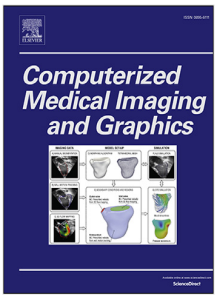
Journal Pre-proof



Prior knowledge-guided vision-transformer-based unsupervised domain adaptation for intubation prediction in lung disease at one week

Junlin Yang, John Anderson Garcia Henao, Nicha Dvornek, Jianchun He, Danielle V. Bower, Arno Depotter, Herkus Bajercius, Aurélie Pahud de Mortanges, Chenyu You, Christopher Gange, Roberta Eufrasia Ledda, Mario Silva, Charles S. Dela Cruz, Wolf Hautz, Harald M. Bonel, Mauricio Reyes, Lawrence H. Staib, Alexander Poellinger, James S. Duncan

PII: S0895-6111(24)00119-8
DOI: https://doi.org/10.1016/j.compmedimag.2024.102442
Reference: CMIG 102442

To appear in: *Computerized Medical Imaging and Graphics*

Received date : 13 June 2024
Revised date : 5 September 2024
Accepted date : 30 September 2024

Please cite this article as: J. Yang, J.A.G. Henao, N. Dvornek et al., Prior knowledge-guided vision-transformer-based unsupervised domain adaptation for intubation prediction in lung disease at one week. *Computerized Medical Imaging and Graphics* (2024), doi: https://doi.org/10.1016/j.compmedimag.2024.102442.

Computerized Medical Imaging and Graphics (2024)

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

Prior Knowledge-Guided Vision-Transformer-Based Unsupervised Domain Adaptation for Intubation Prediction in Lung Disease at One Week

Junlin Yang[a,*], John Anderson Garcia Henao[b], Nicha Dvornek[a,d], Jianchun He[a,d], Danielle V Bower[c], Arno Depotter[c], Herkus Bajercius[c], Aurélie Pahud de Mortanges[c], Chenyu You[e], Christopher Gange[d], Roberta Eufrasia Ledda[g], Mario Silva[f,g], Charles S. Dela Cruz[h], Wolf Hautz[i], Harald M. Bonel[c,j,k], Mauricio Reyes[b,l], Lawrence H. Staib[a,d,e], Alexander Poellinger[c,**], James S. Duncan[a,d,e,**]

[a] *Department of Biomedical Engineering, Yale University, New Haven, CT, USA*
[b] *The ARTORG Center for Biomedical Research, University of Bern, Bern, Switzerland*
[c] *Department of Diagnostic, Interventional, and Pediatric Radiology, Inselspital Bern, University of Bern, Bern, Switzerland*
[d] *Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, CT, USA*
[e] *Department of Electrical Engineering, Yale University, New Haven, CT, USA*
[f] *Section of "Scienze Radiologiche," Diagnostic Department, University Hospital of Parma, Parma, Italy*
[g] *Department of Medicine and Surgery, University of Parma, Italy*
[h] *Section of Pulmonary, Critical Care, and Sleep Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA*
[i] *Department of Emergency Medicine, Inselspital University Hospital, University of Bern, Bern, Switzerland*
[j] *Campusradiologie, Department of Radiological Diagnostics, Lindenhofspital Bern, Bern, Switzerland*
[k] *Campus Stiftung Lindenhof Bern, Bern, Switzerland*
[l] *Department of Radiation Oncology, Inselspital, Bern University Hospital, Bern, Switzerland*

ARTICLE INFO

Article history:

2000 MSC: 41A05, 41A10, 65D05, 65D17

Keywords: Unsupervised Domain Adaptation, Prior Knowledge, 3D/2D, Transformer, Pneumonia, Chest CT, Chest X-ray

ABSTRACT

Data-driven approaches have achieved great success in various medical image analysis tasks. However, fully-supervised data-driven approaches require unprecedentedly large amounts of labeled data and often suffer from poor generalization to unseen new data due to domain shifts. Various unsupervised domain adaptation (UDA) methods have been actively explored to solve these problems. Anatomical and spatial priors in medical imaging are common and have been incorporated into data-driven approaches to ease the need for labeled data as well as to achieve better generalization and interpretation. Inspired by the effectiveness of recent transformer-based methods in medical image analysis, the adaptability of transformer-based models has been investigated. How to incorporate prior knowledge for transformer-based UDA models remains underexplored. In this paper, we introduce a prior knowledge-guided and transformer-based unsupervised domain adaptation (PUDA) pipeline. It regularizes the vision transformer attention heads using anatomical and spatial prior information that is shared by both the source and target domain, which provides additional insight into the similarity between the underlying data distribution across domains. Besides the global alignment of class tokens, it assigns local weights to guide the token distribution alignment via adversarial training. We evaluate our proposed method on a clinical outcome prediction task, where Computed Tomography (CT) and Chest X-ray (CXR) data are collected and used to predict the intubation status of patients in a week. Abnormal lesions are regarded as anatomical and spatial prior information for this task and are annotated in the source domain scans. Extensive experiments show the effectiveness of the proposed PUDA method.

1. Introduction

Data-driven machine learning approaches have been overwhelmingly successful in a variety of medical image analysis tasks Ronneberger et al. (2015); Litjens et al. (2017); Shen et al. (2017) and have proven more powerful and accurate than their model-driven counterparts. However, such impressive achievements rely heavily on massive amounts of labeled data, which is often costly and time-consuming to obtain. Besides, data-driven models, particularly deep learning, often suffer from poor generalization to unseen new data and unclear interpretability. This situation motivates research on semi-supervised learning Van Engelen and Hoos (2020), unsupervised learning Raza and Singh (2021), and unsupervised domain adaptation (UDA) Wang and Deng (2018); Wilson and Cook (2020).

Most UDA work seeks to alleviate domain divergence. The mainstream approaches tend to learn domain-invariant features by performing alignment across different distributions by adversarial learning Ganin and Lempitsky (2015). Despite recent advances, UDA remains a challenging task due to the large domain shifts for many real-world applications.

In many real-world applications, especially medical imaging, prior knowledge is often widely available and can provide insights into the underlying structure of the data across domains. Manual annotation requires prior knowledge of anatomy and clinical expertise regarding the disease. This process exploits anatomical and spatial similarity across patient scans. Explicitly employing this prior knowledge has been explored in CNN-based deep-learning approaches. One of the popular strategies is to add a shape prior constraint Oktay et al. (2017) to encourage segmentation results to match both the ground truth and the shape prior. More recent work often exploits such prior information in deep learning models to reduce the use of labeled data via semi-supervised learning, unsupervised learning, and self-supervised learning Zhou et al. (2019); Dalca et al. (2018); Miao et al. (2022). As for UDA, Sun et al. (2022a) introduces prior knowledge of target class distribution to guide UDA. There are a few works Bateson et al. (2022); Zhang et al. (2022); Yao et al. (2022) to incorporate shape priors for medical image segmentation under the UDA setting.

Most above methods are based on CNN backbones. Recent transformers have achieved great success on various machine learning tasks. Some works investigate the UDA with transformer backbones, such as Yang et al. (2023); Xu et al. (2021); Sun et al. (2022b). How to incorporate prior knowledge for transformer-based UDA models remains under-explored.

