Challenges in Domain Adaptation for Medical Image Segmentation

A Study on Generalization of Hip X-Ray Segmentation for Osteoarthritis

Adam Bayle
Supervisors: Jesse Krijthe, Gijs van Tulder, Myrthe van den Berg
EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Adam Bayle
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Gijs van Tulder, Myrthe van den Berg, Xucong Zhang

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Osteoarthritis is a degenerative disease that affects the aging population by degrading the cartilage in the joints. The early and accurate diagnosis of this disease is key to effective treatment. For an early and accurate diagnosis of this disease, clinicians often use X-ray imaging. This allows medical professionals to manually measure the joint space width (JSW) in X-rays images to determine the progression of the disease. This method however proves to be both time-consuming and variable based on the professional. This research addresses the automation of the measurement of the JSW for the hip, using deep learning techniques, to improve precision and efficiency.

The automated measurement of the JSW is challenged by variations in the imaging conditions across different clinical settings. To address these discrepancies and keep a good performance, domain adaptation techniques are used to counter these domain shifts to ensure a consistent JSW segmentation across different imaging domains.

The study investigates whether a specific domain adaptation technique can enhance the accuracy and robustness of deep learning models specifically for femur segmentation in X-ray images across different datasets. A base deep learning model is developed for femur segmentation, and supervised domain adaptation is applied. The study compares the performance of the adapted model with the base model across two different datasets.

Results indicate that supervised domain adaptation does not significantly improve the model's robustness and accuracy in femur segmentation among two different datasets. These unexpected findings suggest that incorporating domain adaptation techniques may not always lead to a more reliable and efficient diagnosis of osteoarthritis, reducing the manual workload for clinicians.

1 Introduction

Osteoarthritis is one of the most common joint diseases that can occur as people get older [5]. Osteoarthritis causes the cartilage of a joint to break down over time. In order to improve the decision-making process for diagnosing osteoarthritis, X-ray images are often used. They allow for a quantitative measurement of the progression of the disease. For the hip joint, the joint area is the space between the femoral head and the acetabulum. This process is typically performed manually by healthcare professionals. However, this manual evaluation is time-consuming as well as variable and subjective, which can affect diagnosis accuracy.

Methods of automated segmentation have been developed to solve this issue. These methods are however not without limitations: One of the key problems to solve in the automated segmentation of the femur in hip X-rays is the variability in imaging datasets. Models trained on one dataset may perform poorly on unseen data from another dataset [7]. This difference in performance is caused by variations in the imaging technique, imaging devices, and patient demographics. Domain adaptation technique aim to tackle this issue by modifying models to generalize better across different datasets.

This paper will focus specifically on utilizing a specific domain adaptation technique known as supervised domain adaptation. This approach involves training models to learn from the specific characteristics of target data, enhancing performance on unseen data of that domain. Many existing domain adaptation techniques have been developed and applied to medical imaging [9], even specifically to hip joint bone segmentation [1]. However, the aforementioned paper focuses on adapting the domain from CT to MRI, while our focus remains on X-ray imaging exclusively. Our study aims to fill this gap by evaluating the

impact of supervised domain adaptation on femur segmentation in X-ray images, with a focus on the model's ability to generalize across datasets from different hospitals. [2] [10] [3]

We hypothesized that employing supervised domain adaptation will significantly enhance the robustness and accuracy of deep learning models for femoral segmentation in X-ray images. This hypothesis is based on several assumptions: Supervised domain adaptation allows models to learn specific characteristics of target datasets, such as variations in imaging techniques. This targeted learning can improve model performance by making model adapt to the unique features of the data it will encounter. Furthermore, by training on multiple datasets with different imaging conditions, models should be able to perform better on unseen data from both domains. Rather than "transfer learning" [8], our aim is to generalize a single model so it can adapt to different target sources at the same time. This would even allow the model to be effective in domains other than the original and the target datasets, without the need to train it again. The resulting segmentation should be more consistent as well, since it is performed by the same model, rather than 2 different ones.

This paper will look into the effectiveness of this technique in ensuring accurate segmentation of the femur in hip joint X-ray images across two datasets. We first present the methodology employed for automatic segmentation on hip X-rays. Subsequently, we evaluate the performance of our automated femur segmentation method on hip X-rays, conducting a series of experiments with various dataset configurations to determine their effect on segmentation accuracy. Finally, we discuss the conclusions drawn from this study, focusing on the implications of the chosen domain adaptation approach and the insights gained from parameter tuning.

2 Methodology

2.1 Deep learning model selection

In this study, we employed the U-net architecture as a base for our segmentation model. U-net is a convolutional neural network (CNN), widely recognized for its efficacy in medical image segmentation. This network is shown to be very robust for segmenting various anatomical structures, across many different image modalities and application domains [6].

The U-Net architecture inFigure 1 is well suited for medical image inputs thanks to its design. Its use of both down-sampling (contracting) and up-sampling (expansive) paths allows it to capture and use image information at different scales, for a very precise segmentation. The contracting path gradually reduces the dimensions of the input image, extracting high-level features, for example the general shape and surrounding tissues of the femoral head. The expansive path returns the previous result to the size of the original input, to produce a highly detailed segmentation. The U-Net model is also very adaptable and flexible, which allows us to easily modify its parameters to perform segmentation on different structures or image types. This makes the U-Net architecture an ideal choice as base model for our study, as it should ensure a high performing segmentation of the femur for any dataset.

