Bayesian VoxDRN: A Probabilistic Deep Voxelwise Dilated Residual Network for Whole Heart Segmentation from 3D MR Images

Zenglin Shi[1], Guodong Zeng[1], Le Zhang[2], Xiahai Zhuang[3], Lei Li[4], Guang Yang[5], and Guoyan Zheng[1](✉)

[1] Institute of Surgical Technology and Biomechanics, University of Bern, Bern, Switzerland
guoyan.zheng@istb.unibe.ch
[2] Advanced Digital Sciences Center, Illinois at Singapore, Singapore, Singapore
[3] School of Data Science, Fudan University, Shanghai, China
[4] Department of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[5] National Heart and Lung Institute, Imperial College London, London, UK

Abstract. In this paper, we propose a probabilistic deep voxelwise dilated residual network, referred as *Bayesian VoxDRN*, to segment the whole heart from 3D MR images. Bayesian VoxDRN can predict voxelwise class labels with a measure of model uncertainty, which is achieved by a dropout-based Monte Carlo sampling during testing to generate a posterior distribution of the voxel class labels. Our method has three compelling advantages. First, the dropout mechanism encourages the model to learn a distribution of weights with better data-explanation ability and prevents over-fitting. Second, focal loss and Dice loss are well encapsulated into a complementary learning objective to segment both hard and easy classes. Third, an iterative switch training strategy is introduced to alternatively optimize a binary segmentation task and a multi-class segmentation task for a further accuracy improvement. Experiments on the MICCAI 2017 multi-modality whole heart segmentation challenge data corroborate the effectiveness of the proposed method.

1 Introduction

Whole heart segmentation from magnetic resonance (MR) imaging is a prerequisite for many clinical applications including disease diagnosis, surgical planning and computer assisted interventions. Manually delineating all the substructures (SS) of the whole heart from 3D MR images is labor-intensive, tedious and subject to intra- and inter-observer variations. This has motivated numerous research works on automated whole heart segmentation such as atlas-based approaches [1,2], deformable model-based approaches [3], patch-based approaches [2,4] and machine learning based approaches [5]. Although significant

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11073, pp. 569–577, 2018.
https://doi.org/10.1007/978-3-030-00937-3_65

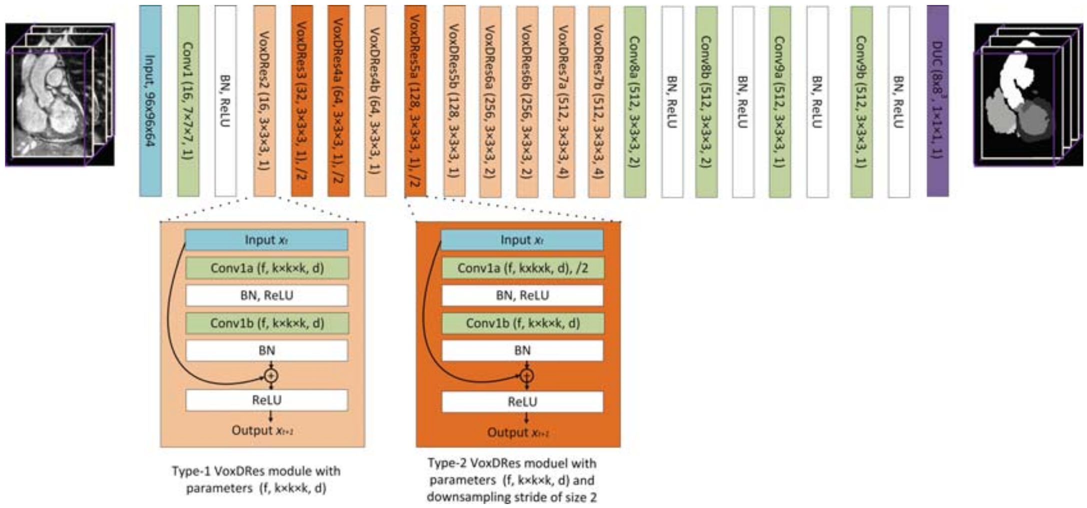

Fig. 1. The architecture of the proposed VoxDRN, consisting of BN layers, ReLU, and dilated convolutional layers N (ConvN) with parameters ($\mathbf{f}$, $\mathbf{k} \times \mathbf{k} \times \mathbf{k}$, $\mathbf{d}$), where $\mathbf{f}$ is the number of channels, $\mathbf{k} \times \mathbf{k} \times \mathbf{k}$ is the filter size, and $\mathbf{d}$ is the dilation size. At the output, we use DUC layer to generate voxel-level prediction. We also illustrate two different types of VoxDRes modules: type-1 without stride downsampling and type-2 with downsampling stride of size 2.

