Anatomy-aware data augmentation techniques in contrastive self-supervised learning for diagnosing hip osteoarthritis in X-ray images

Zhenya Yancheva[1]

Supervisors: Jesse Krijthe[1]**, Gijs van Tulder**[1]

[1]**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 22, 2025

Name of the student: Zhenya Yancheva
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Gijs van Tulder, Michael Weinmann

Abstract

Supervised learning approaches have proven to be useful in diagnosing Osteoarthritis from X-ray images, aiding professionals in an otherwise time-consuming and subjective process. However, in the medical field, labeled data is scarce. For this reason, we investigate a contrastive self-supervised approach, SimCLR, capable of learning useful representations from unlabeled data. Specifically, we explore a core component of this method – the data augmentation techniques. While these augmentations are highly effective in introducing variability in conventional image datasets, they are too aggressive for medical images, often altering their semantic meaning. In this paper, we implement custom anatomy-aware augmentation techniques, which aim to preserve the main region of interest needed for a diagnosis. We evaluate these anatomy-aware augmentations including Gaussian blur, Contrast enhancement, Random resized crop, and Random erasing, against their classical counterparts by training multiple encoders based on different combinations of those augmentations. The findings of our study have shown that utilizing this anatomy-aware approach for all data augmentations a model uses does not lead to a significant improvement in its performance. However, selective use of anatomy-awareness on geometric-based approaches seems to show promising initial results.

1 Introduction

Osteoarthritis (OA) is a degenerative joint disease which causes the bone and connective tissue around a joint to wear down over time. When diagnosing osteoarthritis from patients' X-ray images, a particularly challenging aspect is the subjectivity of the diagnosis, since the stages of the disease may look different between patients. For this reason, the use of machine learning techniques could prove to be incredibly beneficial in automating this tedious task. While supervised methods have been shown to be effective [1], labeled medical data is not easy to acquire in large amounts. In order to avoid this issue, we investigate how to utilize self-supervised methods which would be able to learn from unlabeled data.

Contrastive Self-supervised learning is an approach where feature learning is guided by what is considered "positive pairs" [2] - pairs of the same image that has been augmented in different ways but is still considered to have the same semantic meaning. For this reason, data augmentation is a crucial component of those algorithms. Existing literature suggests that applying data augmentations, which are quite extreme and even introduce unrealistic variation could lead to better performance of contrastive SSL models [3]. However, since medical images are usually taken under standardized protocols and follow a particular format, they have significantly lower variability [2]. For this reason, they differ from the diverse sets of images those models are usually trained on, which benefit from strong data augmentations. The semantic meaning of medical images is usually not identified by the depicted object, but rather by concrete meaningful features, which may easily be erased by random cropping or blurring. One way of avoiding this problem is adapting the data augmentation techniques and making them context-aware. By taking anatomical knowledge into account, those custom augmentations can preserve the important anatomical features an thus - the semantic meaning of the images - the diagnosis.

There have been previous attempts to address the problem of the strength of data augmentations and limit it by prioritizing some areas of the image. Peng et al. have investigated a more curated way of cropping images, which ensures cropped views provide the most relevant information in the image[4]. Using data from radiologists' eye movements, Wang et al. have developed a novel augmentation method and have shown that "semantic-aware augmentation consistently outperforms the conventional way of random augmentation" [5]. Li et al. also utilize a supervised method for detecting zones within the image, which they provide to the classifier as "prior knowledge" [6]. While those approaches have shown to be successful, they rely on supervised models on extract the meaningful part of the image. It is yet to be explored if those regions of interest which the medical professionals pay attention to could be extracted from the images automatically, without the need for collection of additional data or training additional models.

This paper explores the effect of data augmentation methods which preserve the key anatomy structures in a medical image in the context of Contrastive Self-Supervised learning methods. With this goal in mind, we make a comparison between the performance of the same model while using classical data augmentation methods versus anatomy-aware ones, which are more limited in the extent to which they could be applied without erasing the important anatomical features. Additionally, we are interested in finding out which types of data augmentations would benefit from such an anatomically-aware approach.