2.2 Data sets

For this research, we had access to two datasets: the Cohort Hip and Cohort Knee (CHECK) dataset and the Osteoarthritis Initiative (OAI) dataset. Both datasets include X-Ray images of the hip, although the OAI dataset focuses specifically on subjects at risk of OA. This make

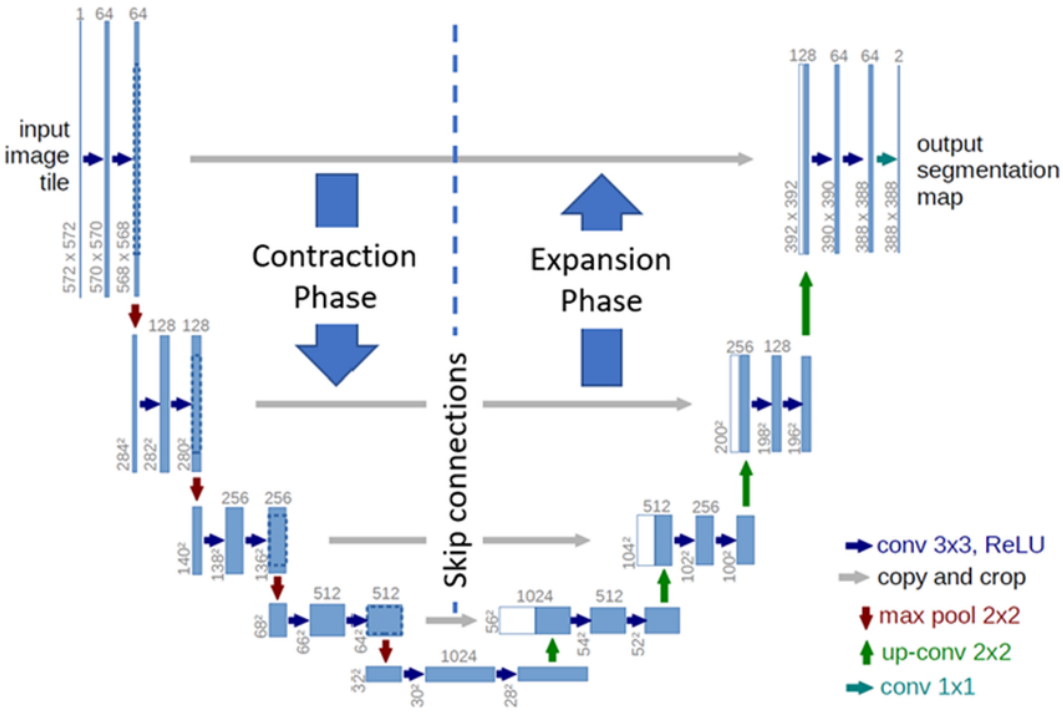

Figure 1: U-Net architecture diagram. Source: Ronneberger et al., 2015.

the datasets perfect for evaluating and improving our segmentation models across different domains.

2.3 Data Preprocessing

Data preprocessing is an important step in preparing datasets for training or testing our segmentation model. We made sure that all input images are in a suitable and comparable format, to reduce some of the high variance in image format. This was done through three steps:

The first step was to normalized the images. To keep pixel intensities consistent across different images, we applied a normalization operation. This involved scaling the pixels intensity values to a fixed range, [0,1]. This allowed the model to focus on anatomical features rather than intensity differences.

The second step was image cropping and resizing. For consistency in input size across any dataset, and to match the input requirements of the U-Net model, all images were resized to 256x256 pixels. The images were also cropped to only include the left side of the hip. This way, the network needs less compute time thanks to lower image size as well as being able to handle input images without doing any modifications.

The third step was to provide each input image with its corresponding ground truth mask. This step was accomplished using the Bonefinder program, which generates points on the bones in the hips to create the ground truth masks. These masks were also cropped and resized to match the input images.

Thanks to these three steps, all the data from both datasets was normalized and resized,

which allows the model to handle input from CHECK and OAI on equal footing. The model also had access to the corresponding ground truth masks for all input images.

2.4 Performance metrics

To evaluate the performance of our segmentation model, we used two metrics: Dice score [11] and Hausdorff distance [4]. The dice score measures the overlap between the predicted and the ground truth segmentations. On the other hand, the Hausdorff distance gives us a way of assessing the scale of the boundary errors. It calculates the maximum distance between points of the predicted and ground truth boundaries. This is especially relevant for medical imaging, where the boundaries are the most important part of the segmentation. Using both these metrics allowed us to measure the overall quality of the segmentation, using overall accuracy and boundary accuracy.

2.5 the Domain shift

After preprocessing the data, and training a simple U-Net model on both CHECK and OAI, it became apparent that the model achieved comparable performance scores on both source and target datasets. This indicated that even after some minimal data preprocessing, the datasets showed little domain differences (domain shifts) in image characteristics. This was measured through two aforementioned metrics, the Dice Score and Hausdorff distance.

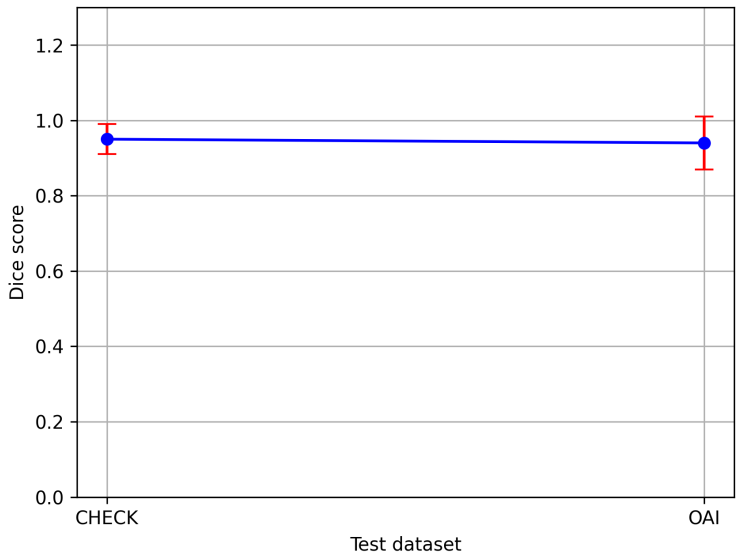

Figure 2: Baseline Dice scores, original datasets

To get these measurements, we trained a model on a train set of CHECK dataset, and tested the performance on a test set of CHECK and a test set of OAI. We observed on this table that the drop in performance on the test set from source domain to target domain is

