The Effect of Cross-Domain Class Imbalance on Distribution Alignment

by

Justin Luu

To obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday November 12, 2025 at 13:00 PM.

Student number:	5088100
Thesis advisor:	Dr. ir. Jesse Krijthe
Thesis daily co-supervisor:	Dr. Gijs van Tulder
External committee member:	Dr. Emir Demirović
Project duration:	February 25, 2025 – November 12, 2025

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

TU Delft

The Effect of Cross-Domain Class Imbalance on Distribution Alignment

Justin Luu
Delft University of Technology

Abstract

Statistical distribution alignment methods for domain adaptation assume similar class distributions across domains, but this assumption cannot always be guaranteed in medical imaging data. This research investigates the effect of cross-domain class imbalance on statistical distribution alignment in unsupervised domain adaptation for medical image classification. Our experiments demonstrate that statistical distribution alignment using MMD performs reliably under mild domain shifts but struggles when both severe cross-domain class imbalance and complex domain shifts are present. To address this, we implement class-conditioned domain alignment with a new weighted minibatch sampling method. Under conditions of extreme domain shift and severe cross-domain class imbalance, combining statistical distribution alignment with more complex sampling strategies results in small improvements compared to alignment with random sampling, suggesting that class-conditioned distribution alignment offers limited practical benefits. The model appears robust to label noise, but since the performance gains are tiny, the choice of sampling strategy could have limited influence on overall performance. In our experiments, we employ the CHECK and OAI hip X-ray datasets to investigate binary osteoarthritis classification under varying levels of domain shift and cross-domain class imbalance.

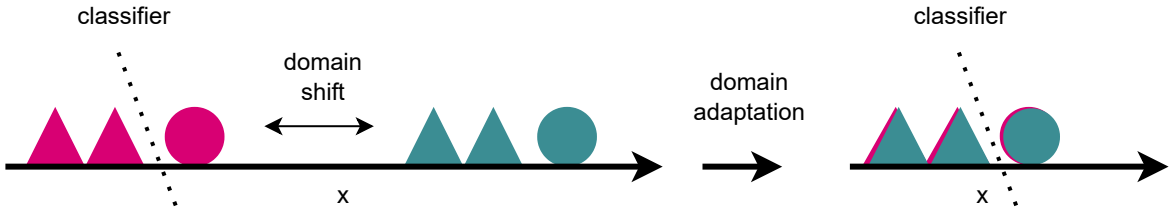
1. Introduction

Deep learning models have achieved significant success in medical image analysis; however, their performance often declines when applied to data obtained from different clinical settings [13]. This degradation can arise from variations in imaging equipment, acquisition protocols, or patient demographics. These discrepancies give rise to domain shift, which occurs when the feature distribution of the training data (source domain) differs from that of the deployment data (target domain). As a result, models that perform well on the source dataset may fail to generalize effectively and produce unreliable predictions when applied to other datasets.

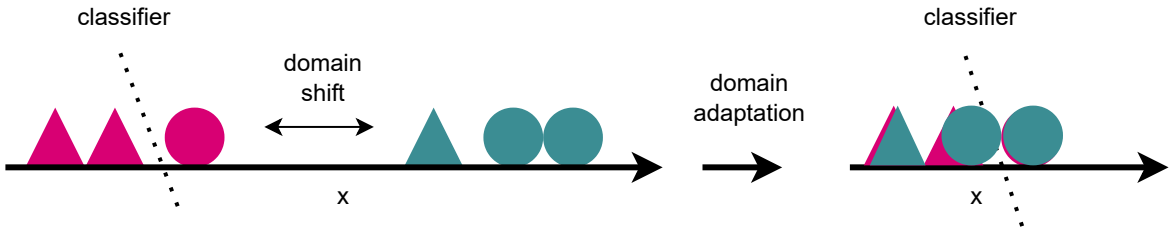
Domain adaptation [29, 12] is a promising solution to this problem by adapting models trained on a source domain to perform well on a different target domain. This process could be performed in a supervised manner, where a model trained on labeled source data is fine-tuned using a small labeled subset of the target domain. However, in medical imaging, obtaining such labeled target data is often impractical due to the high cost and time requirements of expert annotation. To overcome these limitations, an Unsupervised Domain Adaptation approach [20] is often used, which leverages labeled source data to adapt models to unlabeled target domains, thereby eliminating the need for additional target-domain annotations.

A common approach to unsupervised domain adaptation is statistical distribution alignment, which aims to reduce the discrepancy between the feature distributions of the source and target domains. In deep learning, this is typically achieved by training models to learn domain-invariant feature representations that generalize well across datasets with differing distributions. One widely used discrepancy metric for this purpose is the Maximum Mean Discrepancy (MMD) loss [21, 30, 11], which explicitly measures and reduces the distance between the source and target feature distributions during training. This loss is calculated by comparing the mean embeddings of samples in the two domains in this space, representing how dissimilar their distributions are. By minimizing this MMD loss, the model is encouraged to learn feature representations that are indistinguishable in terms of their distribution.

A drawback of statistical distribution alignment methods such as MMD is that they rely on the assumption that the class distribution in the source and target domains are similar, which cannot always be guaranteed. Differences in class distributions between source and target domains are a well-recognized challenge in machine learning [14] and this issue is particularly pronounced in medical imaging, where healthy samples are far more common than unhealthy samples [8, 23]. This asymmetry in class distributions between domains is known as cross-domain class imbalance [15]. When the source and target domains have



(a) Source and target domain have similar class distributions.



(b) Source and target domain have different class distributions

Figure 1: Illustration of a one-dimensional domain shift along a single feature axis. Colours denote different domains, while shapes represent distinct classes. In panel (a), a clear decision boundary remains after distribution alignment. In panel (b), the decision boundary becomes indistinct due to misalignment.

differing class proportions, distribution alignment methods may match samples across domains belonging to different classes, leading to misalignment. Such misalignment could blur class-specific distinctions and weaken the discriminative boundaries for classification, thereby degrading performance on the target domain (Figure 1)

To address misalignment, class-conditioned domain alignment can be used [3], where samples across domains are aligned based on their labels.

While many studies propose methods such as class-conditioned domain alignment to mitigate misalignment, few have empirically examined under which circumstances misalignment actually poses a problem. We aim to fill this gap by systematically investigating the impact of misalignment through controlled experiments under different cross-domain class imbalances and by proposing a different sampling strategy to ensure similar class priors across domains without upsampling the minority class during training.

Our research question is therefore: How does cross-domain class imbalance affect the performance of distribution alignment?

To address this question, we build on the work of Jiang et al. [16] on Implicit Class Conditioned Domain Alignment. In their approach, pseudo-labels are used to estimate target sample classes during training. These pseudo-labels are then used to create minibatches that are uniformly distributed with respect to the label classes, achieving class-conditioned domain alignment.

Uniform minibatch sampling offers advantages such as mitigating bias toward the majority class and maintaining similar class distributions across domains. However, it oversamples minority classes to create uniformly distributed minibatches, increasing the risk of overfitting and reducing model generalization. Instead, we propose an extension to the method by Jiang et al. [16] in which we replace their uniform minibatch sampling with a weighted sampling strategy. This approach ensures consistent class distributions across domains by constructing minibatches that follow the source domain's class distribution, thereby avoiding the need for oversampling within the source domain.

The effectiveness of pseudo-labeling depends on the accuracy of the generated pseudo-labels. If a classifier generalizes poorly to the target domain, it may assign incorrect labels to target samples, introducing pseudo-label noise. This noise can accumulate over training as the model reinforces its own errors. In class-conditioned domain alignment methods, inaccurate pseudo-labels can still lead to the alignment of samples from different classes across domains, reducing class separation in the learned feature space. Therefore, it is important to evaluate the robustness of the proposed method.

To systematically explore the impact of class imbalance and the proposed method, we can further divide the main question into the following three subproblems:

1. How does cross-domain class imbalance affect distri-

bution alignment?

2. How does the choice of sampling strategy influence the classification performance on the target domain?

3. How does pseudo-label noise influence the reliability and effectiveness of class-conditioned sampling?

In our experiments we use two clinical datasets of hip osteoarthritis (OA): the Cohort Hip & Cohort Knee (CHECK) and the Osteoarthritis Initiative (OAI). Our study focuses on binary OA classification using hip X-ray images derived from these datasets.

CHECK and OAI provide complementary experimentation sets for studying binary hip OA classification under domain shift. Differences in demographic characteristics, imaging protocols, and annotation standards introduce natural distributional discrepancies between the datasets, making them well-suited for evaluating domain adaptation methods in medical imaging. The datasets also exhibit very different class priors: in CHECK, there is approximately one diseased case for every two healthy cases, whereas in OAI, the ratio is roughly one diseased case for every nine healthy cases.

The remainder of this paper is organized as follows. In Section 2, we review related work on domain adaptation with a particular emphasis on methods addressing distribution alignment and class imbalance in medical imaging. Section 3 outlines our proposed methodology for handling cross-domain class imbalance in unsupervised domain adaptation. Section 4 describes the experimental setup, including details of the CHECK and OAI hip datasets and preprocessing steps. Section 5 presents the empirical results, followed by a detailed analysis and interpretation in Section 6. Finally, Section 7 summarizes our findings.

2. Related Work

2.1. Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) has been studied through several complementary approaches. While this work does not directly employ image-to-image translation or adversarial distribution alignment methods such as Domain Adversarial Neural Networks, these approaches provide important context for understanding the broader UDA landscape and highlight the complementary role of statistical distribution alignment techniques, which are most relevant to our setting.

