Feasibility of End-To-End Trainable Two-Stage U-Net for Detection of Axillary Lymph Nodes in Contrast-Enhanced CT Based Scans on Sparse Annotations

Hidir Cem Altun[a,b], Grzegorz Chlebus[a,c], Colin Jacobs[c], Hans Meine[b,d], Bram van Ginneken[c,a], and Horst K. Hahn[a,b]

[a]Fraunhofer MEVIS, Institute for Digital Medicine, Bremen/DE
[b]Jacobs University, Bremen/DE
[c]Radboud University Medical Center Nijmegen/NL
[d]University of Bremen/DE

ABSTRACT

Manual detection of lymph nodes by a radiologist is time-consuming, error-prone and suffers from inter-observer variability. We propose a mostly generic computer-aided detection system, which can be trained in an end-to-end fashion, to automatically detect axillary lymph nodes using state of the art fully convolutional neural networks. We aim at a system that can be easily transferred to other body regions such as the mediastinal region. Our pipeline is a two-stage approach, where first a volume of interest (VOI) (axillary region) is localized and then axillary lymph node detection is performed within the VOI. Training was done on 58 CT volumes from 36 patients comprising 300 axillary lymph nodes. On our test dataset, consisting of 75 axillary lymph nodes in the size range 5–10 mm and 17 larger than 10 mm from 30 different patients, we achieved a sensitivity of 83% with 6.7 FPs per volume on average.

Keywords: Lymph Nodes, CNN, U-net, CAD

1. INTRODUCTION

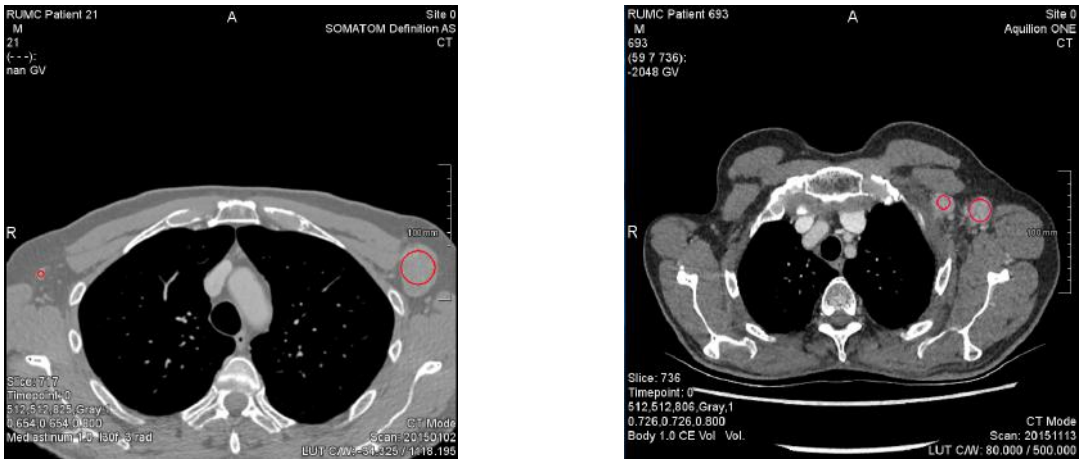
Lymph nodes (LN) are part of the lymphatic system, a network that helps to filter harmful substances and fight against infection.[1] They are essential for monitoring disease progress, therapy success, and diagnosing the disease stage. According to current clinical standards, the detection of enlarged LNs (short-axis diameter $\geq$ 10 mm), which has a higher chance of being malignant,[2] is of major interest. Most commonly, contrast enhanced CT is used for LN imaging. Unfortunately, the annotations of LNs are done manually by a radiologist, which is time-consuming, error-prone and suffers from inter-observer variability. Therefore, a computer-aided detection (CAD) system is needed for speeding up the process and reducing the human error.

Automatic detection of LNs is challenging due to variation in size, shape, orientation, position and low contrast with surroundings, see Figure 1. In the literature there are several works done on CAD systems for LN detection. These works focused on mediastinal,[3–7] abdominal,[3–5] pelvic[8] and axillary[8] LNs. The common pattern in all CAD systems mentioned in this study is that they are multi-stage systems. First VOI proposal and/or LN candidate generation is done. Then the false positives (FP) will be removed via another classifier(s). Roth et al., Seff et al. and, Shin et al. used same publicly available lymph node CT dataset.[9] Roth et al. first obtained LN candidates using,[10,11] the LN candidates are publicly available at,[9] then a 2.5D CNN and random view aggregation was used for false positive reduction.[3] Seff et al. and Shin et al. used same LN candidates as.[3] Instead of training a CNN from scratch, Shin et al. used pre-trained CNN to detect the LNs.[5] On the other hand Seff et al. extracted histogram of oriented gradients (HOG) features from raw CT slices and sketch tokens[12] to train a support vector machine (SVM).[4] Barbu et al. trained multiple classifiers (using 3D Haar features and gradient based features[13]) to detect the LNs.[8] Oda et al. created a VOI by segmenting the lungs.

Further author information: (Send correspondence to H.C.A.)
H.C.A.: E-mail: h.cemaltun1997@gmail.com

Then a 3D U-net is fine tuned to segment the LNs.[7] Lui et al. first created VOI using the lungs and spine. To create LN candidates Lui et al. used spatial, shape and size information to generate LN candidates then SVM committee[14] used to remove false positives.[6]

The proposed CAD systems are not fully end-to-end fashion, especially the VOI proposal step. For example Barbu et al.[8] created fixed size bounding boxes. The upper left corners of the bounding boxes is determined by segmenting region specific organs (lungs, pubic symphysis tip). Our proposed system is a multi-stage system like proposed CAD systems. However, we design our CAD system such that it is transferable to other body regions whit out need of body region specific feature extraction or segmentation of other body parts. A 3D-Unet[15] is used to create a VOI then another 3D U-net detects the lymph node in the proposed VOI, see Figure 2.



