Domain Generalization by Dropping Spurious Information Out

Anonymous CVPR 2021 submission

Paper ID ****

Abstract

Generalization capacity to unseen domains is an essential issue to deploy deep learning algorithms in real-world applications. Domain-invariant representation is a widely used strategy that performs well on out-of-distribution. However, recent studies point out a fundamental tradeoff between distribution alignment and target error minimization from an information-theoretic perspective. To solve this problem, we introduce a mutual information maximization module and explicitly drop superfluous information that is not shared across multiple domains to prevent models from relying on spurious correlations. We further boost the performance by using the class prior-normalized value and self-distillation. Our method can be viewed as an extension to contrast learning in domain generalization, which focuses on the estimated mutual information between the learned representations of images from the same category among multiple domains rather than the multi-view of the same image. We demonstrate the effectiveness of methods on two common domain generalization benchmarks and evaluate our method thoroughly from both theoretical and empirical perspectives.

1. Introduction

Deep learning methods have achieved remarkable success in computer vision tasks. However, the domain shift problem is still challenging since it violates the assumption that training and testing data are independent and identically distributed (i.i.d), which leads to a non-neglectable drop in the performance of a trained model. As the shift in data statistics exists extensively in real-world applications, e.g., self-driving and medical imaging, the research community has proposed a line of works to solve the problems, including multi-domain learning [7, 36], domain adaptation [38, 55, 11], and domain generalization [34, 16].

To deal with the distribution shift problem, domain adversarial training strategy is extensively used for domain-invariant representations [23, 33, 38, 55, 39, 20] and has solid theoretical foundations [6, 55, 24]. Another line of methods explicitly matches feature distributions under different metrics, including the mean and covariance [47], maximum mean discrepancy [32], and Wasserstein distance [58]. However, all of these methods may induce a nontrivial lower bound of the error in the target domain when the marginal label distributions differ between source and target domains [20, 11]. Several works attempt to solve this problem, including estimating importance weights and aligning reweighted feature distributions [35, 3, 11], changing sampling strategy [28], and invariant risk minimization [2].

In this work, we propose a new perspective to learn robust correlations among different domains to promote the generalization ability of models. Based on contrast learning, we explicitly maximize the mutual information between the representation of images from the same category across domains $MI(z_i, z_j|y_i = y_j)$. In this way, we enforce the feature extractor to preserve robust label information across multiple domains and minimizing the superfluous information related to the domain label. Compared with mutual information maximization for unsupervised representation learning [42, 27, 25, 9], our method does not focus on the estimated mutual information between learned representations of the multi-view of the same image. That step is an important extension for contrast learning. To summarize, the contributions of this paper are as follows:

- We propose a mutual information maximization module to explicitly drop superfluous information related to the domain label. This approach promotes the generalization ability of the model to out-of-distribution and avoids the tradeoff between distribution alignment and target error minimization from a new perspective.
- We conduct extensive experiments on domain generalization benchmarks. Compared with state-of-art methods, our method achieves strong performance on all tasks.
- We thoroughly review our methods from a theoretical and empirical perspective, clearly demonstrating the connections and advantages with domain adversarial training and triplet loss.

2. Related work

In this section, we provide a literature review on domain generalization and mutual information estimation approaches.

2.1. Domain generalization

Most existing approaches dealing with domain generalization can be mainly divided into two categories: learning domain-invariant representations and using an episodic training paradigm to simulate the unseen domain.

Domain-invariant representation methods These methods are mainly the extension of domain adaptation. In domain generalization setting, these methods achieve domain-invariant representations within the multiple source domains [32, 33, 40], rather than between source and target domains [37, 23, 38, 33, 39]. The domain-invariant representations can also be achieved by disentangling representation into the domain-specific feature and domain-shared feature [43] and synthesizing data from pseudo-novel domains to augment the source domains [45, 59].

Episodic training paradigm methods Inspired by meta-learning, these methods use episodic training to simulate domain shift. [30] adopt a similar update rule as MAML [21]; [31] use multiple feature extractors and classifiers and train them alternatively to learn robust components; [4] learn a meta regularizer for the classification layer while [34] learn a meta regularizer for the feature extractor; [15] proposes global class alignment and local sample clustering on feature space.

Recently, there are several emerging directions for domain generalization, including through self-supervision and variational information bottleneck principle. Self-supervision pretext helps the model to learn distinctive embeddings between every image in the dataset to avoid supervision collapse, that the model only represents class information and lose the information might be useful to transfer [14, 8]. The variational information bottleneck principle can deviate to a regularization term, the Kullback–Leibler (KL) divergence between distributions of latent encoding of the samples from the same category in multiple source domains [16]. Our approach falls into the domain-invariant representation category while replacing the domain adversarial training and considering the marginal label distribution.

