Cluster Triplet Loss for Unsupervised Domain Adaptation on Histology Images

Ruby Wood[1] Enric Domingo[2] Viktor Hendrik Koelzer[2,3,4]
Timothy S. Maughan[2,5] Jens Rittscher [1]

[1]Department of Engineering Science, University of Oxford, Oxford, UK
[2]Department of Oncology, University of Oxford, Oxford, UK
[3]Department of Pathology and Molecular Pathology, University Hospital Zurich, University of Zurich, Zurich, Switzerland
[4]Institute of Medical Genetics and Pathology, University Hospital Basel, Basel, Switzerland
[5]Department of Molecular and Clinical Cancer Medicine, University of Liverpool, Liverpool, UK
{ruby.wood,jens.rittscher}@eng.ox.ac.uk

Abstract

Deep learning models that predict cancer patient treatment response from medical images need to be generalisable across different patient cohorts. However, this can be difficult due to heterogeneity across patient populations. Here we focus on the problem of predicting colorectal cancer patients' response to radiotherapy from histology images scanned from tumour biopsies, and adapt this prediction model onto a new, visibly different, target cohort of patients. We present a novel unsupervised domain adaptation method with a Cluster Triplet Loss function, using minimal information from the source domain, resulting in an improvement in AUC from 0.544 to 0.818 on the target cohort. We avoid the use of pseudo-labels and class feature centres to avoid adding noise and bias to the adapted model, and perform experiments to verify the preferable performance of our model over such state-of-the-art methods. Our proposed approach can be applied in many complex medical imaging cases, including prediction on large whole slide images, based on combining predictions from smaller, memory-feasible representations of the image extracted from graph neural networks.

1. Introduction

Adapting a deep learning model in the field of medical imaging from one group of patients to another can be challenging, due to the wide variability that can occur between patients. In this work we focus on using deep learning to predict colorectal cancer (CRC) patients' response to radiotherapy from a digital histology image of the pre-treatment tumour tissue, and we attempt to adapt this model to a completely unseen cohort of patients from a different geographic region. In this work we focus on unsupervised domain adaptation (UDA), since for this prediction model to be useful in clinical practice we would need to adapt the model without knowledge of the patient outcomes at time of use.

While much research has been done on using domain adaptation in other fields, application to histology images is more challenging due to complications arising from the size and heterogeneity of this imaging modality [12].

Histology slides are the haematoxylin and eosin (H&E) stained, digitally scanned, tumour tissue slices cut from a biopsy sample. These slices are scanned at very high resolution, resulting in extremely large file sizes. Images must be split into smaller sections to fit into computer memory, and a multiple instance learning (MIL) method is then required to combine the predictions into one prediction per slide. Here we present a method which focuses only on the intermediate feature representation within a model, hence preserving any optional MIL methods on the features for final outputs. Specifically, we make predictions from naturally segmented tissue regions using a graph neural network (GNN) approach, using the features within the GNN to help adapt our model to a new domain.

While socioeconomic factors could influence patients' experience with cancer in different regions or countries [5], batch effects in histology images can commonly develop from the processing of the tumour biopsy once it is removed from the patient. The processing of the tissue samples is performed slightly differently across medical centres, which introduces an inherent domain shift into the data [29]. We train and validate our method on patient cohorts from three different medical centres, all of which have different tissue processing practices.

We approach our binary prediction problem with a generalised view, avoiding pseudo-labels by focusing only on adapting the underlying features to a new domain, and preserving the original classification branches. By avoiding the use of pseudo-labels, unlike many other UDA approaches [19, 30, 35, 40–43, 45], we avoid adding bias and noise from our source model into our predictions.

Furthermore, we avoid the use of class-based clustering to find a cluster representative for each class label, as many in the literature have done [10, 15, 19, 30, 43], to allow for more variance within each class label by clustering on the whole feature set at once, allowing for a natural number of clusters that is not constrained by the number of class labels in the dataset. This approach works much better particularly for binary outcome data since it allows for more than two clusters to represent the entire source dataset.

In this paper we develop a feature-alignment UDA technique to transfer our trained clinical model onto an unseen target cohort without the use of any target labels. We propose a novel approach, defining a loss function to be used in a 'source-supervised' training manner for domain adaptation. This loss only requires a lightweight representation of the source data to guide the learning of a new, domain-adapted, target model. Our method allows for distributed training of a cohort-tuned model without requiring any training or updating of the original model, therefore providing a secure federated learning technique that can protect patient confidentiality between locations. Rather than confusing the results with all the dataset permutations, we focus on the dataset which is most dissimilar as our target dataset, as this is the biggest challenge. This also mimics application in clinical practice where we would need to transfer frozen pre-trained models onto to new cohort domains to better predict patient outcomes, without advance knowledge of a patient's response to treatment. This method requires no batch or cohort assumptions and can be applied to even a single new data point.

2. Related Work

2.1. Unsupervised Domain Adaptation

Clustering While many papers have explored the use of clustering for domain adaptation, with various methods of aligning source and domain distributions using contrastive or adversarial loss approaches [13, 16, 22, 43], to the best of our knowledge none have used the lightweight clustering approach we suggest here.