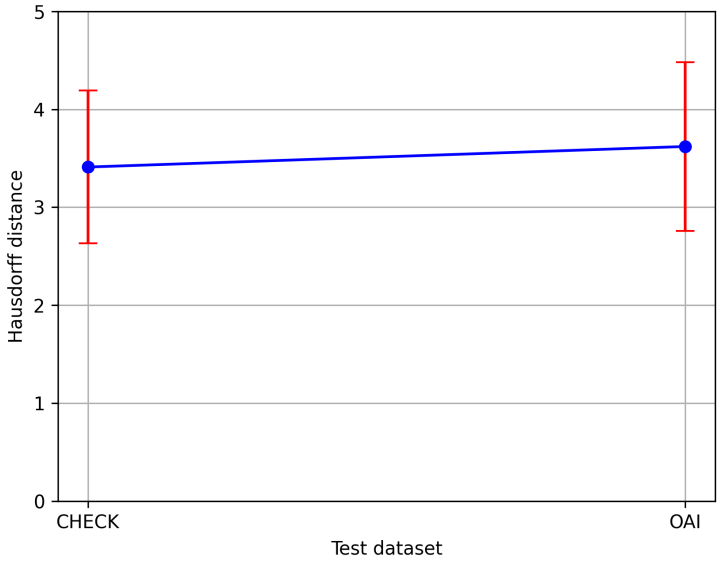

Figure 3: Baseline Hausdorff distances, original datasets

insignificant <2% . This meant that conducting a domain adaptation experiment on two datasets that already share so many features would be unnecessary.

To resolve this issue, we decided to create a controlled domain shift ourselves. to this end, we employed the gamma transformation, which alters the intensity values of an image to simulate different imaging conditions. This transformation adjusts the brightness and contrast of the images by remapping pixel intensities to different values. The gamma transformation function is defined as:

$$I_{\text{out}}(x, y) = \left(\frac{I_{\text{in}}(x, y)}{255} \right)^{\gamma} \cdot 255$$

where:

$$\begin{aligned} I_{\text{in}}(x, y) &\text{ is the input intensity value at pixel } (x, y), \\ I_{\text{out}}(x, y) &\text{ is the resulting output intensity value at pixel } (x, y), \\ \gamma &\text{ is the gamma parameter.} \end{aligned}$$

This means we can alter the look of the input images without changing the maximum and minimum intensities of them, effectively keeping them normalized in the [0,1] range.

We chose values of 0.25, 0.5 and 1. This is firstly to keep the value in the [0,1] range, effectively making the images darker for all inputs as we can see in Figure 7, Figure 5, and Figure 6. Secondly, we did not use values under 0.25, because it resulted in images where the femur was too difficult to outline, even manually, as seen in Figure 7. As we can observe in Figure 9 and Figure 10, by applying gamma values of 0.25, 0.5, and 1.0, we can generate datasets with high, medium and very low domain shifts respectively. This translated as a drop in mean dice score, and an increase in the standard deviation of the Hausdorff distance.

Figure 4: Adjusted image with gamma=1.0

Figure 5: Adjusted image with gamma=0.5

Figure 6: Adjusted image with gamma=0.25

Figure 7: Adjusted image with gamma=0.10

We can also observe that even for highly modified OAI datasets, training and testing on them still yields high performance scores, as seen on Figure 8. This means the features on images of these datasets are still learnable, and the drop in performance is not entirely caused by a degradation in overall image quality.

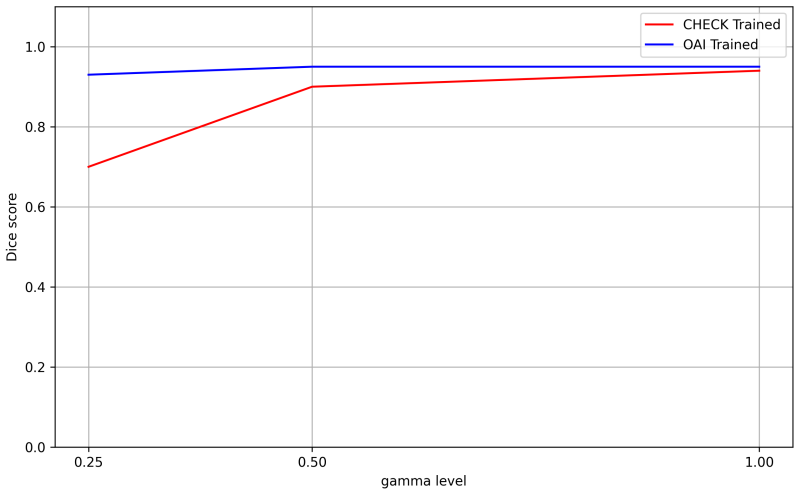

Figure 8: Dice scores of training and testing on OAI for different gamma values

This method allowed us to artificially introduce domain differences that were not present in the original datasets, enabling us to evaluate the model's ability to adapt to different degrees of domain shift.

