Medical Image Analysis 97 (2024) 103261

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

ALFREDO: Active Learning with FeatuRe disEntangelement and DOmain adaptation for medical image classification

Dwarikanath Mahapatra [a,c,*], Ruwan Tennakoon [b], Yasmeen George [c], Sudipta Roy [d], Behzad Bozorgtabar [e], Zongyuan Ge [c], Mauricio Reyes [f,g]

[a] Inception Institute of AI, Abu Dhabi, United Arab Emirates
[b] School of Computing Technologies, RMIT University, Melbourne, Australia
[c] Faculty of IT, Monash University, Melbourne, Australia
[d] Jio Institute, Navi Mumbai, India
[e] Lausanne University Hospital (CHUV), Lausanne, Switzerland
[f] ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland
[g] Department of Radiation Oncology, University Hospital Bern, University of Bern, Switzerland

ARTICLE INFO

Keywords:
Active learning
Domain adaptation
Feature disentanglement
X-ray
Histopathology

ABSTRACT

State-of-the-art deep learning models often fail to generalize in the presence of distribution shifts between training (source) data and test (target) data. Domain adaptation methods are designed to address this issue using labeled samples (supervised domain adaptation) or unlabeled samples (unsupervised domain adaptation). Active learning is a method to select informative samples to obtain maximum performance from minimum annotations. Selecting informative target domain samples can improve model performance and robustness, and reduce data demands. This paper proposes a novel pipeline called **ALFREDO** (**A**ctive **L**earning with **F**eatu**R**e dis**E**ntangelement and **DO**main adaptation) that performs active learning under domain shift. We propose a novel feature disentanglement approach to decompose image features into domain specific and task specific components. Domain specific components refer to those features that provide source specific information, e.g., scanners, vendors or hospitals. Task specific components are discriminative features for classification, segmentation or other tasks. Thereafter we define multiple novel cost functions that identify informative samples under domain shift. We test our proposed method for medical image classification using one histopathology dataset and *two* chest X-ray datasets. Experiments show our method achieves state-of-the-art results compared to other domain adaptation methods, as well as state of the art active domain adaptation methods.

1. Introduction

Deep neural networks (DNNs) demonstrate state-of-the-art (SOTA) results for many medical image analysis applications. Although they excel at learning from large labeled datasets, it is challenging for DNNs to generalize the learnt knowledge to new target domains (Saenko et al., 2010; Torralba and Efros, 2011). This limits their real-world utility, as it is impractical to collect large datasets for every novel application with the aim of retraining the network. Although, in a hypothetical situation, organizations may have a large budget for data annotation, it is impractical to annotate every data point and include them as part of the training set since annotating medical images requires high clinical expertise. In a time-efficient active learning setup the goal would be to automate the annotations as much as possible while reducing manual annotations needed to improve the model. However, this does not eliminate the need for annotations in the target domains. Rather, a more realistic paradigm is where radiologists will perform monitoring and corrections of AI-based results, as opposed to annotating every case without any assistance. This will significantly reduce annotations.

The annotation problem is further exacerbated when annotating images from a different domain (or the target class). Considering a supervised domain adaptation setting, all available samples from the target class are not equally informative and annotating random samples may result in a waste of time and effort. Under these circumstances, it makes sense to select most informative target domain samples for labeling as it can reduce annotation cost and training time, and also improve efficiency of model training. Additionally, in an unsupervised domain adaptation setting where sample annotation is not required, informative sample selection identifies the most important samples

* Corresponding author.
E-mail address: dwarikanath.mahapatra@inceptioniai.org (D. Mahapatra).

https://doi.org/10.1016/j.media.2024.103261
Received 25 October 2023; Received in revised form 5 June 2024; Accepted 26 June 2024
Available online 4 July 2024

which are used for model training leading to improved performance compared to conventional approaches.

Active Learning (AL) methods enable an expert to select informative samples and add them to a training set for incremental training (Li and Guo, 2013). This allows a model to obtain high performance with minimal labeled samples (i.e., high learning rates) and is particularly suitable for medical image analysis tasks where AL methods must adapt to varying conditions like device vendor, imaging protocol, machine learning model, etc. While conventional AL methods (Ash et al., 2019; Ducoffe and Precioso, 2018; Sener and Savarese, 2018) have extensively studied the problem of identifying informative instances for labeling, they typically choose samples from the same domain and hence do not account for domain shift. Thus conventional AL models are not very effective for domain adaptation applications. In many practical scenarios, models are trained on a source domain and deployed in a different target domain. This can be a common occurrence for medical image analysis applications where target domain data is scarce, or as a result of different image capturing protocols, parameters, devices, scanner manufacturers, etc. Domain shift is also observed when images of the dataset are from multiple facilities. Consequently, domain adaptation techniques have to be used for such scenarios (Ganin and Lempitsky, 2015; Hoffman et al., 2018; Saenko et al., 2010).

Despite the existence of domain adaptation methods, the primary challenge of obtaining labeled data is accessing domain experts and annotating data in reasonable time. In this scenario it is appropriate that we make optimal use of the experts' time and use AL to obtain maximal information from minimal annotations. Hence it is beneficial to have a technique that can choose informative samples despite the observed domain shift. This will lead to different trained models being adapted for a wide variety of tasks.

In this work, we study the problem of active learning under such a domain shift, called Active Domain Adaptation (Kirsch et al., 2019) (ADA). Given (i) labeled data in a source domain, (ii) unlabeled data in a target domain, and (iii) the ability to obtain labels for a fixed budget of target instances, the goal of ADA is to select target instances for labeling and learn a model with high accuracy on the target test set. We specifically apply our method to medical imaging datasets since domain shift is a common problem for medical image computing tasks.

2. Prior work

2.1. Domain adaptation in medical image analysis

Domain adaptation (DA) has attracted increasing attention of researchers, and has emerged as an important research topic in machine learning based medical image analysis (Ghafoorian et al., 2017; Raghu et al., 2019; Kamnitsas et al., 2017b). In a practical use case, a reliable DA model trained to, for instance, segment cardiac structures from MR images can be also used on cardiac CT images (Zhuang and Shen, 2016). Another useful case of DA is stain normalization of histopathology images where models trained with images from one hospital can be used to analyze images from a different one (Bandi et al., 2019; Mahapatra et al., 2022). In DA, it is assumed that the domain feature spaces and tasks remain the same, i.e, the set of labels is the same for source and target domain tasks, while the marginal distributions (e.g., source of data collection or modality) are different.

DA can be categorized into different types based on clinical scenarios, constraints and algorithms. The survey paper of Guan and Liu (2021) categorizes DA methods under 6 types. However, since our work focuses on supervised and unsupervised DA, we briefly review related works and refer the reader to Guan and Liu (2021) for more details. In terms of target domain label availability, existing DA methods can be divided into supervised DA (SDA), semi-supervised DA (sSDA), and unsupervised DA (UDA). In SDA, a small number of labeled data in the target domain are available for model training while sSDA has some labeled data and lots of unlabeled target domain data. In UDA, only unlabeled target data are available.

2.1.1. Supervised domain adaptation

One of the first SDA methods for medical images (Kumar et al., 2017) used ResNet as the feature extractor and applied to mammography images. They evaluate three shallow domain adaptation methods and provide empirical results. Huang et al. (2017) propose to use LeNet-5 to extract features of histological images from different domains for epithelium-stroma classification, project them onto a subspace (via PCA) and align them for adaptation to the target domain. Ghafoorian et al. (2017) evaluate the impact of fine-tuning strategies on brain lesion segmentation, by using CNN models pre-trained on brain MRI scans. Their experimental results reveal that using only a small number of target training examples for fine-tuning can improve the transferability of models. They further evaluate the influence of the training set size and different network architectures on the adaptation performance. Based on similar findings, numerous methods have been proposed to leverage CNNs that are well pre-trained on ImageNet to tackle medical image analysis problems.

2.1.2. Unsupervised domain adaptation

UDA for medical image analysis has gained significance in recent years since it does not require labeled target domain data. Prior works in UDA focused on medical image classification (Ahn et al., 2020), object localization, lesion segmentation (Heimann et al., 2013; Kamnitsas et al., 2017a), and histopathology stain normalization (Chang et al., 2021). Heimann et al. (2013) used GANs to increase the size of training data and demonstrated improved localization in X-ray fluoroscopy images. Likewise, Kamnitsas et al. (2017a) used GANs for improved lesion segmentation in magnetic resonance imaging (MRI). Ahn et al. (2020) use a hierarchical unsupervised feature extractor to reduce reliance on annotated training data. Chang et al. (2021) propose a novel stain mix-up for histopathology stain normalization and subsequent UDA for classification. Graph networks for UDA (Ma et al., 2019; Wu et al., 2020) have been used in medical imaging applications (Ahmedt-Aristizabal et al., 2021) such as brain surface segmentation (Gopinath et al., 2020) and brain image classification (Hong et al., 2019a,b). However, none of them explore DA from the active learning perspective, i.e., they do not investigate if DA can be equally effective by identifying informative samples from the target domain and minimizing the need for annotated samples. Mahapatra and Ge (2020) achieve training data independent image registration using generative adversarial networks and domain adaptation.

Recent works on UDA for medical image analysis also include super resolution of wireless capsule endoscopy images (Liu et al., 2023), source-free approaches for segmentation using prototypes and contrastive learning (Yu et al., 2023), brain tumor segmentation (Alefsen et al., 2023), spectral adversarial mixup for few shot UDA (Zhang et al., 2023), anatomical landmark detection (Jin et al., 2023), and automated quality assessment of transoesophageal echocardiography images (Xu et al., 2023a). Xu et al. (2023b) propose a UDA framework based on appearance and structure consistency for segmentation by constraining the consistency before and after a frequency-based image transformation. Ghamsarian et al. (2023) propose a semi-supervised learning strategy for domain adaptation termed transformation-invariant self-training (TI-ST). The method assesses pixel-wise pseudo-labels' reliability and filters out unreliable detections during self-training. Lin et al. (2023) propose a multi-target domain adaptation through implicit feature adaptation and prompt learning for medical image segmentation.

