Medical Image Analysis 69 (2021) 101980

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Segmentation evaluation with sparse ground truth data: Simulating true segmentations as perfect/imperfect as those generated by humans

Jieyu Li[a,b], Jayaram K. Udupa[b,*], Yubing Tong[b], Lisheng Wang[a], Drew A. Torigian[b]

[a] Institute of Image Processing and Pattern Recognition, Department of Automation, Shanghai Jiao Tong University, 800 Dongchuan RD, Shanghai, 200240, China

[b] Medical Image Processing Group, Department of Radiology, University of Pennsylvania, 602 Goddard building, 3710 Hamilton Walk, Philadelphia, PA, 19104, United States

ARTICLE INFO

Article history:
Received 4 February 2020
Revised 19 January 2021
Accepted 20 January 2021
Available online 26 January 2021

Keywords:
Medical image segmentation
Ground truth generation
Inter-segmenter variability
Segmentation evaluation

ABSTRACT

Fully annotated data sets play important roles in medical image segmentation and evaluation. Expense and imprecision are the two main issues in generating ground truth (GT) segmentations. In this paper, in an attempt to overcome these two issues jointly, we propose a method, named SparseGT, which exploit variability among human segmenters to maximally save manual workload in GT generation for evaluating actual segmentations by algorithms. Pseudo ground truth (p-GT) segmentations are created by only a small fraction of workload and with human-level perfection/imperfection, and they can be used in practice as a substitute for fully manual GT in evaluating segmentation algorithms at the same precision.

p-GT segmentations are generated by first selecting slices sparsely, where manual contouring is conducted only on these sparse slices, and subsequently filling segmentations on other slices automatically. By creating p-GT with different levels of sparseness, we determine the largest workload reduction achievable for each considered object, where the variability of the generated p-GT is statistically indistinguishable from inter-segmenter differences in full manual GT segmentations for that object. Furthermore, we investigate the segmentation evaluation errors introduced by variability in manual GT by applying p-GT in evaluation of actual segmentations by an algorithm.

Experiments are conducted on ~500 computed tomography (CT) studies involving six objects in two body regions, Head & Neck and Thorax, where optimal sparseness and corresponding evaluation errors are determined for each object and each strategy. Our results indicate that creating p-GT by the concatenated strategy of uniformly selecting sparse slices and filling segmentations via deep-learning (DL) network show highest manual workload reduction by ~80-96% without sacrificing evaluation accuracy compared to fully manual GT. Nevertheless, other strategies also have obvious contributions in different situations. A non-uniform strategy for slice selection shows its advantage for objects with irregular shape change from slice to slice. An interpolation strategy for filling segmentations can achieve ~60-90% of workload reduction in simulating human-level GT without the need of an actual training stage and shows potential in enlarging data sets for training p-GT generation networks. We conclude that not only over 90% reduction in workload is feasible without sacrificing evaluation accuracy but also the suitable strategy and the optimal sparseness level achievable for creating p-GT are object- and application-specific.

1. Introduction

1.1. Background

Numerous 3D anatomy segmentation methods have emerged since the advent of tomographic imaging modalities in the 1970s. Early methods were based purely on the information available in the image to be segmented (Herman et al., 1979; Liu, 1977). Since they did not harvest information available via anatomic priors, they needed ground truth (or reference) segmentations only for segmentation evaluation. Although such approaches continue to seek new frontiers, methods that exploit priors in various forms have emerged during the past 2-3 decades and have shown significant gain in segmentation robustness and accuracy. These later methods may be generically referred to as model-based since they employ some form of model to encode prior information. Such models include shape and geographic models (Cootes et al., 1995; Pizer et al., 2003; Shen et al., 2011; Staib and Duncan, 1992; Udupa et al.,

* Corresponding author: Medical Image Processing Group, Department of Radiology, 3710 Hamilton Walk, 6th Floor, Rm 602W, Philadelphia PA 19104, United States.
E-mail address: jay@pennmedicine.upenn.edu (J.K. Udupa).

https://doi.org/10.1016/j.media.2021.101980
1361-8415/

2014), atlases (Ashburner and Friston, 2009; Christensen et al., 1994; Chu et al., 2013; Gee et al., 1993; Shi et al., 2017), and deep neural network models (Agn et al., 2019; Cerrolaza et al., 2019; Drozdzal et al., 2018; Moeskops et al., 2016; Wang et al., 2019a). However, for these methods, fully annotated ground truth (GT) segmentations that capture the very variability over a human population of focus they purport to encode is of fundamental importance, some of them requiring a large number of such data sets for robust model building alone, not to mention for evaluation as well.

There are two main issues with generating GT segmentations: expense and imprecision. GT reference is most typically provided by manual (human expert) contouring of anatomical objects in medical imaging. Thus, generating large fully manual GT sets becomes impractical and *expensive* (Schipaanboord et al., 2018). Unsupervised and semi-supervised methods may be utilized that do not require fully annotated data sets which can ease somewhat the expense issue. However, accuracy and convergence in learning are superior with supervised learning than with semi-supervised or unsupervised methods (Park et al., 2019). Crowdsourced non-expert annotation can be another solution to the expense issue. However, the second issue – *imprecision* – that naturally and commonly exists among expert annotations becomes more pronounced when non-expert annotations are employed. Several factors (Joskowicz et al., 2019) lead to imprecision in manually annotated GT segmentations including image quality issues, lack of standard ways of defining objects or variations in the interpretation of (pseudo) standards (when they exist), human subjectivity in interpreting boundaries in images, and institution- and application-specific vagaries in clinical contouring culture. The magnitude of these imprecisions is object-specific and application-specific. Small, non-compact, and sparse objects entail larger degrees of imprecision inversely proportionate to their size compared to large, well-defined, and compact objects. The Expectation-Maximization-based STAPLE (Warfield et al., 2004) framework and its extensions are a series of methods commonly used to generate consensus GT from multiple manual segmentations. However, unsupervised/semi-supervised methods and generating consensus segmentations deal with only one of the two key issues and not both simultaneously.

1.2. Related work

Numerous works have investigated solutions to the issues of expense and imprecision in generating GT segmentations. Bounding box (Rajchl et al., 2016) or partial annotation via partial slices or scribble strategies (Can et al., 2018; Çiçek et al., 2016; Koch et al., 2017) are used in semi-supervised learning based algorithms in medical image segmentation, while it has been pointed out that unsupervised and semi-supervised methods cannot surpass the accuracy of fully supervised methods (Rajchl et al., 2016; Papandreou et al., 2015; Wang et al., 2019b). Cost-effective annotation methods deal with the expense issue by active learning or interactive segmentation to iteratively improve performance of segmentation models (Tajbakhsh et al., 2020), where active learning methods select most representative and uncertain samples to generate as few GT segmentations as possible (Yang et al, 2017), and interactive segmentation methods modify the auto-segmentation manually and fine-tune the model in a sample-specific manner (Wang et al., 2018). They are not suitable to generate a large GT set which is needed in evaluation.

Machine learning and deep learning methods have been utilized to conduct segmentation evaluation and estimate evaluation metrics without ground truth. A SVM regressor (Kohlberger et al., 2012) trained by a space of shape and appearance features or regression neural networks trained directly from binary and intensity images (Robinson et al., 2018) or the difference images of intensity images reconstructed from binary segmentations (Zhou et al., 2020) are utilized to yield metrics without knowing GT delineation. Reverse testing (Bhaskaruni et al., 2018) also contributed to segmentation evaluation without GT in medical images. (Valindria et al., 2017) introduce a reverse classification accuracy (RCA) framework which took predicted segmentations on test samples as pseudo ground truth and train a classifier to reversely test on labeled training samples. The predicted segmentations are considered as of good quality if the trained classifier works well on at least part of the training samples. Extensions of RCA are also utilized to estimate Dice evaluation metric values in cardiac magnetic resonance (MR) images in (Robinson et al., 2017,2018, 2019). Instead of directly generating metrics for a single method or a single segmentation, there are also unsupervised methods to compare effectiveness of different segmentation algorithms by computational statistical measures (Chabrier et al., 2006) or region-correlation matrix (Sikka and Deserno, 2010).

Besides the expense issue, the imprecision issue also naturally and commonly exists in the human-drawn ground truth segmentations (Joskowicz et al., 2019; Park et al., 2019; Sharp et al., 2014; Yang et al., 2018) and will lead to different segmentation results and metric values when the samples are evaluated on ground truth generated by different algorithms (Lampert et al., 2016) or annotated by segmenters different from those employed for training samples (Shwartzman et al., 2019). (Heller et al., 2018) investigated the influence the errors in labels may have on the segmentation quality. (Bø et al., 2017) demonstrated that even for radiologists, intra-segmenter differences still exist depending on their familiarity with the segmentation tool. To minimize inter-segmenter disagreements and generate more precise consensus ground truth, segmentations from multiple human segmenters are utilized by averaging (Cheplygina and Pluim, 2018), majority voting (Nowak and Rüger, 2010; O'Neil et al., 2017), maximizing topological agreements (Yang and Choe, 2011), and Expectation-Maximization-based STAPLE (Warfield et al., 2004) method and its extensions (Gordon et al., 2009; Li et al., 2011; Schlesinger et al., 2017; Shwartzman et al., 2019). Since generating expert-level high quality annotations is a time-consuming and expensive task, methods have been investigated to evaluate the quality of crowd-sourced non-expert annotations and fuse crowd-sourced labels (Gurari et al., 2015). (Cheplygina and Pluim, 2018) observed that although the agreement from crowd-sourced annotations is best when utilized in medical image analysis, the disagreement of segmenters is also informative. Inter-segmenter differences provide good reference for algorithm evaluation (Joskowicz et al., 2019; Park et al., 2019; Popović and Thomas, 2017) and also can be utilized to estimate regions with variability (Zhou et al., 2020) or uncertainty (Jungo et al., 2018) for helping in making clinical decisions.

Expense is purely a labor/cost issue. *Imprecision*, however, raises several conceptual issues. Although great strides have been made in examining these dual issues in the literature separately as delineated above, they have not been examined jointly or one as a function of the other. This area calls for a lot more attention in view of the promises suggested by deep neural network models. Most importantly, the practical question of the savings that ensue in cost as a function of the imprecision in GT data as a result of its "sparsification" has not been examined so far. In other words, is it feasible to simulate full GT segmentations from sparse GT data such that the simulated GT is as perfect/ imperfect as, but not worse than, the GT generated by human experts? The cost saving then will directly depend on the degree of sparsity affordable for the sparse GT data and will be tied with the second issue of imprecision. In this work, keeping the expense issue in mind and recognizing the fact that perfect ground truth does not exist, we take inter-segmenter variability as reference to create segmentations as

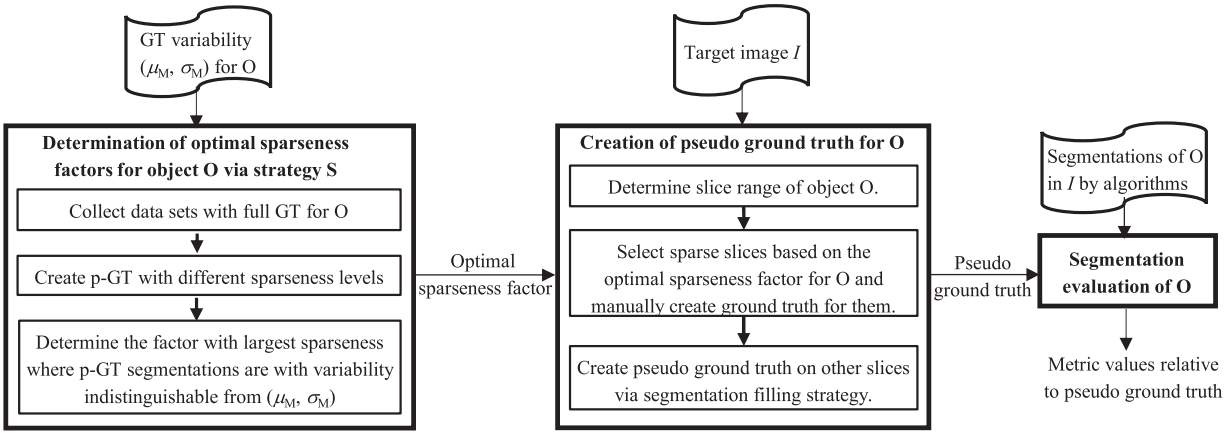

Fig. 1. A schematic representation of the SparseGT method.

perfect/imperfect as those produced by expert segmenters with the attendant reduction of manual workload via proper strategies for sparse slice selection and segmentation filling. The created segmentations should be able to replace full manual GT in segmentation evaluation. There is no work in the published literature to address this important and very practical issue.

