72
Proposed	$\mathbf{87.7 \pm 3.7}$	$\mathbf{93.4 \pm 5.5}$	$\mathbf{86.5 \pm 5.2}$	$\mathbf{1.39 \pm 0.36}$	1.34 ± 1.75	$\mathbf{1.32 \pm 0.50}$

Comparison with the State-of-the-Art Methods. Table 3 compares our proposed HD-Net with several state-of-the-art deep learning algorithms. Among these methods, U-Net [5] achieved the best performance on ISBI 2012 EM challenge dataset; DCAN [6] has won the 1st prize in 2015 MICCAI Grand Segmentation Challenge 2 and 2015 MICCAI Nuclei Segmentation Challenge; DenseSeg [7] has won the first prize in the 2017 MICCAI grand challenge on 6-month infant brain MRI segmentation.

Table 3 shows the segmentation results of U-Net [5], DCAN [6], DenseSeg [7], as well as our proposed network. As can be seen, all the results from the compared algorithms are reasonably well on predicting the global contour of the target organs; however, our proposed algorithm still outperforms the state-of-the-art methods by approximately 1% in Dice ratio and nearly **10%** in average surface distance for prostate and rectum. By visualizing the segmentation results of a representative sample in Fig. 4, we can see that the improvement mainly comes from the better boundary localization.

Table 3. Comparison with the state-of-the-art deep learning algorithms.

Networks	Prostate	Bladder	Rectum	Prostate	Bladder	Rectum
	DSC (%)			ASD (mm)		
U-Net [5]	86.0 ± 5.2	91.7 ± 5.9	85.5 ± 5.1	1.53 ± 0.49	1.77 ± 1.85	1.47 ± 0.53
DCAN [6]	86.8 ± 4.3	92.7 ± 7.1	84.8 ± 5.8	1.55 ± 0.55	1.72 ± 2.59	1.85 ± 1.13
DenseSeg [7]	86.5 ± 3.8	92.5 ± 7.0	85.2 ± 5.5	1.58 ± 0.53	1.37 ± 1.30	1.53 ± 0.76
Proposed	$\mathbf{87.7 \pm 3.7}$	$\mathbf{93.4 \pm 5.5}$	$\mathbf{86.5 \pm 5.2}$	$\mathbf{1.39 \pm 0.36}$	$\mathbf{1.34 \pm 1.75}$	$\mathbf{1.32 \pm 0.50}$

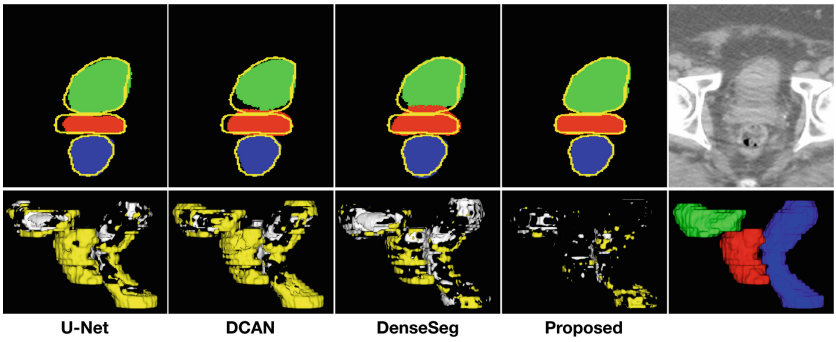

Fig. 4. Illustration of segmentation results. The first row visualizes the axial segmentation results and the corresponding intensity image (yellow curves denote the ground-truth contours). The second row is the 3D difference between the estimated and the ground-truth segmentation results. In these sub-figures, yellow and white portions denote the false positive and false negative predictions, respectively. The last sub-figure shows the 3D ground-truth contours.

4 Conclusion

In this paper, to address the adverse effect of blurry boundaries and also conduct fine-grained segmentation for medical images, we proposed to extract multiple high-resolution semantic information. To this end, we first replace downsampling with dilation for receptive field enlargement for accurate location prediction. Then, by absorbing both the advantages of residual blocks and dense blocks, we propose a new module with better mid-term and long-term information flow and less redundancy, i.e., hierarchical dilated module. Finally, by further integrating several HD-module with different resolutions using our newly defined late fusion module in parallel, we propose our hierarchical dilated network. Experimental results, based on a CT pelvic dataset, demonstrate the superior segmentation performance of our method, especially on localizing the blurry boundaries.

References

1. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
2. Veit, A., Wilber, M.J., Belongie, S.: Residual networks behave like ensembles of relatively shallow networks. In: NIPS, pp. 550–558 (2016)
3. Huang, G., Liu, Z., Weinberger, K.Q., et al.: Densely connected convolutional networks. In: CVPR, vol. 1, no. 2, p. 3 (2017)
4. Greff, K., Srivastava, R.K., Schmidhuber, J. Highway and residual networks learn unrolled iterative estimation. arXiv preprint arXiv:1612.07771 (2016)
5. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

6. Chen, H., Qi, X., Yu, L., et al.: DCAN: deep contour-aware networks for object instance segmentation from histology images. Med Image Anal. **36**, 135–146 (2017)
7. Bui, T.D., Shin, J., Moon, T.: 3D densely convolution networks for volumetric segmentation. arXiv preprint arXiv:1709.03199 (2017)
8. Chen, Y., Li, J., Xiao, H., et al.: Dual path networks. In: NIPS, pp. 4470–4478 (2017)
9. Oktay, O., et al.: Anatomically constrained neural networks (ACNN): application to cardiac image enhancement and segmentation. In: TMI (2017)