Combining Data Augmentation and Domain Distance Minimisation to Reduce Domain Generalisation Error

Hoang Son Le, Rini Akmeliawati, Gustavo Carneiro
The University of Adelaide
Email: {hoangson.le, rini.akmeliawati, gustavo.carneiro}@adelaide.edu.au

***Abstract*—Domain generalisation represents the challenging problem of using multiple training domains to learn a model that can generalise to previously unseen target domains. Recent papers have proposed using data augmentation to produce realistic adversarial examples to simulate domain shift. Under current domain adaptation/generalisation theory, it is unclear whether training with data augmentation alone is sufficient to improve domain generalisation results. We propose an extension of the current domain generalisation theoretical framework and a new method that combines data augmentation and domain distance minimisation to reduce the upper bound on domain generalisation error. Empirically, our algorithm produces competitive results when compared with the state-of-the-art methods in the domain generalisation benchmark PACS. We have also performed an ablation study of the technique on a real-world chest x-ray dataset, consisting of a subset of CheXpert, Chest14, and PadChest datasets. The result shows that the proposed method works best when the augmented domains are realistic, but it can perform robustly even when domain augmentation fails to produce realistic samples.**

I. Introduction

Machine learning techniques have demonstrated outstanding results under controlled environments that guarantee the i.i.d. assumption between training and testing domains [1]. However, violations of this condition are common due to changes in latent factors between the training and testing environments. This introduces domain shift, where a model is tested on unseen domains with statistical distributions[1] significantly different from the training domains. For instance, in medical image analysis, classification systems trained with images from a finite number of source (or training) domains can be tested in different target (or testing) domains, with images acquired under different imaging protocols, using devices from different vendors, and from different patient populations [2]. In general, strong domain shifts may severely reduce the classification accuracy at testing. [1].

Domain adaptation and domain generalisation [3, 4, 5, 6] aim to develop robust models in non-i.i.d settings. While domain adaptation utilises examples from testing domains at training [7, 3, 8, 5, 6], domain generalisation has no access to any information from testing domains [9, 10, 11, 2, 6]. As such, several assumptions were made to make the problem tractable: that the marginal distribution of the input changes between domains, and that the conditional distribution remains invariant (invariant labelling function).

Recent papers [11, 2] proposed to address domain generalisation by generating synthetic training data via data augmentation. While preliminary results were encouraging, the method suffers from a lack of robustness. More precisely, indiscriminate applications of data augmentation can produce inconsistencies between the object and the label, rendering the invariant labelling assumption invalid. Data augmentation can avoid such problems only if domain knowledge is present, which is not a guarantee for domain generalisation tasks.

In this paper, we address the shortcomings of pure data augmentation approaches for domain generalisation by introducing a combined loss function that enforces closeness between training domains and augmented domains generated by data augmentation. Based on domain generalisation theory, this approach works well for target domains that can be represented as a convex combination of the training and augmented domains. Unlike common data augmentation procedures for contrastive learning [12], our data augmentation is more aggressive to simulate possible realistic domain shifts. Empirically, we demonstrate the effectiveness of the proposed method hereby referred as DASCL on the benchmark dataset PACS and on a multi-domain medical chest x-ray dataset built from Chexpert, Chest14, and Padchest. The result shows that the proposed method works best when the augmented domains are realistic, but it can perform robustly even when domain augmentation fails to produce realistic samples.

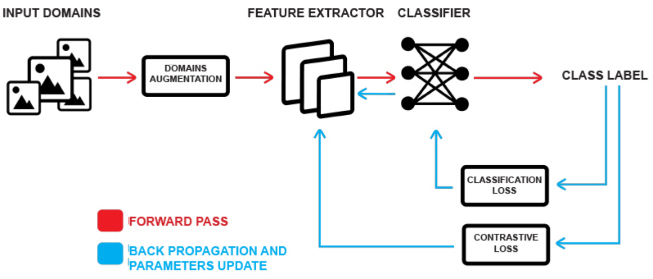

Fig. 1. Main stages of the proposed domain generalisation with domain-augmented supervised contrastive learning algorithm.

[1] In this paper, we use source and training, target and testing, and distribution and domain interchangeably.

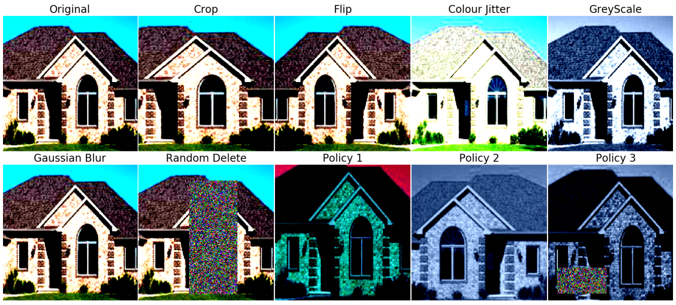

Fig. 2. Domain augmentation functions.

