Domain Generalization for Medical Imaging Classification with Linear-Dependency Regularization

Haoliang Li[1]   **YuFei Wang**[1]   **Renjie Wan**[1]   **Shiqi Wang**[2]   **Tie-Qiang Li**[3,4]   **Alex C. Kot**[1]

[1]Rapid-Rich Object Search Lab, Nanyang Technological University, Singapore
[2]Department of Computer Science, City University of Hong Kong, China
[3]Department of Clinical Science, Intervention, and Technology, Karolinska Institute, Sweden
[4]Department of Medical Radiation and Nuclear Medicine, Karolinska University Hospital, Sweden
{lihaoliang,yufei001,rjwan,eackot}@ntu.edu.sg   shiqwang@cityu.edu.hk   tie-qiang.li@ki.se

Abstract

Recently, we have witnessed great progress in the field of medical imaging classification by adopting deep neural networks. However, the recent advanced models still require accessing sufficiently large and representative datasets for training, which is often unfeasible in clinically realistic environments. When trained on limited datasets, the deep neural network is lack of generalization capability, as the trained deep neural network on data within a certain distribution (e.g. the data captured by a certain device vendor or patient population) may not be able to generalize to the data with another distribution. In this paper, we introduce a simple but effective approach to improve the generalization capability of deep neural networks in the field of medical imaging classification. Motivated by the observation that the domain variability of the medical images is to some extent compact, we propose to learn a representative feature space through variational encoding with a novel linear-dependency regularization term to capture the shareable information among medical data collected from different domains. As a result, the trained neural network is expected to equip with better generalization capability to the "unseen" medical data. Experimental results on two challenging medical imaging classification tasks indicate that our method can achieve better cross-domain generalization capability compared with state-of-the-art baselines.

1   Introduction

Due to the breakthrough in machine learning and deep learning, recent years have witnessed numerous significant successes in various medical imaging tasks. However, one of the limitations of deep learning is that it lacks generalization capability when the number of training data is not sufficient [43]. In practice, it is often the case that the testing data (a.k.a. target domain) can be dissimilar to the training data (a.k.a. source domain) in terms of many factors, such as imaging protocol, device vendors and patient populations. Such domain shift problem can lead to a significantly negative impact on the performance of medical imaging classification. To tackle such domain shift problem, domain adaptation [30] aims to transfer the knowledge from a source domain to a different but relevant target domain. Recently, many studies have been conducted to improve the transferable capability in the field of medical imaging classification with domain adaptation by assuming that target domain data are accessible [42, 8].

In many cases, requiring to access the target domain data in advance may not be feasible. For example, in the real-time clinical application scenario, it is difficult to collect sufficient target domain data to

---

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

help with network training. For another example, it is also difficult to access the target domain data as many medical data are protected by privacy regulation. Thus, it is natural to ask whether we can still learn a generalized deep neural network without any prior knowledge regarding the target domain. Domain generalization has been proposed to tackle this problem by assuming to have no access to the target information but utilizing multiple source domains' information to better generalize to the "unseen" new domain for testing.

Generally speaking, current research regarding domain generalization in the field of medical imaging classification can be categorized into two streams. The first stream aims at conducting data augmentation based on medical imaging data in terms of image quality, image appearance and spatial shape [41]. Although the variation of medical images turns out to be more compact as the capturing environment can be fixed in advanced compared with the images captured in our daily life, it may be difficult to choose suitable augmentation types and magnitudes for clinical deployment purposes in a certain environment. The other stream leverages the advantage of domain alignment or meta-learning methods for feature representation learning [38, 7]. However, the learned feature representation may still suffer from the overfitting problem, as the feature representations are only shareable among multiple source domains which may not be able to generalize to target.

In this work, we propose to marriage the advantage of data augmentation and domain alignment to tackle the domain generalization problem for medical imaging classification. Instead of directly conducting augmentation in the image domain through some linear transformations with pre-defined parameters [41], we assume that there exists linear dependency in a latent space among various domains. To model such linear dependency, we propose to train a deep neural network with a novel rank regularization term on latent feature space by setting the rank of latent feature to be the number of categories. Meanwhile, we also propose to restrict the distribution of latent features to follow a pre-defined prior distribution through variational encoding. We theoretically prove that an upper bound on the empirical risk of any "unseen" but related target domain can be achieved under our formulation, such that the overfitting problem can be alleviated. Experimental results on two challenging medical imaging classification tasks, including imbalanced-category based skin lesion classification as well as spinal cord gray matter segmentation (which can be treated as pixel-wise classification), indicate that our proposed method can achieve much better generalization capability compared with other state-of-the-art baselines. The code is available at https://github.com/wyf0912/LDDG.

2   Related Works

Domain Adaptation and Generalization.   To tackle the domain-shift problem between source and target domain data, traditional domain adaptation approaches focused on either subspace learning or instance re-weighting [16, 29, 40, 11]. Deep learning methods are also proved to be effective for domain adaptation task through either distribution alignment (e.g. Maximum Mean Discrepancy) [25] or adversarial learning through feature level [10, 34] or pixel level [3]. Recently, it has been shown that by considering pixel level adaptation and feature level adaptation together, better adaptation performance can be achieved [15, 23].

