Automatic Multiorgan Segmentation via Multiscale Registration and Graph Cut

Razmig Kéchichian, Sébastien Valette, Michel Desvignes

Automatic Multiorgan Segmentation via Multiscale Registration and Graph Cut

Razmig Kéchichian*, Sébastien Valette, and Michel Desvignes

***Abstract*—We propose an automatic multiorgan segmentation method for 3D radiological images of different anatomical content and modality. The approach is based on a simultaneous multilabel Graph Cut optimization of location, appearance and spatial configuration criteria of target structures. Organ location is defined by target-specific probabilistic atlases (PA) constructed from a training dataset using a fast (2+1)D SURF-based multiscale registration method involving a simple 4-parameter transformation. PAs are also used to derive target-specific organ appearance models represented as intensity histograms. The spatial configuration prior is derived from shortest-path constraints defined on the adjacency graph of structures. Thorough evaluations on Visceral project benchmarks and training dataset, as well as comparisons with the state of the art confirm that our approach is comparable to and often outperforms similar approaches in multiorgan segmentation, thus proving that the combination of multiple suboptimal but complementary information sources can yield very good performance.**

***Index Terms*—Segmentation, Registration, Atlases, Abdomen, Magnetic resonance imaging (MRI), X-ray imaging and computed tomography, thorax, keypoints, spatial prior, graph cut**

I. Introduction and Related Work

CLINICAL practice and medical research today generate vast numbers of images of high dimensions, especially in whole-body CT and MR imaging. This is due to the increasing need for more accurate and less invasive procedures, made possible by ongoing advances in acquisition and storage technologies. The increasing number and complexity of images increases the workload of radiologists. It also makes tasks of data access, analysis and visualization challenging, especially in distributed environments involving web terminals for remote access and visualization [1]. Automatic image analysis methods, such as detection, segmentation and registration, have become part of radiological practice, underpinning algorithms that make computer-aided diagnosis and treatment possible.

The majority of medical image analysis methods, and those of segmentation in particular, target a single anatomical structure or pathology [2], [3]. The reason is that many such methods do not scale up to higher numbers of structures, or may be bound to the specificities of a certain modality or anatomy. Single-organ methods leave the potential of organ cosegmentation in large field-of-view (FOV) images unexplored.

Multiorgan segmentation is necessary in several situations. Some clinical procedures, such as radiotherapy [4] and detection of metastasis [5] require a simultaneous scrutiny of several anatomies. Other applications include the creation of patient-specific models and the semantic navigation of anatomy [6]. Speaking of methods, multiorgan segmentation expressed as a sequence of single-organ segmentations is prone to the propagation of errors between different stages. This is usually alleviated by post-segmentation correction, which is difficult to generalize [7]. Intrinsically multiorgan methods raise questions neither on segmentation sequence nor on error propagation. We thus focus on anatomy-independent multiorgan methods that can be applied to a wide range of medical images.

A. Multiorgan segmentation methods

The basic procedure that underlies multiatlas segmentation (MAS) methods [8]–[19] is the following: a number of images are selected from a dataset and registered onto the target structure, then corresponding annotations are transfered to the target and merged to localize and segment it. An atlas designates the pair of the intensity image and its annotation. A recent survey [2] shows that MAS methods are applied to brain MRI more often than to thoracic-abdominal CT images. The reason is that the relatively low inter-subject variation in overall brain shape and structure locations allows good alignment. In contrast, abdominal structures exhibit high location and shape variability and are challenging for registration [20].

A related family of methods are those that rely on statistical shape models (SSM) [21]–[23]. In a nutshell, SSM methods create a statistical model for a structure describing its mean shape or appearance and its variabilities. The model is then matched to a target in order to localize and segment it. Note that such mean models represent variation in the particular dataset they are created from. Therefore, when a target significantly differs from the mean, specificity can be low. Moreover, due to wide inter-subject differences, the mean shape of some structures may have no anatomical meaning. We refer the reader to the following review for further details [24].

The shortcomings of SSM are addressed in some atlas-based methods, e.g. [13], [18], by the creation of subject-specific shape models, called probabilistic atlases (PA), constructed by merging registered annotations into a spatial probability of organ location, used subsequently as a prior to segment the target. The construction of such target-specific PAs can however be costly since several atlases are required to build a PA for a specific target that cannot be used to segment another.

R. Kéchichian and S. Valette are with Creatis research center: Université de Lyon; CNRS UMR 5220; Inserm U1044; INSA-Lyon; Université Lyon 1, Villeurbanne, France (e-mail: firstname.lastname@creatis.insa-lyon.fr).

M. Desvignes is with Gipsa-lab: CNRS UMR 5216; Grenoble-INP; Université Joseph Fourier; Université Stendhal, Saint-Martin-d'Hères, France (e-mail: michel.desvignes@gipsa-lab.grenoble-inp.fr).

Digital Objet Identifier

Patch-based segmentation methods [25], [26] avoid the computationally expensive step of, often non-rigid, atlas registration in MAS methods. Instead of registering organ-level atlases, the latter are split into a large number of patches, and patch selection is done according to a similarity metric in local neighborhoods around the voxels of the target.

Machine learning has had important contributions to multiorgan segmentation. Since classifiers, when applied as they are, label image elements (voxels, superpixels, etc.) independently, researchers have taken local context into account by using long-range spatial features [27], [28], by introducing location variables into the classifier's objective [29] or by using random fields in conjunction with classifiers [30]. With the emergence of deep learning, neural networks have regained ground in medical image analysis [3]. Deep neural networks of fully convolutional [31]–[33] and especially of the U-net variant [34], [35] have achieved very good results in multiorgan segmentation. The effectiveness of these methods depends on the availability of large annotated datasets and GPU-based computing to optimize empirically defined network architectures that can have millions of parameters [35].

Regularization techniques are widely resorted to in multiorgan segmentation to favor the spatial consistency of labels assigned to individual image elements by a voting or a probability maximization mechanism. Level-set [35], [36] and random walk methods [28] have been used to regularize initial segmentations produced by classifiers. The use of Graph Cut regularization is arguably the most common. It is employed by many methods such as [10]–[13], [18], [23], [25], [26]. Graph Cut optimization has the advantage of theoretical and empirical optimality [37]. For a particular class of objective functions frequently arising in segmentation, Graph Cut algorithms produce global optima in single-object and provably-good approximate solutions in multiobject segmentation [38].

B. Keypoint-based registration and segmentation

Widely used by the computer vision community [39], keypoint-based image description and matching methods, such as SIFT [40] and SURF [41], have found relatively few applications in medical image analysis. These methods first detect a number of interest points (edges, ridges, blobs etc.) in the image, then compute feature vectors describing local neighborhoods around these points and use them as content descriptors. In medical imaging, 3D versions of SIFT have been used for brain MR image matching [42], linear registration of radiation therapy data [43] and deformable registration of thoracic CT [44]. A review of keypoint-based medical image registration can be found in [45]. Taking an approach close to ours, PAs used as priors in Graph Cut segmentation are created in [18] by registering atlases to the target via SIFT keypoints extracted from the target and atlases and used to estimate an affine transformation. A final step of deformable registration is applied on images prior to merging registered atlases to create PAs. Significant speedups in MAS have been achieved in [46] by eliminating dense image registration between target and atlases. Organ labels are transfered to the target image based on sparse correspondences between keypoints identified in the latter and those detected in atlas images that carry labels.

II. Methods

We propose an automatic multiorgan segmentation method for 3D images of different anatomical content and modality. It follows a Bayesian approach and uses location and intensity likelihoods of structures and a prior distribution of their spatial configuration. Location likelihoods are defined by target-specific PAs constructed by registering atlases to the target in shrinking frames using a fast SURF-based registration method that estimates a homothetic transformation (translation + isotropic scale). Confidence regions of PAs are used to derive target-specific intensity likelihoods. The spatial prior is derived from shortest-path constraints defined on the adjacency graph of structures [47]. Likelihoods and the spatial prior define an energy function, optimized by a multilabel Graph Cut algorithm to obtain the multiorgan segmentation. A preliminary, less efficient version of the present work has appeared in [48], [49] where we employed a different registration approach by using a reference image in the frame of which PAs were created and transfered to the target by pairwise registration.

We describe our approach in this section and present evaluations in the subsequent one. The last section concludes the paper pointing out future directions of research.

A. SURF keypoint-based image registration

Image registration is a key component in our segmentation method using PAs. Dense deformable registration methods are robust to shape variability but they exhibit high computational cost [20], which is a problem for the construction of PAs for over two dozen structures. Moreover, as PAs do not constitute the dominant decision factor in segmentation, suboptimal PAs can be nearly as effective as ones built via deformable registration. Imperfections can be balanced out by criteria such as organ appearance and spatial coherence.

We use a fast and robust registration algorithm based on SURF keypoints and a homothetic transformation, capable of registering a pair of images within a few seconds. Keypoints are extracted from each 2D axial slice, however, registration is computed in 3D i.e. the transformation is estimated in the 3D object space, regardless of image spacing, hence the (2+1)D designation. Volumes with partial overlap are well handled, which is important for localizing organs not entirely falling in the image frame. In this paper we improve the registration method detailed in [49] by following a multiscale approach.