The intuition behind our domain adaptation approach builds on the idea of Attracting and Dispersing [40], where the authors aim to bring similar features together and dissimilar features apart in the feature space. This unsupervised method uses k-nearest neighbours and pseudo-labels to maximise consistency of predictions between neighbours, and minimise similarity of dissimilar feature predictions. A similar method, Structurally Regularized Deep Clustering (SRDC) [30], uses KMeans to cluster intermediate network features of the target data, and minimises the Kullback-Leibler (KL) divergence between the distributions of the predicted target labels and the true source labels, as well as the KL divergence between the learnable source and target cluster centres. Another approach using KMeans is the Source Hypothesis Transfer (SHOT) method proposed by Liang *et al.* [19], who freeze the final classifier layer of a source model and use the rest as initialisation for a target model. Their unsupervised approach predicts pseudo-labels and minimises entropy, finding target class centroids in a manner similar to weighted KMeans, and then defining a target sample's pseudo-label by its nearest neighbour class centroid, measured using cosine distance.

Pseudo-labels Most UDA approaches use pseudo-labels to train their model [19, 30, 35, 40–43, 45], which can provide more information in the multi-class classification setting than the binary one. These pseudo-labels are commonly used for masking or as an indicator method to calculate some further statistic for use in a loss function [40, 45]. Methods using pseudo-labels depend heavily on the teacher model having a prior reasonable accuracy on the target domain, which is not always the case, as pointed out by Li *et al.* [18]. Crucially, they also observe that there are no common methods to evaluate the quality of these pseudo-labels. While many papers acknowledge this caveat and propose methods to counteract it [18, 30, 41, 43], it is a clear inherent design flaw that can add unnecessary bias and noise. Zhang *et al.* acknowledge this and regularize their pseudo-labels with weights during training, by measuring distances to feature centroids of classes [41]. The Divide and Contrast method divides the target data into source-like or not, and makes the reasonable assumption that pseudo-labels from source-like target data are more accurate than those from target-specific samples [43]. The SRDC authors also admit that the unreliability of the source model on the target data could lead to some incorrect target predictions, and consequently add an extra term to the loss function using pseudo-labels as an indicator on the predicted labels [30].

Triplet loss The idea of a triplet loss using central features was first introduced by [10] for object retrieval, where they propose a Triplet Centre Loss (TCL) to align features of the same class to a learnable class centre, and repel features from different classes. They use Euclidean distance to measure the difference between the class centre and sample features, as we do here, though for their negative sample in their triplet loss they choose the closest negative centre. They also use class labels to identify the corresponding class centre, so the method is not unsupervised. Other works have used a similar approach using a triplet loss on feature centres [2, 15, 33, 37], across different fields. Most focus on calculating the feature centres from pseudo-labels to find a centre representing each class in a classification problem [2, 15, 37].

The Centroid Triplet Loss proposed by Wieczorek *et al*. for image retrieval [37] uses a traditional triplet loss on the target features with the positive example as the centroid of the class of that target example, and the negative example as the centroid of a negative class, which is similar to what we propose here, but differing in our exclusion of any assumed or known class information. Lagunes-Fortiz *et al*. [15] use a different negative sample in their triplet loss, using a sample from the domain itself instead of a feature centre. The triplet loss has also been used to define target and source clusters as class guided constraints [35], for better class alignment between the domains.

2.2. Histology domain adaptation

Staining In the field of deep learning on histopathology, tissue staining and processing can vary heavily across hospitals and laboratories, and efforts have been made to counter these cohort staining effects [8, 14, 25, 44] beyond traditional colour normalisation methods [20, 31]. However, sometimes this approach alone is not enough to guarantee domain generalisability of a model. Lafarge *et al*. [14] propose a domain-adversarial neural network (DANN) to predict the probability that a sample comes from a particular domain, allowing removal of domain-specific features while maintaining those features which are useful for prediction. They also experiment with traditional staining domain adaptation methods, and their best results are achieved when the DANN is used in addition to colour augmentation or stain normalization.

Feature alignment In this work we focus on feature alignment between the source and target domains. Of the feature alignment approaches in the field of histology that use a cluster-based approach, most use pseudo-labels to find a class prediction which can help to update class-wise feature centres [6, 32]. Distill-SODA [32] is one such source-free UDA method that performs Monte Carlo simulations of its clustering for robustness. Similar to our method, they calculate a cluster centroid to compare with target features in their loss function; however their centroids are not label-agnostic but are constrained to one per class, instead of naturally deriving them from the source domain. Another feature-alignment approach introduced by Jian *et al*. [26] trains a convolutional neural network (CNN) to map target images into the source model feature space, minimising the difference between domains. This method goes further to introduce a Siamese model to encourage patches from the same whole slide image (WSI) to be classified with the same label, but this approach does not account for naturally occurring heterogeneity within the tissue sample. Wang *et al*. [36] focus on using GNN node features for alignment of CRC histology images for nuclei detection using an adversarial loss. Abbet *et al*. [1] use few source labels to train a model for CRC tissue classification.

Binary classification Most research focuses on multi-class classification or segmentation problems, where pseudo-labels or class-centres can provide a higher quantity of information. Some works focus on binary classification problems such as epithelium-stroma classification, with one paper training a single model on source and target at once and adapting the kernels of a CNN to the target domain using a simple vector multiplication of the eigenvectors corresponding to the largest eigenvalues from the target and source domains [11]. Qi *et al*. [24] also work on epithelium-stroma classification and apply a curriculum learning approach, measuring cosine similarity between samples and class centroids to avoid samples that are more likely to give false pseudo-labels, selecting initial training samples based on maximum distance to source domain.

Li *et al*. [17] focus on classifying tumour as benign or malignant on breast, lung and colon cancer histology slides. Despite the lack of outcome classes they do, however, have multiple dataset cohorts, and so their UDA approach trains a separate feature extractor on each source and target domain, and uses the source labels to learn alignment of the feature distributions. Optimal transport has also been used to penalize domain prediction in a binary classification of tumour vs normal tissue [9]. We found no previous research on UDA for models which predict patient treatment response from histology images.

