Automatic Lymph Node Cluster Segmentation Using Holistically-Nested Neural Networks and Structured Optimization in CT Images

Isabella Nogues[1(✉)], Le Lu[1], Xiaosong Wang[1], Holger Roth[1], Gedas Bertasius[2], Nathan Lay[1], Jianbo Shi[2], Yohannes Tsehay[1], and Ronald M. Summers[1]

[1] Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, MD 20892-1182, USA
isabella.nogues@nih.gov
[2] University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract. Lymph node segmentation is an important yet challenging problem in medical image analysis. The presence of enlarged lymph nodes (LNs) signals the onset or progression of a malignant disease or infection. In the thoracoabdominal (TA) body region, neighboring enlarged LNs often spatially collapse into "swollen" lymph node clusters (LNCs) (up to 9 LNs in our dataset). Accurate segmentation of TA LNCs is complexified by the noticeably poor intensity and texture contrast among neighboring LNs and surrounding tissues, and has not been addressed in previous work. This paper presents a novel approach to TA LNC segmentation that combines holistically-nested neural networks (HNNs) and structured optimization (SO). Two HNNs, built upon recent fully convolutional networks (FCNs) and deeply supervised networks (DSNs), are trained to learn the LNC appearance (HNN-A) or contour (HNN-C) probabilistic output maps, respectively. HNN first produces the class label maps with the same resolution as the input image, like FCN. Afterwards, HNN predictions for LNC appearance and contour cues are formulated into the unary and pairwise terms of conditional random fields (CRFs), which are subsequently solved using one of three different SO methods: dense CRF, graph cuts, and boundary neural fields (BNF). BNF yields the highest quantitative results. Its mean Dice coefficient between segmented and ground truth LN volumes is 82.1 % ± 9.6 %, compared to 73.0 % ± 17.6 % for HNN-A alone. The LNC relative volume (cm^3) difference is 13.7 % ± 13.1 %, a promising result for the development of LN imaging biomarkers based on volumetric measurements.

1 Introduction

Lymph node (LN) segmentation and volume measurement play a crucial role in important medical imaging based diagnosis tasks, such as quantitatively evaluating disease progression or the effectiveness of a given treatment or therapy. Enlarged LNs, defined by the widely observed RECIST criterion [14] to have

S. Ourselin et al. (Eds.): MICCAI 2016, Part II, LNCS 9901, pp. 388–397, 2016.
DOI: 10.1007/978-3-319-46723-8_45

a short axis diameter ≥ 10 mm on an axial computed tomography (CT) slice, signal the onset or progression of a malignant disease or an infection. Often performed manually, LN segmentation is highly complex, tedious and time consuming. Previous methods for automatic LN segmentation in CT images fall under several categories, including atlas registration and label fusion [17], 3D deformable surface shape model [6] and statistical 3D image feature learning [1,7], respectively. This paper addresses and solves a novel problem: lymph node cluster (LNC) segmentation in the thoracoabdominal (TA) region. LN volumes are subsequently predicted from our segmentation results.

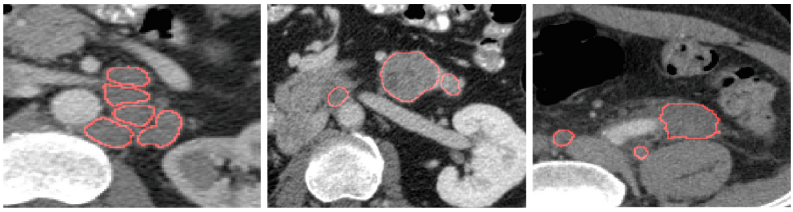


Fig. 1. CT images of thoracoabdominal lymph node clusters with annotated (red) boundaries.

In CT, the TA region exhibits exceptionally poor intensity and texture contrast among neighboring LNs and between LNs and their surrounding tissues. Furthermore, TA LNs often appear in clusters. Weak intensity contrast renders the boundaries of distinct agglomerated LNs ambiguous (Fig. 1). Existing fully-automated methods have been applied to the more contrast-distinctive axillary and pelvic regions [1], as well as the head-and-neck section [6,17]. This paper presents a fully-automated method for TA LNC segmentation. More importantly, the segmentation task is formulated as a flexible, bottom-up image binary classification problem that can be effectively solved using deep convolutional neural networks (CNN) and graph-based structured optimization and inference. Our bottom-up approach can easily handle all variations in LNC size and spatial configuration. By contrast, top-down, model-fitting methods [1,6,7,17] may struggle to localize and segment each LN. LN volume is a more robust metric than short-axis diameter, which is susceptible to high inter-observer variability and human error. Furthermore, our proposed method is well-suited for measuring agglomerated LNs, whose ambiguous boundaries compromise the accuracy of diameter measurement.

This paper addresses a clinically relevant and challenging problem: automatic segmentation and volume measurement of TA LNCs. A publicly available dataset[1] containing 171 TA 3D CT volumes (with manually-annotated LN segmentation masks) [15] is used. The method in this paper integrates HNN learning with structured optimization (SO). Furthermore, it yields remarkable

[1] https://wiki.cancerimagingarchive.net/display/Public/CT+Lymph+Nodes.

quantitative results. The mean Dice similarity coefficient (DSC) between predicted and segmented LN volumes is 82.1 % ± 9.6 % for boundary neural fields (BNF), and 73.0 % ± 17.6 % for HNN-A. The relative volume measurement error is 13.7 % ± 13.1 % for BNF and 32.16 % ± 36.28 % for HNN-A.

2 Methods

Our segmentation framework comprises two stages: holistically-nested neural network (HNN) training/inference and structured optimization (SO). (1) Two HNNs, designed in [18], are trained on pairs of raw CT images from the TA region and their corresponding binary LN appearance (segmentation) or contour (boundary) masks. We denote them as HNN-A and HNN-C, respectively. The HNN merges the CNN frameworks of a fully convolutional network (FCN) [11] and a deeply-supervised network (DSN) [10]. The FCN component is an end-to-end holistic image training-prediction architecture: the output probability label map has the same dimension as the input image. The DSN component performs multi-scale feature learning, in which deep layer supervision informs and refines classification results at multiple convolutional stages. HNN's multi-level contextual architecture and auxiliary cost functions (which assign pixel-wise label penalties) allow for capturing implicit, informative deep features to enhance the segmentation accuracy. (2) However, HNN's large receptive fields and pooling layers potentially lead to segmentation outputs that are not precisely localized along the LN boundaries. Hence, we implement and evaluate explicit Conditional Random Field (CRF) based structured optimization schemes [2,3,9] to refine segmentation results. Particularly, we optimize a CRF energy function that includes a unary term encoding LN area information (HNN-A) and a pairwise term representing the object-specific LN boundary discontinuities (learned by HNN-C via deep supervision). This pairwise energy differs from the conventional intensity contrast-sensitive term from [9].

We note that the integration of boundary and appearance neural networks for automatic segmentation has also been newly exploited in [13]. Both our paper and [13] have been inspired by the visual cognitive science and computer vision literature (namely [12,18]). However, the global methods are different: [13] refines HNN predictions via robust spatial aggregation using random forest, as opposed to structured optimization.

2.1 Holistically-Nested Neural Networks

The holistically-nested neural network (HNN) [18] was first proposed as an image-to-image solution to the long-standing edge detection problem using deep CNN. In this study, we empirically find that the HNN architecture is also highly effective and efficient in predicting the full object segmentation mask, due to its per-pixel label cost formulation. Therefore, we train two separate holistically-nested neural networks to learn the probabilistic label maps of the LN specific binary appearance mask (HNN-A) and contour mask (HNN-C) from raw TA

CT images. The HNN-A prediction map provides the approximate location and shape of the target LNC, while the HNN-C prediction map renders LN boundary cues. By learning boundaries alone, HNN-C generates boundary information with refined spatial and contextual detail, relative to HNN-A. HNN-A and HNN-C results are combined in the Structured Optimization phase (cf. Section: Structured Optimization) to obtain more accurate pixel-wise label predictions.

HNN Training. We adopt the CNN architecture in [18], which derives from a VGGNet model pre-trained on ImageNet [16]. The HNN contains five convolutional stages, with strides 1, 2, 4, 8, and 16, respectively, and different receptive field sizes, all nested in the VGGNet as in [18]. The HNN includes one side layer per convolutional stage, which is associated with an auxiliary classifier. The side outputs, generated by each side layer, are increasingly refined, as they gradually approach the ground truth. Finally, all side outputs are fed into a "weighted-fusion" layer, which generates a global probability map merging the information from all side output scales. During the training phase, HNN seeks to minimize the per-pixel cost of each network stage, by applying stochastic gradient descent to the global objective function

$$(\mathbf{W}, \mathbf{w}, \mathbf{h})^* = \text{argmin}(\mathcal{L}_{\text{side}}(\mathbf{W}, \mathbf{w}) + \mathcal{L}_{\text{fuse}}(\mathbf{W}, \mathbf{w}, \mathbf{h})), \tag{1}$$

where $\mathcal{L}_{\text{side}}$ is the loss function computed at each side-output layer (i.e., auxiliary cost functions), and $\mathcal{L}_{\text{fuse}}$ is the cross-entropy distance between the ground truth and fusion layer output edge maps. The loss function $\mathcal{L}_{\text{side}}$ is a linear combination of the image-level loss functions $\ell_{\text{side}}^{(m)}(\mathbf{W}, \mathbf{w}^{(m)})$. The parameters $\mathbf{W}$ correspond to the set of standard network parameters, and $\mathbf{w}$ to the weights in each side-output layer's classifier. Due to the per-pixel cost setup [11,18], HNN does not require a large number of training images to converge, and thus can be applied to small datasets.

HNN Testing. During the testing phase, the network generates edge map predictions for each layer. The final unified output is a weighted average of all prediction maps:

$$\hat{Y}_{\text{HED}} = \text{Average}(\hat{Y}_{\text{fuse}}, \hat{Y}_{\text{side}}^{(1)}, \ldots, \hat{Y}_{\text{side}}^{(5)}) \tag{2}$$

HNN is highly efficient, requiring a mere 0.4 s per image in the feed-forward testing. Further details on the HNN architecture and training are provided in [18].

2.2 Structured Optimization

Although HNN is a state-of-the-art, image-to-image, semantic pixel-wise labeling method, it tends to produce imprecise segmentations, like many deep CNN models. Its large receptive fields and many pooling layers compromise clarity and spatial resolution in the deep layers. Therefore, we exploit explicit structured optimization techniques to refine HNN-A's segmentation results. We select

a conditional random field (CRF) optimization framework, as it is well-suited for integrating LN predictions with LN boundary cues. As in [2], the boundary cues serve to improve segmentation coherence and object localization. For a given target CT image, the unary potential is a function of the corresponding HNN-A prediction, while the pairwise potential is a function of the corresponding HNN-C prediction. Three structured optimization representations and methods are described and evaluated: dense CRF, graph cuts, and boundary neural fields.

Under all three techniques, segmentation is cast as a binary classification problem. A graph representation for the original CT image is provided, in which vertices correspond to image pixels, and edges to inter-pixel connections. A boundary strength-based affinity function, defined in [2], for distinct pixels i and j is given by:

$$w_{ij} = \exp\left(\frac{-M_{ij}}{\sigma}\right), \tag{3}$$

where M_{ij} is the magnitude of the strongest LN boundary intersecting $\{i, j\}$, and σ is a smoothing hyper-parameter. The boundary strength map is obtained by performing non-maximum suppression on an HNN-C prediction [18]. This boundary-based affinity function better refines the segmentation than would a standard image intensity gradient-based function, as intensity information is highly ambiguous in TA CT images. We set the degree of pixel i to $d_i = \sum_{i \neq j}^{N} w_{ij}$, where N is the total number of pixels.

Dense Conditional Random Field: Our dense conditional random field (dCRF) representation follows the framework in [5,9]. We adopt the CT intensity contrast-sensitive pixel affinities for all possible image pixel pairs, as described in [5]. Finally, the dCRF solver designed in [9] is utilized, as a variation of distributed message passing.

Graph Cuts: The minimum-cut/maximum-flow graph cuts (GC) algorithm described in [4] is applied to the segmentation problem. We optimize an energy function whose unary term is the negative log-likelihood of the HNN-A LN segmentation probability value per pixel and pairwise term is defined using Eq. 3 [2]. All inter-pixel affinities are computed within a 20×20 neighborhood for each pixel location.

Boundary Neural Fields: The LN mask (HNN-A) and boundary (HNN-C) predictions are integrated into a matrix model. We optimize the global energy function:

$$\mathbf{X}^* = \underset{\mathbf{X}}{\operatorname{argmin}} \frac{\mu}{2}\mathbf{D}(\mathbf{X} - \mathbf{D^{-1}f})^T(\mathbf{X} - \mathbf{D^{-1}f}) + \frac{1}{2}\mathbf{X^T}(\mathbf{D} - \mathbf{W})\mathbf{X}, \tag{4}$$

where $\mathbf{X}^*$ is a N x 1 vector representing an optimal continuous *label assignment* for a vectorized input image (with N pixels), $\mathbf{D}$ is the N x N diagonal degree matrix, $\mathbf{W}$ is the N x N pairwise affinity matrix, and $\mathbf{f}$ is a N x 1 vector containing the HNN-A prediction values. Each diagonal entry $d_{i,i}$ is set to d_i, defined above. $\mathbf{W}$ is a sparse weight matrix: the entries w_{ij} are computed only for pixel pairs i, j belonging to the same 20×20 pixel neighborhood. (via Eq. 3 [2]).

The unary energy attempts to find a segmentation assignment $\mathbf{X}$ that deviates little from the HNN-A output. The assignment $\mathbf{X}$ is weighted by $\mathbf{D}$, in order to assign larger unary costs to pixels with many similar neighbors. By contrast, the pairwise energy minimizes the cost assigned to such pixels by weighting the squared distances between segmentation assignments of similar pixel pairs $\{i, j\}$ by their affinity w_{ij}. To balance the unary and pairwise contributions, the unary term is weighted by the hyperparameter μ. As in [2], μ is set to 0.025. The optimal segmentation is given by:

$$\mathbf{X}^* = (\mathbf{D} - \alpha\mathbf{W})^{-1}\beta\mathbf{f}, \tag{5}$$

where $\alpha = \frac{1}{1+\mu}$ and $\beta = \frac{\mu}{1+\mu}$. Note that Eq. 5 is a closed-form solution.

3 Results and Discussion

3.1 Dataset Creation

Our dataset contains 84 abdominal and 87 mediastinal 3D CT scans ($512 \times 512 \times 512$ voxels) (publicly available from [15]). We spatially group the ground truth binary LN masks in 3D to form clusters with a linking distance constraint. All LN clusters, padded by 32 pixels in each direction, are subsequently cropped, resulting in 1∼7 such subvolume regions ($77 \times 76 \times 79$ – $212 \times 235 \times 236$ voxels) per CT volume. All CT axial slices have been extracted from the portal venous phase with slice thickness $1 - 1.25$ mm and manually segmented by an expert radiologist. This yields a total of $39,361$ images ($16,268$ images containing LN pixels) in 411 LN clusters (with 395 abdominal and 295 mediastinal LNs). By extracting all LN contours from the ground truth appearance masks, we obtain the LN contour masks to train HNN-C. Examples of LN CT image ground truth boundaries are shown in Figs. 1 and 2.

3.2 Quantitative Analysis

Segmentation accuracies of HNN-A, BNF, GC, and dCRF are evaluated. The volume-wise means and standard deviations (std.) are computed for three evaluations metrics: Dice similarity coefficient (DSC), Intersection over Union (IoU), and Relative Volume Difference (RVD) (cm^3) between predicted and ground truth LNC 3D masks. The RVD indicates whether volume measurement is accurate enough to be used as a new imaging biomarker, in addition to the diameter-based RECIST criterion [14].

Our experiments are conducted under 4-fold cross-validation, with the dataset split at the patient level. Prior to generating binary segmentation results for HNN-A prediction maps alone (with pixels in the range $[0, 1]$), we remove all pixels below the threshold $\tau = 8.75 \times 10^{-1}$. This value of τ, which is shown to maximize the mean DSC between the HNN-A predictions and ground truth LN masks, is calibrated using the training folds. HNN is run on Caffe [8], using a

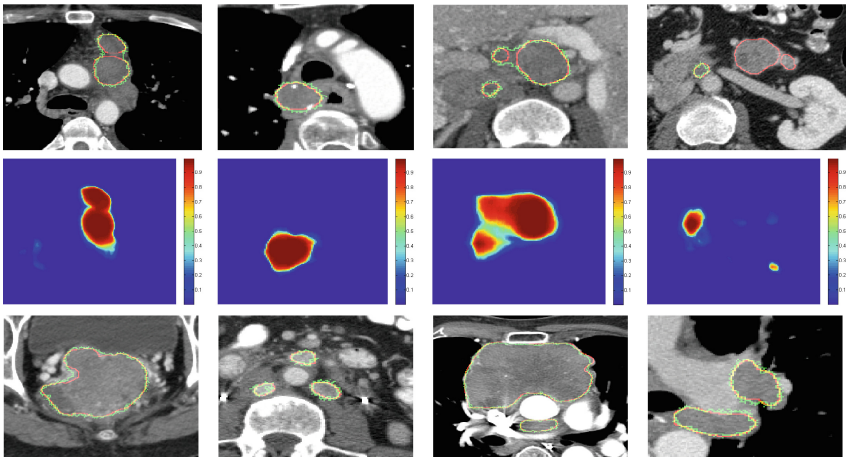
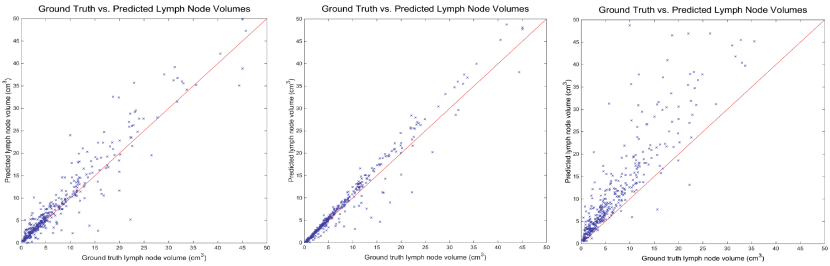


Fig. 2. Examples of LN CT image segmentation. **Top, Bottom**: CT images with ground truth (red) and BNF segmented (green) boundaries. **Center**: HNN-A LN probability maps. **CT images 1–3, 5–8** depict successful segmentation results. **CT image + map 4** present an unsuccessful case.

Nvidia Tesla K40 GPU. The system requires 5 h 40 min for training (30K iterations), and 7 min 43 s for testing (computation of all contour and area maps). BNF is run on MATLAB 2013a ($\sim$ 9 h). dCRF and GC are run using C++ ($\sim$ 3 h; $\sim$ 12 h). The hyperparameters of dCRF and GC (described in [3,9]) are optimized using a randomized search.

BNF yields the highest quantitative segmentation results. Its mean and std per-volume DSC is $82.1 \pm 9.6\%$, above HNN-A's 73.0 ± 17.6 %, dCRF's $69.0 \pm 22.0\%$, and GC's $67.3 \pm 16.8\%$ (cf. Table 1). Additionally, it decreases HNN-A's mean RVD from 32.2 % to 13.7 %. Meanwhile, dCRF marginally decreases the RVD to 29.6 %, and GC increases it to 86.5 %. Figure 2 contains 8 examples of LNC segmentation using BNF. The plots in Fig. 3 compare segmentation and ground truth LNC volume values (in cm^3).

Due to the HNN's deeply nested architecture and auxiliary loss functions on multi-scale side-output layers, the quality of HNN-A segmentation is already high. dCRF's decline in performance relative to HNN-A may be partly attributed to its CT intensity contrast-sensitive pairwise CRF term, defined in [9]. The appearance kernel in this function implies that neighboring pixels of similar intensity are likely to belong to the same class. Though highly relevant in natural images, this idea cannot be extended to TA CT images, in which LN and background pixels may have similar intensities. GC, however, uses a pairwise term defined by the HNN-C boundary cues. Its lower performance may be attributed to its usage of an L_1 norm for the CRF energy minimization. Unlike the L_2 norm, used by dCRF and BNF, the L_1 norm may yield multiple and/or unstable solutions, thus compromising the final segmentation accuracy. Like GC, BNF omits all intensity information from the energy function. However, it utilizes the



(a) HNN-A (b) Boundary Neural Fields (c) Graph Cuts

Fig. 3. Comparison between ground truth and predicted LN volumes.

boundary cues from HNN-C in both the unary and pairwise terms. Hence, one may infer that an emphasis on LN boundary information is crucial to overcome the complexity of TA LNC CT image segmentation. Additionally, the L_2 norm may increase the accuracy of the final segmentation result.