This paper introduces a prior knowledge-guided and transformer-based unsupervised domain adaptation (PUDA) pipeline. To validate the effectiveness of the proposed pipeline, we focus on an important application of clinical outcome prediction. The early prediction of severity in COVID-19 patients is of vital importance for providing rapid and essential care to reduce mortality and optimize the use of medical resources Guner et al. (2021). More specifically, we collected both chest X-ray (CXR) and chest computed tomography (CT) from COVID-19 patients on day 0 and our goal is to achieve UDA on the task to predict if the patient will be intubated in one week from the time of imaging. Intubation labels are binary and generated from electronic health records. Abnormal regions in both CXR and CT scans are assumed to represent the shared prior knowledge.

CXR imaging is among the most commonly used diagnostic tools in clinical practice and has an essential role in the diagnosis of lung diseases, such as pneumonia, tuberculosis, interstitial lung disease, and early lung cancer Qin et al. (2018). Compared to chest CT, CXR is achieved with lower radiation doses and is much more available in almost all clinical settings due to its fast and low-cost acquisition Inui et al. (2021). While CXR is more available, it may be deemed less useful than chest CT due to its low sensitivity in the diagnosis of subtle parenchymal abnormalities and its limited ability to help differentiate parenchymal patterns Schaefer-Prokop and Prokop (2021). Chest CT carries more detailed information in infected regions while the projective nature of CXR causes large overlapping of anatomies, blurry object boundaries, and complex texture patterns. In contrast to CT scans, chest X-rays (CXR) can be performed frequently, even on severely ill patients, as portable CXR can be conducted within the ICU. This capability allows for significantly improved temporal monitoring of disease progression. It is thus of clinical interest to perform domain adaptation across CXR and CT images for various applications. Specifically, in our application of predicting intubation at seven days for COVID-19 patients, though using standardized severity criteria has contributed to reduced ICU overload Carbonell et al. (2021), a key challenge is that not all hospitals possess advanced imaging, such as CT, for this purpose. Certain hospitals, particularly in low-resource settings, depend solely on chest X-rays (CXR) to decide admission. To connect 3D CT scans and CXR scans, we generate digitally reconstructed radiographs (DRRs) from CT scans Unberath et al. (2018). Anatomical regions or lesions related to infection, such as lung, consolidation (CON), and ground glass opacity (GGO) are regarded as useful clinical prior knowledge for domain adaptation in our prediction task.

The contributions of our approach are as follows: (1) We propose a transformer-based UDA framework that utilizes shared anatomical and spatial priors across domains for medical image analysis. (2) The proposed method effectively regularizes the attention heads in the vision transformer guided by prior knowledge and improves both the discriminability and transferability of the learned features. Besides the global alignment of class tokens, the regularized attention guides the adversarial alignment of the distribution of sequential features. (3) Our knowledge-guided model incorporating domain expertise produces a more reasonable attention map along with the prediction results, thus leading to a better understanding of the deep learning model. (4) In addition, the proposed model performs favorably compared with human radiologists on the same intubation prediction task.

*Corresponding author.
**Correspondence to: 300 Cedar St. New Haven CT 06519. E-mail addresses: junlin.yang@yale.edu (J. Yang), james.duncan@yale.edu (J.S. Duncan), alexander.poellinger@insel.ch (A. Poellinger).

The rest of the paper is organized as follows. Related work is summarized in Sec. 2. The details of the method are described in Sec. 3. The experimental design and results are shown in Sec. 4. Finally, the work is concluded in Sec. 5.

2. Related work

2.1. Intubation Prediction

Covid-related machine learning works on medical imaging extensively focus on the diagnosis and assessment of COVID-19 patients with CT scans and CXR scans Bhatele et al. (2022); Alghamdi et al. (2021); Serte and Demirel (2021). Some works for COVID-19 mortality and intubation prediction are based on CT scans Chamberlin et al. (2022) and a few are based on CXR scans Kwon et al. (2020); Nakashima et al. (2023). Specifically, Kwon et al. (2020) utilizes both image features and clinical variables, and Nakashima et al. (2023) uses CXR for mortality prediction based on radiomics features combined with a bone-suppressing approach. In this work, we collect both CT and CXR scans and investigate an unsupervised domain adaptation model for the intubation prediction task by leveraging expert lesion labels across different image modalities.

2.2. Prior Knowledge

Incorporating domain expertise and prior knowledge into data-driven approaches is of particular interest for medical image analysis. Most related work focuses on utilizing anatomical and spatial priors for biomedical segmentation tasks. A popular method is to apply conditional random fields (CRFs) on the output of deep learning models in post-processing to take into account the context of neighboring labels. In particular, as pathology shows that most breast cancer originates from cells in the mammary layer, Huang et al. (2018) exploits the position of tumors and their relative locations with the mammary layer as a new term in a CRF energy function to refine the segmentation result from neural networks. Some supervised learning methods add a shape prior constraint (Oktay et al., 2017) to encourage the segmentation prediction to match the shape prior. More recent works often exploit such prior information to reduce the need for labeled data in various settings. Zhou et al. (2019) utilizes organ prior statistics via a prior-aware loss for partially-supervised organ segmentation. Following classical atlas-based probabilistic segmentation methods, Dalca et al. (2018) proposes a generative model that achieves fast unsupervised segmentation with anatomical priors. Miao et al. (2022) proposes spatial prior attention for better self-supervision training and improves the performance on downstream classification tasks. Inspired by Miao et al. (2022), our prior guided attention regularization embeds anatomical and spatial priors into the attention heads in vision transformers. Instead of self-supervision, we assume this anatomical and spatial domain expertise is not only shared across scans but also across different imaging modalities. Different from Miao et al. (2022), our model requires learning representations that are not only discriminative but also transferable across domains. It is proposed to combine the global and weighted local transfer loss with the spatial prior attention to learn transferable representations.

2.3. Vision Transformer

Attention-based transformers Vaswani et al. (2017) were initially proposed to model sequential data. Dosovitskiy et al. (2020) showed the state-of-the-art performance by transformer-based models on vision tasks. Since then, vision transformers (ViTs) have become increasingly popular and efforts have been made to improve their performance. Many ViTs and their variants are proposed to achieve remarkable performance on various vision tasks, including image classification Chen et al. (2021); Chu et al. (2021); Li et al. (2022b), object detection Beal et al. (2020); Fang et al. (2021); Li et al. (2022a), and semantic segmentation Strudel et al. (2021); Hatamizadeh et al. (2021); Gu et al. (2022). In our work, to better leverage the prior domain expertise knowledge, we utilize the attention mechanism in vision transformer to infuse expert knowledge. More specifically, the attention regularization framework explicitly enforces the embedding to incorporate the shared anatomical and spatial priors.

