DON'T MISS OUR SPECIAL SESSION AT ASTRO

Performing QA Remotely in the Age of COVID

October 26, 11:45 AM EST

At home. On site. **In sync.**

SunCHECK™ enables complete, collaborative **remote QA coverage** for COVID-19 & beyond.

Click to explore:

- Advantages of a centralized Patient & Machine QA solution
- How SunCHECK eased the transition to remote work for users worldwide
- Three new ways we're simplifying Platform adoption

Go to: **sunnuclear.com/getprepared**

SUN NUCLEAR corporation

Patient Safety Starts Here

AAR-LN-DQ: Automatic anatomy recognition based disease quantification in thoracic lymph node zones via FDG PET/CT images without Nodal Delineation

Guoping Xu
School of Electronic Information and Communications, Huazhong University of Science and technology, Wuhan, Hubei 430074, China
Medical Image Processing Group, Department of Radiology, University of Pennsylvania, 602 Goddard building, 3710 Hamilton Walk, Philadelphia, PA 19104, USA

Jayaram K. Udupa[a)#], Yubing Tong, and Dewey Odhner
Medical Image Processing Group, Department of Radiology, University of Pennsylvania, 602 Goddard building, 3710 Hamilton Walk, Philadelphia, PA 19104, USA

Hanqiang Cao
School of Electronic Information and Communications, Huazhong University of Science and technology, Wuhan, Hubei 430074, China

Drew A. Torigian
Medical Image Processing Group, Department of Radiology, University of Pennsylvania, 602 Goddard building, 3710 Hamilton Walk, Philadelphia, PA 19104, USA
Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, PA 19104, USA

(Received 17 January 2020; revised 22 April 2020; accepted for publication 8 May 2020; published 15 June 2020)

Purpose: The derivation of quantitative information from medical images in a practical manner is essential for quantitative radiology (QR) to become a clinical reality, but still faces a major hurdle because of image segmentation challenges. With the goal of performing disease quantification in lymph node (LN) stations without explicit nodal delineation, this paper presents a novel approach for disease quantification (DQ) by automatic recognition of LN zones and detection of malignant lymph nodes within thoracic LN zones via positron emission tomography/computed tomography (PET/CT) images. Named AAR-LN-DQ, this approach decouples DQ methods from explicit nodal segmentation via an LN recognition strategy involving a novel globular filter and a deep neural network called SegNet.

Method: The methodology consists of four main steps: (a) Building lymph node zone models by automatic anatomy recognition (AAR) method. It incorporates novel aspects of model building that relate to finding an optimal hierarchy for organs and lymph node zones in the thorax. (b) Recognizing lymph node zones by the built lymph node models. (c) Detecting pathologic LNs in the recognized zones by using a novel globular filter (g-filter) and a multi-level support vector machine (SVM) classifier. Here, we make use of the general globular shape of LNs to first localize them and then use a multi-level SVM classifier to identify pathologic LNs from among the LNs localized by the g-filter. Alternatively, we designed a deep neural network called SegNet which is trained to directly recognize pathologic nodes within AAR localized LN zones. (d) Disease quantification based on identified pathologic LNs within localized zones. A fuzzy disease map is devised to express the degree of disease burden at each voxel within the identified LNs to simultaneously handle several uncertain phenomena such as PET partial volume effects, uncertainty in localization of LNs, and gradation of disease content at the voxel level. We focused on the task of disease quantification in patients with lymphoma based on PET/CT acquisitions and devised a method of evaluation. Model building was carried out using 42 near-normal patient datasets via contrast-enhanced CT examinations of their thorax. PET/CT datasets from an additional 63 lymphoma patients were utilized for evaluating the AAR-LN-DQ methodology. We assess the accuracy of the three main processes involved in AAR-LN-DQ via fivefold cross validation: lymph node zone recognition, abnormal lymph node localization, and disease quantification.

Results: The recognition and scale error for LN zones were 12.28 mm $\pm$ 1.99 and 0.94 $\pm$ 0.02, respectively, on normal CT datasets. On abnormal PET/CT datasets, the sensitivity and specificity of pathologic LN recognition were 84.1% $\pm$ 0.115 and 98.5% $\pm$ 0.003, respectively, for the g-filter-SVM strategy, and 91.3% $\pm$ 0.110 and 96.1% $\pm$ 0.016, respectively, for the SegNet method. Finally, the mean absolute percent errors for disease quantification of the recognized abnormal LNs were 8% $\pm$ 0.09 and 14% $\pm$ 0.10 for the g-filter-SVM method and the best SegNet strategy, respectively.

Conclusions: Accurate disease quantification on PET/CT images without performing explicit delineation of lymph nodes is feasible following lymph node zone and pathologic LN localization. It is very useful to perform LN zone recognition by AAR as this step can cover most (95.8%) of the

abnormal LNs and drastically reduce the regions to search for abnormal LNs. This also improves the specificity of deep networks such as SegNet significantly. It is possible to utilize general shape information about LNs such as their globular nature via g-filter and to arrive at high recognition rates for abnormal LNs in conjunction with a traditional classifier such as SVM. Finally, the disease map concept is effective for estimating disease burden, irrespective of how the LNs are identified, to handle various uncertainties without having to address them explicitly one by one. © *2020 American Association of Physicists in Medicine* [https://doi.org/10.1002/mp.14240]

Key words: automatic anatomy recognition (AAR), disease quantification (DQ), FDG-PET/CT, thoracic lymph node zones

1. INTRODUCTION

1.A. Background

Accurate assessment of lymph node (LN) involvement via positron emission tomography/computed tomography (PET/CT) plays an important role in the clinical diagnosis, staging, treatment planning, treatment response assessment, and outcome prediction of patients with cancer.[1] However, it is not easy to detect LNs due to their obscure boundaries and low contrast with subjacent tissues on CT images. The International Association for the Study of Lung Cancer (IASLC) has defined a standard way of identifying lymph node zones (or stations) in the thorax for describing the anatomical locations of lymph nodal metastases, which is essential for disease staging and potentially for prognostication of patient outcome.[2] While this is helpful in standardizing a means of interpreting and reporting thoracic lymph node disease sites, it still leaves the radiologist with the arduous task of memorizing and following the definitions and finding the zones on CT images manually. At the same time, labeling individual pathologic[1] LNs in clinical radiology practice is performed manually by qualitative visual assessment on CT and PET/CT scans, which is also time consuming and prone to error in the assignment of lymph nodes to particular nodal zones. The ability for automatic localization of nodal zones, zone-wise disease burden estimation, and zone-wise enumeration of the different pathologic nodes has many potential applications in clinical disease staging, response assessment, response prediction, restaging, etc.

One possible approach to quantify disease in LN zones is to segment pathologic LNs individually on CT or combined CT and PET, identify the zone to which each LN belongs, and then quantify total disease in each zone. However, this approach is difficult to realize in a production-mode implementation of disease quantification mainly because segmenting (delineating) individual LNs is very challenging. Alternatively, if we can directly localize (recognize) LN zones on CT and quantify total disease burden in each zone via PET without explicitly delineating nodes, the rather ill-defined problem of delineating individual nodes can be circumvented. Works on recognizing nodal zones directly in CT imagery are rather sparse[3–8] compared to a much larger body of literature on LN delineation (segmentation), noting that[6–8] require detection or segmentation of LNs for zone localization and[3–5] constitute our own early work. Considering the challenges of segmenting LNs, we take a different approach: Recognizing (localizing) LN zones in CT images first via our Automatic Anatomy Recognition (AAR) method,[9] then recognizing pathologic LNs in each zone without explicitly delineating them, and finally quantifying disease via PET images, thereby accomplishing disease quantification (DQ) without explicitly delineating either the nodal zones or LNs. This approach is inspired by our recent AAR-disease quantification (DQ) methodology for quantifying disease in anatomic organs via PET/CT without the explicit segmentation of organs or pathology.[10] In this paper, we will take this stance to quantify nodal disease via PET/CT and our focus will be the *thoracic body region*. Hence, our review of literature below will be confined to past works related to the thoracic body region only.

1.B. Related work

As mentioned above, approaches focusing on direct LN zonal recognition are very few. There are two approaches to localize zones: Model-based and atlas-based. In Ref. [3], we modified a previously developed fuzzy-model-based body-wide AAR system[9] to automatically localize IASLC-defined lymph node zones on CT images. The LN zones are first modeled in the way anatomic organs are modeled in Refs. [9–11] based on their shape and geographic layout. Subsequently, the AAR approach is utilized to localize LN zones in given patient images. The approach has been extended to other body regions.[5,6] In Ref. [7], the authors utilized spatial priors from a multi-atlas label fusion strategy to detect (all, not necessarily diseased) LNs and map LN stations. Unlike the model-based strategy, the atlas-based approaches do not directly localize LN zones. They achieve a nodal zone inclusion accuracy of 85–88%. In Ref. [8], a multi-atlas organ segmentation approach is utilized to identify IASLC-defined mediastinal and hilar LN zones on CT scans guided by the segmented organs. The previous AAR approaches achieved a zonal localization accuracy of 4–5 voxels which leaves considerable room for improvement for localization accuracy from the perspective of DQ. Our goal in this paper is (a) to

[1]In this paper, we assume that LNs with high FDG uptake relative to the background tissue in PET images are "pathologic" nodes. Although this is true in many disease conditions, we note that not all high uptake nodes necessarily constitute disease involvement by cancer.

bring this error down to 2–3 voxels, and (b) to demonstrate that zonal disease quantification can be subsequently performed accurately without explicit delineation of LNs.

Compared to research on automatic localization of LN zones, most published papers focused on detection and segmentation of LNs on CT images irrespective of whether or not the individual LNs are pathological. LN detection methods can be divided broadly into two groups: Those based on classical pattern recognition techniques and those based on the more recent deep learning strategies. In the first group, hand-crafted features are first selected followed by the detection and delineation of LNs. Features selected include those derived via Hessian analysis,[6,7,12,13] texture properties computed from Haralick gray-level co-occurrence matrix,[14] Haar-like features,[12,15,16] histogram of oriented gradients,[17] and local binary pattern,[18] etc. Once the features are extracted, a classifier is often used to decide if a blob-like entity characterized by the features is a real LN. The classifiers employed include support vector machine (SVM),[7,12,18,19] random forest,[14] etc.

Methods based on deep learning can extract features automatically within the neural network and also perform within the same network the LN detection task. In Ref. [20], the authors developed a 2.5D representation for LN detection by using a convolutional neural network (CNN). The basic idea is to use the CNN as a classifier for each patch extracted from image volumes of interest. In Ref. [21], the authors used three convolutional neural architectures (CifarNet, AlexNet, and GoogLeNet) to compare and evaluate LN detection performance. In Ref. [1], a fully convolutional network is trained to detect lymph node clusters and a conditional random field approach is used subsequently to segment LNs. A 3D U-Net is used in Ref. [22] to segment mediastinal LNs in CT images where other anatomical structures like lungs, airways, and aortic arch, etc., are also segmented in order to improve the performance of LN segmentation. In Ref. [23], a data augmentation approach based on generative adversarial networks is proposed, and the U-Net model is trained for lymph node segmentation. The U-Net and Mask R-CNN architectures are combined for segmentation and detection of mediastinal lymph nodes and anatomical structures in CT data in Ref. [24].

Owing to advantages of the functional information from PET and anatomical structure information from CT, in Ref. [25] a Markov random field model to segment lung tumors is proposed, which encoded the information from both modalities. In Ref. [26], random forest classification within the mixed spatial-spectral space of component-trees modeling PET/CT images was employed for segmentation of lymphoma. In Ref. [27], a fully convolutional neural network (FCN) is used to segment lung cancer utilizing PET and CT. In Ref. [28], the authors proposed two-stream chained deep neural network for esophageal gross tumor volume segmentation that fused the CT and PET modalities. All of these papers focus on detection or segmentation of LNs on CT scans. As far as we know, there are no reports that deal with methods to recognize LNs, particularly pathologic, by using deep learning networks on PET/CT images.