II. Background and Problem Setting

Following the standard nomenclature, we consider the binary classification problem, using $\mathcal{X}$ to represent the image space, $\mathcal{Z}$ to denote the feature space, and $\mathcal{Y}$ the label space, which, without loss of generality, is assumed to be binary, or $\{0, 1\}$. A domain $\mathcal{D}$ is a distribution on $\mathcal{X}$ with the labelling function $g : \mathcal{X} \to [0, 1]$, which corresponds to the probability that the label of $\mathbf{x} \in \mathcal{X}$ is 1. A representation function $f : \mathcal{X} \to \mathcal{Z}$ induces a distribution $\tilde{\mathcal{D}}$ from $\mathcal{D}$ and a corresponding target function $\tilde{g} : \mathcal{Z} \to [0, 1]$ from g. A classifier is a function $h : \mathcal{Z} \to [0, 1]$ drawn from a hypothesis class $\mathcal{H}$ and the classification error w.r.t. the labeling function $\tilde{g}$ of the source domain $\tilde{\mathcal{D}}_S$ is defined by

$$\epsilon_S(h) = \mathbb{E}_{z\sim\tilde{\mathcal{D}}_S}[|\tilde{g}(z) - h(z)|],$$

where $\tilde{\mathcal{D}}_S$ denotes the induced source domain distribution on the feature space $\mathcal{F}$. Similarly, $\epsilon_T(h)$ is the expected error of h in the induced target domain $\tilde{\mathcal{D}}_T$.

Techniques in domain adaptation and generalisation are related, where the main difference, as mentioned before, resides in the assumption about accessibility to unseen domains [9, 10, 11, 2]. The study in [13] establishes the following necessary conditions for domain adaptation: (1) the source and target distributions are close w.r.t the $\mathcal{A}$ distance, (2) the labelling function is invariant to the domain, and (3) there exists a hypothesis in the hypothesis class $\mathcal{H}$ that has low error on both distributions. Even though these conditions are established for domain adaptation, domain generalisation, and particularly multi-source domain generalisation, also require the same conditions.

The goal of a multi-source domain generalisation is to learn a network $h(z)$ that produces low classification errors in an unseen target domain. In this framework, we have access to K datasets sampled from source distributions $\{\mathcal{D}_{S_k}\}_{k=1}^K$, with each dataset being labelled by the corresponding domain labelling function $g_{S_k}(.)$. [14] proposed a generalisation bound for single source domain adaptation that was extended to multi-source domain adaptation in [15] by considering the aggregated source distribution as a convex combination of the realisable source domains (mixed source domain):

$$\mathcal{D}_S = \sum_{k=1}^K \lambda_k \mathcal{D}_{S_k} \quad \text{where} \quad \sum_{k=1}^K \lambda_k = 1.$$

Then for $0 < \delta < 1$, with probability at least $1 - \delta$, and for every $h \in \mathcal{H}$ where $\mathcal{H}$ is a hypothesis class of VC dimension d, the generalisation error is upper-bounded by:

$$\epsilon_T(h) \le \sum_{k=1}^K \lambda_k \left(\hat{\epsilon}_{S_k}(h) + \frac{1}{2} d_{\mathcal{H}\Delta H}(\hat{\mathcal{D}}_{S_k}, \hat{\mathcal{D}}_T)\right) + \theta_\lambda + O\left(\sqrt{\frac{1}{KM}\left(\log\frac{1}{\delta} + d\log\frac{KM}{d}\right)}\right), \tag{1}$$

where $\hat{\epsilon}_{S_k}(h)$ represents the empirical risk for the k^{th} source domain, θ_λ denotes the risk of the optimal hypothesis on the mixture source and target domains, and M is the number of i.i.d. samples from the source and target domains. The upper bound in equation (1) depends on: i) the source domain composition $\{\lambda_k\}_{k=1}^K$, ii) the source classifier empirical errors $\{\hat{\epsilon}_{S_k}(h)\}_{k=1}^K$, iii) the empirical distance (or $\mathcal{A}$ distance) between the source domains and the target domain denoted by $\{d_{\mathcal{H}\Delta\mathcal{H}}(\hat{\mathcal{D}}_{S_k}, \hat{\mathcal{D}}_T)\}_{k=1}^K$, iv) the risk θ_λ, and v) a growth term that depends on number of samples M, number of source domains K, VC-dimension d, and δ. The $\mathcal{A}$ distance in (1) between domains $\mathcal{D}_0$ and $\mathcal{D}_1$ is defined as [14]:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_0, \mathcal{D}_1) = 2 \sup_{A\in\mathcal{H}\Delta\mathcal{H}} |\mathcal{D}_0(A) - \mathcal{D}_1(A)|,$$

where $\mathcal{H}\Delta\mathcal{H}$ is the set of all symmetric differences between elements of $\mathcal{H}$ [13].