progress has been achieved, automated whole heart segmentation remains to be a challenging task due to large anatomical variations among different subjects, ambiguous cardiac borders and similar or even identical intensity distributions between adjacent tissues or SS of the heart.

Recently, with the advance of deep convolutional neural network (CNN)-based techniques [6–10], many CNN-based approaches have been proposed as well for automated whole heart segmentation with superior performance [2,11]. These methods basically follow a fully convolutional downsample-upsample pathway and typically commit to a single prediction without estimating the model uncertainty. Moreover, different SS of the heart vary greatly in volume size, e.g., the left atrium blood cavity and the pulmonary artery often have smaller volume size than others. This can cause learning bias towards the majority class and poor generalization, i.e., the class-imbalance problem. To address such a concern, class-balanced loss functions have been proposed such as weighted cross entropy [2] and Dice loss [10].

This paper proposes a probabilistic deep voxelwise dilated residual network (VoxDRN), referred as *Bayesian VoxDRN*, which is able to predict voxelwise class labels with a measure of the model uncertainty. This involves following key innovations: (1) we extend the dilated residual network (DRN) of [12], previously limited to 2D image segmentation, to 3D volumetric segmentation; (2) inspired by the work of [13,14], we introduce novel architectures incorporating multiple dropout layers to estimate the model uncertainty, where units are randomly inactivated during training to avoid over-fitting. At testing, the posterior distribution of voxel labels is approximated by Monte Carlo sampling of multiple predictions with dropout; (3) we propose to combine focal loss with Dice loss, aiming for a complementary learning to address the class imbalance issue; and (4) we introduce an iterative switching training strategy to alternatively optimize a binary segmentation task and a multi-class segmentation task for a further accuracy improvement. We conduct ablation study to investigate the effectiveness of each proposed component in our method.

2 Methods

We first present our 3D extension to the 2D DRN of [12], referred as *VoxDRN*. Building on it, we then devise new architectures incorporating multiple dropout layers for model uncertainty estimation.

DRN. Dilated residual network [12] is a recently proposed method built on residual connections and dilated convolutions. The rationale behind DRN is to retain high spatial resolution and provide dense output to cover the input field such that back-propagation can learn to preserve detailed information about smaller and less salient objects. This is achieved by dilated convolutions which allow for exponential increase in the receptive field of the network without loss of spatial resolution. Building on the ResNet architecture of [6], Yu et al. [12] devised DRN architecture using dilated convolutions. Additional adaptations were used to eliminate gridding artifacts caused by dilated convolutions [12] via (a) removing max pooling operation from ResNet architecture; (b) adding 2 dilated residual blocks at the end of the network with progressively lower dilation; and (c) removing residual connections of the 2 newly added blocks. DRN works in a fully-convolutional manner to generate pixel-level prediction using bilinear interpolation of the output layer.

VoxDRN. We extend DRN to 3D by substituting 2D operators with 3D ones to create a deep voxelwise dilated residual network (VoxDRN) architecture as shown in Fig. 1. Our architecture consists of stacked voxelwise dilated residual (VoxDRes) modules. We introduce two different types of VoxDRes modules: type-1 without stride downsampling and type-2 with downsampling stride of size 2 as shown in Fig. 1. In each VoxDRes module, the input feature x_l and transformed feature $F_l(x_l, W_l)$ are added together with skip connection, and hence the information can be directly propagated to next layer in the forward and backward passes. There are three type-2 VoxDRes modules with downsampling stride of size 2, which reduce the resolution size of input volume by a factor of 8. We empirically find that such a resolution works well to preserve important information about smaller and less salient objects. The last VoxDRes module is followed by four convolutional layers with progressively reduced dilation to eliminate gridding artifacts. Batch normalization (BN) layers are inserted intermediately to accelerate the training process and improve the performance [15]. We use the rectified linear units (ReLU) as the activation function for non-linear transformation [16].