2 Methodology

2.1 Model Architecture

Multiple different frameworks for implementing contrastive self-supervised learning exist - the most popular among which are SimCLR [7], MoCo [8] and BYOL. All of them use data augmentations in some way and thus, could potentially benefit from the proposed method of making them anatomy-aware. However, SimCLR specifically is more dependent on the strength of those data augmentations, since they are the core of its learning process, while the others implement additional mechanisms to aid it. SimCLR is also most prominent among recent literature and is more simple compared to the other alternatives, making it a good choice for our experiments.

The architecture of SimCLR can be found in Figure 1 and contains the following modules:

- A stochastic **data augmentation module** which takes an image as input and produces two augmented views from it, given a pre-determined set of data augmentations.

- A **neural network base encoder** which takes each of those views and extracts their features into representation vectors. Similarly to the original SimCLR paper, we opt for a ResNet encoder, particularly ResNet18.
- A small neural network **projection head** which maps those representations vectors to a lower dimension space, in which the contrastive loss can be calculated. Again, similarly to the original SimCLR paper, we use a two-layer MLP with ReLU activation in the hidden layer.
- A **contrastive loss function** - normalized temperature-scaled cross entropy loss, which the autors of the SimCLR paper have named NT-Xent. After comparing it to other commonly used contrastive loss functions, they have established it leads to the best performance of the model.

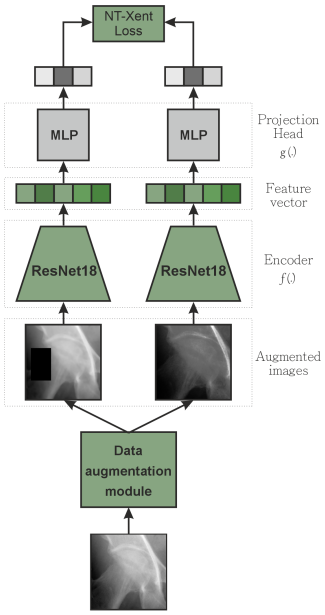

Figure 1: SimCLR training procedure. Images are used as input to a data augmentation module which produces two different augmented views. Each view is encoded and the encoding is then projected to the space where the contrastive loss is applied.

2.2 Data Augmentations in Contrastive Self-Supervised Learning

While numerous data augmentations exist and any of them could be customized or combined, not all of them are suitable for medical imaging. For example, the original SimCLR paper emphasizes that "the combination of random crop and color distortion is crucial to achieve a good performance" [7]. However, the use of color distortion would not be beneficial for medical images which are usually grayscale. It would only introduce a new dimension into the data (additional color channels), which do not carry any semantic meaning, and thus it will not contribute to the learning process.

In this study, we only focus on data augmentations which by their nature carry the risk of concealing important anatomical regions. An example of a data augmentation technique we are not interested in is rotation, since all of the elements of the image remain intact. We will investigate two types of augmentations which we will call geometric and appearance-based. The former remove parts of the image, while the latter preserve the image in its whole but apply filters which impact its appearance.

2.3 Joint Space Segmentation

BoneFinder [9] is a fully automatic software tool which can be used to outline the contours of the skeletal structures from 2D radiographs. It outputs a set of landmark points that trace the curves of the bones as shown in Figure 2. We use this data in order to segment the region in between the femur head and the acetabular roof - it marks the primarily weight-bearing area of the hip joint where the cartilage is more likely to show signs of wear.

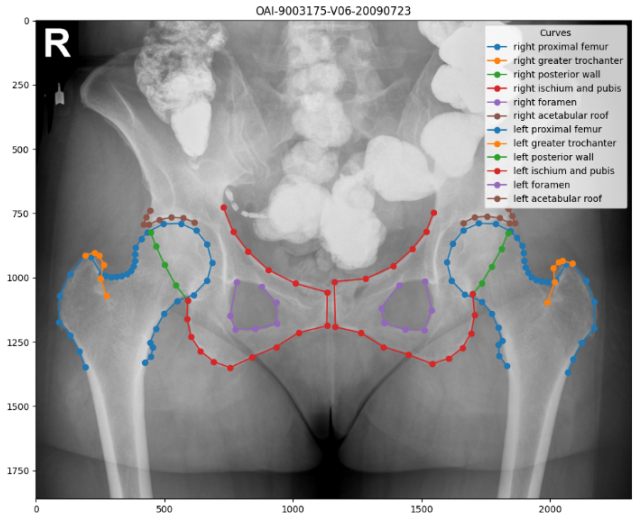

Figure 2: The output of the BoneFinder tool consists of the point coordinates which trace the outline of the bones. Here they have been plotted on the original input image.