[16] considered a special case of domain generalisation in which the target distribution is a mixture distribution of the source domains:

$$\mathcal{D}_T = \sum_{k=1}^K \alpha_k \mathcal{D}_{S_k} \quad \text{where} \quad \sum_{k=1}^K \alpha_k = 1.$$

Under this assumption, at asymptotic condition, the upper bound of the generalisation loss $\epsilon_T(h)$ can be improved by reducing the pairwise distance of source domains [16]. Pairwise domain distance minimisation methods utilise GAN-based or MMD-based adversarial losses [16, 17, 18]. However, it is worth noting that if the target distribution is outside of the convex hull of the training domain, minimising the pairwise distance of source domains does not imply minimising the distance between source and target domains.

Recent work [11, 2] explored data augmentation mechanisms for domain generalisation. For instance, [11] designed a generative model to produce training domain samples at a distance upper bounded by a threshold ρ from the source domain. [2] generated domain samples using a composition of data augmentation functions. It is important to note that data augmentation alone may not enable domain generalisation because it can violate the conditions in [13]. In particular, it may generate training distributions that (i) increase the source to target $d_{\mathcal{H}\Delta\mathcal{H}}$ distance, (ii) violate the covariate-shift assumption, (iii) increase the error of the optimal hypothesis θ_λ, and/or (iv) increase the empirical loss during training $\sum_{k=1}^K \lambda_k \hat{\epsilon}_{S_k}$, all of which will result in an increase in the generalisation bound in (1). Augmentation schemes, similar to the ones proposed in [19, 20], are ”learned” based on performance on a validation set. While this process may help to improve generalisation performance on testing samples drawn from the same source distribution as the training and validation samples (i.i.d. setting), it exerts no control over the distance to target domains, where the i.i.d. assumption is violated.

III. Methodology

Both invariant feature representation learning and data augmentation suffer theoretical and practical challenges when applying to domain generalisation problems. However, the challenges for each method are different in nature:

1) Data augmentation [19] offers no theoretical guarantee but provides the learning model with access to new training data and training domains.
2) Invariant feature representation learning from training domains (Equation 1) offers a theoretical guarantee, but the number of target domains that can be covered is limited by the original training domains.

Hence, the combination of the two approaches may address the weakness of each method, which is demonstrated in Figure 3.

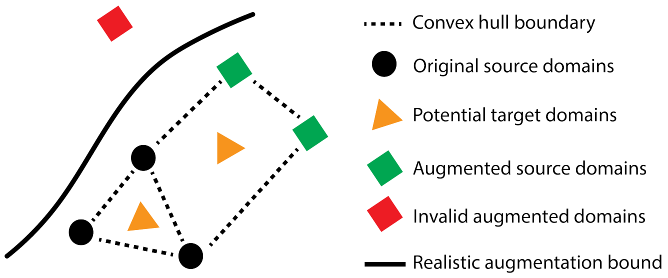

Fig. 3. An intuitive explanation of the motivation for the proposed method. Invariant feature representation alone does not cover target domains outside the original convex hull. Data augmentation alone does not guarantee that the augmented domain is close to the target domain. However, a method that combines both ideas may overcome both limitations. Figure adapted from [21].

More concretely, we include target distributions that are contained inside the convex hull whose vertices are defined by the k sources and $m-k$ augmented domains:

$$\mathcal{D}_T = \sum_{i=1}^{k} \beta_i \mathcal{D}_{S_i} + \sum_{j=k+1}^{m} \beta_j \mathcal{D}_{S'_j} \quad \text{, where } \sum_{i=1}^{m} \beta_i = 1.$$

Ideally, the larger the space covered by the convex hull, the greater the number of target domains that the method can generalise to. This can be achieved by applying (i) a stronger data augmentation transformation (distortion), or (ii) a higher number of successive transformation functions. However, as discussed above, excessive augmentation may result in unrealistic domains (see Figure 4), so it is important to use only augmentations that are realistic.

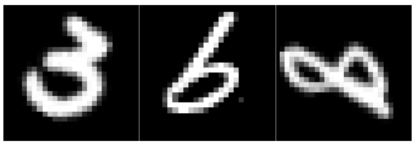

Fig. 4. Left: original label is ’3’, and after data augmentation (rotation 30 degrees), the label is still ’3’ (i.e., image appears to be sampled from the same distribution and label does not change). Middle: original label is ’9’, and after data augmentation (rotation 180 degrees), the label switches to ’6’ (i.e., image appears to be sampled from the same distribution, but label changes). Right: Original label is ’8’, and after data augmentation (rotation 90 degrees), the label is ∞, which is not contained in the sample label set (i.e., image appears to be sampled from a different distribution). Excessive augmentations (Figure 4 right) can create unrealistic images and introduce noise, when the annotation does not match the features.