2.6 Adaptation to the shift

Domain adaptation techniques allow training models to 'adapt' to a target domain where the feature distribution differ significantly. In our study, we chose supervised domain adaptation, as opposed to unsupervised domain adaptation, given its ability to utilise labeled data from the target domain.

The process involved training a model on both a labeled source dataset as well as a smaller part of the target's dataset. By using some of the target's data in the training, the model can directly learn from the target domain's specific characteristics, reducing the effects of the domain shift on the performance.

In many realistic scenarios, acquiring a small amount of labeled data from the target domain is feasible (manually annotated data for example), especially in medical imaging environment. Having access to labels can greatly speed up the model's adaptability to the new features. This way, the model becomes more robust and apt to handle inputs from both the source and target domains, using minimal target data in the training process.

3 Experimental setup and Analysis of results

3.1 Data setup

To evaluate the effectiveness of supervised domain adaptation against domain shifts, we conducted a series of experiments using the CHECK and OAI datasets. Through these experiments, we measured the model's adaptability to various degrees of domain shifts induced by gamma transforms. Each model was also be trained with different levels of domain adaptation, which allows for a general overview of the performance of the method.

We used a set of 700 input images from each dataset. Each dataset was further subdivided into training set, validation set and testing set, at a ratio of 65/15/20 respectively. Every input was normalized to the [0,1] pixel intensity range, and was resized to 256x256. The CHECK dataset was used as the source domain, while the OAI dataset was used as target domain, which was shifted using the gamma transformation. We used values of 0.25, 0.5 and 1.0 of gamma to make 3 different OAI datasets with different domain shifts.

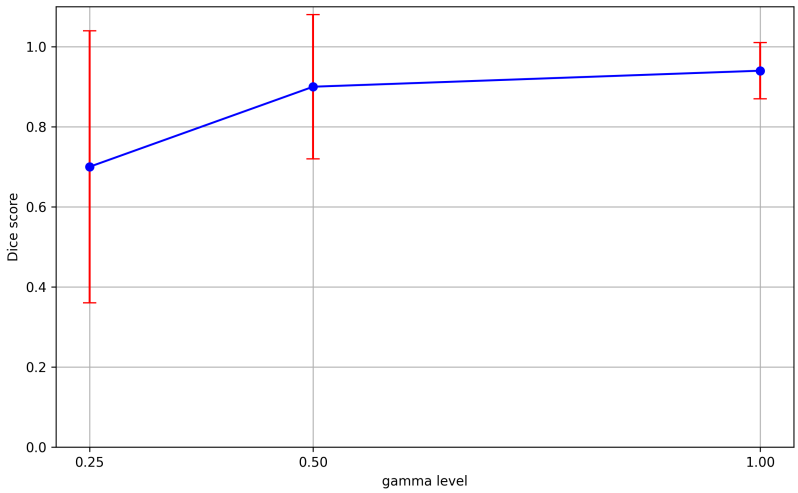

Figure 9: Baseline Dice scores, shifted datasets

3.2 Training process

Firstly, the U-Net model was initially trained on the CHECK dataset. This provided a baseline performance for the source domain. Then, for the domain adaptation part, we introduced batches of target data (OAI dataset) into the training process. We completed this process of training the model using both source and target batches 7 times: Using 0%, 1%, 2%, 4%, 10%, 20% and 50% of its training data from the target dataset. This means for instance that for the 10% OAI DATA model, every time 10 batches from the source dataset was loaded, 1 batch from the target dataset was loaded. We performed this training with all combinations of % of OAI data and gamma values (different target datasets). We then repeated the process using balanced weighting of the batches.

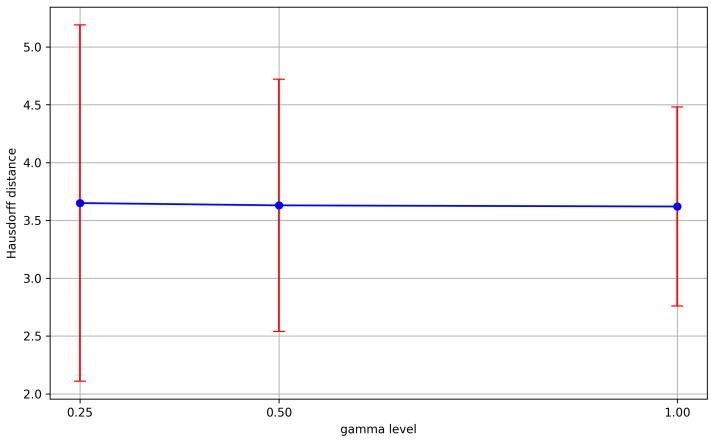

Figure 10: Baseline Hausdorff distances, shifted datasets

Then after 10 epochs, the resulting model was tested on both the test set of CHECK and OAI. This way, we could assess how the performance evolved with different training and target dataset. We trained the models using Root Mean Square Propagation as the loss function, and validated the training step using a combination of Dice score and Binary cross entropy. The model's performance was evaluated by comparing the predicted segmentation to the ground truth, as seen in Figure 12 and Figure 13 using the aforementioned Dice coefficient and Hausdorff distance.

Figure 11: Sample input image from OAI

Figure 12: Ground truth Mask

Figure 13: Predicted segmentation mask

The results of our experiments are summarized in Figure 14. The full table including additional experimental conditions and results is provided in the Appendix.