Compared with domain adaptation, domain generalization is much more challenging, as we assume that we have no access to the target domain. Instead, we aim to train a model that is expected to be generalized to the "unseen" target by assuming that only multiple source domains are available. For example, Yang and Gao [37] proposed to leverage Canonical Correlation Analysis (CCA) to extract shareable information among domains. Muandet *et al.* [28] proposed a Domain Invariant Component Analysis (DICA) algorithm to learn an empirical mapping based on multiple source-domain data where the distribution mismatch across domains was minimized. This idea was further extended by [22, 24] in an autoencoder framework with distribution regularization on latent space. In [36, 20], the low-rank regularization based on classifier and model parameters were explored to extract universal feature representation. Ghifary *et al.* [12] proposed a multi-task autoencoder to learn domain invariant features by reconstructing the latent representation of a given sample from one domain to another domain. Motiian *et al.* [27] proposed to minimize the semantic alignment loss as well as the separation loss based on deep learning models. Carlucci *et al.* [4] proposed to shuffle the image patch to learn generalized feature representation. Recently, Wang *et al.* [35] proposed to extend MixUp [39] to the settings of multiple domains for heterogeneous domain generalization task. As for meta-learning based techniques, Li *et al.* [21] proposed to transfer the idea in [9] to the "unseen" target domain setting by randomly constructing meta-train and meta-test set, which was

further extended by Balaji *et al.* [1] with a scheme to learn a regularization network to improve the scalability of domain generalization.

Cross-Domain Medical Imaging Classification. Due to the various imaging protocols, device vendors and patient populations, we may also encounter the problem of distribution shift in clinical practice. To tackle such domain shift problem, image synthesis can be adopted through Generative Adversarial Networks [13, 44] to mitigate the domain shift problem. For example, Zhang *et al.* [42] proposed to leverage CycleGAN for medical imaging problem to transfer the knowledge from CT images to X-ray images. Chen *et al.* [5] conducted domain translation from MR to CT domain for heart segmentation problem. With few label information available in target domain, Zhang *et al.* [43] proposed to conduct segmentation and data synthesis jointly to segment heart chambers in both CT and MR domain. Dou *et al.* [8] proposed a two parallel domain-specific encoders and a decoder where the weights are shared between domains to boost the performance of training on both single domain and cross-domain scenario. When target domain data are not available, Zhang *et al.* [41] proposed to conduct data augmentation on source domain to achieve better generalization capability for medical imaging classification task. Yoon *et al.* [38] proposed to learn generalized feature representation through classification and contrastive semantic alignment technique [27]. More recently, Dou *et al.* [7] proposed to conduct meta-learning with global class alignment as well as local sample clustering regularization for medical imaging classification task.

3 Methodology

Preliminary. We denote the training samples from multiple source domains on a joint space $\mathcal{X} \times \mathcal{Y}$ as $\mathcal{D} = \{(x_i^k, y_i^k)\}_{i=1}^{N_k}, k \in \{1, 2, ..., K\}$, where x_i^k denotes the ith training sample from the kth source domain, N_k is the number of samples in the kth domain, and y_i^k is the corresponding label groundtruth. The goal of domain generalization is that given a sample x_T from an unseen domain, we aim to predict its output $\hat{y}^T$ through a trained classifier.

We provide a framework named Linear-Dependency Domain Generalization (LDDG) that improves the generalization capability of medical imaging classification. By assuming that there exists linear dependency in the latent feature space among various domains based on a certain task, we propose to regularize the latent feature space by modeling intra-class variation among multiple source domains through rank constraint meanwhile matching the distribution of latent features extracted from multiple source domains to a pre-defined distribution prior, such that the shareable information among domains can be learned. The details of our proposed method are introduced below.

Linear-Dependency Modeling. Directly training a classification network with a task-specific loss (e.g. cross-entropy loss) may not be feasible, as in the field of medical imaging, it is difficult to collect large and diverse datasets, which can lead to poor generalization on a new and "unseen" domain. To improve the generalization capability of medical imaging classification, in [41], based on the observation that medical image domain variability is more compact compared with other image data, a data augmentation approach was proposed based on three different aspects: image quality (e.g., blurriness), image appearance (e.g., brightness) and spatial shape (e.g., rotation, scaling), by assuming other characteristics are supposed to be more consistent. However, we empirically find that it is challenging to choose a suitable augmentation type as well as its magnitude for a specific medical imaging classification task.

