Cloud–based Evaluation of Anatomical Structure Segmentation and Landmark Detection Algorithms: VISCERAL Anatomy Benchmarks

Oscar Alfonso Jiménez–del–Toro[1,2], Henning Müller[1,2], Markus Krenn[3], Katharina Gruenberg[4], Abdel Aziz Taha[5], Marianne Winterstein[4], Ivan Eggel[1], Antonio Foncubierta–Rodríguez[6], Orcun Goksel[6], András Jakab[3], Georgios Kontokotsios[5], Georg Langs[3], Bjoern H. Menze[6], Tomàs Salas Fernandez[7], Roger Schaer[1], Anna Walleyo[4], Marc–André Weber[4], Yashin Dicente Cid[1,2], Tobias Gass[6], Mattias Heinrich[8], Fucang Jia[9], Fredrik Kahl[10], Razmig Kechichian[11], Dominic Mai[12], Assaf B. Spanier[13], Graham Vincent[14], Chunliang Wang[15], Daniel Wyeth[16], Allan Hanbury[5]

Abstract—**Variations in the shape and appearance of anatomical structures in medical images are often relevant radiological signs of disease. Automatic tools can help automate parts of this manual process. A cloud–based evaluation framework is presented in this paper including results of benchmarking current state–of–the–art medical imaging algorithms for anatomical structure segmentation and landmark detection: the VISCERAL Anatomy benchmarks. The algorithms are implemented in virtual machines in the cloud where participants can only access the training data and can be run privately by the benchmark administrators to objectively compare their performance in an unseen common test set. Overall, 120 computed tomography and magnetic resonance patient volumes were manually annotated to create a standard Gold Corpus containing a total of 1295 structures and 1760 landmarks. Ten participants contributed with automatic algorithms for the organ segmentation task, and three for the landmark localization task. Different algorithms obtained the best scores in the four available imaging modalities and for subsets of anatomical structures. The annotation framework, resulting data set, evaluation setup, results and performance analysis from the three VISCERAL Anatomy benchmarks are presented in this article. Both the VISCERAL data set and Silver Corpus generated with the fusion of the participant algorithms on a larger set of non–manually–annotated medical images are available to the research community.**

Index Terms—**Evaluation framework, organ segmentation, landmark detection.**

I. Introduction

Multiple anatomical structures are visually analyzed in medical images as part of the daily work of radiologists. Subtle variations in size, shape or appearance can be used as relevant radiological signs to confirm or discard a particular diagnosis. In the current clinical environment, clinical experts screen through large regions in the full imaging data to detect and interpret these findings. However, manual measurements and personal experience may result in intra– and inter–operator variability when interpreting the images, particularly in difficult or inconclusive cases [1], [2]. Furthermore, the amount of clinical data that have to be analyzed have increased considerably in size and complexity during the past years [3].

Computer aided radiology has proven helpful in facilitating the time consuming and demanding task of handling this large amount of data [4]. Through Computer Aided Diagnosis (CAD) algorithms, multiple organs can be objectively measured and evaluated for robust and repeatable quantification [5]. There are multiple algorithms that have shown promising results in the segmentation and automated identification of different anatomical structures, which is a first necessary step towards CAD. A comprehensive review of different organ segmentation techniques can be found in [6]–[8].

To train and objectively test such systems for diagnostic aid, manually annotated data sets are required. Currently, a first step for annotating data in radiology images is the localization and manual segmentation of the various structures in the images. Performing manual segmentation demands an intensive and time–consuming labour from the radiologists and is subject to variations [9]. Therefore, a frequent bottleneck in the evaluation of segmentation methods is the lack of a common large data set where different algorithms can be tested and compared [10]. This benchmark exercise is fundamental in determining the optimal solution for practical tasks that can then be implemented in a clinical environment, helping to build a comparative analysis of the prevailing state–of–the–art methods [11]. It is still common practice for solutions published in the scientific literature to be evaluated on non–

[1] University of Applied Sciences Western Switzerland, Sierre (HES–SO), Switzerland,
[2] University Hospital and University of Geneva, Switzerland,
[3] Medical University of Vienna, Austria,
[4] University of Heidelberg, Germany,
[5] Vienna University of Technology, Austria,
[6] Swiss Federal Institute of Technology (ETH) Zurich, Switzerland,
[7] Catalan Agency for Health Information, Assessment and Quality, Spain,
[8] University of Lübeck, Germany,
[9] Shenzhen Intitutes of Advanced Technology, Chinese Academy of Sciences, China,
[10] Chalmers University of Technology, Sweden,
[11] University of Lyon, France,
[12] University of Freiburg, Germany,
[13] The Hebrew University of Jerusalem, Israel,
[14] Imorphics, United Kingdom,
[15] KTH–Royal Institute of Technology, Sweden
[16] Toshiba Medical Visualization Systems Europe, United Kingdom,

public data sets. Problematic aspects of this type of evaluation include the use of unsuitable datasets and comparison to poor baselines, leading to "improvements that don't add up" [12], an "illusion of progress" [13] and a lack of reproducibility of the results.

A. Medical data challenges

In recent years, performing public contests with shared data sets and a well-defined task has found widespread use in different fields of research, including medical imaging [1], often involving academic groups as well as companies [2]. Some of the previous challenges concerning the annotation of medical data have focused on:

Anatomical structure segmentation:

- Brain anatomical structures [14] and tumors: MR imaging (MRI) [15]
- Head and neck structures: MRI [3]
- Heart anatomy [4] and motion tracking [16]: MRI and ultrasound (US)
- Airway path [17], lung vessels [18] and lung nodules [19]: CT and CTce
- Prostate and surrounding structures [20]: MRI
- Spine and vertebrae [5]
- Individual abdominal organs (liver [21] [6], pancreas [7]: CT and CTce

Landmark detection

- Head [8]: X–ray
- Lung [22]: CT

However, the final evaluations of these challenges are usually performed providing both the training and testing set to the participants, either in advance [17] or during live competitions [15]. The administrators rely on the participants to not train their algorithms with the test set and to not introduce bias, intended or unintended, in their evaluations [11]. In addition, participant groups can also gain additional advantage in the competitions depending on their lab computation resources, potentially masking limitations when compared to other algorithms.

On the other hand, few of these challenges have addressed multiple structure segmentation [20], targeting single organs instead [21] and, in some cases, in cropped medical images around the region of interests (e.g. abdomen). When clinicians visually inspect medical images searching for radiological signs, the spatial anatomical relations between structures is an important feature. Methods considering multiple structures, when automatically segmenting the anatomy, have shown to improve the segmentation of smaller important structures with higher anatomical variability [23], [24].

B. VISCERAL benchmarks

In the VISual Concept Extraction challenge in RAdioLoGy (VISCERAL[9]) project a cloud–based infrastructure for the evaluation of medical image analysis techniques in Computed Tomography (CT) and Magnetic Resonance (MR) imaging was established. Three benchmarks (Anatomy 1–3) on automated anatomy localization and segmentation of whole–body 3D volumes have been organized. To the best of our knowledge, these are the first benchmarks to evaluate multi–modal medical image analysis techniques using a large amount of data annotated by radiologists. The participant algorithms are installed and executed in identical cloud computing instances and thus fully reproducible. We present a per–anatomy, per–modality evaluation depending on the nature of participating algorithms and the attempted image analysis tasks. The aim of the VISCERAL benchmarks is to create a single, large, and multi–purpose medical image data set and evaluation infrastructure. Through organized benchmarks, research groups can test their specific applications and compare them to other available solutions against the standard manual annotations. This article describes the setup, evaluation metrics, and results of the three VISCERAL Anatomy benchmarks. Main trends in the algorithms and potential future directions of enhancing these segmentation approaches are also discussed herein.

II. VISCERAL EVALUATION FRAMEWORK

A. Cloud infrastructure

Distributing large data sets of terabytes to several participants of a challenge is often not straightforward. Currently, the most common approach is to send the data on hard drives by post or to download the data (training and test set) via online platforms [15]. We developed in VISCERAL a cloud–based infrastructure for the evaluation of medical segmentation algorithms on a large common data set. The scalability of a cloud platform is virtually unlimited in both storage and computation power, enabling the storing of big data sets and subsets with different access permissions. The VISCERAL project was hosted in the Microsoft Azure cloud environment. The Microsoft Azure platform provides a framework for the creation and management of virtual machines (VMs) and data storage containers. Data can be stored centrally complying with privacy requirements for anonymized patient data. In particular, the Azure cloud is HIPAA (Health Information Portability and Accountability Act) certified. Using a shared cloud environment brings the algorithms to the data, avoiding extensive downloads and keeping confidential information access only to algorithms of registered participants and not the

[1] MICCAI Grand Challenges,http://grand-challenge.org/All_Challenges/ *

[2] Kaggle, https://www.kaggle.com/ *

[3] Head and Neck Auto Segmentation Challenge, http://www.imagenglab.com/wiki/mediawiki/index.php?title=2015_MICCAI_Challenge *

[4] Second Annual Data Science Bowl, https://www.kaggle.com/c/second-annual-data-science-bow, *

[5] Computational Methods and Clinical Applications for Spine Imaging, http://csi2015.weebly.com *

[6] Proceedings of SHAPE 2015 Symposium, http://www.shapesymposium.org/proceedings-screen.pdf *

[7] Pancreas Segmentation from 3D Abdominal CT images, http://www.biomedicalimaging.org/2014/program/challenges/ *

[8] Automatic Cephalometric X-Ray Landmark Detection Challenge 2014, http://www-o.ntust.edu.tw/~cweiwang/celph/,*

* as of 1 June 2016

[9] `http://www.visceral.eu/`

participants themselves, avoiding duplication of the confidential data. Participant algorithms can be evaluated independently by the administrators to avoid an unfair exploitation of the test set by participants. The evaluation is therefore more objective, limiting bias in the comparisons. An aspect of the framework that makes it attractive for evaluations on medical data is that the data are stored centrally and are not distributed individually.

Initially, the full data set with both the medical data and additional annotations created by expert radiologists was uploaded to a cloud storage container. Other cloud storage containers were then created in each benchmark to store the training and testing data sets, participant output files and evaluations. Over the course of the project, new images and their annotations were added to the storage containers when required. In order to run the VISCERAL benchmarks, the participants needed access to the stored data and computing instances to execute their algorithms. Virtual machines running on the Microsoft Azure cloud infrastructure were pre–configured to run these tasks. Different templates were configured for 5 operating systems including both Windows and Linux. A virtual machine was provided to each participant, allowing them to access the training data set and upload their algorithms. All the participant VM instances had the same computing specifications and capabilities. Time–restricted read–only access keys were distributed securely to the participants for accessing the training data sets. Participants could remotely access their VMs during the training phase. Moreover, they could install all the tools and libraries needed to run their algorithms. At this stage they could optimize their approaches with the available training set. Specification guidelines were written by the administrators for each benchmark on the usage and permissions applying to the VMs. The platform's web management portal was used for the VISCERAL project to simplify the administrative tasks of handling the VMs. The VISCERAL registration and management system[10], containing all the information needed in the benchmarks, was created (user agreement, specifications, data set lists). Through the participant dashboard in the system, participants received the private access credentials for the their VM and had the option to start it or shut it down during the training phase.

B. Data set

CT and MR scans of the whole body (wb), the whole trunk (CT contrast–enhanced, CTce) or from the abdomen (MR T1 contrast-enhanced, MRT1cefs) were used, in order to have a large variety of anatomical structures and medical imaging modalities in the data set. Having both un-enhanced and enhanced data sets supports the evaluation of segmentation algorithms both on high and low contrast at sufficient resolution for their radiological interpretation. Each modality data set includes a large number of studies that are representative for daily clinical routine work.

Whole body unenhanced imaging in CT (CTwb) was acquired in patients with confirmed bone marrow neoplasms, such as multiple myeloma, in order to detect focal bone lesions (osteolysis). The field of view from these CT scans starts at the head and ends at the knee of the patient. Contrast–enhanced CT scans were acquired from patients with malignant lymphoma. Their field of view starts at about at the corpus mandibulae, i.e. in between the skull base and the neck and ends at the pelvis. These scans were enhanced by an iodine-containing contrast agent that is commonly administered to improve tissue contrast, in order to detect pathological lymph nodes or organ affection of the lymphoma. These studies are usually acquired in patients with multiple myeloma in order to detect affection (either as diffuse infiltration or as (multi–)focal infiltration or both) of the bone marrow and to detect extra osseous involvement, e.g. soft tissue masses. The field of view of these MR scans starts with the head and ends at the feet, as shown in Figure 1. These studies are unenhanced. Nevertheless, most organs can be seen in these MR images. All of these examinations include a coronal T1–weighted and fat-suppressed T2–weighted or STIR (short tau inversion recovery) sequence of the whole body, plus a sagittal T1–weighted and a sagittal T2–weighted sequence of the entire vertebral column. MRI studies of the abdomen, abdomen contrast-enhanced fat-saturated MR T1 (Ab/MRT1cefs), are also included. These images were acquired in oncological patients, who had metastases within the abdomen. The examinations are contrast-enhanced by a gadolinium–chelate. The scans start at the top of the diaphragm and end at the pelvis.

The four imaging modalities had all their data sets taken from the same hospital during clinical practice using the same imaging protocols and the same imaging device for each modality. Its use was subjected to specific regulations according to the Medical Ethics Committee from the hospital where the images were obtained. This committee gave restrictions that controlled the collection, use, distribution of human data and its inclusion in research studies. All work on data collection of humans was conducted under the rules and legislation in place according to the Declaration of Helsinki (Informed consent for participation of human subjects in medical and scientific research, 2004). All the data used in the Anatomy benchmarks was fully anonymized. The radiology reports and meta data were anonymized by removing all patient names, physician names, hospital and institution names and other identifying information. Radiology images were anonymized by blurring face regions but preserving the underlying structure of the face so that it can still be used as reference for image analysis. Any embedded text in the image, and other identifying information such as serial numbers on implants was also removed from the image. The data of the Anatomy benchmarks were available only for non–commercial research and only after participants signed a license agreement that assured the use of the data in its given environment and for its research purpose. The information regarding the format and characteristics of the data set were available to the participants in the project deliverables and benchmark specifications published in the VISCERAL website.

For the creation of the VISCERAL Gold Corpus, considered as the anatomical reference annotation data base for the Anatomy benchmarks, 391 CT and MRI data sets (889 sequences) in total with 20 different organs and 53 landmarks

[10] http://visceral.eu:8080/register/Login.xhtml

TABLE I: Overview of the manually annotated Anatomy Gold Corpus. For each modality the field–of–view is defined as FOV. Both the in–plane resolution range and in–between plane resolution are reported in milimiters. The number of volumes, annotated anatomical structures (Annotations) and located anatomical landmarks are also shown per modality.

Modality	FOV	Contrast	Resolution (mm)	Volumes	Annotations	Landmarks
CT	whole–body	un enhanced	$0.977^2 - 1.405^2 \times 3$	30	384	530
	trunk	contrasted	$0.604^2 - 0.793^2 \times 3$	30	387	440
MR T1w & T2w	whole–body	un enhanced	$1.250^2 \times 5$	30	305	520
	abdomen	contrasted	$0.840^2 - 1.302^2 \times 3 - 8$	30	219	270

were included. Patient scans were disregarded if they were not complete in protocol (i.e. complete T1 and T2 of the whole body for the MRI examination) or had too many artefacts (e.g. due to movement of the patient or breathing artefacts in MRI). For the CTwb, scans with a slice thickness higher than 3 mm were also disregarded. The data set comprises roughly the same number of images from male and female patients (62 male, 69 female); the average patient age is 59.9 years(± 9.79 years standard deviation). A subset of thirty volumes per modality (120 volumes in total) was manually annotated by medical experts for up to twenty anatomical structures of interest. When organs were not visible in a modality they were not annotated and thus for a few organs fewer examples are included in the Gold Corpus. These annotations served as 'ground truth' for the training and testing phases (Table I). Different 3D annotation tools were reviewed to provide efficient and reliable annotations to significantly reduce the amount of time required when compared to slice–by–slice manual annotations. The GeoS annotation tool was dominant in structures with high contrast as semi-automatic, while 3D Slicer was more efficient for small structures with less contrast and weak visual separation from surrounding structures.

A key that ensured an optimal use of the data was the accurate annotation based on detailed written guidelines updated in the course of the project, quality control and choice of annotated examples. A quality control team was created from the VISCERAL consortium with three radiologists and two medical doctors who checked annotations systematically. If annotations did not adhere visually to the project's defined annotation guidelines they were either corrected manually or send back again for re–annotation.

