



Domain Generalization with Small Data

Kecheng Chen¹ · Elena Gal² · Hong Yan¹ · Haoliang Li¹

Received: 19 May 2023 / Accepted: 8 February 2024 / Published online: 6 March 2024
© The Author(s) 2024

Abstract

In this work, we propose to tackle the problem of domain generalization in the context of *insufficient samples*. Instead of extracting latent feature embeddings based on deterministic models, we propose to learn a domain-invariant representation based on the probabilistic framework by mapping each data point into probabilistic embeddings. Specifically, we first extend empirical maximum mean discrepancy (MMD) to a novel probabilistic MMD that can measure the discrepancy between mixture distributions (i.e., source domains) consisting of a series of latent distributions rather than latent points. Moreover, instead of imposing the contrastive semantic alignment (CSA) loss based on pairs of latent points, a novel probabilistic CSA loss encourages positive probabilistic embedding pairs to be closer while pulling other negative ones apart. Benefiting from the learned representation captured by probabilistic models, our proposed method can marriage the measurement on the *distribution over distributions* (i.e., the global perspective alignment) and the distribution-based contrastive semantic alignment (i.e., the local perspective alignment). Extensive experimental results on three challenging medical datasets show the effectiveness of our proposed method in the context of insufficient data compared with state-of-the-art methods.

Keywords Domain generalization · Healthcare · Small data · Medical imaging

1 Introduction

Nowadays, we have witnessed a lot of successes with the application of machine learning techniques in a variety of tasks related to computer vision (Li et al., 2022; Zaidi et al., 2022) and natural language processing (Mridha et al., 2022). Despite many achievements so far, the widely-adopted assumption for most existing methods, i.e., the data are iden-

tically and independently distributed in training and testing, may not always hold in actual applications (Zhou et al., 2022; Liu et al., 2022). In the real-world scenario, it is quite common that the distributions between training and testing data may be different, owing to changed environments. For example, acquired histopathological images of breast cancer from different healthcare centers exhibit significant domain gaps (*a.k.a.*, domain shift, see Fig. 1 for more detail) caused by differences in device vendors and staining methods, which may lead to the catastrophic deterioration of the performance (Qi et al., 2020). To address this issue, *domain generalization* (DG) is developed to learn a model from multiple related yet different domains (*a.k.a.*, source domains) that is able to generalize well on unseen testing domain (*a.k.a.*, target domain).

Recently, researchers proposed several domain generalization approaches, such as data augmentation with randomization (Yue et al., 2019), data generalization with stylization (Verma et al., 2019; Zhou et al., 2021), meta learning-based training schemes (Li et al., 2018; Kim et al., 2021), among which representation learning-based methods are one of the most popular ones. These representation learning-based methods (Balaji et al., 2019) aim to learn domain-invariant feature representation. To be specific, if the discrepancy

Communicated by Jingdong Wang.

✉ Haoliang Li
haoliang.li@cityu.edu.hk
Kecheng Chen
kechuchen3-c@my.cityu.edu.hk
Elena Gal
elena.gal@maths.ox.ac.uk
Hong Yan
h.yan@cityu.edu.hk

¹ Department of Electrical Engineering and Center for Intelligent Multidimensional Data Analysis, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong SAR

² Department of Engineering, University of Oxford, Old Road Campus Research Building, Roosevelt Drive, Oxford OX3 7DQ, UK

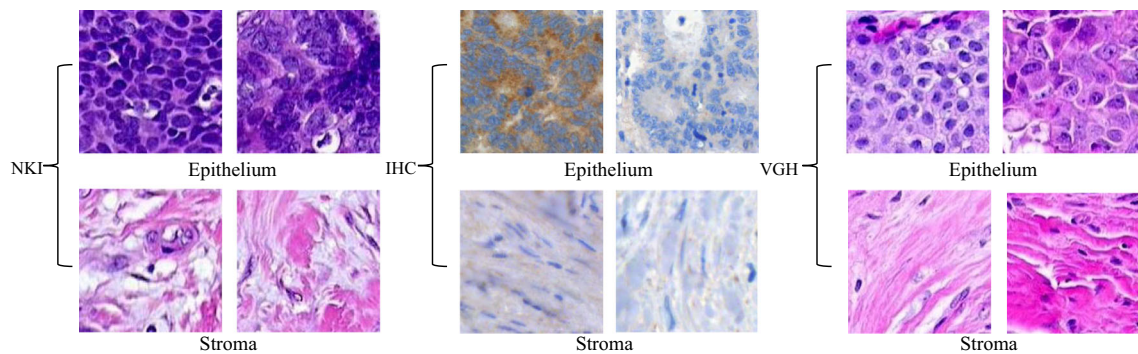


Fig. 1 Histopathological image examples of breast cancer tissue from three different healthcare institutes, including NKI with 626 images, IHC with 645 images, and VGH with 1324 images. There are two dif-

ferent tissue types, including epithelium and stroma. Obvious domain gaps (e.g., the density of tissue and the staining color) can be observed

between source domains in feature space can be minimized, the model is expected to generalize better on unseen target domain, due to learned domain-invariant and transferable feature representation (Ben-David et al., 2006). For instance, an classical contrastive semantic alignment (CSA) loss proposed by Motiian et al. (2017) was to encourage positive sample pairs (with same label) from different domains closer while pulling other negative pairs (with different labels) apart. Dou et al. (2019) introduced the CSA loss which jointly considers *local class alignment loss* (for point-wise domain alignment) and *global class alignment loss* (for distribution-wise alignment).

Despite the progress so far, a reliable contrastive semantic loss with point-wise (or local) perspective usually requires sufficient samples on source domains such that diverse sample-to-sample pairs can be constructed (Sohn, 2016; Khosla et al., 2020). For example, Khosla et al. (2020) proposed a supervised contrastive semantic loss with a considerable volume of batch size on large-scale datasets such that decent performance can be guaranteed. Yao et al. (2022) also emphasized the importance of the number of sample-to-sample pairs influenced by data sizes for contrastive-based loss on DG problem. On the other hand, in the eye of distribution-wise (*a.k.a.*, global) alignment between domains (Dou et al., 2019), a consistent distribution measurement (e.g., Kullback–Leibler (KL) divergence) theoretically relies on sufficient samples for the distribution estimation as discussed by (Bu et al., 2018). However, these sufficient samples from multiple source domains may not always be *available* or *accessible* in the real world. For example, for the medical imaging data, insufficient sample scenarios either exist in *all source domains* (e.g., rare diseases inherently have a small volume of data from all healthcare centers (Lee et al., 2022)) or in *some source domains* (e.g., some specific domains have significantly smaller sample sizes than others, resulting from the differences of the ethnicity (Johnson &

Louis, 2022), the demography (Gurdasani et al., 2019), and the privacy-preserving regulation (Can & Ersoy, 2021)). It is therefore necessary to develop reliable and effective semantic alignments from both local and global perspectives in the context of insufficient samples (*a.k.a.*, small-data scenario) based on the source domains, in order to achieve better domain-invariant representations.

In this paper, we propose to learn domain-invariant representation from multiple source domains to tackle the domain generalization problem in the context of *insufficient samples*. Instead of extracting latent embeddings (i.e., latent points) based on deterministic models (e.g., convolutional neural networks, CNNs), we propose to leverage a probabilistic framework endowed by variational Bayesian inference to map each data point into probabilistic embeddings (i.e., the latent distribution) for domain generalization. Specifically, by following the domain-invariant learning from global (distribution-wise) perspective, we propose to extend empirical maximum mean discrepancy (MMD) to a novel probabilistic MMD (P-MMD) that can empirically measure the discrepancy between mixture distributions (*a.k.a.*, *distributions over distributions*), consisted of a serial of latent distributions rather than latent points. From a local perspective, instead of imposing the CSA loss based on pairs of latent points, a novel probabilistic contrastive semantic alignment (P-CSA) loss with kernel mean embedding is proposed to encourage positive probabilistic embedding pairs closer while pulling other negative ones apart. Extensive experimental results on three challenging medical imaging classification tasks, including epithelium stroma classification on insufficient histopathological images, skin lesion classification, and spinal cord gray matter segmentation, show that our proposed method can achieve better cross-domain performance in the context of insufficient data compared with state-of-the-art methods.

2 Related Works

2.1 Domain Generalization with Medical Images

Existing DG methods can be generally categorized into three different streams, namely data augmentation/generation (Yue et al., 2019; Graves, 2011; Zhou et al., 2021), meta-learning (Li et al., 2018; Kim et al., 2021) and feature representation learning (Li et al., 2018; Gong et al., 2019; Xiao et al., 2021). Among these methods, feature representation learning, which aims to explore invariant feature information that can be shared across domains, demonstrates to be a widely adopted method for the problem of DG. For the feature representation learning-based DG method, (Li et al., 2018) proposed to conduct multi-domain alignment in latent space via a multi-domain MMD distance. Gong et al. (2019) leveraged adversarial training to eliminate the domain discrepancy such that domain-invariant representation can be learned in a manifold space. Due to the varieties of imaging protocol (e.g., the choice of image solution for MRI image), device vendors (e.g., Philips or Siemens CT scanners), and patient populations (the race and age group), the acquired imaging data from different medical sites may exist significant domain shift problem (Liu et al., 2021). Dou et al. (2019) proposed a meta-learning framework to perform local and global semantic alignment for medical image classification. A similar design is also adopted by Li et al. (2022) for tissue image classification. Qi et al. (2020) utilized the curriculum learning scheme to transfer the knowledge for histopathological images classification. Li et al. (2020) combined the data augmentation and domain alignment to achieve decent performance on multiple medical data classification tasks. However, these methods may not focus on learning domain-invariant representation on *insufficient samples* from source domains.

2.2 Probabilistic Neural Networks

Compared with deterministic models, probabilistic neural networks turns to learn a distribution over model parameters, which can integrate the uncertainty in predictive modeling (Kingma et al., 2015; Gal & Ghahramani, 2016). When the data is insufficient, probabilistic models usually can achieve better generalized performance due to its probabilistic property (as an implicit regularization) (Blundell et al., 2015). In the context of insufficient samples, Bayesian neural network (Neal, 2012) (BNN) with variational inference, a representative probabilistic model, not only can improve predictive accuracy as a classifier (Wilson & Izmailov, 2020), but also can build up the quality of low-dimensional embeddings of insufficient data (Mallick et al., 2021), which is a crucial motivation for this paper. Meanwhile, modern analytical approximation techniques (e.g., Variational inference (Blei et

al., 2017), empirical Bayes (Krishnan et al., 2020)) can efficiently infer the posterior distribution of model parameters with stochastic gradient descent method, which can integrate BNN with deterministic DNN conveniently.

In Xiao et al. (2021), the authors proposed to consider the uncertainty of a generalizable model based on BNN, where the distances of positive probabilistic embedding pairs and class distribution are minimized via KL measure. Despite the effectiveness, the dissimilar pairs (i.e., negative pairs) are ignored, which may not benefit feature representation learning. Moreover, they only focused on sample similarity while the distribution information is ignored. Instead, our proposed method comprehensively considers both positive and negative probabilistic embedding pairs via a novel distribution-based contrastive semantic loss. Last but not the least, our proposed method highlights the benefit of the BNN for building up the quality of latent embeddings under insufficient sample scenarios.