2.4. Unsupervised Domain Adaptation

UDA has attracted a lot of attention as it greatly improves the generalization ability of deep learning models to unseen new data. Various deep learning-based UDA methods have been explored Wang and Deng (2018). Discrepancy-based methods diminish the domain shift via fine-tuning the neural networks with unlabeled target data. Commonly used divergence measures include Maximum Mean Discrepancy (MMD) Tzeng et al. (2014) and Correlation Alignment (CORAL) Sun and Saenko (2016). Adversarial-based works utilize domain discriminators to encourage domain confusion through an adversarial objective Ganin and Lempitsky (2015); Tzeng et al. (2017). The pursuit of domain invariance of learned representations might result in negative transfer, i.e. transferring knowledge from the source can have a negative impact on the target learner. Since not all features are equally transferable, recent methods seek to learn more transferable features while preserving the discriminative ability of these features Wang et al. (2019).

In many real-world applications, while UDA remains a challenging task due to the large domain shifts, prior knowledge is often widely available. Such prior knowledge about data across domains provides valuable clues that are complementary to the unlabeled training data. Thus, knowledge-guided UDA has been explored. Sun et al. (2022a) incorporates prior knowledge on target class distribution to guide the UDA. There are different types of prior knowledge for UDA to consider. Specifically in the medical domain, some works primarily focus on the UDA for segmentation tasks with shape priors. For example, Bateson et al. (2022); Zhang et al. (2022); Yao et al. (2022) introduce shape constraints to improve medical image segmentation under UDA settings. Bigalke et al. (2023) proposes to embed anatomy prior knowledge for 3D human pose estimation.

In addition, the above methods are developed based on CNNs backbones. In recent years, the transformers have gained popularity for their effectiveness in modeling long-range dependencies and achieved great success on various machine learning tasks. Some works investigate the UDA with transformer

backbones, such as Yang et al. (2023); Xu et al. (2021); Sun et al. (2022b). For medical imaging applications, Ji and Chung (2023) explores UDA for cross-modality medical images on the segmentation task.

In this paper, we utilize the attention mechanism of the transformer backbones to incorporate spatial prior knowledge for UDA, which guides the learning process to focus on features that are both transferable and discriminative. The model demonstrates the transferability of transformer backbones on the medical image prediction task as well as the performance boost from the prior knowledge.

3. Methods

3.1. Problem Overview

The WHO severity score from 0-10 Marshall et al. (2020) is used to assess the disease status of COVID-19 patients. Patients with a WHO severity score greater than or equal to 7 will be intubated. Baseline medical scans (unpaired 3D CT scans and 2D CXR scans) at day 0 are acquired and will be used to train a model to predict if the patients will need intubation within 7 days. In addition, ground truth prediction labels for intubation or not within 7 days are acquired. In this task, abnormal lesions due to COVID-19 in both CT and CXR scans are regarded as shared anatomical and spatial priors. 3D CT scans are projected into DRRs. For each image DRR or CXR $X \in R^{w \times h}$, manually annotated abnormal lesions are denoted as prior expert knowledge map $K \in R^{w \times h}$. w and h are the width and height of image scans.

In this paper, our goal is to utilize anatomical and spatial priors to facilitate the unsupervised domain adaptation. More specifically, we focus on the task of improving the performance of the intubation prediction on unlabeled target scans using source scans with pixel-wise expert-annotated abnormal lesion maps $K \in R^{w \times h}$.

3.2. Spatial Prior Acquisition

Abnormal lesions (GGO and CON) in both CT and CXR scans are annotated to serve as spatial priors. Of note, for 3D CT scans, lung and abnormal lesions are automatically generated by a segmentation model trained in our previous work on a different CT dataset Henao et al. (2023). DRRs along with lung, CON and GGO regions are generated from projections of labeled CT scans Unberath et al. (2018), as shown in Fig. 2.

Abnormal lesions (GGO and CON) for our training CXR scans are automatically generated by a standard U-Net model Ronneberger et al. (2015) trained on a public CXR dataset [2], as shown in Fig. 2.

Though generated lesion annotations are not perfectly correct, they contain useful domain expertise to serve as spatial priors for our unsupervised domain adaptation tasks.

[2]https://github.com/GeneralBlockchain/covid-19-chest-xray-segmentations-dataset

3.3. Intubation Prediction Network

Fig. 1 shows the proposed model. For simplicity, we describe the methods taking labeled CTs (DRRs) as source data and unlabeled CXRs as target data. The same methods apply for when we switch domains, assigning labeled CXRs as source data and unlabeled CTs (DDRs) as target data.

Vision Transformers (ViT) take in a sequence of N image patches prepended by the $[CLS]$-token. The $[CLS]$-token can be utilized for downstream prediction tasks. Denote source domain DRRs as X_s, source domain intubation prediction binary labels as y_s, transformer encoder as F_e, and classifier as F_c. Cross-entropy loss L_{CE} is optimized for this prediction task:

$$L_1 = L_{CE}(F_c(F_e(X_s)), y_s) = -\frac{1}{n_s} \sum_{x_s^i \in X_s} F_c(F_e(x_s^i)) \log y_s^i, \quad (1)$$

where n_s is the number of samples in the source domain.

Self-attention modules are the key to transformer-based vision models Vaswani et al. (2017). Formally, we have a query Q, a key K and a value V calculated from a sequential input patch, and we calculate the attention as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

where d_k is defined as the dimensionality of Q and K and the attention matrix $softmax(\frac{QK^T}{\sqrt{d_k}}) \in R^{N \times N}$. Investigating self-attention, we extract the attention matrix values of each patch with respect to the $[CLS]$-token of the last layer of each attention head and exclude the attention matrix value for the $[CLS]$-token with itself. This tensor can be upsampled into the shape of the original image resulting in an attention map $A_{map} \in R^{w \times h \times n_h}$ where w and h are the dimensions of X and n_h is the number of attention heads.

3.4. Prior Knowledge-Guided Attention Regularization

As mentioned in the previous subsection 3.2 and subsection 3.3, we have an prior expert knowledge binary map $K \in R^{w \times h}$ and attention map $A_{map} \in R^{w \times h \times n_h}$. With $a_{i,j} \in A_{map}$ and $k_{i,j} \in K$, to incorporate the prior information, we consider $k_{i,j} = 1$ to signify that the patch at location i, j is a more important region. Thus, we encode the expertise knowledge into an attention regularization term, defined as

$$\sum_i \sum_j a_{i,j} k_{i,j} - \sum_i \sum_j a_{i,j}(1 - k_{i,j}). \quad (3)$$

In the experiment, we use the above regularization term to regularize one attention head in the transformer and embed the prior information from expert knowledge into the model. It encourages the transformer to pay attention to the important regions across domains.