1) Annotated anatomical structures: A representative selection of major and minor structures that can be detected in a large set of CT or MRI examinations is included in the data set. The selection includes 20 structures of 15 organs: left/right kidney, spleen, liver, left/right lung, urinary bladder, rectus abdominis muscle, 1^{st} lumbar vertebra, pancreas, left/right psoas major muscle, gallbladder, sternum, aorta, trachea, left/right adrenal gland. Not all structures can be located in MR images due to the lower resolution compared to CT, a lack of contrast for the skeletal structures, and due to partial volume artifacts that occur because of relatively thicker slices. Breathing, pulsation, and other motion artifacts are also common, making some smaller organs particularly difficult to delineate accurately. If a structure could not be detected or annotated with sufficient certainty, it was not segmented. Volume annotations were expressed numerically by assigning each voxel a binary value (0,1), where 1 corresponds to the annotated structure.

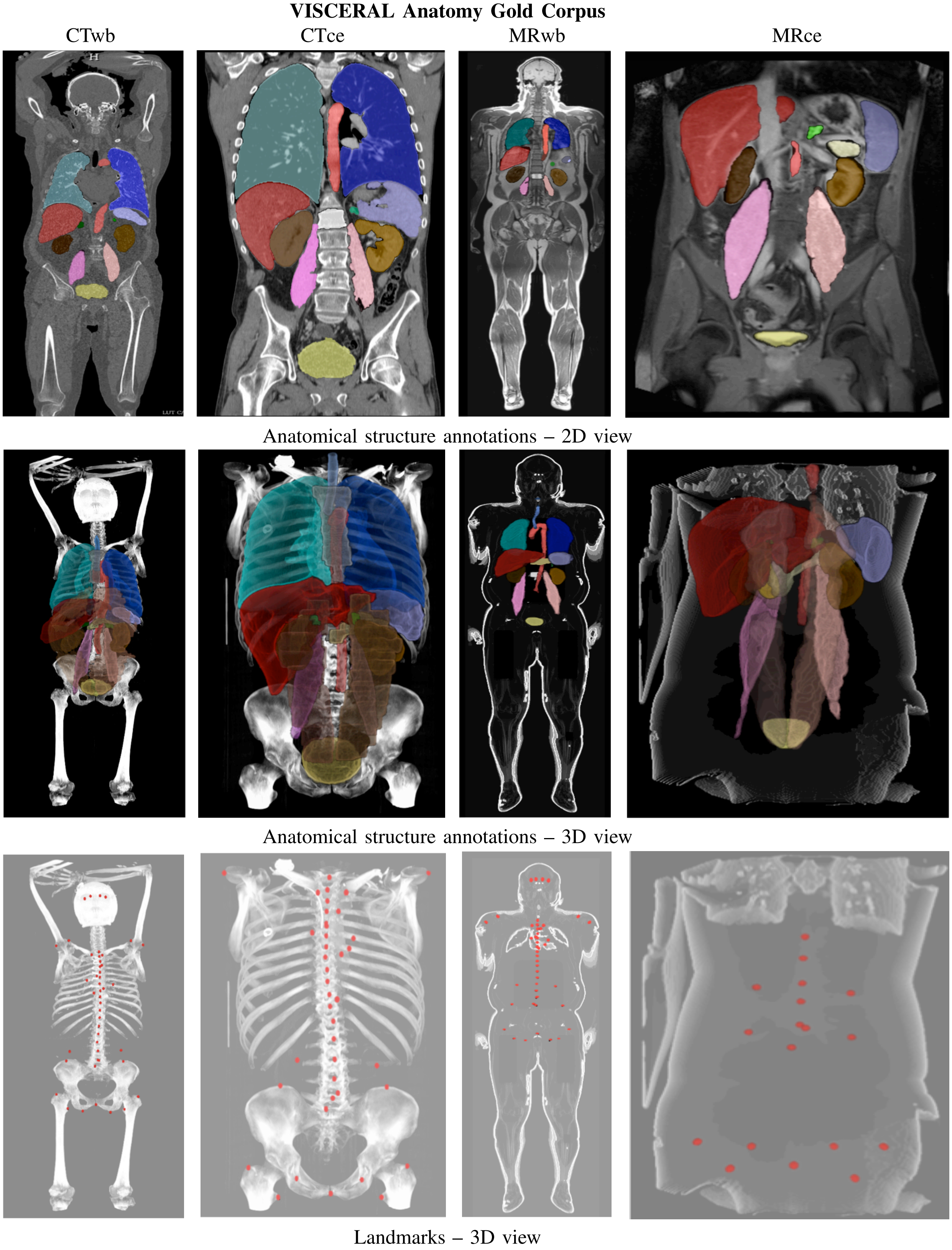
2) Landmark localization: Anatomical landmarks are the locations of selected anatomical structures that can be identified in different image sequences. Their universal nature makes them important, e.g. as a first step in parsing image content or for triangulating other more specific anatomical structures. Being invariant to the field of view, they are of particular importance for image retrieval tasks. Landmarks are stored in text–based comma–separated CSV files with each column holding an ID that identifies a specific landmark together with the coordinates of that landmark.

C. Anatomy benchmark setup

1) Anatomy1: A clear split of training and test images was used. Only the training images were seen by the participants who prepared an executable in a given format that was then used by the organizers to run the trained algorithms on the test data. Participants registered in the VISCERAL registration system uploading a signed agreement on data usage. The registered participants had access to a VM and a training set of 28 annotated scans (7 per modality) with their corresponding annotated structures within a cloud storage container. All the images and annotations were available as individual anonymized files in NIfTI (Neuroimaging Informatics Technology Initiative) format without any additional cropping or pre–processing from their raw DICOM format. Participants could then implement and train their algorithms in the cloud computing instances with 4–core CPU and 8GB RAM. At the deadline of the Anatomy 1 benchmark, these VMs were submitted and participants had no longer access to their VM. The algorithm executables in the VMs were run automatically by the administrators on a test set of 51 manually annotated patient scans (27 CT, 24 MR).

2) Anatomy2: In Anatomy2, the size of the training set was increased to 20 volumes per modality with their corresponding organ annotations. The computation power of the participants VMs was also doubled from 4 to 8 core CPU with 16GB of RAM.

3) Anatomy3 continuous evaluation: For Anatomy3, a continuous evaluation system was implemented where participants could submit their algorithms iteratively, at most once a week. The Benchmark is currently still running and results on the test set can be obtained interactively at any time beyond the end of the project. A public leaderboard was launched on the VISCERAL website where participants may choose to make their (best) results public. A snapshot of the results



Anatomical structure annotations – 2D view

Anatomical structure annotations – 3D view

Landmarks – 3D view

Fig. 1: Examples of patient volumes in the VISCERAL Anatomy Gold Corpus with their corresponding anatomical structures and landmarks. A 2D coronal section from each of the four modalities is presented in the first row. Annotated structures have been overlaid in different colors on top of the original images. In the second row, the structures are shown in 3D with the bone structure (not manually annotated), in CT images, and the body contour (not manually annotated), in MR images, added for spatial reference. In the final row, 3D views of the landmarks (red dots) present in each volume are shown. Bone structure and body contour are also shown in the background for spatial reference.

was taken to be presented during the ISBI 2015 workshop. The Anatomy3 continuous evaluation benchmark and public Leaderboard are currently open [11] as of 1 June 2016.

D. Evaluation metrics

The output files from the participant algorithms were evaluated with an efficient evaluation tool implemented for the VISCERAL project using a consistent set of metrics [25]. The algorithms used to calculate the metrics were selected and optimized to achieve high efficiency in speed and memory necessary to meet the challenging requirements of evaluating volumes with large grid sizes. For the landmark localization task, the Euclidean distance and percentage contribution of landmarks for each method were computed. For the organ segmentation task, binary and fuzzy segmentations using 20 evaluation metrics were compared. A more detailed analysis of the selected metrics is presented in [25]. These metrics were categorized based on their nature and the equivalence between some of them to help find a reasonable combination when more than one metric is to be considered. For brevity, only the DICE coefficient and average Hausdorff distance are shown for the results of the benchmarks. The complete results for the three benchmarks with all the evaluation metrics are available on the VISCERAL website.

The Dice coefficient [26] (DICE), also called the overlap index, is the most frequently used metric in validating medical volume segmentations. $DICE$ is defined by

$$DICE = \frac{2.|S_g^1 \cap S_t^1|}{|S_g^1| + |S_t^1|} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (1)$$

Spatial distance based metrics are widely used in the evaluation of image segmentation as dissimilarity measures. They are recommended when the overall segmentation accuracy, e.g. the boundary delineation (contour), of the segmentation is of importance [27].

The Average Distance, or the Average Hausdorff Distance (AVD), is the Hausdorff distance (HD) averaged over all points. The AVD is known to be stable and less sensitive to outliers than the HD. It is defined by

$$AVD(A,B) = max(d(A,B), d(B,A)) \quad (2)$$

where $d(A,B)$ is the directed Average Hausdorff distance that is given by

$$d(A,B) = \frac{1}{N} \sum_{a \in A} \min_{b \in B} ||a - b|| \quad (3)$$

III. Anatomy benchmarks

For the three Anatomy benchmarks including the ISBI 2014 [28] and ISBI 2015 [29] Anatomy challenges, there were 164 participants registered in the VISCERAL registration system. The dataset was accessed by 61 participants who signed the license agreement and were provided with virtual machines. There were 22 algorithms submitted by 12 research groups for the segmentation and landmark detection tasks. Participants did not need to segment all the structures provided, but could attempt segmenting any single anatomical structure or a set of them.

The number of volumes included in the training set and test set of Anatomy1 changed in Anatomy2 and Anatomy3. The manually annotated data set was extended in the latter two benchmarks and thus the results from Anatomy2 and 3 are the main focus of the analysis in this paper. Since Anatomy3 is open for submissions at the time of writing this paper, we discuss the results only from algorithms that are currently (early 2016) published in the online Leaderboard.

A short description of the submitted participant algorithms can be found in Table II. Further information on the participant methods can be found in the Appendix available in the supplementary materials and the cited publications.

IV. Results

Altogether, 518 evaluation runs were performed on the VISCERAL Anatomy test set. Each run corresponds to the anatomical segmentations of 10 volumes per modality, computed by a participant algorithm on an unseen test set. The modality with most submissions was CTce with 245 runs and MRT1cefs was the one with least, at 44 runs. The most frequently evaluated anatomical structure in the four modalities was the liver with a total of 43 runs. The one with the fewest submissions was the rectus abdominis muscle with 8 runs each side (left and right). In this section, we first present the quantitative results from the segmentation tasks (for CT and MR data) in the Anatomy benchmarks. Inter–annotator agreement (InA) scores obtained during the manual annotation process by medical experts are shown together with the participant results. A qualitative evaluation performed by medical experts on a subset of the output segmentations from the participant algorithms is then addressed. Finally, the landmark detection task results are shown.

A. Anatomical segmentation task

*1) **CT segmentation**: Anatomy1 benchmark.* Six algorithms participated in the Anatomy1 CT segmentation task: 6 in CTce, 2 in CTwb [12]. Two methods (Ga1 [38] and Ke1 [43]) segmented all the structures available in CTce. The method with the highest number of top results was Ga1 [38]. However, the structures with highest participation (liver, lungs, kidneys) in CTce were better segmented by Wa1 [35], Ji1 [40] and Sp1 [31]. Regarding CTwb, the results of Wa1 [35] are higher when compared to those obtained by Ga1 [38], although the latter was implemented for more anatomical structures. The best DICE overlap scores between the same structures are similar for both modalities (CTwb and CTce) with the most significant differences seen in the first lumbar vertebra (lVert1) and gallbladder with lower DICE scores in CTwb (see Table. III). *Anatomy2 and 3 benchmarks.* Thirteen algorithms contributed with at least one structure to the Anatomy2–3

[11] Anatomy3 public Leaderboard, http://visceral.eu:8080/register/Leaderboard.xhtml

[12] http://www.visceral.eu/closed-benchmarks/benchmark-1/benchmark-1-results/, as of 1 June 2016

TABLE II: Overview of the participant algorithms from the Anatomy benchmarks 1–3. A detailed description of their implementation can be found in the Appendix and in the VISCERAL ISBI 2014 [28] and ISBI 2015 workshop proceedings [29]. The participant segmentation algorithms are organized according to their segmentation approach. The total number of organs included per modality (Organs) in the final Gold Corpus test set was: CTwb and CTce (20), MRwb (17), MRce (15). * The testing runtime is shown per patient volume.

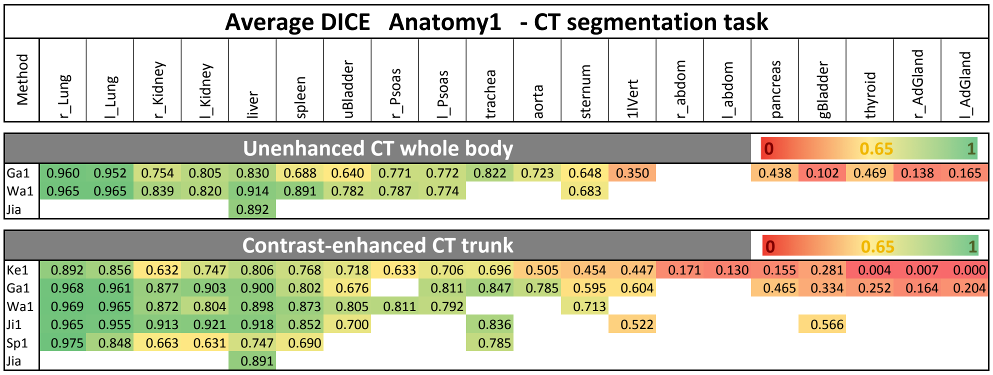
VISCERAL Anatomy benchmarks organ segmentation

Abbrev	Method	Description	Organs	CT	MR	A1	A2	A3	Runtime*
Intensity–based clustering									
Dic	Dicente et al. [30]	K–means clustering and geometric techniques	2	wb,ce	–	–	–	✓	8m
Rule–based									
Sp1,Sp2	Spanier et al. [31]	Rule–based segmentation w/region growing	7	ce	–	✓	✓	–	3h
Shape and appearance models									
Jia,Li,He	Jia et al. [32], [33]	Multi–boost learning and SSM search	6	wb,ce	–	✓	✓	✓	25m
Vin	Vincent [34]	Active appearance models	8	wb,ce	–	-	✓	–	1h45m
Wa1,Wa2,Wa3	Wang et al. [35], [36]	Model based level–set and hierarchical shape priors	10	wb,ce	–	✓	✓	✓	1h
Multi–atlas registration									
Ga1,Ga2	Gass et al. [37], [38]	Multi–atlas registration via Markov Random Field	18	wb,ce	wb,ce	✓	✓	–	4h30m
Hei	Heinrich et al. [39]	Multi–atlas seg. w/discrete optimisation and self–similarities	7	ce	ce	–	–	✓	40m
Ji1,Ji2	Jiménez et al. [40], [41]	Multi–atlas registration, anatomical spatial correlations	20	wb,ce	–	✓	✓	–	12h
Kah	Kahl et al. [42]	RANSAC registration, random forest classifier, graph cut	20	wb	–	–	–	✓	13h
Ke1,Ke2,Ke3	Kéchichian et al. [43]	Atlas registration, clustering, graph cut w/spatial relations	20	ce	–	✓	✓	✓	2h

VISCERAL Anatomy benchmarks landmark detection

-	Gass et al. [37], [38]	Template based approach	–	wb,ce	wb,ce	✓	✓	NA	30m
-	Mai et al. [44]	Histogram of Gradients for landmark detection	–	wb,ce	wb,ce	–	✓	NA	7m
-	Wyeth et al. [45]	Classification forests trained at voxel–level	–	wb	–	✓	–	NA	2m

TABLE III: Tables showing the average Dice results from Anatomy1 in CTce and CTwb. The scores are colored according to the reference range shown on the top left corner of the tables. Ga1= Gass et al., Jia= Jia et al., Ji1= Jiménez et al., Ke1= Kéchichian et al., Sp1= Spanier et al., Wa1= Wang et al.