1) Feature extraction and matching: We currently extract SURF keypoints [41], but our method is generic and could use other off-the-shelf descriptors as well. To reduce computation time, we first isotropically resample the volume so that its second longest dimension is equal to a desired resolution R. Next, we extract SURF features from each axial slice. Fig. 1 shows features found in a pair of slices in Visceral training dataset images `10000108_1_CTce_ThAb` and `10000109_1_CTce_ThAb`. The total number of features is 11500 and 9400, respectively. Lastly, extracted features are matched using the second closest ratio criterion [40]. Fig. 1 shows the 9 matching keypoint pairs found in both slices.

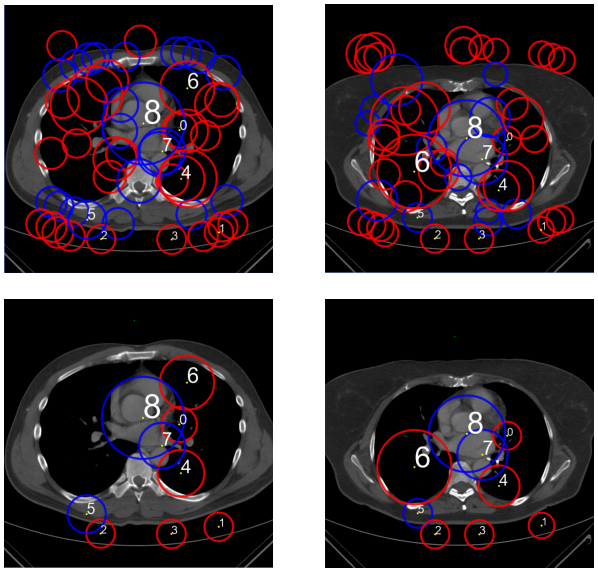


Fig. 1. (top) Features extracted from axial slices in a pair of Visceral training images, (bottom) the 9 matching features. A feature is represented by a circle of radius proportional to the scale at which it is detected, blue and red circles correspond to positive and negative Laplacian values [41]. Reproduced from Fig. 1 in [49].

2) (2+1)D registration: Once keypoint pairs are found, we proceed with volume registration. One problem with keypoint-based registration is that the number of true matching keypoint pairs is relatively low. Hence, for robustness, we use a simple homothetic transformation model:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = s \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} . \quad (1)$$

We estimate s, t_x, t_y and t_z on keypoint pairs similarly to the RANSAC method [50]. The use of a single scale factor preserves the intrinsic shape of an organ, which might otherwise be compromised with higher degrees of freedom. Rotation parameters are unnecessary because we currently process images with consistent patient orientations.

3) Registration in shrinking frames: To improve the robustness of organ localization proposed in [49], we follow a multiscale approach seeking a balance between registration domain size and accuracy according to the following heuristic: a large image contains many keypoints, therefore its registration would be robust, however its structures will all be registered by the same transformation limiting the accuracy of individual structure registrations. In contrast, a smaller image contains fewer keypoints, therefore its registration would be less robust but might be locally more accurate. Therefore, when registering a source patient organ with its counterpart in the target, we start by registering patients on the entire image frame, then progressively shrink registration frames while converging towards the bounding box of the source organ. To register an organ O, our algorithm performs the following steps:

- Set the source frame F_1 to the whole image of patient 1, and the target frame F_2 to that of patient 2.
- Register F_1 onto F_2.

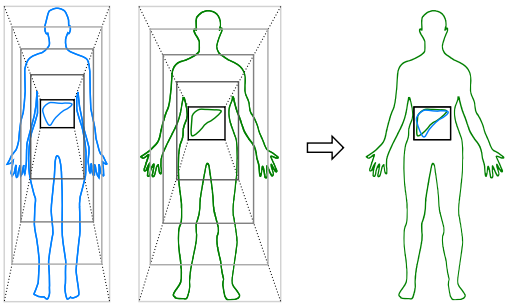


Fig. 2. Multiscale registration in shrinking frames. Source (blue) and target (green) patients are initially set to entire images. The algorithm alternates registration and shrinking for a number of iterations, or until registration fails.

- If registration is successful, i.e. the number of RANSAC inliers is above a threshold θ, shrink F_1 and F_2 towards the bounding box of source O, and recompute the transform, otherwise use the previous transform and terminate.

In our experiments, we use 5 shrinking steps setting θ to 20. Fig. 2 illustrates the procedure. We note that our global-to-local registration strategy is closer to the MRF-based multiscale registration approach taken in the MAS method [16] compared to MAS methods [15], [51] that proceed according to anatomical hierarchies, performing affine registration on thorax-abdomen and deformable registration on organ levels. A related 2-step regression-based method is proposed in [52].

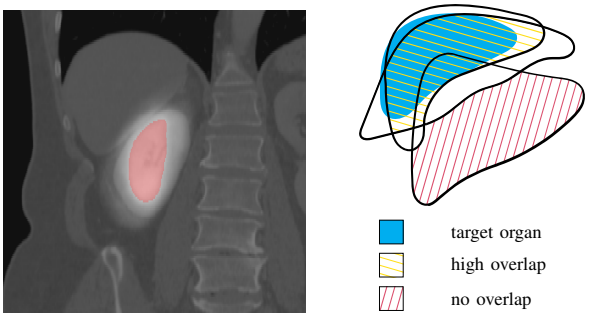
B. Organ probabilistic atlas construction

A target-specific PA of an organ introduces a bias for certain locations in the image for the shape it represents allowing to rule out locations that are unlikely to be part of the organ. Fig. 3a illustrates such a PA. The reference-frame PA creation approach in our earlier work [49], while having lower computational cost, fails to exploit the full variability of the dataset because only a single pairwise registration, that of the reference and target images, is attempted. Moreover the use of a dataset-representative reference introduces bias for it.

Using images and annotations from the Visceral training dataset as atlases [53], we construct target-specific PAs for the 20 structures listed in Table I. We also create PAs for three additional regions: background (BKG), thorax and abdomen (THAB) and body envelope (ENV) from annotations generated automatically as explained in [49]. Full-image modeling by the introduction of these labels gives more discriminative power to our spatial prior and allows to label corresponding regions in the target with the same segmentation algorithm as the structures of interest, rather than resorting to target-dependent preprocessing to remove undesired regions as in [51].

We create a PA separately for each structure in a target image of a given modality according to the following steps:

- Register dataset images of the same modality as the target onto the latter via the procedure described in Section II-A3. Use bounding boxes of organ annotations in dataset images to guide the frame shrinking process.
- Apply obtained transformations to corresponding annotation images (essentially binary mask images).



(a) (b)

Fig. 3. (a) Probabilistic atlas of right kidney (transparent white) and its confidence region (transparent red) overlaid on the organ in a contrasted CT image (Section II-D2). (b) Illustration of annotation ranking: two registered annotations with good overlap having good chances of selection for PA construction, and one poorly registered annotation highly likely to be eliminated.

- Rank registered annotations according to their mutual overlap as measured by the Dice similarity metric. Annotations with greater mean overlap (agreement) with others are ranked higher. Refer to Fig. 3b.
- Select highest ranking annotations eliminating ones with mean overlap inferior to a predefined threshold τ.
- Accumulate selected annotations in a 3D histogram of target image dimensions and normalize it to produce a spatial probability distribution representing the PA.

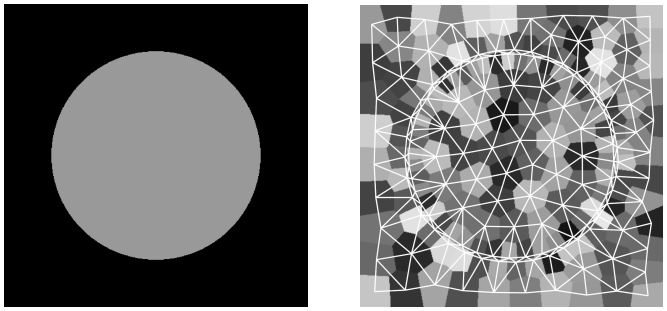
Our ranking scheme is similar to the approach taken in [13], although the latter uses the overlap measure to weigh registered annotations in the PA rather than selecting better and eliminating poorer ones. In practice, we set $\tau = 0.20$.

C. Image clustering

The full-resolution voxel representation is often redundant because objects usually comprise many similar pixels that could be grouped. Therefore, we simplify the image prior to segmentation by an image-adaptive centroidal Voronoi tessellation (CVT) achieving a good balance between cluster compactness and object boundary adherence. We have shown that the clustering improves the runtime and memory footprint of the segmentation up to an order of magnitude without compromising the quality of the result [47], as it remains largely stable across practical settings of clustering resolution. Note however that imprecise tessellations cannot be corrected, as we apply no post-segmentation correction or refinement. We shall leave out algorithmic details of the method and define the graph of a CVT, illustrated in Fig. 4b. Denote the surface of a cluster C_i by ∂C_i. Given a clustering $\mathcal{C}$, let the set $\mathcal{S}$ index its clusters, and let $\mathcal{G} = \langle \mathcal{S}, \mathcal{E} \rangle$ be an undirected graph on cluster centroids where pairs of clusters sharing a surface define the set of edges $\mathcal{E} = \big\{\{i,j\} \mid i,j \in \mathcal{S},\ |\partial C_i \cap \partial C_j| \neq 0\big\}$. Consequently, the neighborhood of a node $i \in \mathcal{S}$ is defined as $\mathcal{N}_i = \big\{j \mid j \in \mathcal{S}, \exists\, \{i,j\} \in \mathcal{E}\big\}$.