2.2. Mutual information estimation

Recently, there have been many promising results achieved by maximizing mutual information for unsupervised representation learning [42, 25, 9]. InfoMax principle and the properties of mutual information have been well understood for a long time. The main breakthrough of that line of work is that they find a tractable lower bound of mutual information and use a neural network to estimate it since mutual information is notoriously difficult to calculate. For example, MINE [5] uses the Donsker-Vardhan representation for KL divergence and samples for the joint distribution and marginal distribution to unbiasedly estimate the mutual information. InforNCE [42] is defined as the expectation sampled from the joint distribution and used to maximize the mutual information between the context and the prediction. DeepInfoMax [27] defines a Jensen-Shannon estimator to maximize the global features and local features in one image. JS estimator is insensitive to the number of negative samples compared with the aforementioned method.

In this paper, we maximize the estimated mutual information in the supervised learning paradigm, between the images of the same category from the mixture of multiple source domains, which is different from the multi-view of the same image used in the unsupervised paradigm and the work for enhancing discriminability of domain-specific information [20].

3. Proposed Method

Here, we present details of our approaches. Section 3.1 illustrates the tradeoff between distribution alignment and target error minimization from an information-theoretic perspective. Section 3.2 demonstrates the motivation and framework of instance-based mutual information maximization. We describe the overview of our method in Section 3.3. Further, we demonstrate mutual information loss in Section 3.4, and class prior-normalized value in Section 3.5. Finally, in Section 3.6, we introduce how knowledge distillation helps domain generalization.

3.1. Preliminaries

Generalization bound for the unseen domains can be viewed as an extension to the well-studied generalization bound for domain adaptation. In the seminal work [6], the H-divergence was proposed to measure the discrepancy between source and target domain. That leads to the generalization bound:

Let H is a hypothesis space of Vapnik–Chervonenkis (VC) dimension d, $\hat{D}_S, \hat{D}_T$ are samples of size m from source and target domains. For any $\delta \in (0, 1)$, with probability at least $1 - \delta, \forall h \in H$:

$$\begin{aligned} \epsilon_T(h) \leq & \epsilon_S(h) + \frac{1}{2} d_{H \Delta H}(\hat{D}_S, \hat{D}_T) + \lambda \\ & + 4\sqrt{\frac{2d \log(2m) + \log(\frac{2}{\delta})}{m}}. \end{aligned} \tag{1}$$

A recent work [1] extends this bound for unseen domains under the assumption that the distributions in multiple domains are in the convex hull. They devise the domain adversarial training to minimize pair-wise domain divergences in multiple source domains. However, domain-invariant learning suffers from a theoretical challenge when the marginal label distributions differ between source and target domains. [56] suggests λ is not negligible. The upper bound is also correlated with the distance between the labeling functions from the source and target domains. Besides, the lower bound was firstly proposed in [56], and further extend to k-class classification and conditional adversarial training by [11]. The low bound is as follow:

Let D_{JS} denote the Jensen-Shannon divergence between two distributions. Suppose Markov chain: $X \xrightarrow{g} Z \xrightarrow{h} Y$ holds and $D_{JS}(D_S^Y \parallel D_T^Y) \geq D_{JS}(D_S^Z \parallel D_T^Z)$, then:

$$\begin{aligned}\epsilon_T(h) + \epsilon_S(h) \geq \frac{1}{2}(&\sqrt{D_{JS}(D_S^Y \parallel D_T^Y)} \\ &- \sqrt{D_{JS}(D_S^Z \parallel D_T^Z)})^2.\end{aligned} \tag{2}$$

The insight from the lower bound demonstrates when the marginal label distributions differ between source and target domains, achieving domain-invariant representations, and minimizing the empirical risk can hinder the algorithm from successfully transfer across domains. Rather than explicitly align the distribution among different domains, We aim to learn the robust correlations among multi-domains by explicitly dropping superfluous information that is related to the domain label.

3.2. Mutual information maximization for dropping superfluous information

Assume I_{d1} and I_{d2} to be two images from the same category among different domains, we aim to force the feature extractor to encode images I_{d1} and I_{d2} to feature Z_{d1} and Z_{d2} containing robust and necessary label information while dropping all the superfluous information of the domain label. We can formulate the objective as:

$$MI(I_{d1}, Z_{d1}|I_{d2}) = 0 \Rightarrow MI(I_{d1}, Z_{d1}) = MI(I_{d2}, Z_{d1}). \tag{3}$$

To better understand the objective, We can use the chain rules of mutual information to subdivide $MI(I_{d1}, Z_{d1})$ into two components:

$$MI(I_{d1}, Z_{d1}) = \underbrace{MI(I_{d1}, Z_{d1}|I_{d2})}_{superfluous information} + \underbrace{MI(I_{d2}, Z_{d1})}_{predictive information} \tag{4}$$

I_{d1} contains more information related to domain $d1$ than I_{d2} and vice versa. The optimal feature extractor aims to minimize that information related to domain label and maximize the mutual information between I_{d2} and Z_{d1} to capture the robust correlations related to the label. [18] proposed $MI(I_{d2}, Z_{d1})$ is the upper bound of $MI(z_{d1}, Z_{d2})$. So our final objective to maximize the mutual information between the learned representations of images from the same category among multiple domains.