2.1.1 Image-to-Image Translation

Image-to-image translation methods [9, 10] aim to reduce low-level appearance discrepancies by transforming source images to match the style of the target domain. A common approach to match styles of different domains are Generative Adversarial Networks (GANs). They enable translation between domains without requiring paired data [17, 34, 22]. Through such translations, source-domain images can be rendered in the target style and vice versa. This way, labeled source data can be used while incorporating the appearance characteristics of the target domain, bridging the domain gap. This allows one to train a model on source images rendered in the target style, ensuring that supervision is maintained while adapting the model to the target distribution [22]. However, a practical limitation of such methods is their computational cost, as GAN-based translation models are often expensive to train and require substantial time and resources [24]. Another disadvantage is the potential loss of low-level information. While the goal is to modify style while preserving semantic content, GAN-based translations may inadvertently alter clinically relevant features such as fine bone textures or joint space width. These features, though low-level in appearance, can be critical for diagnostic tasks like osteoarthritis grading.

2.1.2 Domain-Adversarial Neural Networks

Domain-Adversarial Neural Networks (DANN) [7, 2] aim to learn feature representations that are both task-discriminative and domain-invariant. Unlike image-to-image translation methods, DANN operates directly in the feature space, thereby avoiding the risk of altering low-level appearance cues that may be critical for diagnosis. The DANN model architecture comprises a feature extractor, a domain discriminator, and a task-specific classifier. The feature extractor is trained to minimize the task loss on labeled source data while simultaneously maximizing the domain discriminator's loss. The domain discriminator attempts to distinguish whether extracted features originate from the source or target domain, whereas the feature extractor learns to confuse it, thereby reducing domain shift through adversarial training. This is typically implemented using a gradient reversal layer, which inverts gradients during backpropagation, allowing the feature extractor to be optimized for both task performance and domain invariance. By aligning feature distributions across domains, DANN enables better generalization to the unlabeled target domain. However, adversarial training is known to be unstable [28], often requiring extensive hyperparameter tuning and careful optimization, which can make DANN computationally demanding in practice.

2.1.3 Statistical Distribution Alignment Methods

Statistical distribution alignment methods aim to reduce domain shift by explicitly minimizing a discrepancy loss between source and target feature distributions. Several discrepancy measures could be used for this purpose [27, 33].

A well studied metric is the Maximum Mean Discrepancy loss. In unsupervised domain adaptation, this loss is frequently employed as a regularization term in the training objective of neural networks [21, 30, 11]. By minimizing the discrepancy between the learned feature distributions of the source and target domains, the model is encouraged to learn domain-invariant representations. This alignment reduces distributional shift, thereby enabling better transfer of knowledge from labeled source data to unlabeled target data. The approach is conceptually simple, computationally fast, and quick to implement. In practice, though, its effectiveness is strongly dependent on the choice of kernel function, and selecting an appropriate kernel for a given dataset remains a challenge, as it determines the sensitivity to discrepancies between domains.

2.2. Class-Conditioned Domain Alignment

UDA methods implicitly assume that the source and target domains share similar label distributions. In practice, however, this assumption often does not hold. As a result, distribution alignment may inadvertently match samples from different classes, leading to suboptimal representations. A natural solution is to try and enforce alignment on samples that have the same label, this is called Class-Conditioned Domain Alignment [3, 16]. This domain alignment can be guided using pseudolabels to approximate class information in the target domain.

2.2.1 Explicit and Implicit Class-Conditioned Domain Alignment

Within class-conditioned domain alignment, a distinction can be made between explicit and implicit approaches. In explicit class-conditioned domain alignment, pseudolabels directly contribute to the optimization objective by training the model to classify target data using its own predicted labels as supervision which could lead to an accumulation of errors [5]. They may be used to compute class-wise domain discrepancy losses, such as clustering losses [6, 18], or to reweight samples during training in accordance with estimated target class priors [32]. Here, the pseudo-labels actively guide the optimization process.

In contrast, implicit alignment does not use pseudolabels in the loss function; instead, they are employed to guide the sampling process for minibatch creation, indirectly encouraging class-consistent alignment [16]. In this approach, pseudo-labels ensure uniformly distributed minibatches across domains. This means that features from samples belonging to the same class across domains are aligned ensuring class-wise distribution alignment across domains. Since pseudolabels in implicit class-conditioned domain alignment do not contribute to the optimization objective directly, implicit class-conditioned domain alignment suffers less from error accumulation [16].

3. Methodology

We begin by describing our overall research design and problem formulation. We then introduce our proposed method, which extends the Implicit Class-Conditioned Domain Alignment framework with a weighted sampling strategy.

3.1. Problem Formulation

We consider the setting of unsupervised domain adaptation for binary classification. Let the source domain be denoted as $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, where $x_i^s \in \mathcal{X}$ are input samples, $y_i^s \in \{0, 1\}$ are corresponding labels, and N_s is the number of labeled source samples. The target domain is denoted as $\mathcal{D}_t = \{x_j^t\}_{j=1}^{M_t}$, consisting of M_t unlabeled samples drawn from a related but distinct distribution. Both domains share the same label space, but differ in their data distributions, i.e., $P_s(x, y) \neq P_t(x, y)$.

The objective of UDA is to learn a classifier $f_\theta : \mathcal{X} \to \{0, 1\}$ that generalizes well to the target domain, despite the absence of target labels during training. Domain alignment methods attempt to reduce the discrepancy between source and target feature representations, under the assumption that the label distributions across domains are similar.

In practice, however, this assumption often does not hold. Real-world datasets frequently exhibit cross-domain class imbalance, where the class priors in the source and target domains differ. In such cases, distribution alignment may inadvertently match samples from different classes, which could potentially lead to degraded performance.

3.2. Maximum Mean Discrepancy

In this work, we employ Maximum Mean Discrepancy, a widely studied metric for domain adaptation [11]. It compares the average feature representations of the distributions in a reproducing kernel Hilbert space (RKHS[1]), capturing differences in their overall statistics. This property makes MMD particularly effective for tasks such as domain adaptation, where the goal is to reduce distributional shifts between source and target domains.

Formally, given two distributions $P(x)$ and $Q(y)$, and a kernel function $k(\cdot, \cdot)$, the squared MMD is defined as:

$$\mathrm{MMD}^2(P, Q) = \|\mathbb{E}_{x\sim P}[\phi(x)] - \mathbb{E}_{y\sim Q}[\phi(y)]\|^2_{\mathcal{H}}, \quad (1)$$

where $\phi(\cdot)$ denotes the feature mapping induced by the kernel into the RKHS $\mathcal{H}$.

Intuitively, MMD measures the distance between the mean embeddings of two distributions in the feature space defined by the kernel. If the kernel is characteristic, then

[1] An RKHS is a Hilbert space of functions where inner products can be evaluated implicitly through a kernel function.

$\text{MMD}(P, Q) = 0$ if and only if $P = Q$, making it a fitting tool for distribution alignment. In practice, commonly used kernels such as the Gaussian radial basis function (RBF[2]) allow MMD to capture discrepancies at multiple levels of granularity.

To compute the MMD loss in our experiments, input images are first passed through a feature extractor network to obtain embeddings from both the source and target domains. The MMD loss is then calculated on these embeddings at the batch level and combined with the classification loss to form the overall training objective. Both losses are weighted, with the MMD term scaled by a factor λ, (see 2). By minimizing this loss during training, the feature extractor is encouraged to produce representations in which source and target distributions are closely aligned, thereby facilitating knowledge transfer across domains.

$$\text{Loss}(P, Q) = \text{Classificationloss(P)} + \lambda * MMD(P, Q), \tag{2}$$

3.3. Implicit Class-Conditioned Domain Alignment Method

In this paper we build upon the Implicit Class Conditioned Domain Alignment (ICCDA) method by Jiang et al. [16]. It performs class-conditioned domain alignment at the minibatch level through a sampling strategy. The training procedure can be summarized as follows. Given a labeled source dataset $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and an unlabeled target dataset $\mathcal{D}_t = \{x_j^t\}_{j=1}^{M_t}$, the goal is to align source and target feature representations while preserving discriminative structure across classes.

At each training epoch, pseudolabels are first predicted for the target samples using the current state of the classifier f_θ. A minibatch is then constructed by sampling a fixed number of classes from the label space according to a predefined distribution $p(y)$. Jiang et al. use a uniform distribution. For each selected class, K samples are drawn from both the source and target domains, conditioned on the class label in the source and on the pseudolabel in the target. This results in a class-balanced minibatch containing matched examples from both domains.

The constructed minibatch is then used to perform domain adaptation training, where alignment losses are applied jointly with the classification loss on labeled source data.

A uniform sampler in ICCDA has certain limitations: to generate evenly balanced minibatches, it is necessary to oversample the minority class. Oversampling the minority class artificially increases its weight in the source training set by replicating samples. In mathematical terms, this forces

$$p(y) \to \text{Uniform}(\mathcal{Y}), \tag{3}$$

which diverges from both the true $p_S(y)$ and realistic $p_T(y)$. This risks overfitting, as the model repeatedly trains on a small set of minority samples.

3.3.1 Proposed Minibatch Sampling Based on Source Domain Priors

Instead of using a uniform sampling distribution $p(y)$ as in ICCDA, we propose a minibatch sampler that constructs minibatches according to the class priors observed in the source domain. Formally, if the source class prior is $p_S(y)$, then the probability of sampling a target-domain sample of class y is defined as

$$\Pr\left(y_T = y\right) = p_S(y). \tag{4}$$

This ensures that the sampling distribution of the target domain matches the source class prior, allowing the adaptation process to proceed under the theoretical assumption that class priors across domains are similar [32].

A key distinction between our sampling strategy and ICCDA's uniform sampling method is that it restricts oversampling to the target domain. This is important because the classification loss is computed on the labeled source data, and oversampling source classes would make the model more prone to overfitting.

3.3.2 Proposed Weighted Random Sampling Method

Instead of sampling minibatches directly, we could employ a weighted random sampler that assigns sampling probabilities to target-domain examples so that, in expectation, the class distribution of the sampled target data aligns with the class priors observed in the source domain. In theory, this approach should achieve the same effect as constructing minibatches directly according to source priors and it also does not require us to oversample the source domain. However, unlike the minibatch-based strategy, weighted random sampling does not guarantee that every minibatch contains samples from all classes. As a result, individual minibatches may occasionally consist of examples from a single class, which can introduce variability in the alignment process.