1.3. Outline of approach

In this paper, we propose a method, named *SparseGT*, to deal with the expense and imprecision issues jointly by exploiting natural variability existing among human segmenters, generating pseudo ground truth (p-GT for short) from sparse manual segmentations, and utilizing p-GT in actual segmentation evaluation. The method is fully described in Section 2 and is schematically presented in Fig. 1. p-GT segmentation is generated in two sequential steps: selecting slices sparsely to conduct manual annotation and filling segmentations for other slices. Each of the two steps is investigated by two strategies, including equal interval based uniform strategy and shape change function based non-uniform strategy for slice selection, and shape-based interpolation strategy (Raya and Udupa, 1990) and object-specific 2D U-net (Ronneberger et al., 2015) based deep learning strategy for segmentation filling. The largest sparseness levels where p-GT segmentations for the object under consideration are statistically indistinguishable from the full manual GT, taking into account the variability existing among a group of expert segmenters, are considered as the object-specific optimal sparseness to determine the cost saving that is feasible for the SparseGT method. The created p-GT can generate accurate metric values as full manual GT did in segmentation evaluation.

In Section 3, we describe the experimental set up, anatomic objects and data sets utilized, results, and their analysis. Experiments are conducted on the four possible combinations of strategies for sparse slice selection and segmentation filling. The optimal sparseness factors are object-, strategy- and application-specific, and the optimal combinations of strategies are determined by examining the yielded actual workload reduction and their actual availability in practice. Our conclusions, gaps remaining in this work, and avenues for potential improvements are discussed in Section 4.

2. Method

Our description will follow the schematic in Fig. 1.

Notation:

$\mathcal{B}$: Body region of interest.

O: Anatomic object of interest in $\mathcal{B}$.

$\mathcal{I} = \{I_1, \ldots, I_N\}$: Image data sets of $\mathcal{B}$ for which complete GT delineations for O are available.

$\mathcal{I}_b = \{I_{1,b}, \ldots, I_{N,b}\}$: Binary images representing complete GT delineations of O in $\mathcal{I}$.

$\mathcal{J}_b = \{J_{1,b}, \ldots, J_{N,b}\}$: Binary images representing complete segmentations of O in $\mathcal{I}$ by a segmentation algorithm A.

$\mathcal{P}_b = \{P_{1,b}, \ldots, P_{N,b}\}$: Optimally generated p-GT corresponding to $\mathcal{I}_b$.

Abbreviations and acronyms commonly used in this paper are listed in Table 1 for ease of reference.

2.0. Determining slice range for object O

Let N_O be the number of slices covering an object O (such as the mandible) in a patient image that includes O. The coverage of O should be decided by expert segmenters to guarantee that a proper and consistent anatomic definition of O is adhered to. Without loss of generality, we assume that the slices are transaxial with respect to $\mathcal{B}$ and denote the cranio-caudal direction orthogonal to the slice plane by z.

2.1. Sparse slice selection

One of the main aims of the SparseGT method is to reduce manual workload needed in creating expert-quality p-GT for segmentation algorithm evaluation, where only a few out of all slices are selected to conduct manual annotation and segmentations on other slices are automatically created via segmentation filling strategies. Two strategies are investigated to determine sparse slice positions, as illustrated in Fig. 2. Along the z axis, slices of object O are sparsely selected in a *uniform* or *non-uniform* manner.

(1) *Uniform strategy for selecting sparse slices*

We use a parameter t to represent the degree of uniform sparseness, where one slice is selected every $(t+1)$ slices within the range of N_O slices occupied by the object sample, i.e., t slices are skipped without performing manual annotation between two adjacent selected sparse slices. An ideal t with least workload would be $t_3 = \lfloor (N_O - 3)/2 \rfloor$, where the middle slice and two end slices are selected and all slices in between are divided into two skipped blocks with t_3 slices. More generally, if we set $1 \le t_n \le t_3$, $N_S = \lfloor (N_O - 1)/(t_n + 1) \rfloor + 1$ slices are selected with uniform spacing. Fig. 2(a) demonstrates an example for uniform sparse slice selection, where the object sample occupies $N_O = 23$ slices and $t = 4$ is determined as the sparseness factor, and only $N_S = \lfloor (23-1)/(4+1) \rfloor + 1 = 5$ slices are selected to conduct manual annotation.

(2) *Non-uniform strategy for selecting sparse slices*

Table 1
Abbreviations used in the paper.

Abbreviation	Full Name	Abbreviation	Full Name
CtEs	Cervical esophagus	GT	Ground truth
Mnd	Mandible	p-GT	Pseudo ground truth
OHPh	Orohypopharynx constrictor muscle	Tr	Training data set
Hrt	Heart	Te	Testing data set
TB	Trachea & proximal Bronchi	VOI	Volume of interest
RLg	Right lung	ROI	Region of interest
DC	Dice coefficient	SI	Shape-based interpolation
JI	Jaccard index	DL	Deep learning
ASD	Average symmetric surface distance	U	Uniform slice selection
EOR	Exclusive or	N	Non-uniform slice selection
PCC	Pearson correlation coefficient		

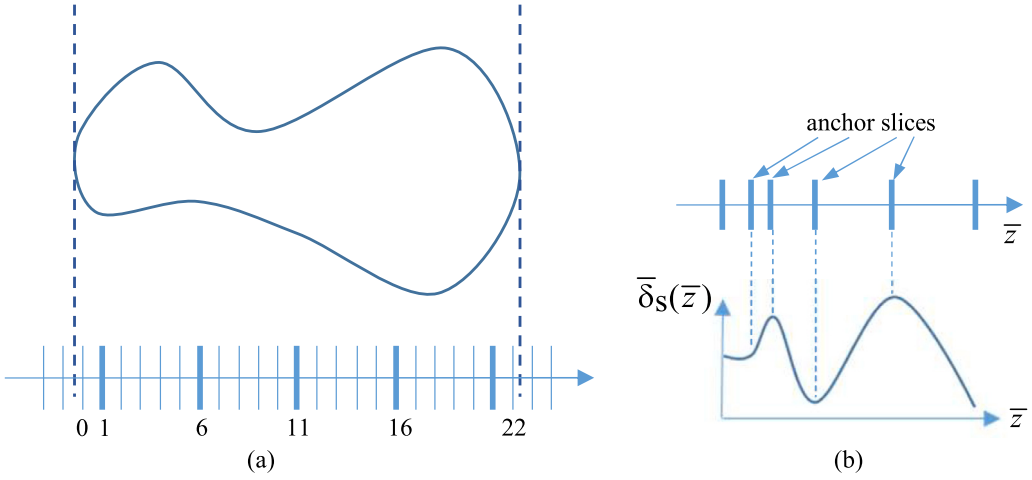

Fig. 2. Illustration of the process of selecting positions of sparse slices for the given GT data. (a) Uniform strategy. Manual contouring is performed on every $(t+1)^{\text{th}}$ slice. In the figure, $t=4$, and bold vertical lines indicate the selected sparse slices. (b) Non-uniform strategy. The four positions with local extrema in shape-change function constitute anchor slices for the object. These positions together with the two end slices constitute 6 anchor slices selected for manual contouring.

In the uniform strategy, we selected sparse slices along z independent of the variation of object shape along z. In the non-uniform strategy, our idea is to select slices strategically – more densely where shape changes more rapidly from slice to slice and more sparsely where this change is small. This is accomplished by first selecting *anchor slices* at locations (z positions) where a shape-change function shows local extrema, and subsequently selecting $k \geq 0$ slices uniformly positioned between every pair of successive anchor slices, as illustrated in Fig. 2(b). In the figure, 4 such local extrema are depicted. Together with the two end slices of the object sample, there will be 6 anchor slices in this case. Parameter k is used to represent the degree of non-uniform sparseness, where $k=0$ if only anchor slices are selected as sparse slices. Although the selected k positions are evenly distributed between each pair of successive anchor slices, since anchors themselves are non-uniformly distributed along the z axis, the sparse slices selected will be non-uniformly distributed guided by the shape change. In order to systematize the selection of anchor slices independent of the individual samples of O, we map the z range of each sample of O to a normalized range [0, 1], denote the position on this normalized range by $\bar{z}$, and determine an average shape-change function $\bar{\delta}_S(\bar{z})$ for O based on a population S of the GT samples of O. The steps involved in selecting the non-uniformly distributed slices are as follows.

Step 1. Estimate a shape change measure $M_{SC}(s_i)$ for each slice position.

Let s_i denote a slice through object O in a GT binary image I in I_b at position z_i. To estimate the change in shape of O in I at s_i, a neighborhood is first defined around s_i to measure local shape change at s_i. A factor R is used to define this neighborhood, where $N_n = max(\lfloor N_O/R \rfloor, 1)$ slices previous and next to s_i are considered in defining $M_{SC}(s_i)$. The shape change is measured by two metrics: exclusive or (EOR) and Jaccard index (JI). $M_{SC}(s_i)$ is calculated as a weighted average of shape changes between s_i and each slice in its neighborhood, where the weight factor is determined based on the distance of the neighboring slices from s_i, as shown in (1):

$$M_{SC}(s_i) = \frac{\sum_{j \in \{i-N_n,\ldots,i-1,i+1,\ldots,i+N_n\}} (N_n + 1 - |i-j|) \times m(s_i, s_j)}{\sum_{j \in \{i-N_n,\ldots,i-1,i+1,\ldots,i+N_n\}} (N_n + 1 - |i-j|)}, \quad (1)$$

where $m(s_i, s_j)$ stands for the shape change measured by one of the metrics (EOR or JI) between slices s_i and s_j at locations z_i and z_j, and the nearer neighboring slices are assigned larger weights.

Step 2. Normalize slice positions and $M_{SC}(s_i)$ both to range [0, 1].

Since different object samples are likely to contain different numbers of slices and different magnitudes for $M_{SC}(s_i)$, to devise a standardized shape-change function that avoids interpolation of images, the slice positions of different object samples are normalized to a fixed slice range and the magnitudes of $M_{SC}(s_i)$ are also normalized. A sufficiently large parameter N_R is determined such that $N_R > N_O$ for all considered object samples, and the normalized range [0, 1] is discretized into N_R positions with $N_R - 1$ intervals. Then, all slices of the object samples are assigned with proportional discrete positions, where only the slice positions are interpolated into a unified range and the actual image slices are not changed. We denote the normalized positions by the variable $\bar{z}$. Besides slice positions, $M_{SC}(s_i)$ also shows different magnitude ranges in different object samples irrespective of whether the

shape change is measured based on EOR or JI. To facilitate deriving the mean shape-change function$\overline{\delta}_S(\bar{z})$, we also normalize $M_{SC}(s_i)$ by min-max normalization:

$$M_{SCn}(s_i) = (M_{SC}(s_i) - M_{\min})/(M_{\max} - M_{\min}), \quad (2)$$

where $M_{\max}$ and $M_{\min}$ denote the maximum and minimum shape change of each object sample.

Step 3. Estimate standardized mean shape-change function $\overline{\delta}_S(\bar{z})$.

The value of the shape-change function $\delta_S(\bar{z})$at each discrete position $\bar{z}$ for each object sample in a population of object samples S is estimated as in (3). The mean shape-change function $\overline{\delta}_S(\bar{z})$is then derived by averaging $\delta_S(\bar{z})$ over the population S. It is possible that some of the positions on the whole range of $\bar{z}$ are not accounted for by data from any of the subjects, so values at those positions are set as null. This will not influence the modeling of the average shape-change function $\overline{\delta}_S(\bar{z})$.

$$\begin{aligned} &\bar{z}_i = \lfloor i/(N_O - 1) \times (N_R - 1) \rfloor/(N_R - 1), i = 0, ...N_O - 1, \\ &\delta_s(\bar{z}_i) = M_{SCn}(s_i). \end{aligned} \quad (3)$$

2.2. Segmentation filling on non-selected slices

Two segmentation filling approaches are investigated in this work: one is a straight-forward strategy of shape-based interpolation (SI) and the other is a deep-learning (DL) based strategy.

(1) *Shape-based interpolation (SI) strategy*

Given two adjacent sparse slices l_1 and l_2 with manual annotations, the segmentation on a slice l_i at any position in between is estimated by using the shape-based interpolation approach (Raya and Udupa, 1990) by following the slice-to-slice shape of the object. Signed (2D) distance transform is first applied to l_1 and l_2 with the convention that distances from the object boundary are positive for pixels inside the object and negative for outside pixels. Then, the distance map for l_i is estimated by linearly interpolating the distance maps for l_1 and l_2, and the binary mask is finally determined by pixels with positive distance values on slice l_i.

(2) *Deep-learning (DL) based strategy*

The deep-learning based approach is developed using the 2D U-net architecture (Ronneberger et al., 2015), where the input and output images are designed in different manners for the two sparse slice selection strategies. For the t-based uniform selection strategy, as shown in Fig. 3(a), the range of continuous four sparsely selected slices contains a total of $3t+4$ slice positions with four selected sparse slices and three blocks of non-selected slices in between. The intensity image slices and binary slices are combined into $2 \times (3t+4) = 6t+8$ channel images as the input to the network, where only the selected sparse slices contain annotated segmentations while the binary image slices on non-selected positions are all empty or with 0 value. The training target is to estimate GT segmentations for the central block of non-selected t slices, and the loss is calculated between predicted GT results and actual GT on the central block of slices. The other two blocks are for giving contextual information.

