Domain Generalization via Entropy Regularization

通过熵正则化的领域泛化

Shanshan Zhao
赵珊珊
The University of Sydney
悉尼大学
Australia
澳大利亚
szha4333@uni.sydney.edu.au
Mingming Gong
龚明明
University of Melbourne
墨尔本大学
Australia
澳大利亚
mingming.gong@unimelb.edu.au
Tongliang Liu
刘同亮
The University of Sydney
悉尼大学
Australia
澳大利亚
tongliang.liu@sydney.edu.au
Huan Fu
傅欢
Alibaba Group
阿里巴巴集团
China
中国
fuhuan.fh@alibaba-inc.com
Dacheng Tao
陶大成
The University of Sydney
悉尼大学
Australia
澳大利亚
dacheng.tao@sydney.edu.au

Abstract

摘要

Domain generalization aims to learn from multiple source domains a predictive model that can generalize to unseen target domains. One essential problem in domain generalization is to learn discriminative domain-invariant features. To arrive at this, some methods introduce a domain discriminator through adversarial learning to match the feature distributions in multiple source domains. However, adversarial training can only guarantee that the learned features have invariant marginal distributions, while the invariance of conditional distributions is more important for prediction in new domains. To ensure the conditional invariance of learned features, we propose an entropy regularization term that measures the dependency between the learned features and the class labels. Combined with the typical task-related loss, e.g., cross-entropy loss for classification, and adversarial loss for domain discrimination, our overall objective is guaranteed to learn conditional-invariant features across all source domains and thus can learn classifiers with better generalization capabilities. We demonstrate the effectiveness of our method through comparison with state-of-the-art methods on both simulated and real-world datasets. Code is available at: https://github.com/sshan-zhao/DG_via_ER.

领域泛化旨在从多个源领域学习一个可以推广到未见目标领域的预测模型。领域泛化中的一个基本问题是学习具有区分性的领域不变特征。为此，一些方法通过对抗学习引入领域鉴别器，以匹配多个源领域中的特征分布。然而，对抗训练只能保证学习到的特征具有不变的边际分布，而条件分布的不变性对于在新领域中的预测更为重要。为了确保学习到的特征的条件不变性，我们提出了一种熵正则化项，用于衡量学习特征与类别标签之间的依赖关系。结合典型的任务相关损失，例如分类的交叉熵损失和领域鉴别的对抗损失，我们的总体目标是确保在所有源领域中学习条件不变特征，从而能够学习具有更好泛化能力的分类器。我们通过与最先进的方法在模拟和真实数据集上的比较，展示了我们方法的有效性。代码可在以下链接获取:https://github.com/sshan-zhao/DG_via_ER。

1 Introduction

1 引言

Recent years have witnessed the remarkable success of modern machine learning techniques in various applications. However, a fundamental problem machine learning suffers from is that the model learned from training data often does not generalize well on data sampled from a different distribution, due to the existence of data bias [1, 2] between the training and test data. To tackle this issue, a significant effort has been made in domain adaptation, which reduces the discrepancy between source and target domains [3-8]. The main drawback of this approach is that one has to repeat training for each new dataset, which can be time-consuming. Therefore, domain generalization [9] is proposed to learn generalizable models by leveraging information from multiple source domains [10-13].

近年来，现代机器学习技术在各种应用中取得了显著成功。然而，机器学习面临的一个根本问题是，从训练数据学习的模型往往无法很好地在从不同分布中抽样的数据上进行泛化，这主要是由于训练数据和测试数据之间存在数据偏差 [1, 2] 。为了解决这个问题，领域适应方面进行了大量努力，旨在减少源领域和目标领域之间的差异 [3-8]。这种方法的主要缺点是必须为每个新数据集重复训练，这可能非常耗时。因此，提出了领域泛化 [9]，旨在通过利用来自多个源领域的信息来学习可泛化的模型 [10-13]。

Since there is no prior information about the distribution of the target domain during training, it is difficult to match the distributions between source and target domains, which makes domain generalization more challenging. To improve the generalization capabilities of learned models, various solutions have been developed from different perspectives. A classic but effective solution to domain generalization is learning a domain-invariant feature representation [11,12,14,10,15,14] across source domains. Muandet et al. [10] presented a kernel-based optimization algorithm, called Domain-Invariant Component Analysis, to learn an invariant transformation by minimizing the dissimilarity across domains. Ghifary et al. [11] proposed to learn features robust to variations across domains by introducing multi-task auto-encoders. Another line of research explores various data augmentation strategies [16-18]. For example, Shankar et al. [16] presented a gradient-based domain perturbation strategy to perturb the input data. By augmenting the original feature space, Blanchard et al. [19] viewed the problem of domain generalization as a kind of supervised learning problem. Then, they developed a kernel-based method that predicts classifiers from the augmented feature space. To make theoretical complementary to these empirically supported approaches, Deshmukh et al. [20] proved the first known generalization error bound for multi-class domain generalization through studying a kernel-based learning algorithm. Apart from the clues aforementioned, some recent works [21-24] attempted to exploit meta-learning for domain generalization. A latest work, MASF [21], proposed a model-agnostic episodic learning procedure to regularize the semantic structure of the feature space.

由于在训练期间没有关于目标领域分布的先前信息，因此很难匹配源领域和目标领域之间的分布，这使得领域泛化变得更加具有挑战性。为了提高学习模型的泛化能力，已经从不同的角度开发了各种解决方案。一个经典但有效的领域泛化解决方案是学习跨源领域的领域不变特征表示 [11,12,14,10,15,14]。Muandet 等人 [10] 提出了一个基于核的优化算法，称为领域不变成分分析，通过最小化领域之间的不相似性来学习不变变换。Ghifary 等人 [11] 提出了通过引入多任务自编码器来学习对领域变化具有鲁棒性的特征。另一条研究方向探索了各种数据增强策略 [16-18]。例如，Shankar 等人 [16] 提出了基于梯度的领域扰动策略来扰动输入数据。通过增强原始特征空间，Blanchard 等人 [19] 将领域泛化问题视为一种监督学习问题。然后，他们开发了一种基于核的方法，从增强的特征空间中预测分类器。为了在理论上补充这些经验支持的方法，Deshmukh 等人 [20] 通过研究一种基于核的学习算法证明了多类领域泛化的第一个已知泛化误差界限。除了上述线索，一些近期的工作 [21-24] 尝试利用元学习进行领域泛化。最新的工作 MASF [21] 提出了一个模型无关的情景学习程序，以规范特征空间的语义结构。

In this paper, we revisit the domain-invariant feature representation learning methods. Most of existing methods assume that the marginal distribution $P(X)$ changes while the conditional distribution

$P(Y \mid X)$ stays stable across domains. Therefore, significant effort has been made in learning a feature representation $F(X)$ that has invariant $P(F(X))$, either by traditional moment matching [25] or modern adversarial training [15,14]. To ensure the universality of $F(X)$ and also make it discriminative, a joint classification model is trained on all the source domains and can be used for prediction in new datasets. However, the stability of $P(Y \mid X)$ is often violated in real applications, leading to sub-optimal solutions. Li et al. [14] proposed to learn invariant class-conditional distribution ($P(F(X) \mid Y)$) by doing adversarial training for each class. However, the method becomes less effective as the number of classes increases.

在本文中，我们重新审视了领域不变特征表示学习方法。现有的大多数方法假设边际分布 $P(X)$ 发生变化，而条件分布 $P(Y \mid X)$ 在各个领域保持稳定。因此，已经进行了大量工作以学习具有不变性 $P(F(X))$ 的特征表示 $F(X)$ ，这可以通过传统的时刻匹配 [25] 或现代对抗训练 [15,14] 来实现。为了确保 $F(X)$ 的普遍性并使其具有区分性，我们在所有源领域上训练了一个联合分类模型，并可以用于新数据集的预测。然而，实际应用中 $P(Y \mid X)$ 的稳定性往往受到破坏，导致次优解。Li 等 [14] 提出了通过对每个类别进行对抗训练来学习不变的类别条件分布 ($P(F(X) \mid Y)$) 。然而，随着类别数量的增加，该方法的有效性降低。

To tackle the aforementioned issues, we propose an entropy-regularization approach which directly learns features that have invariant $P(Y \mid F(X))$ across domains. In specific, the conditional entropy term $H(Y \mid F(X))$ measures the dependency between $F(X)$ and class label Y , and we aim to minimize the dependency by maximizing the conditional entropy. We show theoretically that our entropy-regularization together with the cross-entropy classification loss effectively minimize the divergence between $P(Y \mid F(X))$ in all source domains. In addition, we show that $H(Y \mid F(X))$ can be effectively estimated by assuming a multinomial distribution for $P(Y \mid F(X))$, which is a weak assumption for discrete class labels. Together with the adversarial training on $P(F(X))$, our approach can guarantee the invariance of the joint distribution $P(F(X), Y)$ and thus has a better generalization capability. We demonstrate the effectiveness of our approach through conducting comprehensive experiments on several benchmark datasets.

为了解决上述问题，我们提出了一种熵正则化方法，该方法直接学习在各个领域具有不变性 $P(Y \mid F(X))$ 的特征。具体而言，条件熵项 $H(Y \mid F(X))$ 衡量 $F(X)$ 与类别标签 Y 之间的依赖关系，我们的目标是通过最大化条件熵来最小化这种依赖关系。我们从理论上证明，我们的熵正则化与交叉熵分类损失有效地最小化了所有源领域中 $P(Y \mid F(X))$ 之间的散度。此外，我们展示了可以通过假设 $P(Y \mid F(X))$ 的多项分布来有效估计 $H(Y \mid F(X))$ ，这对于离散类别标签来说是一个较弱的假设。结合对 $P(F(X))$ 的对抗训练，我们的方法可以保证联合分布 $P(F(X), Y)$ 的不变性，从而具有更好的泛化能力。我们通过在多个基准数据集上进行全面实验来证明我们方法的有效性。

2 Method

2 方法

2.1 Problem Definition

2.1 问题定义

Let $\mathcal{X}$ and $\mathcal{Y}$ be the feature and label spaces, respectively. In the domain generalization subject, there are K source domains $\{\mathcal{D}_i\}_{i=1}^{K}$ and L target domains [1]$\{\mathcal{D}_i\}_{i=K+1}^{L+K}$. The goal is to generalize the model learned using data samples of source domains to unseen target domains. In the following, we denote the joint distribution of domain i by $P_i(X, Y)$ (defined on $\mathcal{X} \times \mathcal{Y}$). During the training process, there are K datasets $\{S_i\}_{i=1}^{K}$ available, where $S_i = \left\{\left(\mathbf{x}_j^{(i)}, y_j^{(i)}\right)\right\}_{j=1}^{N_i}$. Here, N_i is the number of samples of S_i , which are sampled from the i^{th} domain. In the test stage, we evaluate the generalization capabilities of the learned model on L datasets sampled from the L target domains, respectively. This paper mainly studies domain generalization for image classification, where the label space $\mathcal{Y}$ contains C discrete labels $\{1, 2, \cdots, C\}$.