3.5. Global Transfer Loss

To obtain domain-invariant features for the downstream prediction task, denote a domain discriminator applied to the output state of the class tokens of the source and target images as

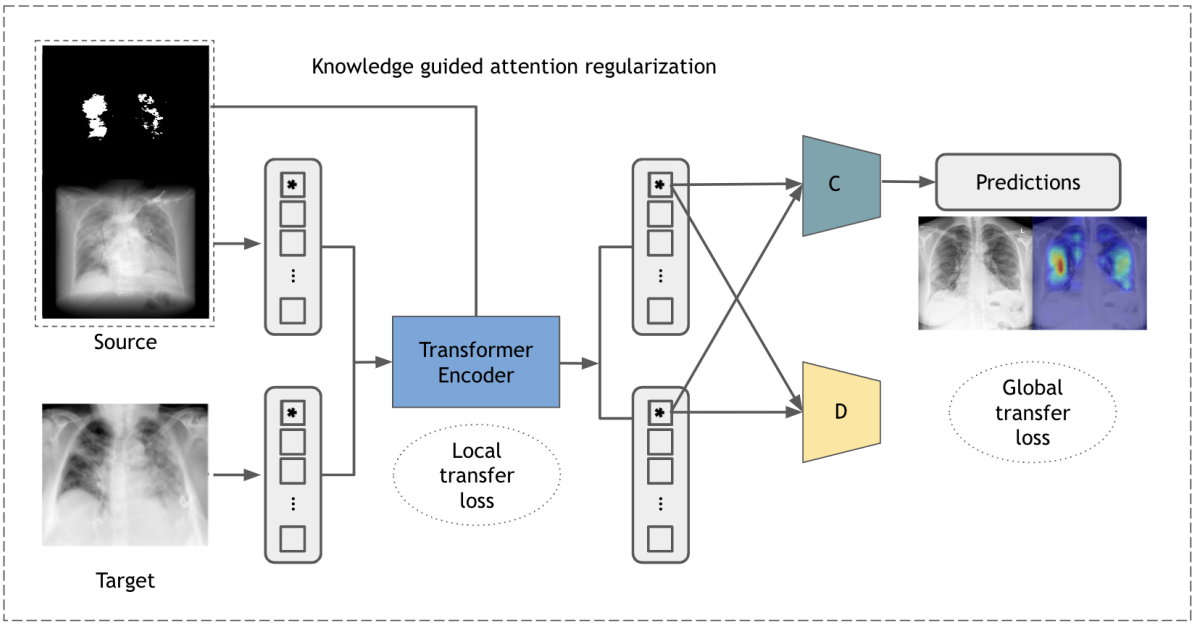


Fig. 1. Prior Knowledge-guided Unsupervised Domain Adaptation (PUDA) framework. It is composed of several individual components, including the transformer backbone, classification head, knowledge guided attention regularization, global transfer loss and weighted local transfer loss.

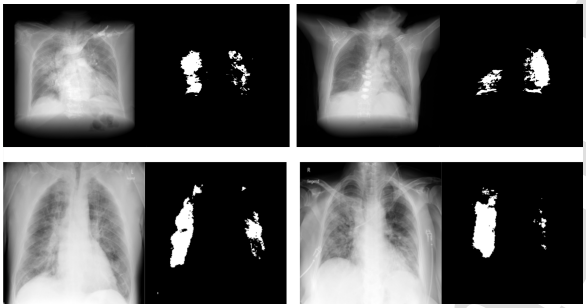


Fig. 2. The first row are two examples of DRR images and abnormal region masks (ground glass opacity and consolidation). They are generated from labeled 3D CTs. The second row are two examples of CXR images with abnormal region masks (ground glass opacity and consolidation).

D_g and target domain CXRs as X_t. The adversarial loss is defined as

$$L_{gt} = -\frac{1}{n_s}\sum_{x_s^i \in X_s} L_{CE}(D_g(F_e(x_s^i)), y_s^i) - \frac{1}{n_t}\sum_{x_t^i \in X_t} L_{CE}(D_g(F_e(x_t^i)), y_t^i) \quad (4)$$

where y_*^i denotes the domain label ($y_s^i = 1$ for source and $y_t^i = 0$ for target) and n_t is the number of samples in the target domain.

3.6. Weighted Local Transfer Loss

The global transfer loss above aims to effectively close global domain gaps. As the learning of the class token embedding depends on the learning of the image token sequence embeddings, we also optimize a local transfer loss to achieve token-wise sequence feature alignment to address the domain shift caused by local texture and style. Each token embedding in the encoder sequence is fed into a domain classifier D_l for adversarial feature alignment. The adversarial local transfer loss is then defined as:

$$L_{lt} = -\frac{1}{n_s N}\sum_{x_s^i \in X_s}\sum_{n \in N} L_{CE}(D_l(F_e(x_s^{in})), y_s^i) - \frac{1}{n_t N}\sum_{x_t^i \in X_t}\sum_{n \in N} L_{CE}(D_l(F_e(x_t^{in})), y_t^i) \quad (5)$$

$n \in N$, where N is the number of fixed patches in each input image.

In addition, tokens contribute differently to the prediction results. Simply aligning tokens across domains can neglect to match the key tokens and make the local transfer less efficient. Motivated by this, we assign higher weights to those tokens during adversarial training according to the attention weight in the regularized attention heads. The weighted local transfer loss is then defined as

$$L_{wlt} = (1 + w_n)L_{lt} \quad (6)$$

where w_n is the weight for the $n-$th token based on the attention weight in the regularized attention heads.

4. Experiment and Evaluation

4.1. Experimental Data

The collected dataset consists of CT and CXR imaging exams and CXR from COVID-19 patients with acute lung disease from three medical centers: Inselspital Bern, University of Bern in Switzerland (IBE), Lindenhofspital Bern in Switzerland (SLB), and Yale New Haven Hospital in the USA (UYA). The clinical data for intubation prediction labels were obtained during routine clinical workup and retrospectively collected and anonymized. The study was approved by the Ethics Commission of the Canton of Bern (ID: 2020-02614, ID: 2020-00954), the Ethics Committee at Yale University (ID: 2000027839), and the Ethics Committee at the University Hospital of Parma (ID: 1398/2020/OSS/AOUPR). All patients in the study gave consent for their data to be used for research. The subjects included in the study had to have a positive COVID-19 PCR test and CT scan or CXR scan. We retrospectively collected patients' medical imaging and clinical data from which a subset of the available cases was selected using the criteria that (1) the patient had a CT scan or CXR scan with clinical lab data available on both day 0 and day 7 of image scan acquisition and (2) clinical lab data used to manually label WHO severity scores indicated the patient had a WHO severity score in the range of 3-6 on day 0 and 3-10 on day 7. Both CT and CXR images were automatically segmented with our deep-learning COVID lung and lesion segmentation models to generate pseudo labels for lung, GGO, and CON. We split the above CT and X-ray scans into the training, validation, and test sets by patient, resulting in 692, 166, and 195 CTs and 601, 132, and 176 X-rays, respectively.

4.2. Spatial Priors

To validate the relevance of abnormal regions in the scans to the intubation prediction task, radiomics features, including first-order, shape, and texture, were extracted from both lung regions and abnormal regions in 3D CT scans. Pyradiomics package was used for implementation Van Griethuysen et al. (2017). These radiomics features were selected and fed into a linear classifier and random forest classifier for the day 7 intubation prediction task. The best performance with an F1 score of 72.7 and an AUC score of 79.8 is achieved by combining GGO and CON radiomics features with a random forest classifier. This result serves as a reference for us to choose these two types of lesions as our prior knowledge for this task.