D. Multiorgan image segmentation

We formulate image segmentation as a Bayesian labeling problem, defined as the optimal assignment of a label from a set of labels L, representing the structures to be segmented,



(a) (b)

Fig. 4. Adaptive CVT clustering and its graph (b) for a circle image (a). Reproduced from Fig. 5 in [49].

to each of the variables in a set of n variables, indexed by $\mathcal{S}$. Assume that each variable $i \in \mathcal{S}$ represents a cluster of a CVT-clustered image and is associated with the corresponding node in the CVT graph $\mathcal{G}$. An assignment of labels to all variables, denoted by $\ell \in \mathcal{L}$, is called a configuration. An assignment of a label to a single variable is denoted by ℓ_i. In order to find the optimal segmentation, we follow a maximum a posteriori approach and compute the optimal configuration by minimizing the energy of a posterior distribution of ℓ, defined by:

$$E(\ell) = t_1 \sum_{i \in \mathcal{S}} D_i(\ell_i) + t_2 \sum_{i \in \mathcal{S}} P_i(\ell_i) + \frac{1}{2} \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{N}_i} V_{i,j}(\ell_i, \ell_j)\ . \tag{2}$$

In (2), t_1 and t_2 are temperature parameters, $\mathcal{N}_i$ is the neighborhood of the variable i. The first and second sums in (2) correspond respectively to organ intensity and location (PA) likelihood energies, and the third is the energy of a prior distribution of label configurations expressed as a Markov random field (MRF) with respect to $\mathcal{G}$. We minimize (2) via the expansion moves multilabel Graph Cut algorithm [38].

1) Spatial configuration prior: Pairwise terms of (2) encode prior information on interactions between labels assigned to pairs of neighboring variables. They favor the spatial consistency of the labeling with respect to a reference model by favoring valid and penalizing invalid but possible solutions. We define these terms according to the piecewise-constant vicinity prior model we proposed in [47]. Let $\mathcal{A} = \langle L, W \rangle$ be a weighted undirected graph on L where W is the set of unit-weight edges linking pairs of nodes representing adjacent structures in the image. We define the pairwise term in (2) by:

$$V_{i,j}\big(\ell_i, \ell_j\big) = |\partial C_i \cap \partial C_j|\, \omega\big(a, b\big), \quad \ell_i = a,\, \ell_j = b\ . \tag{3}$$

where $\omega\big(a,b\big)$ is the shortest-path weight from a to b in $\mathcal{A}$, $|\partial C_i \cap \partial C_j|$ is the area of the common surface of clusters C_i, C_j ensuring that (2) is independent of clustering resolution. We use the graph given in Fig. 5 to define the spatial prior in experiments involving the Visceral dataset.

2) Intensity and location likelihoods: Unary terms of (2) measure the cost of label assignments. They are defined by:

$$D_i(\ell_i) = -\ln \prod_{v \in C_i} \Pr(I_v \mid \ell_i)\ , \tag{4a}$$

$$P_i(\ell_i) = -\ln \prod_{v \in C_i} \Pr(X_v \mid \ell_i)\ , \tag{4b}$$

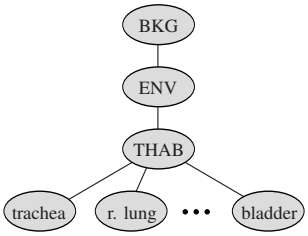


Fig. 5. The adjacency graph used to define the spatial prior in our experiments. Reproduced from Fig. 6 in [49].

where I_v and X_v denote the intensity and the coordinates of voxel v. Location likelihoods $\Pr(X|l)$ are defined from PAs. The intensity likelihood $\Pr(I|l)$ for a given l is estimated as a Gauss-smoothed normalized intensity histogram derived from voxels inside regions of high probability in the corresponding PA. We define confidence regions according to a probability threshold. Fig. 3a gives an illustration for the right kidney.

III. Evaluation Results

We have evaluated our method in Visceral benchmarks [54]. Visceral provides training and test datasets of annotated radiological images and a cloud-based platform allowing participants to develop and evaluate segmentation algorithms in identical environments on the same test dataset. Since the latter comprises few images and is not available to participants, in order to make statistically valid conclusions and to explore parameter settings, we have evaluated our method independently on the training dataset as well. In addition to comparisons with benchmarks participants, we compare our method with state of the art methods and conclude this section by comparing our results for abdominal organs with those reported by specialized single-organ MAS methods.

A. Data

The Visceral dataset [53] is divided into training and testing subsets comprising, respectively, 10 and 20 annotated large FOV images of 4 modalities with the following mean image dimensions and voxel sizes: contrast-enhanced thoracic-abdominal CT (`CTce_ThAb`, $512\times512\times438$ voxels, $0.71\times0.71\times1.5$ mm), unenhanced whole-body CT (`CT_wb`, $512\times512\times877$ voxels, $0.84\times0.84\times1.5$ mm), contrast-enhanced abdominal MRI (`MRT1cefs_Ab`, $313\times76\times384$ voxels, $1.25\times3\times1.25$ mm) and unenhanced whole-body MRI (`MRT1_wb`, $391\times29\times1469$ voxels, $1.26\times6\times1.26$ mm). Annotations for all structures listed in Table I are provided for all modalities except abdominal MRI, which does not carry annotations for thoracic structures. We discard these structures in segmentation. Details on subjects and acquisition conditions are given in the Visceral project documentation [55].

B. Performance metrics

We measure segmentation quality with respect to ground truth via two metrics, the well-known Dice similarity metric (DSM) and mean surface distance (MSD). These metrics are complementary in that DSM measures volume overlap whereas MSD measures the accuracy of boundary delineation.

Visceral benchmarks follow a per-anatomy evaluation strategy that is more suited to single-organ or sequential multiorgan segmentation methods. Simultaneous multiorgan methods, producing segmentations for all organs in a single run with a single parameterization ought to be evaluated additionally on the entire image. Thus, in addition to DSM and MSD for individual structures, we employ "overall" metrics calculated from respective mean weighted measures for all structures.

Let S_l and T_l represent the sets of voxels labeled with $l\in L$ in the segmented image $\mathcal{I}$ and the ground-truth annotation $\mathcal{T}$ respectively, and denote the DSM for a structure $l\in L$ by $\mathrm{DSM}_l(\mathcal{I},\mathcal{T})$. We define the "overall" DSM metric by:

$$\underset{L}{\mathrm{DSM}}(\mathcal{I},\mathcal{T})=\frac{\sum_{l\in L}\mathrm{DSM}_l(\mathcal{I},\mathcal{T})\;|T_l|}{|\mathcal{T}|}, \tag{5}$$

where $|\cdot|$ denotes the structure size in voxels. Note that the weighting mechanism allows larger structures to dominate (5) biasing it against smaller ones.

Let $M^l_{\mathcal{S}}$ and $M^l_{\mathcal{T}}$ be the surfaces of structure volumes labeled by $l\in L$ in the segmented image $\mathcal{I}$ and the ground-truth annotation $\mathcal{T}$. The MSD for $l\in L$ is given by:

$$\underset{l}{\mathrm{MSD}}=\max\big(d(M^l_{\mathcal{S}},M^l_{\mathcal{T}}),d(M^l_{\mathcal{T}},M^l_{\mathcal{S}})\big), \tag{6}$$

where $d(A,B)$ is the directed mean distance in millimeters. For a pair of surfaces A and B, $d(A,B)$ is given by:

$$d(A,B)=\frac{1}{M}\sum_{a\in A}\min_{b\in B}\|a-b\|. \tag{7}$$

The overall MSD metric is defined similarly to (5).

C. Qualitative evaluation

Fig. 6 gives 3D illustrations of multiorgan segmentations produced by our method on 2 images of different modalities from the Visceral training dataset. Corresponding 2D coronal views are given in Fig. 7. Further qualitative and quantitative results are given in the supplement. In 3D views, segmentations are represented by surfaces extracted from labeled volumes, overlaid on coronal cross sections of corresponding images and rendered with transparencies to allow occluded structures to be visible. Fig. 6 is organized in 2 groups of 3 views presenting 3 multiorgan segmentations for the same image. The leftmost view in each group presents the best image segmentation as measured by (5), the middle presents an aggregate of best organ segmentations as measured by DSM and the rightmost view presents the ground-truth annotation. It is obvious that most structures in globally or individually evaluated segmentations have similar segmentation quality, especially larger ones such as the lungs and the liver. Differences can be observed on smaller organs, e.g. the oversegmentation of the bladder in `CTce_ThAb` (left).