设 $\mathcal{X}$ 和 $\mathcal{Y}$ 分别为特征空间和标签空间。在领域泛化的主题中，有 K 个源领域 $\{\mathcal{D}_i\}_{i=1}^{K}$ 和 L 个目标领域 [1]$\{\mathcal{D}_i\}_{i=K+1}^{L+K}$ 。目标是将使用源领域的数据样本学习到的模型推广到未见过的目标领域。在下面，我们用 $P_i(X, Y)$ 表示领域 i 的联合分布 (定义在 $\mathcal{X} \times \mathcal{Y}$ 上)。在训练过程中，有 K 个数据集 $\{S_i\}_{i=1}^{K}$ 可用，其中 $S_i = \left\{\left(\mathbf{x}_j^{(i)}, y_j^{(i)}\right)\right\}_{j=1}^{N_i}$ 。这里，N_i 是从 i^{th} 领域中抽样的 S_i 的样本数量。在测试阶段，我们

评估学习到的模型在分别从 L 个目标领域抽样的 L 个数据集上的泛化能力。本文主要研究图像分类的领域泛化，其中标签空间 $\mathcal{Y}$ 包含 C 个离散标签 $\{1, 2, \cdots, C\}$ 。

2.2 Domain Generalization Through Adversarial Learning

2.2 通过对抗学习进行领域泛化

We first present how domain generalization can be learned in an adversarial learning framework.

我们首先介绍如何在对抗学习框架中学习领域泛化。

For the classification subject, the model consists of one feature extractor F parameterized by θ and one classifier T parameterized by ϕ . We can optimize θ and ϕ on the K source datasets by minimizing a cross-entropy loss:

对于分类主题，模型由一个特征提取器 F 和一个分类器 T 组成，前者由 θ 参数化，后者由 ϕ 参数化。我们可以通过最小化交叉熵损失来优化 θ 和 ϕ 在 K 源数据集上的表现:

$$\min_{F,T} \mathcal{L}_{cls}(\theta, \phi) = -\sum_{i=1}^{K} \mathop{\mathbb{E}}_{(X,Y)\sim P_i(X,Y)} \left[\log\left(Q^T(Y \mid F(X))\right)\right] \tag{1}$$

$$= -\sum_{i=1}^{K}\sum_{j=1}^{N_i} \mathbf{y}_j^{(i)} \cdot \log\left(T\left(F\left(\mathbf{x}_j^{(i)}\right)\right)\right)$$

where $\mathbf{y}_j^{(i)}$ is the one-hot vector of the class label $y_j^{(i)}$," represents the dot product operation, and $Q^T(Y \mid F(X))$ denotes the predicted label distribution (conditioned on $F(X)$) corresponding to domain i .

其中 $\mathbf{y}_j^{(i)}$ 是类别标签 $y_j^{(i)}$ 的独热向量，"代表点积操作，$Q^T(Y \mid F(X))$ 表示与领域 i 对应的预测标签分布 (以 $F(X)$ 为条件)。

However, optimized by the classification loss solely, the model cannot learn domain-invariant features, and thus shows limitations in generalizing to the unseen domains. By exploiting the adversarial learning [26], we can alleviate the issue. Specifically, we further introduce a domain discriminator D parameterized by ψ , and train D and F in a minimax game as follows:

然而，仅通过分类损失进行优化，模型无法学习到领域不变特征，因此在推广到未见领域时表现出局限性。通过利用对抗学习 [26]，我们可以缓解这个问题。具体而言，我们进一步引入一个由 ψ 参数化的领域鉴别器 D ，并在以下的极小极大博弈中训练 D 和 F :

$$\min_{F}\max_{D} \mathcal{L}_{adv}(\theta, \psi) = \sum_{i=1}^{K} \mathop{\mathbb{E}}_{X\sim P_i(X)} \left[\log D(F(X))\right] \tag{2}$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{N_i} \mathbf{d}_j^{(i)} \cdot \log\left(D\left(F\left(\mathbf{x}_j^{(i)}\right)\right)\right)$$

where $\mathbf{d}_j^{(i)}$ is the one-hot representation of the domain label i .

其中 $\mathbf{d}_j^{(i)}$ 是领域标签 i 的独热表示。

lthough optimizing Eq. 2 can lead to invariant marginal distributions i.e., $P_1(F(X)) = P_2(F(X)) = \cdots = P_K(F(X))$, it cannot guarantee the conditional distribution $P(Y \mid F(X))$ is invariant across domains. This would degrade the generalization capabilities of the model. Even though the classifier attempts to cluster the samples from the same category together in the feature space, which benefits to the learning of the invariant conditional distribution, there still exists an issue. We take the simulated data for example. Firstly, we sample data from two 2D-distributions (shown in Figure 1) as the Domain_0 and Domain_1, respectively. The marginal distributions of the first dimension (x_0) in the two domain are the same, while the second (x_1) comes from different marginal distributions. Each domain consists of three components. We take each dimension as the input to train a classifier using Eq. 1 and Eq. 2, and we find that the classifier distinguishes the second dimension better than the first (loss: -0.34 v.s. -0.16).

[1] Source/Target: seen/unseen during training.

[1] 源/目标: 在训练期间的已见/未见。

This indicates that the classifier might not select the domain-invariant feature, but select the features easier to discriminate. Therefore, it is challenging for the typical classification loss to achieve a balance between learning domain-invariant features and discriminative features.

尽管优化方程 2 可以导致不变的边际分布，即 $P_1(F(X)) = P_2(F(X)) = \cdots = P_K(F(X))$ ，但它无法保证条件分布 $P(Y \mid F(X))$ 在不同领域之间是不变的。这将降低模型的泛化能力。即使分类器试图在特征空间中将来自同一类别的样本聚集在一起，这有助于学习不变的条件分布，但仍然存在一个问题。我们以模拟数据为例。首先，我们从两个二维分布中采样数据 (如图 1 所示)，分别作为 Domain_0 和 Domain_1。两个领域的第一维的边际分布 (x_0) 是相同的，而第二维 (x_1) 来自不同的边际分布。每个领域由三个组成部分构成。我们将每个维度作为输入，使用方程 1 和方程 2 训练分类器，我们发现分类器对第二维的区分能力优于第一维 (损失:-0.34 对 -0.16)。这表明分类器可能没有选择领域不变特征，而是选择了更容易区分的特征。因此，典型的分类损失在学习领域不变特征和区分特征之间实现平衡是具有挑战性的。

2.3 Entropy Regularization

2.3 熵正则化

Description. To address the issues aforementioned, we propose to regularize the distributions of the features by minimizing the KL divergence between the conditional distribution $P_i(Y \mid F(X))$ in the i^{th} domain and the conditional distribution $Q^T(Y \mid X)$. $P_i(Y \mid F(X))$ denotes the predicted label distribution conditioned on the learned features. By matching any conditional distribution $P_i(Y \mid F(X))$ to a common distribution $Q^T(Y \mid F(X))$, we can obtain the domain-invariant conditional distribution $P(Y \mid F(X))$. For the purpose, we define an optimization problem as follows:

描述。为了解决上述问题，我们提出通过最小化条件分布 $P_i(Y \mid F(X))$ 在 i^{th} 域与条件分布 $Q^T(Y \mid X)$. $P_i(Y \mid F(X))$ 之间的 KL 散度来规范特征的分布，其中条件分布 $Q^T(Y \mid X)$. $P_i(Y \mid F(X))$ 表示基于学习特征的预测标签分布。通过将任何条件分布 $P_i(Y \mid F(X))$ 匹配到一个共同分布 $Q^T(Y \mid F(X))$ ，我们可以获得域不变的条件分布 $P(Y \mid F(X))$ 。为此，我们定义如下优化问题:

$$\min_{F,T} \sum_{i=1}^{K} KL\left(P_i(Y \mid F(X)) \parallel Q^T(Y \mid F(X))\right). \tag{3}$$

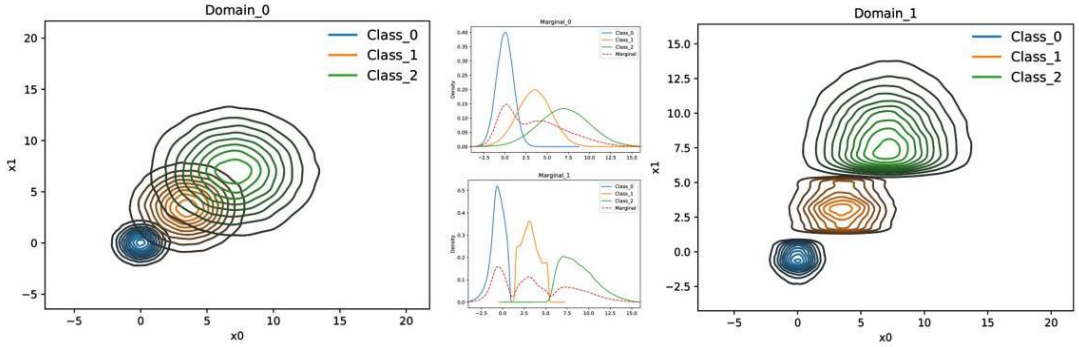

Figure 1: Simulated data. We create two domains from the two 2D-distributions (left and right), respectively. The data in Domain_0 and Domain_1 is two-dimensional. In specific, the first dimensions in two domains are both sampled from Marginal_0 (top-middle), while the second dimension in Domain_0 and Domain_1 is sampled from Marginal_0 and Marginal_1 (bottom-middle), respectively.