(a) LNs that are different size. **(b)** LNs that are blended with the background.

Figure 1: Example of lymph nodes. The CT scans used in this study obtained from RUMC dataset. Red circles indicate lymph node.

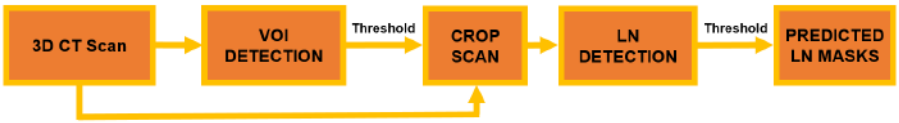


Figure 2: LN detection pipeline.

2. METHODS

2.1 Architecture

3D U-net was employed to train VOI proposal and LN detection networks. Our proposed VOI and LN detection models have the same network architecture. Since the VOI network has some knowledge about CT scans and the VOI annotations are created from the same sparse LN annotations, we tested, whether it is beneficial to use this as a starting point for training the LN detection network.

Due to the small amount of training data, we use dropout before each convolution layer in the bottleneck and up-scaling path and data augmentation (rotating respect to the axial axis, zooming in/out, flipping the sagittal axis and adding Gaussian noise) to cope with overfitting. Additionally, batch normalization is applied after each convolution to improve learning stability. The ReLu activation is used for non-linearity. The final convolution layer consists of 1x1x1 convolution followed by ReLu and Softmax activations, see 3.

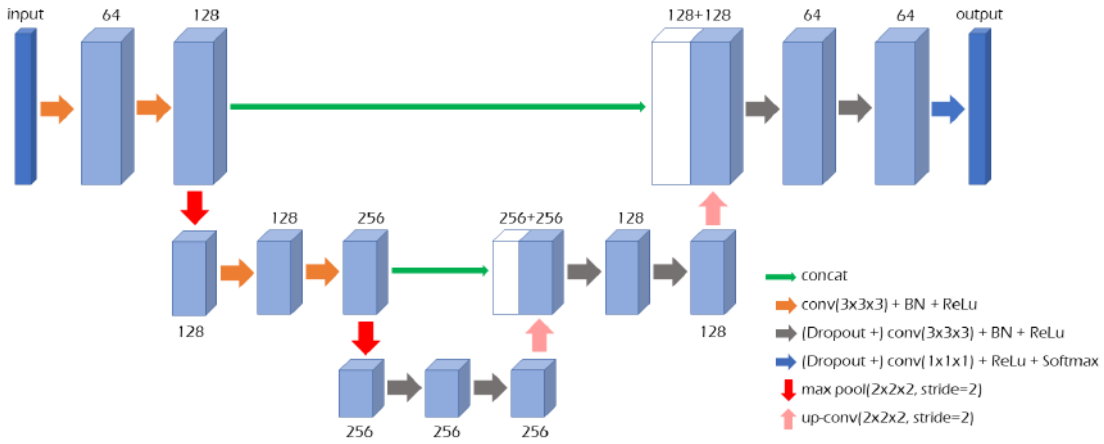
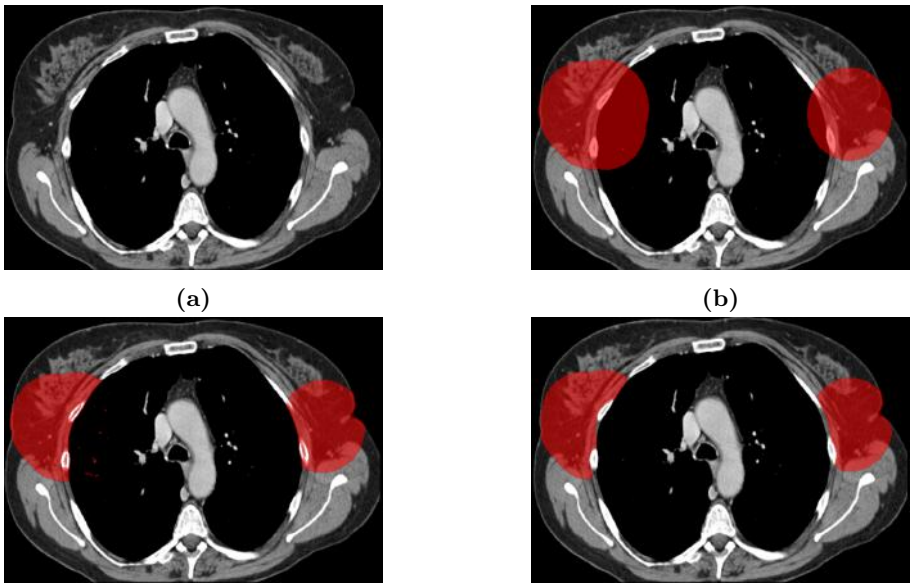


Figure 3: Three resolution level 3D-Unet architecture. Its approximated receptive field is 44x44x44 voxels and total number of trainable parameters are 8,516,034. Each block represents the feature maps and number of channels denoted top/bottom of the each block.

2.2 Volume of Interest (VOI) Localization

The input masks for VOI localization network are solely relying on the LN annotations. First, an ensemble of spheres are drawn around the centers of the annotated LNs with an empirically chosen radius of 5 cm, which is in many cases large enough to cover the axillary region (4b). Then air (<-600 HU) and bones (>300 HU) are filtered via simple thresholding (4c). Finally, to smooth the masks, morphological opening with a spherical 11x11x11 kernel is applied (4d). The first-stage network is applied to a downsampled version of the CT volumes (isotropic voxel size 3x3x3 mm^3, Lanczos-3 resampling).



(a) **(b)**

(c) **(d)**

Figure 4: Pre-processing steps of VOI mask.