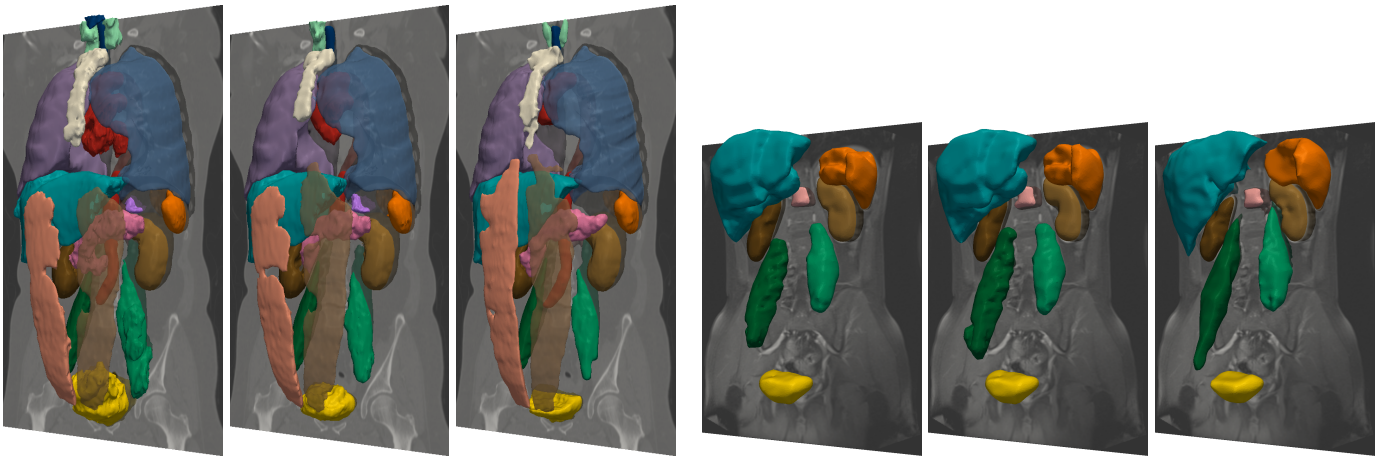


Fig. 6. 3D rendering of multiorgan segmentations of Visceral training dataset images `10000109_1_CTce_ThAb` (left) and `10000331_4_MRT1cefs_Ab` (right). In each group, the leftmost view gives the best image segmentation, the middle an aggregate of best organ segmentations and the rightmost the ground-truth annotation.

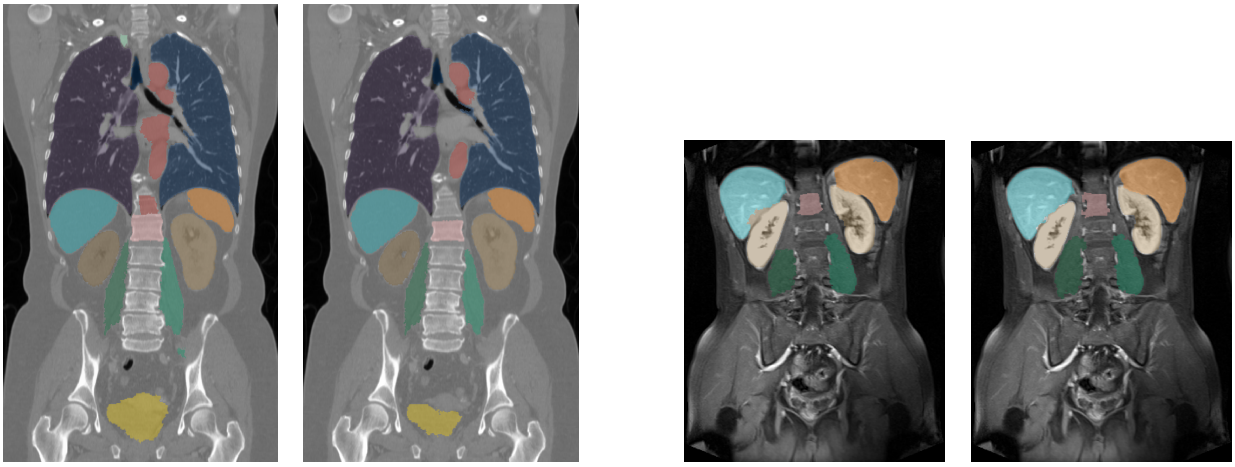


Fig. 7. 2D coronal views of multiorgan segmentations of Visceral training dataset images `10000109_1_CTce_ThAb` (left) and `10000331_4_MRT1cefs_Ab` (right). In each group, the left view gives the best image segmentation, the right an aggregate of best organ segmentations.

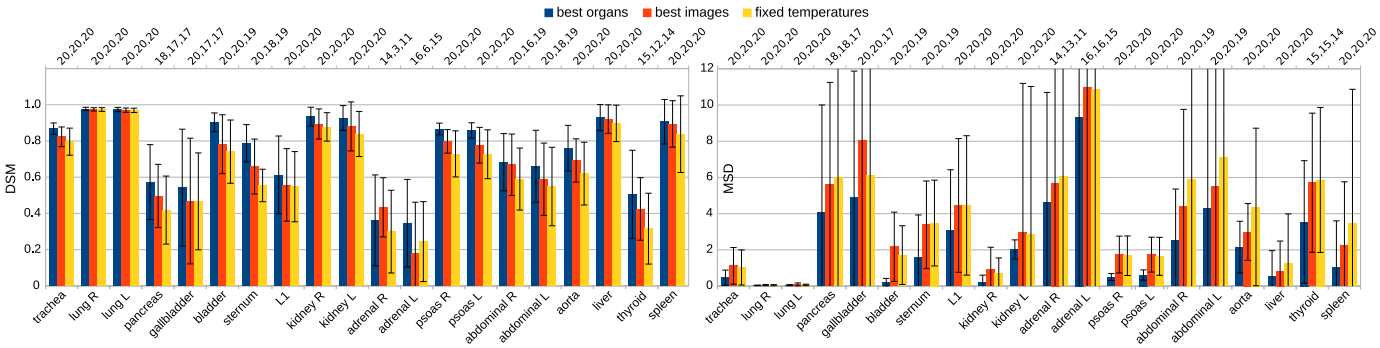


Fig. 8. Comparison of best organ segmentations vs. organs coming from best image segmentations vs. organ segmentations produced by fixed parameters on the Visceral training dataset `CTce_ThAb`. Sample sizes are given above the upper horizontal axis. Refer to Section III-D for details.

D. Results on the Visceral training dataset

For all images in the dataset, we follow a leave-one-out approach by selecting a target and using all remaining annotated images in the same modality to construct PAs. We set the registration resolution parameter R to 150 voxels. 100 segmentations are attempted per image, corresponding to combinations of energy parameters t_1 and t_2 (2) uniformly sampled on empirically determined ranges. For CT modalities $t_1 \in [0.1, 1.0]$, and for MRI $t_1 \in [0.5, 1.4]$. Irrespective of modality, t_2 is set to $c \times t_1$, where $c \in [0.1, 1.0]$. Due to differ-

TABLE I
QUANTITATIVE LEAVE-ONE-OUT EVALUATION RESULTS OF THE PROPOSED METHOD ON ALL FOUR MODALITIES OF THE VISCERAL TRAINING DATASET.

	CTce_ThAb			CT_wb			MRT1cefs_Ab			MRT1_wb		
Structures	#	DSM	MSD	#	DSM	MSD	#	DSM	MSD	#	DSM	MSD
trachea	20	0.868 ± 0.031	0.465 ± 0.419	20	0.878 ± 0.047	0.349 ± 0.327	0	-	-	20	0.556 ± 0.300	3.114 ± 8.672
lung R	20	0.978 ± 0.009	0.037 ± 0.019	20	0.974 ± 0.011	0.043 ± 0.025	0	-	-	20	0.883 ± 0.037	0.312 ± 0.154
lung L	20	0.974 ± 0.012	0.049 ± 0.034	20	0.972 ± 0.011	0.046 ± 0.031	0	-	-	19	0.866 ± 0.081	0.404 ± 0.349
pancreas	18	0.573 ± 0.207	4.066 ± 5.940	20	0.441 ± 0.151	4.717 ± 2.943	11	0.338 ± 0.265	4.624 ± 4.296	5	0.137 ± 0.147	8.798 ± 5.041
gallbladder	20	0.542 ± 0.323	4.882 ± 6.993	18	0.179 ± 0.200	15.96 ± 15.52	10	0.133 ± 0.245	21.06 ± 21.01	6	0.093 ± 0.227	59.34 ± 89.44
bladder	20	0.903 ± 0.052	0.215 ± 0.207	19	0.747 ± 0.124	1.232 ± 0.930	20	0.673 ± 0.269	1.952 ± 3.246	20	0.700 ± 0.277	1.806 ± 3.064
sternum	20	0.788 ± 0.103	1.557 ± 2.371	20	0.772 ± 0.098	1.035 ± 1.194	0	-	-	5	0.280 ± 0.139	4.401 ± 2.749
L1	20	0.613 ± 0.215	3.067 ± 3.358	20	0.486 ± 0.149	3.388 ± 1.984	13	0.375 ± 0.196	4.462 ± 3.241	19	0.495 ± 0.264	2.900 ± 3.162
kidney R	20	0.935 ± 0.052	0.208 ± 0.397	20	0.762 ± 0.175	1.651 ± 2.415	17	0.783 ± 0.191	1.771 ± 2.740	20	0.784 ± 0.155	0.704 ± 0.870
kidney L	20	0.941 ± 0.069	0.190 ± 0.530	20	0.852 ± 0.107	0.670 ± 1.072	19	0.844 ± 0.181	1.859 ± 5.073	19	0.745 ± 0.236	0.927 ± 1.315
adrenal R	14	0.361 ± 0.251	4.621 ± 6.077	13	0.193 ± 0.153	5.127 ± 3.232	3	0.001 ± 0.002	16.83 ± 10.04	2	0.006 ± 0.008	2.770 ± 0.862
adrenal L	16	0.346 ± 0.242	9.340 ± 20.26	14	0.219 ± 0.111	4.370 ± 2.868	5	0.166 ± 0.227	9.975 ± 8.625	6	0.006 ± 0.014	14.33 ± 20.01
psoas R	20	0.866 ± 0.033	0.496 ± 0.192	20	0.792 ± 0.101	1.085 ± 0.776	20	0.749 ± 0.061	0.831 ± 0.301	20	0.747 ± 0.168	1.200 ± 2.099
psoas L	20	0.858 ± 0.042	0.610 ± 0.284	20	0.792 ± 0.083	1.049 ± 0.693	20	0.709 ± 0.086	1.318 ± 0.921	20	0.728 ± 0.248	9.779 ± 39.65
abdominal R	20	0.683 ± 0.158	2.509 ± 2.849	20	0.510 ± 0.213	4.442 ± 3.647	3	0.063 ± 0.057	14.42 ± 11.56	2	0.132 ± 0.185	5.457 ± 1.173
abdominal L	20	0.661 ± 0.199	4.262 ± 8.083	20	0.545 ± 0.263	5.058 ± 6.670	4	0.197 ± 0.164	6.592 ± 2.810	2	0.075 ± 0.103	6.175 ± 0.003
aorta	20	0.760 ± 0.126	2.150 ± 1.433	20	0.621 ± 0.109	2.674 ± 1.150	3	0.360 ± 0.288	6.440 ± 3.623	20	0.540 ± 0.083	1.741 ± 0.710
liver	20	0.929 ± 0.072	0.546 ± 1.410	20	0.889 ± 0.039	0.689 ± 0.508	20	0.868 ± 0.052	0.512 ± 0.426	19	0.818 ± 0.038	0.696 ± 0.339
thyroid	15	0.505 ± 0.244	3.543 ± 3.383	17	0.444 ± 0.223	3.475 ± 3.018	0	-	-	13	0.283 ± 0.266	3.047 ± 2.138
spleen	20	0.906 ± 0.123	1.030 ± 2.570	20	0.898 ± 0.053	0.376 ± 0.603	20	0.817 ± 0.113	0.717 ± 0.812	20	0.739 ± 0.102	0.680 ± 0.629