Consider an augmented domain and its induced labelling function $(\tilde{\mathcal{D}}, g_{\tilde{\mathcal{D}}})$, constructed by applying augmentation functions on the images without changing the labels, we call an augmented domain realistic if the induced labelling function $g_{\tilde{\mathcal{D}}}$ agrees with the latent underlying labelling function g. More concretely, an augmented domain is realistic if the distance between the underlying and the induced labelling functions is close:

$$\mathbb{E}_Z[|g_{\tilde{\mathcal{D}}}(z) - g(z)|] < \epsilon \tag{2}$$

where $\epsilon \in (0,1)$ is a threshold value, and $g_{\tilde{\mathcal{D}}} : \mathcal{Z} \rightarrow \mathcal{Y}$ is a mapping function that labels instances in domain $\tilde{\mathcal{D}}$. The expectation is computed over the mixture distribution of Z, which include all the observable and unknown domains. Practically, this measure is difficult to compute, given that g is unknown, and only a limited number of source domains is available. However, we can relax the problem by approximating the true labelling function using a neural network to learn a hypothesis h on the original training domain. Assuming that h and g are defined on the same support set, we can bound the distance in equation 2 by requiring that the induced labelling function be close to the trained hypothesis, and that the trained hypothesis be close to the true labelling function:

$$\mathbb{E}_Z[|g_{\tilde{\mathcal{D}}}(z) - g(z)|] \le \mathbb{E}_Z[|g(z) - h(z)|] + \mathbb{E}_Z[|g_{\tilde{\mathcal{D}}}(z) - h(z)|] \tag{3}$$

Equation 3 can be relaxed further similar to [14] by taking the expectation over realisable sets - i.e. the first expectation on the RHS computed over the original domain and the second expectation on the RHS computed over the augmented domain. Therefore, we can heuristically select a realistic data augmentation by (1) training a classifier on the original domain until convergence, and (2) running the classifier on augmented data to compute the distance in equation 3 and rejecting augmentations who distance is above a certain threshold.

A. Data Augmentation

For simplicity, we consider a data augmentation scheme similar to [2, 19, 20, 12], consisting of a set of augmentation functions, including random cropping, random horizontal flip, random color jitter, random gray scale, random Gaussian blur, normalisation, random erasing, each having a set of parameters that control the level of distortion. To increase the diversity of augmentation, each magnitude parameter is uniformly sampled from a pre-defined range of values. Similarly, each augmentation is given an application probability. For example, one composite function to be drawn is random cropping with magnitude 0.8 followed by random horizontal flip, while another can be random cropping with magnitude 0.9. Given that there is a chance that none of the augmentation functions are triggered, the augmentation set can vary from an identity transformation to the maximum distortion in which all the augmentation functions with the strongest distortion magnitude parameters are applied. More precisely, the distance between the new and original domains vary from zero to the maximum distance defined by the particular augmentation scheme.

The parameters for the augmentation function is tuned as described in the previous section. The selected parameters are presented as follows:

```
color_jitter = transforms.ColorJitter(0.8, 0.8, 0.8, 0.8)
self.transform = transforms.Compose([
transforms.RandomResizedCrop(224, scale=(0.8, 1.0)),
transforms.RandomHorizontalFlip(),
transforms.RandomApply([color_jitter], p=0.8),
transforms.RandomGrayscale(p=0.2),
GaussianBlur(kernel_size=int(0.1 * 224)),
transforms.ToTensor(),
transforms.Normalize([0.485, 0.456, 0.406],
                     [0.229, 0.224, 0.225])])
```

B. Invariant Feature Learning

We minimise the distance between original and augmented domains using a supervised constrastive learning loss that combines contrastive and cross-entropy losses [12, 22]. More specifically, we treat pairs of samples from the same class as positive, and pairs of samples from different classes as negative. The contrastive loss minimises the distance between positive pairs and maximises the distance between negative pairs [23] based on class labels. Intuitively, the loss function reduces the distances between empirical distributions of the training domains for each class label to achieve invariant representation while increasing the distances between class clusters to achieve better classification. This effect is shown in Figures 5 and 6 in the ablation study.