2.3 Probabilistic Embedding

Compared with deterministic point embeddings, probabilistic embeddings aim to characterize the data with a distribution. Due to its high robustness and effective representation (Nguyen et al., 2017), probabilistic representation has been applied to several fields, such as video representation learning (Park et al., 2022), image representation learning (Oh et al., 2018), face recognition (Shi & Jain, 2019; Chang et al., 2020), speaker diarization (Silnova et al., 2020), and human pose estimation (Sun et al., 2020). Recently, some researchers further leveraged the probabilistic embeddings to bridge the gap between data modalities (Chun et al., 2021; Neculai et al., 2022; Chun, 2023). For example, Chun et al. (2021) found that the probabilistic representation can lead to a richer embedding space for the challengeable relation reasoning between the images and their captions. These probabilistic embedding-based approaches either inherit the inherent distribution property of data (e.g., the multiple frames of a video) or tackle the one-to-many correspondences through distributional representation. Instead, our proposed method imposes the Bayesian neural network to generate probabilistic embedding. As such, the representative capacity of the data in the small-data regime can be enhanced. Moreover, we devise a novel probabilistic MMD to measure the discrepancy between mixture probabilistic embeddings for domain-invariant learning.

3 Methodology

Preliminary Assume that there are K domains from different collected environments. The samples in each domain can be represented as $\mathbf{X}_l = \{\mathbf{x}_{l_1}, \dots, \mathbf{x}_{l_{n_l}}\}$, where $l \in$

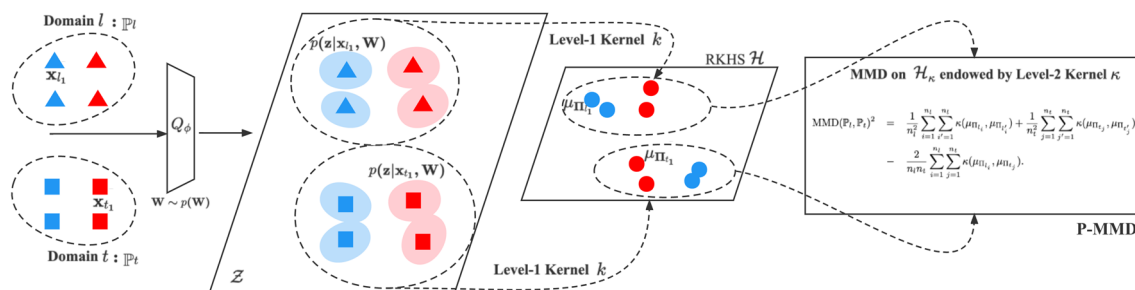


Fig. 2 A visualized computational process for probabilistic MMD (P-MMD) on two source domains. The same color for samples in different domains denotes the same label

$\mathbb{N}^+ : \{1, \dots, K\}$, $\mathbf{x}_{l_i} \in \mathbb{R}^{d \times 1}$ denotes a sample with the d dimension vector in the l -th domain. n_l is the total number of samples in the l -th domain. The corresponding labels of samples \mathbf{X}_l in each domain can be denoted as $\mathbf{Y}_l = \{y_{l_1}, \dots, y_{l_{n_l}}\}$, where $y_{l_i} \in \mathbb{R}^{m \times 1}$ is the form of one-hot encoding with m classes in total. For the setting of domain generalization, the source domain data represented as $\{\mathbf{X}_l^S, \mathbf{Y}_l^S\}_{l=1}^K$, can be available in the training phase only, whereas the target domain data, denoted by \mathbf{X}^T , are only seen in test phase.

Overall We provide a framework that can learn better domain-invariant representation when there is insufficient source domain data. The probabilistic neural network is imposed to enable high-quality and powerful feature representation in the context of insufficient samples. To effectively perform global perspective alignment, a novel probabilistic MMD is proposed to empirically measure the discrepancy between distributions over distributions based on reproducing kernel Hilbert space. We also propose a probabilistic contrastive semantic alignment to adapt probabilistic embeddings with local perspective. The details of our proposed method are discussed as below.

Probabilistic Embedding of Insufficient Data Compared with deterministic models, the probabilistic models can learn a distribution over model weights, which has shown a better capacity to represent latent embeddings under insufficient sample scenario (Mallick et al., 2021). In this work, Bayesian neural network (BNN) (Blei et al., 2017) is utilized to extract the low-dimensional embeddings from high-dimensional inputs. By feeding the inputs into BNN with a parameter $\mathbf{W} \sim p(\mathbf{W})$, the samples $\mathbf{X}_l = \{\mathbf{x}_{l_1}, \dots, \mathbf{x}_{l_{n_l}}\}$ of each domain can be represented by a set of probabilistic embeddings (i.e., latent distributions), i.e., $p(\mathbf{Z}|\mathbf{X}_l) = \{p(\mathbf{z}|\mathbf{x}_{l_1}, \mathbf{W}), \dots, p(\mathbf{z}|\mathbf{x}_{l_{n_l}}, \mathbf{W})\}$ where $\mathbf{W} \sim p(\mathbf{W})$ is sampled stochastically. The variational inference is used to approximate the posterior distribution of \mathbf{W} with the evidence lower bound (ELBO) (more details can be found in ‘‘Appendix A.1’’). By using Monte Carlo (MC) estimators with T stochastic sampling operations from \mathbf{W} , the predictive distribution of each $p(\mathbf{z}|\mathbf{x})$ can be an unbiased approximation.

3.1 Distribution Alignment via Probabilistic Maximum Mean Discrepancy

In this section, we introduce an approach to learning domain-invariant representation from a global perspective by minimizing the discrepancy among domains. Among various distribution distance metrics, Maximum Mean Discrepancy (MMD) is widely adopted (Long et al., 2017; Li et al., 2018) which aims to measure the distance between two probability distributions in a non-parametric manner. Specifically, assume that latent embeddings $\mathbf{Z}_l = \{\mathbf{z}_{l_1}, \dots, \mathbf{z}_{l_{n_l}}\}$ and $\mathbf{Z}_t = \{\mathbf{z}_{t_1}, \dots, \mathbf{z}_{t_{n_t}}\}$ are drawn from two unknown distributions \mathbb{P}_l and \mathbb{P}_t . The probability measure \mathbb{P} can be mapped into a reproducing kernel Hilbert space (RKHS) \mathcal{H} as a element by setting,

$$\mu_{\mathbb{P}} := \mathbb{E}_{\mathbf{z} \sim \mathbb{P}}[\phi(\mathbf{z})] = \int_{\mathcal{Z}} k(\mathbf{z}, \cdot) d\mathbb{P} = \mathbb{E}_{\mathbf{z} \sim \mathbb{P}}[k(\mathbf{z}, \cdot)], \quad (1)$$

where a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ and corresponding feature map $\phi : \mathcal{X} \rightarrow \mathcal{H}$ are defined. Let the kernel k is characteristic such that the map $\mu : \mathbb{P} \rightarrow \mu_{\mathbb{P}}$ is injective. In this case the MMD can be defined as the distance $\|\mu_{\mathbb{P}_l} - \mu_{\mathbb{P}_t}\|_{\mathcal{H}}$ in \mathcal{H} between mean embeddings and it can be used as a measure of distance between the distributions \mathbb{P}_l and \mathbb{P}_t (Borgwardt et al., 2006; Gretton et al., 2012). The explicit computation of MMD can be derived by unbiased empirical estimation of mean map (Gretton et al., 2012), i.e.,

$$\text{MMD}(\mathbb{P}_l, \mathbb{P}_t)^2 = \left\| \frac{1}{n_l} \sum_{i=1}^{n_l} \phi(\mathbf{z}_{l_i}) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi(\mathbf{z}_{t_j}) \right\|_{\mathcal{H}}^2 \quad (2)$$

The idea of using MMD for domain generalization has been explored in several works (e.g., (Li et al., 2018; Hu et al., 2020)).

In the probabilistic framework, instead of the individual latent embeddings \mathbf{z}_{l_1}, \dots , we have latent probabilistic embeddings $\Pi_{l_1} := p(\mathbf{z}|\mathbf{x}_{l_1}, \mathbf{W}), \dots$. For a source domain D_l , we have the associated *distribution over distributions* $\mathbb{P}_l = \{\Pi_{l_1}, \dots, \Pi_{l_{n_l}}\}$. For this scenario, we propose to extend the existing *point-based* empirical MMD estimate to

a *distribution-based* empirical probability MMD (P-MMD) estimate. P-MMD utilizes empirical estimation by kernels on distributions to measure the discrepancy between mixture distributions \mathbb{P}_l and \mathbb{P}_t under the probabilistic framework.

Specifically, we first represent latent probabilistic embeddings as elements in RKHS \mathcal{H}_k using the kernel k , that we coin a *level-1* kernel in the sequel, e.g., $\mu_{\Pi_{l_1}} := \mathbb{E}_{\mathbf{z} \sim \Pi_{l_1}}[\phi(\mathbf{z})] = \mathbb{E}_{\mathbf{z} \sim \Pi_{l_1}}[k(\mathbf{z}, \cdot)]$, which is an analog to the Eq. (1). The kernel mean embedding $\mu_{\Pi_{l_1}}$ can be regarded as a new feature map for a variety of tasks (Yoshikawa et al., 2014). Here, to enable *non-linear* learning on distributions, we introduce a *level-2* kernel K (Muandet et al., 2012). Consider a level-1 kernel κ on \mathcal{H} and its reproducing kernel Hilbert space (RKHS) \mathcal{H}_κ . Define K as

$$K(\Pi_{l_i}, \Pi_{t_j}) = \kappa(\mu_{\Pi_{l_i}}, \mu_{\Pi_{t_j}}) = \langle \psi(\mu_{\Pi_{l_i}}), \psi(\mu_{\Pi_{t_j}}) \rangle_{\mathcal{H}_\kappa}, \quad (3)$$

where K and its explicit form on kernel mean embeddings κ are p.d. kernels (Berlinet & Thomas-Agnan, 2011). We define a novel *probabilistic MMD* (P-MMD) empirical estimation method using the *level-2* kernel K :

$$\begin{aligned} \text{P-MMD}(\mathbb{P}_l, \mathbb{P}_t)^2 &= \left\| \frac{1}{n_l} \sum_{i=1}^{n_l} \psi(\mu_{\Pi_{l_i}}) - \frac{1}{n_t} \sum_{j=1}^{n_t} \psi(\mu_{\Pi_{t_j}}) \right\|_{\mathcal{H}_\kappa}^2 \\ &= \frac{1}{n_l^2} \sum_{i=1}^{n_l} \sum_{i'=1}^{n_l} K(\Pi_{l_i}, \Pi_{l_{i'}}) + \frac{1}{n_t^2} \sum_{j=1}^{n_t} \sum_{j'=1}^{n_t} K(\Pi_{t_j}, \Pi_{t_{j'}}) \\ &\quad - \frac{2}{n_l n_t} \sum_{i=1}^{n_l} \sum_{j=1}^{n_t} K(\Pi_{l_i}, \Pi_{t_j}). \end{aligned} \quad (4)$$