Inspired by the observation that most of the aforementioned augmentation processes can be conducted through linear transformation, we assume that there exists linear dependency on the latent feature space. To be more specific, by assuming that we have a medical image batch collected from K different domains with label c as $\{x_{i_1,c}^1, x_{i_2,c}^2, ..., x_{i_K,c}^K\}$, there exists a set of parameters $\{\alpha_1, \alpha_2, ..., \alpha_K\}$ such that the corresponding latent features $\{z_{i_1,c}^1, z_{i_2,c}^2, ..., z_{i_K,c}^K\}$ hold the property that $z_{i_j,c}^j = \alpha_1 z_{i_1,c}^1 + \alpha_2 z_{i_2,c}^2 + ... + \alpha_{j-1} z_{i_{j-1},c}^{j-1} + \alpha_{j+1} z_{i_{j+1},c}^{j+1} + ... + \alpha_K z_{i_K,c}^K$ for different j. In other words, there exists a dominant eigenvalue capturing the category information of the matrix $[z_{i_1,c}^1, z_{i_2,c}^2, ..., z_{i_K,c}^K]$. Therefore, given a sample mini-batch denoted by $\mathcal{X} = \{x_i^k\}$, we can obtain the corresponding latent features as $\mathcal{Z}$ through a posterior $q(z|x)$ parameterized by an encoder. By further conducting mode-1 flattening $\mathcal{Z}$ as $\mathbf{Z}$[1], our proposed rank regularization can be given as $rank(\mathbf{Z}) = C$, where C is the number of categories of a specific task.

---

[1] We assume that the first dimension is associated with sample index.

Setting the rank of $\mathbf{Z}$ to C is equivalent to minimize the $(C+1)$th singular value of $\mathbf{Z}$. By denoting it as σ_{C+1}, we can reformulate the rank loss and compute its sub-gradient as

$$\mathcal{L}_{rank} = \sigma_{C+1}, \qquad \frac{\partial \sigma_{C+1}}{\partial \mathbf{Z}} = \mathbf{U}_{:,C+1}\mathbf{V}_{:,C+1}^{\top}, \tag{1}$$

where $\mathbf{U}$ and $\mathbf{V}$ are obtained through SVD as $\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$.

Noted that it is quite common to impose low-rank regularization in the final objective for domain generalization task [36, 20]. Our proposed method is different from these methods in two folds, 1) we impose rank regularization based on the latent feature space while the existing works imposed low-rank regularization on classifier parameters, which are not computational efficiency; 2) we set the rank to be a specific number instead of simply conducting low-rank regularization, we show in the experimental section that it can lead to better performance.

Distribution Alignment. In addition to modeling the linear dependency in latent feature space, we further propose to extract shareable information among multiple domains, such that a more transferable feature representation can be learned which can benefit generalization capability of deep neural networks. Existing techniques aimed to either minimize domain variance through distribution alignment between domain pairs [28, 22] or conduct local sample clustering through contrastive loss or triplet loss [27, 7]. However, based on our empirical analysis, the aforementioned technique may suffer from overfitting problem to the source domains, which is not surprising as the distribution of "unseen" target domain may not match the distribution of multiple source domains. Thus, simply minimizing domain variance or local sample cluttering on source domains may not be able to generalize well to the "unseen" one. To this end, we propose to conduct variational encoding [18] by adopting Kullback-Leibler (KL) divergence, which aims to match the latent features from multiple source domains to a pre-defined prior distribution. In our work, we adopt Gaussian distribution $\mathcal{N} \sim (0,1)$ as the prior distribution, which is computationally tractable through reparameterization trick [18]. The KL divergence can be formulated as $KL(q(\mathcal{Z}|\mathcal{X})||\mathcal{N} \sim (0,1))$, where $\mathcal{Z}$ is the latent features defined in the previous section. We show in the next section that by jointly conducting linear-dependency modeling and distribution regularization through KL divergence can lead to an upper bound of empirical risk from any "unseen" but related domains.

Theoretical Analysis. In this section, we provide the theoretical analysis of our proposed framework. In particular, We show that our framework can lead to an upper bound of expected loss on "unseen" but related target domain. We first make the following assumptions:

Assumption 1. *For any latent feature belong to domain T with label c, it can be represented by data from other related domains, i.e., $q(z_{i_T,c}^T|x_{i_T,c}^T) = \sum_{j=1}^{K} \beta_j q(z_{i_j,c}^j|x_{i_j,c}^j)$, where $\beta_j >= 0$ and $\|\beta\| \leq M$, $\{x_{i_1,c}^1, x_{i_2,c}^2, ..., x_{i_K,c}^K\}$ belong to the same category as x_T.*

Noted that Assumption 1 is a mild assumption in the field of medical imaging classification task and is also reasonable in our setting as we restrict the rank of latent features to be the number of category, such that there exists linear dependency based on the latent features belonging to the same category.

We further make assumption on the loss function $\mathcal{L}$ based on the output of classifier.

Assumption 2. *(1) $\mathcal{L}$ is non-negative and bounded. (2) $\mathcal{L}$ is convex: $\mathcal{L}(\sum_j \lambda_j y_j, y) \leq \sum_j \lambda_j \mathcal{L}(y_j, y)$, where $\lambda_j \geq 0$ and $\sum_j \lambda_j = 1$.*

Note that this assumption is easy to be satisfied for several standard loss functions (e.g. cross-entropy loss).

Under these assumptions, we have the following theorems.

