Domain Adaptation for MRI Organ Segmentation using Reverse Classification Accuracy

Vanya V. Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitas, Eric O. Aboagye, Andrea G. Rockall, Daniel Rueckert, and Ben Glocker

1 Biomedical Image Analysis Group, Imperial College London, UK
2 Comprehensive Cancer Imaging Centre, Hammersmith Hospital, Imperial College London, UK
3 The Royal Marsden NHS Foundation Trust, London, UK
v.valindria15@imperial.ac.uk

Abstract
The variations in multi-center data in medical imaging studies have brought the necessity of domain adaptation. Despite the advancement of machine learning in automatic segmentation, performance often degrades when algorithms are applied on new data acquired from different scanners or sequences than the training data. Manual annotation is costly and time consuming if it has to be carried out for every new target domain. In this work, we investigate automatic selection of suitable subjects to be annotated for supervised domain adaptation using the concept of reverse classification accuracy (RCA). RCA predicts the performance of a trained model on data from the new domain and different strategies of selecting subjects to be included in the adaptation via transfer learning are evaluated. We perform experiments on a two-center MR database for the task of organ segmentation. We show that subject selection via RCA can reduce the burden of annotation of new data for the target domain.

1 Introduction
Machine learning has led to significant advances in medical imaging, particularly with big improvements in medical image segmentation. Performance, however, depends on the availability of sufficient amounts of labeled samples for supervised learning, and also whether the test data is coming from the same domain as the training data. In clinical practice, the source domain (S) on which the classifier is trained might be different from the target domain (T) with clinical data. The images from these domains are samples from different appearance distributions. The mismatch of distributions is caused by various factors such as the use of different scanners, types of sequences, or biases in patient cohorts - often causing a trained algorithm to perform poorly on new data. In a scenario where the tasks are the same, but the source and target domains are different, domain adaptation is usually performed to address the domain disparity problem [6].

Domain adaptation can be categorized into three settings, supervised, semi-supervised, and unsupervised [7]. Our work focuses on supervised domain adaptation methods, which uses labeled data from the target domain. In the context of convolutional neural networks (CNNs), supervised domain adaptation can be approached by training from scratch or fine-tuning a network pre-trained on the source domain [8].

An approached called DALSA [3] explored a supervised domain adaptation by introducing a weighting scheme in Random Forests and SVMs for segmentation. This approach preserves the segmentation quality even though only sparse annotated data is used. Another approach tackles the segmentation difficulty of images from different scanners and imaging protocols by weighting different SVM-based...
classifiers for transfer learning [12]. More recent work has tried to explore transfer learning in CNNs in medical imaging [9, 10], which confirmed the potential of fine-tuned and fully trained CNNs.

However, the importance of subject selection in transfer learning has not been widely studied yet. Most works [1, 7, 14] attempt to integrate active learning with domain adaptation. Intuitively, active learning is chosen since it develops a criterion to determine the “value” of a candidate for annotation. In active domain adaptation, the classifier selects among the limited labels on target data [7] by combining hybrid oracles. However, obtaining the oracles is costly. Instead of selecting instances to be added to the training set, [1] selects a bag of instances by self-training. But, this approach is more effective when the source and target domains differ substantially. An appealing work is a recent application of active learning in domain adaptation for biomedical images [14]. Active learning chooses the candidates with higher entropy and higher diversity, which are expected to improve the current performance. In [14], a pre-trained CNN is further fine-tuned continuously by incorporating newly annotated samples in each iteration to enhance the performance incrementally. Although they can reduce the annotation cost, the iterative scheme can be time consuming, especially if applied to volumetric data. Previous work on subject selection with active learning [1, 7, 14] requires iterations and only deals with classification of 2D images.

Instead of using iterative active learning, we propose a framework to select the “most valuable” samples from the unlabeled target domain to be annotated - using reverse classification accuracy (RCA) [11]. We address the question whether RCA can be employed to select few subjects to reduce the cost of annotation in supervised domain adaptation. To answer this question, we systematically conducted several experiments. Our contributions are: (1) we proposed the effective use of RCA as a selector for \( n \)-subjects in target domain to be incorporated into the training dataset, (2) we compare different strategies for supervised domain adaptation with the RCA selection, (3) we study how the training size and the combination of target samples affects segmentation performance.