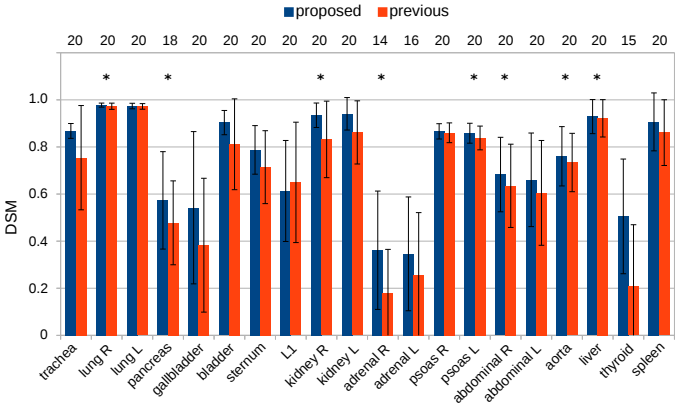


Fig. 9. Comparison of the proposed method with our previous [49] on the Visceral training dataset CTce_ThAb. Sample sizes are given above the upper horizontal axis, significant improvements at $p = 0.01$ indicated via "*".

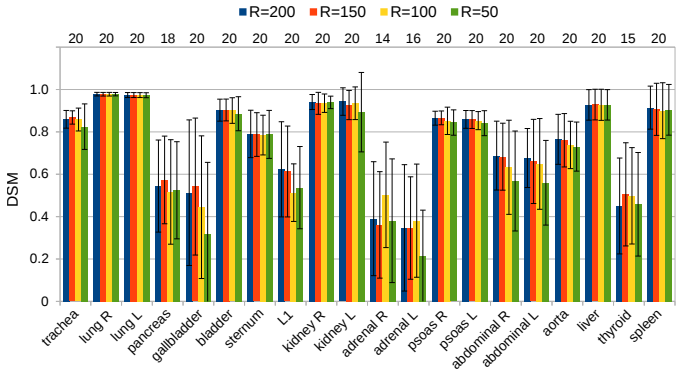


Fig. 10. Evaluation of the impact of the registration resolution parameter R on the quality of segmentation as measured by structure mean DSM on the Visceral training dataset CTce_ThAb. Sample sizes are at the top.

ences in mean image dimensions, we use a different clustering resolution for each modality to allow maximum resolution given memory capacity available to the algorithm. Numbers of CVT clusters for CTce_ThAb, CT_wb, MRT1cefs_Ab and MRT1_wb are set respectively to 5%, 3%, 20% and 20% of image voxel count. In intensity likelihood estimation for all structures, we fix the confidence threshold to 0.75 times that of the maximum probability of corresponding PAs.

In Fig. 8, we present the three quantitative evaluation strategies we have followed on the Visceral training dataset CTce_ThAb modality only. Results on other modalities follow similar patterns. Blue bars (leftmost bars in each group of three) give mean DSM and MSD values calculated over best structure segmentations produced by a parameter setting. Orange bars (middle) give mean DSM and MSD values calculated on organs coming from best image segmentations, as measured by overall DSM (5) and MSD, produced by a per-image parameter setting. Yellow bars (rightmost) give mean DSM and MSD values calculated on structure segmentations produced by a fixed setting for all images, $t_1 = 0.2$ and $t_2 = 0.12$, found empirically to produce good results for most organs in this modality in terms of organ presence in results and acceptable segmentation quality. Naturally, blue bars indicate the best performance, however orange ones corresponding to overall performance are quite similar, especially on major thoracic and abdominal organs, while falling short on smaller ones due to the bias in the overall metric. Lastly, yellow bars indicate performance implications of using "default" parameters in situations where seeking an optimal setting would be impractical. The mean DSM value on all structures in the presented modality is about 10% lower in the fixed setting compared to optimized settings.

Table I presents the results of quantitative evaluation on all 4 modalities of the Visceral training dataset. For each structure in the leftmost column, we give mean DSM and MSD values calculated over best structure segmentations produced by a parameter setting. We give the number of structures for each modality under the "#" column. Since the Visceral dataset does not provide annotations for all structures in all images, this number is not always equal to 20.

In Fig. 9 we compare the proposed method with our

TABLE II
VISCERAL BENCHMARK EVALUATION RESULTS OF THE PROPOSED AND OTHER MULTIORGAN SEGMENTATION METHODS ON THE VISCERAL TEST DATASET CTCE_THAB. BEST PERFORMANCES ARE HIGHLIGHTED IN BOLD FACE.

		Proposed		Gass et al. [16]		Jimenez et al. [15]		Wang & Smedby [51]	
Structures	#	DSM	MSD	DSM	MSD	DSM	MSD	DSM	MSD
trachea	10	0.834 ± 0.050	0.538 ± 0.319	0.847 ± 0.050	0.378 ± 0.515	**0.855 ± 0.022**	**0.223 ± 0.046**	-	-
lung R	10	**0.973 ± 0.016**	**0.049 ± 0.030**	0.965 ± 0.013	0.069 ± 0.035	0.963 ± 0.013	0.065 ± 0.032	0.971 ± 0.014	0.070 ± 0.034
lung L	10	**0.972 ± 0.015**	**0.050 ± 0.029**	0.961 ± 0.011	0.121 ± 0.107	0.959 ± 0.010	0.071 ± 0.022	**0.972 ± 0.013**	0.076 ± 0.061
pancreas	4	**0.585 ± 0.132**	4.459 ± 1.885	0.460 ± 0.159	**3.472 ± 2.270**	0.423 ± 0.136	3.804 ± 2.867	-	-
gallbladder	8	**0.673 ± 0.220**	**2.433 ± 3.134**	0.381 ± 0.208	6.314 ± 7.680	0.484 ± 0.132	3.603 ± 2.910	-	-
bladder	10	0.848 ± 0.097	0.629 ± 0.644	0.683 ± 0.090	1.514 ± 0.639	0.679 ± 0.142	1.879 ± 1.192	**0.866 ± 0.070**	**0.375 ± 0.284**
sternum	8	**0.784 ± 0.112**	**0.801 ± 0.755**	0.635 ± 0.148	1.257 ± 0.941	0.721 ± 0.058	0.899 ± 0.388	0.762 ± 0.092	0.993 ± 0.649
L1	8	0.584 ± 0.233	7.601 ± 6.271	**0.624 ± 0.356**	**3.228 ± 5.710**	0.523 ± 0.301	4.504 ± 5.509	-	-
kidney R	10	0.950 ± 0.013	0.087 ± 0.037	0.914 ± 0.027	0.199 ± 0.116	0.889 ± 0.026	0.243 ± 0.097	**0.959 ± 0.011**	**0.072 ± 0.030**
kidney L	10	**0.947 ± 0.014**	**0.092 ± 0.042**	0.913 ± 0.029	0.335 ± 0.403	0.910 ± 0.015	0.172 ± 0.046	0.945 ± 0.027	0.137 ± 0.127
adrenal R	4	0.290 ± 0.205	3.180 ± 1.910	0.213 ± 0.139	3.035 ± 1.588	**0.342 ± 0.148**	**2.660 ± 1.437**	-	-
adrenal L	4	0.304 ± 0.283	8.632 ± 8.740	0.250 ± 0.159	3.900 ± 2.906	**0.331 ± 0.176**	**3.115 ± 1.965**	-	-
psoas R	10	0.818 ± 0.024	0.989 ± 0.390	-	-	0.799 ± 0.025	0.757 ± 0.230	**0.845 ± 0.026**	**0.671 ± 0.321**
psoas L	10	0.797 ± 0.075	1.036 ± 0.673	0.813 ± 0.046	0.622 ± 0.277	0.794 ± 0.049	0.742 ± 0.298	**0.830 ± 0.074**	**0.638 ± 0.321**
abdominal R	6	**0.633 ± 0.176**	**4.763 ± 4.905**	-	-	0.453 ± 0.173	6.600 ± 5.901	-	-
abdominal L	5	**0.703 ± 0.137**	**3.276 ± 3.255**	-	-	0.474 ± 0.180	6.068 ± 7.420	-	-
aorta	10	0.681 ± 0.130	6.219 ± 7.854	**0.785 ± 0.042**	**1.011 ± 0.619**	0.762 ± 0.039	1.094 ± 0.508	-	-
liver	10	**0.950 ± 0.012**	0.182 ± 0.068	0.908 ± 0.021	0.646 ± 0.378	0.887 ± 0.019	0.514 ± 0.179	0.949 ± 0.010	**0.174 ± 0.075**
thyroid	5	0.375 ± 0.170	4.427 ± 2.568	0.184 ± 0.166	5.847 ± 2.749	**0.410 ± 0.157**	**3.337 ± 1.295**	-	-
spleen	10	**0.911 ± 0.069**	**0.557 ± 1.364**	0.781 ± 0.075	1.530 ± 1.144	0.730 ± 0.116	2.005 ± 1.967	0.909 ± 0.069	0.573 ± 1.210