In this work, the level-1 and level-2 kernels, k and K , are both Gaussian RBF kernel due to its impressive performance on a limited amount of distribution data (Muandet et al., 2012). Namely, $K = K_{Gau}(\Pi_{l_i}, \Pi_{t_j}) = \kappa(\mu_{\Pi_{l_i}}, \mu_{\Pi_{t_j}})$ can be represented as

$$\begin{aligned} \kappa(\mu_{\Pi_{l_i}}, \mu_{\Pi_{t_j}}) &= \exp\left(-\frac{\lambda}{2} \|\mu_{\Pi_{l_i}} - \mu_{\Pi_{t_j}}\|_{\mathcal{H}_\kappa}^2\right) \\ &= \exp\left(-\frac{\lambda}{2} (\langle \mu_{\Pi_{l_i}}, \mu_{\Pi_{l_i}} \rangle_{\mathcal{H}_\kappa} - 2\langle \mu_{\Pi_{l_i}}, \mu_{\Pi_{t_j}} \rangle_{\mathcal{H}_\kappa} \right. \\ &\quad \left. + \langle \mu_{\Pi_{t_j}}, \mu_{\Pi_{t_j}} \rangle_{\mathcal{H}_\kappa})\right) \\ &= \exp\left(-\frac{\lambda}{2} \left(\frac{1}{m_l^2} \sum_{i=1}^{m_l} \sum_{i'=1}^{m_l} k(\mathbf{z}_{l_i}, \mathbf{z}_{l_{i'}}) \right. \right. \\ &\quad \left. \left. - \frac{2}{m_l m_t} \sum_{i=1}^{m_l} \sum_{j=1}^{m_t} k(\mathbf{z}_{l_i}, \mathbf{z}_{t_j}) + \frac{1}{m_t^2} \sum_{j=1}^{m_t} \sum_{j'=1}^{m_t} k(\mathbf{z}_{t_j}, \mathbf{z}_{t_{j'}}) \right)\right), \end{aligned} \quad (5)$$

where m_l and m_t are determined by sampling times T . The kernel mean embedding using the level-1 kernel k creates *distributions* $\mu(\mathbb{P}_1), \dots, \mu(\mathbb{P}_N)$ represented by the samples

$\{\mu_{\Pi_{l_1}}, \dots, \mu_{\Pi_{l_n}}\}$ for $l = 1, \dots, N$ respectively in the RKHS \mathcal{H}_k . The underlying strategy of P-MMD is to apply the classic MMD to these distributions (with respect to the kernel κ). To access the effect that the minimization of P-MMD has on the original latent probability distributions across different domains, we recall the following:

Theorem 1 Muandet et al. (2012) *Let $\mathbb{P}_1, \dots, \mathbb{P}_N$ be probability distributions and $\hat{\mathbb{P}} := \frac{1}{N} \sum_{i=1}^N \mathbb{P}_i$. Then the distributional variance given by $\frac{1}{N} \sum \|\mu_{\mathbb{P}_i} - \mu_{\hat{\mathbb{P}}}\|^2$ is 0 iff $\mathbb{P}_1 = \mathbb{P}_2 = \dots = \mathbb{P}_N$.*

Corollary 3.1 Li et al. (2018) *The upper bound of the distributional variance can be written as*

$$\frac{1}{K^2} \sum_{1 \leq i, j \leq K} \text{MMD}(\mathbb{P}_i, \mathbb{P}_j)^2.$$

In our setting Theorem 1 and Corollary 3.1 along with the fact that k is a characteristic kernel imply the following:

Corollary 3.2 *iff all moments of latent distributions Π_l associated to points of domain D_l for $l = 1, \dots, N$ are distributed identically across domains, $\frac{1}{K^2} \sum_{1 \leq i, j \leq K} \text{P-MMD}(\mathbb{P}_i, \mathbb{P}_j)^2 = 0$ holds.*

Following Corollary 3.2 we define the following loss function:

$$\mathcal{L}_{global} = \frac{1}{K^2} \sum_{1 \leq i, j \leq K} \text{P-MMD}(\mathbb{P}_i, \mathbb{P}_j)^2. \quad (6)$$

Corollary 3.2 implies that as Eq. 6 tends to 0 so does the distance between the distributions of means, variances and higher moments of the distributions Π_l associated to points of different domains.

Remark 1 In Sect. 4.5, we compare the P-MMD approach to simply taking the mean (i.e., first moment) of latent probabilistic embeddings Π_l i.e. taking $\Pi_l \rightarrow \mathbf{m}_{\Pi_l} = \mathbb{E}_{\mathbf{x} \sim \Pi_l[\mathbf{x}]}$, and then minimizing the associated “vanilla” MMD. Although this scheme is more efficient computationally over our proposed method, it discards most information about high-level statistics as discussed by Muandet et al. (2017). We empirically verify that our approach has better performance across domains. The visualized computation of P-MMD is shown in Fig. 2.

Although we focus on the scenario of insufficient samples, the computational consumption from Eqs. (4) and (5) may be still prohibitive as the calculation of MMD distance between distributions can scale at least quadratically with the increasing of sample size (especially for image segmentation task), i.e., $O(n^2)$ in a domain. Here, by following the *linear statistic theory* of MMD, the unbiased estimate can

be derived by drawing pairs from two domains with replacement, i.e., $P\text{-MMD}(\mathbb{P}_l, \mathbb{P}_r)^2 \approx \frac{2}{n_l} \sum_{i=1}^{\frac{n_l}{2}} [K(\Pi_{l_{2i}}, \Pi_{l'_{2i+1}}) + K(\Pi_{l_{2i}}, \Pi_{l'_{2i+1}}) - K(\Pi_{l_{2i}}, \Pi_{l_{2i+1}}) - K(\Pi_{l_{2i+1}}, \Pi_{l_{2i}})]$, where assuming $n_l = n_r$ for simplicity. Borgwardt et al. (2006) gives proof about the unbiased property of the *linear statistic* of MMD and shows that statistic power does not be sacrificed too much.

3.2 Probabilistic Contrastive Semantic Alignment

To learn domain-invariant representation from a local perspective, a popular idea is to encourage positive pairs with same label closer together, while pulling other negative ones with different labels further apart (Motiian et al., 2017; Dou et al., 2019). These methods usually measure the Euclidean distance between samples in the embedding space. However, this scheme may not satisfy our probabilistic framework due to its probabilistic embeddings.

To this end, we propose a probabilistic contrastive semantic alignment (P-CSA) loss that can utilize the empirical MMD to measure the discrepancy between probabilistic embeddings. The proposed P-CSA loss \mathcal{L}_{local} consists of two components, including the positive probabilistic contrastive loss and negative probabilistic contrastive loss. The former aims to minimize the distance between the intra-class distributions from different domains, i.e.,

$$\mathcal{L}_{local}^{pos} = \frac{1}{2} \left\| \frac{1}{T} \sum_{i=1}^T \phi(M_{\Theta}(\mathbf{z}_{n_i})) - \frac{1}{T} \sum_{j=1}^T \phi(M_{\Theta}(\mathbf{z}_{q_j})) \right\|_{\mathcal{H}}^2, \tag{7}$$

where $M_{\Theta}(\cdot)$ denotes the embedding network of metric learning, which will contribute to learn the distance between features better (Dou et al., 2019). Note that $\mathbf{y}_n = \mathbf{y}_q$ needs to be satisfied. Then, the negative probabilistic contrastive loss is denoted by

$$\begin{aligned} \mathcal{L}_{local}^{neg} &= \frac{1}{2} \max[0, \xi - \text{MMD}(\Pi_n, \Pi_q)^2] = \frac{1}{2} \max[0, \xi \\ &\quad - \left\| \frac{1}{T} \sum_{i=1}^T \phi(M_{\Theta}(\mathbf{z}_{n_i})) - \frac{1}{T} \sum_{j=1}^T \phi(M_{\Theta}(\mathbf{z}_{q_j})) \right\|_{\mathcal{H}}^2], \end{aligned} \tag{8}$$

where ξ is a distance margin that can guarantee an appropriate repulsion range. Note that $\mathbf{y}_n \neq \mathbf{y}_q$ needs to be satisfied.

Model Training Our proposed framework consists of three modules, a BNN-based probabilistic extractor Q_{ϕ} , a BNN-based classifier C_{ω} , and a metric network $M_{\Theta}(\cdot)$. For the Q_{ϕ} , we only add a Bayesian layer with ReLU layer on the bottom of a pretrained deterministic model (e.g., ResNet18 by removing fully-connected layers) by following Xiao et al.

(2021). For the C_{ω} , a Bayesian layer is also introduced to adapt the classification on insufficient sample better. More implement details of BNN can be found in ‘‘Appendix A.1’’. The structure of M_{Θ} is the same as Dou et al. (2019). The images $\mathcal{X} = \{\mathbf{x}_i\}$ conduct T stochastic forward passes on the Q_{ϕ} and C_{ω} by MC sampling to obtain probabilistic predicts $\{\hat{y}_i^j\}_{j=1}^T$, where the outputs (i.e., probabilistic embeddings) of Q_{ϕ} serve as the inputs for the calculations of \mathcal{L}_{global} and \mathcal{L}_{local} . The final predicts $\{\hat{y}_i^j\}$ are the expectation of $\{\hat{y}_i^j\}_{j=1}^T$. The total objectives can be summarized as follows,

$$\begin{aligned} \mathcal{L}_{total} &= \sum_{l,i} \mathcal{L}_c(\hat{y}_i, y_i) + \text{KL}[q_{\theta}(Q_{\phi}) \| p(Q_{\phi})] \\ &\quad + \text{KL}[q_{\theta}(C_{\omega}) \| p(C_{\omega})] + \beta_1 \mathcal{L}_{local} + \beta_2 \mathcal{L}_{global}. \end{aligned} \tag{9}$$

Discussion The rationale that our proposed method can benefit DG performance on small-data scenario can come from two aspects. First, *BNN can be adaptive to insufficient data well compared with deterministic models* (Graves, 2011; Mallick et al., 2021). For our proposed method, BNN is introduced to the DG problem in the context of insufficient data, where BNN-based feature extractor and classification layers can take both consistent improvements (see 2nd and 3rd columns in Table 4). More importantly, *domain-invariant representation learning under this probabilistic framework from global and local perspectives* contributes to more robust cross-domain performance (see 4th and 5th columns in Table 4; Fig. 3 in Sect. 4.5 for the effectiveness of P-MMD).

4 Experiments

We evaluate our proposed method on three medical imaging tasks: (1) epithelium stroma classification, (2) skin lesion classification, (3) spinal cord gray matter segmentation. The used datasets in these tasks are collected from different healthcare institutes and suffer from the domain shift problem in the context of insufficient samples, *i.e. insufficient sample scenarios exist either in all or some source domains.*

4.1 Epithelium Stroma Classification

Epithelium stroma classification is a fundamental step for the prognostic analysis of the tumor. The public histopathological image datasets for binary classification (epithelium or stroma) are collected from three healthcare centers with different staining types and tissue densities¹: IHC, NKI, and VGH. After the patching operation, IHC, NKI and VGH

¹ <http://fimm.webmicroscope.net/supplements/epistroma>.