## 2 Materials and Methods

### 2.1 Datasets

**Source domain (S):** The dataset is obtained from our MALIBO study (MAchine Learning In whole Body Oncology) and includes abdominal T1-weighted MR Dixon images of 35 healthy subjects. We consider this dataset as the source domain used to train the initial classifier in a supervised manner as manual organ annotations are available for all subjects. The images have size of \((256 \times 208 \times 202)\) and resolution \((1.64 \times 1.64 \times 5)\) mm.

**Target domain (T):** Data for the target domain is obtained from the UK Biobank (Application 12579). We use 45 subjects with manually annotated T1-weighted Dixon MR images which have been acquired with a similar protocol as the MALIBO data. The main obvious differences are the image size \((224 \times 168 \times 366)\) and resolution \((2.23 \times 2.23 \times 3)\) mm.

Source and target dataset are acquired at different centers. UK Biobank data is acquired with a Siemens 1.5T MAGNETOM Aera scanner while in MALIBO a Siemens 1.5T MAGNETOM Avanto was used. For our study, we use the T1-weighted in-phase images from the Dixon protocol. We focus on organ segmentation within whole-body scans. As a pre-processing, image intensities in both datasets are normalized to zero-mean and unit-variance. Images are resampled to the same size and physical resolution. Visual examples from the source and target database are depicted in Figure 1. Despite the similarity of the scanning protocol and same scanner manufacturer, the drop in segmentation accuracy when applying a model trained on MALIBO and tested on UK Biobank data is striking, as we will show in our experiments. The images seem to encode a significant bias in their appearance which is not obvious upon visual inspection.

### 2.2 Reverse Classification Accuracy

Reverse classification accuracy (RCA) is a recent method for predicting the segmentation accuracy in the absence of ground truth [11]. RCA can accurately predict the quality of a segmentation on a case-by-case basis without the need for labeled testing set. Here we apply RCA with the single-atlas registration classifier as described in [11]. A set of training images are registered to the test-image, then those transformations are used to warp the corresponding segmentation labels in the training
Figure 1: Examples of whole-body scans from source (MALIBO) and target (UK Biobank) database.

Figure 2: Overview of the strategies we tried for \( n \)-subject selection in DARCA: 1. Training from scratch. 2. Fine Tuning. 3. Iterative DARCA

set. The Dice similarity coefficient (DSC) is calculated between the warped segmentations and the predicted test segmentation, and the maximum DSC is the predicted DSC of the test segmentation. In this scheme, RCA acts as a selector for picking up subjects with high and low confidence of segmentation accuracy. Our hypothesis is that transfer learning with specific \( n \)-subject selection is better than picking-up random subjects from target domain. Therefore, fewer subjects are needed to be manually labeled in the target domain. In the following, we call our approach domain adaption using RCA, or DARCA.

2.3 Supervised Domain Adaptation

In our experiments, we employ DeepMedic [4] as the base network for 3D organ segmentation. We use the default 11-layers deep, multi-scale, parallel convolutional pathways architecture. For more detail\(^1\), please refer to [4].

The main approaches of supervised domain adaptation with CNNs are either training from scratch or fine-tuning [10]. Using RCA as a selector, selected \( n \)-subjects from the target domain are added to the training data. With the new training data containing source and \( n \)-target (S+T) subjects, we train the network from scratch. Meanwhile, we can transfer the parameters from a pre-trained network and fine-tune on another database. To build the model, we fine-tune the pre-trained network (from S-only training), with the \( n \)-target subjects selected by RCA. Based on the results in [2], fine-tuning the last layer achieves the best performance compared to using more convolutional layers fine-tuning. Therefore, we fine-tuned only the last layer of the pre-trained networks and used the same optimization but with fewer epochs.

\(^1\)Source code available at [https://github.com/Kamnitsask/deepmedic](https://github.com/Kamnitsask/deepmedic)
3 Experiments and Results

We present results for using different strategies to investigate the effect of RCA-based subject selection for domain adaptation, as shown in Figure 2. In all our experiments, we use 3-fold cross-validation in both domains. As the baseline, we trained the network with all S data and tested it to segment the T images. Then, we predict the DSCs of all target segmentation using RCA. After we sort their DSCs (lowest to highest confidence), we select \( n \)-subjects from T domain to be included with their corresponding manual annotations in the training set.