previous [49]. We limit the presentation to the CTce_ThAb modality using only the DSM metric. We can see that the proposed method improves on the previous for all structures except one. Statistical testing using one-tailed paired t-test confirms that improvements are significant on the $p = 0.05$ level for 18 structures, and for 8 structures on $p = 0.01$. No improvements were made for the 1st lumbar vertebra (L1). This can be explained as follows. The proposed method performs several registrations to construct the structure PA and has greater chances of missing the target L1 and registering against nearby vertebrae compared to the previous approach that performs a single registration between the target and a reference image onto which PAs are constructed in advance.

We close this section by an evaluation of the segmentation quality vs. runtime trade-off induced by settings of the registration resolution parameter R on the training dataset CTce_ThAb. Fig. 10 gives mean DSM values of best structure segmentations for 4 settings of R: 200, 150, 100 and 50. Table III summarizes corresponding mean runtimes. All algorithms are implemented in C++ and run in a 64-bit Linux environment on a laptop computer with a CPU speed of 2.1 GHz and a RAM capacity of 16 GB. Table III shows that a threefold reduction of PA construction time is possible via $R = 50$, and that PA construction time dominates total runtime. We construct PAs sequentially, which is not mandatory since PAs are independent, therefore a speedup proportional to the number of CPUs is possible by parallel PA creation. The current per-organ runtime however is close to that of the recent patch-based method [26] where 2 h are needed to segment 5 abdominal structures using 20 atlases, i.e. 0.4 h per organ. In comparison, using 20 atlases to segment 23 structures, our method requires about 0.57 h per organ for $R = 100$. Fig. 10 confirms that the segmentation quality remains relatively stable for larger organs, such as the lungs and major abdominal organs, and even for the thin and elongated but well contrasted trachea. However important deteriorations in segmentation quality can be observed for smaller or thinner poorly contrasted structures such as abdominal muscles and the gallbladder. We think that the quality–runtime trade-off can be resolved only in the context of an application. We discuss fundamental approaches to runtime reduction in Section IV.

TABLE III
MEAN (PER IMAGE) MEMORY FOOTPRINT AND RUNTIME FIGURES OF PROPOSED ALGORITHMS MEASURED ON THE VISCERAL DATASET CTCE_THAB FOR VARYING REGISTRATION RESOLUTIONS.

R	Mem. (MB)	PA constr.*(s)	Clust. (s)	Seg. (s)	Total (h)
50	10542	27683	3827	1327	9.12
100		42669		1072	13.21
150		90116		1010	26.38
200		127965		1257	36.96

* sequential

E. Visceral benchmark results and comparisons

We present evaluation results obtained on the Visceral test dataset during Visceral benchmarks. We mention that out of a dozen participating groups, only 2, other than ours, attempted to segment all structures in all modalities. The cloud-based platform, evaluation conditions and results of participants up to 2016 are described in [54]. Results are also published online on a leaderboard [57], this, however is left to the discretion of participants. Therefore we refer to both [54] and the leaderboard to obtain complete figures for comparisons.

Table II lists our results along with those reported by participants who have segmented at least half of the structures in the CTce_ThAb modality. Comparisons on other modalities are given in the supplement. For each structure, we report mean DSM and MSD values along with the number of produced segmentations. Unfortunately we do not have access to the latter figures for other participants. Due to the impracticality of searching best settings of segmentation parameters t_1 and t_2 in the virtual machine execution environment, in benchmark runs we use temperature settings that have produced best overall

TABLE IV
COMPARISON WITH STATE OF THE ART METHODS. DSM IS GIVEN AS A PERCENTAGE, MSD IN MILLIMETERS.

Method	Subjects	Metric	Lungs	Liver	Kidneys	Spleen	Pancreas	Bladder	Gallbladder	Aorta
Proposed method ‡ best parameter settings	20	DSM	97.6 ± 1.1	92.9 ± 7.2	93.8 ± 6.1	90.6 ± 12.3	57.3 ± 20.7	90.3 ± 5.2	54.2 ± 32.3	76.0 ± 12.6
		MSD	0.1 ± 0.0	0.6 ± 1.4	0.2 ± 0.5	1.0 ± 2.6	4.1 ± 5.9	0.2 ± 0.2	4.9 ± 7.0	2.2 ± 1.4
Proposed method ‡ fixed parameter setting	20	DSM	97.2 ± 1.1	89.7 ± 10.0	85.8 ± 10.16	83.8 ± 21.2	41.8 ± 18.8	74.1 ± 17.5	46.7 ± 26.8	62.0 ± 17.3
		MSD	0.1 ± 0.0	1.2 ± 2.8	1.8 ± 4.5	3.5 ± 7.4	6.0 ± 6.4	1.7 ± 1.6	6.1 ± 8.2	4.3 ± 4.4
Oliveira et al. [19] ‡	20	DSM	97.7 ± 1.2	93.6 ± 2.8	93.2 ± 7.2	91.0 ± 9.4	57.2 ± 14.7	65.7 ± 17.3	51.8 ± 27.5	86.2 ± 6.6
		MSD	0.9 ± 0.4	2.7 ± 1.1	1.9 ± 3.0	2.7 ± 3.5	6.7 ± 5.6	6.4 ± 2.4	7.9 ± 10.9	2.1 ± 1.2
Heinrich et al. [28] ‡	20	DSM	-	90.9 ± 5.1	89.9 ± 5.4	83.3 ± 17.3	-	62.3 ± 18.8	-	-
Pawlowski et al. [34] *	20	DSM	-	95.7	91.1	92.5	72.2	-	67.6	87.5
Larsson et al. [31] *	20	DSM	-	94.6	88.8	93.1	60.2	-	62.4	86.1
		MSD	-	1.7	2.0	1.9	4.5	-	8.7	5.0
Heinrich et al. [14], [17] *	20	DSM	-	94.8	90.8	91.9	74.0	-	60.4	85.7
		MSD	-	1.5	1.1	1.6	2.3	-	7.0	3.6
Wang et al. [14] *	20	DSM	-	94.6	88.7	92.8	65.1	-	68.0	83.4
		MSD	-	1.5	1.7	1.4	4.9	-	∞	5.2
Tong et al. [26] †	150	DSM	-	94.9 ± 1.9	93.6 ± 3.8	92.5 ± 6.5	71.1 ± 14.7	-	-	-
Wolz et al. [13] †	150	DSM	-	94.0 ± 2.8	92.5 ± 7.2	92.0 ± 9.2	69.6 ± 16.7	-	-	-
		MSD	-	2.0 ± 2.8	2.3 ± 3.4	2.3 ± 3.0	3.7 ± 4.4	-	-	-
Okada et al. [23]	134	DSM	-	94.1 ± 2.4	91.6 ± 14.3	92.1 ± 8.1	72.5 ± 17.6	-	63.5 ± 28.5	85.0 ± 18.6
		MSD	-	1.7 ± 0.9	1.5 ± 2.9	1.2 ± 1.6	3.0 ± 3.1	-	5.2 ± 5.5	2.1 ± 3.3
Chu et al. [12]	100	DSM	-	95.1 ± 1.0	90.1 ± 5.0	91.4 ± 5.7	69.1 ± 15.3	-	-	-
		MSD	-	1.2 ± 0.2	1.3 ± 0.4	0.9 ± 0.4	1.9 ± 0.6	-	-	-
Bagci et al. [22]	20	DSM	-	95.8 ± 0.6	96.5 ± 0.7	96.5 ± 0.8	-	-	-	-
Linguraru et al. [10]	20	DSM	-	95.6 ± 0.6	92.6 ± 2.3	91.8 ± 1.5	-	-	-	
		MSD	-	1.1 ± 0.4	0.8 ± 0.4	1.0 ± 0.5	-	-	-	
Kohlberger et al. [36]	120,100,20	MSD	1.6 ± 0.5	2.9 ± 1.7	1.2 ± 1.0	-	-	-	-	-
Seifert et al. [30]	226,335,203,,53	MSD	-	1.3 ± 0.5	1.1 ± 0.4	2.1 ± 1.2	-	1.4 ± 0.8	-	-

‡ Visceral training dataset (public) * Beyond the Cranial Vault challenge dataset (public) [56] † Nagoya University Hospital dataset (private)

segmentations as measured by (5) on the Visceral training dataset. We have also tested few nearby settings in an attempt to produce better segmentations for smaller structures.

