GENERALIZING TO UNSEEN DOMAINS VIA DISTRIBUTION MATCHING

泛化到未见领域通过分布匹配

Isabela Albuquerque [1,*] , João Monteiro [1] , Mohammad Darvishi [2] , Tiago H. Falk [1] , Ioannis Mitliagkas [3]

Isabela Albuquerque [1,*] , João Monteiro [1] , Mohammad Darvishi [2] , Tiago H. Falk [1] , Ioannis Mitliagkas [3]

[1] INRS-EMT, Université du Québec
[1] INRS-EMT, 魁北克大学
[2] Faubert Lab, Université de Montréal
[2] Faubert 实验室, 蒙特利尔大学
[3] Mila &DIRO, Université de Montréal
[3] Mila & DIRO, 蒙特利尔大学

ABSTRACT

摘要

Supervised learning results typically rely on assumptions of i.i.d. data. Unfortunately, those assumptions are commonly violated in practice. In this work, we tackle such problem by focusing on domain generalization: a formalization where the data generating process at test time may yield samples from never-before-seen domains (distributions). Our work relies on the following lemma: by minimizing a notion of discrepancy between all pairs from a set of given domains, we also minimize the discrepancy between any pairs of mixtures of domains. Using this result, we derive a generalization bound for our setting. We then show that low risk over unseen domains can be achieved by representing the data in a space where (i) the training distributions are indistinguishable, and (ii) relevant information for the task at hand is preserved. Minimizing the terms in our bound yields an adversarial formulation which estimates and minimizes pairwise discrepancies. We validate our proposed strategy on standard domain generalization benchmarks, outperforming a number of recently introduced methods. Notably, we tackle a real-world application where the underlying data corresponds to multi-channel electroencephalography time series from different subjects, each considered as a distinct domain.

监督学习的结果通常依赖于独立同分布 (i.i.d.) 数据的假设。不幸的是，这些假设在实践中常常被违反。在本研究中，我们通过关注领域泛化来解决此问题: 一种形式化，其中测试时的数据生成过程可能产生来自前所未见领域 (分布) 的样本。我们的工作依赖于以下引理: 通过最小化给定领域集合中所有对之间的差异，我们也最小化任何领域混合对之间的差异。利用这一结果，我们推导出我们设置的泛化界限。然后，我们展示了通过在一个空间中表示数据可以实现对未见领域的低风险，其中 (i) 训练分布是不可区分的，并且 (ii) 与当前任务相关的信息得以保留。最小化我们界限中的项产生了一种对抗性形式，估计并最小化成对差异。我们在标准领域泛化基准上验证了我们提出的策略，超越了许多最近提出的方法。值得注意的是，我们解决了一个实际应用，其中基础数据对应于来自不同受试者的多通道脑电图时间序列，每个受试者被视为一个独特的领域。

1 Introduction

1 引言

The main assumption within the empirical risk minimization framework is that all examples used for training and testing predictors are independently drawn from a fixed distribution, i.e. the i.i.d. assumption. A number of generalization guarantees were derived upon that assumption and those results induced several algorithms for the solution of supervised learning problems. However, important limitations in this setting can be highlighted: i) the i.i.d. property is unverifiable [1] given that one doesn't have access to the data distribution, and ii) it doesn't account for distribution shifts which often occur in practice. Representative examples of these distribution shifts include changes in data acquisition conditions, such as illumination in images for object segmentation, or new data sources such as unseen speakers when performing speech recognition.

在经验风险最小化框架内的主要假设是，所有用于训练和测试预测器的示例都是从固定分布中独立抽取的，即 i.i.d. 假设。基于这一假设推导出了一些泛化保证，这些结果引导了几种解决监督学习问题的算法。然而，在这种设置中可以突出一些重要的局限性:i) i.i.d. 属性是不可验证的 [1]，因为无法访问数据分布；ii) 它没有考虑在实践中经常发生的分布变化。这些分布变化的代表性例子包括数据采集条件的变化，例如用于物体分割的图像中的光照变化，或在进行语音识别时出现的新数据源，例如未见过的说话者。

A number of alternative settings was then introduced in order to better cope with more realistic cases. Risk minimization under the domain adaptation setting, for instance, relaxes part of the i.i.d. assumption by allowing a source distribution (or domain) 1 as well as a different target distribution observed at test time. The domain adaptation results introduced in [2] showed that the generalization gap in terms of risk difference across the two considered distributions for a fixed predictor is upper bounded by a notion of distance measured between the training and testing domains. While less restrictive than the previous setting, the domain adaptation case is still limited in that only pairs of distributions seen during training are expected to yield low risk, and shifts beyond those domains will likely induce poor performance. Moreover, algorithms devised for this setting rely on access at training time to an unlabeled sample from the target distribution so that representations can be learned inducing invariance across train and target domains [3]. This is a limiting factor for practical applications where target domain data may be inaccessible; for example, a speech recognition service cannot be (re)trained on data obtained from every new speaker it observes.

为了更好地应对更现实的情况，随后引入了一些替代设置。例如，在领域适应设置下的风险最小化放宽了部分 i.i.d. 假设，允许在测试时观察到一个源分布 (或领域)1，以及一个不同的目标分布。[2] 中引入的领域适应结果表明，对于固定预测器，考虑的两个分布之间的风险差异的泛化差距由训练和测试领域之间测量的距离的概念上限。尽管比之前的设置限制更少，但领域适应案例仍然有限，因为仅期望在训练期间看到的分布对产生低风险，而超出这些领域的变化可能会导致性能不佳。此外，为该设置设计的算法依赖于在训练时访问来自目标分布的未标记样本，以便学习表示，从而在训练和目标领域之间诱导不变性 [3]。这对于实际应用是一个限制因素，因为目标领域数据可能无法访问；例如，语音识别服务无法在从每个新说话者获得的数据上进行 (重新) 训练。

A more general setting is often referred to as domain generalization [4]. In this case, it is assumed that a set of distributions over the data is available at training time. At test time, however, both observed distributions as well as unseen novel domains might appear, and a low risk is expected regardless of the underlying domain. More importantly, unlike domain adaptation in which the goal is to find a representation that aligns training data distributions with a specific target domain, domain generalization strategies aim at finding a representation space that yields good performance on novel distributions, unknown at training time. Recent work on domain generalization has included the use of data augmentation [5, 6] at training time, meta-learning to simulate domain shift [7], adding a self-supervised task to encourage an encoder to learn robust representations [8, 9], and learning domain-invariant representations [10], among other approaches.

更一般的设置通常被称为领域泛化 [4]。在这种情况下，假设在训练时可获得一组数据的分布。然而，在测试时，可能会出现观察到的分布以及未见过的新领域，并且无论基础领域如何，预期风险都很低。更重要的是，与领域适应不同，领域适应的目标是找到一种表示，使训练数据分布与特定目标领域对齐，而领域泛化策略旨在找到一种表示空间，以便在训练时未知的新分布上获得良好的性能。近期关于领域泛化的研究包括在训练时使用数据增强 [5, 6]、元学习以模拟领域转变 [7]、添加自监督任务以鼓励编码器学习鲁棒表示 [8, 9]，以及学习领域不变的表示 [10]，等等。

In this paper, we propose an innovation within the domain generalization setting. We first argue and prove that, given a set of distributions over data, if the distances measured between any pair of such distributions is small, so is the distance between mixtures obtained from the same set. This leads to the development of a bound on the risk measured against any distribution, and further shows that generalization can be expected if one considers distributions on the neighborhood of the "convex hull" [3] defined by the set of domains accessible during training. Inspired by these findings, we define an approach so that an encoder is enforced to map the data to a space where domain-dependent cues are filtered away while relevant information to the task of interest is conserved. While doing so, unlike standard domain adaptation approaches, no data from test distributions is observed.

---

*Correspondence to isabelamcalbuquerque@gmail.com
* 联系邮箱:isabelamcalbuquerque@gmail.com

[2] We use the terms domain, data distribution, and data source interchangeably throughout the text.

[2] 在本文中，我们将术语领域、数据分布和数据源互换使用。

在本文中，我们在领域泛化的设置中提出了一项创新。我们首先论证并证明，给定一组数据的分布，如果任何一对这样的分布之间测量的距离很小，那么从同一组获得的混合物之间的距离也很小。这导致了对任何分布的风险测量的界限的发展，并进一步表明，如果考虑在训练期间可访问的“凸包” [3] 的邻域上的分布，可以预期泛化。受到这些发现的启发，我们定义了一种方法，使编码器被强制映射数据到一个空间，在该空间中，领域相关的线索被过滤掉，而与感兴趣任务相关的信息得以保留。在此过程中，与标准领域适应方法不同，未观察到来自测试分布的数据。

We summarize our contributions in the following:

我们在以下内容中总结了我们的贡献:

1. We introduce assumptions on the data generating process tailored to the domain generalization setting, which we argue are more general than standard i.i.d. requirements and more likely to hold in practice. In other words, given a data sample, it is more likely that our assumptions will hold compared to the more restrictive i.i.d. property;

1. 我们引入了针对领域泛化设置的数据生成过程的假设，我们认为这些假设比标准的独立同分布 (i.i.d.) 要求更为一般，并且在实践中更可能成立。换句话说，给定一个数据样本，我们的假设成立的可能性比更严格的 i.i.d. 性质更高；

2. We prove a generalization bound for the risk over unseen domains and show that generalization can be expected for domains on the neighborhood of a notion of convex hull of distributions observed at training time;

2. 我们证明了未见领域风险的泛化界限，并展示了在训练时观察到的分布的凸包概念邻域内的领域可以期待泛化；

3. Aiming to minimize the terms of the introduced bound, we devise an adversarial approach so that pairwise domain divergences are estimated and minimized. In order to do so, several practical improvements are proposed on top of previous approaches for domain adaption including the use of random projection layers prior to domain discriminators.

3. 为了最小化引入的界限的各项，我们设计了一种对抗性方法，以便估计和最小化成对领域的差异。为此，我们在先前的领域适应方法基础上提出了若干实际改进，包括在领域判别器之前使用随机投影层。

4. We provide evidence through empirical evaluation showing that the proposed approach yields improvements relative to alternative methods across scenarios where different assumptions over the observed domains hold, including realistic cases where the labeling functions might shift.

4. 我们通过实证评估提供证据，表明所提出的方法在不同假设的观察领域场景中相对于替代方法取得了改进，包括标签函数可能发生变化的现实案例。

The remainder of this paper is organized as follows: In Section 2 we introduce the notation adopted within this work, and review related work and generalization guarantees. In section 3 we define the domain generalization setting and present our main results, as well as the resulting algorithm. Section 4 provides the experiment descriptions and their respective results. Section 5 concludes the paper.

本文的其余部分组织如下: 在第 2 节中，我们介绍了本工作中采用的符号，并回顾相关工作和泛化保证。在第 3 节中，我们定义了领域泛化设置，并呈现我们的主要结果以及相应的算法。第 4 节提供实验描述及其结果。第 5 节对本文进行总结。

2 Background

2 背景

2.1 Notation

2.1 符号

Let the data be represented by $\mathcal{X} \subset \mathbb{R}^D$, while $\mathcal{Y}$ denotes the label space. Examples correspond to a pair $(x, y) : x \in \mathcal{X}, y \in \mathcal{Y}$, such that $y = f_{\mathcal{D}}(x)$, and $f_{\mathcal{D}} : \mathcal{X} \to \mathcal{Y}$ is a deterministic labeling function.

设数据表示为 $\mathcal{X} \subset \mathbb{R}^D$, 而 $\mathcal{Y}$ 表示标签空间。示例对应于一对 $(x, y) : x \in \mathcal{X}, y \in \mathcal{Y}$, 使得 $y = f_{\mathcal{D}}(x)$, 并且 $f_{\mathcal{D}} : \mathcal{X} \to \mathcal{Y}$ 是一个确定性的标记函数。

A domain is defined as a tuple $\langle \mathcal{D}, f_{\mathcal{D}} \rangle$ where $\mathcal{D}$ corresponds to a probability distribution over $\mathcal{X}$. Moreover, we define a mapping $h : \mathcal{X} \to \mathcal{Y}$, such that $h \in \mathcal{H}$, where $\mathcal{H}$ is a set of candidate hypothesis, and finally define the risk R associated with a given hypothesis h on domain $\langle \mathcal{D}, f_{\mathcal{D}} \rangle$ as:

域被定义为一个元组 $\langle \mathcal{D}, f_{\mathcal{D}} \rangle$, 其中 $\mathcal{D}$ 对应于 $\mathcal{X}$ 上的概率分布。此外，我们定义一个映射 $h : \mathcal{X} \to \mathcal{Y}$, 使得 $h \in \mathcal{H}$, 其中 $\mathcal{H}$ 是一组候选假设，最后定义与给定假设 h 在域 $\langle \mathcal{D}, f_{\mathcal{D}} \rangle$ 上相关的风险 R 为:

$$R[h] = \mathbb{E}_{x\sim\mathcal{D}}\ell\left[h\left(x\right), f_{\mathcal{D}}\left(x\right)\right] \tag{1}$$

where the loss $\ell : \mathcal{Y} \times \mathcal{Y} \to R_+$ quantifies how different $h\left(x\right)$ is from the true labeling function $y = f_{\mathcal{D}}\left(x\right)$ for a given instance(x, y).

其中损失 $\ell : \mathcal{Y} \times \mathcal{Y} \to R_+$ 量化了给定实例 (x, y) 的 $h\left(x\right)$ 与真实标记函数 $y = f_{\mathcal{D}}\left(x\right)$ 之间的差异。

2.2 Related work

2.2 相关工作

A number of contributions under the domain generalization setting borrowed tools from causal inference to enforce the learned representations to be invariant across the different domains presented to the model at training time [11, 12, 13]. Other contributions such as [8] and [9] proposed different strategies to leverage self-supervised tasks to improve the out-of-distribution performance of a given model. Inspired by the domain adaptation literature [14, 2, 3], previous work on domain generalization also proposed to add a regularization term based on the minimization of a notion of divergence between the source domains to the empirical loss computed on the training data. This is the case of CIDDG [10], where class-specific domain classifiers are employed to induce the encoder to learn representations where the mismatch between the labels conditional distributions is minimized. Moreover, [15] proposed MMD-AAE, an approach that relies on an adversarial autoencoder used along with a maximum mean discrepancy penalty [16] to remove domain-specific information.

在域泛化设置下，许多贡献借用了因果推断的工具，以强制学习到的表示在训练时呈现给模型的不同域之间保持不变 [11, 12, 13]。其他贡献如 [8] 和 [9] 提出了不同的策略，以利用自监督任务来提高给定模型的分布外性能。受到域适应文献 [14, 2, 3] 的启发，之前关于域泛化的工作还建议在训练数据上计算的经验损失中添加一个基于源域之间的某种发散度最小化的正则化项。这就是 CIDDG [10] 的情况，其中使用类特定的域分类器来诱导编码器学习表示，从而最小化标签条件分布之间的不匹配。此外，[15] 提出了 MMD-AAE，这是一种依赖于对抗自编码器的方法，结合最大均值差异惩罚 [16] 来去除域特定信息。

Recent work has proposed settings where domain-shifts are simulated at training time by splitting the source domains into meta-train and meta-test sets [7, 17, 18, 19]. Strategies based on learning domain-invariant representations [4], data augmentation [6, 5], and on decomposing the model's parameters into domain-agnostic and domain-specific components [20] have also been introduced. Work on other settings with more restrictive assumptions than domain generalization are also related to our contribution. For example, recent work on multi-domain learning [21], a setting where multiple domains are available at training time and test data is drawn from the same distributions seen during training [22], also leveraged an adversarial approach to perform $\mathcal{H}$ -divergence minimization.

最近的研究提出了一种设置，其中通过将源领域分成元训练集和元测试集来模拟训练时的领域转移 [7, 17, 18, 19]。还引入了基于学习领域不变表示 [4]、数据增强 [6, 5] 和将模型参数分解为领域无关和领域特定组件 [20] 的策略。其他具有比领域泛化更严格假设的设置的研究也与我们的贡献相关。例如，最近关于多领域学习的研究 [21]，这是一个在训练时可以使用多个领域并且测试数据来自于训练期间看到的相同分布的设置 [22]，也利用了一种对抗性方法来执行 $\mathcal{H}$ -散度最小化。

A straightforward approach to extend the empirical risk minimization (ERM) setting for domain generalization would be to learn h minimizing the empirical risk $\widehat{R}[h]$ measured over all N_S source domains and hope generalization would be achieved to the target data, i.e.:

扩展领域泛化的经验风险最小化 (ERM) 设置的一个直接方法是学习 h ，以最小化在所有 N_S 源领域上测量的经验风险 $\widehat{R}[h]$ ，并希望能够实现对目标数据的泛化，即:

$$h = \arg\min \widehat{R} = \frac{1}{N_S}\sum_{j=1}^{N_S}\frac{1}{M_j}\sum_{i=1}^{M_j}\ell\left[h\left(x_i\right), f\left(x_i\right)\right]. \tag{2}$$

In fact, as will be discussed in more detail in next sections, such a rather simplistic approach often yields strong baselines.

实际上，正如将在接下来的部分中更详细讨论的那样，这种相对简单的方法通常会产生强大的基准。

[3] i.e., the set of all mixtures obtained from given distributions.

[3] 即，从给定分布中获得的所有混合的集合。

2.3 Generalization guarantees for domain adaptation

2.3 领域适应的泛化保证

We now state results from the domain adaptation literature which are relevant in the context of this work. The discussion in [23] established the theoretical foundations for studying cross-domain generalization properties for domain adaptation problems. They showed that, given a source domain $\mathcal{D}_S$ and a target domain $\mathcal{D}_T$, the risk of a given hypothesis $h \in \mathcal{H}$, $h : \mathcal{X} \to \{0, 1\}$, on the target is bounded by:

现在我们陈述与本工作相关的领域适应文献中的结果。[23] 中的讨论建立了研究领域适应问题的跨领域泛化特性的理论基础。他们表明，给定源领域 $\mathcal{D}_S$ 和目标领域 $\mathcal{D}_T$ ，给定假设 $h \in \mathcal{H}$ 、$h : \mathcal{X} \to \{0, 1\}$ 在目标上的风险是有界的:

$$R_T[h] \leq R_S[h] + d_{\mathcal{H}\Delta\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T] + \lambda, \tag{3}$$

where λ corresponds to the minimal total risk over both domains which can be achieved within a given hypothesis class $\mathcal{H}$. The term $d_{\mathcal{H}\Delta\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T]$ corresponds to the $\mathcal{H}$ -divergence introduced in [14] for a hypothesis class $\mathcal{H}\Delta\mathcal{H} = \{h(x) \oplus h'(x) \mid h, h' \in \mathcal{H}\}$, where $\oplus$ is the XOR function. The $\mathcal{H}$ -divergence between two distributions $\mathcal{D}_S$ and $\mathcal{D}_T$ is

其中 λ 对应于在给定假设类 $\mathcal{H}$ 内可以实现的两个领域的最小总风险。术语 $d_{\mathcal{H}\Delta\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T]$ 对应于在 [14] 中为假设类 $\mathcal{H}\Delta\mathcal{H} = \{h(x) \oplus h'(x) \mid h, h' \in \mathcal{H}\}$ 引入的 $\mathcal{H}$ -散度，其中 $\oplus$ 是异或函数。两个分布 $\mathcal{D}_S$ 和 $\mathcal{D}_T$ 之间的 $\mathcal{H}$ -散度是

defined as:

定义为:

$$d_{\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T] = 2 \sup_{\eta \in \mathcal{H}} \left| \Pr_{x \sim \mathcal{D}_S}[\eta(x) = 1] - \Pr_{x \sim \mathcal{D}_T}[\eta(x) = 1] \right|. \tag{4}$$

As discussed in [2], an estimate of $d_{\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T]$ can be directly computed from the error of a binary classifier trained to distinguish domains.

正如在 [2] 中讨论的，可以直接从训练用于区分领域的二元分类器的错误中计算 $d_{\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T]$ 的估计值。

3 Learning domain agnostic representations for domain generalization

3 学习领域无关的表示以实现领域泛化

3.1 Formalizing domain generalization

3.1 正式化领域泛化

We start by defining a set of assumptions over the data generating process considering the domain generalization case as well as the notion of risk we are concerned with. We then define $\mathfrak{D}$, referred to as meta-distribution, corresponding to a probability distribution over a countable set of possible domains. Under this view, a query for a data example consists of: i) sampling a domain from the meta-distribution, and ii) sampling a data point according to that particular domain. Such process is repeated m times so as to yield a training sample $(x^m \sim \mathfrak{D}^m, y^m)$. We remark the described model of data generating processes is sufficiently general so as to include the i.i.d. case (the meta-distribution yields a single domain) as well as the domain adaptation setting (if two domains are allowed), but further supports several other cases where multiple domains exist.

我们首先定义一组关于数据生成过程的假设，考虑领域泛化的情况以及我们所关心的风险概念。然后我们定义 $\mathfrak{D}$ ，称为元分布，对应于可数个可能领域上的概率分布。在这种视角下，对数据示例的查询包括:i) 从元分布中抽样一个领域，以及 ii) 根据该特定领域抽样一个数据点。这样的过程重复 m 次，以产生一个训练样本 $(x^m \sim \mathfrak{D}^m, y^m)$ 。我们注意到，所描述的数据生成过程模型足够一般，可以包括 i.i.d. 情况 (元分布产生单个领域) 以及领域适应设置 (如果允许两个领域)，但进一步支持多个领域存在的其他情况。

Once a finite train sample is collected, a set of N_S domains is observed. Each distribution $\mathcal{D}_S^i, i \in [N_S]$, in such set will be referred to as source domain. At test time, however, drawing samples from $\mathfrak{D}$ might yield data distributed according to new unseen domains. We then introduce extra notation and represent the set of possible domains unobserved while train data is acquired by $\mathcal{D}_U^j, j \in [N_U]$. The labeling rules corresponding to each domain are denoted as f_{S_i} and f_{U_i} , for the source and unseen domains, respectively. For the sake of clarity, we hereinafter omit the index from the notation corresponding to unseen domains whenever it can be inferred from the context.

一旦收集到有限的训练样本，就会观察到一组 N_S 域。在该集合中，每个分布 $\mathcal{D}_S^i, i \in [N_S]$ 将被称为源域。然而，在测试时，从 $\mathfrak{D}$ 中抽取样本可能会产生根据新的未见域分布的数据。我们接着引入额外的符号，并用 $\mathcal{D}_U^j, j \in [N_U]$ 表示在获取训练数据时未观察到的可能域的集合。与每个域对应的标记规则分别表示为 f_{S_i} 和 f_{U_i} ，对应于源域和未见域。为了清晰起见，以下我们在符号中省略未见域的索引，除非可以从上下文推断出。

We proceed and define a risk minimization framework similar to that corresponding to the i.i.d. setting: find the predictor $h^* \in \mathcal{H}$ that minimizes the meta-risk $R_{\mathfrak{D}}[h]$ defined as follows:

我们继续定义一个风险最小化框架，类似于对应于独立同分布 (i.i.d.) 设置的框架: 寻找能够最小化以下定义的元风险 $R_{\mathfrak{D}}[h]$ 的预测器 $h^* \in \mathcal{H}$ 。

$$h^* \in \operatorname*{argmin}_{h \in \mathcal{H}} R_{\mathfrak{D}}[h] \tag{5}$$

$$R_{\mathfrak{D}}[h] = \mathbb{E}_{\mathcal{D} \sim \mathfrak{D}}\left[\mathbb{E}_{x \sim \mathcal{D}}\left[\ell\left(h\left(x\right), f_{\mathcal{D}}\left(x\right)\right)\right]\right].$$

However, within the domain generalization setting, no information regarding possible test distributions is available at training time, which renders estimating $R_{\mathfrak{D}}[h]$ uninformative for a practical number of source domains. Moreover, we argue that no-free-lunch type of impossibility results may be used to conclude that it is impossible to generalize to any possible unknown distribution 4, so that one must assume something about the test domains in order to enable generalization. In the following results, we tackle this issue and introduce generalization guarantees for a particular set of domains lying close to the set of mixtures of source distributions, i.e., those observed once train data is collected.

然而，在域泛化设置中，训练时没有关于可能测试分布的信息，这使得对实际数量的源域估计 $R_{\mathfrak{D}}[h]$ 变得无信息。此外，我们认为无免费午餐类型的不可能性结果可以用来得出结论，即不可能泛化到任何可能的未知分布，因此必须对测试域做出某种假设，以便实现泛化。在以下结果中，我们解决了这个问题，并为一组接近源分布混合集合的特定域引入了泛化保证，即那些在收集训练数据后观察到的域。

3.2 Matching distributions in the convex hull

3.2 在凸包中匹配分布

Let a set S of source domains such that $|S| = N_S$ be denoted by $\mathcal{D}_S^i, i \in [N_S]$. The convex hull Λ_S of S is defined as the set of mixture distributions given by: $\Lambda_S = \left\{\overline{\mathcal{D}} : \overline{\mathcal{D}}\left(\cdot\right) = \sum_{i=1}^{N_S} \pi_i \mathcal{D}_S^i\left(\cdot\right), \pi_i \in \Delta_{N_S - 1}\right\}$, where Δ_{N_S-1} is the $N_S - 1$ -th dimensional simplex. The following lemma shows that for any pair of domains such that $\mathcal{D}', \mathcal{D}'' \in \Lambda_S^2$, the $\mathcal{H}$ -divergence between $\mathcal{D}'$ and $\mathcal{D}''$ is upper-bounded by the largest $\mathcal{H}$ -divergence measured between elements of S

设一组源领域 S ，使得 $|S| = N_S$ 表示为 $\mathcal{D}_S^i, i \in [N_S]$ 。源领域 S 的凸包 Λ_S 定义为由以下混合分布给出的集合: $\Lambda_S = \left\{\overline{\mathcal{D}} : \overline{\mathcal{D}}\left(\cdot\right) = \sum_{i=1}^{N_S} \pi_i \mathcal{D}_S^i\left(\cdot\right), \pi_i \in \Delta_{N_S - 1}\right\}$ ，其中 Δ_{N_S-1} 是 $N_S - 1$ 维单纯形。以下引理表明，对于任何一对领域，使得 $\mathcal{D}', \mathcal{D}'' \in \Lambda_S^2$ ，$\mathcal{H}$ -散度在 $\mathcal{D}'$ 和 $\mathcal{D}''$ 之间的上界由 S 的元素之间测量的最大 $\mathcal{H}$ -散度给出。

Lemma 1 (Bounding the $\mathcal{H}$ -divergence between domains in the convex hull of the sources). Let $d_{\mathcal{H}}\left[\mathcal{D}_S^i, \mathcal{D}_S^k\right] \leq \epsilon, \forall i, k \in [N_S]$. The following inequality holds for the $\mathcal{H}$ -divergence between any pair of domains $\mathcal{D}', \mathcal{D}'' \in \Lambda_S^{\mathcal{I}}$:

引理 1 (界定源领域凸包中领域之间的 $\mathcal{H}$ -散度)。设 $d_{\mathcal{H}}\left[\mathcal{D}_S^i, \mathcal{D}_S^k\right] \leq \epsilon, \forall i, k \in [N_S]$ 。对于任何一对领域 $\mathcal{D}', \mathcal{D}'' \in \Lambda_S^{\mathcal{I}}$ ，以下不等式成立:

$$d_{\mathcal{H}}\left[\mathcal{D}', \mathcal{D}''\right] \leq \epsilon \tag{6}$$

Proof. C.f. Supplementary material.

证明。参见补充材料。

We thus argue that if one minimizes the maximum pairwise $\mathcal{H}$ -divergence between source domains, which can be achieved by an encoding process that filters away domain discriminative cues, the $\mathcal{H}$ -divergence between any two domains in Λ_S also decreases.

因此我们认为，如果最小化源领域之间的最大成对 $\mathcal{H}$ -散度，这可以通过一种过滤掉领域区分线索的编码过程来实现，那么 Λ_S 中任何两个领域之间的 $\mathcal{H}$ -散度也会减少。

3.3 Generalizing to unseen domains

3.3 推广到未见领域

Now we turn our attention to the set of unseen distributions $\mathcal{D}_U^j, j \in [N_U]$, i.e., those in the support of the meta-distribution but not observed within the training sample. Given an unseen domain $\mathcal{D}_U$, we further introduce $\mathcal{D}_U$, the element within Λ_S which is closest to $\mathcal{D}_U$, i.e., $\overline{\mathcal{D}}_U$ is given by $\text{argmin}_{\pi_1,\ldots,\pi_{N_S}} d_{\mathcal{H}}\left[\mathcal{D}_U, \sum_{i=1}^{N_S} \pi_i \mathcal{D}_S^i\right]$. We now use Lemma 1 and previously proposed generalization bounds for the domain adaptation setting [24, 25] to derive a generalization bound for the risk $R_U[h]$.

现在我们将注意力转向未见分布的集合 $\mathcal{D}_U^j, j \in [N_U]$ ，即那些在元分布的支持中但未在训练样本中观察到的分布。给定一个未见领域 $\mathcal{D}_U$ ，我们进一步引入 $\mathcal{D}_U$ ，即在 Λ_S 中与 $\mathcal{D}_U$ 最近的元素，即 $\overline{\mathcal{D}}_U$ 由 $\text{argmin}_{\pi_1,\ldots,\pi_{N_S}} d_{\mathcal{H}}\left[\mathcal{D}_U, \sum_{i=1}^{N_S} \pi_i \mathcal{D}_S^i\right]$ 给出。我们现在使用引理 1 和之前提出的领域适应设置的推广界限 [24, 25] 来推导风险 $R_U[h]$ 的推广界限。

Theorem 1 (Upper-bounding the risk on unseen domains). Given the previous setup, let S be the set of source domains and $\mathcal{Y} = [0,1]$. The risk $R_U[h], \forall h \in \mathcal{H}$, for any unseen domain $\mathcal{D}_U$ such that $d_{\mathcal{H}}\left[\overline{\mathcal{D}}_U, \mathcal{D}_U\right] = \gamma$ is bounded as:

定理 1(对未见领域的风险上界)。在之前的设置下，设 S 为源领域的集合，设 $\mathcal{Y} = [0,1]$ 。对于任何未见领域 $\mathcal{D}_U$ ，其风险 $R_U[h], \forall h \in \mathcal{H}$ 被界定为:

$$R_U[h] \leq \sum_{i=1}^{N_S} \pi_i R_S^i[h] + \gamma + \epsilon + \min\left\{\mathbb{E}_{\overline{\mathcal{D}}_U}[|f_{S_\pi} - f_U|], \mathbb{E}_{\mathcal{D}_U}[|f_U - f_{S_\pi}|]\right\}, \tag{7}$$

where ϵ is the highest pairwise $\widetilde{\mathcal{H}}$ -divergence measured between pairs within $S, \widetilde{\mathcal{H}} = \{\text{sign}(|h(x) - h'(x)| - t) \mid h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$ and $f_{S_\pi}(x) = \sum_{i=i}^{N_S} \pi_i f_{S_i}(x)$ is the labeling function for any $x \in \text{Supp}\left(\overline{\mathcal{D}}_U\right)$ resulting from combining all f_{S_i} with weights $\pi_i, i \in [N_S]$, determined by $\overline{\mathcal{D}}_U$.

其中 ϵ 是在 $S, \widetilde{\mathcal{H}} = \{\text{sign}(|h(x) - h'(x)| - t) \mid h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$ 内部对每对之间测量的最高成对 $\widetilde{\mathcal{H}}$ -散度，$f_{S_\pi}(x) = \sum_{i=i}^{N_S} \pi_i f_{S_i}(x)$ 是由所有 f_{S_i} 结合权重 $\pi_i, i \in [N_S]$ 产生的任何 $x \in \text{Supp}\left(\overline{\mathcal{D}}_U\right)$ 的标记函数，由 $\overline{\mathcal{D}}_U$ 确定。

Proof. C.f. Supplementary material.

证明。参见补充材料。

Remark 1: Notice that the right-most term in Theorem 1 accounts for the mismatch between the labeling functions f_{S_π} and f_U , which reduces to 0 in most adopted scenarios within domain adaptation/generalization applications, since it is often considered that the covariate shift assumption holds [26]. Under such setting, the labeling functions are the same across all domains in the support of $\mathfrak{D}$, i.e. $f_{S_i} = f_{U_i} = f$ for all $i \in [N_S]$ and $j \in [N_U]$. Besides the covariate shift assumption, previous work on multi-source domain adaptation [27] considered the case where the unseen domain $\mathcal{D}_U$ can be represented as a mixture of the sources with weights $\pi_i, i \in [N_S]$, i.e. $\mathcal{D}_U = \overline{\mathcal{D}}_U$. When such assumption holds, the term indicated by γ in Theorem 1 will vanish. We thus re-state in Corollary 1 the previous result under such simplifying assumptions.

注 1: 注意到定理 1 中最右侧的项考虑了标记函数 f_{S_π} 和 f_U 之间的不匹配，在大多数领域适应/泛化应用中，这会减少到 0，因为通常认为协变量转移假设成立 [26]。在这种设置下，标记函数在 $\mathfrak{D}$ 的支持下的所有领域都是相同的，即对于所有 $i \in [N_S]$ 和 $j \in [N_U]$ ，都有 $f_{S_i} = f_{U_i} = f$ 。除了协变量转移假设，之前关于多源领域适应的工作 [27] 考虑了未见领域 $\mathcal{D}_U$ 可以表示为源的加权混合的情况，即 $\mathcal{D}_U = \overline{\mathcal{D}}_U$ 。

[4] For a fixed h , one can always define a distribution yielding high risk.

[4] 对于固定的 h ，总是可以定义一个产生高风险的分布。

当这种假设成立时，定理 1 中由 γ 指示的项将消失。因此，我们在推论 1 中重新陈述在这种简化假设下的先前结果。

Corollary 1 (Generalization to unseen domains within Λ_S under the covariate shift assumption). Let all domains within the the support of the meta-distribution $\mathfrak{D}$ have labeling function f . Let S be set of source domains and its convex-hull be denoted as Λ_S . The risk $R_U[h]$ of a hypothesis h on an unseen domain $\mathcal{D}_U \in \Lambda_S$, is upper-bounded by:

推论 1(在协变量转移假设下对未见领域的推广)。设元分布 $\mathfrak{D}$ 支持下的所有领域都有标记函数 f 。设 S 为源领域的集合，其凸包记作 Λ_S 。假设 h 在未见领域 $\mathcal{D}_U \in \Lambda_S$ 上的风险 $R_U[h]$ 被上界为:

$$R_U[h] \leq \sum_{i=1}^{N_S} \pi_i R_S^i[h] + \epsilon. \tag{8}$$

Proof. C.f. Supplementary material.

证明。参见补充材料。

Remark 2: Based on the introduced results we define in the following an algorithm relying solely on source data, unlike domain adaptation approaches. While the total source risk can be minimized as usual, ϵ can be minimized by encoding source data to a space where source domains are hard to distinguish. We remark that we empirically found (c.f. Sections 4.1.1 and 4.2) the proposed algorithm was able to succeed even in scenarios where the considered assumptions are not likely to hold.

备注 2: 基于所介绍的结果，我们在以下定义了一种仅依赖于源数据的算法，这与领域适应方法不同。虽然总源风险可以像往常一样被最小化，但 ϵ 可以通过将源数据编码到一个源领域难以区分的空间来最小化。我们注意到，我们在经验上发现 (参见第 4.1.1 节和第 4.2 节) 所提出的算法能够在考虑的假设不太可能成立的情况下成功。

Remark 3: We further highlight that the introduced results also provide insights regarding the importance of acquiring diverse datasets in practice when targeting domain generalization (and hint as to why data augmentation is often helpful). The more diverse a dataset is regarding the number of domains present at training time, more likely it is that an unseen distribution lies within the convex hull of the source domains (i.e. $\gamma \to 0$). Therefore, not only the amount of data is important to achieve better generalization on unseen domains, but also the diversity of the training data is crucial.

备注 3: 我们进一步强调，所介绍的结果还提供了关于在实践中获取多样化数据集的重要性的见解，尤其是在针对领域泛化时 (并暗示了数据增强为何通常有帮助)。数据集在训练时所包含的领域数量越多，未见分布位于源领域的凸包 (即 $\gamma \to 0$) 内的可能性就越大。因此，不仅数据量对在未见领域上实现更好的泛化很重要，训练数据的多样性也是至关重要的。

Remark 4: Another practical aspect worth remarking is that, even though our domain generalization setting is more general than ERM, Theorem 1 suggests that source domain labels should also be available, since they are required to estimate ϵ , which is not the case for ERM. However, collecting domain labels is inherent to the data acquisition procedure for several tasks and commonly available as meta-data in cases such as, speech recognition, where different speakers or recording devices can be viewed as different domains.

备注 4: 另一个值得注意的实际方面是，尽管我们的领域泛化设置比经验风险最小化 (ERM) 更为一般，但定理 1 表明源领域标签也应该是可用的，因为它们是估计 ϵ 所必需的，而这在 ERM 中并不适用。然而，收集领域标签是多个任务数据获取过程的固有部分，并且在某些情况下，如语音识别中，不同的说话者或录音设备可以视为不同的领域，通常作为元数据可用。

3.4 Practical contributions

3.4 实际贡献

Motivated by the previous results, we propose to design algorithms that minimize the terms in the bound in (14) that can be estimated even if only source data is observed, i.e., ϵ as well as the risks over the train sample. We thus aim at learning an encoder $E : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z} \subset \mathbb{R}^d$ preserves information relevant for separating classes, while removing domain-specific cues in such a way that it is harder to distinguish examples from different domains in comparison to the original space $\mathcal{X}$.

受到之前结果的启发，我们提出设计算法，以最小化 (14) 中可以估计的界限中的项，即使仅观察到源数据，也就是 ϵ 以及训练样本上的风险。因此，我们的目标是学习一个编码器 $E : \mathcal{X} \to \mathcal{Z}$ ，其中 $\mathcal{Z} \subset \mathbb{R}^d$ 保留与分类分离相关的信息，同时以这样的方式去除领域特定的线索，使得与原始空间 $\mathcal{X}$ 相比，更难区分来自不同领域的示例。

Efficiently estimating ϵ : Previous work on domain adaptation introduced strategies based on minimizing the empirical $\mathcal{H}$ -divergence between sources and a given target domain [3,24]. Instead, as per the discussion following Theorem 1, the domain generalization setting requires estimating pairwise $\mathcal{H}$ -divergences across all available sources, not considering target data of any sort. Naively extending previous methods to our case would require $\mathcal{O}\left(N_S^2\right)$ estimators, which is unpractical given real-world cases where several source domains are available. We thus propose to use one-vs-all classifiers. In this case, there is one domain discriminator per source domain and the k -th discriminator estimates $\sum_{l \neq k} d_{\mathcal{H}}\left[\mathcal{D}_S^k, \mathcal{D}_S^l\right]$, and improves the method to a number of $\mathcal{H}$ -divergence estimators linear on N_S .

高效估计 ϵ : 之前关于领域适应的工作引入了基于最小化源与给定目标领域之间的经验 $\mathcal{H}$ -散度的策略 [3,24]。然而，根据定理 1 后面的讨论，领域泛化设置要求估计所有可用源之间的成对 $\mathcal{H}$ -散度，而不考虑任何形式的目标数据。简单地将之前的方法扩展到我们的情况将需要 $\mathcal{O}\left(N_S^2\right)$ 估计器，这在现实世界中存在多个源领域的情况下是不切实际的。因此，我们建议使用一对多分类器。在这种情况下，每个源领域都有一个领域判别器，第 k 个判别器估计 $\sum_{l \neq k} d_{\mathcal{H}}\left[\mathcal{D}_S^k, \mathcal{D}_S^l\right]$ ，并将该方法改进为数量上与 N_S 成线性的 $\mathcal{H}$ -散度估计器。

Training: The proposed approach contains three main modules, all parameterized by neural networks: an encoder E with parameters ϕ , a task classifier C with parameters θ_C , and a set of $\mathcal{H}$ -divergence estimators D_k with parameters θ_k , $k \in [N_S]$. Intuitively, E attempts to minimize a classification loss $\mathcal{L}_C\left(\cdot ; \theta_C\right)$ (standard cross-entropy in our case) and empirical $\mathcal{H}$ -divergences, which is achieved through the maximization of domain discrimination losses, denominated $\mathcal{L}_k$. Each domain discriminator, on the other hand, aims at minimizing $\mathcal{L}_k$. The procedure for estimating ϕ, θ_T , and all θ_k ' s can be thus formulated as the following multiplayer minimax game:

训练: 所提出的方法包含三个主要模块，均由神经网络参数化: 一个编码器 E ，其参数为 ϕ ，一个任务分类器 C ，其参数为 θ_C ，以及一组 $\mathcal{H}$ -散度估计器 D_k ，其参数为 θ_k 、$k \in [N_S]$ 。直观上，E 试图最小化分类损失 $\mathcal{L}_C\left(\cdot ; \theta_C\right)$ (在我们的情况下为标准交叉熵) 和经验 $\mathcal{H}$ -散度，这通过最大化领域区分损失 $\mathcal{L}_k$ 来实现。另一方面，每个领域鉴别器旨在最小化 $\mathcal{L}_k$ 。因此，估计 ϕ, θ_T 的过程以及所有 θ_k 可以被表述为以下多玩家最小最大博弈:

$$\min_{\phi, \theta_C} \max_{\theta_1, \ldots, \theta_{N_S}} \mathcal{L}_C\left(C\left(E\left(x ; \phi\right) ; \theta_C\right), y_C\right) - \sum_{k=1}^{N_S} \mathcal{L}_k\left(D_k\left(E\left(x ; \phi\right) ; \theta_k\right), y_k\right) \tag{9}$$

where y_C corresponds to the task label for the example x , and y_k is equal to 1 in case $x \sim \mathcal{D}_S^k$, or 0 otherwise. Training is carried out with alternate updates. A pseudocode describing the training procedure is presented in Algorithm I. To further illustrate our proposed approach, we provide in Figure 2 (in the Supplementary material) a diagram showing the main components of the model in a case where three domains are available at training time.

其中 y_C 对应于示例 x 的任务标签，而 y_k 在 $x \sim \mathcal{D}_S^k$ 的情况下等于 1，否则为 0。训练通过交替更新进行。描述训练过程的伪代码在算法 I 中给出。为了进一步说明我们提出的方法，我们在图 2(在补充材料中) 中提供了一个图示，展示了在训练时可用三个领域的模型主要组件。

Algorithm 1 Generalizing to unseen Domains via Distribution Matching

算法 1 通过分布匹配推广到未见领域

1: Requires: classifier and encoder learning rate $\left(\beta_C\right)$, domain discriminators learning rate $\left(\beta_D\right)$, scaling (α) , mini-batch size(m).

1: 需要: 分类器和编码器学习率 $\left(\beta_C\right)$ ，领域鉴别器学习率 $\left(\beta_D\right)$ ，缩放 (α) ，小批量大小 (m)。

2: Initialize $\phi, \theta_C, \theta_1, \ldots, \theta_{N_S}$ as $\phi^0, \theta_C^0, \theta_1^0, \ldots, \theta_{N_S}^0$.

2: 将 $\phi, \theta_C, \theta_1, \ldots, \theta_{N_S}$ 初始化为 $\phi^0, \theta_C^0, \theta_1^0, \ldots, \theta_{N_S}^0$ 。

for $t = 1, \ldots$, number of iterations do

对于 $t = 1, \ldots$ ，迭代次数 do

Sample one mini-batch from each source domain $\left\{\left(x_1^i, y_C^i, y_1^i, \ldots, y_{N_S}^i\right)\right\}_{i=1}^m$

从每个源领域 $\left\{\left(x_1^i, y_C^i, y_1^i, \ldots, y_{N_S}^i\right)\right\}_{i=1}^m$ 中抽取一个小批量

Update domain discriminators

更新领域鉴别器

for $k = 1, \ldots, N_S$ do

对于 $k = 1, \ldots, N_S$ do

$\theta_k^t \leftarrow \theta_k^{t-1} + \frac{\beta_D}{N_S \cdot m} \sum_{i=1}^{N_S \cdot m} \nabla_{\theta_k} \mathcal{L}_k \left(D_k \left(E \left(x^i; \phi^{t-1} \right); \theta_k^{t-1} \right), y_k^i \right)$

end for

end for

Update task classifier

更新任务分类器

$\theta_C^t \leftarrow \theta_C^{t-1} + \frac{\beta_C}{N_S \cdot m} \sum_{i=1}^{N_S \cdot m} \nabla_{\theta_C} \mathcal{L}_C \left(C \left(E \left(x^i; \phi^{t-1} \right); \theta_C^{t-1} \right), y_C^i \right)$

Update encoder

更新编码器

$\phi^t \leftarrow \phi^{t-1} + \frac{\beta_C}{N_S \cdot m} (\sum_{i=1}^{N_S \cdot m} \alpha \nabla_\phi \mathcal{L}_C \left(C \left(E \left(x^i; \phi^{t-1} \right); \theta_C^{t-1} \right), y_C^i \right)$

$- (1 - \alpha) \nabla_{\theta_k} \mathcal{L}_k \left(D_k \left(E \left(x^i; \phi^{t-1} \right); \theta_k^t \right), y_k^i \right))$

end for

结束

Improving training stability: Previous work on domain adaptation/generalization [3, 10] proposed solving the problem stated in [9] using a gradient reversal layer [28]. We empirically observed such approach to be heavily dependent on the choice of hyperparameters in order for training to converge. We propose to augment the described adversarial approach using strategies originally utilized for stabilizing the training of generative adversarial networks with multiple discriminators [29, 30]. Namely, we include a random projection layer in the input of each domain discriminator with the goal of making examples from different distributions harder to be distinguished. In addition, we use the negative log-hypervolume instead of the summation in the game represented in (9) in order to assign more preference to solutions which decrease all pairwise divergences uniformly even in cases where there is a trade-off in their minimization. We refer to the proposed approach as G2DM (Generalizing to unseen Domains via Distribution Matching).

提高训练稳定性: 之前关于领域适应/泛化的工作 [3, 10] 提出了使用梯度反转层 [28] 来解决 [9] 中提出的问题。我们经验性地观察到，这种方法在训练收敛时对超参数的选择高度依赖。我们建议通过使用最初用于稳定具有多个鉴别器的生成对抗网络训练的策略来增强所描述的对抗方法 [29, 30]。具体而言，我们在每个领域鉴别器的输入中包含一个随机投影层，目的是使来自不同分布的样本更难以区分。此外，我们在 (9) 中的游戏表示中使用负对数超体积，而不是求和，以便更偏向于那些即使在最小化时存在权衡的情况下也能均匀减少所有成对散度的解决方案。我们将所提出的方法称为 G2DM(通过分布匹配推广到未见领域)。

Differences to multi-source domain adaptation: We further remark the differences between G2DM and previous adversarial approaches which are often employed in domain adaptation. Essentially, G2DM compares examples only from source domains to learn domain-agnostic representations, i.e., there is no notion of target distribution. Other settings such as [31, 32] are more restricted in that a particular distribution is targeted and data from that distribution is required, besides the source data we use in our case. Moreover, those approaches do not aim at matching source distributions and only consider $\mathcal{H}$ -divergences computed between each source domain and the given target. In the case of G2DM, on the other hand, the goal is to match source domain distributions to decrease ϵ , and thus only pairwise discrepancies between training domains are considered.

与多源领域适应的区别: 我们进一步指出 G2DM 与之前在领域适应中常用的对抗方法之间的区别。基本上，G2DM 仅比较源领域中的样本，以学习领域无关的表示，即没有目标分布的概念。其他设置如 [31, 32] 更为严格，因为它们针对特定分布，并且除了我们在本案例中使用的源数据外，还需要来自该分布的数据。此外，这些方法并不旨在匹配源分布，仅考虑在每个源领域与给定目标之间计算的 $\mathcal{H}$ -散度。另一方面，在 G2DM 的情况下，目标是匹配源领域分布以减少 ϵ ，因此仅考虑训练领域之间的成对差异。

4 Experimental Setup and Results

4 实验设置与结果

We design our empirical evaluation to validate G2DM in conditions where different assumptions are satisfied. In the first scenario, we chose experimental conditions such that the covariate shift assumption holds. For that, we employ G2DM on object recognition tasks. In this case, we aim to answer the following research questions: i) Can G2DM perform better than standard ERM under i.i.d. assumptions by using information of source domains only? ii) Where does G2DM performance stand in comparison to previously proposed domain generalization strategies? iii) Is G2DM indeed enforcing distribution matching across source and unseen domains? And iv) What is the effect on the resulting performance given by different access models to test distributions during training?

我们设计了实证评估，以验证 G2DM 在不同假设满足的条件下的有效性。在第一个场景中，我们选择了实验条件，使得协变量转移假设成立。为此，我们在物体识别任务中应用 G2DM。在这种情况下，我们旨在回答以下研究问题:i) G2DM 是否能够在独立同分布假设下，仅使用源域的信息，表现得比标准的经验风险最小化 (ERM) 更好? ii) G2DM 的性能与之前提出的领域泛化策略相比处于什么水平? iii) G2DM 是否确实在源域和未见域之间强制执行分布匹配? iv) 在训练过程中，不同的测试分布访问模型对最终性能的影响是什么?

We then evaluate whether G2DM is able to attain good out-of-domain performance even in the challenging scenario where the covariate shift assumption is likely to be violated. For that, we consider a real-world task that involves classifying electroencephalography (EEG) time series for affective state prediction, a burgeoning area within the human-machine systems field. In applications involving EEG data, subjects are often considered as distinct domains with different labeling functions [33]. Such shift can be attributed to environmental and anatomic factors such as device placement and scalp characteristics [34, 35].

然后我们评估 G2DM 是否能够在协变量转移假设可能被违反的挑战性场景中获得良好的域外性能。为此，我们考虑一个涉及对脑电图 (EEG) 时间序列进行分类以预测情感状态的真实世界任务，这是人机系统领域内一个新兴的研究方向。在涉及 EEG 数据的应用中，受试者通常被视为具有不同标记函数的独立域 [33]。这种转移可以归因于环境和解剖因素，例如设备放置和头皮特征 [34, 35]。

4.1 Evaluation under covariate shift

4.1 在协变量转移下的评估

The VLCS benchmark [36] is composed by examples from five overlapping classes from the VOC2007 [37], LabelMe [38], Caltech-101 [39], and SUN [40] datasets. PACS [20], in turn, consists of images distributed into seven classes from four different datasets: photo (P), art painting (A), cartoon (C), and sketch (S). We compare the performance of our proposed approach with a model trained with no mechanism to enforce domain generalization (referred to as ERM throughout this Section). Moreover, we consider the recently introduced invariant risk minimization (IRM) strategy [11] and include results reported in the literature achieved by Epi-FCR [19], JiGen [8] along with the ERM results they provided (referred to as ERM-JiGen), and MMD-AAE [15]. Finally, the adaptation of DANN for domain generalization reported in [19] was also considered. All such methods have as encoder the convolutional stack of AlexNet [41]. Further implementation details can be found in the Supplementary material.

VLCS 基准 [36] 由来自 VOC2007 [37]、LabelMe [38]、Caltech-101 [39] 和 SUN [40] 数据集的五个重叠类别的示例组成。PACS [20] 则由来自四个不同数据集的图像分布到七个类别中: 照片 (P)、艺术画作 (A)、卡通 (C) 和素描 (S)。我们将所提出的方法的性能与没有强制领域泛化机制的模型进行比较 (在本节中称为 ERM)。此外，我们考虑了最近提出的不变风险最小化 (IRM) 策略 [11]，并包括文献中由 Epi-FCR [19]、JiGen [8] 以及他们提供的 ERM 结果 (称为 ERM-JiGen) 和 MMD-AAE [15] 所报告的结果。最后，我们还考虑了 [19] 中报告的 DANN 在领域泛化中的适应。所有这些方法的编码器均为 AlexNet [41] 的卷积堆栈。更多实施细节可在补充材料中找到。

In Tables 1 and 2, we report the average best accuracy across three runs with different random seeds on the test partition of the unseen domain under a leave-one-domain-out validation scheme. Results show that G2DM outperforms ERM in terms of average performance across the unseen domains for both benchmarks, and supports the claim that leveraging source domain information as done by G2DM provides an improvement on generalization to unseen distributions in comparison to simply considering the i.i.d. requirement is satisfied. G2DM further presented better average performance when compared

to our implementation of IRM, as well as results from other methods previously reported in the literature. We finally highlight that G2DM showed an improvement in performance in more challenging domains [20], such as LabelMe and Sketch.

在表 1 和表 2 中，我们报告了在未见领域的测试分区上，使用不同随机种子进行三次运行的平均最佳准确率，采用的是留一领域验证方案。结果表明，G2DM 在两个基准的未见领域的平均性能上优于 ERM，并支持了 G2DM 通过利用源领域信息提供了相较于简单考虑独立同分布 (i.i.d.) 要求的改进。与我们实现的 IRM 以及文献中先前报告的其他方法的结果相比，G2DM 还表现出更好的平均性能。最后，我们强调 G2DM 在更具挑战性的领域 [20] 中表现出性能提升，例如 LabelMe 和 Sketch。

Table 1: Classification accuracy (%) on VLCS datasets for Table 2: Classification accuracy (%) on PACS datasets for

表 1: 在 VLCS 数据集上的分类准确率 (%) 表 2: 在 PACS 数据集上的分类准确率 (%)

models trained with leave-one-domain-out validation.

使用留一领域验证训练的模型。

models trained with leave-one-domain-out validation.

使用留一领域验证训练的模型。

Unseen domain (→)	V	L	C	S	Average	Unseen domain (→)	P	A	C	S	Average
DANN	66.40	64.00	92.60	63.60	71.70	DANN	88.10	63.20	67.50	57.00	69.00
MMD-AAE	67.70	62.60	94.40	64.40	72.28	Epi-FCR	86.10	64.70	72.30	65.00	72.00
Epi-FCR	67.10	64.30	94.10	65.90	72.90	JiGen	89.00	67.63	71.71	65.18	73.38
JiGen ERM - JiGen	70.62 71.96	60.90 59.18	96.93 96.93	64.30 62.57	73.19 72.66	ERM - JiGen	89.98	66.68	69.41	60.02	71.52
IRM	72.16	62.36	98.35	67.82	75.17	IRM	89.97	64.84	71.16	63.63	72.39
ERM	73.44	60.44	97.88	67.92	74.92	ERM	90.02	64.86	70.18	61.40	71.61
G2DM	71.14	67.63	95.52	69.37	75.92	G2DM	88.12	66.60	73.36	66.19	73.55

未见领域 (→)	V	L	C	S	平均	未见领域 (→)	P	A	C	S	平均
DANN	66.40	64.00	92.60	63.60	71.70	DANN	88.10	63.20	67.50	57.00	69.00
MMD-AAE	67.70	62.60	94.40	64.40	72.28	Epi-FCR	86.10	64.70	72.30	65.00	72.00
Epi-FCR	67.10	64.30	94.10	65.90	72.90	JiGen	89.00	67.63	71.71	65.18	73.38
JiGen ERM - JiGen	70.62 71.96	60.90 59.18	96.93 96.93	64.30 62.57	73.19 72.66	ERM - JiGen	89.98	66.68	69.41	60.02	71.52
IRM	72.16	62.36	98.35	67.82	75.17	IRM	89.97	64.84	71.16	63.63	72.39
ERM	73.44	60.44	97.88	67.92	74.92	ERM	90.02	64.86	70.18	61.40	71.61
G2DM	71.14	67.63	95.52	69.37	75.92	G2DM	88.12	66.60	73.36	66.19	73.55

4.1.1 Checking $\mathcal{H}$ -divergences across sources and unseen domains

4.1.1 检查 $\mathcal{H}$ -发散在源和未见领域之间

We now investigate whether cross-domain $\mathcal{H}$ -divergences are being in fact reduced by G2DM. We use ERM as a baseline as it does not include any mechanism to enforce distribution matching. We estimate $\mathcal{H}$ -divergences by computing the proxy pairwise $\mathcal{A}$ -distance [2] for each pair of domains on the PACS benchmark. Classifiers are trained on top of the representations $\mathcal{Z}$ obtained with ERM and G2DM. We show in Figures 1 the differences in estimated discrepancies between ERM and G2DM for each unseen domain. Each entry corresponds to a pair of domains indicated in the row and the column and positive values indicate that G2DM decreased the corresponding pairwise $\mathcal{A}$ -distance in comparison to ERM. Notice that the diagonals are left blank as we do not compute the classification accuracy between the same domains.

我们现在研究 G2DM 是否确实减少了跨领域的 $\mathcal{H}$ -发散。我们使用 ERM 作为基线，因为它不包括任何强制分布匹配的机制。我们通过计算 PACS 基准上每对领域的代理成对 $\mathcal{A}$ -距离 [2] 来估计 $\mathcal{H}$ -发散。分类器是在 ERM 和 G2DM 获得的表示 $\mathcal{Z}$ 之上训练的。我们在图 1 中展示了 ERM 和 G2DM 在每个未见领域之间估计的差异。每个条目对应于行和列中指示的一对领域，正值表示 G2DM 相较于 ERM 减少了相应的成对 $\mathcal{A}$ -距离。请注意，对角线留空，因为我们不计算相同领域之间的分类准确率。

We observe that, apart from the case where 'photo' is the test domain, G2DM was in fact able to better match most of the source distributions, thus yielding smaller ϵ which favours generalization as predicted by Theorem 1. Notably, we highlight that although our proposed approach has no access to data from the unseen domain at training time and, therefore, does not directly implement a strategy to decrease the divergence between the unseen domain and the convex hull of the sources (i.e. γ), the results presented in Figure 1 show that the estimated pairwise $\mathcal{H}$ -divergence between the unseen domain and sources also decreased in most of the considered cases.

我们观察到，除了“照片”是测试领域的情况外，G2DM 确实能够更好地匹配大多数源分布，从而产生较小的 ϵ ，这有利于如定理 1 所预测的泛化。值得注意的是，我们强调，尽管我们提出的方法在训练时无法访问未见领域的数据，因此并未直接实施减少未见领域与源的凸包之间的发散 (即 γ) 的策略，但图 1 中呈现的结果显示，在大多数考虑的情况下，未见领域与源之间的估计成对 $\mathcal{H}$ -发散也减少了。

In fact, the only mechanism the encoder has in order to reduce ϵ corresponds to learning how to filter domain information from the data, in the sense that once samples from two distinct distributions are encoded, one cannot distinguish from which distribution each sample came from. Observed results thus suggest that such encoder also removes domain information from the unseen distributions observed at test time, preventing the learning algorithm to yield a high γ .

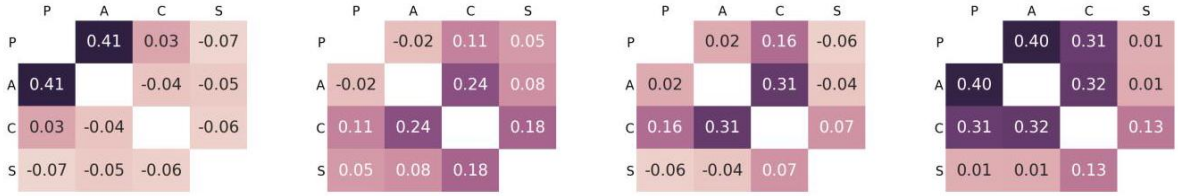
事实上，编码器减少 ϵ 的唯一机制在于学习如何从数据中过滤领域信息，这意味着一旦来自两个不同分布的样本被编码，就无法区分每个样本来自哪个分布。因此，观察到的结果表明，这种编码器还会从在测试时观察到的未见分布中去除领域信息，从而防止学习算法产生高 γ 。

4.1.2 The effect of different access methods to test data during training

4.1.2 训练期间不同访问方法对测试数据的影响

Results of previous experiments correspond to an optimistic scenario where data from the unseen domain is made available for selecting the best performing model. This is not the case in practice since varying unseen distributions might appear. In Table 3, we compare results obtained further considering different access methods to the test data. Namely, we consider the case where no access to the unseen distribution is allowed and only source data can be used

先前实验的结果对应于一个乐观的情景，即未见领域的数据可用于选择表现最佳的模型。然而，实际上可能会出现不同的未见分布。在表 3 中，我们比较了在考虑不同访问测试数据的方法后获得的结果。具体而言，我们考虑不允许访问未见分布，仅可以使用源数据的情况。



(a) Photo. (b) Art. (c) Cartoon. (d) Sketch.

Figure 1: Differences between estimated pairwise $\mathcal{H}$ -divergences under ERM and G2DM on PACS (captions denote unseen domains). Higher values indicate that G2DM better matched domains. Overall, G2DM is able to decrease pairwise discrepancies.

图 1: 在 PACS 上，ERM 和 G2DM 下估计的成对 $\mathcal{H}$ -散度之间的差异 (标题表示未见领域)。较高的值表明 G2DM 更好地匹配了领域。总体而言，G2DM 能够减少成对差异。

in order to define stopping criteria. In such cases, both validation accuracy and training loss computed on left-out in-domain data are employed (referred in Table 3 as source accuracy and source loss, respectively). Moreover, as a reference of performance, we further report the accuracy achieved assuming access to unseen domain data during training in order to select the best model (referred in Table [3] as unseen accuracy). For comparison, we further present the performance reported by [10] for CIDDG, since a stopping criterion using solely data from source domains was employed in that case. We observe that, when using the training loss as stopping criterion, our strategy outperforms CIDDG for almost all domains, while the baseline performance severely degrades when 'sketch' is the unseen domain.

为了定义停止标准。在这种情况下，使用在保留的领域内数据上计算的验证准确率和训练损失 (在表 3 中分别称为源准确率和源损失)。此外，作为性能的参考，我们进一步报告在训练期间假设可以访问未见领域数据所获得的准确率，以选择最佳模型 (在表 [3] 中称为未见准确率)。为了比较，我们进一步呈现 [10] 对 CIDDG 报告的性能，因为在该情况下仅使用源领域数据的停止标准。我们观察到，当使用训练损失作为停止标准时，我们的策略在几乎所有领域上都优于 CIDDG，而当 “草图” 是未见领域时，基线性能严重下降。

As an alternative to AlexNet, we further evaluate the performance of the proposed approach using the convolutional stack of a ResNet-18 [42], since it has shown promising results in recent work [8]. We compare our approach with JiGer[5] adopting the same previously-discussed test data access methods for both approaches. We further report in Table 3 the performance obtained by JiGen as reported in [8] although it is unclear which stopping criteria were adopted for that case. We observe that replacing

AlexNet by ResNet-18 yields a more stable average performance across stopping criteria. Based on the results obtained with AlexNet, we remark that results might be too optimistic/pessimistic depending on the assumed access method to unseen distributions, and as such, in order to allow fair comparison between different approaches, the performance across different access methods should be reported.

作为 AlexNet 的替代方案，我们进一步评估了使用 ResNet-18 的卷积堆栈 [42] 的提议方法的性能，因为它在最近的工作中显示出了良好的结果 [8]。我们将我们的方法与 JiGer[5] 进行比较，采用相同的先前讨论过的测试数据访问方法。我们在表 3 中进一步报告了 JiGen 的性能，如 [8] 中所述，尽管不清楚在该情况下采用了哪些停止标准。我们观察到，用 ResNet-18 替换 AlexNet 使得在停止标准下的平均性能更加稳定。基于使用 AlexNet 获得的结果，我们指出，结果可能过于乐观/悲观，具体取决于假设的对未见分布的访问方法，因此，为了允许不同方法之间的公平比较，应报告不同访问方法下的性能。

Table 3: Accuracy (%) on PACS with different stopping criteria.

表 3: 在不同停止标准下 PACS 的准确率 (%)。

Method	Criterion	P	A	C	S	Average
AlexNet						
CIDDG 10	From 10	78.65	62.70	69.73	64.45	68.88
G2DM	Source acc.	85.33	57.76	69.71	49.45	65.56
	Source loss	87.37	66.70	70.26	50.98	68.82
	Unseen acc.	88.80	66.70	73.29	65.03	73.45
ResNet-18						
JiGen [8]	Source acc.	95.83	78.52	73.31	69.14	79.20
	Source loss	95.83	78.89	73.32	70.73	79.69
	Unseen acc.	96.11	79.56	74.25	71.00	80.23
	From [8]	96.03	79.42	75.25	71.35	80.51
G2DM	Source acc.	93.70	79.22	76.34	75.14	81.10
	Source loss	93.75	77.78	75.54	77.58	81.16
	Unseen acc.	94.63	81.44	79.35	79.52	83.34

方法	标准	P	A	C	S	平均
AlexNet						
CIDDG 10	从 10	78.65	62.70	69.73	64.45	68.88
G2DM	源准确率	85.33	57.76	69.71	49.45	65.56
	源损失	87.37	66.70	70.26	50.98	68.82
	未见准确率	88.80	66.70	73.29	65.03	73.45
ResNet-18						
JiGen [8]	源准确率	95.83	78.52	73.31	69.14	79.20
	源损失	95.83	78.89	73.32	70.73	79.69
	未见准确率	96.11	79.56	74.25	71.00	80.23
	来源于 [8]	96.03	79.42	75.25	71.35	80.51
G2DM	源准确率	93.70	79.22	76.34	75.14	81.10
	源损失	93.75	77.78	75.54	77.58	81.16
	未见准确率	94.63	81.44	79.35	79.52	83.34

4.2 Evaluation beyond the covariate shift assumption

4.2 超越协变量转移假设的评估

We proceed to evaluate G2DM on a unfavorable scenario where the covariate shift is unlikely to hold. The goal of the selected task is to perform affective state estimation with three classes (positive, neutral, or negative) based on EEG signals from the SEED dataset [43] collected from 15 subjects. We use the architecture described in [44] for both G2DM and ERM. For each subject left out for testing, we use 10 out of the remaining 14 domains for training and use the other 4 as validation data.

我们继续评估 G2DM 在一个不利场景中的表现，在该场景中，协变量转移不太可能成立。所选任务的目标是基于来自 15 名受试者的 SEED 数据集 [43] 的 EEG 信号进行情感状态估计，分为三类 (积极、中性或消极)。我们使用 [44] 中描述的架构来实现 G2DM 和 ERM。对于每个被排除用于测试的受试者，我们使用剩余 14 个领域中的 10 个进行训练，另 4 个作为验证数据。

We report in Table 4 the classification accuracy (%) averaged across all unseen subjects and three independent training runs. Under source data validation, the reported performance was computed on the epoch of highest accuracy on the

我们在表 4 中报告了所有未见受试者和三次独立训练运行的分类准确率 (%) 的平均值。在源数据验证下，报告的性能是在最高准确率的时期计算得出的。

Table 4: Average accuracy (%) on the SEED dataset across 15 subjects. Semi-privileged approaches correspond to the best performing model under the domain generalization setting. Privileged baselines (domain adaptation setting) have access to unseen domain data at training time.

表 4: 在 15 名受试者中，SEED 数据集的平均准确率 (%)。半特权方法对应于在领域泛化设置下表现最佳的模型。特权基线 (领域适应设置) 在训练时可以访问未见领域数据。

Setting	Method	Average accuracy (%)
Domain generalization	Source data validation	
	ERM	51.98
	G2DM	55.77
	Semi-privileged	
	ERM	56.82
	G2DM	60.26
Privileged baselines		
Domain adaptation	DAN 145.46	50.28
	DANN 13.46	55.87
	MDAN [24.46	56.65
	MDMN [46]	60.59

设置	方法	平均准确率 (%)
域泛化	源数据验证	
	风险管理	51.98
	G2DM	55.77
	半特权	
	风险管理	56.82
	G2DM	60.26
特权基线		
领域适应	DAN 145.46	50.28
	DANN 13.46	55.87
	MDAN [24.46	56.65
	MDMN [46]	60.59

validation partition. The results under semi-privileged were obtained on the epoch of highest accuracy on the unseen subject data. The comparison between G2DM and ERM shows that even in this challenging case where the mismatch between labeling functions is not negligible, G2DM is able to successfully leverage the available domain information (which in this case comes with no additional effort at the data collection) and presents an improvement of more than 3.4% in accuracy in comparison to ERM in both considered scenarios for the DG setting.

验证分区。半特权下的结果是在未见受试者数据的最高准确率时期获得的。G2DM 和 ERM 之间的比较表明，即使在标签函数之间的差异不可忽视的情况下，G2DM 也能够成功利用可用的领域信息 (在这种情况下，数据收集没有额外的努力)，并在 DG 设置的两种考虑场景中，与 ERM 相比，准确率提高了超过 3.4% 。

Comparison with domain adaptation strategies: We further report in Table 4 results obtained by domain adaptation strategies (DA). Such methods, reported in Table 4 under privileged baselines, are privileged in the sense that unlabeled data belonging to the unseen domain (unknown in our case) is used to adapt representations at training time in order to yield subject-specific models. When comparing the DA strategies with our proposed domain generalization (DG) approach, we remark that DG strategies aim to obtain domain-agnostic models, as opposed to DA methods which target a specific distribution. As such, one would expect DA approaches to achieve better performance than DG. However, we observe G2DM's performance to be on par with, or even better than, some of the considered DA strategies. We conjecture a larger number of source domains available at training time would decrease the gap between DG and DA even further; i.e., it would be more likely that unseen domains are exactly represented in the convex hull of the sources yielding low γ (c.f. Theorem 1).

[5] Results are generated using JiGen authors' source code (https://github.com/fmcarlucci/JigenDG).

[5] 结果是使用 JiGen 作者的源代码生成的 (https://github.com/fmcarlucci/JigenDG)。

与领域适应策略的比较: 我们在表 4 中进一步报告了通过领域适应策略 (DA) 获得的结果。这些方法在表 4 中被列为特权基线，之所以被称为特权，是因为使用了属于未见领域 (在我们案例中未知) 的未标记数据来在训练时调整表示，以便产生特定于主题的模型。在将 DA 策略与我们提出的领域泛化 (DG) 方法进行比较时，我们注意到 DG 策略旨在获得领域无关的模型，而 DA 方法则针对特定的分布。因此，人们可以预期 DA 方法的性能会优于 DG。然而，我们观察到 G2DM 的性能与一些考虑的 DA 策略相当，甚至更好。我们推测，在训练时可用的源领域数量较多将进一步缩小 DG 和 DA 之间的差距；即，未见领域更有可能在源的凸包中被准确表示，从而产生低 γ (参见定理 1)。

5 Conclusion

5 结论

In this paper, we tackled the domain generalization problem and showed that generalization can be achieved in the neighborhood of the set of mixtures of distributions observed during training. Based on this result, we introduced G2DM, an efficient approach in yielding invariant representations across unseen distributions. Our method employs multiple one-vs-all domain discriminators, such that pairwise divergences between source distributions are estimated and minimized at training time. We provide empirical evidence supporting the claim that making use of domain information improves performance relative to standard settings relying on i.i.d. requirements. Moreover, the introduced approach outperformed recent methods which also leverage domain labels. We further showed that our proposed method resulted in strong results on a realistic setting, with performance comparable to privileged systems tailored to test distributions.

在本文中，我们解决了领域泛化问题，并展示了在训练期间观察到的分布混合集的邻域中可以实现泛化。基于这一结果，我们引入了 G2DM，这是一种在未见分布中产生不变表示的有效方法。我们的方法采用多个一对多的领域鉴别器，以便在训练时估计并最小化源分布之间的成对散度。我们提供了实证证据，支持利用领域信息相对于依赖于独立同分布 (i.i.d.) 要求的标准设置提高性能的主张。此外，所提出的方法在性能上优于最近的也利用领域标签的方法。我们进一步展示了我们提出的方法在现实设置中取得了强劲的结果，其性能可与专门针对测试分布的特权系统相媲美。

References

参考文献

[1] J. Langford, "Tutorial on practical prediction theory for classification," Journal of machine learning research, vol. 6, no. Mar, pp. 273-306, 2005.

[2] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in Advances in neural information processing systems, 2007, pp. 137-144.

[3] 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, vol. 17, no. 1, pp. 2096-2030, 2016.

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

[5] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi, "Generalizing across domains via cross-gradient training," in International Conference on Learning Representations, 2018. [Online]. Available: https://openreview.net/forum?id=r1Dx7fbCW

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

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

[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 Conference on Computer Vision and Pattern Recognition, 2019, pp. 2229-2238.

[9] I. Albuquerque, N. Naik, J. Li, N. Keskar, and R. Socher, "Improving out-of-distribution generalization via multi-task self-supervised pretraining," arXiv preprint arXiv:2003.13525, 2020.

[10] 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), 2018, pp. 624-639.

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

[12] D. Mahajan, S. Tople, and A. Sharma, "Domain generalization using causal matching," arXiv preprint arXiv:2006.07500, 2020.

[13] K. Ahuja, K. Shanmugam, K. Varshney, and A. Dhurandhar, "Invariant risk minimization games," arXiv preprint arXiv:2002.04692, 2020.

[14] D. Kifer, S. Ben-David, and J. Gehrke, "Detecting change in data streams," in Proceedings of the Thiriteth international conference on Very large data bases-Volume 30. VLDB Endowment, 2004, pp. 180-191.

[15] H. Li, S. Jialin Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5400-5409.

[16] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," The Journal of Machine Learning Research, vol. 13, no. 1, pp. 723-773, 2012.

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

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

[19] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, "Episodic training for domain generalization," arXiv preprint arXiv:1902.00113, 2019.

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

[21] A. Schoenauer-Sebag, L. Heinrich, M. Schoenauer, M. Sebag, L. F. Wu, and S. J. Altschuler, "Multi-domain adversarial learning," arXiv preprint arXiv:1903.09239, 2019.

[22] M. Dredze, A. Kulesza, and K. Crammer, "Multi-domain learning by confidence-weighted parameter combination," Machine Learning, vol. 79, no. 1-2, pp. 123-149, 2010.

[23] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," Machine learning, vol. 79, no. 1-2, pp. 151-175, 2010.

[24] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," in Advances in Neural Information Processing Systems, 2018, pp. 8559-8570.

[25] H. Zhao, R. T. Des Combes, K. Zhang, and G. Gordon, "On learning invariant representations for domain adaptation," in International Conference on Machine Learning, 2019, pp. 7523-7532.

[26] S. Ben-David, T. Lu, T. Luu, and D. Pál, "Impossibility theorems for domain adaptation," in International Conference on Artificial Intelligence and Statistics, 2010, pp. 129-136.

[27] J. Hoffman, M. Mohri, and N. Zhang, "Algorithms and theory for multiple-source adaptation," in Advances in Neural Information Processing Systems, 2018, pp. 8246-8256.

[28] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in International conference on machine learning. PMLR, 2015, pp. 1180-1189.

[29] B. Neyshabur, S. Bhojanapalli, and A. Chakrabarti, "Stabilizing gan training with multiple random projections," arXiv preprint arXiv:1705.07831, 2017.

[30] I. Albuquerque, J. Monteiro, T. Doan, B. Considine, T. Falk, and I. Mitliagkas, "Multi-objective training of generative adversarial networks with multiple discriminators," in International Conference on Machine Learning, 2019, pp. 202-211.

[31] S. Sun, H. Shi, and Y. Wu, "A survey of multi-source domain adaptation," Information Fusion, vol. 24, pp. 84-92, 2015.

[32] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1406-1415.

[33] I. Albuquerque, J. Monteiro, O. Rosanne, A. Tiwari, J.-F. Gagnon, and T. H. Falk, "Cross-subject statistical shift estimation for generalized electroencephalography-based mental workload assessment," in 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). IEEE, 2019, pp. 3647-3653.

[34] D. Wu, V. J. Lawhern, and B. J. Lance, "Reducing bci calibration effort in rsvp tasks using online weighted adaptation regularization with source domain selection," in 2015 International Conference on

Affective Computing and Intelligent Interaction (ACII). IEEE, 2015, pp. 567-573.

[35] C.-S. Wei, Y.-P. Lin, Y.-T. Wang, C.-T. Lin, and T.-P. Jung, "A subject-transfer framework for obviating inter-and intra-subject variability in eeg-based drowsiness detection," NeuroImage, vol. 174, pp. 407-419, 2018.

[36] C. Fang, Y. Xu, and D. N. Rockmore, "Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias," in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1657-1664.

[37] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," International journal of computer vision, vol. 88, no. 2, pp. 303-338, 2010.

[38] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," International journal of computer vision, vol. 77, no. 1-3, pp. 157-173, 2008.

[39] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007.

[40] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, "Exploiting hierarchical context on a large database of object categories," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010, pp. 129-136.

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

[42] 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, 2016, pp. 770-778.

[43] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," IEEE Transactions on Autonomous Mental Development, vol. 7, no. 3, pp. 162-175, 2015.

[44] Y. Li, K. Dzirasa, L. Carin, D. E. Carlson et al., "Targeting eeg/lfp synchrony with neural nets," in Advances in Neural Information Processing Systems, 2017, pp. 4620-4630.

[45] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in International Conference on Machine Learning, 2015, pp. 97-105.

[46] Y. Li, D. E. Carlson et al., "Extracting relationships by multi-domain matching," in Advances in Neural Information Processing Systems, 2018, pp. 6798-6809.

[47] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision (IJCV), vol. 115, no. 3, pp. 211-252, 2015.

[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818-2826.

Supplementary Material

补充材料

A Proof of Lemma 1

引理 1 的证明

Lemma 1. Let $d_{\mathcal{H}}\left[\mathcal{D}_S^i, \mathcal{D}_S^k\right] \leq \epsilon, \forall i, k \in [N_S]$. The following inequality holds for the $\mathcal{H}$ -divergence between any pair of domains $\mathcal{D}', \mathcal{D}'' \in \Lambda_S^2$:

引理 1. 设 $d_{\mathcal{H}}\left[\mathcal{D}_S^i, \mathcal{D}_S^k\right] \leq \epsilon, \forall i, k \in [N_S]$ 。以下不等式适用于任意一对领域 $\mathcal{D}', \mathcal{D}'' \in \Lambda_S^2$ 之间的 $\mathcal{H}$ -散度:

$$d_{\mathcal{H}}\left[\mathcal{D}', \mathcal{D}''\right] \leq \epsilon \tag{10}$$

Proof. Consider two unseen domains, $\mathcal{D}'_U$ and $\mathcal{D}''_U$ on the convex-hull Λ_S of N_S source domains with support Ω . Consider also $\mathcal{D}'_U(\cdot) = \sum_{k=1}^{N_S} \pi_k \mathcal{D}_S^k(\cdot)$ and $\mathcal{D}''_U(\cdot) = \sum_{l=1}^{N_S} \pi_l \mathcal{D}_S^l(\cdot)$ The $\mathcal{H}$ -divergence between $\mathcal{D}'_U$ and $\mathcal{D}''_U$ can be written as:

证明。考虑两个未见领域, $\mathcal{D}'_U$ 和 $\mathcal{D}''_U$ 在支持 Ω 的源领域的凸包 Λ_S 上。同时考虑 $\mathcal{D}'_U(\cdot) = \sum_{k=1}^{N_S} \pi_k \mathcal{D}_S^k(\cdot)$ 和 $\mathcal{D}''_U(\cdot) = \sum_{l=1}^{N_S} \pi_l \mathcal{D}_S^l(\cdot)$ 。$\mathcal{D}'_U$ 和 $\mathcal{D}''_U$ 之间的 $\mathcal{H}$ -散度可以写为:

$$
\begin{aligned}
d_{\mathcal{H}}\left[\mathcal{D}'_U, \mathcal{D}''_U\right] &= 2 \sup_{h \in \mathcal{H}} \left| \Pr_{x \sim \mathcal{D}'_U}\left[h(x) = 1\right] - \Pr_{x \sim \mathcal{D}''_U}\left[h(x) = 1\right] \right|, \\
&= 2 \sup_{h \in \mathcal{H}} \left| \mathbb{E}_{x \sim \mathcal{D}'_U}\left[\mathbf{I}(h(x))\right] - \mathbb{E}_{x \sim \mathcal{D}''_U}\left[\mathbf{I}(h(x))\right] \right|, \\
&= 2 \sup_{h \in \mathcal{H}} \left| \int_{\Omega} \mathcal{D}'_U(x)\, \mathbf{I}(h(x))\, dx - \int_{\Omega} \mathcal{D}''_U(x)\, \mathbf{I}(h(x))\, dx \right|, \\
&= 2 \sup_{h \in \mathcal{H}} \left| \int_{\Omega} \sum_{k=1}^{N_S} \pi_k \mathcal{D}_S^k(x)\, \mathbf{I}(h(x))\, dx - \int_{\Omega} \sum_{l=1}^{N_S} \pi_l \mathcal{D}_S^l(x)\, \mathbf{I}(h(x))\, dx \right|, \qquad (11) \\
&= 2 \sup_{h \in \mathcal{H}} \left| \int_{\Omega} \sum_{l=1}^{N_S} \sum_{k=1}^{N_S} \pi_l \pi_k \mathcal{D}_S^k(x)\, \mathbf{I}(h(x))\, dx - \int_{\Omega} \sum_{l=1}^{N_S} \sum_{k=1}^{N_S} \pi_l \pi_k \mathcal{D}_S^l(x)\, \mathbf{I}(h(x))\, dx \right|, \\
&= 2 \sup_{h \in \mathcal{H}} \left| \sum_{l=1}^{N_S} \sum_{k=1}^{N_S} \pi_l \pi_k \left(\int_{\Omega} \mathcal{D}_S^k(x)\, \mathbf{I}(h(x))\, dx - \int_{\Omega} \mathcal{D}_S^l(x)\, \mathbf{I}(h(x))\, dx \right) \right|.
\end{aligned}
$$

Using the triangle inequality, we can write:
利用三角不等式，我们可以写成:

$$
d_{\mathcal{H}}\left[\mathcal{D}'_U, \mathcal{D}''_U\right] \leq 2 \sup_{h \in \mathcal{H}} \sum_{l=1}^{N_S} \sum_{k=1}^{N_S} \pi_l \pi_k \left| \int_{\Omega} \mathcal{D}_S^k(x)\, \mathbf{I}(h(x))\, dx - \int_{\Omega} \mathcal{D}_S^l(x)\, \mathbf{I}(h(x))\, dx \right|. \qquad (12)
$$

Finally, using the sub-additivity of the sup:
最后，利用上确界的次可加性:

$$
\begin{aligned}
d_{\mathcal{H}}\left[\mathcal{D}'_U, \mathcal{D}''_U\right] &\leq \sum_{l=1}^{N_S} \sum_{k=1}^{N_S} \pi_l \pi_k 2 \sup_{h \in \mathcal{H}} \left| \int_{\Omega} \mathcal{D}_S^k(x)\, \mathbf{I}(h(x))\, dx - \int_{\Omega} \mathcal{D}_S^l(x)\, \mathbf{I}(h(x))\, dx \right|, \qquad (13) \\
&= \sum_{l=1}^{N_S} \sum_{k=1}^{N_S} \pi_l \pi_k d_{\mathcal{H}}\left[\mathcal{D}_S^k, \mathcal{D}_S^l\right]
\end{aligned}
$$

Given $d_{\mathcal{H}}\left[\mathcal{D}_S^k, \mathcal{D}_S^l\right] \leq \epsilon \forall k, l \in [N_S]$:
给定 $d_{\mathcal{H}}\left[\mathcal{D}_S^k, \mathcal{D}_S^l\right] \leq \epsilon \forall k, l \in [N_S]$:

$$
d_{\mathcal{H}}\left[\mathcal{D}'_U, \mathcal{D}''_U\right] \leq \epsilon
$$

□

B Proof of Theorem 1

B 定理 1 的证明

Theorem 1. Let S be the set of source domains and $\mathcal{Y} = [0, 1]$. The risk $R_U[h], \forall h \in \mathcal{H}$, for any unseen domain $\mathcal{D}_U$ such that $d_{\mathcal{H}}\left[\overline{\mathcal{D}}_U, \mathcal{D}_U\right] = \gamma$, is bounded as:

定理 1. 设 S 为源领域的集合，$\mathcal{Y} = [0, 1]$ 。对于任何未见领域 $\mathcal{D}_U$ ，使得 $d_{\mathcal{H}}\left[\overline{\mathcal{D}}_U, \mathcal{D}_U\right] = \gamma$ ，风险 $R_U[h], \forall h \in \mathcal{H}$ 被界定为:

$$
R_U[h] \leq \sum_{i=1}^{N_S} \pi_i R_S^i[h] + \gamma + \epsilon + \min\left\{ \mathbb{E}_{\overline{\mathcal{D}}_U}\left[|f_{S_\pi} - f_U|\right], \mathbb{E}_{\mathcal{D}_U}\left[|f_U - f_{S_\pi}|\right] \right\}, \qquad (14)
$$

where ϵ is the highest pairwise $\widetilde{\mathcal{H}}$ -divergence measured between pairs of domains within $S, \widetilde{\mathcal{H}} = \{\text{sign}(|\ h(x) - \ h'(x) \mid -t) \mid h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$ and $f_{S_\pi}(x) = \sum_{i=i}^{N_S} \pi_i f_{S_i}(x)$ is the labeling function for any $x \in \text{Supp}\left(\overline{\mathcal{D}}_U\right)$ resulting from combining all f_{S_i} with weights $\pi_i, i \in [N_S]$, determined by $\overline{\mathcal{D}}_U$.

其中 ϵ 是在 $S, \widetilde{\mathcal{H}} = \{\text{sign}(|\ h(x) - \ h'(x) \mid -t) \mid h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$ 内部的领域对之间测量的最高成对 $\widetilde{\mathcal{H}}$ -散度，$f_{S_\pi}(x) = \sum_{i=i}^{N_S} \pi_i f_{S_i}(x)$ 是通过将所有 f_{S_i} 结合并用权重 $\pi_i, i \in [N_S]$ 确定的标记函数，由 $\overline{\mathcal{D}}_U$ 确定。

Proof of Theorem 1. Let the source and target domains be $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$, respectively. For the single-source, single-target domain adaptation case, it was previously shown that the risk of any $h \in \mathcal{H}, h : \mathcal{X} \to [0, 1]$ is bounded by

定理 1 的证明。设源领域和目标领域分别为 $\langle \mathcal{D}_S, f_S \rangle$ 和 $\langle \mathcal{D}_T, f_T \rangle$ 。对于单源单目标领域适应情况，之前已证明任何 $h \in \mathcal{H}, h : \mathcal{X} \to [0, 1]$ 的风险是被界定的:[25]:

$$R_T[h] \leq R_S[h] + d_{\widetilde{\mathcal{H}}}[\mathcal{D}_S, \mathcal{D}_T] + \min\left\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_T - f_S|]\right\}, \tag{15}$$

where $\widetilde{\mathcal{H}} = \{\text{sign}\left(|h(x) - h'(x)| - t\right) \mid h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$.

其中 $\widetilde{\mathcal{H}} = \{\text{sign}\left(|h(x) - h'(x)| - t\right) \mid h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$ 。

In order to devise a generalization bound for the risk on any unseen domain in terms of quantities related to the distributions seen at training time, we start by writing (15) considering $\mathcal{D}_U$ and its "projection" onto the convex-hull of the sources $\overline{\mathcal{D}}_U = \text{argmin}_{\pi_1,\ldots,\pi_{N_S}} d_{\mathcal{H}}\left[\mathcal{D}_U, \sum_{i=1}^{N_S} \pi_i \mathcal{D}_S^i\right]$. For that, we introduce the labeling function $f_{S_\pi}(x) = \sum_{i=i}^{N_S} \pi_i f_{S_i}(x)$, which is an ensemble of the respective labeling functions from each source domain weighted by the mixture coefficients that determine $\mathcal{D}_U$.$R_U[h]$ can thus be bounded as:

为了为任何未见领域的风险制定一个泛化界限，我们开始写 (15)，考虑 $\mathcal{D}_U$ 及其在源的凸包上的 "投影" $\overline{\mathcal{D}}_U = \text{argmin}_{\pi_1,\ldots,\pi_{N_S}} d_{\mathcal{H}}\left[\mathcal{D}_U, \sum_{i=1}^{N_S} \pi_i \mathcal{D}_S^i\right]$ 。为此，我们引入标记函数 $f_{S_\pi}(x) = \sum_{i=i}^{N_S} \pi_i f_{S_i}(x)$ ，它是来自每个源领域的相应标记函数的集合，按混合系数加权，因此可以界定为:

$$R_U[h] \leq R_{\bar{U}}[h] + d_{\widetilde{\mathcal{H}}}\left[\overline{\mathcal{D}}_U, \mathcal{D}_U\right] + \min\left\{\mathbb{E}_{\overline{\mathcal{D}}_U}[|f_{S_\pi} - f_U|], \mathbb{E}_{\mathcal{D}_U}[|f_U - f_{S_\pi}|]\right\}.$$

Similarly to the proof of our Lemma 1 for the case where $\mathcal{D}' = \mathcal{D}_U$ and $\mathcal{D}'' = \overline{\mathcal{D}}_U$ (and to [24]), it follows that:

类似于我们引理 1 的证明，对于 $\mathcal{D}' = \mathcal{D}_U$ 和 $\mathcal{D}'' = \overline{\mathcal{D}}_U$ 的情况 (以及 [24])，可以得出:

$$R_U[h] \leq \sum_{i=1}^{N_S} \pi_i R_S^i[h] + \sum_{i=1}^{N_S} \pi_i d_{\widetilde{\mathcal{H}}}\left[\mathcal{D}_S^i, \mathcal{D}_U\right] + \min\left\{\mathbb{E}_{\overline{\mathcal{D}}_U}[|f_{S_\pi} - f_U|], \mathbb{E}_{\mathcal{D}_U}[|f_U - f_{S_\pi}|]\right\}. \tag{16}$$

Using the triangle inequality for the $\mathcal{H}$ -divergence along with Lemma 1, we can bound the $\widetilde{\mathcal{H}}$ -divergence between $\mathcal{D}_U$ and any source domain $\mathcal{D}_S^i, d_{\widetilde{\mathcal{H}}}\left[\mathcal{D}_U, \mathcal{D}_S^i\right]$, according to:

利用三角不等式对于 $\mathcal{H}$ -散度以及引理 1,我们可以根据以下公式界定 $\mathcal{D}_U$ 和任何源领域 $\mathcal{D}_S^i, d_{\widetilde{\mathcal{H}}}\left[\mathcal{D}_U, \mathcal{D}_S^i\right]$ 之间的 $\widetilde{\mathcal{H}}$ -散度:

$$\begin{aligned} d_{\widetilde{\mathcal{H}}}\left[\mathcal{D}_U, \mathcal{D}_S^i\right] &\leq d_{\widetilde{\mathcal{H}}}\left[\mathcal{D}_U, \overline{\mathcal{D}}_U\right] + d_{\widetilde{\mathcal{H}}}\left[\overline{\mathcal{D}}_U, \mathcal{D}_S^i\right] \\ &\leq \gamma + \epsilon, \end{aligned}$$

where $\gamma = d_{\widehat{\mathcal{H}}}\left[\mathcal{D}_U, \overline{\mathcal{D}}_U\right]$. Using this result, we can now upper-bound $\sum_{i=1}^{N_S} \pi_i d_{\widehat{\mathcal{H}}}\left[\mathcal{D}_U, \mathcal{D}_S^i\right]$ by $\gamma + \epsilon$ and finally re-write (16) as:

其中 $\gamma = d_{\widehat{\mathcal{H}}}\left[\mathcal{D}_U, \overline{\mathcal{D}}_U\right]$ 。利用这个结果，我们现在可以将 $\sum_{i=1}^{N_S} \pi_i d_{\widehat{\mathcal{H}}}\left[\mathcal{D}_U, \mathcal{D}_S^i\right]$ 上界为 $\gamma + \epsilon$ ，并最终将 (16) 重写为:

$$R_U[h] \leq \sum_{i=1}^{N_S} \pi_i R_S^i[h] + \gamma + \epsilon + \min\left\{\mathbb{E}_{\overline{\mathcal{D}}_U}[|f_{S_\pi} - f_U|], \mathbb{E}_{\mathcal{D}_U}[|f_U - f_{S_\pi}|]\right\}.$$

C Proof of Corollary 1

C 引理 1 的证明

Corollary 1. Let all domains within the the support of the meta-distribution $\mathfrak{D}$ have labeling function f . Let S be set of source domains and its convex-hull be denoted as Λ_S . The risk $R_U[h]$ of a hypothesis h on an unseen domain $\mathcal{D}_U \in \Lambda_S$, is upper-bounded by:

引理 1. 设所有在元分布 $\mathfrak{D}$ 支持内的领域具有标记函数 f 。设 S 为源领域的集合，其凸包记作 Λ_S 。假设 h 在未见领域 $\mathcal{D}_U \in \Lambda_S$ 上的风险 $R_U[h]$ ，上界为:

$$R_U[h] \leq \sum_{i=1}^{N_S} \pi_i R_S^i[h] + \epsilon. \tag{17}$$

Proof of Corollary 1. The right-most term of (14) accounts for the mismatch between the labeling functions of $\mathcal{D}_U$ and $\overline{\mathcal{D}}_U$. Since all domains within $\mathfrak{D}$ have the same labeling function, this term is equal to 0 . As $\mathcal{D}_U \in \Lambda_S, \mathcal{D}_U = \overline{\mathcal{D}}_U$, which results in $d_{\widetilde{\mathcal{H}}}\left[\mathcal{D}_U, \overline{\mathcal{D}}_U\right] = \gamma = 0$. □

引理 1 的证明。(14) 的最右项考虑了 $\mathcal{D}_U$ 和 $\overline{\mathcal{D}}_U$ 的标记函数之间的不匹配。由于 $\mathfrak{D}$ 中的所有领域具有相同的标记函数，因此该项等于 0。由于 $\mathcal{D}_U \in \Lambda_S, \mathcal{D}_U = \overline{\mathcal{D}}_U$ ，这导致 $d_{\widetilde{\mathcal{H}}}\left[\mathcal{D}_U, \overline{\mathcal{D}}_U\right] = \gamma = 0$ 。□

D One-vs-all $\mathcal{H}$ -divergence estimation

D 一对多 $\mathcal{H}$ -散度估计

We illustrate the estimation of $\mathcal{H}$ -divergences using one-vs-all discriminators by considering an example in which 3 source domains are available. Consider samples of size M from $N_S = 3$ source domains which are available at

我们通过考虑一个有 3 个源领域可用的示例，说明了使用一对多判别器估计 $\mathcal{H}$ -散度。考虑来自 $N_S = 3$ 源领域的样本大小为 M 。

training time. The loss $\mathcal{L}_1$ for the domain discriminator D_1 accounting for estimating $d_{\mathcal{H}}[\mathcal{D}_1, \mathcal{D}_2]$ and $d_{\mathcal{H}}[\mathcal{D}_1, \mathcal{D}_3]$ can be written as:

在训练时，领域判别器 D_1 的损失 $\mathcal{L}_1$ 用于估计 $d_{\mathcal{H}}[\mathcal{D}_1, \mathcal{D}_2]$ 和 $d_{\mathcal{H}}[\mathcal{D}_1, \mathcal{D}_3]$ 可以写成:

$$\mathcal{L}_1 = \frac{1}{3M} \sum_{i=1}^{3M} \ell\left(D_1(x_i), y_1\right) \tag{18}$$

$$= \frac{1}{M} \sum_{i=1}^{M} \ell\left(D_1(x_i), y_1\right) + \frac{1}{M} \sum_{i=M+1}^{2M} \ell\left(D_1(x_i), y_1\right) + \frac{1}{M} \sum_{i=2M+1}^{3M} \ell\left(D_1(x_i), y_1\right),$$

where ℓ represents a loss function (e.g. 0-1 loss) and each term accounts for the loss provided by examples from one domain. Splitting the first term in two parts and replacing the domain labels y_1 by their corresponding values, we obtain:

其中 ℓ 代表损失函数 (例如 0-1 损失)，每个项对应于来自一个领域的示例所提供的损失。将第一项分为两部分，并用其对应值替换领域标签 y_1 ，我们得到:

$$\mathcal{L}_1 = \frac{1}{M} \sum_{i=1}^{M/2} \ell\left(D_1(x_i), 1\right) + \frac{1}{M} \sum_{i=M+1}^{2M} \ell\left(D_1(x_i), 0\right) \tag{19}$$

$$+ \frac{1}{M} \sum_{i=\frac{M}{2}+1}^{M} \ell\left(D_1(x_i), 1\right) + \frac{1}{M} \sum_{i=2M+1}^{3M} \ell\left(D_1(x_i), 0\right).$$

The first two terms from Eq 19 account for $d_{\mathcal{H}}[\mathcal{D}_1, \mathcal{D}_2]$ and the last two terms account for $d_{\mathcal{H}}[\mathcal{D}_1, \mathcal{D}_3]$.

Eq 19 中的前两项对应于 $d_{\mathcal{H}}[\mathcal{D}_1, \mathcal{D}_2]$ ，最后两项对应于 $d_{\mathcal{H}}[\mathcal{D}_1, \mathcal{D}_3]$ 。

E Illustration

E 说明

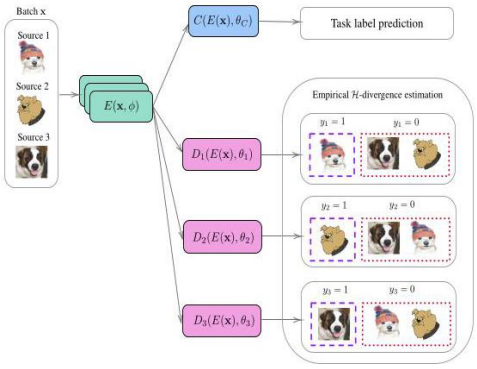


Figure 2: Proposed approach illustration.
图 2: 提出的方法说明。

F Extra experiments

F 额外实验

F.1 Impact of source domains diversity on unseen domain accuracy

F.1 源领域多样性对未见领域准确性的影响

In this experiment, we verify whether removing examples from one source domain impacts the performance on the target domain. We evaluate each target domain on models trained using all possible combinations of the remaining domains as sources. The ERM baseline is also included for reference. Results presented in Table 5 show that for all unseen domains, decreasing the number of source domains from 3 (see Table 1) to 2 hurt the classification performance for almost all combinations of source domains. We notice that in some cases, excluding a particular source from the training severely decreases the target loss. As an example, for the Caltech-101, excluding from training examples from the VOC dataset decreased the accuracy in more than 10% for the proposed approach, as well as for ERM.

在这个实验中，我们验证从一个源领域移除示例是否会影响目标领域的性能。我们在使用所有可能组合的剩余领域作为源的模型上评估每个目标领域。ERM 基线也包括在内以供参考。表 5 中呈现的结果显示，对于所有未见领域，将源领域的数量从 3(见表 1) 减少到 2 几乎对所有源领域组合的分类性能造成了损害。我们注意到在某些情况下，排除特定源的训练会严重降低目标损失。例如，对于 Caltech-101，从训练中排除 VOC 数据集的示例使得提出的方法以及 ERM 的准确率下降了超过 10% 。

F.2 Effect of random projection size

F.2 随机投影大小的影响

We further investigate the effectiveness on providing a more stable training of the random projection layer in the input of each discriminator. For that, we run experiments with 7 different projection sizes, as well as directly using the output of the feature extractor model. Besides the random projection size, we use the same hyperparameters values (the same used in the previous experiment) and initialization for all models. We report in Figure 3 the best target accuracy achieved with all random projection sizes on the PACS benchmark considering the Sketch dataset as unseen domain. Overall, we observed that the random projection layer has indeed an impact on the generalization of the learned representation and that the best result was achieved with a size equal to 1000 . Moreover, we notice that, in this case, having

a smaller (500) random projection layer is less hurtful for the performance than using a larger one. We also found that removing the random projection layer did not allow the training to converge with this experimental setting.

我们进一步研究在每个判别器输入中提供随机投影层的更稳定训练的有效性。为此，我们进行了 7 种不同投影大小的实验，并直接使用特征提取模型的输出。除了随机投影大小外，我们对所有模型使用相同的超参数值 (与之前实验中使用的相同) 和初始化。我们在图 3 中报告了在 PACS 基准上使用所有随机投影大小所达到的最佳目标准确率，考虑到草图数据集作为未见领域。总体而言，我们观察到随机投影层确实对学习表示的泛化产生了影响，并且最佳结果是在大小为 1000 时取得的。此外，我们注意到，在这种情况下，使用较小的 (500) 随机投影层对性能的影响小于使用较大的随机投影层。我们还发现，去除随机投影层并未使训练在此实验设置下收敛。

Table 5: Impact of decreasing the number of source domains on VLCS. Rows represent the two source domains used.

表 5: 减少源领域数量对 VLCS 的影响。行表示使用的两个源领域。

		Source					
Target	Method	VC	VL	VS	LC	LS	CS
V	ERM	-	-	-	66.14	72.16	69.89
	Ours	-	-	-	62.39	69.89	67.23
L	ERM	58.32	-	62.11	-	-	59.85
	Ours	65.37	-	65.87	-	-	64.37
C	ERM	-	98.82	98.58	-	84.67	-
	Ours	-	95.75	96.70	-	81.84	-
S	ERM	69.04	66.29	-	59.80	-	-
	Ours	69.54	68.43	-	57.06	-	-

		源					
目标	方法	VC	VL	VS	LC	LS	CS
V	ERM	-	-	-	66.14	72.16	69.89
	我们的方法	-	-	-	62.39	69.89	67.23
L	ERM	58.32	-	62.11	-	-	59.85
	我们的方法	65.37	-	65.87	-	-	64.37
C	ERM	-	98.82	98.58	-	84.67	-
	我们的方法	-	95.75	96.70	-	81.84	-
S	ERM	69.04	66.29	-	59.80	-	-
	我们的方法	69.54	68.43	-	57.06	-	-

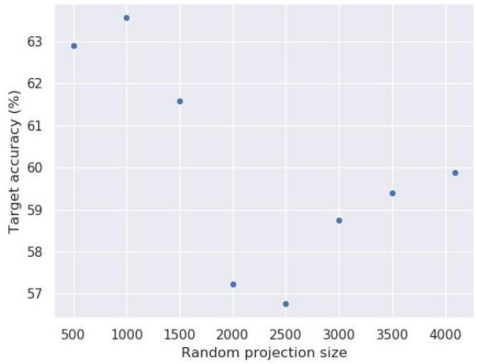


Figure 3: Accuracy obtained on the PACS benchmark using Sketch as target domain.

图 3: 在 PACS 基准上使用草图作为目标领域获得的准确率。

G Domain generalization benchmarks

G 领域泛化基准

The VLCS benchmark is composed by 4 datasets with 5 common classes, namely, bird, car, chair, dog, and person. The number of data points per dataset is detailed as follows. We split each dataset in 80%/20% train/test partitions.

VLCS 基准由 4 个数据集组成，具有 5 个共同类别，即鸟、汽车、椅子、狗和人。每个数据集的数据点数量详细如下。我们将每个数据集分为 80%/20% 训练/测试分区。

- Pascal VOC2007: 3376;
- Pascal VOC2007: 3376；
- LabelMe: 2656;
- LabelMe: 2656；
- Caltech-101: 1415;
- Caltech-101: 1415；
- SUN09: 3282.
- SUN09: 3282。

The PACS benchmark is composed by 4 datasets with 7 common classes, namely, dog, elephant, giraffe, guitar, horse, house, and person. The number of data points per dataset is detailed as follows. We use the original train/validation partitions provided by the benchmark authors.

PACS 基准由 4 个数据集组成，具有 7 个共同类别，即狗、大象、长颈鹿、吉他、马、房子和人。每个数据集的数据点数量详细如下。我们使用基准作者提供的原始训练/验证分区。

- Photos: 1670;
- 照片: 1670；
- Art painting: 2048;
- 艺术绘画:2048；

- Cartoon: 2344;

- Sketch: 3929.

H Implementation details

H 实现细节

H.1 VLCS and PACS benchmarks

H.1 VLCS 和 PACS 基准测试

In order to obtain a consistent comparison with the aforementioned baseline models, we follow previous work and employ the weights of a pre-trained AlexNet [41] and ResNet-18 [42] as the initialization for the feature extractor model on the experiments. The last layer is discarded and the representation of size 4096 for AlexNet and 512 for ResNet-18 is used as input for the task classifier and the domain discriminators. The domain discriminator architecture with AlexNet, consists of a four-layer fully-connected neural network of size 4096 $\rightarrow$ random projection size $\rightarrow$ 1024 $\rightarrow$ 1 and five-layer fully connected network of size 512 $\rightarrow$ random projection size $\rightarrow$ 512 $\rightarrow$ 256 $\rightarrow$ 1 for ResNet-18 . The random projection layer is implemented as a linear layer with weights normalized to have unitary L2-norm. The task classifier is a one-layer fully-connected network of size 4096 $\rightarrow$ number of classes in the case of AlexNet and 512 $\rightarrow$ number of classes in the case of ResNet. Following previous work on domain generalization [20, 19], we use models pre-trained on the ILSVRC dataset [47] as initialization. For fair comparison, all models we implemented were given a budget of 200 epochs. We use label smoothing [48] on the task classifier in order to prevent overfitting. Models were trained using SGD with Polyak's acceleration. One epoch

corresponds to the length of the largest source domain training sample. The learning rate was "warmed-up" for a number of training iterations equal to nw . Hyperparameter tuning was performed through random search over a pre-defined grid so as to find the best values for the learning rate (lr), momentum, weight decay, label smoothing parameter ls, nw , random projection size [6] learning rate reduction factor, and weighting (α) . Each model was run with three different initializations (random seeds 1,10, and 100 selected a priori) and the average best accuracy on the test partition of the target domain is reported. Details of the hyperparameters grid used in the search are provided in the Supplementary material. For our ERM we used the same hyperparameters as in [8], while for IRM we employed the same hyperparameter values reported in the authors implementation of the colored MNIST experiments.

为了与上述基线模型进行一致的比较，我们遵循之前的工作，并采用预训练的 AlexNet [41] 和 ResNet-18 [42] 的权重作为实验中特征提取模型的初始化。最后一层被丢弃，AlexNet 的 4096 维表示和 ResNet-18 的 512 维表示被用作任务分类器和领域鉴别器的输入。与 AlexNet 相关的领域鉴别器架构由一个四层全连接神经网络组成，大小为 4096 → 随机投影大小 → 1024 → 1 ，而 ResNet-18 的领域鉴别器则由一个五层全连接网络组成，大小为 512 → 随机投影大小 → 512 → 256 → 1 。随机投影层被实现为一个线性层，其权重经过归一化以具有单位 L2 范数。任务分类器是一个一层全连接网络，大小为 AlexNet 的 4096 → 类别数和 ResNet 的 512 → 类别数。遵循之前关于领域泛化的工作 [20, 19]，我们使用在 ILSVRC 数据集 [47] 上预训练的模型作为初始化。为了公平比较，我们实现的所有模型都给定了 200 个周期的预算。我们在任务分类器上使用标签平滑 [48] 以防止过拟合。模型使用带有 Polyak 加速的 SGD 进行训练。一个周期对应于最大源领域训练样本的长度。学习率在等于 nw 的训练迭代次数内进行了 "预热"。超参数调优通过在预定义网格上进行随机搜索，以找到学习率 (lr)、动量、权重衰减、标签平滑参数 ls, nw 、随机投影大小 [6] 、学习率减少因子和加权 (α) 的最佳值。每个模型都使用三个不同的初始化 (随机种子 1、10 和 100，选择 a 先验) 进行运行，并报告目标领域测试分区上的平均最佳准确率。搜索中使用的超参数网格的详细信息在补充材料中提供。对于我们的 ERM，我们使用了与 [8] 中相同的超参数，而对于 IRM，我们采用了作者在彩色 MNIST 实验中实现的相同超参数值。

The grids used on the hyperparameter search for each hyperparameter are presented in the following. A budget of 200 runs was considered and for each combination of hyperparameters each model was trained for 200 and 30 epochs in the case of AlexNet and ResNet-18, respectively. The best hyperparamters values for AlexNet on PACS and VLCS benchmarks are respectively denoted by $^{*}, ^{\dagger}$. For the ResNet-18 experiments on PACS we indicate the hyperparameters by $^{+}$. Moreover, in the case of ResNet-18, we aggregated the discriminators losses by computing the corresponding hypervolume as in [30], with a nadir slack equal to 2.5. All experiments were run considering a minibatch size of 64 (training each iteration took into account 64 examples from each source domain) on single GPU hardware (either an NVIDIA V100 or NVIDIA GeForce GTX 1080Ti).

用于每个超参数搜索的网格如下所示。考虑到 200 次运行的预算，对于每组超参数，AlexNet 和 ResNet-18 模型分别训练了 200 和 30 个周期。AlexNet 在 PACS 和 VLCS 基准上的最佳超参数值分别用 $^{*}, ^{\dagger}$ 表示。对于在 PACS 上进行的 ResNet-18 实验，我们用 $^{+}$ 表示超参数。此外，在 ResNet-18 的情况下，我们通过计算相应的超体积来聚合判别器损失，如 [30] 所示，最低松弛等于 2.5。所有实验均在单个 GPU 硬件 (NVIDIA V100 或 NVIDIA GeForce GTX 1080Ti) 上进行，考虑到小批量大小为 64(每次迭代训练时考虑来自每个源域的 64 个示例)。

- Learning rate for the task classifier and feature extractor: $\left\{0.01^{*,+}, 0.001^{\dagger}, 0.0005\right\}$;
- 任务分类器和特征提取器的学习率: $\left\{0.01^{*,+}, 0.001^{\dagger}, 0.0005\right\}$;
- Learning for the domain classifiers: $\left\{0.0005^{*}, 0.001, 0.005^{\dagger,+}\right\}$;
- 域分类器的学习率: $\left\{0.0005^{*}, 0.001, 0.005^{\dagger,+}\right\}$;
- Weight decay: $\left\{0.0005^{*}, 0.001, 0.005^{\dagger+}\right\}$;
- 权重衰减: $\left\{0.0005^{*}, 0.001, 0.005^{\dagger+}\right\}$;
- Momentum: $\left\{0.5, 0.9^{*,\dagger,+}\right\}$
- 动量: $\left\{0.5, 0.9^{*,\dagger,+}\right\}$
- Label smoothing: $\left\{0.0^{+}, 0.1, 0.2^{*,\dagger}\right\}$;
- 标签平滑: $\left\{0.0^{+}, 0.1, 0.2^{*,\dagger}\right\}$;

- Losses weighting (α) : $\left\{0.35, 0.8^{*,\dagger,+}\right\}$;
- 损失加权 (α) : $\left\{0.35, 0.8^{*,\dagger,+}\right\}$;
- Random projection size: $\left\{1000^{*}, 3000, 3500^{\dagger},\ \text{None}^{+}\right\}$;
- 随机投影大小: $\left\{1000^{*}, 3000, 3500^{\dagger},\ \text{None}^{+}\right\}$;
- Task classifier and feature extractor learning rate warm-up iterations: $\left\{1, 300^{*,\dagger}, 500^{+}\right\}$;
- 任务分类器和特征提取器学习率预热迭代: $\left\{1, 300^{*,\dagger}, 500^{+}\right\}$;
- Warming-up threshold: $\left\{0.00001^{*}, 0.0001^{\dagger,+}, 0.001\right\}$;
- 预热阈值: $\left\{0.00001^{*}, 0.0001^{\dagger,+}, 0.001\right\}$;
- Learning rate schedule patience: $\left\{25^{+}, 60^{\dagger}, 80^{*}\right\}$;
- 学习率调度耐心: $\left\{25^{+}, 60^{\dagger}, 80^{*}\right\}$;
- Learning rate schedule decay factor: $\left\{0.1^{+}, 0.3^{\dagger}, 0.5^{*}\right\}$.
- 学习率调度衰减因子: $\left\{0.1^{+}, 0.3^{\dagger}, 0.5^{*}\right\}$.

H.2 Affective state prediction

H.2 情感状态预测

We use SyncNet [44] as the encoder for the experiments with the SEED dataset. We follow previous work and apply a simple pre-processing that consists of clipping artifacts with amplitude 5 times higher than the mean of the channel signal and windowing data with chunks of 60 seconds. Each window was normalized to have zero mean and unit variance. For the encoder network, we adopt an one layer parameterized convolutional filter with 2 filters (designed to extract synchrony coherence which interpretable features based on the previous neuroscience literature [44]). We train all models for 100 epochs using SGD with Polyak's acceleration. The learning rate was "warmed-up" for a number of training iterations equal to 500 .

我们使用 SyncNet [44] 作为 SEED 数据集实验的编码器。我们遵循之前的工作，应用简单的预处理，包括裁剪幅度比通道信号均值高 5 倍的伪影，并将数据分成 60 秒的窗口。每个窗口被归一化为零均值和单位方差。对于编码器网络，我们采用一个层的参数化卷积滤波器，具有 2 个滤波器 (旨在提取基于之前神经科学文献 [44] 的同步一致性可解释特征)。我们使用带有 Polyak 加速的 SGD 训练所有模型 100 个周期。学习率在 500 次训练迭代内进行了 "预热"。

The output of the encoder with size 602 is used as input for the task classifier and the domain discriminators. The domain discriminator architecture consists of a four-layer fully-connected neural network of size 602 $\rightarrow$ random projection size $\rightarrow$ 256 $\rightarrow$ 128 $\rightarrow$ 2 . The random projection layer is implemented as a linear layer with weights normalized to have unitary L2-norm. The task classifier is a two-layer fully-connected network of size 602 $\rightarrow$ 100 $\rightarrow$ number of classes.

编码器输出的大小为 602，作为任务分类器和领域鉴别器的输入。领域鉴别器架构由一个四层全连接神经网络组成，大小为 602 $\rightarrow$ 随机投影大小 $\rightarrow$ 256 $\rightarrow$ 128 $\rightarrow$ 2 。随机投影层实现为一个线性层，其权重被归一化为单位 L2 范数。任务分类器是一个两层全连接网络，大小为 602 $\rightarrow$ 100 $\rightarrow$ 类别数。

The summary of parameters is presented in the following.

参数汇总如下。

- Window size: 60 seconds
- 窗口大小: 60 秒

[6] The option of not having the random projection layer is included in the grid search.

[6] 不包含随机投影层的选项包含在网格搜索中。

- Number of filters: 2
- 滤波器数量: 2
- Filters length: 40
- 滤波器长度: 40
- Pooling size: 40
- 池化大小: 40
- Input drop out rate: 0.2
- 输入丢弃率: 0.2
- Initial learning rate task classifier: 9.963e-04
- 初始学习率任务分类器:9.963e-04
- Initial learning rate discriminator: 9.963e-05
- 初始学习率鉴别器:9.963e-05
- Random projection size: 602
- 随机投影大小:602

H.3 Proxy $\mathcal{A}$ -distance estimation

H.3 代理 $\mathcal{A}$ - 距离估计

We implement the domain discriminators using tree ensemble classifiers with 100 estimators. We thus report the average classification accuracy using 5-fold cross-validation independently run for each domain pair. Each domain is represented by a random sample of size 500 .

我们使用具有 100 个估计器的树集成分类器实现领域鉴别器。因此，我们报告使用 5 折交叉验证独立运行每个领域对的平均分类准确率。每个领域由大小为 500 的随机样本表示。