图 1: 模拟数据。我们分别从两个二维分布 (左侧和右侧) 创建两个域。Domain_0 和 Domain_1 中的数据是二维的。具体而言，两个域中的第一维均来自 Marginal_0(顶部中间)，而 Domain_0 和 Domain_1 中的第二维分别来自 Marginal_0 和 Marginal_1(底部中间)。

By using the definition of the KL divergence, we have:

通过使用 KL 散度的定义，我们有:

$$\min_{F,T} \sum_{i=1}^{K} KL\left(P_i(Y \mid F(X)) \parallel Q^T(Y \mid F(X))\right) = \sum_{i=1}^{K} \mathop{\mathbb{E}}_{(X,Y)\sim P_i(X,Y)} \left[\log \frac{P_i(Y \mid F(X))}{Q^T(Y \mid F(X))}\right] \tag{4}$$

$$= \sum_{i=1}^{K} \mathop{\mathbb{E}}_{(X,Y)\sim P_i(X,Y)} \left[\log P_i\left(Y \mid F\left(X\right)\right)\right] - \sum_{i=1}^{K} \mathop{\mathbb{E}}_{(X,Y)\sim P_i(X,Y)} \left[\log Q^{T}\left(Y \mid F\left(X\right)\right)\right].$$

The second term is actually the cross-entropy classification loss (Eq. 1), while the first one is the sum of K negative conditional entropy terms $\sum_{i=1}^{K} -H_{P_i}\left(Y \mid F\left(X\right)\right)$. However, it is difficult to optimize $-H_{P_i}\left(Y \mid F\left(X\right)\right)$ directly, since we do not know the conditional distribution $P_i\left(Y \mid F\left(X\right)\right)$. To overcome this issue, we first provide the following theorem to exploit the relationship between the negative conditional entropy term and the Jensen-Shannon divergence (JSD) between the conditional distributions $\{P_i\left(F\left(X\right) \mid Y=c\right)\}_{c=1}^{C}$.

第二项实际上是交叉熵分类损失 (方程 1)，而第一项是 K 负条件熵项 $\sum_{i=1}^{K} -H_{P_i}\left(Y \mid F\left(X\right)\right)$ 的总和。然而，由于我们不知道条件分布 $P_i\left(Y \mid F\left(X\right)\right)$，直接优化 $-H_{P_i}\left(Y \mid F\left(X\right)\right)$ 是困难的。为了解决这个问题，我们首先提供以下定理，以利用负条件熵项与条件分布之间的 Jensen-Shannon 散度 (JSD) 之间的关系 $\{P_i\left(F\left(X\right) \mid Y=c\right)\}_{c=1}^{C}$。

Theorem 1. Assuming that all classes are equally likely, minimizing $-H_{P_i}\left(Y \mid F\left(X\right)\right)$ is equivalent to minimizing the JSD between the conditional distributions $\{P_i\left(F\left(X\right) \mid Y=c\right)\}_{c=1}^{C}$. The global minimum is achieved if and only if $P_i\left(F\left(X\right) \mid Y=1\right) = P_i\left(F\left(X\right) \mid Y=2\right) = \cdots = P_i(F\left(X\right) \mid Y=C)$. Note that, if the dataset is balanced, it is easy to make the assumption satisfied. Otherwise, we can enforce it through biased batch sampling.

定理 1。假设所有类别的可能性相等，最小化 $-H_{P_i}\left(Y \mid F\left(X\right)\right)$ 等价于最小化条件分布之间的 JSD $\{P_i\left(F\left(X\right) \mid Y=c\right)\}_{c=1}^{C}$。全局最小值仅在 $P_i\left(F\left(X\right) \mid Y=1\right) = P_i\left(F\left(X\right) \mid Y=2\right) = \cdots = P_i(F\left(X\right) \mid Y=C)$ 时达到。注意，如果数据集是平衡的，满足该假设是容易的。否则，我们可以通过偏向批次采样来强制满足该假设。

The proof is given in Sec. S1 of the Supplementary Materials. Inspired by Theorem 1 and the minimax game proposed in GAN [26] and conditional GAN [27], we introduce K additional classifiers $\{T_i'\}_{i=1}^{K}$, and then present the following minimax game:

证明见补充材料的第 S1 节。受到定理 1 和 GAN [26] 及条件 GAN [27] 中提出的极小极大博弈的启发，我们引入 K 额外的分类器 $\{T_i'\}_{i=1}^{K}$，然后提出以下极小极大博弈:

$$\min_{F} \max_{\{T_i'\}_{i=1}^{K}} V\left(F, T_1', T_2', \cdots, T_K'\right) = \sum_{i=1}^{K} \mathop{\mathbb{E}}_{(X,Y)\sim P_i(X,Y)} \left[\log Q_i^{T_i'}\left(Y \mid F\left(X\right)\right)\right], \tag{5}$$

where T_i' parameterized by ϕ_i' represents a classifier trained on data sampled from domain $\mathcal{D}_i$, and $Q_i^{T_i'}\left(Y \mid F\left(X\right)\right)$ denotes the conditional distribution induced by T_i'. The following theorem (the proof can be found in Sec. S2 of the Supplementary Materials) shows that the minimax game is equal to minimizing the JSD between the conditional distributions $\{P_i\left(F\left(X\right) \mid Y=c\right)\}_{c=1}^{C}$. According to Theorem 1, we can thus achieve the optimization of $\sum_{i=1}^{K} -H_{P_i}\left(Y \mid F\left(X\right)\right)$.

其中 T_i' 由 ϕ_i' 参数化，表示在从域 $\mathcal{D}_i$ 中抽样的数据上训练的分类器，而 $Q_i^{T_i'}\left(Y \mid F\left(X\right)\right)$ 表示由 T_i' 引发的条件分布。以下定理 (证明见补充材料的第 S2 节) 表明，极小极大博弈等同于最小化条件分布之间的 JSD $\{P_i\left(F\left(X\right) \mid Y=c\right)\}_{c=1}^{C}$。根据定理 1，我们因此可以实现 $\sum_{i=1}^{K} -H_{P_i}\left(Y \mid F\left(X\right)\right)$ 的优化。

Theorem 2. If $U\left(F\right)$ is the maximum value of $V\left(F, T_1', T_2', \cdots, T_K'\right)$, i.e.,

定理 2。如果 $U\left(F\right)$ 是 $V\left(F, T_1', T_2', \cdots, T_K'\right)$ 的最大值，即，

$$U\left(F\right) = \max_{\{T_i'\}_{i=1}^{K}} V\left(F, T_1', T_2', \cdots, T_K'\right), \tag{6}$$

the global minimum of the minimax game is attained if and only if $P_i\left(F\left(X\right) \mid Y=1\right) = P_i\left(F\left(X\right) \mid Y=2\right) = \cdots = P_i\left(F\left(X\right) \mid Y=C\right)$. At this point, $U\left(F\right)$ attains the value $-KC\log C$.

当且仅当 $P_i\left(F\left(X\right) \mid Y=1\right) = P_i\left(F\left(X\right) \mid Y=2\right) = \cdots = P_i\left(F\left(X\right) \mid Y=C\right)$ 时，极小极大博弈的全局最小值被达到。在此时，$U\left(F\right)$ 达到值 $-KC\log C$。

Therefore, our proposed entropy regularization loss can be defined as:

因此，我们提出的熵正则化损失可以定义为:

$$\min_{F}\max_{\{T_i'\}_{i=1}^{K}}\mathcal{L}_{er}\left(\theta,\{\phi_i'\}_{i=1}^{K}\right)=\sum_{i=1}^{K}\mathop{\mathbb{E}}_{(X,Y)\sim P_i(X,Y)}\left[\log Q_i^{T_i'}\left(Y\mid F\left(X\right)\right)\right]. \tag{7}$$

Combining Eq. 7 with the classification loss (Eq. 1) and the domain discrimination loss (Eq. 2), we obtain the training objective:

将方程 7 与分类损失 (方程 1) 和领域区分损失 (方程 2) 结合，我们得到训练目标:

$$\min_{F,T}\max_{D,\{T_i'\}_{i=1}^{K}}\mathcal{L}\left(\theta,\phi,\psi,\{\phi_i'\}_{i=1}^{K}\right)=\mathcal{L}_{cls}\left(\theta,\phi\right)+\alpha_1\mathcal{L}_{adv}\left(\theta,\psi\right)+\alpha_2\mathcal{L}_{er}\left(\theta,\{\phi_i'\}_{i=1}^{K}\right), \tag{8}$$

where α_1 and α_2 are trade-off parameters.

其中 α_1 和 α_2 是权衡参数。

Algorithm. In our experiments, we observed that directly optimizing the loss Eq. 8 may show instability, since the minimax game in Eq. 7 encourages the learned features not to be distinguished by the classifiers. That may impede the optimization of the classification loss. To alleviate this issue, we introduce additional classifiers $\{T_i\}_{i=1}^{K}$ and add a new cross-entropy loss $\mathcal{L}_{\text{cel}}$:

算法。在我们的实验中，我们观察到直接优化损失方程 8 可能会显示不稳定性，因为方程 7 中的极小极大博弈鼓励学习到的特征不被分类器区分。这可能会妨碍分类损失的优化。为了解决这个问题，我们引入额外的分类器 $\{T_i\}_{i=1}^{K}$ 并添加一个新的交叉熵损失 $\mathcal{L}_{\text{cel}}$:

$$\begin{aligned}\min_{F,\{T_i\}_{i=1}^{K}}\mathcal{L}_{cel}\left(\theta,\{\phi_i\}_{i=1}^{K}\right)=&-\sum_{i=1}^{K}\mathop{\mathbb{E}}_{(X,Y)\sim P_i(X,Y)}\left[\log Q_i^{T_i}\left(Y\mid\bar{F}\left(X\right)\right)\right]\\&-\sum_{i=1}^{K}\sum_{j=1,j\neq i}^{K}\mathop{\mathbb{E}}_{(X,Y)\sim P_j(X,Y)}\left[\log Q_i^{\bar{T}_i}\left(Y\mid F\left(X\right)\right)\right],\end{aligned} \tag{9}$$

where $Q_i^{T_i}\left(Y\mid F\left(X\right)\right)$ denotes the conditional distribution induced by T_i . Here, $\bar{F}$ and $\bar{T}_i$ mean that we fix the parameters of F and T during the training procedure, respectively. Specifically, we feed the learned features in the i^{th} domain into T_i to optimize its parameters ϕ_i . Additionally, we expect the feature extractor can map the data in domains $\{\mathcal{D}_j\}_{j=1,j\neq i}^{K}$ to a representation, which can be distinguished by T_i accurately. This strategy, on the one hand, can impose regularization on the feature distribution of domains $\{\mathcal{D}_j\}_{j=1,j\neq i}^{K}$. On the other hand, the new loss can be considered as a complementary of $\mathcal{L}_{cls}$.