In order to achieve volumetric dense prediction, we need to recover full resolution at output. Conventional method such as bilinear upsampling [12] is

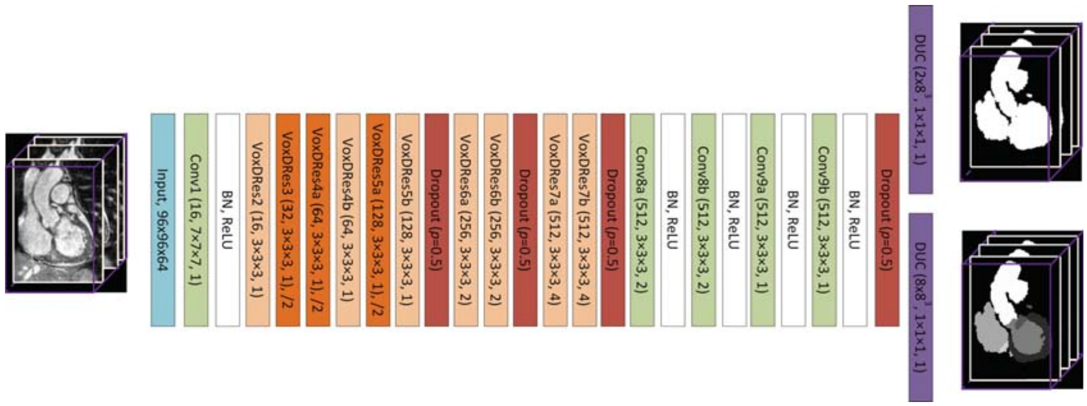

Fig. 2. The architecture of our Bayesian VoxDRN.

not attractive as the upsampling parameters are not learnable. Deconvolution could be an alternative but, unfortunately, it can easily lead to "uneven overlap", resulting in checkerboard artifacts. In this paper, we propose to use Dense Upsampling Convolution (DUC) of [17] to get the voxel-level prediction at the output where the final layer has Cr^3 channels, r being the upsampling rate and C being the number of classes. The DUC operation takes an input of shape $h \times w \times d \times Cr^3$ and remaps voxels from different channels into different spatial locations in the final output, producing a $rh \times rw \times rd \times C$ image, where h, w, and d denote height, width and depth. The mapping is done in 3D with $O(F)_{i,j,k,c} = F_{[i/r],[j/r],[k/r],r^3\cdot c+\bmod(i,r)+r\cdot\bmod(j,r)+r^2\cdot\bmod(k,r)}$ where F is the pre-mapped feature responses and O is the output image. DUC is equivalent to a learned interpolation that can capture and recover fine-detailed information with the advantages to avoid checkerboard artifacts of deconvolution.

Bayesian VoxDRN. Gal and Ghahramani [13] demonstrated that Bayesian CNN offered better robustness to over-fitting on small data than traditional approaches. Given our observed training data $\mathbf{X}$ and labels $\mathbf{Y}$, Bayesian CNN requires to find the posterior distribution $p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$ over the convolutional weights, $\mathbf{W}$. In general, this posterior distribution is not tractable. Gal and Ghahramani [13] suggested to use variational dropout to tackle this problem for neural networks. Inspired by the work of [13,14], we devise a new architecture incorporating dropout layers as shown in Fig. 2, referred as *Bayesian VoxDRN*, to enable estimation of the model uncertainty, where subsets of units are inactivated with a dropout probability of 0.5 during training to avoid over-fitting. Applying dropout after each convolution layer may slow down the learning process. This is because the shallow layers of a CNN, which aims at extracting low-level features such as edges can be better modeled with deterministic weights [14]. We insert four dropout layers in the higher layers of VoxDRN to learn Bayesian weights on higher level features such as shape and contextual information. At testing, we sample the posterior distribution over the weights using dropout to obtain the posterior distribution of softmax class probabilities. The final segmentation is obtained by conducting majority voting on these samples. We use the variance to obtain model uncertainty for each class. In our experiments, following the suggestion in [13,14], we used 10 samples in majority voting to have a better accuracy and efficiency trade-off.