### 3.1 Training from scratch

We set our baseline (the lower bound) as the T segmentation from S-only trained network, whereas the upper-bound is the T segmentation from T-only trained network on liver segmentation, as shown in Table 1.

In Figure 2-1 showed that RCA selects \( n \)-subjects from T domain, to be manually labeled and incorporated into training dataset. For this experiment, we compared best-/worst-5 subjects selected by RCA and the real best-/worst-5 (real DSCs from the target ground-truth). Besides, we also run random-5 subject selection (repeated with three different random combination and taking the average) to be trained from scratch.

In training-from-scratch strategy, we train S+T data simultaneously with the same optimization, update-rule, number of epochs, loss function, and regularization techniques as in the baseline. From Table 2 column one, we can see that picking up 5 subjects from target domain has already improved the accuracy, compared to the baseline. Table 2 shows that the segmentation accuracy using the best-5 or worst-5 subject selection outperformed the random selection. Moreover, RCA selection gives a relatively similar segmentation accuracy to the real selection.

As we increase the number of annotated target images to be incorporated in training from scratch, the accuracy does improve, as shown in Table 3. As we combined 5 target subjects with highest confidence and 5 subjects with lowest confidence into the training. The result shows that this combination achieves similar accuracy to the one that incorporated all annotated T subjects. Hence, we reduce the “less valuable” samples for training by RCA selection. Unfortunately, training from scratch requires more time since the networks must learn from the beginning.

### 3.2 Fine tuning (FT) a pre-trained network

In Figure 2-2, we fine-tune the pre-trained network (from S), with \( n \)-subjects selected by RCA. Fine-tuning requires less effort and time than training from scratch. Here, we fine-tuned with three different selections (random, real, and by RCA) at different set size (2, 5, 10, 15, and all of the T-data). From Table 3, the accuracy results when we fine-tuned using all of the T data is very similar (DSC: 0.830) to when we trained from scratch (DSC: 0.831).

However, RCA seems worse to predict the real segmentation accuracy of fine-tuning. There is a gap between best-/worst-5 real selection and best-/worst-5 RCA accuracy (Table 2). One of the reasons could be the under-estimation of RCA prediction on the baseline segmentation accuracies on T, with 0.88 correlation and 0.15 MAE. As explained in the experiments of multiple-organs segmentation [11], the RCA prediction for organs with the real DSCs between (0.6 - 0.8) is not as accurate as in organs with real DSCs above 0.8. In our case, the average real DSCs of baseline liver segmentation is 0.639, but RCA under-estimated the mean of predicted DSCs to be 0.497.

As we did in Section 3.1, we also combined best-5 and worst-5 in T domain to be annotated and incorporated in fine-tuning. This 10 combined subjects give much better accuracy than when we choose only best-10 subjects (Table 3). Best-5 and worst-5 annotated samples with real selection

<table>
<thead>
<tr>
<th>Train on S (baseline)</th>
<th>DSC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.639 (0.149)</td>
</tr>
<tr>
<td>Train on T (upper-bound)</td>
<td>0.873 (0.046)</td>
</tr>
</tbody>
</table>

Table 1: Baseline and upper-bound accuracies
<table>
<thead>
<tr>
<th>Strategies</th>
<th>Training from scratch</th>
<th>Fine tuning</th>
<th>Iterative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.639</td>
<td>0.639</td>
<td>0.639</td>
</tr>
<tr>
<td>All T</td>
<td>0.831</td>
<td>0.830</td>
<td>n/a</td>
</tr>
<tr>
<td>Random 5</td>
<td>0.720</td>
<td>0.710</td>
<td>n/a</td>
</tr>
<tr>
<td>Worst 5 (real)</td>
<td>0.797</td>
<td>0.619</td>
<td>n/a</td>
</tr>
<tr>
<td>Worst 5 (RCA)</td>
<td>0.799</td>
<td>0.771</td>
<td><strong>0.828</strong></td>
</tr>
<tr>
<td>Best 5 (real)</td>
<td>0.747</td>
<td>0.723</td>
<td>n/a</td>
</tr>
<tr>
<td>Best 5 (RCA)</td>
<td>0.755</td>
<td>0.687</td>
<td>0.777</td>
</tr>
<tr>
<td>Best 5 and Worst 5 (Real)</td>
<td>0.823</td>
<td><strong>0.842</strong></td>
<td>n/a</td>
</tr>
<tr>
<td>Best 5 and Worst 5 (RCA)</td>
<td><strong>0.831</strong></td>
<td>0.835</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 2: Strategies of DARCA on liver segmentation: Training from scratch, fine-tuning, and iterative scheme (only for 2nd iteration with RCA selection) with different choice of subject selection.