Currently, several commercial vendors offer software for disease measurement (for example, Refs. [29,30]); however, they all operate under the paradigm of first manually performing recognition of diseased regions by manually specifying a region of interest (ROI), subsequently automatically delineating lesions by making use of information from PET alone or from both PET and CT, and finally measuring disease burden in the form of volume and PET standardized uptake value (SUV) statistics within the lesion region. Although numerous papers have been published as we discussed above[1,15,16,19,23] their focus has been LN segmentation and not disease burden estimation within LN zones. To the best of our knowledge, no methods have been reported for LN disease quantification aside from the manually guided methods[29,30] mentioned above.

In summary, a complete automated system for LN disease quantification (LN-DQ) on PET/CT images within well-defined LN zones has numerous clinical applications. Although several individual components of such a system have been worked on and published, none of them has reached the final goal of disease measurement in zones. Most methods focused on LN segmentation and did not show how disease measurement can be accomplished continuing beyond nodal segmentation. Methods that demonstrated LN zone localization also fell short and did not demonstrate how DQ can be performed within each zone. In this paper, we present a methodology for a complete LN-DQ system which, given a PET-CT image, reports LN-zone-wise disease burden. It bypasses the challenging and somewhat ill-defined problem of explicit nodal delineation but rather localizes the LN zones and recognizes the individual pathologic nodes within the recognized zone for estimating total disease burden within each zone without explicitly delineating either the zone or the nodes they contain.

We will refer to our methodology as AAR-LN-DQ, an abbreviation for Automatic-Anatomy-Recognition-Lymph-Node-Disease-Quantification. Overall, it consists of four modules as illustrated schematically in Fig. 1. (a) Building a fuzzy anatomy model of the LN zones in a body region of focus (in our case, thorax) following AAR principles[9]; (b) Performing LN zone recognition; (c) Recognizing diseased LNs within the localized zone; and (d) Performing disease quantification within each zone. In Section 2.A, we briefly summarize the IASLC zonal definitions, including our adaptations to make them computationally unambiguous, and our approach to model LN zones by treating them as 3D anatomic objects. Our approach to zonal recognition (localization), described in Section 2.B, also follows AAR principles but with a key novelty. We select several anatomic organs as "anchor" objects to locate LN zones relative to them and determine with a comprehensive search the best anchor organ for each zone and the best overall hierarchy in which to arrange anchor organs and the zones related to them. Following delineationless AAR-DQ principles formulated recently,[10] in Section 2.D we present an approach to directly quantify disease burden within each zone. It makes use of the fuzzy object model mask resulting from zonal recognition and consists of four key steps: (a) Recognizing high uptake

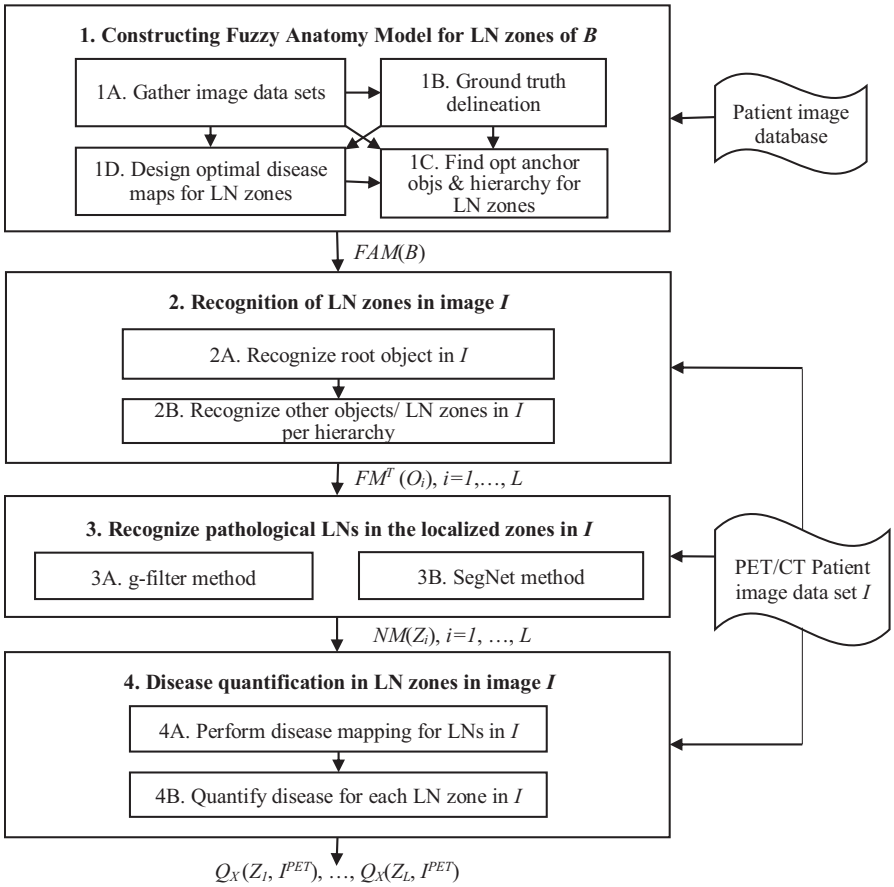

FIG. 1. A schematic representation of the AAR-LN-DQ approach.

confounding objects (such as heart); (b) recognizing pathologic nodes (Section 2.C) using a novel concept of a globular filter (g-filter) or a deep network SegNet; (c) estimating disease severity at each voxel in the zone by considering its SUV, its fuzzy object mask membership, its confounding object membership, and its g-filter output; and (d) estimating disease burden within each zone. In Section 3, we present our results from experiments involving 42 near-normal diagnostic CT images (used for model building) and 63 PET/CT datasets from patients with lymphoma. Our concluding remarks are summarized in Section 4.

The AAR-LN-DQ approach has the following unique features: (a) It treats LN zones as any anatomic 3D object, which makes the general AAR approach become directly applicable in their localization. (b) It decouples dependence on explicit segmentation (delineation) of lymph nodes from disease measurement, and performs disease quantification directly from zonal and nodal (rough) localization information found automatically. This in turn makes the disease quantification process robust, efficient, and practical. (c) It takes a fuzzy approach to handle uncertainties throughout: for LN zone modeling, zone and node recognition, disease mapping, and disease quantification. (d) It creatively combines explicit modellable high-level knowledge encoded via AAR with the ability of deep networks to harness exquisite low-level details to build a practical system for measuring nodal disease burden. (e) By the characteristics of the AAR approach, AAR-LN-DQ is not tied to any specific body region or object(s), and hence it is applicable body-wide although in this paper we focus on the thorax.

A preliminary version of the LN zone and node recognition aspect of this paper appeared in the SPIE Medical Imaging Conference of 2014 and 2018.[3,5] The current paper differs from those conference presentations in major ways. (a) The present paper gives a comprehensive literature review. (b) It describes all involved steps fully with detailed algorithmic steps including the above recognition steps. (c) It describes a method to find an optimal way of determining the anchor objects and the associated hierarchy. (d) It describes an approach to quantify disease within zones by considering both zonal and nodal recognition. (e) It presents results involving a much larger set of patient image data. (f) Most importantly, based on earlier experience, numerous improvements have been made, and a complete AAR-LN-DQ approach is designed and presented.

2. MATERIALS AND METHODS

2.A. AAR-LN-DQ approach

We will follow the scheme in Fig. 1 to describe our approach, summarize briefly previous methods for completeness when used unaltered (please refer to earlier AAR papers, specifically,[9,11] for details), and describe AAR-LN-DQ-specific new advances in detail. We will follow the notation used in

TABLE I. Notations used in this paper.

Notation	Definition
$O_1,\ldots,O_L$	Our *body region* of interest B is the thorax L 3D *objects* of B that are considered in our study which are organs and lymph node zones in B
$\mathcal{I}^m = \{I_1^m, \cdots, I_N^m\}$	A set of *training images* in modality m of B from N near-normal subjects, where $m \in \{d\text{CT}, l\text{CT}, \text{PET}\}$. dCT and lCT represent diagnostic contrast-enhanced CT and low-dose CT
$(I^{\text{CT}}, I^{\text{PET}})$	The image pair in a PET/CT acquisition from a given patient
$\mathcal{I}_b = \{I_{n,l} : 1 \le n \le N \& 1 \le l \le L\}$	The set of all *binary images* of the objects of B which are used for model building, $I_{n,l}$ being the binary image representing object O_l in image I_n^m
$\mathcal{D}^m = \{D_1^m, \cdots, D_k^m\}$	A set of *training images* of B in modality m of *patients with disease*
$\mathcal{C}^m = \{C_1^m, \cdots, C_M^m\}$	A set of *test images* of B in modality m of *patients with disease*
$FM(O_l)$	*Fuzzy model* of object O_l derived from the set of all binary images of O_l
$d_Z(x)$	*Disease map* associated with LN zone Z. It maps SUV x at a voxel v within Z to disease severity at v on a [0, 1] scale
$FAM(B)$	*Fuzzy anatomy model* of the whole object assembly in B which includes all prior information gathered about objects such as the hierarchical arrangement of objects, their SUV properties, disease maps, object relationships, fuzzy models, etc
$FM^T(O)$	Transformed $FM(O)$ corresponding to a state when O is recognized in a patient image
$NM(Z)$	A fuzzy nodal mask
$Q_X(Z, I^{PET})$	A set of quantitative measures[2] describing the disease burden within LN zone Z

previous AAR publications closely, but will need some new terminology as well which we will introduce as we progress. In Table I, we summarize the terminology used in the paper. Our *body region* of interest B in this paper is the thorax.

2.B. Constructing fuzzy anatomy model for lymph node zones

2.B.1. Gathering image data

This retrospective study was conducted following approval from the Institutional Review Board at the Hospital of the University of Pennsylvania along with a Health Insurance Portability and Accountability Act waiver. For the near-normal set $\mathcal{I}^m$, contrast-enhanced diagnostic chest CT scans of 42 near-normal subjects (radiologically normal with exception of minimal incidental focal abnormalities) are selected. For abnormal sets $\mathcal{D}^m$ and $\mathcal{C}^m$, we selected whole-body ^{18}F-fluorodeoxyglucose (FDG)-PET/CT scans of 63 patients with Hodgkin lymphoma or Diffuse large B-cell lymphoma. Only the thoracic portion of these scans was utilized in our study. All PET/CT scans had previously been acquired on scanners with time-of-flight capabilities (Gemini TF, Philips Medical Systems, Bothell, WA). 3D PET data had been acquired ~60 min after intravenous administration of ~555 MBq of FDG for ~3 min per bed position. Image reconstruction had been performed at 4 mm nominal slice thickness in the axial plane. Voxel size in PET images was much larger than in the lCT counterpart, as such PET images were interpolated to make their voxel size equal that of lCT images. The two image datasets are summarized in Table II.

2.B.2. Defining and delineating objects

As per AAR methodology, anatomic body region of B and its organs are precisely defined first (see Ref. [9]).

For LN zones, we followed the IASLC definitions[2] but adapted them to our goal with some changes as deemed necessary. For example, in implementing those definitions computationally, we found that in some cases some zones became empty based on the spatial relationships of the anatomical structures that determine the boundaries of the zones. For similar reasons, we split Zone 3a into two zones — 3a-sup, which is equivalent to Zone 3a based on the IASLC definitions, and 3a-inf, which is a new zone we created to cover the inferior portion of the pre-vascular mediastinum that was not addressed by the IASLC definitions. Every zone is specified with a boundary in each of anterior, posterior, superior, inferior, left, and right directions to express the limits of the zone anatomically. These limits are defined by planes which are not necessarily parallel to the image coordinate planes. In our implementation, for modeling the zones as 3D objects, we express each zone as a polyhedron by following the definitions. Figure 2 shows an example using Zone 3p for illustration. See Supplementary Material for a compact description of the definitions of the LN zones.

For generating ground truth binary masks, all objects (organs and LN zones) are delineated following their definitions. This step generates the set of binary images $\mathcal{I}_b$ from the input set of images $\mathcal{I}^m$. The tracings are done on the dCT images of this set. Table III lists all organs and zones considered in this study and their acronyms used throughout this paper.