Theorem 1. *Given a sample $x_{i_T,c}^T$ from target domain T where the distribution of its latent variable is represented as $q(z_{i_T,c}^T|x_{i_T,c}^T) = \sum_{j=1}^{K} \beta_j q(z_{i_j,c}^j|x_{i_j,c}^j)$, its latent variable is within the manifold of $\mathcal{N} \sim (0,1)$.*

4

Proof. For simplicity, we first denote the distribution on latent variables as $q_i(z), i = \{1, 2, ..., K, T\}$ as well as the Gaussian prior $\mathcal{N} \sim (0, 1)$ as $q_*(z)$. We can obtain the following upper bound,

$$KL(q_T(z)||q_*(z)) = \sum_{j=1}^{K} \beta_j \int_z q_j(z) \log \frac{q_T(z)}{q_*(z)} dz$$

$$= \sum_{j=1}^{K} \beta_j \int_z q_j(z) \log \frac{q_j(z)[1 + (q_T(z)/q_j(z) - 1)]}{q_*(z)} dz \leq \sum_{j=1}^{K} \beta_j KL(q_j(z)||q_*(z)), \quad (2)$$

where we use $\log(1 + x) \leq x$ and $\int q_T(z)dz = \int q_j(z)dz = 1$. As $KL(q_j(z)||q_*(z))$ is minimized according to our proposed distribution alignment, $KL(q_T(z)||q_*(z))$ is then minimized.

This completes the proof. $\square$

Theorem 1 shows that the latent feature of any unseen but related domain lies in the manifold of pre-defined prior. With the help of Theorem 1, we can further derive the upperbound of empirical risk of target domain.

Theorem 2. *Given data from K source domains, where the empirical risk of domain j is given as $\mathcal{L}(\hat{y}^j, y) = \epsilon_j \leq \epsilon$, the expected loss $\mathcal{L}(\hat{y}^T, y)$ is at most $M\epsilon + \log C$, where C denotes the number of category given a task, if the classification layer is linear with softmax normalization trained by $\mathcal{L}$ which is a cross-entropy loss.*

Proof. Based on Theorem 1, we have $q_T(z) = q_*(z)$. Thus, we have the following upper bound,

$$\int_z \mathcal{L}(\hat{y}^T, y)q_T(z)dz = \int_z \mathcal{L}(\sum_{j=1}^{K} \beta_j \hat{y}^j, y)q_*(z)dz = \int_z \mathcal{L}(\|\beta\| \sum_{j=1}^{K} \frac{\beta_j}{\|\beta\|} \hat{y}^j, y)q_*(z)dz$$

$$\leq \sum_{j=1}^{K} \frac{\beta_j}{\|\beta\|} \int_z \mathcal{L}(\|\beta\|\hat{y}^j, y)q_j(z)dz \leq \sum_{j=1}^{K} \frac{\beta_j}{\|\beta\|} M\epsilon + \log C = M\epsilon + \log C, \quad (3)$$

where $\mathcal{L}(\cdot)$ denotes the cross-entropy loss with softmax operation. Noted that we only adopt a linear layer for classifier, thus $\hat{y}^T = \sum_{j=1}^{K} \beta_j \hat{y}^j$ holds based on Assumption 1. We also utilize the bounds of Log-Sum-Exp function $f(a) \leq \max\{a_1, a_2, \ldots, a_n\} + \log n$, where $f(a) = \log \sum_{i=1}^{n} \exp(a_i)$. In our work, $\{a_1, a_2, \ldots a_n\}$ corresponds to the softmax output of n different nodes. Thus, the second line of proof holds.

This completes the proof. $\square$

Theorem 2 shows that our proposed method has a good generalization capability. If the empirical risks on source domains are small, the empirical risk on target domain is also expected to be small.

Model Training. Our proposed architecture consists of three part, a feature extractor Q_θ, a variational encoding network F_ω, and a classification network T_ϕ. Regarding the classification network, we only adopt a linear module (e.g. convolutional layer, linear layer) without any non-linear processing, such that our assumption can be satisfied. Images $\mathcal{X} = \{x_i^k\}$ are first fed into the feature extractor Q_θ to obtain the latent features, and then the latent features are resampled [18] through variational encoding network F_ω, finally the classification network T_ϕ outputs the corresponding prediction $\{\hat{y}_i^k\}$. A cross-entropy loss together with rank and distribution regularization to penalize the difference between $\{\hat{y}_i^k\}$ and the groundtruth label $\{y_i^k\}$, the distribution difference between latent features and the Gaussian prior as well as the rank of the latent features. In summary, our model can be trained by minimizing the following objective as

$$\mathcal{L}_{obj} = \sum_{i,k} \mathcal{L}_c(\hat{y}_i^k, y_i^k) + \lambda_1 \mathcal{L}_{rank} + \lambda_2 KL(q(\mathcal{Z}|\mathcal{X})||\mathcal{N} \sim (0, 1)), \quad (4)$$

where $\mathcal{L}_c(\hat{y}_i^k, y_i^k)$ denotes the cross-entropy loss with softmax operation, $\mathcal{L}_{rank}$ is the rank loss defined in Equation 1.

Table 1: Domain generalization results on the skin lesion classification task. We repeat experiment for 5 times for each technique and report the mean value and standard deviation.