The contrastive loss is defined by

$$L_{i,j}(\psi) = -\log \frac{\exp(\mathrm{sim}(f_\psi(x_i), f_\psi(x_j))/\tau)}{\sum_{k=1,k\neq i}^{N} \exp(\mathrm{sim}(f_\psi(x_i), f_\psi(x_k))/\tau)}, \tag{4}$$

where $f_\psi(.)$ is the feature representation network parameterised by ψ, τ is the distillation temperature [24], and $\mathrm{sim}(a, b)$ denotes the cosine similarity measure. For a batch size of N samples, the total contrastive loss is computed as follows:

$$L_{con}(\psi) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1,j\neq i}^{N} 1(y_i = y_j) L_{i,j}(\psi), \tag{5}$$

where $1(.,.)$ is the indicator function. This class of contrastive loss, referred to as normalised temperature-scaled cross entropy loss (NT-Xent), has been used in the more recent papers [25, 26, 22] and was demonstrated [12] to outperform other common contrastive loss functions. Given a classifier $h_\theta(.)$, parameterised by θ, the classification loss is defined by

$$L_{class}(\psi, \theta) = -\frac{1}{N} \sum_{i=1}^{N} \log(h_\theta(y_i | f_\psi(x_i))), \tag{6}$$

with $h_\theta(.)$ being an MLP parameterised by θ that maps $f_\psi(x_i)$ to an output that has the probability of each class in $\mathcal{Y}$, and the total loss being optimised is

$$L_{total}(\psi, \theta) = L_{class}(\psi, \theta) + \lambda L_{con}(\psi), \tag{7}$$

where λ is a penalty term that balances between the classification and contrastive losses. After every iteration t, the parameters of the feature extractor $f_\psi(.)$ and the classifier $h_\theta(.)$ are updated with:

$$\begin{aligned} \psi^{(t+1)} &= \psi^{(t)} - \alpha_\psi \nabla_\psi L_{total}(\psi^{(t)}, \theta^{(t)}), \\ \theta^{(t+1)} &= \theta^{(t)} - \alpha_\theta \nabla_\theta L_{class}(\psi^{(t)}, \theta^{(t)}), \end{aligned} \tag{8}$$

where $\alpha_\psi, \alpha_\theta$ are the learning rates and ∇ is the gradient operator. Note that this method requires the label distribution to be roughly equal across the domains (balanced dataset).

This method has a perceived advantage of not having to rely on the original domain split (i.e., domain labels) for training - that is, it is domain agnostic. Given that for most datasets, domain labels are assigned arbitrarily based on the acquisition sources, using the original domain split may not be optimal. Moreover, we avoid the costly process outlined in MMLD [17], which automatically assigns new labels based on image style by clustering. Additionally, unlike [12, 22] we do not use any projection layer, as our experiments show that the performance improves without it – we show this result in the experiment section below.

IV. Experiments

A. Datasets

We evaluate our method on a standard benchmark for domain generalisation, and a new medical domain generalisation benchmark that we propose, consisting of the chest x-ray datasets. The dataset PACS [27] contains four domains (Photos, Arts, Cartoon, Sketch) with seven shared classes, and is considered one of the most challenging benchmarks [27] given its severe domain shift and sparse data, particularly for the more difficult domain (Sketch). Our proposed medical benchmark dataset benchmark is built from three publicly available datasets, namely Chexpert [28], Chest14 [29], and Padchest [30]. Given that diseases diagnosis may utilise secondary information, and that the general classification task for these datasets is multi-label, the solution to which is beyond the scope of this paper, we focus solely on the Cardiomegaly binary classification problem, which can be detected based on radiographs alone. Since Cardiomegaly is typically diagnosed from anteroposterior/ posteroanterior or AP/PA view images, we sample 4000 AP/AP view images from each of the three datasets. For each sampled domain, the label distributions are similar, with roughly one-in-ten samples having the disease.

B. Implementation Details

To ensure evaluation consistency for the PACS benchmark, we adopt the same training setup and hyper-parameter values described in previous papers [9, 17]. This includes withholding one domain for testing and training the network with the remaining domains. During training, the aggregated training domains (three remaining domains) are split 0.9/0.1 to form the training and validation datasets. The best performing model is selected based on the accuracy on the validation dataset. The test results on the withheld datasets (generalisation result) of the selected models are recorded. This is to simulate real-life scenarios, in which we have no access to the latent target domain, and have to rely on the validation set as the best alternative.

Similarly to [9, 17], we train the model with the same stochastic gradient descent optimizer using a momentum of 0.9, weight decay of 5e-4, mini-batch size of 128, and Nesterov acceleration, for 30 epochs. We also adopt the same initial learning rate of 1e-3 [17], with a step scheduler that reduces the initial learning rate by a factor of 10 after 24 epochs. We use the augmentation function defined in the previous section. The contrastive penalty weight λ in Equation 7 is set at 0.02. This weight is chosen to keep the overall training loss (cross entropy + contrastive) comparable in magnitude to the cross entropy loss alone to ensure the empirical loss in Equation 1 does not change much when the regularisation term is introduced. The temperature parameter τ in Equation 4 is set to 0.07 [12]. Following the evaluation framework in [31, 10, 17], we demonstrate the performance and scalability of our method using Caffe Alexnet and Resnet18 models on PACS. The ERM and our proposed DASCL are trained using the same network, same training set size, following the same training scheme and using the same parameters, but different augmentation functions and without the contrastive loss term. ERM uses the common, more conservative augmentation found in previous papers [17]. The set up for the proposed medical benchmark dataset benchmark is the same as for PACS, with a similar model, augmentation, and training parameters. Due to class imbalance issues, the evaluation criteria for the medical benchmark dataset benchmark is based on the area under the receiver operating characteristic curve (AUC).