For the k-based non-uniform strategy, we will follow the above spirit of including information from four consecutive selected sparse slices to design a 10-channel input as illustrated in Fig. 3(b) and (c). M_a, M_b, M_c, and M_d are four consecutive sparse slice positions. When filling segmentations in the skipped block between slices at positions M_b and M_c, intensity and binary images on M_a, M_b, F_1, M_c, and M_d compose the 10-channel input at first, where F_1 denotes the first skipped slice position next to M_b. Subsequently, after the slice at F_1 is filled with segmentation predicted by the network in Fig. 3(b), the input to the network is updated with the composition of images at positions M_a, F_1, F_2, M_c, and M_d, where F_2 is the target slice next to F_1. The second slice position is always the slice previous to the target slice with manual or predicted binary mask. This process continues iteratively until all skipped slices in this block are filled with segmentations.

The networks are composed of convolutional layers with 3×3 kernels followed by Batch Normalization and ReLU nonlinear activation, and the output layers are activated by a sigmoid function. Down-sampling is achieved by convolutional layers with stride 2 while others are with stride 1. Adam optimizer is used to minimize the cross-entropy loss function in training the networks.

2.3. Determining optimal sparseness factor based on GT variability

Individual differences among human expert segmenters employed to create ground truth will always exist naturally and can never be eliminated owing to the variabilities in their clinical knowledge, perceived boundaries in the image, annotation experience, definitions adopted for the objects, and the software used for GT segmentations[1]. Fig. 4 presents two examples of GT delineations by four dosimetrists in our health system of objects in two different body regions and with different segmentation difficulties: (i) Heart: a relatively large blob-like non-sparse object in the thoracic body region with well-defined boundary contrast; and (ii) Cervical esophagus: a thin tube-like sparse object with low boundary contrast in the head & neck body region. Variables e_1 and e_2 here denote two dosimetrists (experts) who segmented the heart in the same image, and e_3 and e_4 denote two other dosimetrists segmenting the esophagus.

The SparseGT method aims to generate pseudo GT that is as good as the actual GT data generated by expert human segmenters with all of their imperfections. This is really the central tenet of our method. Three metrics are employed to demonstrate our approach including two region-based metrics – Dice coefficient (DC) and Jaccard index (JI) – and a boundary-based metric – average symmetric surface distance (ASD). For each object O whose segmentations output by some algorithm A are to be evaluated, we obtain the variability of these metric values among a group G of expert segmenters by having them create GT segmentations of O on a given set V of images. Typically, V required for the SparseGT approach in deriving inter-segmenter variability information is much smaller than the size of the data sets required for training model-based segmentation algorithms, as well as the potential data sets which will be created in clinical practice and segmented by algorithm A. More importantly, this variability needs to be established only once for each fixed O and for each kind of clinical application. In this paper, we employ segmentations from 4 dosimetrists to establish this variability for demonstration purposes, although the object samples for a given O are segmented by two experts, see further explanation later in this section. For each metric M, we describe its variability by a pair (μ_M, σ_M), where μ_M denotes the mean value and σ_M denotes the standard deviation of M over all samples of V among all combinations of expert segmenters in G taken two at a time, where one is taken as the reference segmentation with respect to which the other expert's segmentations are evaluated via M.

Analogously, we determine the variability (μ_{Mp}, σ_{Mp}) of the p-GT generated by SparseGT from a separate training data set Tr by taking experts in G as reference GT for assessing M, where μ_{Mp} denotes the mean value of M and σ_{Mp} is the standard deviation of M over all experts in G considered as reference. In this paper, for demonstration purposes and clinical reasons as stated below, we have considered only one expert in G for each patient sample.

[1] Many software tools allow manually-driven automatic delineation for GT drawing. Such tools introduce their own variability in addition to the variability due to subjective human input.

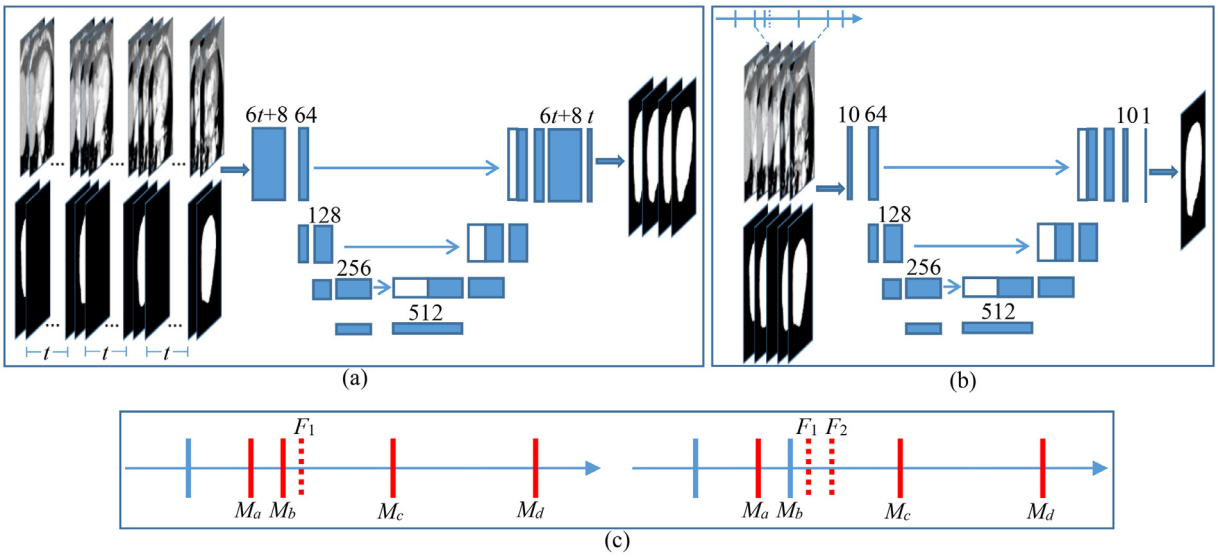

Fig. 3. Illustration of deep-learning-based segmentation filling strategies. The networks are constructed based on U-net. (a) Filling segmentations between uniformly selected sparse slices, for the case of t=4 in this example. (b) Segmentations at slice positions in between non-uniformly selected sparse slices are filled one at a time. Four annotated slices at sparsely selected positions with a blank slice corresponding to the first slice at F_1 to be filled in compose the ten-channel input to the network. (c) The manner of updating the second slice at position F_2 to be filled in.

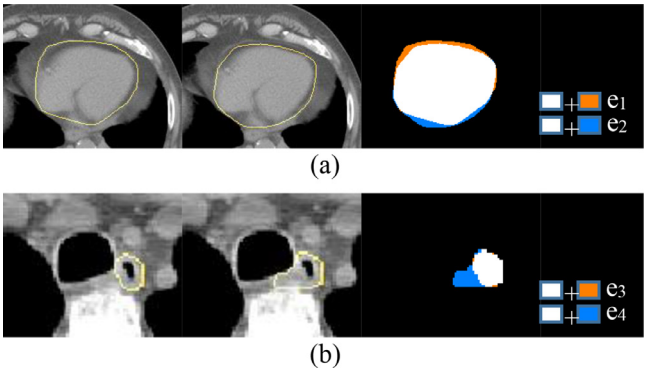

Fig. 4. Illustration of imprecision in GT delineations. Top row: Segmentations (by two dosimetrists e_1 and e_2) of heart – a relatively large blob-like non-sparse object in the thoracic body region. Bottom row: Segmentations (by two additional experts e_3 and e_4) of esophagus – a thin tube-like sparse object with low boundary contrast in the head & neck body region. Substantial differences can be seen between the two segmentations, where white regions represent inter-segmenter agreements, and orange and blue regions denote inter-segmenter differences.

We then determine the optimal sparseness factor t_O in the uniform strategy and k_O in the non-uniform strategy as the largest value of t, or the smallest value of k, where the deviation (μ_{Mp}, σ_{Mp}) of the p-GT becomes statistically indistinguishable from the variability (μ_M, σ_M) among experts in the actual full GT. Note that $t_O \in [0, t_3]$, where $t_3 = \lfloor (N_{OP} - 3)/2 \rfloor$, and N_{OP} represents a reasonably small number such that $N_{OP} \leq N_O$ for most patient images covering O over a population of patient images I, and $k_O \in [0, k_{max}]$, where k_{max} represents the case wherein all slices are selected.

We then assess the applicability of the determined t_O and k_O by using a separate (testing) set of image data Te which is disjoint from Tr. In practice, when realistic GT data are available because contouring of objects is done for clinical reasons, such as for radiation therapy (RT) planning as in our application, it is impossible to guarantee that the same expert contoured all samples of all objects or even all samples of the same object in Tr and Te since it is quite common for dosimetrists employed in radiation therapy planning departments to share workload depending on clinical demand. For the same reasons, it is impossible to guarantee that any experts in G employed for estimating (μ_M, σ_M) would have performed GT contouring of the naturally available data sets. This scenario is more realistic, practical, and sound than a situation where one expert in G or all experts in G performed GT contouring of the data sets in Tr and Te. In our SparseGT method, 4 dosimetrists were involved for establishing inter-segmenter variability and several experts were involved for creating one GT segmentation for each of other samples in the data set in this manner.

2.4. Estimating parameters of the SparseGT method

There are two parameters in the whole method – N_R and R. They both relate to the non-uniform sparse slice selection strategies and are for deciding the proportional positions of anchor slices. N_R represents the number of discrete positions defined in the unified slice range, and the ratio factor R is used in defining the neighborhood in calculating shape change measure. As indicated in Table 4 in the next section, among the 6 objects we have considered for demonstration of the SparseGT method in 498 images, most object samples occupy less than 100 slices and the largest sample occupies less than 200 slices, so we set $N_R = 201$; i.e., 200 intervals are defined on the unified range [0, 1] with increments of 0.05. Also, we take {5, 10, 20, 100} as candidate values for R, where the neighborhood is defined as $N_n = max(\lfloor N_O/R \rfloor, 1)$ slices neighboring the considered slice on each side. Intuitively, different values of R define different windows for smoothing the shape change measure. $R = 5$ represents a coarse neighborhood window containing $\sim 0.4 N_O$ slices and draws the most general pattern of the shape change function, while $R = 100$ represents a fine neighborhood window with only 3 slices and the determined shape change function will be sensitive to local changes.

The difference in sensitivity of different R can be verified by investigating the Pearson correlation coefficient (PCC) in metric $\bar{\delta}_s(\bar{z})$ for different choices of R. For illustration of the method of selecting R, we take typical objects with differing shapes CtEs, Hrt, and Mnd (see Table 4 for object definitions) as examples, and exclusive or (EOR) as the shape change measure. Behavior for other objects and metrics was similar and hence not presented here. To determine the proper R for forming shape change functions, we investigate PCC among $\bar{\delta}_s(\bar{z})$ generated by four candidate R values on the training set. These correlations are listed in the first three rows of Table 2. As an example of object CtEs depicted in Fig. 5, the shape

Table 2
Pearson correlation coefficient between $\overline{\delta}_s(\overline{z})$ generated by different choices for R for the training samples (first three rows) and by the training and testing samples (last row).

	CtEs				Hrt				Mnd			
	R=5	R=10	R=20	R=100	R=5	R=10	R=20	R=100	R=5	R=10	R=20	R=100
PCC in R=10	0.973				0.982				0.982			
$\overline{\delta}_s(\overline{z})$within R=20	0.938	0.98			0.956	0.988			0.958	0.992		
Tr R=100	0.867	0.909	0.952		0.924	0.963	0.989		0.936	0.974	0.99	
PCC in $\overline{\delta}_s(\overline{z})$between Tr and Te	0.91	0.851	0.772	0.612	0.965	0.946	0.916	0.874	0.989	0.98	0.965	0.937

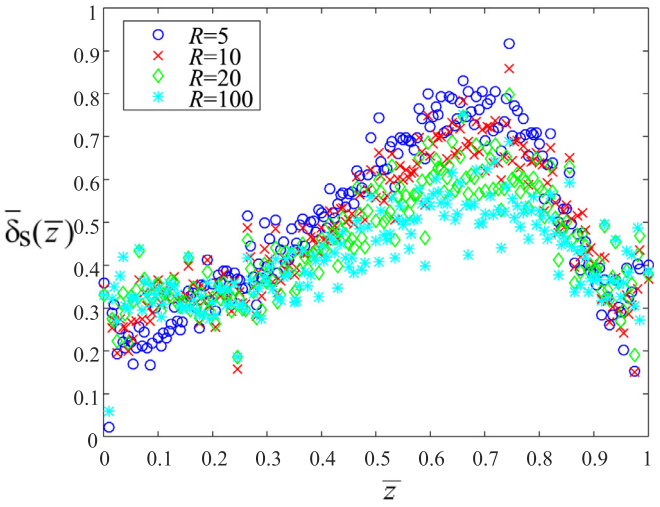

Fig. 5. Illustration of the average shape change function resulting from choosing different values for R for CtEs. The average shape change functions with smaller R are sharper but blurred with larger R.

change functions $\delta_s(\overline{z})$ with smaller R show more general patterns and $\overline{\delta}_s(\overline{z})$ are sharper, whereas functions $\overline{\delta}_s(\overline{z})$ with larger R are blurred because of the sensitivity of $\delta_s(\overline{z})$ to local changes. In most cases in the first three rows, $\overline{\delta}_s(\overline{z})$ with $R=10$ or 20 shows more balanced correlation to functions with other R values, which means this value is a balanced choice representing both general pattern and local details. To verify how good a choice for R is, we check the PCC between $\overline{\delta}_s(\overline{z})$ generated for the same value of R by the training and test sets Tr and Te. These PCC values are listed in the last row of Table 2. From the PCC values in the last row of Table 2, we observe that$\overline{\delta}_s(\overline{z})$values with smaller R are more general and yield PCC relatively close to 1, and those with larger R are more sensitive to local changes. Therefore, we empirically set $R=10$ for all experiments related to non-uniform sparse selection strategies.