Target	DeepAll	MASF [7]	MLDG [21]	CCSA [38]	LDDG (Ours)
DMF	0.2492±0.0127	0.2692±0.0146	0.2673±0.0452	0.2763±0.0263	**0.2793**±0.0244
D7P	0.5680±0.0181	0.5678±0.0361	0.5662±0.0212	0.5735±0.0227	**0.6007**±0.0208
MSK	0.6674±0.0083	0.6815±0.0122	0.6891±0.0167	0.6826±0.0131	**0.6967**±0.0193
PH2	0.8000±0.0167	0.7833±0.0101	0.8016±0.0096	0.7500±0.0419	**0.8167**±0.0096
SON	0.8613±0.0296	0.9204±0.0227	0.8817±0.0198	0.9045±0.0128	**0.9272**±0.0117
UDA	0.6264±0.0312	0.6538±0.0196	0.6319±0.0284	0.6758±0.0138	**0.6978**±0.0110
Avg	0.6287	0.6460	0.6396	0.6438	**0.6697**

4 Experiments

In this section, we evaluate our proposed method based on two different medical imaging classification tasks: skin lesion classification task and gray matter segmentation task of spinal cord. The detail of architectures and experimental settings can be found in supplementary materials.

4.1 Skin lesion classification

We adopt seven public skin lesion datasets, including HAM10000 [33], Dermofit (DMF) [2], Derm7pt (D7P) [17], MSK [6], PH2 [26], SONIC (SON) [6], and UDA [6], which contain skin lesion images collected from different equipments. We follow the protocol in [38] by choosing seven-category subset from these datasets, including melanoma (mel), melanocytic nevus (nv), dermatofibroma (df), basal cell carcinoma (bcc), vascular lesion (vasc), benign keratosis (bkl), and actinic keratosis (akiec). Each dataset is randomly divided into 50% training set, 20% validation set and 30% testing set, where the relative class proportions are maintained across dataset partitions. As suggested in [38], for each setting, we use one dataset from DMF, D7P, MSK, PH2, SON and UDA as target domain and the remaining datasets together with HAM10000 as source domains. We use a ResNet18 model [14] pretrained on ImageNet as the backbone for our proposed method as well as other baselines.

Results: We compare our method with state of the art domain generalization methods, including MASF [7], MLDG [21], and CCSA [38], which have shown the capability to generalize across domains in the field of medical imaging. We report the baseline results by tuning the hyper-parameters in a wide range. We also adopt the baseline by directly training the model with the classification loss, which is referred as "DeepAll". The results are shown in Table 1.

As we can see, all the domain generalization based techniques can outperform the DeepAll by directly training on source domains with classification loss. Among the domain generalization methods, MASF can achieve relatively better performance compared with MLDG and CCSA, as it is built upon meta-learning mechanism by considering both global and local based domain alignment. Compared with all baselines, our proposed algorithm can achieve better performance in a clear margin, which is reasonable as our proposed training mechanism leverage the advantage of both data augmentation and domain alignment, which is less likely to suffer from overfitting problem. We can also observe that in some cases, all algorithms can perform relatively well, which we conjecture that the domain gap between source and target domain is relatively small. However, the performances are not desired in some cases (e.g. when using DMF as target domain), which may be due to the large domain gap between source and target domain. This observation is also consistent with the results reported in [38]. We also experiment using the data augmentation based technique BigAug [41] by considering a wide-range of augmentation spaces but find it cannot yield competitive performance and the results are omitted here for brevity. We conjecture the reason that the augmentation types may not be suitable for skin lesion classification task.

In practice, it is also likely that we only have the data from one single domain during training. To further analyze the effectiveness of our proposed method under this scenario, we consider HAM10000 for training and the others for testing for skin lesion classification task. Besides directly training with classification loss (DeepAll), we also compare with CCSA [38] and MixUp [39]. Noted that other domain generalization baselines are not applicable in this case, as they require multiple domains available to simulate domain shift. The results are shown in Table 2. In most of the cases, our proposed method can outperform the DeepAll, CCSA as well as MixUp. For CCSA, as the contrastive loss is applied on only one source domain while target domain is unseen, it is likely to suffer from overfitting

Table 2: Domain generalization results with HAM10000 as source domain.

	DMF	D7P	MSK	PH2	SON	UDA	Average
DeepAll	0.3003	0.4972	0.1667	0.4945	0.5025	0.4945	0.4093
CCSA [38]	0.2762	0.5082	0.4652	0.4667	0.5275	0.5055	0.4582
MixUp [39]	**0.3514**	0.4029	0.3000	0.4333	0.6296	0.4615	0.4298
Ours	0.2943	**0.5191**	**0.5087**	**0.5500**	**0.6949**	**0.5714**	**0.5231**

Table 3: Domain generalization results on gray matter segmentation task.

(a) DeepAll

source	target	DSC	CC	JI	TPR	ASD
2,3,4	1	0.8560	65.34	0.7520	0.8746	0.0809
1,3,4	2	0.7323	26.21	0.5789	0.8109	0.0992
1,2,4	3	0.5041	-209	0.3504	0.4926	1.8661
1,2,3	4	0.8775	71.92	0.7827	0.8888	0.0599
Average		0.7425	-11.4	0.6160	0.7667	0.5265

(b) Probabilistic U-Net [19]

source	target	DSC	CC	JI	TPR	ASD
2,3,4	1	0.8387	59.94	0.7276	0.8943	0.1820
1,3,4	2	0.8067	51.53	0.6778	0.7555	0.0580
1,2,4	3	0.5113	-188	0.3550	0.5638	2.0866
1,2,3	4	0.8782	72.18	0.7833	0.8910	0.2183
Average		0.7587	-1.09	0.6359	0.7762	0.6362

(c) MASF [7]

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**	**0.8991**	0.0366
Average		0.7710	23.52	0.6502	0.7803	0.5505

(d) MLDG [21]

source	target	DSC	CC	JI	TPR	ASD
2,3,4	1	0.8585	64.57	0.7489	0.8520	0.0573
1,3,4	2	0.8008	49.65	0.6696	0.7696	0.0745
1,2,4	3	0.5269	-108	0.3668	0.5066	1.7708
1,2,3	4	0.8837	73.60	0.7920	0.8637	0.0451
Average		0.7675	19.96	0.6443	0.7480	0.4869

(e) CCSA [38]

source	target	DSC	CC	JI	TPR	ASD
2,3,4	1	0.8061	50.15	0.6801	0.8703	0.1678
1,3,4	2	0.8009	50.04	0.6687	0.8141	0.0939
1,2,4	3	0.5012	-112	0.3389	0.5444	1.5480
1,2,3	4	0.8686	69.61	0.7684	0.8926	0.0449
Average		0.7442	14.45	0.6140	0.7804	0.4637

(f) LDDG (Ours)

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**	**0.8485**	**0.0416**
1,2,4	3	**0.5543**	**-71.6**	**0.3889**	**0.5923**	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**

problem. For MixUp, the combination is conducted only in a convex manner, which may not be able to generalize well to the out-of-distribution target domain.

4.2 Spinal cord gray matter segmentation

We then consider the task of gray matter segmentation of spinal cord based on magnetic resonance imaging (MRI) to evaluate our proposed method. In particular, we adopt the data from spinal cord gray matter segmentation challenge [31], which are collected from four different medical centers with different MRI systems (Philips Achieva, Siemens Trio, Siemens Skyra). The voxel size resolutions are ranging from $0.25 \times 0.25 \times 2.5$mm to $0.5 \times 0.5 \times 5$mm. To evaluate the generalization capability of our proposed method, we consider the data collected from one medical center as a domain, which leads to four different domain, namely "site1", "site2", "site3" and "site4", where one domain is adopted as target domain and the remaining are considered as source domains. We adopt 2D-UNet [32] as the backbone network by considering the MRI axial slice as input[2]. Due to the imbalance of the number of voxels belonging to spinal cord gray matter and background in the MRI image, we follow [31] to consider a two-stage strategy in a coarse-to-fine manner: 1) segment the spinal cord area (where the groundtruth of spinal cord is available), 2) segment the gray matter area from the output of 1) for our proposed method as well as baselines for comparison.

Results: We compare our method with state of the art domain generalization methods, including MASF [7], MLDG [21], CCSA [38], by tuning the baseline hyper-parameters in a wide range. Moreover, we also compare with the Probabilistic U-Net [19] which automatically learned a prior for medical imaging segmentation task. For quantitative evaluation, we use a number of metrics to validate the effectiveness of our method. In particular, the metrics include three overlapping metrics: Dice Similarity Coefficient (DSC), Jaccard Index (JI) and Conformity Coefficient (CC); One

---

[2]Noted that we have tried 3D-UNet as suggested in [31] but find the performances are similar to 2D-UNet in cross-domain scenario.
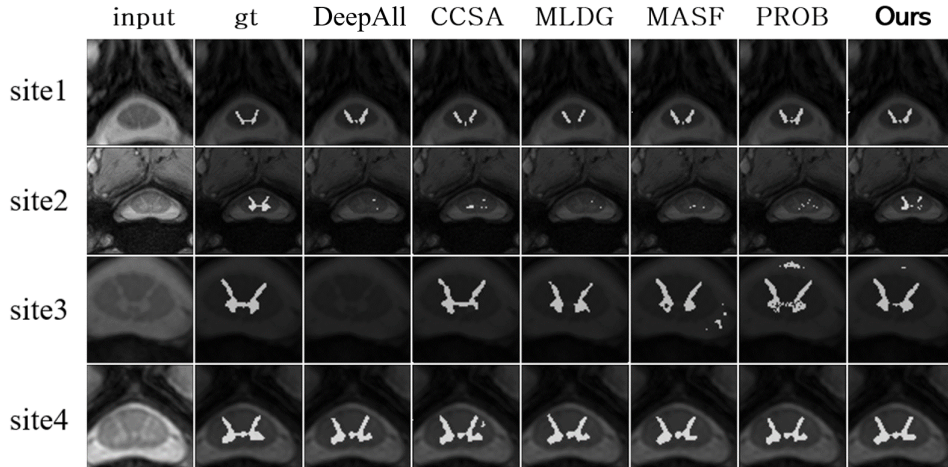
Figure 1: Qualitative comparisons. Each row represents a sample from a specific domain. Each column denotes the input, ground truth (gt) or different methods including DeepAll, CCSA [38], MLDG [21], MASF [7], Probabilistic U-Net [19] (abbreviated as PROB here), respectively. As the area of interest in the original samples is very small, all the samples are center cropped for better visualization.

statistical based metrics: Sensitivity (a.k.a. True Positive Rate (TPR))[3]; One distance based metric: Average surface distance (ASD), which are all performed in 3D. The results are shown in Table 3.

As we can observe, our method achieve the best results when using "site1", "site2" and "site3" as target domain based on all metrics, and achieve the best results in "site4" under ASD, which shows the effectiveness of our proposed method. Among all other domain generalization based methods, MASF can achieve better performance compared with CCSA and MLDG. Such results are consistent with the performance for skin lesion classification task. We also observe that probabilistic U-Net can achieve relatively better performance compared with the "DeepAll" baseline by directly training on source domain with classification loss. However, it may still suffer from overfitting problem as the learned prior may not generalize well to "unseen" target domain. Again, we also evaluate the method proposed in [41] by considering a wide range of data augmentation parameters but find the results are not desired. We conjecture that the default augmentation types are not suitable for this task. Some results of adopting [41] can be found in the supplementary materials.

We further show some qualitative results in Figure 1. As we can see, while the "DeepAll" baseline as well as other domain generalization based methods fail to segment the gray matter (e.g. when using "site2" as target domain) or over-segment a large portion of gray matter by extending the segmentation maps to the white matter (e.g. when using "site3" as target domain), our proposed method can generally achieve better performance compared with all the methods for comparison.

4.3 Ablation Study

we first conduct experiments on skin lesion classification task to understand the impact of different components of our proposed algorithm by considering UDA as target domain. The results are shown in Table 4, where "Rank" and "KL" denote our proposed rank regularization and distribution alignment through KL divergence minimization, respectively. "LR" denotes the rank regularization with nuclear norm minimization, which is a popular way to conduct low-rank constraint. We have the following observations: 1) both rank regularization and distribution alignment through KL can benefit the generalization capability for medical imaging classification task, which is reasonable as adopting these two terms jointly can theoretically lead to a empirical risk upper bound on target domain; 2) our proposed rank regularization can outperform the low-rank regularization by directly

---

[3]We omit True Negative Rate in our case due to the imbalance of the number of voxels belonging to spinal cord gray matter and background in the MRI image.

Rank	-	LR	LR	-	✓	✓
KL	-	-	✓	✓	-	✓
accuracy	0.6264	0.6319	0.6703	0.6703	0.6813	0.6978

Table 4: Ablation study on key components of our method. We choose the skin lession classification task and use UDA as the target domain.
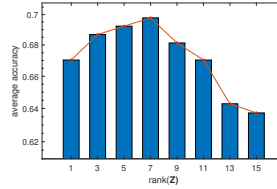

Figure 2: The model performance with different $rank(\mathbf{Z})$. UDA from skin lesion classification task is selected as the target domain.

minimizing the nuclear norm, which is reasonable as the discriminative category specific information can be explored by enforcing the value of rank to be the number of category.

We then evaluate the effectiveness of our proposed rank regularization by varying the rank values of latent features by considering UDA as target domain. The results are shown in Figure 2. As we can observe, the average classification accuracy has an ascending trend first, and then drop. In particular, the accuracy reaches its peak when $rank(\mathbf{Z}) = 7$, which is also the number of category in our task. However, we also observe that the performance drops when $rank(\mathbf{Z})$ gets larger, which is reasonable as increasing the value of $rank(\mathbf{Z})$ may leads to noise information which can have negative impact to the task. Noted that we can still achieve better performance compared with only using low-rank regularization with nuclear norm as shown in Table 4, which is reasonable as the nuclear norm does not take category information into consideration.

Finally, we are interested in the singular values of our proposed method, which can be computed through SVD. In particular, we conduct experiments on segmentation task for stage 1 and stage 2 by considering "site1" as the target domain. We show the convergence results of singular values in Figure 3. As we can see, our proposed method can converge in 100 epochs for both stage 1 and 2 despite the fact that deep neural networks are highly nonlinear. Regarding the singular value, we find that the magnitude for σ_1 and σ_2 are relatively large while other values are much smaller, which is reasonable as we aim to explore the category specific information. Last but not the least, we find that compared with σ_2, the value of σ_1 is much larger, which we conjecture the reason that the area of gray matter (or spinal cord) is much smaller than others.
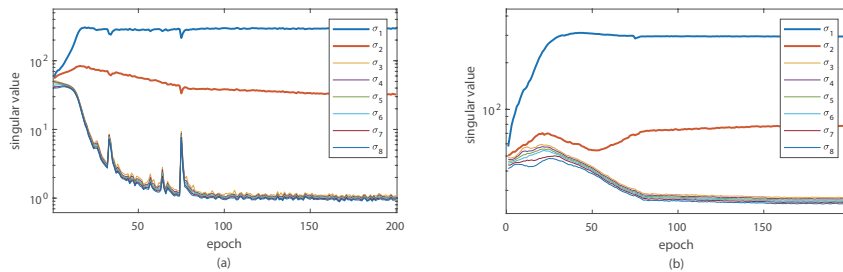

Figure 3: Analysis of singular values, (a) singular values in spinal cord segmentation stage (stage 1), (b) singular values of in gray matter segmentation stage (stage 2).