Table 1 Experiment results of epithelium stroma classification of histopathological images

Method	IHC	NKI	VGH	Average (%)
DeepAll	73.29 ± 0.13	70.60 ± 0.15	79.56 ± 0.11	74.48
MASF (Dou et al., 2019)	80.45 ± 0.10	76.10 ± 0.11	84.44 ± 0.12	80.33
LDDG (Li et al., 2020)	81.19 ± 0.23	73.27 ± 0.25	82.58 ± 0.23	79.01
KDDG (Wang et al., 2021)	83.65 ± 0.19	74.04 ± 0.15	83.13 ± 0.20	80.27
SWAD (Cha et al., 2021)	79.74 ± 0.15	74.84 ± 0.13	84.29 ± 0.12	79.62
BDIL (Xiao et al., 2021)	85.56 ± 0.12	71.89 ± 0.14	85.90 ± 0.18	81.05
DNA (Chu et al., 2022)	83.93 ± 0.18	73.94 ± 0.15	85.57 ± 0.17	81.14
DSU (Li et al., 2022)	81.56 ± 0.14	72.47 ± 0.12	83.94 ± 0.16	79.32
MIRO (Cha et al., 2022)	82.69 ± 0.11	74.93 ± 0.13	84.63 ± 0.11	80.80
Ours (in this paper)	88.82 ± 0.09	76.71 ± 0.10	86.92 ± 0.14	84.06

Each column denotes a cross-domain task. For example, in the second column, we use IHC dataset as the target domain and the remaining datasets as the source domains. Note that all baseline methods adopt the SWAD method (Cha et al., 2021) for weight averaging. The baseline in the sixth row, namely SWAD, denotes the ERM training strategy with the SWAD method

datasets respectively have 1342, 1230, and 1376 patches, which means that the insufficient sample problem exist in *all source domains* compared with large-scale natural images. We randomly split the data of each source domain into a training set (80%) and a test set (20%) and adopt the leave-one-domain-out strategy for evaluation. The pretrained ResNet18 is introduced as the backbone. The structure of Bayesian layer in Q_ϕ is a fully-connected-based BNN with 512×512 . The structure of Bayesian layer in C_ω is also a fully-connected-based BNN with 512×2 . We utilize Adam optimizer with learning rate as 5×10^{-5} for training. The batch size is 32 for each source domain with 4000 iterations. The hyperparameters are selected in a wide range on the validation set, where the β_1 and β_2 are 0.1 and 0.7 for the \mathcal{L}_{local} and the \mathcal{L}_{global} , respectively. For the P-MMD, *level-1* and *level-2* kernels are the Gaussian RBF kernels (the kernel bandwidth is empirically set to 1 for all kernels) by following Muandet et al. (2012). For the P-CSA loss, the distance margin ξ is set to 1. By balancing the performance and computational efficiency, the number of MC sampling in each Bayesian layer (*a.k.a.*, T), is set to 10. We also discuss the influence of using different T in the Sect. 4.5. We report the results based on average value and standard deviation in each target domain by running the experiment for five different times.

Results Table 1 shows the epithelium stroma classification results on different target domains. We compare with five different approaches. All methods are constructed via a SWAD-based (Cha et al., 2021) framework (where a pretrained ResNet18 as the backbone) with the same training schedule. DeepAll is a deterministic version of our proposed method that is trained on all source domains without any DG strategy in the sequel. Some observations can be summarized as following. First, both our proposed method and BDIL (Xiao et al., 2021) achieve promising results on both

IHC and VGH tasks, which may benefit from the positive impact of probabilistic framework on insufficient samples. However, BDIL has an obvious performance drop on the more challenging NKI domain (which has a lower average accuracy compared with other target domains) compared with our proposed method (*i.e.*, 71.89% *v.s.* **76.71%**). This may be due to the introduction of global alignment (*i.e.*, P-MMD, as used by ours), which is more powerful for learning domain-invariant representations than only using local negative pairs (as used by BDIL). Second, compared with contrastive semantic alignment-based method (*i.e.*, MSFA (Dou et al., 2019)), our proposed method achieves a significantly better performance (80.45% *v.s.* **88.82%** on IHC domain), due to a more reliable distribution-based contrastive learning manner on insufficient samples from all source domains. Finally, our proposed method achieves the best average performance with a clear margin compared with other approaches (*i.e.*, LDDG (Li et al., 2020), KDDG (Wang et al., 2021), DNA (Chu et al., 2022), MIRO (Cha et al., 2022), and DSU (Chu et al., 2022)).

4.2 Skin Lesion Classification

Seven public skin lesion datasets² for seven classes of lesions are collected from various institutes using different dermatoscope types: HAM10000 with 10015 images, UDA with 601 images, SON with 9251 images, DMF with 1212 images, MSK with 3551 images, D7P with 1926 images, and PH2 with 200 images. We can observe that the insufficient sample problem exists in *some source domains*, especially in PH2 and UDA domains. Following previous work (Li et al., 2020), each domain is randomly split into a 50% training set, 30% test set, and 20% validate set, respectively. As adopted

² <https://challenge.isic-archive.com/landing/2018/47/>.

Table 2 Domain generalization results on skin lesion classification

Method	DMF	D7P	MSK	PH2	SON	UDA	Average
DeepAll	0.2492 ± 0.0127	0.5680 ± 0.0181	0.6674 ± 0.0083	0.8000 ± 0.0167	0.8613 ± 0.0296	0.6264 ± 0.0312	0.6287
MASF (Dou et al., 2019)	0.2692 ± 0.0146	0.5678 ± 0.0361	0.6815 ± 0.0122	0.7833 ± 0.0101	0.9204 ± 0.0227	0.6538 ± 0.0196	0.6460
LDDG (Li et al., 2020)	0.2793 ± 0.0244	0.6007 ± 0.0187	0.6967 ± 0.0211	0.8167 ± 0.0209	0.9272 ± 0.0117	0.6978 ± 0.0182	0.6697
KDDG (Wang et al., 2021)	0.3189 ± 0.0256	0.5829 ± 0.0212	0.7014 ± 0.0178	0.9021 ± 0.0314	0.9398 ± 0.0213	0.6882 ± 0.0139	0.6889
SWAD (Cha et al., 2021)	0.3582 ± 0.0234	0.5491 ± 0.0231	0.6842 ± 0.0156	0.9167 ± 0.0121	0.9824 ± 0.0012	0.7240 ± 0.0251	0.7024
BDIL (Xiao et al., 2021)	0.2985 ± 0.0452	0.6204 ± 0.0212	0.7059 ± 0.0145	0.8967 ± 0.0096	0.9860 ± 0.0198	0.7219 ± 0.0284	0.7049
DNA (Chu et al., 2022)	0.3532 ± 0.0133	0.5581 ± 0.0178	0.7120 ± 0.0194	0.9333 ± 0.0045	0.9851 ± 0.0032	0.7314 ± 0.0141	0.7122
DSU (Li et al., 2022)	0.3830 ± 0.0267	0.5739 ± 0.0147	0.6935 ± 0.0165	0.8833 ± 0.0231	0.9841 ± 0.0098	0.7201 ± 0.0121	0.7063
MIRO (Cha et al., 2022)	0.3432 ± 0.0092	0.5863 ± 0.0113	0.6919 ± 0.0101	0.9300 ± 0.0021	0.9659 ± 0.0292	0.7328 ± 0.0233	0.7084
Ours (in this paper)	0.3781 ± 0.0136	0.6120 ± 0.0115	0.7276 ± 0.0201	0.9416 ± 0.0103	0.9889 ± 0.0041	0.7486 ± 0.0123	0.7328

Each column denotes a cross-domain task. For example, in the second column, we use DMF dataset as the target domain and the remaining datasets as the source domains. The best and second-best performance on each target domain are bolded and underlined, respectively. Note that all baseline methods adopt the SWAD method (Cha et al., 2021) for weight averaging. The baseline in the sixth row, namely SWAD, denotes the ERM training strategy with the SWAD method

Table 3 Domain generalization results on gray matter segmentation task

(a) MASF						
Source	Target	DSC	CC	JI	TPR	ASD
2,3,4	1	0.8502	64.22	0.7415	0.8903	0.2274
1,3,4	2	0.8115	53.04	0.6844	0.8161	0.0826
1,2,4	3	0.5285	−99.3	0.3665	0.5155	1.8554
1,2,3	4	0.8938	76.14	0.8083	<u>0.8991</u>	0.0366
Average		0.7710	23.52	0.6502	0.7803	0.5505
(b) KDDG						
Source	Target	DSC	CC	JI	TPR	ASD
2,3,4	1	<u>0.8745</u>	<u>70.75</u>	<u>0.7795</u>	0.8949	0.0539
1,3,4	2	0.8229	56.71	0.6997	0.8226	0.0490
1,2,4	3	0.5676	−63.1	0.3866	0.5904	<u>1.2805</u>
1,2,3	4	0.8894	75.06	0.8011	0.9222	0.0377
Average		0.7886	34.86	0.6667	0.8075	0.3553
(c) LDDG						
Source	Target	DSC	CC	JI	TPR	ASD
2,3,4	1	0.8708	69.29	0.7753	0.8978	0.0411
1,3,4	2	0.8364	60.58	0.7199	<u>0.8485</u>	0.0416
1,2,4	3	0.5543	−71.6	<u>0.3889</u>	<u>0.5923</u>	1.5187
1,2,3	4	0.8910	75.46	0.8039	0.8844	0.0289
Average		0.7881	33.43	0.6720	0.8058	0.4076
(d) SWAD						
Source	Target	DSC	CC	JI	TPR	ASD
2,3,4	1	0.8726	70.23	0.7702	0.8995	0.0502
1,3,4	2	0.8378	60.71	0.7230	0.8176	0.0424
1,2,4	3	0.5388	−99.0	0.3789	0.5083	1.4789
1,2,3	4	0.8903	<u>75.89</u>	0.8026	0.8859	<u>0.0302</u>
Average		0.7849	26.96	0.6687	0.7778	0.4002
(e) DSU						
Source	Target	DSC	CC	JI	TPR	ASD
2,3,4	1	0.8739	70.32	0.7794	<u>0.9210</u>	0.0793
1,3,4	2	0.8474	63.58	0.7367	0.8502	0.0494
1,2,4	3	0.5574	−70.4	0.3923	0.6097	1.5049
1,2,3	4	0.8897	75.10	0.8018	0.9225	0.0415
Average		0.7921	34.65	0.6775	0.8225	0.4362
(f) Ours						
Source	Target	DSC	CC	JI	TPR	ASD
2,3,4	1	0.8786	71.57	0.7873	0.9293	<u>0.0422</u>
1,3,4	2	0.8485	63.78	0.7389	0.8401	0.0401
1,2,4	3	<u>0.5634</u>	−68.0	0.3992	0.6103	1.2239
1,2,3	4	<u>0.8921</u>	75.69	<u>0.8058</u>	0.9245	0.0362
Average		0.7957	35.76	0.6828	0.8260	0.3356

For the DSC, CC, TPR, and JI, the higher the better. For the ASD, the lower the better. Note that all baseline methods adopt the SWAD method (Cha et al., 2021) for weight averaging. The baseline, namely SWAD, denotes the ERM training strategy with the SWAD method

in Li et al. (2020), one domain from DMF, D7P, MSK, PH2, SON and UDA is as target domain and the remaining domains together with HAM10000 as source domains. The pretrained ResNet18 is introduced as the backbone. The structure of Bayesian layer in Q_ϕ is a fully-connected-based BNN with 512×512 . The structure of Bayesian layer in C_ω is also a fully-connected-based BNN with 512×7 . The Adam optimizer is employed with learning rate of 5×10^{-5} for 2000 iterations. The hyperparameters are selected in a wide range on the validation set, where the β_1 and β_2 are 0.1 and 0.7 for the \mathcal{L}_{local} and the \mathcal{L}_{global} , respectively. The batch size is 32 for each source domain. The remaining settings are the same with epithelium stroma classification. For the results, the average value and standard deviation are reported by running five times.