2.5. Evaluating the performance of SparseGT

We will denote the four strategies investigated for creating p-GT in SparseGT by $S_{U\text{-}SI}$, $S_{N\text{-}SI}$, $S_{U\text{-}DL}$, and $S_{N\text{-}DL}$, where U and N denote uniform and non-uniform sparse slice selection and SI and DL represent shape-based interpolation and deep learning filling strategies, respectively. The optimal sparseness factors are determined by using the training set Tr. The SI strategy does not need training, while for DL we employ 2-fold cross validation. The number of sparse slices, N_S, selected for optimal outcome for each object may be different for each of the four strategies. The strategy that can achieve maximum sparseness is determined by the real reduction in workload. The workload itself is given by the ratio N_S/N_O.

We will utilize two measures to evaluate the strategies investigated in the SparseGT method as described below. The first measure is used for determining the optimal sparseness that can be achieved by each strategy for each object and the second assesses the difference in the evaluations carried out by GT and p-GT of an actual segmentation output by an algorithm by each strategy for each object.

(i) For a given metric α (one of DC, JI, and ASD), object O, and sparseness factor x (x = t or x = k), consider the metric values $\alpha(I_{ib}, P_{ib}, O, x)$, $i = 1, ..., N$, where the metric is evaluated to assess the deviation (from real GT) of the p-GT obtained corresponding to the sparseness level specified by x. Let μ_α and σ_α denote the mean and standard deviation, respectively, of the $\alpha(I_{ib}, P_{ib}, O, x)$ values obtained over $i = 1, ..., N$.

(ii) For a given metric α (one of DC, JI, and ASD), object O, and segmentation algorithm A, let $\varepsilon(\alpha, O, A)$ denote the root mean squared error between the metric values in using the optimal p-GT $\mathcal{P}_b = \{P_{1,b}, ..., P_{N,b}\}$ and the true GT $\mathcal{I}_b = \{I_{1,b}, ..., I_{N,b}\}$ in evaluating the output $\mathcal{J}_b = \{J_{1,b}, ..., J_{N,b}\}$ of A corresponding to the input image set $\mathcal{I} = \{I_1, ..., I_N\}$:

$$\varepsilon(\alpha, O, A) = \sqrt{\frac{1}{N}([\alpha(I_{1b}, J_{1b}) - \alpha(P_{1b}, J_{1b})]^2 + , ..., + [\alpha(I_{Nb}, J_{Nb}) - \alpha(P_{Nb}, J_{Nb})]^2)}. \quad (5)$$

For demonstration purposes, the segmentation algorithm A we evaluated is the method reported in (Wu et al., 2019).

3. Experiments, results, and discussion

3.1. Data sets and experiments

This 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. Experiments are conducted on computed tomography (CT) images of two body regions, Head & Neck and Thorax, with three objects in each body region – cervical esophagus (CtEs), mandible (Mnd), and orohypopharynx constrictor muscle (OHPh) in the Head & Neck body region, and heart (Hrt), trachea & proximal bronchi (TB), and right lung (RLg) in the Thorax body region. The objects are chosen to represent different shape and size characteristics and different degrees of challenges for segmentation (RLg and Mnd: least challenging, Hrt and TB: moderately challenging, CtEs and OHPh: most challenging). Our set $\mathcal{I}$ of images consists of 498 3D images – 298 in Head & Neck region and 200 in Thorax region. For the above 6 objects, GT data are available to us which constitute real clinical data as contoured by several dosimetrists (as explained in the previous section) for the routine RT planning of patients with Head & Neck or Thoracic cancer. For 81 of the Head & Neck studies, each of two dosimetrists contoured the above 3 objects separately. Similarly, another two dosimetrists contoured the 3 Thoracic objects separately on 87 Thoracic studies. These data sets constitute the training set we denoted by V previously and were utilized for estimating (μ_M, σ_M).

We note that not every study in the cohort $\mathcal{I}$ necessarily has all 6 objects contoured since the actual objects contoured in clinical

Table 3
The relationship of data sets V, Tr and Te to the full cohort $\mathcal{I}$ and their roles in different stages of parameter estimation.

Type of parameters	Division of data sets	
Inter-segmenter variability (μ_M, σ_M)	Training set: V (V$\subset\mathcal{I}$)	
Determining R in $\overline{\delta}_s(\overline{z})$	Training set: Tr Verifying set: Te ($\mathcal{I}$ = Tr + Te)	
Variability (μ_{Mp}, σ_{Mp}) between p-GT and manual GT	Determining t_O and k_O on Tr	$S_{U\text{-}SI}$ and $S_{N\text{-}SI}$: No training is needed. $S_{U\text{-}DL}$ and $S_{N\text{-}DL}$: 2-fold cross validation on Tr.
	Verifying t_O and k_O on Te	$S_{U\text{-}SI}$ and $S_{N\text{-}SI}$: No training is needed. $S_{U\text{-}DL}$ and $S_{N\text{-}DL}$: DL networks are trained on Tr and tested on Te.

Table 4
Number of object samples and object sizes in the Head & Neck and Thorax data sets.

Objects	Number of samples(total/ Tr/ Te)	Object sizes(in voxels)	DL-input size(in pixels)
CtEs	283/ 225/ 58	$45\times77\times(18\sim71)$	64×96
Mnd	292/ 232/ 60	$139\times126\times(17\sim86)$	160×144
OHPh	266/ 211/ 55	$64\times55\times(27\sim113)$	80×80
Hrt	199/ 160/ 39	$181\times169\times(20\sim70)$	208×192
TB	197/ 157/ 40	$155\times130\times(36\sim109)$	176×160
RLg	189/ 152/ 37	$167\times241\times(56\sim159)$	192×272

Table 5
Proportional anchor positions found for the different objects from $\overline{\delta}_s(\overline{z})$ curves.

Object	Number of anchor positions	Anchor positions
CtEs	3	0, 0.69, 1
Mnd	6	0, 0.215, 0.395, 0.64, 0.83, 1
OHPh	7	0, 0.21, 0.39, 0.625, 0.77, 0.93, 1
Hrt	5	0, 0.1, 0.42, 0.825, 1
TB	5	0, 0.22, 0.51, 0.83, 1
RLg	5	0, 0.2, 0.665, 0.845, 1

practice for RT planning depend on the location and size of the tumor in each study. The data sets in the two body regions are separately divided into disjoint sets of training and test samples, where about 20% samples are randomly selected to compose the test set Te, and the remaining samples compose the training set Tr to determine optimal sparseness factors t_O and k_O, which are verified on Te to check if the test set can also yield indistinguishable deviation with respect to expert variability. The voxel size in our data sets varies from $0.93\times0.93\times1.5$ mm^3 to $1.6\times1.6\times3$ mm^3. The object sizes are also variable among subjects, where object samples of different subjects occupy varying numbers of slices. The bounding box size is determined for each object based on the largest occupied range among its samples. To fit the input size to the DL network, the 2D region of interest (ROI) of object samples is set with size in multiples of 8, since there are three convolutional layers with stride 2 in both of the designed U-net based networks. The splitting of data sets and their roles in different stages of SparseGT are listed in Table 3 for clarity, and key information about data sets is summarized in Table 4 for ready reference.

3.2. Results and discussion

(1) *Shape change function and anchors*

The proportional anchors found on the average shape change functions $\overline{\delta}_s(\overline{z})$ with $R=10$ are listed in Table 5, where numbers and positions of anchors are object-specific.

(2) *Optimal sparseness*

In Fig. 6, we display mean values of $\alpha(I_{ib}, P_{ib}, O, x)$ as a function of t for the uniform strategies for O = Hrt and as a function of k for the non-uniform strategies for O = CtEs, where μ_M and σ_M are also marked as well as t or k values which show deviation of p-GT from GT that is statistically insignificant (marked by a cross). Optimal values of t and k are determined on the training set and are verified on the test set (marked by triangles) to check if the test set can also yield indistinguishable deviation with respect to GT variability. In Table 6, we list μ_α and σ_α for all considered objects at t_O and k_O (optimal values of t and k, respectively).

Image examples are illustrated in Fig. 7 for all six considered objects including GT from four expert segmenters e_1, e_2, e_3, and e_4 (all are dosimetrists), and optimal p-GT created by the four strategies. Binary masks of GT/p-GT are overlaid on CT images and overlaid also by the reference contours from e_1 or e_3. The corresponding surface renditions are shown as well. Among our objects, CtEs and OHPh have poor contrast; Hrt and TB have confusing boundaries; Mnd has sharp shape changes; and RLg is affected by lesions in the chest wall near the object boundary. Small, non-compact, and sparse objects entail larger degrees of imprecision inversely proportional to their size compared to large, well-defined, and compact objects. The degree of segmentation challenge is also an influencing factor in producing optimal p-GT; for example, TB is a sparse object; however, it poses moderate segmentation challenge due to its rather well-defined boundaries. Its level of imprecision compared to another similarly sparse object, CtEs, is smaller, while CtEs poses great segmentation challenges due to its sparseness and poor boundary contrast.

Among all considered objects, OHPh has the highest imprecision due to the fact that it is an object with extreme sparseness, low contrast, multiple intensity ranges within the object, and given the variability that may exist in object interpretation. Low μ_M and large σ_M values lead to optimal factors denoting greater sparseness. The created p-GT in such cases may sometimes have unacceptable deviations, as shown in surface renditions for $S_{U\text{-}SI}$ and $S_{N\text{-}SI}$ of OHPh. Although we have utilized the most common region-based and boundary-based metrics to describe similarity or dissimilarity among segmentations, none of them is able to

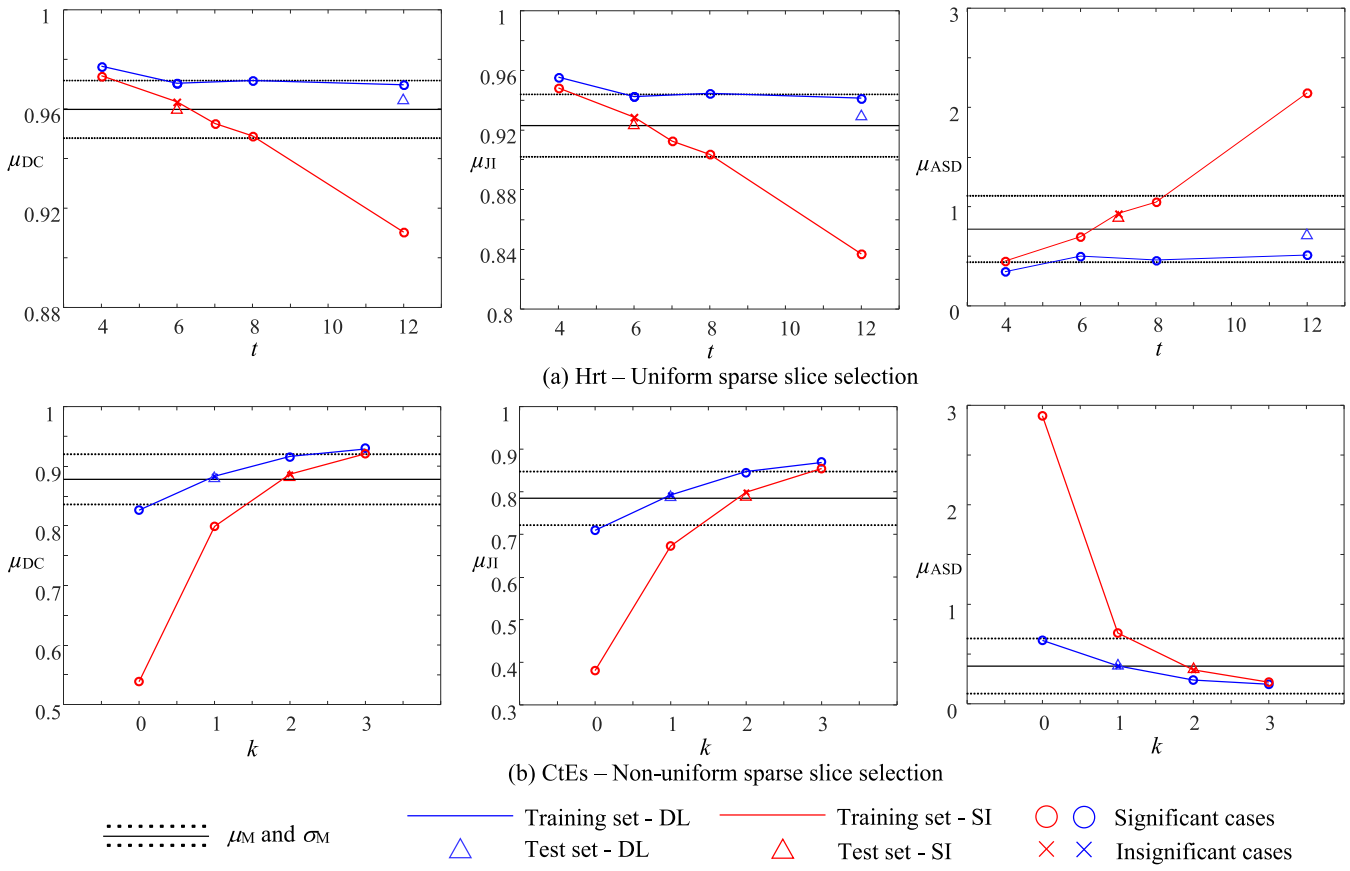

Fig. 6. Mean of metric values $\alpha(I_{ib}, P_{ib}, O, x)$ as a function of t for the uniform strategies for O = Hrt and as a function of k for the non-uniform strategies for O = CtEs and for SI and DL filling strategies. GT mean and standard deviation μ_M and σ_M are also marked. The optimal sparseness factor t_O or k_O is determined on the training set by the largest t or smallest k which does not demonstrate statistically significant difference compared to (μ_M, σ_M) and is verified on the test set where deviation on the generated p-GT is also demonstrated without significant difference compared to (μ_M, σ_M) .

perfectly describe the great inter-segmenter difference prevailing for this extremely sparse and low-contrast object.