Triplet loss on histology Very little research has applied triplet loss for domain adaptation on histology, and even less for unsupervised approaches. Sikaroudi *et al*. [28] use triplet loss in their efforts to learn hospital-agnostic histology representations, again focusing on the class-conditional shift across domains. They take a supervised approach with a cross entropy loss on the target predictions, as well as KL divergence to align feature domains, and a metric loss to separate classes.

3. Methods

This work assumes we already have a pre-trained source model which we wish to adapt to a new domain. We describe the source model below, which is building on a similar previous model in this field [38], and then explain how we train a new model (using the weights of the source model at initialisation) to adapt the prediction to a new domain. We explain the clustering approach used on the source data to extract a lightweight representation of the source data, which is then used in our proposed Cluster Triplet Loss function to train and adapt the new model.

We first introduce some terminology. The source model is the model that we are starting with before any domain adaptation is applied. The source model was previously trained and validated on the source data, x_s. The target data, x_t, is the new unseen dataset that we wish to adapt the source model to. The target model is an updated version of the source model that has been adapted to the target data.

3.1. Source model

Our source model is a GNN with three Graph Isomorphism Network layers [39] of feature sizes 64, 32 and 16. Instead of feeding our WSI straight into this GNN, we first apply a superpixel method on the WSI and then calculate superpixel features from patch features in the same region (size $[1, 768]$) [38], extracted using the self-supervised pretrained large histology model CTransPath [34]. From these superpixel features we construct a graph representation of each WSI, where the nodes and node features are defined from the superpixels and the edges of the graph are defined by nearest neighbours using Delaunay triangulation. These graphs are then used as input to the GNN, which is trained in a semi-supervised manner to predict a patient's response to radiotherapy.

On the source validation dataset the source model achieved metrics of 0.931 AUC, 0.803 balanced accuracy and 0.885 weighted F1. Evidently our source model can perform well on the source cohorts, and while efforts were made to generalise this model in training, the application of this model on an unseen test cohort demonstrates the inadequate generalisability of the model with metrics of 0.544 AUC, 0.500 balanced accuracy and 0.840 weighted F1, as seen in Table 2. Efforts made to avoid overfitting on the training cohorts include extensive data augmentation on the training images prior to extracting features, heavy dropout in the GNN and classification branches ($p = 0.5$), training on more than one geographic cohort of patients, and applying a multi-task learning approach to ensure the final feature set includes information on molecular traits and spatial tissue architecture as well [38].

For this work we are only concerned with the intermediate feature representation, not the final prediction stage of the model. When training our new domain-adapted model we freeze the classification branches on our target model (of which there are multiple due to a multitask learning approach with the source model, where one of these branches predicts the patient's response to radiotherapy), and we train only on the GNN layers before this. Hence in this work we focus on the node-level features of our dataset, rather than the slide-level features. We refer to the node feature extractor part of the source model as $\mathcal{F}_s$, and the classifiers after this remain constant across the source and target models.

3.2. Clustering

We use clustering on the source data to extract a lightweight, high-level representation of the source data feature set. GNNs provide us with the node-level predictions from the superpixel nodes, providing an intuitive, natural representation of tissue segments within the tumour. We extract the features of these nodes from the final layer in our GNN before we split into three prediction branches for the multi-task learning approach.

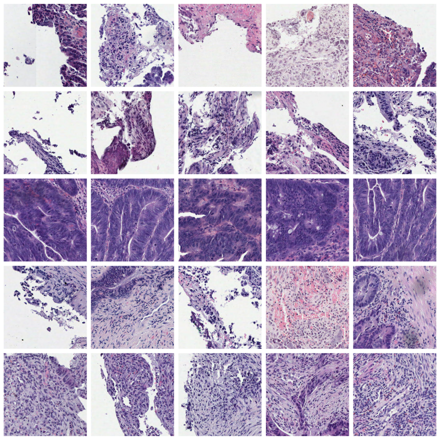

Figure 1. Nearest neighbour tissue segments for each of the five optimal clusters found on the source data. Each row represents a single cluster centre, containing the five nearest neighbours when comparing the optimal cluster centres C to the source data x_s.

We apply our clustering approach to the normalised concatenated set of node feature vectors from the source data cohorts seen in training. The concatenated feature vectors are of size $[N, 16]$, where $N = 134,132$ is the total number of nodes and 16 is the number of features per node. To find the optimal number of clusters, k_{opt}, we calculate the silhouette width [27] of the clustering for the number of clusters $k = 2, \ldots, 20$. We select the number of clusters as the cluster in this range with the highest silhouette width and Calinski-Harabasz index [4], and lowest David Bouldin score [7] for the most distinct clusters in an unsupervised setting. Due to the large sample size we use the KMeans MiniBatch approach, implemented in the Python library sklearn.cluster (version 1.1.3) [23]. We fit the MiniBatch Kmeans on a subsample ($n = 10,000$ node features) of our source dataset for efficiency, using the optimal number of clusters. We extract the resulting cluster centres C of size $[k_{opt}, 16]$.

3.3. Cluster Triplet Loss

To train and adapt our model onto the target dataset, we propose the Cluster Triplet Loss, which makes use of the source clustering from the previous section.