To avoid class imbalance, dice loss[?] used as the objective function, which is used as default objective function during the study, and the batches are created to comprise 50% background and 50% foreground. The dataset is divided into the three groups for testing (51 scans), training (36 scans), and validation (12 scans). Some patients had multiple CT scans thus, splitting of the dataset is done patient-wise.

The best model is picked with the Jaccard index (intersection over union score). The Adam optimizer with initial learning rate 1e−3 is used for training and the learning rate halved if no improvement obtained after 10 validation step. The training is ended, if no improvement obtained after 50 validation step, every 100 iterations a validation step computed. During the training, the input patch size was 72x72x72 voxels/216x216x216 mm^3 and the output patch was 32x32x32 voxels/96x96x96 mm^3.

2.3 Lymph Node (LN) Detection

Again, we construct spheres around LN annotations (using center-point & short-axis diameter) to yield a voxel mask volume to be considered foreground when training our U-net. These spheres, however, won't cover all parts of the respective LNs, which typically deviate from the spherical shape. As a consequence, it would be problematic to consider all voxels outside of these masks as background. Instead, we ignore all voxels in a shell around these sphere during training and validation by setting their weights to zero. The shell is constructed through an external morphological gradient by dilating the combined sphere voxel mask for each volume using an empirically chosen 11x11x11 octagon type kernel (cf. Figure 5). We referred this strategy as shell masking.

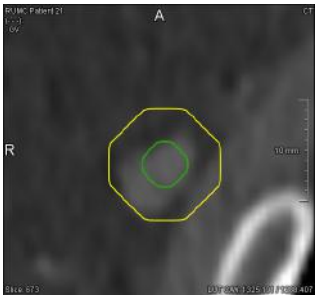


Figure 5: Shell Masking. The area between the inner circle (LN) and the outer circle loss are not computed.

In our experiments, the network was biased towards large LNs since finding a large LN makes a stronger contribution than a small LN due to the voxel-wise classification. In addition, finding small LNs is harder because they have a larger proportion of boundary voxels. To enhance the learning capacity of our network for small LNs, we computed normalized weights, dividing each LN foreground mask with a factor proportional to the size of that LN. Since non-zero weights doesn't have an effect on dice loss, except the zero weights, categorical cross entropy loss used for training the network, see Appendix A.

The same training strategies and hyperparameters are used as for VOI localization. The CT volumes downsampled to 1x1x1 mm^3 using Lanczos-3 resampling, which makes the networks approximated receptive field of the network is 44x44x44 mm^3 which is sufficient enough to see all the LNs. Due to resampling one LN annotation is lost, Figure 6. The dataset is divided into three groups for testing (30 scans), training (58 scans) and validation (10 scans). Test scans are picked from VOI localization test scans in order to avoid an information propagation bias.

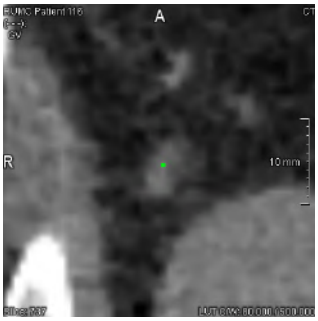


Figure 6: Annotation of LN lost due to resampling. The image was taken before resampling.

2.4 Detection Criteria

We use detection criteria inspired by,[7] considering a LN to be successfully detected if its sphere mask intersects with the network prediction (see Equation 1). Additionally, we propose an alternative detection criterion, considering a LN to be detected if either the center of the LN annotation is contained in the prediction or the center of mass of a connected blob in our prediction is contained in the sphere mask (see Equation 2).

$$|G \cap P| \geq 1 \tag{1}$$

$$CoM(G) \in P \parallel CoM(P) \in G \tag{2}$$

where: G is a reference sphere mask
P is a connected component of the thresholded network prediction
$CoM(A)$ is center of mass of object A

3. EXPERIMENTS AND RESULTS

3.1 Dataset

The dataset used in this study provided by Radboud University Medical Center, Nijmegen, Netherlands. It consists of 99 full body CT scans from 76 different patients and the total number of annotated axillary LNs is 492. Most of the annotated LNs are not medically relevant, see Figure 7. In our dataset, LNs that are smaller than 10 mm was (n=393, i.e. 81%) included, because we believe that not enlarged LNs can be clinically relevant but are often ignored due to an efficiency tradeoff necessary in manual reading. The total number of enlarged/medically relevant LNs ($\geq$ 10 mm in shortest-axis) is 92 and 28 of them are measurable lymph nodes ($\geq$ 15 mm in shortest-axis) after cleaning the dataset.

We cleaned the dataset where seven LNs are removed. There are five LNs that have a negative diameter, another LN is identified no signs of being a LN after visual assessment (Figure 8). Additionally, one LN, which is abnormally large (9 cm diameter), is also removed. The CT slice thickness varies from 0.7 mm to 1.5 mm and slice resolution varies from 0.6 mm to 1 mm.

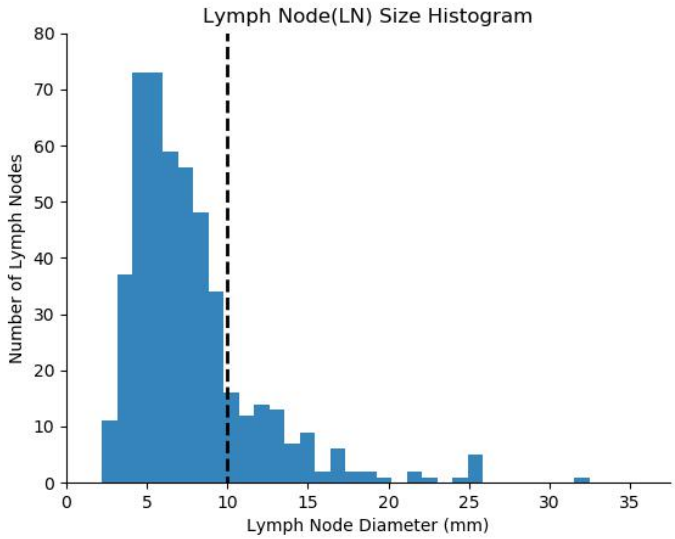


Figure 7: Histogram of lymph node size in mm on the curated version of the dataset. The dashed line indicates the 10 mm mark.