To implement the weighted random sampling strategy, we first estimate the class proportions in both the source and target domains. Let $p_S(y)$ and $p_T(y)$ denote the empirical class priors in the source and target datasets, respectively. These are computed as the fraction of samples belonging to each class in the corresponding domain.

[2]The Gaussian RBF kernel maps samples into a potentially infinite-dimensional space where similarity is measured in terms of distance.

Next, a reweighting factor for each class is computed as the ratio of the source prior to the target prior:

$$w(y) = \frac{p_S(y)}{p_T(y) + \epsilon}, \tag{5}$$

where ϵ is a small constant to prevent division by zero. This factor ensures that underrepresented classes in the target domain are sampled more frequently, while overrepresented classes are sampled less frequently, aligning the target sampling distribution with the source class priors.

Finally, each target sample is assigned a weight corresponding to the reweighting factor of its pseudolabel. These sample-level weights are then used by a weighted random sampler to construct minibatches for domain-adaptive training, ensuring that the class distribution in each minibatch reflects the source priors while avoiding oversampling of the source dataset.

3.4. Overview of the Proposed Method

Our approach extends the Implicit Class-Conditioned Domain Alignment framework [16] with a weighted minibatch sampling strategy instead of the uniform sampling method. Unlike the uniform sampling in ICCDA, our methods weigh target samples according to source class priors, avoiding oversampling of the source domain while aligning class distributions across domains. This approach directly enforces distribution alignment's assumption of similar class distributions across domains without the need to upsample the source domain. The implementation of this method can be found in our GitHub repository [3].

4. Experiments

4.1. Hypothesis

From the problem formulation in section 3.1, we derive the following hypotheses for our study:

1. Severe class imbalance across domains adversely affects distribution alignment performance, limiting the benefits of MMD.

2. Incorporating MMD with a weighted sampling strategy under severe cross-domain class imbalance improves generalization to the target domain over MMD with random sampling.

3. Implicit Class-conditioned domain alignment is robust to pseudolabel noise.

Together, these hypotheses allow us to examine the interaction between distribution alignment, cross-domain class imbalance, and pseudolabel robustness in the context of binary hip osteoarthritis classification.

[3] https://github.com/LuuJustin/MScThesis

4.2. Model

We adopt the ResNet-18 architecture as our baseline backbone. In our setup, we remove the final fully connected classification layer of the network and instead use the feature representation (a 512-dimensional vector after global average pooling) as our feature embedding. This transforms the network into a feature extractor that provides representations to calculate the MMD between source and target domain for adaptation.

On top of the feature extractor, we attach a lightweight linear classifier tailored to our binary classification setting. This classifier is a single fully connected layer of dimensions 512 , which maps the feature vector to a scalar logit. We use the `BCEWithLogitsLoss` function during training. This loss function combines a sigmoid activation with the binary cross-entropy loss in a numerically stable formulation, which is why we chose it for our binary classification tasks.

4.3. Data

We use the CHECK and Osteoarthritis Initiative (OAI) hip datasets in this research. Both datasets contain annotated hip X-ray images collected from different clinical centers.

The CHECK study is a population-based, observational, multicenter cohort established in the Netherlands, consisting of 1002 individuals with early symptomatic OA of the hip [31]. Participants are from general practices and hospitals across ten centers, ensuring a diverse sampling of patients in the early stages of disease. The study was designed to investigate the natural course and progression of OA, with extensive follow-up including clinical assessments, imaging, and patient-reported outcomes. The hip X-ray data from CHECK provide valuable insight into early disease manifestation.

OAI is a large-scale, multi-center, longitudinal study sponsored by the U.S. National Institutes of Health [1]. It followed roughly 4800 men and women over a ten-year period, collecting clinical, and imaging data with a focus on risk factors and progression of hip OA. The imaging component includes a rich set of radiographs, systematically graded for OA severity by expert readers.

These datasets differ in imaging protocols, scanner types, population demographics, and disease severity distributions, making them well-suited for examining domain shifts in medical imaging. Furthermore, both datasets exhibit class imbalance typical of OA studies, where the prevalence of healthy subjects outweighs severe cases. This combination of inter-domain variability and cross-domain class imbalance makes the CHECK–OAI pair a valuable testing environment for researching cross-domain class imbalance and domain adaptation strategies.

Figure 2: Hip from the CHECK dataset.

Figure 3: Hip from the CHECK dataset after applying a gamma correction on it. (γ=0.65)

Figure 4: Hip from the CHECK dataset after applying a pixel inversion transformation on it.

4.4. Preprocessing

To focus on the relevant anatomical region, all DICOM images from the CHECK and OAI datasets were cropped to 224×224 pixels around the femoral head. Anatomical landmarks were obtained using BoneFinder [19], a model–based tool for detecting bone structures in radiographic images.

For each hip joint, BoneFinder provides a set of landmark coordinates delineating key points on the femoral head. We used these landmarks to calculate the geometric center of the femoral head. A square region of 224×224 pixels centered at this point was then extracted from each image. This approach ensures consistent localization of the region of interest across all samples while minimizing irrelevant background information.

Prior to input into the Resnet-18 model, each grayscale image was normalized to have zero mean and unit variance and then duplicated across three channels to satisfy the network's input requirements. KL grades were binarized such that grades 0–1 were labeled as "non-OA" and grades 2–4 as "OA." Images corresponding to hip replacements (KL grade 5) or with missing labels were excluded. Left hip radiographs were horizontally flipped to match the orientation of right hips. Minimal data augmentation was applied in the form of random rotations of up to $\pm 10^{\circ}$ to improve model robustness.

To prevent data leakage, all splits were performed at the patient level, ensuring that radiographs from the same individual were assigned to the same subset. Patients were randomly shuffled and allocated into training, validation, and test sets in an 80/10/10 ratio. These patient-level splits were consistently applied across all experiments, allowing for direct comparison between different sampling strategies, including MMD + random sampler, MMD + uniform minibatch sampler, MMD + weighted minibatch sampler, and MMD + weighted random sampler.

The resulting datasets were balanced as such: the OAI dataset comprised 90% non-OA and 10% OA samples, whereas CHECK contained 68% non-OA and 32% OA.

4.5. Synthetic Datasets

The domain shift between the OAI and CHECK datasets arises not only from differences in scanner protocols, but also from variations in demographic and clinical characteristics. To enable controlled experiments, we have chosen to utilise the CHECK dataset to create our synthetic dataset, since it contains a higher number of OA images compared to OAI. This allows us to systematically manipulate class distributions and create cross-domain imbalances for our study.

To isolate the effect of domain shift, we construct synthetic datasets in which the shift is introduced in a controlled manner via gamma correction or pixel inversion of image intensities.

4.5.1 Synthetic Dataset using Gamma Correction

Gamma correction applies a non-linear remapping of pixel intensities, altering global image appearance while preserving structural integrity. Given an input image I with pixel values normalized to $[0, 1]$, gamma correction is applied as:

$$I' = I^{\gamma}, \tag{6}$$

where $\gamma > 0$ controls the degree of intensity transformation.

We use Gamma correction to reflect real-world variability in image acquisition. For example, the same patient imaged on different machines may produce identical anatomical structures with differing grayscale distributions. By introducing a controlled domain shift without removing critical features, gamma correction provides a meaningful synthetic scenario for evaluating domain adaptation methods.

To confirm that gamma correction induces a domain gap, we trained models independently on the original and gamma corrected datasets and evaluated them via cross-domain testing. The observed performance drop in the

Figure 5: Domain shift when applying pixel inversion.

cross-domain setting indicates that gamma correction introduces a pronounced covariate shift, making it a suitable choice for constructing synthetic domains in controlled experiments. The selected gamma value (0.65) was derived by testing multiple values deviating from 1.0 in increments of 0.2. Experiments on how we found this value can be found in the appendix. In this paper we will refer to the synthetic gamma-corrected CHECK dataset as the CHECK Gamma dataset.

4.5.2 Synthetic Dataset using Pixel Inversion

To create a more complex domain shift, we constructed a synthetic target dataset by inverting the grayscale values of the radiographs. This operation preserves the underlying anatomical structures but drastically changes the visual appearance, producing a domain that is the visual inverse of the source. From the model's perspective, this represents a severe domain shift: the intensity distribution is completely reversed while spatial structures remain intact (Figure 5).

Although pixel inversion is not a clinically realistic transformation, it serves as a stress test for domain adaptation methods. By maximizing differences in global pixel statistics, this transformation forces the model to rely primarily on structural features rather than intensity patterns. This experiment was designed to explore whether combining MMD with implicit class-conditioned sampling offers any advantage over MMD alone under extreme distributional shifts, as prior observations by us showed little difference.

4.6. Experimental Setup: Cross-Domain Class Imbalance Scenarios

To investigate how class imbalance influences domain adaptation performance across domains, we evaluate the four model configurations using the following six source–target configurations:

1. Balanced source $\rightarrow$ Balanced target: Both domains have a 50/50 class distribution.

2. Balanced source $\rightarrow$ Unbalanced target: Source is balanced; target is skewed.

3. Balanced source $\rightarrow$ Inversely unbalanced target: Source has a 50/50 split; target has the inverse skew (30/70), creating a strong mismatch in class priors.

4. Unbalanced source $\rightarrow$ Unbalanced target: Both domains share the same 70/30 class distribution.

5. Unbalanced source $\rightarrow$ Balanced target: Source is skewed; target is balanced.

6. Unbalanced source $\rightarrow$ Inversely unbalanced target: Source follows a 70/30 distribution, while the target has the inverse 30/70 split, representing the strongest mismatch condition.