shows the best results (DSC: 0.842), and RCA selection also gives similar result (DSC: 0.835). Additionally, this best-5 and worst-5 combination (real and RCA selection) performs better than when we use all of the annotated T-subjects (DSC: 0.830). Hence, we cut the cost of annotation by 67%, using only 10 selected subjects instead of 30 subjects and achieve a higher accuracy.

### 3.2.1 Pseudo-labels in Fine-tuning

In this experiment we investigate the use of pseudo ground-truth labels in a semi-supervised way. The work of [13] states that deep neural networks can easily fit random labels, so we challenged the roles of pseudo-label (PL) training in domain adaptation. In Section 3.2, we incorporated the \( n \)-subject by fine-tuning with their ground-truth. What if, instead of using the real annotations, we use the predicted labels as pseudo ground-truth, which are the baseline segmentation results - for training. In the previous work by [5], pseudo-labels are used for semi-supervised learning with pre-trained and fine-tuning scheme. Pseudo-labels are defined as labels that have maximum predicted probability and seen as equivalent to entropy regularization, which encourages low density separation between classes.

However, from Table 3.3, it is clear that using pseudo-labels cannot improve the segmentation performance on target domain. Fine-tuning with all of the pseudo-labeled T subjects gives the worst result amongst all. The more noisy data (pseudo-labels) are introduced to the networks, the poorer the segmentation performance. Hence, PL training is not suitable for domain adaptation, since it assumes the baseline classifier is of high quality, while by default it should be considered to be severely suboptimal.

### 3.3 Iterative DARCA

Different from the previous strategies, here, we wish to mimic the active learning domain adaptation [7], where at each iteration, RCA chooses \( n \)-subjects from target domain to fine-tune the baseline networks (see Figure 2-3). At the first iteration, we fine-tuned the baseline network with best-5 subjects selected by RCA. This best-5 RCA networks segment the test images from target domain, which then are predicted the DSCs by RCA. At the second iteration, we fine-tuned the best-5 RCA networks with the the best-5 and worst-5 selected subjects. The result in Table 3.2 shows that second iteration FT with worst-5 subjects gives a higher accuracies than FT with best-5 RCA. Surprisingly, again combination of best-5 (1st iteration FT), and worst-5 (2nd iteration FT) by RCA performs almost the same (DSC: 0.828) as fine-tuning using all of the labeled target data (DSC: 0.830). Additionally, 2nd iteration with worst-5 RCA selection generally improves the 1st iteration (by best-5 RCA selection) accuracies. Hence, with less labeled data, we can save more time with similar results.

### 3.4 Fine-tuning in Right Kidney Segmentation

From the three different strategies of DARCA in liver segmentation, we can see that fine-tuning with DARCA gives better results, less time-consuming (compared to training from scratch), and no iterative scheme needed. Also, from the results in liver segmentation (Table 2), combination of best-5 and worst-5 subjects always gives better or similar results than using all of the T subjects, in
Figure 3: Plot for different n-selection training size

Table 3: Different n-subject selection size in fine-tuning (FT), fine-tuning with pseudo-labels (PL), and training from scratch (S+T)