In addition to the binary masks described above used for model building, we also created ground truth delineations of all pathologic nodes in the patient PE/CT datasets $\mathcal{D}^m$ and $\mathcal{C}^m$. Each such node was identified on the PET images manually by the radiologist (Torigian) and was delineated on PET by first thresholding and subsequently by verification on lCT and manual correction as needed. These masks will be used as ground truth for estimating disease maps in LN zones and for determining the accuracy of recognizing pathologic nodes and the accuracy of disease quantification in LN zones.

[2]Notation Q_X is fashioned after notations D_X and R_X commonly used for diagnostics and therapeutics, and is intended to denote quantitative disease analytics, as employed in[10].

TABLE II. A description of the image datasets used in this study.

Dataset #	Number of subjects	Modality	Imaging protocol	Image size, resolution	Use, train/test division
DS1	42	Diagnostic (near-normal) CT; used for building *FAM*(*B*) and testing zone recognition	*dCT*: Contrast-enhanced, axial, breath-hold	512 × 512 × 45–68, 0.77 × 0.77 × 5.0 mm^3	Zone recognition, 30/12 ($\mathcal{D}^m/\mathcal{C}^m$), sixfold
DS2	63	Patient PET/CT; Used for node recognition and disease quant	*l*CT: Unenhanced, axial, quiet breathing	512 × 512 × 52–92, CT: 1.14 × 1.14 × 3.75 mm^3 PET: 4 × 4 × 4 mm^3	Node recognition & disease quantification, 53/10 ($\mathcal{D}^m/\mathcal{C}^m$), fivefold. Total number of pathological LNs = 214

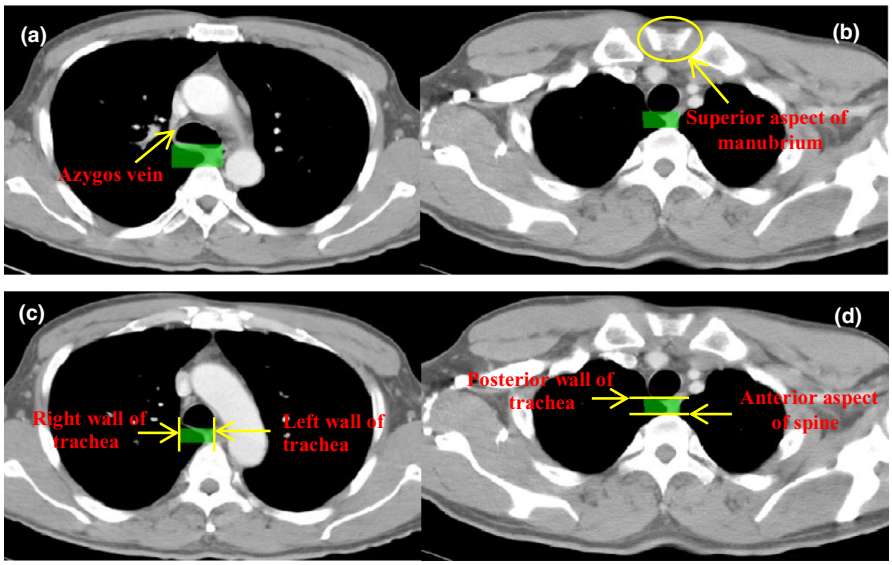

FIG. 2. Definition of lymph node Zone 3p. (a) Inferior boundary: Inferior aspect of the horizontal portion of azygos vein. (b) Superior boundary: Superior aspect of manubrium. (c) Right & left boundaries: right & left wall of trachea. (d) Anterior & posterior boundaries: anterior & posterior wall of trachea. The zone is displayed as green overlay. [Color figure can be viewed at wileyonlinelibrary.com]

2.B.3. Find optimal anchor objects and hierarchy for LN zones

In the AAR set up, the *fuzzy anatomy model FAM*(*B*) of *B*, with all its organs and LN zones of interest is defined by 5 elements.

$$FAM(B) = (H, M, \rho, \lambda, \eta). \quad (1)$$

For a detailed description of these parameters, see Refs. [9,11]. Briefly, H is a hierarchy of objects in B, represented as a tree. M is a set of fuzzy models, one model for each of the L objects in B, $M = \{FM(O_k): k = 1, ..., L\}$. ρ describes the parent-to-offspring relationship in H over the population. λ is a family of scale factor ranges. η denotes a set of measurements pertaining to the object assembly in B including intensity properties and all learned parameters that are needed for object recognition and disease quantification. Note that the object ensemble considered for *FAM*(*B*) includes organs and LN zones. Our interest in organs per se is secondary here; they are used as anchor objects of reference for LN zones, the latter being our primary objects of focus. Achieving high accuracy of recognition of LN zones is crucial for guaranteeing high accuracy of disease quantification in zones. Recognition accuracy in turn is determined by the anchor objects chosen for the zones and the constructed hierarchy H. This is where AAR-LN-DQ differs significantly from previous AAR strategies in the model building process, as we explain below. Please see Ref. [9] for details on the remaining parameters M, ρ, λ, and η of *FAM*(*B*).

The idea underlying H in *FAM*(*B*) is to facilitate locating an LN zone Z in a given image I based solely on prior information. The prior information is encoded in the form of the relationship of Z with a reference organ O. For each zone Z, our goal is to select that organ O with respect to which Z has the steadiest relationship, so that once O is recognized in I accurately, Z can be placed (localized) in I with least error. To achieve this goal, we set aside TSkn as the root object in H since it is easy to recognize in I. We consider each of the remaining organs listed in Table III as a potential reference anchor object for each LN zone. Additionally, we allow composite organs created by the union of organs taken two at a time as reference anchors. Denoting the set of organs in

TABLE III. Organs and LN zones in the thorax considered in this study and their abbreviations.

Organs	Description	Zones	Description
TSkn	The outer skin boundary of the thoracic body region	Z1	Zone 1
TSk	Thoracic skeleton	Z2R	The right part of Zone 2
AS	Arterial system	Z2L	The left part of Zone 2
IMS	Internal mediastinum	Z12	Z1 ∪ Z2R ∪ Z2L
LPS	Left lung (pleural sac)	Z3a-sup	The superior part of Zone 3a
RPS	Right lung (pleural sac)	Z3a-inf	The inferior part of Zone 3a
RS	Respiratory system = LPS + RPS + TB	Z3p	The posterior part of Zone 3
SCord	Thoracic spinal cord	Z4R	The right part of Zone 4
TB	Trachea and bronchi	Z4L	The left part of Zone 4
PC	Pericardium boundary representing the heart	Z4	Z4R ∪ Z4L
E	Thoracic esophagus	Z5	Zone 5
Ao	Aorta	Z6	Zone 6
VS	Venous system	Z56	Z5 ∪ Z6
TV	TB + VS	Z7	Zone 7
AD	AS + SCord	Z89	Z8 ∪ Z9
AR	AS + RS	Z10R	The right part of Zone 10
LR	LPS + RPS	Z10L	The left part of Zone 10

Table III (excluding TSkn) by $\mathcal{O}$, the set of reference anchor organs considered becomes $\mathcal{O}_A = \mathcal{O} \cup \{O_i \cup O_j: O_i \neq O_j \,\&\, O_i, O_j \in \mathcal{O}\}$. From the 14 zones listed in Table III, we create ten zones by merging some of the smaller zones into a single zone. This was done mainly to make sure that we catch a sufficient number of pathologic nodes in our datasets so that testing the accuracy of disease quantification becomes statistically meaningful. Otherwise, many zones will remain empty or catch very few pathologic nodes. Specifically, the set of zones we considered is $\mathcal{Z}$ = {Z12, Z3a-inf, Z3a-sup, Z3p, Z4, Z56, Z7, Z89, Z10R, Z10L}, where Z12 = Z1 ∪ Z2R ∪ Z2L, Z4 = Z4R ∪ Z4L, Z56 = Z5 ∪ Z6, and Z89 = Z8 ∪ Z9.

To find H, we take the following approach. For each zone Z_i in $\mathcal{Z}$ and organ O in $\mathcal{O}_A$, we form a mini hierarchy as illustrated in Fig. 3(a) and create $FAM(B)$ by using a subset of dataset DS1 (Table I). We use this model to test the error $\theta(Z_i, O)$ of recognition of zone Z_i in a second disjoint subset of DS1[3]. We then choose for Z_i that organ O_i in $\mathcal{O}_A$ that yields the least recognition error among all members of $\mathcal{O}_A$. Finally, we form the hierarchy H as illustrated in Fig. 3(b) by making

[3]We actually divide DS1 into 3 subsets — a training set used to build models, a validation set to determine optimal hierarchy, and a test set to determine the accuracy of zonal recognition. This we do in a multi-fold manner as explained in Section 3.

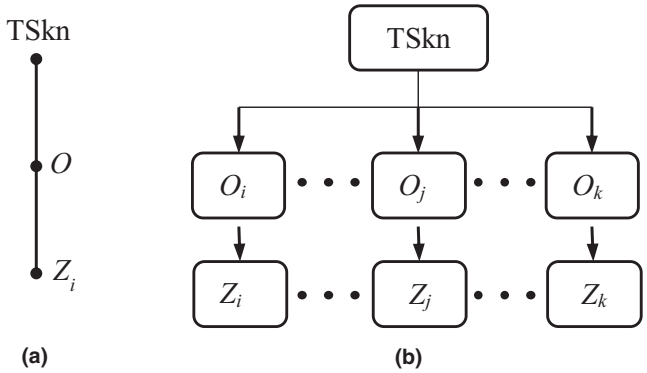

FIG. 3. (a) Mini hierarchy considered for determining optimal anchor organ to be paired with each LN zone Z_i. (b) Optimal hierarchy formed after an optimal organ O_i is found for each LN zone Z_i.

the best organ selected in this manner to be the parent for each zone. Note that there are $|\mathcal{Z}| \times |\mathcal{O}_A|$ experiments of the type described above involved in our search for optimal hierarchy, where |x| denotes the cardinality of set x. For our case, the number of experiments is 580.

For recognition error $\theta(Z_i, O)$, we utilized the error in localizing zone Z_i, which is expressed as the 3D distance between the true geometric center of Z_i and the geometric center of the fuzzy model $FM^T(Z_i)$ that is localized in the image under question. If O_i yields the smallest error $\theta(Z_i, O_i)$, then this suggests that Z_i has the tightest (least variable) positional relationship with respect to organ O_i.

2.B.4. Design optimal disease maps for LN zones

The process of disease quantification involves a training part, which belongs to the model building stage of the AAR-LN-DQ process, and an actual disease estimation part. For the sake of continuity, we will present both parts in Section 2.D — Disease quantification.

2.C. Recognition of lymph node zones

Once $FAM(B)$ is built for the objects (organs and LN zones) in B (thorax) following the procedure described in Section 2.B, it can be employed to recognize LN zones included in the model in any given image[4] I of B. The procedure for recognizing zones follows the same process as described in previous AAR publications for organ recognition. There are some minor differences due to the fact that, now, the primary objects of interest are LN zones, and not organs. We will explain these differences below; please see Refs. [9,11] for details on the basic recognition algorithms.

AAR organ recognition starts off by first recognizing the root object TSkin (the outer boundary of skin of the thoracic body region in our case) of H in I following the method described in Ref. [9]. Subsequently, organ recognition in I

[4]Note that I may be a dCT image from data set DS1 or a lCT image I^{CT} from DS2.

proceeds in two stages following the objects in H in a breadth-first order. In the first stage, called *one-shot recognition*, the child organ O in H is located (recognized) purely from the parent-to-child relationship information ρ encoded in $FAM(B)$. This strategy already places the transformed fuzzy model $FM^T(O)$ of the child O in I in the close proximity of the true location (pose) of O in I. In the second stage of *refined recognition*, the result from the first stage is fine-tuned based on image intensity properties (including known actual pixel value ranges for O, O's known texture properties, etc.) in the vicinity of $FM^T(O)$ in I by maximizing the agreement of these properties in I relative to those in the region defined by $FM^T(O)$. In AAR-LN-DQ, for organs in H, we follow exactly the same process. However, unlike organs, LN zones are conceptually defined regions of space without any observable intensity (or texture) boundaries in images, as such we perform only one-shot recognition for them. This is the reason that accurate localization of their parent organs becomes crucial as well as determination of the optimal parent to be assigned to each zone as an anchor organ. Thus, after the root object is located, we localize the anchor organs by using the two-stage process, and subsequently their offspring zones are localized using the one-shot strategy. Finally, the result of the recognition process is the adjusted fuzzy model $FM^T(O)$ in I for each object O (organ or zone).

2.D. Lymph node recognition

We have designed two strategies for LN recognition, the first based on a new concept of *globular filter* or *g-filter* and the second based on a deep neural network SegNet. The g-filter approach looks for blob-like objects within the zonal mask using spherical-ball templates of varying sizes that best match the intensity and uptake properties in I^{CT} and I^{PET} that are expected for pathologic nodes. Subsequently, it employs machine learning techniques to identify nodes by iteratively relaxing/refining the classification strategy from ball level to the slice level to the voxel level. In this process, the fuzzy model masks from Hrt (heart) and TSk (skeleton of thorax) are excluded from the zonal mask. The deep network approach trains a SegNet network to identify pathologic nodes by using the recognized zonal masks and the truly pathologic nodes identified within them together with the I^{CT} and I^{PET} image information within the masks. Both approaches output masks denoting roughly the whereabouts of the pathologic nodes (and not their explicit delineations). The two approaches are described in detail in the rest of this section.

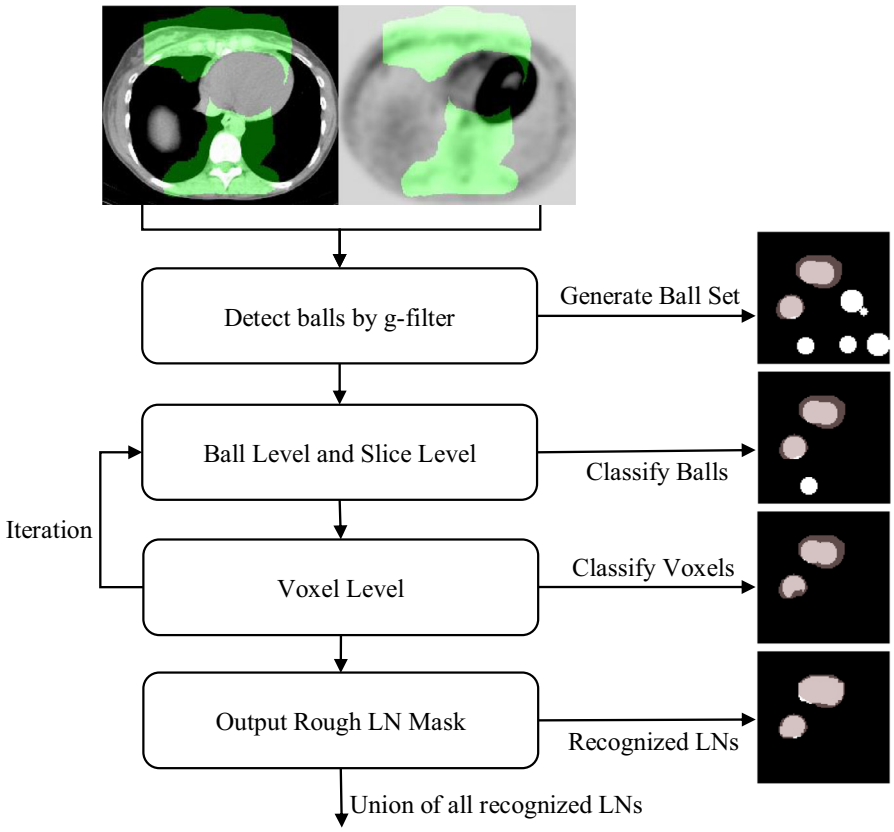

FIG. 4. A schematic representation of the g-filter approach. The light gray mask denotes lymph node ground truth; the bright mask denotes the rough mask obtained at each level. [Color figure can be viewed at wileyonlinelibrary.com]

2.D.1. g-filter approach to node recognition

A schematic representation of the proposed g-filter approach is depicted in Fig. 4. It includes two main parts: generating ball candidates by g-filter and removing false positive balls and voxels by SVM thorough three levels — ball level, slice level, and voxel level. The aim of ball level and slice level is to remove false positive balls. The voxel level plays a role in the refinement of LN recognition. Note that these three levels are executed iteratively.

This approach is delineated as Algorithm g-filter (gF) below. The individual steps in the algorithm are described at an intuitive level in the rest of this section.

Algorithm g-filter (gF)

Input: Image pair (I^{CT}, I^{PET}), fuzzy model $FM^T(Z)$ for zone Z recognized in I^{CT}, fuzzy model $FM^T(O)$ for $O \in$ {Hrt, TSk} recognized in I^{CT}, number of iterations k for Steps 2-4.

Output: A fuzzy nodal mask $NM(Z)$ of the pathologic nodes found in zone Z.

Begin

0. Set $X = FM^T(Z) - (FM^T(\text{Hrt}) \cup FM^T(\text{TSk}))$; /Note that X is a fuzzy set/
1. Apply g-filter separately to I^{CT} and I^{PET}; let the set of all balls produced within the fuzzy region X from both images be β_0;
2. *Ball-level classification*: Select those balls in β_0 that correspond to pathologic nodes; let the resulting set of balls be β_1;
3. *Slice-level classification*: Further discard those balls b in β_1 such that all slices of b are considered not to correspond to a pathologic node; let the resulting set of balls be β_2;
4. *Voxel-level classification*: Set $\beta = \beta_2$ and select voxels in the balls in β that belong to pathologic nodes; let the resulting set of voxels be V_1 and the balls associated with the voxels in V_1 be β_2;
5. Set $\beta_0 = \beta_2$ and repeat Steps 2-4 for k iterations;
6. Output $NM(Z) = [\bigcup_{b \in \beta_2} b] \cap X$;

End

Step 0. Suppressing confounding objects: In this step, the confounding objects with high uptake, namely heart (FM^T(Hrt)) and bone marrow (FM^T(Tsk)) from skeletal structures are removed if zone Z ($FM^T(Z)$) partially overlaps with those objects. Removal is done by fuzzy set subtraction. In Fig. 5, we illustrate examples of high uptake confounding regions arising from the two objects Hrt and TSk.

Step 1. g-filter to detect balls: Let $I^{CT} = (V, f_{CT}(v))$ and $I^{PET} = (V, f_{PET}(v))$, where $f_{CT}(v)$ and $f_{PET}(v)$ denote the image intensity value at voxel v in I^{CT} and I^{PET}, respectively. We will apply the g-filter to each image I^{CT} and I^{PET} separately to detect the most plausible spherical ball with its center at *each voxel* v as described below. For I^{CT}, this will yield a new image denoted $I_g^{CT} = (V, F_{CT}(v))$, where the intensity function $F_{CT}(v) = (\delta_{CT}(v), r_{CT}(v))$ assigns two values to each voxel v: $\delta_{CT}(v)$ denotes the filter response at v (see below) and $r_{CT}(v)$ corresponds to the radius of the most plausible ball centered at v. Similarly, the output of the g-filter for I^{PET} will be a vector-valued image $I_g^{PET} = (V, F_{PET}(v))$, where $F_{PET}(v) = (\delta_{PET}(v), r_{PET}(v))$. Let the fuzzy model $FM^T(Z)$ of the recognized zone Z in I^{CT} (and hence I^{PET}) be denoted as an image $(V, f_{FM}(v))$ where $f_{FM}(v)$ stands for the fuzzy membership of Z at voxel v. Thus, at every voxel v in the given PET/CT image pair (I^{CT}, I^{PET}), we will have seven values: fuzzy mask membership $f_{FM}(v)$, CT intensity $f_{CT}(v)$, PET intensity $f_{PET}(v)$, two filter output values $F_{CT}(v)$ for the CT image, and two filter output values $F_{PET}(v)$ for the PET image.

The filter operation on I^{CT} is as follows (it works exactly identically on I^{PET}). At each voxel v of I^{CT}, each ball b, selected from a series of template balls, centered at v of radius from a pre-determined minimum r_{min} to maximum r_{max} is considered. A t-statistic of the difference in intensity distributions (histograms) inside b and outside[5] b is estimated and Welch's unequal variance t-test is used to find the optimal ball at v as follows. Let $\overline{X}_1$, s_1, and N_1 denote the mean, standard deviation, and the sample size of voxel intensities from inside b and let the corresponding values for outside b be $\overline{X}_2$, s_2, and N_2. Then the t-statistic is given by

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}. \quad (2)$$

The ball that gives the maximum t value at v from among the template set of balls is considered to be the most plausible ball at v, $r_{CT}(v)$ is taken to be the radius of this optimal ball, and the filter response $\delta_{CT}(v)$ is taken to be the statistic t. The above process proposes a ball $b(v)$ at every voxel v in I^{CT} (and similarly in I^{PET}) with radius $r_{CT}(v)$ and response $\delta_{CT}(v)$. At true nodal "centers", we expect δ_{CT} to peak, so we find local maxima as potential nodal locations. Thus, only those balls which satisfy two conditions are selected to be in β_0: (a) $b(v)$ has a locally maximum response $\delta_{CT}(v)$; and (b) the membership value at v in X is greater than 0. In words, β_0 constitutes a set of all balls, confined to the mask of zone Z, with confounding objects (heart and bone marrow) suppressed, such that each ball represents the most plausible sphere that can be fit to the manifestation of an LN that appears within Z in I^{CT}. Note that β_0 constitutes the union of balls found in I^{CT} and I^{PET} separately. β_0 may contain too many false balls that do not correspond to pathologic nodes or even nodes. The remaining steps in gF gradually weed out most false balls as explained below. Figure 6 demonstrates an example of β_0 found in two I^{CT} datasets.

We emphasize that the g-filter is quite different from the popular Hessian-based method[14,31] of deriving features at every voxel about the underlying shape (spherical, cylindrical, etc.). The g-filter explicitly finds the best matching ball within a set of template balls that can exist at every voxel with the voxel as its center. The Hessian method on the other hand finds features for the part of the surface (spherical, cylindrical, etc.) that may exist in the vicinity of the voxel. These features will have to be subsequently put together into a gestalt to find balls in the Hessian method.

[5]Voxels immediately outside b within one voxel distance are considered for finding the histogram of the outside region.

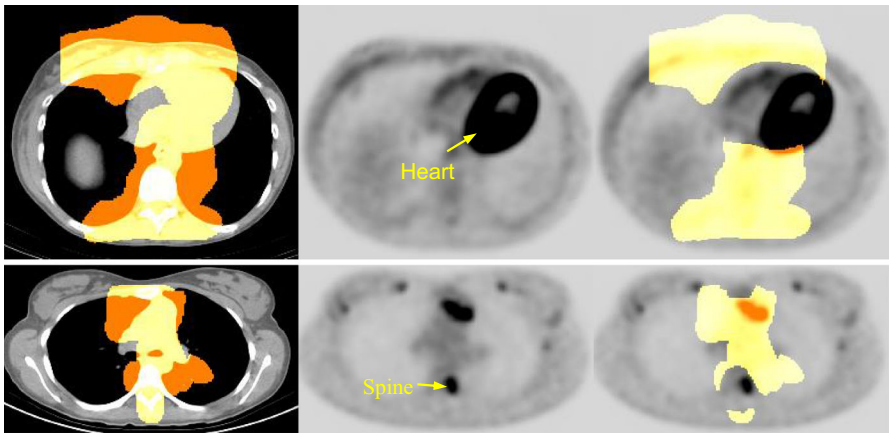

FIG. 5. Examples illustrating Step 0 of Procedure gF. Top row (L to R): A CT slice showing zone *Z89*; the corresponding PET slice where the confounding high uptake region is from Hrt; and PET slice with the overlay of fuzzy set *X* found after suppressing Hrt. Bottom row: Same as top row but for zone *Z89* and confounding region coming from an FDG avid lesion in the TSk. [Color figure can be viewed at wileyonlinelibrary.com]

Steps 2, 3. Ball-level and slice-level classification: For both ball-level and slice-level classifications, we use a set of features derived from within the balls from CT and PET images in addition to the radius and response values in the filter outputs I_g^{CT} and I_g^{PET}. The features associated with a ball b (v) at voxel v in I^{CT} are as follows: mean and standard deviation (SD), maximum value, minimum value and median value of $f_{CT}(v)$, $\delta_{CT}(v)$, $r_{CT}(v)$, and 10 texture properties derived from the gray level co-occurrence matrix obtained from the CT image, namely: energy, entropy, correlation, contrast, variance, inertia, cluster shade, cluster tendency, and homogeneity.[30] Similar features are defined based on I^{PET}. For ball-level classification, the features are derived from the ball region, and for slice-level classification, the features are from the region of the cross section of the ball in the slice under consideration. For both cases, the number of positive samples (balls/slices containing a pathologic LN) is much smaller than that of negative samples. Multiple classifiers were designed to fully use positive samples by training each classifier with balanced sample sets including all positive samples and the randomly selected negative samples for every classifier.[31] Here, ten SVM classifiers are trained by the same positive training balls and different sets of negative training balls selected from the total negative ball set randomly. A voting strategy is used to combine the outputs of the 10 classifiers. Ball-level classification yields a reduced set of balls β_1 which is further reduced to set β_2 by the more conservative slice-level selection.

In summary, the set of balls β_0 produced in Step 1 has two issues: there are too many false balls and some balls that cover pathologic LNs may not cover them fully well. Steps 2 and 3 are designed to drastically reduce the number of false balls. The second issue will be addressed in Step 4 and its iterations.

Steps 4, 5. Voxel-level classification: In these steps, voxels within the union of all balls in the set β_2 are classified (again using a SVM classifier) as either belonging to or not belonging to a pathologic node. These steps iteratively refine the voxel-level recognition of pathologic nodes and also expand the region of containment by taking the union of the balls associated with the resulting voxels (note that there is a ball

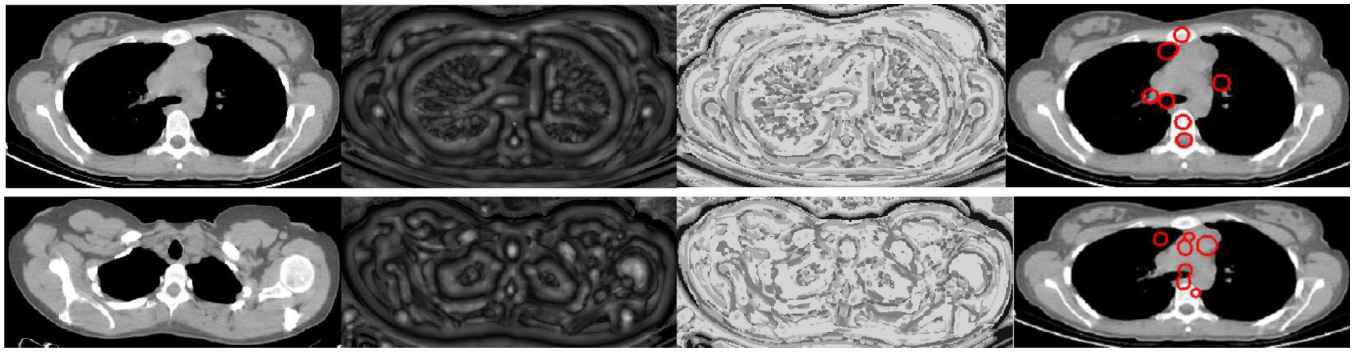

FIG. 6. Examples of balls (set β_0) produced in Step 1 of procedure gF for two CT datasets shown in two rows. L to R: A CT slice, corresponding slice with the response value δCT(v), radius value rCT(v), and cross sections of balls overlaid on the CT slice. Zone *Z* considered here is the union of all zones in Z. [Color figure can be viewed at wileyonlinelibrary.com]

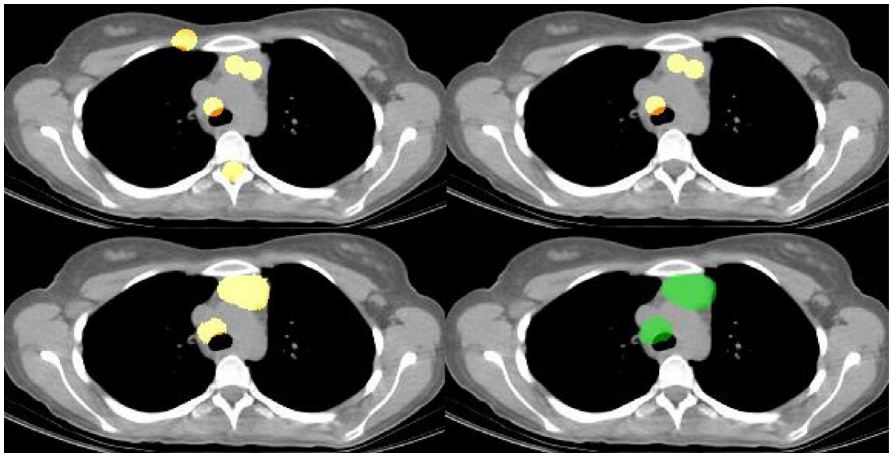

FIG. 7. An example of balls (set β_2) produced finally in Step 5 of procedure gF. Ball cross section is overlaid on a CT slice. Row 1, L to R: After ball-level and slice-level classification. Row 2, L to R: After voxel-level classification with two iterations and the ground truth of pathologic lymph nodes. Zone Z considered here is the union of all zones in Z. [Color figure can be viewed at wileyonlinelibrary.com]

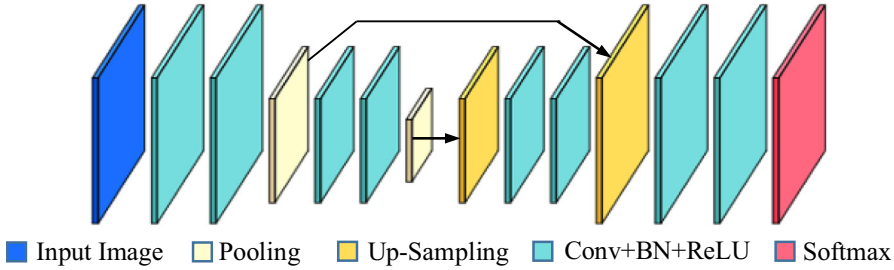

FIG. 8. Network architecture utilized in the SegNet approach to LN recognition. Each encoder and decoder does the operations of convolution, batch normalization, and ReLU. Input to the network: fuzzy model masks from zonal recognition, $FM^T(Z)$ for zones Z, $FM^T(O)$ for O $\in$ {Hrt, TSk}, and other input images (PET/CT, radius image, response image, etc; see text). It outputs a probability map that indicates the likelihood of each voxel in Z being in a pathologic node. [Color figure can be viewed at wileyonlinelibrary.com]

associated with every voxel with voxel as the ball center). Figure 7 illustrates the refinement that takes from ball-level to slice-level to voxel-level in nodal recognition due to Steps 2–4.

Step 6. Output: Finally, a fuzzy mask $NM(Z)$ is output where the fuzzy membership value $f_{\mathrm{FM}}(v)$ associated with each voxel v denotes the belongingness of v in some pathological node within Z. $NM(Z)$ is found by taking a fuzzy intersection between the fuzzy recognition LN zone mask X (with the confounding object recognition masks removed) found in Step 0 and the union of all recognized pathological LNs found in Step 5. In other words, the recognized LNs become fuzzified when multiplied by the fuzzy zone mask X.

2.D.2. SegNet approach to node recognition

The goal for this approach is: Given the results of zonal recognition in the form of the fuzzy model masks $FM^T(Z)$ for zones Z and $FM^T(O)$ for $O \in$ {Hrt, TSk}, to recognize pathologic nodes in Z, the output being a probability map that indicates the likelihood of each voxel in Z being in a node with disease. Since the probability map can also be thought of as being similar to the fuzzy mask $NM(Z)$ output by procedure gF, we will denote the output of SegNet also by $NM(Z)$. The output in this approach may also be thought of as a fuzzy segmentation. The architecture of SegNet[32] employed here consists of encoder and decoder sub-networks, each of which is made up of convolutional, batch normalization, and ReLU layers; see Fig. 8. The novelty of SegNet lies in the decoder part, which makes use of pooling indices computed in the max-pooling step of the corresponding encoder to complete non-linear up-sampling. It is shown in Fig. 8 by arrows.

The encoder performs convolution with a filter bank to generate a set of feature maps. These are processed by batch normalized and element-wise ReLUs. Then, max-pooling with a 2 $\times$ 2 window and stride 2 (non-overlapping window) is applied. Here, the encoder network is designed for LN recognition in low resolution feature maps, and the decoder network is used to recover higher resolution feature details at the deepest encoder output, which utilize the maximum pooling indices from the max-pooling step. The final decoder output is fed to a two-class (true — meaning pathologic, and false node) Softmax classifier to produce class probabilities for each voxel independently.

In summary, we input the LN zones from AAR to the encoder sub-network, and it will extract features with a bank of convolution operations, batch normalization and ReLU layer by layer. Finally, the global features can be learned to complete LN recognition. Then, the resolution of the feature map can be recovered in the decoder sub-network, and the same size of output recognition mask can be obtained after Softmax layer. Compared to FCN or U-Net, the main difference of SegNet is that it captures and stores the max-pooling indices, for example, the locations of the maximum feature value in each pooling window. (Although U-Net can provide more contextual information by long-skip connection, it will increase the number of parameters and memory compared to the SegNet.) Importantly, note that we train the network to recognize pathologic nodes only within the zones and not within the whole image. This is the spirit of AAR recognition which already hones in on the region of interest in the image for the network to look for diseased nodes.

The decoder sub-network performs up-sampling its input feature maps using the memorized max-pooling indices from the corresponding encoder feature maps. This helps to recover higher resolution feature maps and complete the LN recognition task. Compared to the task of semantic pixel-wise segmentation,[32] here the emphasis is on recognizing the pathologic LNs and not delineating the boundary of the LNs since our disease quantification strategy needs only rough localization of the LNs. Note also that since the network is trained to recognize only the pathological nodes excluding confounding objects such as Hrt and TSk, there is no need to separately subtract the fuzzy masks of these objects as in procedure gF. This is a major difference between gF and SegNet.

We used on the average 2166 image patches extracted from lymph node zone regions for training in each fold, with 705 image patches in each fold for testing the performance. We used data augmentation strategies including random translation in the horizontal and vertical directions ranging from -10 to 10 voxels to improve training. The size of the input image is resized to 256×256, and the convolutional kernel size is 3×3 in each convolution layer, where the number of kernels is 64. To train the model, we used stochastic gradient descent with a fixed learning rate of 0.001 and momentum of 0.9 using Matlab implementation of SegNet. We train the models until the training loss converges, in which the maximum epoch number is 20. Before each epoch, the training set is shuffled and each mini-batch (4 images in a batch) is then picked in order, thus ensuring that each image is used only once in an epoch. The standard cross-entropy loss is employed as the objective function for training the network. We select the model which performs highest on the validation dataset. We created three different versions of SegNet — SegNet-2, SegNet-4, and SegNet-6 — by using the same architecture but with 2, 4, and 6 input channels, respectively, as follows: SegNet-2: CT and PET images; SegNet-4: CT and PET images and the respective g-filter response images, SegNet-6: CT and PET images and the associated g-filter response and radius images for both.

We have used three variants of SegNet called SegNet-2, SegNet-4, and SegNet-6 depending on the input images utilized. The input images are as follows (in addition to the fuzzy model masks): PET and CT images for SegNet-2; PET, CT, and radius and response images from CT for SegNet-4, and PET, CT, and radius and response images from both PET and CT for SegNet-6.

2.E. Disease quantification

The goal of this step is: Given a PET/CT image pair (I^{CT}, I^{PET}) and the fuzzy nodal mask $NM(Z)$ of the pathologic nodes found in zone Z by one of the above two methods of nodal recognition, to output disease quantities $Q_X(Z, I^{PET})$. In our case, the disease quantity consists of three elements, $Q_X(Z, I^{PET}) = [SUV_{mean}(Z, I^{PET}), SUV_{max}(Z, I^{PET}), fTLG(Z, I^{PET})]$, where the elements represent, respectively, the mean and maximum SUV within Z and total lesion glycolysis (TLG). TLG is commonly utilized[33] to express the total disease burden of a lesion by taking the product of the (metabolic) lesion volume and mean SUV within the lesion. We extend this concept to the whole LN zone Z in a fuzzy manner where the membership expressed in $NM(Z)$ at each voxel v takes on a fuzzy value accounting for various uncertainties due to partial volume effect in I^{PET}, ill-defined boundaries, non-committal segmentation, and disease severity. Accordingly, we define a fuzzy TLG of Z, denoted by $fTLG(Z, I^{PET})$, as follows.

$$fTLG\left(Z, I^{PET}\right) = |v| \sum_{v} d_Z(I_S(v)) I_S(v) NM(Z)(v). \tag{3}$$

In this equation, $|v|$ denotes the volume of voxel v, $d_Z(I_S(v))$ is the disease severity estimated at v, $NM(Z)(v)$ denotes the membership value of $NM(Z)$[6] at v, and $I_S(v)$ is the SUV at v computed from $I^{\mathrm{PET}}(v)$ using the following equation. (In Eq. (3), all voxels are assumed to be of the same size in I^{PET}. If this is not the case, $|v|$ should be brought inside the summation sign.)

$$I_S(v) = \frac{I^C(v)}{ID/BW}. \tag{4}$$

Here, ID is the injected dose of the radiotracer (expressed in MBq), BW is the body weight of the patient (expressed in g), and $I^C(v)$ denotes the radioactivity concentration (expressed in MBq/cc where we assume 1 cc of tissue weighs 1 g) measured at voxel v of I^{PET} which is corrected for decay from the time of injection to the time of image acquisition.

`The disease quantification process for any LN zone Z consists of two steps. Step 0: Estimating an optimal disease map $d_Z(x)$ to map the SUV $x = I_S(v)$ at a voxel v in Z to disease severity corresponding to that SUV; and Step 1: To estimate $Q_X(Z)$.

[6]We have used a binarized version of $NM(Z)(v)$ in all our experiments where the threshold is set at 0.5.

2.E.1. Step 0. Estimating disease map $d_Z(x)$

The disease map $d_Z(x)$, where $x = I_S(v)$, is a parametric function which indicates disease severity at every voxel v within LNs as a function of the voxel's SUV value x. $d_Z(x)$ is modeled as a Gaussian fuzzy mapping function where SUV $x < (\mu_d - \sigma_d)$ are de-emphasized by Gaussian weight while $x \geq (\mu_d - \sigma_d)$ are set to the maximum value.

$$d_Z(x) = \begin{cases} \exp\left[-(x-(\mu_d-\sigma_d))^2/2\sigma_d^2\right], & \text{if } x<(\mu_d-\sigma_d) \\ 1, & \text{if } x\geq(\mu_d-\sigma_d) \end{cases}. \quad (5)$$

On the PET images in the training datasets $\mathcal{D}^m$ where we have carefully delineated pathological nodes, we estimate the mean μ_d and standard deviation σ_d of the SUVs within the diseased nodes. For any zone Z and voxel v in it, then, the above disease map is employed to define disease severity $d_Z(I_S(v))$ at v. The estimated parameters of the disease map are saved in the fifth element η of the anatomy model $FAM(B)$.

2.E.2. Step 1. Estimating disease quantity $Q_X(Z, I^{PET})$

Given a test image pair $(I^{\text{CT}}, I^{\text{PET}})$ from the set $\mathcal{C}^m$, first the objects are recognized in I^{CT}. Following procedure gF or SegNet, nodal masks $NM(Z)$ are determined for each zone Z. Subsequently, the disease map is retrieved from $FAM(B)$ and $fTLG(Z, I^{PET})$ is computed following Eq. (3). $fTLG(Z, I^{PET})$ (expressed in cc) is a weighted sum of the SUV values of voxels within the mask of the recognized LNs multiplied by the voxel volume $|v|$ (expressed in cc). There are two weights for each voxel — mask weight $NM(Z)(v)$ and disease weight $d_Z(I_S(v))$. The estimation of $SUV_{max}(Z, I^{PET})$ within the fuzzy mask $NM(Z)$ is straightforward: Find the maximum SUV within the mask where the membership $NM(Z)(v)$ is non-zero. For estimating $SUV_{mean}(Z, I^{PET})$, we take a similar approach: the mean is computed overall voxels in $NM(Z)$ where the membership value is non-zero and the SUVs are weighted by the disease map value.

In summary, our whole AAR-LN-DQ approach to quantify LN disease by zones consists of four distinct stages: One-time model building which includes all processes related to collecting prior knowledge; recognition of zones; recognition of pathological nodes within each localized zone; and disease quantification within each zone. All key parameters in the entire process are learned in the model building/training stage. The only parameters that are hand-set are as follows. Related to node recognition via gF: r_{min}, r_{max}, and number of iterations k. These are currently set to 5, 12 , and 3 pixels, respectively. Related to node recognition via SegNet: the initial learning rate, minimal batch size of training samples, and the maximum number of training epochs. These are set to 0.001, 4, and 40, respectively. Thus, the whole gF methodology has three hand-set parameters and SegNet has three additional parameters whose values are fixed as above.

3. RESULTS AND DISCUSSION

The datasets for our experiments listed in Table I are used as follows. DS1 (near-normal CT): Used for building $FAM(B)$ and testing zonal recognition with a train-test division as explained below. DS2 (patient PET/CT): Used for nodal recognition and disease quantification, where a portion of DS2 is used for model building (estimating the parameters of the disease map $d_Z(x)$) and the remainder is used for testing as explained below.

3.A. Lymph node zone recognition

For this experiment, dataset DS1 is utilized. Thirty studies from DS1 are utilized for building $FAM(B)$ and the remaining 12 studies are used for testing recognition accuracy. This entire process is repeated six times each time by selecting a different 30-12 partition to yield a total of 72 test cases. Note that in the first fold, the 30 studies are also used for finding the best hierarchy following the procedure described in Section 2.B.3. Subsequently, this hierarchy is used in all folds and in all other experiments (including node recognition) for testing. The resulting best hierarchy is displayed in Fig. 9.

In Table IV, we summarize the zonal recognition results where the mean and standard deviation of errors over the tested experiments from the known true zonal definitions are listed. We employ two metrics to express this error for a zone Z: Location error (LE), defined as the distance (in mm) of the geometric center of the fuzzy model $FM^T(Z)$ of Z at recognition from the geometric center of the ground truth of Z; and scale error (SE), defined as the ratio of the estimated size of Z at recognition to its true size. Figure 10 displays sample zone recognition results.

We note that zones are localized within 2–3 voxels with respect to the ground truth location and the scale error is mostly close to 1 (ideal value). We feel that this is quite remarkable given that there are no intensity boundaries that define the spatial extent of the zones. Compared to the results reported in Ref. [3], the results have improved considerably by the use of the optimized hierarchy approach. In our visual assessment of the recognition results (Fig. 10), we find generally that the fuzzy model masks at recognition cover the zones very well even when LE is about three voxels. This is all that matters for finding (recognizing) pathological nodes within localized zones.

3.B. Lymph node recognition

Recall that the two fundamental differences between AAR-LN-DQ and other published methods on LN detection are: (a) AAR-LN-DQ takes a global-to-local approach by explicitly modeling and recognizing the zones and then localizing the nodes within the already located zones, and (b) it focuses only on pathologic nodes since our goal is disease quantification within zones. As such, the LN recognition step

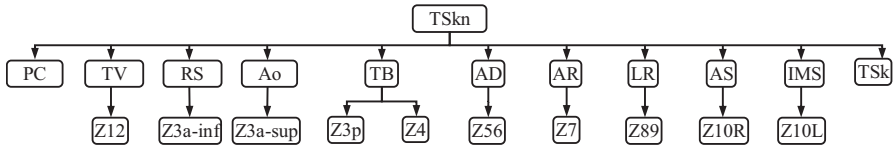

FIG. 9. Optimal hierarchy arrived at by AAR-LN-DQ Approach. For object abbreviations, see Table III.

TABLE IV. Mean and standard deviation of location error (LE) and scale error (SE) in zonal recognition.

	Z12	Z3a-sup	Z3a-inf	Z3p	Z4	Z56	Z7	Z89	Z10R	Z10L	Mean
LE (mm)	11.50	13.22	13.41	9.63	11.27	13.54	17.14	12.17	8.84	12.10	12.28
	1.50	2.27	1.89	1.44	2.18	2.91	3.27	1.83	1.21	1.48	1.99
SE	0.80	0.89	0.96	0.90	0.83	1.07	1.03	1.00	1.00	0.98	0.94
	0.01	0.04	0.02	0.02	0.03	0.02	0.02	0.01	0.02	0.02	0.02

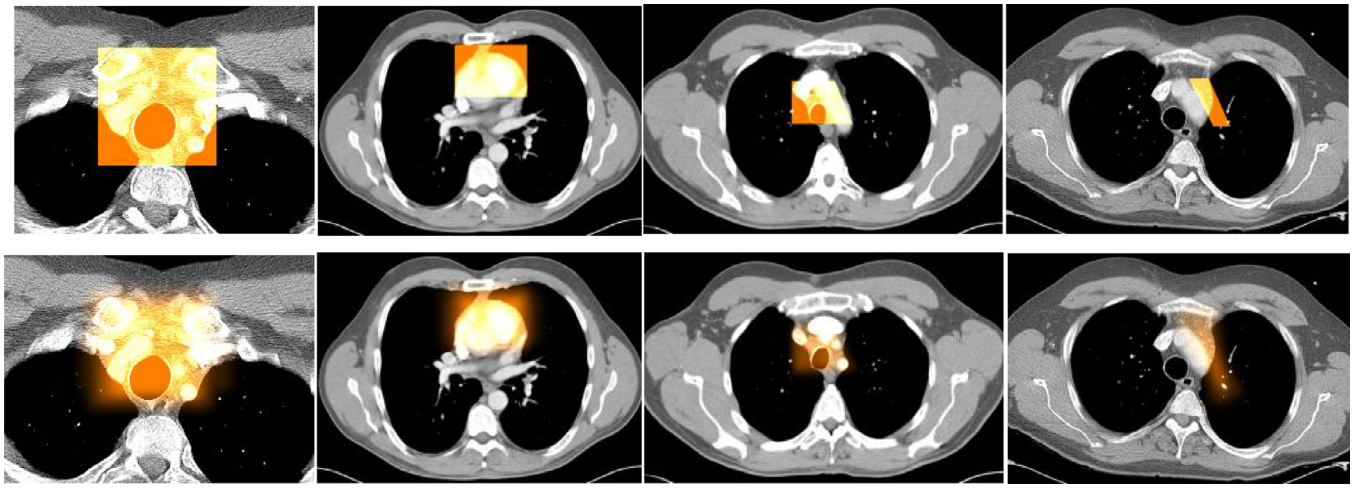

FIG. 10. Sample zone recognition results (2nd row) with the ground truth zonal extent (top row) also shown. The zones are displayed as overlay on the CT slices. Zones shown are (L to R): Z12, Z3a-inf, Z4 (R and L combined), and Z56 (with all its sub-zones combined). Note how small and subtle some zones are. [Color figure can be viewed at wileyonlinelibrary.com]

takes as input the given PET/CT image pair (I^{CT}, I^{PET}) of a patient, localized zonal mask $FM^T(Z)$ of each zone Z, and localized masks $FM^T(O)$ for two special organs, namely, the heart (Hrt) and the skeleton (TSk). The last entity is needed because Hrt and (bone marrow portion of) TSk act as confounding objects in our disease quantification quest, as such these objects have to be suppressed from the recognized zones that contain them, even partially. The output of the LN recognition step is a fuzzy nodal mask $NM(Z)$ where the fuzzy membership indicates the degree of belongingness of every voxel in Z in some pathologic node within Z.

For evaluating node recognition, dataset DS2 is employed (recall that DS1 is composed of studies from near-normal subjects). However, the $FAM(B)$ model built from DS1 in the first fold is utilized for performing zone recognition needed as the first step before nodal recognition on DS2. A fivefold break-up of DS2 is designed: 53 for training procedures, the remaining ten for testing, and the entire process is repeated five times, to yield a total of 50 non-overlapping test cases. Training procedures correspond to training the SVM parameters in the gF procedure and training SegNet. Since the studies in DS2 contained pathological nodes mainly in the mediastinal region and not in Z10R and Z10L, for nodal recognition we will focus on the rest of the zones Z1 through Z9 in our set $\mathcal{Z}$ of all zones. Moreover, since each individual zone in this set did not contain a sufficiently large number of samples, for evaluating nodal recognition, instead of examining each zone separately, we created one composite zone named Z19 by combining all zones: Z19 = Z1 + Z2R + . . . + Z89. Note, however, that zone recognition still followed the process described above; that is, each of the component zones was recognized separately and the resulting fuzzy model masks $FM^T(Z)$ were utilized in procedure gF and SegNet.

We established the ground truth for node recognition via PET/CT image reading by a board-certified radiologist (coauthor Torigian) with > 20 yr of clinical and research experience in thoracic, oncologic, and cross-sectional (CT, PET, magnetic resonance (MR)) imaging. The actual image location of each pathological node identified in this manner was

recorded for the purpose of evaluating node recognition and disease quantification. Our datasets contained a total of 214 pathological nodes in the 63 patient cases.

Sample node recognition results are shown in Fig. 11 for both gF and Seg-Net-2 methods where the recognized zonal model masks as well as the final nodal recognition masks $NM(Z)$ are shown overlaid on sample CT and PET slices. Nodal recognition results from our evaluation are summarized in Table V. We make the following observations from these quantitative results as well as our qualitative examinations. (a) Zonal recognition captures ~96% of the pathological nodes, suggesting that AAR zone recognition is very effective, although it would be ideal if recognition can be improved to reach ~100% capture rate. (b) The g-filter and the iterative strategy in gF seem to help considerably in minimizing false negatives. Beyond the third iteration, improvement in sensitivity was not significant. Therefore, for disease quantification via gF, we take the output at $k = 3$. Interestingly, for the gF method, false positives are not a challenge, achieving a level of <1.5%. This is a real strength of the gF approach. (c) All versions of SegNet achieve higher sensitivity than the gF approach (with statistical significance, $P < 0.001$) for recognizing pathological LNs, reaching over 91–96%, roughly 9% better than the gF approach, while the latter slightly (by about 2.5%) outperforms SegNet in specificity (also with statistical significance, $P < 0.001$). No statistically significant difference was found among the SegNet versions in sensitivity or specificity. (d) For disease quantification, as we will see below, the fuzzy masks $NM(Z)$ found by the gF approach cover the pathological nodes much better than any version of SegNet, resulting in higher accuracy of disease quantification via gF than SegNet.

The g-filter method is natural-intelligence (NI)-driven where we embed into its process prior human knowledge of different sorts explicitly such as shape and location of confounding objects, shape of the LNs, and radius and response information. As such, it achieves high TP rate to detect LNs. Subsequently, SVM effectively removes false positive balls/voxels and refines recognition results. This NI-driven strategy improves LN recognition performance. In addition, the g-filter method has two attributes akin to deep neural networks: layering and iteration. Layering comes from the division of classification task into three levels: ball-level, slice-level, and voxel-level. Iteration represents repetition of these three layered levels to refine LN recognition. With these strategies, the g-filter method shows some advantages in performance over SegNet.

TABLE V. Results for nodal recognition. Mean and standard deviation over the tested cases and folds are shown.

	TP rate	Sensitivity	Specificity
Zone recognition	0.958	0.985	0.752
gF			
$k = 0$	0.907	0.459	0.682
$k = 1$	0.928 ± 0.05	0.383 ± 0.03	0.997 ± 0.01
$k = 2$	0.928 ± 0.05	0.809 ± 0.11	0.988 ± 0.01
$k = 3$	0.928 ± 0.05	0.841 ± 0.11	0.985 ± 0.01
SegNet-2	0.908 ± 0.11	0.913 ± 0.11	0.961 ± 0.02
SegNet-4	0.919 ± 0.06	0.927 ± 0.08	0.958 ± 0.01
SegNet-6	0.936 ± 0.04	0.930 ± 0.06	0.961 ± 0.01

3.C. Disease quantification

For evaluating the disease quantification method, we utilize dataset DS2, although DS1 comes into play in the form of the anatomy model $FAM(B)$ created from it. The only training component specific to disease quantification is related to the estimation of the parameters of the disease map $d_Z(x)$. For this task, we follow the same fivefold cross validation 53-10 train-test division of studies in DS2 as in nodal recognition.

The true disease quantities are described by

$$Q_X^t(Z) = [SUV_{mean}^t(Z, I^{PET}), SUV_{\max}^t(Z, I^{PET}), fTLG^t(Z, I^{PET})]. \quad (6)$$

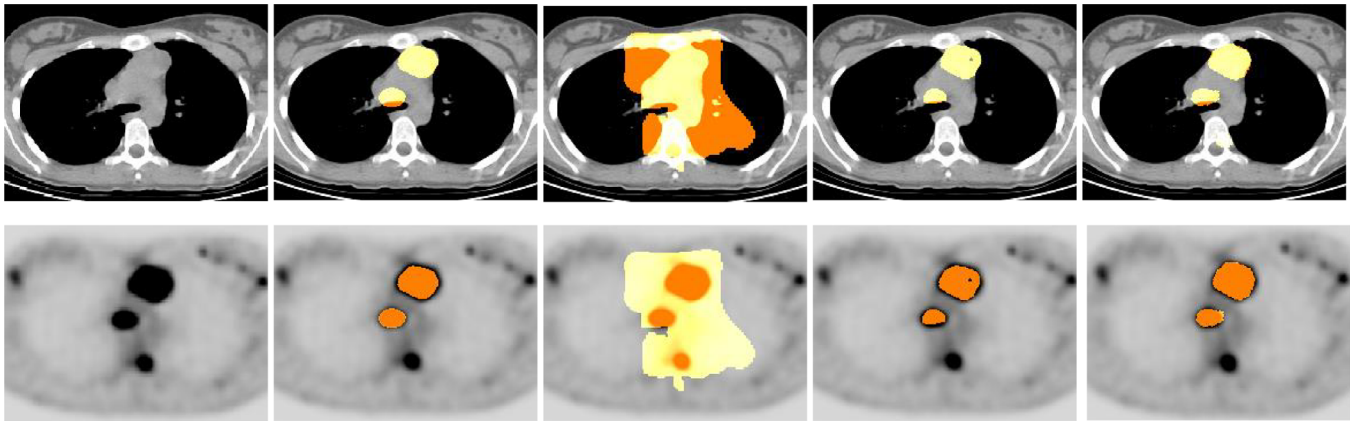

FIG. 11. Sample nodal recognition results for both gF and SegNet methods overlaid on CT (top) and PET (bottom) slices. L to R: A CT/PET slice from one study; CT/PET slice with ground truth masks of pathological nodes; CT/PET slice with recognized zonal mask region; CT/PET slice with the recognized nodal mask $NM(Z)$ output by procedure gF; CT/PET slice with recognized nodal mask output by SegNet-2. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE VI. Mean and standard deviation of the error (ε_1) in estimating SUV_{mean} over all tested cases for the different methods with and without the consideration of disease map and confounding objects Hrt and TSk.

	gF		SegNet-2		SegNet-4		SegNet-6	
Disease map	N	Y	N	Y	N	Y	N	Y
With Hrt & TSk	0.10	0.11	0.31	0.10	0.38	0.34	0.34	0.10
	0.07	0.03	0.14	0.01	0.09	0.08	0.10	0.01
Without Hrt & TSk	0.11	0.10	0.11	0.10	0.09	0.07	0.30	0.05
	0.05	0.05	0.05	0.05	0.02	0.07	0.09	0.07

TABLE VII. Mean and standard deviation of the error (ε_2) in estimating SUV_{max} over all tested cases for the different methods with and without the consideration of disease map and confounding objects Hrt and TSk.

	gF		SegNet-2		SegNet-4		SegNet-6	
Disease map	N	Y	N	Y	N	Y	N	Y
With Hrt & TSk	0.05	0.04	0.16	0.16	0.16	0.16	0.16	0.16
	0.10	0.10	0.07	0.07	0.07	0.07	0.07	0.07
Without Hrt & TSk	0.06	0.05	0.15	0.15	0.15	0.15	0.15	0.14
	0.09	0.09	0.06	0.06	0.06	0.06	0.06	0.06

Since the ground truth masks for the pathological nodes within any zone Z are known, $SUV^t_{mean}(Z, I^{PET})$ and $SUV^t_{\max}(Z, I^{PET})$ can be computed in a straightforward manner. $fTLG^t(Z, I^{PET})$ is computed by using a modified form of Eq. (3):

$$fTLG^t(Z, I^{PET}) = |v| \sum_v I_S(v) NM^t(Z)(v), \tag{7}$$

where $NM^t(Z)$ denotes the union of the binary masks for all pathological nodes within zone Z, and $NM^t(Z)(v)$ is the value of the binary mask (0 or 1) at voxel v within Z.

We express component-wise error in $Q_X(Z, I^{PET})$ as deviation with respect to the ground truth value in that component:

$$\varepsilon(Z, I^{PET}) = [\varepsilon_1(Z, I^{PET}), \varepsilon_2(Z, I^{PET}), \varepsilon_3(Z, I^{PET})],$$

where

$$\varepsilon_1(Z, I^{PET}) = \frac{SUV_{mean}(Z, I^{PET}) - SUV^t_{mean}(Z, I^{PET})}{SUV^t_{mean}(Z, I^{PET})}, \tag{8}$$

TABLE VIII. Mean and standard deviation of the error (ε_3) in estimating $fTLG$ over all tested cases for the different methods with and without the consideration of disease map and confounding objects Hrt and TSk.

	gF		SegNet-2		SegNet-4		SegNet-6	
Disease map	N	Y	N	Y	N	Y	N	Y
With Hrt & TSk	0.31	0.24	0.73	0.35	0.74	0.33	0.75	0.34
	0.11	0.09	0.32	0.13	0.35	0.13	0.49	0.14
Without Hrt & TSk	0.14	0.08	0.42	0.14	0.41	0.12	0.42	0.14
	0.12	0.09	0.26	0.10	0.25	0.09	0.26	0.10

and ε_2 and ε_3 are defined similarly for the second and third components, respectively, of $Q_X(Z, I^{PET})$.

In Tables VI–VIII, we summarize the mean and standard deviation of ε_1, ε_2, and ε_3, respectively, over all tested cases and folds for both gF and the SegNet methods. As for evaluation of nodal recognition and for the same reasons, we performed evaluation of DQ for the combined zone *Z19*. To understand the influence of the confounding objects Hrt and TSk and disease map, we show error statistics with and without the consideration of these two factors. We make the following observations based on Tables V–VII. (a) Overall, gF achieves the best performance with better accuracy for SUV_{max} and *fTLG* than SegNet and an accuracy for SUV_{mean} similar to that of SegNet methods. (b) gF also seems to be overall more stable than other methods with smaller standard deviation, especially in estimating *fTLG*. (c) Removal of confounding objects Hrt and TSk via AAR is one of the key factors responsible for improving DQ accuracy, particularly for estimating *fTLG*, where the improvement is 17–33%. Interestingly, this gain is the least for the gF approach. (d) The use of disease map also boosts accuracy, also specifically for *fTLG*, where the gain (6–29%) is larger for SegNet methods than gF. Notably, disease map also makes *fTLG* estimation substantially more robust by lowering the standard deviation, where again SegNet methods gain more than gF.

3.C.1. Computational considerations

The computing platform consists of an i7-Core CPU with 64 Gbyte RAM and one NVidia Titan XP GPU (12G GDDR5X memory with 3840 CUDA cores) running under Ubuntu 18.04 OS.

Computational times were as follows. Time for AAR model building once the optimal hierarchy is determined: ~4 h. Time for training each network: ~6 h. Time for zone recognition per zone: ~30 s. Time for nodal recognition/zone: ~30 s. Time for DQ per study: ~20 s. Total time for analyzing one study: ~80 s.

4. CONCLUSIONS

Inspired by our recent work[10] on quantifying disease via FDG-PET/CT in organs without explicitly delineating them, in this paper we extended that approach to LN zones and showed its application in the thorax. LN zones differ majorly from organs in that they are not manifest with visually perceptible boundaries in the image but are present as virtual conceptual 3D regions. We devised a recognition strategy tailored to LN zones based on the AAR framework to achieve high enough accuracy so that the subsequent step of DQ affords acceptable accuracy. We creatively combined the high-level AAR model-based strategy to localize zones with a specially designed filter, called globular filter, to recognize only pathological nodes within the already recognized LN zones, but also excluding confounding objects like heart and bone marrow localized via AAR. The DQ operation is then

performed with a disease mapping strategy within the zone excluding the confounding objects. We also used several versions of a deep network trained on the recognized LN zones to localize pathologic nodes. To the best of our knowledge, no demonstrated method exists for automatically estimating the triple disease quantities within LN zones in the thorax, or any other body region. The proposed AAR-LN-DQ approach shows that it is feasible to perform this estimation accurately, robustly, and fully automatically.

Since the method is firmly entrenched in AAR whose generalizability to different body regions has been already demonstrated, AAR-LN-DQ can also be generalized to other body regions (neck, abdomen, pelvis) readily once the anatomy model $FAM(B)$ can be generated for that body region B and the appropriate confounding objects, such as kidneys and bladder, are identified and included in the model.

For methods of LN detection and delineation,[12–19,33] false positives are a challenge. Interestingly, the specificity of locality afforded by AAR via the recognition process mitigates this problem substantially and allows the gF approach to reach high specificity of nodal recognition of 98.5%. This high-level human knowledge is a challenge to deep networks as well, as well as the difficulties posed by the confounding objects. Our approach of marrying model-based high-level AAR zonal recognition excluding the confounding objects with the ability of deep networks to garner low-level details shows how model-based strategies can boost network performance when designed properly with human insight into the problem. The idea of the disease map is another strong feature of AAR-LN-DQ which shows that accuracy and robustness can be both improved for both gF and SegNet approaches.

There is room for improvement of sensitivity for the gF approach which may further improve DQ accuracy. Among the three parameters r_{min}, r_{max}, and number of iterations k, $k = 3$ seems adequate since no improvement is observed for higher values. The current values set, $r_{min} = 5$ pixels and $r_{max} = 12$ pixels, can be changed to expand the set of template balls which may enhance sensitivity. However, it is unknown as to how this may influence the current high specificity of nodal recognition and the resulting DQ accuracy.

Another gap in the present work is that zones have been grouped together for certain operations (mainly DQ). This was necessitated mainly to garner enough statistics to perform meaningful analysis although the whole methodology can be executed zone by zone without any conceptual hurdles. With a sufficiently large number of datasets available with ground truth and with balanced distribution of pathologic nodes by zones, we will be able to gain an understanding of zone-wise accuracy in nodal recognition and DQ.

Finally, in this paper we focused on FDG-PET/CT and assumed that pathologic nodes are manifested by high FDG uptake. As such, lymph nodes that are involved by disease but which do not demonstrate increased radiotracer uptake will lead to false negative results if based on PET alone. Similarly, lymph nodes that are involved by non-neoplastic disease (e.g., inflammation) may manifest with increased FDG uptake, leading to false positive results. These limitations may be mitigated by taking into account CT-based properties of lymph nodes (e.g., size, volume, shape, etc.) to increase specificity. Also, the current work was studied and implemented only in the context of FDG-PET/CT images. The approach may also be extended to the assessment of lymph nodes on PET/CT scans using other radiotracers.

ACKNOWLEDGMENTS

The training of Mr. Guoping Xu in the Medical Image Processing Group, the Department of Radiology, University of Pennsylvania, Philadelphia, for the duration of one year was supported by the China Scholarship Council. His subsequent training was supported by research funds from Torigian.

[#]Main designer of the entire study and main contributor to manuscript preparation.
[a)]Author to whom correspondence should be addressed. Electronic mail: jay@pennmedicine.upenn.edu; Telephone: 215-746-8627.

REFERENCES

1. Nogues I, Lu L, Wang X, et al. Automatic lymph node cluster segmentation using holistically-nested neural networks and structured optimization in CT images. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Berlin: Springer; 2016:388–397.
2. Rusch VW, Asamura H, Watanabe H, Giroux DJ, Rami-Porta R, Goldstraw P. The IASLC lung cancer staging project: a proposal for a new international lymph node map in the forthcoming seventh edition of the TNM classification for lung cancer. *J Thorac Oncol*. 2009;4:568–577.
3. Matsumoto MMS, Beig NG, Udupa JK, Archer S, Torigian DA. Automatic localization of IASLC-defined mediastinal lymph node stations on CT images using fuzzy models. In: Medical Imaging 2014: Computer-Aided Diagnosis. International Society for Optics and Photonics; 2014. page 90350J.Available from: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2044333
4. Liu Y, Udupa JK, Odhner D, et al. Definition and automatic anatomy recognition of lymph node zones in the pelvis on CT images. In: Medical Imaging 2016: Biomedical Applications in Molecular, Structural, and Functional Imaging. International Society for Optics and Photonics; 2016. page 97881J.
5. Xu G, Udupa JK, Tong Y, et al. Thoracic lymph node station recognition on CT images based on automatic anatomy recognition with an optimal parent strategy. In: Medical Imaging 2018: Image Processing. International Society for Optics and Photonics; 2018. page 105742F.
6. Feuerstein M, Glocker B, Kitasaka T, Nakamura Y, Iwano S, Mori K. Mediastinal atlas creation from 3-D chest computed tomography images: application to automated detection and station mapping of lymph nodes. *Med Image Anal*. 2012;16:63–74.
7. Liu J, Zhao J, Hoffman J, et al. Detection and station mapping of mediastinal lymph nodes on thoracic computed tomography using spatial prior from multi-atlas label fusion. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI); 2014. page 1107–10. Available from http://ieeexplore.ieee.org/document/6868068/.
8. Hoffman J, Liu J, Turkbey E, Kim L, Summers RM. Automatic identification of IASLC-defined mediastinal lymph node stations on CT scans using multi-atlas organ segmentation; 2015 (March 2015):94141R. Available from http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2082190

9. Udupa JK, Odhner D, Zhao L, et al. Body-wide hierarchical fuzzy modeling, recognition, and delineation of anatomy in medical images. *Med Image Anal*. 2014;18:752–771.
10. Tong Y, Udupa JK, Odhner D, Wu C, Schuster SJ, Torigian DA. Disease quantification on PET/CT images without explicit object delineation. *Med Image Anal*. 2019;51:169–183.
11. Wu X, Udupa JK, Tong Y, et al. AAR-RT – a system for auto-contouring organs at risk on CT images for radiation therapy planning: principles, design, and large-scale evaluation on head-and-neck and thoracic cancer cases. *Med Image Anal*. 2019;54:45–62.
12. Nakamura Y, Nimura Y, Oda M, et al. Ensemble lymph node detection from CT volumes combining local intensity structure analysis approach and appearance learning approach. In: Progress in Biomedical Optics and Imaging - Proceedings of SPIE; 2016.
13. Oda H, Bhatia KK, Oda M, et al. Automated mediastinal lymph node detection from CT volumes based on intensity targeted radial structure tensor analysis. *J Med Imaging*. 2017;4:1.
14. Liu J, Hoffman J, Zhao J, et al. Mediastinal lymph node detection and station mapping on chest CT using spatial priors and random forest. *Med Phys*. 2016;43:4362–4374.
15. Barbu A, Suehling M, Xu X, Liu D, Zhou SK, Comaniciu D. Automatic detection and segmentation of lymph nodes from CT data. *IEEE Trans Med Imaging*. 2012;31:240–250.
16. Feulner J, Kevin Zhou S, Hammon M, Hornegger J, Comaniciu D. Lymph node detection and segmentation in chest CT data using discriminative learning and a spatial prior. *Med Image Anal*. 2013;17:254–270.
17. Seff A, Lu L, Cherry KM, et al. 2D view aggregation for lymph node detection using a shallow hierarchy of linear classifiers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention; 2014. page 544–552.
18. Cherry KM, Wang S, Turkbey EB, Summers RM. Abdominal lymphadenopathy detection using random forest. In: Medical Imaging 2014: Computer-Aided Diagnosis; 2014: page 90351G.Available from http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2043837
19. Nimura Y, Hayashi Y, Kitasaka T, Furukawa K, Misawa K, Mori K. Automated abdominal lymph node segmentation based on RST analysis and SVM. In: Medical Imaging 2014: Computer-Aided Diagnosis; 2014. page 90352U.Available from http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2043349
20. Roth HR, Lu L, Seff A, et al. A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Berlin: Springer; 2014:520–527.
21. Shin H, Roth HR, Gao M, et al. Deep learning and convolutional neural networks for medical image computing. Adv Comput Vis Pattern Recognit. 2017;113–136. http://link.springer.com/10.1007/978-3-319-42999-1
22. Oda H, Bhatia KK, Roth HR, et al. Dense volumetric detection and segmentation of mediastinal lymph nodes in chest CT images. In: Medical Imaging 2018: Computer-Aided Diagnosis. International Society for Optics and Photonics; 2018:1057502.
23. Tang Y, Oh S, Xiao J, Summers RM, Tang Y. CT-realistic data augmentation using generative adversarial network for robust lymph node segmentation. In: Medical Imaging 2019: Computer-Aided Diagnosis. International Society for Optics and Photonics; 2019:109503V.
24. Bouget D, Jørgensen A, Kiss G, Leira HO, Langø T. Semantic segmentation and detection of mediastinal lymph nodes and anatomical structures in CT data for lung cancer staging. *Int J Comput Assist Radiol Surg*. 2019;14:977–986.
25. Song Q, Bai J, Han D, et al. Optimal co-segmentation of tumor in PET-CT images with context information. *IEEE Trans Med Imaging*. 2013;32:1685–1697.
26. Grossiord E, Talbot H, Passat N, Meignan M, Najman L. Automated 3D lymphoma lesion segmentation from PET/CT characteristics. Proc Int Symp Biomed Imaging; 2017;174–178.
27. Zhao X, Li L, Lu W, Tan S. Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network. *Phys Med Biol*. 2019;64:015011.
28. Jin D, Guo D, Ho T-Y, et al. Accurate esophageal gross tumor volume segmentation in pet/ct using two-stream chained 3D deep network fusion. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Berlin: Springer; 2019:182–191.
29. Hofheinz F, Potzsch C, Jorg VDH. Quantitative 3D ROI volume delineation in PET: algorithm and validation. *J Nucl Med*. 2007;48:407P.
30. Hofheinz F, Dittrich S, Pötzsch C, Van Den Hoff J. Effects of cold sphere walls in PET phantom measurements on the volume reproducing threshold. *Phys Med Biol*. 2010;55:1099.
31. Foruzan AH, Zoroofi RA, Sato Y, Hori M. A Hessian-based filter for vascular segmentation of noisy hepatic CT scans. *Int J Comput Assist Radiol Surg*. 2012;7:199–205.
32. Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell*. 2017;39:2481–2495.
33. Torigian DA, Lopez RF, Alapati S, et al. Feasibility and performance of novel software to quantify metabolically active volumes and 3D partial volume corrected SUV and metabolic volumetric products of spinal bone marrow metastases on 18F-FDG-PET/CT. *Hell J Nucl Med*. 2011;14:8–14.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1. Definitions of lymph node zones.