3.2 VOI Localization Results

Measuring the performance of the VOI network is not a trivial task. Its main task is to capture the complete region where the LNs occurred (here: axillary region) and at the same time to reduce the search space as much as possible. We used the following four metrics to determine the trade-off between LN detection rate and VOI size: sensitivity/true positive rate (TPR) per scan using Equation 2; Jaccard index per scan; the number of connected components; the total number of missed LNs.

We conducted our experiments with and without data augmentation. For the data augmentation patches rotated respect to axial axis it is picked from $\mathcal{N}(0, 5)$, zooming in/out in all three axis the zooming factor is chosen from $\mathcal{U}(-0.2, 0.2)$, the sagittal axis flipped and a small noise, $\mathcal{N}(0, 5)$, added to image. As shown

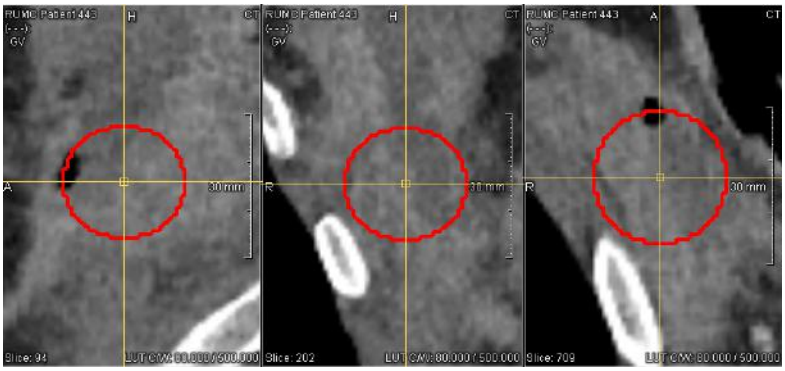


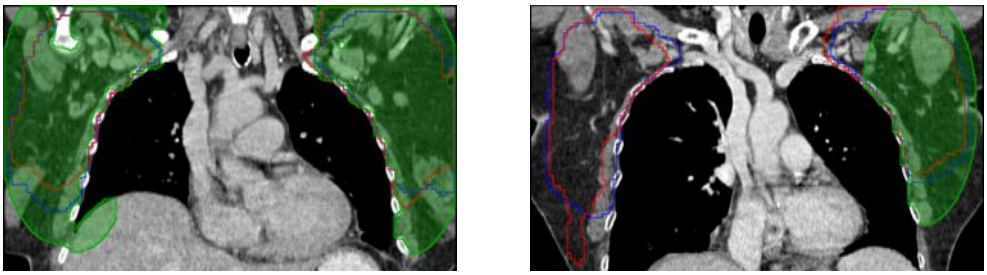
Figure 8: Falsely annotated lymph node. The respective location is shown in three planes (left) sagittal, (middle) coronal and (right) axial.

in Table 1, data augmentation gives a tiny performance boost. However, no statistical significance was found when computing the Wilcoxon signed-rank test for the first three measurements (p-value > 0.05, cf. Table 1). Thus, the data augmentation didn't improve the network, in this case. However, we used network with data augmentation as VOI network since, it missed less LNs and got higher Jaccard index per scan.

Note that the relatively low Jaccard index of 0.46 might be acceptable here. Even though the training VOI masks were noisy/incomplete (leaking into the organs Figure 9a, missing contralateral axillary region Figure 9b, etc.) the network achieved a sensitivity of almost 100% on the independent test cases.

	TPR/Scan	**Jaccard/Scan**	**# of Blobs**	**Miss**
with data aug.	$\mathbf{0.97 \pm 0.13}$	$\mathbf{0.46 \pm 0.17}$	3.66 ± 2.09	**6**
w/o data aug.	0.96 ± 0.14	0.44 ± 0.17	$\mathbf{2.54 \pm 0.86}$	10
p-value	0.17971	0.16015	0.00024	–

Table 1: Results of VOI localization with and without data augmentation. Networks' output thresholded (≥ 0.5) to create the VOI mask. The last row is the result of the Wilcoxon signed-rank test.



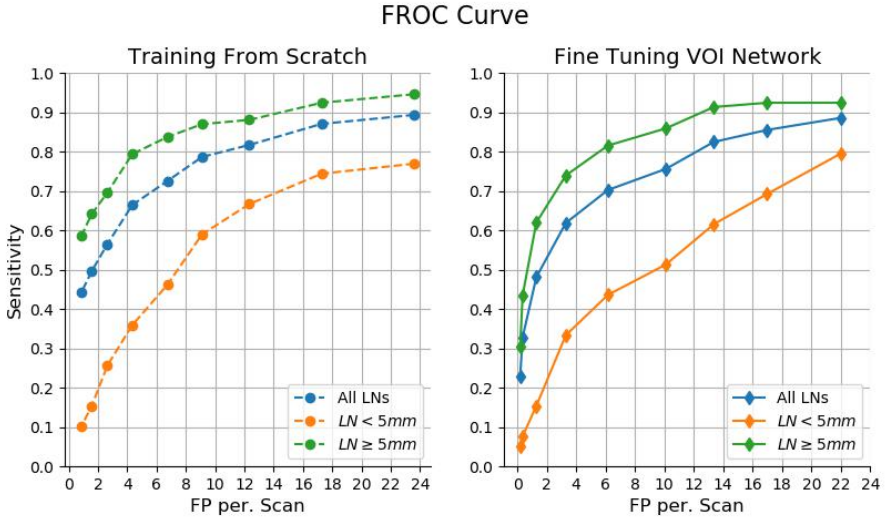
(a) **(b)**

Figure 9: VOI predictions of networks trained with data augmentation (blue line), without data augmentation (red line), and the reference region (green) for two test cases. Predictions are resampled to the scans' original voxel size and for visualization shown in the coronal plane.