3.3. Method overview

We introduce the proposed method under the scenario that domain generalization using a mixture of k source domains $\{(x_n^i, y_n^i)\}_{n=1}^K$. Since domain labels are unknown, the dataset is $D = \{(x^i, y^i)\}$. We split the model as three parts: a feature extractor $f_\phi : X \rightarrow Z$, a classifier $g_\theta : Z \rightarrow \mathbb{R}^c$ and a mutual information estimator $h_\omega : MI = h_\omega(x, y)$. Algorithm 1 provides a summary of our method.

Algorithm 1 Maximizing sample-based mutual information with the class prior-normalized value

Require: mixture of multiple source domains D.
Require: feature extractor ϕ, classifier θ,
Require: embedding network φ, MI estimator ω.
Require: class prior-normalized value α.
Require: hyperparameter β, η.
  Randomly split D into disjoint D_{trn} and D_{val}
  for $k = 1$ to number of iterations **do**
    Sample mini-batch d_{trn} from D_{trn}
    $z_{trn} = f_\phi(d_{trn})$
    Compute mutual information loss:
    $L_{MI} = \alpha \cdot h_\omega(g_\varphi(z_{trn}))$ // Section 3.4,3.5
    Compute cross-entropy loss:
    $L_{task} \leftarrow \sum_{d_{trn}} l^{(CE)}(g_\theta(z_{trn}), y_{trn})$
    update feature extractor, classifier
    $(\phi, \theta) \leftarrow (\phi, \theta) - \eta(\nabla_{\phi,\theta}(L_{task} - \beta \cdot L_{MI}))$
    update embedding network, MI estimator
    $(\varphi, \omega) \leftarrow (\varphi, \omega) - \eta(\nabla_{\varphi,\omega}\beta \cdot -L_{MI})$
  end for

3.4. Maximizing the sample-based mutual information

We explicitly maximize the mutual information of the representation belong to the same category from multiple source domains, i.e. $I(z|y_j, z|y = y_j)$. Adding the objective to the loss function forces the feature extractor to capture the shared and robust label information and reduce the sensitivity caused by domain difference. We adopt the lower bound of mutual information in JSD objective because that objective is insensitive to negative sample strategies [27]:

$$MI(x, y) \geq \mathbb{E}_P[-sp(h_\omega(x, y))] - \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[sp(h_\omega(x', y))], \tag{5}$$

where x is an input sample, x' is an input sampled from $\mathbb{P} = P$, and $sp(z) = \log(1 + e^z)$ is the softplus function.

Recent work [15] and our preliminary experiments reveal applying regularization onto feature Z may too heavy for feature extractor. So we apply the mutual information maximization on the low dimensional embedding e of feature Z through the embedding network.

Firstly, inputs from a mini-batch d_{trn} was encoded by a feature extractor f_ϕ to feature z_{trn}, and further encoded by the first two layer of classifier $g_{\theta'}$ to embedding e_{trn}. Secondly, we form positive pairs $(e_i, e_j | y_i = y_j)$ by independently sampling two images belong to the same category from mini-batch d_{trn} and we sample negative instances belong to other categories from mini-batch d_{trn} to form negative pairs$(e_i, e'_j | y_i \neq y'_j)$.

3.5. Class prior-normalized value

In addition to aligning the class conditional distributions in the embedding space, we further take the marginal label distribution into consideration. In the domain generalization setting, since the data from the target domain is unavailable, we cannot estimate the importance weight $\frac{D_T}{D_S}$ between target and source domains for reweighting source feature distribution [11]. However, we can explicitly calculate class prior-normalized value α_j and apply to the learning process. The joint distribution of $P_D(f_\phi(x) \mid Y) = \mathbb{E}_j[P(f_\phi(x) \mid Y_j)P(Y = j)]$. When $P(Y = j)\alpha_j = \frac{1}{c}$, c is the number of categories, the joint distribution from multiple source domains are well aligned after we align the class conditional distributions in the embedding space. That helps the feature extractor to learn unbiased representation in a class balanced setting. α_j can be obtained as

$$\alpha_j = \frac{1}{L \cdot p(Y = j)} = \frac{N}{c \cdot N_j}, \tag{6}$$

where N denotes the total number of data in D, c denotes the number of category, and N_j denotes the number of data of categories j in D.

3.6. Knowledge distillation

Knowledge distillation [26] is an approach to transfer knowledge embedded in the teacher model or class relationships between different domains. Recent works [50, 15] have demonstrated aligning class relations between different domains can promote model generalization. However, when the multiple-source domain data are mixed, we cannot explicitly transfer the class relationships across domains. Therefore, we adopt sequential self-distillation proposed in [22], the knowledge distillation loss is as follows:

$$L_{kd} = KL(s_{\theta_k,\phi_k}, s_{\theta_{k-1},\phi_{k-1}}), \tag{7}$$

$$s_{\theta_k,\phi_k} = softmax(g_{\theta_k}(f_{\phi_k}(D_{trn}))/\tau), \tag{8}$$

where k is the number we operate self-distillation and s_{θ_i,ϕ_i} is the soft label distributions softmax at temperature $\tau > 0$. At each step, the new generation θ_k, ϕ_k is trained to minimize an auxiliary knowledge distillation loss that is the KL divergence between predictions and soft label predicted by θ_{k-1}, ϕ_{k-1}. We analyze the effectiveness of self-distillation in Section 4.5.

4. Experiments

4.1. Datasets

We evaluate our approach on two datasets for domain generalization. PACS [29] includes four domain data(Photo, Art paintings, Cartoon and Sketches). It includes 9991 images of size 224×224 from 7 categories. VLCS [48] covers 5 shared object categories from PASCAL VOC 2007 [17], LabelMe [44], Caltech101[19] and Sun09 [10].

Following the same experimental protocol in [8, 40], we use three domains as the source domain and the remaining one as the test domain each time. And in testing, we use the accuracy of the validation set (10% in the case of PACS, 30% in case of VLCS from source domain) as the model selection methods.

4.2. Implementation details

We use Alexnet and ResNet-50 pre-trained on ImageNet by removing the last layer as the feature extractor ϕ. As the embedding network, we adopt two fully connected layers ($1024 \to 256$), the same architecture as [15]. As the classifier, we initiate three fully connected layers ($1024 \to 256 \to c$), which shares the parameters in the first two layers with the embedding network since they all encode feature to low dimensional vector and require similar computation. For the MI estimator, we initiate two fully connected layers ($512 \to 1$). We use the same hyper-parameters employed by [8]. That is, we train the model for 30 epochs using Stochastic gradient descent (SGD) optimizer with a momentum = 0.9, a weight decay = 5e-4, and a batch size = 128; the learning rate is initiated as 1e-3 and scale it by a factor of 0.1 after 80% of the training epochs; using random crop, color jittering, random horizontal flip, and normalization as the pre-processing. We further distill our model with temperature $\tau = 4$ and coefficient 0.5 to rescale knowledge distillation loss. We distill the model three times, 10 epochs each time, and use the accuracy on the validation set for model selection.

4.3. Baselines

We compare the performance of our method with the following domain generalization methods. TF [29] proposes a low-rank parameterized neural network. CIDDG [33] aligns the joint distribution in the representation layer by using discriminators for each class and class prior-normalized value. MLDG [3] adopts an episodic training paradigm to simulate domain shift. CCSA [41] proposes to address the semantic distribution alignments for domain adaptation and generalization. MMD-AAE [32] jointly optimizes a multi-domain autoencoder, a discriminator, and a classifier with adversarial learning. SLRC [12] uses a structured lowrank constraint to align domain-specific networks and the domain-invariant one. D-SAM [13] proposes a domain-specific aggregation module to merge generic and specific information in multiple source domains. JiGen [8] combines the self-supervision task, Jigsaw puzzle, to improve the discriminability of the model and perform well on domain generalization. MetaReg [4] generates domain-guided perturbation of input instances. MMLD [40] learns to generate pseudo domain labels for adversarial training and achieve better results without using domain labels. MASF [15] proposes two regularizations on semantic feature space. MetaVIB [16] extends Information Bottleneck to an episodic training paradigm for domain generalization. Since the experimental protocol used in the reported results is different, we report Deep All for a fair comparison. Deep All is the result of training a pre-trained alexnet training by minimizing the cross-entropy loss of all source domains.

4.4. Results

Table 1 and Table 2 summarize the results on PACS and VLCS datasets. The results of our methods are average over three repetitions of each run. For all datasets, our methods achieve results that surpass all of the existing methods that do not use domain labels.

In the PACS dataset, our method shows a significant advantage over Deep All baseline, which proves that maximizing mutual information of the images from the same categories among the mixture of multiple source domains is effective for domain generalization. It is worthwhile to notice our methods achieve great results when the architecture goes deep.

4.5. Ablation analysis

We conduct an extensive study to investigate two key points: 1)the contribution of each component to the performance of our method, 2) how the class prior-normalized value boosts performance of domain generalization under mismatched label distributions. Firstly, we test all combinations of the key components, including mutual information maximization module, class prior-normalized value, and sequential distillation. It should be noticed that the class prior-normalized value is a weight to rescale class conditional alignment, it can only be used together with mutual information maximization module. From table 3 and table 4, we can see the performance gain consistently in all datasets.

Before finding out the benefit brought by class prior-normalized value, we first analyze the mismatch label distributions in the domain generalization setting. It consists two aspect: 1)the mismatched label distributions between multiple source domains and unseen domain, 2)the mismatched label distributions among multiple source domains. We apply symmetric critic, JS divergence, to measure the two discrepancy, which is demonstrated in Table 5. When applying class prior-normalized value in PACS, the performance improves significantly when the unseen domain is Sketch, which suffers more serious than other unseen domains. We further demonstrate the confusion matrix in Figure 1. It can be noticed that class prior-normalized value promote the performance especially on imbalanced classes, e.g., house, person, and dog. Similar results is also shown in VLCS,

	Art.	Cartoon	Sketch	Photo	Avg.
Deep All	63.30	63.13	54.07	87.7	67.05
TF*	62.86	66.97	57.51	89.5	69.21
Deep All	57.55	67.04	58.52	77.98	65.27
CIDDG*	62.70	69.73	64.45	78.65	68.88
Deep All	64.91	64.28	53.08	86.67	67.24
MLDG*	66.23	66.88	58.96	88.00	70.01
Deep All	64.44	72.07	58.07	87.50	70.52
D-SAM*	63.87	70.70	64.66	85.55	71.20
Deep All	66.68	69.41	60.02	89.98	71.52
JiGen	67.63	71.71	65.18	89.00	73.38
Deep All	67.21	66.12	55.32	88.47	69.28
MetaReg*	69.82	70.35	59.26	**91.07**	72.66
Deep All	63.77	66.77	57.27	88.62	72.66
FC*	64.89	71.72	61.85	89.94	72.1
Deep All	68.09	70.23	61.80	88.86	72.25
MMLD	69.27	**72.83**	66.44	88.98	74.38
Deep All	67.60	68.87	61.13	89.20	71.70
MASF*	70.35	72.46	**67.33**	90.68	**75.21**
Deep All	67.66	69.70	63.76	89.88	72.75
Ours	**71.97**	70.09	66.48	90.12	74.67
Deep All	85.4	77.7	69.5	97.8	82.6
MetaReg*	**87.2**	79.2	70.3	97.6	83.6
Deep All	81.41	78.61	69.69	94.83	81.14
MASF*	82.89	**80.49**	72.29	95.01	82.67
Deep All	85.69	75.00	72.54	97.66	82.72
Ours	87.13	77.51	**76.32**	**98.46**	**84.86**

Table 1. Domain generalization results on PACS. The column title indicates the name of the domain used as target. The asterisk indicates the method uses domain labels in the training progress, but Deep ALL, JiGen, MMLD, and our method do not use them. The last three rows use ResNet-50 as the backbone.

	Caltech	Labelme	Pascal	Sun	Avg.
Deep All	85.73	61.28	62.71	59.33	67.26
CIDDG*	88.83	63.06	64.38	62.10	69.59
Deep All	86.10	55.60	59.10	54.60	63.85
CCSA*	92.30	62.10	67.10	59.10	70.15
Deep All	86.67	58.20	59.10	57.86	65.46
SLRC*	92.76	62.34	65.25	63.54	70.97
Deep All	93.40	62.11	68.41	64.16	72.02
TF*	93.63	63.49	69.99	61.32	72.11
Deep All	94.95	57.45	66.06	65.87	71.08
D-SAM*	91.75	56.95	58.59	60.84	67.03
Deep All	96.93	59.18	71.96	62.57	72.66
JiGEN	96.93	60.90	70.62	64.30	73.19
Deep All	95.89	57.88	<u>**72.01**</u>	<u>67.76</u>	73.39
MMLD	96.66	58.77	71.96	**68.13**	73.88
Deep All	92.86	63.10	68.67	64.11	72.19
MASF*	94.78	**64.90**	69.14	67.64	**74.11**
Deep All	96.07	59.35	68.48	62.40	71.58
Ours	**97.34**	62.39	70.58	65.32	73.91

Table 2. Domain generalization results on VLCS. We underline the result which is higher than all the others despite prodeced by the Deep All baseline.

	Art.	Cartoon	Sketch	Photo	Avg.
DeepAll	67.66	69.70	63.76	89.88	72.75
MI	69.11	69.05	65.83	89.04	73.26
$MI + prior$	71.19	68.98	**67.78**	89.4	74.34
$MI + prior + distill$	**71.97**	**70.09**	66.48	**90.12**	**74.67**

Table 3. Ablation study on PACS.

	Caltech	Labelme	Pascal	Sun	Avg.
DeepAll	96.07	59.35	68.48	62.40	71.58
MI	96.54	59.22	69.10	62.71	71.89
$MI + prior$	95.99	61.23	**70.58**	64.47	73.07
$MI + prior + distill$	**97.34**	**62.39**	**70.58**	**65.32**	**73.91**

Table 4. Ablation study on VLCS.

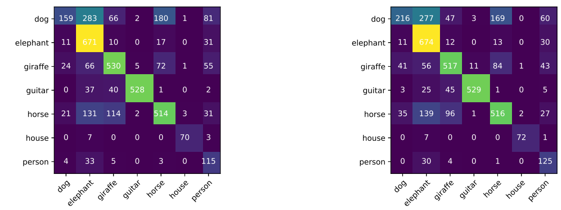
which is shown in Appendix.

4.6. Influence of mutual information ratio β

Figure 6 and Figure 7 report the influence of mutual information ratio β on the performance of PACS and LVCS. In both datasets, our method achieves the best performance when $\beta = 0.05$.

	Art.	Cartoon	Sketch	Photo
$D_s \to D_u$	0.020	0.012	**0.074**	0.031
Among D_s	0.179	**0.187**	0.037	0.152

Table 5. Mismatched label distribution in PACS. Each column title indicates the name of unseen domain. D_s indicates multiple source domains, while D_u indicates the unseen domain.



(a) Confusion matrices when sketches is used as unseen domain.

Figure 1. Class prior-normalized value especially improve the performance on imbalanced classes.

	Art.	Cartoon	Sketch	Photo	Avg.
$\beta = 0.1$	68.49	**70.21**	67.43	**90.32**	74.11
$\beta = 0.05$	**71.19**	68.98	**67.78**	89.4	**74.34**
$\beta = 0.01$	67.72	69.5	66.61	89.46	73.32
$\beta = 0.005$	68.85	69.5	65.51	89.28	73.29
$\beta = 0.001$	68.46	68.88	67.57	89.88	73.70

Table 6. Influence of mutual information ratio β on PACS. Our method obtains best performance for $\beta = 0.05$. All results in the table is the average of three runs.

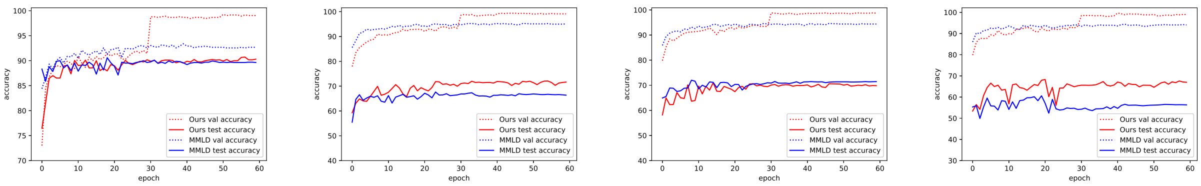
	Caltech	Labelme	Pascal	Sun	Avg.
$\beta = 0.1$	**96.7**	59.56	70.54	**65.05**	72.96
$\beta = 0.05$	95.99	**61.23**	**70.58**	64.47	**73.07**
$\beta = 0.01$	96.46	60.73	69.4	60.81	71.85
$\beta = 0.005$	95.28	59.23	69.4	60.91	71.21
$\beta = 0.001$	95.75	58.22	67.82	64.87	71.67

Table 7. Influence of mutual information ratio β on VLCS.

5. Further study

5.1. Mutual information versus adversarial training

Adversarial training can be viewed as an algorithm that first optimizes the discriminator to approximate the variational upper bound of the mutual information between representation and domain label $MI(f_\phi(x), d)$, then optimizes the feature extractor to minimize that upper bound.



(a) From left to right: the unseen domain is photo, art painting, cartoon, and sketch, respectively.

Figure 2. Comparison with domain adversarial training in PACS.

The adversarial objective function of domain adversarial training can be generally written as :

$$\min_{G} \max_{D} E_{x_{(i)} \sim D} \log[D(E(x_{(i)}))] \tag{9}$$

The domain classifier and the feature extractor can be modeled as $q_D(d|z)$ and $p_E(d, z)$ respectively. We can add a constant value to the object function, the log of label distribution $q(d)$, and rewrite the minmax game as

$$\min_{G} \max_{D} E_{P_E(z,d)}[\log[D(q_D(d|z))] - \log q(d)] \tag{10}$$

This term is a lower bound of an upper bound of the mutual information between representation and domain label $MI(f_\phi(x), d)$ [46, 20]. However, our method focus on the mutual information between the images of the same category from the mixture of multiple domains. It provides a fair new direction for domain generalization. In Figure 2, we also compare our method with the domain adversarial training method, which achieves the state-of-art on the mixture of multiple source domains.

5.2. Mutual information versus triplet loss

Recent work [49] views mutual information maximization from the perspective of metric learning. The lower bound of mutual information is equivalent to triplets loss when maximizing I_{NCE} [42] using symmetric separable critic $f(s, y) = \phi(x)^\top x$ and share the encoder for different view. However, we adopt the JSD objective and use 2 layers fully connected layer as the critic function, which has the connection with asymmetric variants of multi-class K-pair loss [53, 54]. Maximizing mutual information has its benefits since they do not need to carefully choose the negative samples, while the performance of the latter is highly related to semi-hard pair mining or the formulation of the loss function, which inspires some researches in that direction [51, 52, 57].

Table 8 demonstrates the performance with different sample strategies. The column of P and N indicate the positive pairs sample strategy and the negative pairs sample strategy. Hard means the farthest distance among the instances from the same category or the closest instances from other categories. We adopt L2 norm distance as the distance metric in the embedding space. The results demonstrate the significant sample insensitive obtained by sample-based mutual information maximization.

P	N	Art.	Cartoon	Sketch	Photo	Avg.
random	random	68.49	70.21	**67.43**	90.32	**74.11**
random	hard	70.07	69.45	62.46	**90.41**	73.10
hard	random	**70.18**	**70.85**	64.03	89.82	73.72
hard	hard	68.28	69.75	65.52	89.8	73.34

Table 8. Influence of sample strategy on PACS.

6. Conclusion

In this paper, we proposed a mutual information maximization module to take the place of adversarial training in the domain generalization setting. We circumvent the minmax game and the tradeoff between distribution alignment and target error minimization by incorporate class prior-normalized value into the class conditional mutual information estimation. Our methods achieve compatible performance without using domain labels.

References

[1] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H. Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *ArXiv*, abs/1911.00804, 2020. 2

[2] Martín Arjovsky, L. Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *ArXiv*, abs/1907.02893, 2019. 1

[3] Kamyar Azizzadenesheli, Anqi Liu, F. Yang, and Anima Anandkumar. Regularized learning for domain adaptation under label shifts. *ArXiv*, abs/1903.09734, 2019. 1, 5

[4] Y. Balaji, S. Sankaranarayanan, and R. Chellappa. Metareg: Towards domain generalization using meta-regularization. In *NeurIPS*, 2018. 2, 5

[5] Mohamed Ishmael Belghazi, A. Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In *ICML*, 2018. 2

[6] Shai Ben-David, John Blitzer, K. Crammer, A. Kulesza, Fernando C Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2009. 1, 2

[7] Adrian Bulat, Jean Kossaifi, Georgios Tzimiropoulos, and M. Pantic. Incremental multi-domain learning with network latent tensor factorization. In *AAAI*, 2020. 1

[8] Fabio Maria Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. pages 2224–2233, 2019. 2, 4, 5

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020. 1, 2

[10] M. J. Choi, Joseph J. Lim, A. Torralba, and A. Willsky. Exploiting hierarchical context on a large database of object categories. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 129–136, 2010. 4

[11] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and G. Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *ArXiv*, abs/2003.04475, 2020. 1, 3, 4

[12] Z. Ding and Yun Fu. Deep domain generalization with structured low-rank constraint. *IEEE Transactions on Image Processing*, 27:304–313, 2018. 5

[13] Antonio D'Innocente and Barbara Caputo. Domain generalization with domain-specific aggregation modules. *ArXiv*, abs/1809.10966, 2018. 5

[14] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. *ArXiv*, abs/2007.11498, 2020. 2

[15] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, 2019. 2, 4, 5

[16] Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees G. M. Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. *ArXiv*, abs/2007.07645, 2020. 1, 2, 5

[17] M. Everingham, L. Gool, C. K. Williams, J. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2009. 4

[18] Marco Federici, A. Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. *ArXiv*, abs/2002.07017, 2020. 3

[19] Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshops*, 2004. 4

[20] Zeyu Feng, C. Xu, and Dacheng Tao. Self-supervised representation learning from multi-domain data. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3244–3254, 2019. 1, 2, 7

[21] Chelsea Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *ArXiv*, abs/1703.03400, 2017. 2

[22] T. Furlanello, Zachary Chase Lipton, Michael Tschannen, L. Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, 2018. 4

[23] GaninYaroslav, UstinovaEvgeniya, AjakanHana, Germain-Pascal, LarochelleHugo, LavioletteFrançois, Marchand-Mario, and LempitskyVictor. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 2016. 1, 2

[24] Ian J. Goodfellow, Jean Pouget-Abadie, M. Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 1

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. pages 9726–9735, 2020. 1, 2

[26] Geoffrey E. Hinton, Oriol Vinyals, and J. Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015. 4

[27] R. Devon Hjelm, A. Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *ArXiv*, abs/1808.06670, 2019. 1, 2, 3

[28] Xiang Jiang, Qicheng Lao, Stan Matwin, and Mohammad Havaei. Implicit class-conditioned domain alignment for unsupervised domain adaptation. *ArXiv*, abs/2006.04996, 2020. 1

[29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. pages 5543–5551, 2017. 4

[30] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 2

[31] Da Li, J. Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M. Hospedales. Episodic training for domain generalization. pages 1446–1455, 2019. 2

[32] Haoliang Li, Sinno Jialin Pan, S. Wang, and A. Kot. Domain generalization with adversarial feature learning. pages 5400–5409, 2018. 1, 2, 5

[33] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018. 1, 2, 4

[34] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy M. Hospedales. Feature-critic networks for heterogeneous domain generalization. In *ICML*, 2019. 1, 2

[35] Zachary Chase Lipton, Yu-Xiang Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. *ArXiv*, abs/1802.03916, 2018. 1

[36] Yajing Liu, X. Tian, Ya Li, Z. Xiong, and F. Wu. Compact feature learning for multi-domain image classification. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7186–7194, 2019. 1

[37] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 2

[38] Mingsheng Long, Zhangjie Cao, J. Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, 2018. 1, 2

[39] Mingsheng Long, Han Zhu, J. Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. *ArXiv*, abs/1605.06636, 2017. 1, 2

[40] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *AAAI*, 2020. 2, 4, 5

[41] Saeid Motiian, Marco Piccirilli, D. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. pages 5716–5726, 2017. 5

[42] A. Oord, Y. Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 1, 2, 7

[43] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. *ArXiv*, abs/1904.12347, 2019. 2

[44] Bryan C. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173, 2007. 4

[45] S. Shankar, Vihari Piratla, Soumen Chakrabarti, S. Chaudhuri, P. Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *ArXiv*, abs/1804.10745, 2018. 2

[46] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and S. Ermon. Learning controllable fair representations. In *AISTATS*, 2019. 7

[47] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. *ArXiv*, abs/1607.01719, 2016. 1