Comparison to previous work: The proposed bottom-up LN segmentation scheme equals or surpasses previous top-down methods in performance, though applied to a more challenging body region. For head-and-neck segmentation, [17] obtains a mean DSC of 73.0 % on five CT scans. The neck study in [6] yields relative volumetric segmentation error ratios of 38.88 % − 51.75 % for five CT scans. The discriminative learning approach in [7] only reports LN detection results on 54 chest CT scans and no information on LN segmentation accuracy. The more comprehensive study from [1] achieves a mean DSC of 80.0 % ± 12.6 % for 308 axillary LNs and 76.0 % ± 12.7 % for 455 pelvic+abdominal LNs, from a dataset of 101 CT cases.

Table 1. Evaluation of segmentation accuracy: HNN-A, BNF, dCRF, and GC

Method	Evaluation Metric		
	Mean DSC (%)	Mean IoU (%)	Mean RVD (%)
HNN-A	73.0 ± 17.6	60.1 ± 18.8	32.2 ± 46.3
BNF	82.1 ± 9.6	70.6 ± 11.9	13.7 ± 13.1
dCRF	69.0 ± 22.0	56.2 ± 21.6	29.6 ± 45.4
GC	67.3 ± 16.8	53.0 ± 17.9	86.5 ± 107.6

4 Conclusion

To solve a challenging problem with high clinical relevance – automatic segmentation and volume measurement of TA LNCs in CT images, our method integrates HNN learning in both LN appearance and contour channels and exploits different structured optimization methods. BNF (combining HNN-A and

HNN-C via a sparse matrix representation) is the most accurate segmentation scheme. BNF's mean RVD of $13.7 \pm 13.1\%$ is promising for the development of LN imaging biomarkers based on volumetric measurements, which may lay the groundwork for improved RECIST LN measurements.

Acknowledgments. This work was supported by the Intramural Research Program at the NIH Clinical Center.

References

1. Barbu, A., Suehling, M., Xu, X., Liu, D., Zhou, S.K., Comaniciu, D.: Automatic detection and segmentation of lymph nodes from CT data. IEEE Trans. Med. Imaging (2012)
2. Bertasius, G., Shi, J., Torresani, L.: Semantic segmentation with boundary neural fields. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
3. Boykov, Y., Funka-Lea, G.: Graph cuts and efficient ND image segmentation. IJCV (2006)
4. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Trans. Pat. Ana. Mach. Intel. (2004)
5. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: ICLR (2015)
6. Dornheim, J., Seim, H., Preim, B., Hertel, I., Strauss, G.: Segmentation of neck lymph nodes in CT datasets with stable 3D mass-spring models. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4191, pp. 904–911. Springer, Heidelberg (2006). doi:10.1007/11866763_111
7. Feulnera, J., Zhou, S., Hammond, M., Horneggera, J., Comaniciu, D.: Lymph node detection and segmentation in chest CT data using discriminative learning and a spatial prior. In: Medical Image Analysis, pp. 254–270 (2013)
8. Jia, Y.: Caffe: an open source convolutional architecture for fast feature embedding (2013). http://goo.gl/Fo9YO8
9. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with Gaussian edge potentials. In: NIPS (2012)
10. Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: AISTATS (2015)
11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE CVPR, pp. 3431–3440 (2015)
12. Martin, D., Fowlkes, C., Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Trans. Pat. Ana. Mach. Intel. (2004)
13. Roth, H., Lu, L., Farag, A., Sohn, A., Summers, R.: Spatial aggregation of holistically-nested networks for automated pancreas segmentation. In: Ourselin, S., Wells, W.M., Joskowicz, L., Sabuncu, M., Unal, G. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 451–459. Springer, Heidelberg (2016)
14. Schwartz, L., Bogaerts, J., Ford, R., Shankar, L., Therasse, P., Gwyther, S., Eisenhauer, E.: Evaluation of lymph nodes with recist 1.1. Euro. J. Cancer **45**(2), 261–267 (2009)

15. Seff, A., Lu, L., Barbu, A., Roth, H., Shin, H.-C., Summers, R.M.: Leveraging mid-level semantic boundary cues for automated lymph node detection. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9350, pp. 53–61. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24571-3_7
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2014)
17. Stapleford, L., Lawson, J., et al.: Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer. Int. J. Rad. Onc. Bio. Phys. (2010)
18. Xie, S., Tu, Z.: Holistically-nested edge detection. In: IEEE ICCV, pp. 1395–1403 (2015)