3.3 Assessing baseline performances

The partial overview of the results in Figure 14 offers a summary of the performance of 30 different models and their performances on both the CHECK test set and the OAI test set.

batch weight balancing	Test set	Gamma	% of OAI data									
			0		4		10		20		50	
			DCE	HD	DCE	HD	DCE	HD	DCE	HD	DCE	HD
unbalanced	CHECK	1	0.95 ±0.04	3.41±0.78	0.95±0.04	3.53±0.71	0.94±0.04	3.56±0.63	0.94±0.04	3.53±0.71	0.94±0.07	3.66±0.88
		0.5	0.95 ±0.04	3.41±0.79	0.90±0.08	3.89±0.80	0.93±0.05	3.69±0.70	0.94±0.05	3.68±0.70	0.91±0.08	3.83±0.60
		0.25	0.95 ±0.04	3.41±0.80	0.95±0.05	3.50±0.69	0.94±0.07	3.54±0.84	0.91±0.10	3.93±0.76	0.75±0.26	4.07±1.15
	OAI	1	0.94±0.07	3.62±0.86	0.93±0.07	3.94±1.39	0.93±0.10	3.74±1.55	0.92±0.08	3.94±1.46	0.92±0.11	3.98±1.41
		0.5	0.90±0.18	3.63±1.09	0.73±0.29	4.13±1.52	0.88±0.15	3.87±1.01	0.91±0.12	3.99±1.01	0.89±0.12	4.03±0.76
		0.25	0.70±0.34	3.65±1.54	0.73±0.33	3.68±1.68	0.77±0.31	3.63±1.34	0.78±0.24	4.56±1.36	0.76±0.27	4.26±1.37
balanced	CHECK	1	0.95 ±0.04	3.41±0.84	0.94±0.05	3.67±0.95	0.92±0.07	3.82±0.84	0.94±0.05	3.58±0.90	0.94±0.07	3.66±0.88
		0.5	0.95 ±0.04	3.41±0.85	0.91±0.07	3.86±0.79	0.92±0.09	3.71±0.71	0.90±0.07	3.96±0.77	0.93±0.09	3.60±0.83
		0.25	0.95 ±0.04	3.41±0.86	0.93±0.05	3.71±0.81	0.94±0.04	3.52±0.65	0.93±0.05	3.70±0.64	0.67±0.32	3.83±1.23
	OAI	1	0.94±0.07	3.62±0.86	0.92±0.11	3.64±0.80	0.90±0.12	3,97±0.98	0.94±0.07	3.63±0.67	0.92±0.11	3.63±0.79
		0.5	0.90±0.18	3.63±1.09	0.80±0.25	4.39±1.46	0.84±0.21	3.97±1.10	0.79±0.24	4.30±1.33	0.92±0.10	3.68±0.93
		0.25	0.70±0.34	3.65±1.54	0.66±0.34	3.80±1.65	0.77±0.29	3.74±1.44	0.77±0.27	4.28±1.46	0.74±0.28	4.10±1.45

Figure 14: Partial Results Overview

To evaluate the effectiveness of the domain adaptation technique, we first need to assess the baseline performances of the model with no adaptation.

We can once again observe that for gamma values of 1 and no domain adaptation, the Dice score remains stable with a very high mean, consistently around 0.95, which confirms that the original domain shift between the CHECK and OAI dataset is not significant enough. The Hausdorff distances for the same row did show more variability, but the differences are not substantial enough to justify a need for domain adaptation.

This is reflected in Figure 14 in the column with 0% of OAI data. We can observe that on the CHECK test set, the results are as expected very high, 0.95 with very low variance. On the OAI set however, we can see the impact of the domain shift, with scores of 0.94, 0.90 and 0.70 for gamma values of 1, 0.5 and 0.25 respectively. This can be considered as a baseline for the model's adaptability to the OAI features. For the Hausdorff distances, the degradation in adaptability is reflected in the variance of the results. We also have results of training a model exclusively on OAI data, giving us an idea of what an optimal performance on this domain would be, as seen on Figure 8. We expect further results on the OAI set with different degrees of domain adaptation to improve upon the baseline scores, up to this upper optimal threshold.

3.4 Effectiveness of Domain Adaptation

We observe that for gamma values of 0.5, different degrees of domain adaptation, the Dice score did not show any significant improvement of the segmentation on the OAI set. The domain adaptation even significantly worsened the scores when using 4% of target data. The same general pattern is observed for gamma values of 0.25, although it showed a slight improvement of 5% for higher percentages of utilized target data.

As expected for the CHECK test set, injecting more and more of the target's data into the training did consistently result in a drop in Dice score and a rise in Hausdorff distance variance. The drop in performance is most noticeable for a gamma value of 0.25, where the domain shift is the largest, as we can observe in Figure 15 and Figure 16.

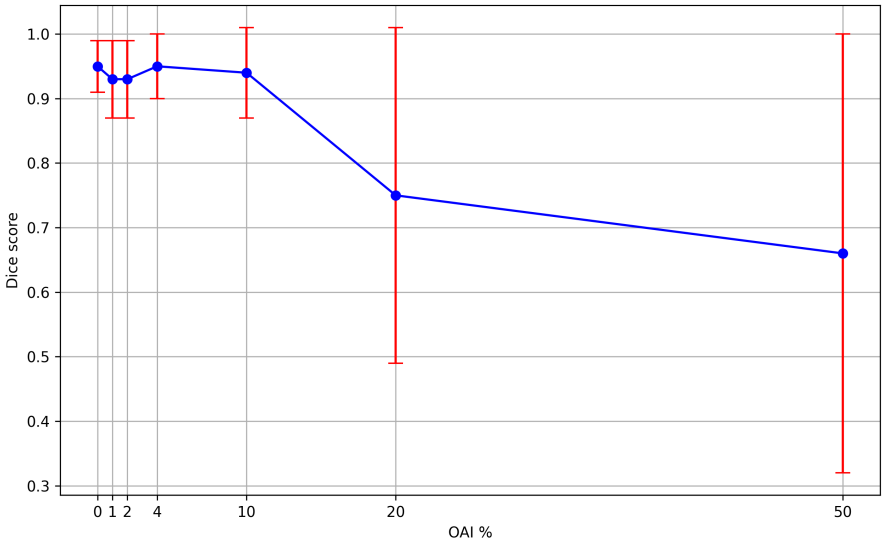

Figure 15: Dice score on CHECK test set, gamma=0.25

3.5 Effectiveness of load balancing

We observed that load balancing (compensated target batch weight) did not provide any performance improvements over using regular batches. For instance, we can observe on Figure 17 and Figure 18 that for gamma=0.5, both the mean dice score and the standard deviation of the Hausdorff distance do not see any significant improvement when tested on the OAI test set.