F. Comparison with the state of the art

It is difficult to compare one's method to the state of the art due to differences in datasets, evaluation strategies and metrics. Nevertheless, in Table IV we present a quantitative comparison of multiorgan segmentation methods representative of the families of approaches discussed in Section I-A. With few exceptions, these methods have been developed and evaluated on contrast-enhanced abdominal CT images. We report results for commonly segmented structures. Methods evaluated on the same dataset appear with a common symbol in the table. We can see that our method achieves comparable performances to those of related methods while opting for simpler algorithms and segmenting 4–5 times as many structures.

G. Comparison with single organ MAS methods

Finally we compare our results to those reported by 3 specialized MAS methods for major abdominal organs. The comparison confirms that our generic method achieves close results to and can even outperform these methods.

1) Liver: In [58], a MAS approach using PAs constructed by nonrigid registration is used in a graph-cut segmentation incorporating an intensity model, a label prior probability and spatial regularizer based on a Finsler metric. On 10 contrast-enhanced CT images, reported mean DSM = 0.973 ± 0.007. In comparison, we obtain 0.929 ± 0.072 on 20 such images.

2) Kidneys: In [59], a two-step approach is followed whereby kidneys are first located via affine registration of atlases in low resolution, then aligned to higher resolution atlases via deformable registration to obtain final segmentations. On 22 kidneys segmented from contrasted angiographic CT images, mean DSM = 0.952 ± 0.018 and MSD = 0.913 ± 1.06 mm. On 40 kidneys in similar images, we achieve mean DSM = 0.938 ± 0.061 and mean MSD = 0.20 ± 0.46 mm.

3) Spleen: In [60], atlas-to-target registrations are used to locate the spleen, the fusion of which provides a shape constraint for a level-set segmentation. On 25 contrasted CT images, mean DSM = 0.83 ± 0.08 and mean MSD = 3.48 ± 1.88 mm. In comparison, we obtain mean DSM = 0.906 ± 0.123 and mean MSD = 1.030 ± 2.570 mm on 20 such cases.

IV. FUTURE WORK

There are several opportunities for near-term improvements on the present method. Table I shows that it performs better on CT than on MR images. The reason is that CT images have consistent appearances, whereas MR images suffer intensity inhomogeneity the correction of which [61] can improve registration and segmentation. Furthermore, the accuracy of PAs for some "stray" structures, e.g. L1, can be improved by initializing registration for such structures from those of more stable neighbors, e.g. psoas major muscles and the kidneys.

Generic data-independent methods are suitable for the unattended processing of large datasets. With similar segmentation performances by most methods, the main challenge today is to achieve significant improvements in runtime with increasing numbers of anatomies while maintaining accuracy levels. Recent approaches, such as patch-based methods, while eliminating deformable registration, are still computationally expensive, leading in practice to the use of fewer atlases thus reducing the robustness of the method. Deep learning approaches require large amounts of annotated data and dedicated hardware to be effective. We think that sparse representations constitute a promising approach to overcome the burden of image data and to accelerate algorithms significantly. For keypoint-based representations [46], the problem of false matches, the number of which can be well above one third of all matches [62], has to be addressed. We think that metric learning methods [63], especially in 3D, can be beneficial in this context.

Another interesting future research direction is the ability to handle images of varying FOV where not all atlas dataset structures are present. This too is beneficial for the batch processing of large heterogeneous datasets. In this case, an initial detection of the image's content would subsequently allow to select suitable atlases to define models only for structures appearing in the image. Content detection could also be carried out by a keypoint-based approach [64].

ACKNOWLEDGMENTS

We thank the Visceral Consortium for allowing us to use the Visceral training dataset in independent evaluations.

REFERENCES

[1] H. Jacinto, R. Kéchichan, M. Desvignes, R. Prost, and S. Valette, "A web interface for 3D visualization and interactive segmentation of medical images," in *17th Int. Conf. on 3D Web Technology*, 2012, pp. 51–58.

[2] J. E. Iglesias and M. Sabuncu, "Multi-atlas segmentation of biomedical images: a survey," *Med. Image Anal.*, vol. 24, no. 1, pp. 205–219, 2015.

[3] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[4] D. Pasquier, T. Lacornerie, M. Vermandel, J. Rousseau, E. Lartigau, and N. Betrouni, "Automatic segmentation of pelvic structures from magnetic resonance images for prostate cancer radiotherapy," *Int. J. of Radiation Oncology*Biology*Physics*, vol. 68, no. 2, pp. 592–600, 2007.

[5] J. H. Moltz *et al.*, "Advanced segmentation techniques for lung nodules, liver metastases, and enlarged lymph nodes in CT scans," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 1, pp. 122–134, 2009.

[6] S. Seifert, M. Kelm, M. Moeller, S. Mukherjee, A. Cavallaro, M. Huber, and D. Comaniciu, "Semantic annotation of medical images," in *SPIE Medical Imaging*, 2010, pp. 762 808–762 808.

[7] O. Camara, O. Colliot, and I. Bloch, "Computational modeling of thoracic and abdominal anatomy using spatial relationships for image segmentation," *Real-Time Imaging*, vol. 10, no. 4, pp. 263–273, 2004.

[8] H. Park, P. H. Bland, and C. R. Meyer, "Construction of an abdominal probabilistic atlas and its application in segmentation," *IEEE Trans. Med. Imag.*, vol. 22, no. 4, pp. 483–492, 2003.

[9] J. Iglesias, E. Konukoglu, A. Montillo, Z. Tu, and A. Criminisi, "Combining generative and discriminative models for semantic segmentation of CT scans via active learning," in *Inf. Process. in Med. Imaging*, 2011, pp. 25–36.

[10] M. G. Linguraru, J. A. Pura, V. Pamulapati, and R. M. Summers, "Statistical 4D graphs for multi-organ abdominal segmentation from multiphase CT," *Med. Image Anal.*, vol. 16, no. 4, pp. 904–914, 2012.

[11] M. Oda, T. Nakaoka, T. Kitasaka, K. Furukawa, K. Misawa, M. Fujiwara, and K. Mori, "Organ segmentation from 3D abdominal CT images based on atlas selection and graph cut," *Abdominal Imaging. Computational and Clinical Applications*, pp. 181–188, 2012.

[12] C. Chu, M. Oda, T. Kitasaka, K. Misawa, M. Fujiwara, Y. Hayashi, Y. Nimura, D. Rueckert, and K. Mori, "Multi-organ segmentation based on spatially-divided probabilistic atlas from 3D abdominal CT images," in *Proc. MICCAI Conf.*, 2013, pp. 165–172.

[13] R. Wolz, C. Chu, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert, "Automated abdominal multi-organ segmentation with subject-specific atlas generation," *IEEE Trans. Med. Imag.*, vol. 32, no. 9, pp. 1723–1730, 2013.

[14] H. Wang and P. Yushkevich, "Multi-atlas segmentation with joint label fusion and corrective learning – an open source implementation," *Frontiers in neuroinformatics*, vol. 7, p. 27, 2013.

[15] O. Jimenez-del Toro and H. Müller, "Hierarchic multi-atlas based segmentation for anatomical structures: Evaluation in the visceral anatomy benchmarks," in *Proc. MICCAI-MCV Workshop*, 2014, pp. 189–200.

[16] T. Gass, G. Szekely, and O. Goksel, "Multi-atlas segmentation and landmark localization in images with large field of view," in *Proc. MICCAI-MCV Workshop*, 2014, pp. 171–180.

[17] M. Heinrich, O. Maier, and H. Handels, "Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities." in *Proc. ISBI Visceral Challenge*, 2015, pp. 27–30.

[18] F. Kahl, J. Alvén, O. Enqvist, F. Fejne, J. Ulén, J. Fredriksson, M. Landgren, and V. Larsson, "Good features for reliable registration in multi-atlas segmentation." in *ISBI Visceral Challenge*, 2015, pp. 12–17.

[19] B. Oliveira, S. Queirós, P. Morais, H. R. Torres, J. Gomes-Fonseca, J. C. Fonseca, and J. L. Vilaça, "A novel multi-atlas strategy with dense deformation field reconstruction for abdominal and thoracic multi-organ segmentation from CT," *Med. Image Anal.*, vol. 45, pp. 108–120, 2018.