To ensure a fair comparison across experimental scenarios, we maintain consistent training set size while varying the class proportions to investigate the effect of cross-domain class imbalance. The total number of training samples is fixed at 1,936, corresponding to the available OA cases (70% of 1,936). To maintain this total, healthy cases are undersampled rather than oversampling OA cases, avoiding duplicated samples that could lead to overfitting. Within this setup, class proportions are systematically varied while keeping the overall sample size constant, allowing us to isolate the effect of label imbalance without confounding effects from dataset size.

These settings allows us to understand how misaligned class priors interact with feature misalignment. In particular, the extremely imbalanced condition #6 is expected to highlight the failure mode of distribution alignment.

It is worth noting that we will not be creating artificial scenarios where the source has a 30/70 class distribution of non-OA vs. OA. In such cases, the limiting factor is primarily class discrimination rather than domain alignment, because the source data would not be representable of the true underlying class distribution in the population used for testing.

For the class conditioned samplers using pseudolabels we do not rely on model-generated predictions in the experiments. Instead, we use the ground-truth labels for the experiments, representing a best-case scenario for evaluating the sampling methods.

4.7. Real-World Dataset Experiments

The cross-domain configurations used in the synthetic experiments cannot be directly replicated for the real-world datasets. In the controlled setting, training set sizes could be adjusted to create different class balances. However, this approach is not feasible for cross-domain experiments between OAI and CHECK. The OAI dataset contains relatively too few positive samples, and performance degrades substantially when additional training data are removed in order to enforce a particular balance, as observed during the research.

To address this limitation, we instead conduct experiments using the full datasets in both directions: models are trained on CHECK and evaluated on OAI, and conversely trained on OAI and evaluated on CHECK. This setup enables a meaningful comparison of MMD-based domain adaptation with the proposed sampling methods under realistic data constraints, because the CHECK and OAI datasets already suffer from cross-domain class imbalance.

4.8. Model Configurations

To evaluate the impact of class-conditioned sampling and different sampling strategies on domain adaptation performance, we consider six model configurations, including two baselines:

1. Source-only baseline: trained on the source domain and directly tested on the target domain without any domain adaptation. This provides a reference for the performance drop due to domain shift.

2. Target-only baseline: trained and tested on the target domain using labels, representing a theoretical upper bound for achievable performance.

3. Baseline MMD: standard Maximum Mean Discrepancy with random sampling.

4. MMD + uniform minibatch sampling: standard MMD with enforced class-balanced minibatches.

5. MMD + source-weighted minibatch sampling (proposed): standard MMD with minibatches sampled to match the source-domain class distribution.

6. MMD + source-based weighted random sampling (proposed): our method, which constructs target-domain minibatches to approximately match the source-domain label distribution without artificially oversampling.

This set of configurations allows us to isolate the effects of domain shift, class imbalance, and sampling strategies, while providing clear references for both the unadapted source model and the ideal target-trained upper bound. Each experimental configuration is repeated five times with different random seeds to reduce variance due to data partitioning. Results are reported in terms of mean performance and standard deviation across repetitions.

4.9. Pseudolabel Robustness

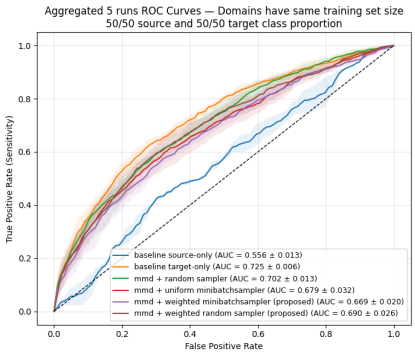
To evaluate the robustness of our method to noisy supervision, we conducted experiments in which a fraction of the ground-truth labels was deliberately corrupted. We select noise thresholds of 10% to 100% with increments of 10 and randomly flip the corresponding proportion of labels in the training set. The flip follows a weighted coin toss, where the probability of selecting a class is determined by the target domain's class priors. This controlled design simulates the presence of inaccurate pseudolabels, as often encountered during training, and allows us to assess the sensitivity of the proposed method to pseudolabel quality.

To assess the reliability of the pseudolabels generated by our model, we also computed calibration curves. These curves illustrate the relationship between the model's predicted confidence and its actual accuracy, providing insight into how well the model's probability estimates reflect true correctness. If the model is poorly calibrated, overly confident yet incorrect predictions can introduce noise into the adaptation process. By analyzing calibration, we can therefore evaluate the trustworthiness of the pseudolabels and determine whether certain sampling or alignment strategies improve model confidence estimation.

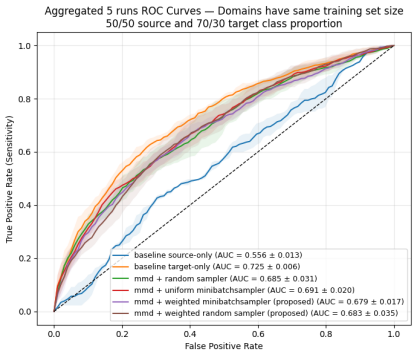
4.10. Evaluation Protocol

4.10.1 AUC Metric

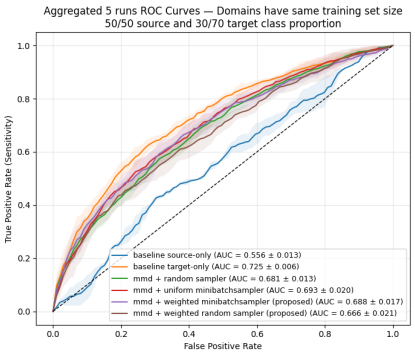
Evaluation of the proposed models is performed using the Area Under the Receiver Operating Characteristic Curve (AUROC). It measures the probability that a randomly chosen positive sample (e.i., a hip image with osteoarthritis) is ranked higher than a randomly chosen negative sample (healthy or low-grade OA). AUROC is especially useful for binary classification tasks with imbalanced datasets, as it



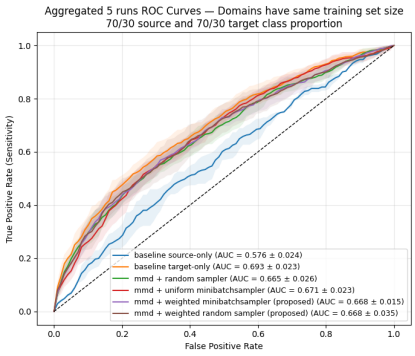
(a) Balanced source and target.



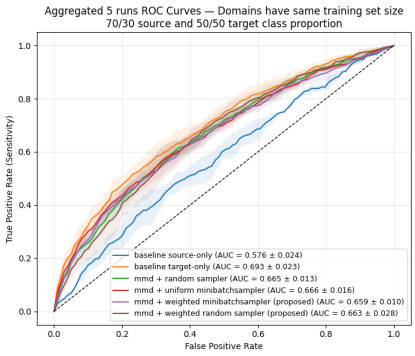
(b) Balanced source and unbalanced target.



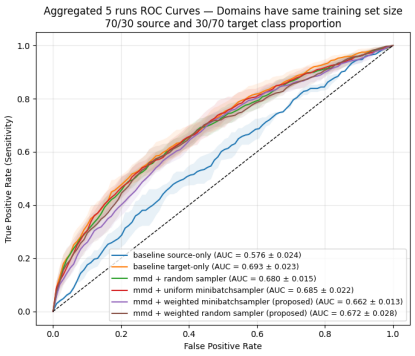
(c) Balanced source and unbalanced target inverted.



(d) Unbalanced source and target.



(e) Unbalanced source and unbalanced target.



(f) Unbalanced source and balanced target inverted.

Figure 6: AUC scores of a trained model after distribution alignment with MMD combined with different sampling methods trained on the CHECK dataset and tested on the Gamma CHECK dataset under varying class balance conditions. The theoretical upper bound of the classifier is represented by the orange line and the theoretical lower bound of the classifier is represented by the blue line.

provides a threshold-independent assessment of model performance. In this study, AUROC is reported for each evaluation to quantify how well the model generalizes.

4.10.2 UMAP

Uniform Manifold Approximation and Projection (UMAP) is a dimensionality reduction technique that projects high-dimensional data into a lower-dimensional space, while preserving both local and global structure of the data. UMAP allows us to qualitatively assess how features from the source and target domains are distributed in the embedding space. By visualizing these embeddings, we can inspect whether samples from the same class cluster together and whether the source and target domains are well-aligned.

In our experiments, UMAP embeddings provide an intuitive visualization of how effectively the proposed sampling and feature alignment strategies bring source and target features closer together at the class level.

4.10.3 Grad-CAM

To gain insight into model decision-making, we apply Gradient-weighted Class Activation Mapping (Grad-CAM) [26]. Grad-CAM generates a heatmap by combining the gradients of the output with the feature maps of the final convolutional layer, highlighting regions that contribute most strongly to the prediction.

We generated Grad-CAM visualizations for both the baseline model that is trained solely on the source domain and the domain-adapted model. The resulting activation maps are superimposed on the original grayscale X-rays using a jet colormap, where regions highlighted in red indicate areas of greater importance to the model's classification decision.

5. Results

In this section, we present the results of our experiments designed to investigate the impact of cross-domain class

imbalance on domain alignment. Specifically, we aimed to determine whether cross-domain class imbalance poses a problem for domain alignment methods, under which conditions it affects performance, and how it can be tackled. We first start by determining whether there actually is a domain gap between our CHECK dataset and CHECK Gamma dataset.

5.1. CHECK Gamma Dataset Domain Adaptation Results

We observe a domain shift between the CHECK and CHECK Gamma dataset by the performance gap between the source-only and target-only baselines (The orange and blue lines in Figure 6). The experiments from Section 4.6 show that incorporating distribution alignment with any sampling strategy into the model improves generalization to the target dataset. This means that distribution alignment was successful in extracting domain invariant features and that it could transfer its domain knowledge from the labeled source domain to the unlabeled target domain.