Hybrid Loss. We propose to combine weighted focal loss [18] with Dice loss [10] to solve class imbalance problem. The weighted focal loss is calculated as $L_{wFL} = \sum_{c \in C} -\alpha_c(1-p_c)^\lambda log(p_c)$, where $|X|$ and $|X_c|$ are the frequency of all classes and that of class c, respectively; $\alpha_c = 1 - \frac{|X_c|}{|X|}$ is designed to adaptively balance the importance of large and small SS of the heart; p_c is the probability of class c and $(1-p_c)^\lambda$ is the scaling factor to reduce the relative loss for well-classified examples such that we can put more focus on hard, misclassified examples. Focal loss often guides networks to preserve complex boundary details but could bring certain amount of noise, while Dice loss tends to generate smoother segmentation. Therefore, we propose to combine these two loss functions with equal weights for a complementary learning.

Iterative Switch Training. We propose a progressive learning strategy to train our Bayesian VoxDRN. The rationale and intuition behind such a strategy are that we would like to first separate foreground from background, and then further segment the foreground into a number of SS of the heart. By doing this, our network is alternatively trained to solve a simpler problem at each step than the original one. To achieve this, as shown in Fig. 2, the Bayesian VoxDRN is modified to have two branches after the last convolution layer: each branch, equipped with its own loss and operated only on images coming from the corresponding dataset, is responsible for estimating the segmentation map therein. During training, we alternatively optimize our network by using binary loss and multi-class loss supervised by binary labels and multi-class labels, respectively. Please note that at any moment of the training, only one branch is trained. More specifically, at each training epoch, we first train the binary branch to learn to separate the foreground from the background. We then train the multi-class branch to put the attention of our model to segment foreground into a few SS of the heart. While at testing, we are only interested in the output from the multi-class branch.

Implementation Details. The proposed method was implemented with Python using TensorFlow framework and trained on a workstation with a 3.6 GHz Intel i7 CPU and a GTX1080 Ti graphics card with 11 GB GPU memory. The network was trained using Adam optimizer with mini-batch size of 1. In total, we trained our network for 5'000 epochs. All weights were randomly initialized. We set initial momentum value to 0.9 and initial learning rate to 0.001. Randomly cropped $96 \times 96 \times 64$ sub-volumes serve as input to train our network. We adopted sliding window and overlap-tiling stitching strategies to generate predictions for the whole volume, and removed small isolated connected components in the final labeling results.

3 Experiments and Results

Data and Pre-processing. We conducted extensive experiments to evaluate our method on the 2017 MM-WHS challenge MR dataset [1,4][1]. There are in total 20 3D MR images for training and another 40 scans for testing. The training dataset contains annotations for seven SS of the heart including blood cavity for the left ventricle (LV), the right ventricle (RV), the left atrium (LA) and the right atrium (RA) as well as the myocardium of the LV (Myo), the ascending aorta (AA) and the pulmonary artery (PA). We resampled all the training data to isotropic resolution and normalized each image as zero mean and unit variance. Data augmentation was used to enlarge the training samples by rotating each image with a random angle in the range of $[-30°, 30°]$ around z axis.

Comparison with Other Methods. The quantitative comparison between our method and other approaches from the participating teams is shown in Table 1. According to the rules of the challenge, methods were ranked based on Dice score on the whole heart segmentation, not on each individual substructure. Although most of the methods are based on CNNs, Heinrich et al. [2] achieved impressive results using discrete nonlinear registration and fast non-local fusion.

Table 1. Comparison (Dice score) with different approaches on MM-WHS 2017 MR dataset. The best result for each category is highlighted with bold font.