2.4 Anatomy-aware data augmentations

X-ray imaging is a standard part of the OA diagnosing process, since it is risk-free, cost-effective and widely available. However, during the early stages of the disease, bone tissue may be unaffected, while the cartilage, which does not show up on X-ray images, is worn out. For this reason, the main way to detect cartilage damage is by the reduction of the space between the bones of the joint [10]. As the disease progresses, other formations withing the bone tissue around the joint - such as osteophytes or cysts, may start to form. This makes the joint space the most important area of the X-ray image for diagnosing OA.

Using the coordinates of the points provided by BoneFinder, which outline the segmented joint space, a bounding box is defined around it. We will consider this area

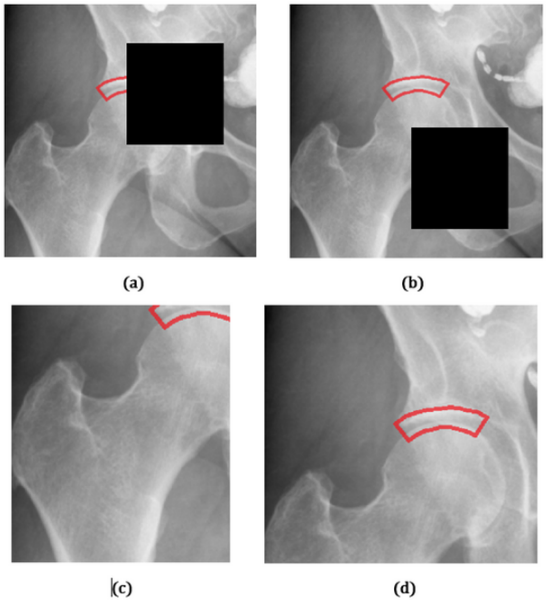

Figure 3: Images (a) and (c) show two views of the same image, which use different augmentation methods (erase and crop respectively) but exclude the joint space segment (marked in red). In (b) and (d) the same methods lead to views that preserve it.

to be the anatomy-relevant space, all of the details of which need to be preserved in order for Contrastive SSL models to better learn the features around the joint area. For this reason, in order to transform a data augmentation technique into an anatomy-aware one, we ensure that the area inside the bounding box remains as in the original image. For geometric data augmentations, such as erase and crop, this would mean ensuring that this region stays in the newly generated image as shown in Figure 3. For appearance-based data augmentations such as Gaussian blur and contrast enhancement, this would entail applying those respective filters over the whole image except inside the bounding box.

3 Experimental Setup

3.1 Dataset and Data Pre-processing

The dataset which will used for this study is taken from the Cohort Hip and Cohort Knee (CHECK) study [11]. It includes X-ray images of 1002 participants from multiple visits over a 10-year period. The data was split in 70% training set / 15% validation set and 15% test set based on the participants IDs. It is important to mention that the number of visits per participant varies due to some of them dropping out early from the study, and thus this ratio for the data split does not exactly represent the ratio between the number of images in each subset. However, this separation based on participant ID is done intentionally in order to ensure independence between the testing and training datasets. It is crucial that images from the same patient do not appear in both, since they may carry certain anatomical features specific to this person and the model may try to learn to differentiate between people rather than detect signs of their disease.

The images have been pre-processed by cropping out the 15 cm by 15 cm region around the hip joint, centered on the center of the femoral head. All images of left hips were flipped in order to reduce variability in the data set that is not relevant to the classification task. All cropped images were resized to 224 by 224 pixels in order to fit the resnet18 input shape. The coordinates of the keypoints found by BoneFinder have also been converted to coordinates in the new cropped images and a bounding box was calculated around those, which outline the hip joint space - specifically points 18 to 22 (for the acetabular roof) and 69 to 74 (for the femoral head), as numbered by the BoneFinder Algorithm. Those selected points can be seen in Figure 4, overlaid on top of the final pre-processed image. After pre-processing, the dataset we will be using consists of 6859 images.

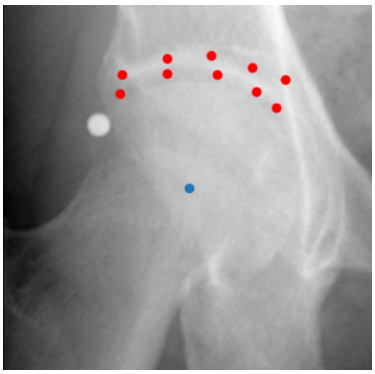

Figure 4: Pre-processed image, centered around the center of the femoral head (in blue). Includes the points (in red) used to segment the bounding box around the joint space.

The original dataset contains a Kellgren-Lawrence (KL) grade [12], which indicates the severity of osteoarthritis based solely on the X-ray images. The 5 grades defined as follows: None (0), Doubtful (1), Minimal (2), Moderate (3), Severe (4). For the purposes of our experiment, those labels have been transformed into two classes - Negative, containing grades 0 and 1, and Positive - containing grades 3 and above. This is done in order to simplify the classification task in order to observe whether the encoders would learn any features relevant to diagnosing OA at all, rather than diagnose the exact stage of the disease - a task which orthopedic surgeons and radiologists often seems to struggle to reach agreement on [13].

3.2 Model Training

In order to compare classic data augmentations to the proposed anatomy-aware ones, four encoders will be pre-trained using different combinations of augmentations among those two classes. We define the following training protocol which ensures the results are a direct reflection of the changes in the data augmentations.

- All 4 models have the exact same architecture as described in Section 2.
- Batch size of 64 is used for loading the data.
- All 4 models have been trained for 100 epochs.

- The Adam optimizer was used with learning rate of 1e-4.

		Base Model	Appearance Model	Geometrical Model	Fully Anatomical Model
Appearance	Contrast enhancement	C	AA	C	AA
	Gaussian Blur	C	AA	C	AA
Geometric	Random Resized Crop	C	C	AA	AA
	Random Erasing	C	C	AA	AA

Figure 5: Data augmentations used for each model. C stands for the Classic data augmentation version, AA stands for the Anatomy-aware alternative. The models have been named based on the subset of augmentations that are anatomy-aware.

The only difference between the models being trained is in the data augmentation techniques they utilize in the first step of pre-training the encoder. The combinations used can be found in Figure 5. The augmentation techniques are chained in the order in which they appear in the table, meaning that the output from the contrast enhancement is fed into the Gaussian blur and so on. To implement the classical augmentations, torchvision.transforms.v2 has been used and the specific methods and parameters are as follows:

- ColorJitter (contrast=0.3)
- GaussianBlur (kernel_size=25, sigma=(0.1, 2.0))
- RandomResizedCrop (size=224, scale=(0.30, 0.90))
- RandomErasing (p=1.0, scale=(0.05, 0.15))

The custom anatomy-aware transforms have been implemented in such a way that they also abide by those parameters. The appearance-based contrast enhancement and Gaussian blur directly utilize the torchvision methods and simply replace the area inside the joint space bounding box with the original image data. Meanwhile, for crop and erase, the source code from the torchvision functions has been adapted to take into account the bounding box and not exclude the region inside it.

3.3 Evaluation

Linear Probing

To evaluate the quality of the learned representations by the different encoders, we will use Linear Probing, which is a technique used often in the field of Self-supervised learning [2][7] and has been shown to best predict the ranking of SSL methods, compared to other protocols [14]. This makes it a suitable choice for our experiment, since we are more interested in comparing the relative performance of the different classifiers, rather than maximizing their accuracy. Linear probing involves freezing the pretrained encoder which we want to be evaluate and training a simple linear classifier on top of it. This process can be seen in Figure 6. The intuitive idea behind it is that if the learned representation is good for the given classification task, then the dataset classes would be are linearly separable. The linear classifier which will be used in this study is a single-layer perceptron. A sigmoid function is applied on its output to obtain the final binary prediction - 0 or 1 representing respectively the labels 'False' (has not OA) and 'True' (has OA) .

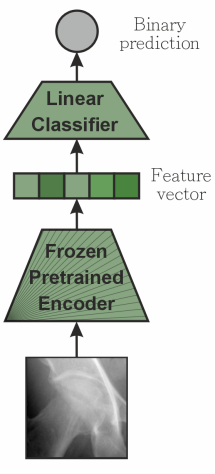

Figure 6: Overview of the Linear probing procedure. A linear classifier is trained on the encodings from a frozen pre-trained encoder.

Evaluation metrics

The main evaluation metric we will use to evaluate the performance of the classifiers is the AUC ROC score (Area Under the Curve of the Receiver Operating Characteristic), which is suited for evaluation of binary classification models and is also a usual choice in medical diagnostic studies [15]. Given the fact that the classes in the dataset are quite unbalanced, it is a suitable choice in our case, since it gives an indication of the model's quality across all possible classification threshold values.

While ROC curves and their corresponding AUC ROC scores present us with a nice visual overview of the performance of the models, they are not suitable for direct quantitative comparison. For this reason, we perform the DeLong's test [16] based on the true labels and the raw probabilistic outputs of the models (the logits before applying the sigmoid function). The test calculates a p-value, wich determines if the difference between two AUC ROC scores is statistically significant ($p < 0.5$) or a result of random chance ($p \geq 0.5$).

4 Results

As we can see from Figure 7, the loss curve for the Base model converges to similar values on both the training and validation datasets, which indicates that the model is generalizing well and not overfitting. However, as we can see from the other three graphs, the validation loss is not as stable for the models which use anatomy-aware data augmentations, which indicates problems with generalization. Since this does not occur in the Base model, we can conclude this is not due to unsuitable learning rate or noise in the validation set.

Figure 8 contains the ROC curves for the 'True' class (has OA) of all four models on the training, validation and testing datasets. AUC ROC values of around 0.7 are generally considered to be on the border between inadequate discrimination and acceptable performance, which indicates that none of the classifiers perform too well, including the one learning from representations extracted from the Base model. This

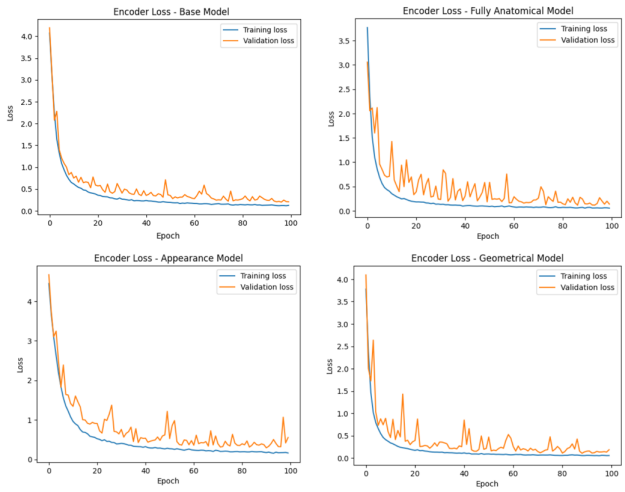

Figure 7: Training and validation loss curves for all 4 models - top to bottom, left to right: Base, Fully Anatomical, Appearance and Geometrical.

suggests that the reason for the poor performance is not due to the differences in the data augmentations of the encoders. It is worth noting, however, that despite the small differences, the classifier learning from representations from the Fully Anatomical Model is the one which has the lowest scores across all 3 datasets.

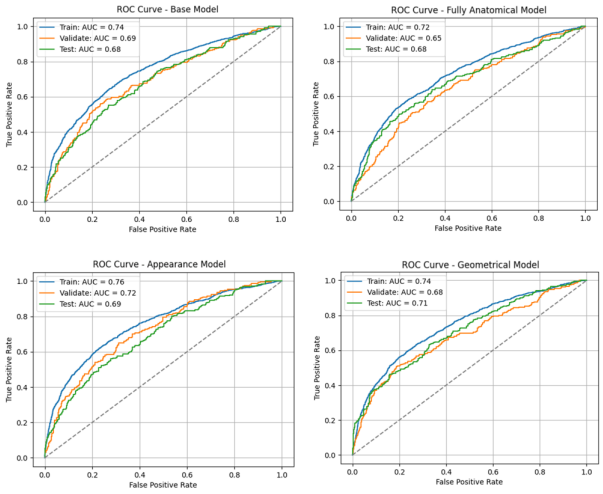

Figure 8: ROC curves on the train, validate and test datasets for all 4 models - top to bottom, left to right: Base, Fully Anatomical, Appearance and Geometrical. AUC ROC scores are included.

In order to investigate whether the poor performance is a consequence of the imbalance in the dataset between the two classes (True and False - for has / has not OA), the recall score was calculated on the True class. This metric was chosen since when it comes to medical diagnosis, we are mostly interested in maximizing the share of patients suffering from a disease which are correctly diagnosed. Indeed, the score for the Base model classifier was very poor - 0.08. After several attempts to train it using different thresholds or adding different weights to positive and negative examples in the loss function, the recall score increased to 0.22 and 0.50 respectively. While most of those encoders did not present significant changes in their ROC curves, it one case (using a threshold of 0.2 instead of 0.5) the accuracy score increased - from the original 76% to 78%.

	Accuracy	**AUC - ROC**
Base Model	76%	0.68
Appearance Model	76%	0.69
Geometrical Model	78%	0.71
Fully Anatomical Model	75%	0.68

Table 1: Accuracy and AUC - ROC scores for all four classifiers.

	P-value from performing DeLong's test
Appearance Model	0.3637
Geometrical Model	0.0392
Fully Anatomical Model	0.8340

Table 2: P-value from performing DeLong's test on all 3 classfiers, trained on the models which utilize anatomy-aware data augmentations, when compared against the classifier trained on the Base Model.

The results from performing DeLong's test on the AUC ROC scores of all three models, utilizing anatomy-aware augmentations, are included in Table 2. They were all compared against the Base model's encoder in order to observe whether the difference in AUC ROC scores is statistically meaningful. Only in the case of the Geometric model the p-value is less than 0.05 which means that there is a significant difference between the ROC curves of two models being compared. Since the only difference between the training procedure of those classifiers is the data augmentations of the encoders, whose output they use as input, we can conclude that the anatomy-aware approach on the geometric augmentations has contributed to a better performance of the Geometrical model. In the case of the Appearance and Fully Anatomical model, the p-values are much greater than 0.05 and thus, the no significant statistical difference is observed in their AUC ROC scores compared to the base model.

5 Discussion

As can be observed by Table 1, the two models, which used the classic data augmentation techniques and our custom anatomy-aware ones respectively, present little to no difference in their performance. In this sense, this study found no benefit from incorporating anatomy-awareness in all data augmentations used in a Contrastive Self-Supervised model. However, including anatomy-awareness partially and selectively - specifically when it comes to geometric data augmentations, seems to show promising initial results. This is in line with the initial hypothesis of this research - that anatomy-awareness would be more valuable for those data augmentations which exclude regions of the images, rather then merely alter their appearance. It is worth noting, however, that a lim-

itation of this study is that performing multiple runs of the experiment was not possible, due to the time constraints of the project. Despite the findings from DeLong's test, drawing conclusions from the performance of a single model would be unreasonable, given the unpredictable nature of deep learning models.

It is evident from Figure 7 that introducing any anatomy-aware augmentations leads to more unstable learning, given the learning curves on the validation set. This could be attributed to the fact that the custom anatomy-aware augmentations could be described as limited and less "strong" compared to their classical counterparts. Variability in between different views is the core driver of the learning process in Contrastive SSL, so it is possible that the augmented images become too similar, in comparison to those produced by the Base model, and the encoder learns to produce very generic feature vectors. What is more, the hyperparameters used for the data augmentations could also be described as leading to "safer" views. For example, the torchvision default scale parameter for RandomResizedCrop is (0.08, 1.0), which indicates the lower and upper bounds for the cropped area respectively. The same values are used in the SimCLR study. The value of 0.08 was considered not suitable for the purpose of this study, since it lead to almost fully gray views. For this reason, the bounds were changed, but determining the perfect parameters is not an easy task and would require a more systematic approach, which would be outside of the scope of this study.

The unstable learning of the encoders also directly correlates to the batch size being used to load the data into the model. The bigger it is, the better the model would be able to generalize, rather than overfitting on each new incoming entry. In our experiments, we used batch size of 64, mostly due to the fact that computational times increase drastically as the batch size does. However, in the experiments in the original SimCLR paper, batch size 4096 is used - which is more than half of the data we have available for this study. The authors also show that the larger the batch size, the better the performance of the model, since this provides more negative pairs for the model to learn from and thus, produce more varied representations.

6 Responsible Research

6.1 Data

In relation to this study, no data has been collected by our team. The data used for training and testing the models is taken from the Cohort Hip and Cohort Knee (CHECK) study, which is a large semi-public dataset intended to aid research in Osteoarthritis and commonly used in the field. The data is anonymized and has been collected with the patients' consent, following procedures approved by medical ethics committees. It has been obtained after a request for access, specifying the intended use. Since the datasets contain sensitive medical data, the experiments have been executed on the university's supercomputer DelftBlue [17] in order to ensure the data is used only for the purposes of this research and stays contained in the university's storage system.

6.2 Reproducibility

In order to ensure reproducibility of the research experiments, the architecture of the selected model has been described in detail and all hyperparameters used for the training phase have been stated. However, sincethe data augmentation methods introduce an element of randomness, the accuracy of the models might vary over different runs of the same training setup.

As previously mentioned, CHECK is a semi-public dataset. Thus, availability of the data may become an obstacle in reproducing this study, since the team at UMC Utrecht who collected the data needs to be contacted in order to obtain it. Furthermore, the dataset only contains X-ray images of elderly patients (45 to 65 years old) who are known to be at risk of developing Osteoarthritis or have previously exhibited symptoms. This may introduce a bias in the model, which would result in it not performing well on data of healthy or young people. This risk, however, would not lead to detrimental outcomes if the model is used as intended – as a tool to assist medical professionals in the diagnosis process. This would mean that the X-ray images which the model assesses would be of patients which a medical practitioner has already decide may have Osteoarthritis. This aligns with the profiles of the people represented in the datasets.

7 Conclusions and Future Work

In this study, we proposed custom data augmentation techniques, which aim to solve the problem , which classical augmentations present on medical data - that they may alter the image too strongly and thus, change its semantic meaning. We compared multiple encoders, whose training procedure was based on different combinations of custom and classic augmentations. Our findings suggest that utilizing this anatomy-aware approach for a larger number of chained image transformations may hinder the learning process and lead to less discriminative representations. However, when used in moderation, this approach could be beneficial. Particularly, geometric transformations, such as crop and erase, could benefit more from it.

Given the very minimal differences between the accuracies and AUC ROC scores of the compared models, multiple additional runs of the experiment on different data splits would be beneficial to draw more general conclusions. In order to more closely replicate the success of the original SimCLR model, the experimental setup could benefit from using its original hyperparameters. This would require the scale of both the models and the training procedure to be increased. Resnet50 could be used instead of resnet18 and the batch size could be drastically increased. For this, more data would be needed and the Osteoarthritis Initiative (OAI) dataset could be utilized, since it is a popular choice in the field. What is more, it has been observed that the current bottleneck of the training procedure is generating the two views of the images, rather than training the model. In the interest of saving time and computational resources, the custom data augmentations could be implemented more efficiently.

While approximating our region of interest - the joint space, by using a bounding box around it seems to work

well enough for geometric augmentations, it could be hurting the performance of the model when used on appearance augmentations. With the current implementation, artifacts appear around the edges of the box - harsh lines, defined by difference in contrast or blur. It is possible that

An interesting area for future research is utilizing those anatomy-aware augmentations in a different context. While they are a crucial part of contrastive self-supervised learning, their main application in machine learning is to augment small datasets. This would be particularly useful in the area of medical imaging, where data is sparse.

References

[1] Haoming Zhao, Liang Ou, Ziming Zhang, Le Zhang, Ke Liu, and Jianjun Kuang. The value of deep learning-based x-ray techniques in detecting and classifying k-l grades of knee osteoarthritis: a systematic review and meta-analysis. *European Radiology*, 35(1):327–340, 2024.

[2] Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P. Lungren, Serena Yeung, and Akshay S. Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digital Medicine*, 6(1), 2023.

[3] Benjamin Billot, Eleanor Robinson, Adrian V Dalca, and Juan Eugenio Iglesias. Partial volume segmentation of brain mri scans of any resolution and contrast. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII 23*, pages 177–187. Springer, 2020.

[4] Xiangyu Peng, Kai Wang, Zheng Zhu, Mang Wang, and Yang You. Crafting better contrastive views for siamese representation learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16010–16019, 2022.

[5] Sheng Wang, Zihao Zhao, Zixu Zhuang, Xi Ouyang, Lichi Zhang, Zheren Li, Chong Ma, Tianming Liu, Dinggang Shen, and Qian Wang. Learning better contrastive view from radiologist's gaze. *Pattern Recognition*, 162:111350, 2025.

[6] Wei Li, Zhongli Xiao, Jin Liu, Jiaxin Feng, Dantian Zhu, Jianwei Liao, Wenjun Yu, Baoxin Qian, Xiaojun Chen, Yijie Fang, and Shaolin Li. Deep learning-assisted knee osteoarthritis automatic grading on plain radiographs: the value of multiview x-ray images and prior knowledge. *Quantitative Imaging in Medicine and Surgery*, 13(6), 2023.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020.

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. 2020.

[9] C. Lindner, S. Thiagarajah, J. M. Wilkinson, The arcOGEN Consortium, G. A. Wallis, and T. F. Cootes. Fully automatic segmentation of the proximal femur using random forest regression voting. *IEEE Transactions on Medical Imaging*, 32(8):1462–1472, 2013.

[10] Matthew F. Koff. Imaging for osteoarthritis: An overview. https://www.hss.edu/health-library/conditions-and-treatments/osteoarthritis-imaging. Accessed: 10.06.2025.

[11] J. Wesseling, Maarten Boers, Max Viergever, Wim Hilberdink, Floris Lafeber, Joost Dekker, and Johannes Bijlsma. Cohort profile: Cohort hip and cohort knee (check) study. *International journal of epidemiology*, 45, 08 2014.

[12] J.H. Kellgren and J.S. Lawrence. Radiological assessment of osteo-arthrosis. *Annals of the Rheumatic Diseases*, 16(4):494–502, 1957.

[13] Justin A. Magnuson, Nihir Parikh, Francis Sirch, Justin R. Montgomery, Raja N. Kyriakos, Arjun Saxena, and Andrew M. Star. Is the interpretation of radiographic knee arthritis consistent between orthopaedic surgeons and radiologists? *Journal of Orthopaedic Experience amp; Innovation*, 5(1), 2024.

[14] Markus Marks, Manuel Knott, Neehar Kondapaneni, Elijah Cole, Thijs Defraeye, Fernando Perez-Cruz, and Pietro Perona. A closer look at benchmarking self-supervised pre-training with image classification, 2024.

[15] Şeref Kerem Çorbacıoğlu and Gökhan Aksel. Receiver operating characteristic curve analysis in diagnostic accuracy studies. *Turkish Journal of Emergency Medicine*, 23(4):195–198, 2023.

[16] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837, 1988.

[17] Delft High Performance Computing Centre (DHPC). *DelftBlue Supercomputer (Phase 2)*, 2024. https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2.