Domain Invariant Representation Learning with Domain Density Transformations

域不变表示学习与域密度变换

A. Tuan Nguyen
A. Tuan Nguyen
University of Oxford; VinAI Research
牛津大学；VinAI 研究所
Oxford, United Kingdom
英国牛津
tuan@robots.ox.ac.uk
Toan Tran
Toan Tran
VinAI Research
VinAI 研究所
Hanoi, Vietnam
越南河内
v.toantm3@vinai.io
Yarin Gal
Yarin Gal
University of Oxford
牛津大学
Oxford, United Kingdom
英国牛津
yarin@cs.ox.ac.uk
Atilim Gunes Baydin
Atilim Gunes Baydin
University of Oxford
牛津大学
Oxford, United Kingdom
英国牛津
gunes@robots.ox.ac.uk

Abstract

摘要

Domain generalization refers to the problem where we aim to train a model on data from a set of source domains so that the model can generalize to unseen target domains. Naively training a model on the aggregate set of data (pooled from all source domains) has been shown to perform suboptimally, since the information learned by that model might be domain-specific and generalize imperfectly to target domains. To tackle this problem, a predominant domain generalization approach is to learn some domain-invariant information for the prediction task, aiming at a good generalization across domains. In this paper, we propose a theoretically grounded method to learn a domain-invariant representation by enforcing the representation network to be invariant under all transformation functions among domains. We next introduce the use of generative adversarial networks to learn such domain transformations in a possible implementation of our method in practice. We demonstrate the effectiveness of our method on several widely used datasets for the domain generalization problem, on all of which we achieve competitive results with state-of-the-art models.

领域泛化指的是我们旨在在一组源领域的数据上训练模型，以使模型能够泛化到未见过的目标领域。简单地在汇总的数据集 (从所有源领域汇总而来) 上训练模型已被证明效果不佳，因为该模型所学习的信息可能是特定于领域的，并且在目标领域的泛化效果不理想。为了解决这个问题，一种主要的领域泛化方法是学习一些领域不变的信息以用于预测任务，旨在实现跨领域的良好泛化。在本文中，我们提出了一种理论基础的方法，通过强制表示网络在领域之间的所有变换函数下保持不变来学习领域不变的表示。接下来，我们介绍了使用生成对抗网络来学习这种领域变换，以便在实践中实现我们的方法。我们在多

个广泛使用的数据集上展示了我们方法的有效性，这些数据集均为领域泛化问题，我们在所有数据集上都取得了与最先进模型相媲美的结果。

1 Introduction

1 引言

Domain generalization refers to the machine learning scenario where the model is trained on multiple source domains so that it is expected to generalize well to unseen target domains. The key difference between domain generalization [25, 37, 18] and domain adaptation [49, 48, 14, 45] is that, in domain generalization, the learner does not have access to data of the target domain, making the problem much more challenging. One of the most common domain generalization approaches is to learn an invariant representation across domains, aiming at a good generalization performance on target domains. For instance, in the representation learning framework, the prediction function $y = f(x)$, where x is data and y is a label, is obtained as a composition $y = h \circ g(x)$ of a deep representation network $z = g(x)$, where z is a learned representation of data x , and a smaller classifier $y = h(z)$, predicting label y given representation z , both of which are shared across domains. With this framework, we can aim to learn an "invariant" representation z across the source domains with the "hope" of a better generalization to the target domain.

域泛化是指机器学习场景，其中模型在多个源域上进行训练，以期在未见过的目标域上具有良好的泛化能力。域泛化 [25, 37, 18] 和域适应 [49, 48, 14, 45] 之间的关键区别在于，在域泛化中，学习者无法访问目标域的数据，这使得问题变得更加具有挑战性。最常见的域泛化方法之一是学习跨域的不变表示，旨在在目标域上获得良好的泛化性能。例如，在表示学习框架中，预测函数 $y = f(x)$ ，其中 x 是数据，y 是标签，是通过深度表示网络 $z = g(x)$ 的组合 $y = h \circ g(x)$ 获得的，其中 z 是数据 x 的学习表示，而较小的分类器 $y = h(z)$ 则在给定表示 z 的情况下预测标签 y ，这两者在各个域之间是共享的。通过这个框架，我们可以旨在学习跨源域的 "不可变" 表示 z ，以 "期望" 在目标域上获得更好的泛化。

Most existing "domain-invariance"-based methods in domain generalization focus on the marginal distribution alignment [37, 1, 44, 43, 32] , which are still prone to distributional shifts when the conditional data distribution is not stable. In particular, the marginal alignment refers to making the representation distribution $p(z)$ to be the same across domains. This is essential since if $p(z)$ for the target domain is different from that of source domains, the classification network $h(z)$ would face out-of-distribution data at test time. Conditional alignment refers to aligning the conditional distribution of the label given the representation $p(y \mid z)$ to expect that the classification network (trained on the source domains) would give accurate predictions at test time. The formal definitions of these two types of alignment are discussed in Section 3.

现有的大多数基于 "域不变性" 的域泛化方法集中于边际分布对齐 [37, 1, 44, 43, 32] ，当条件数据分布不稳定时，这些方法仍然容易受到分布变化的影响。特别地，边际对齐是指使表示分布 $p(z)$ 在各个域之间保持一致。这是至关重要的，因为如果目标域的 $p(z)$ 与源域的不同，则分类网络 $h(z)$ 在测试时将面临分布外数据。条件对齐是指对齐给定表示 $p(y \mid z)$ 的标签的条件分布，以期望分类网络 (在源域上训练) 在测试时能够给出准确的预测。这两种对齐类型的正式定义将在第 3 节中讨论。

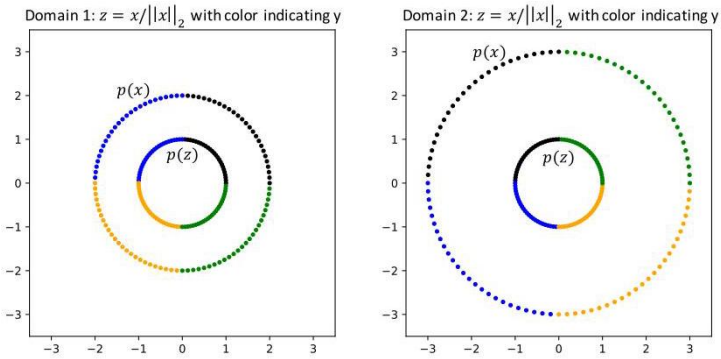


Figure 1: An example of two domains. For each domain, x is uniformly distributed on the outer circle (radius 2 for domain 1 and radius 3 for domain 2), with the color indicating class label y . After the transformation $z = x/ \parallel x\|_2$, the marginal of z is aligned (uniformly distributed on the unit circle for

both domains), but the conditional $p(y \mid z)$ is not aligned. Thus, using this representation for predicting y would not generalize well across domains.

图 1: 两个领域的示例。在每个领域中，x 在外圆上均匀分布 (领域 1 的半径为 2，领域 2 的半径为 3)，颜色表示类别标签 y 。经过变换 $z = x/\parallel x\parallel_2$ 后，z 的边际分布对齐 (在两个领域中均匀分布于单位圆上)，但条件分布 $p(y \mid z)$ 并未对齐。因此，使用该表示法来预测 y 在不同领域之间的泛化效果不佳。

In Figure 1 we illustrate an example where the representation z satisfies the marginal alignment but not the conditional alignment. Specifically, x is distributed uniformly on the circle with radius 2 (and centered at the origin) for domain 1 and distributed uniformly on the circle with radius 3 (centered at the origin) for domain 2 . The representation z defined by the mapping $z = g(x) = x/\parallel x\parallel_2$ will align the marginal distribution $p(z)$, i.e., z is now distributed uniformly on the unit circle for both domains. However, the conditional distribution $p(y \mid z)$ is not aligned between the two domains (y is represented by color), which means using this representation for classification is suboptimal, and in this extreme case would lead to 0% accuracy in the target domain 2 . This is an extreme case of misalignment but it does illustrate the importance of the conditional alignment. Therefore, we need to align both the marginal and the conditional distributions for a domain-invariant representation.