Our proposed Cluster Triplet Loss works on a per-sample basis, meaning it can be used to adapt a model to any size of cohort. For each feature vector provided, it calculates the mean squared error loss between the feature vector and the fixed source cluster centres, akin to one iteration of the traditional KMeans algorithm. From this we select the closest and furthest cluster centres to our input feature vector, and give these as the positive and negative samples in the calcu-

```
Algorithm 1: Training with Cluster Triplet Loss
   Input  : source feature model $\mathcal{F}_s$, source data $x_s$,
            target data $x_t$
 1 Extract source features $\mathcal{F}_s(x_s)$ from final layer of
     GNN before classification;
 2 Run KMeans on $\mathcal{F}_s(x_s)$ for $k = 2, ..., 20$ clusters
     and calculate optimal $k_{opt}$ using silhouette width;
 3 From best KMeans extract $k_{opt}$ cluster centres $C$;
 4 Initialise target model $\mathcal{F}_t$ with weights from source
     model $\mathcal{F}_s$;
 5 while Training do
 6     Extract target features from target model,
         $\mathcal{F}_t(x_t)$;
 7     Calculate Euclidean distance from $\mathcal{F}_t(x_t)$ to
         each cluster centre in $C$ with Eq. (1);
 8     Find closest ($C_{pos}$) and furthest ($C_{neg}$) clusters
         to target features using distances with Eq. (2);
 9     Calculate mean triplet loss for $x_t$ with Eqs. (3)
         and (4) over the batch and backpropagate
10 end
   Output: adapted target feature model $\mathcal{F}_t$
```

lation of the triplet loss, with the input feature vector as the anchor, to move the feature vector onto the cluster domain while simultaneously clustering the sample. We vectorize and apply this method simultaneously on all feature vectors from the model training batch. In our triplet loss implementation we use a margin of 1 and we swap the distance between the input and the negative cluster centre with the distance between the positive and negative cluster centres, as proposed by Baltnas *et al.* [3].

We first define the source model $\mathcal{M}_s = \mathcal{H}_s(\mathcal{F}_s)$, where $\mathcal{H}_s$ is the classifier part of the model and $\mathcal{F}_s$ is the feature part of the model which we adapt to a new domain. We define the target model $\mathcal{M}_t = \mathcal{H}_s(\mathcal{F}_t)$, where we use the same classifier from the source model, $\mathcal{H}_s$, but update the feature part of the source model to get $\mathcal{F}_t$. Hence the source model and target model have the exact same model architecture but different model weights.

In our proposed Cluster Triplet Loss function, we start with the cluster centres, C, from the optimal clustering of the source data. Taking our input target data, x_t, in a batch of size b, we calculate the Euclidean distance d_{ij} between the input and each cluster centre,

$$d_{ij} = \|x_{t_i} - C_j\|^2, \tag{1}$$

where $i \in [1, b]$ denotes each node input within the batch.

We use these distances to find the closest ($C_{j_{pos}}$) and furthest ($C_{j_{neg}}$) cluster centres, using

$$j_{pos_i} = \arg\min_j d_{ij}, \quad j_{neg_i} = \arg\max_j d_{ij}. \tag{2}$$

Cohort	CR	NoCR	% CR/Total	Total
Aristotle	24	97	20%	121
Grampian	61	186	25%	247
Salzburg	6	49	11%	55

Table 1. Slide counts split by outcome (CR - positive, complete response to radiotherapy, NoCR - negative, no complete response to radiotherapy) across patient cohorts.

We use these positive and negative cluster centres in our adjusted triplet loss function, as defined by

$$L_i(x_{t_i}) = \max\{\|x_{t_i} - C_{j_{pos_i}}\|^2 - \|C_{j_{pos_i}} - C_{j_{neg_i}}\|^2 + \mu, 0\} \tag{3}$$

using the margin $\mu = 1$.

Finally we reduce the output by taking the mean over our batch, and backpropagate through the model with the batch loss

$$L_b(x_t; C, \mu) = \frac{1}{b} \sum_i L_i(x_{t_i}, j_{pos_i}, j_{neg_i}; C, \mu), \tag{4}$$

where the cluster centres C and margin μ are fixed, but j_{pos_i} and j_{neg_i} vary depending on Equations (1) and (2).

The algorithm for our whole method can be found in Algorithm 1. Steps 1-3 need only be performed once, and then, given the source data representation C, steps 4 onwards can be used to train any number of target models on different domains.

4. Experiments

4.1. Data

For our experiments we have three private CRC histology datasets, Grampian, Aristotle and Salzburg, all from different geographic locations in Europe. For all datasets we have the digital WSIs of the H&E stained tumour tissue taken from pre-treatment biopsies. For Grampian and Aristotle we have the patients' recorded response to adjuvant radiotherapy treatment, categorised as pathological complete response (CR) if there are no tumour cells remaining after the treatment course is completed, or defined as no complete response (NoCR) if any number of tumour cells remain post-treatment. For the Salzburg data we define complete response to radiotherapy as having a Dvorak tumour regression of 4, post-treatment. In this work we aim to predict the response to radiotherapy as our primary binary outcome. The outcome response counts across cohorts are given in Table 1, where we can see that the ratio of positive to negative outcomes (% CR/Total) is similarly imbalanced across all cohorts.

Two of these cohorts, Grampian and Aristotle, were used for training our original source model, with the WSIs from

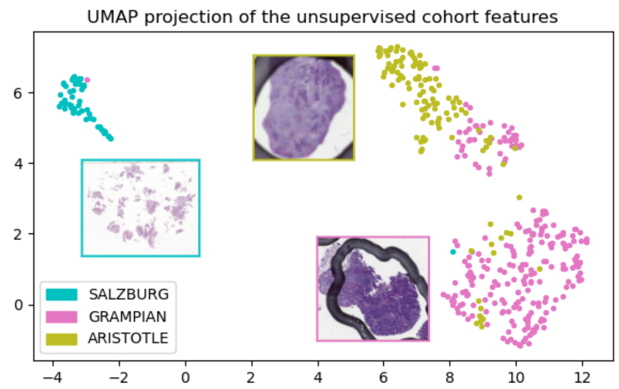

Figure 2. UMAP projections of unsupervised CTransPath features from our different patient cohorts, using the mean features per WSI. Our target dataset in this work, Salzburg, is clearly very different from our source cohorts, Grampian and Aristotle. For each cohort we overlay a section of a randomly sampled WSI in that cohort, shown in a box of the same colour, to help visualise the cohort differences.

roughly 30% patients in each cohort used for validation, and the rest used for semi-supervised training. The third cohort of patients, Salzburg, is introduced for this work as our target dataset, previously completely unseen by our model. Hence we refer to Grampian and Aristotle as our source data, and Salzburg as our target data.

The differences between the cohorts can be visualised in the reduced dimensionality UMAP projection [21] in Figure 2. For each WSI in the cohorts we extracted the unsupervised CTransPath features [34], which we use as input to our models. We fit a UMAP on the mean features per WSI, and plot the resulting embeddings, colouring by cohort. Our target cohort, Salzburg, is clearly very different to our two source cohorts, Grampian and Aristotle, and we observe the trend of sparse biopsy specimens across the Salzburg data.

4.2. Results

Clustering Applying our clustering method to our source data, we find $k_{opt} = 5$ optimal cluster centres in the feature space with the highest silhouette width of 0.28. These clusters can be visualised in Figure 1, where for each of the optimal five clusters we have plotted the five nearest neighbours to the cluster centres from the source data.

Training target model We use the weights from our source model to initialise a new target model, as described in Section 3.1. In training the target model we use heavy training data augmentations using the Pytorch torchvision.transforms library (version 0.13.1) as follows: resize, random vertical flip ($p = 0.5$), random horizontal flip ($p = 0.5$), colour jitter (brightness 0.1, contrast 0.25, saturation 0.5 and hue 0.25), Gaussian blur over a kernel of size 9, random adjust sharpness ($p = 0.2$), random auto contrast ($p = 0.5$), rotation by multiples of 90 degrees and normalizing the colour channels. We use the Adam optimiser with a learning rate of $1e{-}3$ with weight decay $1e{-}4$. We use a batch size of 32 and train for 30 epochs to avoid overfitting to the new domain (the source model was trained for 50).

Method results The results from the proposed method can be seen in Table 2, which shows the mean and standard deviation of metrics from five separate seed rounds of training, each initialised with a different random seed. Where required, we use the unoptimised threshold of 0.5 for metric calculations for fair comparison across experiments. Our domain adapted model achieves an AUC of 0.818 and a balanced accuracy of 0.619 on the target dataset, improving over the source model by +0.274 AUC and +0.119 balanced accuracy, demonstrating the effectiveness of our proposed method on this complex real world dataset.

Visualising domain shift The differences in the intermediate model features before and after domain adaptation can be visualised by plotting a UMAP embedding of the node features in Figure 3. We randomly subsampled the source data for balanced outcomes to better visualise the shift. The feature embeddings are coloured by both domain and outcome, specifying whether the data is from the source or target domain, and whether the patient outcome is a positive CR or a negative NoCR. We have plotted the UMAP embedding of the fixed source cluster centres in black, which can be useful as fiducial markers across the two plots since they aren't updated after domain adaptation. The top scatter plot shows features extracted from the source model, and the bottom scatter plot shows features extracted from our adapted target model.

The top plot in Figure 3 shows the Source CR (purple) and Source NoCR (red) classes are reasonably separated, demonstrating the competency of our source model on the source data. Before adaptation, the target outcomes (orange and green) are mixed in with each other, and the Target CR (green) shows no overlap with the Source CR (purple). However, after domain adaptation, the Target CR (green) features have moved towards the Source CR (purple) domain, better aligning the features for this minority class across domains.

Quantifying domain shift We measured the distance between our target features and our source cluster centres before and after domain adapation. Measuring the distance from the target data to the *closest* cluster centre, the mean distance over the target data decreased from 0.146 to 0.137 after our domain adaptation method (−0.009). However, measuring the distance from the target data to *all* cluster centres, the mean distance increased from 1.100 to 1.141 (+0.041). This highlights how our loss function is designed to both pull the nearest cluster centre closer, but also push other cluster centres further away, helping to create more

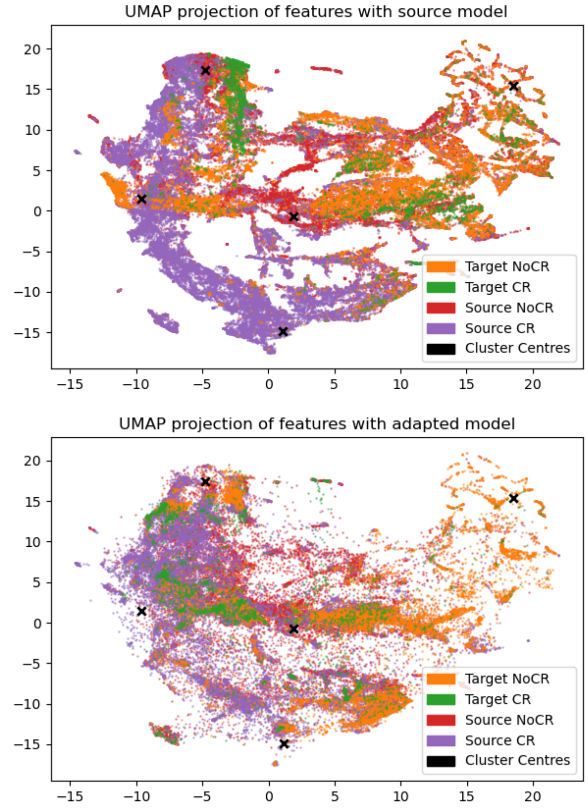

Figure 3. UMAP projections of our intermediate model features before (top) and after (bottom) applying our UDA method. Features are coloured by the source or target domain and positive (CR) or negative outcome (NoCR). The five stationary source cluster centres are overlaid in black. These target features across domains are better aligned after domain adaptation.

distinct clusters in the feature set with the idea of guiding the model to an easier classification decision.

4.3. Comparison with State-of-the-Art (SOTA)

As well as comparing our proposed method to our baseline source model, we also implemented select SOTA UDA methods for further comparison. Due to the intricate nature of most published methods, we chose to implement only those which had their code publicly available online.

We implemented four UDA methods, Distill-SODA (referred to as DS in the results table) [32], SHOT [19], TCL [10] and SRDC [30]. The results can be found in Table 2. Distill-SODA is the only method here which was specifically introduced for histology images, whereas the other methods are for general computer vision or other fields. When choosing the best epochs to evaluate results for each model, we either chose the final epoch as defined in the papers or the epoch with the lowest training loss if unspecified.

Models	k	AUC	BAcc	F1
Source	-	0.544	0.500	0.840
S-Stain	-	0.646	0.573	0.866
DS	-	0.511±0.00	0.461±0.00	0.736±0.00
TCL	-	0.684±0.01	0.605±0.04	0.860±0.02
SHOT	-	0.578±0.00	0.543±0.00	0.830±0.00
SRDC	-	0.498±0.11	0.500±0.00	0.840±0.00
k-abl	3	0.667±0.06	0.512±0.04	0.809±0.03
k-abl	4	0.750±0.05	0.607±0.06	0.847±0.02
k-abl	6	0.596±0.11	0.568±0.05	0.816±0.03
k-abl	7	0.744±0.02	0.611±0.04	0.868±0.02
Excl G	2	0.650±0.04	0.490±0.01	0.830±0.01
Excl A	5	0.573±0.15	0.527±0.04	0.768±0.11
Stain	5	0.757±0.04	0.540±0.04	0.852±0.02
Ours	5	**0.818**±0.04	**0.619**±0.04	**0.878**±0.02

Table 2. Results for our methods: Source model with no domain adaptation; S-Stain, the source model with Vahadane stain normalisation [31] on the target data; comparison SOTA UDA methods on the target dataset; k-abl, an ablation study on changing the number of clusters k used for the cluster centres in our Cluster Triplet Loss function; an ablation study removing source cohorts Grampian (Excl G) and Aristotle (Excl A) from the calculation of the cluster centers used in our loss with optimal number of clusters; Stain, our UDA approach with Vahadane stain normalisation [31] on the target data; our best model using the Cluster Triplet Loss proposed in this paper with the optimal number of clusters $k_{opt} = 5$. Metrics provided are the mean and standard deviation of the AUC, balanced accuracy (BAcc) and weighted F1 score (F1) over five seed rounds. We provide the number of clusters k where our method was used.

Each method had to be adjusted somewhat to work on our problem. The details of the implementations of the SOTA methods are given in Supplementary Material Section 7.

Overall our method has the best metrics compared to all other UDA methods implemented here.

4.4. Ablation Studies

Number of clusters For the following ablation studies we trained each model variation over five different random seeds and averaged the results. We experimented with the number of clusters and scaling the cluster centres before use in the loss function. The results from using different numbers of clusters can be found in Table 2, on the rows named *k-abl*. We found that the optimal number of clusters from our clustering analysis achieved the best results compared to other numbers of clusters. Scaling the cluster centres $\in [0, 1]$ didn't improve results either, achieving 0.747 AUC over five rounds with the optimal number of $k_{opt} = 5$ clusters. An ablation study on the clustering method used can be found in Supplementary Material Section 6.

Removing a source cohort We ran experiments where we removed one of the two source cohorts before calculating the source cluster centres, and then trained our model on the target cohort using the reduced cohort clusters in our loss function. The results averaged over five rounds can be shown in Table 2, in the rows *Excl G* and *Excl A*, where G is Grampian and A is Aristotle. These results demonstrate the importance of including both source datasets in the training set of the source model. We would expect the source model to be more generalisable when trained on more than one cohort domain, and these results show that such a model can be better adapted to new domains using our domain adaptation method.

Staining For comparison, we used Vahadane stain normalisation [31] on the target data, which has been show to be an effective technique in histology domain adaptation [14]. The source model predictions were better with stain normalisation than without (see *S-Stain* results in Table 2), but still do not match the results from our proposed UDA method. We applied our UDA method on the stain-normalised target data (see *Stain* results in Table 2), but it did not show any improvement over the standard implementation.

5. Discussion

5.1. Limitations & Future Work

We acknowledge that our adapted model is only trained up to the point of feature extraction, meaning the classification branches for the prediction of outcomes from these domain-shifted features are not updated. Since we are shifting the feature domain onto that of the original source features, on which the existing classification branches were trained to predict from, this part of the model should adapt without further training. However, there could be some useful cohort-specific information being missed in this final step.

As we demonstrated in our ablation studies (Section 4.4), finding an optimal clustering of the source data is the key to getting the best results from this method. It may be possible to extend this work to test how this approach can generalise onto multiple target cohorts. To imitate real life application, a cumulative approach should be considered to recalculate the cluster centres over each new target domain, and measure how this affects the model adaptability. It could also be explored how the adapted model performance may change on the original source domain.

The power of this approach depends to some degree on how much the disease space is covered by the disease variation in the source data. If we are confident our source model has seen a particular disease variation before, we could be far more aggressive in shifting features, and similarly less aggressive for outliers, introducing some sort of weighted outlier detection approach.

5.2. Conclusion

We propose a novel method that uses graph node features and source cluster centres in a Cluster Triplet Loss function for UDA of a histology deep learning model. Our approach allows for local domain adaptation within the WSI so that different tissue sections in one target image do not have to be 'shifted' by the same amount.

Whilst our proposed method is not entirely source-free, we require only a dense representation of the original source data, which avoids having to store the memory intensive source dataset and would preserve patient data anonymity if implemented in different hospital settings. This method is generalisable across any number of outcome classes and can be applied to multiple different deep learning and MIL approaches.

5.3. Acknowledgements

The Grampian and Aristotle datasets used in this study were available through the S:CORT consortium. The S:CORT consortium was reviewed and approved by the South Cambs Research Ethics committee (REC ref 15/EE/0241; IRAS reference 169363). The Stratification in Colorectal Cancer Consortium (S:CORT) was funded by the Medical Research Council and Cancer Research UK (MR/M016587/1). The Aristotle trial was funded by Cancer Research UK (CRUK/08/032). The funders played no role in the analyses performed or the results presented. This work was supported by the Research Fund of the Paracelsus Medical University Salzburg, Austria (PMU-FFF R-17/03/090-HUW). Financial support: RW - EPSRC Center for Doctoral Training in Health Data Science (EP/S02428X/1), Oxford CRUK Cancer Centre; VHK - Promedica Foundation (F-87701-41-01) and Swiss National Science Foundation (P2SKP3_168322/1, P2SKP3_168322/2); TSM - S:CORT (see above); JR - Oxford NIHR National Oxford Biomedical Research Centre and the PathLAKE consortium (InnovateUK). The computational aspects of this research were funded from the NIHR Oxford BRC with additional support from the Wellcome Trust Core Award Number 203141/Z/16/Z. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

[1] Christian Abbet, Linda Studer, Andreas Fischer, Heather Dawson, Inti Zlobec, Behzad Bozorgtabar, and Jean-Philippe Thiran. Self-rule to multi-adapt: Generalized multi-source feature learning using unsupervised domain adaptation for colorectal cancer tissue detection. *Medical Image Analysis*, 79:102473, 2022. 3

[2] Alaa Alnissany and Yazan Dayoub. Modified centroid triplet loss for person re-identification. *Journal of Big Data*, 10(74), 2023. 2

[3] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *British Machine Vision Conference*, 2016. 5

[4] T. Caliński and J Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. 4

[5] John M. Carethers and Chyke A. Doubeni. Causes of socioeconomic disparities in colorectal cancer and intervention framework and strategies. *Gastroenterology*, 158(2):354–367, 2020. Colorectal Cancer: Recent Advances Future Challenges. 1

[6] Kuo-Sheng Cheng, Qiong-Wen Zhang, Hung-Wen Tsai, Nien-tsu Li, and Pau-Choo Chung. Domain-centroid-guided progressive teacher-based knowledge distillation for source-free domain adaptation of histopathological images. *IEEE Transactions on Artificial Intelligence*, pages 1–14, 2023. 3

[7] David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979. 4

[8] William Dee, Rana Alaaeldin Ibrahim, and Eirini Marouli. Histopathological domain adaptation with generative adversarial networks bridging the domain gap between thyroid cancer histopathology datasets. *bioRxiv*, 2023. 3

[9] Kianoush Falahkheirkhah, Alex Xijie Lu, David Alvarez-Melis, and Grace Huynh. Domain adaptation using optimal transport for invariant learning using histopathology datasets. In *Medical Imaging with Deep Learning*, pages 1765–1782. PMLR, 2024. 3

[10] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. Triplet-center loss for multi-view 3d object retrieval. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1945–1954, 2018. 2, 7, 1

[11] Yue Huang, Han Zheng, Chi Liu, Xinghao Ding, and Gustavo K. Rohde. Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images. *IEEE Journal of Biomedical and Health Informatics*, 21(6):1625–1632, 2017. 3

[12] Mostafa Jahanifar, Manahil Raza, Kesi Xu, Trinh Vuong, Rob Jewsbury, Adam Shephard, Neda Zamanitajeddin, Jin Tae Kwak, Shan E Ahmed Raza, Fayyaz Minhas, and Nasir Rajpoot. Domain generalization in computational pathology: Survey and guidelines, 2023. 1

[13] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9619–9628, 2021. 2

[14] Maxime W. Lafarge, Josien P. W. Pluim, Koen A. J. Eppenhof, and Mitko Veta. Learning domain-invariant representations of histological images. *Frontiers in Medicine*, 6, 2019. 3, 8

[15] Miguel Lagunes-Fortiz, Dima Damen, and Walterio Mayol-Cuevas. Centroids triplet network and temporally-consistent embeddings for in-situ object recognition. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10796–10802, 2020. 2, 3