When comparing distribution alignment methods employing different sampling strategies, we observe that the classification performance achieved using more complex sampling approaches is comparable to that obtained with simple random sampling. Even when the target dataset has an extremely skewed 30/70 class balance in the target domain (Figure 6) the performance differences were not demonstrable.

We observe that models trained on a balanced source distribution achieve higher target-domain performance than those trained on an imbalanced source. This outcome supports the idea that balanced data mitigates overfitting to the dominant class and promotes the learning of more generalizable, class-invariant representations, while simultaneously also guaranteeing distribution alignment's assumption that class priors across domains are similar. Although uniform sampling enforces an artificial balance between classes, it may help reduce biases toward the majority class and encourage more equal feature learning across classes.

We found out that the direction of domain adaptation is important, because the performance gains depend on which dataset is used as the source and which as the target. Specifically, we see a larger improvement in classification performance when adapting from CHECK to CHECK Gamma than when adapting from CHECK Gamma to CHECK (Appendix A. Figure 11). This difference may arise from the fact that CHECK Gamma images exhibit higher overall intensity than CHECK images, and the ResNet-18 model, pre-trained on ImageNet, appears to adapt more readily to darker image domains than to brighter ones. Consequently, domain adaptation is not symmetric—the direction of transfer influences how well the model aligns feature distributions and generalizes across domains. The smaller gain observed when adapting from CHECK Gamma to CHECK further suggests that the distribution shift in this direction is less pronounced, reducing the potential benefit of alignment-based methods such as MMD.

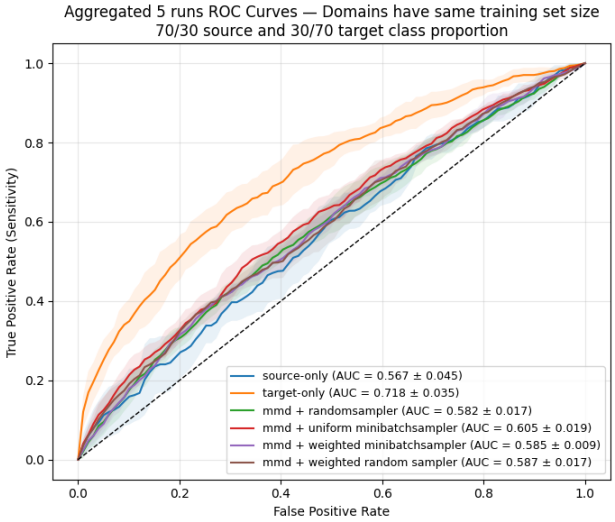
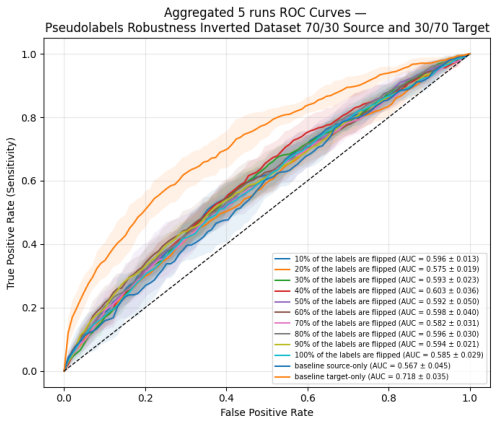


Figure 7: The AUC scores of the upperbound and lowerbound are still similar to the ones from CHECK Gamma. However, the performance gains after distribution alignment are still far from the upperbound (orange line) compared to CHECK Gamma.

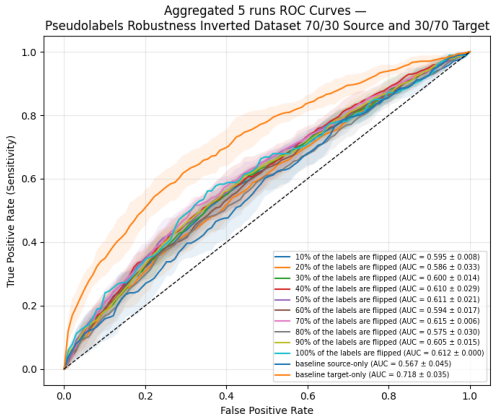
5.2. Pixel-Inverted Synthetic Dataset Domain Adaptation Results

Since the model after distribution alignment showed no sensitivity to cross-domain class imbalance on the CHECK Gamma dataset, we take a look at a more challenging synthetic dataset to further evaluate its robustness. In this dataset, we applied pixel inversion to the target images to induce a more substantial appearance-based domain shift. Under this more complex setting, the model exhibits difficulties in adapting to the target domain, even after applying distribution alignment. Across most scenarios, the different model configurations all achieve comparable performance. The results indicate that pixel inversion introduces a more challenging domain shift than gamma correction, as reflected by the larger gap between the source-only and target-only baselines and the smaller improvement over the baseline when implementing domain adaptation. To aid interpretation, we report only the highly imbalanced scenario where the source has a balance of 70/30 and the target has a balance of 30/70, as this shows performance differences (Figure 7) between feature aligment with the different sampling methods.

Our proposed weighted random sampling and weighted minibatch sampling strategies, along with the uniform mini-



(a) Minibatch sampler tested on different pseudolabel noise thresholds



(b) Weighted random sampler tested on different pseudolabel noise thresholds

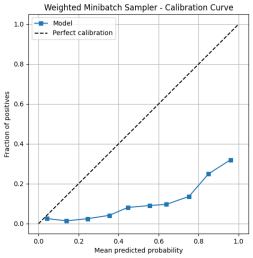
batch sampling strategy, show improvements over standard MMD under conditions of both extreme class imbalance and severe domain shift. Among these, the uniform sampling strategy improves over random sampling the most. Although the gains are minimal, they indicate that the proposed methods can contribute to improving adaptation when both severe class imbalance and complex domain shift are present.

5.3. Pseudolabel Robustness Results

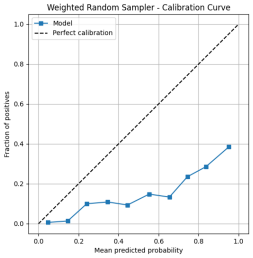
The results of the pseudolabel noise experiments indicate that the model demonstrates a high degree of robustness to random label noise (Figure 8b). Its performance remains relatively stable across all evaluated thresholds, suggesting that the method does not strongly depend on perfect label accuracy when incorporating pseudolabels. This stability is important in practice, as pseudolabels are inherently noisy in real-world applications, and the ability to tolerate such noise reflects the reliability of the proposed approach.

The calibration curves of the pseudolabels in our proposed methods (Figure 8) indicate that the models suffer from poor calibration. In both the weighted minibatch sampler and the weighted random sampler settings, the predicted confidence levels are consistently much higher than the observed accuracy, which indicates that the model is overconfident in its predictions. This mismatch suggests that although the models are able to assign high confidence to their predictions, these probabilities do not reflect the true likelihood of correctness.

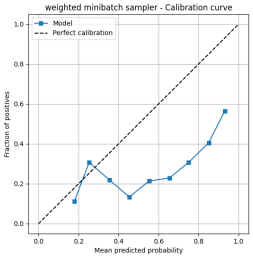
When comparing the two sampling strategies, the weighted minibatch sampler demonstrates comparatively better calibration than the weighted random sampler, but still far from ideal. This relative improvement may stem from the minibatch sampling approach always having both of the classes per batch, which could lead to less extreme shifts in confidence scores. The persistence of overconfidence across both methods highlights the inherent difficulty of maintaining calibrated probabilities under class-imbalanced and domain-shifted conditions, particularly when pseudolabels are used.



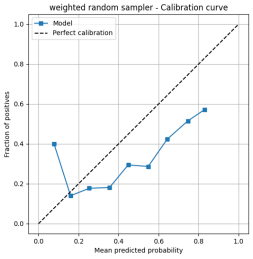
(a) Calibration curve of the weighted minibatch sampler with model trained on CHECK tested on OAI.



(b) Calibration curve of the weighted random sampler with model trained on CHECK tested on OAI.



(c) Calibration curve of the weighted minibatch sampler with model trained on CHECK tested on Gamma CHECK.



(d) Calibration curve of the weighted random sampler with model trained on CHECK tested on Gamma CHECK.

Figure 8: Calibration curves for weighted samplers across domain configurations.

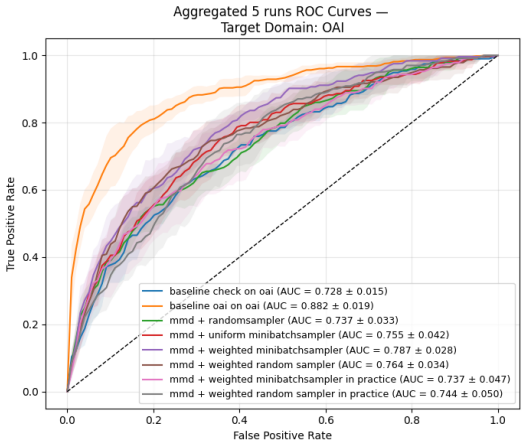


Figure 9: Comparison of sampling methods trained on CHECK and deployed on OAI.

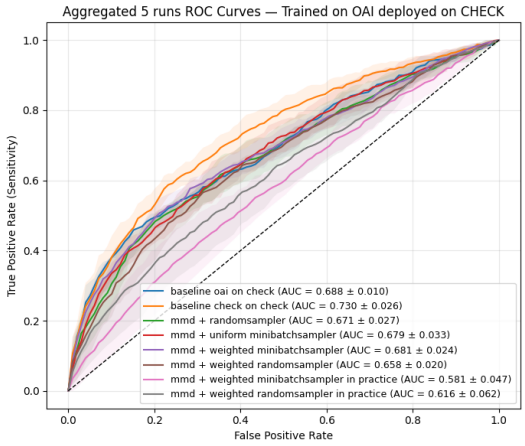
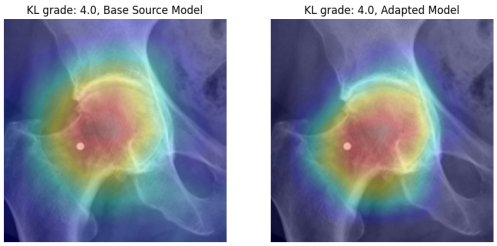


Figure 10: Comparison of sampling methods trained on OAI and deployed on CHECK.