在图 1 中，我们展示了一个示例，其中表示 z 满足边际对齐但不满足条件对齐。具体而言，对于领域 1，x 在半径为 2 的圆上均匀分布 (以原点为中心)，而对于领域 2，x 在半径为 3 的圆上均匀分布 (同样以原点为中心)。由映射 $z = g(x) = x/\parallel x\parallel_2$ 定义的表示 z 将对齐边际分布 $p(z)$ ，即 z 现在在两个领域的单位圆上均匀分布。然而，两个领域之间的条件分布 $p(y \mid z)$ 并未对齐 (y 由颜色表示)，这意味着使用该表示法进行分类并非最佳选择，在这种极端情况下会导致目标领域 2 的 0% 准确率。这是一个对齐失效的极端案例，但确实说明了条件对齐的重要性。因此，我们需要对齐边际和条件分布，以实现领域不变的表示。

Recently, there have been several attempts [33, 34, 50] to align the joint distribution of the representation and the label $p(y, z)$ in a domain generalization problem by aligning the distribution of z across domains for each class, i.e., $p(z \mid y)$ (given that the label distribution $p(y)$ is unchanged across domains). However, the key drawbacks of these methods are that they either do not scale well with the number of classes or have limited performance in real-world computer vision datasets (see Section 5).

最近，有几项尝试 [33, 34, 50] 旨在通过对每个类别的 z 在不同领域之间的分布进行对齐，从而在领域泛化问题中对表示和标签的联合分布 $p(y, z)$ 进行对齐，即 $p(z \mid y)$ (假设标签分布 $p(y)$ 在不同领域之间保持不变)。然而，这些方法的主要缺点是它们要么在类别数量增加时扩展性差，要么在实际计算机视觉数据集中的表现有限 (见第 5 节)。

In this paper, we focus on learning a domain-invariant representation that aligns both the marginal and the conditional distributions in domain generalization problems. We present theoretical results regarding the necessary and sufficient conditions for the existence of a domain-invariant representation; and subsequently propose a method to learn such representations by enforcing the invariance of the representation network under domain density transformation functions. A simple intuition for our approach is that if we enforce the representation to be invariant under the transformations among the source domains, the representation will become more robust under other domain transformations.

在本文中，我们专注于学习一种领域不变的表示，该表示在领域泛化问题中对边际分布和条件分布进行对齐。我们提出了关于领域不变表示存在的必要和充分条件的理论结果；随后提出了一种通过强制表示网络在领域密度变换函数下的不变性来学习这种表示的方法。我们方法的简单直觉是，如果我们强制表示在源领域之间的变换下保持不变，那么该表示在其他领域变换下将变得更加稳健。

Furthermore, we introduce an implementation of our method in practice, in which the domain transformation functions are learned through the training process of generative adversarial networks (GANs) [20, 12]. We conduct extensive experiments on several widely used datasets and observe a significant improvement over relevant baselines. We also compare our methods against other state-of-the-art models and show that our method achieves competitive results.

此外，我们介绍了我们方法在实践中的实现，其中领域变换函数通过生成对抗网络 (GANs)[20, 12] 的训练过程进行学习。我们在几个广泛使用的数据集上进行了大量实验，并观察到相较于相关基线的显著改进。我们还将我们的方法与其他最先进的模型进行了比较，显示出我们的方法达到了具有竞争力的结果。

Our contribution in this work is threefold:

我们在这项工作中的贡献有三方面:

- We revisit the domain invariant representation learning problem and shed some light by providing several observations: a necessary and sufficient condition for the existence of a domain-invariant representation and a connection between domain-independent representation and a marginally-aligned representation.

- 我们重新审视了领域不变表示学习问题，并通过提供几个观察结果来阐明这一问题: 领域不变表示存在的必要和充分条件，以及领域无关表示与边际对齐表示之间的联系。

- We propose a theoretically grounded method for learning a domain-invariant representation based on domain density transformation functions. We also demonstrate that we can learn the domain transformation functions by GANs in order to implement our approach in practice.

- 我们提出了一种基于领域密度变换函数的理论基础方法来学习领域不变表示。我们还证明可以通过 GANs 学习领域变换函数，以便在实践中实现我们的方法。

- We empirically show the effectiveness of our method by performing experiments on widely used domain generalization datasets (e.g., Rotated MNIST, VLCS and PACS) and compare our method with relevant baselines (especially CIDG [33], CIDDG [34] and DGER [50]).

- 我们通过在广泛使用的领域泛化数据集 (例如，旋转 MNIST、VLCS 和 PACS) 上进行实验，实证展示了我们方法的有效性，并将我们的方法与相关基线 (特别是 CIDG [33]、CIDDG [34] 和 DGER [50]) 进行了比较。

2 Related Work

2 相关工作

Domain generalization: Domain generalization is an seminal task in real-world machine learning problems where the data distribution of a target domain might vary from that of the training source domains. Therefore, extensive research has been developed to handle that domain-shift problem, aiming at a better generalization performance in the unseen target domain. A predominant approach for domain generalization is domain invariance [37, 33, 34, 3, 47, 2, 24, 50, 1, 32, 44, 43] that learns a domain-invariant representation (which we define as to align the marginal distribution of the representation or the conditional distribution of the output given the representation or both). We are particularly interested in CIDG [33], CIDDG [34] and DGER [50], which also learn a representation that aligns the joint distribution of the representation and the label given that the class distribution is unchanged across domains. It should be noted that Zhao et al. [50] assume the label is distributed uniformly on all domains, while our proposed method only requires an assumption that the distribution of label is unchanged across domains (and not necessarily uniform). We also show later in our paper that the invariance of the distribution of class label across domains turns out to be the necessary and sufficient condition for the existence of a domain-invariant representation. Moreover, we provide a unified theoretical discussion about the two types of alignment, and then propose a method to learn a representation that aligns both the marginal and conditional distributions via domain density transformation functions for the domain generalization problem. Note that there exist several related works, such as Ajakan et al. [1], Ganin et al. [17], that use adversarial loss with a domain discriminator to align the marginal distribution of representation among domains, but they are different from our approach. In particular, our method only uses GANs or normalizing flows to learn the transformation among domains, and learn a representation that is invariant under these functions, without using an adversarial loss on the representation (which can lead to very unstable training [19, 27]). There also exist works [35, 23, 7, 40] in the domain adaptation literature that use generative modeling to learn a domain transformation function from source to target images, and use the transformed images to train a classifier. Our method differs from these by enforcing the representation to be invariant under the domain transformation, and we show theoretically that the representation learned that way would be domain-invariant marginally and conditionally. Meanwhile, the above works use the domain transformation to transform the images and train the classifier directly on the transformed data, and are not effective or applicable for domain generalization.

领域泛化: 领域泛化是现实世界机器学习问题中的一个重要任务，其中目标领域的数据分布可能与训练源领域的数据分布不同。因此，已经开展了大量研究来处理这一领域转移问题，旨在在未见过的目标领域中实现更好的泛化性能。领域泛化的一个主要方法是领域不变性 [37, 33, 34, 3, 47, 2, 24, 50, 1, 32, 44, 43]，它学习一个领域不变的表示 (我们定义为对表示的边际分布或给定表示的输出的条件分布或两者进行对齐)。我们特别关注 CIDG [33]、CIDDG [34] 和 DGER [50]，这些方法也学习一个表示，使得在类分布在各个领域中保持不变的情况下，对齐表示和标签的联合分布。需要注意的是，Zhao 等人 [50] 假设标签在所有领域中均匀分布，而我们提出的方法仅要求标签在各个领域中的分布保持不变 (并不一定是均匀的)。我们在论文后面还表明，类标签在各个领域中的分布不变性实际上是存在领域不变表示的必要和充分条

件。此外，我们提供了关于这两种对齐类型的统一理论讨论，然后提出了一种方法，通过领域密度变换函数学习一个同时对齐边际和条件分布的表示，以解决领域泛化问题。请注意，存在一些相关工作，例如 Ajakan 等人 [1]、Ganin 等人 [17]，它们使用对抗损失与领域鉴别器来对齐各个领域之间表示的边际分布，但它们与我们的方法不同。特别是，我们的方法仅使用 GAN 或归一化流来学习领域之间的变换，并学习一个在这些函数下不变的表示，而不对表示使用对抗损失 (这可能导致非常不稳定的训练 [19, 27])。在领域适应文献中也存在一些工作 [35, 23, 7, 40]，它们使用生成建模来学习从源图像到目标图像的领域变换函数，并使用变换后的图像来训练分类器。我们的方法与这些方法不同，通过强制表示在领域变换下保持不变，我们理论上证明了以这种方式学习的表示在边际和条件上都是领域不变的。同时，上述工作使用领域变换来变换图像，并直接在变换后的数据上训练分类器，这对于领域泛化并不有效或适用。

Another line of methods that received a recent surge in interest is applying the idea of meta-learning for domain generalization problems [16, 4, 31, 5] . The core idea behind these works is that if we train a model that can adapt among the source domains well, it would be more likely to adapt to unseen target domains. Recently, there are approaches [15, 9, 41] that make use of the domain specificity, together with domain invariance, for the prediction problem. The argument here is that domain invariance, while being generalized well between domains, might be insufficient for the prediction of each specific domain and thus domain specificity is necessary. We would like to emphasize that our method is not a direct competitor of meta-learning based and domain-specificity based methods. In fact, we expect that our method can be used in conjunction with these methods to get the best of both worlds for better performance.

