CrossMoDA 2021 challenge: Benchmark of Cross-Modality Domain Adaptation techniques for Vestibular Schwannoma and Cochlea Segmentation

Reuben Dorent[a,*], Aaron Kujawa[a], Marina Ivory[a], Spyridon Bakas[b,c,d], Nicola Rieke[e], Samuel Joutard[a], Ben Glocker[f], Jorge Cardoso[a], Marc Modat[a], Kayhan Batmanghelich[n], Arseniy Belkov[u], Maria Baldeon Calisto[r], Jae Won Choi[i], Benoit M. Dawant[j], Hexin Dong[h], Sergio Escalera[p], Yubo Fan[j], Lasse Hansen[q], Mattias P. Heinrich[q], Smriti Joshi[p], Victoriya Kashtanova[m], Hyeon Gyu Kim[g], Satoshi Kondo[t], Christian N. Kruse[q], Susana K. Lai-Yuen[s], Hao Li[j], Han Liu[j], Buntheng Ly[m], Ipek Oguz[j], Hyungseob Shin[g], Boris Shirokikh[v,w], Zixian Su[k,l], Guotai Wang[o], Jianghao Wu[o], Yanwu Xu[n], Kai Yao[k,l], Li Zhang[h], Sébastien Ourselin[a], Jonathan Shapey[a,x], Tom Vercauteren[a]

[a]School of Biomedical Engineering & Imaging Sciences, King's College London, London, United Kingdom
[b]Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, USA
[c]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[d]Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[e]NVIDIA
[f]Department of Computing, Imperial College London, Department of Computing, London, United Kingdom
[g]School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
[h]Center for Data Science, Peking University, Beijing, China
[i]Department of Radiology, Armed Forces Yangju Hospital, Yangju, Korea
[j]Vanderbilt University, Nashville, USA
[k]University of Liverpool, Liverpool, United Kingdom
[l]School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China
[m]Inria, Université Côte d'Azur, Sophia Antipolis, France
[n]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, USA
[o]School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China
[p]Artificial Intelligence in Medicine Lab (BCN-AIM) and Human Behavior Analysis Lab (HuPBA), Universitat de Barcelona, Barcelona, Spain
[q]Institute of Medical Informatics, Universität zu Lübeck, Germany
[r]Universidad San Francisco de Quito, Quito, Ecuador
[s]University of South Florida, Tampa, USA
[t]Muroran Institute of Technology, Muroran, Japan
[u]Moscow Institute of Physics and Technology, Moscow, Russia
[v]Skolkovo Institute of Science and Technology, Moscow, Russia
[w]Artificial Intelligence Research Institute (AIRI), Moscow, Russia
[x]Department of Neurosurgery, King's College Hospital, London, United Kingdom

ARTICLE INFO

ABSTRACT

Domain Adaptation (DA) has recently been of strong interest in the medical imaging community. While a large variety of DA techniques have been proposed for image segmentation, most of these techniques have been validated either on private datasets or on small publicly available datasets. Moreover, these datasets mostly addressed single-class problems.

To tackle these limitations, the Cross-Modality Domain Adaptation (crossMoDA) challenge was organised in conjunction with the 24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2021). Cross-MoDA is the first large and multi-class benchmark for unsupervised cross-modality Domain Adaptation. The goal of the challenge is to segment two key brain structures involved in the follow-up and treatment planning of vestibular schwannoma (VS): the VS and the cochleas. Currently, the diagnosis and surveillance in patients with VS are commonly performed using contrast-enhanced T1 (ceT_1) MR imaging. However, there is growing interest in using non-contrast imaging sequences such as high-resolution T2 (hrT_2) imaging. For this reason, we established an unsupervised cross-modality segmentation benchmark. The training dataset provides annotated ceT_1 scans (N=105) and unpaired non-annotated hrT_2 scans (N=105). The aim was to automatically perform unilateral VS and bilateral cochlea segmentation on hrT_2 scans as provided in the

---

*Corresponding author
e-mail:* `reuben.dorent@kcl.ac.uk` (Reuben Dorent)

testing set (N=137). This problem is particularly challenging given the large intensity distribution gap across the modalities and the small volume of the structures.

A total of 55 teams from 16 countries submitted predictions to the validation leaderboard. Among them, 16 teams from 9 different countries submitted their algorithm for the evaluation phase. The level of performance reached by the top-performing teams is strikingly high (best median Dice score - VS: 88.4%; Cochleas: 85.7%) and close to full supervision (median Dice score - VS: 92.5%; Cochleas: 87.7%). All top-performing methods made use of an image-to-image translation approach to transform the source-domain images into pseudo-target-domain images. A segmentation network was then trained using these generated images and the manual annotations provided for the source image.

1. Introduction

Machine learning (ML) has recently reached outstanding performance in medical image analysis. These techniques typically assume that the training dataset (source domain) and test dataset (target domain) are drawn from the same data distribution. However, this assumption does not always stand in clinical practice. For example, the data may have been acquired at different medical centres, with different scanners, and under different image acquisition protocols. Recent studies have shown that ML algorithms, including deep learning ones, are particularly sensitive to data changes and experience performance drops due to domain shifts (van Opbroek et al., 2015; Donahue et al., 2014). This domain shift problem strongly reduces the applicability of ML approaches to real-world clinical settings.

To increase the robustness of ML techniques, a naive approach aims at training models on large-scale datasets that cover large data variability. Therefore, efforts have been made in the computer vision community to collect and annotate data. For example, the Open Images dataset (Kuznetsova et al., 2020) contains 9 million varied images with rich annotations. While natural images can be easily collected from the Internet, access to medical data is often restricted to preserve medical privacy. Moreover, annotating medical images is time-consuming and expensive as it requires the expertise of physicians, radiologists, and surgeons. For these reasons, it is unlikely that large, annotated and open databases will become available for most medical problems.

To address the lack of large amounts of labelled medical data, domain adaptation (DA) has been of strong interest in the medical imaging community. DA is a subcategory of transfer learning that aims at bridging the domain distribution discrepancy between the source domain and the target domain. While the source and target data are assumed to be available at training time, target label availability is either limited (supervised and semi-supervised DA), incomplete (weakly-supervised DA) or missing (unsupervised DA). A complete review of DA for medical image analysis can be found in Guan and Liu (2021). Unsupervised DA (UDA) has especially raised attention as it doesn't require any additional annotations. However, existing UDA techniques have been either tested on private, small or single class datasets. Consequently, there is a need for a public
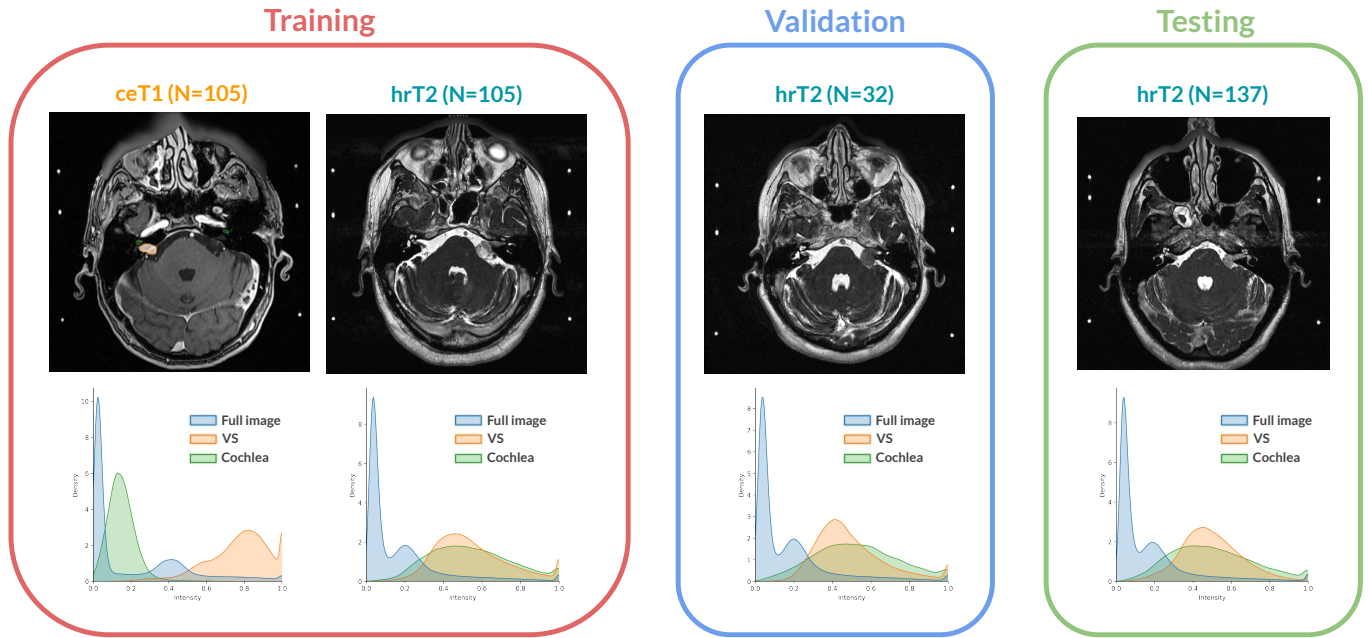
benchmark on a large and multi-class dataset.

To benchmark new and existing unsupervised DA techniques for medical image segmentation, we organised the crossMoDA challenge in conjunction with the 24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2021). The goal of the challenge was to segment two key brain structures involved in the follow-up and treatment planning of vestibular schwannoma (VS): the VS and the cochleas. With data from 379 patients, crossMoDA is the first large and multi-class benchmark for unsupervised cross-modality domain adaptation.

VS is a benign tumour arising from the nerve sheath of one of the vestibular nerves. The incidence of VS has been estimated to be 1 in 1000 (Evans et al., 2005). For smaller tumours, observation using MR imaging is often advised. If the tumour demonstrates growth, management options include conventional open surgery or stereotactic radiosurgery (SRS), which requires the segmentation of VS and the surrounding organs at risk (e.g., the cochlea) (Shapey et al., 2021a). The tumour's maximal linear dimension is typically measured to estimate the tumour growth. However, recent studies (MacKeith et al., 2018; Varughese et al., 2012) have demonstrated that a volumetric measurement is a more accurate and sensitive method of calculating a VS's true size and is superior at detecting subtle growth. For these reasons, automated methods for VS delineation have been recently proposed (Wang et al., 2019; Shapey et al., 2019; Lee et al., 2021; Dorent et al., 2021).

Currently, the diagnosis and surveillance of patients with VS are commonly performed using contrast-enhanced T1 (ceT$_1$) MR imaging. However, there is growing interest in using non-contrast imaging sequences such as high-resolution T2 (hrT$_2$) imaging, as it mitigates the risks associated with gadolinium-containing contrast agents (Khawaja et al., 2015). In addition to improving patient safety, hrT$_2$ imaging is 10 times more cost-efficient than ceT$_1$ imaging (Coelho et al., 2018). For this reason, we proposed a cross-modality benchmark (from ceT$_1$ to hrT$_2$) that aims to automatically perform VS and cochleas segmentation on hrT$_2$ scans.

This paper summarises the 2021 challenge and is structured as follows. First, a review of existing datasets used to assess existing domain adaptation techniques for image segmentation

Fig. 1: Overview of the challenge dataset. Annotations are only available for the training ceT_1 scans. Intensity distribution on each set are shown per structure. The intensity is normalised between 0 and 1 for each volume.

is proposed in Section 2. Then, the design of the crossMoDA challenge is given in Section 3. Section 4 presents the evaluation strategy of the challenge (metrics and ranking scheme). Participating methods are then described and compared in Section 5. Finally, Section 6 presents the results obtained by the participating team and Section 7 provides a discussion and concludes the paper.

2. Related work

We performed a literature review to survey the benchmark datasets used to assess DA techniques for unsupervised medical image segmentation. On the methodological side, as detailed afterwards, a range of methods was used by the participating teams, illustrating a wide breadth of different DA approaches. Nonetheless, a thorough review of DA methodologies is out of the scope of this paper. We refer the interested reader to Guan and Liu (2021) for a recent review of these.

Many domain adaptation techniques for medical image segmentation have been validated on private datasets, for example Kamnitsas et al. (2017); Yang et al. (2019). Given that these datasets used for the experiments are not publicly available, it is not possible to compare new methods with these techniques.

Other authors have used public datasets to validate their methods. Interestingly, these datasets often come from previous medical segmentation challenges that weren't originally proposed for domain adaptation. For this reason, unsupervised problems are generated by artificially removing annotations on subsets of these challenge datasets. We present these open datasets and highlight their limitations for evaluating unsupervised domain adaptation:

- *WMH*: The MICCAI White Matter Hyperintensities (WMH) Challenge dataset (Kuijf et al., 2019) consists of

brain MR images with manual annotations of WMH from three different institutions. Each institution provided 20 multi-modal images for the training set. Domain adaptation techniques have been validated on each set of scans acquired at the same institution (Orbes-Arteaga et al., 2019; Palladino et al., 2020; Sundaresan et al., 2021). Each institution set ($N = 20$) is not only used to assess the methods but also to perform domain adaptation during training. Consequently, the test sets are extremely small ($N \leq 10$ scans), leading to comparisons with low statistical power. Another limitation of this dataset is that it only assesses single-class UDA solutions. Finally, the domain shift is limited as the source and target domains correspond to the same image modalities acquired with 3T MRI scanners.