Results Table 2 shows the skin lesion classification accuracies on different target domains. Six different methods are utilized for comparison. All approaches are implemented using the SWAD-based framework and a pretrained ResNet18 model as the backbone. One has some observations as following. First, compared with contrastive semantic alignment-based method (e.g., MASF), our proposed method provides more reliable distribution-based pairs, leading to a better performance on insufficient samples from *some source domains* (MASF:0.2692 v.s. Ours: **0.3781** on DMF domain, where PH2 and UDA as the parts of source domain). Second, although BDIL slightly outperforms our proposed method on D7P domain (**0.6204** v.s. 0.6120), our proposed method has a significantly better performance on the challenging DMF domain (0.2985 v.s. **0.3781**) and the average results (0.7049 v.s. **0.7328**) and other domains. Third, it seems that other baseline methods (e.g., SWAD, DNS, and DSU) impose respective schemes to relieve the impact of insufficient samples from *some source domains*. For example, DSU achieves the best performance on DMF domain. This may be due to the positive impact of straightforward domain randomization. Yet, it is difficult for DSU to realize consistently better results in multiple domains compared with our proposed method, owing to the lack of explicit domain alignment.

4.3 Spinal Cord Gray Matter Segmentation

The spinal cord gray matter (GM) segmentation (pixel-level classification) is an emergent task for predicting disability via evaluating the atrophy of GM area. The acquired magnetic resonance imaging (MRI) data are collected from four healthcare centers³: “site1” with 30 slices, “site2” with 113 slices, “site3” with 246 slices, and “site4” with 122 slices. One can observe that the insufficient sample problem exists in *some source domains*, especially in “site1” domain. By following previous work (Li et al., 2020), we randomly split

³ <http://niftyweb.cs.ucl.ac.uk/challenge/index.php>.

Table 4 Ablation study on each component of our proposed method for spinal cord gray matter segmentation task (where “site2” is as the target domain)

Backbone (Unet)	Bayesian layers	Local alignment	Global alignment	Bayesian classifier	DSC	CC	JI	TPR	ASD
✓	✗	-	-	✗	0.7223	26.21	0.5789	0.8109	0.0992
✓	✓	-	-	✗	0.7934	47.19	0.6595	0.8133	0.0692
✓	✓	-	-	✓	0.8268	57.52	0.7067	0.8156	0.0501
✓	✓	✓	✗	✓	0.8364	60.72	0.7195	0.8267	0.0486
✓	✓	✗	✓	✓	0.8371	60.57	0.7217	0.8152	0.0510
✓	✓	✓	✓	✓	0.8485	63.78	0.7389	0.8401	0.0401

The model on the first row denotes the basic Unet model

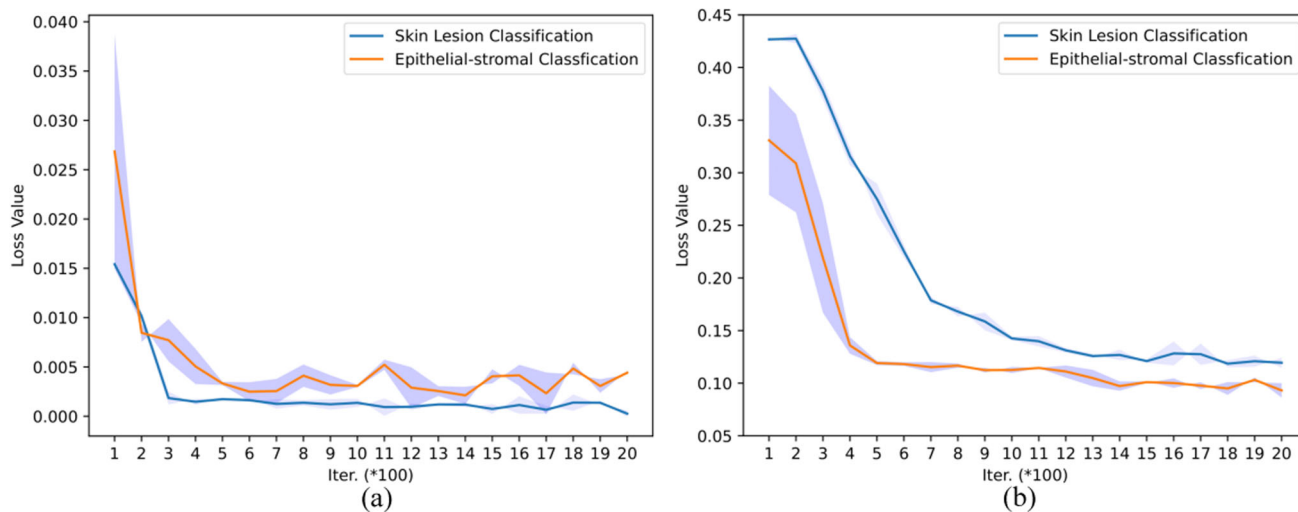


Fig. 3 The loss curve of iteration on skin lesion and epithelial-stromal classification tasks. **a** Global alignment loss. **b** Local alignment loss

the data of each source domain into a training set (80%) and a test set (20%) and adopt the leave-one-domain-out strategy for evaluation. The hyperparameters are selected in a wide range on the validation set, where the β_1 and β_2 are 0.01 and 0.001 for the \mathcal{L}_{local} and the \mathcal{L}_{global} , respectively. The 2D-Unet (Ronneberger et al., 2015) is leveraged as the backbone for all methods. The structures of Q_ϕ and C_ω as well as more experimental details can be found in “Appendix B”. By following Li et al. (2020), the average results in each target domain are reported by running three times.

Results Table 3 shows the spinal cord GM segmentation results on different target domains. Dice Similarity Coefficient (DSC), Jaccard Index (JI), and Conformity Coefficient (CC) are used to measure the accuracy of obtained segmentation results. True Positive Rate (TPR) and Average Surface Distance (ASD) are introduced from statistical and distance-based perspectives. We compare with five different methods (which have effective segmentation performance). Some observations can be found as follows. First, our proposed method outperforms all baseline methods in terms of average results for five quantitative metrics. Second, compared with the contrastive semantic alignment-based method (e.g., MASF), reliable pixel-level pairs constructed by our proposed method also contribute to improving performance in scenarios with insufficient samples from some source domains. Third, as we can see, our proposed method and DSU achieve the best and second-best performance compared with other baselines, which may be reasonable as they can benefit from the modeling of the data uncertainty in the small-data regime using the BNN or the distribution modeling.

4.4 Ablation Analysis

The spinal cord gray matter segmentation task is utilized to explore the effectiveness of each component for our proposed method, *due to its various quantitative metrics*. The results can be shown in Table 4. First, we observe that better performance can be achieved by introducing a probabilistic layer compared with the results that using Unet, which reflects the superiority of probabilistic models. Secondly, we observe that by either introducing local or global alignment for domain-invariant information learning, better performance can be achieved compared with the results of only using the probabilistic layer, which shows the effectiveness of the introduced probabilistic feature regularization term. Last but not least, by imposing domain-invariant learning with both local and global views, the performances are further improved, which justifies the effectiveness of our proposed method by jointly considering local and global alignment.

Effectiveness of domain-invariant loss We are also interested in the impacts of domain-invariant losses on different tasks. Thus, we visualize the learning curves of different task in Fig. 3. As we can observe, for the skin lesion (on DMF) and epithelium-stroma (on IHC) classification tasks, the loss curves with iterations reflect the global discrepancy converges faster than the local discrepancy, while the more challenging cross-domain task converges more slowly on global alignment.

4.5 Further Analyses of Our Proposed Method

Results on Different Fractions of Training Data We are interested in how different fractions of training samples influence

Table 5 Domain generalization results on MSK dataset by randomly picking same proportion of samples from each source domain

Proportion (%)	BDIL	DNA	Ours
100	0.7059± 0.0284	0.7121± 0.0141	0.7276±0.0123
80	0.6625±0.0920	0.6591±0.0022	0.6975±0.0036
60	0.6468±0.0106	0.6149±0.0112	0.6641±0.0114
40	0.6491±0.0171	0.6065±0.0111	0.6579±0.0057
Average (80,60,40) ↑	0.6528	0.6268	0.6732
Average Attenuation Rate ↓	7.67%	11.98%	7.37%

A smaller proportion (< 40%) is unavailable because equal batch sizes cannot be maintained in PH2 dataset

Table 6 Domain generalization results on MSK dataset by randomly picking same number of samples from each class in each domain

Number of sample	BDIL	DNA	Ours
40	0.5897 ± 0.0029	0.5412 ± 0.0143	0.6368 ± 0.0074
30	0.5762 ± 0.0101	0.5132 ± 0.0229	0.6138 ± 0.0291
20	0.5573 ± 0.0011	0.5048 ± 0.0087	0.6037 ± 0.0121
Average (40,30,20) ↑	0.5744	0.5196	0.6183
Average attenuation rate ↓	5.49%	6.72%	5.19%

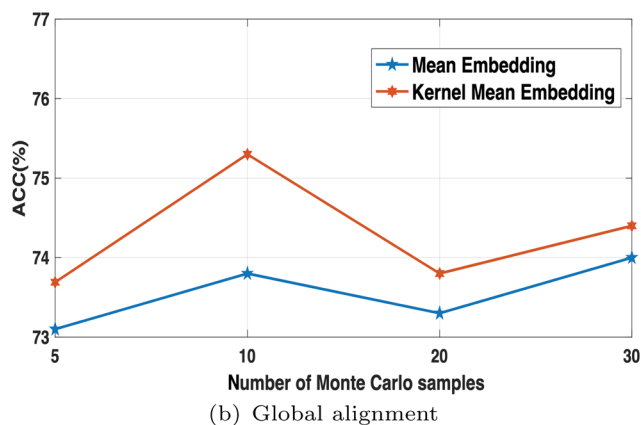
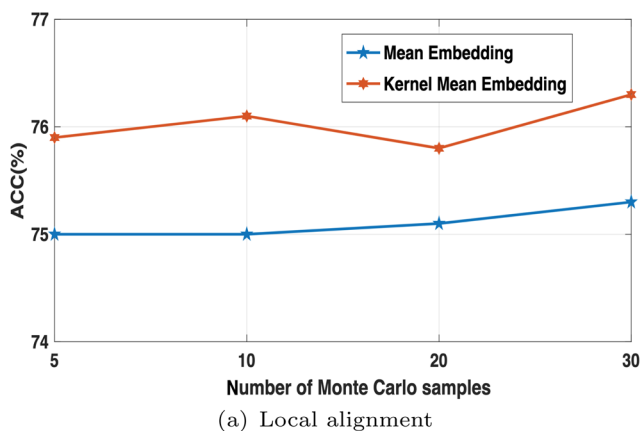


Fig. 4 The performance comparison between mean embedding method and kernel mean embedding method with different Monte Carlo samples T . For each sub-figure, we use only one alignment operation. **a** Local alignment. *Mean Embedding* The mean embedding operation with Euclidean distance is utilized between probabilistic embedding pairs. *Kernel Mean Embedding* The kernel mean embedding with MMD

distance is utilized between probabilistic embedding pairs. **b** Global alignment. *Mean Embedding* The mean embedding operation with MMD distance is utilized between domains (as distributions). *Kernel Mean Embedding* The kernel mean embedding with P-MMD distance is utilized between domains (as distributions over distributions)

the final performance based on small-data scenario. To this end, we adopt skin lesion classification task for evaluation since the number of samples in each domain turns out to be imbalance (some domains such as HAM10000 contain sufficient data while the number of samples in some other domains such as PH2 and UDA is insufficient). As such, we can better simulate the scenario that the issue of insufficient samples either exists in *all source domains* or in *partial source domains*. We choose MSK dataset with the second-highest number of samples as the target domain and the remaining datasets as the source domains (including PH2 and UDA). Table 5 shows the accuracies. As we can see, compared with DNA (with second-best performance on 100% of samples), our proposed method and BDIL have better average results

and lower performance attenuation as the decrease of sample number, due to their probabilistic gain under small data scenarios. Compared with BDIL, our proposed method shows a more robust performance on this challenging small data scenario, due to the integration of local (P-CSA) and global (P-MMD) alignments.