另一种最近受到关注的方法是将元学习的思想应用于领域泛化问题 [16, 4, 31, 5] 。这些工作的核心思想是，如果我们训练一个能够在源领域之间良好适应的模型，它更有可能适应未见过的目标领域。最近，有一些方法 [15, 9, 41] 利用领域特异性和领域不变性来解决预测问题。这里的论点是，领域不变性虽然在领域之间能够很好地泛化，但可能不足以预测每个特定领域，因此领域特异性是必要的。我们想强调的是，我们的方法并不是基于元学习和领域特异性方法的直接竞争者。实际上，我们期望我们的方法可以与这些方法结合使用，以获得更好的性能。

Density transformation between domains: Since our method is based on domain density transformations, we will review briefly some related works here. To transform the data density between domains, one can use several types of generative models. Two common methods are based on GANs [51, 12, 13] and normalizing flows [21]. Although our method is not limited to the choice of the generative model used for learning the domain transformation functions, we opt to use GAN, specifically StarGAN [12], for scalability. This is just an implementation choice to demonstrate the use and effectiveness of our method in practice, and it is unrelated to our theoretical results.

领域之间的密度变换: 由于我们的方法基于领域密度变换，我们将在这里简要回顾一些相关工作。为了在领域之间转换数据密度，可以使用几种类型的生成模型。两种常见的方法基于 GAN [51, 12, 13] 和归一化流 [21]。虽然我们的方法并不局限于用于学习领域变换函数的生成模型的选择，但我们选择使用 GAN，特别是 StarGAN [12]，以便于扩展。这只是一个实现选择，用于展示我们的方法在实践中的使用和有效性，与我们的理论结果无关。

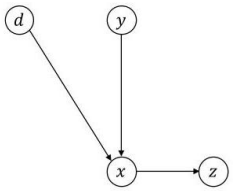


Figure 2: Graphical model. The data distribution is $p(x, y \mid d)$ for each domain d . Our goal is to learn a representation z with a mapping $p(z \mid x)$ from x so that z can be generalized across domains for the prediction task.

图 2: 图形模型。每个领域的数据分布是 $p(x, y \mid d)$ d 。我们的目标是学习一个表示 z ，并通过从 x 的映射 $p(z \mid x)$ 使 z 能够在领域间泛化，以便用于预测任务。

Connection to contrastive learning: Our method can be interpreted intuitively as a way to learn a representation network that is invariant (robust) under domain transformation functions. On the other hand, contrastive learning [10, 11, 36] is also a representation learning paradigm where the model learns images' similarity. In particular, contrastive learning encourages the representation of an input to be similar under different transformations (usually image augmentations). However, the transformations in contrastive learning are not learned and do not serve the purpose of making the representation robust under domain transformations. Our method first learns the transformations between domains and then uses them to learn a representation that is invariant under domain shifts.

连接到对比学习: 我们的方法可以直观地解释为一种学习在域变换函数下不变 (鲁棒) 的表示网络的方法。另一方面，对比学习 [10, 11, 36] 也是一种表示学习范式，其中模型学习图像的相似性。特别是，对比学习鼓励输入的表示在不同变换下相似 (通常是图像增强)。然而，对比学习中的变换并不是学习得来的，也不旨在使表示在域变换下变得鲁棒。我们的方法首先学习域之间的变换，然后利用这些变换学习在域变化下不变的表示。

3 Theoretical Approach

3 理论方法

3.1 Problem Statement

3.1 问题陈述

Let us denote the data distribution for a domain $d \in \mathcal{D}$ by $p(x, y \mid d)$, where the variable $x \in \mathcal{X}$ represents the data and $y \in \mathcal{Y}$ is its corresponding label. The graphical model for our domain generalization framework is depicted in Figure 2, in which the joint distribution is presented as follows:

设我们用 $p(x, y \mid d)$ 表示某一域的数据显示分布 $d \in \mathcal{D}$ ，其中变量 $x \in \mathcal{X}$ 表示数据，$y \in \mathcal{Y}$ 是其对应的标签。我们的域泛化框架的图形模型如图 2 所示，其中联合分布表示如下:

$$p(d, x, y, z) = p(d)\, p(y)\, p(x \mid y, d)\, p(z \mid x)\,. \tag{1}$$

In the domain generalization problem, since the data distribution $p(x, y \mid d)$ varies between domains, we expect the changes in the marginal data distribution $p(x \mid d)$ or the conditional data distribution $p(y \mid x, d)$ or both. In this paper, we assume that $p(y \mid d)$ is invariant across domains, i.e., the marginal distribution of the label y is not dependent on the domain d -this assumption is shown to be the key condition for the existence of a domain-invariant representation (see Remark 1). This is practically reasonable since in many classification datasets, the class distribution can be assumed to be unchanged across domains (usually uniform distribution among the classes, e.g., balanced datasets).

在域泛化问题中，由于数据分布 $p(x, y \mid d)$ 在不同域之间变化，我们预期边际数据分布 $p(x \mid d)$ 或条件数据分布 $p(y \mid x, d)$ 或两者都会发生变化。在本文中，我们假设 $p(y \mid d)$ 在不同域之间是不变的，即标签 y 的边际分布不依赖于域 d - 这一假设被证明是存在域不变表示的关键条件 (见备注 1)。这是在实际中合理的，因为在许多分类数据集中，可以假设类分布在不同域之间保持不变 (通常在类之间是均匀分布，例如，平衡数据集)。

Our aim is to find a domain-invariant representation z represented by the mapping $p(z \mid x)$ that can be used for the classification of label y and be generalized among domains. In practice, this mapping can be deterministic (in that case, $p(z \mid x) = \delta_{g_\theta(x)}(z)$ with some function g_θ , where δ is the Dirac delta distribution) or probabilistic (e.g., a normal distribution with the mean and standard deviation outputted by a network parameterized by θ). For all of our experiments, we use a deterministic mapping for an efficient inference at test time, while in this section, we present our theoretical results with the general case of a distribution $p(z \mid x)$.

我们的目标是找到一个领域不变的表示 z ，通过映射 $p(z \mid x)$ 表示，该表示可用于标签 y 的分类，并能在不同领域之间进行泛化。在实践中，这个映射可以是确定性的 (在这种情况下，$p(z \mid x) = \delta_{g_\theta(x)}(z)$ 与某个函数 g_θ ，其中 δ 是狄拉克 delta 分布) 或概率性的 (例如，正态分布，其均值和标准差由参数化为 θ 的网络输出)。在我们所有的实验中，我们使用确定性映射以实现测试时的高效推理，而在本节中，我们展示了分布 $p(z \mid x)$ 的一般情况的理论结果。

In most existing domain generalization approaches, the domain-invariant representation z is defined using one of the two following definitions:

在现有的大多数领域泛化方法中，领域不变的表示 z 是使用以下两个定义之一来定义的:

Definition 1. (Marginal Distribution Alignment) The representation z is said to satisfy the marginal distribution alignment condition if $p(z \mid d)$ is invariant w.r.t. d .

定义 1. (边际分布对齐) 如果表示 z 对 d 不变，则称其满足边际分布对齐条件 $p(z \mid d)$ 。

Definition 2. (Conditional Distribution Alignment) The representation z is said to satisfy the conditional distribution alignment condition if $p(y \mid z, d)$ is invariant w.r.t. d .