4 Responsible Research

In this study, we made sure to keep patient data confidential. All data is anonimized to protect patient privacy and prevent misuse. Data processing and model training was performed using TU Delft's Delft Blue supercomputer, supplemented with local computations for some smaller scale experiments. This ensured the data remained secure and private during the training process.

To ensure reproducilibty of the experiments, we kept the data splitting by using a fixed random seed. This way, the different model's performances can be fairly compared for a good analysis, and the same sets can be recreated for future studies. However it is important to note that we did not eliminate all randomness in the training process, such as the initial weights of the neural network. This introduced slight variations in results upon replication. Despite this, we believe our method allows for meaningful replication and validation.

As for the biases involved in training an machine learning model, the primary objective was to evaluate the model's ability to generalize across different domains. We did not observe significant improvements using supervised domain adaptation techniques, which may be caused by the very biases that the model could not overcome. The model might

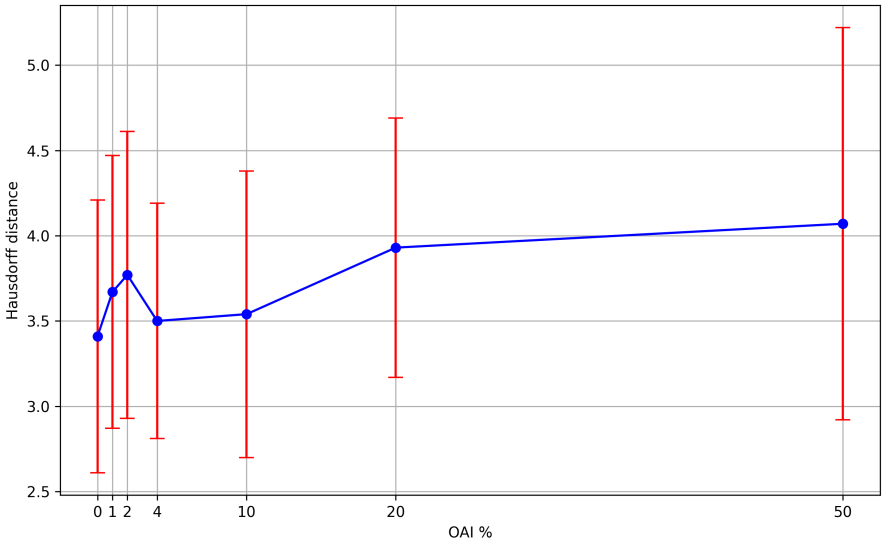

Figure 16: Hausdorff distance on CHECK test set, gamma=0.25

have focused on different features from one dataset to another, leading to performance issues. The use of adaptive batch weight aimed to reduce training bias, but did not yield the expected improvements, and may even have worsened these biases. This underlines the need for more complex or sophisticated approaches to tackle these domain biases.

We acknowledge these limitations, and future research should work to ensure model fairness, especially in the topic of domain adaptation for medical imaging.

5 Discussion

The study aimed to evaluate whether supervised domain adaptation could improve a segmentation model's generalization across two different datasets. The findings show that the implemented technique provided marginal improvements, if not worsened the performance in some cases.

The assumption that the model could generalize across different domains simultaneously may have been too optimistic. The domain specific features significantly impacted the model's performance and learning both was not an effective solution. The limited improvements we observed did not bridge the measured performance gap.

Compared to previous research, which has proven the potential success of supervised domain adaptation, our results show the limitations of the current approach. These limited improvements can be explained. The use of whole batches of target data may have led to conflicting learning directions, making it difficult for the model to learn domain invariant features, focusing instead on one at the time during training. Using mixed batches or other ways of injecting the target data into the training process could provide a smoother learning curve, leading to better performances.

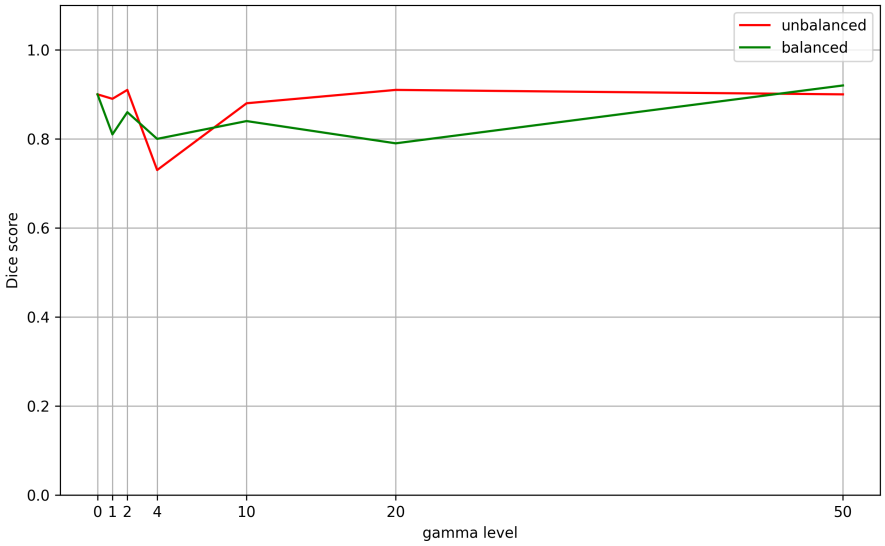

Figure 17: Dice score on OAI test set, gamma=0.5

Furthermore, load balancing did not improve on this issue. Using large weights to multiply the training loss for the target batches may have worsened the learning process. These weights could have caused over fitting on the target domain, making the conflict between learning directions even worse. Resampling target images instead may help fix this issue.