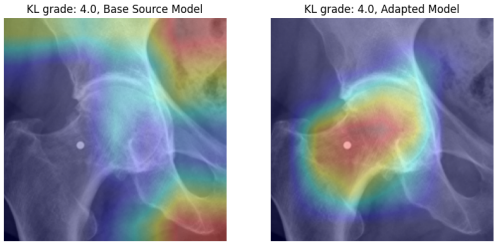
5.4. Real World OAI and CHECK Results

There is a substantial domain shift between OAI and CHECK, one that is larger than the shifts introduced in the synthetic datasets, making this a more complex adaptation problem (Figures 9 and 10). Across model configurations, the results show that generalization to the target domain improves when MMD is applied as a distribution alignment method, this is consistent with our findings on the synthetic datasets. Even after domain adaptation the model performance remains well below the target-only baseline, which represents the upper bound. Using MMD + uniform minibatch sampler yields a little improvement over using MMD + random sampler. Our proposed sampling method provides a modest improvement over standard MMD provided with the truth labels, in practice with pseudolabels from the classifier the performance is comparable to that of MMD + random sampler, which could be attributed to the badly performing pseudolabeler.

The domain gap from OAI to CHECK is relatively small compared to the other direction. When training on OAI and deploying on CHECK, we observe no improvement in performance from domain adaptation and in practice we even see a hard decrease in performance. A similar pattern was observed in our experiments with the CHECK Gamma and CHECK datasets, suggesting that domain adaptation may be less effective when the baseline model already generalizes well. It is possible that the remaining minor differences between the datasets cannot be fully addressed by standard domain adaptation techniques alone. Additionally it can also mean that the CHECK dataset contains more diverse samples, therefore making the model better at generalizing to other domains.



(a) Heatmap visualization of model prior to distribution alignment. Image is taken from the CHECK dataset.



(b) Heatmap visualization of model prior to distribution alignment. Same image as the one above but from the CHECK Gamma dataset.

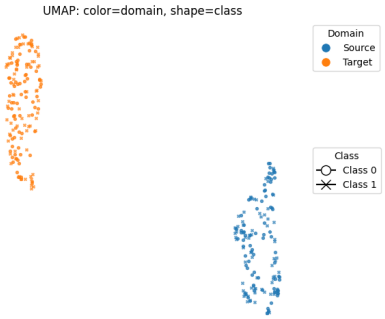
5.5. Grad-CAM Results

To gain insight into model decision-making, we applied Grad-CAM to a sample from the CHECK dataset (Figure 11a). The visualization shows that both the baseline and the adapted models primarily focus on the femoral head when making classifications. We speculate that the baseline model without distribution alignment attends to the femoral head on source-domain data, but its focus may become less interpretable when applied to gamma-corrected target data. Following domain adaptation, we expect the model to preserve attention on the femoral head across both source and target domains.

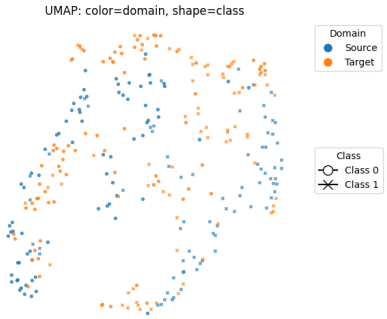
The same sample, after applying a gamma correction, is shown in Figure 11b. In this case, the baseline model exhibits diffuse and poorly localized attention, whereas the model trained with distribution alignment continues to concentrate on the femoral head. This indicates that feature alignment enables the adapted model to maintain relevant focus in the target domain, suggesting that feature alignment facilitates the transfer of discriminative features to the target domain.

5.6. UMAP Results

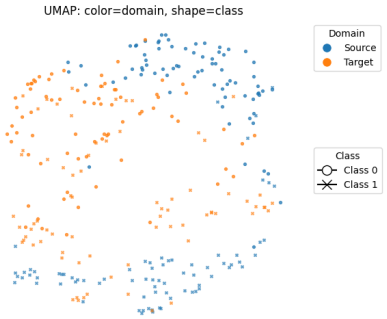
Prior to adaptation, the embeddings reveal a clear separation between the CHECK and CHECK Gamma domains (Figure 11a), reflecting the divergence introduced by the gamma correction transformation. After applying MMD, the target samples align more closely with the source distribution, resulting in better overlap between domains. Both domains now overlap (Figures 11b, 11c, 11d and 11e). This visualization provides an intuitive confirmation that MMD facilitates the learning of domain-invariant features.



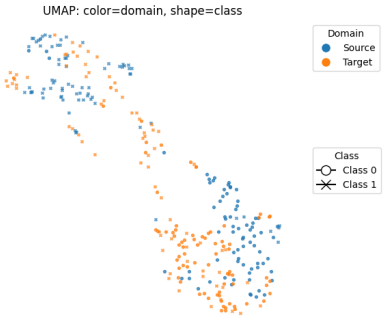
(a) UMAP visualisation of the CHECK and CHECK Gamma datasets of a model trained on CHECK without distribution alignment.



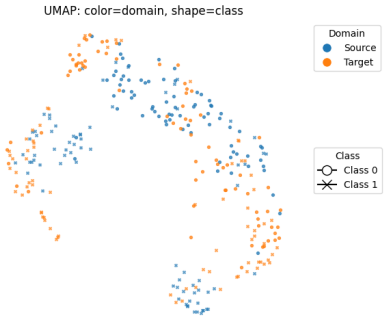
(b) UMAP visualisation of the CHECK and CHECK Gamma datasets of a model trained on CHECK with distribution alignment using random sampling.



(c) UMAP visualisation of the CHECK and CHECK Gamma datasets of a model trained on CHECK with distribution alignment using uniform minibatch sampling.



(d) UMAP visualisation of the CHECK and CHECK Gamma datasets of a model trained on CHECK with distribution alignment using weighted minibatch sampling.



(e) UMAP visualisation of the CHECK and CHECK Gamma datasets of a model trained on CHECK with distribution alignment using weighted random sampling.

Figure 11: UMAP visualizations of model embeddings without distribution alignment and with distribution alignment using different sampling strategies.

6. Discussion

6.1. Cross-Domain Class Imbalance

In our experiments we found that the impact of cross-domain class imbalance starts to become noticable under complex domain shift. Because the images in the dataset itself are unchanged and only the transformation to the dataset images differs, we think that distribution alignment is harder under the pixel inversion domain shift compared to the gamma correction domain shift.

We observe that the effect of cross-domain imbalance is limited under a simple domain shift such as the Gamma corrected dataset. Even in the most challenging scenario, where the source domain is heavily skewed and the target domain exhibits the inverse class imbalance, MMD demonstrates performance comparable to those obtained with sampling-based strategies, suggesting that MMD is relatively robust to shifts in class priors and is not substantially affected by cross-domain imbalance when the domain shift is simple.

Our proposed sampling strategy improves performance relative to random sampling (Figure 7), though the gains are small. MMD with random sampling still provides gains over the source-only baseline, confirming that aligning feature distributions between source and target domains can mitigate the effects of domain shift to some extent. These results suggest that more complex sampling mechanisms can help the model better preserve class relationships across domains and reduce sensitivity to cross-domain class imbalance.

The observed performance differences between the adaptation strategies MMD with random sampling, MMD with a uniform minibatch sampler, MMD with a weighted minibatch sampler, and MMD with a weighted random sampler cannot be fully explained by misalignment caused by cross-domain class imbalance. While the uniform and weighted sampling strategies are specifically designed to mitigate cross-domain class imbalance, no consistent significant improvement was observed across configurations in the CHECK Gamma dataset experiments. We identify two factors that may contribute to this:

1. **Limited complexity of the domain shift.** The synthetic gamma correction introduces smooth, global intensity changes that neural networks can often compensate for with relative ease. Compared to more complex shifts, such as scanner artifacts or textural differences, this transformation may not sufficiently challenge adaptation methods. As a result, advanced sampling strategies could provide limited observable gains in this setting.

2. **Ambiguity in class boundaries.** The distinction between osteoarthritis (OA) and healthy cases is not entirely objective. In this study, we grouped Kellgren–Lawrence (KL) grades 0–1 as "non-OA" and grades 2–4 as "OA." However, the boundary between KL=1 and KL=2 is subtle and may not be reliably identified on radiographs [25]. Consequently, the effective decision boundary between classes is unclear, and misclassifications may be driven as much by label uncertainty as by domain adaptation challenges. This limits the observable impact of class imbalance, as sampling strategies must contend with label uncertainty when constructing minibatches.

A possible explanation for the improvements often reported in the literature for class-conditioned domain alignment methods may relate to the nature of the tasks and datasets typically used in those studies. Most existing works evaluate their methods on multi-class classification problems, such as object recognition across visual domains (e.g., Office-Home, VisDA, or DomainNet). In such settings, the class differences are typically large and visually distinct e.g. distinguishing between categories such as "mouse", "keyboard", and "monitor". Consequently, the decision boundaries between classes are well-defined, and any misalignment between source and target features across classes could have more negative impact on performance [4]. Aligning features in a class-conditional manner therefore mitigates this issue and results in a more prominent performance gain.