定义 2. (条件分布对齐) 如果表示 z 对 d 不变，则称其满足条件分布对齐条件 $p(y \mid z, d)$ 。

However, when the joint data distribution varies between domains, it is crucial to align both the marginal and the conditional distribution of the representation z . To this end, this paper aims to

然而，当联合数据分布在不同领域之间变化时，至关重要的是对表示 z 的边际和条件分布进行对齐。为此，本文旨在

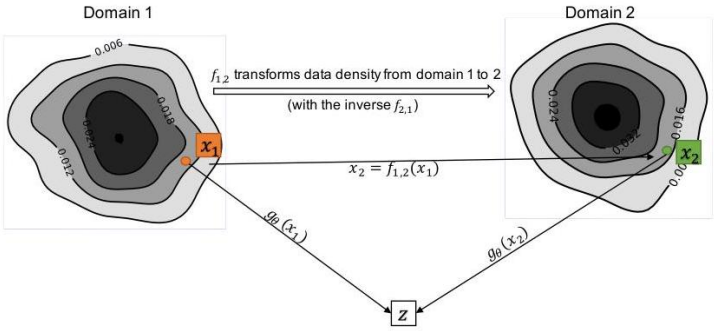


Figure 3: Domain density transformation. If we know the function $f_{1,2}$ that transforms the data density from domain 1 to domain 2, we can learn a domain invariant representation network $g_\theta(x)$ by enforcing it to be invariant under $f_{1,2}$, i.e., $g_\theta(x_1) = g_\theta(x_2)$ for any $x_2 = f_{1,2}(x_1)$.

图 3: 领域密度变换。如果我们知道将数据密度从领域 1 转换到领域 2 的函数 $f_{1,2}$ ，我们可以通过强制其在 $f_{1,2}$ 下不变，即对于任何 $x_2 = f_{1,2}(x_1)$ ，学习一个领域不变的表示网络 $g_\theta(x)$ 。

learn a representation z that satisfies both the marginal and conditional alignment conditions. We justify our assumption of independence between y and d (thus $p(y \mid d) = p(y)$) by the following remark, which shows that this assumption turns out to be the necessary and sufficient condition for learning a domain-invariant representation. Note that this condition is also used in several existing works [50, 33, 34] .

学习一个表示 z ，使其同时满足边际和条件对齐条件。我们通过以下备注来证明我们对 y 和 d 之间独立性的假设 (因此 $p(y \mid d) = p(y)$)，这表明该假设是学习领域不变表示的必要和充分条件。请注意，这一条件在一些现有工作中也被使用 [50, 33, 34] 。

Remark 1. The invariance of $p(y \mid d)$ across domains d is the necessary and sufficient condition for the existence of a domain-invariant representation (that aligns both the marginal and conditional distributions).

备注 1. $p(y \mid d)$ 在领域 d 之间的不变性是存在一个领域不变表示 (同时对齐边际和条件分布) 的必要和充分条件。

Proof. provided in the appendix.

证明。见附录。

It is also worth noting that methods which learn a domain independent representation, for example, [24], only align the marginal distribution. This comes directly from the following remark:

还值得注意的是，学习领域独立表示的方法，例如 [24]，仅对齐边际分布。这直接来自以下备注:

Remark 2. A representation z satisfies the marginal distribution alignment condition if and only if $I(z, d) = 0$, where $I(z, d)$ is the mutual information between z and d .

备注 2. 表示 z 当且仅当 $I(z, d) = 0$ 时满足边际分布对齐条件，其中 $I(z, d)$ 是 z 和 d 之间的互信息。

Proof. provided in the appendix.

证明。见附录。

The question still remains that how we can learn a non-trivial domain invariant representation that satisfies both of the distribution alignment conditions. This will be discussed in the following subsection.

仍然存在一个问题，即我们如何能够学习一个满足两个分布对齐条件的非平凡领域不变表示。这个问题将在以下小节中讨论。

3.2 Learning a Domain-Invariant Representation with Domain Density Transformation Functions

3.2 使用领域密度变换函数学习领域不变表示

To present our method, we will make some assumptions about the data distribution. Specifically, for any two domains d, d' , we assume that there exists an invertible and differentiable function denoted by $f_{d,d'}$ that transforms the density $p(x \mid y, d)$ to $p(x' \mid y, d'), \forall y$. Let $f_{d,d'}$ be the inverse of $f_{d',d}$, i.e.,

为了展示我们的方法，我们将对数据分布做一些假设。具体而言，对于任何两个领域 d, d' ，我们假设存在一个可逆且可微的函数，记作 $f_{d,d'}$ ，它将密度 $p(x \mid y, d)$ 转换为 $p(x' \mid y, d'), \forall y$ 。设 $f_{d,d'}$ 为 $f_{d',d}$ 的逆，即，$f_{d',d} := (f_{d,d'})^{-1}$.

Due to the invertibility and differentiability of f ' s, we can apply the change of variables theorem [39, 6] for the distributions above. In particular, with $x' = f_{d,d'}(x)$ (and thus $x = f_{d',d}(x')$), we

由于 f 的可逆性和可微性，我们可以对上述分布应用变量变化定理 [39, 6] 。特别地，使用 $x' = f_{d,d'}(x)$ (因此 $x = f_{d',d}(x')$)，我们

have:

得到:

$$p(x \mid y, d) = p(x' \mid y, d') \left| \det J_{f_{d',d}}(x') \right|^{-1}, \tag{2}$$

where $J_{f_{d',d}}(x')$ is the Jacobian matrix of the function $f_{d',d}$ evaluated at x' .

其中 $J_{f_{d',d}}(x')$ 是在 x' 处评估的函数 $f_{d',d}$ 的雅可比矩阵。

Multiplying both sides of Eq. 2 with $p(y \mid d) = p(y \mid d')$, we get

将方程 2 的两边都乘以 $p(y \mid d) = p(y \mid d')$ ，我们得到

$$p(x, y \mid d) = p(x', y \mid d') \left| \det J_{f_{d',d}}(x') \right|^{-1}; \tag{3}$$

and marginalizing both sides of the above equation over y gives us

并且对上述方程的两边进行边际化，得到

$$p(x \mid d) = p(x' \mid d') \left| \det J_{f_{d',d}}(x') \right|^{-1}. \tag{4}$$

By using Eq. 2 and Eq. 4, we can prove the following theorem, which offers an efficient way to learn a domain-invariant representation, given the transformation functions f ' s between domains.

通过使用方程 2 和方程 4，我们可以证明以下定理，该定理提供了一种有效的方法来学习领域不变的表示，给定领域之间的变换函数 f 。

Theorem 1. Given an invertible and differentiable function $f_{d,d'}$ (with the inverse $f_{d',d}$) that transforms the data density from domain d to d' (as described above). Assuming that the representation z satisfies:

定理 1. 给定一个可逆且可微的函数 $f_{d,d'}$ (其逆为 $f_{d',d}$)，该函数将数据密度从领域 d 转换为领域 d' (如上所述)。假设表示 z 满足:

$$p(z \mid x) = p(z \mid f_{d,d'}(x)), \forall x, \tag{5}$$

Then it aligns both the marginal and the conditional of the data distribution for domain d and d' .

然后它对领域 d 和 d' 的数据分布的边际和条件进行对齐。

Proof. provided in the appendix.

证明。见附录。

This theorem indicates that, if we can find the functions f that transform the data densities among the domains, we can learn a domain-invariant representation z by encouraging the representation to be invariant under all the transformations f . This idea is illustrated in Figure 3 . We therefore can use the following learning objective to learn a domain-invariant representation $z = g_\theta(x)$:

该定理表明，如果我们能够找到在领域之间转换数据密度的函数 f ，我们可以通过鼓励表示在所有变换 f 下保持不变来学习领域不变的表示 z 。这一思想在图 3 中得到了说明。因此，我们可以使用以下学习目标来学习领域不变的表示 $z = g_\theta(x)$:

$$\mathbb{E}_d \left[\mathbb{E}_{p(x,y|d)} \left[l(y, g_\theta(x)) + \mathbb{E}_{d'} \left[\operatorname{dis}(g_\theta(x), g_\theta(f_{d,d'}(x))) \right] \right] \right] \tag{6}$$

Assume that we have a set of K sources domain $D_s = \{d_1, d_2, \ldots, d_K\}$, the objective function in Eq. 6 becomes:

假设我们有一组 K 源领域 $D_s = \{d_1, d_2, \ldots, d_K\}$ ，方程 6 中的目标函数变为:

$$\mathbb{E}_{d,d' \in D_s, p(x,y|d)} \left[l(y, g_\theta(x)) + \operatorname{dis}(g_\theta(x), g_\theta(f_{d,d'}(x))) \right], \tag{7}$$

where $l(y, g_\theta(x))$ is the prediction loss of a network that predicts y given $z = g_\theta(x)$, and dis is a distance metric to enforce the invariant condition in Eq. 5. In our implementation, we use a squared error distance, e.g., $\operatorname{dis}(g_\theta(x), g_\theta(f_{d,d'}(x))) = \left\| g_\theta(x) - g_\theta(f_{d,d'}(x)) \right\|_2^2$, since it performs the best in

practice. However, we also considered other distances such as constrastive distance, which we discuss in more detail in the appendix.