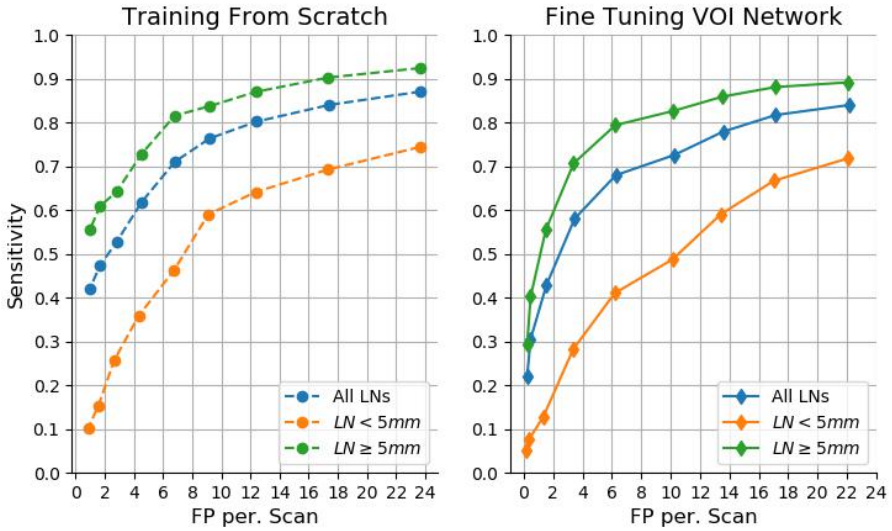
3.3 LN Detection Results

With small datasets, fine tuning of CNN hyperparameters / transfer learning is a common option. Here, we conducted one experiment by training the U-net from scratch and one by fine-tuning the existing VOI network towards LN detection. Note that we have empirically chosen a lower dropout rate (p_d) of 0.25 for random initialization compared to 0.5 when taking the VOI network as starting configuration. Therefore, we cannot know which change was improve/worsen our results since two things changed at once.

We calculated FROC curves to investigate the trade-off between sensitivity and the number of false positive detections per scan, see Figure 10. Thereby, LNs are split into two groups (smaller/larger than 5 mm) to measure how well the networks detect LNs with different size, see Figure 11). For LNs larger than 10 mm, FROC curves weren't calculated due to their small number (n=17). Some exemplary test results of common false negatives (FN), false positives (FP), and true positives (TP) are shown in Figure 12, Figure 13, and Figure 14. The prediction confidence indicated in these figures represent the maximum network output within the predicted blob for TP/FP cases and mean network output under the sphere mask for FN cases.



(a) Detection Criteria 1, Equation 1



(b) Detection Criteria 2, Equation 2

Figure 10: FROC curve of fine-tuning with p_d=0.5 and random initialization with p_d=0.25.

4. DISCUSSION AND CONCLUSION

In this study, a two-stage CAD system for LN detection was introduced by first proposing a VOI and then detecting LNs within the proposed VOI. Both stages were performed using an identical 3D U-net architecture, applied to two different resolutions and with two different annotation processing schemes. An advantage of the proposed CAD system compared to the state-of-the-art is that in each step end-to-end training is done, such

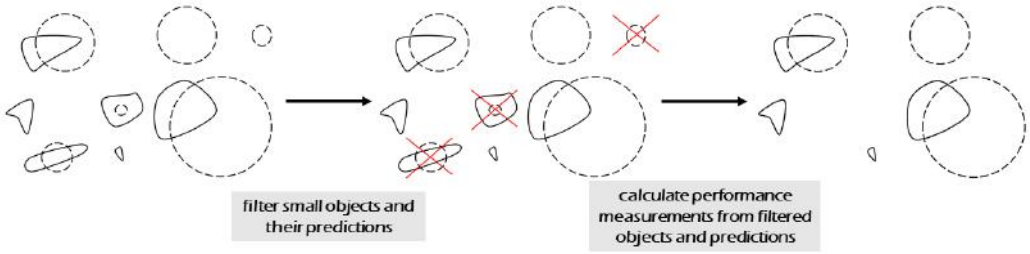


Figure 11: An example of how filtering objects with different sizes is done. First we filter the objects (dashed) smaller/larger than some threshold and their prediction if exist any (solid). Then we calculate the performance measurements, in our case detection rate and number of false positives per scan, from filtered objects and their predictions.

Method	Target Area	# of Cases	TPR	FPs/vol
Roth et al.[3]	Abdomen	86	83/90%	3/6
Shin et al.[5]	Abdomen	86	70%	3
Seff et al.[4]	Abdomen	86	78/89%	3/6
Barbu et al.[8]	Pelvic & Abdomen	54	80%	3.2
Barbu et al.[8]	Axillary	131	83%	1
Oda et al.[7]	Mediastinal	45	88.8/92.6%	8/10.6
Seff A. et al.[4]	Mediastinal	90	78/88%	3/6
Shin et al.[5]	Mediastinal	90	85%	3
Roth et al.[3]	Mediastinal	90	70/84%	3/6
Lui et al.[6]	Mediastinal	70	88%	8
Random init. (eq. 1)	Axillary	99	79/83%	4.3/6.7
Random init. (eq. 2)	Axillary	99	72/81%	4.5/6.8
Fine tuning (eq. 1)	Axillary	99	73/81%	3.3/6.1
Fine tuning (eq. 2)	Axillary	99	70/79%	3.4/6.2

Table 2: Results of our LN detection pipeline compared to other systems from literature.

that only the images and the sparse LN annotations are required as training material. Previous methods were multi-stage systems like ours, however, with only parts of the steps implemented in an end-to-end fashion. Most frequently, the VOI localization included some concepts specific for the respective body region of interest such as lung segmentation (e.g. Oda et al.[7]).