其中 $Q_i^{T_i}\left(Y\mid F\left(X\right)\right)$ 表示由 T_i 引起的条件分布。在这里，$\bar{F}$ 和 $\bar{T}_i$ 意味着我们在训练过程中分别固定 F 和 T 的参数。具体而言，我们将 i^{th} 领域中学习到的特征输入到 T_i 中，以优化其参数 ϕ_i 。此外，我们期望特征提取器能够将 $\{\mathcal{D}_j\}_{j=1,j\neq i}^{K}$ 领域中的数据映射到一个表示中，该表示可以被 T_i 准确区分。这一策略一方面可以对 $\{\mathcal{D}_j\}_{j=1,j\neq i}^{K}$ 领域的特征分布施加正则化，另一方面，新损失可以被视为 $\mathcal{L}_{cls}$ 的补充。

Thus, our final objective is formulated as:

因此，我们的最终目标被表述为:

$$\min_{F,T,\{T_i\}_{i=1}^{K}}\max_{D,\{T_i'\}_{i=1}^{K}}\mathcal{L}\left(\theta,\phi,\psi,\{\phi_i\}_{i=1}^{K},\{\phi_i'\}_{i=1}^{K}\right)=\mathcal{L}_{cls}+\alpha_1\mathcal{L}_{adv}+\alpha_2\mathcal{L}_{er}+\alpha_3\mathcal{L}_{cel}, \tag{10}$$

where α_3 is a weighting factor. To illustrate the training process clearly, we provide the pseudo-code of our algorithm in Alg. 1. We also provide the framework in the Supplementary Materials.

其中 α_3 是一个权重因子。为了清晰地说明训练过程，我们在算法 1 中提供了我们算法的伪代码。我们还在补充材料中提供了框架。

Algorithm 1: Training algorithm for domain generalization via entropy regularization.

算法 1: 通过熵正则化进行领域泛化的训练算法。

Input: $\{S_i\}_{i=1}^{K}$: K source training datasets

输入: $\{S_i\}_{i=1}^{K}$: K 源训练数据集

Input: $\alpha_1,\alpha_2,\alpha_3$: weighting factors

输入: $\alpha_1,\alpha_2,\alpha_3$: 权重因子

Output: F : feature extractor; $T,\{T_i\}_{i=1}^{K},\{T_i'\}_{i=1}^{K}$: classifier; D : discriminator

输出: F : 特征提取器; $T,\{T_i\}_{i=1}^{K},\{T_i'\}_{i=1}^{K}$: 分类器; D : 鉴别器

while training is not end do

当训练未结束时

Sample data from each training dataset respectively
分别从每个训练数据集中抽样数据
Update θ, ϕ , and ψ by optimizing the first and second terms of Eq. 10
通过优化方程 10 的第一和第二项更新 θ, ϕ 和 ψ
for i in $1:K$ do
对于 i 在 $1:K$ 中
Sample data from the i^{th} dataset S_i
来自 i^{th} 数据集 S_i 的示例数据
Update $\{\phi_i\}_{i=1}^K$ by optimizing the forth term of Eq. 10
通过优化方程 10 的第四项来更新 $\{\phi_i\}_{i=1}^K$
Update θ , and $\{\phi_i'\}_{i=1}^K$ by optimizing the third term of Eq. 10
通过优化方程 10 的第三项来更新 θ 和 $\{\phi_i'\}_{i=1}^K$
Sample data from datasets $\{S_j\}_{j=1,j\neq i}^K$
来自数据集 $\{S_j\}_{j=1,j\neq i}^K$ 的示例数据
Update θ by optimizing the forth term of Eq. 10.
通过优化方程 10 的第四项来更新 θ 。
end
结束
end
结束

Discussion. In comparison with the typical classification loss, our entropy regularization loss can push the network to learn domain-invariant features. For instance, in the example of simulated data in Figure 1, the summation of the classification loss, the regularization loss and the domain adversarial loss is -0.16 in classifying the first dimension, and is -0.02 in classifying the second dimension. Therefore, our training objective can enforce the learned features to be domain-invariant.

讨论。与典型的分类损失相比，我们的熵正则化损失可以推动网络学习领域不变特征。例如，在图 1 中模拟数据的例子中，分类损失、正则化损失和领域对抗损失的总和在分类第一维时为 -0.16，在分类第二维时为 -0.02。因此，我们的训练目标可以强制学习到的特征具有领域不变性。

3 Experiments

3 实验

In this section, we study domain generalization on four datasets, including two simulated datasets (i.e., Rotated MNIST [11] and Rotated CIFAR-10) and two real-world datasets (i.e., VLCS [11], PACS [28]). We make comparisons against state-of-the-art methods to demonstrate the effectiveness of the proposed algorithm. We conduct extensive ablations to discuss our method comprehensively.

在本节中，我们研究了四个数据集上的领域泛化，包括两个模拟数据集 (即，旋转 MNIST [11] 和旋转 CIFAR-10) 和两个真实世界数据集 (即，VLCS [11]，PACS [28])。我们与最先进的方法进行比较，以证明所提算法的有效性。我们进行了广泛的消融实验，以全面讨论我们的方法。

Table 1: Results on MNIST dataset with object recognition accuracy (%) averaged over 10 runs.
表 1: 在 MNIST 数据集上以对象识别准确率 (%) 进行的结果，平均经过 10 次运行。

Target	CrossGrad [16]	MetaReg [23]	Reptile [34]	Feature-Critic [30]	DeepAll	Basic-Adv	Ours
M_0	86.03	85.70	87.78	87.04	88.37 ± 1.19	88.88 ± 1.08	90.09 ± 1.25
M_{15}	98.92	98.87	99.44	99.53	99.13 ± 0.41	99.10 ± 0.19	99.24 ± 0.37
M_{30}	98.60	98.32	98.42	99.41	99.28 ± 0.27	99.25 ± 0.14	99.27 ± 0.16
M_{45}	98.39	98.58	98.80	99.52	99.09 ± 0.29	99.25 ± 0.17	99.31 ± 0.21
M_{60}	98.68	98.93	99.03	99.23	99.14 ± 0.28	99.16 ± 0.32	99.45 ± 0.19
M_{75}	88.94	89.44	87.42	91.52	87.48 ± 1.01	89.06 ± 1.54	90.81 ± 1.35
Avg.	94.93	94.97	95.15	96.04	95.42	95.78	96.36

目标	CrossGrad [16]	MetaReg [23]	Reptile [34]	Feature-Critic [30]	DeepAll	Basic-Adv	我们的
M_0	86.03	85.70	87.78	87.04	88.37 ± 1.19	88.88 ± 1.08	90.09 ± 1.25
M_{15}	98.92	98.87	99.44	99.53	99.13 ± 0.41	99.10 ± 0.19	99.24 ± 0.37
M_{30}	98.60	98.32	98.42	99.41	99.28 ± 0.27	99.25 ± 0.14	99.27 ± 0.16
M_{45}	98.39	98.58	98.80	99.52	99.09 ± 0.29	99.25 ± 0.17	99.31 ± 0.21
M_{60}	98.68	98.93	99.03	99.23	99.14 ± 0.28	99.16 ± 0.32	99.45 ± 0.19
M_{75}	88.94	89.44	87.42	91.52	87.48 ± 1.01	89.06 ± 1.54	90.81 ± 1.35
平均值	94.93	94.97	95.15	96.04	95.42	95.78	96.36

Table 2: Results on CIFAR-10 dataset with object recognition accuracy (%) averaged over 5 runs.
表 2:CIFAR-10 数据集上对象识别准确率 (%) 的结果，平均值基于 5 次运行。

Method	M0	M15	M30	M45	M60	M75	$Avg.$
DeepAll	71.28 ± 1.59	97.94 ± 0.32	99.14 ± 0.04	99.06 ± 0.19	99.07 ± 0.40	76.59 ± 0.89	90.51
Basic-Adv	75.85 ± 1.45	99.03 ± 0.18	99.16 ± 0.06	99.14 ± 0.11	99.29 ± 0.13	81.14 ± 1.34	92.27
Ours	77.91 ± 0.83	$\mathbf{99.05 \pm 0.22}$	99.33 ± 0.09	99.39 ± 0.14	99.40 ± 0.29	80.12 ± 0.60	92.53

方法	M0	M15	M30	M45	M60	M75	$Avg.$
DeepAll	71.28 ± 1.59	97.94 ± 0.32	99.14 ± 0.04	99.06 ± 0.19	99.07 ± 0.40	76.59 ± 0.89	90.51
基础-高级	75.85 ± 1.45	99.03 ± 0.18	99.16 ± 0.06	99.14 ± 0.11	99.29 ± 0.13	81.14 ± 1.34	92.27
我们的方法	77.91 ± 0.83	$\mathbf{99.05 \pm 0.22}$	99.33 ± 0.09	99.39 ± 0.14	99.40 ± 0.29	80.12 ± 0.60	92.53

3.1 Simulated Datasets

3.1 模拟数据集

Rotated MNIST. Following the setting in [11], we first randomly choose 100 samples per category (1000 in total) from the original dataset [29] to form the domain M_0 . Then, we create 5 rotating domains $\{M_{15}, M_{30}, M_{45}, M_{60}, M_{75}\}$ by rotating each image in M_0 five times with 15 degrees intervals in clockwise direction. As done by previous works [30, 16], we conduct leave-one-domain-out experiments by selecting one domain to hold out as the target. For fair comparisons, we exploit the standard MNIST CNN, where the feature network consists of two convolutional layers and one fully-connected (FC) layer, and the classifier has one FC layer. We train our model with the learning rate of $1e-4$ (F, T , and D) , and $1e-5$ $\left(\{T_i, T_i'\}_{i=1}^{5}\right)$ for 3,000 iterations. We set the weighting factors to $0.5\,(\alpha_1), 0.005\,(\alpha_2)$, and $0.01\,(\alpha_3)$, respectively. We repeat all of the experiments 10 times, and report the average mean and standard deviation of recognition accuracy in Table 1.