其中 $l\left(y, g_\theta\left(x\right)\right)$ 是一个网络的预测损失，该网络在给定 $z = g_\theta\left(x\right)$ 的情况下预测 y ，而 dis 是一个距离度量，用于强制执行方程 5 中的不变条件。在我们的实现中，我们使用平方误差距离，例如 $\text{dis}\left(g_\theta\left(x\right), g_\theta\left(f_{d,d'}\left(x\right)\right)\right) = \left\|g_\theta\left(x\right) - g_\theta\left(f_{d,d'}\left(x\right)\right)\right\|_2^2$ ，因为它在实践中表现最佳。然而，我们也考虑了其他距离，例如对比距离，我们在附录中进行了更详细的讨论。

This theorem motivates us to learn such domain transformation functions for our domain-invariant representation learning framework. In the next section, we show how one can incorporate this idea into real-world domain generalization problems by learning the transformations with generative adversarial networks.

该定理促使我们学习这样的领域转换函数，以便为我们的领域不变表示学习框架提供支持。在下一节中，我们将展示如何通过使用生成对抗网络学习这些转换，将这一思想融入到现实世界的领域泛化问题中。

4 An Practical Implementation using Generative Adversarial Networks

4 使用生成对抗网络的实际实现

In practice, we can learn the functions f that transform the data distributions between domains by using several generative modeling frameworks, e.g., normalizing flows [21] or GANs [51, 12, 13]. One advantage of normalizing flows is that this transformation is naturally invertible by design of the neural network. However, existing frameworks (e.g., Grover et al. [21]) require two flows to transform between each pair of domains, making it not scalable (scales linearly with the number of domains). Moreover, an initial implementation of our method using AlignFlow shows similar performance with the version using GAN. Therefore, we opt to use GANs for better scalability. In particular, we use the StarGAN [12] model, which is a unified network (only requiring a single network to transform across all domains) designed for image domain transformations. It should be noted that the transformations learned by StarGAN are differentiable everywhere or almost everywhere with typical choices of the activation function (e.g., tanh or ReLU), and the cycle-consistency loss enforces each pair of transformations to approximate a pair of inverse functions.

在实践中，我们可以使用几个生成建模框架来学习在领域之间转换数据分布的函数 f ，例如，归一化流 [21] 或 GAN [51, 12, 13]。归一化流的一个优点是，这种转换在神经网络的设计上是自然可逆的。然而，现有框架 (例如，Grover 等 [21]) 需要两个流在每对领域之间进行转换，这使得其不可扩展 (与领域数量线性相关)。此外，我们的方法的初步实现使用 AlignFlow 显示出与使用 GAN 的版本相似的性能。因此，我们选择使用 GAN 以获得更好的可扩展性。特别地，我们使用 StarGAN [12] 模型，这是一个统一的网络 (只需一个网络即可在所有领域之间转换)，旨在进行图像领域转换。值得注意的是，StarGAN 学习的转换在典型的激活函数选择 (例如，tanh 或 ReLU) 下几乎处处可微，并且循环一致性损失强制每对转换近似一对逆函数。

The goal of StarGAN is to learn a unified network G that transforms the data density among multiple domains. In particular, the network $G\left(x, d, d'\right)$ (i.e., G is conditioned on the image x and the two different domains d, d') transforms an image x from domain d to domain d' . Different from the original StarGAN model that only takes the image x and the desired destination domain d' as its input, in our implementation, we feed both the original domain d and desired destination domain d' together with the original image x to the generator G .

StarGAN 的目标是学习一个统一的网络 G ，该网络在多个领域之间转化数据密度。特别是，该网络 $G\left(x, d, d'\right)$ (即 G 以图像 x 和两个不同的领域 d, d' 为条件) 将图像 x 从领域 d 转换为领域 d' 。与原始的 StarGAN 模型不同，后者仅将图像 x 和期望的目标领域 d' 作为输入，在我们的实现中，我们将原始领域 d 和期望的目标领域 d' 以及原始图像 x 一起输入到生成器 G 中。

The generator' s goal is to fool a discriminator D into thinking that the transformed image belongs to the destination domain d' . In other words, the equilibrium state of StarGAN, in which G completely fools D , is when G successfully transforms the data density of the original domain to that of the destination domain. After training, we use $G\left(., d, d'\right)$ as the function $f_{d,d'}\left(\ .\ \right)$ *describedi* in the previous section and perform the representation learning via the objective function in Eq. 7.

生成器的目标是欺骗判别器 D ，使其认为转换后的图像属于目标领域 d' 。换句话说，StarGAN 的平衡状态是 G 完全欺骗 D ，即 G 成功地将原始领域的数据密度转换为目标领域的数据密度。经过训练

后，我们使用 $G(.,d,d')$ 作为上一节中的函数 $f_{d,d'}(\ .\ )describedi$ ，并通过公式 (7) 中的目标函数执行表示学习。

Three important loss functions of the StarGAN architecture are:

StarGAN 架构的三个重要损失函数是:

- Domain classification loss $\mathcal{L}_{cls}$ that encourages the generator G to generate images that closely belongs to the desired destination domain d' .
- 领域分类损失 $\mathcal{L}_{cls}$ ，鼓励生成器 G 生成与期望目标领域 d' 密切相关的图像。
- The adversarial loss $\mathcal{L}_{adv}$ that is the classification loss of a discriminator D that tries to distinguish between real images and the synthetic images generated by G. The equilibrium state of StarGAN is when G completely fools D , which means the distribution of the generated images (via $G(x,d,d'),x\sim p(x\mid d)$) becomes the distribution of the real images of the destination domain $p(x'\mid d')$. This is our objective, i.e., to learn a function that transforms domains' densities.
- 对抗损失 $\mathcal{L}_{adv}$ ，这是判别器 D 的分类损失，判别器试图区分真实图像和生成器 G 生成的合成图像。StarGAN 的平衡状态是 G 完全欺骗 D ，这意味着生成图像 (通过 $G(x,d,d'),x\sim p(x\mid d)$) 的分布变为目标领域 $p(x'\mid d')$ 的真实图像的分布。这是我们的目标，即学习一个转换领域密度的函数。
- Reconstruction loss $\mathcal{L}_{\text{rec}}=\mathbb{E}_{x,d,d'}\left[\left\|x-G(x',d',d)\right\|_1\right]$ where $x'=G(x,d,d')$ to ensure that the transformations preserve the image's content. Note that this also aligns with our interest since we want $G(.,d',d)$ to be the inverse of $G(.,d,d')$, which minimizes $\mathcal{L}_{\text{rec}}$.
- 重建损失 $\mathcal{L}_{\text{rec}}=\mathbb{E}_{x,d,d'}\left[\left\|x-G(x',d',d)\right\|_1\right]$ 其中 $x'=G(x,d,d')$ 以确保变换保持图像的内容。请注意，这也与我们的兴趣一致，因为我们希望 $G(.,d',d)$ 是 $G(.,d,d')$ 的逆，这样可以最小化 $\mathcal{L}_{\text{rec}}$ 。

We can enforce the generator G to transform the data distribution within the class y (e.g., $p(x\mid y,d)$ to $p(x'\mid y,d')\forall y$) by sampling each minibatch with data from the same class y , so that the discriminator will distinguish the transformed images with the real images from class y and domain d' . However, we found that this constraint can be relaxed in practice, and the generator almost always transforms the image within the original class y .

我们可以强制生成器 G 在类别 y 内部转换数据分布 (例如，通过从同一类别 y 中抽样每个小批量数据，将 $p(x\mid y,d)$ 转换为 $p(x'\mid y,d')\forall y)$)，以便判别器能够区分变换后的图像与来自类别 y 和领域 d' 的真实图像。然而，我们发现这个约束在实践中可以放宽，生成器几乎总是将图像转换为原始类别 y 内部。

As mentioned earlier, after training the StarGAN model, we can use the generator $G(.,d,d')$ as our

如前所述,在训练 StarGAN 模型后,我们可以将生成器 $G(.,d,d')$ 用作我们的 $f_{d,d'}(\ .\ )functionandlearnademariantr$

this implementation of our method DIRT-GAN (Domain Invariant Representation learning with domain Transformations via Generative Adversarial Networks).