Results on Different Numbers of Samples for Training Data We are also interested in how different numbers of samples per class influence the final performance based on small-data scenarios. To this end, we also adopt skin lesion classification task for evaluation. Specifically, we randomly draw T samples from each class in a source domain to represent this domain for training. Here, we set T to 20,30, and 40, respectively, in different experiments.

The results can be found in Table 6. As we can see, our proposed method achieved the best performance among all settings compared with all baseline methods. Meanwhile, it seems that the Bayesian-based DG approaches (e.g., our proposed method and BIDL) have better performance compared with other methods, which is reasonable as the BNN can be adaptive to the small data scenario well. Especially, our proposed method has around 5% improvements compared with the second-best method when T is set to smaller, i.e., 20.

Kernel Mean Embedding (i.e., P-MMD) v.s. Mean Embedding We explore the effect of different schemes for probabilistic embeddings. A straightforward method is to represent probabilistic embeddings with the expectation, which is called the “Mean Embedding”. Then, a probabilistic embedding can be considered as a latent point and the MMD can be used to measure the discrepancy between distributions consisting of latent points.

For the mean embedding-based \mathcal{L}_{global} , the computational process of this scheme for MMD distance can be formulated as

$$MMD(\mathbb{P}_l, \mathbb{P}_r)^2 = \left\| \frac{1}{n_l} \sum_{i=1}^{n_l} \varphi(\mathbb{E}[\Pi_{l_i}]) - \frac{1}{n_r} \sum_{j=1}^{n_r} \varphi(\mathbb{E}[\Pi_{r_j}]) \right\|_{\mathcal{H}}^2. \quad (10)$$

Equation (10) can be further constructed a global alignment loss \mathcal{L}_{global} . For the local alignment loss \mathcal{L}_{local} , the Euclidean distance can be used to compute the distance between latent points, which is similar to the original CAS loss in Motiian et al. (2017). For the positive pairs with the same label, the mean embedding-based positive contrastive loss can be represented as

$$\mathcal{L}_{local}^{pos} = \frac{1}{2} \left\| \frac{1}{T} \sum_{i=1}^T \mathbb{E}[M_{\Theta}(\mathbf{z}_{n_i})] - \frac{1}{T} \sum_{j=1}^T \mathbb{E}[M_{\Theta}(\mathbf{z}_{q_j})] \right\|_2^2. \quad (11)$$

$M_{\Theta}(\cdot)$ denotes the embedding network of metric learning. For the negative pairs with different labels, the negative contrastive loss $\mathcal{L}_{local}^{neg}$ is denoted by

$$\frac{1}{2} \max \left[0, \xi - \left\| \frac{1}{T} \sum_{i=1}^T \mathbb{E}[M_{\Theta}(\mathbf{z}_{n_i})] - \frac{1}{T} \sum_{j=1}^T \mathbb{E}[M_{\Theta}(\mathbf{z}_{q_j})] \right\|_2 \right]. \quad (12)$$

As a result, a mean embedding-based contrastive loss with the view of local alignment can be calculated as

$$\mathcal{L}_{local} = \mathcal{L}_{local}^{pos} + \mathcal{L}_{local}^{neg}. \quad (13)$$

Instead, we can observe from Fig. 2 that our proposed method induces a level-2 kernel-based MMD with empirical estima-

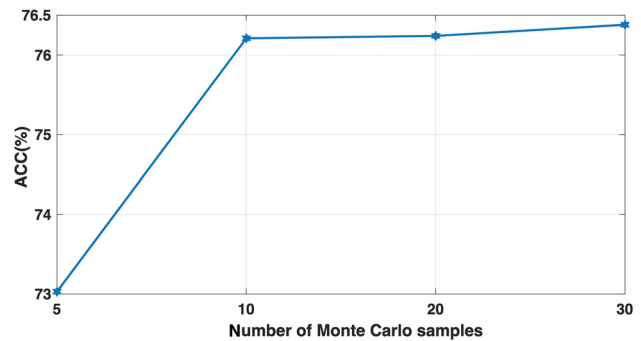


Fig. 5 The performance of our proposed model on the NKI task of Epithelium Stroma classification with different Monte Carlo samples T

tion for probabilistic embeddings. Specifically, our proposed scheme can preserve higher moments of a probabilistic embedding via nonlinear level-1 kernel (see the fourth component in Fig. 2). Moreover, by introducing a level-2 kernel, the similarities between probabilistic embeddings also can be measured based on their own moment information (see the last component in Fig. 2). Benefiting from these virtues, the proposed probabilistic MMD can accurately capture the discrepancy between mixture distributions via an extended empirical MMD fashion.

Here, we validate the effectiveness of different schemes on the NKI task of Epithelium Stroma classification in each aligned view. The experimental settings are similar for different methods. The experimental results can be found in Fig. 4. As we can see, our proposed method achieves consistent improvements in each alignment method with different Monte Carlo samples, which may be reasonable as the kernel mean representation can preserve many statistical components due to the injective property. Second, when the number of MC samples is 10, we can observe an obvious margin in global alignment, which refers to the computation between mixture distributions.

Influence of Different Number of MC Samples It is much important to balance the number of Monte Carlos samples and the computational efficiency. On the one hand, the property of probabilistic embeddings can be affected by the Monte Carlos sampling. On the other hand, too many Monte Carlos samples may suffer from the heavy computational cost. Xiao et al. (2021) suggested that the distributional property and computational cost are both acceptable for the computation of the KL divergence when the number of Monte Carlos samples is chosen appropriately, the practical performance of our proposed method needs to be explored. We conduct the experiments on the NKI task of Epithelium-Stromal classification with different Monte Carlo samples T .

The results are shown in Fig. 5. As we can see, if the number of Monte Carlo samples is too small, it is difficult to capture the property of distribution for probabilistic embeddings. As the increase of T , there is an obvious improvement

Table 7 Out-of-domain accuracies (%) on PACS based on ResNet50

Method	Art	Cartoon	Photo	Sketch	Average (%)
RSC (Huang et al., 2020)	78.9	76.9	94.1	76.8	81.7
L2A-OT (Zhou et al., 2020)	83.3	78.2	96.2	73.6	82.8
MatchDG (Mahajan et al., 2021)	81.2	80.4	96.8	77.2	83.9
pAdaIN (Nuriel et al., 2021)	81.7	76.6	96.3	75.1	82.5
MixStyle (Zhou et al., 2021)	86.8	79.0	96.6	78.5	85.2
SagNet (Nam et al., 2021)	87.4	80.7	97.1	80.0	86.3
ERM (Vapnik, 1999)	84.7	80.8	97.2	79.3	85.5
DNA (Chu et al., 2022)	89.8	83.4	97.7	82.6	88.4
ERM+SWAD (Cha et al., 2021)	89.3	83.4	97.3	82.5	88.1
MIRO+SWAD (Cha et al., 2022)	-	-	-	-	88.4
Ours+SWAD	90.2	85.2	98.7	83.6	89.4

Table 8 Out-of-domain accuracies (%) on OfficeHome based on ResNet50

Algorithm	Art	Clipart	Product	Real	Avg
Mixstyle (Zhou et al., 2021)	51.1	53.2	68.2	69.2	60.4
RSC (Huang et al., 2020)	60.7	51.4	74.8	75.1	65.5
DANN (Ganin et al., 2016)	59.9	53.0	73.6	76.9	65.9
GroupDRO (Sagawa et al., 2019)	60.4	52.7	75.0	76.0	66.0
MTL (Blanchard et al., 2021)	61.5	52.4	74.9	76.8	66.4
VREx (Krueger et al., 2021)	60.7	53.0	75.3	76.6	66.4
MLDG (Balaji et al., 2018)	61.5	53.2	75.0	77.5	66.8
SagNet (Qian et al., 2021)	63.4	54.8	75.8	78.3	68.1
CORAL (Sun & Saenko, 2016)	65.3	54.4	76.5	78.4	68.7
ERM (Vapnik, 1999)	63.1	51.9	77.2	78.1	67.6
DNA (Chu et al., 2022)	67.7	57.7	78.9	80.5	71.2
ERM+SWAD (Cha et al., 2021)	66.1	57.7	78.4	80.2	70.6
MIRO+SWAD (Cha et al., 2022)	-	-	-	-	72.4
Ours+SWAD	68.2	58.9	80.2	80.7	<u>72.0</u>

Table 9 Out-of-domain accuracies (%) on VLCS based on ResNet50

Algorithm	Caltech	LabelMe	SUN	VOC	Avg
Mixstyle (Zhou et al., 2021)	98.3	64.8	72.1	74.3	77.4
RSC (Huang et al., 2020)	97.9	62.5	72.3	75.6	77.1
DANN (Ganin et al., 2016)	99.0	65.1	73.1	77.2	78.6
GroupDRO (Sagawa et al., 2019)	97.3	63.4	69.5	76.7	76.7
MTL (Blanchard et al., 2021)	97.8	64.3	71.5	75.3	77.2
VREx (Krueger et al., 2021)	98.4	64.4	74.1	76.2	78.3
MLDG (Balaji et al., 2018)	97.4	65.2	71.0	75.3	77.2
SagNet (Qian et al., 2021)	97.9	64.5	71.4	77.5	77.8
CORAL (Sun & Saenko, 2016)	98.3	66.1	73.4	77.5	78.8
ERM (Vapnik, 1999)	97.7	64.3	73.4	74.6	77.5
DNA (Chu et al., 2022)	98.8	63.6	74.1	79.5	79.0
ERM+SWAD (Cha et al., 2021)	98.8	63.3	75.3	79.2	79.1
MIRO+SWAD (Cha et al., 2022)	-	-	-	-	79.6
Ours+SWAD	98.9	63.4	75.8	79.8	<u>79.5</u>

for our proposed method. Interestingly, the performance is gradually saturated. As a result, by balancing the number of Monte Carlo samples and the computational efficiency, the number of Monte Carlo samples T in each Bayesian layer is set to 10.

4.6 Scalability to Benchmark Datasets

Here, we introduce three DG benchmarks, namely PACS (Art: 2048 images, Cartoon: 2344 images, Photo: 1670 images, Sketch: 3929 images), OfficeHome (has 15588 samples with 65 classes from four domains) and VLCS (has 10729 samples with 5 classes from four domains), for comparison. Compared with some large-scale benchmarks (e.g., DomainNet and Wilds), these three datasets are more appropriate to explore the effectiveness of different DG models under the scenario of insufficient samples. performance. We adopt pretrained ResNet50 as the backbone for all benchmarks. The structure of the overall framework is similar with the model mentioned in lesion skin classification. Our proposed method as well as baseline methods are all based on DomainBed, where the holdout fraction (the proportion of validation set) rate for DomainBed is set to 0.2 for all methods. A domain is the target domain and the remaining domains are the source domain for training. The testing is on the overall data of a target domain.

Here, our proposed method is optimized by Adam optimizer with learning rate as 5×10^{-5} . The batch size for each source domain is 32. The training steps are set to 20000 for PACS and OfficeHome, and 2000 for VLCS. By following the SWAD framework, the training process will be stopped for our proposed method when the validation loss increases significantly. The hyperparameters are selected in a wide range on the validation set. For the \mathcal{L}_{local} and \mathcal{L}_{global} , the β_1 and the β_2 are set to 0.1 and 1 for all benchmark datasets. Other hyperparameters such as kernel function, kernel bandwidth and distance margin are similar to the settings mentioned before.

