



HAL
open science

Bridging Contrastive Learning and Domain Adaptation: Theoretical Perspective and Practical Application

Gonzalo Iñaki Quintana, Laurence Vancamberg, Vincent Jugnon, Agnès
Desolneux, Mathilde Mougeot

► **To cite this version:**

Gonzalo Iñaki Quintana, Laurence Vancamberg, Vincent Jugnon, Agnès Desolneux, Mathilde Mougeot. Bridging Contrastive Learning and Domain Adaptation: Theoretical Perspective and Practical Application. 2025. hal-04922401

HAL Id: hal-04922401

<https://hal.science/hal-04922401v1>

Preprint submitted on 30 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0
International License

Bridging Contrastive Learning and Domain Adaptation: Theoretical Perspective and Practical Application

Gonzalo Iñaki Quintana^{1,2} Laurence Vancamberg² Vincent Jugnon² Agnès Desolneux¹ Mathilde Mougeot^{1,3}

Abstract

This work studies the relationship between Contrastive Learning and Domain Adaptation from a theoretical perspective. The two standard contrastive losses, NT-Xent loss (Self-supervised) and Supervised Contrastive loss, are related to the Class-wise Mean Maximum Discrepancy (CMMD), a dissimilarity measure widely used for Domain Adaptation. Our work shows that minimizing the contrastive losses decreases the CMMD and simultaneously improves class-separability, laying the theoretical groundwork for the use of Contrastive Learning in the context of Domain Adaptation. Due to the relevance of Domain Adaptation in medical imaging, we focused the experiments on mammography images. Extensive experiments on three mammography datasets - synthetic patches, clinical (real) patches, and clinical (real) images - show improved Domain Adaptation, class-separability, and classification performance, when minimizing the Supervised Contrastive loss.

1. Introduction

Given a source data distribution or domain, we are often interested in transferring the representation learned to a different, albeit related, target domain. This is crucial for leveraging models pre-trained on large annotated datasets, as well as for adapting test and training distributions, which are generally different (de Mathelin et al., 2021). In particular, Domain Adaptation (DA) methods seek to minimize the effects of the domain shift to enable more efficient transfer. This is especially relevant in the medical imaging domain, where high data variability and limited access to large datasets pose significant challenges to the development of Deep Learning (DL)-based solutions, often hindering model generalization and performance across diverse clinical settings (Garrucho et al., 2022).

¹Centre Borelli, CNRS & ENS Paris-Saclay, F-91190 Gif-sur-Yvette, France ²GE HealthCare, 78530 Buc, France ³ENSIE, 91000 Evry, France. Correspondence to: Gonzalo Iñaki Quintana <gonzalo.quintana@ens-paris-saclay.fr>

Contrastive Learning (CL) is a learning paradigm where semantically similar data-points are close to one another in the feature space, enabling to learn representations that are invariant given certain transformations. Intuitively, mapping data points from different domains to the same region in the feature space mirrors the DA problem. In addition CL separates the representations of semantically different data-points, which has been found to be beneficial for downstream task performance, like classification, detection, or segmentation. Contrastive Learning has been widely applied in the medical imaging domain (Chaitanya et al., 2020; Dong & Voiculescu, 2021; Cao et al., 2021; Quintana et al., 2024).

Inspired by the similarity of the tasks that Domain Adaptation and Contrastive Learning pursue, as well as by the growing interest in CL for DA, we analyze both paradigms to provide theoretical justifications for applying CL to DA. Due the relevance of Domain Adaptation in medical imaging, we conduct experiments on mammography images for classification tasks, specifically determining the presence or absence of breast cancer.

1.1. Related work

Domain Adaptation. Let $\mathcal{D}_s = \{\mathcal{X}_s \times \mathcal{Y}_s, \pi_s\}$ be a source domain and $\mathcal{D}_t = \{\mathcal{X}_t \times \mathcal{Y}_t, \pi_t\}$ a target domain, where \mathcal{X} is an instance or covariate space, \mathcal{Y} is the label space, and $\pi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ a joint probability measure. The target domain \mathcal{D}_t is typically unlabeled, i.e., $\mathcal{Y}_t = \emptyset$, contains fewer labels than \mathcal{D}_s , or has a smaller dataset. In Domain Adaptation, we seek to transfer the representations learned in the source domain for solving a source task \mathcal{T}_s to the target domain, while considering that source and target tasks are the same. Various DA strategies have been proposed based on the nature of the domain shift (e.g., covariate shift, prior probability shift, concept shift), the availability of labels in the target domain (supervised, unsupervised, semi-supervised), and the type of models employed (e.g., shallow or deep architectures). In this work we consider the *hidden covariate shift* (de Mathelin et al., 2023) or *covariate observation shift* (Kull & Flach, 2014), a subtype of concept shift where it is assumed that there exists a non-linear transformation of the covariates that eliminates the shift. One

of the most widely used DA method consists in aligning domains by minimizing a domain dissimilarity measure, such as the Mean Maximum Discrepancy (MMD) (Gretton et al., 2006), the Kullback-Leibler divergence, the Wasserstein distance (Damodaran et al., 2018; Lee et al., 2019), or the Bregman divergence (Farahani et al., 2021). Long et al. (2013) first proposed Deep Adaptation Network (DAN), where the MMD is used to minimize the marginal distribution shift, and extended it to matching both the marginal and conditional distributions with Joint Adaptation Network (JAN) (Long et al., 2017). However, minimizing these dissimilarity measures has been reported to attain DA at the expense of reducing feature-label correlation, decreasing class-separability in the feature space and negatively impacting downstream tasks like classification or detection (Wang et al., 2021). Domain-adversarial Neural Network (DANN) (Ganin et al., 2016) and its variants (Long et al., 2018; Shen et al., 2018; Tzeng et al., 2017) consists in jointly training an encoder, classifier, and domain discriminator to obtain domain invariant representations. These adversarial methods require training an additional network with an unstable min-max loss, which often demands extra training time and computational resources (Kouw & Loog, 2021).

Contrastive Learning consists in learning representations where positive pairs of features are dragged to the same region of the feature space, and negative pairs are pushed apart. The definition of positive and negative pairs differs on each application and on the availability of annotations. CL was first introduced as a max-margin loss for dimensionality reduction (Hadsell et al., 2006; Chopra et al., 2005), which later evolved into the triplet (Weinberger & Saul, 2009) and N-pair-mc (Sohn, 2016), which improved convergence and removed the need for negative hard mining. The NT-Xent loss (Chen et al., 2020), a temperature-scaled version of the N-pair-mc loss, is currently one of the most widely used losses for Self-supervised representation learning, both in Computer Vision (Chen et al., 2020; Oord et al., 2018) and in Natural Language Processing (Gao et al., 2021). In this context, positive pairs are typically transformed version of the same instance, while negative pairs are all pairs of features that originate from different instances. It has also been used for multimodal learning with text-image aligned representations, which has applications in zero-shot image classification (Radford et al., 2021; Jia et al., 2021), image retrieval (Huang et al., 2024; Schall et al., 2024), and text-conditioned image generation (Rombach et al., 2022). More recently, the Supervised Contrastive loss was introduced (Khosla et al., 2020) to enable features from different instances to be mapped closely in the feature space. In this case, positive pairs come from instances with the same label, and negative pairs from instances with different labels (Khosla et al., 2020; Li et al., 2022). Recently, Contrastive Learning has started gaining traction as a Domain Adapta-

tion method (Thota & Leontidis, 2021; Darban et al., 2024; Singh, 2021). However, despite promising empirical results, a theoretical understanding of the DA capabilities of Contrastive Learning is lacking.

Mammography image classification is crucial for improving screening or diagnostic workflow and accuracy. Its clinical applications span from triaging normal (lesion-free) and abnormal exams to enhance image reading efficiency, to assessing the likelihood of breast cancer and recommending biopsy procedures (Kyono et al., 2020). Today’s state-of-the-art models rely on Convolutional Neural Network (CNN) patch-based approaches: a Deep Learning model is first pre-trained for patch classification and then extended to full-image classification by adding additional layers and re-training (Shen et al., 2019; Petrini et al., 2022; Quintana et al., 2023). Currently, there is no publicly available reference dataset of digital mammograms, primarily due to the high cost of obtaining sufficiently large and diverse annotated datasets. DL models typically achieve an AUC ranging from 0.75 to 0.81 for benign vs. malignant classification (Bobowicz et al., 2023; Petrini et al., 2022; Shen et al., 2019). Multi-view models increase that range to 0.83-0.89 by leveraging different views of the same breast and bi-lateral asymmetries between left and right breasts, and using ensembling (Wu et al., 2019; Bobowicz et al., 2023; Petrini et al., 2022). In this work, we focus on studying CL and DA and not on establishing a new benchmark performance.

1.2. Main contributions

The main contributions of this work are the following:

- We show that minimizing the NT-Xent loss and the Supervised Contrastive loss decreases the CMMD, thus improving Domain Adaptation.
- We show that minimizing the contrastive losses improves class-separability, by extending the work of Li et al. (2021).
- We validate these theoretical results by conducting experiments in a concrete mammography image classification application, using three distinct datasets and the Supervised Contrastive loss.
- We introduce a synthetic mammography image dataset based on Gaussian textures and simple lesion simulation. The dataset, along with the code for its generation and for reproducing the experiments in this work, can be found in this repository: github.com/gonzaq94/contrastive-da-synthetic-patch.

2. Contrastive Learning and dissimilarity measures

Consider a learning problem with data from two labeled domains $\mathcal{D}_0 = \{\mathcal{X} \times \mathcal{Y}, \pi_0\}$ and $\mathcal{D}_1 = \{\mathcal{X} \times \mathcal{Y}, \pi_1\}$, with $\pi_d : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$ the joint probability measure of the instances $x \in \mathcal{X} \subseteq \mathcal{R}^{n_1 \times n_2}$ and labels $y \in \mathcal{Y}$ of the d -th domain. We denote by $\pi_d^{\mathcal{X}}$ and $\pi_d^{\mathcal{Y}}$ the marginal probability measures on the instances and labels, and by $\pi_{d,c}^{\mathcal{X}|\mathcal{Y}}$, $c \in \mathcal{Y}$, the conditional probability measure on the instances knowing the label is c . We also consider the mixture domain $\mathcal{D}_p = \{\mathcal{X} \times \mathcal{Y}, \pi_p\}$ with joint probability measure $\pi_p := p\pi_1 + (1-p)\pi_0$, where p is the mixture probability. In this work, we mostly consider the equiprobable domains case $p = 0.5$.

Let $\phi : \mathcal{X} \rightarrow \mathcal{Z} = \mathbb{R}^m$ be a feature map parametrized by a neural network, where m is the embedding dimension. ϕ defines a Reproducing Kernel Hilbert Space (RKHS) with kernel k such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{Z}}$, where $\langle \cdot, \cdot \rangle_{\mathcal{Z}}$ denotes the inner product in \mathcal{Z} .

2.1. Contrastive Learning

We recall the definitions of the Normalized Temperature-scaled Cross Entropy (NT-Xent) loss and the Supervised Contrastive loss.

Definition 2.1 (NT-Xent loss). Consider a batch of instances \mathcal{B} and their feature representation z , which are assumed of unitary norm, i.e., $\|z\| = 1$. The NT-Xent loss defined as:

$$\mathcal{L}_{NT-Xent} = -\frac{1}{|\mathcal{B}|} \sum_{i=0}^{|\mathcal{B}|-1} \log \frac{e^{z_i \cdot z_{j(i)}/\tau}}{\sum_{l \in \mathcal{A}(i)} e^{z_i \cdot z_l/\tau}}, \quad (1)$$

where $z_i = \phi(x_i)$ is the feature representation of instance x_i , $z_{j(i)}$ is the positive counterpart of feature z_i , and $\mathcal{A}(i) = \{0, \dots, |\mathcal{B}|-1\} \setminus \{i\}$ is the set of the indices of all features with the exception of z_i , and τ is a temperature parameter.

Definition 2.2 (Supervised Contrastive loss). Given a batch of instances \mathcal{B} , the Supervised Contrastive loss is defined as:

$$\mathcal{L}_{SupContr} = -\frac{1}{|\mathcal{B}|} \sum_{i \in |\mathcal{B}|} \frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \frac{e^{z_i \cdot z_j/\tau}}{\sum_{l \in \mathcal{A}(i)} e^{z_i \cdot z_l/\tau}}, \quad (2)$$

where $z_i = \phi(x_i)$ is the feature representation of instance x_i , and $\mathcal{P}(i) = \{j \in \mathcal{A}(i) : y_j = y_i\}$ is the set of indices of the positive counterparts of feature z_i .

2.2. Contrastive Learning and Domain Adaptation

In the following, we revisit the definition of the CMMD and establish its connection to contrastive losses.

Definition 2.3 (CMMD). Given two labeled domains \mathcal{D}_0 and \mathcal{D}_1 , and the mapping $\phi : \mathcal{X} \rightarrow \mathcal{Z}$. The CMMD is

defined as:

$$\begin{aligned} \text{CMMD}^2(\mathcal{D}_0, \mathcal{D}_1, \phi) \\ = \mathbb{E}_{C \sim \pi^{\mathcal{Y}}} \left[\left\| \mathbb{E}_{X \sim \pi_{0,C}^{\mathcal{X}|\mathcal{Y}}} [\phi(X)] - \mathbb{E}_{X \sim \pi_{1,C}^{\mathcal{X}|\mathcal{Y}}} [\phi(X)] \right\|_{\mathcal{Z}}^2 \right]. \end{aligned} \quad (3)$$

The CMMD calculates the difference between the expected embedding of instances in the two domains, for each class. If ϕ is adjusted so as to minimize the CMDD, then the embeddings $\phi(X)$ are similar regardless of the domain, and the conditional distributions of the embeddings of the two domains will be matched. It can thus be seen as a measure of Domain Adaptation. Definition 2.3 corresponds to the definition of the Weighted Class-wise MMD (WCMMD) and encompasses other definitions of the CMMD found in the literature (Wang et al., 2021) as a particular case when all the classes have the same prior probability. We propose the following lemma that relates Contrastive Learning to the minimization of the CMMD (proof in Appendix C).

Lemma 2.4. *In a high temperature regime, both the Supervised Contrastive loss and the NT-Xent loss can be expressed in terms of the CMMD by the following equation:*

$$\begin{aligned} \tau \mathcal{L}_{Contr} \\ \approx \frac{1}{4} \text{CMMD}^2(\mathcal{D}_0, \mathcal{D}_1, \phi) + \underbrace{\mathbb{E}_{X, X' \sim \pi_{0.5}^{\mathcal{X}}} [k(X, X')]}_A \\ - \frac{1}{2} \underbrace{\mathbb{E}_{C \sim \pi^{\mathcal{Y}}} \left[\mathbb{E}_{X, X' \sim \pi_{0,C}^{\mathcal{X}|\mathcal{Y}}} [k(X, X')] + \mathbb{E}_{X, X' \sim \pi_{1,C}^{\mathcal{X}|\mathcal{Y}}} [k(X, X')] \right]}_B \\ + \frac{1}{2\tau} \underbrace{\mathbb{E}_{X \sim \pi_{0.5}^{\mathcal{X}}} \left[\text{Var}_{X' \sim \pi_{0.5}^{\mathcal{X}}} [k(X, X')] \right]}_C \\ + \mathcal{O} \left(\frac{\mathbb{E}_{X \sim \pi_{0.5}^{\mathcal{X}}} \left[\text{Var}_{X' \sim \pi_{0.5}^{\mathcal{X}}} [k(X, X')]^2 \right]}{\tau^4} \right) \\ + \log(|\mathcal{B}|-1). \end{aligned} \quad (4)$$

Lemma 2.4 suggests that decreasing \mathcal{L}_{contr} decreases the CMMD, which improves Domain Adaptation. Equation (4) also includes other terms, which can be analyzed as follows: A represents the similarity between all pairs of features, while B denotes the similarity between pairs of features from the same class and domain. C is a variance term. The constant term $\log(|\mathcal{B}|-1)$, with $|\mathcal{B}|$ and τ the batch size and the temperature of the Contrastive losses, is irrelevant for the optimization. The last term of Equation (4) refers to the approximation error of the Taylor series used to obtain the equation. The term $A - B/2$ can be interpreted the difference of the similarity between all the features, and the similarity between features of the same class and domain. In a nutshell, the contrastive loss compute the contrast with

respect to all pairs, and the CMMD the contrast with respect to pairs with the same class and domain. This difference is adjusted in Equation (4).

2.3. Contrastive Learning and class-separability

Extending on the work of Li et al. (2021), we can relate the contrastive losses (Supervised and NT-Xent) to an Inter-class MMD (IMMD) through the following lemma.

Lemma 2.5. *By assuming that the kernel k is bounded, i.e., $|k(x, x')| < k^{max}$, $\forall x, x'$, and that the inner product on \mathcal{Y} satisfies $\langle y, y' \rangle_{\mathcal{Y}} = \Delta l \mathbf{1}_{\{y=y'\}} + l_0$, then the Contrastive losses bound the IMMD:*

$$-\frac{1}{\alpha} \text{IMMD}^2 + \gamma \text{HSIC}(X, X) \quad (5)$$

$$+ \mathcal{O}(\text{Var}[k(X, X')]) \leq \mathcal{L}_{Contr},$$

with

$$\text{IMMD}^2 = \mathbb{E}_{C_1, C_2 \sim \pi_{0.5}^{\mathcal{Y}}} \left[\left\| \mathbb{E}_{X \sim \pi_{0.5, C_1}^{X|\mathcal{Y}}} [\phi(X)] \right. \right. \quad (6)$$

$$\left. \left. - \mathbb{E}_{X \sim \pi_{0.5, C_2}^{X|\mathcal{Y}}} [\phi(X)] \right\|_{\mathcal{Z}}^2 \right],$$

where $\text{HSIC}(X, X)$ is the Hilbert-Schmidt Independence Criterion (Gretton et al., 2005), α is a proportionality constant which depends on problem parameters, and $\gamma \in \mathbb{R}$ is a constant satisfying $\max\{2, 2k^{max}\} = (1 + \sqrt{1 - 4\gamma}) / (2\gamma)$. For the Supervised Contrastive loss, $\Delta l = K$ (the number of classes). For the NT-Xent loss, $\Delta l = N$ (the number of instances).

The IMMD computes the difference between embeddings in the mixture domain with different class, and is thus a measure of class-separability. The HSIC is equal to the covariance in the feature space \mathcal{Z} , and it thus measures a non-linear covariance between the instances given the map ϕ . We remark that the condition on $\langle y, y' \rangle_{\mathcal{Y}}$ is satisfied when considering one-hot vectors and the Euclidean inner product on the label space. Li et al. (2021) proved Lemma 5 for the NT-Xent loss. In Appendix D we extend the proof to the Supervised Contrastive loss.

To measure class-separability in the feature space we define another MMD-based quantity, the Different-class MMD (DCMMD), which is more general than the inter-class MMD of Equation (5). The DCMMD measures the difference between the features of two different classes, in the same and different domains.

Definition 2.6 (DCMMD). Given two labeled domains \mathcal{D}_0 ,

\mathcal{D}_1 , and a mixed domain \mathcal{D}_p , the DCMMD is defined as:

$$\text{DCMMD}^2(\mathcal{D}_0, \mathcal{D}_1, \phi)$$

$$= \mathbb{E}_{C_1, C_2 \sim \pi_{p_0}^{\mathcal{Y}}; C_1 \neq C_2; D_1, D_2 \sim \text{Ber}(p)} \quad (7)$$

$$\left[\left\| \mathbb{E}_{X \sim \pi_{D_2, C_1}^{X|\mathcal{Y}}} [\phi(X)] - \mathbb{E}_{X \sim \pi_{D_1, C_2}^{X|\mathcal{Y}}} [\phi(X)] \right\|_{\mathcal{Z}}^2 \right],$$

where $\text{Ber}(p)$ is the Bernoulli distribution. To remain consistent with the CMMD definition, we usually consider equiprobable domains and set $p = 1/2$.

3. Numerical experiments

This section describes the datasets, models, and training settings used for the numerical experiments.

3.1. Datasets

Three types of mammography datasets are considered in this work: a clinical mammography image dataset (GEHC image dataset), a clinical mammography patch dataset (GEHC patch dataset), and a synthetic mammography patch dataset (synthetic patch dataset). The two clinical datasets contain GEHC images from anonymized patients, collected from a single institution in France following the EU General Data Protection Regulation. Additionally, two publicly-available datasets, CBIS-DDSM (Lee et al., 2017) and InBreast (Mora et al., 2012) are also used in this work.

GEHC image dataset. It contains 1300 cases, of which 197 are biopsy-proven cancers and 313 contain benign biopsied lesions. The remaining 790 are normal cases, which are studies in which no suspicious lesion was found in the breasts, and are confirmed by a one-year follow-up exam. The dataset is split in training (936 cases), validation (167 cases), and test (197) subsets in a stratified fashion, which takes into account the case pathology (benign or malignant), the lesions contained in the image (mass and calcification), and the description or sub-type of the lesions (e.g., spiculated mass, oval mass, granular calcification, etc.).

GEHC patch dataset. Ten normal patches, and at least ten lesion 512×512 pixel patches are extracted from each image that contains a lesion (mass or calcification), with two different strategies: “fixed” and “random” extractions. For every lesion, a “fixed” patch centered in the lesion centered is extracted. If the lesion is too large to be entirely contained in the patch, the space covered by the lesion is divided into a grid of $N \times M$ non-overlapping patches, which are incorporated to the patch dataset. This assures that every part of the lesion is represented in the dataset but may introduce an undesirable bias, as most patches coming from

large lesions contain the lesion fragment in the corners. To reduce this bias, the patch dataset is enriched with “random” lesion patches, centered at random positions of the lesion. The extracted patches have an Intersection over Union (IoU) smaller than 0.5 between each other, to avoid generating patches that are too similar.

Synthetic patch dataset. A synthetic patch dataset is created to enable controllable and efficient experiments while maintaining resemblance to real images. Mammography patches are generated by first sampling a Gaussian texture, and then inserting simulated mammography lesions. First, a white Gaussian random field $w \in \mathbb{R}^{N \times M}$ is sampled. Then, a low-pass filter with the following transfer function is applied to w :

$$H(u, v) = \frac{1}{\sqrt{u^2 + v^2}^\beta}, \tag{8}$$

where u and v are the coordinates of the image in the frequency space and β is a non-negative real slope parameter, which can be associated to the breast density (Mainprize et al., 2012). The application of H adds some spatial correlation to the pixels and creates the base texture. Two types of simple breast lesions are generated and inserted in the texture: masses and calcifications. Masses are simulated by randomly-centered Gaussian intensity profiles. Calcifications are modeled as high intensity pixels, clustered in random regions of the texture.

The synthetic patch dataset contains three types of patches: normal (only simulated breast texture), mass (breast texture containing an added synthetic mass), and calcification (breast texture with some synthetic calcifications), and it thus defines a three-class classification problem. A dataset of 1k 256×256 pixel synthetic patches, balanced in terms of classes, is generated. The parameters for generating this dataset are detailed in Appendix B.

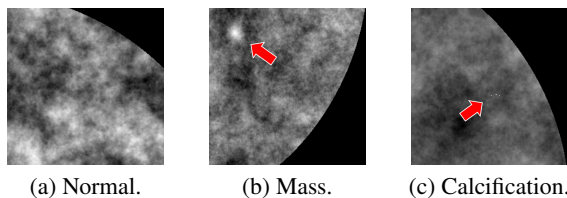


Figure 1. Examples of synthetic patches (arrows signaling the lesions were inserted).

3.2. Image style heterogeneity

In this work, we focus on studying the effect of training a DL model with data with different image styles. In particular, we use the sigmoid Look-Up Table (LUT) function, a contrast enhancement technique commonly used in

mammography (Hernández-Vázquez et al., 2024; on the Evaluation of Cancer-Preventive Interventions, 2016), as proxy transformation to define the two data distributions or domains. However, the methodology developed in this work is applicable to any other image style or contrast transformation. Figure 2 shows an example of the LUT application on a full mammography image.

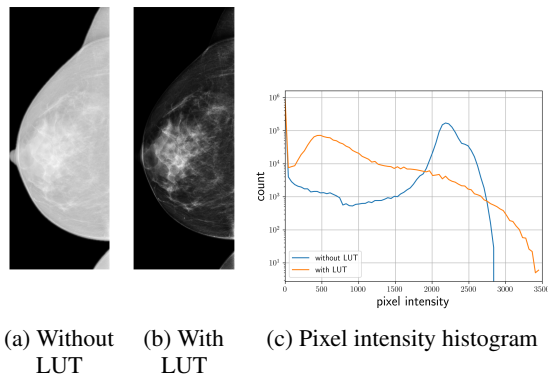


Figure 2. Illustration of a single FFDM image: (a) without LUT application, (b) with LUT application, and (c) the pixel intensity histogram in logarithmic scale for both post-processings.

Two distinct types of datasets are created from the base datasets for training and validation purposes, by splitting mammograms and patches at the case level:

- An *augmented dataset*, which includes two versions of each image: one with the LUT applied and the other without. This dataset can be seen as equivalent to applying a LUT-based data augmentation.
- *Mixed datasets*, where the original dataset was divided into two groups. One group had images with the LUT applied, while the other did not.

For the clinical images and patches, four mixed datasets are constructed. For the synthetic patches, a single mixed dataset is constructed, due to the simplicity of the problem and the less variability observed in the results. Models are trained using each of these datasets, and evaluated in the same test set, which contained the two versions of each image (with and without LUT).

3.3. Model architecture and training methodology

A patch-based model is used for classifying mammography images, which consists in first training a path-classifier and then extending it to a whole image classifier by appending additional residual blocks and re-training on complete images (see Figure 3). In this work, DenseNet-121 (Huang

et al., 2017) is used as backbone. For the clinical patch-classifier and whole image classifier, a Multi-layer Perceptron (MLP) projector is added for training with the Supervised Contrastive loss. The use of a projector is standard in CL, and enables to avoid the training task’s overfitting bias (Balestriero et al., 2023), caused by the fact that the optimal features for the training (in this case, minimizing the Contrastive loss) task may not be optimal for the downstream task (in this case, classification). In the case of the patch-classifier, the projector features two hidden layers with 2048 units each, and an output layer of 1024 units (Figure 3 - top). For the whole image classifier, it consists of one hidden layer of 2048 units, and an output layer of 1024 units (Figure 3 - bottom). The projector is key to avoiding perfect invariance in the features used for classification, which lower the classification performance for clinical images, and it is used solely for the model training phase with Supervised Contrastive Learning (SCL), but not in the inference or evaluation stages. For synthetic patch classification the Supervised Contrastive loss is applied directly to the extracted features, as perfect invariance did not affect classification performance. This is likely due to the simplicity of the problem, which makes the projector unnecessary. To initialize the clinical patch-classifier, two methods are explored: one using the ImageNet dataset and the other using the CBIS-DDSM dataset. As CBIS-DDSM is a 2D mammography image dataset, it is more similar to the GEHC dataset than ImageNet, and is thus expected to provide a better initialization. For obtaining the whole image classifier, only the patch-classifier initialized on CBIS-DDSM is used.

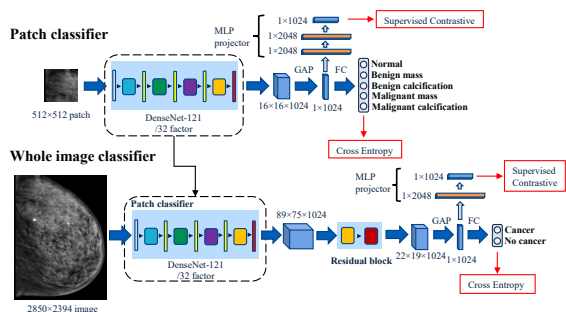


Figure 3. Patch-classifier and whole image classifier architectures for training with the Cross Entropy loss and with the Supervised Contrastive loss. GAP: Global Average Pooling, FC: Fully Connected layer.

A model trained solely with the Cross Entropy (CE) loss, and thus without Domain Adaptation, was compared to a model trained using the Supervised Contrastive loss. As suggested by Khosla et al. (2020), the model is first trained with the Supervised Contrastive loss to extract domain-invariant features. Then, the feature extraction layers are frozen, and only the final linear classification layer is trained with CE.

This training strategy, also known as Linear Classification Protocol (LCP) is standard in Contrastive Learning (He et al., 2020; Wu et al., 2021). The resulting model is denoted as SupContr+LCP. As a third training strategy, the SupContr+LCP model is fully re-trained using the Cross Entropy loss (without freezing the feature extraction layers), resulting in a model denoted as SupContr+CE. These three training strategies (CE, SupContr+LCP, and SupContr+CE) are compared for patch, both synthetic and clinical, and whole image classification. The experimental settings and hyperparameters are detailed in Appendix A.

In addition, the generalization capability of the whole image classifier is assessed on InBreast, a publicly available dataset of mammography images. For this, the InBreast dataset is split into training (288 cases), validation (46 cases) and testing (75 cases) sets, with the same stratification strategy used for the GEHC dataset. To keep the learned representation fixed, the feature extractor is frozen during InBreast fine-tuning, and only the output linear layer is updated.

For synthetic patches, the CE, SupContr+LCP, and SupContr+CE models are trained on the mixed synthetic dataset. For clinical patches and full images, the models are trained on the four mixed datasets, and on the augmented dataset. The results of the four trainings on the mixed datasets are aggregated to calculate a mean performance and 95% Confidence Intervals (CI), as well as p-values, of a one-sided Welch’s t-test. For the synthetic patch-classifier and the trainings on the clinical augmented datasets, Bootstrapping was used for obtaining the 95% CI and p-values.

4. Results

The numerical results are organized into three sections: first, an illustration of Lemma 2.4, followed by a quantitative analysis, and finally, a qualitative analysis.

4.1. Illustration of Lemma 2.4

Figure 4 shows the evolution of the individual terms in Equation (4) during training with synthetic patches. We observe that the CMMD and the similarity between all pairs, given by term A, decrease while minimizing the Supervised Contrastive loss, following Equation (4). The similarity between pairs of features with the same class and domain increases, as predicted by Equation (4).

To quantify the trends observed, we analyze the correlation between the derivatives of each of the terms, and the derivative of the Supervised Contrastive loss for different temperature values τ . This enables to numerically assess if the quantities are moving in the same, or opposite direction

Bridging Contrastive Learning and Domain Adaptation: Theoretical Perspective and Practical Application

Model	CMMD ↓	DCMMD ↑	AUC ↑	AUC (OvO) ↑	AUC (OvR) ↑	Accuracy ↑
Patch-classifier (synthetic - 3 classes)						
CE	0.348 ± 0.010	0.405 ± 0.005	-	0.998 ± 0.002*	0.995 ± 0.005*	0.981 ± 0.018*
SupContr+LCP	0.239 ± 0.032	0.417 ± 0.010	-	1.000 ± 0.000	0.999 ± 0.001	0.985 ± 0.015*
SupContr+CE	0.226 ± 0.029	0.438 ± 0.007	-	0.998 ± 0.003*	0.996 ± 0.006*	0.969 ± 0.02
Patch-classifier (clinical - TL from ImageNet - 5 classes)						
CE	0.120 ± 0.041	0.094 ± 0.041	-	0.728 ± 0.017	0.684 ± 0.020	0.391 ± 0.037
SupContr+LCP	0.092 ± 0.016*	0.185 ± 0.020*	-	0.847 ± 0.011*	0.793 ± 0.019*	0.497 ± 0.044*
SupContr+CE	0.092 ± 0.016*	0.183 ± 0.021*	-	0.846 ± 0.012*	0.792 ± 0.020*	0.508 ± 0.019*
Patch-classifier (clinical - TL from CBIS-DDSM - 5 classes)						
CE	0.092 ± 0.020*	0.205 ± 0.042	-	0.915 ± 0.016*	0.880 ± 0.014*	0.627 ± 0.035*
SupContr+LCP	0.081 ± 0.019*	0.267 ± 0.011	-	0.878 ± 0.014	0.845 ± 0.012	0.569 ± 0.029
SupContr+CE	0.069 ± 0.009	0.278 ± 0.011	-	0.918 ± 0.010*	0.880 ± 0.008*	0.628 ± 0.005*
Whole image classifier (2 classes)						
CE	0.118 ± 0.029	0.087 ± 0.042	0.718 ± 0.043	-	-	0.609 ± 0.084
SupContr+LCP	0.050 ± 0.007*	0.174 ± 0.041*	0.759 ± 0.033*	-	-	0.674 ± 0.023*
SupContr+CE	0.049 ± 0.011*	0.151 ± 0.061*	0.776 ± 0.009*	-	-	0.698 ± 0.046*

Table 1. Results on the mixed datasets.

Model	CMMD ↓	DCMMD ↑	AUC ↑	AUC (OvO) ↑	AUC (OvR) ↑	Accuracy ↑
Patch-classifier (clinical - TL from ImageNet - 5 classes)						
CE	0.152 ± 0.004	0.163 ± 0.018*	-	0.871 ± 0.006	0.817 ± 0.010	0.547 ± 0.022*
SupContr+LCP	0.027 ± 0.005	0.229 ± 0.005	-	0.878 ± 0.005	0.846 ± 0.010	0.538 ± 0.020
SupContr+CE	0.037 ± 0.006	0.171* ± 0.004	-	0.887 ± 0.005	0.805 ± 0.011	0.542 ± 0.021*
Patch-classifier (clinical - TL from CBIS-DDSM - 5 classes)						
CE	0.091 ± 0.004	0.226 ± 0.002	-	0.927 ± 0.004	0.896 ± 0.007	0.656 ± 0.021
SupContr+LCP	0.034 ± 0.004*	0.243 ± 0.003	-	0.880 ± 0.005	0.842 ± 0.009	0.551 ± 0.019
SupContr+CE	0.032 ± 0.006*	0.283 ± 0.003	-	0.919 ± 0.004	0.881 ± 0.008	0.599 ± 0.020
Whole image classifier (GEHC dataset - 2 classes)						
CE	0.108 ± 0.009	0.099 ± 0.018	0.745 ± 0.050	-	-	0.625 ± 0.061
SupContr+LCP	0.040 ± 0.016*	0.127 ± 0.022	0.763 ± 0.058	-	-	0.671 ± 0.043*
SupContr+CE	0.066 ± 0.029*	0.213 ± 0.030	0.816 ± 0.042	-	-	0.728 ± 0.073*

Table 2. Results on the augmented datasets.

to the Supervised Contrastive loss. Figure 5 shows the evolution of the Pearson correlation coefficient of the derivatives ρ with the temperature τ for different temperature values, spanning from $\tau = 0.01$ to $\tau = 5.0$. While the Pearson correlation coefficient for the CMMD and term A is positive, which confirms that the two quantities move in the same direction, it is negative for terms B and C. We observe that the magnitude of the Pearson correlation coefficient increases with increasing temperature, and reaches a *plateau* between $\tau = 0.2$ and $\tau = 0.5$. This is explained by the fact that the Taylor approximation used for proving Lemma 2.4 is valid under relatively large temperatures. However, it is important to note that a certain level of correlation is observed, even at lower temperatures.

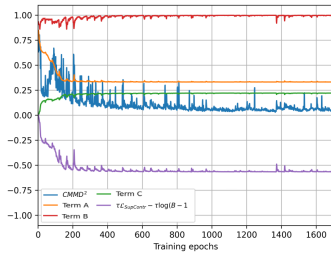


Figure 4. Evolution of the terms of Equation (4) during training.

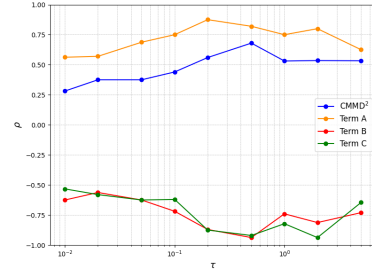


Figure 5. Evolution of the Pearson coefficients with the temperature.

4.2. Quantitative analysis

Table 1 shows the results for the mixed datasets. Domain Adaptation is measured in terms of the CMMD and class-separability is measured by the DCMMD. Classification performance is evaluated with the accuracy and AUC for the binary whole image classifier, and with the accuracy, One vs. One (OvO) AUC, and One vs. Rest (OvR) AUC for the patch classifiers. We observe that in all the classification problems, the models trained with the Supervised Contrastive loss (SupContr+LCP, SupContr+CE) achieve higher Domain Adaptation and class-separability than the CE models. This translates into a higher downstream clas-

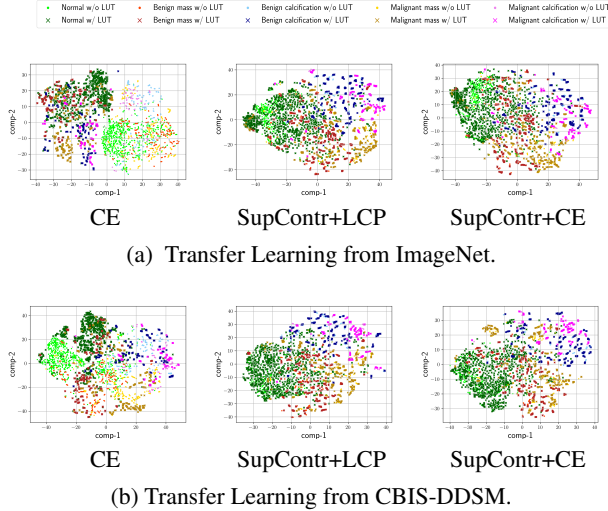


Figure 6. t-SNE plots of the features from the patch-classifier, indicating class and domain.

sification performance for the whole image classifier and patch-classifier with synthetic patches, and clinical patches with TL from ImageNet. On the contrary, when weights are initialized from CBIS-DDSM (scanned mammography films), SupContr+CE matches CE performance but fails to outperform it. In this case, the pre-training dataset is closer to the GEHC dataset than ImageNet, leading to improved domain adaptation, class separability, and classification performance. This reduces the negative impacts of fine-tuning with images from different domains.

Table 2 presents the results for the augmented datasets under consideration, demonstrating that the conclusions drawn from the mixed datasets remain valid. In addition, by comparing the two tables for the whole image classifier, we can see that the contrastive-based models trained on the mixed datasets outperform the CE model trained on the augmented dataset, despite the latter having been trained on twice the amount of data.

Finally, Table 3 shows the results on the publicly available InBreast dataset. The SupContr+CE model exhibits superior generalization by achieving a 13% AUC increase with respect to the CE model. We argue that this is a measure of generalization capabilities of the representations, as images from InBreast are only used for fine-tuning the linear classification layer while freezing the feature extraction.

Model	AUC	Accuracy
CE	$0.733 \pm 0.096^*$	0.571 ± 0.067
SupContr+LCP	$0.746 \pm 0.083^*$	0.647 ± 0.061
SupContr+CE	0.831 ± 0.071	0.703 ± 0.060

Table 3. Results on InBreast.

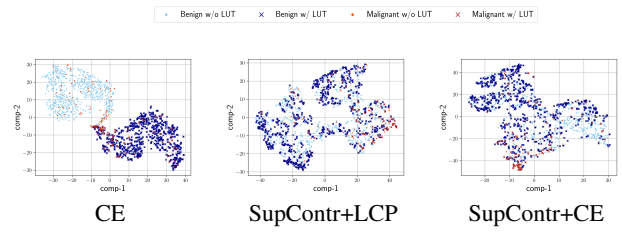


Figure 7. t-SNE plots of the features from the whole image classifier, indicating class and domain.

4.3. Qualitative analysis

We now perform a qualitative evaluation of the extracted feature space. Figure 6 shows the t-SNE plot of the extracted features for clinical patch-classifier, trained with the three losses (CE, SupContr+LCP, SupContr+CE), with weights initialized from ImageNet (Figure 6a) and CBIS-DDSM (Figure 6b). When Transfer Learning from ImageNet is used, the CE model features are more separated by domain than by class, while the SupContr+LCP and SupContr+CE models are domain-invariant (Figure 6a). When Transfer Learning from CBIS-DDSM is used (Figure 6b) it can be seen that the CE model has already some degree of domain invariance, especially for non-normal classes. We hypothesize that, in this case, the similarity of the CBIS-DDSM dataset to the GEHC images is leveraged by the DL-model, virtually increasing the training dataset and increasing the robustness of the learned features. As will be seen later, this decreases the impact of Domain Adaptation on classification performance. The SupContr+LCP and SupContr+CE models attain domain invariance for all the classes, including Normal patches.

Figure 7 shows the features t-SNE plot for the whole image classifier. It can be seen that the features of the CE model can be easily separated by domain, despite the features of the CE patch-classifier being domain-invariant for most classes (we recall that the whole image classifier was obtained by extending the patch-classifier, pre-trained on CBIS-DDSM). We hypothesize that this is caused by the maladaptation of the normal patches for the CE model in Figure 6b, as every mammography image contains many normal regions. On the other hand, the features of the SupContr+LCP and SupContr+CE models are domain-invariant.

5. Conclusions

In this work, we mathematically showed that minimizing two standard contrastive losses - NT-Xent loss and Supervised Contrastive loss - decreases the CMMD and thus performs Domain Adaptation. Moreover, it improves class-separability in the feature space, which is often associated to higher downstream task performance. These findings

offer a theoretical foundation for the growing adoption of Contrastive Learning as an effective approach for Domain Adaptation. Our theoretical results were further validated through numerical experiments, which demonstrated that minimizing the Supervised Contrastive Loss consistently improved Domain Adaptation and class separability, leading to enhanced classification performance in most cases. Considering these theoretical and empirical results, we conclude that Contrastive Learning can be effectively used for attaining Domain Adaptation while maintaining or improving class-separability in the feature space. Future research should explore the boundaries of these improvements in classification performance, particularly regarding the impact of weight initialization and the role of Transfer Learning.

References

- Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., Schwarzschild, A., Wilson, A. G., Geiping, J., Garrido, Q., Fernandez, P., Bar, A., Pirsiavash, H., LeCun, Y., and Goldblum, M. A cookbook of self-supervised learning, 2023.
- Bobowicz, M., Rygusik, M., Buler, J., Buler, R., Ferlin, M., Kwasigroch, A., Szurowska, E., and Grochowski, M. Attention-based deep learning system for classification of breast lesions—multimodal, weakly supervised approach. *Cancers*, 15(10):2704, 2023.
- Cao, Z., Yang, Z., Tang, Y., Zhang, Y., Han, M., Xiao, J., Ma, J., and Chang, P. Supervised contrastive pre-training for mammographic triage screening models. In de Bruijne, M., Cattin, P. C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., and Essert, C. (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 129–139, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87234-2.
- Chaitanya, K., Erdil, E., Karani, N., and Konukoglu, E. Contrastive learning of global and local features for medical image segmentation with limited annotations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12546–12558. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/949686ecef4ee20a62d16b4a2d7ccca3-Paper.pdf.
- Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.
- Chopra, S., Hadsell, R., and LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 539–546 vol. 1, 2005. doi: 10.1109/CVPR.2005.202.
- Damodaran, B. B., Kellenberger, B., Flamary, R., Tuia, D., and Courty, N. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 447–463, 2018.
- Darban, Z. Z., Yang, Y., Webb, G. I., Aggarwal, C. C., Wen, Q., and Salehi, M. Dacad: Domain adaptation contrastive learning for anomaly detection in multivariate time series. *arXiv preprint arXiv:2404.11269*, 2024.
- de Mathelin, A., Deheeger, F., Richard, G., Mougeot, M., and Vayatis, N. Adapt : Awesome domain adaptation python toolbox, 2021.
- de Mathelin, A., Deheeger, F., Mougeot, M., and Vayatis, N. *From Theoretical to Practical Transfer Learning: The ADAPT Library*, pp. 283–306. Springer International Publishing, Cham, 2023. ISBN 978-3-031-11748-0. doi: 10.1007/978-3-031-11748-0_12. URL https://doi.org/10.1007/978-3-031-11748-0_12.
- Dong, N. and Voiculescu, I. Federated contrastive learning for decentralized unlabeled medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 378–387. Springer, 2021.
- Farahani, A., Voghoei, S., Rasheed, K., and Arabnia, H. R. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.
- Gao, T., Yao, X., and Chen, D. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP 2021-2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2021.
- Garrucho, L., Kushibar, K., Jouide, S., Diaz, O., Igual, L., and Lekadir, K. Domain generalization in deep learning based mass detection in mammography: A large-scale multi-center study. *Artificial Intelligence in Medicine*, 132:102386, 2022. ISSN 0933-3657. doi: <https://doi.org/10.1016/j.artmed.2022.102386>. URL <https://www.sciencedirect.com/science/article/pii/S0933365722001415>.

- Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In Jain, S., Simon, H. U., and Tomita, E. (eds.), *Algorithmic Learning Theory*, pp. 63–77, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31696-1.
- Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel method for the two-sample-problem. In Schölkopf, B., Platt, J., and Hoffman, T. (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/e9fb2eda3d9c55a0d89c98d6c54b5b3e-Paper.pdf.
- Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742, 2006. doi: 10.1109/CVPR.2006.100.
- He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.
- Hernández-Vázquez, M. A., Hernández-Rodríguez, Y. M., Cortes-Rojas, F. D., Bayareh-Mancilla, R., and Cigarroa-Mayorga, O. E. Hybrid feature mammogram analysis: detecting and localizing microcalcifications combining gabor, prewitt, glcm features, and top hat filtering enhanced with cnn architecture. *Diagnostics*, 14(15):1691, 2024.
- Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.
- Huang, W., Wu, A., Yang, Y., Luo, X., Yang, Y., Hu, L., Dai, Q., Dai, X., Chen, D., Luo, C., et al. Llm2clip: Powerful language model unlock richer visual representation. *arXiv preprint arXiv:2411.04997*, 2024.
- Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.
- Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.
- Kouw, W. M. and Loog, M. A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):766–785, mar 2021. doi: 10.1109/tpami.2019.2945942. URL <https://doi.org/10.1109%2Ftpami.2019.2945942>.
- Kull, M. and Flach, P. Patterns of dataset shift. In *First international workshop on learning over multiple contexts (LMCE) at ECML-PKDD*, volume 5, 2014.
- Kyono, T., Gilbert, F. J., and van der Schaar, M. Improving workflow efficiency for mammography using machine learning. *Journal of the American College of Radiology*, 17(1):56–63, 2020.
- Lee, C.-Y., Batra, T., Baig, M. H., and Ulbricht, D. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10285–10295, 2019.
- Lee, R., Gimenez, F., Hoogi, A., Miyake, K., Gorovoy, M., and Rubin, D. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, 4:170177, 12 2017. doi: 10.1038/sdata.2017.177.
- Li, S., Xia, X., Ge, S., and Liu, T. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 316–325, 2022.
- Li, Y., Pogodin, R., Sutherland, D. J., and Gretton, A. Self-supervised learning with kernel dependence maximization. *Advances in Neural Information Processing Systems*, 34:15543–15556, 2021.
- Long, M., Wang, J., Ding, G., Sun, J., and Yu, P. S. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE international conference on computer vision*, pp. 2200–2207, 2013.
- Long, M., Zhu, H., Wang, J., and Jordan, M. I. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pp. 2208–2217. PMLR, 2017.
- Long, M., Cao, Z., Wang, J., and Jordan, M. I. Conditional adversarial domain adaptation. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/ab88b15733f543179858600245108dd8-Paper.pdf.

- Loshchilov, I. and Hutter, F. SGDR: stochastic gradient descent with restarts. *CoRR*, abs/1608.03983, 2016. URL <http://arxiv.org/abs/1608.03983>.
- Mainprize, J. G., Tyson, A. H., and Yaffe, M. J. The relationship between anatomic noise and volumetric breast density for digital mammography. *Medical Physics*, 39(8):4660–4668, 2012.
- Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., and Cardoso, J. S. Inbreast: Toward a full-field digital mammographic database. *Academic Radiology*, 19(2):236–248, 2012. ISSN 1076-6332. doi: <https://doi.org/10.1016/j.acra.2011.09.014>. URL <https://www.sciencedirect.com/science/article/pii/S107663321100451X>.
- on the Evaluation of Cancer-Preventive Interventions, I. W. G. Breast cancer screening. 2(Screening Techniques): 1691, 2016.
- Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- Petrini, D. G., Shimizu, C., Roela, R. A., Valente, G. V., Folgueira, M. A. A. K., and Kim, H. Y. Breast cancer diagnosis in two-view mammography using end-to-end trained efficientnet-based convolutional network. *Ieee access*, 10:77723–77731, 2022.
- Quintana, G. I., Li, Z., Vancamberg, L., Mougeot, M., Desolneux, A., and Muller, S. Exploiting patch sizes and resolutions for multi-scale deep learning in mammogram image classification. *Bioengineering*, 10(5), 2023. ISSN 2306-5354. doi: 10.3390/bioengineering10050534. URL <https://www.mdpi.com/2306-5354/10/5/534>.
- Quintana, G. I., Jugnon, V., Vancamberg, L., Desolneux, A., and Mougeot, M. Contrastive learning: an efficient domain adaptation strategy for 2d mammography image classification. In *2024 IEEE 21st International Symposium on Biomedical Imaging (ISBI)*, 2024.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Schall, K., Barthel, K. U., Hezel, N., and Jung, K. Optimizing clip models for image retrieval with maintained joint-embedding alignment. In *International Conference on Similarity Search and Applications*, pp. 97–110. Springer, 2024.
- Shen, J., Qu, Y., Zhang, W., and Yu, Y. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., and Sieh, W. Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports*, 9(1), aug 2019. doi: 10.1038/s41598-019-48995-4. URL <https://doi.org/10.1038%2Fs41598-019-48995-4>.
- Singh, A. Clda: Contrastive learning for semi-supervised domain adaptation. *Advances in Neural Information Processing Systems*, 34:5089–5101, 2021.
- Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf.
- Thota, M. and Leontidis, G. Contrastive domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2209–2218, June 2021.
- Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.
- Wang, W., Li, B., Yang, S., Sun, J., Ding, Z., Chen, J., Dong, X., Wang, Z., and Li, H. A unified joint maximum mean discrepancy for domain adaptation, 2021.
- Weinberger, K. and Saul, L. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 02 2009.
- Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzębski, S., Févray, T., Katsnelson, J., Kim, E., et al. Deep neural networks improve radiologists’ performance in breast cancer screening. *IEEE transactions on medical imaging*, 39(4):1184–1194, 2019.

Wu, Y., Zeng, D., Wang, Z., Shi, Y., and Hu, J. Federated contrastive learning for volumetric medical image segmentation. In de Bruijne, M., Cattin, P. C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., and Essert, C. (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 367–377, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87199-4.

A. Numerical experiments details

For each dataset, all the models were trained during the same number of epochs, which was set by making sure that all the models had converged. The patch-classifiers with clinical patches were trained for 250 epochs, the patch-classifiers with synthetic patches for 100 epochs, and the whole image classifiers for 200 epochs. The models with the best performance on the validation set (AUC for binary classifiers and AUC OvO for multi-class classifiers) were retained after each training round.

All models were optimized using Stochastic Gradient Descent (SGD) without momentum, with the learning rate scheduled by a cosine annealing scheduler (Loshchilov & Hutter, 2016) with period $T = 4$ epochs. The base learning rate was set to 10^{-4} for clinical images and patches, and to 10^{-3} for synthetic patches. To achieve balanced batches, less-represented classes, such as malignant ones, were oversampled. The batch size was set to 8 for the whole image classifier (the maximum value that fit into the GPU RAM) and 30 for the patch classifiers. All models used a weight decay of 10^{-4} and did not employ dropout. Data augmentation for patch classifiers included vertical and horizontal flips, as well as rotations by 90° , 180° , and 270° . For the whole image classifier, only horizontal flips were used. All images were used at their original resolutions, as resizing complicates the detection of small lesions (Quintana et al., 2023). For training the whole image classifier, the first three dense blocks of the DenseNet backbone were frozen to reduce GPU RAM usage, which did not negatively impact performance.

For training with the Supervised Contrastive loss, the temperature was set to $\tau = 0.5$ for synthetic patches. For clinical patches and full images, it was maintained at $\tau = 0.5$ for the first 50 epochs, then linearly decreased over the next 100 epochs to $\tau = 0.1$, where it remained constant for the final epochs (100 epochs for the patch-classifier and 50 for the whole image classifier). When fine-tuning linear layers, i.e., LCP, 20 epochs were used.

Each training was conducted on a 24 GB Nvidia Quadro RTX 6000 GPU. Training each patch-classifier with clinical patches took approximately 3 days, while training each whole image classifier took about 2.5 days.

B. Parameters for generating the synthetic patch dataset

The patch size is set to 256×256 pixels, and β varies between 1.2 and 1.6 (see Equation (8)). Calcifications are organized into clusters containing 5 to 12 instances within square areas with side lengths ranging from 15 to 60 pixels. Calcification intensity spans 90% to 100% of the maximum image intensity. Masses are represented as Gaussian-shaped profiles with radii ranging from 5 to 45 pixels. The radii may differ along the two axes, resulting in oval or circular shapes. The intensity at the center of each mass is between 90% and 100% of the maximum image intensity. All parameters are adjustable using the code provided in the supplementary material.

C. Proof of Lemma 2.4

Proof. Starting with the definition of the Contrastive loss from Lemma E.3, we seek to linearize the first term of Equation (16).

By using the 2-nd order Taylor development of the exponential around $\mathbb{E}_{X' \sim \pi_p}[k(X, X')/\tau]$, we can write

$$\begin{aligned}
 e^{k(X, X')/\tau} &\approx \\
 &e^{\mathbb{E}_{X' \sim \pi_p}[k(X, X')/\tau]} \left(1 + \frac{1}{\tau} k(X, X') - \frac{1}{\tau} \mathbb{E}_{X' \sim \pi_p}[k(X, X')] \right. \\
 &\left. + \frac{1}{2\tau^2} (k(X, X') - \mathbb{E}_{X' \sim \pi_p}[k(X, X')])^2 + \mathcal{O}\left(\frac{(k(X, X') - \mathbb{E}_{X' \sim \pi_p}[k(X, X')])^3}{\tau^3}\right) \right).
 \end{aligned}$$

Then, we have

$$\begin{aligned}
 & \mathbb{E}_{X \sim \pi_p} \left[\log \mathbb{E}_{X' \sim \pi_p} \left[e^{k(X, X')/\tau} \right] \right] = \\
 & \mathbb{E}_{X \sim \pi_p} \left[\log \mathbb{E}_{X' \sim \pi_p} \left[e^{\mathbb{E}_{X' \sim \pi_p} [k(X, X')]/\tau} \left(1 + \frac{1}{\tau} k(X, X') - \frac{1}{\tau} \mathbb{E}_{X' \sim \pi_p} [k(X, X')] + \right. \right. \right. \\
 & \left. \left. \left. \frac{1}{2\tau^2} (k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^2 + \mathcal{O} \left(\frac{(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^3}{\tau^3} \right) \right) \right] \right], \\
 \\
 & \mathbb{E}_{X \sim \pi_p} \left[\log \mathbb{E}_{X' \sim \pi_p} \left[e^{k(X, X')/\tau} \right] \right] = \\
 & \mathbb{E}_{X \sim \pi_p} \left[\log \left(e^{\mathbb{E}_{X' \sim \pi_p} [k(X, X')]/\tau} \left(1 + \frac{1}{\tau} \mathbb{E}_{X' \sim \pi_p} [k(X, X')] - \frac{1}{\tau} \mathbb{E}_{X' \sim \pi_p} [k(X, X')] + \right. \right. \right. \\
 & \left. \left. \left. \mathbb{E}_{X' \sim \pi_p} \left[\frac{1}{2\tau^2} (k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^2 \right] + \mathbb{E}_{X' \sim \pi_p} \left[\mathcal{O} \left(\frac{(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^3}{\tau^3} \right) \right] \right) \right) \right], \\
 \\
 & \mathbb{E}_{X \sim \pi_p} \left[\log \mathbb{E}_{X' \sim \pi_p} \left[e^{k(X, X')/\tau} \right] \right] = \\
 & \mathbb{E}_{X \sim \pi_p} \left[\log \left(e^{\mathbb{E}_{X' \sim \pi_p} [k(X, X')]/\tau} \left(1 + \mathbb{E}_{X' \sim \pi_p} \left[\frac{1}{2\tau^2} (k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^2 \right] + \right. \right. \right. \\
 & \left. \left. \left. \mathcal{O} \left(\frac{\mathbb{E}_{X' \sim \pi_p} [(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^3]}{\tau^3} \right) \right) \right) \right], \\
 \\
 & \mathbb{E}_{X \sim \pi_p} \left[\log \mathbb{E}_{X' \sim \pi_p} \left[e^{k(X, X')/\tau} \right] \right] = \\
 & \mathbb{E}_{X \sim \pi_p} \left[\log \left(e^{\mathbb{E}_{X' \sim \pi_p} [k(X, X')]/\tau} \left(1 + \mathcal{O} \left(\frac{\mathbb{E}_{X' \sim \pi_p} [(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^3]}{\tau^3} \right) \right) \right) \right] \quad (9) \\
 & + e^{\mathbb{E}_{X' \sim \pi_p} [k(X, X')]/\tau} \frac{1}{2\tau^2} \mathbb{E}_{X' \sim \pi_p} [(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^2] \left. \right].
 \end{aligned}$$

They 1-st order Taylor expansion of $f(x) = \log(a + bx)$ around zero is given by:

$$\log(a + bx) \approx \log a + \frac{b}{a}x + \mathcal{O} \left(\frac{b^2}{a^2}x^2 \right).$$

In Equation (9), we have:

$$\begin{cases}
 a = e^{\mathbb{E}_{X' \sim \pi_p} [k(X, X')]/\tau} \left(1 + \mathcal{O} \left(\frac{\mathbb{E}_{X' \sim \pi_p} [(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^3]}{\tau^3} \right) \right) \\
 b = \frac{1}{2} e^{\mathbb{E}_{X' \sim \pi_p} [k(X, X')]/\tau} \\
 x = \mathbb{E}_{X' \sim \pi_p} \left[\frac{(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^2}{\tau^2} \right],
 \end{cases}$$

which implies

$$\begin{aligned}
 \log a &= \log \left(e^{\mathbb{E}_{X' \sim \pi_p} [k(X, X')]/\tau} \left(1 + \mathcal{O} \left(\frac{\mathbb{E}_{X' \sim \pi_p} [(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^3]}{\tau^3} \right) \right) \right) \\
 &= \frac{1}{\tau} \mathbb{E}_{X' \sim \pi_p} [k(X, X')] + \mathcal{O} \left(\log \left(1 + \frac{\mathbb{E}_{X' \sim \pi_p} [(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^3]}{\tau^3} \right) \right),
 \end{aligned}$$

and

$$\frac{b}{a} = \frac{1}{2} \frac{1}{1 + \mathcal{O}\left(\frac{\mathbb{E}_{X' \sim \pi_p} [(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^3]}{\tau^3}\right)}.$$

We thus have

$$\begin{aligned} & \mathbb{E}_{X \sim \pi_p} \left[\log \mathbb{E}_{X' \sim \pi_p} \left[e^{k(X, X')/\tau} \right] \right] \approx \\ & \mathbb{E}_{X \sim \pi_p} \left[\frac{1}{\tau} \mathbb{E}_{X' \sim \pi_p} [k(X, X')] + \mathcal{O}\left(\log \left(1 + \frac{\mathbb{E}_{X' \sim \pi_p} [(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^3]}{\tau^3} \right) \right) \right. \\ & \quad \left. + \frac{1}{2} \frac{1}{1 + \mathcal{O}\left(\frac{\mathbb{E}_{X' \sim \pi_p} [(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^3]}{\tau^3}\right)} \mathbb{E}_{X' \sim \pi_p} \left[\frac{(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^2}{\tau^2} \right] + \right. \\ & \quad \left. \mathcal{O}\left(\frac{1}{\left(1 + \mathcal{O}\left(\frac{\mathbb{E}_{X' \sim \pi_p} [(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^3]}{\tau^3}\right)\right)^2} \frac{(\mathbb{E}_{X' \sim \pi_p} [(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^2])^2}{\tau^4} \right) \right]. \end{aligned}$$

However, in the vicinity of $\mathbb{E}_{X' \sim \pi_p} [k(X, X')]$, where the Taylor approximation is valid, we have that $|\mathbb{E}_{X' \sim \pi_p} [k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')]| < \tau$ and

$$\begin{aligned} & \mathbb{E}_{X \sim \pi_p} \left[\log \mathbb{E}_{X' \sim \pi_p} \left[e^{k(X, X')/\tau} \right] \right] \\ & \approx \mathbb{E}_{X \sim \pi_p} \left[\frac{1}{\tau} \mathbb{E}_{X' \sim \pi_p} [k(X, X')] + \frac{1}{2} \mathbb{E}_{X' \sim \pi_p} \left[\frac{(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^2}{\tau^2} \right] + \right. \\ & \quad \left. \mathcal{O}\left(\frac{(\mathbb{E}_{X' \sim \pi_p} [(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^2])^2}{\tau^4} \right) \right] \\ & \approx \frac{1}{\tau} \mathbb{E}_{X, X' \sim \pi_p} [k(X, X')] + \frac{1}{2\tau^2} \mathbb{E}_{X \sim \pi_p} \left[\mathbb{E}_{X' \sim \pi_p} [(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^2] \right] + \\ & \quad \mathcal{O}\left(\frac{\mathbb{E}_{X \sim \pi_p} \left[(\mathbb{E}_{X' \sim \pi_p} [(k(X, X') - \mathbb{E}_{X' \sim \pi_p} [k(X, X')])^2])^2 \right]}{\tau^4} \right) \\ & \approx \frac{1}{\tau} \mathbb{E}_{X, X' \sim \pi_p} [k(X, X')] + \frac{1}{2\tau^2} \mathbb{E}_{X \sim \pi_p} [\text{Var}_{X' \sim \pi_p} [k(X, X')]] + \\ & \quad \mathcal{O}\left(\frac{\mathbb{E}_{X \sim \pi_p} [\text{Var}_{X' \sim \pi_p} [k(X, X')]^2]}{\tau^4} \right), \end{aligned}$$

and Equation (16) can then be re-written as

$$\begin{aligned} \mathcal{L}_{Contr} & \approx \frac{1}{\tau} \mathbb{E}_{X, X' \sim \pi_p} [k(X, X')] + \frac{1}{2\tau^2} \mathbb{E}_{X \sim \pi_p} [\text{Var}_{X' \sim \pi_p} [k(X, X')]] \\ & \quad - \frac{1}{\tau} \mathbb{E}_{X, X' \sim P_{os}} [k(X, X')] + \log(|\mathcal{B}| - 1) + \mathcal{O}\left(\frac{\mathbb{E}_{X \sim \pi_p} [\text{Var}_{X' \sim \pi_p} [k(X, X')]^2]}{\tau^4} \right). \end{aligned}$$

By solving for $\mathbb{E}_{X, X' \sim P_{os}}[k(X, X')]$,

$$\begin{aligned} \mathbb{E}_{X, X' \sim P_{os}}[k(X, X')] &\approx \mathbb{E}_{X, X' \sim \pi_p}[k(X, X')] + \frac{1}{2\tau} \mathbb{E}_{X \sim \pi_p} [\text{Var}_{X' \sim \pi_p}[k(X, X')]] \\ &+ \tau \log(|\mathcal{B}|-1) - \tau \mathcal{L}_{Contr} + \mathcal{O}\left(\frac{\mathbb{E}_{X \sim \pi_p} [\text{Var}_{X' \sim \pi_p}[k(X, X')]^2]}{\tau^3}\right). \end{aligned} \quad (10)$$

As in the CMMD the positive pairs are sampled from a mixture distribution with equiprobable domains, we need to set $p = 1/2$ in Equation (10). This means that if the observed mixture distribution does not have equiprobable domains, the batch size has to be artificially constructed to have them. By setting $p = 1/2$ and using Equation (10) with Lemma E.4, we have

$$\begin{aligned} CMMD^2(\mathcal{D}_0, \mathcal{D}_1) &\approx 2\mathbb{E}_{C \sim Q} \left[\mathbb{E}_{X, X' \sim \pi_{0.5}^{X|y}} [k(X, X')] + \mathbb{E}_{X, X' \sim \pi_{1,C}^{X|y}} [k(X, X')] \right] \\ &- 4 \left(\mathbb{E}_{X, X' \sim \pi_{0.5}^X} [k(X, X')] + \frac{1}{2\tau} \mathbb{E}_{X \sim \pi_{0.5}^X} [\text{Var}_{X' \sim \pi_{0.5}^X} [k(X, X')]] + \tau \log(|\mathcal{B}|-1) \right. \\ &\left. - \tau \mathcal{L}_{Contr} + \mathcal{O}\left(\frac{\mathbb{E}_{X \sim \pi_{0.5}^X} [\text{Var}_{X' \sim \pi_{0.5}^X} [k(X, X')]^2]}{\tau^3}\right) \right), \end{aligned}$$

and by re-organizing

$$\begin{aligned} \mathcal{L}_{Contr} &\approx \frac{1}{4\tau} CMMD^2(\mathcal{D}_0, \mathcal{D}_1) + \frac{1}{\tau} \mathbb{E}_{X, X' \sim \pi_{0.5}^X} [k(X, X')] - \frac{1}{2\tau} \mathbb{E}_{C \sim Q} \left[\mathbb{E}_{X, X' \sim \pi_{0,C}^{X|y}} [k(X, X')] \right. \\ &\quad \left. + \mathbb{E}_{X, X' \sim \pi_{1,C}^{X|y}} [k(X, X')] \right] + \frac{1}{2\tau^2} \mathbb{E}_{X \sim \pi_{0.5}^X} [\text{Var}_{X' \sim \pi_{0.5}^X} [k(X, X')]] + \log(|\mathcal{B}|-1) \\ &+ \mathcal{O}\left(\frac{\mathbb{E}_{X \sim \pi_{0.5}^X} [\text{Var}_{X' \sim \pi_{0.5}^X} [k(X, X')]^2]}{\tau^4}\right). \end{aligned}$$

Multiplying the two sides by τ proves the relationship. □

D. Proof of Lemma 2.5

Lemma 2.5 has been proven by Li et al. (2021) for the NT-Xent loss. Here, we prove it is also valid for the Supervised Contrastive loss.

Proof. Using Lemmas E.3 and E.5 to write the contrastive losses (Supervised and NT-Xent) and $\text{HSIC}(X, Y)$ in terms of the expectations, and applying Theorem B.1 of Li et al. (2021) we obtain:

$$-\text{HSIC}(X, Y) + \gamma \text{HSIC}(X, X) + \mathcal{O}(\text{Var}[k(X, X')]) \leq \mathcal{L}_{Contr}. \quad (11)$$

Theorem B.1 gives the conditions for the kernels and for γ . Finally, in the Appendix B, Li et al. (2021) prove the following relationship between the inter-class MMD and $\text{HSIC}(X, Y)$:

$$\underbrace{\mathbb{E}_{C_1, C_2 \sim \pi_{0.5}^Y} \left[\left\| \mathbb{E}_{X \sim \pi_{0.5, C_1}^{X|y}} [\phi(X)] - \mathbb{E}_{X \sim \pi_{0.5, C_2}^{X|y}} [\phi(X)] \right\|^2 \right]}_{\text{inter-class MMD}} = \alpha \text{HSIC}(X, Y), \quad (12)$$

where α is a proportionality constant which depends on problem parameters, such as the number of classes and the kernels used. Combining Equations (12) and (11) proves the lemma. \square

E. Useful lemmas

Lemma E.1. *Let $l(y, y') = \langle y, y' \rangle_{\mathcal{Y}}$ be a kernel over \mathcal{Y} , which is assumed to be a function of $y \cdot y'$ or $\|y - y'\|$. Then, it can be written as:*

$$l(y, y') = \begin{cases} l_1 & \text{if } y = y' \\ l_0 & \text{otherwise} \end{cases} = \Delta l \mathbf{1}_{\{y=y'\}} + l_0, \quad (13)$$

where $\Delta = l_1 - l_0$ and $\mathbf{1}_{\{\cdot\}}$ is the indicator function.

Proof. Assuming that the label of each sample y is in one-hot format, any kernel that is a function of $y \cdot y'$ or $\|y - y'\|$ can take only two possible values: l_1 when the two data points share the same label, i.e., $y = y'$, and l_0 when they have different labels, i.e., $y \neq y'$. \square

Lemma E.2. *Given two domains $\mathcal{D}_0 = \{\mathcal{X} \times \mathcal{Y}, \pi_0\}$ and $\mathcal{D}_1 = \{\mathcal{X} \times \mathcal{Y} \times \mathcal{Y}, \pi_1\}$. The expectation of an arbitrary integrable function $g : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ on the mixture domain $\mathcal{D}_p = \{\mathcal{X} \times \mathcal{Y}, \pi_p\}$ is given by:*

$$\begin{aligned} \mathbb{E}_{X, X' \sim \pi_p^{\mathcal{X}}} [g(X, X')] &= p^2 \mathbb{E}_{X, X' \sim \pi_1^{\mathcal{X}}} [g(X, X')] + p(1-p) \mathbb{E}_{X \sim \pi_1^{\mathcal{X}}, X' \sim \pi_0^{\mathcal{X}}} [g(X, X')] \\ &\quad + p(1-p) \mathbb{E}_{X \sim \pi_0^{\mathcal{X}}, X' \sim \pi_1^{\mathcal{X}}} [g(X, X')] + (1-p)^2 \mathbb{E}_{X, X' \sim \pi_0^{\mathcal{X}}} [g(X, X')]. \end{aligned} \quad (14)$$

If the variables X and X' are interchangeable in $g(X, X')$, then the expectation is given by:

$$\mathbb{E}_{X, X' \sim \pi_p^{\mathcal{X}}} [g(X, X')] = p^2 \mathbb{E}_{X, X' \sim \pi_1^{\mathcal{X}}} [g(X, X')] + 2p(1-p) \mathbb{E}_{X \sim \pi_1^{\mathcal{X}}, X' \sim \pi_0^{\mathcal{X}}} [g(X, X')] + (1-p)^2 \mathbb{E}_{X, X' \sim \pi_0^{\mathcal{X}}} [g(X, X')]. \quad (15)$$

Proof. By defining a binary hidden variable $Z \sim \text{Ber}(p)$ that determines the original distribution from which X is sampled ($Z \in \{1, 2\}$), we have:

$$\begin{aligned} &\mathbb{E}_{X, X' \sim \pi_p^{\mathcal{X}}} [g(X, X')] \\ &= \mathbb{E}_{Z, Z' \sim \text{Ber}(p)} \left[\mathbb{E}_{X, X' \sim \pi_p^{\mathcal{X}}} [g(X, X') | Z = z, Z' = z'] \right] \\ &= p^2 \mathbb{E}_{X, X' \sim \pi_p^{\mathcal{X}}} [g(X, X') | Z = 1, Z' = 1] \\ &\quad + p(1-p) \mathbb{E}_{X, X' \sim \pi_p^{\mathcal{X}}} [g(X, X') | Z = 1, Z' = 2] + (1-p)p \mathbb{E}_{X, X' \sim \pi_p^{\mathcal{X}}} [g(X, X') | Z = 2, Z' = 1] \\ &\quad + (1-p)^2 \mathbb{E}_{X, X' \sim \pi_p^{\mathcal{X}}} [g(X, X') | Z = 2, Z' = 2] \\ &= p^2 \mathbb{E}_{X, X' \sim \pi_1^{\mathcal{X}}} [g(X, X')] + p(1-p) \mathbb{E}_{X \sim \pi_1^{\mathcal{X}}, X' \sim \pi_0^{\mathcal{X}}} [g(X, X')] + (1-p)p \mathbb{E}_{X \sim \pi_0^{\mathcal{X}}, X' \sim \pi_1^{\mathcal{X}}} [g(X, X')] \\ &\quad + (1-p)^2 \mathbb{E}_{X, X' \sim \pi_0^{\mathcal{X}}} [g(X, X')]. \end{aligned}$$

If $g(X, X') = g(X', X)$ we have that $\mathbb{E}_{X \sim \pi_1^{\mathcal{X}}, X' \sim \pi_0^{\mathcal{X}}} [g(X, X')] = \mathbb{E}_{X \sim \pi_0^{\mathcal{X}}, X' \sim \pi_1^{\mathcal{X}}} [g(X, X')]$ and thus

$$\mathbb{E}_{X, X' \sim \pi_p^{\mathcal{X}}} [g(X, X')] = p^2 \mathbb{E}_{X, X' \sim \pi_1^{\mathcal{X}}} [g(X, X')] + 2p(1-p) \mathbb{E}_{X \sim \pi_1^{\mathcal{X}}, X' \sim \pi_0^{\mathcal{X}}} [g(X, X')] + (1-p)^2 \mathbb{E}_{X, X' \sim \pi_0^{\mathcal{X}}} [g(X, X')].$$

\square

Lemma E.3. *The NT-Xent loss and the Supervised Contrastive loss can be written in terms of the expectation by the following equation, with $k(X, X') = \phi(X)^T \phi(X') = Z^T Z'$:*

$$\mathcal{L}_{Contr} \approx \mathbb{E}_{X \sim \pi_p^X} \left[\log \mathbb{E}_{X' \sim \pi_p^X} \left[e^{k(X, X')/\tau} \right] \right] - \frac{1}{\tau} \mathbb{E}_{X, X' \sim Pos} [k(X, X')] + \log(|\mathcal{B}| - 1), \quad (16)$$

where π_p^X is the probability measure of the mixture of the two domains, with a mixture probability p , and where \mathcal{L}_{Contr} represents any of the two Contrastive losses.

Proof (NT-Xent loss). From Equation (1), and by replacing $k(X, X') = \phi(X)^T \phi(X') = Z^T Z'$

$$\mathcal{L}_{Contr} = \frac{1}{|\mathcal{B}|} \sum_{x_i \in \mathcal{B}} \log \sum_{l \in \mathcal{A}(i)} e^{k(x_i, x_l)/\tau} - \frac{1}{|\mathcal{B}| \tau} \sum_{x_i \in \mathcal{B}} k(x_i, x_{j(i)}).$$

By denoting as $\hat{\mathbb{E}}$ the estimation of the expectation, we can write

$$\mathcal{L}_{Contr} = \hat{\mathbb{E}}_{X \sim \pi_p^X} \left[\log \left((|\mathcal{B}| - 1) \hat{\mathbb{E}}_{X' \sim \pi_p^X, X' \neq X} [e^{k(X, X')/\tau}] \right) \right] - \frac{1}{\tau} \hat{\mathbb{E}}_{X, X' \sim Pos} [k(X, X')]. \quad (17)$$

By using the product property of the logarithmic we get

$$\mathcal{L}_{Contr} = \hat{\mathbb{E}}_{X \sim \pi_p^X} \left[\log \hat{\mathbb{E}}_{X' \sim \pi_p^X, X' \neq X} [e^{k(X, X')/\tau}] \right] - \frac{1}{\tau} \hat{\mathbb{E}}_{X, X' \sim Pos} [k(X, X')] + \log(|\mathcal{B}| - 1).$$

For $|\mathcal{B}|$ sufficiently large, $\hat{\mathbb{E}}_{X \sim \pi_p^X} [\dots] \approx \mathbb{E}_{X \sim \pi_p^X} [\dots]$, $\hat{\mathbb{E}}_{X, X' \sim Pos} [\dots] \approx \mathbb{E}_{X, X' \sim Pos} [\dots]$ and

$\hat{\mathbb{E}}_{X' \sim \pi_p^X, X' \neq X} [\dots] \approx \mathbb{E}_{X' \sim \pi_p^X, X' \neq X} [\dots] = \mathbb{E}_{X' \sim \pi_p^X} [\dots]$, as $\pi_p^X(X = X') = 0 \quad \forall X, X' \in \mathcal{X} \times \mathcal{X}$. Thus,

$$\mathcal{L}_{Contr} \approx \mathbb{E}_{X \sim \pi_p^X} \left[\log \mathbb{E}_{X' \sim \pi_p^X} [e^{k(X, X')/\tau}] \right] - \frac{1}{\tau} \mathbb{E}_{X, X' \sim Pos} [k(X, X')] + \log(|\mathcal{B}| - 1).$$

□

Proof (Supervised Contrastive loss). From Equation (2), and by replacing $k(X, X') = \phi(X)^T \phi(X') = Z^T Z'$

$$\mathcal{L}_{Contr} = \frac{1}{|\mathcal{B}|} \sum_{i \in |\mathcal{B}|} \frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \sum_{l \in \mathcal{A}(i)} e^{k(x_i, x_l)/\tau} - \frac{1}{|\mathcal{B}| \tau} \sum_{i \in |\mathcal{B}|} \frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} k(x_i, x_j).$$

By rewriting in terms of the expectations, we obtain

$$\mathcal{L}_{Contr} = \hat{\mathbb{E}}_{X, X' \sim Pos} \left[\log \left((|\mathcal{B}| - 1) \hat{\mathbb{E}}_{X'' \sim \pi_p^X, X'' \neq X} [e^{k(X, X'')/\tau}] \right) \right] - \frac{1}{\tau} \hat{\mathbb{E}}_{X, X' \sim Pos} [k(X, X')].$$

As the first term does not depend on X' , we can rewrite

$$\mathcal{L}_{Contr} = \hat{\mathbb{E}}_{X \sim \pi_p^X} \left[\log \left((|\mathcal{B}| - 1) \hat{\mathbb{E}}_{X' \sim \pi_p^X, X' \neq X} [e^{k(X, X')/\tau}] \right) \right] - \frac{1}{\tau} \hat{\mathbb{E}}_{X, X' \sim Pos} [k(X, X')],$$

where we have renamed X'' as X' , and have assumed a balanced batch in terms of the classes. This is Equation (17), and we can then proceed as in the Self-supervised Learning case to obtain Equation (16).

□

Lemma E.4. *By considering a mapping of the type $\phi : \mathcal{X} \rightarrow \mathcal{Z} \subseteq \mathbb{R}^m$, the square of the CMMD can be written in terms of the expectation by the following equation:*

$$CMMD^2(\mathcal{D}_0, \mathcal{D}_1, \phi) = 2\mathbb{E}_{C \sim \pi^Y} \left[\mathbb{E}_{X, X' \sim \pi_{0,C}^{X|Y}} [k(X, X')] + \mathbb{E}_{X, X' \sim \pi_{1,C}^{X|Y}} [k(X, X')] \right] - 4\mathbb{E}_{X, X' \sim Pos} [k(X, X') | 0.5], \quad (18)$$

where $k(X, X') = \phi(X)^T \phi(X')$, $\mathbb{E}_{C \sim \pi^Y} \left[\mathbb{E}_{X, X' \sim \pi_{0,C}^{X|Y}} [k(X, X')] + \mathbb{E}_{X, X' \sim \pi_{1,C}^{X|Y}} [k(X, X')] \right]$ is the mean similarity in each class $C \in \{1, \dots, c\}$ and domain $D \in \{1, 2\}$, $X, X' \sim Pos$ indicates that X and X' are positive pairs (instances with the same class, and same or different domain), and p is the domain mixture probability. The fact that $p = 1/2$ states that for the CMMD definition, the two domains are equiprobable.

Proof. As $\phi : \mathcal{X} \rightarrow \mathcal{Z} \subseteq \mathbb{R}^m$, $\langle \phi(X), \phi(X') \rangle = \phi(X)^T \phi(X')$ and by using the linearity of the expectation and inner product in Equation (3), we obtain

$$\begin{aligned}
 & \text{CMMD}^2(\mathcal{D}_0, \mathcal{D}_1, \phi) \\
 &= \mathbb{E}_{C \sim \pi^{\mathcal{Y}}} \left[\langle \mathbb{E}_{X \sim \pi_{0,C}^{\mathcal{X}|Y}} [\phi(X)] - \mathbb{E}_{X \sim \pi_{1,C}^{\mathcal{X}|Y}} [\phi(X)], \mathbb{E}_{X \sim \pi_{0,C}^{\mathcal{X}|Y}} [\phi(X)] - \mathbb{E}_{X \sim \pi_{1,C}^{\mathcal{X}|Y}} [\phi(X)] \rangle \right] \\
 &= \mathbb{E}_{C \sim \pi^{\mathcal{Y}}} \left[\langle \mathbb{E}_{X \sim \pi_{0,C}^{\mathcal{X}|Y}} [\phi(X)], \mathbb{E}_{X \sim \pi_{0,C}^{\mathcal{X}|Y}} [\phi(X)] \rangle - 2 \langle \mathbb{E}_{X \sim \pi_{0,C}^{\mathcal{X}|Y}} [\phi(X)], \mathbb{E}_{X \sim \pi_{1,C}^{\mathcal{X}|Y}} [\phi(X)] \rangle + \langle \mathbb{E}_{X \sim \pi_{1,C}^{\mathcal{X}|Y}} [\phi(X)], \mathbb{E}_{X \sim \pi_{1,C}^{\mathcal{X}|Y}} [\phi(X)] \rangle \right] \\
 &= \mathbb{E}_{C \sim \pi^{\mathcal{Y}}} \left[\langle \mathbb{E}_{X, X' \sim \pi_{0,C}^{\mathcal{X}|Y}} [\langle \phi(X), \phi(X') \rangle] - 2 \mathbb{E}_{X \sim \pi_{1,C}^{\mathcal{X}|Y}, X' \sim \pi_{1,C}^{\mathcal{X}|Y}} [\langle \phi(X), \phi(X') \rangle] + \mathbb{E}_{X, X' \sim \pi_{1,C}^{\mathcal{X}|Y}} [\langle \phi(X), \phi(X') \rangle] \right] \\
 &= \mathbb{E}_{C \sim \pi^{\mathcal{Y}}} \left[\mathbb{E}_{X, X' \sim \pi_{0,C}^{\mathcal{X}|Y}} [k(X, X')] - 2 \mathbb{E}_{X \sim \pi_{1,C}^{\mathcal{X}|Y}, X' \sim \pi_{1,C}^{\mathcal{X}|Y}} [k(X, X')] + \mathbb{E}_{X, X' \sim \pi_{1,C}^{\mathcal{X}|Y}} [k(X, X')] \right],
 \end{aligned}$$

where it was used that $k(x, y) = \langle \phi(x), \phi(y) \rangle$. The notation was simplified by dropping the explicit dependence on the space in the inner products and norms, i.e., $\langle \cdot, \cdot \rangle_{\mathcal{Z}} = \langle \cdot, \cdot \rangle$ and $\| \cdot \|_{\mathcal{Z}} = \| \cdot \|$. By adding and subtracting the intra-domain similarities, we have

$$\begin{aligned}
 \text{CMMD}^2(\mathcal{D}_0, \mathcal{D}_1, \phi) &= 2 \mathbb{E}_{C \sim \pi^{\mathcal{Y}}} \left[\mathbb{E}_{X, X' \sim \pi_{0,C}^{\mathcal{X}|Y}} [k(X, X')] + \mathbb{E}_{X, X' \sim \pi_{1,C}^{\mathcal{X}|Y}} [k(X, X')] \right] \\
 &\quad - \mathbb{E}_{C \sim \pi^{\mathcal{Y}}} \left[\mathbb{E}_{X, X' \sim \pi_{0,C}^{\mathcal{X}|Y}} [k(X, X')] + 2 \mathbb{E}_{X \sim \pi_{1,C}^{\mathcal{X}|Y}, X' \sim \pi_{1,C}^{\mathcal{X}|Y}} [k(X, X')] + \mathbb{E}_{X, X' \sim \pi_{1,C}^{\mathcal{X}|Y}} [k(X, X')] \right].
 \end{aligned}$$

From Lemma E.2, if the mixture probability is $p = 1/2$, we have:

$$\begin{aligned}
 & \text{CMMD}^2(\mathcal{D}_0, \mathcal{D}_1, \phi) = \\
 & 2 \mathbb{E}_{C \sim \pi^{\mathcal{Y}}} \left[\mathbb{E}_{X, X' \sim \pi_{0,C}^{\mathcal{X}|Y}} [k(X, X')] + \mathbb{E}_{X, X' \sim \pi_{1,C}^{\mathcal{X}|Y}} [k(X, X')] \right] - 4 \mathbb{E}_{C \sim \pi^{\mathcal{Y}}} \left[\mathbb{E}_{X, X' \sim \pi_{m,0.5,C}^{\mathcal{X}|Y}} [k(X, X')] \right],
 \end{aligned} \tag{19}$$

as the similarity is symmetric with respect to X and X' . The first term of Equation (19) contains the mean similarity inside each cluster (label and domain). The second term of Equation (19) contains the mean similarity between features that share the same label, with different and same domain. In the Contrastive Learning literature, these are commonly denoted as positive pairs. Equation (19) can then be rewritten in terms of the expectation:

$$\text{CMMD}^2(\mathcal{D}_0, \mathcal{D}_1, \phi) = 2 \mathbb{E}_{C \sim \pi^{\mathcal{Y}}} \left[\mathbb{E}_{X, X' \sim \pi_{0,C}^{\mathcal{X}|Y}} [k(X, X')] + \mathbb{E}_{X, X' \sim \pi_{1,C}^{\mathcal{X}|Y}} [k(X, X')] \right] - 4 \mathbb{E}_{X, X' \sim Pos} [k(X, X') | p = 1/2],$$

where p is the mixture probability. □

Lemma E.5. *In a learning setting with N data points sampled independently with the same probability, the $HSIC(X, Y)$ can be written as*

$$HSIC(X, Y) = \beta \mathbb{E}_{X, X' \sim Pos} [k(X, X')] - \beta \mathbb{E} [k(X, X')], \tag{20}$$

where $X, X' \sim Pos$ means that the features sampled are positive pairs and β is a constant. For the Supervised Contrastive loss $\beta = \frac{\Delta}{K}$, with K the number of equiprobable classes and Δ a kernel-related constant, whereas for the Self-supervised Contrastive loss $\beta = \frac{\Delta}{N}$.

Proof (NT-Xent loss). Refer to Theorem A.1 of Li et al. (2021). □

Proof (Supervised Contrastive loss). We use Equation (2) of Li et al. (2021) to write $HSIC(X, Y)$ as

$$HSIC(X, Y) = \mathbb{E} [k(X, X') l(Y, Y')] - 2 \mathbb{E} [k(X, X') l(Y, Y'')] + \mathbb{E} [k(X, X')] \mathbb{E} [l(Y, Y')]. \tag{21}$$

Using Lemma E.1, the first term of Equation (21) yields:

$$\begin{aligned}
 \mathbb{E} [k(X, X')l(Y, Y')] &= \Delta l \mathbb{E} [k(X, X') \mathbf{1}_{\{Y=Y'\}}] + l_0 \mathbb{E} [k(X, X')] \\
 &= \Delta l \mathbb{E}_{Y, Y'} [\mathbb{E}_{X, X'} [k(X, X') \mathbf{1}_{\{y=y'\}} | Y=y, Y'=y']] + l_0 \mathbb{E} [k(X, X')] \\
 &= \Delta l \mathbb{P}(Y=Y') \mathbb{E}_{X, X'} [k(X, X') | Y=Y'] + l_0 \mathbb{E} [k(X, X')] \\
 &= \frac{\Delta l}{M} \mathbb{E}_{X, X' \sim Pos} [k(X, X')] + l_0 \mathbb{E} [k(X, X')],
 \end{aligned} \tag{22}$$

as $\mathbb{P}(Y=Y') = 1/M$ and $\mathbb{E}_{X, X'} [k(X, X') | Y=Y] = \mathbb{E}_{X, X' \sim Pos} [k(X, X')]$. The second one, using the independence between X' and Y'' , yields:

$$\begin{aligned}
 \mathbb{E} [k(X, X')l(Y, Y'')] &= \mathbb{E}_{X, Y} \left[\mathbb{E}_{X'} \left[k(X, X') \left(\Delta l \underbrace{\mathbb{E}_{Y''} [\mathbf{1}_{\{Y=Y''\}}]}_{1/M} + l_0 \right) \right] \right] \\
 &= \left(\frac{\Delta l}{M} + l_0 \right) \mathbb{E}_{X, Y} [\mathbb{E}_{X'} [k(X, X')]] \\
 &= \left(\frac{\Delta l}{M} + l_0 \right) \mathbb{E} [k(X, X')].
 \end{aligned} \tag{23}$$

Finally, as $\mathbb{E} [l(Y, Y'')] = \frac{\Delta l}{M} + l_0$, the last term is equal to the second one, and it gets cancelled. We thus have

$$\begin{aligned}
 \text{HSIC}(X, Y) &= \frac{\Delta l}{M} \mathbb{E}_{X, X' \sim Pos} [k(X, X')] + l_0 \mathbb{E} [k(X, X')] - \left(\frac{\Delta l}{M} + l_0 \right) \mathbb{E} [k(X, X')] \\
 &= \frac{\Delta l}{M} \mathbb{E}_{X, X' \sim Pos} [k(X, X')] - \frac{\Delta l}{M} \mathbb{E} [k(X, X')].
 \end{aligned} \tag{24}$$

□