方法的实现名称为 DIRT-GAN(通过生成对抗网络进行领域不变表示学习与领域变换)。

5 Experiments

5 实验

5.1 Datasets

5.1 数据集

To evaluate our method, we perform experiments in three datasets that are commonly used in the literature for domain generalization.

为了评估我们的方法，我们在三个常用于文献中的数据集上进行实验，以进行领域泛化。

Rotated MNIST [18]: In this dataset, 1,000 MNIST images (100 per class) [29] are chosen to form the first domain (denoted $\mathcal{M}_0$), then counter-clockwise rotations of $15°, 30°, 45°, 60°$ and $75°$ are applied to create five additional domains, denoted $\mathcal{M}_{15}, \mathcal{M}_{30}, \mathcal{M}_{45}, \mathcal{M}_{60}$ and $\mathcal{M}_{75}$. The task is classification with ten classes (digits 0 to 9).

旋转的 MNIST [18]: 在该数据集中，选择了 1,000 张 MNIST 图像 (每类 100 张) [29] 形成第一个领域 (记为 $\mathcal{M}_0$)，然后对 $15°, 30°, 45°, 60°$ 和 $75°$ 进行逆时针旋转，以创建五个额外的领域，记为 $\mathcal{M}_{15}, \mathcal{M}_{30}, \mathcal{M}_{45}, \mathcal{M}_{60}$ 和 $\mathcal{M}_{75}$ 。任务是分类，共有十个类别 (数字 0 到 9)。

VLCS [18]: contains 10,729 images from four domains, each domain is a subdataset. The four datasets are VOC2007 (V), LabelMe (L), Caltech-101 (C), and SUN09 (S). The task is classification with five classes.

VLCS [18]: 包含来自四个领域的 10,729 张图像，每个领域是一个子数据集。这四个数据集是 VOC2007(V)、LabelMe(L)、Caltech-101(C) 和 SUN09(S)。任务是分类，共有五个类别。

PACS [30]: contains 9,991 images from four different domains: art painting, cartoon, photo, sketch. The task is classification with seven classes.

PACS [30]: 包含来自四个不同领域的 9,991 张图像: 艺术绘画、卡通、照片和素描。任务是进行七类分类。

5.2 Experimental Setting

5.2 实验设置

For all datasets, we perform "leave-one-domain-out" experiments, where we choose one domain as the target domain, train the model on all remaining domains and evaluate it on the chosen domain. Following standard practice, we use 90% of available data as training data and 10% as validation data, except for the Rotated MNIST experiment where we do not use a validation set and just report the performance of the last epoch.

对于所有数据集，我们执行 "留一领域法" 实验，其中选择一个领域作为目标领域，在所有剩余领域上训练模型，并在所选领域上进行评估。遵循标准做法，我们使用可用数据的 90% 作为训练数据，使用 10% 作为验证数据，除了旋转 MNIST 实验外，我们不使用验证集，仅报告最后一个周期的性能。

For the Rotated MNIST dataset, we use a network of two 3×3 convolutional layers and a fully connected layer as the representation network g_θ to get a representation z of 64 dimensions. A single

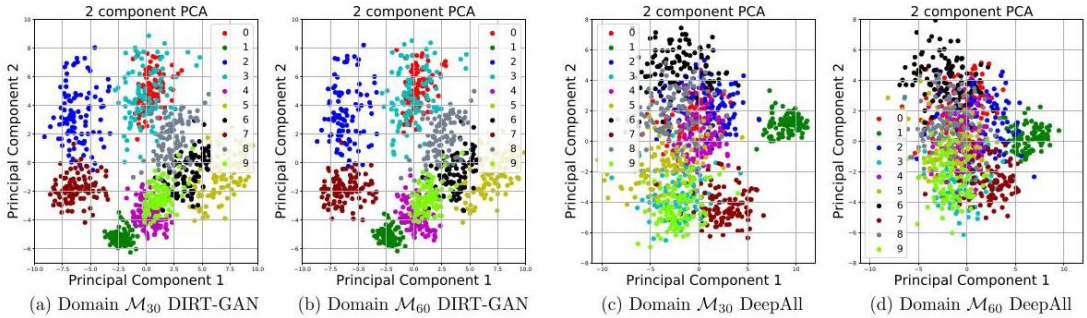
对于旋转 MNIST 数据集，我们使用两个 3×3 卷积层和一个全连接层作为表示网络 g_θ ，以获得 64 维的表示 z 。一个单一的线性层随后用于将表示 3×3 映射到十个输出类别。这种架构是 Ilse 等人 [24] 使用的网络的确定性版本。我们使用 Adam 优化器 [26] 训练网络 500 个周期，学习率为 0.001，最小批量大小为 64，并在最后一个周期后报告测试领域的性能。

Table 1: Rotated Mnist. Reported numbers are mean accuracy and standard deviation among 5 runs

表 1: 旋转 MNIST。报告的数字是 5 次运行的平均准确率和标准差。

Model	Domains						Average
	$\mathcal{M}_0$	$\mathcal{M}_{15}$	$\mathcal{M}_{30}$	$\mathcal{M}_{45}$	$\mathcal{M}_{60}$	$\mathcal{M}_{75}$	
HIR [47]	90.34	99.75	99.40	96.17	99.25	91.26	96.03
DIVA [24]	93.5	99.3	99.1	99.2	99.3	93.0	97.2
DGER [50]	90.09	99.24	99.27	99.31	99.45	90.81	96.36
DA [17]	86.7	98.0	97.8	97.4	96.9	89.1	94.3
LG [42]	89.7	97.8	98.0	97.1	96.6	92.1	95.3
HEX [46]	90.1	98.9	98.9	98.8	98.3	90.0	95.8
ADV [46]	89.9	98.6	98.8	98.7	98.6	90.4	95.2
DIRT-GAN (ours)	97.2(±0.3)	99.4(±0.1)	99.3(±0.1)	99.3(±0.1)	99.2(±0.1)	97.1(±0.3)	98.6

模型	领域						平均
	$\mathcal{M}_0$	$\mathcal{M}_{15}$	$\mathcal{M}_{30}$	$\mathcal{M}_{45}$	$\mathcal{M}_{60}$	$\mathcal{M}_{75}$	
HIR [47]	90.34	99.75	99.40	96.17	99.25	91.26	96.03
DIVA [24]	93.5	99.3	99.1	99.2	99.3	93.0	97.2
DGER [50]	90.09	99.24	99.27	99.31	99.45	90.81	96.36
DA [17]	86.7	98.0	97.8	97.4	96.9	89.1	94.3
LG [42]	89.7	97.8	98.0	97.1	96.6	92.1	95.3
HEX [46]	90.1	98.9	98.9	98.8	98.3	90.0	95.8
ADV [46]	89.9	98.6	98.8	98.7	98.6	90.4	95.2
DIRT-GAN(我们的)	97.2(±0.3)	99.4(±0.1)	99.3(±0.1)	99.3(±0.1)	99.2(±0.1)	97.1(±0.3)	98.6



(a) Domain $\mathcal{M}_{30}$ DIRT-GAN (b) Domain $\mathcal{M}_{60}$ DIRT-GAN (c) Domain $\mathcal{M}_{30}$ DeepAll (d) Domain $\mathcal{M}_{60}$ DeepAll

Figure 4: Visualization of the representation space. Each point indicates a representation z of an image x in the two dimensional space and its color indicates the label y . Two left figures are for our method DIRT-GAN and two right figures are for the naive model DeepAll.

图 4: 表示空间的可视化。每个点表示二维空间中图像 x 的表示 z ，其颜色表示标签 y 。左侧的两个图是我们的方法 DIRT-GAN，右侧的两个图是简单模型 DeepAll。

linear layer is then used to map the representation z to the ten output classes. This architecture is the deterministic version of the network used by Ilse et al. [24]. We train our network for 500 epochs with the Adam optimizer [26], using the learning rate 0.001 and minibatch size 64, and report performance on the test domain after the last epoch.