- *SCGM*: The Spinal Cord Gray Matter Challenge (SCGM) dataset is a collection of cervical MRI from four institutions (Prados et al., 2017). Each site provided unimodal images from 20 healthy subjects along with manual segmentation masks. Various unsupervised domain adaptation techniques have been tested on this dataset (Perone et al., 2019; Liu et al., 2021b; Shanis et al., 2019). Again, the main limitation of this dataset is the small size of the test sets ($N \leq 20$ scans). Moreover, the problem is single-class, and the domain shift is limited (intra-modality UDA).

- *IVDM3Seg*: The Automatic Intervertebral Disc Localization and Segmentation from 3D Multi-modality MR Images (IVDM3Seg) is a collection of 16 manually annotated 3D multi-modal MR scans of the lower spine. Domain adaptation techniques were validated on this dataset (Bateson et al., 2019, 2020). The test set is extremely small ($N = 4$), and it is a single-class segmentation task.

- *MM-WHS*: The Multi-Modality Whole Heart Segmentation (MM-WHS) Challenge 2017 dataset (Zhuang et al., 2019) is a collection of MRI and CT volumes for cardiac segmentation. Specifically, the training data consist of 20 MRI and 20 unpaired CT volumes with ground truth masks. This dataset has been used to benchmark most multi-classes cross-modality domain adaptation techniques (Dou et al., 2018; Ouyang et al., 2019; Cui et al., 2021; Zou et al., 2020). While the task is challenging, the very limited size of the test set ($N = 4$) strongly reduced the statistical power of comparisons.

- *CHAOS*: The Combined (CT-MR) Healthy Abdominal Organ Segmentation (CHAOS) dataset (Kavur et al., 2021) corresponds to 20 MR volumes and 30 unpaired CT volumes. Cross-modality domain adaptation techniques have been tested on this dataset (Chen et al., 2020a; Jiang and Veeraraghavan, 2020). 4 and 6 scans are respectively used as test sets for the MR and CT domains. Consequently, the test sets are particularly small.

- *BraTS*: The Brain Tumor Segmentation (BraTS) benchmark (Menze et al., 2014; Bakas et al., 2017, 2019) is a popular dataset for the segmentation of brain tumour sub-regions. While images were collected from a large number of medical institutions with different imaging parameters, the origin of the imaging data is not specified for each case. Instead, unsupervised pathology domain adaptation (high to low grades) has been tested on this dataset (Shanis et al., 2019), which is a different problem than ours. Alternatively, BraTS has been used for cross-modality domain adaptation (Zou et al., 2020). However, the problem is artificially generated by removing image modalities and has limited clinical relevance.

In conclusion, test sets used to assess segmentation methods for unsupervised domain adaptation are either private, small or single-class.

3. Challenge description

3.1. Overview

The goal of the crossMoDA challenge was to benchmark new and existing unsupervised cross-modality domain adaptation techniques for medical image segmentation. The proposed segmentation task focused on two key brain structures involved in the follow-up and treatment planning of vestibular schwannoma (VS): the tumour and the cochleas. Participants were invited to submit algorithms designed for inference on high-resolution T2 (hrT_2) scans. Participants had access to a training set of high-resolution T2 scans without their manual annotations. Conversely, manual annotations were provided for an unpaired training set of contrast-enhanced T1 (ceT_1) scans. Consequently, the participants had to perform unsupervised cross-modality domain adaptation from ceT_1 (source) to hrT_2 (target) scans.

3.2. Data description

3.2.1. Data overview

The dataset for the crossMoDA challenge is an extension of the publicly available Vestibular-Schwannoma-SEG collection released on The Cancer Imaging Archive (TCIA) (Shapey et al., 2021b; Clark et al., 2013). To ensure that no data in the test set was accessible to the participants, no publicly available scan was included in the test set. The open Vestibular-Schwannoma-SEG dataset was used for training and validation, while an extension was kept private and used as the test set.

The complete crossMoDA dataset (training, validation and testing) contained a set of MR images collected on 379 consecutive patients (Male:Female 166:214; median age: 56 yr, range: 24 yr to 84 yr) with a single sporadic VS treated with Gamma Knife stereotactic radiosurgery (GK SRS) between 2012 and 2021 at a single institution. For each patient, contrast-enhanced T1-weighted (ceT_1) and high-resolution T2-weighted (hrT_2) scans were acquired in a single MRI session prior to and typically on the day of the radiosurgery. 75 patients had previously undergone surgery. Data were obtained from the Queen Square Radiosurgery Centre (Gamma Knife). All contributions to this study were based on approval by the NHS Health Research Authority and Research Ethics Committee (18/LO/0532) and were conducted in accordance with the 1964 Declaration of Helsinki.

The scans acquired between October 2012 and December 2017 correspond to the publicly available Vestibular-Schwannoma-SEG dataset on TCIA (242 patients). The Vestibular-Schwannoma-SEG dataset was randomly split into three sets: the source training set (105 annotated ceT_1 scans), the target training set (105 non-annotated hrT_2 scans) and the target validation set (32 non-annotated hrT_2 scans).

The scans acquired between January 2018 and March 2021 were used to make up the test set (137 non-annotated hrT_2 scans). The test set remained private to the challenge participants and accessible only to the challenge organisers, even during the evaluation phase.

As shown in Table 1, the target training, validation, and test sets have a similar distribution of features (age, gender and operative status of patients; slice thickness and in-plane resolution of hrT_2).

3.2.2. Image acquisition

All images were obtained on a 32-channel Siemens Avanto 1.5T scanner using a Siemens single-channel head coil. Contrast-enhanced T1-weighted imaging was performed with an MP-RAGE sequence (in-plane resolution=0.47×0.47mm, matrix size=512×512, TR=1900ms, TE=2.97ms) and slice thickness of 1.0-1.5mm. High-resolution T2-weighted imaging was performed with either a Constructive Interference Steady State (CISS) sequence (in-plane resolution=0.47×0.47mm, matrix size=448×448, TR=9.4ms, TE=4.23ms) or a Turbo Spin Echo (TSE) sequence (in-plane resolution=0.55×0.55mm, matrix size=384×384, TR=750ms, TE=121ms) and slice thickness of 1.0-1.5mm. The details of the dataset are given in Table 1, and sample cases from the source and target sets are illustrated in Fig 1.

Table 1: Summary of data characteristics of the crossMoDA sets

	Training			Validation		Test	
	Source	Target		Target		Target	
Sequence	MP-RAGE ceT$_1$	CISS hrT$_2$	TSE hrT$_2$	CISS hrT$_2$	TSE hrT$_2$	CISS hrT$_2$	TSE hrT$_2$
Number of scans	105	83	22	28	4	132	5
Number of patients	105	83	22	28	4	132	5
Available annotations	VS + Cochleas	×	×	×	×	×	×
In-plane matrix	512 × 512	448 × 448	384 × 384	448 × 448	384 × 384	448 × 448 (96%) 512 × 512 (4%)	384 × 384 (60%) 512 × 512 (40%)
Average axial slice number	123 ± 11	80 ± 1	39 ± 4	80 ± 0	40 ± 0	80 ± 4	35 ± 8
In-plane resolution in mm	0.41 × 0.41	0.46 × 0.46	0.55 × 0.55	0.46 × 0.46	0.55 × 0.55	0.46 × 0.46	0.55 × 0.55
Slice thickness in mm	1.0 (7%) 1.5 (93%)	1.0 (10%) 1.5 (90%)	1.5	1.0 (7%) 1.5 (93%)	1.5	1.0 (7%) 1.5 (93%)	1.5
Male:Female	44% : 56%	36% : 64%		34% : 66%		51% : 49%	
Post-operative cases	26%	20%		16%		15%	
Age in years - Median [Q1-Q3]	54 [47-66]	56 [44-64]		58 [51-66]		56 [45-66]	

3.2.3. Annotation protocol

All imaging datasets were manually segmented following the same annotation protocol.

The tumour volume (VS) was manually segmented by the treating neurosurgeon and physicist using both the ceT$_1$ and hrT$_2$ images. All VS segmentations were performed using the Leksell GammaPlan software that employs an in-plane semi-automated segmentation method. Using this software, delineation was performed on sequential 2D axial slices to produce 3D models for each structure.

The adjacent cochlea (hearing organ) is the main organ at risk during VS radiosurgery. In the crossMoDA dataset, patients have a single sporadic VS. Consequently, only one cochlea per patient - the closest one to the tumour - was initially segmented by the treating neurosurgeon and physicist. Preliminary results using a fully-supervised approach (Isensee et al., 2021) showed that considering the remaining cochlea as part of the background leads to poor performance for cochlea segmentation. Given that tackling this challenging issue is beyond the scope of the challenge, both cochleas were manually segmented by radiology fellows with over 3 years of clinical experience in general radiology using the ITK-SNAP software (Yushkevich et al., 2019). hrT$_2$ images were used as reference for cochlea segmentation. The basal turn with osseous spiral lamina was included in the annotation of every cochlea to keep manual labels consistent. In addition, modiolus, a small low-intensity area (on hrT$_2$) within the centre of the cochlea, was included in the segmentation as well.

3.2.4. Data curation

The data was fully de-identified by removing all health information identifiers and defaced (Milchenko and Marcus, 2013). Details can be found in Shapey et al. (2021b). Since the data was acquired consistently (similar voxel spacing, same scanner), no further image pre-processing was employed. Planar contour lines (DICOM RT-objects) of the VS were converted into label maps using SlicerRT (Pinter et al., 2012).

Images and segmentation masks were distributed as compressed NIfTI files (.nii.gz). The training and validation data was made available on zenodo[1]. As we expect this dataset to be used for other purposes in addition to cross-modality domain adaptation, the data was released under a permissive copyright-license (CC-BY-4.0), allowing for data to be shared, distributed and improved upon.

3.3. Challenge setup

The validation phase was hosted on Grand Challenge[2], a well-established challenge platform, allowing for automated validation leaderboard management. Participant submissions are automatically evaluated using the `evalutils`[3] and `MedPy`[4] Python packages. To mitigate the risk that participants select their model hyper-parameters in a supervised manner, i.e. by computing the prediction accuracy, only one submission per day was allowed on the validation leaderboard. The validation phase was held between the 5th of May 2021 and the 15th of August 2021.

Following the best practice guidelines for challenge organisation (Maier-Hein et al., 2020), the test set remained private to reduce the risk of cheating. Participants had to containerise their methods with Docker following guidelines[5] and submit their Docker container for evaluation on the test set. Only one submission was allowed. Docker containers were run on a Ubuntu (20.04) desktop with 64GB RAM, an Intel Xeon CPU E5-1650 v3 and an NVIDIA TITAN X GPU with 12 GB memory. To test the quality of the predictions performed on the local machine, predictions on the validation set were computed and compared with the ones obtained using participants' machines. *In fine*, all participant containers passed the quality control test.

4. Metrics and evaluation

The choice of the metrics used to assess the performance of the participants' algorithm and the ranking strategy are keys for

---

[1] https://zenodo.org/record/4662239
[2] https://grand-challenge.org/
[3] https://evalutils.readthedocs.io/en/latest/
[4] https://loli.github.io/medpy/
[5] https://crossmoda.grand-challenge.org/submission/

adequate interpretation and reproducibility of results (Maier-Hein et al., 2018). In this section, we follow the BIAS best practice recommendations for assessing challenges (Maier-Hein et al., 2020)

4.1. Choice of the metrics

The algorithms' main property to be optimised is the accuracy of the predictions. As relying on a single metric for the assessment of segmentations leads to less robust rankings, two metrics were chosen: the Dice similarity coefficient (DSC) and the Average symmetric surface distance (ASSD). DSC and ASSD have frequently been used in previous challenges (Kavur et al., 2021; Antonelli et al., 2021) because of their simplicity, their rank stability and their ability to assess segmentation accuracy.

Let S_k be the predicted binary segmentation mask of the region k where $k \in \{VS, Cochleas\}$. Let G_k be the manual segmentation of the region k. The Dice Score coefficient quantifies the similarity of two masks S_k and G_k by normalising the size of their intersection over the average of their sizes:

$$\mathrm{DSC}(S_k, G_k) = \frac{2 \sum_i S_{k,i} G_{k,i}}{\sum_i S_{k,i} + \sum_i G_{k,i}} \qquad (1)$$

Let B_{S_k} and B_{G_k} be the boundaries of the segmentation mask S_k and the manual segmentation G_k. The average symmetric surface distance (ASSD) is the average of all the Euclidean distances (in mm) from points on the boundary B_{S_k} to the boundary B_{G_k} and from points on the boundary of B_{G_k} to the boundary B_{S_k}:

$$\mathrm{ASSD}(S_k, G_k) = \frac{\sum_{s_i \in B_{S_k}} d(s_i, B_{G_k}) + \sum_{s_i \in B_{G_k}} d(s_i, B_{S_k})}{|B_{S_k}| + |B_{G_k}|} \qquad (2)$$

where d is the Euclidean distance.

Note that if predictions only contain background, i.e. for all voxels i, $S_{k,i} = 0$, then the ASSD is set as the maximal distance between voxels in the test set (350mm).

4.2. Ranking scheme

We used a standard ranking scheme that has previously been employed in other challenges with satisfactory results, such as the BraTS (Bakas et al., 2019) and the ISLES (Maier et al., 2017) challenges. Participating teams are ranked for each testing case, for each evaluated region (i.e., VS and cochleas), and for each measure (i.e., DSC and ASSD). The lowest rank of tied values is used for ties (equal scores for different teams). Rank scores are then calculated by firstly averaging across all these individual rankings for each case (i.e., cumulative rank) and then averaging these cumulative ranks across all patients for each participating team. Finally, teams are ranked based on their rank score. This ranking scheme was defined, released prior to the start of the challenge, and available on the dedicated Grand Challenge page[6] and the crossMoDA website[7].

---

[6] https://crossmoda.grand-challenge.org/
[7] https://crossmoda-challenge.ml/

To analyse the stability of the ranking scheme, we employed the bootstrapping method detailed in Wiesenfarth et al. (2021). One bootstrap sample consists of N=137 test cases randomly drawn with replacement from the test set of size N=137. On average, 63% of distinct cases are retained in a bootstrap sample. A total of 1,000 of these bootstrap samples were drawn, and the proposed ranking scheme was applied to each bootstrap sample. The original ranking computed on the test set was then pairwise compared to the rankings based on the individual bootstrap samples. The correlation between these pairs of rankings was computed using Kendall's τ, which provides values between -1 (for reverse ranking order) and 1 (for identical ranking order).

5. Participating methods

A total of 341 teams registered to the challenge, allowing them to download the data. 55 teams from 16 different countries submitted predictions to the validation leaderboard. Among them, 16 teams from 9 different countries submitted their containerised algorithm for the evaluation phase.

In this section, we provide a summary of the methods used by these 16 teams. Each method is assigned a unique colour code used in the tables and figures. Brief comparisons of the proposed techniques in terms of methodology and implementation details (training strategy, pre-, post-processing, data augmentation) are presented in Table 2.

To bridge the domain gap between the source and target images, proposed techniques can be categorised into three groups that use:

1. Image-to-image translation approaches such as CycleGAN and its extensions to transform ceT_1 scans into pseudo-hrT_2 scans (● ● ● ● ● ● ● ● ● ● ●) or hrT_2 scans into pseudo-ceT_1 scans (●). Then, one or multiple segmentation networks are trained on the pseudo-scans using the manual annotations and used to perform image segmentation on the target images.

2. MIND cross-modal features (Heinrich et al., 2012) to translate the target and source images in a modality-agnostic feature space. These features are either used to propagate labels using image registration (●) or to train an ensemble of segmentation networks (●).

3. discrepancy measurements either based on discriminative losses (● ●) or minimal-entropy correlation (●) to align features extracted from the source and target images.

Each of the methods is now succinctly described with reference to a corresponding paper whenever available.

● ***Samoyed (1st place, Shin et al.).*** The proposed model is based on target-aware domain translation and self-training (Shin et al., 2022). Labelled ceT_1 scans are first converted to pseudo-hrT_2 scans using a modified version of CycleGAN (Zhu et al., 2017), where an additional decoder is attached to the shared encoder to perform vestibular schwannoma and cochleas segmentation simultaneously with domain conversion, thereby preserving the shape of vestibular schwannoma and cochleas

Table 2: Metrics values and corresponding scores of submission. Median and interquartile values are presented. The best results are given in bold. Arrows indicate favourable direction of each metric.

	Methodology						Training Strategy (Segmentation network)				Inference Strategy	
	Segmentation network	Feature alignment	Cross-modal descriptors	Image-to-image translation	Self-training	Cropping	Data Augmentation	Loss Function(s)	Optimisation	Pre-processing voxel size (mm)	Ensembling	Post-processing
Samoyed	3D and 2D nnU-Net	×	×	CycleGAN w. segm. decoder	✓	Fixed size	nnU-Net augm.[1]	Dice + C-E	SGD Batch:2	nnU-Net pre[2]	5× 3D nnU-Net 5× 2D nnU-Net	VS: Largest CC
PKU_BIALAB	3D nnU-Net	×	×	NiceGAN (2D)	✓	Fixed size	nnU-Net augm.[1]	Dice + C-E	SGD Batch:4	nnU-Net pre[2]	5× 3D nnU-Net	VS: Largest CC
jwc-rad	3D nnU-Net	×	×	CUT (2D)	×	Fixed size	nnU-Net augm.[1] VS: intensity augm.	Dice + C-E	SGD Batch: 2	nnU-Net pre[2] 0.6×0.6×1.0	5× 3D nnU-Net	VS: Largest CC
MIP	2.5D U-Net w. Attention	×	×	CycleGAN (2D+3D) CUT (3D)	×	Manual ROI+ rigid registration	Int. shift, contrast, affine def.	Dice + C-E	Adam Batch: 1	[0,1] norm. 0.46×0.46×1.5	3× 2.5D U-Net	VS: largest CC Coch.: 2 largest CC
PremiLab	DAR-U-Net	Content-Style	×	Content-Style GAN (3D)	×	×	Affine+Elastic deformation	Dice + Focal	Adabelief Batch: 2	[−1,−1] norm. 0.41×0.41×1.5	Fusing: clean-lab	VS: Largest CC
Epione-Liryc	3D U-Net	×	×	CycleGAN w. Pair-Loss	×	MNI registration label-based ROI	Flipping, rotation, Int. Noise	Dice	Adam Batch: 1	[0,1] norm. 0.5×0.5×0.5	×	K-means & mean-shift + CRF
MedICL	2.5D U-Net 3D CNN	×	×	CycleGAN (2D)	×	MNI registration label-based ROI	Affine+Elastic Deformation + IT[3]	Dice	Adam Batch: 2	[0,1] norm. 0.38×0.45×1.5	2× 2.5D U-Net 2×3D CNN	VS: Largest CC
DBMI_pitt	3D U-Net w. attention	×	×	CUT (3D)	×	×	Int. shift, resizing, affine def.	Attention Dice	Adam Batch:4	[0,4000] norm 1.0×1.0×1.0	×	VS: Largest CC Hole filling
Hi-Lib	2.5D U-Net	×	×	CycleGAN (2D) CUT (2D)	×	Label-based ROI	Flipping	Dice	Adam Batch:4	[−1,1] norm (X-Y) 0.41×0.41	×	Small CC removal + CRF
smriti161096	2D nnU-Net	×	×	CycleGAN (2D)	✓	Label-based ROI	nnU-Net augm.[1]	Dice + C-E	Adam Batch:51	Resizing (X-Y): 192×224	5× 2D nnU-Net	VS: Largest CC
IMI	3D nnU-Net	Registration-based	MIND	×	×	Fixed size	nnU-Net augm.[1]	Dice + C-E	Adam Batch:1	nnU-Net pre[2]	5× 3D nnU-Net	CRF
GapMIND	3D DeepLab w. MobileNetV2	×	MIND	×	×	Half size (X-Y)	Int. noise, affine def.	Weighted C-E	Adam Batch:1	z-score norm (X-Y) 0.5×0.5	×	×
gabybaldeon	2D U-Net	Adversarial loss on outputs	×	CycleGAN (2D)	×	×	Flipping, affine+ elastic def.	Dice + C-E	Adam Batch: 1	[0,1] norm. 0.46×0.46×1.5	×	×
SEU_Chen	3D nnU-Net	Entropy	×	×	×	×	nnU-Net augm.[1]	Weighted C-E	Nesterov Batch: 2	nnU-Net pre[2]	2× 3D nnU-Net (source + adapted)	×
skip	3D E-Net per-structure	Gradient Reversal Layer	×	×	×	×	×	Dice + C-E	Adam Batch: 4+16	z-score norm 0.5×0.5×0.5	×	×
IRA	3D U-Net	Gradient Reversal Layer	×	×	×	Fixed size	Int. Gamma-transform	Dice + C-E	Adam Batch: 2	[0,1] norm. 0.5×0.5×0.5	×	×

[1]: Rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma correction and mirroring.

[2]: Cropping: cropped the background regions of the images so that the images could fit the brains. Resampling: In-plane with third-order spline, out-of-plane with nearest neighbour Intensity normalisation: z-score normalisation.

[3] Intensity Augmentation (IT): random apply multi-channel Contrast Limited Adaptive Equalization (mCLAHE) and gamma correction, Gaussian blur, Gaussian noise and image sharpening.

def.: deformations; augm.: augmentation; Int.: intensity; C-E: cross-entropy; CC: connected component.

in the generated pseudo-hrT_2 scans. Next, self-training is employed, which consists of 1) training segmentation with labelled pseudo-hrT_2 scans, 2) inferring pseudo-labels on unlabelled real hrT_2 scans by using the trained model, and 3) retraining segmentation with the combined data of labelled pseudo-hrT_2 scans and pseudo-labelled real hrT_2 scans. For self-training, nnU-Net (Isensee et al., 2021) is used as the backbone segmentation model. 2D and 3D models are ensembled, and all-but-largest-component-suppression is applied to vestibular schwannoma.

● *PKU_BIALAB (2nd place, Dong et al.)*. Dong et al. (2021) proposed an unsupervised cross-modality domain adaptation approach based on pixel alignment and self-training (PAST). During training, pixel alignment is applied to transfer ceT_1 scans to hrT_2 modality to alleviate the domain shift. The synthesised hrT_2 scans are then used to train a segmentation model with supervised learning. To fit the distribution of hrT_2 scans, self-training (Yu et al., 2021) is applied to adapt the decision boundary of the segmentation network. The model in the pixel alignment stage relies on NiceGAN (Chen et al., 2020b) (i.e., an extension method of CycleGAN), which improves the efficiency and effectiveness of training by reusing discriminators for encoding. For 3D segmentation, the nnU-Net (Isensee et al., 2021) framework is used with the default 3D full resolution configuration.

● *jwc-rad (3rd place, Choi)*. The proposed method (Choi, 2021) is based on out-of-the-box deep learning frameworks for unpaired image translation and image segmentation. For domain adaptation, CUT (Park et al., 2020), a model for unpaired image-to-image translation based on patch-wise contrastive learning and adversarial learning, is used. CUT was implemented using the default configurations of the framework except that no resizing or cropping was performed, and the number of epochs with the initial learning rate and the number of epochs with decaying learning rate were both set to 25. For the segmentation task, nnU-Net (Isensee et al., 2021) is used with the default 3D full resolution configuration of the framework, except that the total number of epochs for training is set to 250. Data augmentation for the segmentation task is performed by generating additional training data with lower tumour signals by reducing the signal intensity of the labelled vestibular schwannomas by 50%.

● *MIP (4th place, Liu et al.)*. This team proposed to minimise the domain divergence by image-level domain alignment (Liu et al., 2021a). The target domain pseudo-images are synthesised and used to train a segmentation model with source domain labels. Three image translation models are trained, including 2D/3D CycleGANs and a 3D Contrastive Unpaired Translation (CUT) model. The segmentation backbone follows the same architecture proposed in Wang et al. (2019). To improve the segmentation performance, the segmentation model is fine-tuned using both labelled pseudo-T2 images and unlabelled real T2 images via a semi-supervised learning method called Mean Teacher (Tarvainen and Valpola, 2017). Lastly, the

predictions from three models are fused by a noisy label correction method named CLEAN-LAB (Northcutt et al., 2021). Specifically, the softmax output from one model is converted to a one-hot encoded mask, which is considered to be a "noisy label" and corrected by the softmax output from another model. For pre-processing, the team manually determined a bounding box around the cochleas as a ROI in an atlas (randomly selected volume) and obtained ROIs in the other volumes by rigid registration. For post-processing, the tumour components with centres 15 pixels superior to the centres of cochleas are considered to be false positive and thus removed. Moreover, 3D connected components analysis was utilised to ensure that only two cochleas and one tumour are remained.

● *PremiLab (5th place, Yao et al.)*. The proposed framework consists of a content-style disentangled GAN for style transfer and a modified 3D version ResU-Net along with two types of attention modules for segmentation (DAR-U-Net). Specifically, content is extracted from both modalities using the same encoder, while style is extracted using modality-specific encoders. A discriminative approach is adopted to align the content representations of the source and target domain. Once the GAN is trained, ceT_1-to-hrT_2 images are generated with diverse styles to later train the segmentation network, which can imitate the diversity of hrT_2 domain. Different from the original 2D ResU-Net (Diakogiannis et al., 2020), 2.5D structure and group normalisation are employed for computation efficiency of 3D images. Meanwhile, Voxel-wise Attention Module (VAM) and Quartet Attention Module (QAM) are implemented in each level of the decoder and each residual block, respectively. VAM enhances the essential areas of the feature maps in the decoder using the encoder feature, while QAM captures the inter-dimensional dependencies to improve networks with low computation cost. Code available at: `https://github.com/Kaiseem/DAR-UNet`.

● *Epione-Liryc (6th place, Ly et al.)*. The team proposed a regularised image-to-image translation approach. First, input images are spatially normalised to MNI space using SPM12[8], allowing for the identification of a global region of interest bounding box using the simple addition of the ground truth labels. The cross-modality domain adaptation is then performed using a CycleGAN model (Zhu et al., 2017) to translate the ceT_1 to pseudo-hrT_2. To improve the performance of the CycleGAN, a supervised regularisation technique is added to control the training process, called the Pair-Loss. This loss is calculated as the MSE loss between the pairs of closest 2D-slice images, which are semi-automatically selected using the cross-entropy metric. The segmentation model is built using the 3D U-Net architecture and trained using both ceT_1 and pseudo-hrT_2 data. At the inference stage, the segmentation output is reverted to the original spatial domain and refined using majority voting between the segmented mask and the k-Means and Mean-Shift derived masks. Finally, Dense Conditional Random Field (CRF)

---

[8]https://www.fil.ion.ucl.ac.uk/spm/software/spm12/

(Krähenbühl and Koltun, 2011) is applied to further improve the segmentation result.

● ***MedICL (7th place, Li et al.)***. This framework proposed by Li et al. (2021a) consists of two components: Synthesis and segmentation. For the synthesis component, the Cycle-GAN pipeline is used for unpaired image-to-image translation between ceT_1 and hrT_2 MRIs. For the segmentation component, the generated hrT_2 MRIs are fed into two 2.5D U-Net models (Wang et al., 2019) and two 3D U-Net models Li et al. (2021b). The 2.5D models contain both 2D and 3D convolutions, as well as an attention module. Residual blocks and deep supervision are used in the 3D CNN models. Furthermore, various data augmentation schemes are applied during training to cope with MRIs from different scanners, including spatial, image appearance, and image quality augmentations. Different parameter settings are used for the two 2.5D CNN models. The difference between the two 3D CNN models is that only one of them had an attention module. Finally, the models are ensembled to obtain the final segmentation result.

● ***DBMI_pitt (8th place, Zu et al.)***. The proposed framework use 3D image-to-image translation to generate pseudo-data used to train a segmentation network. To perform image-to-image translation, authors extended the 2D CUT model (Park et al., 2020) to 3D translation. The translation model consists of a generator G, a discriminator D and a feature extractor F. The generator G is built upon the 2.5D attention U-Net proposed in Wang et al. (2019), where two down-sampling layers are removed. For the discriminator D, the PatchGAN discriminator is selected Isola et al. (2017). The model structure of D is the 6 layers of Resnet, and when feeding images to the D, the images are divided into 16 equal-size patches, which is faster for model feed-forward without sacrificing any performance. The feature extractor F is a simple multi-layer full-connected network as Park et al. (2020). The segmentation network is a 2.5D (Wang et al., 2019). To perform image segmentation, the attention 2.5 U-Net (Wang et al., 2019) is used as backbone architecture. A detection module is built upon it. Finally, post-processing is performed using standard morphological operations (hole filling) and the largest component selection for VS. Code is available at: `https://github.com/chkzhao/crossMoDA.git`.

● ***Hi-Lib (9th place, Wu et al.)***. The proposed approach is based on the 2.5D attention U-Net (Wang et al., 2019), where a GAN-based data augmentation strategy is employed to eliminate the instability of unpaired image-to-image translation. Specifically, a source-domain image is sent to the trained CycleGAN (Zhu et al., 2017) and CUT (Park et al., 2020) to obtain two different pseudo-target images, and then they are converted back to source domain-like images so that each source domain image shares the same label with its augmented versions. To pre-process the data, the team calculated the largest bounding box based on the labelled training images and used it to crop all the images. The training process is done in PyMIC, where intensity normalisation, random flipping and cropping are used in training. Each test image is translated into source domain-like by CycleGAN and sent to the trained 2.5D segmentation

network. After inference, conditional random fields and removing small-connected regions are used for post-processing. Code available at: `https://github.com/JianghaoWu/FPL-UDA.git`.

● ***smriti161096 (10th place, Joshi et al.)***. This approach is based on existing frameworks. The method follows three main steps: pre-processing, image-to-image translation and image segmentation. First, axial slices from MRI are selected using maximum coordinates of bounding boxes across all available segmentation masks and resizing them to a uniform size. Then, pseudo-hrT_2 images are generated using the 2D Cycle-GAN (Zhu et al., 2017) architecture. Finally, the 3D nnU-Net (Isensee et al., 2021) framework is trained using the generated images and their manual annotations. To improve the segmentation, the segmentation model is pre-trained using the ceT_1 images. Moreover, multiple checkpoints are selected from CycleGAN training to generate images with varying representations of tumours. In addition to this, self-training is employed: pseudo-labels for hrT_2 data are generated with the trained network and then used to further train the network with real images in hrT_2 modality. Lastly, 3D component analysis is applied as a post-processing step to keep the largest connected component for the tumour label.

● ***IMI (11th place, Hansen et al.)***. The proposed approach is based on robust deformable multi-modal multi-atlas registration to bridge the domain gap between T1 (source) and T2 (target) weighted MRI scans. The source and target domain are resampled to isotropic 1 mm resolution and cropped with an ROI of size 64×64×96 voxels within the left and right hemispheres. 30 source training images are randomly selected and automatically registered to a subset of the target training scans both linearly and non-rigidly. Registration is performed using the discrete optimisation framework deeds Heinrich et al. (2013a) with multi-modal feature descriptors (MIND-SSC Heinrich et al. (2013b)). The propagated source labels are fused using the popular STAPLE algorithm. For fast inference, a nnU-Net model is trained on the noisy labels in the target domain. Based on the predicted segmentations, an automatic centre crop of 48×48×48 mm is chosen (on both hemispheres using the centre-of-mass) with a 0.5 mm resolution. The described process (multi-atlas registration, label propagation and fusion using STAPLE, nnU-Net training) is repeated on the refined crops.

● ***GapMIND (12th place, Kruse et al.)***. The proposed approach uses modality-independent neighbourhood descriptors (MIND) (Heinrich et al., 2012) to obtain a domain-invariant representation of the source and target data. MIND features describe each voxel with the intensity relations between the surrounding image patches. To perform image segmentation, a Deeplab segmentation pipeline with a MobileNetV2 backbone (Sandler et al., 2018) is then trained on the annotated source MIND feature maps obtained from the annotated source images. To improve the performance of the segmentation network, pseudo-labels are generated for the MIND feature maps from

the target data and used during training. Specifically, the authors used the approach from the IMI team (●) based on image registration and STAPLE fusion to obtain noisy labels for these target MIND feature maps.

● *gabybaldeon (13th place, Calisto et al.).* This team implemented an image- and feature-level adaptation method (Baldeon-Calisto and Lai-Yuen, 2021). First, images from the source domain are translated to the target domain via a Cycle-GAN model (Zhu et al., 2017). Then, a 2D U-Net model is trained to segment the target domain images in two stages. In the first stage, the U-Net is trained using the translated source images and their annotations. In the second phase, the feature-level adaptation is achieved through an adversarial learning scheme. The pre-trained U-Net network takes the role of the generator and predicts the segmentations for the target and translated source domain images. Inspired by Li et al. (2020), the discriminator takes as input the concatenation of the predicted segmentations, the element-wise multiplication of the predicted segmentation and the original image, and the contour of the predicted segmentation by applying a Sobel operator. This input provides information about the shape, texture, and contour of the segmented region to force the U-Net to be boundary and semantic-aware.

● *SEU_Chen (14th place, Xiaofei et al.).* The proposed approach employs minimal-entropy correlation alignment (Morerio et al., 2018) to perform domain adaptation. Two segmentation models are trained using the 3D nnU-Net (Isensee et al., 2021) framework. Firstly, the annotated source training set is used to train a 3D nnU-net framework. Secondly, another 3D nnU-net framework is trained with domain adaptation. Specifically, DA is performed by minimising the weighted cross-entropy on the source domain and the weighted entropy on the target domain. At the inference stage, VS segmentation is performed using the adapted nnU-Net framework. In contrast, the two models are ensembled for the cochleas task.

● *skjp (15th place, Kondo).* The proposed approach uses Gradient Reversal Layer (GRL) (Ganin et al., 2016) to perform domain adaption. First, a 3D version of ENet (Paszke et al., 2016) is trained with the annotated source domain dataset. Then, domain adaptation is performed. Specifically, feature maps extracted in the encoder part of the network are fed to GRL. The GRL's output is then used as input of a three fully-connected layers domain classifier. The segmentation network, GRL, and domain classification network are trained with samples from both source and target domains using adversarial learning. Two separate networks are trained for VS and cochleas segmentation.

● *IRA (16th place, Belkov et al.).* Gradient Reversal Layer (GRL) (Ganin et al., 2016) was utilised to perform domain adaption. Two slightly modified 3D U-Net architectures are used to solve the binary segmentation task for cochleas and VS, respectively. These models are trained using ceT_1 data and adapted using pairs of ceT_1 and hrT_2 scans. The adversarial head is used to align the domain features as in Ganin et al.

(2016). Contrary to the original implementation, the Gradient Reversal Layer is attached to the earlier network blocks based on the layer-wise domain shift visualisation from Zakazov et al. (2021).

6. Results

Participants submissions were required to submit their Docker container by 15th August 2021. Winners were announced during the crossMoDA event at the MICCAI 2021 conference. This section presents the results obtained by the participant teams on the test set and analyses the stability and robustness of the proposed ranking scheme.

6.1. Overall segmentation performance

The final scores for the 16 teams are reported in Table 3 in the order in which they ranked. Figures 2 and 3 show the boxplots for each structure (VS and cochleas) and are colour-coded according to the team. The performance distribution is given for each metric (DSC and ASSD). Qualitative results are shown in Figure 4.
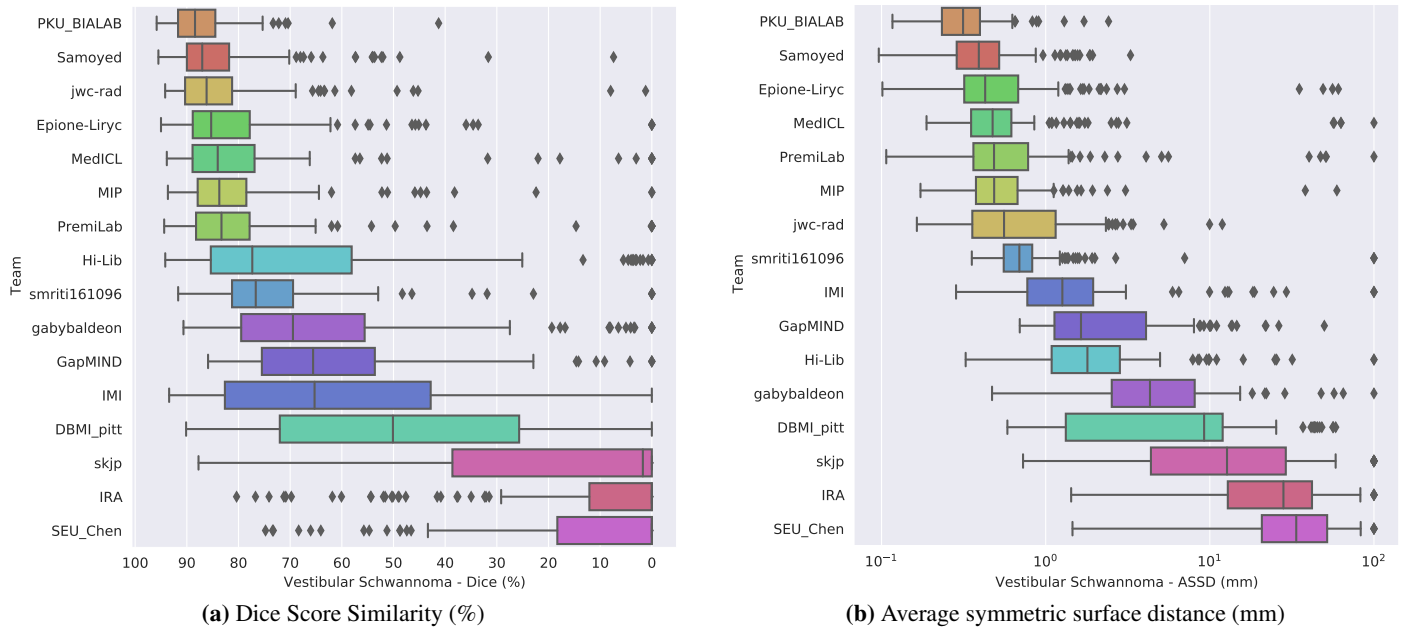
The winner of the crossMoDA challenge is Samoyed with a rank score of 2.7. Samoyed is the only team that reached a median DSC greater than 85% for both structures. Other teams in the top five also obtained outstanding results with a median DSC greater than 80% for each structure. In contrast, the low DSC and ASSD scores of the three teams with the lowest rank highlight the complexity of the cross-modality domain adaptation task.

The top ten teams all used an approach based on image-to-image translation. As shown in Table 3, the medians are significantly higher, and the interquartile ranges (IQRs) are smaller compared to other approaches. Approaches using MIND cross-modality features obtained the following places (eleventh and twelfth ranks), while those aiming at aligning the distribution of the features extracted from the source and target images obtained the last positions. This highlights the effectiveness of using CycleGAN and its extensions to bridge the gap between ceT_1 and hrT_2 scans.
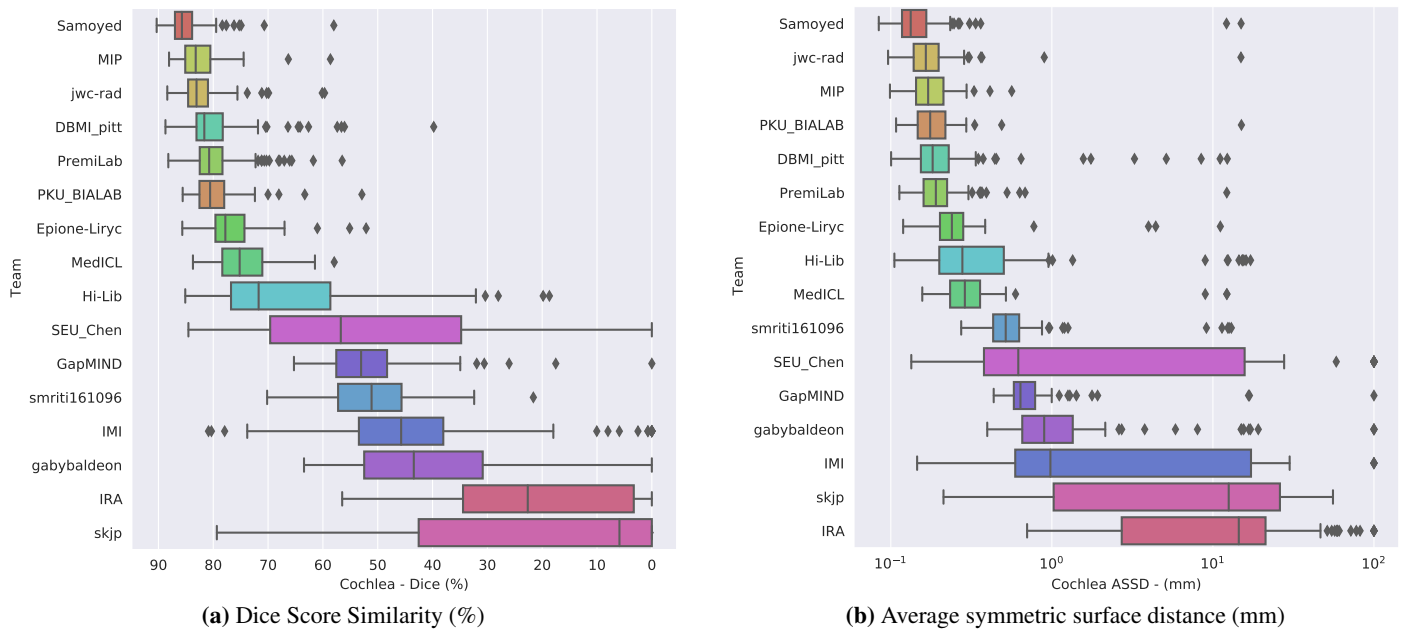
6.2. Evaluation per structure and impact on the rank

The level of robustness and performance of the proposed techniques highly depends on the structure, impacting the ranking.

Examining the distribution of the scores is crucial for analysing the robustness of the proposed methods. More variability can be observed in terms of algorithm performance for the tumour than for the cochleas. On average, the IQRs of the top 10 performing teams for the DSC and ASSD are respectively 2.6 and 16 times larger for VS than cochleas. Moreover, Figures 3 and 2 show that there are more outliers for VS than for cochleas. This suggests that the proposed algorithms are less robust on VS than on cochleas. For example, the winning team obtained a relatively poor DSC (< 60%) for respectively 8% (N=11) and 1% (N=1) of the testing set on the tumour and the cochleas. This can be explained by the fact that cochleas

(a) Dice Score Similarity (%)　　　　　**(b)** Average symmetric surface distance (mm)

Fig. 2: Box plot of the method's segmentation performance for the vestibular schwannoma in terms of (a) DSC and (b) ASSD.

(a) Dice Score Similarity (%)　　　　　**(b)** Average symmetric surface distance (mm)

Fig. 3: Box plot of the method's segmentation performance for the cochleas in terms of (a) DSC and (b) ASSD.

Table 3: Metrics values and corresponding scores of submission. Median and interquartile values are presented. The best results are given in bold. Arrows indicate favourable direction of each metric.

	Challenge Rank		Vestibular Schwannoma		Cochleas	
	Global Rank ↓	Rank Score ↓	DSC (%) ↑	ASSD (mm) ↓	DSC (%) ↑	ASSD (mm) ↓
Full-supervision	-	-	92.5 [89.2 - 94.2]	0.20 [0.14 - 0.29]	87.7 [85.8 - 89.3]	0.10 [0.09 - 0.13]
● Samoyed	**1**	**2.7**	87.0 [81.8 - 90.0]	0.39 [0.29 - 0.52]	**85.7 [83.9 - 87.0]**	**0.13 [0.12 - 0.17]**
● PKU_BIALAB	2	3.4	**88.4 [84.5 - 91.7]**	**0.31 [0.23 - 0.4]**	80.6 [78.0 - 82.5]	0.18 [0.15 - 0.22]
● jwc-rad	3	4.2	86.2 [81.2 - 90.3]	0.56 [0.36 - 1.15]	83.1 [81.0 - 84.6]	0.17 [0.14 - 0.2]
● MIP	4	4.5	83.7 [78.5 - 87.9]	0.49 [0.38 - 0.67]	83.2 [80.5 - 85.1]	0.17 [0.14 - 0.21]
● PremiLab	5	5.2	83.3 [77.8 - 88.2]	0.48 [0.36 - 0.78]	80.7 [78.3 - 82.5]	0.19 [0.16 - 0.22]
● Epione-Liryc	6	6.0	85.3 [77.8 - 88.9]	0.43 [0.32 - 0.68]	77.8 [74.3 - 79.6]	0.24 [0.2 - 0.28]
● MedICL	7	6.6	84.0 [76.9 - 88.9]	0.48 [0.35 - 0.62]	75.2 [71.1 - 78.4]	0.29 [0.23 - 0.36]
● DBMI_pitt	8	8.4	50.1 [25.7 - 72.0]	9.23 [1.33 - 11.98]	81.6 [78.3 - 83.1]	0.18 [0.15 - 0.23]
● Hi-Lib	9	8.8	77.3 [58.1 - 85.4]	1.8 [1.09 - 2.83]	71.7 [58.6 - 76.8]	0.28 [0.2 - 0.5]
● smriti161096	10	9.7	76.7 [69.4 - 81.2]	0.69 [0.56 - 0.83]	51.1 [45.7 - 57.2]	0.52 [0.43 - 0.63]
● IMI	11	11.0	65.3 [42.8 - 82.6]	1.26 [0.77 - 1.95]	45.7 [38.0 - 53.4]	0.98 [0.59 - 17.24]
● GapMIND	12	11.2	65.6 [53.6 - 75.5]	1.64 [1.13 - 4.09]	53.0 [48.3 - 57.6]	0.64 [0.58 - 0.79]
● gabybaldeon	13	11.9	69.5 [55.6 - 79.5]	4.32 [2.53 - 8.09]	43.4 [30.8 - 52.5]	0.9 [0.65 - 1.35]
● SEU_Chen	14	13.2	0.0 [0.0 - 18.3]	33.62 [20.75 - 51.86]	56.7 [34.8 - 69.6]	0.62 [0.38 - 15.74]
● skjp	15	13.9	1.7 [0.0 - 38.6]	12.73 [4.38 - 29.03]	5.9 [0.0 - 42.5]	12.54 [1.03 - 26.07]
● IRA	16	14.7	0.0 [0.0 - 12.1]	28.09 [12.87 - 41.92]	22.6 [3.3 - 34.4]	14.48 [2.72 - 21.23]

are more uniform in terms of location, volume size and intensity distribution than tumours. This could also be the reason why techniques using feature alignment collapsed on VS task.

Conversely, the level of performance for the cochleas task had a stronger impact on the rank scores. Table 3 shows that the top seven teams obtained a comparable performance on the VS task (median DSC - min: 83.3%; max: 88.4%), while more variability is observed on the cochleas task (median DSC - min: 75.2%; max: 85.7%). Table 4 shows the distribution of the individual cumulative ranks for each structure (VS and cochleas). It can be observed that the winner significantly outperformed all the other teams on the cochleas. In contrast, while the second team obtained the best performance on the VS task (see Table 3), it didn't rank high enough on the cochleas task to win the challenge. Similarly, the fourth to seventh teams, which obtained comparable cumulative ranks on the VS task (median between 5 and 5.5), are ranked in the same order as their median cumulative ranks on the cochleas task. This shows that the performance of the top-performing algorithms on the cochleas was the most discriminative for the final ranking.

6.3. Remarks about the ranking stability

It has been shown that challenge rankings can be sensitive to various design choices, such as the test set used for validation, the metrics chosen for assessing the algorithms' performance and the scheme used to aggregate the values (Maier-Hein et al., 2018). In this section, we analyse and visualise the ranking stability with respect to these design choices.
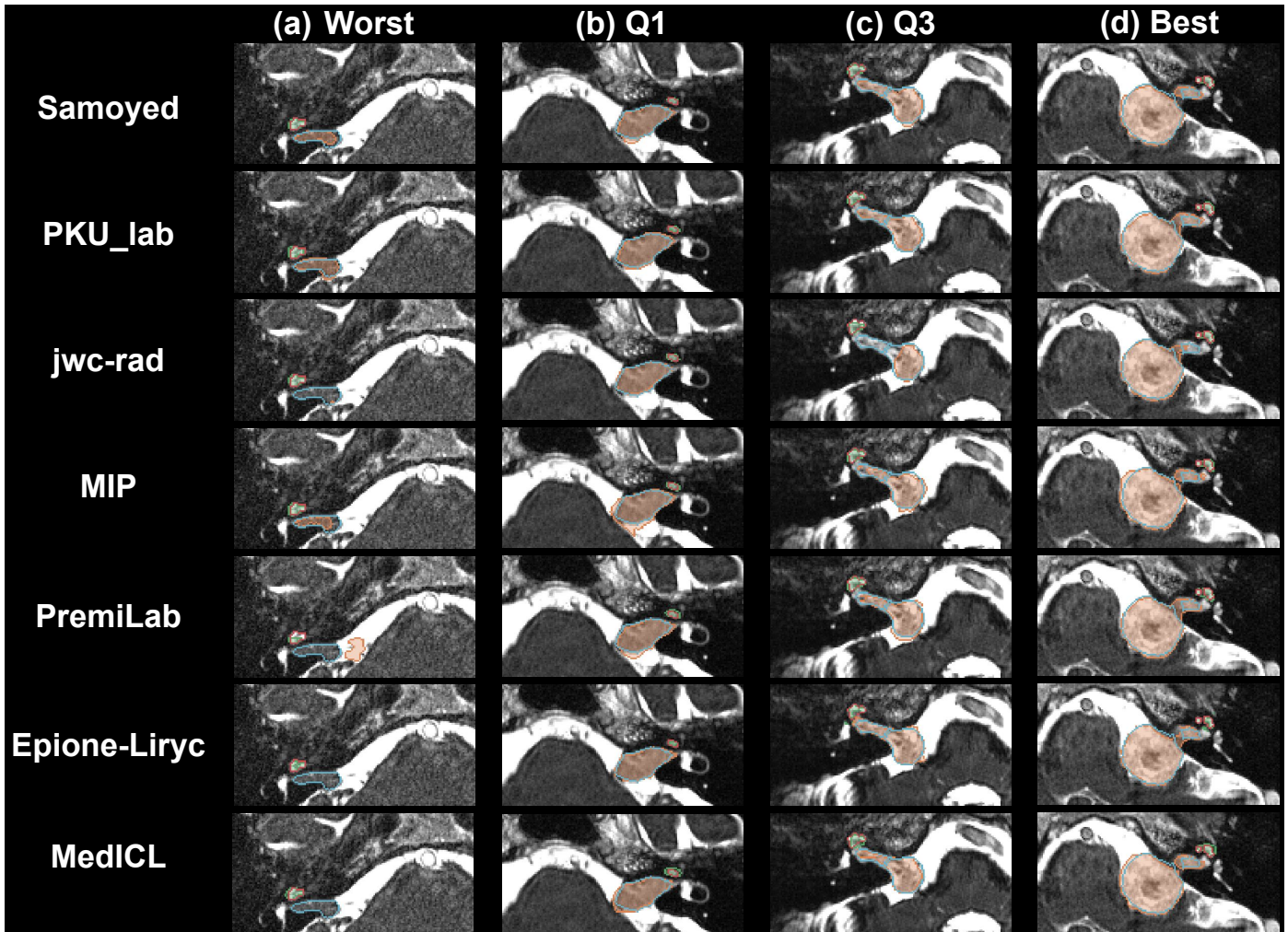
A recent work proposed techniques to assess the stability of rankings with respect to sampling variability (Wiesenfarth et al., 2021). Following their recommendations, we performed bootstrapping (1,000 bootstrap samples) to investigate the ranking uncertainty and stability of the proposed ranking scheme with

Table 4: Distribution of the individual cumulative ranks for each structure. Median and interquartile values are presented.

	Challenge Rank	Vestibular Schwannoma	Cochleas
● Samoyed	1	3.0 [2.0 - 4.5]	1.0 [1.0 - 2.0]
● PKU_BIALAB	2	1.5 [1.0 - 2.5]	5.0 [3.5 - 6.0]
● jwc-rad	3	5.0 [3.0 - 7.0]	3.0 [2.0 - 4.5]
● MIP	4	5.0 [4.0 - 7.0]	3.5 [2.0 - 4.5]
● PremiLab	5	5.5 [3.5 - 7.0]	5.0 [3.5 - 6.0]
● Epione-Liryc	6	5.0 [3.0 - 6.5]	7.0 [6.5 - 8.0]
● MedICL	7	5.0 [3.5 - 7.0]	8.0 [7.0 - 9.0]
● DBMI_pitt	8	13.0 [10.5 - 13.5]	5.0 [3.5 - 6.0]
● Hi-Lib	9	9.0 [7.5 - 10.5]	8.5 [7.5 - 9.5]
● smriti161096	10	8.0 [7.5 - 9.0]	11.0 [10.5 - 12.5]
● IMI	11	10.0 [8.5 - 11.0]	13.0 [12.0 - 14.5]
● GapMIND	12	11.0 [10.0 - 12.0]	11.5 [11.0 - 13.0]
● gabybaldeon	13	11.0 [10.0 - 12.0]	13.0 [12.0 - 14.0]
● SEU_Chen	14	15.0 [14.5 - 15.5]	11.5 [10.0 - 14.0]
● skjp	15	14.0 [13.0 - 15.0]	15.0 [13.0 - 16.0]
● IRA	16	14.5 [14.0 - 15.5]	15.0 [14.5 - 15.5]

respect to sampling variability. To this end, the ranking strategy is performed repeatedly on each bootstrap sample. To quantitatively assess the ranking stability, the agreement of the challenge ranking and the ranking lists based on the individual bootstrap samples was determined via Kendall's τ, which provides values between -1 (for reverse ranking order) and 1 (for identical ranking order). The median [IQR] Kendall's τ was 1 [1-1], demonstrating the perfect stability of the ranking scheme. Figure 5 shows a blob plot of the bootstrap rankings. The same conclusion can be drawn: the ranking stability of the proposed scheme is excellent. In particular, the winning team is first-ranked for all the bootstrap samples.

To evaluate the stability of the ranking with respect to the

Fig. 4: Qualitative comparison of the top 7 performing teams. Selected cases correspond to the (a) lowest, (b) lower-quartile, (c) upper-quartile, and (d) highest mean Dice score (averaged over the top 7 team and over the two structures).

choice of the metrics, we compared the stability of single-metric (DSC or ASSD) ranking schemes with our multi-metric (DSC and ASSD) ranking scheme. Specifically, bootstrapping was used to compare the stability of the ranking for the three sets of metrics and Kendall's τ were computed to compare the ranking list computed on the full assessment data and the individual bootstrap samples. Violin plots shown in Figure 6 illustrate bootstrap results for each metric. It can be observed that Kendall's τ are more dispersed across the bootstrap samples when using only one metric. Median Kendall's τ are respectively 0.98, 0.98 and 1 using DSC, ASSD and the combination of both as metric. This demonstrates that the ranking stability is higher when multiple metrics are used.

Finally, we compared our ranking scheme with other ranking methods with different aggregation methods. The most prevalent approaches are:

- Aggregate-then-rank: metric values across all test cases are first aggregated (e.g., with the mean, median) for each structure and each metric. Ranks per structure and per metric are then computed for each team. Ranking scores correspond to the aggregation (e.g., with mean, median) of

these ranks and are used for the final ranking.

- Rank-then-aggregate: algorithms' ranks are computed for each test case, for each metric and each structure and then aggregated (e.g., with the mean, median). Then, the aggregated rank score is used to rank algorithms.

Our ranking scheme corresponds to a rank-then-aggregate approach with the mean as aggregation technique. We compared our approach with: 1/ a rank-then-aggregate approach using another aggregation technique (the median); 2/ aggregate-then-rank approaches using either the mean and the median for metric aggregation. Ranking robustness across these different ranking methods is shown on the line plots in Figure 7. It can be seen that the ranking is robust to these different ranking techniques. In particular, the first seven ranks are the same for all ranking scheme variations. Note that the aggregate-then-rank approach using the mean is less robust due to the presence of outliers for the ASSD metric caused by missing segmentation for a given structure. This demonstrates that the ranking of the challenge is stable and can be interpreted with confidence.

Table 5: Impact of self-supervision on top 2 team methods. Median and interquartile values are presented. The best results are given in bold. Arrows indicate the favourable direction of each metric.

	Self-Supervision	Vestibular Schwannoma		Cochleas	
		DSC (%) ↑	ASSD (mm) ↓	DSC (%) ↑	ASSD (mm) ↓
● Samoyed	✓	**87.0 [81.8 - 90.0]**	**0.39 [0.29 - 0.52]**	**85.7 [83.9 - 87.0]**	**0.13 [0.12 - 0.17]**
	✗	85.3 [79.7 - 89.1]	0.42 [0.33 - 0.59]	84.7 [81.9 - 86.3]	0.15 [0.12 - 0.2]
● PKU_BIALAB	✓	**88.4 [84.5 - 91.7]**	**0.31 [0.23 - 0.4]**	**80.6 [78.0 - 82.5]**	**0.18 [0.15 - 0.22]**
	✗	86.5 [80.6 - 90.4]	0.36 [0.27 - 0.5]	75.7 [71.4 - 78.8]	0.22 [0.18 - 0.28]

Fig. 5: Stability of the proposed ranking scheme for 1000 bootstrap samples.

Fig. 6: Stability of the ranking scheme with respect to the choice of the metrics. 1000 bootstrap samples are used.

7. Discussion

In this study, we introduced the crossMoDA challenge in terms of experimental design, evaluation strategy, proposed methods and final results. In this section, we discuss the main insights and limitations of the challenge.

Fig. 7: Line plots visualising rankings robustness across different ranking methods for the brain task. Each algorithm is represented by one coloured line. For each ranking method encoded on the x-axis, the height of the line represents the corresponding rank. The lowest rank of tied values is used for ties (equal scores for different teams)

7.1. Performance of automated segmentation methods

To compare the level of performance reached by the top-performing teams with a fully-supervised approach, we trained a nnU-Net framework (Isensee et al., 2021) on the hrT$_2$ scans paired with the ceT$_1$ scans from the source training set ($N = 105$) and their manual annotations. Segmentation performances on VS and cochleas are reported in Table 3. It can be observed that full supervision significantly outperforms all participating teams on the two structures. The performance gap between the best performing team on VS and the fully-supervised model is 4.1% (median DSC) and 0.11mm (median ASSD). The performance gap between the best performing team on cochleas and the fully-supervised model is 2% (median DSC) and 0.03mm (median ASSD). Moreover, a fully-supervised approach obtained tighter IQRs, demonstrating better robustness. This shows that even though top-performing teams obtained a high level of performance, full supervision still outperforms these proposed approaches. Note that the top performing teams reached a level of performance that is higher than the one reported in another study with a weakly-supervised approach trained using scribbles on the VS and on a different split of the dataset Dorent et al. (2020).

7.2. Analysis of the top-ranked methods

Image-to-image translation using CycleGAN and its extension was the most successful approach to bridge the gap between the source and target images. Except for one team, 2D image-to-image translation was performed on 2D axial slices. Teams used CycleGAN and two of its extensions (NiceGAN, CUT).

However, the results in Table 3 do not demonstrate the advantage of using one image-to-image translation approach over another. For example, the third and ninth teams, which used the same implementation of the same approach (2D CUT), respectively obtained a median DSC of 83.1% and 71.7% on the cochleas. This shows that the segmentation performance depends on other parameters, such as the pre-processing step (cropping, image resampling, image normalisation) and the segmentation network. However, the current study does not allow for the identification of the optimal combination of parameters.

Except for three teams, all teams have used U-Net as segmentation backbone. In particular, the top three teams used the nnU-Net framework, a deep segmentation method that automatically configures itself, including pre-processing, network architecture, training and post-processing based on heuristic rules. This suggests the effectiveness of this framework for VS and cochleas segmentation.

Finally, it can be seen that the two top-performing teams used self-training. To study the impact of self-supervision on the framework performance, an ablation study was performed. Specifically, the two top-performing teams were asked to train their framework without the self-supervision component. Results are shown in Table 5. It can be seen that self-supervision leads to statically significant performance improvement provided by a Wilcoxon test ($p < 0.01$) for both teams. Self-training for domain adaptation has been previously proposed for medical segmentation problems. However, it is the first time that self-training has been successfully used in the context of large domain gaps. In practice, image-to-image translation and self-training are used sequentially: first pseudo-target images are generated and used to train segmentation networks, and then self-training is used to fine-tune the trained networks to manage the (small) domain gap between pseudo- and real target images.

7.3. Limitations and future directions for the challenge

The lack of robustness to unseen situations is a key problem for deep learning algorithms in clinical practice. We created this challenge to benchmark new and existing domain adaptation techniques on a large and multi-class dataset. In this challenge, the domain gap between the source and target images is large, as it corresponds to different imaging modalities. However, the lack of robustness can also occur when the same image modalities are acquired in different settings (e.g., hospital, scanner). This problem is not addressed in this challenge. Indeed, images within a domain (target or source) have been acquired at a unique medical centre with the same scanner. Moreover, 96% of the test set has been acquired with the same sequence parameters. For this reason, we plan to diversify the challenge dataset by adding data from other institutions. In particular, different hrT_2 appearances are likely to occur, making it challenging for

image-to-image translation approaches, which assume that the relationship between the target and source domains is a bijection.

8. Conclusion

The crossMoDA challenge was introduced to propose the first benchmark of domain adaptation techniques for medical image segmentation. The level of performance reached by the top-performing teams is surprisingly high and close to full supervision. Top performing teams all used an image-to-image translation approach to transform the source images into pseudo-target images and then train a segmentation network using these generated images and their manual annotations. Self-training has been shown to lead to performance improvements.

CRediT authorship contribution statement

Reuben Dorent: Conceptualization, Methodology, Software, Formal analysis, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Aaron Kujawa:** Conceptualization, Methodology, Data curation, Writing - review & editing. **Spyridon Bakas:** Conceptualization, Methodology. **Nicola Rieke:** Conceptualization, Methodology. **Samuel Joutard:** Conceptualization, Methodology. **Ben Glocker:** Conceptualization, Methodology. **Jorge Cardoso:** Conceptualization, Methodology. **Marc Modat:** Conceptualization, Methodology. **Kayhan Batmanghelich:** Methodology, Software. **Arseniy Belkov:** Methodology, Software. **Maria Baldeon Calisto:** Methodology, Software. **Jae Won Choi:** Methodology, Software. **Benoit M. Dawant:** Methodology, Software. **Hexin Dong:** Methodology, Software. **Sergio Escalera:** Methodology, Software. **Yubo Fan:** Methodology, Software. **Lasse Hansen:** Methodology, Software. **Mattias P. Heinrich:** Methodology, Software. **Smriti Joshi:** Methodology, Software. **Victoriya Kashtanova:** Methodology, Software. **Hyeon gyu Kim:** Methodology, Software. **Satoshi Kondo:** Methodology, Software. **Christian N. Kruse:** Methodology, Software. **Susana K. Lai-Yuen:** Methodology, Software. **Hao Li:** Methodology, Software. **Han Liu:** Methodology, Software. **Buntheng Ly:** Methodology, Software. **Ipek Oguz:** Methodology, Software. **Hyungseob Shin:** Methodology, Software. **Boris Shirokikh:** Methodology, Software. **Zixian Su:** Methodology, Software. **Guotai Wang:** Methodology, Software. **Jianghao Wu:** Methodology, Software. **Yanwu Xul:** Methodology, Software. **Kai Yao:** Methodology, Software. **Li Zhang:** Methodology, Software. **Sébastien Ourselin:** Supervision, Funding acquisition. **Jonathan Shapey:** Conceptualization, Methodology, Data curation, Writing - review & editing, Funding acquisition. **Tom Vercauteren:** Project administration, Conceptualization, Methodology, Formal analysis, Resources, Writing - original draft, Writing - review & editing, Funding acquisition.