6 Conclusions and Future Work

This research had for aim to ease, make more objective, and speed-up the process of diagnosing osteoarthritis, a common and sever disease. Using deep learning techniques, an automatic segmentation of the joint space on X-ray images of the hip allows for fast and reliable measurements of the joint space width. To improve the model's generalization across diverse patient datasets, we explored the application of supervised domain adaptation techniques, specifically focusing on whether these methods could address domain shifts between datasets. Our findings indicate that our supervised domain adaptation techniques, including adversarial training and supervised domain adaptation, did not significantly enhance model performance for this task.

While previous research demonstrated the potential benefits of supervised domain adaptation, our study suggests that our approach may not be sufficient for the complexity of the task. This could be due to the simplistic nature of the adaptation technique, or potential implementation flaws. While this experiment definitely showed room for improvements for the specific case of segmentation on Hip X-ray images, there are many areas for future studies to explore. First, the integration of large datasets from other hospitals coming from different medical machinery could be useful in validating the robustness of the solutions. Additionally, exploring the use of other completely different imaging mediums, such as MRI

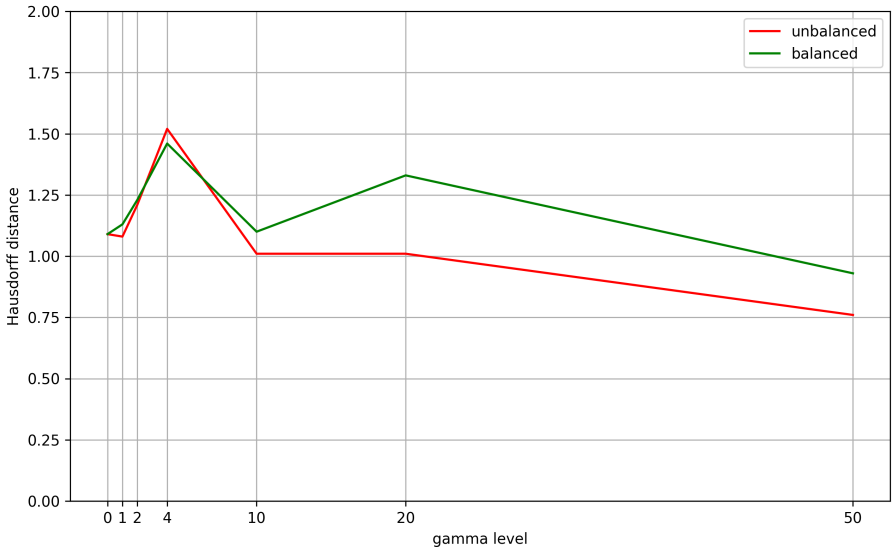

Figure 18: Hausdorff distance on OAI test set, gamma=0.5

or ultrasound, together with X-ray information, could give even more accurate segmentation.

In the case of more successful implementations of supervised adaptation, future research could explore the optimal percentage of target data for model generalization. Knowing what the minimal amount of target data is required to achieve the best performance can guide future implementations and make the adaptation process more efficient. This is because another very important part of the segmentation problem is the inaccuracy or even the lack of ground truth, annotated data. Looking into the training of segmentation models for clinical purposes using minimal annotated data could be an interesting path to take. This way, there would be less need for clinicians to manually annotate enormous amounts of data, saving a lot of time.

Future research could also test the effectiveness of these techniques in segmentation tasks on other joint spaces affected by osteoarthritis, such as the knees and hands. While our current study focuses on femoral segmentation in hip X-ray images, the principles and methodologies employed, particularly in deep learning and domain adaptation, may generalize to similar segmentation tasks in other joint areas.

In conclusion, this study demonstrates the limitations of supervised domain adaptation techniques to enhance the accuracy and robustness of deep learning models for femoral segmentation in hip X-ray images, as our results indicate that the current approach may not be sufficient. Future research could also explore other strategies for training segmentation models with even less annotated data. By continuing to experiment in this domain, we can go towards a more efficient and accessible diagnosis tool for osteoarthritis, benefiting patients and healthcare professionals.