然后使用一个线性层将表示 z 映射到十个输出类别。这种架构是 Ilse 等人 [24] 使用的网络的确定性版本。我们使用 Adam 优化器 [26] 训练网络 500 个周期，学习率为 0.001，最小批量大小为 64，并在最后一个周期后报告测试领域的性能。

For the VLCS and PACS datasets, for a fair comparison against our main baselines, we use the most common choices of backbone networks for those datasets in existing works as the representation networks g_θ , i.e., Alexnet [28] for VLCS and Resnet18 [22] for PACS. We replace the last fully connected layer of the backbone with a linear layer of dimension 256 so that our representation has 256 dimensions. As with the Rotated MNIST experiment, we use a single layer to map from the representation z to the output. We train the network for 100 epochs with plain stochastic gradient descent (SGD) using learning rate 0.001, momentum 0.9, minibatch size 64, and weight decay 0.001. Data augmentation is also standard practice for real-world computer vision datasets like VLCS and PACS, and during the training we augment our data as follows: crops of random size and aspect ratio, resizing to 224×224 pixels, random horizontal flips, random color jitter, randomly converting the image tile to grayscale with 10% probability, and normalization using the ImageNet channel means and standard deviations.

对于 VLCS 和 PACS 数据集，为了与我们的主要基线进行公平比较，我们使用现有工作中这些数据集的最常见骨干网络选择作为表示网络 g_θ ，即，VLCS 使用 Alexnet [28]，PACS 使用 Resnet18 [22]。我们用一个维度为 256 的线性层替换骨干网络的最后一个全连接层，以便我们的表示具有 256 个维度。与旋转 MNIST 实验一样，我们使用单层将表示 z 映射到输出。我们使用学习率 0.001、动量 0.9、小批量大小 64 和权重衰减 0.001，采用普通随机梯度下降 (SGD) 训练网络 100 个周期。数据增强也是 VLCS 和 PACS 等现实世界计算机视觉数据集的标准做法，在训练过程中，我们对数据进行如下增强: 随机大小和纵横比的裁剪，调整大小到 224×224 像素，随机水平翻转，随机颜色抖动，以 10% 概率随机将图像块转换为灰度，以及使用 ImageNet 通道均值和标准差进行归一化。

The StarGAN [12] model implementation is taken from the authors' original source code with no significant modifications. For each set of source domains, we train the StarGAN model for 100,000 iterations with a minibatch of 16 images per iteration.

StarGAN [12] 模型的实现来自作者的原始源代码，没有重大修改。对于每组源域，我们训练 StarGAN 模型 100,000 次迭代，每次迭代的小批量为 16 张图像。

Our code is available at https://github.com/atuannguyen/DIRT.We train our model on a NVIDIA Quadro RTX 6000.

我们的代码可在 https://github.com/atuannguyen/DIRT 获取。我们在 NVIDIA Quadro RTX 6000 上训练我们的模型。

5.3 Results

5.3 结果

Rotated MNIST Experiment. Table 1 shows the performance of our model on the Rotated MNIST dataset. The main baselines we consider in this experiment are HIR [47], DIVA [24] and DGER [50], which are domain invariance based methods. Our method recognizably outperforms those, illustrating the effectiveness of our method in learning a domain-invariant representation over the existing works. We also include other best-performing models for this dataset in the second half of the table. To the best of our knowledge, we set a new state-of-the-art performance on this Rotated MNIST dataset.

旋转 MNIST 实验。表 1 显示了我们模型在旋转 MNIST 数据集上的表现。我们在此实验中考虑的主要基线是基于领域不变性的方法 HIR [47]、DIVA [24] 和 DGER [50]。我们的模型显著优于这些基线，展示了我们的方法在学习领域不变表示方面的有效性。我们还在表的后半部分包含了该数据集的其他最佳表现模型。根据我们所知，我们在这个旋转 MNIST 数据集上设定了新的最先进的性能。

Table 2: VLCS. Reported numbers are mean accuracy and standard deviation among 5 runs

表 2: VLCS。报告的数字是 5 次运行的平均准确率和标准差。

Model	Backbone	VLCS				
		V	L	C	S	Average
CIDG [33]	Alexnet	65.65	60.43	91.12	60.85	69.51
CIDDG [34]	Alexnet	64.38	63.06	88.83	62.10	69.59
DGER [50]	Alexnet	73.24	58.26	96.92	69.10	74.38
HIR [47]	Alexnet	69.10	62.22	95.39	65.71	73.10
JiGen [8]	Alexnet	70.62	60.90	96.93	64.30	73.19
DIRT-GAN (ours)	Alexnet	72.1(±1.0)	64.0(±0.9)	97.3(±0.2)	72.2(±1.1)	76.4

模型	主干	VLCS				
		V	L	C	S	平均
CIDG [33]	Alexnet	65.65	60.43	91.12	60.85	69.51
CIDDG [34]	Alexnet	64.38	63.06	88.83	62.10	69.59
DGER [50]	Alexnet	73.24	58.26	96.92	69.10	74.38
HIR [47]	Alexnet	69.10	62.22	95.39	65.71	73.10
JiGen [8]	Alexnet	70.62	60.90	96.93	64.30	73.19
DIRT-GAN(我们的)	Alexnet	72.1(±1.0)	64.0(±0.9)	97.3(±0.2)	72.2(±1.1)	76.4

Table 3: PACS. Reported numbers are mean accuracy and standard deviation among 5 runs

表 3: PACS。报告的数字是 5 次运行的平均准确率和标准差。

Model	Backbone	PACS				
		Art Painting	Cartoon	Photo	Sketch	Average
DGER [50]	Resnet18	80.70	76.40	96.65	71.77	81.38
JiGen [8]	Resnet18	79.42	75.25	96.03	71.35	79.14
MLDG [31]	Resnet18	79.50	77.30	94.30	71.50	80.70
MetaReg [4]	Resnet18	83.70	77.20	95.50	70.40	81.70
CSD [38]	Resnet18	78.90	75.80	94.10	76.70	81.40
DMG [9]	Resnet18	76.90	80.38	93.35	75.21	81.46
DIRT-GAN (ours)	Resnet18	82.56(± 0.4)	76.37(± 0.3)	95.65(± 0.5)	79.89(± 0.2)	83.62

模型	骨干	PACS				
		艺术绘画	卡通	照片	草图	平均
DGER [50]	Resnet18	80.70	76.40	96.65	71.77	81.38
JiGen [8]	Resnet18	79.42	75.25	96.03	71.35	79.14
MLDG [31]	Resnet18	79.50	77.30	94.30	71.50	80.70
MetaReg [4]	Resnet18	83.70	77.20	95.50	70.40	81.70
CSD [38]	Resnet18	78.90	75.80	94.10	76.70	81.40
DMG [9]	Resnet18	76.90	80.38	93.35	75.21	81.46
DIRT-GAN(我们的)	Resnet18	82.56(± 0.4)	76.37(± 0.3)	95.65(± 0.5)	79.89(± 0.2)	83.62

We further analyze the distribution of the representation z by performing principal component analysis to reduce the dimension of z from 64 to two principal components. We visualize the representation space for two domains $\mathcal{M}_{30}$ and $\mathcal{M}_{60}$, with each point indicating the representation z of an image x in the two-dimensional space and its color indicating the label y . Figures 4a and 4 b show the representation space of our method (in domains $\mathcal{M}_{30}$ and $\mathcal{M}_{60}$ respectively). It is clear that both the marginal (judged by the general distribution of the points) and the conditional (judged by the positions of colors) are relatively aligned. Meanwhile, Figures 4c and 4d show the representation space with naive training (in domains $\mathcal{M}_{30}$ and $\mathcal{M}_{60}$ respectively), showing the misalignment in the marginal distribution (judged by the general distribution of the points) and the conditional distribution (for example, the distributions of blue points and green points).

我们进一步通过主成分分析来分析表示 z 的分布，以将 z 的维度从 64 降至两个主成分。我们可视化了两个领域 $\mathcal{M}_{30}$ 和 $\mathcal{M}_{60}$ 的表示空间，每个点表示二维空间中图像 x 的表示 z ，其颜色表示标签 y 。图 4a 和 4 b 显示了我们方法的表示空间 (在领域 $\mathcal{M}_{30}$ 和 $\mathcal{M}_{60}$ 中分别)。显然，边际分布 (通过点的一般分布判断) 和条件分布 (通过颜色的位置判断) 都是相对对齐的。同时，图 4c 和 4d 显示了使用简单训练的表示空间 (在领域 $\mathcal{M}_{30}$ 和 $\mathcal{M}_{60}$ 中分别)，显示了边际分布 (通过点的一般分布判断) 和条件分布 (例如，蓝点和绿点的分布) 中的不对齐。

VLCS and PACS. Tables 2 and 3 show the results for the VLCS and PACS datasets. In these real-world computer vision datasets, we consider HIR [47], CIDG [33], CIDDG [34] and DGER [50] as our main domain-invariance baselines. We also include other approaches (meta-learning based or domain-specificity based) in the second half of the tables for references. Our method significantly ourperforms other invariant-representataion baselines, namely CIDG, CIDDG and DGER, with the same backbone architectures, showing the effectiveness of our representation alignment method.