There are open-source tools with functions for automatic and semiautomatic segmentation such as our own CAVASS platform[2] and 3D Slicer[3], both of which contain functions for filling segmentations between slices. The interpolation algorithm in 3D Slicer is based on (Albu et al., 2008), and the shape-based interpolation (SI) strategy used in this paper is available in CAVASS and was introduced to the literature by us in 1990. However, those tools do not provide any guidance on how to determine the variability that naturally exists in human-provided GT, how to analyze this in an object-specific manner, and how to exploit the variability optimally to determine the number and locations of the key sparse slices to be selected. The SI based results shown in this paper can be repeated in CAVASS. Since the 3D Slicer does not use SI, the optimal factors may not be exactly the same as what we found but should be very similar since the same type of algorithm is applied. Although we did not do detailed analysis on which factors will be optimal in 3D Slicer, the optimal sparseness factor we found for $S_{U\text{-}SI}$ is able to generate acceptable p-GT in 3D Slicer as illustrated by an example shown in Fig. 8. The SparseGT methodology is general and hence any software system can be used in the proposed framework to determine optimal factors to save maximum workload based on the segmentation filling algorithms available in the software system.

(3) *Evaluation of actual segmentations by optimal pseudo ground truth*

Segmentations created by the Automatic Anatomy Recognition-Radiation Therapy (AAR-RT) method (Wu et al., 2019) are utilized to demonstrate the effectiveness of the SparseGT method in practical segmentation evaluation. Table 7 summarizes the $\varepsilon(\alpha, O, A)$ values for all considered objects and metrics under all four strategies.

We observe from the quantitative results that p-GT created by the DL strategies shows best capability to replace manual ground truth in that it generates least error in evaluation measures compared to the SI strategies.

The $\varepsilon(\alpha, O, A)$ values associated with a p-GT strategy also show the influence the inter-segmenter difference may have on segmentation evaluation. Specific to practical usage, if the data set for training or model building and the test data set are contoured by different GT segmenters, taking DC as an example, there will be an error of about 0.03, 0.02, 0.05, 0.01, 0.015, and 0.007 for CtEs, Mnd, OHPh, Hrt, TB, and RLg, respectively. This error may be blamed on inter-segmenter differences but not on the real capability of the trained model or the algorithm.

Inter-segmenter differences show object-dependent upper bounds for how accurate the automatic segmentation algorithms can become. Beyond those bounds it is doubtful, if directly verified on other sources of data sets, that the algorithms will be able to yield segmentations with as good evaluation measures as with the training data sets. With the explosive development of deep learning architectures, we believe that there are several algorithms that are able to reach or surpass this upper bound for objects like Mnd

[2] http://www.mipg.upenn.edu/Vnews/mipg_software.html.
[3] https://www.slicer.org/.

Table 6
Mean and standard deviation μ_α and σ_α of $\alpha(I_{ib}, P_{ib}, O, x)$ are listed for all objects for all 4 strategies for both training (Tr) and test (Te) data sets.

Object	Metric	μ_M, σ_M	Strategies							
			$S_{U\text{-}SI}$		$S_{U\text{-}DL}$		$S_{N\text{-}SI}$		$S_{N\text{-}DL}$	
			Tr	Te	Tr	Te	Tr	Te	Tr	Te
CtEs	Optimal sparseness		$t_O = 5$		$t_O = 14$		$k_O = 2$		$k_O = 1$	
	DC	0.878	0.894	0.884	0.877	0.875	0.887	0.881	0.885	0.879
		0.042	0.046	0.051	0.028	0.038	0.042	0.046	0.03	0.041
	JI	0.785	0.811	0.796	0.781	0.78	0.799	0.79	0.795	0.787
		0.063	0.07	0.077	0.043	0.057	0.066	0.072	0.047	0.061
	ASD	0.379	0.293	0.308	0.396	0.393	0.342	0.342	0.381	0.376
	(mm)	0.277	0.112	0.118	0.141	0.196	0.146	0.152	0.135	0.138
Mnd	Optimal sparseness		$t_O = 2$		$t_O = 16$		$k_O = 2$		$k_O = 0$	
	DC	0.909	0.939	0.942	0.934	0.934	0.951	0.947	0.945	0.943
		0.019	0.019	0.017	0.019	0.022	0.013	0.013	0.019	0.019
	JI	0.834	0.886	0.891	0.877	0.877	0.907	0.9	0.896	0.893
		0.033	0.033	0.03	0.032	0.037	0.023	0.024	0.033	0.034
	ASD	0.354	0.259	0.236	0.285	0.263	0.204	0.215	0.226	0.232
	(mm)	0.109	0.091	0.079	0.121	0.1	0.077	0.07	0.12	0.107
OHPh	Optimal sparseness		$t_O = 14$		$t_O = 24$		$k_O = 0$		$k_O = 0$	
	DC	0.654	0.712	0.704	0.771	0.78	0.805	0.807	0.819	0.808
		0.062	0.123	0.136	0.05	0.06	0.062	0.07	0.041	0.046
	JI	0.488	0.565	0.559	0.63	0.643	0.678	0.681	0.695	0.681
		0.067	0.14	0.153	0.065	0.077	0.085	0.094	0.058	0.063
	ASD	0.813	0.841	0.965	0.623	0.566	0.465	0.486	0.465	0.518
	(mm)	0.258	0.569	0.838	0.193	0.171	0.173	0.22	0.134	0.137
Hrt	Optimal sparseness		$t_O = 6$		$t_O = 12$		$k_O = 1$		$k_O = 0$	
	DC	0.96	0.963	0.96	0.97	0.963	0.979	0.977	0.968	0.967
		0.012	0.012	0.014	0.009	0.01	0.008	0.009	0.009	0.01
	JI	0.923	0.929	0.923	0.941	0.929	0.959	0.956	0.938	0.937
		0.021	0.022	0.026	0.017	0.019	0.015	0.017	0.017	0.019
	ASD	0.775	0.697	0.843	0.512	0.71	0.319	0.374	0.569	0.617
	(mm)	0.334	0.318	0.401	0.217	0.251	0.176	0.193	0.237	0.272
TB	Optimal sparseness		$t_O = 1$		$t_O = 4$		$k_O = 4$		$k_O = 2$	
	DC	0.938	0.949	0.951	0.935	0.937	0.934	0.934	0.941	0.949
		0.018	0.01	0.01	0.013	0.013	0.015	0.016	0.012	0.012
	JI	0.883	0.903	0.906	0.879	0.882	0.876	0.877	0.889	0.903
		0.031	0.018	0.017	0.023	0.024	0.026	0.028	0.021	0.021
	ASD	0.389	0.139	0.133	0.251	0.213	0.225	0.231	0.213	0.184
	(mm)	0.255	0.041	0.036	0.229	0.112	0.104	0.13	0.093	0.102
RLg	Optimal sparseness		$t_O = 5$		$t_O = 32$		$k_O = 3$		$k_O = 0$	
	DC	0.974	0.972	0.972	0.974	0.973	0.976	0.976	0.979	0.98
		0.044	0.007	0.005	0.012	0.012	0.005	0.004	0.011	0.01
	JI	0.953	0.945	0.946	0.949	0.948	0.952	0.953	0.959	0.961
		0.077	0.012	0.01	0.022	0.021	0.009	0.008	0.02	0.019
	ASD	0.831	0.679	0.653	0.653	0.638	0.557	0.55	0.604	0.545
	(mm)	1.478	0.183	0.147	0.396	0.381	0.137	0.148	0.421	0.345

and RLg, while there is still considerable room for improvement for sparse and challenging objects like CtEs (Chan et al., 2019; Tong et al., 2019).

(4) *Workload reduction*

The primary goal of this work is to reduce the manual workload needed for creating GT segmentations. The ratio N_S/N_O expresses the fraction of the full workload needed for the optimal sparseness achieved by the strategies in the SparseGT method or $[1 - (N_S/N_O)] \times 100$ describes the % reduction achieved in workload. In Table 8, we summarize the mean and standard deviation of N_S and N_S/N_O as well as the associated optimal sparseness factors t_O and k_O for achieved the four strategies for each object. Note that $N_S = \lfloor (N_O - 1)/(t_O + 1) + 1 \rfloor$ for the uniform strategy and $N_S = k_a + k_O \times (k_a - 1)$ for the non-uniform strategy where k_a denotes the number of anchors derived from $\bar{\delta}_s(\bar{z})$and $N_S = k_a$ when $k_O = 0$. Thus, N_S/N_O has larger standard deviation in the non-uniform strategies than in uniform strategies.

We observe that the uniform strategies yield larger workload savings in most cases, except for TB with SI strategies. Observe that the optimal sparseness factors for SI strategies show how regular the object is along the axis orthogonal to the scanning plane. For TB, its shape changes greatly from slice to slice with $t_O = 1$ and $k_O = 4$. Instead of sampling the shape change by uniformly selecting slices, the non-uniform strategy formulates shape change functions and points out the important proportional positions with local maximum or minimum shape changes as the anchors which contain more shape change information compared to other random uniform positions. However, an imperfection in the current non-uniform strategy is that, with increasing k values, N_S changes more rapidly compared to uniform strategies with large t values. So, for regular objects, the uniform strategies need less sparse slices but the variability is greater in $S_{U\text{-}SI}$ compared to $S_{N\text{-}SI}$ for Mnd, OHPh, and Hrt. TB presents with the most shape change from slice to slice with $t_O = 1$, and the advantage of the non-uniform strategy shows

Fig. 7. Image examples for all six considered objects. GT samples for each object are generated by two expert segmenters – e_1 and e_2 for Head & Neck objects, and e_3 and e_4 for Thorax objects. p-GT samples are generated from sparse GT by e_1 or e_3 and the optimal t_O and k_O via SI and DL strategies, respectively. 2D binary masks are overlaid on the CT intensity images and overlaid by e_1 or e_3 contours, and the corresponding surface models are presented as well.

here that, following anchors, the workload reduction improves by ~10%.

The strength of the DL strategies is demonstrated in Table 8 where it is shown that p-GT with human-level accuracy can be created with less than 20% (less than 10% for regular and non-sparse objects) of the workload needed in creating full manual GT. The sparseness for all objects, except TB, approaches the ideal value with the slices selected just on the two ends and in the middle of the object.