A Appendix

batch weight balancing	Test set	Gamma	% of OAI data													
			0		1		2		4		10		20		50	
			DCE	HD	DCE	HD	DCE	HD	DCE	HD	DCE	HD	DCE	HD	DCE	HD
unbalanced	CHECK	1	0.95 ±0.04	3.41±0.78	0.95±0.05	3.44±0.93	0.93±0.06	3.65±0.84	0.95±0.04	3.53±0.71	0.94±0.04	3.56±0.63	0.94±0.04	3.53±0.71	0.94±0.07	3.66±0.88
		0.5	0.95 ±0.04	3.41±0.79	0.94±0.05	3.58±0.68	0.95±0.04	3.44±0.79	0.90±0.08	3.89±0.80	0.93±0.05	3.69±0.70	0.94±0.05	3.68±0.70	0.91±0.08	3.83±0.60
		0.25	0.95 ±0.04	3.41±0.80	0.93±0.06	3.67±0.80	0.93±0.06	3.77±0.84	0.95±0.05	3.50±0.69	0.94±0.07	3.54±0.84	0.91±0.10	3.93±0.76	0.75±0.26	4.07±1.15
	OAI	1	0.94±0.07	3.62±0.86	0.94±0.06	3.66±1.60	0.91±0.10	4.16±1.62	0.93±0.07	3.94±1.39	0.93±0.10	3.74±1.55	0.92±0.08	3.94±1.46	0.92±0.11	3.98±1.41
		0.5	0.90±0.18	3.63±1.09	0.89±0.14	3.94±1.08	0.91±0.14	3.83±1.21	0.73±0.29	4.13±1.52	0.88±0.15	3.87±1.01	0.91±0.12	3.99±1.01	0.89±0.12	4.03±0.76
		0.25	0.70±0.34	3.65±1.54	0.67±0.35	3.83±1.64	0.62±0.35	3.68±1.74	0.73±0.33	3.68±1.68	0.77±0.31	3.63±1.34	0.78±0.24	4.56±1.36	0.76±0.27	4.26±1.37
balanced	CHECK	1	0.95 ±0.04	3.41±0.84	0.95±0.04	3.46±0.91	0.93±0.06	3.67±0.87	0.94±0.05	3.67±0.95	0.92±0.07	3.82±0.84	0.94±0.05	3.58±0.90	0.94±0.07	3.66±0.88
		0.5	0.95 ±0.04	3.41±0.85	0.91±0.06	3.89±0.76	0.94±0.06	3.70±+0.66	0.91±0.07	3.86±0.79	0.92±0.09	3.71±0.71	0.90±0.07	3.96±0.77	0.93±0.09	3.60±0.83
		0.25	0.95 ±0.04	3.41±0.86	0.95±0.04	3.64±0.96	0.94±0.04	3.62±0.62	0.93±0.05	3.71±0.81	0.94±0.04	3.52±0.65	0.93±0.05	3.70±0.64	0.67±0.32	3.83±1.23
	OAI	1	0.94±0.07	3.62±0.86	0.94±0.05	3.76±0.96	0.90±0.11	3.91±0.92	0.92±0.11	3.64±0.80	0.90±0.12	3,97±0.98	0.94±0.07	3.63±0.67	0.92±0.11	3.63±0.79
		0.5	0.90±0.18	3.63±1.09	0.81±0.86	3.93±1.13	0.86±0.22	3.91±1.23	0.80±0.25	4.39±1.46	0.84±0.21	3.97±1.10	0.79±0.24	4.30±1.33	0.92±0.10	3.68±0.93
		0.25	0.70±0.34	3.65±1.54	0.65±0.35	3.74±1.53	0.66±0.34	4.04±1.68	0.66±0.34	3.80±1.65	0.77±0.29	3.74±1.44	0.77±0.27	4.28±1.46	0.74±0.28	4.10±1.45

Figure 19: Complete Results Overview

References

[1] Li Guan and Yizhou Liu. Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 2022.

[2] Konstantinos Kamnitsas and et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *Information Processing in Medical Imaging*, 2017.

[3] A. L. Martel and et al. *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020*. Springer Science+Business Media, 2020.

[4] Anam Nazir, Muhammad Nadeem Cheema, Bin Sheng, Huating Li, Ping Li, Po Yang, Younhyun Jung, Jing Qin, Jinman Kim, and David Dagan Feng. Off-enet: An optimally fused fully end-to-end network for automatic dense volumetric 3d intracranial blood vessels segmentation. *IEEE Transactions on Image Processing*, 29:7192–7202, July 2020. Funding Information: Manuscript received September 25, 2019; revised March 7, 2020 and April 30, 2020; accepted June 1, 2020. Date of publication June 9, 2020; date of current version July 8, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61872241 and Grant 61572316, in part by the Science and Technology Commission of Shanghai Municipality under Grant 18410750700, Grant 17411952600, and Grant 16DZ0501100, in part by the Hong Kong Research Grants Council under Grant PolyU 152035/17E, and in part by The Hong Kong Polytechnic University under Grant P0030419 and Grant P0030929. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xudong Jiang. (Corresponding authors: Bin Sheng; Huating Li.) Anam Nazir, Muhammad Nadeem Cheema, and Bin Sheng are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: shengbin@sjtu.edu.cn). Publisher Copyright: © 1992-2012 IEEE.

[5] World Health Organization. Osteoarthritis, 2023. Retrieved from `https://www.who.int/news-room/fact-sheets/detail/osteoarthritis`.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[7] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176. IEEE, 2017.

[8] Annegreet van Opbroek, M. Arfan Ikram, Meike W. Vernooij, and Marleen de Bruijne. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE Transactions on Medical Imaging*, 34(5):1018–1030, 2015.

[9] Sulaiman Vesal, Nishant Ravikumar, and Andreas Maier. Automated multi-sequence cardiac mri segmentation using supervised domain adaptation. In Mihaela Pop, Maxime Sermesant, Oscar Camara, Xiahai Zhuang, Shuo Li, Alistair Young, Tommaso Mansi, and Avan Suinesiaputra, editors, *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*, pages 300–308, Cham, 2020. Springer International Publishing.

[10] G. Zeng and et al. Icmsc: Intra- and cross-modality semantic consistency for unsupervised domain adaptation on hip joint bone segmentation. *arXiv preprint arXiv:2012.12570*, 2020.

[11] Kelly H. Zou, Simon K. Warfield, Aditya Bharatha, Clare M.C. Tempany, Michael R. Kaus, Steven J. Haker, William M. Wells, Ferenc A. Jolesz, and Ron Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. *Academic Radiology*, 11(2):178–189, 2004.