C. Benchmark Results

The evaluation results of our method DASCL are shown in Tables I, II, and III for Alexnet PACS, Resnet18 PACS, and Alexnet Medical respectively. The results consist of the mean and standard deviation classification accuracy (for PACS) and AUC (for medical) over at least five runs for each domain, while the DASCL Alexnet results for PACS are acquired over fifteen runs. For comparison, we show the published results of the most recent domain generalisation techniques. The baseline ERM is represented by Alexnet PACS and Resnet18 PACS models with pre-trained weights. Epi-FCR [10] extends ERM with an episodic training scheme that fine-tunes the feature extractors and classifiers on single domains sequentially, encouraging robust performance on novel domains. Jigen [9] regularizes the classification loss with a jigsaw loss calculated on images with shuffled parts to encourage learning of spatial relationships. MMLD [17] is an adversarial-learning approach that assumes the existence of latent domains and uses a clustering technique to assign latent domain labels based on style, before passing downstream to a domain discriminator-classifier pair that enforces domain-invariant representation. MASF [31] explores the meta-learning framework with a global loss that preserves inter-domain class-concept relationship and a local metric-learning loss that clusters embedding features based on class labels.

Overall, our method outperforms the current state of the art (sota) in all benchmarks. For the medical benchmark dataset results shown in Table III, both DASCL and MMLD (the method most similar to ours in training scheme) show an improvement over the baseline approach. Additionally, DASCL consistently outperforms MMLD across all domains.

TABLE I
Mean and standard deviation of the classification accuracy results (in %) on PACS using Alexnet. Targets are domains withheld from training, sources are the non-withheld domains used for training. The results of other methods are acquired from their papers. The best results are highlighted in bold (up to the decimal value).

Target	P	A	C	S	Average
Epi-FCR	86.1	64.7	**72.3**	65.0	72.03
Jigen	89.00	67.63	71.71	65.18	73.38
MMLD	88.98	69.27	**72.83**	66.44	74.38
MASF	**90.68**	70.35	**72.46**	67.33	75.21
ERM	88.89	68.14	70.19	61.07	72.06
DASCL	89.80	**71.71**	71.55	**72.77**	**76.41**
Std.dev	0.81	1.17	1.14	0.98	0.39

TABLE II
Mean and standard deviation of the classification accuracy results (in %) on PACS using ResNet18. The hyper-parameters for Resnet18 DASCL are the same as those in the Alexnet counterpart (optimiser, scheduling, contrastive weight, etc). The best results are highlighted in bold (up to the decimal value).

Target	P	A	C	S	Average
Epi-FCR	93.9	82.1	77.0	73.0	81.5
Jigen	96.03	79.41	75.25	71.35	80.51
MMLD	96.09	81.28	77.16	72.29	81.83
MASF	94.99	80.29	77.17	71.69	81.04
ERM	**96.32**	77.95	74.88	66.81	78.99
DASCL	96.25	**82.26**	**78.16**	**76.60**	**83.32**
Std.dev	0.13	1.10	0.69	0.41	0.23

D. Ablation Study

We conduct an ablation study on the relative contribution of each component on the PACS and the medical benchmark datasets in Tables III and IV. It is worth noting that in spite of its simplicity, the chosen augmentation performs well for the PACS dataset, bringing the highest contribution to the overall improvement (Table IV). On the other hand, for the medical benchmark dataset, using data augmentation alone produces just a small AUC improvement (Table III). This

TABLE III
MEAN AND STANDARD DEVIATION OF THE CLASSIFICATION AUC RESULTS (IN %) ON THE MEDICAL BENCHMARK DATASET USING ALEXNET. ALL PARAMETERS FOR ERM AND DASCL ARE SIMILAR TO THOSE IN THE EXPERIMENT WITH PACS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD (UP TO THE DECIMAL VALUE).

Target	Chexpert	Chest14	Padchest	Average
ERM	76.53	85.88	83.13	81.85
Std.dev	1.22	2.28	1.35	0.93
MMLD	75.40	87.70	84.99	82.70
Std.dev	2.30	1.53	0.29	1.14
ERM+Aug	75.65	87.24	81.70	81.92
Std.dev	0.29	0.77	2.38	0.88
ERM+CL	77.34	87.92	85.45	83.57
Std.dev	0.55	0.93	1.16	0.21
DASCL	**77.54**	**88.83**	**87.31**	**84.55**
Std.dev	1.15	0.92	1.21	0.62

TABLE IV
CLASSIFICATION ACCURACY RESULTS (%) FROM EACH COMPONENT OF DASCL TRAINED USING CAFFE ALEXNET ON PACS DATASET.