Average DICE Anatomy1 - CT segmentation task

Method	r_Lung	l_Lung	r_Kidney	l_Kidney	liver	spleen	uBladder	r_Psoas	l_Psoas	trachea	aorta	sternum	1lVert	r_abdom	l_abdom	pancreas	gBladder	thyroid	r_AdGland	l_AdGland
Unenhanced CT whole body (0 – 0.65 – 1)																				
Ga1	0.960	0.952	0.754	0.805	0.830	0.688	0.640	0.771	0.772	0.822	0.723	0.648	0.350			0.438	0.102	0.469	0.138	0.165
Wa1	0.965	0.965	0.839	0.820	0.914	0.891	0.782	0.787	0.774			0.683								
Jia					0.892															
Contrast-enhanced CT trunk (0 – 0.65 – 1)																				
Ke1	0.892	0.856	0.632	0.747	0.806	0.768	0.718	0.633	0.706	0.696	0.505	0.454	0.447	0.171	0.130	0.155	0.281	0.004	0.007	0.000
Ga1	0.968	0.961	0.877	0.903	0.900	0.802	0.676		0.811	0.847	0.785	0.595	0.604			0.465	0.334	0.252	0.164	0.204
Wa1	0.969	0.965	0.872	0.804	0.898	0.873	0.805	0.811	0.792			0.713								
Ji1	0.965	0.955	0.913	0.921	0.918	0.852	0.700			0.836			0.522				0.566			
Sp1	0.975	0.848	0.663	0.631	0.747	0.690				0.785										
Jia					0.891															

benchmarks for CT, with all the anatomical structures having at least two methods to compare [11] [13]. In CTwb, the algorithm by Kah [42] segmented the largest number of structures (12) with the highest DICE overlap scores. It was followed by Ji2 [40] with 6 structures with the top DICE scores, particularly for those of a smaller size (thyroid, adrenal glands) but with worse overlap and average distance errors. The best score for CTwb liver was obtained by Wa3 [35] (DICE 0.936, avgdist 0.19). Eleven CTwb structures had a DICE overlap >0.8, with the best overlap scores obtained for the lung (DICE 0.975) and the worst for the gallbladder (DICE 0.276).

There were four methods with multiple top ranking positions in CTce: Wa3 [35], Vin [34], Ke3 [43] and Ji2 [40]. For the structures with most algorithm submissions (lungs, liver, kidneys and spleen) the overlap scores were relatively close between the different approaches, with a small advantage for the algorithm by Wa3 [35](Anat3) or by Vin [34]. The highest overlap was obtained in lungs (DICE 0.974) and the lowest in the adrenal glands (DICE 0.331). For structures where the highest DICE overlap scores were smaller than 0.75 (8 out of 20), the AVD was higher than 1 voxel (see Table.V).

2) ***MR segmentation:*** The algorithm by Ga1 and Ga2 [38] was the only one that generated segmentations for both MR modalities. Hei [39] contributed with 7 organs segmented in MRce. Only five structures (right lung, liver, left psoas muscle, and both kidneys) out of 18 obtained an overlap >0.8 in MRT1wb. Only the average distance metric from the spleen, left psoas and aorta in MRT1wb, and the left psoas in MRT1cefs, were smaller than those of the inter–annotator agreement (see Figure 4, and 5). The correlation was high between DICE and AVD with the extreme cases being the gallbladder and the sternum with an overlap of 0 and avgdist>200 (see Table IV and Table V).

[13] http://www.visceral.eu/closed-benchmarks/anatomy2/anatomy2-results/, as of 1 June 2016

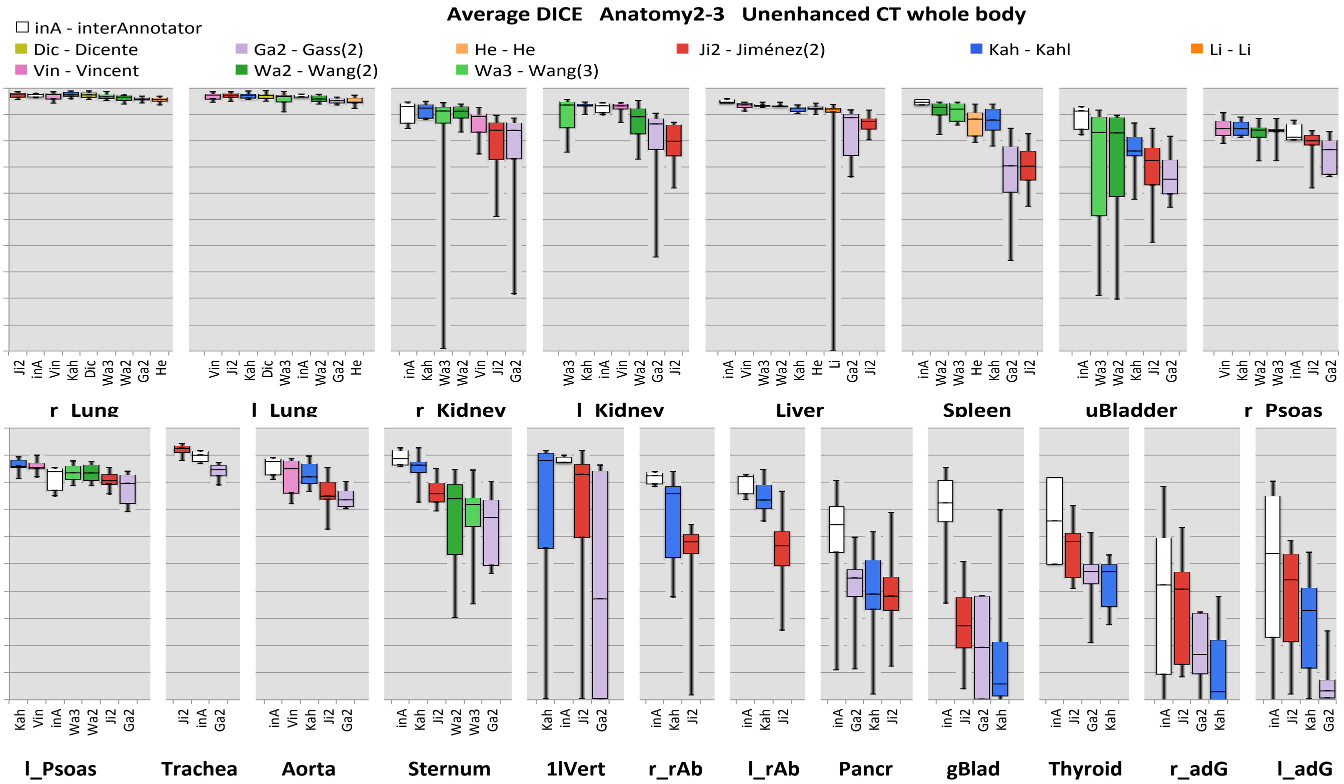


Fig. 2: Anatomy2–3 Boxplot chart of the Dice scores in CTwb from the participants and inter–annotator agreement. The scores are organized according to the median Dice obtained (horizontal black bar inside the box). The quartile ranges (Q1,Q3) of the scores on the final Anatomy Gold Corpus test set are outlined below and above the median. The participant algorithms color code and name abbreviation are shown on top. Additional evaluation metrics can be found on the the Anatomy Leaderboard [11].

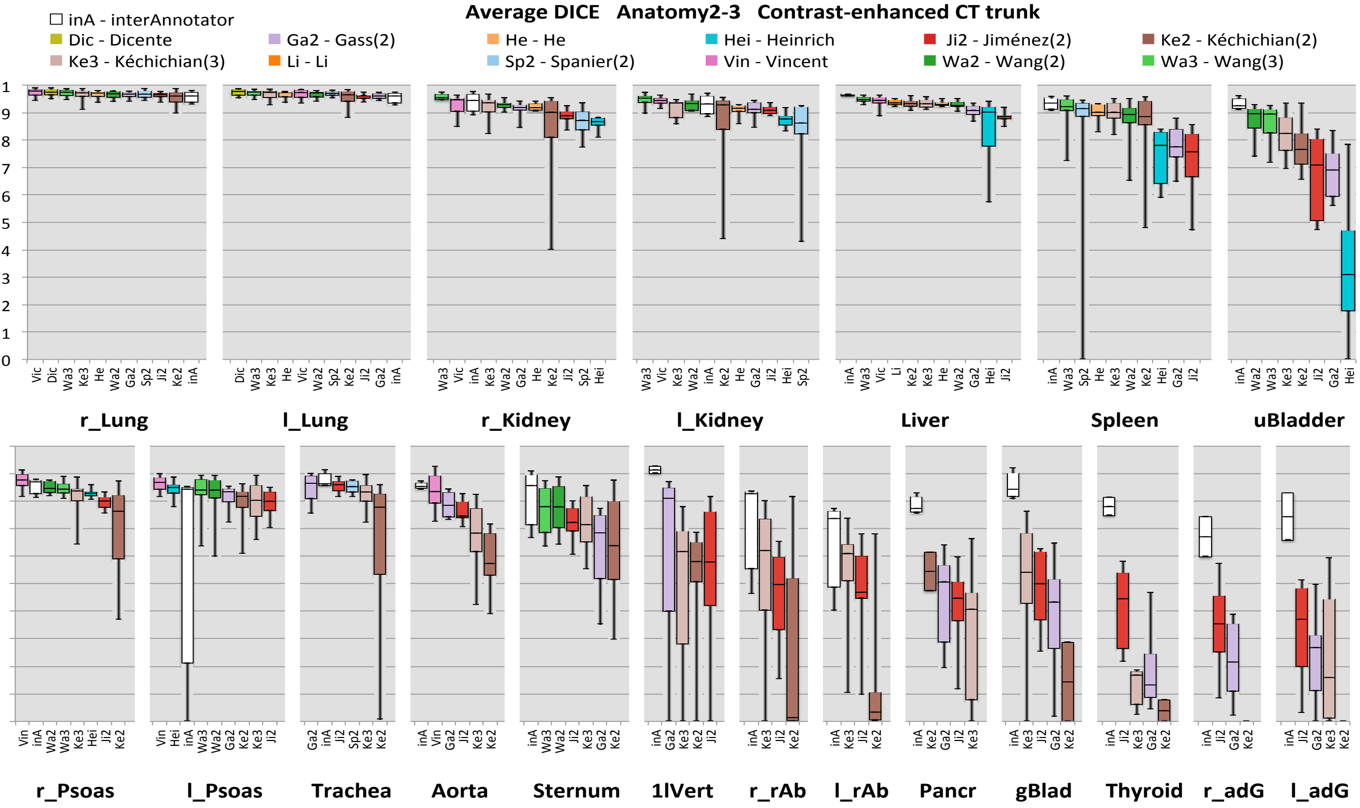


Fig. 3: Anatomy2–3 Boxplot chart of the Dice scores in CTce from the participants and inter–annotator agreement. The scores are organized according to the median Dice obtained (horizontal black bar inside the box). The quartile ranges (Q1,Q3) of the scores on the final Anatomy Gold Corpus test set are outlined below and above the median. Additional evaluation metrics can be found on the the Anatomy Leaderboard [11].

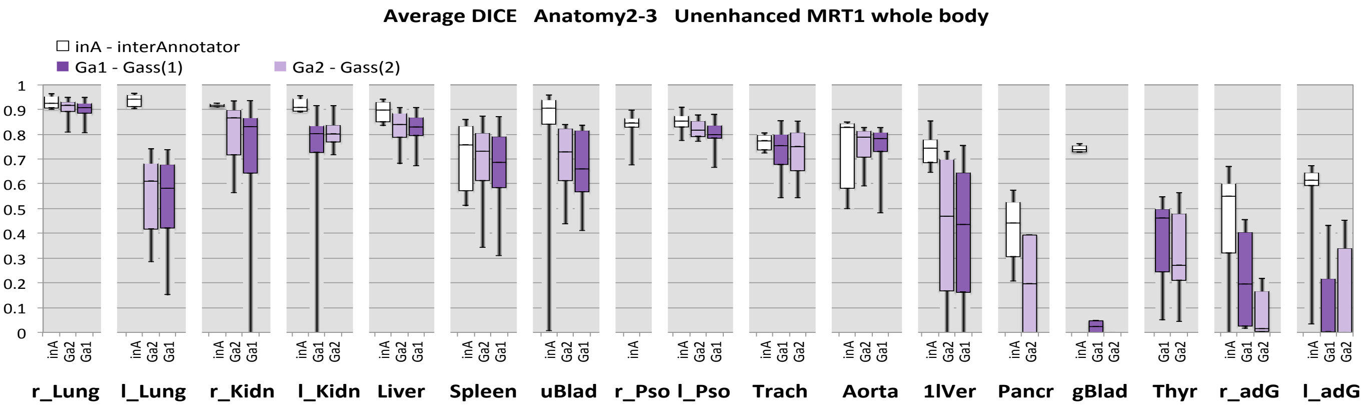


Fig. 4: Anatomy2–3 Boxplot chart of the Dice scores in MRwb from the participants and inter–annotator agreement. The scores are organized according to the median Dice obtained (horizontal black bar inside the box). The quartile ranges (Q1,Q3) of the scores on the final Anatomy Gold Corpus test set are outlined below and above the median. The participant algorithm color code and name abbreviation are shown on top. Additional evaluation metrics can be found on the the Anatomy Leaderboard [11]

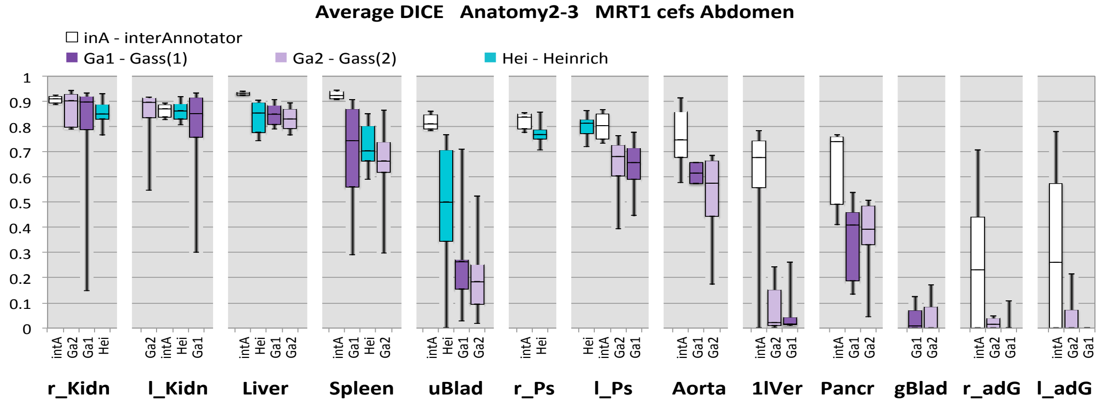


Fig. 5: Anatomy2–3 Boxplot chart of the Dice scores in MRce from the participants and inter–annotator agreement. The scores are organized according to the median Dice obtained (horizontal black bar inside the box). The quartile ranges (Q1,Q3) of the scores on the final Anatomy Gold Corpus test set are outlined below and above the median. The participant algorithms color code and name abbreviation are shown on top. Additional evaluation metrics can be found on the the Anatomy Leaderboard [11]

B. *Qualitative Evaluation*

Selecting a suitable metric to assess the accuracy and the quality of the segmentation algorithms is not a trivial task. Since manual rankings provide a reference for judging metrics and evaluation methods, two radiologists independently ranked the output segmentations from six organs in a double blind fashion, by visually inspecting a subset of the output segmentations from the participant algorithms. A total of 483 output segmentations from 110 Gold Corpus structures in CTwb and CTce were visually inspected and manually ranked according to a point–based system (score 1–5) defined through a medical interpretation of the results. Severe deviation to other organs, crossing of an organ border, missing parts or optimal segmentation were included in the ranking criteria. Rankings were considered per segmentation, which allowed for multiple segmentations potentially having the same score. The top five algorithms from the Anatomy2 benchmark with the best dice overlap scores for left lung, liver, right kidney, urinary bladder, aorta and pancreas were evaluated. These organs were selected as a representation of various organ shapes and sizes available in the VISCERAL data set. Pearson's correlation between the two manual rankings was 0.62, which revealed a moderate inter–rater correlation with significant discrepancies between the rankers. At system level, when all output segmentations are considered for the same organ for each algorithm, Pearson's correlation was 0.81 for the DICE metric when compared to manual ranking by the first rater. This was, together with five other metrics, the highest correlation among the 20 evaluated metrics, therefore indicating the suitability of DICE representing the preference of expert radiologists.

Qualitative segmentation results are shown in Fig. 6 and Fig. 7. The sections and outlined segmentations show regions of conflict between the different participating algorithms and highlight the corresponding manually annotated ground truth.

C. *Landmark detection task*

This task was present only in the Anatomy1 and Anatomy2 benchmarks, with a much larger number of landmark locations (12 vs. 53) evaluated in the latter test set. Three algorithms participated with their results shown in Table VI. The landmarks that had the highest mean Euclidean distance errors

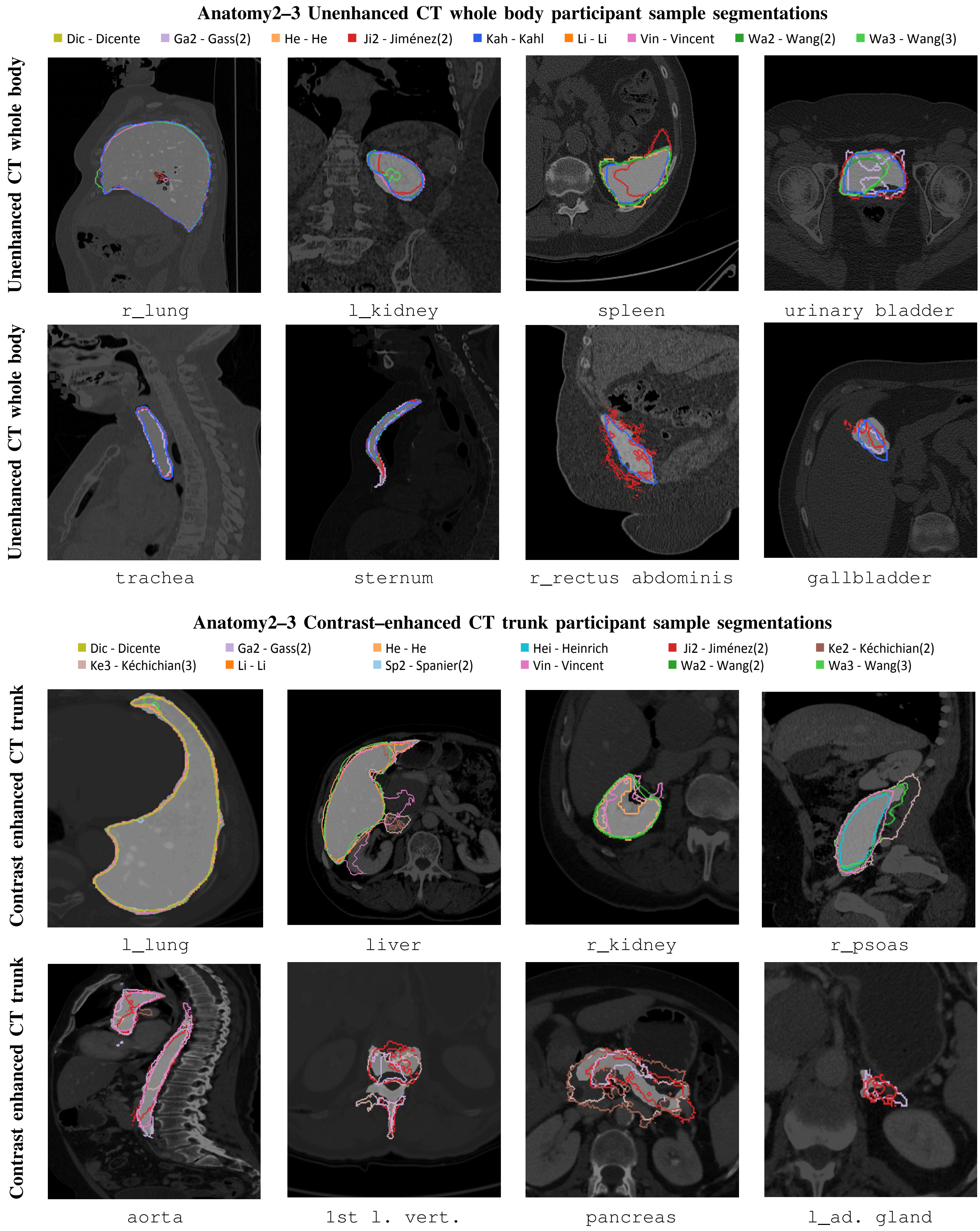


Fig. 6: Sample Anatomy2–3 CT output segmentations in the Gold Corpus test set. Only the top five algorithms with the best DICE scores are shown per structure (name below). The ground truth is highlighted in white from the volume sections, while the rest is darkened. The segmentation contours are color coded and overlaid according to the mean Dice score for the corresponding structure. Views vary between patients in order to show areas of conflict between the algorithm segmentations.

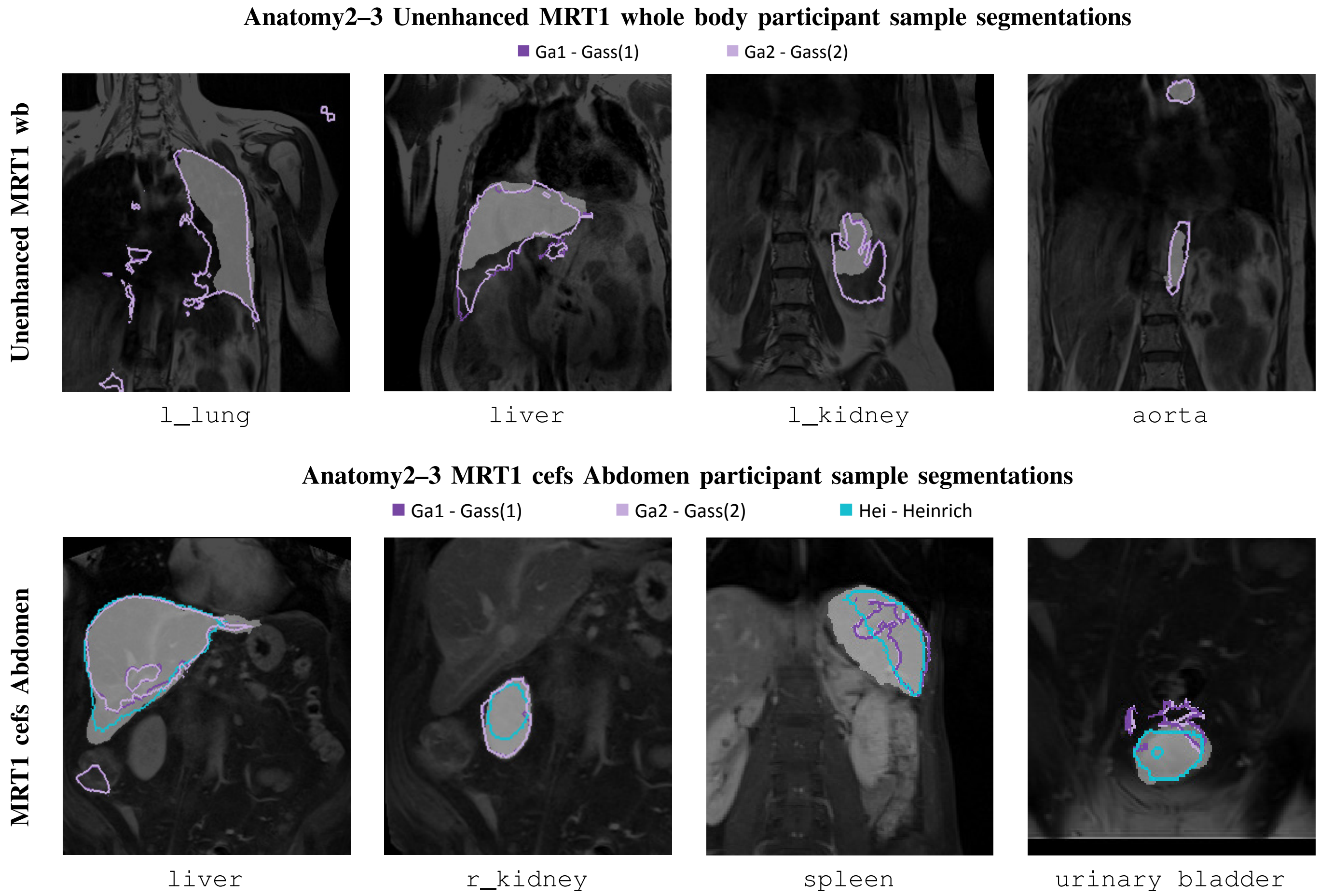
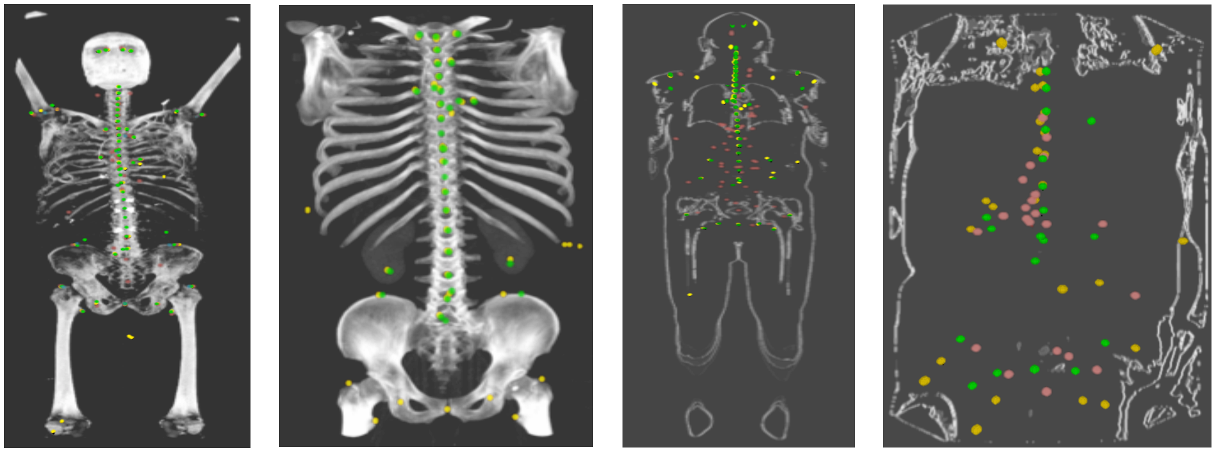


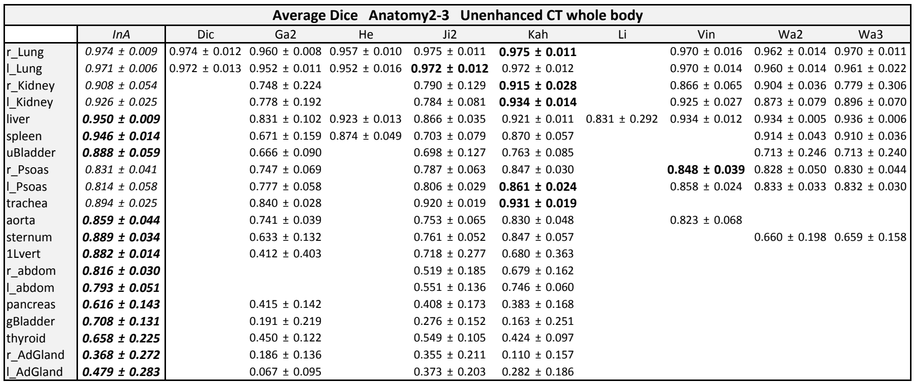
Fig. 7: Sample Anatomy1–3 MR output segmentations in the Gold Corpus test set. All participant algorithms are shown per structure (name below). The ground truth is highlighted in white from the volume sections, while the rest is darkened. The segmentation contours are color coded and overlaid according to the mean Dice score for the corresponding structure. Views vary between patients in order to show areas of conflict between the algorithm segmentations.



(a) **CT wb** (b) **CT ce** (c) **MR wb** (d) **MR ce**

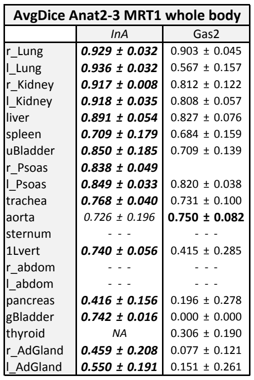
Fig. 8: Sample landmark localization results in the Gold Corpus test set. A 3D volume is shown per modality with the ground truth landmarks displayed as green dots. The output from participant Mai et al. are shown as yellow dots, Gass et al. landmarks are red dots and Wyeth et al. results (for a smaller set of landmarks from Anatomy1 CTwb), are displayed as blue dots. The bone structure and anatomical contours are shown in the background for spatial reference.

TABLE IV: Average Dice results from Anatomy2–3 in the four available modalities: CTwb, MRwb, CTce, MRce. The participants are listed with the name abbreviation used in Fig. 2, 3, 4 and 5. The inter–anotator agreement scores (InA) obtained during the manual annotation phase are also shown. Anatomical structures are listed according to the obtained overlap, from first to last. The highest score is highlighted in bold font. NA indicates annotators declaring poor visibility of the structure. Anatomical structures outside the field–of–view of the 3D MR volumes were marked as empty (-).



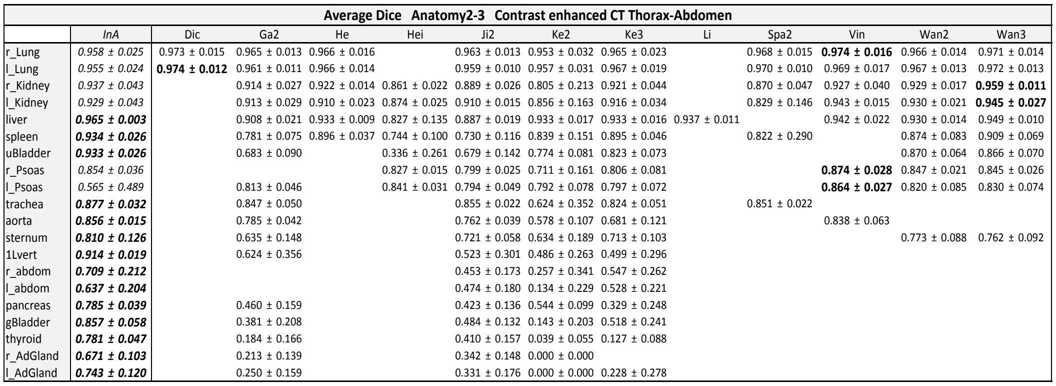
Average Dice   Anatomy2-3   Unenhanced CT whole body

	InA	Dic	Ga2	He	Ji2	Kah	Li	Vin	Wa2	Wa3
r_Lung	*0.974 ± 0.009*	0.974 ± 0.012	0.960 ± 0.008	0.957 ± 0.010	0.975 ± 0.011	**0.975 ± 0.011**		0.970 ± 0.016	0.962 ± 0.014	0.970 ± 0.011
l_Lung	*0.971 ± 0.006*	0.972 ± 0.013	0.952 ± 0.011	0.952 ± 0.016	**0.972 ± 0.012**	0.972 ± 0.012		0.970 ± 0.014	0.960 ± 0.014	0.961 ± 0.022
r_Kidney	*0.908 ± 0.054*		0.748 ± 0.224		0.790 ± 0.129	**0.915 ± 0.028**		0.866 ± 0.065	0.904 ± 0.036	0.779 ± 0.306
l_Kidney	*0.926 ± 0.025*		0.778 ± 0.192		0.784 ± 0.081	**0.934 ± 0.014**		0.925 ± 0.027	0.873 ± 0.079	0.896 ± 0.070
liver	***0.950 ± 0.009***		0.831 ± 0.102	0.923 ± 0.013	0.866 ± 0.035	0.921 ± 0.011	0.831 ± 0.292	0.934 ± 0.012	0.934 ± 0.005	0.936 ± 0.006
spleen	***0.946 ± 0.014***		0.671 ± 0.159	0.874 ± 0.049	0.703 ± 0.079	0.870 ± 0.057			0.914 ± 0.043	0.910 ± 0.036
uBladder	***0.888 ± 0.059***		0.666 ± 0.090		0.698 ± 0.127	0.763 ± 0.085			0.713 ± 0.246	0.713 ± 0.240
r_Psoas	*0.831 ± 0.041*		0.747 ± 0.069		0.787 ± 0.063	0.847 ± 0.030		**0.848 ± 0.039**	0.828 ± 0.050	0.830 ± 0.044
l_Psoas	*0.814 ± 0.058*		0.777 ± 0.058		0.806 ± 0.029	**0.861 ± 0.024**		0.858 ± 0.024	0.833 ± 0.033	0.832 ± 0.030
trachea	*0.894 ± 0.025*		0.840 ± 0.028		0.920 ± 0.019	**0.931 ± 0.019**				
aorta	***0.859 ± 0.044***		0.741 ± 0.039		0.753 ± 0.065	0.830 ± 0.048		0.823 ± 0.068		
sternum	***0.889 ± 0.034***		0.633 ± 0.132		0.761 ± 0.052	0.847 ± 0.057			0.660 ± 0.198	0.659 ± 0.158
1Lvert	***0.882 ± 0.014***		0.412 ± 0.403		0.718 ± 0.277	0.680 ± 0.363				
r_abdom	***0.816 ± 0.030***				0.519 ± 0.185	0.679 ± 0.162				
l_abdom	***0.793 ± 0.051***				0.551 ± 0.136	0.746 ± 0.060				
pancreas	***0.616 ± 0.143***		0.415 ± 0.142		0.408 ± 0.173	0.383 ± 0.168				
gBladder	***0.708 ± 0.131***		0.191 ± 0.219		0.276 ± 0.152	0.163 ± 0.251				
thyroid	***0.658 ± 0.225***		0.450 ± 0.122		0.549 ± 0.105	0.424 ± 0.097				
r_AdGland	***0.368 ± 0.272***		0.186 ± 0.136		0.355 ± 0.211	0.110 ± 0.157				
l_AdGland	***0.479 ± 0.283***		0.067 ± 0.095		0.373 ± 0.203	0.282 ± 0.186				



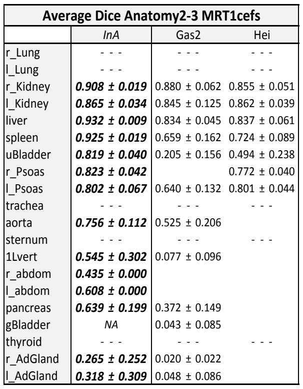
AvgDice Anat2-3 MRT1 whole body

	InA	Gas2
r_Lung	***0.929 ± 0.032***	0.903 ± 0.045
l_Lung	***0.936 ± 0.032***	0.567 ± 0.157
r_Kidney	***0.917 ± 0.008***	0.812 ± 0.122
l_Kidney	***0.918 ± 0.035***	0.808 ± 0.057
liver	***0.891 ± 0.054***	0.827 ± 0.076
spleen	***0.709 ± 0.179***	0.684 ± 0.159
uBladder	***0.850 ± 0.185***	0.709 ± 0.139
r_Psoas	***0.838 ± 0.049***	
l_Psoas	***0.849 ± 0.033***	0.820 ± 0.038
trachea	***0.768 ± 0.040***	0.731 ± 0.100
aorta	*0.726 ± 0.196*	**0.750 ± 0.082**
sternum	- - -	- - -
1Lvert	***0.740 ± 0.056***	0.415 ± 0.285
r_abdom	- - -	- - -
l_abdom	- - -	- - -
pancreas	***0.416 ± 0.156***	0.196 ± 0.278
gBladder	***0.742 ± 0.016***	0.000 ± 0.000
thyroid	*NA*	0.306 ± 0.190
r_AdGland	***0.459 ± 0.208***	0.077 ± 0.121
l_AdGland	***0.550 ± 0.191***	0.151 ± 0.261



Average Dice   Anatomy2-3   Contrast enhanced CT Thorax-Abdomen

	InA	Dic	Ga2	He	Hei	Ji2	Ke2	Ke3	Li	Spa2	Vin	Wan2	Wan3
r_Lung	*0.958 ± 0.025*	0.973 ± 0.015	0.965 ± 0.013	0.966 ± 0.016		0.963 ± 0.013	0.953 ± 0.032	0.965 ± 0.023		0.968 ± 0.015	**0.974 ± 0.016**	0.966 ± 0.014	0.971 ± 0.014
l_Lung	*0.955 ± 0.024*	**0.974 ± 0.012**	0.961 ± 0.011	0.966 ± 0.014		0.959 ± 0.010	0.957 ± 0.031	0.967 ± 0.019		0.970 ± 0.010	0.969 ± 0.017	0.967 ± 0.013	0.972 ± 0.013
r_Kidney	*0.937 ± 0.043*		0.914 ± 0.027	0.922 ± 0.014	0.861 ± 0.022	0.889 ± 0.026	0.805 ± 0.213	0.921 ± 0.044		0.870 ± 0.047	0.927 ± 0.040	0.929 ± 0.017	**0.959 ± 0.011**
l_Kidney	*0.929 ± 0.043*		0.913 ± 0.029	0.910 ± 0.023	0.874 ± 0.025	0.910 ± 0.015	0.856 ± 0.163	0.916 ± 0.034		0.829 ± 0.146	0.943 ± 0.015	0.930 ± 0.021	**0.945 ± 0.027**
liver	***0.965 ± 0.003***		0.908 ± 0.021	0.933 ± 0.009	0.827 ± 0.135	0.887 ± 0.019	0.933 ± 0.017	0.933 ± 0.016	0.937 ± 0.011		0.942 ± 0.022	0.930 ± 0.014	0.949 ± 0.010
spleen	***0.934 ± 0.026***		0.781 ± 0.075	0.896 ± 0.037	0.744 ± 0.100	0.730 ± 0.116	0.839 ± 0.151	0.895 ± 0.046		0.822 ± 0.290		0.874 ± 0.083	0.909 ± 0.069
uBladder	***0.933 ± 0.026***		0.683 ± 0.090		0.336 ± 0.261	0.679 ± 0.142	0.774 ± 0.081	0.823 ± 0.073				0.870 ± 0.064	0.866 ± 0.070
r_Psoas	*0.854 ± 0.036*				0.827 ± 0.015	0.799 ± 0.025	0.711 ± 0.161	0.806 ± 0.081			**0.874 ± 0.028**	0.847 ± 0.021	0.845 ± 0.026
l_Psoas	*0.565 ± 0.489*		0.813 ± 0.046		0.841 ± 0.031	0.794 ± 0.049	0.792 ± 0.078	0.797 ± 0.072			**0.864 ± 0.027**	0.820 ± 0.085	0.830 ± 0.074
trachea	***0.877 ± 0.032***		0.847 ± 0.050			0.855 ± 0.022	0.624 ± 0.352	0.824 ± 0.051		0.851 ± 0.022			
aorta	***0.856 ± 0.015***		0.785 ± 0.042			0.762 ± 0.039	0.578 ± 0.107	0.681 ± 0.121			0.838 ± 0.063		
sternum	***0.810 ± 0.126***		0.635 ± 0.148			0.721 ± 0.058	0.634 ± 0.189	0.713 ± 0.103				0.773 ± 0.088	0.762 ± 0.092
1Lvert	***0.914 ± 0.019***		0.624 ± 0.356			0.523 ± 0.301	0.486 ± 0.263	0.499 ± 0.296					
r_abdom	***0.709 ± 0.212***					0.453 ± 0.173	0.257 ± 0.341	0.547 ± 0.262					
l_abdom	***0.637 ± 0.204***					0.474 ± 0.180	0.134 ± 0.229	0.528 ± 0.221					
pancreas	***0.785 ± 0.039***		0.460 ± 0.159			0.423 ± 0.136	0.544 ± 0.099	0.329 ± 0.248					
gBladder	***0.857 ± 0.058***		0.381 ± 0.208			0.484 ± 0.132	0.143 ± 0.203	0.518 ± 0.241					
thyroid	***0.781 ± 0.047***		0.184 ± 0.166			0.410 ± 0.157	0.039 ± 0.055	0.127 ± 0.088					
r_AdGland	***0.671 ± 0.103***		0.213 ± 0.139			0.342 ± 0.148	0.000 ± 0.000						
l_AdGland	***0.743 ± 0.120***		0.250 ± 0.159			0.331 ± 0.176	0.000 ± 0.000	0.228 ± 0.278					



Average Dice Anatomy2-3 MRT1cefs

	InA	Gas2	Hei
r_Lung	- - -	- - -	- - -
l_Lung	- - -	- - -	- - -
r_Kidney	***0.908 ± 0.019***	0.880 ± 0.062	0.855 ± 0.051
l_Kidney	***0.865 ± 0.034***	0.845 ± 0.125	0.862 ± 0.039
liver	***0.932 ± 0.009***	0.834 ± 0.045	0.837 ± 0.061
spleen	***0.925 ± 0.019***	0.659 ± 0.162	0.724 ± 0.089
uBladder	***0.819 ± 0.040***	0.205 ± 0.156	0.494 ± 0.238
r_Psoas	***0.823 ± 0.042***		0.772 ± 0.040
l_Psoas	***0.802 ± 0.067***	0.640 ± 0.132	0.801 ± 0.044
trachea	- - -	- - -	- - -
aorta	***0.756 ± 0.112***	0.525 ± 0.206	
sternum	- - -	- - -	- - -
1Lvert	***0.545 ± 0.302***	0.077 ± 0.096	
r_abdom	***0.435 ± 0.000***		
l_abdom	***0.608 ± 0.000***		
pancreas	***0.639 ± 0.199***	0.372 ± 0.149	
gBladder	*NA*	0.043 ± 0.085	
thyroid	- - -	- - -	- - -
r_AdGland	***0.265 ± 0.252***	0.020 ± 0.022	
l_AdGland	***0.318 ± 0.309***	0.048 ± 0.086	

TABLE VI: Landmark detection results from the Anatomy benchmarks. The total count of landmarks detected (Count) and Euclidean distance error measurements in voxels are presented in the table. For the Euclidean distance the median, mean and standard deviation (Std) are shown.

Method	Benchmark	Modality	Count	Median	Mean ± Std.
Gass et al.	Anatomy1	CTwb	12	**8.784**	**10.90** ±9.491
Wyeth et al.		CTwb	12	9.592	11.11± **5.052**
Gass et al.		MRT1cefs	8	62.22	65.91 ± 20.09
Mai et al.	Anatomy2	CTwb	53	**10.34**	**20.10** ±29.99
Gass et al. (2)		CTwb	53	16.85	25.29± **22.60**
Mai et al.		CTce	44	11.41	13.01 ± 12.71
Mai et al.		MRT1wb	52	**19.00**	**99.47** ±217.4
Gass et al. (2)		MRT1wb	52	90.75	109.80± **82.85**
Mai et al.		MRT1cefs	21	**35.59**	**42.57** ± **34.95**
Gass et al. (2)		MRT1cefs	27	70.94	93.69 ± 54.24

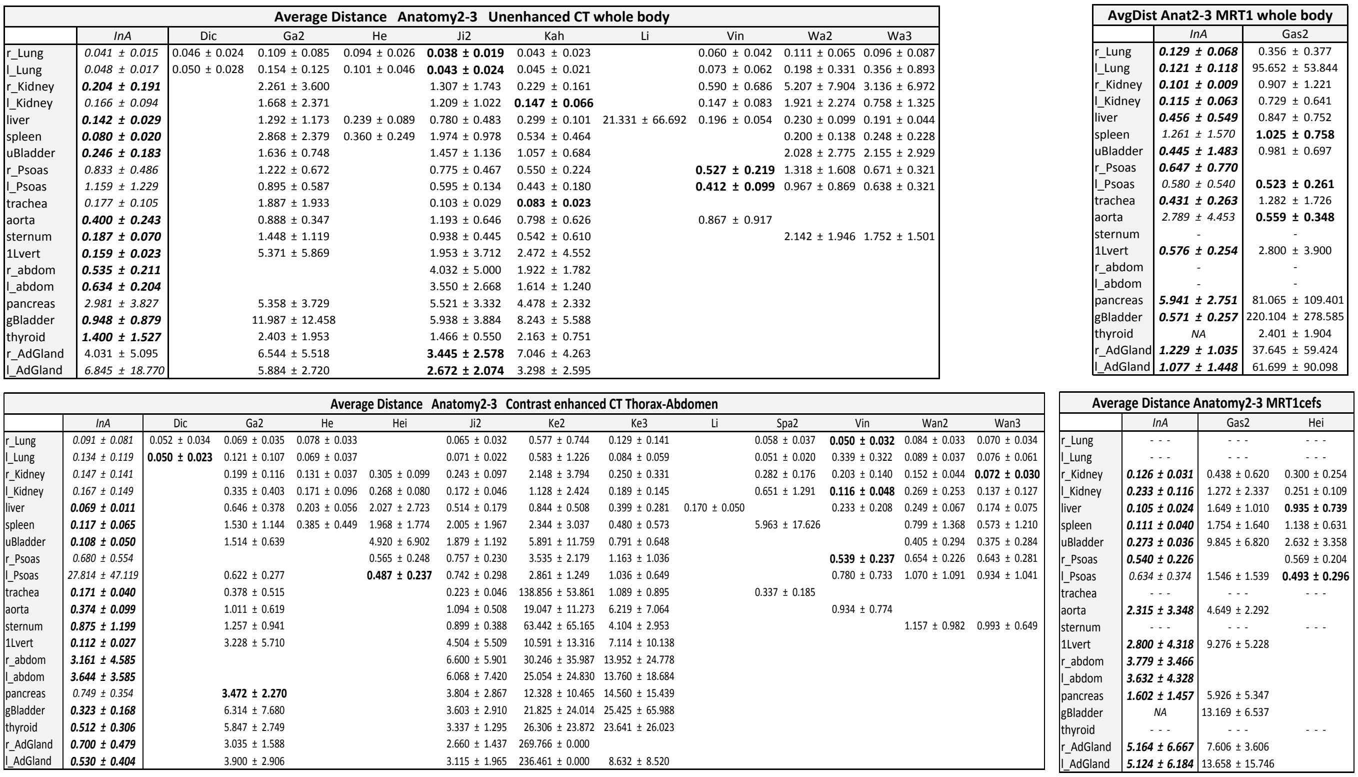
were the xyphoideus, e.g. 228 voxels in MRT1wb, and the thorax vertebrae (Th6–Th10), e.g. 189 voxels in MRT1cefs. The landmarks with lowest mean Euclidean distance error were the trachea bifurcation and right eye, both with error of 2 voxels in CTwb. Overall, the highest distance errors were computed on the MRT1wb volumes. Qualitative sample results are presented in Fig. 8.

V. Discussion

A. Challenges in biomedical image analysis

Evaluation campaigns aim to objectively compare existing methods in the search of an optimal solution for a given clinical task. The VISCERAL Anatomy benchmarks focused on the detection and segmentation of anatomical structures through the processing of large–scale 3D radiology images. Unlike previous organ segmentation benchmarks with a restrictive field–of–view and oriented towards a single anatomical target (e.g. liver [21], lung [22]) the Anatomy benchmarks use 3D clinical scans with a large field–of–view, showing either the trunk or the whole body, with up to 20 different manually annotated organs and 53 landmarks. A multi–modal gold corpus was created through the manual annotations of medical experts providing a training set that participants accessed via a cloud platform, and a private test set. This platform is capable of hosting larger data sets than those distributed to the participants through hard disks or via download, as it is currently done in other challenges. Twelve research groups submitted fully automatic algorithms for one or more of the tasks available in the benchmarks. The results are publicly available on the VISCERAL website and through a participant leaderboard. Another particularity of the Anatomy benchmarks is their innovative use of a cloud infrastructure for the creation of a Silver Corpus, running and evaluation of the challenges

TABLE V: Average Distance results (in voxels) from Anatomy2–3 in the four available modalities: CTwb, MRwb, CTce, MRce. The participants are listed with the name abbreviation used in Fig. 2, 3, 4 and 5 . The inter–anotator agreement scores (InA) obtained during the manual annotation phase are also shown. Anatomical structures are listed according to the lowest distance error, from first to last. The highest score is highlighted in bold font.



Average Distance   Anatomy2-3   Unenhanced CT whole body

	InA	Dic	Ga2	He	Ji2	Kah	Li	Vin	Wa2	Wa3
r_Lung	*0.041 ± 0.015*	0.046 ± 0.024	0.109 ± 0.085	0.094 ± 0.026	**0.038 ± 0.019**	0.043 ± 0.023		0.060 ± 0.042	0.111 ± 0.065	0.096 ± 0.087
l_Lung	*0.048 ± 0.017*	0.050 ± 0.028	0.154 ± 0.125	0.101 ± 0.046	**0.043 ± 0.024**	0.045 ± 0.021		0.073 ± 0.062	0.198 ± 0.331	0.356 ± 0.893
r_Kidney	***0.204 ± 0.191***		2.261 ± 3.600		1.307 ± 1.743	0.229 ± 0.161		0.590 ± 0.686	5.207 ± 7.904	3.136 ± 6.972
l_Kidney	*0.166 ± 0.094*		1.668 ± 2.371		1.209 ± 1.022	**0.147 ± 0.066**		0.147 ± 0.083	1.921 ± 2.274	0.758 ± 1.325
liver	***0.142 ± 0.029***		1.292 ± 1.173	0.239 ± 0.089	0.780 ± 0.483	0.299 ± 0.101	21.331 ± 66.692	0.196 ± 0.054	0.230 ± 0.099	0.191 ± 0.044
spleen	***0.080 ± 0.020***		2.868 ± 2.379	0.360 ± 0.249	1.974 ± 0.978	0.534 ± 0.464			0.200 ± 0.138	0.248 ± 0.228
uBladder	***0.246 ± 0.183***		1.636 ± 0.748		1.457 ± 1.136	1.057 ± 0.684			2.028 ± 2.775	2.155 ± 2.929
r_Psoas	*0.833 ± 0.486*		1.222 ± 0.672		0.775 ± 0.467	0.550 ± 0.224		**0.527 ± 0.219**	1.318 ± 1.608	0.671 ± 0.321
l_Psoas	*1.159 ± 1.229*		0.895 ± 0.587		0.595 ± 0.134	0.443 ± 0.180		**0.412 ± 0.099**	0.967 ± 0.869	0.638 ± 0.321
trachea	*0.177 ± 0.105*		1.887 ± 1.933		0.103 ± 0.029	**0.083 ± 0.023**				
aorta	***0.400 ± 0.243***		0.888 ± 0.347		1.193 ± 0.646	0.798 ± 0.626		0.867 ± 0.917		
sternum	***0.187 ± 0.070***		1.448 ± 1.119		0.938 ± 0.445	0.542 ± 0.610			2.142 ± 1.946	1.752 ± 1.501
1Lvert	***0.159 ± 0.023***		5.371 ± 5.869		1.953 ± 3.712	2.472 ± 4.552				
r_abdom	***0.535 ± 0.211***				4.032 ± 5.000	1.922 ± 1.782				
l_abdom	***0.634 ± 0.204***				3.550 ± 2.668	1.614 ± 1.240				
pancreas	*2.981 ± 3.827*		5.358 ± 3.729		5.521 ± 3.332	4.478 ± 2.332				
gBladder	***0.948 ± 0.879***		11.987 ± 12.458		5.938 ± 3.884	8.243 ± 5.588				
thyroid	***1.400 ± 1.527***		2.403 ± 1.953		1.466 ± 0.550	2.163 ± 0.751				
r_AdGland	4.031 ± 5.095		6.544 ± 5.518		**3.445 ± 2.578**	7.046 ± 4.263				
l_AdGland	*6.845 ± 18.770*		5.884 ± 2.720		**2.672 ± 2.074**	3.298 ± 2.595				

AvgDist Anat2-3 MRT1 whole body

	InA	Gas2
r_Lung	***0.129 ± 0.068***	0.356 ± 0.377
l_Lung	***0.121 ± 0.118***	95.652 ± 53.844
r_Kidney	***0.101 ± 0.009***	0.907 ± 1.221
l_Kidney	***0.115 ± 0.063***	0.729 ± 0.641
liver	***0.456 ± 0.549***	0.847 ± 0.752
spleen	*1.261 ± 1.570*	**1.025 ± 0.758**
uBladder	***0.445 ± 1.483***	0.981 ± 0.697
r_Psoas	***0.647 ± 0.770***	
l_Psoas	*0.580 ± 0.540*	**0.523 ± 0.261**
trachea	***0.431 ± 0.263***	1.282 ± 1.726
aorta	*2.789 ± 4.453*	**0.559 ± 0.348**
sternum	-	-
1Lvert	***0.576 ± 0.254***	2.800 ± 3.900
r_abdom	-	-
l_abdom	-	-
pancreas	***5.941 ± 2.751***	81.065 ± 109.401
gBladder	***0.571 ± 0.257***	220.104 ± 278.585
thyroid	*NA*	2.401 ± 1.904
r_AdGland	***1.229 ± 1.035***	37.645 ± 59.424
l_AdGland	***1.077 ± 1.448***	61.699 ± 90.098

Average Distance   Anatomy2-3   Contrast enhanced CT Thorax-Abdomen

	InA	Dic	Ga2	He	Hei	Ji2	Ke2	Ke3	Li	Spa2	Vin	Wan2	Wan3
r_Lung	*0.091 ± 0.081*	0.052 ± 0.034	0.069 ± 0.035	0.078 ± 0.033		0.065 ± 0.032	0.577 ± 0.744	0.129 ± 0.141		0.058 ± 0.037	**0.050 ± 0.032**	0.084 ± 0.033	0.070 ± 0.034
l_Lung	*0.134 ± 0.119*	**0.050 ± 0.023**	0.121 ± 0.107	0.069 ± 0.037		0.071 ± 0.022	0.583 ± 1.226	0.084 ± 0.059		0.051 ± 0.020	0.339 ± 0.322	0.089 ± 0.037	0.076 ± 0.061
r_Kidney	*0.147 ± 0.141*		0.199 ± 0.116	0.131 ± 0.037	0.305 ± 0.099	0.243 ± 0.097	2.148 ± 3.794	0.250 ± 0.331		0.282 ± 0.176	0.203 ± 0.140	0.152 ± 0.044	**0.072 ± 0.030**
l_Kidney	*0.167 ± 0.149*		0.335 ± 0.403	0.171 ± 0.096	0.268 ± 0.080	0.172 ± 0.046	1.128 ± 2.424	0.189 ± 0.145		0.651 ± 1.291	**0.116 ± 0.048**	0.269 ± 0.253	0.137 ± 0.127
liver	***0.069 ± 0.011***		0.646 ± 0.378	0.203 ± 0.056	2.027 ± 2.723	0.514 ± 0.179	0.844 ± 0.508	0.399 ± 0.281	0.170 ± 0.050		0.233 ± 0.208	0.249 ± 0.067	0.174 ± 0.075
spleen	***0.117 ± 0.065***		1.530 ± 1.144	0.385 ± 0.449	1.968 ± 1.774	2.005 ± 1.967	2.344 ± 3.037	0.480 ± 0.573		5.963 ± 17.626		0.799 ± 1.368	0.573 ± 1.210
uBladder	***0.108 ± 0.050***		1.514 ± 0.639		4.920 ± 6.902	1.879 ± 1.192	5.891 ± 11.759	0.791 ± 0.648				0.405 ± 0.294	0.375 ± 0.284
r_Psoas	*0.680 ± 0.554*				0.565 ± 0.248	0.757 ± 0.230	3.535 ± 2.179	1.163 ± 1.036			**0.539 ± 0.237**	0.654 ± 0.226	0.643 ± 0.281
l_Psoas	*27.814 ± 47.119*		0.622 ± 0.277		**0.487 ± 0.237**	0.742 ± 0.298	2.861 ± 1.249	1.036 ± 0.649			0.780 ± 0.733	1.070 ± 1.091	0.934 ± 1.041
trachea	***0.171 ± 0.040***		0.378 ± 0.515			0.223 ± 0.046	138.856 ± 53.861	1.089 ± 0.895		0.337 ± 0.185			
aorta	***0.374 ± 0.099***		1.011 ± 0.619			1.094 ± 0.508	19.047 ± 11.273	6.219 ± 7.064			0.934 ± 0.774		
sternum	***0.875 ± 1.199***		1.257 ± 0.941			0.899 ± 0.388	63.442 ± 65.165	4.104 ± 2.953				1.157 ± 0.982	0.993 ± 0.649
1Lvert	***0.112 ± 0.027***		3.228 ± 5.710			4.504 ± 5.509	10.591 ± 13.316	7.114 ± 10.138					
r_abdom	***3.161 ± 4.585***					6.600 ± 5.901	30.246 ± 35.987	13.952 ± 24.778					
l_abdom	***3.644 ± 3.585***					6.068 ± 7.420	25.054 ± 24.830	13.760 ± 18.684					
pancreas	*0.749 ± 0.354*		**3.472 ± 2.270**			3.804 ± 2.867	12.328 ± 10.465	14.560 ± 15.439					
gBladder	***0.323 ± 0.168***		6.314 ± 7.680			3.603 ± 2.910	21.825 ± 24.014	25.425 ± 65.988					
thyroid	***0.512 ± 0.306***		5.847 ± 2.749			3.337 ± 1.295	26.306 ± 23.872	23.641 ± 26.023					
r_AdGland	***0.700 ± 0.479***		3.035 ± 1.588			2.660 ± 1.437	269.766 ± 0.000						
l_AdGland	***0.530 ± 0.404***		3.900 ± 2.906			3.115 ± 1.965	236.461 ± 0.000	8.632 ± 8.520					

Average Distance Anatomy2-3 MRT1cefs

	InA	Gas2	Hei
r_Lung	- - -	- - -	- - -
l_Lung	- - -	- - -	- - -
r_Kidney	***0.126 ± 0.031***	0.438 ± 0.620	0.300 ± 0.254
l_Kidney	***0.233 ± 0.116***	1.272 ± 2.337	0.251 ± 0.109
liver	***0.105 ± 0.024***	1.649 ± 1.010	**0.935 ± 0.739**
spleen	***0.111 ± 0.040***	1.754 ± 1.640	1.138 ± 0.631
uBladder	***0.273 ± 0.036***	9.845 ± 6.820	2.632 ± 3.358
r_Psoas	***0.540 ± 0.226***		0.569 ± 0.204
l_Psoas	*0.634 ± 0.374*	1.546 ± 1.539	**0.493 ± 0.296**
trachea	- - -	- - -	- - -
aorta	***2.315 ± 3.348***	4.649 ± 2.292	
sternum	- - -	- - -	- - -
1Lvert	***2.800 ± 4.318***	9.276 ± 5.228	
r_abdom	***3.779 ± 3.466***		
l_abdom	***3.632 ± 4.328***		
pancreas	***1.602 ± 1.457***	5.926 ± 5.347	
gBladder	*NA*	13.169 ± 6.537	
thyroid	- - -	- - -	- - -
r_AdGland	***5.164 ± 6.667***	7.606 ± 3.606	
l_AdGland	***5.124 ± 6.184***	13.658 ± 15.746	

and storing the participant outputs and VMs with their self–installed executables. Bringing the algorithms to the data set is a shift in the common approach of distributing large amounts of data through hard disks or as downloads from the web. Although other challenges have been run through an online platform [14], [15], in the Anatomy benchmarks the participants install functional executables inside their provided cloud VMs which are then submitted and tested by the administrators. Downloading large data sets can hamper the fairness of evaluating algorithms due to the time restriction of performing a live challenge, sometimes grouping the participating methods under different conditions [22]. The scalability and storage capacity of a cloud infrastructure is virtually unlimited. This has allowed the interaction of the VISCERAL participants with a training data set of over 2000 patient volumes, manually annotated labels and radiologic reports. During the testing phase participants are restricted from accessing their VMs allowing the administrators to run their algorithms on unseen data, thus promoting an unbiased evaluation using the same computation power for each participant and a common large data set [11]. Creating and sustaining individual data sets for each anatomical structure is a more complex task and could hamper the collaboration between different groups working on similar topics. Additionally, once the evaluation of these algorithms has been performed during the challenge, no further usage can be given to the participating methods, limiting the reproducibility and exploitation of the results. In [17] the results from the algorithms were considered for the definition of the ground truth in the challenge data set. After the VISCERAL Anatomy benchmarks, the administrators ran the participant algorithms in their VMs, creating an much larger number of "lower quality" annotations with the consensus estimates in previously non annotated images. This opens the possibility to use this kind of data, which had not been exploited in previous challenges. Since the data are stored centrally, and not distributed outside the cloud environment, the legal and ethical requirements of such data sets can also be satisfied. This allows the benchmarking of algorithms on confidential data sets, with only a smaller training set accessed by participants [46].

B. Anatomical structure segmentation

In the VISCERAL Anatomy benchmarks the scores are presented per–modality per–structure, therefore defining a single winning algorithm is not straightforward. The aim behind creating a large multi–modal data set where different algorithms can test their methods foments an open discussion on whether a specific algorithm can target a certain clinical task better. However, there were clear trends in the Anatomy benchmarks of the participating algorithms and a large evaluation of the results that is publicly available in the VISCERAL Leaderboard.

Ten algorithms participated in the organ segmentation task from the Anatomy benchmarks. The most common approach was multi–atlas segmentation, with five algorithms Ga2 [38], Hei [39], Ji2 [40], Kah [42], and Ke3 [43] implementing a variation of this method. There were three approaches that attempted to segment all the 20 structures available in one or more modalities: Ga2 [38], Ji2 [40], and Kah [42]. The generalization of this approach for multiple organs with different shapes and intensities, makes it a reliable option either as a complete approach (Ga2 [38], Ji2 [40]) or as a first step requiring refinement of the results (Hei [39], Kah [42], Ke3 [43]). In the Anatomy benchmarks these additional refinement steps, like graph–cut (Kah [42], Ke3 [43]), gave an overall advantage in the segmentation scores when compared to simpler approaches based only on majority voting or weighted label fusion (Ga2 [38], Ji2 [40]). This was particularly clear in CTwb where the scores from Kah [42] were generally higher for this method with much sharper edge definition and shape resemblance to the manual ground truth annotations (see Fig.6). Still, these methods have all a long runtime per volume when compared to other approaches. The fastest multi–atlas segmentation method was Hei [39], that used a discrete deformable registration framework and was able to segment 7 structures per volume in 40 minutes. Their scores are still competitive in the participating structures (see Table IV and V).

Intensity–based clustering Dic [30] and 'rule–based' approaches Sp2 [31] gave particularly good results for structures with high contrast (e.g. lungs) and a much faster segmentation per volume. These methods are not based on predefined shape–models and are faster to execute for some organs (e.g. lungs). Nevertheless, both methods are hard to generalize and, notably for 'rule–based', are more prone to leaking errors, leading to failed segmentations in complicated cases.

Shape and appearance segmentation models were also a popular choice among the participating groups: Vin [34], Wa3 [36], He [33]. The best scores for the structures with a higher number of participants, were obtained by these methods. Both of them provide a good trade–off between a lower computation time than atlas registration methods and accurate segmentations. Unfortunately, these methods were not tested for all the available structures included in the Gold Corpus. This suggests that their implementation and generalization are not as straightforward as atlas–based registration methods. Nevertheless, previous studies have demonstrated that the approaches using shape and appearance models attain more accurate segmentation compared with atlas-based registration methods [47] in smaller structures with higher variability (e.g. pancreas). The method of Wa3 [36] included in Anatomy3 a shape model guided local phase analysis that improved their scores even further from their method used in Anatomy2 (Wa2 [35]). Both Vin [34] and Wa3 [36] start their models in low resolution, computing simple threshold and mathematical operations. Wa3 [36] has a faster implementation (1 hour per volume, 10 structures) using an effective technique that focuses the registration of the model only on pre–defined 'trusted zones' in the patient volume. These initial image correspondences are then refined by registering their models to the target image.

1) CT segmentation task: Computed tomography segmentations were the most popular and successful tasks with algorithms obtaining the best scores for most structures in the Gold Corpus test set. In 15 out of 40 CT structures, the inter–annotator agreement scores were reached by at least one of the participant algorithms. Although the evaluation overlaps vary strongly depending on the anatomical structure, the results achieved by the top algorithms are close to the range shown in the inter–annotator agreement. In the CTwb modality, where no tissue contrast is added and the large field–of–view includes the whole body, this is particularly challenging both for annotators and segmentation methods. Structures with high tissue contrast such as the lungs are well segmented in both CT modalities. Bone structures (sternum and 1^{st} lumbar vertebra) were better segmented in CTwb than in CTce. The advantage of added tissue contrast (CTce) is clear in structures like the urinary and gallbladder, with higher scores both in the inter–annotator agreement and in the output algorithm segmentations. However, structures with low tissue contrast in CTwb like the thyroid and adrenal glands show similar scores in both CT modalities, even though CTce has a much higher inter–annotator agreement. This could be the result of a more stable spatial location of these structures in the human body that is better detected by approaches with emphasis on the relative position of these structures (e.g. multi–atlas segmentation).

2) MR segmentation task: MR segmentation methods were uncommon among participants, with only 3 algorithms: Ga1 [37], Ga2 [38], and Hei [39], addressing these modalities (MRT1wb and MRT1cefs). Ga2 [38] participated in MR images with the same multi–atlas segmentation method used for segmenting CT scans, with moderate results. All the structures had a lower overlap in MR images and bigger distance errors (see Table IV and Table V). Isolated segmented regions with no relation to the target structure and failure to detect the structure borders, are common errors when the qualitative results are inspected for this algorithm in MRwb (see Fig. 7). Still, it provides competitive results for tubular structures like the trachea and the aorta with an average overlap similar to the inter–annotator agreement.

Hei [39] participated in Anatomy3 with MRT1cefs segmentations and obtained overall a lower average distance error and better overlap scores than Ga2 [38](Anat2), with a smaller number of structures segmented (7 vs. 12). The registration method of Hei [39] has a regularisation parameter and computes a global minimum that ends up generating more stable spatial deformations, with the output segmentations mimicking more closely the anatomical structures than those from Ga2 [38]. It is also a faster method with a runtime per structure of 6 min vs. 16 mins of Ga2 [38]. The overlap results obtained in MRT1cefs were closer to the inter–annotator agreement than those from MRwb. MR segmentation and hence its validation have been rare in the literature, except for prostate and brain structure segmentation (e.g. PROMISE12 challenge [20], BRATS challenge [15]). In recent years, new approaches have been proposed for some trunk organs such as the lungs [48], liver [49] and thyroid [50] with promising results. The advantage of a common MR data set with more organs can help to improve the organ detection and localization of structures for relevant clinical tasks, e.g., radiotherapy planning [51] and surgical follow–up [52].

C. Landmark localization task

Localization of anatomical landmarks is an important process in intra and interpatient registration and study location and navigation. Three methods participated in the landmark localization task from Anatomy1 and Anatomy2. Wyeth et al. [45] and Mai et al. [44] both used machine learning classification approaches per landmark, random forests in the former and a support vector machine for the latter. They performed a fast localization with a runtime of <10 seconds per volume. On the other hand, Gass et al. [38] performed a patch-based block search using cross-correlation within a large search region in the target image. The best matches are then fused taking the median of location coordinates. The process is performed for all the available landmarks in the volume in an average runtime of 30 mins.

The median and mean Euclidean distance errors from Mai et al. [44] are lower across all the modalities when compared to Gass et al. [38] and Wyeth et al. [45]. However, the Gass et al. [38] results have a lower standard deviation for the whole body modalities (CTwb and MRwb). The results show that the search regions created in Gass et al. [38], are able to limit the localization errors to smaller areas than the location normalization from Mai et al. [44]. An example error in the test set showing this feature, was seen in the localization of the vertebrae, where the method from Mai et al. [44] was able to locate the body of the thoracic vertebrae from Th4 and Th6 with less than 3 voxels of error but the intermediate vertebrae Th5 had a considerable error of 215 voxels. Nevertheless, Mai et al. [44] had consistently smaller distance errors compared to the other participants, with clear localization errors in fewer cases.

D. Anatomy benchmark series buildout

Although a subset of the test set was different in Anatomy1, almost all structures had better results in the following benchmarks. This supports the motivation behind these benchmarks of having strong baseline comparisons to target an optimal solution from the participant algorithms. For example, the total DICE average in CTce structures went from 0.662 in Anatomy1 to 0.731 in Anatomy3, when only the best results are considered. This might have been caused by a larger training set provided to participants in Anatomy3. A targeted optimization of the results is now encouraged, in the currently active Anatomy 3 continuous evaluation benchmark. The analysis of the results shows that multiple algorithms can obtain already robust organ segmentations for popular structures like the liver and kidneys. There are still many important structures like the pancreas and adrenal glands, where anatomical variability requires larger training sets for more robust shape models.

E. VISCERAL Anatomy limitations

For the Anatomy benchmarks participants could select all or just a subset of the presented tasks. The aim of a multi–modal multi–structure benchmark was the inclusion of different research groups in the benchmarks and to compare multiple segmentation methods both for target–oriented structures (e.g. Li [32], Dic [30]) and more general approaches that could be applied to different structures (e.g. Ga2 [37], Kah [42]). A weakness of the Anatomy benchmarks was the small number of volumes tested per structure per modality, which resulted from the distribution of the expert manual annotation time between all the possible structures in four modalities. It was difficult to have more data manually annotated, but new follow up proposals are being submitted to hopefully extend the data set in the future. An alternative direction is to include data annotated by other groups, but with the same protocols to collaboratively extend the data set.

Even though there were algorithms in all the available tasks, a large number of tasks (organs) resulted sometimes in a small number of participants. This restricted the number of methods compared, limiting the selection of a winning methodology (e.g. for MR segmentation). Several registrants for the Anatomy benchmarks, who were given access to the training data and the cloud platform, ultimately did not submit an algorithm for evaluation. Feedback from these research groups indicated as a common concern the high time investment required to access, train and implement their algorithms in the cloud platform. Having a continuous cycle of benchmarks (e.g., annual events) could motivate more participants to invest this time and submit their results for future benchmarks.

F. Silver Corpus

In order to put these benchmark results to a long–term scalable use, a Silver Corpus was generated with the fusion of the participants algorithms output segmentations in a larger set of unannotated data. Although this could generate less accurate annotations than those created manually, the fusion of approaches on the test set was able to overall produce better segmentations than any algorithm in the benchmark, particularly for structures with low DICE scores. Such a Silver Corpus was made possible by having executables in the VMs of the participants. After testing multiple fusion techniques, the SIMPLE approach [53] initialized with performance estimate weights, was the best performing method. The highest increase was seen in the thyroid where the fused estimation produced an increase in the mean DICE score of 0.3 compared to the best participant algorithm: 0.41 vs 0.71 in CTce. The complete results and detailed description of the experiments can be found in reference [54].

The VISCERAL Anatomy Silver Corpus consists of 264 patient volumes, in one of the four VISCERAL modalities (CT and MR), and 4323 segmentations of their anatomical structures (CTce 1227, CTwb 1122, MRT1wb 1095 and MRT1cefs 879). All the volumes and anatomical structure segmentations are publicly available for the research community.

VI. Conclusion

The VISCERAL project organized three Anatomy benchmarks on processing large–scale 3D radiology image data. It developed an innovative cloud–based evaluation approach, where all the participants algorithms share a common test set using identical computing instances without reseachers having access to the test data. Twenty segmentation algorithms and three landmark detection algorithms were submitted for evaluation. Different algorithms obtained the best scores in each of the available four imaging modalities and for subsets of anatomical structures. The algorithms were implemented on individual virtual machines that enable their further usage for comparison on other data sets and also for the creation of a much larger Silver Corpus through the fusion of various output segmentations. Even though the VISCERAL project is officially finished, the data set, the evaluation framework and the Silver Corpus are now available free of charge via the VISCERAL registration system.

Contributions

IE, AFR, OG, KG, AH, AJ, OAJT, GK, MK, GL, BHM, HM, TSF, RS, AAT, AW, MAW and MW organized the VISCERAL Anatomy benchmarks. All other authors contributed as participants of the three Anatomy benchmarks. OAJT, AAT, MK, KG, MW and OG analyzed the results from the benchmarks. OAJT and AAT wrote the manuscript.

Acknowledgment

The research leading to these results received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 318068 (VISCERAL) and 257528 (KHRESMOI). We would like to thank Microsoft research for their support on using the Azure cloud for our benchmarks. We also thank all research groups contributing to this work by participating at the VISCERAL Anatomy benchmarks.

References

[1] K. Marten, F. Auer, S. Schmidt, G. Kohl, E. J. Rummeny, and C. Engelke, "Inadequacy of manual measurements compared to automated CT volumetry in assessment of treatment response of pulmonary metastases using RECIST criteria," *European Radiology*, vol. 16, no. 4, pp. 781–790, 2006.

[2] N. Sharma and L. M. Aggarwal, "Automated medical image segmentation techniques," *Journal of medical physics/Association of Medical Physicists of India*, vol. 35, no. 1, p. 3, 2010.

[3] K. Doi, "Current status and future potential of computer–aided diagnosis in medical imaging," *British Journal of Radiology*, vol. 78, pp. 3–19, 2005.

[4] H. Kobatake, "Future CAD in multi-dimensional medical images: â€" project on multi-organ, multi-disease CAD system â€"," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4–5, pp. 258–266, 2007.

[5] H. Yoshida and A. Dachman, "CAD techniques, challenges, andcontroversies in computed tomographic colonography," *Abdominal Imaging*, vol. 30, no. 1, pp. 26–41, 2004.

[6] 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*, pp. 725902–725902, International Society for Optics and Photonics, 2009.

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

[8] A. Criminisi, D. Robertson, E. Konukoglu, J. Shotton, S. Pathak, S. White, and K. Siddiqui, "Regression forests for efficient anatomy detection and localization in computed tomography scans," *Medical Image Analysis*, vol. 17, no. 8, pp. 1293–1303, 2013.

[9] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[10] 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 *CLEF conference*, Springer Lecture Notes in Computer Science, 2012.

[11] G. Langs, A. Hanbury, B. Menze, and H. Müller, "Visceral: Towards large data in medical imaging — challenges and directions," in *Medical Content-Based Retrieval for Clinical Decision Support* (H. Greenspan, H. Müller, and T. Syeda-Mahmood, eds.), Lecture Notes in Computer Science, (Berlin, Heidelberg), pp. 92–98, Springer, 2012.

[12] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel, "Improvements that don't add up: ad-hoc retrieval results since 1998," in *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pp. 601–610, ACM, 2009.

[13] D. J. Hand, "Classifier technology and the illusion of progress," *Statistical science*, vol. 21, no. 1, pp. 1–14, 2006.

[14] D. W. Shattuck, G. Prasad, M. Mirza, K. L. Narr, and A. W. Toga, "Online resource for validation of brain segmentation methods," *NeuroImage*, vol. 45, no. 2, pp. 431 – 439, 2009.

[15] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, C. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *Medical Imaging, IEEE Transactions on*, vol. 34, no. 10, pp. 1993–2024, 2015.

[16] O. Camara, T. Mansi, M. Pop, K. Rhode, M. Sermesant, and A. Young, *Statistical Atlases and Computational Models of the Heart: Imaging and Modelling Challenges: Third International Workshop, STACOM 2012, Held in Conjunction with MICCAI 2012, Nice, France, October 5, 2012, Revised Selected Papers*, vol. 7746. Springer, 2013.

[17] P. Lo, B. Van Ginneken, J. M. Reinhardt, T. Yavarna, P. A. De Jong, B. Irving, C. Fetita, M. Ortner, R. Pinho, J. Sijbers, *et al.*, "Extraction of airways from ct (exact'09)," *Medical Imaging, IEEE Transactions on*, vol. 31, no. 11, pp. 2093–2107, 2012.

[18] R. D. Rudyanto, S. Kerkstra, E. M. Van Rikxoort, C. Fetita, P.-Y. Brillet, C. Lefevre, W. Xue, X. Zhu, J. Liang, İ. Öksüz, *et al.*, "Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: the vessel12 study," *Medical image analysis*, vol. 18, no. 7, pp. 1217–1232, 2014.

[19] B. van Ginneken, S. G. Armato, B. de Hoop, S. van Amelsvoort-van de Vorst, T. Duindam, M. Niemeijer, K. Murphy, A. Schilham, A. Retico, M. E. Fantacci, *et al.*, "Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the anode09 study," *Medical image analysis*, vol. 14, no. 6, pp. 707–722, 2010.

[20] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, R. Strand, F. Malmberg, Y. Ou, C. Davatzikos, M. Kirschner, F. Jung, J. Yuan, W. Qiu, Q. Gao, P. E. Edwards, B. Maan, F. van der Heijden, S. Ghose, J. Mitra, J. Dowling, D. Barratt, H. Huisman, and A. Madabhushi, "Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge," *Medical Image Analysis*, vol. 18, no. 2, pp. 359–373, 2014.

[21] T. Heimann, B. Van Ginneken, M. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, *et al.*, "Comparison and evaluation of methods for liver segmentation from CT datasets," *Medical Imaging, IEEE Transactions on*, vol. 28, no. 8, pp. 1251–1265, 2009.

[22] K. Murphy, B. Van Ginneken, J. M. Reinhardt, S. Kabus, K. Ding, X. Deng, K. Cao, K. Du, G. E. Christensen, V. Garcia, *et al.*, "Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge,"

Medical Imaging, IEEE Transactions on, vol. 30, no. 11, pp. 1901–1920, 2011.

[23] M. G. Linguraru, J. A. Pura, A. S. Chowdhury, and R. M. Summers, "Multi-organ segmentation from multi-phase abdominal CT via 4D graphs using enhancement, shape and location optimization," in *Medical Image Computing and Computer-Assisted Intervention " MICCAI 2010* (T. Jiang, N. Navab, J. P. Pluim, and M. A. Viergever, eds.), vol. 6363 of *Lecture Notes in Computer Science*, pp. 89–96, Springer Berlin Heidelberg, 2010.

[24] T. Okada, M. G. Linguraru, Y. Yoshida, M. Hori, R. M. Summers, Y.-W. Chen, N. Tomiyama, and Y. Sato, "Abdominal multi-organ segmentation of CT images based on hierarchical spatial modeling of organ interrelations," in *Abdominal Imaging. Computational and Clinical Applications*, pp. 173–180, Springer, 2011.

[25] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection and tool," *BMC Medical Imaging*, vol. 15, no. 1, p. 29, 2015.

[26] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[27] A. Fenster and B. Chiu, "Evaluation of segmentation algorithms for medical imaging," in *Conference Proceedings IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 7, pp. 7186–7189, 2005.

[28] O. A. Jiménez del Toro, O. Goksel, B. Menze, H. Müller, G. Langs, M.-A. Weber, I. Eggel, K. Gruenberg, M. Holzer, G. Kotsios-Kontokotsios, M. Krenn, R. Schaer, A. A. Taha, M. Winterstein, and A. Hanbury, "VISCERAL – VISual Concept Extraction challenge in RAdioLogy: ISBI 2014 challenge organization," in *Proceedings of the VISCERAL Challenge at ISBI* (O. Goksel, ed.), no. 1194 in CEUR Workshop Proceedings, (Beijing, China), pp. 6–15, May 2014.

[29] O. Goksel, A. Foncubierta-Rodríguez, O. A. Jiménez-del Toro, H. Müller, G. Langs, M.-A. Weber, B. Menze, I. Eggel, K. Gruenberg, M. Winterstein, M. Holzer, M. Krenn, G. Kontokotsios, S. Metallidis, R. Schaer, A. A. Taha, A. Jakab, T. Salas Fernandez, and A. Hanbury, "Overview of the VISCERAL challenge at ISBI 2015," in *Proceedings of the VISCERAL Challenge at ISBI* (O. Goksel *et al.*, eds.), no. 1390 in CEUR Workshop Proceedings, pp. 6–11, Apr 2015.

[30] Y. Dicente Cid, O. A. Jiménez-del Toro, A. Depeursinge, and H. Müller, "Efficient and fully automatic segmentation of the lungs in CT volumes," in *Proceedings of the VISCERAL Challenge at ISBI* (O. G. et al., ed.), no. 1390 in CEUR Workshop Proceedings, Apr 2015.

[31] A. B. Spanier and L. Joskowicz, "Rule–based ventral cavity multi–organ automatic segmentation," in *Medical Computer Vision: Algorithms for Big Data: International Workshop, MCV 2014*, vol. 8848, pp. 163–170, Springer, 2014.

[32] X. Li, C. Huang, F. Jia, Z. Li, C. Fang, and Y. Fan, "Automatic liver segmentation using statistical prior models and free–form deformation," in *Medical Computer Vision: Algorithms for Big Data: International Workshop, MCV 2014*, vol. 8848, pp. 181–188, Springer, 2014.

[33] B. He, C. Huang, and F. Jia, "Fully automatic multi–organ segmentation based on multi–boost learning and statistical shape model search," in *Proceedings of the VISCERAL Challenge at ISBI* (O. Goksel *et al.*, eds.), no. 1390 in CEUR Workshop Proceedings, pp. 18–21, Apr 2015.

[34] G. Vincent, G. Guillard, and M. Bowes, "Fully automatic segmentation of the prostate using active appearance models," in *MICCAI Workshop: Prostate Cancer Imaging: The PROMISE12 Prostate Segmentation Challenge*, 2012.

[35] C. Wang and O. Smedby, "Automatic multi–organ segmentation using fast model based level set method and hierarchical shape priors," in *Proceedings of the VISCERAL Challenge at ISBI* (O. Goksel, ed.), no. 1194 in CEUR Workshop Proceedings, (Beijing, China), pp. 25–31, May 2014.

[36] C. Wang and O. Smedby, "Multi–organ segmentation using shape model guided local phase analysis," in *Medical Image Computing and Computer–Assisted Intervention â€" MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), vol. 9351 of *Lecture Notes in Computer Science*, pp. 149–156, Springer International Publishing, 2015.

[37] O. Goksel, T. Gass, and G. Szekely, "Segmentation and landmark localization based on multiple atlases," in *Proceedings of the VISCERAL Challenge at ISBI* (O. Goksel, ed.), no. 1194 in CEUR Workshop Proceedings, (Beijing, China), pp. 37–43, May 2014.

[38] T. Gass, G. Szekely, and O. Goksel, "Multi–atlas segmentation and landmark localization in images with large field of view," in *Medical Computer Vision: Algorithms for Big Data: International Workshop, MCV 2014*, vol. 8848, pp. 171–180, Springer, 2014.

[39] M. P. Heinrich, O. Maier, and H. Handels, "Multi–modal multi–atlas segmentation using discrete optimisation and self–similarities," in *Proceedings of the VISCERAL Challenge at ISBI* (O. Goksel *et al.*, eds.), no. 1390 in CEUR Workshop Proceedings, Apr 2015.

[40] O. A. Jiménez-del Toro and H. Müller, "Hierarchic multi–atlas based segmentation for anatomical structures: Evaluation in the VISCERAL Anatomy benchmarks," in *Medical Computer Vision: Algorithms for Big Data: International Workshop, MCV 2014*, vol. 8848, pp. 189–200, Springer, 2014.

[41] O. A. Jiménez-del Toro, Y. Dicente Cid, A. Depeursinge, and H. Müller, "Hierarchic anatomical structure segmentation guided by spatial correlations (AnatSeg–Gspac): VISCERAL Anatomy3," in *Proceedings of the VISCERAL Challenge at ISBI* (O. Goksel *et al.*, eds.), no. 1390 in CEUR Workshop Proceedings, pp. 22–26, Apr 2015.

[42] 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 *Proceedings of the VISCERAL Challenge at ISBI* (O. Goksel *et al.*, eds.), no. 1390 in CEUR Workshop Proceedings, pp. 12–17, Apr 2015.

[43] 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 *Medical Computer Vision: Algorithms for Big Data: International Workshop, MCV 2014*, vol. 8848, pp. 201–210, Springer, 2014.

[44] D. Mai, P. Fischer, T. Blein, J. Dürr, K. Palme, T. Brox, and O. Ronneberger, "Discriminative detection and alignment in volumetric data," in *Pattern Recognition (GCPR 2013)*, vol. 8142, pp. 205–214, Springer, 2013.

[45] M. A. Dabbah, S. Murphy, H. Pello, R. Courbon, E. Beveridge, S. Wiseman, D. Wyeth, and I. Poole, "Detection and location of 127 anatomical landmarks in diverse CT datasets," in *Proceedings SPIE 9034, Medical Imaging 2014: Image Processing*, pp. 903415–903415, International Society for Optics and Photonics, 2014.

[46] B. Elger, J. Iavindrasana, L. Lo Iacono, H. Müller, N. Roduit, P. Summers, and J. Wright, "Strategies for health data exchange for secondary, cross–institutional clinical research," *Computer Methods and Programs in Biomedicine*, vol. 99, pp. 230–251, September 2010.

[47] A. Shimizu, T. Kimoto, H. Kobatake, S. Nawano, and K. Shinozaki, "Automated pancreas segmentation from three–dimensional contrast–enhanced computed tomography," *International Journal of Computer Assisted Radiology and Surgery*, vol. 5, no. 1, pp. 85–98, 2010.

[48] P. Kohlmann, J. Strehlow, B. Jobst, S. Krass, J.-M. Kuhnigk, A. Anjorin, O. Sedlaczek, S. Ley, H.-U. Kauczor, and M. O. Wielpütz, "Automatic lung segmentation method for mri-based lung perfusion studies of patients with chronic obstructive pulmonary disease," *International journal of computer assisted radiology and surgery*, vol. 10, no. 4, pp. 403–417, 2014.

[49] H. T. Huynh, I. Karademir, A. Oto, and K. Suzuki, "Computerized liver volumetry on MRI by using 3D geodesic active contour segmentation," *American Journal of Roentgenology*, vol. 202, no. 1, p. 152, 2014.

[50] A. Alkan, S. A. Tuncer, and M. Gunay, "Comparative MR image analysis for thyroid nodule detection and quantification," *Measurement*, vol. 47, pp. 861–868, 2014.

[51] V. Pekar, T. R. McNutt, and M. R. Kaus, "Automated model-based organ delineation for radiotherapy planning in prostatic region," *International Journal of Radiation Oncology Biology Physics*, vol. 60, no. 3, pp. 973–980, 2004.

[52] C. W. Jeong, H. K. Park, S. K. Hong, S. Byun, H. J. Lee, and S. E. Lee, "Comparison of prostate volume measured by transrectal ultrasonography and mri with the actual prostate volume measured after radical prostatectomy," *Urologia internationalis*, vol. 81, no. 2, pp. 179–185, 2008.

[53] T. R. Langerak, U. Van Der Heide, A. N. Kotte, M. Viergever, M. Van Vulpen, and J. P. Pluim, "Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE)," *Medical Imaging, IEEE Transactions on*, vol. 29, no. 12, pp. 2000–2008, 2010.

[54] M. Krenn, M. Dorfer, O. A. Jiménez-del Toro, H. Müller, B. Menze, M.-A. Weber, A. Hanbury, and G. Langs, "Creating a Large–Scale Silver Corpus from Multiple Algorithmic Segmentations," in *MICCAI Workshop on Medical Computer Vision: Algorithms for Big Data, MCV 2015*, vol. 8848, pp. 163–170, Springer, 2014.

Appendix

Organ segmentation methods

- **DICENTE et al.:** ***K–means clustering and geometric techniques*** *(A3)* The lung segmentation approach by Dicente et al. [30], was evaluated in CT and CTce scans. It starts by segmenting the full respiratory system from the image (trachea and lungs), using mathematical morphology to 'fill' the darker areas inside the body and then selecting the biggest 3D connected dark region with K–Means clustering. The trachea and primary bronchi are removed by defining the farthest 2D connected components from the center of mass of the mask as the lungs and removing the closest unconnected regions below a distance threshold. An algorithm involving a K–nearest neighbor classifier defines the best boundary between right and left lung for regions of conflict in the output mask, where the division between the lungs is unclear. Finally the holes inside the masks are filled using mathematical morphology 'filling'. Both lungs in a single CT volume were segmented in 8 minutes.
- **GASS et al.:** ***Markov random field based Multi–atlas registration*** *(A1, A2)*. A modality and anatomy independent technique was presented by Gass et al. [38] for all the modalities and almost all the available structures in the Anatomy data set. Multiple atlases are registered individually to a target image and the propagated atlas labels are fused at voxel level using a weighted majority vote approach. The deformable registrations are computed by minimizing the registration energy, decomposed into a local similarity metric and prior assumptions over the displacement, through Markov random fields. Control points are displaced in a multi–resolution cubic B–spline framework and ensure diffeomorphic deformations. The quality of the information obtained from each atlas is weighted computing the local normalized cross correlation between the target image and the deformed atlases. The approach took approximately 4 hours and 30 min. to compute all the segmentations in a single volume.
- **HEINRICH et al.:** ***Multi–atlas segmentation with discrete optimisation and self–similarities*** *(A3)*. Heinrich et al. [39] segmented a subset of 7 structures in CTce and MRce. Their discrete deformable registration framework defines a graphical model with control points in a B–spline grid, an appropriate range of displacements and a regularisation parameter. Edges between selected nodes are modelled by a minimum spanning tree reducing computational costs. Robust self–similarity descriptors are based on local patch distances within the image and are used to calculate a dissimilarity metric using the local patch coordinates in the Hamming space. Before deformable registration, a block–matching based linear registration using the same similarity metric is employed. Afterwards, the global minimum is found with belief propagation on the created mimimun spanning tree. The method took around 40 minutes per volume and 20 training atlases.
- **JIMENEZ–DEL–TORO et al.:** ***Hierarchic multi–atlas registration using anatomical correlations*** *(A1, A2, A3)*. All the anatomical structures in CT scans were segmented with the approach of Jiménez–del–Toro et al. [40], [41]. A multi–atlas segmentation hierarchy is pre–defined according to the size and tissue contrast from the various anatomical structures, refining local regions of interest after each registration step. The optimization of the registrations is done with an adaptive stochastic gradient descent approach in a multi–scale framework. After an initial global affine registration of the atlases and target image, local estimations of the bigger structures like liver and lungs are generated fusing the transformed output labels. The image registration is independently optimized per structure with a local affine registration and then through non–rigid B–spline registration. With the affine coordinate transformations of the bigger structures as initialization, smaller and less defined structures like the pancreas and gallbladder are then registered and segmented. Final transformed labels are re–fused performing majority voting. All the anatomical structures were segmented for a single CT volume in 12h.
- **KAHL et al.:** ***RANSAC registration, random forest classifier and graph cut*** *(A3)*. The multi–atlas segmentation method by Kahl et al. [42] segmented all the available structures in CTwb. It segments each organ independently using a three step pipeline: 1. Feature–based registration with a Random sample consensus (RANSAC), 2. Label fusion with a random forest classifier and 3. Graph cut segmentation with a Potts model. After establishing the best transformation for each atlas and each organ in the training set with 8000–10000 sparse features, similar to SIFT, the 300 top features are selected for each organ for the matching with a symmetric neighbour approach. Outliers are removed with standard RANSAC, and optimized before reaching a threshold of 30 mm. Initially an affine transformation is obtained, later refined with thin plate splines for the remaining correspondances, and for some organs with a standard intensity–based method. An average map, a distance map and voxel intensity features are computed with the transferred labels and used to train a random forest classifier to improve the local appearence around the target organ. The final label is a regularized computing graph–cuts that searches for voxels with probabilities >0.5 in a 6–connected neighborhood no farther than 20 voxels from the original segmentation margin. The segmentation of the 20 structures from a single CT volume required around 13h.
- **KÉCHICHIAN et al.:** ***Clustering, Graph Cut and hierarchic atlas registration*** *(A1, A2, A3)*. The generic method proposed by Kéchichian et al. [43] segmented all the 20 structures in CTce images. Their multiorgan segmentation method is based on multilabel Graph Cut optimization and uses location and intensity likelihoods of organs and prior information of their spatial configuration. The spatial prior is derived from shortest-path constraints defined on the adjacency graph of structures, and location likelihoods are defined by probabilistic atlases constructed from the training data set using a (2+1)D rigid

registration method based on SURF keypoints. To create atlases, a representative image is used as a reference onto which remaining images are registered. In addition to structures in the training set, probabilistic atlases for three additional body regions were created from automatically generated annotations: background, thorax and abdomen, and body envelope. All atlases are registered to the target image via the aforementioned method in a hierarchical fashion starting at the full image, then on an intermediate level corresponding to the thorax and abdomen region, and finally on individual organs. Prior to segmentation, the target image is simplified by an image-adaptive centroidal Voronoi tessellation to reduce subsequent optimization time and memory footprint. The multiorgan segmentation is obtained by minimizing an energy function defined according to organ intensity and location likelihood energies, and the energy of the prior distribution of organ spatial configurations. It is optimized via the Expansion Moves multilabel Graph Cut algorithm. The method segmented all the structures in about 2 hours.

- **JIA et al.:** ***Multi–boost learning and Statitical shape model (SSM) search*** *(A1, A2, A3)*. Li et al. [32] and He et al. [33] segmented the liver in the first two Anatomy benchmarks and 6 structures in Anatomy3 in CT volumes. A statistical shape model is created with a KNN classifier that is trained on intensity and organ boundary gradient profiles establishing the local appearance. Organ regions of interest are extracted employing template matching in a top–down order. These locations as well as other image features like intensity are used to train a multi–boost classifier for the various organs in the Anatomy3 version of the algorithm. The output segmentation image is used as the final segmentation of the lung and kidneys and as a reference to compute a distance map to which the SSM is registered. The previously trained kNN–classifier optimizes the landmark displacements iteratively to the their best positions. Finally, the output segmentation boundary points are optimized with a constrained free–form deformation to improve the local specific variations of the organs. The runtime for liver segmentation was around 5 minutes, and for the 6 structures 25 minutes.
- **SPANIER et al.:** ***Rule–based segmentation using region growing*** *(A1, A2)*. Spanier et al. [31] proposed a method that segments seven organs in CT scans through a pipeline of rules that rely on intensity and anatomical location priors. It extends a cognition network liver segmentation method previously proposed. The patient's body is isolated from the background using these priors in the preprocessing step. Then, a four–step approach is implemented both for the breathing system (lungs and trachea) and for the liver, kidneys and spleen. The lungs and trachea are located starting from the top of the scan in the consecutive slices that are contain values below 300 Houndsfield units (HU). A line that passes through the spinal column at 45° defines a ROI to the left for the left kidney and spleen, and a ROI to the right for the liver and right kidney. The largest selected component is selected from the ROI of the breathing system. An intensity range is computed based on the intensities of the heart enclosed between the lungs, to threshold the ROIs of the abdominal organs. For each organ, a 2D slice with the largest axial cross is selected from the thresholded ROI as a 2D–seed. Region growing from the 2D–seed is performed along the axial direction preserving smoothness and curvature constraints between adjacent slices. The computation time for this method ranged from 2.5 to 3 hrs per volume.
- **VINCENT et al.:** ***Active Appearance Models (AAM) and image registration*** *(A2)*. Vincent et al. [34] presented a framework for the segmentation of eight structures in CTwb and CTce images from the VISCERAL Anatomy 2 benchmark. This framework was also evaluated in MR, for bone and soft tissue in the hands, knee, prostate, among others. A Minimum Description Length groupwise image registration method finds correspondences used in building the AAM. The AAMs are matched to the data through multi-–start optimisation. Initially, this scheme fits low density low resolution models but ends in a robust matching of detailed high resolution models. Finally, the voxels contained in the uncertainty region, defined in a halo around the model boundary, are assigned with a non–linear regression function. The training of this function is performed with a PAC–learning method. This was the only method during the VISCERAL benchmarks that produced fuzzy segmentations (intensities in segmentations correspond to a probability of membership). Results were computed both for the fuzzy segmentations as well as for thresholded binary images at 0.5 like the rest of the participating methods. The segmentation of the aorta, kidneys, liver, lungs and psoas major muscles took around 1.5 to 2 hrs per volume.
- **WANG et al.:** ***Hierarchical shape priors for level set segmentation method*** *(A1, A2, A3)*. Wang et al. [35] presented a multi–organ segmentation pipeline for 10 structures in CTwb and CTce. The target images are initially pre–processed removing the skin and subcutaneous fat tissue using threshold–based level set segmentation and mathematical morphology operations. Then, a selected standard subject volume is rigidly registered to the target. The largest torso cross–section area is estimated in CTwb to focus the processing only on this area. All the training organ labels are registered to the standard subject and statistical shape priors are generated. The statistical mean shapes are then registered to a trust zone in the upper–level structure space. The ventral cavity is first segmented and is then divided into thoracic and abdominopelvic cavity. Individual structures are segmented from left to right guided by empirically defined likelihood. An intensity mapping function estimates the intensity range of a few organs through an iterative approach. The model–based level set method is extended with a coherent propagation method that speeds up the propagation and reduces the frequency of the shape–prior registration. The segmentation accuracy is further improved by using shape model guided local phase analysis [36]. The ten anatomical structures were segmented in aproximately 1 hour.

Landmark detection methods

- **WYETH et al.:** ***Classification forests trained at voxel–level*** *(A1)*. Wyeth et al. contributed with a method that detects and localizes anatomical landmarks in unseen volumetric CT data [45]. The images are first aligned in DICOM Patient Coordinates and then downsampled with a Gaussian smoothing. A pool of 15,625 possible features is built using densities in Houndsfield units at chosen random boxes from the neighborhood of each landmark. Single (4mm downsampled) density features showed the best values over other features during the optimization phase. Forty data sets, from 369 data sets available, are randomly selected for training each classification tree. Both landmark neighborhood samples and background samples are weighted as a Gaussian function of distance from the landmark. The voxel with the highest normalized likelihood for a landmark is selected and a sub–voxel result is obtained with Brent interpolation. This method overcomes the problem of measuring voxels outside the dataset or close to the edge by assigning the sample 50/50 to each branch, in both training and detection. An optimized trade–off was selected between accuracy, detection AUC and runtime, which was 0.5–2 minutes per volume.
- **GASS et al.:** ***Multi–atlas template–matching with consensus–based fusion*** **(A1, A2)**. Gass et al. fused multiple patch location estimates to perform landmark detection in three modalities: CTwb, MRwb and MRce [38]. The templates created for each landmark are localized individually in the training set atlases using a box–shaped image region with an empirically set half–width of 20 mm around the landmark coordinates. A large search region is defined in the normalized voxel coordinates of the target image, selecting approximately the same physically isotropic region as in the training atlases. The atlas templates are then compared to the candidate locations through convolution, computing two similarity metrics: sum of squared differences and normalized cross–correlation. All location estimates are fused finding the median location coordinates. The runtime of their landmark detection approach was 30 min. for a single volume.
- **MAI et al.:** ***Histograms of Oriented Gradients descriptor for landmark detection*** *(A2)*. Mai et al. generated a single discriminative detection filter per anatomical landmark using features from a variant of the histogram of oriented gradients (HOG) [44]. Positive sample patches from the 3D volume surrounding the landmark location were rigidly aligned to normalize the location. Random sampling of negative examples that do not contain the sought landmark were also included in the kernel matrix in a 5:1 proportion to the positive samples. The 3D HOG descriptors included 20 orientation bins that define a direction histogram binning function of equally distributed units. A linear support vector machine is trained with the selected samples for each landmark to cope with the high dimensionality of the data. Detections were computed using a sliding window approach which was efficiently computed as a convolution operation in Fourier space. The runtime per volume was 6–8 minutes in the VISCERAL Anatomy2 VMs.

Runtime per segmentation

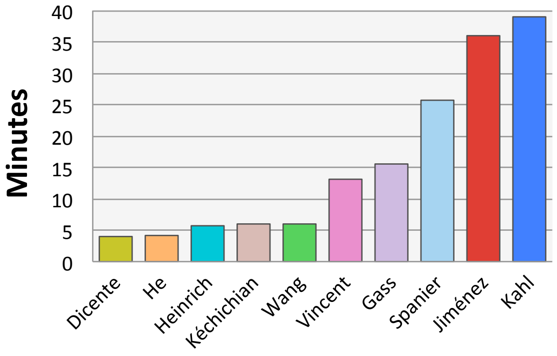


Fig. 9: Chart showing approximate runtime per structure segmented.

Runtime per landmark

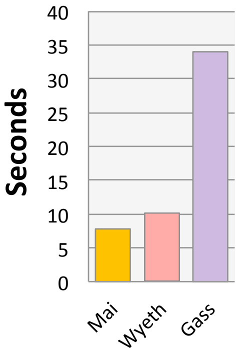


Fig. 10: Chart showing approximate runtime per landmark detected.