旋转 MNIST。根据 [11] 中的设置，我们首先从原始数据集 [29] 中随机选择每个类别 100 个样本 (共 1000 个) 以形成域 M_0 。然后，我们通过将 M_0 中的每个图像顺时针旋转 15 度间隔五次来创建 5 个旋转域 $\{M_{15}, M_{30}, M_{45}, M_{60}, M_{75}\}$ 。如之前的研究 [30, 16] 所做，我们进行留一域实验，选择一个域作为目标域。为了公平比较，我们利用标准的 MNIST CNN，其中特征网络由两个卷积层和一个全连接 (FC) 层组成，分类器有一个 FC 层。我们以学习率 $1e-4$ (F, T , and D) 和 $1e-5$ $\left(\{T_i, T_i'\}_{i=1}^{5}\right)$ 训练模型，迭代 3000 次。我们将权重因子分别设置为 $0.5\,(\alpha_1), 0.005\,(\alpha_2)$ 和 $0.01\,(\alpha_3)$ 。我们重复所有实验 10 次，并在表 1 中报告识别准确率的平均值和标准偏差。

Rotated CIFAR-10. We randomly choose 500 samples per category (5000 in total) from the original CIFAR-10 dataset [31], and then create additional 5 domains using the same strategy as stated in Rotated MNIST. We use AlexNet [32] as our backbone network. In specific, the feature extractor F consists of the top layers of AlexNet model till the POOL5 layer, while T contains FC6, FC7, and an additional FC layer. For $\{T_i, T_i'\}_{i=1}^{5}$ and D , we use a similar architecture to T . We train the whole network from scratch with the learning rate of $1e-3$ (F, T , and D) and $1e-4$ $\left(\{T_i, T_i'\}_{i=1}^{5}\right)$ using the Adam optimizer [33] for 2000 iterations. The weighting factors $(\alpha_1, \alpha_2, \alpha_3)$ are set to0.5,0.001, and 0.1, respectively. We repeat all experiments 5 times, and provide the results in Table 2.

旋转的 CIFAR-10。我们从原始 CIFAR-10 数据集中随机选择每个类别 500 个样本 (共 5000 个)，然后使用与旋转 MNIST 中所述相同的策略创建额外的 5 个领域。我们使用 AlexNet 作为我们的主干网络。具体而言，特征提取器 F 由 AlexNet 模型的顶层构成，直到 POOL5 层，而 T 包含 FC6、FC7 和一个额外的 FC 层。对于 $\{T_i, T_i'\}_{i=1}^{5}$ 和 D ，我们使用与 T 相似的架构。我们从头开始训练整个网络，学习率为 $1e-3$ (F, T , and D) 和 $1e-4$ $\left(\{T_i, T_i'\}_{i=1}^{5}\right)$ ，使用 Adam 优化器 [33] 进行 2000 次迭代。权重因子 $(\alpha_1, \alpha_2, \alpha_3)$ 分别设置为 0.5、0.001 和 0.1。我们重复所有实验 5 次，并在表 2 中提供结果。

Results. We make comparisons against several recent works, e.g., CrossGrad [16], MetaReg [23], Reptile [34], and Feature-Critic [30], on Rotated MNIST. To better illustrate the generalization capabilities of our model, we also evaluate the performance of two additional models, i.e., DeepAll and Basic-Adv, on both Rotated MNIST and Rotated CIFAR-10. DeepAll trains F and T on all of the source domains without performing any domain generalization (Eq. 1), while Basic-Adv is the basic solution through adversarial learning (Eq. 1 and Eq. 2). We can find all of the algorithms perform well on Rotated MNIST from Table 1, which means the generated domains have similar distributions. Nevertheless, our approach still performs better than existing approaches. Furthermore, the higher accuracy compared with DeepAll and Basic-Adv on both Rotated MNIST and Rotated CIFAR-10 shows the better generalization capabilities of the proposed algorithm.

结果。我们对几个近期的工作进行了比较，例如 CrossGrad [16]、MetaReg [23]、Reptile [34] 和 Feature-Critic [30]，均在旋转 MNIST 上进行评估。为了更好地说明我们模型的泛化能力，我们还评估了两个额外模型的性能，即 DeepAll 和 Basic-Adv，分别在旋转 MNIST 和旋转 CIFAR-10 上进行评估。DeepAll 在所有源域上进行训练 F 和 T ，而不执行任何领域泛化 (公式 1)，而 Basic-Adv 是通过对抗学习 (公式 1 和公式 2) 得到的基本解决方案。我们可以从表 1 中发现，所有算法在旋转 MNIST 上表现良好，这意味着生成的域具有相似的分布。然而，我们的方法仍然优于现有的方法。此外，与 DeepAll 和 Basic-Adv 在旋转 MNIST 和旋转 CIFAR-10 上的较高准确率显示了所提出算法更好的泛化能力。

3.2 Real-World Datasets

3.2 真实世界数据集

VLCS. VLCS [11] contains images from four well-known datasets, i.e., Pascal VOC2007 (V) [37], LabelMe (L) [38], Caltech (C) [39], and SUN09 (S) [40]. There are five categories, including bird, car, chair, dog, and person. Following previous works [11,22,21], we randomly split each domain data into training (70%) and test (30%) sets, and do the leave-one-out evaluation. For the configuration of the network, we consider two cases, i.e., MLP and E2E. In specific, in MLP, we use the pre-extracted DeCAF6 features (4096-dimensional vector) as the input, and F consists of two FC layers with latent

VLCS。VLCS [11] 包含来自四个著名数据集的图像，即 Pascal VOC2007 (V) [37]、LabelMe (L) [38]、Caltech (C) [39] 和 SUN09 (S) [40]。共有五个类别，包括鸟、车、椅子、狗和人。遵循之前的工作 [11,22,21]，我们随机将每个域的数据分为训练 (70%) 和测试 (30%) 集，并进行留一法评估。对于网络的配置，我们考虑两种情况，即 MLP 和 E2E。具体而言，在 MLP 中，我们使用预提取的 DeCAF6 特征 (4096 维向量) 作为输入，并且 F 由两个全连接层和潜在层组成。

Table 3: Results on VLCS dataset with object recognition accuracy (%) averaged over 20 runs.

表 3: 在 VLCS 数据集上的结果，物体识别准确率 (%) 在 20 次运行中取平均。

Method	Pascal VOC2007	LabelMe	Caltech	SUN09	Average
MLP					
D-MATE [11]	63.90	60.13	89.05	61.33	68.60
DBADG [28]	65.58	58.74	92.43	61.85	69.65
CCSA [35]	67.10	62.10	92.30	59.10	70.15
MetaReg [23]	65.00	60.20	92.30	64.20	70.43
CrossGrad [16]	65.50	60.00	92.00	64.70	70.55
DANN [36]	66.40	64.00	92.60	63.60	71.65
MMD-AAE [12]	67.70	62.60	94.40	64.40	72.28
MLDG [24]	67.70	61.30	94.40	65.90	72.33
Epi-FCR [22]	67.10	64.30	94.10	65.90	72.85
DeepAll	70.07 ± 0.79	60.54 ± 1.02	93.83 ± 1.08	65.95 ± 1.13	72.60
Basic-Adv	70.47 ± 0.59	60.94 ± 0.94	93.84 ± 1.00	66.05 ± 0.91	72.82
Ours	70.54 ± 0.55	60.81 ± 1.38	94.44 ± 0.98	66.11 ± 0.75	72.97
E2E					
DBADG [28]	69.99	63.49	93.64	61.32	72.11
JiGen [18]	70.62	60.90	96.93	64.30	73.19
MMLD [15]	71.96	58.77	96.66	68.13	73.88
CIDDG [14]	73.00	58.30	97.02	68.89	74.30
DeepAll	73.11 ± 0.67	58.07 ± 0.52	97.15 ± 0.40	68.79 ± 0.44	74.28
Basic-Adv	72.79 ± 0.67	58.53 ± 0.69	97.00 ± 0.50	68.70 ± 0.69	74.26
Ours	73.24 ± 0.49	58.26 ± 0.82	96.92 ± 0.40	69.10 ± 0.46	74.38