VLCS 和 PACS。表 2 和表 3 显示了 VLCS 和 PACS 数据集的结果。在这些现实世界的计算机视觉数据集中，我们将 HIR [47]、CIDG [33]、CIDDG [34] 和 DGER [50] 视为我们的主要领域不变基线。我们还在表的后半部分包含了其他方法 (基于元学习或领域特异性的方法) 以供参考。我们的方法显著优于其他不变表示基线，即 CIDG、CIDDG 和 DGER，使用相同的主干架构，显示了我们表示对齐方法的有效性。

6 Conclusion

6 结论

To conclude, in this work we propose a theoretically grounded approach to learn a domain-invariant representation for the domain generalization problem by using domain transformation functions. We also provide insights into domain-invariant representation learning with several theoretical observations. We then introduce an implementation for our method in practice with the domain transformations learned by a StarGAN architecture and empirically show that our approach outperforms other domain-invariance based methods. Our method also achieves competitive results on several datasets when compared to other state-of-the-art models. A potential limitation of our method is that we need to train an additional network (StarGAN) to learn to transform data density among domains.

总之，在这项工作中，我们提出了一种理论基础的方法，通过使用领域转换函数来学习领域不变表示，以解决领域泛化问题。我们还通过几个理论观察提供了对领域不变表示学习的见解。然后，我们介绍了在实践中实现我们的方法，利用 StarGAN 架构学习的领域转换，并实证表明我们的方法优于其他基于领域不变性的方法。与其他最先进的模型相比，我们的方法在多个数据集上也取得了具有竞争力的结果。我们方法的一个潜在限制是，我们需要训练一个额外的网络 (StarGAN) 来学习在领域之间转换数据密度。

However, this network is only used during training, and the required computation at test time is still the same as other models. In the future, we plan to incorporate our method into meta-learning based and domain-specificity based approaches for improved performance. We also plan to extend the domain-invariant representation learning framework to the more challenging scenarios, for example, where domain information is not available (i.e., we have a dataset pooled from multiple source domains but do not know the domain identification of each data instance).

然而，这个网络仅在训练期间使用，测试时所需的计算仍与其他模型相同。未来，我们计划将我们的方法纳入基于元学习和领域特异性的方法，以提高性能。我们还计划将领域不变表示学习框架扩展到更具挑战性的场景，例如，当领域信息不可用时 (即，我们有一个来自多个源领域的数据集，但不知道每个数据实例的领域识别)。

References

参考文献

[1] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand. Domain-adversarial

[1] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, 和 M. Marchand. 领域对抗

neural networks. arXiv preprint arXiv:1412.4446, 2014.

[2] K. Akuzawa, Y. Iwasawa, and Y. Matsuo. Adversarial invariant feature learning with accuracy constraint for domain generalization. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 315-331. Springer, 2019.

[3] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. arXiv preprint arXiv:1907.02893, 2019.

[4] Y. Balaji, S. Sankaranarayanan, and R. Chellappa. Metareg: Towards domain generalization using meta-regularization. Advances in Neural Information Processing Systems, 31:998-1008, 2018.

[5] H. Behl, A. G. Baydin, and P. H. Torr. Alpha maml: Adaptive model-agnostic meta-learning. In 6th ICML Workshop on Automated Machine Learning, Thirty-sixth International Conference on Machine Learning (ICML 2019), Long Beach, CA, US, 2019.

[6] V. I. Bogachev. Measure theory, volume 1. Springer Science & Business Media, 2007.

[7] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3722-3731, 2017.

[8] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi. Domain generalization by solving jigsaw puzzles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2229-2238, 2019.

[9] P. Chattopadhyay, Y. Balaji, and J. Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In European Conference on Computer Vision, pages 301-318. Springer, 2020.

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597-1607. PMLR, 2020.

[11] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029, 2020.

[12] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8789-8797, 2018.

[13] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Diverse image synthesis for multiple domains. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8188-8197, 2020.

[14] R. T. d. Combes, H. Zhao, Y.-X. Wang, and G. Gordon. Domain adaptation with conditional distribution matching and generalized label shift. arXiv preprint arXiv:2003.04475, 2020.

[15] Z. Ding and Y. Fu. Deep domain generalization with structured low-rank constraint. IEEE Transactions on Image Processing, 27(1):304-313, 2017.

[16] Y. Du, J. Xu, H. Xiong, Q. Qiu, X. Zhen, C. G. Snoek, and L. Shao. Learning to learn with variational information bottleneck for domain generalization. In European Conference on Computer Vision, pages 200-216. Springer, 2020.

[17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. The Journal of Machine Learning Research, 17(1):2096-2030, 2016.

[18] M. Ghifary, W. B. Kleijn, M. Zhang, and D. Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In Proceedings of the IEEE International Conference on Computer Vision, pages 2551-2559, 2015.

[19] I. Goodfellow. Nips 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160, 2016.

[20] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. arXiv preprint arXiv:1406.2661, 2014.

[21] A. Grover, C. Chute, R. Shu, Z. Cao, and S. Ermon. Alignflow: Cycle consistent learning from multiple domains via normalizing flows. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 4028-4035, 2020.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770-778, 2016.

[23] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In International conference on machine learning, pages 1989-1998. PMLR, 2018.

[24] M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling. Diva: Domain invariant variational autoencoders. In Medical Imaging with Deep Learning, pages 322-348. PMLR, 2020.

[25] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba. Undoing the damage of dataset bias. In European Conference on Computer Vision, pages 158-171. Springer, 2012.

[26] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[27] N. Kodali, J. Abernethy, J. Hays, and Z. Kira. On convergence and stability of gans. arXiv preprint arXiv:1705.07215, 2017.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25:1097-1105, 2012.

[29] Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010. URL http://yann.1e cun. com/exdb/mnist/.

[30] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales. Deeper, broader and artier domain generalization. In International Conference on Computer Vision, 2017.

[31] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales. Learning to generalize: Meta-learning for domain generalization. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

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

[33] Y. Li, M. Gong, X. Tian, T. Liu, and D. Tao. Domain generalization via conditional invariant representations. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

[34] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao. Deep domain generalization via conditional invariant adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV), pages 624-639, 2018.

[35] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. arXiv preprint arXiv:1606.07536, 2016.

[36] I. Misra and L. v. d. Maaten. Self-supervised learning of pretext-invariant representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6707-6717, 2020.

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

[38] V. Piratla, P. Netrapalli, and S. Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In International Conference on Machine Learning, pages 7728-7738. PMLR, 2020.

[39] W. Rudin. Real and complex analysis. Tata McGraw-hill education, 2006.

[40] P. Russo, F. M. Carlucci, T. Tommasi, and B. Caputo. From source to target and back: symmetric bi-directional adaptive gan. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8099-8108, 2018.

[41] S. Seo, Y. Suh, D. Kim, J. Han, and B. Han. Learning to optimize domain specific normalization for domain generalization. arXiv preprint arXiv:1907.04275, 3(6):7, 2019.

[42] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi. Generalizing across domains via cross-gradient training. arXiv preprint arXiv:1804.10745, 2018.

[43] J. Shen, Y. Qu, W. Zhang, and Y. Yu. Wasserstein distance guided representation learning for domain adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

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

[45] A. K. Tanwani. Domain-invariant representation learning for sim-to-real transfer. arXiv preprint arXiv:2011.07589, 2020.

[46] H. Wang, Z. He, Z. C. Lipton, and E. P. Xing. Learning robust representations by projecting superficial statistics out. arXiv preprint arXiv:1903.06256, 2019.

[47] Z. Wang, M. Loog, and J. van Gemert. Respecting domain relations: Hypothesis invariance for domain generalization. arXiv preprint arXiv:2010.07591, 2020.

[48] Y. Zhang, T. Liu, M. Long, and M. Jordan. Bridging theory and algorithm for domain adaptation. In International Conference on Machine Learning, pages 7404-7413. PMLR, 2019.

[49] H. Zhao, R. T. Des Combes, K. Zhang, and G. Gordon. On learning invariant representations for domain adaptation. In International Conference on Machine Learning, pages 7523-7532. PMLR, 2019.

[50] S. Zhao, M. Gong, T. Liu, H. Fu, and D. Tao. Domain generalization via entropy regularization. Advances in Neural Information Processing Systems, 33, 2020.

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