[48] A. Torralba and Alexei A. Efros. Unbiased look at dataset bias. *CVPR 2011*, pages 1521–1528, 2011. 4

[49] Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning. *ArXiv*, abs/1907.13625, 2020. 7

[50] E. Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, 2015. 4

[51] B. Yu, T. Liu, M. Gong, Changxing Ding, and D. Tao. Correcting the triplet selection bias for triplet loss. In *ECCV*, 2018. 7

[52] B. Yu and Dacheng Tao. Deep metric learning with tuplet margin loss. pages 6489–6498, 2019. 7

[53] Hong-Xing Yu, Ancong Wu, and W. Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. pages 994–1002, 2017. 7

[54] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, A. Elgammal, and Mohamed Elhoseiny. Large-scale visual relationship understanding. *ArXiv*, abs/1804.10660, 2019. 7

[55] Yuchen Zhang, T. Liu, Mingsheng Long, and Michael I. Jordan. Bridging theory and algorithm for domain adaptation. In *ICML*, 2019. 1

[56] Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J. Gordon. On learning invariant representations for domain adaptation. In *ICML*, 2019. 3

[57] Wenzhao Zheng, Zhaodong Chen, Jiwen Lu, and J. Zhou. Hardness-aware deep metric learning. pages 72–81, 2019. 7

[58] Fan Zhou, Zhuqing Jiang, Changjian Shui, B. Wang, and B. Chaib-draa. Domain generalization with optimal transport and metric learning. *ArXiv*, abs/2007.10573, 2020. 1

[59] Kaiyang Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. *ArXiv*, abs/2007.03304, 2020. 2

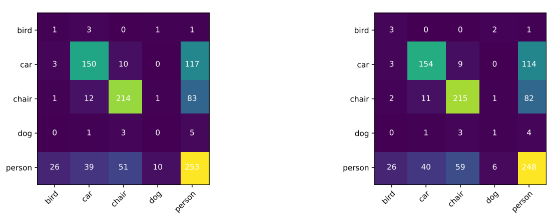
7. Appendix

7.1. Evaluation of class prior-normalized value on VLCS

We provide here some further analysis on the effectiveness of class prior-normalized value on VLCS. Table 9 shows the discrepancy between two kinds of marginal label distribution. Applying class prior-normalized value in VLCS improves the performance significantly when the unseen domain is Sun, which suffers the most serious generalization discrepancy than other unseen domains. We further demonstrate the confusion metrics in Figure 3. Class prior-normalized value promotes the performance especially on imbalanced classes, e.g., bird, dog, and car.

	Caltech	Labelme	Pascal	Sun
$D_s \to D_u$	0.0028	0.0052	0.0078	**0.0124**
Among D_s	**0.0438**	0.0370	0.032	0.0219

Table 9. Mismatched label distribution in VLCS. Each column title indicates the name of unseen domain. D_s indicates multiple source domains, while D_u indicates the unseen domain.



(a) Confusion matrices when Pascal is used as unseen domain. Left: applying mutual information without class prior-normalized value. Right: applying mutual information with class prior-normalized value.

Figure 3. Class prior-normalized value especially improve the performance on imbalanced classes.

7.2. Influence of the number of negative pairs

We compare the influence of the number of negative samples when calculating the mutual information between images of the same category from the mixture of multiple domains. Table 10 and Table 11 demonstrate our method is very insensitive to the number of negative samples.

Num	Art.	Cartoon	Sketch	Photo	Avg.
1	68.49	**70.21**	**67.43**	**90.32**	**74.34**
2	**68.58**	69.41	66.96	90.04	73.75
5	68.53	70.16	66.91	89.74	73.84

Table 10. Influence of the number of negative samples on PACS (accuracy, %).

Num	Art.	Cartoon	Sketch	Photo	Avg.
1	95.99	**61.23**	**70.58**	64.47	**73.07**
2	95.28	59.68	67.62	65.46	72.01
5	**96.15**	60.14	66.66	**66.23**	72.29

Table 11. Influence of the number of negative samples on VLCS (accuracy, %).

7.3. Influence of the number of pairs in a mini-batch

Here we analyze the influence of the number of pairs we sampled from one mini-batch to compute pair-wise mutual information. Even in an extremely small number of sample pairs in one mini-batch, our method can still promote the performance against the Deep All baseline. The result is shown in Table 12 and Table 13.

Num	Art.	Cartoon	Sketch	Photo	Avg.
1	69.45	69.74	67.09	90.14	74.10
5	67.48	69.28	66.28	89.52	72.89
10	68.99	**71.89**	64.62	**90.48**	74.00
50	**71.19**	68.98	**67.78**	89.4	**74.34**

Table 12. Influence of the number of pairs in a mini-batch on PACS (accuracy, %).

Num	Caltech	Labelme	Pascal	Sun	Avg.
1	**96.15**	59.09	64.97	69.07	72.32
5	95.84	68.76	66.8	63.15	71.14
10	93.63	**61.98**	64.17	**64.97**	71.27
50	95.99	61.23	**70.58**	64.47	**73.07**

Table 13. Influence of the number of pairs in a mini-batch on VLCS (accuracy, %).