In contrast, binary osteoarthritis classification presents a more subtle and complex scenario. The classes represent different degrees of joint degeneration rather than distinct object categories. This means that the underlying visual features exist along a continuous spectrum from healthy to non-healthy. As a result, the distinction between classes is less discrete, and a certain degree of feature overlap is expected. In such cases, misalignment between classes could have a less pronounced effect on model performance, since the boundary between classes is inherently ambiguous.

Therefore, while class-conditioned domain adaptation can be effective in some contexts, its success could depend on the nature of the task and the underlying domain characteristics. Our results suggest that its benefits are not generally guaranteed, as its effectiveness may diminish when domain differences are subtle or when class boundaries are less distinct. Rather than assuming it will generalize universally, we should apply it with considerations of the specific task and data.

6.2. Sampling Strategy

Incorporating MMD consistently improves model generalization to the target domain across all experiments. We have observed that random sampling may occasionally yield minibatches dominated by a single class purely by chance, whereas structured samplers guarantee that sam-

ples from both classes are represented in each minibatch. While more complex sampling strategies do not always produce substantial performance gains over random sampling, they can offer benefits by ensuring greater class diversity within minibatches during domain alignment. A more balanced composition of minibatches can promote more stable and effective feature alignment. Similar observations have been reported by Jiang et al. [16], who found that their uniform minibatch sampling approach performed better when a larger number of different classes were included in each minibatch.

6.3. Pseudolabel Noise

We hypothesized that class-conditioned sampling could improve domain alignment by minimizing the negative impact of pseudolabel noise, it should help the model remain robust even when some pseudolabels are incorrect.

However, when we introduced random noise into the pseudolabels, we detected minimal variation in model performance. This observation highlights an important nuance: the insensitivity to pseudolabel noise does not necessarily demonstrate resilience of the method, but may instead reflect that the sampler plays only a marginal role in influencing alignment outcomes in our experiments. This is also reflected in our results, where the improvement between the random sampler and other sampling strategies are minimal. Similar considerations have been discussed by Jiang et al. [16], though their setting revealed clearer benefits of class-conditioned sampling than we observe here. In our case, the minimal performance difference suggests that the method can tolerate some level of pseudolabel noise, largely because the sampling strategy itself contributes only modestly to overall alignment performance.

In the synthetic pixel inverted dataset our proposed sampling methods already had minimal improvement over the random sampling method. This means that with perfect labeling, we could improve current domain adaptation methods by implementing class conditioned domain alignment.

However, throughout our experiments, the proposed sampling strategy did not yield any improvements in practice. This indicates that class-conditioned domain alignment contributes little to overall performance gains in practice.

6.4. Feature Visualization

UMAP and Grad-CAM provide an intuitive illustration to see if domain adaptation has worked. Ideally, UMAP would not only show alignment between domains but also reveal a distinct separation of classes within the shared representation space. This would mean that our model would have achieved both domain invariance and class discriminability. However, in our case, the embeddings primarily highlight the former: while the domains appear to be well aligned, the class boundaries remain less clearly defined. However, the lack of separation between the classes in the 2D embedding space does not necessarily imply that the model fails to capture class-discriminative features. It is also possible that the relevant decision boundary exists in a higher-dimensional feature space that cannot be fully represented in a two-dimensional projection.

6.5. Limitations

Several factors limit the generalizability of our research. First, the problem was formulated as a binary classification task, distinguishing only between healthy and osteoarthritic cases. While this simplifies the learning problem, it does not capture the full spectrum of disease severity present in osteoarthritis, potentially overlooking subtle variations that could influence domain alignment behavior.

Second, osteoarthritis classification is inherently challenging due to the gradual and heterogeneous nature of disease progression and subtle radiographic features. As a result, the complexity of the task may obscure the effects of distribution alignment and sampling strategies. Using a dataset with clearer class separation or more pronounced visual differences between classes might produce more distinct results and make the impact of cross-domain class imbalance easier to observe.

Third, our datasets are relatively small, which constrains the experimental configurations we can evaluate. To conduct controlled experiments, we had to undersample portions of the data, which can negatively affect model performance. Limited data also restricts the diversity of examples available for training, potentially limiting the effectiveness of both MMD-based adaptation and the proposed sampling strategies.

The combination of small dataset size, binary class structure and a gradual classification problem may collectively restrict the performance gains achievable through both domain adaptation and more complex sampling strategies.

6.6. Future Works

Future work could aim to design experiments that expose the conditions under which domain adaptation methods begin to reach their limit, such as highly cross-domain imbalanced settings or more complex but still clinically realistic distribution shifts. Stress-testing in these regimes would allow clearer demonstrations of when and why domain adaptation with sampling strategies are necessary, and when simple baselines may suffice.

7. Conclusion

In this study, we extensively showed through controlled experiments that cross-domain class imbalance does indeed pose a problem in extremely skewed cross-domain imbal-

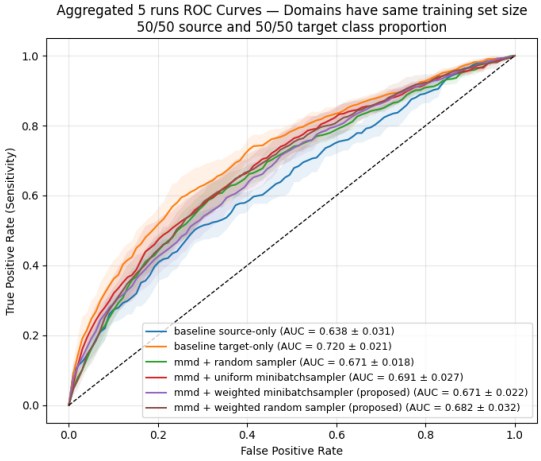
ances and complex domain shifts. To tackle this problem we introduced a weighted minibatch sampling strategy extending the implicit class-conditioned domain alignment method of Jiang et al. [16] for cross-domain osteoarthritis classification from X-ray images. We compared this with another approach that uses a random weighted sampling method applied to the target domain, and the results indicate that both strategies achieve roughly comparable performance.

Our experiments show that the impact of cross-domain class imbalance becomes noticeable only when the domain shift is sufficiently complex and the class imbalance is substantial. In cases where the domain shift is relatively simple, even under extreme class imbalance, models using MMD with random sampling achieved performance similar to that of more complex sampling variants. This suggests that, for binary OA classification, cross-domain class imbalance may not be a primary limiting factor for distribution alignment with MMD. The proposed sampling strategies yielded modest improvements over random sampling only in the most extreme scenarios, even when using ground-truth labels, indicating that class-conditioned sampling would not be of help in a real world unsupervised setting.

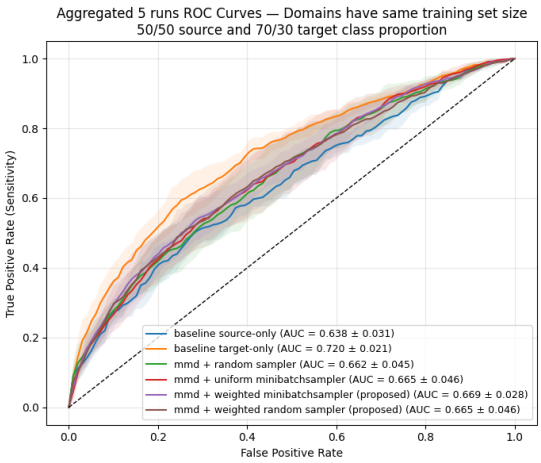
We further investigated the effect of pseudolabel noise in the scenarios where a performance difference between random and weighted sampling was observed. Performance degradation with increasing pseudolabel corruption was minimal. However, since the performance gains achieved with the proposed sampling methods were minimal, the results do not provide sufficient evidence to confirm that implicit class-conditioned domain alignment is robust to pseudolabel noise.

Taken together, our findings highlight that MMD remains an effective approach for mitigating domain shift and is robust to cross-domain class imbalance in simple domain shifts. It starts to reach its limits when confronted with a skewed cross-domain class imbalance and complex domain shift. Extending ICCDA with weighted uniform sampling provides little additional benefit under extreme domain shifts and extreme cross-domain class imbalance in the context of binary OA classification.

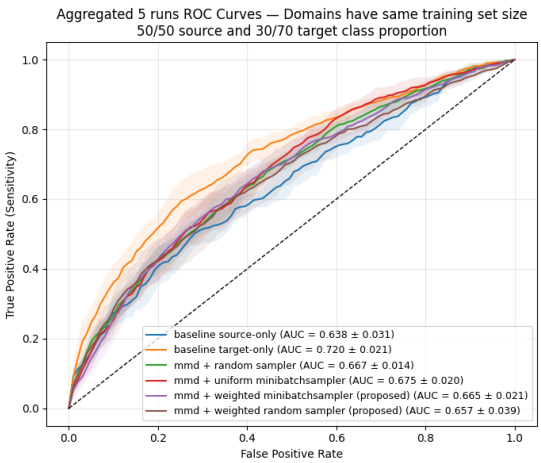
A. Trained on CHECK Gamma results



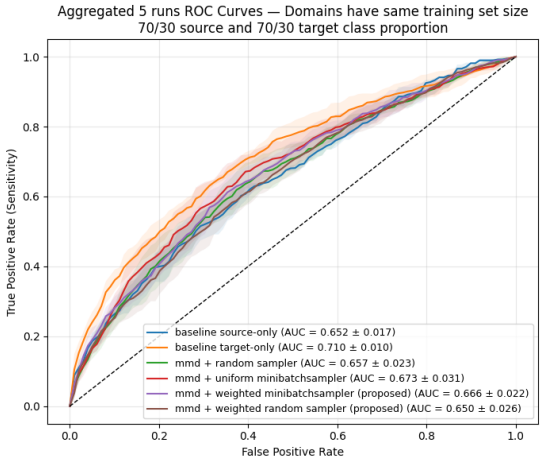
(a) Balanced source and target.



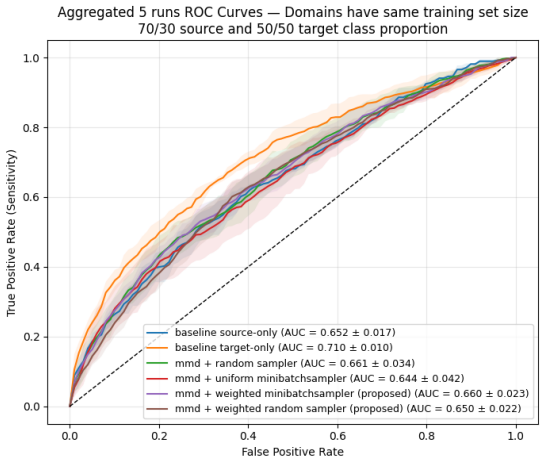
(b) Balanced source and unbalanced target.



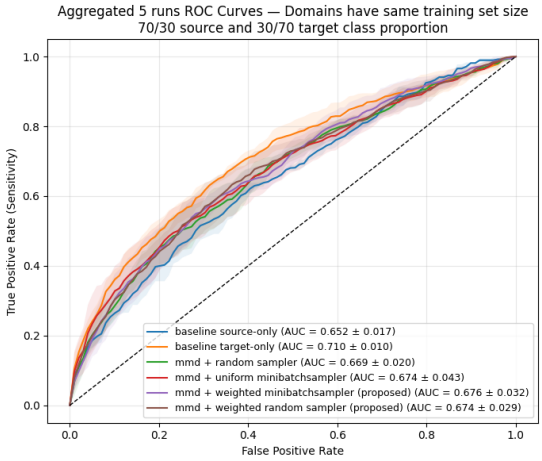
(c) Balanced source and unbalanced target inverted.



(d) Unbalanced source and target.



(e) Unbalanced source and unbalanced target.



(f) Unbalanced source and unbalanced target inverted.

Figure 11: We observe lower performance gains when training on CHECK Gamma and deploying on CHECK.

B. Finding Gamma Value results

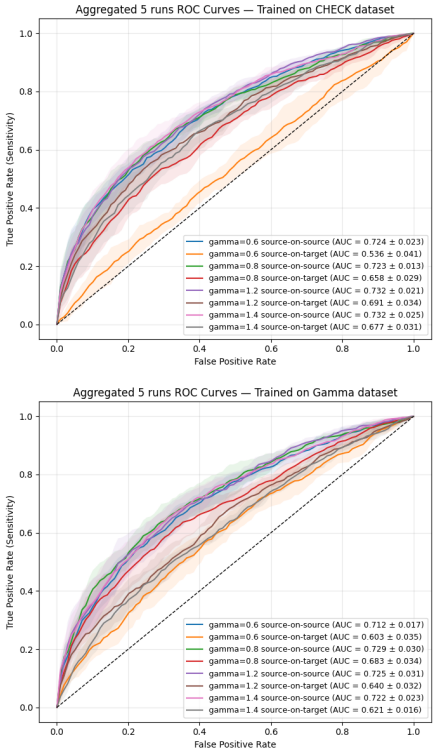
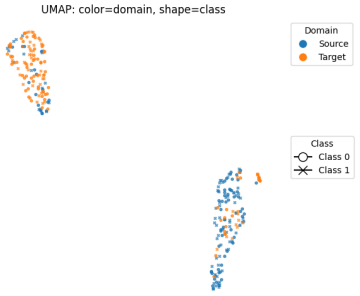
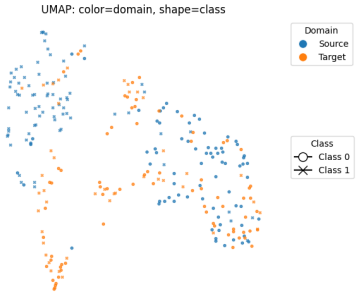


Figure 12: Domain shift when applying different gamma corrections.

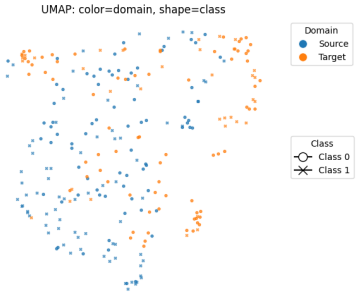
C. UMAP CHECK and OAI



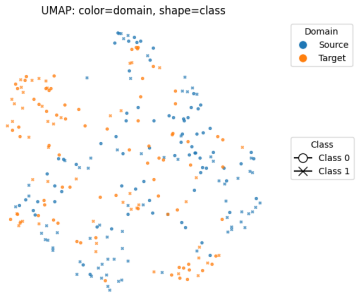
(a) Classifier without distribution alignment



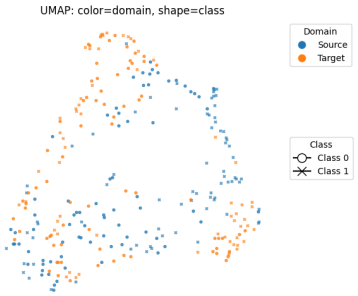
(b) Classifier with MMD and random sampling



(c) Classifier with MMD and uniform mini-batch sampling



(d) Classifier with MMD and weighted mini-batch sampling



(e) Classifier with MMD and weighted random sampling

Figure 12: UMAP visualizations of classifier embeddings before and after different MMD-based distribution alignment strategies.

References

[1] Osteoarthritis initiative. https://nda.nih.gov/oai.

[2] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang. Recent advances in adversarial training for adversarial robustness. *CoRR*, abs/2102.01356, 2021.

[3] A. Belal, A. Meethal, F. P. Romero, M. Pedersoli, and E. Granger. Attention-based class-conditioned alignment for multi-source domain adaptation of object detectors, 2024.

[4] A. Belal, A. Meethal, F. P. Romero, M. Pedersoli, and E. Granger. Multi-source domain adaptation for object detection with prototype-based mean-teacher, 2024.

[5] C. Chen, W. Xie, T. Xu, W. Huang, Y. Rong, X. Ding, Y. Huang, and J. Huang. Progressive feature alignment for unsupervised domain adaptation. *CoRR*, abs/1811.08585, 2018.

[6] Z. Deng, Y. Luo, and J. Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. *CoRR*, abs/1903.09980, 2019.

[7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks, 2016.

[8] L. Gao, L. Zhang, C. Liu, and S. Wu. Handling imbalanced medical image data: A deep-learning-based one-class classification approach. *Artificial Intelligence in Medicine*, 108:101935, 2020.

[9] M. García-Domínguez, C. Domínguez, J. Heras, E. J. Mata, and V. Pascual. Neural style transfer and unpaired image-to-image translation to deal with the domain shift problem on spheroid segmentation. *CoRR*, abs/2112.09043, 2021.

[10] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[11] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

[12] H. Guan and M. Liu. Domain adaptation for medical image analysis: A survey. *CoRR*, abs/2102.09508, 2021.

[13] H. Guan and M. Liu. Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2022.

[14] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou. On the class imbalance problem. *Fourth International Conference on Natural Computation, ICNC '08*, Vol. 4, 10 2008.

[15] T. M. H. Hsu, W. Y. Chen, C.-A. Hou, Y.-H. H. Tsai, Y.-R. Yeh, and Y.-C. F. Wang. Unsupervised domain adaptation with imbalanced cross-domain data. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4121–4129, 2015.

[16] X. Jiang, Q. Lao, S. Matwin, and M. Havaei. Implicit class-conditioned domain alignment for unsupervised domain adaptation. *CoRR*, abs/2006.04996, 2020.

[17] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018.

[18] J. Li, G. Li, Y. Shi, and Y. Yu. Cross-domain adaptive clustering for semi-supervised domain adaptation. *CoRR*, abs/2104.09415, 2021.

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

[20] X. Liu, C. Yoo, F. Xing, H. Oh, G. E. Fakhri, J.-W. Kang, and J. Woo. Deep unsupervised domain adaptation: A review of recent advances and perspectives, 2022.

[21] M. Long and J. Wang. Learning transferable features with deep adaptation networks. *CoRR*, abs/1502.02791, 2015.

[22] Z. Murez, S. Kolouri, D. J. Kriegman, R. Ramamoorthi, and K. Kim. Image to image translation for domain adaptation. *CoRR*, abs/1712.00479, 2017.

[23] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura. Handling imbalanced medical datasets: review of a decade of research. *Artificial Intelligence Review*, 57(10):273, 09 2024.

[24] D. Saxena and J. Cao. Generative adversarial networks (gans): Challenges, solutions, and future directions. *CoRR*, abs/2005.00065, 2020.

[25] D. Schiphof, B. de Klerk, H. Kerkhof, A. Hofman, B. Koes, M. Boers, and S. Bierma-Zeinstra. Impact of different descriptions of the kellgren and lawrence classification criteria on the diagnosis of knee osteoarthritis. *Annals of the Rheumatic Diseases*, 70(8):1422–1427, 2011.

[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct. 2019.

[27] B. Sun, J. Feng, and K. Saenko. Correlation alignment for unsupervised domain adaptation, 2016.

[28] Y. Sun, E. Tzeng, T. Darrell, and A. A. Efros. Unsupervised domain adaptation through self-supervision, 2019.

[29] M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[30] W. Wang, H. Li, Z. Ding, F. Nie, J. Chen, X. Dong, and Z. Wang. Rethinking maximum mean discrepancy for visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1):264–277, 2023.

[31] J. Wesseling, M. Boers, M. A. Viergever, W. K. Hilberdink, F. P. Lafeber, J. Dekker, and J. W. Bijlsma. Cohort profile: Cohort hip and cohort knee (check) study. *International Journal of Epidemiology*, 45(1):36–44, 08 2014.

[32] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation, 2017.

[33] Y. Zhang, T. Liu, M. Long, and M. I. Jordan. Bridging theory and algorithm for domain adaptation. *CoRR*, abs/1904.05801, 2019.

[34] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017.