4.3. Model Implementation Details

In our experiments, we used the base Vision Transformer (ViT-B) Dosovitskiy et al. (2020) with patch size 32 as the backbone of our models. The input image size in our experiments is 224x224. We use the stochastic gradient descent algorithm with weight decay ratio of 1e-4 and momentum of 0.9 to optimize the training process. There are 12 attention heads in total for the vision transformer backbone. The batch size is set to 32. The model is trained on an NVIDIA TITAN RTX GPU. Pytorch 1.8 is used for the implementation of our model. We run models three times with different random seeds.

Table 1. Intubation prediction task results via unsupervised domain adaptation from labeled CTs (DRRs) to unlabeled CXRs

Test on XR	F1	std	AUC	std
ResNet	55.6	3.6	62.8	3.5
ViT	56.0	1.4	65.7	4.2
DANN Ganin and Lempitsky (2015)	61.2	1.8	68.2	2.6
MDD Zhang et al. (2019)	62.1	2.8	67.4	4.3
SCDA Li et al. (2021)	62.4	4.0	68.3	3.4
ViT_{adv}	61.3	3.1	69.3	2.1
TVT Yang et al. (2023)	65.9	1.7	72.2	0.9
CDTrans Xu et al. (2021)	64.6	2.4	71.4	3.4
SSRT Sun et al. (2022b)	69.2	1.5	71.1	1.7
PUDA (Ours)	**69.7**	0.8	**73.4**	2.2

As for other methods for comparison, DANN Ganin and Lempitsky (2015) proposes to play the min-max game with a domain discriminator. MDD Zhang et al. (2019) introduces the margin disparity discrepancy to reduce the distribution discrepancy with a rigorous generalization bound. SCDA Li et al. (2021) encourages the model to focus on the most principal features via the pair-wise adversarial alignment of prediction distributions. The above models use ResNet-50 as the backbone in our experiments. TVT Yang et al. (2023) exploits the transferability of ViT for domain adaptation to extract both transferable and discriminative features. CDTrans Xu et al. (2021) adopts a vision transformer and proposes a two-way center-aware labeling algorithm to produce pseudo labels for target samples. SSRT Sun et al. (2022b) exploits predictions of perturbed target domain data to refine the prediction model. The same vision transformer backbone ViT-B with patch size 32 is used for the above models. Models are implemented with the Pytorch package.

4.4. Results

4.4.1. CTs to CXRs

For the task of unsupervised domain adaptation from CTs (DRRs) to CXRs, a pretrained ViT Dosovitskiy et al. (2020) and a pretrained ResNet50 model He et al. (2016) are finetuned with labeled CXRs for intubation prediction to serve as the supervised learning baselines for comparison. ViT achieves a performance of an F1 score of 70.5 and an AUC score of 72.6 and ResNet achieves a performance of an F1 score of 71.4 and an AUC score of 74.9.

A ViT model finetuned with labeled DRRs and tested on CXRs results in a performance of an F1 score of 55.6 and an AUC of 62.8. In addition, a ViT model with adversarial training using the loss in Sec. 3.5 to simply align the class tokens of the source and the target domains ViT_{adv} is considered another UDA baseline method for comparison. Recent work on UDA with both ResNet and transformer models such as DANN, MDD, SCDA, TVT, CDTrans, and SSRT are also considered for comparison. Our proposed method PUDA outperforms the above methods with an F1 score of 69.7 and an AUC score of 73.4. Please refer to Table 1 for details.

Table 2. Intubation prediction task results via unsupervised domain adaptation from labeled CXRs to unlabeled CTs (DRRs)

Test on DRR	F1	std	AUC	std
ResNet	59.3	1.7	67.9	3.2
ViT	59.0	3.1	70.6	4.4
DANN Ganin and Lempitsky (2015)	61.9	1.2	75.0	2.3
MDD Zhang et al. (2019)	63.8	1.5	78.1	2.4
SCDA Li et al. (2021)	62.6	0.9	77.4	2.1
ViT$_{adv}$	61.5	1.2	77.1	1.3
TVT Yang et al. (2023)	62.9	1.7	76.9	1.8
CDTrans Xu et al. (2021)	63.7	1.4	78.8	1.7
SSRT Sun et al. (2022b)	63.3	0.8	77.1	2.0
PUDA (Ours)	**65.6**	1.3	**80.6**	1.5

Table 3. Domain adaptation ML model compared with two radiologists predicting intubation at seven days from the first-day CXR imaging

Test on CXR	F1	AUC
Radiologist1	67.6	n/a
Radiologist2	64.9	n/a
PUDA (Ours)	69.7	73.4

4.4.2. CXRs to CTs

Similarly, for the task of unsupervised domain adaptation from CXRs to CTs (DRRs), both pretrained ViT and ResNet50 models are finetuned with labeled DRRs for intubation prediction to serve as the supervised learning baselines for comparison. They achieve an F1 score of 67.8 and 69.3 and an AUC score of 83.5 and 84.1, respectively. The vanilla ViT model, baseline ViT$_{adv}$ model, and more recent DANN, MDD, SCDA, TVT, CDTrans and SSRT models are trained for comparison purposes. Our proposed method PUDA achieves the best performance among all methods with an F1 score of 65.6 and an AUC score of 80.6, as shown in Table 2.

4.4.3. Comparison to radiologists

Two experienced radiologists manually perform the same intubation prediction task independently from the machine learning model for comparison. As shown in Table 3, our model performs better in terms of F1 score compared to the two radiologists.

4.5. Model Analysis

As shown in Table 1 and Table 2, our proposed method PUDA performs the best compared to other methods. In addition to the quantitative results, we visualize the attention map from the regularized attention head of our method to have a better understanding of our model. Please refer to Fig. 3 for examples of DRRs, attention maps, and abnormal region masks. It shows that the saliency regions in the generated attention map from the regularized attention head are aligned with the abnormal region masks in the lung regions. During testing, in addition to the prediction results, the model generates attention maps on the test CXRs which reveals the important regions for the model to make the decisions. As shown in Fig. 4, these saliency regions are assumed to be abnormal regions in these CXRs.

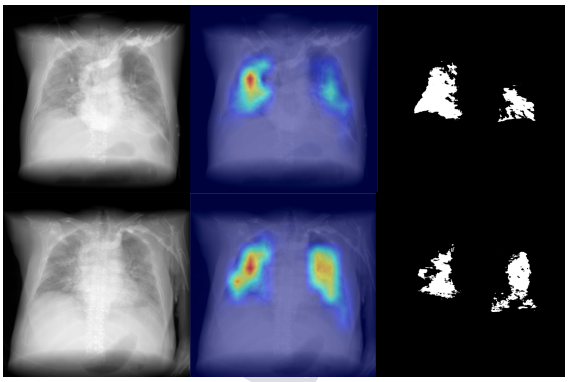


Fig. 3. Two examples of DRRs, attention maps from our PUDA model, and abnormal region masks (ground glass opacity and consolidation).