Target	P	A	C	S	Average
ERM	88.89	68.14	70.19	61.07	72.06
ERM+Aug	89.40	70.21	70.49	71.11	75.41
ERM+CL	89.22	70.16	70.86	64.25	73.63
DASCL	89.80	71.71	71.55	72.77	76.41

is because some transformations (e.g., excessive cropping or flipping) do not produce realistic results. Despite this issue, when combined with the contrastive loss, the result with augmentation (DASCL) is better than without augmentation (ERM+CL) by a large amount.

We further assess on the medical dataset the performance of the method when the augmentation strength varies from the weakest - W, to the strongest - S. For the medical dataset, the weakest augmentation W has a disagreement between the original and induced labels within $\pm 1\%$ (Equation 3), the tuned augmentation produces a disagreement result within $\pm 5\%$, and augmentation S produces a disagreement result more than $\pm 10\%$ from the baseline. The results when DASCL is applied for different augmentation setting is shown in Table V. The ERM result for augmentation S shows that unrealistic augmentation produces poorer predictions on all domains. However, the full method still produces AUC results better than the baseline (ERM with standard augmentation) even when unrealistic augmentation is applied, demonstrating the robustness of this method. It is worth noting that the tuned augmentation is by no means optimal, and a more diverse and realistic set of augmentations may produce better results.

Given that minimising the contrastive loss reduces the distances between the empirical distributions of the training domains, as explained in Section III-B, we address the question of whether DASCL reduces the $\mathcal{A}$ distance between domains. Computing the $\mathcal{A}$ distance between two domains requires training a binary classifier on the task of discriminating between samples from the two domains. This is an approximation for the actual $\mathcal{A}$ distance since finding the optimal classifier is a non-trivial task. We train a linear binary classifier with the extracted features, and evaluate our approach (DASCL) against ERM, since ERM does not explicitly regularise the $\mathcal{A}$ distance.

TABLE V
MEAN AND STANDARD DEVIATION OF THE CLASSIFICATION AUC RESULTS (IN %) ON THE MEDICAL BENCHMARK DATASET USING ALEXNET. ALL PARAMETERS FOR ERM AND DASCL ARE SIMILAR TO THOSE IN THE EXPERIMENT WITH PACS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD (UP TO THE DECIMAL VALUE).

Target	Chexpert	Chest14	Padchest	Average
ERM (Augmentation W)	76.50	86.67	85.17	82.7
Std.dev	1.00	0.77	1.167	0.80
DASCL (Augmentation W)	76.78	87.93	87.01	83.90
Std.dev	0.62	0.89	1.03	0.30
ERM (Tuned Augmentation)	75.65	87.24	81.70	81.92
Std.dev	0.29	0.77	2.38	0.88
DASCL (Tuned Augmentation)	**77.54**	**88.83**	**87.31**	**84.55**
Std.dev	1.15	0.92	1.21	0.62
ERM (Augmentation S)	74.26	83.62	85.52	81.39
Std.dev	0.50	1.30	0.7	0.60
DASCL (Augmentation S)	75.76	86.52	86.18	82.81
Std.dev	0.41	0.53	0.62	0.40

The result shown in Figure 6 illustrates that our approach does indeed reduce the pairwise domain distance using the PACS dataset. This is further demonstrated in Figure 5 by using the t-SNE visualisations of the ERM and DASCL methods on the PACS dataset using the feature $f_\psi(x)$.

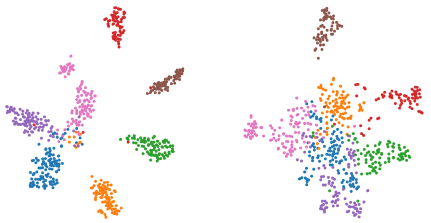

Fig. 5. Left: t-SNE plot of DASCL model, Right: t-SNE plot of ERM model. Both are trained on the PACS dataset using the feature $f_\psi(x)$. Different colours represent different classes. The contrastive approach clusters training samples based on class labels irrespective of domains to reduce distances between domains. The class-clusters are better separated to allow for better classification results.

We also test if an additional projection layer outlined in [12] is beneficial. We conducted an experiment with several projector dimensions as shown in Figure 7. The results show that a better result can be obtained without the projection layer. This contradicts the observation in [12], and a possible explanation can be due to the size of the dataset. This question remains to be addressed in future work.