[16] Issam H. Laradji and Reza Babanezhad. M-ADDA: unsupervised domain adaptation with deep metric learning. *CoRR*, abs/1807.02552, 2018. 2

[17] Xiangning Li, Chen Pan, Lingmin He, and Xinyu Li. Unsupervised domain adaptation for cross-domain histopathology image classification. *Multimedia Tools and Applications*, 83: 1–21, 2023. 3

[18] Yundong Li, Longxia Guo, and Yizheng Ge. Pseudo labels for unsupervised domain adaptation: A review. *Electronics*, 12(15), 2023. 2

[19] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020. 2, 7

[20] Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110, 2009. 3

[21] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. 6

[22] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. *CoRR*, abs/1709.10190, 2017. 2

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 4

[24] Qi Qi, Xin Lin, Chaoqi Chen, Weiping Xie, Yue Huang, Xinghao Ding, Xiaoqing Liu, and Yizhou Yu. Curriculum feature alignment domain adaptation for epithelium-stroma classification in histopathological images. *IEEE Journal of Biomedical and Health Informatics*, 25(4):1163–1172, 2021. 3

[25] Geetank Raipuria, Anu Shrivastava, and Nitin Singhal. Stain-aglr: Stain agnostic learning for computational histopathology using domain consistency and stain regeneration loss. In *Domain Adaptation and Representation Transfer: 4th MICCAI Workshop, DART 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, page 33–44. Springer-Verlag, 2022. 3

[26] Jian Ren, Ilker Hacihaliloglu, Eric A. Singer, David J. Foran, and Xin Qi. Unsupervised domain adaptation for classification of histopathology whole-slide images. *Frontiers in Bioengineering and Biotechnology*, 7, 2019. 3

[27] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. 4

[28] Milad Sikaroudi, Shahryar Rahnamayan, and H. R. Tizhoosh. Hospital-agnostic image representation learning in digital pathology, 2022. 3

[29] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundstrom. Measuring domain shift for deep learning in histopathology. *IEEE Journal of Biomedical and Health Informatics*, PP:1–1, 2020. 1

[30] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8722–8732, 2020. 2, 7, 1

[31] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Transactions on Medical Imaging*, 35(8):1962–1971, 2016. 3, 7, 8

[32] Guillaume Vray, Devavrat Tomar, Behzad Bozorgtabar, and Jean-Philippe Thiran. Distill-soda: Distilling self-supervised vision transformer for source-free open-set domain adaptation in computational pathology. *IEEE Transactions on Medical Imaging*, pages 1–1, 2024. 3, 7, 1

[33] Xiaodong Wang and Feng Liu. Triplet loss guided adversarial domain adaptation for bearing fault diagnosis. *Sensors*, 20(1), 2020. 2

[34] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022. 4, 6

[35] Xiaoshun Wang, Yunhan Li, and Xiangliang Zhang. Improved triplet loss for domain adaptation. *IET Computer Vision*, 18(1):84–96, 2024. 2, 3

[36] Zhi Wang, Kai Fan, Xiaoya Zhu, Honglei Liu, Gang Meng, Minghui Wang, and Ao Li. Cross-domain nuclei detection in histopathology images using graph-based nuclei feature alignment. *IEEE Journal of Biomedical and Health Informatics*, 28(1):78–88, 2024. 3

[37] Mikołaj Wieczorek, Barbara Rychalska, and Jacek Dabrowski. On the unreasonable effectiveness of centroids in image retrieval. In *Neural Information Processing*, pages 212–223, Cham, 2021. Springer International Publishing. 2, 3

[38] Ruby Wood, Enric Domingo, Korsuk Sirinukunwattana, Maxime W. Lafarge, Viktor H. Koelzer, Timothy S. Maughan, and Jens Rittscher. Joint prediction of response to therapy, molecular traits, and spatial organisation in colorectal cancer biopsies. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 758–767, Cham, 2023. Springer Nature Switzerland. 3, 4

[39] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *CoRR*, abs/1810.00826, 2018. 4

[40] Shiqi Yang, Yaxing Wang, Kai Wang, Shangling Jui, and Joost van de Weijer. Attracting and dispersing: A simple approach for source-free domain adaptation. In *Advances in Neural Information Processing Systems*, pages 5802–5815. Curran Associates, Inc., 2022. 2

[41] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. *CoRR*, abs/2101.10979, 2021. 2

[42] Yexun Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. Domain-invariant adversarial learning for unsupervised domain adaption. *CoRR*, abs/1811.12751, 2018.

[43] Ziyi Zhang, Weikai Chen, Hui Cheng, Zhen Li, Siyuan Li, Liang Lin, and Guanbin Li. Divide and contrast: Source-free domain adaptation via adaptive contrastive learning. In *Advances in Neural Information Processing Systems*, pages 5137–5149. Curran Associates, Inc., 2022. 2

[44] Huihui Zhou, Yan Wang, Benyan Zhang, Chunhua Zhou, Maxim S. Vonsky, Lubov B. Mitrofanova, Duowu Zou, and Qingli Li. Unsupervised domain adaptation for histopathology image segmentation with incomplete labels. *Computers in Biology and Medicine*, 171:108226, 2024. 3

[45] Meng Zhou, Zhe Xu, and Raymond Kai yu Tong. Superpixel-guided class-level denoising for unsupervised domain adaptive fundus image segmentation without source data. *Computers in Biology and Medicine*, 162:107061, 2023. 2