方法	Pascal VOC2007	LabelMe	加州理工学院	SUN09	平均
多层感知器					
D-MATE [11]	63.90	60.13	89.05	61.33	68.60
DBADG [28]	65.58	58.74	92.43	61.85	69.65
CCSA [35]	67.10	62.10	92.30	59.10	70.15
MetaReg [23]	65.00	60.20	92.30	64.20	70.43
CrossGrad [16]	65.50	60.00	92.00	64.70	70.55
DANN [36]	66.40	64.00	92.60	63.60	71.65
MMD-AAE [12]	67.70	62.60	94.40	64.40	72.28
MLDG [24]	67.70	61.30	94.40	65.90	72.33
Epi-FCR [22]	67.10	64.30	94.10	65.90	72.85
DeepAll	70.07 ± 0.79	60.54 ± 1.02	93.83 ± 1.08	65.95 ± 1.13	72.60
Basic-Adv	70.47 ± 0.59	60.94 ± 0.94	93.84 ± 1.00	66.05 ± 0.91	72.82
我们的方法	70.54 ± 0.55	60.81 ± 1.38	94.44 ± 0.98	66.11 ± 0.75	72.97
E2E					
DBADG [28]	69.99	63.49	93.64	61.32	72.11
JiGen [18]	70.62	60.90	96.93	64.30	73.19
MMLD [15]	71.96	58.77	96.66	68.13	73.88
CIDDG [14]	73.00	58.30	97.02	68.89	74.30
DeepAll	73.11 ± 0.67	58.07 ± 0.52	97.15 ± 0.40	68.79 ± 0.44	74.28
Basic-Adv	72.79 ± 0.67	58.53 ± 0.69	97.00 ± 0.50	68.70 ± 0.69	74.26
我们的方法	73.24 ± 0.49	58.26 ± 0.82	96.92 ± 0.40	69.10 ± 0.46	74.38

Table 4: Results on PACS dataset with object recognition accuracy (%) averaged over 5 runs.
表 4: 在 PACS 数据集上的结果，物体识别准确率 (%) 在 5 次运行中取平均。

Method	Art Painting	Cartoon	Photo	Sketch	Average
D-MATE [11]	60.27	58.65	91.12	47.68	64.48
CrossGrad [16]	61.00	67.20	87.60	55.90	67.93
DBADG [28]	62.86	66.97	89.50	57.51	69.21
MLDG [24]	66.23	66.88	88.00	58.96	70.01
Epi-FCR [22]	64.70	72.30	86.10	65.00	72.03
Feature-Critic [30]	64.89	71.72	89.94	61.85	71.20
CIDDG [14]	66.99	68.62	90.19	62.88	72.20
MetaReg [23]	69.82	70.35	91.07	59.26	72.62
JiGen [18]	67.63	71.71	89.00	65.18	73.38
MMLD [15]	69.27	72.83	88.98	66.44	74.38
MASF [21]	70.35	72.46	90.68	67.33	75.21
DeepAll	68.35 ± 0.80	70.14 ± 0.87	90.83 ± 0.32	64.98 ± 1.92	73.57
Basic-Adv	71.34 ± 0.81	70.11 ± 1.18	88.86 ± 0.50	70.91 ± 0.94	75.31
Ours	71.34 ± 0.87	70.29 ± 0.77	89.92 ± 0.42	71.15 ± 1.01	75.67

方法	艺术绘画	卡通	照片	草图	平均
D-MATE [11]	60.27	58.65	91.12	47.68	64.48
CrossGrad [16]	61.00	67.20	87.60	55.90	67.93
DBADG [28]	62.86	66.97	89.50	57.51	69.21
MLDG [24]	66.23	66.88	88.00	58.96	70.01
Epi-FCR [22]	64.70	72.30	86.10	65.00	72.03
Feature-Critic [30]	64.89	71.72	89.94	61.85	71.20
CIDDG [14]	66.99	68.62	90.19	62.88	72.20
MetaReg [23]	69.82	70.35	91.07	59.26	72.62
JiGen [18]	67.63	71.71	89.00	65.18	73.38
MMLD [15]	69.27	72.83	88.98	66.44	74.38
MASF [21]	70.35	72.46	90.68	67.33	75.21
DeepAll	68.35 ± 0.80	70.14 ± 0.87	90.83 ± 0.32	64.98 ± 1.92	73.57
基础-进阶	71.34 ± 0.81	70.11 ± 1.18	88.86 ± 0.50	70.91 ± 0.94	75.31
我们的方法	71.34 ± 0.87	70.29 ± 0.77	89.92 ± 0.42	71.15 ± 1.01	75.67

dimensions of 1024 and 128. For the classifiers T and $\{T_i, T_i'\}_{i=1}^3$, we use one FC layer, respectively. For the discriminator D , we utilize three FC layers with the output dimensions of 128,64, and 3 (the

number of source domains). In this case, we train our model with the learning rate of $1e-3$ for 30 epochs using the SGD optimizer. We set all trade-off parameters to 0.1 . In another setting (E2E), we employ the same network configuration as used on Rotated CIFAR-10, but use the model pre-trained on ImageNet [32]. We set the learning rate to $1e-4$, and the weighting factors α_1, α_2 , and α_3 to 0.1, 0.001, and 0.05, respectively. We train the model with the batch size of 64 for each source domain for 60 epochs and repeat all of the experiments 20 times.

维度为 1024 和 128。对于分类器 T 和 $\{T_i, T_i'\}_{i=1}^3$ ，我们分别使用一个全连接层。对于鉴别器 D ，我们利用三个全连接层，输出维度为 128、64 和 3(源域的数量)。在这种情况下，我们使用学习率 $1e-3$ 训练模型 30 个周期，采用 SGD 优化器。我们将所有权衡参数设置为 0.1。在另一种设置 (E2E) 中，我们采用与在旋转 CIFAR-10 上使用的相同网络配置，但使用在 ImageNet [32] 上预训练的模型。我们将学习率设置为 $1e-4$ ，权重因子 α_1, α_2 和 α_3 分别设置为 0.1、0.001 和 0.05。我们对每个源域以批量大小 64 训练模型 60 个周期，并将所有实验重复 20 次。

PACS. PACS [28] is proposed specially for domain generalization. It contains four domains, i.e., Photo (P), Art Painting (A), Cartoon (C), and Sketch (S), and seven categories: dog, elephant, giraffe, guitar, house, horse, and person. For a fair comparison, we use the same training and validation split as presented in [28]. Our network configuration is the same as that used for VLCS (E2E), and we set the weighting factors to $0.5\,(\alpha_1), 0.01\,(\alpha_2)$, and $0.05\,(\alpha_3)$, respectively. Then we train the model with the learning rate of $1e-3\,(F, T, D)$ and $1e-4\left(\{T_i, T_i'\}_{i=1}^3\right)$ for 60 epochs. We repeat all experiments 5 times, and report the results in Tabel 4.

PACS。PACS [28] 是专门为领域泛化提出的。它包含四个领域，即照片 (P)、艺术画 (A)、卡通 (C) 和素描 (S)，以及七个类别: 狗、大象、长颈鹿、吉他、房子、马和人。为了公平比较，我们使用与 [28] 中相同的训练和验证划分。我们的网络配置与用于 VLCS(E2E) 的配置相同，我们将权重因子设置为 $0.5\,(\alpha_1), 0.01\,(\alpha_2)$ 和 $0.05\,(\alpha_3)$ 。然后，我们以 $1e-3\,(F, T, D)$ 和 $1e-4\left(\{T_i, T_i'\}_{i=1}^3\right)$ 的学习率训练模型，训练 60 个周期。我们重复所有实验 5 次，并在表 4 中报告结果。

Results. As shown in Table 3, although the baselines (DeepAll and Basic-Adv) are competitive with previous methods in both cases (MLP and E2E), our proposed entropy regularization still improves the performance further on VLCS. Furthermore, the highest average score and the highest score on several domains of PACS can also demonstrate the effectiveness of our approach. For example, Table 4 shows that our method improves the average accuracy by 2.1% on PACS over DeepAll, and improves 6.17% and 2.99% on Sketch and Art Painting, respectively. In addition, from the results in Table 3 and Table 4, we can observe that the performance (Ours v.s. DeepAll and Basic-Adv v.s. DeepAll) gains obtained by our regularization policy on PACS are more notable than those on VLCS. A possible reason we guess is that only one domain (C) in VLCS is object-centric, while others are

结果。如表 3 所示，尽管基线 (DeepAll 和 Basic-Adv) 在两种情况下 (MLP 和 E2E) 与以前的方法具有竞争力，但我们提出的熵正则化仍然进一步提高了 VLCS 的性能。此外，PACS 中的最高平均分和多个领域的最高分也可以证明我们方法的有效性。例如，表 4 显示我们的方法在 PACS 上相对于 DeepAll 提高了平均准确率 2.1% ，并在素描和艺术画上分别提高了 6.17% 和 2.99% 。此外，从表 3 和表 4 的结果中，我们可以观察到我们在 PACS 上的正则化策略所获得的性能 (我们与 DeepAll 以及 Basic-Adv 与 DeepAll 的比较) 比在 VLCS 上更为显著。我们猜测的一个可能原因是，VLCS 中只有一个领域 (C) 是以对象为中心的，而其他领域则不是。

Table 5: Results with different weighting factors on PACS.

表 5: 在 PACS 上使用不同权重因子的结果。

$\alpha_1, \alpha_2, \alpha_3$	Art Painting	Cartoon	Photo	Sketch	Average
- , - , -	68.35 ± 0.80	70.14 ± 0.87	90.83 ± 0.32	64.98 ± 1.92	73.57
$1.0, -, -$	64.46 ± 3.80	64.07 ± 3.01	83.48 ± 1.39	66.70 ± 2.64	69.68
$0.5, -, -$	71.35 ± 0.81	70.11 ± 1.18	88.86 ± 0.50	70.91 ± 0.94	75.31
$0.1, -, -$	68.22 ± 0.89	70.13 ± 0.67	90.60 ± 0.37	64.61 ± 1.93	73.39
$0.5, 0.05, -$	70.83 ± 1.35	70.06 ± 0.98	89.25 ± 0.38	71.34 ± 0.82	75.37
0.5 , 0.01 , -	71.05 ± 1.62	70.29 ± 0.88	89.44 ± 0.36	70.06 ± 1.80	75.21
0.5 , 0.001 , -	71.72 ± 0.77	69.84 ± 1.65	88.88 ± 0.42	70.85 ± 0.83	75.32
$0.5, -, 0.5$	68.92 ± 0.59	69.62 ± 0.51	89.99 ± 0.38	70.04 ± 0.63	74.74
0.5 , - , 0.1	71.04 ± 0.96	69.78 ± 0.98	89.68 ± 0.51	70.95 ± 0.81	75.36
0.5 , - , 0.05	71.59 ± 1.01	68.97 ± 1.42	89.57 ± 0.23	69.81 ± 3.45	74.99
0.5,0.05,0.1	71.09 ± 1.10	69.55 ± 0.54	89.56 ± 0.33	71.31 ± 0.90	75.37
0.5,0.01,0.1	70.91 ± 0.81	70.05 ± 1.33	89.80 ± 0.44	71.46 ± 0.46	75.56
0.5 , 0.005 , 0.1	70.95 ± 0.77	69.78 ± 0.91	89.56 ± 0.64	71.00 ± 1.12	75.32
0.5 , 0.05 , 0.05	70.55 ± 1.17	69.57 ± 1.14	89.33 ± 0.55	70.40 ± 2.88	74.96
0.5 , 0.01 , 0.05	71.34 ± 0.87	70.29 ± 0.77	89.92 ± 0.42	71.15 ± 1.02	75.67
0.5 , 0.005 , 0.05	70.51 ± 2.26	69.60 ± 0.58	89.69 ± 0.39	71.51 ± 0.84	75.33