<table>
<thead>
<tr>
<th>Strategies</th>
<th>0</th>
<th>2</th>
<th>5</th>
<th>15</th>
<th>30</th>
<th>30 (all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT - Random</td>
<td>0.639</td>
<td>0.665</td>
<td>0.710</td>
<td>0.765</td>
<td>0.803</td>
<td>0.830</td>
</tr>
<tr>
<td>FT - Real</td>
<td>0.639</td>
<td>0.684</td>
<td>0.723</td>
<td>0.780</td>
<td>0.750</td>
<td>0.830</td>
</tr>
<tr>
<td>FT - RCA</td>
<td>0.639</td>
<td>0.631</td>
<td>0.687</td>
<td>0.753</td>
<td>0.722</td>
<td>0.830</td>
</tr>
<tr>
<td>PL - Real</td>
<td>0.639</td>
<td>0.625</td>
<td>0.639</td>
<td>0.640</td>
<td>0.589</td>
<td>0.553</td>
</tr>
<tr>
<td>PL - RCA</td>
<td>0.639</td>
<td>0.614</td>
<td>0.640</td>
<td>0.619</td>
<td>0.632</td>
<td>0.553</td>
</tr>
<tr>
<td>S+T Real</td>
<td>0.639</td>
<td>0.692</td>
<td>0.747</td>
<td>0.763</td>
<td>0.786</td>
<td>0.831</td>
</tr>
<tr>
<td>S+T RCA</td>
<td>0.639</td>
<td>0.711</td>
<td>0.755</td>
<td>0.776</td>
<td>0.797</td>
<td>0.831</td>
</tr>
</tbody>
</table>

Figure 4: DARCA fine-tuning in liver segmentation. Combination of lowest and highest RCA prediction can give a better result than fine-tuning with random selection and with all of the target subjects.

Similarly, the best results of left kidney segmentation with FT is achieved when we combined best-5 and worst-5 subject selection (see Table 4). The result (DSC: 0.716 with RCA selection) is better than when we fine-tune with all of the T subjects (DSC: 0.658), and similar to when we train from scratch using all T subjects. This means we could cut the processing and annotation time. Figure 5 depicts some examples on how DARCA-FT with combination of best-5 and worst-5 subject selection improves the baseline and gives the best segmentation accuracies.

4 Discussion and Conclusion

Set size and subject selection are important in domain adaptation, where usually labels are not available in one of the domains. Thus, we explored whether it will be useful to select only the “valuable” subjects by RCA to be annotated. Pseudo-labels, which normally are used in semi-supervised learning and regularization, seem not to be useful in DARCA. Adding pseudo-labels will
introduce more noise confusing the training of the networks and make them incapable to be applied to the new domain.

All of our strategies in DARCA (training from scratch, fine tuning, and iterative) show a consistent result, combination of best-5 and worst-5 subject selection yields best results. RCA selection of those combined subjects also results in a similar accuracy to the real selection, compared to a bigger gap between RCA and real selection in fine-tuning with only best or worst subject selection.

In this scheme, DARCA shows its potential to leverage the highest and lowest confidence subjects, to be incorporated in the domain adaptation process. We demonstrated that DARCA with few labeled data can perform similarly and/or better to full-size labeled target data. In the examples of Figure 4 and Figure 5, DARCA with best-5 and worst-5 subjects shows consistent results across different tasks (liver and kidney segmentation).

In the case of real DSCs between (0.6, 0.8), RCA underestimated the predictions [11]. This led to a different subject selection for transfer learning procedure. In the future, an improvement of RCA prediction for medium level DSCs (0.6-0.8) needs to be investigated so that it can work more accurately. Our study only focuses at a predefined number of selected subjects, and a more thorough exploration needs to be done in future work. Traditional active learning may have a more flexibility regarding the number of “valuable” samples to be included, but it requires iterations, which is time consuming. DARCA only needs RCA to predict the DSCs from the baseline segmentation where we can select the highest and lowest predicted subjects to be incorporated in a quick fine-tuning procedure. Therefore, we can conclude that DARCA could save processing time (no iteration needed) and annotation time with promising results avoiding the need for a large annotated target database.

Acknowledgments

V. Valindria is supported by the Indonesia Endowment for Education (LPDP)- Indonesian Presidential PhD Scholarship programme. K. Kamnitsas is supported by the Imperial College President’s PhD Scholarship programme. B. Glocker received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 757173, project MIRA, ERC-2017-STG).

The MRI data has been collected as part of the MALIBO project funded by the Efficacy and Mechanism Evaluation (EME) Programme, an MRC and NIHR partnership (EME project 13/122/01). The views expressed in this publication are those of the author(s) and not necessarily those of the MRC, NHS, NIHR or the Department of Health. This research has been conducted using the UK Biobank Resource under Application Number 12579.

References


Figure 5: DARCA fine-tuning in left kidney segmentation. Combination of lowest and highest RCA prediction can give a better result than fine-tuning with all of the target subjects.