V. Discussion and Conclusion

The contributions of our paper include: (1) a new framework that combines domain augmentation and invariant feature learning to solve domain generalisation problems, (2) a simple learning algorithm to determine realistic domain augmentations and combine the augmentations with contrastive learning, and (3) a demonstration of the effectiveness with empirical results and ablation studies. The main limitation with our approach is the heuristic process used to define the augmentation functions, which stems from (i) the lack of a theoretical framework for defining covariate shift measures, and (ii) the

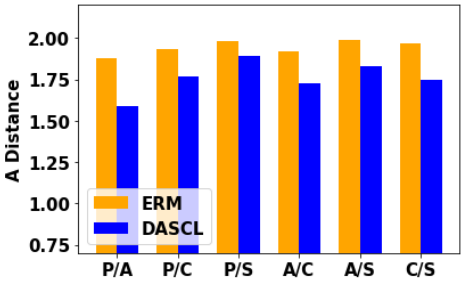

Fig. 6. Empirical $\mathcal{A}$ distance between each possible pair of domains from PACS, under the feature representations of ERM (orange) and DASCL (blue). A lower value implies a lower $\mathcal{A}$ distance between two domains – closer to achieving invariant feature representation. The results are obtained by training a linear classifier with L1 loss on the task of discriminating between domains [14], with features acquired by passing each input image x through the feature representation $f_{\psi}(x)$ of the respective networks (ERM vs DASCL). The feature extractors are acquired from the trained model in Table I, with the target domain being Photo.

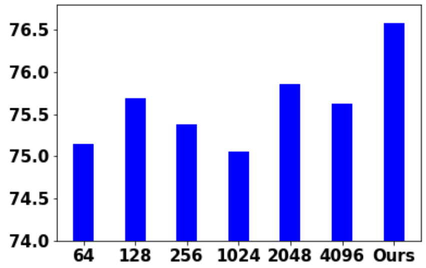

Fig. 7. Five-run-average test accuracy on the PACS dataset for Alexnet with different projector dimensions (X-axis). Ours uses no projector.

lack of a practical augmentation method tailored for non-i.i.d settings. This topic will be explored in future work.

Acknowledgement

This research is supported by Australian Research Council through grants DP180103232 and FT190100525.

References

[1] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. The MIT Press, 2009.

[2] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, H. Roth, A. Myronenko, D. Xu, and Z. Xu, "When unseen domain generalization is unnecessary? rethinking data augmentation," *arXiv preprint arXiv:1906.03347*, 2019.

[3] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014.

[4] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *International Conference on Machine Learning*, 2013, pp. 10–18.

[5] 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*, 2017, pp. 7167–7176.

[6] M. M. Rahman, C. Fookes, M. Baktashmotlagh, and S. Sridharan, "Correlation-aware adversarial domain adaptation and generalization," *Pattern Recognition*, vol. 100, p. 107124, 2020.

[7] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," *arXiv preprint arXiv:1702.05374*, 2017.

[8] R. Shu, H. H. Bui, H. Narui, and S. Ermon, "A dirt-t approach to unsupervised domain adaptation," *arXiv preprint arXiv:1802.08735*, 2018.

[9] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2229–2238.

[10] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, "Episodic training for domain generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1446–1455.

[11] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," in *Advances in Neural Information Processing Systems*, 2018, pp. 5334–5344.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[13] S. B. David, T. Lu, T. Luu, and D. Pál, "Impossibility theorems for domain adaptation," in *AISTAT 2010*, 2010, pp. 129–136.

[14] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira *et al.*, "Analysis of representations for domain adaptation," *Advances in neural information processing systems*, vol. 19, p. 137, 2007.

[15] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," in *Advances in neural information processing systems*, 2018, pp. 8559–8570.

[16] I. Albuquerque, J. Monteiro, M. Darvishi, T. H. Falk, and I. Mitliagkas, "Generalizing to unseen domains via distribution matching," 2019.

[17] T. Matsuura and T. Harada, "Domain generalization using a mixture of multiple latent domains," *arXiv preprint arXiv:1911.07661*, 2019.

[18] 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*, 2018, pp. 5400–5409.

[19] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *CVPR 2019*, 2019, pp. 113–123.

[20] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Ran-

daugment: Practical data augmentation with no separate search," *arXiv preprint arXiv:1909.13719*, 2019.

[21] H. S. Le, R. Akmeliawati, and G. Carneiro, "Domain generalisation with domain augmented supervised contrastive learning," *Association for the Advancement of Artificial Intelligence – Student Abstract*, 2020.

[22] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *arXiv preprint arXiv:2004.11362*, 2020.

[23] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[24] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[25] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in neural information processing systems*, 2016, pp. 1857–1865.

[26] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742.

[27] 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*, 2017, pp. 5542–5550.

[28] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *AAAI 2019*, vol. 33, 2019, pp. 590–597.

[29] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *CVPR 2017*, 2017, pp. 3462–3471.

[30] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "Padchest: A large chest x-ray image dataset with multi-label annotated reports," *Medical Image Analysis*, p. 101797, 2020.

[31] Q. Dou, D. C. de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," in *NIPS 2019*, 2019, pp. 6447–6458.