$\alpha_1, \alpha_2, \alpha_3$	艺术绘画	卡通	照片	草图	平均值
- , - , -	68.35 ± 0.80	70.14 ± 0.87	90.83 ± 0.32	64.98 ± 1.92	73.57
$1.0, -, -$	64.46 ± 3.80	64.07 ± 3.01	83.48 ± 1.39	66.70 ± 2.64	69.68
$0.5, -, -$	71.35 ± 0.81	70.11 ± 1.18	88.86 ± 0.50	70.91 ± 0.94	75.31
$0.1, -, -$	68.22 ± 0.89	70.13 ± 0.67	90.60 ± 0.37	64.61 ± 1.93	73.39
$0.5, 0.05, -$	70.83 ± 1.35	70.06 ± 0.98	89.25 ± 0.38	71.34 ± 0.82	75.37
0.5 , 0.01 , -	71.05 ± 1.62	70.29 ± 0.88	89.44 ± 0.36	70.06 ± 1.80	75.21
0.5 , 0.001 , -	71.72 ± 0.77	69.84 ± 1.65	88.88 ± 0.42	70.85 ± 0.83	75.32
$0.5, -, 0.5$	68.92 ± 0.59	69.62 ± 0.51	89.99 ± 0.38	70.04 ± 0.63	74.74
0.5 , - , 0.1	71.04 ± 0.96	69.78 ± 0.98	89.68 ± 0.51	70.95 ± 0.81	75.36
0.5 , - , 0.05	71.59 ± 1.01	68.97 ± 1.42	89.57 ± 0.23	69.81 ± 3.45	74.99
0.5,0.05,0.1	71.09 ± 1.10	69.55 ± 0.54	89.56 ± 0.33	71.31 ± 0.90	75.37
0.5,0.01,0.1	70.91 ± 0.81	70.05 ± 1.33	89.80 ± 0.44	71.46 ± 0.46	75.56
0.5 , 0.005 , 0.1	70.95 ± 0.77	69.78 ± 0.91	89.56 ± 0.64	71.00 ± 1.12	75.32
0.5 , 0.05 , 0.05	70.55 ± 1.17	69.57 ± 1.14	89.33 ± 0.55	70.40 ± 2.88	74.96
0.5 , 0.01 , 0.05	71.34 ± 0.87	70.29 ± 0.77	89.92 ± 0.42	71.15 ± 1.02	75.67
0.5 , 0.005 , 0.05	70.51 ± 2.26	69.60 ± 0.58	89.69 ± 0.39	71.51 ± 0.84	75.33

Table 6: Results of deeper networks on PACS dataset with object recognition accuracy (%) averaged over 5 runs.

表 6: 在 PACS 数据集上更深网络的结果，基于 5 次运行的对象识别准确率 (%) 的平均值。

Method	Art Painting	Cartoon	Photo	Sketch	Average
ResNet-18					
DeepAll	78.93 ± 0.46	75.02 ± 0.89	96.60 ± 0.16	70.48 ± 0.84	80.25
Basic-Adv	80.54 ± 1.71	75.21 ± 0.92	96.67 ± 0.21	70.65 ± 1.91	80.77
Ours	80.70 ± 0.71	76.40 ± 0.34	96.65 ± 0.21	71.77 ± 1.27	81.38
ResNet-50					
DeepAll	86.18 ± 0.34	76.79 ± 0.33	98.14 ± 0.15	74.66 ± 0.93	83.94
Basic-Adv	87.11 ± 1.08	78.65 ± 1.13	98.22 ± 0.17	76.48 ± 1.09	85.11
Ours	87.51 ± 1.03	79.31 ± 1.40	98.25 ± 0.12	76.30 ± 0.65	85.34

方法	艺术绘画	卡通	照片	草图	平均
ResNet-18					
DeepAll	78.93 ± 0.46	75.02 ± 0.89	96.60 ± 0.16	70.48 ± 0.84	80.25
基础-进阶	80.54 ± 1.71	75.21 ± 0.92	96.67 ± 0.21	70.65 ± 1.91	80.77
我们的方法	80.70 ± 0.71	76.40 ± 0.34	96.65 ± 0.21	71.77 ± 1.27	81.38
ResNet-50					
DeepAll	86.18 ± 0.34	76.79 ± 0.33	98.14 ± 0.15	74.66 ± 0.93	83.94
基础-进阶	87.11 ± 1.08	78.65 ± 1.13	98.22 ± 0.17	76.48 ± 1.09	85.11
我们的方法	87.51 ± 1.03	79.31 ± 1.40	98.25 ± 0.12	76.30 ± 0.65	85.34

all scene-centric. This makes the generalization of the model difficult, although the domain shifts in VLCS are small [28]. In contrast, the images in all domains of PACS are mostly object-centric, and objects in different domains mainly have different styles and shapes. This can better evaluate the generalization capabilities of the model.

所有场景都是以场景为中心的。这使得模型的泛化变得困难，尽管 VLCS 中的领域转移很小 [28]。相比之下，PACS 所有领域中的图像主要是以对象为中心的，不同领域中的对象主要具有不同的风格和形状。这可以更好地评估模型的泛化能力。

3.3 Ablation Studies

3.3 消融研究

The experimental results above have demonstrated the effectiveness of our proposed algorithm for domain generalization. Here, we provide the ablation studies on the designed loss and network backbone to analyze the contributions of the proposed entropy regularization further.

上述实验结果已经证明了我们提出的算法在领域泛化方面的有效性。在这里，我们提供了关于设计的损失和网络骨干的消融研究，以进一步分析所提出的熵正则化的贡献。

Different Weighting Factors. We conduct various experiments with different weighting factors on PACS to examine their impacts. We report the average accuracy of 5 trials in Table 5. The results marked by the "gray" color correspond to the results reported in Table 4. "-" means the corresponding loss term is ignored. As shown in Table 5, in most cases, our proposed conditional entropy regularization ($\alpha_2 \neq 0$) can yield some improvements. Besides, by optimizing the full objective, our approach can further improve the generalization capabilities of the model.

不同的权重因子。我们在 PACS 上进行各种实验，使用不同的权重因子来检查它们的影响。我们在表 5 中报告了 5 次试验的平均准确率。用“灰色”标记的结果对应于表 4 中报告的结果。“-”表示相应的损失项被忽略。如表 5 所示，在大多数情况下，我们提出的条件熵正则化 ($\alpha_2 \neq 0$) 可以带来一些改进。此外，通过优化完整目标，我们的方法可以进一步提高模型的泛化能力。

Deeper Networks. We further study the generalization capabilities of our model by taking deeper networks, e.g., ResNet-18 and RestNet-50 [41], as the backbone network. The models are pre-trained on ImageNet, and fine-tuned on PACS using the proposed loss. In specific, we take the last FC layer as our task network T , and other layers as the feature extractor F . We use three FC layers with output dimensions of 1024,256, and the number of source domains / categories to construct the discriminator D and classifiers $\{T_i, T_i'\}_{i=1}^3$, respectively. For both ResNet-18 and ResNet-50, we use the same hyper-parameters, i.e., $\alpha_1 = 0.1, \alpha_2 = 0.001, \alpha_3 = 0.05$, and the learning rate of $1e-3\,(F, T, D)$ and $1e-4\left(\{T_i, T_i'\}_{i=1}^3\right)$. We learn models for 100 epochs, and report the average scores of 5 trials. As shown in Table 6, even though we take deeper networks as our backbones, our approach still yield higher scores than the two baselines.

更深的网络。我们通过采用更深的网络，例如 ResNet-18 和 ResNet-50 [41]，作为主干网络，进一步研究我们模型的泛化能力。模型在 ImageNet 上进行预训练，并在 PACS 上使用提出的损失进行微调。具体而言，我们将最后的全连接层作为我们的任务网络 T ，其他层作为特征提取器 F 。我们使用三个全连接层，输出维度分别为 1024、256 和源域/类别的数量，以构建鉴别器 D 和分类器 $\{T_i, T_i'\}_{i=1}^3$ 。对于 ResNet-18 和 ResNet-50，我们使用相同的超参数，即 $\alpha_1 = 0.1, \alpha_2 = 0.001, \alpha_3 = 0.05$ ，学习率为 $1e-3\,(F, T, D)$ 和 $1e-4\left(\{T_i, T_i'\}_{i=1}^3\right)$ 。我们训练模型 100 个周期，并报告 5 次试验的平均分数。如表 6 所示，尽管我们采用更深的网络作为主干，但我们的方法仍然比两个基线模型获得更高的分数。

Class Imbalance. We address the class imbalance issue by using the weighted cross-entropy loss according to the number of each class in each batch. If not using the weighted loss i.e., setting the weight to 1 for each class, the model yields a lower average accuracy of 75.58% (weighted loss used: 75.67%) on PACS, but still has better generalization capabilities.