2.2. Active learning in medical image analysis

Informative sample selection techniques are key to active learning frameworks and an excellent survey of active learning related to medical image analysis can be found in Budd et al. (2021) and Wang et al. (2024) . Different sample selection approaches have been investigated for deep learning based medical image analysis, including sample entropy (Zhu and Bento, 2017), model uncertainty (Mahapatra

et al., 2018; Gal et al., 2017), Fisher information (Sourati et al., 2019), visual saliency (Mahapatra and Buhmann, 2016) and clustering-based sample selection (Zheng et al., 2019). Wang et al. (2017a) use sample entropy, and margin sampling to select informative samples while Zhou et al. (2016) use GANs to synthesize samples close to the decision boundary and annotate it by human experts. Mayer and Timofte (2018) use GANs to generate high entropy samples, which are used as a proxy to find the most similar samples from a pool of real annotated samples.

Yang et al. (2017) propose a two-step sample selection approach based on uncertainty estimation and maximum set coverage similarity metric. Test-time Monte-Carlo dropout (Gal et al., 2017) has been used to estimate sample uncertainty, and consequently select the most informative ones for label annotation (Gal et al., 2017; Bozorgtabar et al., 2019). The state-of-the-art in active learning is mostly dominated by methods relying on uncertainty estimations. However, the reliability of uncertainty estimations has been questioned for deep neural networks used in computer vision and medical imaging applications due to model calibration issues (Abdar et al., 2021; Jungo et al., 2020). Recent work (Budd et al., 2021; Mahapatra et al., 2021) has highlighted the importance of interpretability in active learning setups. Interpretability has been shown to improve a AL method's ability to select informative samples leading to greater performance gain with fewer annotations (Mahapatra et al., 2021, 2023).

2.3. Active domain adaptation

ADA can be cost-effective solution when the quantity or cost of labeling in the target domain is prohibitive. Despite its practical utility, ADA is challenging and has seen limited exploration since its introduction (Chattopadhyay et al., 2013; Kirsch et al., 2019). Kirsch et al. (2019) first applied ADA to sentiment classification from text data by sampling instances based on model uncertainty and a learned domain separator. Chattopadhyay et al. (2013) select target instances and learn importance weights for source points through a convex optimization problem. However no ADA methods have been proposed for medical image analysis.

In a traditional AL setting informative sample selection does not focus on addressing domain shift. Thus, AL methods based on uncertainty or diversity sampling are less effective for ADA. Uncertainty sampling selects instances that are highly uncertain under the model's beliefs (Gal et al., 2017), which under a domain shift leads to miscalibration and selection of uninformative, outlier, or redundant samples for expert labeling (Ovadia et al., 2019).

AL based on diversity sampling selects instances dissimilar to one another (Gissin and Shalev-Shwartz, 2019; Sener and Savarese, 2018; Sinha et al., 2019). In ADA this can lead to sampling uninformative instances from regions of the feature space that are already well-aligned across domains (Prabhu et al., 2021). While using uncertainty or diversity sampling exclusively is suboptimal for ADA, their combination can be very effective as shown by Su et al. (2020a). Ash et al. (2019) perform clustering in a hallucinated "gradient embedding" space, but rely on distance-based clustering in high-dimensional spaces, which often leads to suboptimal results. Prabhu et al. (2021) propose a label acquisition strategy, termed as Clustering Uncertainty-weighted Embeddings (CLUE), for ADA that combines uncertainty and diversity sampling without the need for complex gradient or domain discriminator-based diversity measures.

2.4. Our contributions

While domain adaptation and active learning have been well studied in medical image analysis, their combination has not been explored. Our work is one of the first to look at active domain adaptation in the medical image analysis setting. This paper makes the following contributions:

1. We propose one of the first applications of active domain adaptation in medical image analysis, denoted as **ALFREDO** (**A**ctive **L**earning with **F**eatu**R**e dis**E**ntangelement and **DO**main adaptation).
2. We propose a feature disentanglement approach that extracts domain specific and task specific features from a given image. Domain specific components refer to the source specific part, e.g., scanners, vendors or hospitals from where the data originates. Task specific components refer to classification, segmentation or other tasks which are the focus of the method. The combination of these features are used for active learning based sample selection and classification.
3. Using the different feature components we propose novel metrics to quantify the informativeness of samples across different domains. Thus we demonstrate that the novel feature disentanglement components are able to identify informative samples in the presence of domain shift.
4. We evaluate our method's effectiveness by using it for classification of 3 publicly available medical imaging datasets. We also benchmark our method against multiple domain adaptation and active domain adaptation methods used for computer vision applications.

3. Method

We aim to tackle the problem of active domain adaptation and show its applicability to both supervised domain adaptation (SDA) and unsupervised domain adaptation (UDA). While active SDA (ASDA) requires selecting the target samples to be labeled, in active UDA (AUDA) we select the unlabeled target samples that will go into the pool for training along with labeled samples from the source domain. Recently, UDA has grown in prominence for the medical image analysis field because of the large numbers of medical images being acquired, and the ensuing difficulty in annotating them due to shortage of experts. Both ASDA and AUDA are related tasks requiring the selection of informative samples, and hence can be solved within a common framework. Although the individual components of ADA – addressing domain shift and informative sample selection – have been explored in detail, their combination presents a different challenge. In prior work much effort has gone into exploring properties such as transferable diversity, uncertainty, etc, wherein the common criteria for informative sample selection is adapted to ensure it does equally well for samples from a new domain. In our approach we propose to use feature disentanglement to extract different types of features from the labeled source data and unlabeled target data such that samples with the same label are projected to the same region of the new feature space. They are used to devise a domain agnostic sample informativeness metric.

Prior work on feature disentanglement for domain adaptation decompose the latent feature vector into domain specific and domain agnostic features and model training requires aligning the domain agnostic features. This helps to learn a set of features that are consistent across different domains and ensures that the learned model has similar performance for different datasets. Conventional approaches to feature disentanglement is sub-optimal for ADA because : (1) the domain agnostic features are usually segmentation maps that encode the structural information (Park et al., 2020). The structural information provided by segmentation maps is constant across data from different domains having a shared label space. However selecting informative images from a different domain on the basis of structural segmentation map information can be challenging due to different appearances and field of views captured by the target domain images. (2) Domain specific features usually encode information such as texture, intensity distributions, etc. Domain specific features of one domain are not generally useful in selecting informative samples from a different domain.

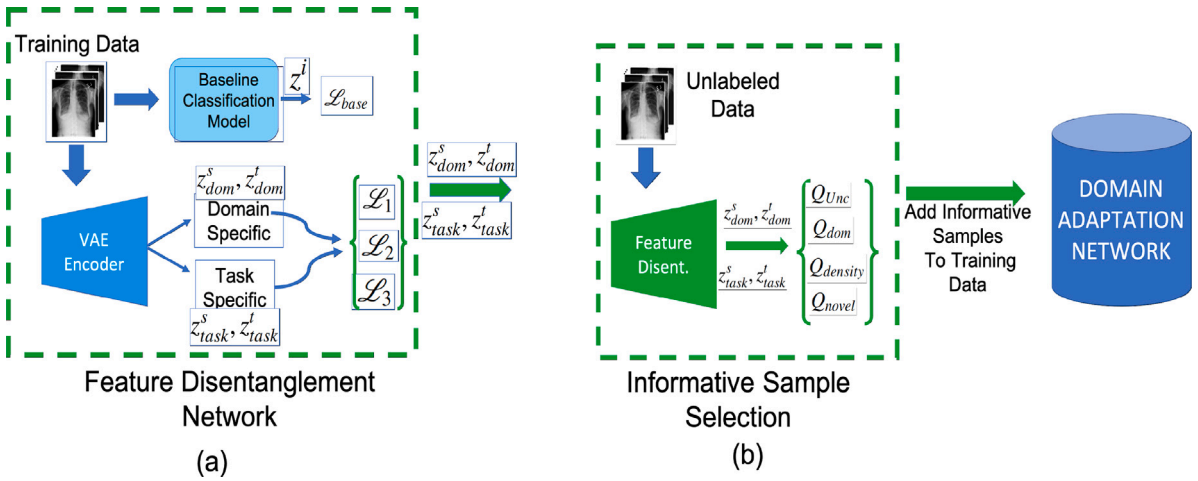

Fig. 1. Workflow of the proposed method. (a) **Feature Disentanglement:** Training data goes through an autoencoder to obtain different components z_{dom}, z_{task} and they are used to calculate different loss functions. After training is complete we get the domain and task specific features. (b) For **informative sample selection** we obtain disentangled feature representations of the unlabeled data and calculate the informativeness score of each sample in the batch. Thereafter the most informative samples are labeled (for supervised domain adaptation) added to the labeled samples to initiate the domain adaptation steps.

In our approach we use feature disentanglement to obtain a set of features that are consistent across different domains, and at the same time give high performance (e.g. classification accuracy) . Given source and target domains S and T, we hypothesize that an ideal domain agnostic feature is one which will give classification accuracy on domain S that is close to those obtained using the original images' features. At the same time this feature should be similar for samples having same labels from domains S, T. The resulting feature space can be used to train a classifier on domain S, and use principles of active learning to select informative samples from domain T by operating on this new feature space. Such an approach allows conventional active learning techniques to be used with minor modifications without having to factor in transferrability.

In Active DA, the learning algorithm has access to labeled instances from the source domain (X_S, Y_S), unlabeled instances from the target domain X_{UT}, and a budget B which is much smaller than the amount of unlabeled target data. The learning algorithm may query an oracle to obtain labels for at most B instances from X_{UT}, and add them to the set of labeled target instances X_{LT}. The entire target domain data is $X_T = X_{LT} \bigcup X_{UT}$. The task is to learn a function $h : X \longrightarrow Y$ (a convolutional neural network (CNN) parameterized by θ) that achieves good predictive performance on the target domain. The samples $x_S \in X_S$ and $x_T \in X_T$ are images, and labels $y_S \in Y_S, y_T \in Y_T$ are categorical variables $y \in 1, 2, \ldots, C$.

3.1. Feature disentanglement network

The primary challenge of domain shift is the inability of learned source domain features to transfer to the target domain. As a result we are unable to replicate the source domain performance on target domain data. An often used approach is to map data from both domains to a common feature space such that a model trained on one domain can perform well on another domain. The MMD (maximum man discrepancy) method (Bermudez-Chacon et al., 2018) was one of the first to take advantage of such a scenario. However, MMD-like approaches present the following challenges for active domain adaptation: (1) It does not ensure that the new feature space provides optimum results for the source domain data in its original feature space. (2) Due to the sub-optimal nature of the feature space the informative samples selected may not be ideal for active learning.

Fig. 1 shows the workflow of our proposed method. The feature disentanglement network (FDN) (Fig. 2) consists of an autoencoder each for source and target domains. The FDN consists of two encoders

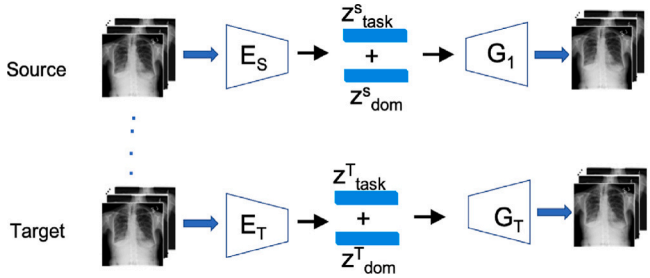

Fig. 2. Architecture of feature disentanglement network. Given training images from different classes of the same domain, we disentangle features into domain-specific and task-specific features using autoencoders. The different feature components are used to define the different loss terms.

$(E_S(\cdot), E_T(\cdot))$ and two decoder networks $(G_S(\cdot), G_T(\cdot))$, for the source and target domains respectively. Similar to a classic autoencoder, each encoder, $E_\bullet(\cdot)$, produces a latent code z_i for image $x_i^\bullet \sim p_\bullet$. Each decoder, $G_\bullet(\cdot)$, reconstructs the original image from z_i. Furthermore, we divide the latent code, z_i, into two components: a domain specific component, z_{dom}, and a task specific component, z_{task} by having two heads for the latent code.. The disentanglement network is trained using the following loss function:

$$\mathscr{L}_{Disent} = \mathscr{L}_{Rec} + \lambda_1 \mathscr{L}_1 + \lambda_2 \mathscr{L}_2 + \lambda_3 \mathscr{L}_3 + \lambda_{base} \mathscr{L}_{base} \quad (1)$$

$\mathscr{L}_{Rec}$, is the commonly used image reconstruction loss and is defined as:

$$\mathscr{L}_{Rec} = \mathbb{E}_{x_i^S \sim p_S} \left[\left\| x_i^S - G_S(E_S(x_i^S)) \right\| \right] + \mathbb{E}_{x_j^T \sim p_T} \left[\left\| x_j^T - G_T(E_T(x_j^T)) \right\| \right]. \quad (2)$$

The input data consists of images from the source and target domains (Fig. 1(a)). The disentangled features from both domains are denoted as z^s_{dom}, z^s_{task} for the source domain and as z^t_{dom}, z^t_{task} for the target domain. z^s_{dom}, z^s_{task} are then combined and input to the source decoder G_S to reconstruct the original source domain image, while z^t_{dom}, z^t_{task} are combined and given as input into the source decoder G_T to reconstruct the original target domain image.

Since domain-specific features encode information unique to the domain, they will be different for source and target domains. Hence, the semantic similarity between z^t_{dom} and z^s_{dom} will be low. This is captured using the following loss term:

$$\mathscr{L}_1 = \langle z^s_{dom}, z^t_{dom} \rangle. \quad (3)$$

Where $\langle\cdot\rangle$ denotes the cosine similarity of the two feature vectors.

Additionally, we expect task-specific features of the two domains to have high similarity because they solve the same task of identifying the common labels of different domain images. This is incorporated using the following loss term

$$\mathscr{L}_2 = 1 - \langle z^s_{task}, z^t_{task}\rangle \tag{4}$$

We want the two components of the features to be as dissimilar as possible in order to capture mutually complementary information, and this is achieved using the following loss term

$$\mathscr{L}_3 = \langle z^s_{dom}, z^s_{task}\rangle + \langle z^t_{dom}, z^t_{task}\rangle \tag{5}$$

To ensure that performance is not affected when extracting task specific features we enforce their similarity with features extracted without the task-specific constraint. While using the task specific features to train a classifier it should give performance levels similar to those obtained using the original image features. We first train a baseline classification model M_{base} using the original image features of source domain data (since labels are available only for the source domain data). For a given sample i from the source domain, we pass it through M_{base} to obtain a feature vector z^{base}. The task-specific component is denoted as $z^{s,i}_{task}$. Their cosine similarity should be high to ensure it captures highly relevant information. The corresponding loss term is

$$\mathscr{L}_{base} = 1 - \langle z^{s,i}_{task}, z^{base}\rangle \tag{6}$$

Note that the feature vector z^{base} is obtained from a pre-trained classifier and can be considered as the optimal feature vector (depending upon the chosen classifier model M_{base}) whereas the task specific feature vector $z^{s,i}_{task}$ is obtained as part of the training process. Our objective is to ensure that $z^{s,i}_{task}$ is very close to z^{base} in terms of semantic similarity. $\mathscr{L}_{base}$ is named as such to denote its comparison with M_{base}. We use a DenseNet-121 as M_{base}.

3.2. Informative sample selection

Having trained a feature disentanglement network and after obtaining the different sets of features, we train two classifiers, M^{source}_{task} on z^s_{task}, and M^{source}_{dom} on z^s_{dom}. The source domain features z^s_{task} have high similarity with the corresponding target domain task specific feature set z^t_{task}. As a result this ensures that the classifier M^{source}_{task} trained with source domain features z^s_{task} can be used with z^t_{task} to obtain similar performance levels, and make it easier to identify informative samples. We define multiple score functions to quantify the informativeness based on multiple criteria and the final informativeness score of a sample is the combination of these scores.

We use the following criteria to choose informative samples: (1) **Uncertainty:** We take model M^{source}_{task} trained on the source domain features and use it to calculate the uncertainty of the target domain samples using the task specific features z^t_{task}. Since the source domain and target domain features are similar, the model M^{source}_{task} can reliably determine the most uncertain samples. To measure informativeness we use predictive entropy $\mathscr{H}(Y|x)$ (Wang and Shang, 2014) which for C-way classification, is defined as:

$$Q_{Unc} = \mathscr{H}(Y|x) = -\sum_{c=1}^{C} p_\theta(Y=c|x)\log p_\theta(Y=c|x) \tag{7}$$

(2) **Domainness:** determines whether a sample is from the same domain as a reference sample (e.g., source domain) or belongs to a different domain. Recent active learning or ADA methods (Huang et al., 2018; Su et al., 2020b) consider samples with higher distinctiveness from source domain samples as informative since they capture the unique characteristics in the target domain. However, outliers exist in the target domain, which are not informative for target classification. In the original feature space with domain shift, both normal target samples, different from the source domain, and target outliers are far from the source domain and there is no clear way to exclude outliers.

For a given target domain sample we obtain its disentangled feature representations z^t_{task}, z^t_{dom}. The domain specific features z^t_{dom} are compared with the domain specific features of source domain data z^s_{dom} of each label. If the cosine similarity is below a threshold then the sample is determined as different from the source domain data, and hence not considered for labeling. Fu et al. (2021) show that very low similarity scores between source and target domain samples denotes outliers. On the other hand high similarity scores indicate uninformative samples since the target domain sample has already been included within the training set. Hence we consider a lower threshold η_1 for the cosine similarity below which the sample is considered an outlier and an upper threshold η_2 above which the sample is considered as uninformative. Thus we define a score

$$Q_{dom} = \begin{cases} 0 & \text{if } \langle z^s_{dom}, z^t_{dom}\rangle < \eta_1 \\ \langle z^s_{dom}, z^t_{dom}\rangle & \text{if } \eta_1 \le \langle z^s_{dom}, z^t_{dom}\rangle \le \eta_2 \\ 0 & \text{if } \langle z^s_{dom}, z^t_{dom}\rangle > \eta_2 \end{cases} \tag{8}$$

To set the thresholds η_1 and η_2 we plot a distribution of the cosine similarity values and η_1 equals the 30th percentile value while η_2 corresponds to the 75th percentile.

(3) **Density:** determines whether a sample represents other samples which are similar in the feature space. One way to reduce the number of annotations is to choose samples which are representative of many other samples. If a sample lies in a dense region of the feature space then it is representative of many other samples. We cluster the target domain samples into N clusters using the task specific features z^t_{task} where N is the number of classes of the source domain data. For each sample we calculate the feature similarity with respect to other samples in the batch, and take the average similarity over the top K closest samples. A higher average feature similarity indicates that the sample is more similar to other samples and is in a dense region of the feature space. By obtaining the label of one such sample we, in effect, obtain the labels of more samples. Thus we define a density score as this average feature similarity:

$$Q_{density} = \frac{1}{K}\sum_{k=1,\ldots,K} \langle z^i_{task}, z^k_{task}\rangle \tag{9}$$

In our experiments we set $K = 20$.

(4) **Novelty:** This criterion checks whether the target sample being selected for labeling is similar or different to previously selected target samples. The similarity with respect to previously selected samples can be quantified in different ways - e.g., distance in feature space, or location of cluster, etc. However we find that the similarity of the samples based on z_{task} is the best criteria. For a given pair of feature vectors their similarity depends a lot on the task in hand. If the downstream task changes then their similarity will also change. Hence for a given target domain sample i with feature vector z^i_{task} we calculate its cosine similarity with previously annotated samples z^j_{task}. If the similarity is high then the informativeness score of sample i is low and vice-versa. Thus we define a "novelty-score" as

$$Q_{novel} = \sum_j 1 - \langle z^i_{task}, z^j_{task}\rangle \tag{10}$$

The cosine similarities of i with other previously annotated samples j are summed to get the "novelty-score". The final informativeness score of a sample is the sum of all the above scores and is defined as

$$Q_{Inf} = \lambda_{Unc}Q_{Unc} + \lambda_{Dom}Q_{Dom} + \lambda_{Density}Q_{Density} + \lambda_{Novel}Q_{Novel} \tag{11}$$

Higher values of Q_{Inf} indicates greater informativeness. The top N informative samples are selected in every batch and added to the training set, and the classifier is updated. Informative sample selection continues till there is no further change in the performance of a separate validation set. The different stages of our method is summarized in Algorithm 1.

Algorithm 1 ALFREDO

Require: Pretrained model M_0, Feature Disentanglement Network $FDN(\cdot)$, $\mathbb{I}_{validation}$, AUC_{target}

$M \leftarrow M_0$ (trained using 2% training data)
repeat
  $\mathbb{I}_{in} \leftarrow \{I_{in}\}$ ▷ define set of input testing images
  $\mathbb{S}_{in} \leftarrow \{FDN(\mathbb{I}_{in}, M)\}$ ▷ disentangled features given input set and FDN
  $\{scores\}_{in} \leftarrow Q_{inf}(\mathbb{FDN}_{in})$ ▷ calculate informativeness scores using FDN outputs
  $\mathbb{I}_{sort} \leftarrow sort(\mathbb{I}_{in}, \{scores\}_{in})$ ▷ sort $\mathbb{I}_{in}$ in decreasing order by scores
  $\mathbb{I}_{train} \leftarrow \mathbb{I}_{sort}\{i = 1, ..., top\_n\}$ ▷ select top-n ranked samples
  $\mathbb{L}_{train} \leftarrow expert\_query(\mathbb{I}_{train})$ ▷ label querying of selected samples
  $M_{new} \leftarrow train(M, \mathbb{I}_{train}, \mathbb{L}_{train})$ ▷ train new model
until $AUC(M_{new}, \mathbb{I}_{validation}) \geq AUC_{target}$ ▷ Repeat until target AUC is attained or $AUC(M_{new}, \mathbb{I}_{validation})$ does not change
return M_{new}

4. Experimental results

4.1. Baseline methods

We compare our proposed method,denoted as **ALFREDO** (**A**ctive **L**earning with **F**eatu**R**e dis**E**ntangelement and **DO**main adaptation), against several state-of-the art methods for Active DA and Active Learning such as:

1. AADA: Active Adversarial Domain Adaptation (AADA) (Su et al., 2020a) which performs alternate rounds of active sampling and adversarial domain adaptation via Domain-Adversarial Training of Neural Networks (DANN) (Ganin et al., 2016a).
2. Entropy based Uncertainty approach (Unc) (Wang and Shang, 2014): Selects instances for which the model has highest predictive entropy.
3. Batch Active learning by Diverse Gradient Embeddings (BADGE) (Ash et al., 2019): BADGE proposes a state-of-the-art active learning strategy that constructs diverse batches by running KMeans++ on "gradient embeddings" that incorporate model uncertainty and diversity.
4. The CLUE (Clustering Uncertainty-weighted Embeddings) method of Prabhu et al. (2021) - that performs uncertainty-weighted clustering to identify target instances for labeling that are both uncertain under the model and diverse in feature space.
5. (Fu et al., 2021)'s method using transferable uncertainty by combining transferable committee, transferable uncertainty, and transferable domainness.
6. The transformation invariant approach of Ghamsarian et al. (2023) that uses self training for UDA.

4.2. Experimental settings

Since our goal is to demonstrate the effectiveness of our active learning method under domain shift and not to propose a new domain adaptation method, we show results of our method integrated with existing SOTA methods for SDA and UDA. Our end task is to perform classification through active domain adaptation. Given a source domain dataset we train our feature disentanglement method on it. With the provided target domain dataset we use the pre-trained feature disentanglement network to obtain z^t_{task} and z^t_{dom}. We then select most informative samples from the target domain and add to the training set which initially consists of only the source domain samples. After each addition the classifier is updated and evaluated on a separate test set from the target domain.

Annotating samples from the target domain is a classic case of supervised domain adaptation. In the unsupervised setting there are no samples to label. Labeled source domain data and unlabeled target domain data are trained together to minimize the discrepancy in some feature space. We adapt our active domain adaptation method such that instead of using the entire unlabeled dataset, we select informative samples from target domain to use for training the classifier. We adopt the following experimental setup:

1. Train a benchmark method using a network trained in a fully-supervised manner on the training set from the same domain, i.e., training, validation and test data are from the same hospital/dataset. This setting gives the upper-bound performance expectation for a SDA model, and this upper bound depends upon the used network architecture. We refer to this benchmark as FSL−SD (fully supervised learning based method on same domain data).
2. We also train a SOTA domain adaptation method (either supervised DA or unsupervised DA) using the available source and target domain data. Here the entire dataset is used and there is no informative sample selection. We refer to this as the SDA$_{SOTA}$ (SOTA supervised domain adaptation) or UDA$_{SOTA}$ (SOTA unsupervised domain adaptation) methods. Note that for UDA we do not label the samples but add them to the training set.
3. In all cases we find that the SDA$_{SOTA}$ is obtained by taking the FSL−SD network architecture of the source data and finetuning the last layers using the labeled samples of the target domain data.
4. We use our active domain adaptation method, **ALFREDO**, to select informative samples and incrementally add to the training set. We investigate the effectiveness of our active learning based methods in selecting the best possible set of samples for labeling, and explore their degree of success in reducing the required number of annotated samples. As we add samples to the training set we report the test accuracy for every 10% increase of the training set.
5. For the UDA setting informative samples are selected by first disentangling the features and then choosing the most informative samples. We argue that informative sample selection is important in the UDA setting in order to reduce the data annotation requirements and improve model robustness by feeding it high quality data for training. Active learning works clearly establish that using high quality data for training leads to better performance with fewer labeled samples. We test the hypothesis that use of informative samples will lead to better UDA performance, which is supported by the results showed in later sections.
6. We also report the performance of other active learning methods, including domain adaptation based and conventional methods that do not address the domain shift.

4.3. Results on histopathology datasets

Dataset Description: We use the CAMELYON17 dataset (Bandi et al., 2019) to evaluate the performance of the proposed method on tumor/normal classification. In this dataset, a total of 500 $H\&E$ stained WSIs are collected from five medical centers (denoted as $C1, C2, C3, C4, C5$ respectively). A total of 50 of these WSIs include lesion-level annotations. All positive and negative WSIs are randomly split into training/validation/test sets and provided by the organizers in a 50/30/20% split for the individual medical centers to obtain the following split: $C1$:37/22/15, $C2$: 34/20/14, $C3$: 43/24/18, $C4$: 35/20/15, $C5$: 36/20/15. 256 × 256 image patches are extracted from the annotated tumors for positive patches and from tissue regions of WSIs without tumors for negative patches.

Since the images have been taken from different medical centers their appearance varies despite sharing the same disease labels. This

Table 1
Classification results in terms of AUC measures for different domain adaptation methods on the CAMELYON17 dataset. Note: FSL−SD is a fully-supervised model trained on target domain data.

	No ADA	MMD	CycleGAN	Chang et al. (2021) (UDA_{SOTA})	FSL−SD	SDA_{SOTA}
$C1$	0.8068	0.8742	0.9010	0.964	0.976	0.969
$C2$	0.7203	0.6926	0.7173	0.933	0.957	0.941
$C3$	0.7027	0.8711	0.8914	0.931	0.95	0.938
$C4$	0.8289	0.8578	0.8811	0.95	0.971	0.957
$C5$	0.8203	0.7854	0.8102	0.927	0.942	0.933
$Avg.$	0.7758	0.8162	0.8402	0.941	0.959	0.948

Table 2
Active domain adaptation results for Camelyon17 dataset. AUC values for different baselines and proposed approach along with ablation studies.

	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	p-
FSL-SD	0.959	0.959	0.959	0.959	0.959	0.959	0.959	0.959	0.959	0.959	<0.001
Random	0.693	0.71	0.75	0.794	0.821	0.858	0.891	0.914	0.928	0.938	<0.001
Unc	0.706	0.733	0.772	0.812	0.845	0.891	0.922	0.931	0.939	0.943	<0.001
AADA	0.712	0.742	0.791	0.841	0.872	0.903	0.924	0.939	0.945	0.948	0.001
BADGE	0.707	0.728	0.768	0.803	0.847	0.885	0.903	0.924	0.932	0.940	0.005
CLUE	0.715	0.746	0.786	0.839	0.878	0.911	0.921	0.934	0.941	0.947	0.007
Fu et al. (2021)	0.714	0.739	0.775	0.813	0.849	0.883	0.914	0.925	0.935	0.944	0.001
Ghamsarian et al. (2023)	0.721	0.754	0.793	0.825	0.858	0.889	0.914	0.929	0.941	0.95	0.02
$ALFREDO_{ASDA}$	0.73	0.775	0.801	0.831	0.872	0.895	0.927	0.937	0.946	0.964	0.04
$ALFREDO_{AUDA}$	0.721	0.762	0.793	0.828	0.863	0.893	0.915	0.927	0.941	0.952	–
Ablation studies											
$ALFREDO_{No-FeatDisent}$	0.711	0.737	0.777	0.815	0.85	0.895	0.924	0.934	0.941	0.945	<0.001
Feature disentanglement											
$AUDA_{wo\mathscr{L}_1}$	0.702	0.734	0.772	0.842	0.872	0.885	0.896	0.902	0.911	0.921	0.001
$AUDA_{wo\mathscr{L}_2}$	0.711	0.729	0.765	0.802	0.854	0.867	0.881	0.898	0.914	0.928	0.005
$AUDA_{wo\mathscr{L}_3}$	0.692	0.724	0.768	0.813	0.843	0.869	0.884	0.896	0.901	0.914	0.0009
$AUDA_{wo\mathscr{L}_{base}}$	0.671	0.703	0.734	0.771	0.81	0.848	0.866	0.881	0.895	0.908	0.0008
Informative sample selection											
$AUDA_{woQ_{Unc}}$	0.705	0.74	0.778	0.852	0.881	0.898	0.906	0.913	0.924	0.932	0.001
$AUDA_{woQ_{dom}}$	0.691	0.724	0.761	0.812	0.857	0.884	0.898	0.904	0.916	0.923	0.001
$AUDA_{woQ_{density}}$	0.693	0.719	0.753	0.788	0.814	0.861	0.878	0.896	0.908	0.919	0.0001
$AUDA_{woQ_{novel}}$	0.682	0.711	0.746	0.779	0.817	0.856	0.869	0.882	0.897	0.912	0.0001

is due to slightly different protocols of $H\&E$ staining. Stain normalization has been a widely explored topic which aims to standardize the appearance of images across all centers, which is equivalent to domain adaptation. Recent approaches to stain normalization/domain adaptation favor use of GANs and other deep learning methods. We compare our approach to recent approaches and also with Chang et al. (2021) which explicitly performs UDA using MixUp. The method by Chang et al. (2021) is denoted as UDA_{SOTA}

To evaluate our method's performance: (1) We use $C1$ as the source dataset and train a ResNet-101 classifier (He et al., 2016) ($ResNet_{C1}$). Each remaining dataset from the other centers are, separately, taken as the target dataset. We select informative samples add them to training set and update $ResNet_{C1}$. As a baseline, we perform the experiment without domain adaptation denoted as No-ADA where $ResNet_{C1}$ is used to classify images from other centers. We report results for a network trained in a fully-supervised manner on the training set from the same domain (FSL−SD) to give an upper-bound expectation, where a ResNet-101 is trained on the training images and used to classify test images, all from the same hospital. All the above experiments are repeated using each of $C2, C3, C4, C5$ as the source dataset.

We report in Table 1 a center wise and also an average performance for different UDA methods. The results in Table 1 show that UDA methods are better than conventional stain normalization approaches as evidenced by the superior performance of Chang et al. (2021). In Table 2 we report performance of different active domain adaptation methods. The results are compared against the average numbers for all 5 centers. Our *ALFREDO* approach when applied to supervised domain adaptation ($ALFREDO_{ASDA}$) outperforms the FSL-SD approach while our method's unsupervised domain adaptation approach, $ALFREDO_{AUDA}$, approaches the same performance of FSL−SD, the theoretical maximum performance. Since the FSL−SD approach is already at a high performance level our $ALFREDO_{ASDA}$ outperforms it at closer to 95% of labeled data.

The different ablation studies show the importance of different components of our method. The $\mathscr{L}_{base}$ term has the single biggest contribution in the performance of our feature disentanglement step justifying our approach to distil knowledge from a fully supervised classifier. Other components have significant roles to play. Similarly, for the informative sample selection components, Q_{novel} has the single most important contribution followed by significant contributions of other components.

4.4. Results on chest Xray dataset

Dataset Description: We use the following chest Xray datasets: **NIH Chest Xray** Dataset: The NIH ChestXray14 dataset (Wang et al., 2017b) has 112,120 expert-annotated frontal-view X-rays from 30,805 unique patients and has 14 disease labels. Original images were resized to 256 × 256. **CheXpert** Dataset: This dataset (Irvin et al., 2019) has 224,316 chest radiographs of 65,240 patients labeled for the presence of 14 common chest conditions. The validation ground-truth is obtained using majority voting from annotations of 3 board-certified radiologists. Original images were resized to 256 × 256. These two datasets have the same set of disease labels.

We divide both datasets into train/validation/test splits on the patient level at 70/10/20 ratio, such that images from one patient are in only one of the splits. Then we train a DenseNet-121 (Rajpurkar

Table 3
Classification results on the CheXpert dataset's test split using NIH data as the source domain. Note: FSL−SD is a fully-supervised model trained on target domain data.

	Atel.	Card.	Eff.	Infil.	Mass	Nodule	Pneu.	Pneumot.	Consol.	Edema	Emphy.	Fibr.	PT.	Hernia	Mean
No DA	0.697	0.814	0.761	0.652	0.739	0.694	0.703	0.781	0.704	0.792	0.815	0.719	0.728	0.811	0.752
MMD	0.741	0.851	0.801	0.699	0.785	0.738	0.748	0.807	0.724	0.816	0.831	0.745	0.754	0.846	0.769
CycleGANs	0.765	0.874	0.824	0.736	0.817	0.758	0.769	0.832	0.742	0.838	0.865	0.762	0.773	0.864	0.781
DANN	0.792	0.902	0.851	0.761	0.849	0.791	0.802	0.869	0.783	0.862	0.894	0.797	0.804	0.892	0.837
FSL−SD	*0.849*	*0.954*	*0.903*	*0.814*	*0.907*	*0.825*	*0.844*	*0.928*	*0.835*	*0.928*	*0.951*	*0.847*	*0.842*	*0.941*	0.831
SDA$_{SOTA}$	0.854	0.965	0.914	0.824	0.918	0.835	0.856	0.937	0.845	0.936	*0.963*	0.861	0.852	0.952	0.882
GCN2 (UDA$_{SOTA}$)	0.809	0.919	0.870	0.765	0.871	0.807	0.810	0.882	0.792	0.883	0.921	0.817	0.812	0.914	0.848

Table 4
For NIH data as the source domain. AUC values for different baselines and proposed approach along with ablation studies. We focus on **Infiltration** condition.

	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	p-
FSL-SD	0.814	0.814	0.814	0.814	0.814	0.814	0.814	0.814	0.814	0.814	<0.001
Random	0.639	0.671	0.709	0.741	0.775	0.784	0.797	0.810	0.818	0.821	<0.001
Unc	0.648	0.687	0.725	0.763	0.797	0.809	0.819	0.835	0.842	0.851	<0.001
AADA	0.655	0.694	0.735	0.773	0.808	0.829	0.845	0.858	0.876	0.88	<0.001
BADGE	0.643	0.678	0.716	0.757	0.789	0.81	0.824	0.843	0.849	0.858	0.005
CLUE	0.648	0.688	0.729	0.763	0.793	0.815	0.837	0.849	0.863	0.869	0.007
Fu et al. (2021)	0.652	0.689	0.732	0.775	0.805	0.827	0.845	0.855	0.872	0.879	0.001
Ghamsarian et al. (2023)	0.656	0.688	0.732	0.779	0.810	0.823	0.843	0.861	0.869	0.881	0.02
ALFREDO$_{ASDA}$	0.669	0.702	0.743	0.787	0.825	0.851	0.872	0.889	0.899	0.914	0.039
ALFREDO$_{AUDA}$	0.661	0.694	0.735	0.777	0.818	0.837	0.861	0.873	0.883	0.898	–
Ablation studies											
ALFREDO$_{No-FeatDisent}$	0.649	0.689	0.728	0.767	0.799	0.811	0.821	0.836	0.843	0.853	<0.001
Feature disentanglement											
AUDA$_{wo\mathscr{L}_1}$	0.615	0.639	0.687	0.719	0.781	0.809	0.819	0.832	0.843	0.851	0.01
AUDA$_{wo\mathscr{L}_2}$	0.621	0.649	0.698	0.725	0.788	0.816	0.824	0.836	0.849	0.859	0.02
AUDA$_{wo\mathscr{L}_3}$	0.606	0.637	0.678	0.707	0.772	0.796	0.808	0.819	0.828	0.841	0.009
AUDA$_{wo\mathscr{L}_{base}}$	0.604	0.629	0.664	0.685	0.731	0.77	0.785	0.806	0.818	0.829	0.008
Informative sample selection											
AUDA$_{woQ_{Unc}}$	0.625	0.657	0.699	0.729	0.795	0.821	0.832	0.839	0.852	0.866	0.01
AUDA$_{woQ_{dom}}$	0.618	0.635	0.689	0.714	0.778	0.813	0.821	0.828	0.841	0.851	0.008
AUDA$_{woQ_{density}}$	0.610	0.631	0.685	0.717	0.77	0.805	0.812	0.822	0.831	0.846	0.009
AUDA$_{woQ_{novel}}$	0.600	0.624	0.682	0.710	0.767	0.801	0.809	0.818	0.829	0.842	0.004

Table 5
Classification results on the NIH Xray dataset's test split using CheXpert data as the source domain. Note: $FSL-SD$ is a fully-supervised model trained on target domain data.

	Atel.	Card.	Eff.	Infil.	Mass	Nodule	Pneu.	Pneumot.	Consol.	Edema	Emphy.	Fibr.	PT	Hernia	Mean
No DA	0.718	0.823	0.744	0.730	0.739	0.694	0.683	0.771	0.712	0.783	0.803	0.711	0.710	0.785	0.752
MMD	0.734	0.846	0.762	0.741	0.785	0.738	0.709	0.793	0.731	0.801	0.821	0.726	0.721	0.816	0.769
CycleGANs	0.751	0.861	0.785	0.761	0.817	0.758	0.726	0.814	0.746	0.818	0.837	0.741	0.737	0.836	0.778
DANN	0.773	0.882	0.819	0.785	0.837	0.791	0.759	0.838	0.770	0.836	0.863	0.766	0.762	0.861	0.811
FSL−SD	*0.814*	*0.929*	*0.863*	*0.821*	*0.869*	*0.825*	*0.798*	*0.863*	*0.805*	*0.872*	*0.904*	*0.802*	*0.798*	*0.892*	0.856
SDA$_{SOTA}$	0.801	0.913	0.844	0.807	0.851	0.809	0.779	0.848	0.790	0.849	0.891	0.789	0.781	0.873	0.829
UDA$_{SOTA}$	0.786	0.906	0.833	0.789	0.831	0.802	0.763	0.835	0.774	0.837	0.868	0.768	0.763	0.860	0.781

et al., 2017) classifier on one dataset (e.g. NIH's train split). Here the NIH dataset serves as the source data and CheXpert is the target dataset. We then apply the trained model on the training split of the NIH dataset and tested on the test split of the same domain, the results are denoted as $FSL-Same$. When we apply this model to the test split of the CheXpert data without domain adaptation, the results are reported under No-UDA.

Tables 3 and 5 show classification results for different DA techniques where, respectively, the NIH and CheXpert dataset were the source domain and the performance metrics are for, respectively, CheXpert and NIH dataset's *test split* (the target domain). We observe that UDA methods perform worse than FSL−SD. This is expected since it is very challenging to perfectly account for domain shift. However all UDA methods perform better than fully supervised methods trained on one domain and applied on another without domain adaptation. The DANN architecture (Ganin et al., 2016b) outperforms MMD and cycleGANs, and is on par with graph convolutional methods GCAN (Ma et al., 2019) and GCN2 (Hong et al., 2019b). However ALFREDO outperforms all compared methods which can be attributed to the combination of active learning and domain adaptation.

4.5. Ablation studies

We also show in different tables the results for ablation studies where different components of ALFREDO are excluded and the corresponding performance numbers are calculated. We show results for the following methods:

1. AUDA$_{wo\mathscr{L}_1}$: Our proposed method used for AUDA without the loss term $\mathscr{L}_1$ in Eq. (3)
2. AUDA$_{wo\mathscr{L}_2}$: AUDA without the loss term $\mathscr{L}_2$ in Eq. (4)
3. AUDA$_{wo\mathscr{L}_3}$: AUDA without the loss term $\mathscr{L}_3$ in Eq. (5)
4. AUDA$_{wo\mathscr{L}_{base}}$: AUDA without the loss term $\mathscr{L}_{base}$ in Eq. (6)
5. AUDA$_{woQ_{Unc}}$: AUDA without the informativeness term Q$_{Unc}$ in Eq. (7)
6. AUDA$_{woQ_{dom}}$: AUDA without the domainness term Q$_{dom}$ in Eq. (8)
7. AUDA$_{woQ_{density}}$: AUDA without the density term Q$_{density}$ in Eq. (9)
8. AUDA$_{woQ_{novel}}$: AUDA without the novelty term Q$_{novl}$ in Eq. (10)

AUDA indicates that we show the results for ALFREDO applied to active unsupervised domain adaptation.

Table 6
For CheXpert data as the source domain. AUC values for different baselines and proposed approach along with ablation studies. We focus on **Infiltration** condition.

	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	p-
FSL-SD	0.821	0.821	0.821	0.821	0.821	0.821	0.821	0.821	0.821	0.821	<0.001
Random	0.632	0.663	0.702	0.736	0.768	0.775	0.792	0.803	0.811	0.814	<0.001
Unc	0.641	0.678	0.719	0.757	0.789	0.801	0.812	0.825	0.836	0.843	<0.001
AADA	0.649	0.686	0.728	0.768	0.80	0.821	0.837	0.851	0.867	0.873	<0.001
BADGE	0.638	0.672	0.714	0.751	0.785	0.804	0.817	0.834	0.843	0.851	0.005
CLUE	0.641	0.68	0.721	0.761	0.789	0.812	0.830	0.843	0.859	0.862	0.007
Fu et al. (2021)	0.649	0.686	0.728	0.768	0.80	0.821	0.837	0.851	0.867	0.873	0.001
Ghamsarian et al. (2023)	0.651	0.683	0.725	0.773	0.802	0.818	0.835	0.857	0.866	0.877	0.02
ALFREDO$_{ASDA}$	0.661	0.696	0.737	0.78	0.817	0.843	0.865	0.881	0.891	0.907	0.039
ALFREDO$_{AUDA}$	0.657	0.689	0.730	0.772	0.811	0.829	0.855	0.869	0.878	0.892	–
Ablation studies											
ALFREDO$_{No-FeatDisent}$	0.645	0.682	0.721	0.758	0.792	0.804	0.814	0.828	0.839	0.845	<0.001
Feature disentanglement											
AUDA$_{wo\mathscr{L}_1}$	0.611	0.634	0.681	0.714	0.775	0.803	0.814	0.828	0.838	0.847	0.01
AUDA$_{wo\mathscr{L}_2}$	0.618	0.645	0.692	0.721	0.784	0.811	0.82	0.831	0.844	0.853	0.02
AUDA$_{wo\mathscr{L}_3}$	0.603	0.632	0.673	0.702	0.767	0.791	0.804	0.814	0.822	0.835	0.009
AUDA$_{wo\mathscr{L}_{base}}$	0.601	0.628	0.662	0.683	0.735	0.772	0.789	0.801	0.813	0.826	0.008
Informative sample selection											
AUDA$_{woQ_{Unc}}$	0.621	0.654	0.695	0.725	0.792	0.819	0.826	0.836	0.849	0.861	0.01
AUDA$_{woQ_{dom}}$	0.612	0.633	0.685	0.711	0.772	0.809	0.816	0.825	0.837	0.849	0.008
AUDA$_{woQ_{density}}$	0.605	0.628	0.681	0.712	0.767	0.801	0.809	0.818	0.828	0.841	0.009
AUDA$_{woQ_{novel}}$	0.595	0.621	0.677	0.706	0.762	0.795	0.801	0.813	0.824	0.836	0.004

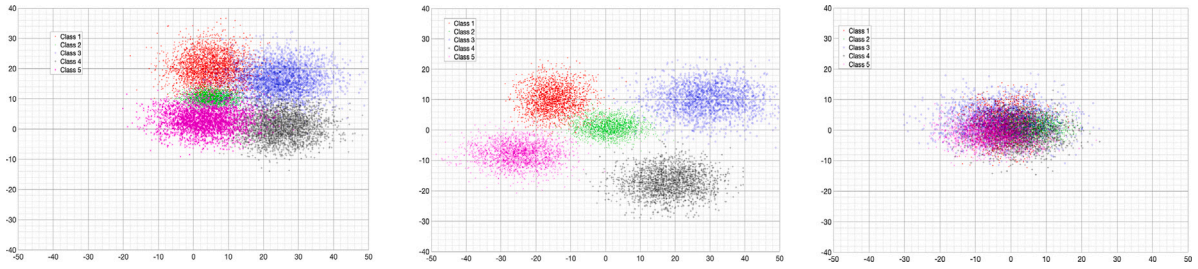

Fig. 3. T-sne results comparison between original image features and feature disentanglement output of source domain features. (a) Original image features; (b) Task specific features; (c) Domain specific features. Each color in the cluster refers to different classes.

Similar to the results for histopathology images the results for xray images show a similar trend regarding the contribution of different components of feature disentanglement and informative sample selection. $\mathscr{L}_{base}$ has the single biggest contribution in the performance of feature disentanglement justifying our approach to distil knowledge from a fully supervised classifier. Of the other three disentanglement components $\mathscr{L}_3$ has the second most important contribution. We hypothesize that this could be due to the fact that $\mathscr{L}_3$ enforces the task specific and domain specific components to be complementary to each other, and thus assimilate more information. In comparison, $\mathscr{L}_1, \mathscr{L}_2$ have similar contributions due to their focus on getting the domain specific and task specific features.

Similarly, for the informative sample selection components, Q_{novel} has the single most important contribution followed by significant contributions of other components. Q_{Unc} is based on conventional label uncertainty, and is a widely used method for determining sample informativeness. The domainness (Q_{Dom}) and density ($Q_{Density}$) components by themselves have similar contributions. All four terms combined contribute to the better performance of our method.

In another set of experiments, to check the effect of feature disentanglement we conduct experiments directly on the image features without resorting to feature disentanglement and denote this method as ALFREDO$_{No-FeatDisent}$ in Tables 2, 4, 6. This reduces the method to being an informative sample selection method without taking into account the domain shift. Consequently, we calculate $Q_{Unc}, Q_{dom}, Q_{density}$ and Q_{novel} using the image features. The results are shown in each of the tables, and are slightly better than using only uncertainty ("Unc"). This is explained by the fact that we use some additional information with uncertainty. However, due to the fact that the domain shift is not considered, the results are much worse than the full ALFREDO method.

4.6. T-SNE visualizations

Fig. 3(a) shows the t-sne plots of image features (taken from the fully connected layer of a DenseNet-121 trained for image classification) while Fig. 3(b) shows the plot using the task-specific features. The plots of the original features shows different image class clusters that overlap and that makes it challenging to have good classification. On the other hand, the clusters obtained using the task-specific features are well separated and there is less overlap between different clusters. Fig. 3(c) shows the output of using domain specific features where a significant overlap is observed among classes. This is due to the fact that domain specific features of samples from different classes (but of the same domain are very similar). This clearly demonstrates the efficacy of our feature disentanglement method, i.e., the task-specific and domain specific features fulfill their desired objectives. In the example in Fig. 3, the features are taken from images belonging to 5 classes (Atelectasis, Consolidation, Effusion, Infiltration and Nodule) from the NIH dataset.

4.7. Robustness and generalization

To test the robustness of the proposed approach, we added simulated gaussian noise of $\mu = 0$ and different $\sigma \in \{0.005, 0.01, 0.015, 0.05, 0.1\}$ and run our UDA pipeline. Fig. 4 shows the AUC values for the baseline performance of UDA and different σ. The results are close to

Table 7
Values of hyperparameters for different datasets used in our experiments.

	Feature disentanglement				Informativeness				Q_{dom} thresholds (Eq. (8))	
	λ_1	λ_2	λ_3	λ_{base}	λ_{Unc}	λ_{Dom}	$\lambda_{Density}$	λ_{Novel}	η_1	η_2
NIH	0.85	1.1	0.95	1.2	1.1	0.9	1.05	1.25	0.24	0.78
CheXpert	0.95	1.0	1.1	1.0	1.2	1.1	0.95	1.0	0.29	0.8
Camelyon17	0.8	1.05	0.85	0.95	0.9	0.75	1.0	1.0	0.21	0.82

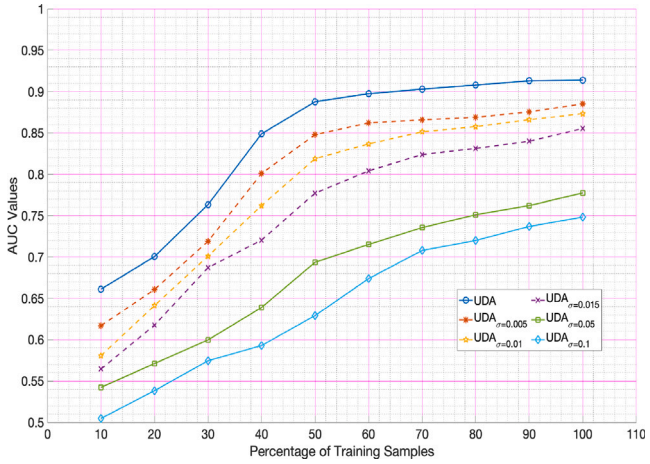

Fig. 4. AUC measures for different features for added Gaussian noise of $\mu = 0$ and different σ.

UDA for $\sigma = 0.005, 0.01$, but start to degrade significantly for noise levels above $\sigma = 0.01$, which we term as noise threshold. These results demonstrate that our method is robust to a reasonable level of added noise.

4.8. Hyperparameter settings

For our method we have two sets of hyperparameter values: for the feature disentanglement (Eq. (1)) and for informative sample selection (Eq. (11)). To set the hyperparameters for feature disentanglement we adopt the following steps using the NIH X-ray dataset. For λ_1 we varied the values from $[0, 1.3]$ in steps of 0.05, keeping $\lambda_2 = 0.45, \lambda_3 = 0.5, \lambda_{base} = 0.6$. The best results were obtained for $\lambda_1 = 0.85$, which was our final value. Then we vary λ_2 in a similar range with constant values of $\lambda_1 = 0.85, \lambda_3 = 0.5, \lambda_{base} = 0.6$ to get the best results for $\lambda_2 = 1.1$. We repeat the above steps to get $\lambda_3 = 0.95, \lambda_{base} = 1.2$. We repeat the entire sequence of steps for the parameters of Eq. (11) and finally set $\lambda_{Unc} = 1.1, \lambda_{Dom} = 0.9, \lambda_{Density} = 1.05, \lambda_{Novel} = 1.25$. Table 7 provides the values for each parameter for the different datasets that we used.

5. Discussion and conclusion

In this paper, we present a novel approach for active domain adaptation that combines active learning and domain adaptation. The key motivation in combining active learning with domain adaptation is to reduce the annotation cost for supervised settings, and in the case of unsupervised domain adaptation, reduce the number of samples required for training an accurate system. Unlike most current works that deal with domain adaptation and active learning separately we combine both approaches and leverage their respective advantages. We propose a novel feature disentanglement approach where an image's feature representation is separated into task specific and domain specific features. The task specific features of source and target domain are projected to a common space such that a classifier trained on one domain features can perform equally well on the other domain. The advantage of task specific features is active learning strategies can be used to select informative target domain samples using a classifier trained on source domain samples. We propose a novel informativeness score that selects informative samples based on the criteria of uncertainty, domainness, density and novelty.

Our proposed method yields better results than SOTA methods for active learning and domain adaptation. Subsequent ablation studies also highlight the importance of each term in the loss function and justifies their inclusion. In future work, we aim to test our model on other medical image datasets. We also aim to test its robustness and generalizability to different classification architectures.

CRediT authorship contribution statement

Dwarikanath Mahapatra: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Ruwan Tennakoon:** Conceptualization, Writing – original draft, Writing – review & editing. **Yasmeen George:** Conceptualization, Writing – original draft, Writing – review & editing. **Sudipta Roy:** Conceptualization, Writing – original draft, Writing – review & editing. **Behzad Bozorgtabar:** Conceptualization, Data curation, Writing – original draft, Writing – review & editing. **Zongyuan Ge:** Conceptualization, Writing – original draft, Writing – review & editing. **Mauricio Reyes:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets are publicly available.

Acknowledgments

This work was supported by the Swiss National Foundation grant number 212939, and Innosuisse, Switzerland grant number 31274.1

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf. Fusion 76, 243–297. http://dx.doi.org/10.1016/j.inffus.2021.05.008.

Ahmedt-Aristizabal, D., Armin, M.A., Denman, S., Fookes, C., Petersson, L., 2021. Graph-based deep learning for medical diagnosis and analysis: Past, present and future. Sensors 21 (14), 47–58.

Ahn, E., Kumar, A., Fulham, M., Feng, D., Kim, J., 2020. Unsupervised domain adaptation to classify medical images using zero-bias convolutional auto-encoders and context-based feature augmentation. IEEE TMI 39 (7), 2385–2394.

Alefsen, M., Vorontsov, E., Kadoury, S., 2023. M-GenSeg: Domain adaptation for target modality tumor segmentation with annotation-efficient supervision. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Springer Nature Switzerland, Cham, pp. 141–151.

Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A., 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. CoRR abs/1906.03671.

Bandi, P., et al., 2019. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. IEEE Trans. Med. Imaging 38 (2), 550–560.

Bermudez-Chacon, R., Marquez-Neila, P., Salzmann, M., Fua, P., 2018. A domain-adaptive two-stream U-Net for electron microscopy image segmentation. In: IEEE ISBI. pp. 400–404.

Bozorgtabar, B., Mahapatra, D., von Teng, H., Pollinger, A., Ebner, L., Thiran, J.-P., Reyes, M., 2019. Informative sample generation using class aware generative adversarial networks for classification of chest xrays. Comput. Vis. Image Underst. 184, 57–65.

Budd, S., Robinson, E.C., Kainz, B., 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Med. Image Anal. 71, 102062. http://dx.doi.org/10.1016/J.MEDIA.2021.102062.

Chang, J.-R., Wu, M.-S., et al., 2021. Stain Mix-Up: Unsupervised domain generalization for histopathology images. In: MICCAI 2021. pp. 117–126.

Chattopadhyay, R., Fan, W., Davidson, I., Panchanathan, S., Ye, J., 2013. Joint transfer and batch-mode active learning. In: Dasgupta, S., McAllester, D. (Eds.), Proceedings of the 30th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, Vol. 28, (3), PMLR, Atlanta, Georgia, USA, pp. 253–261, URL: https://proceedings.mlr.press/v28/chattopadhyay13.html.

Ducoffe, M., Precioso, F., 2018. Adversarial active learning for deep networks: a margin based approach. CoRR abs/1802.09841.

Fu, B., Cao, Z., Wang, J., Long, M., 2021. Transferable query selection for active domain adaptation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 7268–7277. http://dx.doi.org/10.1109/CVPR46437.2021.00719.

Gal, Y., Islam, R., Ghahramani, Z., 2017. Deep Bayesian active learning with image data. In: Proc. International Conference on Machine Learning.

Ganin, Y., Lempitsky, V., 2015. Unsupervised domain adaptation by backpropagation.

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

Ganin, Y., Ustinova, E., et al., 2016b. Domain-adversarial training of neural networks.

Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., Guttmann, C.R.G., de Leeuw, F., Tempany, C.M., van Ginneken, B., Fedorov, A., Abolmaesumi, P., Platel, B., III, W.M.W., 2017. Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation. CoRR abs/1702.07841.

Ghamsarian, N., Gamazo Tejero, J., Márquez-Neila, P., Wolf, S., Zinkernagel, M., Schoeffmann, K., Sznitman, R., 2023. Domain adaptation for medical image segmentation using transformation-invariant self-training. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Springer Nature Switzerland, Cham, pp. 331–341.

Gissin, D., Shalev-Shwartz, S., 2019. Discriminative active learning. CoRR abs/1907.06347.

Gopinath, K., Desrosiers, C., Lombaert, H., 2020. Graph domain adaptation for alignment-invariant brain surface segmentation.

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

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: In Proc. CVPR.

Heimann, T., Mountney, P., John, M., Ionasec, R., 2013. Learning without labeling: Domain adaptation for ultrasound transducer localization. In: MICCAI 2013. pp. 49–56.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., Darrell, T., 2018. CyCADA: Cycle-consistent adversarial domain adaptation. In: Dy, J., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, Vol. 80, PMLR, pp. 1989–1998, URL: https://proceedings.mlr.press/v80/hoffman18a.html.

Hong, Y., Chen, G., Yap, P.-T., Shen, D., 2019a. Multifold acceleration of diffusion MRI via deep learning reconstruction from slice-undersampled data. In: IPMI. pp. 530–541.

Hong, Y., Kim, J., Chen, G., Lin, W., Yap, P.-T., Shen, D., 2019b. Longitudinal prediction of infant diffusion MRI data via graph convolutional adversarial networks. IEEE Trans. Med. Imaging 38 (12), 2717–2725.

Huang, S.-J., Zhao, J.-W., Liu, Z.-Y., 2018. Cost-effective training of deep CNNs with active model adaptation. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '18, Association for Computing Machinery, New York, NY, USA, pp. 1580–1588. http://dx.doi.org/10.1145/3219819.3220026.

Huang, Y., Zheng, H., Liu, C., Ding, X., Rohde, G.K., 2017. Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images. IEEE J. Biomed. Health Inf. 21 (6), 1625–1632. http://dx.doi.org/10.1109/JBHI.2017.2691738.

Irvin, J., Rajpurkar, P., et al., 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv preprint arXiv:1901.07031.

Jin, H., Che, H., Chen, H., 2023. Unsupervised domain adaptation for anatomical landmark detection. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Springer Nature Switzerland, Cham, pp. 695–705.

Jungo, A., Balsiger, F., Reyes, M., 2020. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Front. Neurosci. 14, 282.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., Glocker, B., 2017b. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Information Processing in Medical Imaging. pp. 597–609.

Kamnitsas, K., Baumgartner, C., et al., 2017a. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: IPMI. pp. 597–609.

Kirsch, A., van Amersfoort, J., Gal, Y., 2019. BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning. CoRR abs/1906.08158.

Kumar, D., Kumar, C., Shao, M., 2017. Cross-database mammographic image analysis through unsupervised domain adaptation. In: 2017 IEEE International Conference on Big Data (Big Data). pp. 4035–4042. http://dx.doi.org/10.1109/BigData.2017.8258419.

Li, X., Guo, Y., 2013. Adaptive active learning for image classification. In: Proc. CVPR.

Lin, Y., Nie, D., Liu, Y., Yang, M., Zhang, D., Wen, X., 2023. Multi-target domain adaptation with prompt learning for medical image segmentation. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Springer Nature Switzerland, Cham, pp. 717–727.

Liu, T., Chen, Z., Li, Q., Wang, Y., Zhou, K., Xie, W., Fang, Y., Zheng, K., Zhao, Z., Liu, S., Yang, W., 2023. MDA-SR: Multi-level domain adaptation super-resolution for wireless capsule endoscopy images. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Springer Nature Switzerland, Cham, pp. 518–527.

Ma, X., Zhang, T., Xu, C., 2019. GCAN: Graph convolutional adversarial network for unsupervised domain adaptation. In: IEEE CVPR. pp. 8258–8268.

Mahapatra, D., Bozorgtabar, B., Thiran, J.-P., Reyes, M., 2018. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In: In Proc. MICCAI. pp. 580–588.

Mahapatra, D., Buhmann, J., 2016. Visual saliency-based active learning for prostate magnetic resonance imaging segmentation. SPIE J. Med. Imaging 3 (1), 014003.

Mahapatra, D., Ge, Z., 2020. Training data independent image registration using generative adversarial networks and domain adaptation. Pattern Recognit. 100, 107109. http://dx.doi.org/10.1016/j.patcog.2019.107109, URL: https://www.sciencedirect.com/science/article/pii/S0031320319304108.

Mahapatra, D., Korevaar, S., Bozorgtabar, B., Tennakoon, R.B., 2022. Unsupervised domain adaptation using feature disentanglement and GCNs for medical image classification. In: ECCV 2022 Workshops. In: Lecture Notes in Computer Science, Vol. 13807, Springer, pp. 735–748.

Mahapatra, D., Poellinger, A., Reyes, M., 2023. Graph node based interpretability guided sample selection for active learning. IEEE Trans. Med. Imaging 42 (3), 661–673. http://dx.doi.org/10.1109/TMI.2022.3215017.

Mahapatra, D., Poellinger, A., Shao, L., Reyes, M., 2021. Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. IEEE TMI 40 (10), 2548–2562.

Mayer, C., Timofte, R., 2018. Adversarial sampling for active learning. arXiv preprint arXiv:1808.06671.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B., Snoek, J., 2019. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift. Curran Associates Inc., Red Hook, NY, USA.

Park, T., Zhu, J.-Y., Wang, O., Lu, J., Shechtman, E., Efros, A.A., Zhang, R., 2020. Swapping autoencoder for deep image manipulation. In: Advances in Neural Information Processing Systems.

Prabhu, V., Chandrasekaran, A., Saenko, K., Hoffman, J., 2021. Active domain adaptation via clustering uncertainty-weighted embeddings. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 8485–8494. http://dx.doi.org/10.1109/ICCV48922.2021.00839.

Raghu, M., Zhang, C., Kleinberg, J.M., Bengio, S., 2019. Transfusion: Understanding transfer learning with applications to medical imaging. CoRR abs/1902.07208.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M.P., Ng, A., 2017. Chexnet: Radiologist-level pneumonia detection on chest X-Rays with deep learning. arXiv preprint arXiv:1711.05225.

Saenko, K., Kulis, B., Fritz, M., Darrell, T., 2010. Adapting visual category models to new domains. In: Daniilidis, K., Maragos, P., Paragios, N. (Eds.), Computer Vision – ECCV 2010. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 213–226.

Sener, O., Savarese, S., 2018. Active learning for convolutional neural networks: A core-set approach. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=H1aIuk-RW.

Sinha, S., Ebrahimi, S., Darrell, T., 2019. Variational adversarial active learning. In: 2019 IEEE/CVF International Conference on Computer Vision. ICCV, IEEE Computer Society, Los Alamitos, CA, USA, pp. 5971–5980. http://dx.doi.org/10.1109/ICCV.2019.00607, URL: https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00607.

Sourati, J., Gholipour, A., Dy, J.G., Tomas-Fernandez, X., Kurugol, S., Warfield, S.K., 2019. Intelligent labeling based on fisher information for medical image segmentation using deep learning. IEEE Trans. Med. Imaging 38 (11), 2642–2653.

Su, J., Tsai, Y., Sohn, K., Liu, B., Maji, S., Chandraker, M., 2020a. Active adversarial domain adaptation. In: 2020 IEEE Winter Conference on Applications of Computer Vision. WACV, IEEE Computer Society, Los Alamitos, CA, USA, pp. 728–737. http://dx.doi.org/10.1109/WACV45572.2020.9093390.

Su, J.-C., Tsai, Y.-H., Sohn, K., Liu, B., Maji, S., Chandraker, M., 2020b. Active adversarial domain adaptation.

Torralba, A., Efros, A.A., 2011. Unbiased look at dataset bias. In: CVPR 2011. pp. 1521–1528.

Wang, H., Jin, Q., Li, S., Liu, S., Wang, M., Song, Z., 2024. A comprehensive survey on deep active learning in medical image analysis. Med. Image Anal. 95, 103201. http://dx.doi.org/10.1016/j.media.2024.103201, URL: https://www.sciencedirect.com/science/article/pii/S1361841524001269.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R., 2017b. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: In Proc. CVPR.

Wang, D., Shang, Y., 2014. A new active labeling method for deep learning. In: 2014 International Joint Conference on Neural Networks. IJCNN, pp. 112–119. http://dx.doi.org/10.1109/IJCNN.2014.6889457.

Wang, K., Zhang, D., Li, Y., Zhang, R., Lin., L., 2017a. Cost-effective active learning for deep image classification. IEEE Trans. CSVT 27 (12), 2591–2600.

Wu, M., Pan, S., Zhou, C., Chang, X., Zhu, X., 2020. Unsupervised domain adaptive graph convolutional networks. In: Proceedings of the Web Conference 2020. pp. 1457–1467.

Xu, Z., Gong, H., Wan, X., Li, H., 2023b. ASC: Appearance and structure consistency for unsupervised domain adaptation in fetal brain MRI segmentation. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Springer Nature Switzerland, Cham, pp. 325–335.

Xu, J., Jin, Y., Martin, B., Smith, A., Wright, S., Stoyanov, D., Mazomenos, E.B., 2023a. Regressing simulation to real: Unsupervised domain adaptation for automated quality assessment in transoesophageal echocardiography. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Springer Nature Switzerland, Cham, pp. 154–164.

Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D., 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: Proc. MICCAI. pp. 399–407.

Yu, Q., Xi, N., Yuan, J., Zhou, Z., Dang, K., Ding, X., 2023. Source-free domain adaptation for medical image segmentation via prototype-anchored feature alignment and contrastive learning. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Springer Nature Switzerland, Cham, pp. 3–12.

Zhang, J., Chao, H., Dhurandhar, A., Chen, P.-Y., Tajer, A., Xu, Y., Yan, P., 2023. Spectral adversarial mixup for few-shot unsupervised domain adaptation. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Springer Nature Switzerland, Cham, pp. 728–738.

Zheng, H., Yang, L., Chen, J., Han, J., Zhang, Y., Liang, P., Zhao, Z., Wang, C., Chen, D.Z., 2019. Biomedical image segmentation via representative annotation. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33, pp. 5901–5908.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: Proc. CVPR. pp. 2921–2929.

Zhu, J.-J., Bento, J., 2017. Generative adversarial active learning. arXiv preprint arXiv:1702.07956.

Zhuang, X., Shen, J., 2016. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. Med. Image Anal. 31, 77–87.