Comparing our results directly with state of the art (Table 2) would be not accurate since we are using different dataset, detection criteria and, detecting LNs from different regions. Additionally, when comparing our results to the literature, note that the median LN diameter in our dataset was only approx. 7 mm. In addition, a close inspection of the FPs and FNs raises questions whether our reference annotations might be erroneous in some of these cases. If this would be true, the measured accuracy would further increase.

In the future, we aim at extending our work in order to demonstrate that our generic approach will be transferable to other body regions with no or minimal methodological alterations. Additionally, we need to pick a new detection criteria since the current ones don't have any constraint for prediction size, i.e. predicting all the CT scan as LN, will give a perfect detection score. A suitable detection criteria might be Jaccard index but calculating the FROC curve would be complicated since we have to draw the FROC curve respect to Jaccard index and network confidence. Finaly, we see the two-stage approach comes with a dependency drawback. LNs will only be detected if within the proposed VOI from the first stage. Our results indicate that a single stage network could eventually solve the LN detection problem. We hypothesize that a similar U-net with a receptive field as large as for our stage-one VOI localization network (132x132x132 mm^3) or more, if computationally feasible, could achieve even higher levels of accuracy. Finally, our method contains a small number of ad-hoc decisions (sphere radii, thresholds), which we plan to validate or generalize in the future.

APPENDIX A. THE EFFECT OF NORMALIZED WEIGHTS

Categorical Cross Entropy Lets consider the loss of single LN without any weights,

$$CCE = -\sum_{i \in LN} log(\hat{y}_i + \epsilon) \tag{3}$$

Assume the network's prediction is same for each voxel in LN then Equation 3 can be re-write as,

$$= -|LN| \cdot C \tag{4}$$

Applying normalized weights to Equation 4,

$$= -\frac{|LN_{mean}|}{|LN|}|LN| \cdot C \Rightarrow -|LN_{mean}| \cdot C \tag{5}$$

Thus, each LN have the same contribution to loss regardless of their size.

Dice Loss Weighted dice loss defined as,

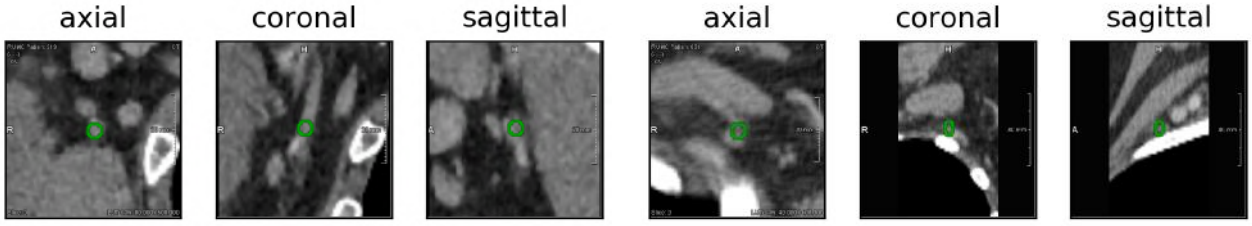
$$\text{WeightedDice}(y, \hat{y}, w) = 1 - \frac{2\sum_{i=0}^{N-1} w_i \cdot y_i \cdot \hat{y}_i + \epsilon}{\sum_{i=0}^{N-1} w_i(y_i + \hat{y}_i) + \epsilon} \tag{6}$$

Since ϵ is a small number, it is used for numerical stability, it can be ignored for non zero weights. So, we can re-write Equation 6,

$$= 1 - \frac{2\sum_{i=0}^{N-1} w_i \cdot y_i \cdot \hat{y}_i}{\sum_{i=0}^{N-1} w_i(y_i + \hat{y}_i)} \Rightarrow 1 - \frac{2\sum_{i=0}^{N-1} y_i \cdot \hat{y}_i}{\sum_{i=0}^{N-1} y_i + \hat{y}_i} \tag{7}$$

Hence, non zero weights have no effect on dice loss.

APPENDIX B. LN DETECTION EXAMPLARY RESULTS

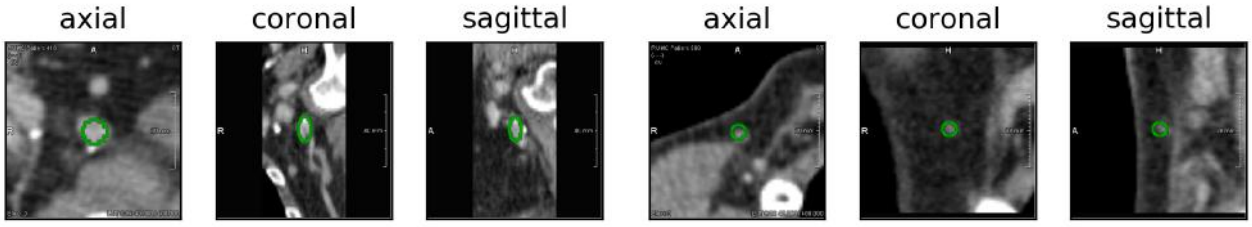


Fine Tuning - 0.0003

Random Init - 0.0001

Fine Tuning - 0.0003

Random Init - 0.0004



Fine Tuning - 0.0000

Random Init - 0.0015

Fine Tuning - 0.0001

Random Init - 0.0001

Figure 12: Exemplary missed LNs (false negatives) by both network. Both models' confidence written at the bottom of each figure.