The experimental results on PACS, OfficeHome and VLCS can be shown in Tables 7, 8, and 9. As we can see, compared with domain-invariant-based approaches (e.g., DANN), our proposed method has a significant improvement due to the introduction of a probabilistic framework. Our proposed method also outperforms the data augmentation-based approach (e.g., Mixstyle, Manifold Mixup, and CutMix). Although data generation methods (e.g., MixStyle) can effectively tackle the insufficient sample problem via additional generative samples, the lack of effective domain-invariant learning may hamper the improvement of the performance. Our proposed method achieves better performances compared with feature disentanglement-based (e.g., pAdaI). Last but not least, compared with the state-of-the-art domain generalization method (e.g., MIRO), our proposed method

achieves comparable performance on small-scale benchmark datasets, which demonstrates the scalability of our proposed method in a small-data regime.

5 Discussion

Limitation (1) MC sampling Our probabilistic embeddings derive from the distribution over the model weights. Similar to widely-used BNN-based models (Blundell et al., 2015; Mallick et al., 2021; Xiao et al., 2021), the predictive distribution of probabilistic embeddings needs to be approximated by MC sampling. This may result in more computational consumption compared with the original deterministic model-based scheme. Although heavy computational consumption can be alleviated in our settings (i.e., DG problem in the context of small data), the computational cost needs to be reduced more on very large-scale datasets in the future. *(2) Approximate posterior* Although the mean-field variational inference (MFVI) (we used) is effective in rendering an approximate posterior of BNN. The potential amortization gap due to the fully factorized Gaussian assumption of MFVI would limit further performance improvement on very challenging datasets (Cremer et al., 2018). This limitation will be explored by replacing it with more expressive approximate posteriors (such as normalizing flows) in the future.

Tradeoff In this paper, we carefully make a tradeoff between *effectiveness and computational complexity*. Specifically, due to the introduction of probabilistic embeddings for the DG problem in the context of insufficient data, the probabilistic MMD is proposed based on the level-2 kernel for more high-level statistics, which may be more complex than other baseline methods with “vanilla” MMD. However, it shows improved effectiveness in addressing the problem of global semantic alignment between domains, consisting of a series of probabilistic embeddings. Other simpler methods may not achieve the same level of performance using just the first moment (i.e., the mean embedding). Moreover, we adopt some strategies, such as the unbiased estimate of MMD with linear complexity and the number selection of appropriate MC sampling, to offset the extra computational complexity compared with other baseline methods.

Usage of our proposed method Compared with a deterministic framework, where a source domain as a distribution consists of a set of point embeddings, a source domain generated by our probabilistic framework includes a set of distributions, i.e., the so-called distribution over distributions or the probability of probability. The vanilla MMD may not directly cope with this situation. Instead, our proposed P-MMD leverages the level-2 kernel to measure the discrepancy between mixture distributions based on the empirical MMD framework and preserve most information about high-level statistics.

Table 10 Ablation study on epithelium stroma classification in histopathological images

Method	IHC	NKI	VGH	Average (%)
Baseline	79.74 ± 0.15	74.84 ± 0.13	84.29 ± 0.12	79.62
Baseline + PE	81.99 ± 0.12	74.90 ± 0.14	85.01 ± 0.18	80.63

PE denotes the introduction of probabilistic embeddings

In this paper, we focus on the DG problem in the context of insufficient data. Especially, for the medical imaging data, insufficient sample scenarios either exist in all source domains or in some source domains. If the number of training samples from all source domains is sufficient, it is reasonable to choose another DG method. For instance, the Domain-Net datasets for natural images have six source domains, whereas the source domain (“clipart”) with the smallest number of training samples still has 50,000 training samples. Moreover, our proposed method shows good scalability on benchmark datasets with a larger number of samples, although it is designed based on the DG scenario of insufficient data.

Gains of Probabilistic Embeddings To further show the gain of introducing probabilistic embeddings for insufficient data, we conducted an ablation study on the Epithelium Stroma Classification and Skin Lesion Classification. We only add a Bayesian layer (for probabilistic embeddings) with ReLU layer on the bottom of a deterministic baseline model (pre-trained ResNet-18).

The results can be found in Table 10. As we can see, consistent improvements can be observed in these small-data tasks due to the introduction of probabilistic embeddings.

6 Conclusion

In this work, we address the DG problem in the context of insufficient data, which can occur in *all or some source domains*. To this end, we introduce a probabilistic framework into the DG problem to derive probabilistic embeddings (which can be adaptive to insufficient samples better compared with deterministic models) for domain-invariant learning. Under this probabilistic framework, an extension of MMD called P-MMD is proposed for measuring the *distribution over distributions*. Moreover, a probabilistic CSA loss is proposed for local alignment. Extensive experiments on insufficient cross-domain medical imaging data show the effectiveness of this method.

Acknowledgements This work is supported in part by Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA), the Research Grant Council (RGC) of Hong Kong through Early Career Scheme (ECS) under the Grant 21200522, CityU Applied Research Grant (ARG) 9667244, and Sichuan Science and Technology Program 2022NSFSC0551.

Funding Open access publishing enabled by City University of Hong Kong Library’s agreement with Springer Nature.

Data availability For this paper only publicly available datasets were used. The links of Epithelium Stroma Classification, Skin Lesion Classification, and Spinal Cord Gray Matter Segmentation can be found in <http://fimm.webmicroscope.net/supplements/epistroma>, <https://challenge.isic-archive.com/landing/2018/47/>, <http://niftyweb.cs.ucl.ac.uk/challenge/index.php>.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix A: Details of Bayesian Neural Network

For our proposed method, the Bayesian layer refers to the probabilistic extractor Q_ϕ and the probabilistic classifier C_ω . Here, a simple and convenient PyTorch library, namely BayesianTorch (Krishnan et al., 2022), is utilized to construct the Bayesian neural network. The log evidence lower bound (ELBO) cost function, i.e.,

$$\mathcal{L} := \int q_\theta \log(y|x, w) dw - \text{KL}[q_\theta(w)|p(w)], \quad (\text{A1})$$

can be calculated automatically. By using BayesianTorch, arbitrary deterministic models can be converted into the Bayesian layers easily. In this paper, mean-field variational inference (MFVI) (Graves, 2011) is adopted, where the parameters of the model are characterized by fully factorized Gaussian distribution endowed by variational parameters μ and σ , i.e.,

$$q_\theta(w) := \mathcal{N}(w|\mu, \sigma). \quad (\text{A2})$$

By using stochastic gradient descent method with ELBO cost, the variational distribution $q_\theta(w)$ as the approximation

of the posterior distribution, and corresponding parameters (μ and σ) and can be learned conveniently.

For the settings of Bayesian layer, we follow the model priors with empirical Bayes using DNN (MOPED) method for the parameter settings of weights prior, each weight is sampled from the Gaussian distribution independently (Krishnan et al., 2020),

$$w \sim \mathcal{N}(w_{\text{DNN}}, \delta | w_{\text{DNN}}|), \quad (\text{A3})$$

where w_{DNN} denotes the mean of prior distribution from the maximum likelihood estimates of weights from deterministic deep neural network. δ , a hyperparameter, is set to the initial perturbation factor for the percentage of the pretrained deterministic weight values. The variational layer is modeled using reparameterization trick. The MOPED can realize better training convergence for complex models (Krishnan et al., 2020), which is beneficial to our proposed method. In this paper, we follow the setting in (Krishnan et al., 2020) to set the initial perturbation factor δ for the weight to 0.1.

Appendix B: Implementation Details of Experiments

B.1 Epithelium Stroma Classification

Implement Details. There are two types of basic tissues, i.e., the epithelium and the stroma. Due to the differences of the scanner, the staining type, and the population, the color of the background and the morphological structure among different histopathological image datasets are diverse. The extract epithelial or stromal patches are resized into 224×224 . The classification objective is the Cross-entropy loss with Softmax function. All baseline methods are trained with the same training scheme. We tune their hyperparameters in a wide range on the validation set. The testing results are reported using the best model on validation set.

B.2 Skin Lesion Classification

Implement details There are seven classes of skin lesions, including melanoma (*mel*), melanocytic nevus (*nv*), dermatofibroma (*df*), basal cell carcinoma (*bcc*) benign keratosis (*bkl*), vascular lesion (*vasc*), and actinic keratosis (*akiec*). For inputs, all images are resized into 224×224 for all methods. Due to the class imbalance problem, the focal loss (Lin et al., 2017) as the classification objective is introduced for all methods.

B.3 Spinal Cord Gray Matter Segmentation

Implementation details By following (Li et al., 2018), the 3D MRI data are split into 2D slices in axial view. Then, these obtained 2D slices are centered cropped to 160×160 and randomly cropped to 144×144 for training. The 2D-Unet (Ronneberger et al., 2015) is leveraged as the backbone for all methods. For our proposed method, probabilistic extractor Q_ϕ is constructed by two Bayesian-based 1×1 convolutional layers. The input and output channels in the first convolutional layer are both 64. After a ReLU layer, the input and output channels in the second convolutional layer are 64 and 1, respectively. The BayesianTorch can enable to convert ordinary convolutional layer into Bayesian convolutional neural network easily. The Bayesian neural network adopts MFVI to approximate the posterior distribution of weights. The parameters of Bayesian layer are the same as aforementioned settings. The structure of the Bayesian layer in the probabilistic classifier C_ω is a Bayesian-based 1×1 convolutional layers. The input and output channels are 64 and 1, respectively. The construction of C_ω is the same as that of Q_ϕ . Here, all methods adopt a two-stage scheme for coarse-to-fine segmentation, as used in (Li et al., 2020). Specifically, we first conduct preliminary segmentation to obtain the spinal cord area from the original 2D slice. Then, we perform elaborative segmentation on obtained spinal cord results to derive gray matter results.

The Adam optimizer is utilized with learning rate as 1×10^{-4} , weight decay as 1×10^{-8} . The batch size is 8 for each source domain. The total epochs are 200, where the learning rate will be decreased every 80 epochs with a factor of 10. Other hyperparameters such as kernel function, kernel bandwidth and distance margin are similar to the settings in skin lesion classification and epithelium-stroma classification. The segmentation can be regarded as the pixel-level classification. For the $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{global}}$, we follow (Motiian et al., 2017) to randomly sample some positive and negative pairs from two domains such that the computational efficiency can be improved significantly. Here, we randomly sample 400 positive and negative pixel pairs from two domains in a mini-batch for the computation of $\mathcal{L}_{\text{local}}$, respectively. By leveraging selected pixels of a domain in $\mathcal{L}_{\text{local}}$, we further utilize these pixels to calculate the $\mathcal{L}_{\text{global}}$, which may induce a more accurate measurement owing to the balanced class distribution, as well as reducing the computational cost.