Methods	LV	Myo	RV	LA	RA	AA	PA	Whole heart
Our method	0.914	**0.811**	**0.880**	0.856	0.873	0.857	0.794	**0.871**
Heinrich et al. [2]	**0.918**	0.781	0.871	**0.886**	0.873	**0.878**	**0.804**	0.870
Payer et al. [2]	0.916	0.778	0.868	0.855	**0.881**	0.838	0.731	0.863
Mortazi et al. [2]	0.871	0.747	0.830	0.811	0.759	0.839	0.715	0.818
Galisot et al. [2]	0.897	0.763	0.819	0.765	0.808	0.708	0.685	0.817
Yang et al. [2]	0.836	0.721	0.805	0.742	0.832	0.821	0.697	0.797
Wang et al. [2]	0.855	0.728	0.760	0.832	0.782	0.771	0.578	0.792
Yu et al. [2]	0.750	0.658	0.750	0.826	0.859	0.809	0.726	0.783
Liao et al. [2]	0.702	0.623	0.680	0.676	0.654	0.599	0.470	0.670

Table 2. Ablation study results [x 100%].

Methods	Dice		Jaccard		Specificity		Recall	
	WH	SS	WH	SS	WH	SS	WH	SS
HighRes3DNet [19]	88.17 ± 0.25	80.42 ± 0.29	79.21 ± 0.63	68.85 ± 0.48	93.96 ± 0.02	87.54 ± 0.20	83.37 ± 0.65	76.63 ± 0.58
3D U-net [9]	88.33 ± 0.35	81.67 ± 0.36	79.59 ± 0.84	70.91 ± 0.62	94.17 ± 0.02	89.04 ± 0.12	83.79 ± 0.91	78.16 ± 0.78
Bayesian VoxDRN+Dice	89.38 ± 0.09	82.58 ± 0.30	80.94 ± 0.25	71.25 ± 0.61	92.97 ± 0.06	85.84 ± 0.30	87.18 ± 0.42	81.13 ± 0.45
Bayesian VoxDRN+Hybrid	90.15 ± 0.10	83.12 ± 0.26	82.23 ± 0.29	72.27 ± 0.49	91.81 ± 0.08	85.53 ± 0.21	88.91 ± 0.43	82.92 ± 0.34
Our method	90.83 ± 0.06	84.39 ± 0.19	83.30 ± 0.19	73.75 ± 0.42	91.99 ± 0.06	85.62 ± 0.17	89.93 ± 0.28	89.93 ± 0.28

[1] One can find details about the MICCAI 2017 MM-WHS challenge at: http://www.sdspeople.fudan.edu.cn/zhuangxiahai/0/mmwhs/.

The other non-CNN approach was introduced by Galisot et al. [2], which was based on local probabilistic atlases and a posterior correction.

Ablation Analysis. In order to evaluate the effectiveness of different components in the proposed method, we performed a set of ablation experiments. Because the ground truth of the testing dataset is held out by the organizers and the challenge organizers only allow resubmission of substantially different methods, we conducted experiments via a standard 2-fold cross-validation study on the training dataset. We also implemented two other state-of-the-art 3D CNN approaches, 3D U-net [9] and HighRes3DNet [19], for comparison. We compared these two methods with following variants of the proposed method: (1) Bayesian VoxDRN trained with Dice loss (Bayesian VoxDRN+Dice); (2) Bayesian VoxDRN trained with our hybrid loss but without using the iterative switch training strategy (Bayesian VoxDRN+Hybrid); and (3) Bayesian VoxDRN trained with our hybrid loss using the iterative switch training strategy (Our method). We evaluated these methods using Dice, Jaccard, specificity and recall for the whole heart (WH) segmentation as well as for segmentation of all SS. The quantitative comparison can be found in Table 2. As observed, our method and its variants achieved better performance than the other two methods under limited training data. Moreover, each component in our method helped to improve the performance. Qualitative results are shown in Fig. 3, where we (A) visually compared the results obtained by different methods; (B) visualized the uncertainty map; and (C) depicted the relationship between the segmentation accuracy and the uncertainty threshold. From Fig. 3(B), one can see that the model is uncertain at object boundaries and with difficult and ambiguous SS.

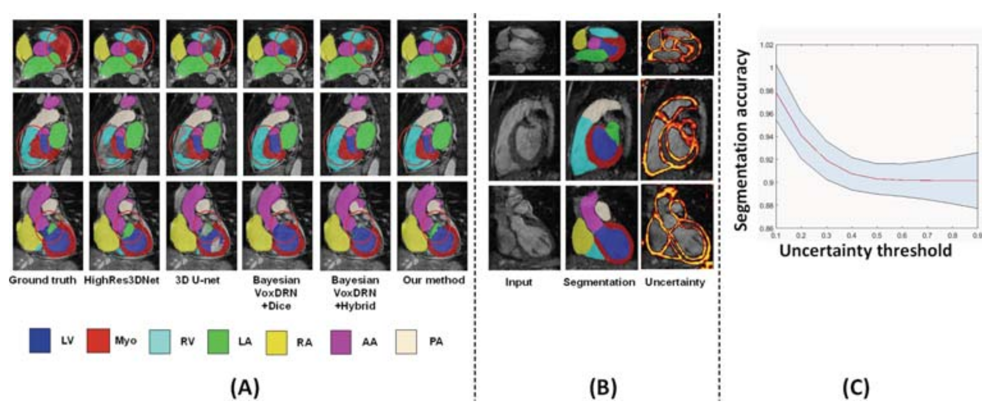

Fig. 3. Qualitative results. (A) qualitative comparison of different methods. Red circles highlight the major differences among various methods; (B) visualization of uncertainty, where the brighter the color, the higher the uncertainty; and (C) the relationship between the segmentation accuracy and the uncertainty threshold. The shaded area represents the standard errors.

4 Conclusion

In this study, we proposed the Bayesian VoxDRN, a probabilistic deep voxelwise dilated residual network with a measure of the model uncertainty, for automatic whole heart segmentation from 3D MR images. The proposed Bayesian VoxDRN models uncertainty by incorporating variational dropouts for an approximated Bayesian inference. In addition, it works well in imbalanced dataset by using both focal loss and Dice loss. Finally, a further improvement on performance is achieved by employing an iterative switch training strategy to train the Bayesian VoxDRN. Comprehensive experiments on an open challenge dataset demonstrated the efficacy of our method in dealing with whole heart segmentation under limited training data. Our network architecture shows promising generalization and can be potentially extended to other applications.

Acknowlegement. The project is partially supported by the Swiss National Science Foundation Project 205321_169239.

References

1. Zhuang, X., Rhode, K., et al.: A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI. IEEE Trans. Med. Imaging **29**, 1612–1625 (2010)
2. Pop, M., et al. (eds.): STACOM 2017. LNCS, vol. 10663. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75541-0
3. Peters, J., et al.: Optimizing boundary detection via simulated search with applications to multi-modal heart segmentation. Med. Image Anal. **14**, 70–84 (2009)
4. Zhuang, X., Shen, J.: Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. Med. Image Anal. **31**, 77–87 (2016)
5. Zheng, Y., Barbu, A., et al.: Four-chamber heart modeling and automatic segmentation for 3-D cardiac CT volumes using marginal space learning and steerable features. IEEE Trans. Med. Imaging **27**(11), 1668–1681 (2008)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings CVPR, pp. 770–778 (2016)
7. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings CVPR, pp. 3431–3440 (2015)
8. 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
9. Çiçek, Ö., et al.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., et al. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
10. Milletari, F., Navab, M., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. arXiv:1606.04797 (2016)
11. Yu, L., et al.: Automatic 3D cardiovascular MR segmentation with densely-connected volumetric ConvNets. In: Descoteaux, M., et al. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 287–295. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_33

12. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Proceedings CVPR, pp. 636–644 (2017)
13. Gal, Y., Ghahramani, Z.: Bayesian convolutional neural networks with bernoulli approximate variational inference. arXiv:1506.02158 (2015)
14. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv:1511.02680 (2016)
15. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings ICML, pp. 448–456 (2015)
16. Krizhevsky, A., Ilya, S., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the NIPS, pp. 1097–1105 (2012)
17. Wang, P., Chen, P., et al.: Understanding convolution for semantic segmentation. arXiv:1702.08502 (2017)
18. Lin, T.Y., et al.: Focal loss for dense object detection. In: Proceedings ICCV (2017)
19. Li, W., et al.: On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 348–360. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_28