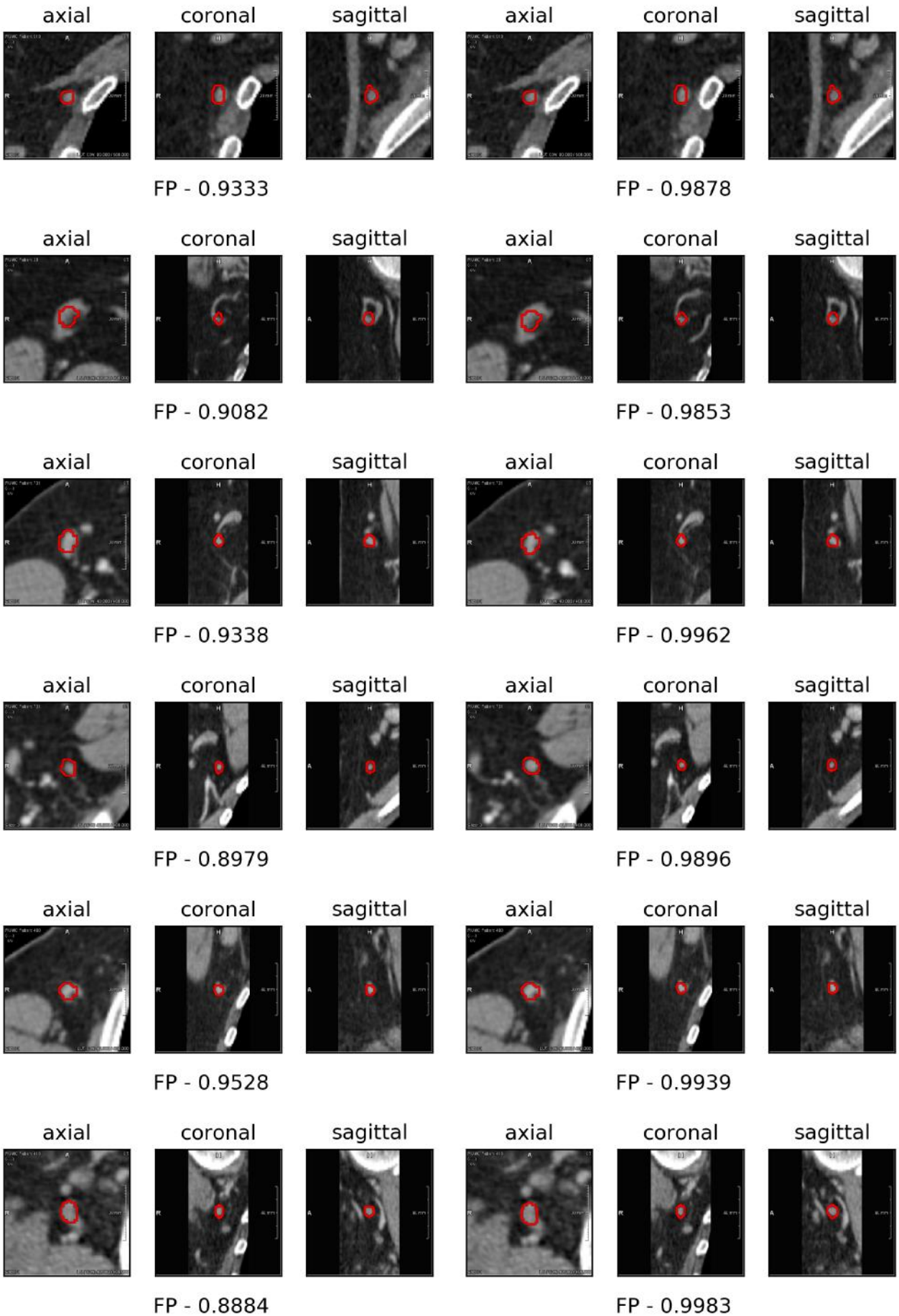


Figure 13: Common false positives that are detected by both network with high confidence (¿0.88). The confidence value of each network written at the bottom of the figures.