We conducted two further experiments to explore whether the workload for creating p-GT can be further reduced:

(E1) $S_{U\text{-}SI\text{-}DL}$: Utilizing p-GT created via $S_{U\text{-}SI}$ to enlarge the training data for $S_{U\text{-}DL}$. Although DL strategies can reach maximum workload reduction, they still need a proper set of samples for network training, while SI strategies do not need a training stage to create p-GT segmentations. Instead of fully manually annotating the whole training set necessary for DL strategies, GT segmentations on part of the training samples can be semi-automatically created by SI strategies. Specific to our experiments, first, we create p-GT for all samples in Tr, except for the samples with annotations from two expert segmenters which we assume as irreplaceable samples to derive natural imprecision measures among human segmenters (these samples constituted our set V mentioned earlier), based on the optimal sparseness factor for $S_{U\text{-}SI}$ (denoted by $t_{U\text{-}SI}$). Then, sparse slice selection is conducted, with the optimal sparseness factor for $S_{U\text{-}DL}$ (denoted by $t_{U\text{-}DL}$), and the DL network is trained on the mixed set of samples with manual GT segmentations or $S_{U\text{-}SI}$-created p-GT segmentations. Finally, sparse slice selection and manual contouring are conducted on the test set with $t_{U\text{-}DL}$ and the skipped slices are filled with pseudo segmentations by the p-GT-trained network. Experiments are conducted for all six considered objects, and, from the verification purpose, the network

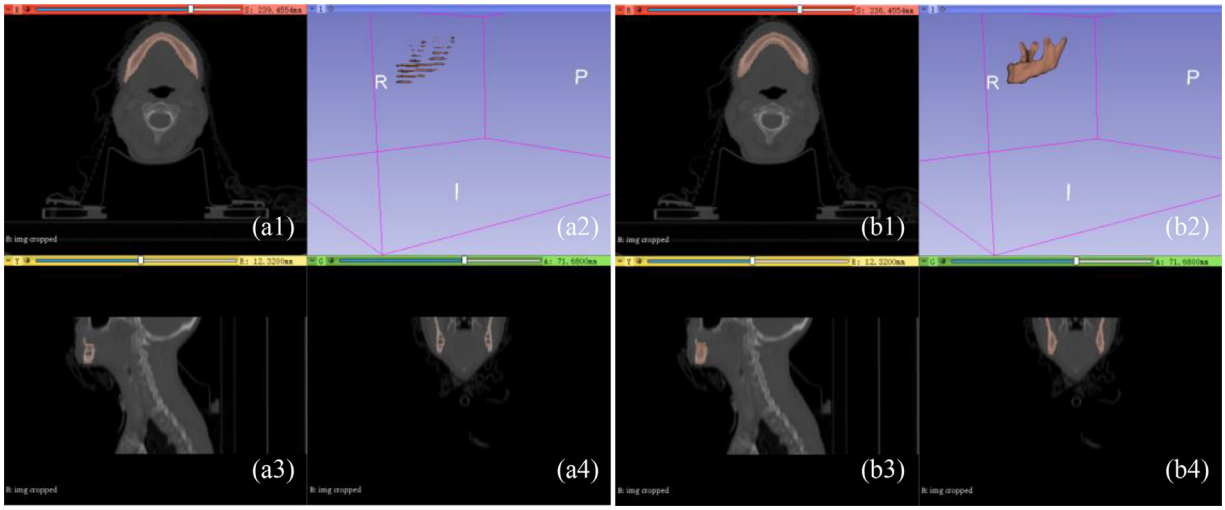

Fig. 8. Generating p-GT for Mnd with 3D Slicer following $t_O = 2$ in $S_{U\text{-}SI}$. (a1) Manual annotation on a selected slice; (a2) - (a4) sparse GT shown in 3D and two other 2D views. (b1) Filled segmentation on a non-selected slice (one slice next to (a1)); (b2) - (b4) p-GT in the three views.

Table 7
Root mean squared errors $\varepsilon(\alpha, O, A)$ of evaluation metric values measured by p-GT in comparison with GT in assessing an actual segmentation algorithm AAR-RT for different objects and strategies. For each object and each metric, the smallest error achieved over all strategies for training and testing data sets is shown in bold.

Object	Metric	Strategies							
		$S_{U\text{-}SI}$		$S_{U\text{-}DL}$		$S_{N\text{-}SI}$		$S_{N\text{-}DL}$	
		Tr	Te	Tr	Te	Tr	Te	Tr	Te
CtEs	Optimal sparseness	$t_O = 5$		$t_O = 14$		$k_O = 2$		$k_O = 1$	
	DC	0.033	0.04	**0.025**	0.028	0.031	0.033	0.028	**0.024**
	JI	0.032	0.04	**0.023**	0.024	0.029	0.032	0.025	**0.021**
	ASD	0.576	0.477	0.386	0.46	**0.309**	**0.256**	0.328	0.346
Mnd	Optimal sparseness	$t_O = 2$		$t_O = 16$		$k_O = 2$		$k_O = 0$	
	DC	0.022	**0.017**	0.019	0.018	**0.015**	0.018	0.018	0.021
	JI	0.031	**0.023**	0.027	0.025	**0.022**	0.025	0.026	0.029
	ASD	0.22	0.185	0.177	**0.145**	0.167	0.197	**0.145**	0.169
OHPh	Optimal sparseness	$t_O = 14$		$t_O = 24$		$k_O = 0$		$k_O = 0$	
	DC	0.088	0.084	0.051	0.056	0.047	**0.049**	**0.042**	0.052
	JI	0.07	0.065	0.045	0.05	**0.037**	**0.037**	0.038	0.047
	ASD	1.307	0.886	0.485	0.267	0.412	0.493	**0.233**	**0.264**
Hrt	Optimal sparseness	$t_O = 6$		$t_O = 12$		$k_O = 1$		$k_O = 0$	
	DC	0.028	0.022	**0.008**	0.01	0.008	0.01	0.01	**0.008**
	JI	0.038	0.031	**0.01**	0.014	0.01	0.014	0.014	**0.012**
	ASD	1.169	1.717	0.401	0.362	0.381	0.47	**0.375**	**0.321**
TB	Optimal sparseness	$t_O = 1$		$t_O = 4$		$k_O = 4$		$k_O = 2$	
	DC	0.021	0.023	0.015	0.013	0.02	0.021	**0.014**	**0.008**
	JI	0.026	0.027	**0.018**	0.015	0.024	0.025	0.019	**0.01**
	ASD	0.802	0.784	0.773	0.742	0.892	0.73	**0.604**	**0.353**
RLg	Optimal sparseness	$t_O = 5$		$t_O = 32$		$k_O = 3$		$k_O = 0$	
	DC	0.007	0.006	0.009	0.005	**0.007**	0.007	0.008	**0.005**
	JI	**0.011**	0.011	0.015	0.008	0.012	0.01	0.013	**0.008**
	ASD	0.369	0.288	0.491	0.459	**0.328**	**0.253**	0.524	0.372

is trained on the whole Tr and tested on Te. We verified that $(\mu_\alpha, \sigma_\alpha)$ of $S_{U\text{-}SI\text{-}DL}$ created p-GT in all cases is lower or insignificantly greater, with *p-value* > 0.05, than inter-segmenter difference (μ_M, σ_M) as presented in Table 9. We also compare $(\mu_\alpha, \sigma_\alpha)$ of p-GT created by $S_{U\text{-}SI\text{-}DL}$ and $S_{U\text{-}DL}$ to explore the coupled error of deviations in $S_{U\text{-}SI}$-created p-GT with systematic errors of the DL network. $(\mu_\alpha, \sigma_\alpha)$ in most comparisons are with *p-value* > 0.05, while one exception is Mnd where the DL strategy is powerful enough to learn the delineation manner from the training set and the error of $S_{U\text{-}SI}$-created p-GT becomes significantly degraded. In spite of possible degradation, we can still infer the potential ability of $S_{U\text{-}SI}$-created p-GT to partially replace manual GT in model training and further reduce manual workload.

(E2) $S_{VOI\text{-}DL}$: Pseudo ground truth creation by the DL strategy with only annotated volume (3D region) of interest (VOI). Most objects have achieved or approached the ideal sparseness defined in this work with the strength of DL strategies. During the process of manual contouring on the selected sparse slices in any strategies considered in this work, the VOI is actually implicitly defined by the rough 3D region occupied by the sparse GT seg-

Table 8
Actual number of selected sparse slices (N_S) and its ratio to the number of slices occupied by the target object (N_S/N_O) under different strategies with their optimal sparseness factors. Mean and standard deviation values are listed. The $S_{U\text{-}DL}$ strategy shows maximum workload reduction. For each object, N_S and N_S/N_O values for the strategies that achieved the maximum workload are shown in bold.

Object	Strategies											
	$S_{U\text{-}SI}$			$S_{U\text{-}DL}$			$S_{N\text{-}SI}$			$S_{N\text{-}DL}$		
	t_O	N_S	N_S/N_O	t_O	N_S	N_S/N_O	k_O	N_S	N_S/N_O	k_O	N_S	N_S/N_O
CtEs	5	7.141	0.202	14	**3.316**	**0.082**	2	7	0.207	1	5	0.132
		1.343	0.012		**0.566**	**0.007**			0.047			0.019
Mnd	2	14.216	0.36	16	**3.101**	**0.074**	2	16	0.42	0	6	0.144
		2.612	0.01		**0.343**	**0.007**			0.091			0.018
OHPh	14	4.917	0.093	24	**3.058**	**0.052**	0	7	0.136	0	7	0.136
		0.797	0.008		**0.281**	**0.004**			0.03			0.03
Hrt	6	5.573	0.185	12	**3.036**	**0.098**	1	9	0.304	0	5	0.162
		0.727	0.011		**0.265**	**0.009**			0.041			0.017
TB	1	26.249	0.515	4	**10.569**	**0.207**	4	21	0.418	2	13	0.259
		3.603	0.005		**1.471**	**0.006**			0.048			0.03
RLg	5	13.799	0.184	32	**3.045**	**0.039**	3	17	0.23	0	5	0.065
		1.846	0.004		**0.236**	**0.003**			0.027			0.006

Table 9
Mean and standard deviation values of $\alpha(I_{ib}, P_{ib}, O, t_O)$ and $\varepsilon(\alpha, O, A)$ for strategies with further increasing sparseness. Mean and standard deviation values on the test set are listed. Strategies with best metric values or least evaluation error are marked in bold.

Object	Metric	t_O	μ_M, σ_M	Strategies					
				$S_{U\text{-}DL}$		$S_{U\text{-}SI\text{-}DL}$		$S_{VOI\text{-}DL}$	
				$\alpha(.)$	$\varepsilon(.)$	$\alpha(.)$	$\varepsilon(.)$	$\alpha(.)$	$\varepsilon(.)$
CtEs	DC	$t_{U\text{-}SI}=5$	0.878	0.875	0.028	**0.876**	**0.028**	0.858	0.033
		$t_{U\text{-}DL}=14$	0.042	0.038		**0.039**		0.051	
	JI		0.785	0.78	**0.024**	**0.782**	0.025	0.755	0.03
			0.063	0.057		**0.059**		0.074	
	ASD		0.379	0.393	**0.46**	**0.384**	0.49	0.425	0.502
	(mm)		0.277	0.196		**0.222**		0.186	
Mnd	DC	$t_{U\text{-}SI}=2$	0.909	**0.934**	**0.018**	0.924	0.019	0.934	0.022
		$t_{U\text{-}DL}=16$	0.019	**0.022**		0.024		0.025	
	JI		0.834	**0.877**	**0.025**	0.859	0.027	0.876	0.03
			0.033	**0.037**		0.04		0.043	
	ASD		0.354	**0.263**	**0.145**	0.345	0.311	0.32	0.198
	(mm)		0.109	**0.1**		0.146		0.2	
OHPh	DC	$t_{U\text{-}SI}=14$	0.654	**0.78**	0.056	0.767	0.062	0.737	**0.049**
		$t_{U\text{-}DL}=24$	0.062	**0.06**		0.083		0.068	
	JI		0.488	**0.643**	0.05	0.628	0.055	0.588	**0.044**
			0.067	**0.077**		0.101		0.081	
	ASD		0.813	**0.566**	**0.267**	0.64	0.353	0.772	0.561
	(mm)		0.258	**0.171**		0.274		0.336	
Hrt	DC	$t_{U\text{-}SI}=6$	0.96	**0.963**	**0.01**	0.962	0.013	0.948	0.03
		$t_{U\text{-}DL}=12$	0.012	**0.01**		0.01		0.027	
	JI		0.923	**0.929**	**0.014**	0.926	0.018	0.902	0.044
			0.021	**0.019**		0.019		0.046	
	ASD		0.775	**0.71**	**0.362**	0.767	0.517	1.383	1.473
	(mm)		0.334	**0.251**		0.273		0.963	
TB	DC	$t_{U\text{-}SI}=1$	0.938	**0.937**	**0.013**	0.937	0.017	0.904	0.048
		$t_{U\text{-}DL}=4$	0.018	**0.013**		0.014		0.025	
	JI		0.883	**0.882**	**0.015**	0.882	0.02	0.825	0.058
			0.031	**0.024**		0.024		0.041	
	ASD		0.389	**0.213**	**0.742**	0.223	0.835	0.545	1.434
	(mm)		0.255	**0.112**		0.156		0.476	
RLg	DC	$t_{U\text{-}SI}=5$	0.974	0.973	0.005	0.97	**0.004**	**0.982**	0.012
		$t_{U\text{-}DL}=32$	0.044	0.012		0.009		**0.009**	
	JI		0.953	0.948	0.008	0.942	**0.006**	**0.965**	0.011
			0.077	0.021		0.017		**0.017**	
	ASD		0.831	0.638	0.459	0.718	**0.313**	**0.584**	0.384
	(mm)		1.478	0.381		0.274		**0.568**	

mentations. An interesting question is whether the manual workload can be further reduced by only annotating the proper region (VOI) of the target object without actually contouring segmentations on any slices. This is equivalent to providing only manual recognition information without the detailed delineations in specifying GT. Experimental results, shown in Table 9, indicate that, among the four objects (Mnd, OHPh, Hrt, and RLg) which reached extreme sparseness with $S_{U\text{-}DL}$, only results for Hrt show significantly greater variability (μ_α, σ_α) than inter-segmenter difference (μ_M, σ_M) with *p-value* < 0.05. However, although it seems that only VOI is enough to yield human-level segmentations for some objects, the standard deviations, especially of boundary-based metric ASD, are obviously larger in most cases. That means, without proper definition of the segmenter's behavior of contouring, the quality of created segmentations is unstable. Furthermore, (μ_α, σ_α) is greater with $S_{VOI\text{-}DL}$ compared to that from $S_{U\text{-}DL}$ in many cases. In this sense, this experiment suggests that not only the human recognition act but also defining the contouring behavior is

Table 10
A summary of the manual help required by the different strategies of the SparseGT method.

	Training					Evaluation				
	End-slices	Full GT	Sparse GT	VOI	Pseudo GT	End-slices	Full GT	Sparse GT	VOI	Pseudo GT
$S_{U\text{-}SI}$	Y	Y	N	N	N	Y	N	Y	N	N
$S_{U\text{-}DL}$	Y	Y	N	N	N	Y	N	Y	N	N
$S_{N\text{-}SI}$	Y	Y	N	N	N	Y	N	Y	N	N
$S_{N\text{-}DL}$	Y	Y	N	N	N	Y	N	Y	N	N
$S_{U\text{-}SI\text{-}DL}$	Y	N	Y	N	N	Y	N	Y	N	N
$S_{VOI\text{-}DL}$	Y	Y	N	Y	N	Y	N	N	Y	N

of fundamental importance in creating human-level pseudo ground truth!

(5) *Comparison with other methods*

The lack of large sample sets with fully annotated GT and the associated inter-segmenter differences in manual contouring are two common difficulties faced by anatomy segmentation and evaluation algorithms. Several algorithms have been proposed to avoid (evade) such challenges, where, as has been stated in Section 1.2, semi- and un-supervised learning based segmentation methods and methods of evaluation without ground truth are proposed to deal with the lack of GT, and the STAPLE algorithm and its extensions have been proposed to estimate the implicit GT from multiple manual contours in an attempt to eliminate the inter-segmenter differences. Different from such methods, to deal with the lack of absolute GT for segmentation evaluation, we propose the SparseGT method to exploit the ubiquitous inter-segmenter differences intelligently, instead of trying to eliminate them, as reference of possible errors and deviations in the imperfect GT, and simulate GT in a semi-automatic manner by first conducting manual contouring on selected sparse slices and then automatically filling segmentations on other skipped slices. To our knowledge, no such work exists in the literature. Although (Valindria et al., 2017) mentions pseudo GT, that work just represents algorithm-created segmentations and does not guarantee human-level accuracy. Thus, their meaning of *pseudo GT* is totally different from the sense implied in our work. The SparseGT created pseudo GT segmentations are object-dependent, guaranteed to reach human-level accuracy, and can greatly reduce manual workload. They can be used in practice as substitutes for fully manual GT in evaluating segmentations via actual algorithms.

(6) *Practical usage*

The practical use of the SparseGT method involves two stages – a training stage and a p-GT generation stage for evaluation. The training stage has to be conducted first for each object of interest to determine the strategy that is optimum for that object to achieve maximum workload reduction without compromising GT accuracy. The amount of manual help needed for the two stages for that object will then depend on the object itself and the optimum strategy for it. Table 10 summarizes the manual help needed for the six strategies examined in this paper for both stages. Estimation of the natural variability (μ_M, σ_M) that exists in GT is an essential step for all strategies and not listed in the table. This requires full GT on a much smaller data set than what is needed for the training stage for a DL strategy. Note that end-slices need to be specified manually in both stages for all strategies. Some strategies do not need full GT for training although all of them require GT on the selected sparse slices. When needed, VOI is determined automatically in the training stage (see Table 4).

Computational considerations: SparseGT was implemented on a computer with the following specifications: 6-core Intel i7-7800X CPU 3.5GHz with 64 GB RAM, NVIDIA TITAN XP GPU with 12 GB of memory and GeForce GTX 1070 GPU with 8 GB of memory, and running the Linux operating system. Generating p-GT for an object sample by $S_{U\text{-}DL}$ or $S_{N\text{-}DL}$ costs less than 2 s. In the network training stage, $S_{U\text{-}DL}$ is trained in the unit of block while $S_{N\text{-}DL}$ is trained slice by slice, so $S_{U\text{-}DL}$ generally takes more time than $S_{N\text{-}DL}$ for training. The size of ROI is another factor influencing training time. Typically, CtEs with the smallest ROI requires ∼21 min in $S_{U\text{-}DL}$ training and ∼1 h in $S_{N\text{-}DL}$ training. RLg with the largest ROI requires ∼3 h in $S_{U\text{-}DL}$ and ∼4 h in $S_{N\text{-}DL}$. The computational efficiency for the SI strategies depends on strategies of sparse slice selection, optimal sparseness factors, and the numbers of slices occupied by the target objects, where, generally, objects with smaller number of slices and optimal factors of greater sparseness require less time to generate p-GT segmentations. $S_{U\text{-}SI}$ is more efficient requiring ∼1.6 s for Hrt to ∼8 s for TB than $S_{N\text{-}SI}$ which requires ∼3.3 s for Hrt to ∼8.5 s for TB in generating p-GT for an object sample. No actual training stage is needed for the SI strategies.

4. Concluding remarks

In this paper, our goal was to address a gap that currently exists in segmentation evaluation, namely, to seek an answer to the question "Is it possible to create machine-generated ground truth (GT) from sparse human annotated data sets such that the generated pseudo GT (p-GT) is just as good as full manual GT?" We investigated a novel method named SparseGT, which provides guidance on how to exploit inter-segmenter differences derived from natural imprecision in human-drawn GT as reference and create p-GT vastly more efficiently for segmentation evaluation than the full manual GT. We have shown that the created optimal p-GT is statistically indistinguishable from the real full GT and works at least as well as the full GT in terms of evaluation accuracy, but requiring only a fraction of the manual workload needed for creating full GT. No such work currently exists.

p-GT data are created in two steps, sparse slice selection to conduct manual annotation and segmentation filling between sparse slices, each of which is investigated by two strategies, including uniform and non-uniform slice selection, and shape-based-interpolation and deep-learning based segmentation filling. Different strategy combinations are evaluated by the actual workload reduction and the evaluation accuracy achieved by p-GT compared to full manual GT. Experiments are conducted utilizing ∼500 CT studies of the Head & Neck and Thorax involving 6 objects of different segmentation challenges, and actual algorithmic segmentations for testing the SparseGT method are generated by the AAR-RT method in the literature.

We summarize our conclusions as follows. (i) Overall, the combined strategy $S_{U\text{-}DL}$ of uniform sparse slice selection coupled with DL-based segmentation filling is able to yield the highest manual workload reduction (∼80-96%!) compared to other strategies for all six objects and all three segmentation evaluation metrics considered. (ii) The root mean squared errors in segmentation evaluation metric values $\varepsilon(\alpha, O, A)$ show a potential practical insight offered by the SparseGT method: If evaluated by manual GT created by different segmenters, there may be some errors emanating from inter-segmenter differences which may be confused as arising due to the actual segmentation algorithm A. (iii) For both DL and SI

filling strategies, uniform sparse slice selection outperforms non-uniform selection in most cases. Nonetheless, non-uniform selection shows its advantages for irregular objects (e.g., TB) with sharp changes from slice to slice when utilizing the SI strategy to create p-GT. (iv) Although the SI strategy generally supports a lower level of sparseness in slice selection compared to DL, it is a straight forward strategy without the need for a training stage in creating p-GT. We have preliminarily demonstrated (under $S_{U\text{-}SI\text{-}DL}$) that it shows potential to enlarge the training data sets for DL when only a small cohort of fully annotated data sets is available. (v) Experimental results show that if further reduction in manual workload by annotating only a 3D region of interest is attempted ($S_{VOI\text{-}DL}$), the accuracy of the generated p-GT becomes unstable. This suggests that the intricate behavior associated with delineation of the GT is essential for accurate p-GT generation and just object recognition help alone is insufficient.

We note that, although we investigate variability of manual segmentation in RT planning and a few strategies for sparse slice selection and segmentation filling, other clinical practice and readily available strategies can take their places and determine object-, strategy- and application- specific optimal sparseness to generate GT for segmentation evaluation following the framework shown in Fig. 1.

There are several gaps and further challenges to be addressed in this investigation. Firstly, although the metrics utilized in this paper are the most commonly used, they are incapable of expressing subtle and local deviations between segmentations. For example, consider OHPh, which is a sparse object with subtle thickness, low contrast, and implicit variability in object interpretation. Its inter-segmenter differences measured by currently-used metrics are disproportionately greater than the meaning expressed by the metrics. Thus, in conjunction with the question raised in this study, the deficiencies associated with the metrics also need to be overcome (Li et al., 2020) and considered.

Second, the shape change function designed for non-uniform sparse selection can potentially be improved. The current approach treats all anchors and the intervals between anchors with equal importance. The k parameter can be potentially made to vary with the degree of shape change with larger values chosen where shape change is more rapid. This approach may then show improvement over uniform selection.

Third, the non-uniform strategy has shown its strength on Trachea and Bronchi which has up to three levels of branches. This strategy can potentially handle more general tree-like objects such as airway trees or pulmonary arteries and veins with greater change in topology in the z-direction. Such objects generally show much greater variation in ground truth than non-tree-like structures. For such objects, the shape-change function can be modified to include an additional variable to indicate the branch-points, which should be treated as anchors, and additional sparse positions may be selected according to the shape variation in each non-branching segment. We surmise that one anchor selected between each pair of branch-point anchors should suffice to handle tree-like objects. This clearly requires further work.

Forth, the current slice selection strategies in the SparseGT method are strongly slice-oriented and depend on the z-axis which is chosen to be the cranio-caudal direction. From the goal of workload reduction, there may be an object-specific optimal axis. This is worth exploring, notwithstanding the fact that this may raise other issues such as interpolation and the associated errors. The actual segmentation itself can be carried out in the native slices while optimal slice orientations can be used just for segmentation evaluation only.

Finally, in this paper we demonstrated the potential of the SparseGT method, but its real applicability in routine segmentation evaluation needs to be independently established. This can be accomplished by conducting large-scale evaluations involving multiple body regions and image modalities and numerous objects where p-GT and GT evaluations are compared. For use in real clinical applications, we may also employ reader studies to determine the degree of *acceptability* of the segmentation for the application as described in (Li et al., 2020). If p-GT and GT evaluations both suggest the same acceptability, then the validity of p-GT is established for that application. We are working toward this direction for the practical use of the SparseGT method.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Jieyu Li: Formal analysis, Writing - original draft, Software, Conceptualization, Investigation, Methodology, Funding acquisition, Validation, Visualization. **Jayaram K. Udupa:** Conceptualization, Supervision, Project administration, Funding acquisition, Methodology, Formal analysis, Writing - original draft, Investigation, Resources, Writing - review & editing. **Yubing Tong:** Data curation, Project administration, Writing - review & editing. **Lisheng Wang:** Supervision. **Drew A. Torigian:** Investigation, Resources, Writing - review & editing.

Acknowledgements

The research reported here is supported party by an NIH grant R42 CA199735. Jieyu Li's training at the Medical Image Processing Group was supported partly by China Scholarship Council.

References

Agn, M., Rosenschöld, af, P.M., Puonti O., Lundemann, M.J., Mancini, L., Papadaki, A., Thust, S., Ashburner, J., Law., I., Van Leemput, k., 2019. A modality-adaptive method for segmenting brain tumors and organs-at-risk in radiation therapy planning. Med. Image Anal. 54, 220–237. doi:10.1016/j.media.2019.03.005.

Albu, A.B., Beugeling, T., Laurendeau, D., 2008. A morphology-based approach for interslice interpolation of anatomical slices from volumetric images. IEEE Trans. Biomed. Eng. 55 (8), 2022–2038. doi:10.1109/TBME.2008.921158.

Ashburner, J., Friston, K.J., 2009. Computing average shaped tissue probability templates. Neuroimage 45 (2), 333–341. http://doi.org/10.1016/j.neuroimage.2008.12.008.

Bhaskaruni, D., Moss, F.P., Lan, C., 2018. Estimating prediction qualities without ground truth: a revisit of the reverse testing framework. In: 2018 24th International Conference on Pattern Recognition, pp. 49–54.

Bø, H.K., Solheim, O., Jakola, A.S., Kvistad, K.A., Reinertsen, I., Berntsen, E.M., 2017. Intra-rater variability in low-grade glioma segmentation. J. Neuro-Oncol. 131 (2), 393–402. doi:10.1007/s11060-016-2312-9.

Can, Y.B., Chaitanya, K., Mustafa, B., Koch, L.M., Konukoglu, E., Baumgartner, C.F., 2018. Learning to segment medical images with scribble-supervision alone. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 236–244.

Cerrolaza, J.J., Picazo, M.L., Humbert, L., Sato, Y., Rueckert, D., Ballester, M.Á.G., Linguraru, M.G., 2019. Computational anatomy for multi-organ analysis in medical imaging: A review. Med. Image Anal. 56, 44–67. http://doi.org/10.1016/j.media.2019.04.002.

Chabrier, S., Emile, B., Rosenberger, C., Laurent, H., 2006. Unsupervised performance evaluation of image segmentation. EURASIP J. Appl. Signal Proc.. http://doi.org/10.1155/ASP/2006/96306. 217-217.

Chan, J.W., Kearney, V., Haaf, S., Wu, S., Bogdanov, M., Reddick, M., Dixit, N., Sudhyadhom, A., Chen, J., Yom, S.S., Solberg, T.D., 2019. A convolutional neural network algorithm for automatic segmentation of head and neck organs at risk using deep lifelong learning. Med. Phys. 46 (5), 2204–2213. http://doi.org/10.1002/mp.13495.

Cheplygina, V., Pluim, J.P., 2018. Crowd disagreement about medical images is informative. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, pp. 105–111.

Christensen, G.E., Rabbitt, R.D., Miller, M.I., 1994. 3D brain mapping using a deformable neuroanatomy. Phys. Med. Biol. 39, 609–618. http://doi.org/10.1088/0031-9155/39/3/022.

Chu, C., Oda, M., Kitasaka, T., Misawa, K., Fujiwara, M., Hayashi, Y., Wolz, R., Rueckert, D., Mori, K., 2013. Multi-organ segmentation from 3D abdominal CT images using patient-specific weighted-probabilistic atlas. SPIE Med. Imaging. 86693Y-86691-86693Y-86697 http://doi.org/10.1117/12.2007601 .

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 424–432. doi:10.1007/978-3-319-46723-8_49.

Cootes, T.F., Taylor, C.J., Cooper, D.H., 1995. Active shape models-their training and application. Comput. Vis. Image Underst. 61 (1), 38–59. http://doi.org/10.1006/cviu.1995.1004.

Drozdzal, M., Chartrand, G., Vorontsov, E., Shakeri, M., Jorio, L.D., Tang, A., Romero, A., Bengio, Y., Pal, C., Kadoury, S., 2018. Learning normalized inputs for iterative estimation in medical image segmentation. Med. Image Anal. 44, 1–13. http://doi.org/10.1016/j.media.2017.11.005.

Gee, J.C., Reivich, M., Bajcsy, R., 1993. Elastically deforming 3D atlas to match anatomical brain images. J. Comput. Assist. Tomogr. 17, 225–236. http://doi.org/10.1097/00004728-199303000-00011.

Gordon, S., Lotenberg, S., Long, R., Antani, S., Jeronimo, J., Greenspan, H., 2009. Evaluation of uterine cervix segmentations using ground truth from multiple experts. Computerized Med. Imaging Graphics 33 (3), 205–216. http://doi.org/10.1016/j.compmedimag.2008.12.002.

Gurari, D., Theriault, D., Sameki, M., Isenberg, B., Pham, T.A., Purwada, A., Solski, P., Walker, M., Zhang, C., Wong, J.Y., Betke, M., 2015. How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms. In: 2015 IEEE Winter Conference on Applications of Computer Vision, pp. 1169–1176.

Heller, N., Dean, J., Papanikolopoulos, N., 2018. Imperfect segmentation labels: How much do they matter? In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, pp. 112–120.

Herman, G.T., Srihari, S., Udupa, J., 1979. Detection of changing boundaries in two- and three-dimensions. In: Badler, N.I., Aggarwal, J.K. (Eds.), Proceedings of the Workshop on Time Varying Imagery. University of Pennsylvania, Philadelphia, Pennsylvania, pp. 14–16.

Joskowicz, L., Cohen, D., Caplan, N., Sosna, J., 2019. Inter-observer variability of manual contour delineation of structures in CT. Eur. Radiol. 29 (3), 1391–1399. http://doi.org/10.1007/s00330-018-5695-5.

Jungo, A., Meier, R., Ermis, E., Blatti-Moreno, M., Herrmann, E., Wiest, R., Reyes, M., 2018. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 682–690.

Koch, L.M., Rajchl, M., Bai, W., Baumgartner, C.F., Tong, T., Passerat-Palmbach, J., Aljabar, P., Rueckert, D., 2017. Multi-atlas segmentation using partially annotated data: methods and annotation strategies. IEEE Trans. Pattern Anal. Mach. Intell. 40 (7), 1683–1696. http://doi.org/10.1109/TPAMI.2017.2711020.

Kohlberger, T., Singh, V., Alvino, C., Bahlmann, C., Grady, L., 2012. Evaluating segmentation error without ground truth. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 528–536.

Lampert, T.A., Stumpf, A., Gançarski, P., 2016. An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation. IEEE Trans. Image Process. 25 (6), 2557–2572. http://doi.org/10.1109/TIP.2016.2544703.

Li, X., Aldridge, B., Fisher, R., Rees, J., 2011. Estimating the ground truth from multiple individual segmentations incorporating prior pattern analysis with application to skin lesion segmentation. In: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1438–1441.

Li, J., Udupa, J.K., Tong, Y., Wang, L., Torigian, D.A., 2020. LinSEM: Linearizing segmentation evaluation metrics for medical images. Med. Image Anal. 60, 101601. http://doi.org/10.1016/j.media.2019.101601.

Liu, H.K., 1977. Two- and three-dimensional boundary detection. Comput. Graph. Image Process. 6, 123–134. http://doi.org/10.1016/S0146-664X(77)80008-7.

Moeskops, P., Viergever, M.A., Mendrik, A.M., De Vries, L.S., Benders, M.J., Išgum, I., 2016. Automatic segmentation of MR brain images with a convolutional neural network. IEEE Trans. Med. Imaging 35 (5), 1252–1261. http://doi.org/10.1109/TMI.2016.2548501.

Nowak, S., Rüger, S., 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: Proceedings of the International Conference on Multimedia Information Retrieval, pp. 557–566.

O'Neil, A.Q., Murchison, J.T., van Beek, E.J., Goatman, K.A., 2017. Crowdsourcing labels for pathological patterns in CT lung scans: can non-experts contribute expert-quality ground truth? In: Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, pp. 96–105.

Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L., 2015. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1742–1750.

Park, S., Chu, L.C., Fishman, E.K., Yuille, A.L., Vogelstein, B., Kinzler, K.W., K.M., Horton., Hruban, R.H., Zinreich, E.S., Fadaei Fouladi, D., Shayesteh, S., Graves, J., Kawamoto, S., 2019. Annotated normal CT data of the abdomen for deep learning: Challenges and strategies for implementation. Diagn. Interv. Imaging 101 (1), 35–44. http://doi.org/10.1016/j.diii.2019.05.008.

Pizer, S.M., Fletcher, P.T., Joshi, S., Thall, A., Chen, J.Z., Fridman, Y., Fritsch, D.S., Gash, A.G., Glotzer, J.M., Jiroutek, M.R., Lu, C.L., Muller, K.E., Tracton, G., Yushkevich, P., Chaney, E.L., 2003. Deformable m-reps for 3D medical image segmentation. Int. J. Comput. Vis. 55 (2-3), 85–106. doi:10.1023/A:1026313132218.

Popović, Z.B., Thomas, J.D., 2017. Assessing observer variability: a user's guide. Cardiovasc. Diagn. Therapy 7 (3), 317–324. http://doi.org/10.21037/cdt.2017.03.12.

Rajchl, M., Lee, M.C., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M.A., Hajnal, J.V., Kainz, B., Rueckert, D., 2016. Deepcut: object segmentation from bounding box annotations using convolutional neural networks. IEEE Trans. Med. Imaging 36 (2), 674–683. http://doi.org/10.1109/TMI.2016.2621185.

Raya, S.P., Udupa, J.K., 1990. Shape-based interpolation of multidimensional objects. IEEE Trans. Med. Imaging MI 9 (1), 32–42. http://doi.org/10.1109/42.52980.

Robinson, R., Oktay, O., Bai, W., Valindria, V.V., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., Fung, K., Lukaschuk, E., Lee, A.M., Carapella, V., Kimm, Y.J., Kainz, B., Peichnik, S.K., Neubauer, S., Petersen, S.E., Page, C., Reuckert, D., Glocker, B., 2018. Real-time prediction of segmentation quality. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 578–585. doi:10.1007/978-3-030-00937-3_66.

Robinson, R., Valindria, V.V., Bai, W., Oktay, O., Kainz, B., Suzuki, H., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., Fung, K., Lukaschuk, E., Lee, A.M., Carapella, V., Kim, Y.J., 2019. Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study. J. Cardiovasc. Magn. Resonance 21 (1), 18. http://doi.org/10.1186/s12968-019-0523-x.

Robinson, R., Valindria, V.V., Bai, W., Suzuki, H., Matthews, P.M., Page, C., Rueckert, D., Glocker, B., 2017. Automatic quality control of cardiac MRI segmentation in large-scale population imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 720–727.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing And Computer-Assisted Intervention, pp. 234–241.

Schipaanboord, B., Boukerroui, D., Peressutti, D., van Soest, J., Lustberg, T., Kadir, T., Dekker, A., van Elmpt, W., Gooding, M., 2018. Can atlas-based auto-segmentation ever be perfect? Insights from extreme value theory. IEEE Trans. Med. Imaging 38 (1), 99–106. http://doi.org/10.1109/TMI.2018.2856464.

Schlesinger, D., Jug, F., Myers, G., Rother, C., Kainmüller, D., 2017. Crowd sourcing image segmentation with iaSTAPLE. In: 2017 IEEE 14th International Symposium on Biomedical Imaging, pp. 401–405.

Sharp, G., Fritscher, K.D., Pekar, V., Peroni, M., Shusharina, N., Veeraraghavan, H., Yang, J., 2014. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. Med. Phys. 41 (5). doi:10.1118/1.4871620.

Shen, T., Li, H., Huang, X., 2011. Active volume models for medical image segmentation. IEEE Trans. Med. Imaging 30 (3), 774–791. http://doi.org/10.1109/TMI.2010.2094623.

Shi, C., Cheng, Y., Wang, J., Wang, Y., Mori, K., Tamura, S., 2017. Low-rank and sparse decomposition based shape model and probabilistic atlas for automatic pathological organ segmentation. Med. Image Anal. 38, 30–49. http://doi.org/10.1016/j.media.2017.02.008.

Shwartzman, O., Gazit, H., Shelef, I., Riklin-Raviv, T., 2019. The Impact Of An Inter-Rater Bias On Neural Network Training arXiv preprint arXiv:1906.11872.

Sikka, K., Deserno, T.M., 2010. Comparison of algorithms for ultrasound image segmentation without ground truth. Medical Imaging 2010: Image Perception, Observer Performance, and Technology Assessment 7627, 76271C http://doi.org/10.1117/12.844504 .

Staib, L.H., Duncan, J.S., 1992. Boundary finding with parametrically deformable models. IEEE Trans. Pattern Anal. Mach. Intell. 14, 1061–1075. http://doi.org/10.1109/34.166621.

Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X., 2020. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Med. Image Anal., 101693. http://doi.org/10.1016/j.media.2020.101693.

Tong, N., Gou, S., Yang, S., Cao, M., Sheng, K., 2019. Shape constrained fully convolutional DenseNet with adversarial training for multiorgan segmentation on head and neck CT and low-field MR images. Med. Phys. 46 (6), 2669–2682. http://doi.org/10.1002/mp.13553.

Udupa, J.K., Odhner, D., Zhao, L., Tong, Y., Matsumoto, M.M., Ciesielski, K.C., Falcao, A.X., Vaideeswaran, P., Ciesielski, V., Saboury, B., Mohammadianrasanani, S., 2014. Body-wide hierarchical fuzzy modeling, recognition, and delineation of anatomy in medical images. Med. Image Anal. 18 (5), 752–771. http://doi.org/10.1016/j.media.2014.04.003.

Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B., 2017. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. IEEE Trans. Med. Imaging 36 (8), 1597–1606. http://doi.org/10.1109/TMI.2017.2665165.

Wang, S., He, K., Nie, D., Zhou, S., Gao, Y., Shen, D., 2019a. CT male pelvic organ segmentation using fully convolutional networks with boundary sensitive representation. Med. Image Anal. 54, 168–178. http://doi.org/10.1016/j.media.2019.03.003.

Wang, D., Li, M., Ben-Shlomo, N., Corrales, C.E., Cheng, Y., Zhang, T., Jayender, J., 2019b. Mixed-supervised dual-network for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 192–200.

Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., Vercauteren, T., 2018. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Trans. Med. Imaging 37 (7), 1562–1573. http://doi.org/10.1109/TMI.2018.2791721.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmenta-

tion. IEEE Trans. Med. Imaging 23 (7), 903–921. http://doi.org/10.1109/TMI.2004.828354.

Wu, X., Udupa, J.K., Tong, Y., Odhner, D., Pednekar, G.V., Simone II, C.B., McLaughlin, D., Apinorasethkul, C., Lukens, J., Mihailidis, C., Shammo, G., James, P., Tiwari, A., Wojtowicz, L., Camaratta, J., Torigian, D.A., 2019. 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. 54, 45–62. http://doi.org/10.1016/j.media.2019.01.008.

Yang, H.F., Choe, Y., 2011. Ground truth estimation by maximizing topological agreements in electron microscopy data. In: International Symposium on Visual Computing, pp. 371–380.

Yang, J., Veeraraghavan, H., Armato III, S.G., Farahani, K., Kirby, J.S., Kalpathy-Kramer, J., van Elmpt, W., Dekker, A., Han, X., Feng, X., Aljabar, P., Oliveira, B., van der Heyden, B., Zamdborg, L., Lam, D., Gooding, M., Sharp, G.C., 2018. Autosegmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017. Med. Phys. 45 (10), 4568–4581. http://doi.org/10.1002/mp.13141.

Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z., 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: International Conference On Medical Image Computing And Computer-Assisted Intervention, pp. 399–407.

Zhou, L., Deng, W., Wu, X., 2020. Robust image segmentation quality assessment. In Medical Imaging with Deep Learning.