类别不平衡。我们通过根据每个批次中每个类别的数量使用加权交叉熵损失来解决类别不平衡问题。如果不使用加权损失，即将每个类别的权重设置为 1，则模型在 PACS 上的平均准确率较低，为 75.58% (使用加权损失时:75.67%)，但仍具有更好的泛化能力。

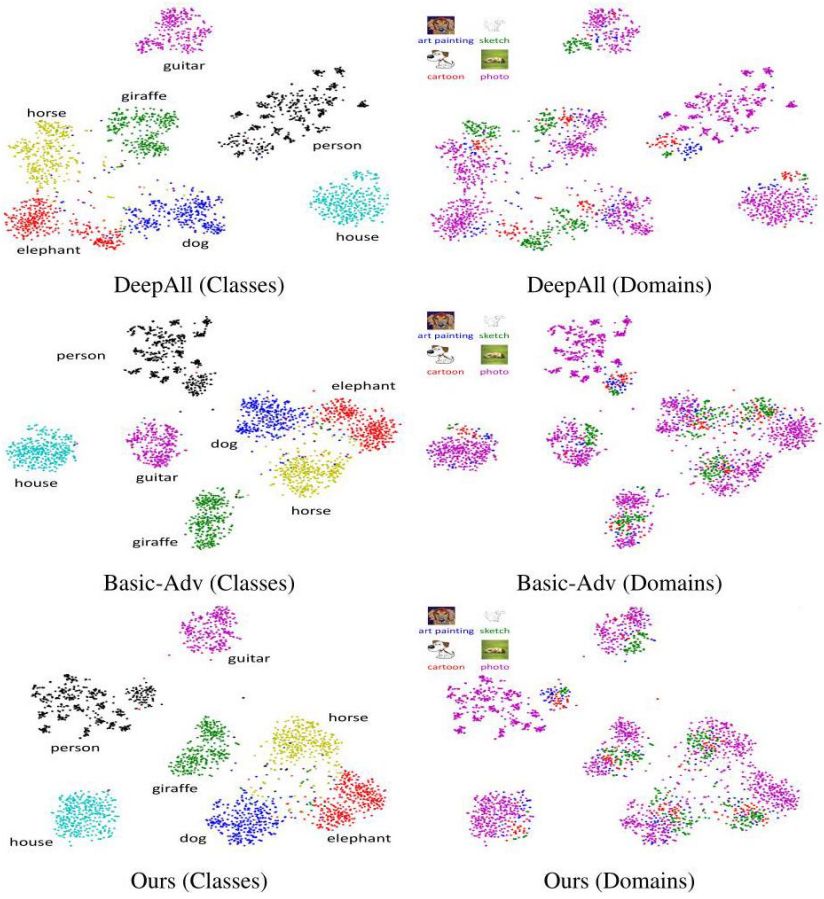

Figure 2: Feature visualization. Left: different colors represent different classes; Right: different colors indicate different domains (Target: Photo). Best viewed in color (Zoom in for details).

图 2: 特征可视化。左: 不同颜色代表不同类别；右: 不同颜色表示不同域 (目标: 照片)。最佳效果为彩色显示 (放大以查看细节)。

Feature Visualization. To better understand the distribution of the learned features, we exploit t-SNE [42] to analyze the feature space learned by DeepAll, Basic-Adv, and Ours, respectively. We conduct this study on PACS, and in specific, we take the Photo dataset as the target, and others as the source. As shown in Figure 2, both Ours and Basic-Adv are capable of minimizing the distance between the distributions of the domains. For example, in DeepAll (Domains), we can observe that the Sketch (Green) is far away from other domains, while in Ours (Domains) and Basic-Adv (Domain), domains are clustered better. Furthermore, the comparison between Ours (Classes, Domains) and Basic-Adv (Classes, Domains) can show that our approach also discriminates the data from different categories better than Basic-Adv.

特征可视化。为了更好地理解学习特征的分布，我们利用 t-SNE [42] 分析 DeepAll、Basic-Adv 和我们的模型所学习的特征空间。我们在 PACS 上进行这项研究，具体来说，我们将 Photo 数据集作为目标，其它数据集作为源。如图 2 所示，我们的模型和 Basic-Adv 都能够最小化领域之间分布的距离。例如，在 DeepAll(领域) 中，我们可以观察到 Sketch(绿色) 与其他领域相距较远，而在我们的模型 (领域) 和 Basic-Adv(领域) 中，领域的聚类效果更好。此外，我们的模型 (类别，领域) 与 Basic-Adv(类别，领域) 之间的比较表明，我们的方法在区分不同类别的数据方面优于 Basic-Adv。

4 Conclusion

4 结论

In this paper, we aim at learning the domain-invariant conditional distribution, which the basic adversarial learning based solutions cannot reach. We analyze the issues existed in related works, and propose an entropy regularization term, i.e., the conditional entropy $H(Y \mid F(X))$, as the remedy. Our approach

can produce domain-invariant features by optimizing the proposed regularization term coupled with the cross-entropy loss and the domain adversarial loss, and thus has a better generalization capability. The experimental results on both simulated and real-world datasets demonstrate the effectiveness of our proposed method. In the future, we can extend our approach to other challenging tasks, like semantic segmentation.

在本文中，我们旨在学习领域不变的条件分布，而基本的对抗学习解决方案无法达到这一目标。我们分析了相关工作中存在的问题，并提出了一种熵正则化项，即条件熵 $H(Y \mid F(X))$ ，作为补救措施。我们的方法通过优化所提出的正则化项，结合交叉熵损失和领域对抗损失，能够生成领域不变的特征，从而具有更好的泛化能力。在模拟和真实世界数据集上的实验结果证明了我们提出的方法的有效性。在未来，我们可以将我们的方法扩展到其他具有挑战性的任务，如语义分割。

5 Acknowledgement

5 致谢

This research was supported by Australian Research Council Projects FL-170100117, DP-180103424, IH-180100002, and DE190101473.

本研究得到了澳大利亚研究委员会项目 FL-170100117、DP-180103424、IH-180100002 和 DE190101473 的支持。

Broader Impact

更广泛的影响

Model generalization is a significant subject, since it is almost impossible for us to train a model for each scenario. However, due to the domain bias, the model trained on a domain often performs worse on other domains. Through exploiting the domain generalization techniques, we can train a model on the publicly available datasets, and then deploy it on other related scenarios directly or with few adaptations. Therefore, the industries can reduce their costs in repeating training the models. On the other hand, since the model is trained on multiple datasets sampled from different domains, the domain generalization techniques can reduce over-fitting, and thus courage the model generate fair results. Based on our knowledge, our work may not have an adverse impact on ethical aspects and future societal consequences.

模型泛化是一个重要的主题，因为我们几乎不可能为每个场景训练一个模型。然而，由于领域偏差，在一个领域上训练的模型在其他领域上的表现往往较差。通过利用领域泛化技术，我们可以在公开可用的数据集上训练模型，然后直接或通过少量适应将其部署到其他相关场景。因此，行业可以减少重复训练模型的成本。另一方面，由于模型是在从不同领域抽样的多个数据集上训练的，领域泛化技术可以减少过拟合，从而鼓励模型生成公平的结果。根据我们的知识，我们的工作可能不会对伦理方面和未来社会后果产生不利影响。

References

参考文献

[1] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In CVPR 2011, pages 1521-1528. IEEE, 2011.

[2] Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. Rethinking importance weighting for deep learning under distribution shift. arXiv preprint arXiv:2006.04662, 2020.

[3] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In International Conference on Machine Learning, pages 819-827, 2013.

[4] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474, 2014.

[5] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In ICML, 2015.

[6] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In European conference on computer vision, pages 443-450. Springer, 2016.

[7] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In Advances in neural information processing systems, pages 343-351, 2016.

[8] Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. In Advances in neural information processing systems, pages 8559-8570, 2018.

[9] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In Advances in neural information processing systems, pages 2178-2186, 2011.

[10] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In International Conference on Machine Learning, pages 10-18, 2013.

[11] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In Proceedings of the IEEE international conference on computer vision, pages 2551-2559, 2015.

[12] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5400-5409, 2018.

[13] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. arXiv preprint arXiv:1907.02893, 2019.

[14] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV), pages 624-639, 2018.

[15] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In AAAI, 2020.

[16] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In International Conference on Learning Representations (ICLR), 2018.

[17] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In Advances in Neural Information Processing Systems, pages 5334-5344, 2018.

[18] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2229-2238, 2019.

[19] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. arXiv preprint arXiv:1711.07910, 2017.

[20] Aniket Anand Deshmukh, Yunwen Lei, Srinagesh Sharma, Urun Dogan, James W Cutler, and Clayton Scott. A generalization error bound for multi-class domain generalization. arXiv preprint arXiv:1905.10392, 2019.

[21] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In Advances in Neural Information Processing Systems, pages 6447-6458, 2019.

[22] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In Proceedings of the IEEE International Conference on Computer Vision, pages 1446-1455, 2019.

[23] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In Advances in Neural Information Processing Systems, pages 998-1008, 2018.

[24] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In AAAI Conference on Artificial Intelligence, 2018.

[25] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In Proceedings of the IEEE International Conference on Computer Vision, pages 1406-1415, 2019.

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672-2680, 2014.

[27] Mingming Gong, Yanwu Xu, Chunyuan Li, Kun Zhang, and Kayhan Batmanghelich. Twin auxilary classifiers gan. In Advances in Neural Information Processing Systems, pages 1328- 1337, 2019.

[28] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In Proceedings of the IEEE international conference on computer vision, pages 5542-5550, 2017.

[29] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278-2324, 1998.

[30] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalisation. In The Thirty-sixth International Conference on Machine Learning, 2019.

[31] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097-1105, 2012.

[33] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. International Conference on Learning Representations, 12 2014.

[34] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999, 2018.

[35] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In Proceedings of the IEEE International Conference on Computer Vision, pages 5715-5725, 2017.

[36] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. The Journal of Machine Learning Research, 17(1):2096-2030, 2016.

[37] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2):303-338, 2010.

[38] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. International journal of computer vision, 77(1-3):157-173, 2008.

[39] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In 2004 conference on computer vision and pattern recognition workshop, pages 178-178. IEEE, 2004.

[40] Myung Jin Choi, Joseph J Lim, Antonio Torralba, and Alan S Willsky. Exploiting hierarchical context on a large database of object categories. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 129-136. IEEE, 2010.

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770-778, 2016.

[42] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579-2605, 2008.