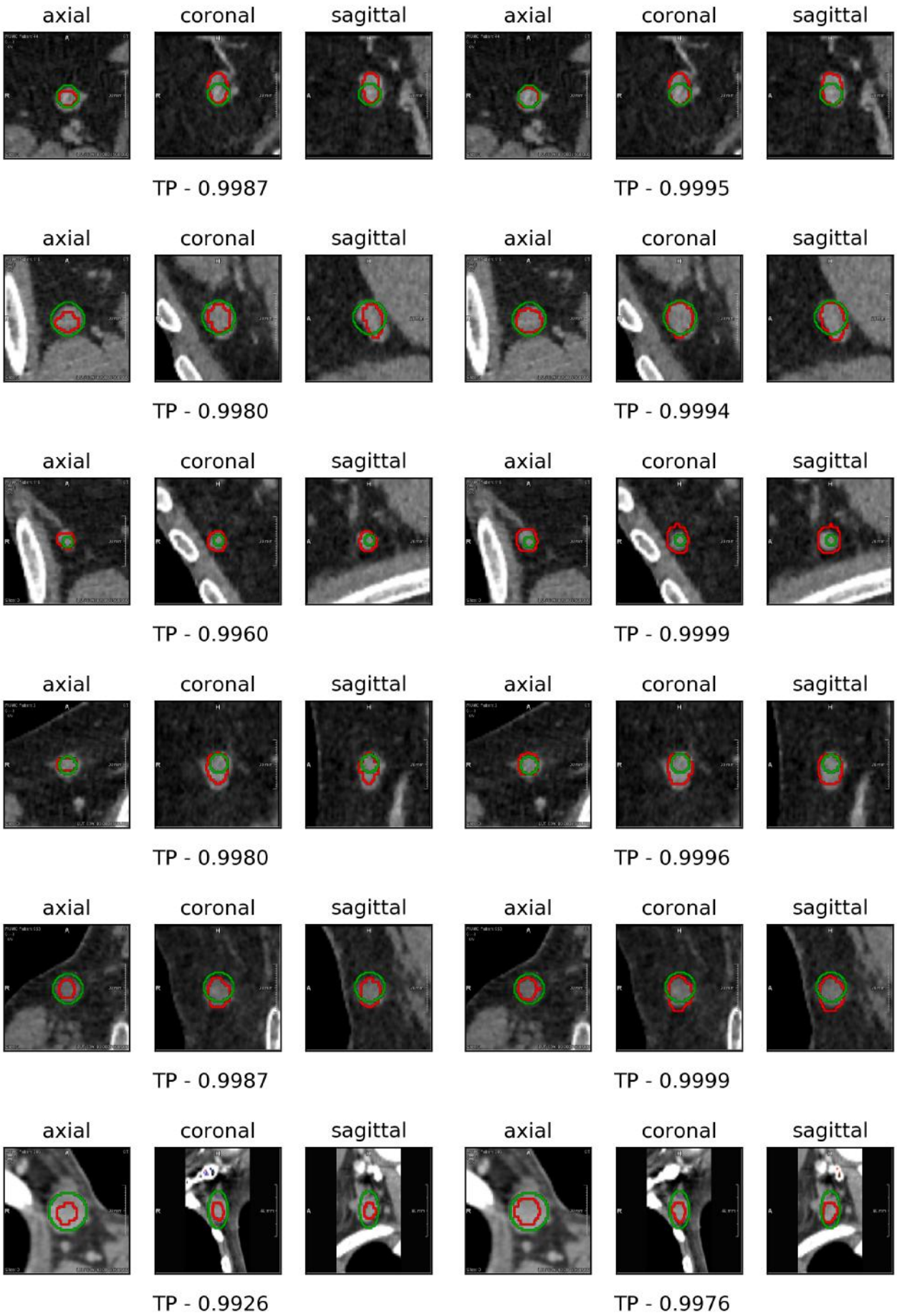


Figure 14: Common true positives that are detected by both network with high confidence (¿0.88). The confidence value of each network written at the bottom of the figures.

REFERENCES

[1] Zimmer, A. K., "Lymphatic System: Facts, Functions & Diseases." Live Science, 25 October 2015 https://www.livescience.com/26983-lymphatic-system.html. (Accessed: 15 January 2018).

[2] Schwartz, L. H., Bogaerts, J., Ford, R., Shankar, L., Therasse, P., Gwyther, S., and Eisenhauer, E. A., "Evaluation of lymph nodes with recist 1.1," *European Journal of Cancer* (2009).

[3] Roth, H. R., Le Lu, A. S., Cherry, K. M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., and Summers, R. M., "A new 2.5d representation for lymph node detection using random sets of deep convolutional neural network observations," *Medical Image Computing and Computer-Assisted Intervention* (2018).

[4] Seff, A., Lu, L., Barbu, A., Roth, H., Shin, H.-C., and Summers, R. M., "Leveraging mid-level semantic boundary cues for automated lymph node detection," *Medical Image Computing and Computer-Assisted Intervention* (2015).

[5] Shin, H., Roth, H., Gao, M., Lu, L., Xu, Z., Yao, J., Mollura, D., and Summers, R., "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging* (2016).

[6] Liu, J., Hoffman, J., Zhao, J., Yao, J., Lu, L., Kim, L., Turkbey, E. B., and Summers, R. M., "Mediastinal lymph node detection and station mapping on chest ct using spatial priors and random forest," *Medical Physics* (2016).

[7] Oda, H., Roth, H. R., Bhatia, K. K., Oda, M., Kitasaka, T., Iwano, S., Homma, H., Takabatake, H., Mori, M., Natori, H., Schnabel, J. A., and Mori, K., "Dense volumetric detection and segmentation of mediastinal lymph nodes in chest ct images," *SPIE, Medical Imaging: Computer-Aided Diagnosis* (2018).

[8] Barbu, A., Suehling, M., Xu, X., Liu, D., Zhou, S., and Comaniciu, D., "Automatic detection and segmentation of lymph nodes from ct data," *IEEE Transactions on Medical Imaging* (2012).

[9] Holger, R., Lu, L., Seff, A., Cherry, K. M., Hoffman, J., Wang, S., and Summer, R. M., "A new 2.5 d representation for lymph node detection in ct. the cancer imaging archive." (2015) http://doi.org/10.7937/K9/TCIA.2015.AQIIDCNM.

[10] Liu, J., Zhao, J., Hoffman, J., Yao, J., Zhang, W., Turkbey, E. B., Wang, S., Kim, C., and Summers, R. M., "Mediastinal lymph node detection on thoracic ct scans using spatial prior from multi-atlas label fusion," *SPIE, Medical Imaging: Computer-Aided Diagnosis* (2014).

[11] Cherry, K. M., Wang, S., Turkbey, E. B., and Summers, R. M., "Abdominal lymphadenopathy detection using random forest," *SPIE, Medical Imaging: Computer-Aided Diagnosis* (2014).

[12] Lim, J. J., Zitnick, C. L., and Dollár, P., "Sketch tokens: A learned mid-level representation for contour and object detection," *IEEE Conference on Computer Vision and Pattern Recognition* (2013).

[13] Zheng, Y., Barbu, A., amd M. Scheuering, B. G., and Comaniciu, D., "Medical image computing and computer-assisted intervention (miccai)," *3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation* (2016).

[14] Yao, J., Summers, R. M., and Hara, A. K., "Optimizing the support vector machines (svm) committee configuration in a colonic polyp cad system," *SPIE, Medical Imaging: Computer-Aided Diagnosis* (2005).

[15] Çiçek, O., Abdulkadir, A., Lienkamp, S., Brox, T., and Ronneberger, O., "3d u-net: Learning dense volumetric segmentation from sparse annotation," *Medical Image Computing and Computer-Assisted Intervention* (2016).