5 Conclusion

In this paper, we tackle the generalization problem in medical imaging classification task. Our proposed method takes the advantage of both linear-dependency modeling and domain alignment. In particular, we propose to learn a representative feature space for medical imaging classification task through variational encoding with linear-dependency regularization with a novel rank regularization term. Our theoretical analysis shows that an empirical risk upper bound on target domain can be achieved under our formulation. Experimental results on skin lesion classification task and spinal cord gray matter segmentation task show the effectiveness of our proposed method.

Broader Impact. Our proposed method shows reasonable potential in the application of clinically realistic environments especially under the scenarios where only limited training samples are available and the capturing vendors and environments are diverse. In the short-term, the potential beneficiary of the proposed research lies in that it could significantly alleviate the domain shift problem in medical image analysis, as evidenced in this paper. In the long term, it is expected that the principled methodology could offer new insights in intelligent medical diagnostic systems. One concrete example is that the medical imaging classification functionality can be incorporated into different types of smartphones (with different capturing sensors, resolutions, etc.) to assess risk of skin disease (e.g. skin cancer in suspicious skin lesions) such that the terminal stage of skin cancer can be avoided. However, the medical data can be protected by privacy regulation such that the protected attributes (e.g. gender, ethnicity) may not be released publicly for training purpose. In this sense, the trained model may lack of fairness, or worse, may actively discriminate against a specific group of people (e.g. ethnicity with relatively small proportion of people). In the future, the proposed methodology can be feasibly extended to improve the algorithm fairness for numerous medical image analysis tasks and meanwhile guarantee the privacy of the protected attributes.

References

[1] Y. Balaji, S. Sankaranarayanan, and R. Chellappa. Metareg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems*, pages 998–1008, 2018.

[2] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees. A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*, pages 63–86. Springer, 2013.

[3] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 7, 2017.

[4] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi. Domain generalization by solving jigsaw puzzles. *arXiv preprint arXiv:1903.06864*, 2019.

[5] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 865–872, 2019.

[6] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[7] Q. Dou, D. C. de Castro, K. Kamnitsas, and B. Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, pages 6447–6458, 2019.

[8] Q. Dou, C. Ouyang, C. Chen, H. Chen, B. Glocker, X. Zhuang, and P.-A. Heng. Pnp-adanet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation. *arXiv preprint arXiv:1812.07907*, 2018.

[9] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017.

[10] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[11] M. Ghifary, D. Balduzzi, W. B. Kleijn, and M. Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1):1–1, 2017.

[12] M. Ghifary, W. Bastiaan Kleijn, M. Zhang, and D. Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *ICML*, 2018.

[16] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, 2006.

[17] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2):538–546, 2018.

[18] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[19] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. A. Eslami, D. J. Rezende, and O. Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems*, pages 6965–6975, 2018.

[20] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5542–5550, 2017.

[21] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[22] H. Li, S. Jialin Pan, S. Wang, and A. C. Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018.

[23] H. Li, R. Wan, S. Wang, and A. C. Kot. Unsupervised domain adaptation in the wild via disentangling representation learning. *International Journal of Computer Vision*, pages 1–17, 2020.

[24] H. Li, S. Wang, R. Wan, and A. K. Chichung. Gmfad: Towards generalized visual recognition via multi-layer feature alignment and disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[25] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

[26] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5437–5440. IEEE, 2013.

[27] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, 2017.

[28] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

[29] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[30] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[31] F. Prados, J. Ashburner, C. Blaiotta, T. Brosch, J. Carballido-Gamio, M. J. Cardoso, B. N. Conrad, E. Datta, G. Dávid, B. De Leener, et al. Spinal cord grey matter segmentation challenge. *Neuroimage*, 152:312–329, 2017.

[32] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[33] P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161, 2018.

[34] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[35] Y. Wang, H. Li, and A. C. Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3622–3626. IEEE, 2020.

[36] Z. Xu, W. Li, L. Niu, and D. Xu. Exploiting low-rank structure from latent domains for domain generalization. In *ECCV*. 2014.

[37] P. Yang and W. Gao. Multi-view discriminant transfer learning. In *IJCAI*, 2013.

[38] C. Yoon, G. Hamarneh, and R. Garbi. Generalizable feature learning in the presence of data bias and domain class imbalance with application to skin lesion classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 365–373. Springer, 2019.

[39] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[40] K. Zhang, M. Gong, and B. Schölkopf. Multi-source domain adaptation: A causal view. In *AAAI*, volume 1, pages 3150–3157, 2015.

[41] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, H. Roth, A. Myronenko, D. Xu, and Z. Xu. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging*, 2020.

[42] Y. Zhang, S. Miao, T. Mansi, and R. Liao. Task driven generative modeling for unsupervised domain adaptation: Application to x-ray image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 599–607. Springer, 2018.

[43] Z. Zhang, L. Yang, and Y. Zheng. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 9242–9251, 2018.

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.