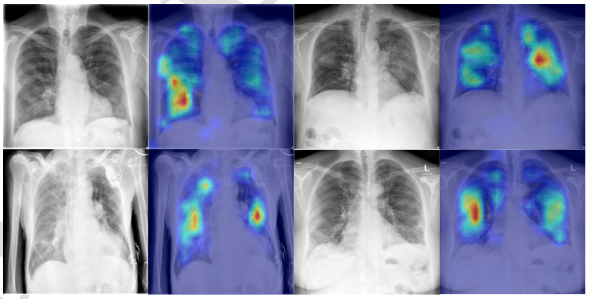


Fig. 4. Four examples of test CXRs and corresponding attention maps from our PUDA model.

For comparison, we also visualize the attention map of the same attention head in the baseline unsupervised domain adaptation model ViT$_{adv}$, where no prior information is used to regularize the attention map. As shown in Fig. 5, without prior guided regularization, the saliency regions in the attention map from the ViT$_{adv}$ model fall outside the lung area.

Furthermore, the attention maps along with the prediction results can be regarded as the interpretation of the prediction model. To evaluate the faithfulness of the model explanation by the attention maps, inspired by the pixel-flipping experiment proposed by Bach et al. (2015), we flip the top 5 percent pixels in input images according to the attention maps and evaluate the impact of these flips on the prediction scores. More specifically, if we flip a pixel, $flipped = pixel \times (-1)$. We use attention maps from PUDA and ViT$_{adv}$ to do the flip individually on the input images. The F1 score and AUC score drops by 13.2% and 19.4% for PUDA and by 7.1% and 8.9% for VIT$_adv$. It shows that the attention maps from PUDA highlight more important regions that are relevant to the prediction task, thus providing a better interpretation for the prediction model.

4.6. Ablation study and analysis

Ablation studies are performed to evaluate the effectiveness of the proposed method. For the task of unsupervised domain

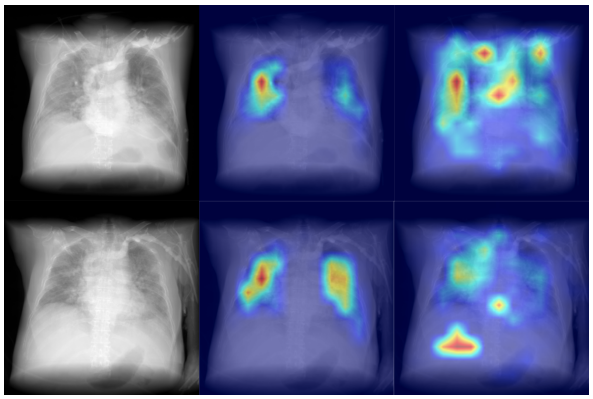


Fig. 5. Two examples of DRRs and generated attention maps from PUDA and ViT$_{adv}$.

Table 4. Ablation analysis

Test on XR	F1	std	AUC	std
V_1	61.3	3.1	69.3	2.1
V_2	67.2	1.8	71.5	2.5
V_3	68.3	2.3	71.9	2.8
PUDA	**69.7**	0.8	**73.4**	2.2

Table 5. Use the whole lung segmentation mask as the prior knowledge map

Test on XR	F1	std	AUC	std
PUDA	69.7	0.8	73.4	2.2
PUDA (lung mask)	67.9	1.3	72.8	1.7

adaptation from CTs (DRRs) to CXRs, we remove the prior knowledge guided regularization and weighted local transfer loss respectively. As shown in Table 4, V_1 denotes the proposed model without prior knowledge guided regularization and weighted local transfer loss, which reduces to the baseline model ViT$_{adv}$. In addition, V_2 denotes the proposed model without the weighted local transfer loss term and V_3 denotes the proposed model without weights for local transfer loss. Each additional component of the proposed model contributes to improved prediction performance, with the full PUDA model performing the best as assessed by both F1 and AUC metrics.

To have a better understanding of the influence on the prediction accuracy from the segmentation accuracy, we first replaced the specific lesion segmentation of GGO and CON with the whole lung segmentation mask. As shown in Table 5, the performance of the proposed PUDA model with the whole lung mask as the prior knowledge map is worse than the original PUDA with GGO and CON lesion masks.

To assess the robustness of the performance against the prior knowledge map, corruption, such as random crop, blur, and elastic transformation, is performed on the prior knowledge segmentation mask. As shown in Table 6, results with blur and elastic transformation as corruption on the prior knowledge map are at the same level as the original PUDA. As for random crop, the performance drops a lot with the random crop 50% corruption. Of noting, random crop x% in the table means to randomly crop from an image of size (h, w) and the cropped region size is (x%h, x%w).

Table 6. Mask corruption on the prior knowledge map

Test on XR	F1	std	AUC	std
Random crop (50 %)	66.2	5.6	70.4	4.8
Random crop (25 %)	69.3	2.7	72.5	2.1
Random crop (10 %)	69.8	2.3	72.7	2.6
Blur	70.3	0.8	73.1	1.4
Elastic transformation	69.5	1.4	73.2	1.6

4.7. Discussion

As shown in the previous sections, our model PUDA outperforms the existing ResNet-based and transformer-based methods on both UDA tasks from CTs (DRRs) to CXRs and from CXRs to CTs (DRRs). An ablation study was performed to evaluate the effectiveness of the proposed components and showed improved model performance with each added component. To further understand the benefits of the proposed method of incorporating prior knowledge, we tested incorporating the prior knowledge map into the CNN approach DANN Ganin and Lempitsky (2015) by adding the segmented image as another input channel. Under the CTs to CXRs UDA task, the DANN-prior model achieves a performance of an F1 score of 65.2 and AUC score of 71.3 while our PUDA model exploits the attention mechanism of transformer models to utilize the shared prior information across domains and improves the UDA performance to an F1 score of 69.7 and 73.4.

While the proposed PUDA model outperforms other baseline and state-of-the-art UDA methods, there are also some limitations to the approach. As the expert anatomical and lesion labels on the training CT scans and CXR scans are generated automatically by deep learning segmentation models, errors by the automated segmentation models could be harmful to the performance of the prediction task. More precise prior knowledge maps with more detailed lesion segmentations might help to further improve the performance of the model.

Both CT and CXR play important roles in the diagnosis and treatment of lung diseases. While CXR is more readily available and can be performed frequently (daily or even multiple times a day) and CT carries more detailed information, they provide complementary characteristics of the same anatomical structures. This work proposes an unsupervised domain adaptation (UDA) tool to transfer knowledge across imaging modalities for clinical prediction. The tool adapts deep learning models from the label-rich source modality (e.g., CT) to the unlabeled target modality (e.g., CXR). By performing a CT scan at approximately the same time as a CXR, the tool can introduce a wealth of information into the CXR interpretation, allowing subsequent CXR interpretations to be made with a higher degree of confidence. On the other hand, the tool can also introduce information from CXR into CT scan interpretation to achieve reasonable performance on the downstream prediction task. The proposed UDA tool aims to alleviate the burden of

data annotation by enabling knowledge transfer across modalities without the need for extensive manual annotation. Additionally, paired CT and CXR datasets are rare, which can pose a significant obstacle in developing effective cross-modal models. The proposed method is designed to work with unpaired CT and CXR data, making it more widely applicable in real-world scenarios. More importantly, it introduces the use of spatial prior knowledge from domain experts to boost UDA performance across domains, which is widely available in medical imaging.