References

- Balaji, Y., Chellappa, R., & Feizi, S. (2019). Normalized Wasserstein for mixture distributions with applications in adversarial learning and

- domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6500–6508).
- Balaji, Y., Sankaranarayanan, S., & Chellappa, R. (2018). Metareg: Towards domain generalization using meta-regularization. *Advances in Neural Information Processing Systems*, *31*, 1–11.
- Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems*, *19*, 151–175.
- Berlinet, A., & Thomas-Agnan, C. (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer.
- Blanchard, G., Deshmukh, A. A., Dogan, Ü., Lee, G., & Scott, C. (2021). Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, *22*(1), 46–100.
- Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, *112*(518), 859–877.
- Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural network. In *International conference on machine learning* (pp. 1613–1622). PMLR.
- Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., & Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, *22*(14), 49–57.
- Bu, Y., Zou, S., Liang, Y., & Veeravalli, V. V. (2018). Estimation of KL divergence: Optimal minimax rate. *IEEE Transactions on Information Theory*, *64*(4), 2648–2674.
- Can, Y. S., & Ersoy, C. (2021). Privacy-preserving federated deep learning for wearable IoT-based biomedical monitoring. *ACM Transactions on Internet Technology (TOIT)*, *21*(1), 1–17.
- Cha, J., Lee, K., Park, S., & Chun, S. (2022). Domain generalization by mutual-information regularization with pre-trained models. In *Computer vision—ECCV 2022: 17th European conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII* (pp. 440–457). Springer.
- Cha, J., Chun, S., Lee, K., Cho, H.-C., Park, S., Lee, Y., & Park, S. (2021). SWAD: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, *34*, 22405–22418.
- Chang, J., Lan, Z., Cheng, C., & Wei, Y. (2020). Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5710–5719).
- Chu, X., Jin, Y., Zhu, W., Wang, Y., Wang, X., Zhang, S., & Mei, H. (2022). DNA: Domain generalization with diversified neural averaging. In *International conference on machine learning* (pp. 4010–4034). PMLR.
- Chun, S. (2023). Improved probabilistic image-text representations. arXiv preprint [arXiv:2305.18171](https://arxiv.org/abs/2305.18171)
- Chun, S., Oh, S. J., De Rezende, R. S., Kalantidis, Y., & Larlus, D. (2021). Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8415–8424).
- Cremer, C., Li, X., & Duvenaud, D. (2018). Inference suboptimality in variational autoencoders. In *International conference on machine learning* (pp. 1078–1086). PMLR.
- Dou, Q., Castro, D., Kamnitsas, K., & Glocker, B. (2019). Domain generalization via model-agnostic learning of semantic features. *Advances in neural information processing systems* (Vol. 32, pp. 1–12).
- Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning* (pp. 1050–1059). PMLR.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, *17*(1), 2096–3030.
- Gong, R., Li, W., Chen, Y., & Gool, L. V. (2019). Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2477–2486).
- Graves, A. (2011). Practical variational inference for neural networks. *Advances in Neural Information Processing Systems*, *24*, 1–9.
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, *13*(1), 723–773.
- Gurdasani, D., Barroso, I., Zeggini, E., & Sandhu, M. S. (2019). Genomics of disease risk in globally diverse populations. *Nature Reviews Genetics*, *20*(9), 520–535.
- Hu, S., Zhang, K., Chen, Z., & Chan, L. (2020). Domain generalization via multidomain discriminant analysis. In *Uncertainty in artificial intelligence* (pp. 292–302). PMLR.
- Huang, Z., Wang, H., Xing, E. P., & Huang, D. (2020). Self-challenging improves cross-domain generalization. In *European conference on computer vision* (pp. 124–140). Springer.
- Johnson, J. D., & Louis, J. M. (2022). Does race or ethnicity play a role in the origin, pathophysiology, and outcomes of preeclampsia? An expert review of the literature. *American Journal of Obstetrics and Gynecology*, *226*(2), 876–885.
- Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. *Advances in Neural Information Processing Systems*, *33*, 18661–18673.
- Kim, J., Lee, J., Park, J., Min, D., & Sohn, K. (2021). Self-balanced learning for domain generalization. In *2021 IEEE international conference on image processing (ICIP)* (pp. 779–783). IEEE.
- Kingma, D. P., Salimans, T., & Welling, M. (2015). Variational dropout and the local reparameterization trick. *Advances in Neural Information Processing Systems*, *28*, 1–9.
- Krishnan, R., Esposito, P., & Subedar, M. (2022). *Bayesian-Torch: Bayesian neural network layers for uncertainty estimation*. <https://doi.org/10.5281/zenodo.5908307>
- Krishnan, R., Subedar, M., & Tickoo, O. (2020). Specifying weight priors in Bayesian deep neural networks with empirical bayes. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, pp. 4477–4484).
- Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., & Courville, A. (2021). Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning* (pp. 5815–5826). PMLR.
- Lee, J., Liu, C., Kim, J., Chen, Z., Sun, Y., Rogers, J. R., Chung, W. K., & Weng, C. (2022). Deep learning for rare disease: A scoping review. medRxiv.
- Li, X., Dai, Y., Ge, Y., Liu, J., Shan, Y., & Duan, L.-Y. (2022). Uncertainty modeling for out-of-distribution generalization. arXiv preprint [arXiv:2202.03958](https://arxiv.org/abs/2202.03958)
- Li, H., Pan, S. J., Wang, S., & Kot, A. C. (2018). Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5400–5409).
- Li, D., Yang, Y., Song, Y.-Z., & Hospedales, T. (2018). Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32).
- Li, M., Huang, B., & Tian, G. (2022). A comprehensive survey on 3D face recognition methods. *Engineering Applications of Artificial Intelligence*, *110*, 104669.
- Li, C., Lin, X., Mao, Y., Lin, W., Qi, Q., Ding, X., Huang, Y., Liang, D., & Yu, Y. (2022). Domain generalization on medical imaging classification using episodic training with task augmentation. *Computers in Biology and Medicine*, *141*, 105144.

- Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).
- Liu, Q., Chen, C., Qin, J., Dou, Q., & Heng, P.-A. (2021). FeddG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1013–1023).
- Liu, X., Yoo, C., Xing, F., Oh, H., El Fakhri, G., Kang, J.-W., & Woo, J. (2022). Deep unsupervised domain adaptation: A review of recent advances and perspectives. *APSIPA Transactions on Signal and Information Processing*, 11(1), 1–51.
- Li, H., Wang, Y., Wan, R., Wang, S., Li, T.-Q., & Kot, A. (2020). Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in Neural Information Processing Systems*, 33, 3118–3129.
- Long, M., Zhu, H., Wang, J., & Jordan, M. I. (2017). Deep transfer learning with joint adaptation networks. In *International conference on machine learning* (pp. 2208–2217). PMLR.
- Mahajan, D., Tople, S., & Sharma, A. (2021). Domain generalization using causal matching. In *International conference on machine learning* (pp. 7313–7324). PMLR.
- Mallick, A., Dwivedi, C., Kailkhura, B., Joshi, G., & Han, T. Y.-J. (2021). Deep kernels with probabilistic embeddings for small-data learning. In *Uncertainty in artificial intelligence* (pp. 918–928). PMLR.
- Motiian, S., Piccirilli, M., Adjero, D. A., & Doretto, G. (2017). Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision* (pp. 5715–5725).
- Mridha, M. F., Ohi, A. Q., Hamid, M. A., & Monowar, M. M. (2022). A study on the challenges and opportunities of speech recognition for Bengali language. *Artificial Intelligence Review*, 55(4), 3431–3455.
- Muandet, K., Fukumizu, K., Dinuzzo, F., & Schölkopf, B. (2012). Learning from distributions via support measure machines. *Advances in Neural Information Processing Systems*, 25, 1–9.
- Muandet, K., Fukumizu, K., Sriperumbudur, B., & Schölkopf, B. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1–2), 1–141.
- Nam, H., Lee, H., Park, J., Yoon, W., & Yoo, D. (2021). Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8690–8699).
- Neal, R. M. (2012). *Bayesian learning for neural networks* (Vol. 118). Springer.
- Neculai, A., Chen, Y., & Akata, Z. (2022). Probabilistic compositional embeddings for multimodal image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4547–4557).
- Nguyen, D. Q., Nguyen, D. Q., Modi, A., Thater, S., & Pinkal, M. (2017). A mixture model for learning multi-sense word embeddings. arXiv preprint [arXiv:1706.05111](https://arxiv.org/abs/1706.05111)
- Nuriel, O., Benaim, S., & Wolf, L. (2021). Permuted AdaIN: Reducing the bias towards global statistics in image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9482–9491).
- Oh, S. J., Murphy, K., Pan, J., Roth, J., Schroff, F., & Gallagher, A. (2018). Modeling uncertainty with hedged instance embedding. arXiv preprint [arXiv:1810.00319](https://arxiv.org/abs/1810.00319)
- Park, J., Lee, J., Kim, I.-J., & Sohn, K. (2022). Probabilistic representations for video contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14711–14721).
- Qian, H., Pan, S. J., & Miao, C. (2021). Latent independent excitation for generalizable sensor-based cross-person activity recognition. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, pp. 11921–11929).
- Qi, Q., Lin, X., Chen, C., Xie, W., Huang, Y., Ding, X., Liu, X., & Yu, Y. (2020). Curriculum feature alignment domain adaptation for epithelium–stroma classification in histopathological images. *IEEE Journal of Biomedical and Health Informatics*, 25(4), 1163–1172.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.
- Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint [arXiv:1911.08731](https://arxiv.org/abs/1911.08731)
- Shi, Y., & Jain, A. K. (2019). Probabilistic face embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6902–6911).
- Silnova, A., Brümmer, N., Rohdin, J., Stafylakis, T., & Burget, L. (2020). Probabilistic embeddings for speaker diarization. arXiv preprint [arXiv:2004.04096](https://arxiv.org/abs/2004.04096)
- Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. *Advances in Neural Information Processing Systems*, 29, 1–9.
- Sun, B., & Saenko, K. (2016). Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision* (pp. 443–450). Springer.
- Sun, J. J., Zhao, J., Chen, L.-C., Schroff, F., Adam, H., & Liu, T. (2020). View-invariant probabilistic embedding for human pose. In *Computer vision—ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16* (pp. 53–70). Springer.
- Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5), 988–999.
- Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., & Bengio, Y. (2019). Manifold mixup: Better representations by interpolating hidden states. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning. Proceedings of machine learning research* (Vol. 97, pp. 6438–6447). PMLR. <https://proceedings.mlr.press/v97/verma19a.html>
- Wang, Y., Li, H., Chau, L.-p., & Kot, A. C. (2021). Embracing the dark knowledge: Domain generalization using regularized knowledge distillation. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 2595–2604).
- Wilson, A. G., & Izmailov, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. *Advances in Neural Information Processing Systems*, 33, 4697–4708.
- Xiao, Z., Shen, J., Zhen, X., Shao, L., & Snoek, C. (2021). A bit more Bayesian: Domain-invariant learning with uncertainty. In *International conference on machine learning* (pp. 11351–11361). PMLR.
- Yao, X., Bai, Y., Zhang, X., Zhang, Y., Sun, Q., Chen, R., Li, R., & Yu, B. (2022). PCL: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7097–7107).
- Yoshikawa, Y., Iwata, T., & Sawada, H. (2014). Latent support measure machines for bag-of-words data classification. *Advances in Neural Information Processing Systems*, 27, 1–9.
- Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A., Keutzer, K., & Gong, B. (2019). Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2100–2110).

- Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M., & Lee, B. (2022). A survey of modern deep learning based object detection models. *Digital Signal Processing*, *126*, 103514.
- Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2022). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(4), 4396–4415.
- Zhou, K., Yang, Y., Hospedales, T., & Xiang, T. (2020). Learning to generate novel domains for domain generalization. In *European conference on computer vision* (pp. 561–578). Springer.
- Zhou, K., Yang, Y., Qiao, Y., & Xiang, T. (2021). Domain generalization with mixstyle. arXiv preprint [arXiv:2104.02008](https://arxiv.org/abs/2104.02008)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.