Acknowledgements

References

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., AnnetteKopp-Schneider, Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., Bilello, M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M.J., Heckers, S.H., Huisman, H., Jarnagin, W.R., McHugo, M.K., Napel, S., Pernicka, J.S.G., Rhode, K., Tobon-Gomez, C., Vorontsov, E., Huisman, H., Meakin, J.A., Ourselin, S., Wiesenfarth, M., Arbelaez, P., Bae, B., Chen, S., Daza, L., Feng, J., He, B., Isensee, F., Ji, Y., Jia, F., Kim, N., Kim, I., Merhof, D., Pai, A., Park, B., Perslev, M., Rezaiifar, R., Rippel, O., Sarasua, I., Shen, W., Son, J., Wachinger, C., Wang, L., Wang, Y., Xia, Y., Xu, D., Xu, Z., Zheng, Y., Simpson, A.L., Maier-Hein, L., Cardoso, M.J., 2021. The medical segmentation decathlon. arXiv:2106.05735.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C., 2017. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data 4, 1–13.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., Prastawa, M., Alberts, E., Lipkova, J., Freymann, J., Kirby, J., Bilello, M., Fathallah-Shaykh, H., Wiest, R., Kirschke, J., Wiestler, B., Colen, R., Kotrotsou, A., Lamontagne, P., Marcus, D., Milchenko, M., Nazeri, A., Weber, M.A., Mahajan, A., Baid, U., Gerstner, E., Kwon, D., Acharya, G., Agarwal, M., Alam, M., Albiol, A., Albiol, A., Albiol, F.J., Alex, V., Allinson, N., Amorim, P.H.A., Amrutkar, A., Anand, G., Andermatt, S., Arbel, T., Arbelaez, P., Avery, A., Azmat, M., B., P., Bai, W., Banerjee, S., Barth, B., Batchelder, T., Batmanghelich, K., Battistella, E., Beers, A., Belyaev, M., Bendszus, M., Benson, E., Bernal, J., Bharath, H.N., Biros, G., Bisdas, S., Brown, J., Cabezas, M., Cao, S., Cardoso, J.M., Carver, E.N., Casamitjana, A., Castillo, L.S., Catà, M., Cattin, P., Cerigues, A., Chagas, V.S., Chandra, S., Chang, Y.J., Chang, S., Chang, K., Chazalon, J., Chen, S., Chen, W., Chen, J.W., Chen, Z., Cheng, K., Choudhury, A.R., Chylla, R., Clérigues, A., Colleman, S., Colmeiro, R.G.R., Combalia, M., Costa, A., Cui, X., Dai, Z., Dai, L., Daza, L.A., Deutsch, E., Ding, C., Dong, C., Dong, S., Dudzik, W., Eaton-Rosen, Z., Egan, G., Escudero, G., Estienne, T., Everson, R., Fabrizio, J., Fan, Y., Fang, L., Feng, X., Ferrante, E., Fidon, L., Fischer, M., French, A.P., Fridman, N., Fu, H., Fuentes, D., Gao, Y., Gates, E., Gering, D., Gholami, A., Gierke, W., Glocker, B., Gong, M., González-Villá, S., Grosges, T., Guan, Y., Guo, S., Gupta, S., Han, W.S., Han, I.S., Harmuth, K., He, H., Hernández-Sabaté, A., Herrmann, E., Himthani, N., Hsu, W., Hsu, C., Hu, X., Hu, X., Hu, Y., Hu, Y., Hua, R., Huang, T.Y., Huang, W., Huffel, S.V., Huo, Q., HV, V., Iftekharuddin, K.M., Isensee, F., Islam, M., Jackson, A.S., Jambawalikar, S.R., Jesson, A., Jian, W., Jin, P., Jose, V.J.M., Jungo, A., Kainz, B., Kamnitsas, K., Kao, P.Y., Karnawat, A., Kellermeier, T., Kermi, A., Keutzer, K., Khadir, M.T., Khened, M., Kickingereder, P., Kim, G., King, N., Knapp, H., Knecht, U., Kohli, L., Kong, D., Kong, X., Koppers, S., Kori, A., Krishnamurthi, G., Krivov, E., Kumar, P., Kushibar, K., Lachinov, D., Lambrou, T., Lee, J., Lee, C., Lee, Y., Lee, M., Lefkovits, S., Lefkovits, L., Levitt, J., Li, T., Li, H., Li, W., Li, H., Li, X., Li, Y., Li, H., Li, Z., Li, X., Li, Z., Li, X., Li, W., Lin, Z.S., Lin, F., Lio, P., Liu, C., Liu, B., Liu, X., Liu, M., Liu, J., Liu, L., Llado, X., Lopez, M.M., Lorenzo, P.R., Lu, Z., Luo, L., Luo, Z., Ma, J., Ma, K., Mackie, T., Madabushi, A., Mahmoudi, I., Maier-Hein, K.H., Maji, P., Mammen, C., Mang, A., Manjunath, B.S., Marcinkiewicz, M., McDonagh, S., McKenna, S., McKinley, R., Mehl, M., Mehta, S., Mehta, R., Meier, R., Meinel, C., Merhof, D., Meyer, C., Miller, R., Mitra, S., Moiyadi, A., Molina-Garcia, D., Monteiro, M.A.B., Mrukwa, G., Myronenko, A., Nalepa, J., Ngo, T., Nie, D., Ning, H., Niu, C., Nuechterlein, N.K., Oermann, E., Oliveira, A., Oliveira, D.D.C., Oliver, A., Osman, A.F.I., Ou, Y.N., Ourselin, S., Paragios, N., Park, M.S., Paschke, B., Pauloski, J.G., Pawar, K., Pawlowski, N., Pei, L., Peng, S., Pereira, S.M., Perez-Beteta, J., Perez-Garcia, V.M., Pezold, S., Pham, B., Phophalia, A., Piella, G., Pillai, G.N., Piraud, M., Pisov, M., Popli, A., Pound, M.P., Pourreza, R., Prasanna, P., Prkovska, V., Pridmore, T.P., Puch, S., Élodie Puybareau, Qian, B., Qiao, X., Rajchl, M., Rane, S., Rebsamen, M., Ren, H., Ren, X., Revanuru, K., Rezaei, M., Rippel, O., Rivera, L.C., Robert, C., Rosen, B., Rueckert, D., Safwan, M., Salem, M., Salvi, J., Sanchez, I., Sánchez, I., Santos, H.M., Sartor, E., Schellingerhout, D., Scheufele, K., Scott, M.R., Scussel, A.A., Sedlar, S., Serrano-Rubio, J.P., Shah, N.J., Shah, N., Shaikh, M., Shankar, B.U., Shboul, Z., Shen, H., Shen, D., Shen, L., Shen, H., Shenoy, V., Shi, F., Shin, H.E., Shu, H., Sima, D., Sinclair, M., Smedby, O., Snyder, J.M., Soltaninejad, M., Song, G., Soni, M., Stawiaski, J., Subramanian, S., Sun, L., Sun, R., Sun, J., Sun, K., Sun, Y., Sun, G., Sun, S., Suter, Y.R., Szilagyi, L., Talbar, S., Tao, D., Tao, D., Teng, Z., Thakur, S., Thakur, M.H.,

Tharakan, S., Tiwari, P., Tochon, G., Tran, T., Tsai, Y.M., Tseng, K.L., Tuan, T.A., Turlapov, V., Tustison, N., Vakalopoulou, M., Valverde, S., Vanguri, R., Vasiliev, E., Ventura, J., Vera, L., Vercauteren, T., Verrastro, C.A., Vidyaratne, L., Vilaplana, V., Vivekanandan, A., Wang, G., Wang, Q., Wang, C.J., Wang, W., Wang, D., Wang, R., Wang, Y., Wang, C., Wang, G., Wen, N., Wen, X., Weninger, L., Wick, W., Wu, S., Wu, Q., Wu, Y., Xia, Y., Xu, Y., Xu, X., Xu, P., Yang, T.L., Yang, X., Yang, H.Y., Yang, J., Yang, H., Yang, G., Yao, H., Ye, X., Yin, C., Young-Moxon, B., Yu, J., Yue, X., Zhang, S., Zhang, A., Zhang, K., Zhang, X., Zhang, L., Zhang, X., Zhang, Y., Zhang, L., Zhang, J., Zhang, X., Zhang, T., Zhao, S., Zhao, Y., Zhao, X., Zhao, L., Zheng, Y., Zhong, L., Zhou, C., Zhou, X., Zhou, F., Zhu, H., Zhu, J., Zhuge, Y., Zong, W., Kalpathy-Cramer, J., Farahani, K., Davatzikos, C., van Leemput, K., Menze, B., 2019. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv:1811.02629.

Baldeon-Calisto, M., Lai-Yuen, S.K., 2021. C-mada: Unsupervised cross-modality adversarial domain adaptation framework for medical image segmentation. arXiv preprint arXiv:2110.15823 .

Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., Ayed, I.B., 2019. Constrained domain adaptation for segmentation, in: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing, Cham. pp. 326–334.

Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., Ben Ayed, I., 2020. Source-relaxed domain adaptation for image segmentation, in: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Springer International Publishing, Cham. pp. 490–499.

Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A., 2020a. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. IEEE Transactions on Medical Imaging 39, 2494–2505. doi:10.1109/TMI.2020.2972701.

Chen, R., Huang, W., Huang, B., Sun, F., Fang, B., 2020b. Reusing discriminators for encoding: Towards unsupervised image-to-image translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8165–8174.

Choi, J.W., 2021. Using out-of-the-box frameworks for contrastive unpaired image translation for vestibular schwannoma and cochlea segmentation: An approach for the crossmoda challenge. arXiv:2110.01607.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al., 2013. The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of digital imaging 26, 1045–1057.

Coelho, D.H., Tang, Y., Suddarth, B., Mamdani, M., 2018. Mri surveillance of vestibular schwannomas without contrast enhancement: Clinical and economic evaluation. The Laryngoscope 128, 202–209. doi:https://doi.org/10.1002/lary.26589, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/lary.26589.

Cui, H., Yuwen, C., Jiang, L., Xia, Y., Zhang, Y., 2021. Bidirectional cross-modality unsupervised domain adaptation using generative adversarial networks for cardiac image segmentation. Computers in Biology and Medicine 136, 104726. URL: https://www.sciencedirect.com/science/article/pii/S0010482521005205, doi:https://doi.org/10.1016/j.compbiomed.2021.104726.

Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C., 2020. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing 162, 94–114.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T., 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition, in: Xing, E.P., Jebara, T. (Eds.), Proceedings of the 31st International Conference on Machine Learning, PMLR, Bejing, China. pp. 647–655.

Dong, H., Yu, F., Zhao, J., Dong, B., Zhang, L., 2021. Unsupervised domain adaptation in semantic segmentation based on pixel alignment and self-training. arXiv:2109.14219.

Dorent, R., Joutard, S., Shapey, J., Bisdas, S., Kitchen, N., Bradford, R., Saeed, S., Modat, M., Ourselin, S., Vercauteren, T., 2020. Scribble-based domain adaptation via co-segmentation. MICCAI .

Dorent, R., Joutard, S., Shapey, J., Kujawa, A., Modat, M., Ourselin, S., Vercauteren, T., 2021. Inter extreme points geodesics for end-to-end weakly supervised image segmentation, in: de Bruijne, M., Cattin, P.C., Cotin, S.,

Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham. pp. 615–624.

Dou, Q., Ouyang, C., Chen, C., Chen, H., Heng, P.A., 2018. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), pp. 691–697.

Evans, D.G.R., Moran, A., King, A., Saeed, S., Gurusinghe, N., Ramsden, R., 2005. Incidence of vestibular schwannoma and neurofibromatosis 2 in the north west of england over a 10-year period: Higher incidence than previously thought. Otology & Neurotology 26.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. The journal of machine learning research 17, 2096–2030.

Guan, H., Liu, M., 2021. Domain adaptation for medical image analysis: A survey. IEEE Transactions on Biomedical Engineering , 1–1doi:10.1109/tbme.2021.3117407.

Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, S.M., Schnabel, J.A., 2012. Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. Medical Image Analysis 16, 1423–1435. doi:https://doi.org/10.1016/j.media.2012.05.008. special Issue on the 2011 Conference on Medical Image Computing and Computer Assisted Intervention.

Heinrich, M.P., Jenkinson, M., Brady, S.M., Schnabel, J.A., 2013a. MRF-based deformable registration and ventilation estimation of lung CT. IEEE Transaction on Medical Imaging (TMI) 32, 1239–48.

Heinrich, M.P., Jenkinson, M., Papież, B.W., Brady, M., Schnabel, J.A., 2013b. Towards realtime multimodal fusion for image-guided interventions using self-similarities, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 187–194.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 18, 203–211.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. CVPR .

Jiang, J., Veeraraghavan, H., 2020. Unified cross-modality feature disentangler for unsupervised multi-domain mri abdomen organs segmentation, in: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Springer International Publishing, Cham. pp. 347–358.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., Glocker, B., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks, in: Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.T., Shen, D. (Eds.), Information Processing in Medical Imaging, Springer International Publishing, Cham. pp. 597–609.

Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., Baydar, B., Lachinov, D., Han, S., Pauli, J., Isensee, F., Perkonigg, M., Sathish, R., Rajan, R., Sheet, D., Dovletov, G., Speck, O., Nürnberger, A., Maier-Hein, K.H., Bozdağı Akar, G., Ünal, G., Dicle, O., Selver, M.A., 2021. Chaos challenge - combined (ct-mr) healthy abdominal organ segmentation. Medical Image Analysis 69, 101950. URL: https://www.sciencedirect.com/science/article/pii/S1361841520303145, doi:https://doi.org/10.1016/j.media.2020.101950.

Khawaja, A.Z., Cassidy, D.B., Al Shakarchi, J., McGrogan, D.G., Inston, N.G., Jones, R.G., 2015. Revisiting the risks of mri with gadolinium based contrast agents—review of literature and guidelines. Insights into Imaging 6, 553–558. URL: https://doi.org/10.1007/s13244-015-0420-2, doi:10.1007/s13244-015-0420-2.

Krähenbühl, P., Koltun, V., 2011. Efficient inference in fully connected crfs with gaussian edge potentials, in: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2011/file/beda24c1e1b46055dff2c39c98fd6fc1-Paper.pdf.

Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., Collins, D.L., Dadar, M., Georgiou, A., Ghafoorian, M., Jin, D., Khademi, A., Knight, J., Li, H., Lladó, X., Luna, M., Mahmood, Q., McKinley, R., Mehrtash, A., Ourselin, S., Park, B.Y., Park, H., Park, S.H., Pezold, S., Puybareau, E.,

Rittner, L., Sudre, C.H., Valverde, S., Vilaplana, V., Wiest, R., Xu, Y., Xu, Z., Zeng, G., Zhang, J., Zheng, G., Chen, C., van der Flier, W., Barkhof, F., Viergever, M.A., Biessels, G.J., 2019. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. IEEE Transactions on Medical Imaging 38, 2556–2568. doi:10.1109/TMI.2019.2905770.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V., 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. IJCV .

Lee, C.c., Lee, W.K., Wu, C.C., Lu, C.F., Yang, H.C., Chen, Y.W., Chung, W.Y., Hu, Y.S., Wu, H.M., Wu, Y.T., Guo, W.Y., 2021. Applying artificial intelligence to longitudinal imaging analysis of vestibular schwannoma following radiosurgery. Scientific Reports 11, 3106. doi:10.1038/s41598-021-82665-8.

Li, H., Hu, D., Zhu, Q., Larson, K.E., Zhang, H., Oguz, I., 2021a. Unsupervised cross-modality domain adaptation for segmenting vestibular schwannoma and cochlea with data augmentation and model ensemble. arXiv preprint arXiv:2109.12169 .

Li, H., Loehr, T., Sekuboyina, A., Zhang, J., Wiestler, B., Menze, B., 2020. Domain adaptive medical image segmentation via adversarial learning of disease-specific spatial patterns. arXiv preprint arXiv:2001.09313 .

Li, H., Zhang, H., Johnson, H., Long, J.D., Paulsen, J.S., Oguz, I., 2021b. Mri subcortical segmentation in neurodegeneration with cascaded 3d cnns, in: Medical Imaging 2021: Image Processing, International Society for Optics and Photonics. p. 115960W.

Liu, H., Fan, Y., Cui, C., Su, D., McNeil, A., Dawant, B.M., 2021a. Crossmodality domain adaptation for vestibular schwannoma and cochlea segmentation. arXiv:2109.06274 .

Liu, L., Zhang, Z., Li, S., Ma, K., Zheng, Y., 2021b. S-cuda: Self-cleansing unsupervised domain adaptation for medical image segmentation. Medical Image Analysis 74, 102214. URL: https://www.sciencedirect.com/science/article/pii/S1361841521002590, doi:https://doi.org/10.1016/j.media.2021.102214.

MacKeith, S., Das, T., Graves, M., Patterson, A., Donnelly, N., Mannion, R., Axon, P., Tysome, J., 2018. A comparison of semi-automated volumetric vs linear measurement of small vestibular schwannomas. European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery 275, 867–874. doi:10.1007/s00405-018-4865-z. 29335780[pmid].

Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., Christiaens, D., Dutil, F., Egger, K., Feng, C., Glocker, B., Götz, M., Haeck, T., Halme, H.L., Havaei, M., Iftekharuddin, K.M., Jodoin, P.M., Kamnitsas, K., Kellner, E., Korvenoja, A., Larochelle, H., Ledig, C., Lee, J.H., Maes, F., Mahmood, Q., Maier-Hein, K.H., McKinley, R., Muschelli, J., Pal, C., Pei, L., Rangarajan, J.R., Reza, S.M., Robben, D., Rueckert, D., Salli, E., Suetens, P., Wang, C.W., Wilms, M., Kirschke, J.S., Krämer, U.M., Münte, T.F., Schramm, P., Wiest, R., Handels, H., Reyes, M., 2017. ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Medical Image Analysis 35, 250–269. doi:https://doi.org/10.1016/j.media.2016.07.009.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., Feldmann, C., Frangi, A.F., Full, P.M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B.A., März, K., Maier, O., Maier-Hein, K., Menze, B.H., Müller, H., Neher, P.F., Niessen, W., Rajpoot, N., Sharp, G.C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A.A., van der Sommen, F., Wang, C.W., Weber, M.A., Zheng, G., Jannin, P., Kopp-Schneider, A., 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. Nature Communications 9, 5217. doi:10.1038/s41467-018-07619-7.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., Saez-Rodriguez, J., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., 2020. Bias: Transparent reporting of biomedical image analysis challenges. Medical Image Analysis 66, 101796. doi:https://doi.org/10.1016/j.media.2020.101796.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2014. The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging 34, 1993–2024.

Milchenko, M., Marcus, D., 2013. Obscuring surface anatomy in volumetric imaging data. Neuroinformatics 11, 65–75. URL: https://doi.org/10.1007/s12021-012-9160-3, doi:10.1007/s12021-012-9160-3.

Morerio, P., Cavazza, J., Murino, V., 2018. Minimal-entropy correlation alignment for unsupervised deep domain adaptation, in: International Conference on Learning Representations.

Northcutt, C., Jiang, L., Chuang, I., 2021. Confident learning: Estimating uncertainty in dataset labels. J. Artif. Int. Res. 70, 1373–1411. doi:10.1613/jair.1.12125.

van Opbroek, A., Ikram, M.A., Vernooij, M.W., de Bruijne, M., 2015. Transfer Learning Improves Supervised Image Segmentation Across Imaging Protocols. IEEE Transactions on Medical Imaging 34, 1018–1030. doi:10.1109/TMI.2014.2366792.

Orbes-Arteaga, M., Varsavsky, T., Sudre, C.H., Eaton-Rosen, Z., Haddow, L.J., Sørensen, L., Nielsen, M., Pai, A., Ourselin, S., Modat, M., et al., 2019. Multi-domain adaptation in brain mri through paired consistency and adversarial learning, in: Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data. Springer, pp. 54–62.

Ouyang, C., Kamnitsas, K., Biffi, C., Duan, J., Rueckert, D., 2019. Data efficient unsupervised domain adaptation for cross-modality image segmentation, in: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing, Cham. pp. 669–677.

Palladino, J.A., Slezak, D.F., Ferrante, E., 2020. Unsupervised domain adaptation via CycleGAN for white matter hyperintensity segmentation in multicenter MR images, in: Brieva, J., Lepore, N., Linguraru, M.G., M.D., E.R.C. (Eds.), 16th International Symposium on Medical Information Processing and Analysis, International Society for Optics and Photonics. SPIE. pp. 1 – 10. URL: https://doi.org/10.1117/12.2579548, doi:10.1117/12.2579548.

Park, T., Efros, A.A., Zhang, R., Zhu, J.Y., 2020. Contrastive learning for unpaired image-to-image translation. arXiv:2007.15651 .

Paszke, A., Chaurasia, A., Kim, S., Culurciello, E., 2016. Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147 .

Perone, C.S., Ballester, P., Barros, R.C., Cohen-Adad, J., 2019. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. NeuroImage 194, 1–11. URL: https://www.sciencedirect.com/science/article/pii/S1053811919302034, doi:https://doi.org/10.1016/j.neuroimage.2019.03.026.

Pinter, C., Lasso, A., Wang, A., Jaffray, D., Fichtinger, G., 2012. Slicerrt - radiation therapy research toolkit for 3d slicer. Medical Physics 39, 6332/7.

Prados, F., Ashburner, J., Blaiotta, C., Brosch, T., Carballido-Gamio, J., Cardoso, M.J., Conrad, B.N., Datta, E., Dávid, G., Leener, B.D., Dupont, S.M., Freund, P., Wheeler-Kingshott, C.A.G., Grussu, F., Henry, R., Landman, B.A., Ljungberg, E., Lyttle, B., Ourselin, S., Papinutto, N., Saporito, S., Schlaeger, R., Smith, S.A., Summers, P., Tam, R., Yiannakas, M.C., Zhu, A., Cohen-Adad, J., 2017. Spinal cord grey matter segmentation challenge. NeuroImage 152, 312–329. URL: https://www.sciencedirect.com/science/article/pii/S1053811917302185, doi:https://doi.org/10.1016/j.neuroimage.2017.03.010.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520.

Shanis, Z., Gerber, S., Gao, M., Enquobahrie, A., 2019. Intramodality domain adaptation using self ensembling and adversarial training, in: Wang, Q., Milletari, F., Nguyen, H.V., Albarqouni, S., Cardoso, M.J., Rieke, N., Xu, Z., Kamnitsas, K., Patel, V., Roysam, B., Jiang, S., Zhou, K., Luu, K., Le, N. (Eds.), Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data, Springer International Publishing, Cham. pp. 28–36.

Shapey, J., Kujawa, A., Dorent, R., Saeed, S.R., Kitchen, N., Obholzer, R., Ourselin, S., Vercauteren, T., Thomas, N.W., 2021a. Artificial intelligence opportunities for vestibular schwannoma management using image segmentation and clinical decision tools. World Neurosurgery 149, 269–270. doi:https://doi.org/10.1016/j.wneu.2021.03.010.

Shapey, J., Kujawa, A., Dorent, R., Wang, G., Dimitriadis, A., Grishchuk, D., Paddick, I., Kitchen, N., Bradford, R., Saeed, S.R., Bisdas, S., Ourselin, S., Vercauteren, T., 2021b. Segmentation of vestibular schwannoma from mri,

an open annotated dataset and baseline algorithm. Scientific Data 8, 286. doi:10.1038/s41597-021-01064-w.

Shapey, J., Wang, G., Dorent, R., Dimitriadis, A., Li, W., Paddick, I., Kitchen, N., Bisdas, S., Saeed, S.R., Ourselin, S., et al., 2019. An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced t1-weighted and high-resolution t2-weighted mri. Journal of neurosurgery 134, 171–179.

Shin, H., Kim, H., Kim, S., Jun, Y., Eo, T., Hwang, D., 2022. COSMOS: Cross-Modality Unsupervised Domain Adaptation for 3D Medical Image Segmentation based on Target-aware Domain Translation and Iterative Self-Training. URL: https://arxiv.org/abs/2203.16557, doi:10.48550/ARXIV.2203.16557.

Sundaresan, V., Zamboni, G., Dinsdale, N.K., Rothwell, P.M., Griffanti, L., Jenkinson, M., 2021. Comparison of domain adaptation techniques for white matter hyperintensity segmentation in brain mr images. Medical Image Analysis 74, 102215. URL: https://www.sciencedirect.com/science/article/pii/S1361841521002607, doi:https://doi.org/10.1016/j.media.2021.102215.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf.

Varughese, J.K., Breivik, C.N., Wentzel-Larsen, T., Lund-Johansen, M., 2012. Growth of untreated vestibular schwannoma: a prospective study: Clinical article. Journal of Neurosurgery JNS 116, 706 – 712. URL: https://thejns.org/view/journals/j-neurosurg/116/4/article-p706.xml, doi:10.3171/2011.12.JNS111662.

Wang, G., Shapey, J., Li, W., Dorent, R., Dimitriadis, A., Bisdas, S., Paddick, I., Bradford, R., Zhang, S., Ourselin, S., et al., 2019. Automatic segmentation of vestibular schwannoma from T2-weighted MRI by deep spatial attention with hardness-weighted loss, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 264–272.

Wiesenfarth, M., Reinke, A., Landman, B.A., Eisenmann, M., Saiz, L.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2021. Methods and open-source toolkit for analyzing and visualizing challenge results. Scientific Reports 11, 2369. doi:10.1038/s41598-021-82017-6.

Yang, J., Dvornek, N.C., Zhang, F., Chapiro, J., Lin, M., Duncan, J.S., 2019. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation, in: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing, Cham. pp. 255–263.

Yu, F., Zhang, M., Dong, H., Hu, S., Dong, B., Zhang, L., 2021. Dast: Unsupervised domain adaptation in semantic segmentation based on discriminator attention and self-training. Proceedings of the AAAI Conference on Artificial Intelligence 35, 10754–10762.

Yushkevich, P.A., Pashchinskiy, A., Oguz, I., Mohan, S., Schmitt, J.E., Stein, J.M., Zukić, D., Vicory, J., McCormick, M., Yushkevich, N., et al., 2019. User-guided segmentation of multi-modality medical imaging datasets with itk-snap. Neuroinformatics 17, 83–102.

Zakazov, I., Shirokikh, B., Chernyavskiy, A., Belyaev, M., 2021. Anatomy of domain shift impact on u-net layers in mri segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham. pp. 211–220.

Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232.

Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Örjan Smedby, Bian, C., Yang, X., Heng, P.A., Mortazi, A., Bagci, U., Yang, G., Sun, C., Galisot, G., Ramel, J.Y., Brouard, T., Tong, Q., Si, W., Liao, X., Zeng, G., Shi, Z., Zheng, G., Wang, C., MacGillivray, T., Newby, D., Rhode, K., Ourselin, S., Mohiaddin, R., Keegan, J., Firmin, D., Yang, G., 2019. Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge. Medical Image Analysis 58, 101537. URL: https://www.sciencedirect.com/science/article/pii/S1361841519300751, doi:https://doi.org/10.1016/j.media.2019.101537.

Zou, D., Zhu, Q., Yan, P., 2020. Unsupervised domain adaptation with dual-scheme fusion network for medical image segmentation, in: Bessiere, C. (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, International Joint Conferences on Artificial Intelligence Organization. pp. 3291–3298. doi:10.24963/ijcai.2020/455. main track.