DRRs generated from 3D CTs work well for the prediction task in our paper. However, it is interesting to model 2D to 3D scenarios, e.g. from Xrays to 3D CTs. There exist some works that can perform 2D-to-3D reconstruction Karade and Ravi (2015); Henzler et al. (2018); Jackson et al. (2017). Specifically, Shi et al. (2024) demonstrates the feasibility of reconstructing 3-D lung surfaces from a single 2-D chest x-ray image via a vision transformer. For future work, it could be interesting to explore Xray to CT domain adaptation with a different task that requires 3D CT, e.g. 3D tissue segmentation. In addition, 3D to 3D domain adaptation could be also interesting for certain clinical applications in practice. 3D cross-modality (e.g. between CT and MRI) segmentation has been explored Guo et al. (2023). There are other cross-modality clinical tasks between 3D imaging, such as the assessment of respiratory disease in cystic fibrosis (CF) with chest imaging (between CT and MRI) . For patients with CF, 3D CT provides higher resolution than 3D MRI while MRI is radiation free.

The proposed model is designed to to introduce the prior knowledge for unsupervised domain adaptation and evaluated on the clinical outcome prediction downstream task. For future work, on one hand, there are other relevant clinical applications such as object detection, disease quantification, and segmentation, and it is interesting to evaluate our proposed model on various different downstream tasks and observe the generalization ability. On the other hand, there are other types of prior knowledge besides the spatial prior knowledge such as class ratio of the target domain, it is also worth exploring to generalize the proposed model to other types of prior knowledge.

5. Conclusion

In this paper, we study how to exploit prior knowledge for unsupervised domain adaptation in medical image analysis and evaluate our method on a clinical outcome prediction task. Specifically, anatomical and spatial priors across domains are embedded into the regularized attention heads of vision transformers. Besides the global alignment of class tokens, it guides sequential feature alignment via local weights. The experimental results for the task of intubation prediction show the effectiveness of the proposed methods.

Acknowledgments

The authors thank the support provided by the Swiss National Science Foundation through the National Research Programme ”COVID-19” (NRP 78) under grant number 198388, as well as Campus Stiftung Lindenhof Bern and the Swiss Institute for Translational and Entrepreneurial Medicine for their valuable support throughout this research project.

References

Alghamdi, H.S., Amoudi, G., Elhag, S., Saeedi, K., Nasser, J., 2021. Deep learning approaches for detecting covid-19 from chest x-ray images: A survey. Ieee Access 9, 20235–20254.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one 10, e0130140.

Bateson, M., Lombaert, H., Ben Ayed, I., 2022. Test-time adaptation with shape moments for image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 736–745.

Beal, J., Kim, E., Tzeng, E., Park, D.H., Zhai, A., Kislyuk, D., 2020. Toward transformer-based object detection. arXiv preprint arXiv:2012.09958 .

Bhatele, K.R., Jha, A., Tiwari, D., Bhatele, M., Sharma, S., Mithora, M.R., Singhal, S., 2022. Covid-19 detection: A systematic review of machine and deep learning-based approaches utilizing chest x-rays and ct scans. Cognitive Computation , 1–38.

Bigalke, A., Hansen, L., Diesel, J., Hennigs, C., Rostalski, P., Heinrich, M.P., 2023. Anatomy-guided domain adaptation for 3d in-bed human pose estimation. Medical Image Analysis 89, 102887.

Carbonell, R., Urgelés, S., Rodríguez, A., Bodí, M., Martín-Loeches, I., Solé-Violán, J., Díaz, E., Gómez, J., Trefler, S., Vallverdú, M., et al., 2021. Mortality comparison between the first and second/third waves among 3,795 critical covid-19 patients with pneumonia admitted to the icu: A multicentre retrospective cohort study. The Lancet Regional Health–Europe 11.

Chamberlin, J.H., Aquino, G., Schoepf, U.J., Nance, S., Godoy, F., Carson, L., Giovagnoli, V.M., Gill, C.E., McGill, L.J., O'Doherty, J., et al., 2022. An interpretable chest ct deep learning algorithm for quantification of covid-19 lung disease and prediction of inpatient morbidity and mortality. Academic Radiology 29, 1178–1188.

Chen, C.F.R., Fan, Q., Panda, R., 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 357–366.

Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., Shen, C., 2021. Twins: Revisiting the design of spatial attention in vision transformers. Advances in Neural Information Processing Systems 34, 9355–9366.

Dalca, A.V., Guttag, J., Sabuncu, M.R., 2018. Anatomical priors in convolutional networks for unsupervised biomedical segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9290–9299.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

Fang, Y., Liao, B., Wang, X., Fang, J., Qi, J., Wu, R., Niu, J., Liu, W., 2021. You only look at one sequence: Rethinking transformer in vision through object detection. Advances in Neural Information Processing Systems 34, 26183–26197.

Ganin, Y., Lempitsky, V., 2015. Unsupervised domain adaptation by backpropagation, in: International conference on machine learning, PMLR. pp. 1180–1189.

Gu, J., Kwon, H., Wang, D., Ye, W., Li, M., Chen, Y.H., Lai, L., Chandra, V., Pan, D.Z., 2022. Multi-scale high-resolution vision transformer for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12094–12103.

Guner, R., Kayaaslan, B., Hasanoglu, I., Aypak, A., Bodur, H., Ates, I., Akinci, E., Erdem, D., Eser, F., Izdes, S., et al., 2021. Development and validation of nomogram to predict severe illness requiring intensive care follow up in hospitalized covid-19 cases. BMC Infectious Diseases 21, 1–13.

Guo, S., Liu, X., Zhang, H., Lin, Q., Xu, L., Shi, C., Gao, Z., Guzzo, A., Fortino, G., 2023. Causal knowledge fusion for 3d cross-modality cardiac image segmentation. Information Fusion 99, 101864.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D., 2021. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, in: International MICCAI Brainlesion Workshop, Springer. pp. 272–284.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Henao, J.A.G., Depotter, A., Bower, D.V., Bajercius, H., Todorova, P.T., Saint-James, H., de Mortanges, A.P., Barroso, M.C., He, J., Yang, J., et al., 2023. A multiclass radiomics method–based who severity scale for improving covid-19 patient assessment and disease characterization from ct scans. Investigative radiology , 10–1097.

Henzler, P., Rasche, V., Ropinski, T., Ritschel, T., 2018. Single-image tomography: 3d volumes from 2d cranial x-rays, in: Computer Graphics Forum, Wiley Online Library. pp. 377–388.

Huang, K., Cheng, H.D., Zhang, Y., Zhang, B., Xing, P., Ning, C., 2018. Medical knowledge constrained semantic breast ultrasound image segmentation, in: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE. pp. 1193–1198.

Inui, S., Gonoi, W., Kurokawa, R., Nakai, Y., Watanabe, Y., Sakurai, K., Ishida, M., Fujikawa, A., Abe, O., 2021. The role of chest imaging in the diagnosis, management, and monitoring of coronavirus disease 2019 (covid-19). Insights into imaging 12, 1–14.