[20] Z. Xu, C. P. Lee, M. P. Heinrich, M. Modat, D. Rueckert, S. Ourselin, R. G. Abramson, and B. A. Landman, "Evaluation of six registration methods for the human abdomen on clinically acquired ct," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 8, pp. 1563–1572, 2016.

[21] X. Chen, J. K. Udupa, U. Bagci, and J. Yao, "Medical image segmentation by combining graph cuts and oriented active appearance models," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2035–2046, 2012.

[22] U. Bagci, X. Chen, and J. K. Udupa, "Hierarchical scale-based multiobject recognition of 3-d anatomical structures," *IEEE Trans. Med. Imag.*, vol. 31, no. 3, pp. 777–789, 2012.

[23] T. Okada, M. G. Linguraru, M. Hori, R. M. Summers, N. Tomiyama, and Y. Sato, "Abdominal multi-organ segmentation from CT images using conditional shape–location and unsupervised intensity priors," *Med. Image Anal.*, vol. 26, no. 1, pp. 1–18, 2015.

[24] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3D medical image segmentation: a review," *Med. Image Anal.*, vol. 13, no. 4, pp. 543–563, 2009.

[25] Z. Wang, K. K. Bhatia, B. Glocker, A. Marvao, T. Dawes, K. Misawa, K. Mori, and D. Rueckert, "Geodesic patch-based segmentation," in *Proc. MICCAI Conf.*, 2014, pp. 666–673.

[26] T. Tong, R. Wolz, Z. Wang, Q. Gao, K. Misawa, M. Fujiwara, K. Mori, J. V. Hajnal, and D. Rueckert, "Discriminative dictionary learning for abdominal multi-organ segmentation," *Med. Image Anal.*, vol. 23, no. 1, pp. 92–104, 2015.

[27] A. Montillo, J. Shotton, J. Winn, J. E. Iglesias, D. Metaxas, and A. Criminisi, "Entangled decision forests and their application for semantic segmentation of CT images," in *Inf. Process. in Med. Imaging*, 2011, pp. 184–196.

[28] M. Heinrich and M. Blendowski, "Multi-organ segmentation using vantage point forests and binary context features." in *Proc. MICCAI Conf.*, 2016, pp. 598–606.

[29] B. Glocker, O. Pauly, E. Konukoglu, and A. Criminisi, "Joint classification-regression forests for spatially structured multi-object segmentation," *Computer Vision–ECCV*, pp. 870–881, 2012.

[30] S. Seifert, A. Barbu, S. K. Zhou, D. Liu, J. Feulner, M. Huber, M. Suehling, A. Cavallaro, and D. Comaniciu, "Hierarchical parsing and semantic navigation of full body CT data," in *SPIE Medical Imaging*, 2009.

[31] M. Larsson, Y. Zhang, and F. Kahl, "Robust abdominal organ segmentation using regional convolutional neural networks," in *Scandinavian Conf. on Image Anal.* Springer, 2017, pp. 41–52.

[32] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. R. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt, "Towards image-guided pancreas and biliary endoscopy: Automatic multi-organ segmentation on abdominal ct with dense dilated networks," in *Proc. MICCAI Conf.*, 2017, pp. 728–736.

[33] I. Lavdas, B. Glocker, K. Kamnitsas, D. Rueckert, H. Mair, A. Sandhu, S. A. Taylor, E. O. Aboagye, and A. G. Rockall, "Fully automatic, multi-organ segmentation in normal whole body magnetic resonance imaging (MRI), using classification forests (CFs), convolutional neural networks (CNNs) and a multi-atlas (MA) approach," *Med. Phys.*, 2017.

[34] N. Pawlowski, S. I. Ktena, M. C. Lee, B. Kainz, D. Rueckert, B. Glocker, and M. Rajchl, "DLTK: State of the art reference implementations for deep learning on medical images," *arXiv preprint 1711.06853*, 2017.

[35] P. Hu, F. Wu, J. Peng, Y. Bao, F. Chen, and D. Kong, "Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 3, pp. 399–411, 2017.

[36] T. Kohlberger, M. Sofka, J. Zhang, N. Birkbeck, J. Wetzl, J. Kaftan, J. Declerck, and S. K. Zhou, "Automatic multi-organ segmentation using learning-based segmentation and level set optimization," in *Proc. MICCAI Conf.*, 2011, pp. 338–345.

[37] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1068–1080, 2008.

[38] Y. Boykov and O. Veksler, "Graph cuts in vision and graphics: Theories and applications," in *Handbook of mathematical models in computer vision*. Springer, 2006, pp. 79–96.

[39] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2008.

[40] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. of Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[41] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[42] W. Cheung and G. Hamarneh, "n-SIFT: n-dimensional scale invariant feature transform," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 2012–2021, 2009.

[43] S. Allaire, J. J. Kim, S. L. Breen, D. A. Jaffray, and V. Pekar, "Full orientation invariance and improved feature selectivity of 3D SIFT with application to medical image analysis," in *IEEE CVPRW*, 2008, pp. 1–8.

[44] M. Urschler, J. Bauer, H. Ditt, and H. Bischof, "SIFT and shape context for feature-based nonlinear registration of thoracic CT images," in *Proc. ECCV-CVAMIA Workshop*, 2006, pp. 73–84.

[45] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1153–1190, 2013.

[46] C. Wachinger, M. Toews, G. Langs, W. Wells, and P. Golland, "Keypoint transfer segmentation," in *Inf. Process. in Med. Imaging*, vol. 9123, 2015, p. 233.

[47] R. Kéchichian, S. Valette, M. Desvignes, and R. Prost, "Shortest-path constraints for 3D multi-object semi-automatic segmentation via clustering and graph cut," *IEEE Trans. Image Process.*, vol. 22, no. 11, pp. 4224–4236, 2013.

[48] R. Kéchichian, S. Valette, M. Sdika, and M. Desvignes, "Automatic 3D multiorgan segmentation via clustering and graph cut using spatial relations and hierarchically-registered atlases," in *Proc. MICCAI-MCV Workshop*, 2014, pp. 201–209.

[49] R. Kéchichian, S. Valette, and M. Desvignes, "Automatic multiorgan segmentation using hierarchically registered probabilistic atlases," in *Cloud-Based Benchmarking of Medical Image Analysis*. Springer, 2017, pp. 185–201.

[50] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[51] C. Wang and O. Smedby, "Automatic multi-organ segmentation using fast model based level set method and hierarchical shape priors," *Proc. ISBI Visceral Challenge*, vol. 1194, pp. 25–31, 2014.

[52] R. Gauriau, R. Cuingnet, D. Lesage, and I. Bloch, "Multi-organ localization with cascaded global-to-local regression and shape prior," *Med. Image Anal.*, vol. 23, no. 1, pp. 70–83, 2015.

[53] A. Hanbury, H. Müller, G. Langs, M. A. Weber, B. H. Menze, and T. S. Fernandez, "Bringing the algorithms to the data: cloud–based benchmarking for medical image analysis," in *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, 2012, pp. 24–29.

[54] O. Jimenez-del Toro *et al.*, "Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks," *IEEE Trans. Med. Imag.*, vol. 35, no. 11, pp. 2459–2475, 2016.

[55] M. Winterstein, M.-A. Weber, K. Grünberg, B. Menze, and G. Langs. (2013) Data set for first competition. [Online]. Available: http://www.visceral.eu/assets/Uploads/Deliverables/VISCERAL-D2.3.1.pdf

[56] Multi-atlas labeling beyond the cranial vault - workshop and challenge. Accessed: 2017-03-01. [Online]. Available: https://www.synapse.org/#!Synapse:syn3193805/wiki/89480

[57] Visceral Project. Segmentation Leaderboard. Accessed: 2017-10-01. [Online]. Available: http://visceral.eu:8080/register/Leaderboard.xhtml

[58] C. Platero and M. C. Tobar, "A multiatlas segmentation using graph cuts with applications to liver segmentation in CT scans," *Computational and mathematical methods in medicine*, vol. 2014, 2014.

[59] G. Yang, J. Gu, Y. Chen, W. Liu, L. Tang, H. Shu, and C. Toumoulin, "Automatic kidney segmentation in CT images based on multi-atlas image registration," in *Proc. IEEE EMBC conf.*, 2014, pp. 5538–5541.

[60] Z. Xu, B. Li, S. Panda, A. J. Asman, K. L. Merkle, P. L. Shanahan, R. G. Abramson, and B. A. Landman, "Shape-constrained multi-atlas segmentation of spleen in CT," in *Proc. of SPIE-Int. Soc. for Opt. Eng.*, vol. 9034, 2014, p. 903446.

[61] U. Vovk, F. Pernus, and B. Likar, "A review of methods for correction of intensity inhomogeneity in MRI," *IEEE Trans. Med. Imag.*, vol. 26, no. 3, pp. 405–421, 2007.

[62] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," *Computer Vision-ECCV 2002*, pp. 128–142, 2002.

[63] L. Zheng, S. Duffner, K. Idrissi, C. Garcia, and A. Baskurt, "Pairwise identity verification via linear concentrative metric learning," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 324–335, 2018.

[64] R. Agier, S. Valette, L. Fanton, P. Croisille, and R. Prost, "Hubless 3D medical image bundle registration," in *Proc. VISAPP Conf.*, vol. 3, 2016, pp. 265–272.