Jackson, A.S., Bulat, A., Argyriou, V., Tzimiropoulos, G., 2017. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression, in: Proceedings of the IEEE international conference on computer vision, pp. 1031–1039.

Ji, W., Chung, A.C., 2023. Unsupervised domain adaptation for medical image segmentation using transformer with meta attention. IEEE Transactions on Medical Imaging .

Karade, V., Ravi, B., 2015. 3d femur model reconstruction from biplane x-ray images: a novel method based on laplacian surface deformation. International journal of computer assisted radiology and surgery 10, 473–485.

Kwon, Y.J., Toussie, D., Finkelstein, M., Cedillo, M.A., Maron, S.Z., Manna, S., Voutsinas, N., Eber, C., Jacobi, A., Bernheim, A., et al., 2020. Combining initial radiographs and clinical variables improves deep learning prognostication in patients with covid-19 from the emergency department. Radiology: Artificial Intelligence 3, e200098.

Li, S., Xie, M., Lv, F., Liu, C.H., Liang, J., Qin, C., Li, W., 2021. Semantic concentration for domain adaptation, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 9102–9111.

Li, Y., Mao, H., Girshick, R., He, K., 2022a. Exploring plain vision transformer backbones for object detection, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX, Springer. pp. 280–296.

Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C., 2022b. Mvitv2: Improved multiscale vision transformers for classification and detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4804–4814.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88.

Marshall, J.C., Murthy, S., Diaz, J., Adhikari, N., Angus, D.C., Arabi, Y.M., Baillie, K., Bauer, M., Berry, S., Blackwood, B., et al., 2020. A minimal common outcome measure set for covid-19 clinical research. The Lancet Infectious Diseases 20, e192–e197.

Miao, K., Gokul, A., Singh, R., Petryk, S., Gonzalez, J., Keutzer, K., Darrell, T., 2022. Prior knowledge-guided attention in self-supervised vision transformers. arXiv preprint arXiv:2209.03745 .

Nakashima, M., Uchiyama, Y., Minami, H., Kasai, S., 2023. Prediction of covid-19 patients in danger of death using radiomic features of portable chest radiographs. Journal of Medical Radiation Sciences 70, 13–20.

Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., De Marvao, A., Dawes, T., O'Regan, D.P., et al., 2017. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. IEEE transactions on medical imaging 37, 384–395.

Qin, C., Yao, D., Shi, Y., Song, Z., 2018. Computer-aided detection in chest radiography based on artificial intelligence: a survey. Biomedical engineering online 17, 1–23.

Raza, K., Singh, N.K., 2021. A tour of unsupervised deep learning for medical image analysis. Current Medical Imaging 17, 1059–1077.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer. pp. 234–241.

Schaefer-Prokop, C., Prokop, M., 2021. Chest radiography in covid-19: no role in asymptomatic and oligosymptomatic disease.

Serte, S., Demirel, H., 2021. Deep learning for diagnosis of covid-19 using 3d ct scans. Computers in biology and medicine 132, 104306.

Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annual review of biomedical engineering 19, 221–248.

Shi, Z., Geng, K., Zhao, X., Mahmoudi, F., Haas, C.J., Leader, J.K., Duman, E., Pu, J., 2024. Xraywizard: Reconstructing 3-d lung surfaces from a single 2-d chest x-ray image via vision transformer. Medical Physics 51, 2806–2816.

Strudel, R., Garcia, R., Laptev, I., Schmid, C., 2021. Segmenter: Transformer for semantic segmentation, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 7262–7272.

Sun, B., Saenko, K., 2016. Deep coral: Correlation alignment for deep domain adaptation, in: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14, Springer. pp. 443–450.

Sun, T., Lu, C., Ling, H., 2022a. Prior knowledge guided unsupervised domain adaptation, in: European Conference on Computer Vision, Springer. pp. 639–655.

Sun, T., Lu, C., Zhang, T., Ling, H., 2022b. Safe self-refinement for transformer-based domain adaptation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7191–7200.

Tzeng, E., Hoffman, J., Saenko, K., Darrell, T., 2017. Adversarial discriminative domain adaptation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7167–7176.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T., 2014. Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 .

Unberath, M., Zaech, J.N., Lee, S.C., Bier, B., Fotouhi, J., Armand, M., Navab, N., 2018. Deepdrr–a catalyst for machine learning in fluoroscopy-guided procedures, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11, Springer. pp. 98–106.

Van Engelen, J.E., Hoos, H.H., 2020. A survey on semi-supervised learning. Machine learning 109, 373–440.

Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J., 2017. Computational radiomics system to decode the radiographic phenotype. Cancer research 77, e104–e107.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Wang, M., Deng, W., 2018. Deep visual domain adaptation: A survey. Neurocomputing 312, 135–153.

Wang, Z., Dai, Z., Póczos, B., Carbonell, J., 2019. Characterizing and avoiding negative transfer, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11293–11302.

Wilson, G., Cook, D.J., 2020. A survey of unsupervised deep domain adaptation. ACM Transactions on Intelligent Systems and Technology (TIST) 11, 1–46.

Xu, T., Chen, W., Wang, P., Wang, F., Li, H., Jin, R., 2021. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. arXiv preprint arXiv:2109.06165 .

Yang, J., Liu, J., Xu, N., Huang, J., 2023. Tvt: Transferable vision transformer for unsupervised domain adaptation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 520–530.

Yao, Y., Liu, F., Zhou, Z., Wang, Y., Shen, W., Yuille, A., Lu, Y., 2022. Unsupervised domain adaptation through shape modeling for medical image segmentation, in: International Conference on Medical Imaging with Deep Learning, PMLR. pp. 1444–1458.

Zhang, F., Li, S., Deng, J., 2022. Unsupervised domain adaptation with shape constraint and triple attention for joint optic disc and cup segmentation. Sensors 22, 8748.

Zhang, Y., Liu, T., Long, M., Jordan, M., 2019. Bridging theory and algorithm for domain adaptation, in: International conference on machine learning, PMLR. pp. 7404–7413.

Zhou, Y., Li, Z., Bai, S., Wang, C., Chen, X., Han, M., Fishman, E., Yuille, A.L., 2019. Prior-aware neural network for partially-supervised multi-organ segmentation, in: Proceedings of the IEEE/CVF international conference on

computer vision, pp. 10672–10681.

Highlights

(1) We propose a transformer-based UDA framework that utilizes shared anatomical and spatial priors across domains for medical image analysis.

(2) The proposed method effectively regularizes the attention heads in the vision transformer guided by prior knowledge and improves both the discriminability and transferability of the learned features. Besides the global alignment of class tokens, the regularized attention guides the adversarial alignment of the distribution of sequential features.

(3) Our knowledge-guided model incorporating domain expertise produces a more reasonable attention map along with the prediction results, thus leading to a better understanding of the deep learning model.

(4) In addition, the proposed model performs favorably compared with human radiologists on the same intubation prediction task.

Declaration of interests

☒ 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.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: