Unified Deep Supervised Domain Adaptation and Generalization

统一深度监督领域适应与泛化

Saeid Motiian Marco Piccirilli Donald A. Adjeroh Gianfranco Doretto

Saeid Motiian Marco Piccirilli Donald A. Adjeroh Gianfranco Doretto

West Virginia University
西弗吉尼亚大学
Morgantown, WV 26508
莫根敦，WV 26508
{samotiian, mpiccir1, daadjeroh, gidoretto}@mix.wvu.edu

{samotiian, mpiccir1, daadjeroh, gidoretto}@mix.wvu.edu

Abstract

摘要

This work provides a unified framework for addressing the problem of visual supervised domain adaptation and generalization with deep models. The main idea is to exploit the Siamese architecture to learn an embedding subspace that is discriminative, and where mapped visual domains are semantically aligned and yet maximally separated. The supervised setting becomes attractive especially when only few target data samples need to be labeled. In this scenario, alignment and separation of semantic probability distributions is difficult because of the lack of data. We found that by reverting to point-wise surrogates of distribution distances and similarities provides an effective solution. In addition, the approach has a high ”speed” of adaptation, which requires an extremely low number of labeled target training samples, even one per category can be effective. The approach is extended to domain generalization. For both applications the experiments show very promising results.

本文提供了一个统一框架，以解决深度模型下视觉监督领域适应和泛化的问题。主要思想是利用孪生网络架构学习一个具有判别性的嵌入子空间，在该空间中，映射的视觉领域在语义上对齐且最大程度分离。监督设置尤其吸引人，特别是在仅需标记少量目标数据样本的情况下。在这种情况下，由于数据不足，语义概率分布的对齐和分离变得困难。我们发现，通过回归到分布距离和相似性的逐点替代，可以提供有效的解决方案。此外，该方法具有较高的适应“速度”，只需极少量的标记目标训练样本，甚至每个类别一个样本也能有效。该方法扩展到领域泛化。对于这两种应用，实验结果显示出非常有希望的结果。

1. Introduction

1. 引言

Many computer vision applications require enough labeled data (target data) for training visual classifiers to address a specific task at hand. Whenever target data is either not available, or it is expensive to collect and/or label it, the typical approach is to use available datasets (source data), representative of a closely related task. Since this practice is known for leading to suboptimal performance, techniques such as domain adaptation [6] and/or domain generalization [5] have been developed to address the issue. Domain adaptation methods require target data, whereas domain generalization methods do not. Domain adaptation can be either supervised [60, 33] , unsupervised [38, 38, 61] , or semi-supervised [26, 29, 67] . Unsupervised domain adaptation (UDA) is attractive because it does not require target data to be labeled. Conversely, supervised domain adaptation (SDA) requires labeled target data.

许多计算机视觉应用需要足够的标记数据 (目标数据) 来训练视觉分类器，以解决手头的特定任务。每当目标数据不可用，或收集和/或标记目标数据的成本较高时，典型的方法是使用可用的数据集 (源数据)，这些数据集代表了一个密切相关的任务。由于这种做法通常会导致次优性能，因此开发了诸如领域适应 [6] 和/或领域泛化 [5] 的技术来解决这个问题。领域适应方法需要目标数据，而领域泛化方法则不需要。领域适应可以是监督的 [60, 33] 、无监督的 [38, 38, 61] 或半监督的 [26, 29, 67] 。无监督领域适应 (UDA) 具有吸引力，因为它不需要标记目标数据。相反，监督领域适应 (SDA) 需要标记的目标数据。

UDA expects large amounts of target data in order to be effective, and this is emphasized even more when using deep models. Moreover, given the same amount of target data, SDA typically outperforms

UDA, as we will later explain. Therefore, especially when target data is scarce, it is more attractive to use SDA, also because limited amounts of target data are likely to not be very expensive to label.

UDA 期望有大量的目标数据以发挥其有效性，尤其在使用深度模型时这一点更加明显。此外，在相同数量的目标数据下，SDA 通常优于 UDA，正如我们稍后将解释的那样。因此，特别是在目标数据稀缺的情况下，使用 SDA 更具吸引力，因为有限数量的目标数据标记成本可能并不高。

In the absence of target data, domain generalization (DG) exploits several cheaply available datasets (sources), representing different specific but closely related tasks. It then attempts to learn by combining data sources in a way that produces visual classifiers that are less sensitive to the specific target data that will need to be processed.

在缺乏目标数据的情况下，领域泛化 (DG) 利用几种廉价可得的数据集 (源)，这些数据集代表不同的特定但密切相关的任务。然后，它试图通过以一种方式组合数据源来学习，从而产生对需要处理的特定目标数据不那么敏感的视觉分类器。

In this work, we introduce a supervised approach for visual recognition that can be used for both SDA and DG. The SDA approach requires very few labeled target samples per category in training. Indeed, even one sample can significantly increase performance, and a few others bring it closer to a peak, showing a remarkable ”speed” of adaptation. Moreover, the approach is also robust to adapting to categories that have no target labeled samples. Although domain adaptation and generalization are closely related, adaptation techniques are not directly applied to DG, and viceversa. However, we show that by making simple changes to our proposed training loss function, and by maintaining the same architecture, our SDA approach very effectively extends to DG.

在本研究中，我们介绍了一种用于视觉识别的监督方法，该方法可用于 SDA 和 DG。SDA 方法在训练中每个类别只需很少的标记目标样本。实际上，甚至一个样本就能显著提高性能，而再加上几个样本则能使其接近峰值，显示出显著的“适应速度”。此外，该方法在适应没有目标标记样本的类别时也具有鲁棒性。尽管领域适应和泛化密切相关，但适应技术并不直接应用于 DG，反之亦然。然而，我们表明，通过对我们提出的训练损失函数进行简单的修改，并保持相同的架构，我们的 SDA 方法可以非常有效地扩展到 DG。

Using basic principles, we analyze how visual classification is extended to handle UDA by aligning a source domain distribution to a target domain distribution to make the classifier domain invariant. This leads to observing that SDA approaches improve upon UDA by making the alignment semantic, because they can ensure the alignment of semantically equivalent distributions from different domains. However, we go one step ahead by suggesting that semantic distribution separation should further increase performance, and this leads to the introduction of a classification and contrastive semantic alignment (CCSA) loss.

我们利用基本原理分析了如何通过将源领域分布与目标领域分布对齐来扩展视觉分类以处理 UDA，从而使分类器具有领域不变性。这导致我们观察到 SDA 方法通过使对齐语义化而改善 UDA，因为它们可以确保来自不同领域的语义等价分布的对齐。然而，我们进一步提出，语义分布分离应进一步提高性能，这引入了一种分类和对比语义对齐 (CCSA) 损失。

We deal with the limited size of target domain samples by observing that the CCSA loss relies on computing distances and similarities between distributions (as typically done in adaptation and generalization approaches). Those are difficult to represent with limited data. Thus, we revert to point-wise surrogates. The resulting approach turns out to be very effective as shown in the experimental section.

我们通过观察 CCSA 损失依赖于计算分布之间的距离和相似性 (如适应和泛化方法中通常所做的) 来处理目标领域样本的有限数量。这些在有限数据下难以表示。因此，我们回归到逐点替代物。结果证明，该方法在实验部分中表现出非常有效的效果。

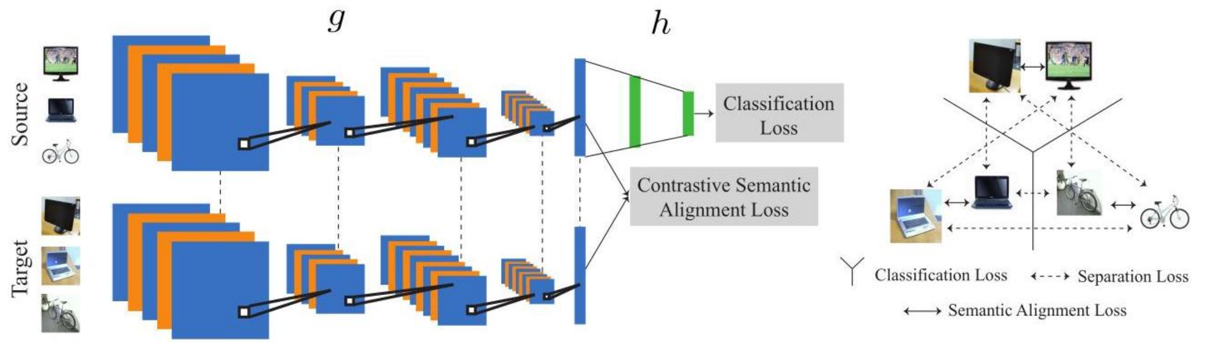

Figure 1. Deep supervised domain adaptation. In training, the semantic alignment loss minimizes the distance between samples from different domains but the same class label and the separation loss maximizes the distance between samples from different domains and class labels. At the same time, the classification loss guarantees high classification accuracy.

图 1. 深度监督领域适应。在训练中，语义对齐损失最小化来自不同领域但具有相同类别标签的样本之间的距离，而分离损失最大化来自不同领域和类别标签的样本之间的距离。同时，分类损失保证了高分类准确性。

2. Related work

2. 相关工作

Domain adaptation. Visual recognition algorithms are trained with data from a source domain, and when they are tested on a target domain with marginal distribution that differs from the one of the sources, we experience the visual domain adaptation (DA) problem (also known as dataset bias [48, 59, 58] , or covariate shift [54]), and observe a performance decrease.

领域适应。视觉识别算法是在源领域的数据上训练的，当它们在目标领域上进行测试，而目标领域的边际分布与源领域不同时，我们会遇到视觉领域适应 (DA) 问题 (也称为数据集偏差 [48, 59, 58] 或协变量偏移 [54])，并观察到性能下降。

Traditional DA methods attempt to directly minimize the shift between source and target distributions. We divide them in three categories. The first one includes those that try to find a mapping between source and target distributions [53, 34, 27, 26, 19, 57] . The second one seeks to find a shared latent space for source and target distributions [40, 2, 44, 21, 22, 47, 43] . The third one regularizes a classifier trained on a source distribution to work well on a target distribution [4, 1, 66, 15, 3, 12] . UDA approaches fall in the first and second categories, while SDA methods could fall either in the second or third category or sometimes both. Recently, [7, 42] have addressed UDA when an auxiliary data view [36, 43] , is available during training, which is beyond the scope of this work.

传统的领域适应方法试图直接最小化源分布和目标分布之间的偏移。我们将它们分为三类。第一类包括那些试图找到源分布和目标分布之间映射的方法 [53, 34, 27, 26, 19, 57] 。第二类寻求为源分布和目标分布找到一个共享的潜在空间 [40, 2, 44, 21, 22, 47, 43] 。第三类则对在源分布上训练的分类器进行正则化，以便在目标分布上表现良好 [4, 1, 66, 15, 3, 12] 。无监督领域适应 (UDA) 方法属于第一类和第二类，而监督领域适应 (SDA) 方法可能属于第二类或第三类，有时两者都有。最近，[7, 42] 在训练期间处理了当有辅助数据视图 [36, 43] 可用时的无监督领域适应，这超出了本工作的范围。

Here, we are interested in finding a shared subspace for source and target distributions. Among algorithms for subspace learning, Siamese networks [11] work well for different tasks [14, 55, 35, 63, 9] . Recently, Siamese networks have been used for domain adaptation. In [60], which is an SDA approach, unlabeled and sparsely labeled target domain data are used to optimize for domain invariance to facilitate domain transfer while using a soft label distribution matching loss. In [56], which is a UDA approach, unlabeled target data are used to learn a nonlinear transformation that aligns correlations of layer activations in deep neural networks. Some approaches went beyond the Siamase weight-sharing and used couple networks for DA. [33] uses two CNN streams, for source and target, fused at the classifier level. [50] uses a two-streams architecture, for source and target, with related but not shared weights. Here we use a Siamese network to learn an embedding such that samples from the same class are mapped as close as possible to each other. This semantic alignment objective is similar to other deep approaches, but unlike them, we explicitly model and introduce cross-domain class separation forces. Moreover, we do so with very few training samples, which makes the problem of characterizing distributions challenging, and this is why we propose to use point-wise surrogates.

在这里，我们关注于寻找源分布和目标分布的共享子空间。在子空间学习算法中，Siamese 网络 [11] 在不同任务中表现良好 [14, 55, 35, 63, 9] 。最近，Siamese 网络已被用于领域适应。在 [60] 中，作为一种 SDA 方法，使用未标记和稀疏标记的目标领域数据来优化领域不变性，以促进领域转移，同时使用软标签分布匹配损失。在 [56] 中，作为一种 UDA 方法，使用未标记的目标数据来学习非线性变换，以对齐深度神经网络中层激活的相关性。一些方法超越了 Siamese 权重共享，使用了耦合网络进行领域适应。[33] 使用两个 CNN 流，分别针对源和目标，在分类器级别融合。[50] 使用双流架构，针对源和目标，具有相关但不共享的权重。在这里，我们使用 Siamese 网络学习嵌入，使得来自同一类别的样本尽可能靠近彼此。这个语义对齐目标与其他深度方法类似，但与它们不同的是，我们明确建模并引入跨领域类别分离力量。此外，我们在训练样本非常少的情况下这样做，这使得表征分布的问题变得具有挑战性，这就是为什么我们建议使用逐点替代。

Domain generalization. Domain generalization (DG) is a less investigated problem and is addressed in two ways. In the first one, all information from the training domains or datasets is aggregated to learn a shared invariant representation. Specifically, [5] pulls all of the training data together in one dataset, and learns a single SVM classifier. [44] learns an invariant transformation by minimizing the dissimilarity across domains. [23], which can be used for SDA too, finds a representation that minimizes the mismatch between domains and maximizes the separability of data. [24] learns features that are robust to variations across domains.

域泛化。域泛化 (DG) 是一个较少研究的问题，主要通过两种方式来解决。第一种方式是将所有来自训练域或数据集的信息汇聚在一起，以学习一个共享的不变表示。具体而言，[5] 将所有训练数据汇聚到一个数据集中，并学习一个单一的支持向量机 (SVM) 分类器。[44] 通过最小化域间的不相似性来学习一个不变变换。[23] 也可以用于 SDA，找到一个最小化域间不匹配并最大化数据可分性的表示。[24] 学习对域间变化具有鲁棒性的特征。

The second approach to DG is to exploit all information from the training domains to train a classifier or regulate its weights [32, 17, 65, 45, 46]. Specifically, [32] adjusts the weights of the classifier to work well on an unseen dataset, and [65] fuses the score of exemplar classifiers given any test sample. While most works use the shallow models, here we approach DG as in the first way, and extend the proposed SDA approach by training a deep Siamese network to find a shared invariant representation where semantic alignment as well as separation are explicitly accounted for. To the best of our knowledge, [24] is the only DG approach using deep models, and our method is the first deep method that solves both adaptation and generalization.

DG 的第二种方法是利用来自训练域的所有信息来训练分类器或调整其权重 [32, 17, 65, 45, 46]。具体而言，[32] 调整分类器的权重，以便在未见数据集上表现良好，而 [65] 则根据任何测试样本融合示例分类器的得分。虽然大多数研究使用的是浅层模型，但在这里我们将 DG 视为第一种方式，并通过训练一个深度孪生网络来扩展提出的 SDA 方法，以找到一个共享的不变表示，其中语义对齐和分离被明确考虑。据我们所知，[24] 是唯一使用深度模型的 DG 方法，而我们的方法是第一个同时解决适应和泛化的深度方法。

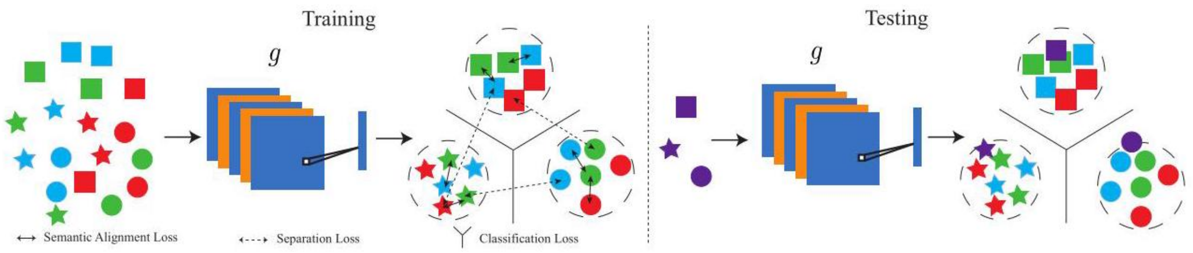

Figure 2. Deep domain generalization. In training, the semantic alignment loss minimizes the distance between samples from different domains but the same class label and the separation loss maximizes the distance between samples from different domains and class labels. At the same time, the classification loss guarantees high classification accuracy. In testing, the embedding function embeds samples from unseen distributions to the domain invariant space and the prediction function classifies them (right). In this figure, different colors represent different domain distributions and different shapes represent different classes.

图 2. 深度域泛化。在训练中，语义对齐损失最小化来自不同域但具有相同类别标签的样本之间的距离，而分离损失最大化来自不同域和类别标签的样本之间的距离。同时，分类损失保证了高分类准确性。在测试中，嵌入函数将来自未见分布的样本嵌入到域不变空间中，预测函数对其进行分类 (右侧)。在该图中，不同的颜色代表不同的域分布，不同的形状代表不同的类别。

3. Supervised DA with Scarce Target Data

3. 有限目标数据的监督领域适应

In this section we describe the model we propose to address supervised domain adaptation (SDA), and in the following Section 4 we extend it to address the domain generalization problem. We are given a training dataset made of pairs $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^N$. The feature $x_i^s \in \mathcal{X}$ is a realization from a random variable X^s , and the label $y_i^s \in \mathcal{Y}$ is a realization from a random variable Y . In addition, we are also given the training data $\mathcal{D}_t = \{(x_i^t, y_i^t)\}_{i=1}^M$, where $x_i^t \in \mathcal{X}$ is a realization from a random variable X^t , and the labels $y_i^t \in \mathcal{Y}$. We assume that there is a covariate shift [54] between X^s and X^t , i.e., there is a difference between the probability distributions $p(X^s)$ and $p(X^t)$. We say that X^s represents the source domain and that X^t represents the target domain. Under this settings the goal is to learn a prediction function $f : \mathcal{X} \to \mathcal{Y}$ that during testing is going to perform well on data from the target domain.

在本节中，我们描述了我们提出的模型，以解决监督领域适应 (SDA) 问题，在接下来的第 4 节中，我们将其扩展以解决领域泛化问题。我们给定一个由对 $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^N$ 组成的训练数据集。特征 $x_i^s \in \mathcal{X}$ 是来自随机变量 X^s 的一个实现，标签 $y_i^s \in \mathcal{Y}$ 是来自随机变量 Y 的一个实现。此外，我们还给定了训练数据 $\mathcal{D}_t = \{(x_i^t, y_i^t)\}_{i=1}^M$ ，其中 $x_i^t \in \mathcal{X}$ 是来自随机变量 X^t 的一个实现，以及标签 $y_i^t \in \mathcal{Y}$ 。我们假设 X^s 和 X^t 之间存在协变量偏移 [54]，即概率分布 $p(X^s)$ 和 $p(X^t)$ 之间存在差异。我们称 X^s 为源领域，称 X^t 为目标领域。在这种设置下，目标是学习一个预测函数 $f : \mathcal{X} \to \mathcal{Y}$ ，该函数在测试时能够在目标领域的数据上表现良好。

The problem formulated thus far is typically referred to as supervised domain adaptation. In this work we are especially concerned with the version of this problem where only very few target labeled samples per class are available. We aim at handling cases where there is only one target labeled sample, and there can even be some classes with no target samples at all.

到目前为止所提出的问题通常被称为监督领域适应。在本研究中，我们特别关注这种问题的一个版本，即每个类别仅有非常少量的目标标记样本可用。我们的目标是处理仅有一个目标标记样本的情况，甚至可能存在一些类别根本没有目标样本。

3.1. Deep SDA

3.1. 深度监督领域适应

In the absence of covariate shift a visual classifier f is trained by minimizing a classification loss

在没有协变量偏移的情况下，通过最小化分类损失来训练视觉分类器 f 。

$$\mathcal{L}_C(f) = E\left[\ell\left(f\left(X^s\right), Y\right)\right] \tag{1}$$

where $E[\cdot]$ denotes statistical expectation and ℓ could be any appropriate loss function (for example categorical cross-entropy for multi-class classification). When the distributions of X^s and X^t are different, a deep model f_s trained with $\mathcal{D}_s$ will have reduced performance on the target domain. Increasing it would be trivial by simply training a new model f_t with data $\mathcal{D}_t$. However, $\mathcal{D}_t$ is small and deep models require large amounts of labeled data.

其中 $E[\cdot]$ 表示统计期望，ℓ 可以是任何适当的损失函数 (例如，多类分类的类别交叉熵)。当 X^s 和 X^t 的分布不同时，使用 $\mathcal{D}_s$ 训练的深度模型 f_s 在目标领域的性能将降低。通过简单地使用数据 $\mathcal{D}_t$ 训练一个新模型 f_t 来增加性能是微不足道的。然而，$\mathcal{D}_t$ 的数据量较小，而深度模型需要大量标记数据。

In general, f could be modeled by the composition of two functions, i.e., $f = h \circ g$. Here $g : \mathcal{X} \to \mathcal{Z}$ would be an embedding from the input space $\mathcal{X}$ to a feature or embedding space $\mathcal{Z}$, and $h : \mathcal{Z} \to \mathcal{Y}$ would be a function for predicting from the feature space. With this notation we would have $f_s = h_s \circ g_s$ and $f_t = h_t \circ g_t$, and the SDA problem would be about finding the best approximation for g_t and h_t , given the constraints on the available data.

一般来说，f 可以通过两个函数的组合来建模，即 $f = h \circ g$ 。在这里，$g : \mathcal{X} \to \mathcal{Z}$ 将是从输入空间 $\mathcal{X}$ 到特征或嵌入空间 $\mathcal{Z}$ 的嵌入，而 $h : \mathcal{Z} \to \mathcal{Y}$ 将是从特征空间进行预测的函数。使用这种符号，我们将有 $f_s = h_s \circ g_s$ 和 $f_t = h_t \circ g_t$ ，而 SDA 问题将是关于在可用数据的约束下找到 g_t 和 h_t 的最佳近似。

The unsupervised DA paradigm (UDA) assumes that $\mathcal{D}_t$ does not have labels. In that case the typical approach assumes that $g_t = g_s = g$, and f minimizes (1), while g also minimizes

无监督领域适应 (UDA) 范式假设 $\mathcal{D}_t$ 没有标签。在这种情况下，典型的方法假设 $g_t = g_s = g$ ，并且 f 最小化 (1)，而 g 也最小化。

$$\mathcal{L}_{CA}(g) = d\left(p\left(g\left(X^s\right)\right), p\left(g\left(X^t\right)\right)\right). \tag{2}$$

The purpose of (2) is to align the distributions of the features in the embedding space, mapped from the source and the target domains. d is meant to be a metric between distributions that once aligned, they will no longer allow to tell whether a feature is coming from the source or the target domain. For that reason, we refer to (2) as the confusion alignment loss. A popular choice for d is the Maximum Mean Discrepancy [28]. In the embedding space $\mathcal{Z}$, features are assumed to be domain invariant. Therefore, UDA methods say that from the feature to the label space it is safe to assume that $h_t = h_s = h$.

(2) 的目的是对来自源域和目标域的特征在嵌入空间中的分布进行对齐。d 被视为分布之间的度量，一旦对齐，就无法判断特征是来自源域还是目标域。因此，我们将 (2) 称为混淆对齐损失。对于 d 的一个流行选择是最大均值差异 [28]。在嵌入空间 $\mathcal{Z}$ 中，特征被假设为领域不变。因此，UDA 方法认为，从特征到标签空间可以安全地假设 $h_t = h_s = h$ 。

Since we are interested in visual recognition, the embedding function g would be modeled by a convolutional neural network (CNN) with some initial convolutional layers, followed by some fully connected layers. In addition, the training architecture would have two streams, one for source and the other for target samples. Since $g_s = g_t = g$, the CNN parameters would be shared as in a Siamese architecture. In addition, the source stream would continue with additional fully connected layers for modeling h . See Figure 1.

由于我们对视觉识别感兴趣，嵌入函数 g 将通过一个卷积神经网络 (CNN) 建模，该网络具有一些初始卷积层，后面跟着一些全连接层。此外，训练架构将有两个流，一个用于源样本，另一个用于目标样本。由于 $g_s = g_t = g$ ，CNN 参数将像在孪生网络架构中一样共享。此外，源流将继续增加全连接层以建模 h 。见图 1。

From the above discussion it is clear that in order to perform well, UDA needs to align effectively. This can happen only if distributions are represented by a sufficiently large dataset. Therefore, UDA approaches are in a position of weakness because we assume $\mathcal{D}_t$ to be small. Moreover, UDA approaches have also another intrinsic limitation, which is that even with perfect confusion alignment, there is no guarantee that samples from different domains but the same class label, would map nearby in the embedding space. This lack of semantic alignment is a major source of performance reduction.

从上述讨论可以清楚地看出，为了表现良好，UDA 需要有效对齐。这只有在分布由足够大的数据集表示时才能发生。因此，UDA 方法处于弱势，因为我们假设 $\mathcal{D}_t$ 较小。此外，UDA 方法还有另一个内在限制，即使在完美的混淆对齐下，也无法保证来自不同域但具有相同类别标签的样本在嵌入空间中会相邻映射。这种缺乏语义对齐是性能降低的主要来源。

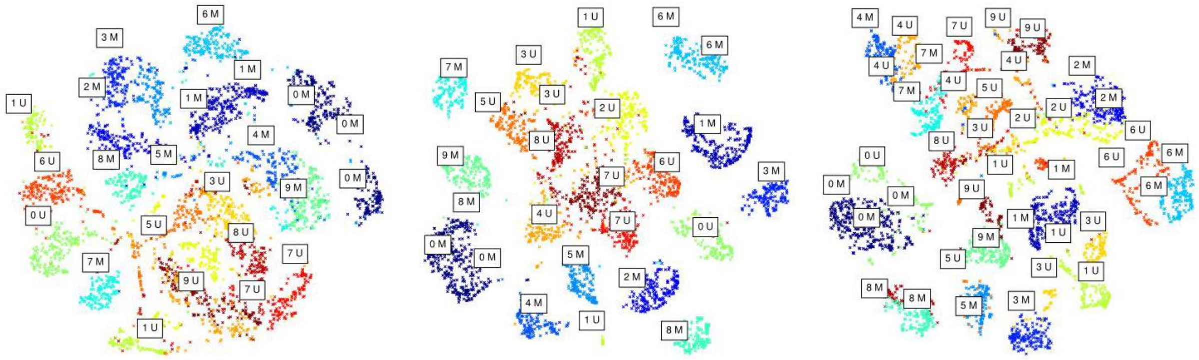

Figure 3. Visualization of the MNIST-USPS datasets. Left: 2D visualization of the row images of the MNIST-USPS datasets. The samples from the same class and different domains lie far from each other on the 2D subspace. Middle: 2D visualization of the embedded images using our base model (without domain adaptation). The samples from the same class and different domains still lie far from each other on the 2D subspace. Right: 2D visualization of the embedded images using our SDA model. The samples from the same class and different domains lie very close to each other on the 2D subspace.

图 3. MNIST-USPS 数据集的可视化。左侧:MNIST-USPS 数据集行图像的 2D 可视化。同一类别和不同领域的样本在 2D 子空间中相距较远。中间: 使用我们的基础模型 (不进行领域适应) 的嵌入图像的 2D 可视化。同一类别和不同领域的样本在 2D 子空间中仍然相距较远。右侧: 使用我们的 SDA 模型的嵌入图像的 2D 可视化。同一类别和不同领域的样本在 2D 子空间中非常接近。

SDA approaches easily address the semantic alignment problem by replacing (2) with

SDA 方法通过将 (2) 替换为来轻松解决语义对齐问题

$$\mathcal{L}_{SA}(g) = \sum_{a=1}^{C} d\left(p\left(g\left(X_a^s\right)\right), p\left(g\left(X_a^t\right)\right)\right), \tag{3}$$

where C is the number of class labels, and $X_a^s = X^s \mid \{Y = a\}$ and $X_a^t = X^t \mid \{Y = a\}$ are conditional random variables. d instead is a suitable distance mesure between the distributions of X_a^s and X_a^t in the embedding space. We refer to (3) as the semantic alignment loss, which clearly encourages samples from different domains but the same label, to map nearby in the embedding space.

其中 C 是类别标签的数量，$X_a^s = X^s \mid \{Y = a\}$ 和 $X_a^t = X^t \mid \{Y = a\}$ 是条件随机变量。d 则是嵌入空间中 X_a^s 和 X_a^t 的分布之间的合适距离度量。我们将 (3) 称为语义对齐损失，它显然鼓励来自不同领域但具有相同标签的样本在嵌入空间中相互靠近。

While the analysis above clearly indicates why SDA provides superior performance than UDA, it also suggests that deep SDA approaches have not considered that greater performance could be achieved by encouraging class separation, meaning that samples from different domains and with different labels, should be mapped as far apart as possible in the embedding space. This idea means that, in principle, a semantic alignment less prone to errors should be achieved by adding to (3) the following term

虽然上述分析清楚地表明 SDA 提供的性能优于 UDA，但它也表明深度 SDA 方法并未考虑通过鼓励类别分离来实现更高的性能，这意味着来自不同领域且具有不同标签的样本应该在嵌入空间中尽可能远离。这一思想意味着，原则上，通过向 (3) 添加以下项，可以实现一种不易出错的语义对齐

$$\mathcal{L}_S(g) = \sum_{a,b \mid a \neq b} k\left(p\left(g\left(X_a^s\right)\right), p\left(g\left(X_b^t\right)\right)\right), \tag{4}$$

where k is a suitable similarity mesure between the distributions of X_a^s and X_b^t in the embedding space, which adds a penalty when the distributions $p\left(g\left(X_a^s\right)\right)$ and $p\left(g\left(X_b^t\right)\right)$ come close, since they would lead to lower classification accuracy. We refer to (4) as the separation loss.

其中 k 是嵌入空间中 X_a^s 和 X_b^t 的分布之间的合适相似度度量，当分布 $p\left(g\left(X_a^s\right)\right)$ 和 $p\left(g\left(X_b^t\right)\right)$ 彼此接近时会增加惩罚，因为这会导致较低的分类准确率。我们将 (4) 称为分离损失。

Finally, we suggest that SDA could be approached by learning a deep model $f = h \circ g$ such that

最后，我们建议通过学习一个深度模型 $f = h \circ g$ 来接近 SDA。

$$\mathcal{L}_{CCSA}(f) = \mathcal{L}_C(h \circ g) + \mathcal{L}_{SA}(g) + \mathcal{L}_S(g). \tag{5}$$

We refer to (5) as the classification and contrastive semantic alignment loss. This would allow to set $g_s = g_t = g$. The classification network h is trained only with source data, so $h_s = h$. In addition, to improve performance on the target domain, h_t could be obtained via fine-tuning based on the few samples in $\mathcal{D}_t$, i.e.,

我们将 (5) 称为分类和对比语义对齐损失。这将允许设置 $g_s = g_t = g$ 。分类网络 h 仅使用源数据进行训练，因此 $h_s = h$ 。此外，为了提高在目标领域的性能，h_t 可以通过基于 $\mathcal{D}_t$ 中的少量样本进行微调获得，即，

$$h_t = \text{ fine-tuning } \left(h \mid \mathcal{D}_t\right). \tag{6}$$

Note that the network architecture remains the one in Figure 1, only with a different loss, and training procedure.

请注意，网络架构仍然是图 1 中的架构，只是损失和训练过程不同。

3.2. Handling Scarce Target Data

3.2. 处理稀缺的目标数据

When the size of the labeled target training dataset $\mathcal{D}_t$ is very small, minimizing the loss (5) becomes a challenge. The problem is that the semantic alignment loss as well as the separation loss rely on computing distances and similarities between distributions, and those are very difficult to represent with as few as one data sample.

当标记的目标训练数据集 $\mathcal{D}_t$ 的大小非常小时，最小化损失 (5) 变得具有挑战性。问题在于，语义对齐损失以及分离损失依赖于计算分布之间的距离和相似性，而这些在只有一个数据样本的情况下非常难以表示。

Rather than attempting to characterize distributions with statistics that require enough data, because of the reduced size of $\mathcal{D}_t$, we compute the distance in the semantic alignment loss (3) by computing average pairwise distances between points in the embedding space, i.e., we compute

与其尝试用需要足够数据的统计特征来表征分布，由于 $\mathcal{D}_t$ 的大小减少，我们通过计算嵌入空间中点之间的平均成对距离来计算语义对齐损失 (3) 中的距离，即，我们计算

$$d\left(p\left(g\left(X_a^s\right)\right), p\left(g\left(X_a^t\right)\right)\right) = \sum_{i,j} d\left(g\left(x_i^s\right), g\left(x_j^t\right)\right), \tag{7}$$

where it is assumed $y_i^s = y_j^t = a$. The strength of this approach is that it allows even a single labeled target sample to be paired with all the source samples, effectively trying to semantically align the entire source data with the few target data. Similarly, we compute the similarities in the separation loss (4) by computing average pairwise similarities between points in the embedding space, i.e., we compute

其中假设 $y_i^s = y_j^t = a$ 。这种方法的优势在于，即使是单个标记的目标样本也可以与所有源样本配对，有效地尝试将整个源数据与少量目标数据进行语义对齐。同样，我们通过计算嵌入空间中点之间的平均成对相似性来计算分离损失 (4) 中的相似性，即，我们计算

$$k\left(p\left(g\left(X_a^s\right)\right), p\left(g\left(X_b^t\right)\right)\right) = \sum_{i,j} k\left(g\left(x_i^s\right), g\left(x_j^t\right)\right), \tag{8}$$

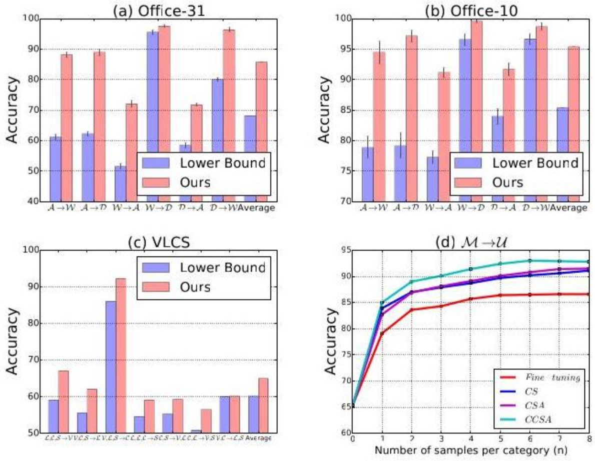

Figure 4. (a), (b), (c): Improvement of CCSA over the base model. (d): Average classification accuracy for the $\mathcal{M} \to \mathcal{U}$ task for different number of labeled target samples per category(n). It shows that our model provides significant improvement over baselines.

图 4. (a), (b), (c): CCSA 相对于基础模型的改进。(d): 针对每个类别标记目标样本数量 (n) 的 $\mathcal{M} \to \mathcal{U}$ 任务的平均分类准确率。结果表明，我们的模型相对于基线提供了显著的改进。

where it is assumed that $y_i^s = a \neq y_j^t = b$.

假设 $y_i^s = a \neq y_j^t = b$ 。

Moreover, our implementation further assumes that

此外，我们的实现进一步假设

$$d\left(g\left(x_i^s\right), g\left(x_j^t\right)\right) = \frac{1}{2}\left\|g\left(x_i^s\right) - g\left(x_j^t\right)\right\|^2, \tag{9}$$

$$k\left(g\left(x_i^s\right), g\left(x_j^t\right)\right) = \frac{1}{2}\max\left(0, m - \left\|g\left(x_i^s\right) - g\left(x_j^t\right)\right\|\right)^2 (10)$$

where $\|\cdot\|$ denotes the Frobenius norm, and m is the margin that specifies the separability in the embedding space. Note that with the choices outlined in (9) and (10), the loss $\mathcal{L}_{SA}(g) + \mathcal{L}_S(g)$ becomes the well known contrastive loss as defined in [30]. Finally, to balance the classification versus the contrastive semantic alignment portion of the loss (5), (7) and (8) are normalized and weighted by $1-\gamma$ and (1) by γ .

其中 $\|\cdot\|$ 表示弗罗贝尼乌斯范数，m 是指定嵌入空间中可分离性的边际。注意，在 (9) 和 (10) 中概述的选择下，损失 $\mathcal{L}_{SA}(g) + \mathcal{L}_S(g)$ 变为在 [30] 中定义的著名对比损失。最后，为了平衡损失 (5)、(7) 和 (8) 中的分类与对比语义对齐部分，这些损失被归一化并通过 $1-\gamma$ 和 (1) 通过 γ 加权。

4. Extension to Domain Generalization

4. 扩展到领域泛化

In visual domain generalization (DG), D labeled datasets $\mathcal{D}_{s_1}, \cdots, \mathcal{D}_{s_D}$, representative of D distinct source domains are given. The goal is to learn from them a visual classifier f that during testing is going to perform well on data $\mathcal{D}_t$, not available during training, thus representative of an unknown target domain.

在视觉领域泛化 (DG) 中，给定了代表 D 不同源领域的 $\mathcal{D}_{s_1}, \cdots, \mathcal{D}_{s_D}$ 标记数据集。目标是从中学习一个视觉分类器 f ，该分类器在测试时能够在训练期间不可用的数据 $\mathcal{D}_t$ 上表现良好，从而代表一个未知的目标领域。

The SDA method in Section 3 treats source and target datasets $\mathcal{D}_s$ and $\mathcal{D}_t$ almost symmetrically. In particular, the embedding g aims at achieving semantic alignment, while favoring class separation. The only asymmetry is in the prediction function h that is trained only on the source, to be then fine-tuned on the target.

第 3 节中的 SDA 方法几乎对称地处理源数据集和目标数据集 $\mathcal{D}_s$ 和 $\mathcal{D}_t$ 。特别是，嵌入 g 旨在实现语义对齐，同时促进类别分离。唯一的不对称在于预测函数 h ，该函数仅在源上训练，然后在目标上进行微调。

In domain generalization, we are not interested in adapting the classifier to the target domain, because it is unknown. Instead, we want to make sure that the embedding g maps to a domain invariant space. To do so we consider every distinct unordered pair of source domains(u, v), represented by $\mathcal{D}_{s_u}$ and $\mathcal{D}_{s_v}$, and, like in SDA, impose the semantic alignment loss (3) as well as the separation loss (4). Moreover, the losses are summed over every pair in order to make the map g as domain invariant as possible. Similarly, the classifier h should be as accurate as possible for any of the mapped samples, to maximize performance on an unseen target. This calls for having a fully symmetric learning for h by training it on all the source domains, meaning that the classification loss (1) is summed over every domain s_u . See Figure 2.

在领域泛化中，我们并不关注将分类器适应于目标领域，因为它是未知的。相反，我们希望确保嵌入 g 映射到一个领域不变的空间。为此，我们考虑每一对不同的无序源领域 (u, v)，由 $\mathcal{D}_{s_u}$ 和 $\mathcal{D}_{s_v}$ 表示，并像在 SDA 中一样，施加语义对齐损失 (3) 以及分离损失 (4)。此外，这些损失在每一对上求和，以使映射 g 尽可能地领域不变。同样，分类器 h 应该对任何映射样本尽可能准确，以最大化在未见目标上的性能。这要求对 h 进行完全对称的学习，通过在所有源领域上训练它，这意味着分类损失 (1) 在每个领域 s_u 上求和。见图 2。

The network architecture is still the one in Figure 1, and we have implemented it with the same choices for distances and similarities as those made in Section 3.2. However, since we are summing the losses (3) and (4) over every unordered pair of source domains, there is a quadratic growth of paired training samples. So, if necessary, rather than processing every paired sample, we select them randomly.

网络架构仍然是图 1 中的架构，我们已经使用与第 3.2 节中所做的距离和相似性相同的选择进行了实现。然而，由于我们在每一对无序源领域上求和损失 (3) 和 (4)，因此成对训练样本的数量呈二次增长。因此，如果有必要，我们会随机选择样本，而不是处理每一个成对样本。

5. Experiments

5. 实验

We divide the experiments into two parts, domain adaptation and domain generalization. In both sections, we use benchmark datasets and compare our domain adaptation model and our domain generalization model, both indicated as CCSA, with the state-of-the-art.

我们将实验分为两个部分，领域适应和领域泛化。在这两个部分中，我们使用基准数据集，并将我们的领域适应模型和领域泛化模型 (均表示为 CCSA) 与最先进的技术进行比较。

5.1. Domain Adaptation

5.1. 领域适应

We present results using the Office dataset [53], the MNIST dataset [37], and the USPS dataset [31].

我们展示了使用 Office 数据集 [53]、MNIST 数据集 [37] 和 USPS 数据集 [31] 的结果。

5.1.1 Office Dataset

5.1.1 Office 数据集

The office dataset is a standard benchmark dataset for visual domain adaptation. It contains 31 object classes for three domains: Amazon, Webcam, and DSLR, indicated as $\mathcal{A}$, $\mathcal{W}$, and $\mathcal{D}$, for a total of 4,652 images. We consider six domain shifts using the three domains ($\mathcal{A} \rightarrow \mathcal{W}, \mathcal{A} \rightarrow \mathcal{D}$, $\mathcal{W} \rightarrow \mathcal{A}, \mathcal{W} \rightarrow \mathcal{D}, \mathcal{D} \rightarrow \mathcal{A}$, and $\mathcal{D} \rightarrow \mathcal{W}$). We performed several experiments using this dataset.

办公室数据集是一个用于视觉领域适应的标准基准数据集。它包含三个领域的 31 个对象类别: 亚马逊、网络摄像头和单反相机，分别表示为 $\mathcal{A}$ 、$\mathcal{W}$ 和 $\mathcal{D}$ ，总共有 4,652 张图像。我们考虑使用这三个领域的六种领域转移 ($\mathcal{A} \rightarrow \mathcal{W}, \mathcal{A} \rightarrow \mathcal{D}$ 、$\mathcal{W} \rightarrow \mathcal{A}, \mathcal{W} \rightarrow \mathcal{D}, \mathcal{D} \rightarrow \mathcal{A}$ 和 $\mathcal{D} \rightarrow \mathcal{W}$ 。我们使用该数据集进行了几项实验。

First experiment. We followed the setting described in [60]. All classes of the office dataset and 5 train-test splits are considered. For the source domain, 20 examples per category for the Amazon domain, and 8 examples per category for the DSLR and Webcam domains are randomly selected for training for each split. Also, 3 labeled examples are randomly selected for each category in the target domain for training for each split. The rest of the target samples are used for testing. Note that we used the same splits generated by [60]. We also report the classification results of the SDA algorithm presented in [39] and [33]. In addition to the SDA algorithms, we report the results of some recent UDA algorithms. They follow a different experimental protocol compared to the SDA algorithms, and use all samples of the target domain in training as unlabeled data together with all samples of the source domain.

第一次实验。我们遵循了 [60] 中描述的设置。考虑了办公室数据集的所有类别和 5 个训练-测试拆分。对于源领域，随机选择亚马逊领域每个类别 20 个示例，以及单反相机和网络摄像头领域每个类别 8 个示例进行训练。对于目标领域，每个类别随机选择 3 个标记示例进行训练。其余的目标样本用于测试。请注意，我们使用了 [60] 生成的相同拆分。我们还报告了 [39] 和 [33] 中提出的 SDA 算法的分类结果。除了 SDA 算法外，我们还报告了一些最新 UDA 算法的结果。它们与 SDA 算法相比遵循不同的实验协议，并将目标领域的所有样本作为未标记数据与源领域的所有样本一起用于训练。

Table 1. Office dataset. Classification accuracy for domain adaptation over the 31 categories of the Office dataset. $\mathcal{A}, \mathcal{W}$, and $\mathcal{D}$ stand for Amazon, Webcam, and DSLR domain. Lower Bound is our base model without adaptation.

表 1. 办公室数据集。针对办公室数据集 31 个类别的领域适应分类准确率。$\mathcal{A}, \mathcal{W}$ 、$\mathcal{D}$ 分别代表亚马逊、网络摄像头和单反相机领域。下限是我们没有适应的基础模型。

		Unsupervised			Supervised		
	LOWER BOUND	[62]	[39]	[25]	[60]	[33]	CCSA
$\mathcal{A} \rightarrow \mathcal{W}$	61.2 ± 0.9	61.8 ± 0.4	68.5 ± 0.4	68.7 ± 0.3	82.7 ± 0.8	84.5 ± 1.7	88.2 ± 1.0
$\mathcal{A} \rightarrow \mathcal{D}$	62.3 ± 0.8	64.4 ± 0.3	67.0 ± 0.4	67.1 ± 0.3	86.1 ± 1.2	86.3 ± 0.8	89.0 ± 1.2
$\mathcal{W} \rightarrow \mathcal{A}$	51.6 ± 0.9	52.2 ± 0.4	53.1 ± 0.3	54.09 ± 0.5	65.0 ± 0.5	65.7 ± 1.7	72.1 ± 1.0
$\mathcal{W} \rightarrow \mathcal{D}$	95.6 ± 0.7	98.5 ± 0.4	99.0 ± 0.2	99.0 ± 0.2	97.6 ± 0.2	97.5 ± 0.7	97.6 ± 0.4
$\mathcal{D} \rightarrow \mathcal{A}$	58.5 ± 0.8	52.1 ± 0.8	54.0 ± 0.4	56.0 ± 0.5	66.2 ± 0.3	66.5 ± 1.0	71.8 ± 0.5
$\mathcal{D} \rightarrow \mathcal{W}$	80.1 ± 0.6	95.0 ± 0.5	96.0 ± 0.3	96.4 ± 0.3	95.7 ± 0.5	95.5 ± 0.6	96.4 ± 0.8
Average	68.2	70.6	72.9	73.6	82.21	82.68	85.8

		无监督			有监督		
	下界	[62]	[39]	[25]	[60]	[33]	CCSA
$\mathcal{A} \rightarrow \mathcal{W}$	61.2 ± 0.9	61.8 ± 0.4	68.5 ± 0.4	68.7 ± 0.3	82.7 ± 0.8	84.5 ± 1.7	88.2 ± 1.0
$\mathcal{A} \rightarrow \mathcal{D}$	62.3 ± 0.8	64.4 ± 0.3	67.0 ± 0.4	67.1 ± 0.3	86.1 ± 1.2	86.3 ± 0.8	89.0 ± 1.2
$\mathcal{W} \rightarrow \mathcal{A}$	51.6 ± 0.9	52.2 ± 0.4	53.1 ± 0.3	54.09 ± 0.5	65.0 ± 0.5	65.7 ± 1.7	72.1 ± 1.0
$\mathcal{W} \rightarrow \mathcal{D}$	95.6 ± 0.7	98.5 ± 0.4	99.0 ± 0.2	99.0 ± 0.2	97.6 ± 0.2	97.5 ± 0.7	97.6 ± 0.4
$\mathcal{D} \rightarrow \mathcal{A}$	58.5 ± 0.8	52.1 ± 0.8	54.0 ± 0.4	56.0 ± 0.5	66.2 ± 0.3	66.5 ± 1.0	71.8 ± 0.5
$\mathcal{D} \rightarrow \mathcal{W}$	80.1 ± 0.6	95.0 ± 0.5	96.0 ± 0.3	96.4 ± 0.3	95.7 ± 0.5	95.5 ± 0.6	96.4 ± 0.8
平均	68.2	70.6	72.9	73.6	82.21	82.68	85.8

Table 2. Office dataset. Classification accuracy for domain adaptation over the Office dataset when only the labeled target samples of 15 classes are available during training. Testing is done on all 31 classes. $\mathcal{A}, \mathcal{W}$, and $\mathcal{D}$ stand for Amazon, Webcam, and DSLR domain. Lower Bound is our base model without adaptation.

表 2. 办公室数据集。当训练期间仅提供 15 个类别的标记目标样本时，针对办公室数据集的领域适应分类准确率。测试在所有 31 个类别上进行。$\mathcal{A}, \mathcal{W}$ 、$\mathcal{D}$ 分别代表亚马逊、网络摄像头和单反相机领域。下限是我们没有适应的基础模型。

	LOWER BOUND	[60]	CCSA
$\mathcal{A} \to \mathcal{W}$	52.1 ± 0.6	59.3 ± 0.6	63.3 ± 0.9
$\mathcal{A} \to \mathcal{D}$	61.6 ± 0.8	68.0 ± 0.5	70.5 ± 0.6
$\mathcal{W} \to \mathcal{A}$	34.5 ± 0.9	40.5 ± 0.2	43.6 ± 1.0
$\mathcal{W} \to \mathcal{D}$	95.1 ± 0.2	97.5 ± 0.1	96.2 ± 0.3
$\mathcal{D} \to \mathcal{A}$	40.1 ± 0.3	43.1 ± 0.2	42.6 ± 0.6
$\mathcal{D} \to \mathcal{W}$	89.7 ± 0.8	90.0 ± 0.2	90.0 ± 0.2
Average	62.26	66.4	67.83

	下限	[60]	CCSA
$\mathcal{A} \to \mathcal{W}$	52.1 ± 0.6	59.3 ± 0.6	63.3 ± 0.9
$\mathcal{A} \to \mathcal{D}$	61.6 ± 0.8	68.0 ± 0.5	70.5 ± 0.6
$\mathcal{W} \to \mathcal{A}$	34.5 ± 0.9	40.5 ± 0.2	43.6 ± 1.0
$\mathcal{W} \to \mathcal{D}$	95.1 ± 0.2	97.5 ± 0.1	96.2 ± 0.3
$\mathcal{D} \to \mathcal{A}$	40.1 ± 0.3	43.1 ± 0.2	42.6 ± 0.6
$\mathcal{D} \to \mathcal{W}$	89.7 ± 0.8	90.0 ± 0.2	90.0 ± 0.2
平均值	62.26	66.4	67.83

For the embedding function g , we used the convolutional layers of the VGG-16 architecture [55] followed by 2 fully connected layers with output size of 1024 and 128, respectively. For the prediction function h , we used a fully connected layer with softmax activation. Similar to [60], we used the weights pre-trained on the ImageNet dataset [51] for the convolutional layers, and initialized the fully connected layers using all the source domain data. We then fine-tuned all the weights using the train-test splits.

对于嵌入函数 g ，我们使用了 VGG-16 架构的卷积层 [55]，后接两个全连接层，输出大小分别为 1024 和 128。对于预测函数 h ，我们使用了一个具有 softmax 激活的全连接层。与 [60] 类似，我们使用了在 ImageNet 数据集 [51] 上预训练的卷积层权重，并使用所有源域数据初始化全连接层。然后，我们使用训练-测试划分微调所有权重。

Table 1 reports the classification accuracy over 31 classes for the Office dataset and shows that CCSA has better performance compared to [60]. Since the difference between $\mathcal{W}$ domain and $\mathcal{D}$ domain is not considerable, unsupervised algorithms work well on $\mathcal{D} \to \mathcal{W}$ and $\mathcal{W} \to \mathcal{D}$. However, in the cases when target and source domains are very different $(\mathcal{A} \to \mathcal{W}, \mathcal{W} \to \mathcal{A}, \mathcal{A} \to \mathcal{D}$, and $\mathcal{D} \to \mathcal{A})$, CCSA shows larger margins compared to the second best. This suggests that CCSA will provide greater alignment gains when there are bigger domain shifts. Figure 4(a) instead, shows how much improvement can be obtained with respect to the base model. This is simply obtained by training g and h with only the classification loss and source training data, so no adaptation is performed.

表 1 报告了 Office 数据集 31 个类别的分类准确率，并显示 CCSA 相较于 [60] 具有更好的性能。由于 $\mathcal{W}$ 域和 $\mathcal{D}$ 域之间的差异不大，无监督算法在 $\mathcal{D} \to \mathcal{W}$ 和 $\mathcal{W} \to \mathcal{D}$ 上表现良好。然而，在目标和源域非常不同的情况下 $(\mathcal{A} \to \mathcal{W}, \mathcal{W} \to \mathcal{A}, \mathcal{A} \to \mathcal{D}$, and $\mathcal{D} \to \mathcal{A})$ ，CCSA 显示出比第二好的方法更大的边际。这表明，当存在更大的领域转移时，CCSA 将提供更大的对齐增益。图 4(a) 则显示了相对于基础模型可以获得多少改进。这是通过仅使用分类损失和源训练数据训练 g 和 h 获得的，因此没有进行适应。

Second experiment. We followed the setting described in [60] when only 10 target labeled samples of 15 classes of the Office dataset are available during training. Similar to [60], we compute the accuracy on the remaining 16 categories for which no target data was available during training. We used the same network structure as in the first experiment and the same splits generated by [60].

第二个实验。我们遵循了 [60] 中描述的设置，当只有 10 个目标标记样本和 15 个类别的 Office 数据集可用于训练时。与 [60] 类似，我们计算了在训练期间没有目标数据可用的其余 16 个类别的准确率。我们使用了与第一个实验相同的网络结构和 [60] 生成的相同划分。

Table 2 shows that CCSA is effective at transferring information from the labeled classes to the unlabeled target classes. Similar to the first experiment, CCSA works well when shifts between domains are larger.

表 2 显示 CCSA 在将信息从标记类别转移到未标记目标类别方面是有效的。与第一个实验类似，当领域之间的转移较大时，CCSA 表现良好。

Third experiment. We used the original train-test splits of the Office dataset [53]. The splits are generated in a similar manner to the first experiment but here instead, only 10 classes are considered (backpack, bike, calculator, headphones, keyboard, laptop-computer, monitor, mouse, mug, and projector). In order to compare our results with the state-of-the-art, we used DeCaF-fc6 features [14] and 800-dimension SURF features as input. For DeCaF-fc6 features (SURF features) we used 2 fully connected layers with output size of 1024 (512) and 128 (32) with ReLU activation as the embedding function, and one fully connected layer with softmax activation as the prediction function. The features and splits are available on the Office dataset webpage [1] .

第三个实验。我们使用了 Office 数据集的原始训练-测试划分 [53]。这些划分的生成方式与第一次实验相似，但这里仅考虑 10 个类别 (背包、自行车、计算器、耳机、键盘、笔记本电脑、显示器、鼠标、杯子和投影仪)。为了将我们的结果与最先进的技术进行比较，我们使用了 DeCaF-fc6 特征 [14] 和 800 维的 SURF 特征作为输入。对于 DeCaF-fc6 特征 (SURF 特征)，我们使用了 2 个全连接层，输出大小分别为 1024(512) 和 128(32)，激活函数为 ReLU，作为嵌入函数，并使用一个全连接层，激活函数为 softmax，作为预测函数。这些特征和划分可在 Office 数据集网页上获得 [1] 。

We compared our results with three UDA (GFK [26], mSDA [8], and RTML [13]) and one SDA (CDML [64]) algorithms under the same settings. Table 3 shows that CCSA provides an improved accuracy with respect to the others. Again, greater domain shifts are better compensated by CCSA . Figure 4(b) shows the improvement of CCSA over the base model using DeCaF-fc6 features.

我们在相同设置下将我们的结果与三种 UDA 算法 (GFK [26]、mSDA [8] 和 RTML [13]) 以及一种 SDA 算法 (CDML [64]) 进行了比较。表 3 显示，CCSA 在准确性方面优于其他算法。此外，CCSA 对更大的领域转移具有更好的补偿能力。图 4(b) 展示了 CCSA 在使用 DeCaF-fc6 特征时相对于基础模型的改进。

5.1.2 MNIST-USPS Datasets

5.1.2 MNIST-USPS 数据集

The MNIST(M)and USPS(U)datasets have recently been used for domain adaptation [20, 50] . They contain images

MNIST(M) 和 USPS(U) 数据集最近被用于领域适应 [20, 50] 。它们包含图像

Table 3. Office dataset. Classification accuracy for domain adaptation over the 10 categories of the Office dataset. $\mathcal{A}, \mathcal{W}$, and $\mathcal{D}$ stand for Amazon, Webcam, and DSLR domain. Lower Bound is our base model with no adaptation.

表 3. Office 数据集。针对 Office 数据集 10 个类别的领域适应分类准确性。$\mathcal{A}, \mathcal{W}$ 和 $\mathcal{D}$ 分别代表亚马逊、网络摄像头和单反相机领域。Lower Bound 是我们的基础模型，没有进行适应。

	LOWER BOUND	GFK [26]	mSDA [8]	CDML [64]	RTML [13]	CCSA
	SURF					
$\mathcal{A} \rightarrow \mathcal{W}$	26.5 ± 3.1	39.9 ± 0.9	35.5 ± 0.5	37.3 ± 0.7	43.4 ± 0.9	71.2 ± 1.3
$\mathcal{A} \rightarrow \mathcal{D}$	17.5 ± 1.2	36.2 ± 0.7	29.7 ± 0.7	35.3 ± 0.5	43.3 ± 0.6	74.2 ± 1.3
$\mathcal{W} \rightarrow \mathcal{A}$	25.9 ± 1.0	29.8 ± 0.6	32.1 ± 0.8	32.4 ± 0.5	37.5 ± 0.7	42.9 ± 0.9
$\mathcal{W} \rightarrow \mathcal{D}$	46.9 ± 1.1	80.9 ± 0.4	56.6 ± 0.4	77.9 ± 0.9	91.7 ± 1.1	85.1 ± 1.0
$\mathcal{D} \rightarrow \mathcal{A}$	19.3 ± 1.9	33.2 ± 0.6	33.6 ± 0.8	29.4 ± 0.8	36.3 ± 0.3	28.9 ± 1.3
$\mathcal{D} \rightarrow \mathcal{W}$	48.0 ± 2.1	79.4 ± 0.6	68.6 ± 0.7	79.4 ± 0.6	90.5 ± 0.7	77.3 ± 1.6
Average	30.6	43.5	38.4	43.5	49.8	63.2
DeCaF-fc6						
$\mathcal{A} \rightarrow \mathcal{W}$	78.9 ± 1.8	73.1 ± 2.8	64.6 ± 4.2	75.9 ± 2.1	79.5 ± 2.6	94.5 ± 1.9
$\mathcal{A} \rightarrow \mathcal{D}$	79.2 ± 2.1	82.6 ± 2.1	72.6 ± 3.5	81.4 ± 2.6	83.8 ± 1.7	97.2 ± 1.0
$\mathcal{W} \rightarrow \mathcal{A}$	77.3 ± 1.1	82.6 ± 1.3	71.4 ± 1.7	86.3 ± 1.6	90.8 ± 1.6	91.2 ± 0.8
$\mathcal{W} \rightarrow \mathcal{D}$	96.6 ± 1.0	98.8 ± 0.9	99.5 ± 0.6	99.4 ± 0.4	100 ± 0.0	99.6 ± 0.5
$\mathcal{D} \rightarrow \mathcal{A}$	84.0 ± 1.3	85.4 ± 0.7	78.8 ± 0.5	88.4 ± 0.5	90.6 ± 0.5	91.7 ± 1.0
$\mathcal{D} \rightarrow \mathcal{W}$	96.7 ± 0.9	91.3 ± 0.4	97.5 ± 0.4	95.1 ± 0.5	98.6 ± 0.3	98.7 ± 0.6
Average	85.4	85.63	80.73	87.75	90.55	95.4

[1] https://cs.stanford.edu/ ∼ jhoffman/domainadapt/

[1] https://cs.stanford.edu/ ∼ jhoffman/domainadapt/

	下限	GFK [26]	mSDA [8]	CDML [64]	RTML [13]	CCSA
	SURF					
$\mathcal{A} \rightarrow \mathcal{W}$	26.5 ± 3.1	39.9 ± 0.9	35.5 ± 0.5	37.3 ± 0.7	43.4 ± 0.9	71.2 ± 1.3
$\mathcal{A} \rightarrow \mathcal{D}$	17.5 ± 1.2	36.2 ± 0.7	29.7 ± 0.7	35.3 ± 0.5	43.3 ± 0.6	74.2 ± 1.3
$\mathcal{W} \rightarrow \mathcal{A}$	25.9 ± 1.0	29.8 ± 0.6	32.1 ± 0.8	32.4 ± 0.5	37.5 ± 0.7	42.9 ± 0.9
$\mathcal{W} \rightarrow \mathcal{D}$	46.9 ± 1.1	80.9 ± 0.4	56.6 ± 0.4	77.9 ± 0.9	91.7 ± 1.1	85.1 ± 1.0
$\mathcal{D} \rightarrow \mathcal{A}$	19.3 ± 1.9	33.2 ± 0.6	33.6 ± 0.8	29.4 ± 0.8	36.3 ± 0.3	28.9 ± 1.3
$\mathcal{D} \rightarrow \mathcal{W}$	48.0 ± 2.1	79.4 ± 0.6	68.6 ± 0.7	79.4 ± 0.6	90.5 ± 0.7	77.3 ± 1.6
平均	30.6	43.5	38.4	43.5	49.8	63.2
DeCaF-fc6						
$\mathcal{A} \rightarrow \mathcal{W}$	78.9 ± 1.8	73.1 ± 2.8	64.6 ± 4.2	75.9 ± 2.1	79.5 ± 2.6	94.5 ± 1.9
$\mathcal{A} \rightarrow \mathcal{D}$	79.2 ± 2.1	82.6 ± 2.1	72.6 ± 3.5	81.4 ± 2.6	83.8 ± 1.7	97.2 ± 1.0
$\mathcal{W} \rightarrow \mathcal{A}$	77.3 ± 1.1	82.6 ± 1.3	71.4 ± 1.7	86.3 ± 1.6	90.8 ± 1.6	91.2 ± 0.8
$\mathcal{W} \rightarrow \mathcal{D}$	96.6 ± 1.0	98.8 ± 0.9	99.5 ± 0.6	99.4 ± 0.4	100 ± 0.0	99.6 ± 0.5
$\mathcal{D} \rightarrow \mathcal{A}$	84.0 ± 1.3	85.4 ± 0.7	78.8 ± 0.5	88.4 ± 0.5	90.6 ± 0.5	91.7 ± 1.0
$\mathcal{D} \rightarrow \mathcal{W}$	96.7 ± 0.9	91.3 ± 0.4	97.5 ± 0.4	95.1 ± 0.5	98.6 ± 0.3	98.7 ± 0.6
平均	85.4	85.63	80.73	87.75	90.55	95.4

Table 4. MNIST-USPS datasets. Classification accuracy for domain adaptation over the MNIST and USPS datasets. $\mathcal{M}$ and $\mathcal{U}$ stand for MNIST and USPS domain. Lower Bound is our base model without adaptation. CCSA - n stands for our method when we use n labeled target samples per category in training. of digits from 0 to 9 . We considered two cross-domain tasks, $\mathcal{M} \rightarrow \mathcal{U}$ and $\mathcal{U} \rightarrow \mathcal{M}$, and followed the experimental setting in $[20, 50]$, which involves randomly selecting 2000 images from MNIST and 1800 images from USPS. Here, we randomly selected n labeled samples per class from target domain data and used them in training. We evaluated our approach for n ranging from 1 to 8 and repeated each experiment 10 times (we only show the mean of the accuracies because the standard deviation is very small).

表 4. MNIST-USPS 数据集。针对 MNIST 和 USPS 数据集的领域适应分类准确率。$\mathcal{M}$ 和 $\mathcal{U}$ 分别代表 MNIST 和 USPS 领域。Lower Bound 是我们未进行适应的基础模型。CCSA - n 代表我们在训练中使用每个类别的 n 标记目标样本时的方法。数字范围为 0 到 9。我们考虑了两个跨领域任务，$\mathcal{M} \rightarrow \mathcal{U}$ 和 $\mathcal{U} \rightarrow \mathcal{M}$ ，并遵循了 $[20, 50]$ 中的实验设置，该设置涉及从 MNIST 随机选择 2000 张图像，从 USPS 随机选择 1800 张图像。在这里，我们从目标领域数据中随机选择了 n 个标记样本，并在训练中使用它们。我们评估了我们的方法，范围从 n 为 1 到 8，并重复每个实验 10 次 (我们仅显示准确率的均值，因为标准差非常小)。

Method	$\mathcal{M} \rightarrow \mathcal{U}$	$\mathcal{U} \rightarrow \mathcal{M}$	Average
ADDA [61]	89.4	90.1	89.7
COGAN [38]	91.2	89.1	90.1
LOWER BOUND	65.4	58.6	62.0
CCSA-1	85.0	78.4	81.7
CCSA-2	89.0	82.0	85.5
CCSA-3	90.1	85.8	87.9
CCSA-4	91.4	86.1	88.7
CCSA-5	92.4	88.8	90.1
CCSA-6	93.0	89.6	91.3
CCSA-7	92.9	89.4	91.1
CCSA-8	92.8	90.0	91.4

方法	$\mathcal{M} \rightarrow \mathcal{U}$	$\mathcal{U} \rightarrow \mathcal{M}$	平均
ADDA [61]	89.4	90.1	89.7
COGAN [38]	91.2	89.1	90.1
下界	65.4	58.6	62.0
CCSA-1	85.0	78.4	81.7
CCSA-2	89.0	82.0	85.5
CCSA-3	90.1	85.8	87.9
CCSA-4	91.4	86.1	88.7
CCSA-5	92.4	88.8	90.1
CCSA-6	93.0	89.6	91.3
CCSA-7	92.9	89.4	91.1
CCSA-8	92.8	90.0	91.4

Table 5. VLCS dataset. Classification accuracy for domain generalization over the 5 categories of the VLCS dataset. LB (Lower Bound) is our base model trained without the contrastive semantic alignment loss. 1NN stands for first nearest neighbor.

表 5. VLCS 数据集。针对 VLCS 数据集 5 个类别的领域泛化分类准确率。LB(Lower Bound) 是我们未使用对比语义对齐损失训练的基础模型。1NN 代表第一个最近邻。

	LOWER BOUND			Domain Generalization			
	1NN	SVM	LB	UML [17]	LRE-SVM [65]	SCA [23]	CCSA
$\mathcal{L},\mathcal{C},\mathcal{S} \to \mathcal{V}$	57.2	58.4	59.1	56.2	60.5	64.3	67.1
$\mathcal{V},\mathcal{C},\mathcal{S} \to \mathcal{L}$	52.4	55.2	55.6	58.5	59.7	59.6	62.1
$\mathcal{V},\mathcal{L},\mathcal{S} \to \mathcal{C}$	90.5	85.1	86.1	91.1	88.1	88.9	92.3
$\mathcal{V},\mathcal{L},\mathcal{C} \to \mathcal{S}$	56.9	55.2	54.6	58.4	54.8	59.2	59.1
$\mathcal{C},\mathcal{S} \to \mathcal{V},\mathcal{L}$	55.0	55.5	55.3	56.4	55.0	59.5	59.3
$\mathcal{C},\mathcal{L} \to \mathcal{V},\mathcal{S}$	52.6	51.8	50.9	57.4	52.8	55.9	56.5
$\mathcal{V},\mathcal{C} \to \mathcal{L},\mathcal{S}$	56.6	59.9	60.1	55.4	58.8	60.7	60.2
Average	60.1	60.1	60.2	61.5	61.4	64.0	65.0

	下界			领域泛化			
	1NN	支持向量机 (SVM)	LB	UML [17]	LRE-SVM [65]	SCA [23]	CCSA
$\mathcal{L},\mathcal{C},\mathcal{S} \to \mathcal{V}$	57.2	58.4	59.1	56.2	60.5	64.3	67.1
$\mathcal{V},\mathcal{C},\mathcal{S} \to \mathcal{L}$	52.4	55.2	55.6	58.5	59.7	59.6	62.1
$\mathcal{V},\mathcal{L},\mathcal{S} \to \mathcal{C}$	90.5	85.1	86.1	91.1	88.1	88.9	92.3
$\mathcal{V},\mathcal{L},\mathcal{C} \to \mathcal{S}$	56.9	55.2	54.6	58.4	54.8	59.2	59.1
$\mathcal{C},\mathcal{S} \to \mathcal{V},\mathcal{L}$	55.0	55.5	55.3	56.4	55.0	59.5	59.3
$\mathcal{C},\mathcal{L} \to \mathcal{V},\mathcal{S}$	52.6	51.8	50.9	57.4	52.8	55.9	56.5
$\mathcal{V},\mathcal{C} \to \mathcal{L},\mathcal{S}$	56.6	59.9	60.1	55.4	58.8	60.7	60.2
平均	60.1	60.1	60.2	61.5	61.4	64.0	65.0

Similar to [37], we used 2 convolutional layers with 6 and 16 filters of 5×5 kernels followed by max-pooling layers and 2 fully connected layers with size 120 and 84 as the embedding function g , and one fully connected layer with softmax activation as the prediction function h . We compare our method with 2 recent UDA methods. Those methods use all target samples in their training stage, while we only use very few labeled target samples per category in training.

类似于 [37]，我们使用了 2 个卷积层，分别具有 6 和 16 个 5×5 核的滤波器，后接最大池化层和 2 个大小为 120 和 84 的全连接层作为嵌入函数 g ，以及一个具有 softmax 激活的全连接层作为预测函数 h 。我们将我们的方法与 2 种最近的 UDA 方法进行了比较。这些方法在训练阶段使用了所有目标样本，而我们在训练中仅使用了每个类别的少量标记目标样本。

Table 4 shows the average classification accuracy of the MNIST-USPS datasets. CCSA works well even when only one target sample per category $(n = 1)$ is available in training. Also, we can see that by increasing n , the accuracy quickly converges to the top.

表 4 显示了 MNIST-USPS 数据集的平均分类准确率。即使在训练中每个类别只有一个目标样本 $(n = 1)$ 可用时，CCSA 仍然表现良好。此外，我们可以看到，通过增加 n ，准确率迅速收敛到最高值。

Ablation study. We consider three baselines to compare with CCSA for the $\mathcal{M} \to \mathcal{U}$ task. First, we train the network with source data and then fine-tune it with available target data. Second, we train the network using the classification and semantic alignment losses $(\mathcal{L}_{CSA}(f) = \mathcal{L}_C(h \circ g) + \mathcal{L}_{SA}(g))$. Third, we train the network using the classification and separation losses $(\mathcal{L}_{CS}(f) = \mathcal{L}_C(h \circ g) + \mathcal{L}_S(g))$. Figure 4(d) shows the average accuracies over 10 repetitions. It shows that CSA and CS improve the accuracy over fine-tuning, while using the proposed CCSA loss shows the best performance.

消融研究。我们考虑三个基线与 CCSA 进行比较，以完成 $\mathcal{M} \to \mathcal{U}$ 任务。首先，我们使用源数据训练网络，然后用可用的目标数据进行微调。其次，我们使用分类和语义对齐损失 $(\mathcal{L}_{CSA}(f) = \mathcal{L}_C(h \circ g) + \mathcal{L}_{SA}(g)$ 训练网络。第三，我们使用分类和分离损失 $(\mathcal{L}_{CS}(f) = \mathcal{L}_C(h \circ g) + \mathcal{L}_S(g))$ 训练网络。图 4(d) 显示了 10 次重复的平均准确率。结果表明，CSA 和 CS 在微调的基础上提高了准确率，而使用提议的 CCSA 损失则显示出最佳性能。

Visualization. We show how samples lie on the embedding space using CCSA . First, we considered the row images of the MNIST and USPS datasets and plotted 2D visualization of them using t-SNE [41]. As Figure 3(Left) shows the row images of the same class and different domains lie far away from each other in the 2D subspace. For example, the samples of the class zero of the USPS dataset(0U)are far from the class zero of the MNIST dataset(0M). Second, we trained our base model with no adaptation on the MNIST dataset. We then plotted the 2D visualization of the MNIST and USPS samples in the embedding space (output of g , the last fully connected layer). As Figure 3(Middle) shows, the samples from the same class and different domains still lie far away from each other in the 2D subspace. Finally, we trained our SDA model on the MNIST dataset and 3 labeled samples per class of the USPS dataset. We then plotted the 2D visualization of the MNIST and USPS samples in the embedding space (output of g). As Figure 3(Right) shows, the samples from the same class and different domains now lie very close to each other in the 2D subspace. Note however, that this is only a 2D visualization of high-dimensional data, and Figure 3(Right) may not perfectly reflect how close is the data from the same class, and how classes are separated.

可视化。我们展示了样本在嵌入空间中的分布，使用了 CCSA。首先，我们考虑了 MNIST 和 USPS 数据集的行图像，并使用 t-SNE [41] 绘制了它们的 2D 可视化。如图 3(左) 所示，同一类别和不同领域的行图像在 2D 子空间中相距甚远。例如，USPS 数据集的零类样本 (0U) 与 MNIST 数据集的零类样本 (0M) 相距较远。其次，我们在 MNIST 数据集上训练了没有适应的基础模型。然后，我们绘制了嵌入空间中 MNIST 和 USPS 样本的 2D 可视化 (g 的输出，即最后一个全连接层)。如图 3(中) 所示，同一类别和不同领域的样本在 2D 子空间中仍然相距甚远。最后，我们在 MNIST 数据集上以及 USPS 数据集中每个类别的 3 个标记样本上训练了我们的 SDA 模型。然后，我们绘制了嵌入空间中 MNIST 和 USPS 样本的 2D 可视化 (g 的输出)。如图 3(右) 所示，同一类别和不同领域的样本现在在 2D 子空间中非常接近。然而，请注意，这仅仅是高维数据的 2D 可视化，图 3(右) 可能并不能完美反映同一类别数据的接近程度，以及类别之间的分离情况。

Weight sharing: There is no restriction on whether or not g_t and g_s should share weights. Not sharing weights likely leads to overfitting, given the reduced amount of target training data, and weight-sharing acts as a regularizer. For instance, we repeated the experiment for the $\mathcal{M} \to \mathcal{U}$ task with $n = 4$. Not sharing weights provides an average accuracy of 88.6 over 10 repetitions, which is less than the average accuracy with weight-sharing (see Table 4). A similar behavior is observable in other experiments.

权重共享: 对于 g_t 和 g_s 是否应共享权重没有限制。不共享权重可能导致过拟合，因为目标训练数据量减少，而权重共享则充当正则化器。例如，我们对 $\mathcal{M} \to \mathcal{U}$ 任务进行了重复实验，使用了 $n = 4$ 。不共享权重的情况下，10 次重复的平均准确率为 88.6，低于共享权重的平均准确率 (见表 4)。在其他实验中也可以观察到类似的行为。

5.2. Domain Generalization

5.2. 域泛化

We evaluate CCSA on different datasets. The goal is to show that CCSA is able to learn a domain invariant embedding subspace for visual recognition tasks.

我们在不同的数据集上评估 CCSA。目标是展示 CCSA 能够为视觉识别任务学习一个域不变的嵌入子空间。

5.3. VLCS Dataset

5.3. VLCS 数据集

In this section, we use images of 5 shared object categories (bird, car, chair, dog, and person), of the PASCAL VOC2007 ($\mathcal{V}$) [16], LabelMe ($\mathcal{L}$) [52], Caltech-101 (C)[18], and SUN09(S)[10] datasets, which is known as VLCS dataset [17].

在本节中，我们使用 PASCAL VOC2007 ($\mathcal{V}$) [16]、LabelMe ($\mathcal{L}$) [52]、Caltech-101 (C) [18] 和 SUN09 (S) [10] 数据集中 5 个共享对象类别 (鸟、车、椅子、狗和人) 的图像，这被称为 VLCS 数据集 [17]。

[24, 23] have shown that there are covariate shifts between the above 4 domains and have developed a DG method to minimize them. We followed their experimental setting, and randomly divided each domain into a training set (70%) and a test set (30%) and conducted a leave-one-domain-out evaluation (4 cross-domain cases) and a leave-two-domain-out evaluation (3 cross-domain cases). In order to compare our results with the state-of-the-art, we used DeCaF-fc6 features which are publicly available [2] , and repeated each cross-domain case 20 times and reported the average classification accuracy.

[24, 23] 已经表明上述 4 个领域之间存在协变量转移，并开发了一种 DG 方法来最小化它们。我们遵循了他们的实验设置，随机将每个领域分为训练集 (70%) 和测试集 (30%) ，并进行了留一域评估 (4 个跨域案例) 和留二域评估 (3 个跨域案例)。为了将我们的结果与最先进的技术进行比较，我们使用了公开可用的 DeCaF-fc6 特征 [2] ，并重复每个跨域案例 20 次，报告平均分类准确率。

We used 2 fully connected layers with output size of 1024 and 128 with ReLU activation as the embedding function g , and one fully connected layer with softmax activation as the prediction function h . To create positive and negative pairs for training our network, for each sample of a source domain we randomly selected 5 samples from each remaining source domain, and help in this way to avoid overfitting. However, to train a deeper network together with convolutional layers, it is enough to create a large amount of positive and negative pairs.

我们使用了两个全连接层，输出大小分别为 1024 和 128，并采用 ReLU 激活作为嵌入函数 g ，以及一个带有 softmax 激活的全连接层作为预测函数 h 。为了为训练我们的网络创建正负样本对，对于每个源领域的样本，我们随机从其余源领域中选择 5 个样本，以此方式帮助避免过拟合。然而，为了训练一个更深的网络与卷积层结合，创建大量的正负样本对是足够的。

Table 6. MNIST dataset. Classification accuracy for domain generalization over the MNIST dataset and its rotated domains.

表 6. MNIST 数据集。MNIST 数据集及其旋转领域的领域泛化分类准确率。

	CAE [49]	MTAE [24]	CCSA
$M_{15^\circ}, M_{30^\circ}, M_{45^\circ}, M_{60^\circ}, M_{75^\circ} \to M$	72.1	82.5	84.6
$M, M_{30^\circ}, M_{45^\circ}, M_{60^\circ}, M_{75^\circ} \to M_{15^\circ}$	95.3	96.3	95.6
$M, M_{15^\circ}, M_{45^\circ}, M_{60^\circ}, M_{75^\circ} \to M_{30^\circ}$	92.6	93.4	94.6
$M, M_{15^\circ}, M_{30^\circ}, M_{60^\circ}, M_{75^\circ} \to M_{45^\circ}$	81.5	78.6	82.9
$M, M_{15^\circ}, M_{30^\circ}, M_{45^\circ}, M_{75^\circ} \to M_{60^\circ}$	92.7	94.2	94.8
$M, M_{15^\circ}, M_{30^\circ}, M_{45^\circ}, M_{60^\circ} \to M_{75^\circ}$	79.3	80.5	82.1
Average	85.5	87.5	89.1

	CAE [49]	MTAE [24]	CCSA
$M_{15^\circ}, M_{30^\circ}, M_{45^\circ}, M_{60^\circ}, M_{75^\circ} \to M$	72.1	82.5	84.6
$M, M_{30^\circ}, M_{45^\circ}, M_{60^\circ}, M_{75^\circ} \to M_{15^\circ}$	95.3	96.3	95.6
$M, M_{15^\circ}, M_{45^\circ}, M_{60^\circ}, M_{75^\circ} \to M_{30^\circ}$	92.6	93.4	94.6
$M, M_{15^\circ}, M_{30^\circ}, M_{60^\circ}, M_{75^\circ} \to M_{45^\circ}$	81.5	78.6	82.9
$M, M_{15^\circ}, M_{30^\circ}, M_{45^\circ}, M_{75^\circ} \to M_{60^\circ}$	92.7	94.2	94.8
$M, M_{15^\circ}, M_{30^\circ}, M_{45^\circ}, M_{60^\circ} \to M_{75^\circ}$	79.3	80.5	82.1
平均值	85.5	87.5	89.1

We report comparative results in Table 5, where all DG methods work better than the base model, emphasizing the need for domain generalization. Our DG method has higher average performance. Also, note that in order to compare with the state-of-the-art DG methods, we only used 2 fully connected layers for our network and precomputed features as input. However, when using convolutional layers on row images, we expect our DG model to provide better performance. Figure 4(c) shows the improvement of our DG model over the base model using DeCaF-fc6 features.

我们在表 5 中报告了比较结果，所有领域泛化 (DG) 方法的表现均优于基础模型，强调了领域泛化的必要性。我们的 DG 方法具有更高的平均性能。此外，请注意，为了与最先进的 DG 方法进行比较，我们仅为我们的网络使用了两个全连接层，并将预计算特征作为输入。然而，当在原始图像上使用卷积层时，我们期望我们的 DG 模型能提供更好的性能。图 4(c) 显示了我们的 DG 模型在使用 DeCaF-fc6 特征时相对于基础模型的改进。

5.4. MNIST Dataset

5.4. MNIST 数据集

We followed the setting in [24], and randomly selected a set M of 100 images per category from the MNIST dataset (1000 in total). We then rotated each image in M five times with 15 degrees intervals, creating five new domains $M_{15^\circ}, M_{30^\circ}, M_{45^\circ}, M_{60^\circ}$, and M_{75° . We conducted a leave-one-domain-out evaluation (6 cross-domain cases in total). We used the same network of Section 5.1.2, and we repeated the experiments 10 times. To create positive and negative pairs for training our network, for each sample of a source domain we randomly selected 2 samples from each remaining source domain. We report comparative average accuracies for CCSA and others in Table 6, showing again a performance improvement.

我们遵循了 [24] 中的设置，并随机从 MNIST 数据集中每个类别选择了一组 M 的 100 张图像 (总共 1000 张)。然后，我们将每张图像旋转 M 五次，每次间隔 15 度，创建了五个新领域 $M_{15^\circ}, M_{30^\circ}, M_{45^\circ}, M_{60^\circ}$

和 $M_{75^{\circ}}$。我们进行了留一领域评估 (总共 6 个跨领域案例)。我们使用了第 5.1.2 节中的相同网络，并重复进行了 10 次实验。为了为我们的网络创建正负样本对，对于每个源领域的样本，我们随机从其余源领域中选择 2 个样本。我们在表 6 中报告了 CCSA 和其他方法的比较平均准确率，再次显示了性能的提升。

6. Conclusions

6. 结论

We have introduced a deep model in combination with the classification and contrastive semantic alignment (CCSA) loss to address SDA. We have shown that the CCSA loss can be augmented to address the DG problem without the need to change the basic model architecture. However, the approach is general in the sense that the architecture sub-components can be changed. We found that addressing the semantic distribution alignments with pointwise surrogates of distribution distances and similarities for SDA and DG works very effectively, even when labeled target samples are very few. In addition, we found the SDA accuracy to converge very quickly as more labeled target samples per category are available.

我们引入了一种深度模型，结合分类和对比语义对齐 (CCSA) 损失来解决 SDA 问题。我们展示了 CCSA 损失可以增强以解决 DG 问题，而无需更改基本模型架构。然而，这种方法是通用的，因为架构的子组件可以更改。我们发现，使用分布距离和相似性的逐点替代品来解决 SDA 和 DG 的语义分布对齐非常有效，即使标记的目标样本非常少。此外，我们发现，随着每个类别可用的标记目标样本增多，SDA 的准确率收敛得非常快。

Acknowledgments

致谢

This material is based upon work supported, in part, by the Center for Identification Technology Research and the National Science Foundation under Grant No. 1066197.

本材料部分基于由识别技术研究中心和国家科学基金会支持的工作，资助号为 1066197。

References

参考文献

[1] Y. Aytar and A. Zisserman. Tabula rasa: Model transfer for object category detection. In Computer Vision (ICCV), 2011 IEEE International Conference on, pages 2252-2259. IEEE, 2011.

[2] M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann. Unsupervised domain adaptation by domain invariant projection. In IEEE ICCV, pages 769-776, 2013.

[3] C. J. Becker, C. M. Christoudias, and P. Fua. Non-linear domain adaptation with boosting. In Advances in Neural Information Processing Systems, pages 485-493, 2013.

[4] A. Bergamo and L. Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In Advances in Neural Information Processing Systems, pages 181-189, 2010.

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

[6] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In Proceedings of the 2006 conference on empirical methods in natural language processing, pages 120-128. Association for Computational Linguistics, 2006.

[7] L. Chen, W. Li, and D. Xu. Recognizing RGB images by learning from RGB-D data. In CVPR, pages 1418-1425, June 2014.

[8] M. Chen, Z. Xu, K. Weinberger, and F. Sha. Marginalized de-noising autoencoders for domain adaptation. arXiv preprint arXiv:1206.4683, 2012.

[2] http://www.cs.dartmouth.edu/ ∼ chenfang/proj-page /FXR_iccv13/index.php

[2] http://www.cs.dartmouth.edu/ ∼ chenfang/proj-page /FXR_iccv13/index.php

[9] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5315-5324, 2015.

[10] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In Computer vision and pattern recognition (CVPR), 2010 IEEE conference on, pages 129-136. IEEE, 2010.

[11] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 539-546. IEEE, 2005.

[12] H. Daume III and D. Marcu. Domain adaptation for statistical classifiers. Journal of Artificial Intelligence Research, 26:101-126, 2006.

[13] Z. Ding and Y. Fu. Robust transfer metric learning for image classification. IEEE Transactions on Image Processing, 26(2):660-670, 2017.

[14] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: a deep convolutional activation feature for generic visual recognition. In arXiv:1310.1531, 2013.

[15] L. Duan, I. W. Tsang, D. Xu, and S. J. Maybank. Domain transfer svm for video concept detection. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 1375-1381. IEEE, 2009.

[16] 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, 88(2):303- 338, 2010.

[17] C. Fang, Y. Xu, and D. N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In International Conference on Computer Vision, 2013.

[18] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Computer vision and Image understanding, 106(1):59-70, 2007.

[19] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In IEEE ICCV, pages 2960-2967, 2013.

[20] B. Fernando, T. Tommasi, and T. Tuytelaarsc. Joint cross-domain classification and subspace learning for unsupervised adaptation. Pattern Recogition Letters, 2015.

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

[22] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59):1–35, 2016.

[23] M. Ghifary, D. Balduzzi, W. B. Kleijn, and M. Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

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

[25] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In European Conference on Computer Vision, pages 597-613. Springer, 2016.

[26] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 2066-2073. IEEE, 2012.

[27] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In IEEE ICCV, pages 999-1006, 2011.

[28] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In NIPS, 2006.

[29] Y. Guo and M. Xiao. Cross language text classification via subspace co-regularized multi-view learning. In Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012, 2012.

[30] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In Computer vision and pattern recognition, 2006 IEEE computer society conference on, volume 2, pages 1735-1742. IEEE, 2006.

[31] J. J. Hull. A database for handwritten text recognition research. IEEE Transactions on pattern analysis and machine intelligence, 16(5):550-554, 1994.

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

[33] P. Koniusz, Y. Tas, and F. Porikli. Domain adaptation by mixture of alignments of second-or higher-order scatter tensors. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[34] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 1785-1792. IEEE, 2011.

[35] B. Kumar, G. Carneiro, I. Reid, et al. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5385-5394, 2016.

[36] M. Lapin, M. Hein, and B. Schiele. Learning using privileged information: SVM+ and weighted SVM. Neural Networks, 53:95-108, 2014.

[37] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278-2324, 1998.

[38] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In Advances in Neural Information Processing Systems, pages 469-477, 2016.

[39] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In ICML, pages 97-105, 2015.

[40] M. Long, G. Ding, J. Wang, J. Sun, Y. Guo, and P. S. Yu. Transfer sparse coding for robust image representation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 407-414, 2013.

[41] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. Journal of Machine Learning Research, 9(Nov):2579-2605, 2008.

[42] S. Motiian and G. Doretto. Information bottleneck domain adaptation with privileged information for visual recognition. In European Conference on Computer Vision, pages 630-647. Springer, 2016.

[43] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto. Information bottleneck learning using privileged information for visual recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1496-1505, 2016.

[44] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In ICML (1), pages 10-18, 2013.

[45] L. Niu, W. Li, and D. Xu. Multi-view domain generalization for visual recognition. In Proceedings of the IEEE International Conference on Computer Vision, pages 4193-4201, 2015.

[46] L. Niu, W. Li, D. Xu, and J. Cai. An exemplar-based multiview domain generalization framework for visual recognition. IEEE Transactions on Neural Networks and Learning Systems, 2017.

[47] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. IEEE TNN, 22(2):199–210, 2011.

[48] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, et al. Dataset issues in object recognition. In Toward category-level object recognition, pages 29–48. Springer, 2006.

[49] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In Proceedings of the 28th international conference on machine learning (ICML-11), pages 833-840, 2011.

[50] A. Rozantsev, M. Salzmann, and P. Fua. Beyond sharing weights for deep domain adaptation. arXiv preprint arXiv:1603.06432, 2016.

[51] 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. IJCV, 2015.

[52] 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, 77(1- 3):157-173, 2008.

[53] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In ECCV, pages 213- 226, 2010.

[54] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference, 90(2):227-244, 2000.

[55] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.

[56] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In Computer Vision-ECCV 2016 Workshops, pages 443-450. Springer, 2016.

[57] T. Tommasi, M. Lanzi, P. Russo, and B. Caputo. Learning the roots of visual domain shift. In Computer Vision-ECCV 2016 Workshops, pages 475-482. Springer, 2016.

[58] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars. A deeper look at dataset bias. In German Conference on Pattern Recognition, pages 504-516. Springer, 2015.

[59] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 1521-1528, 2011.

[60] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In ICCV, 2015.

[61] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[62] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474, 2014.

[63] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human reidentification. In European Conference on Computer Vision, pages 135-153. Springer, 2016.

[64] H. Wang, W. Wang, C. Zhang, and F. Xu. Cross-domain metric learning based on information theory. In AAAI, pages 2099-2105, 2014.

[65] Z. Xu, W. Li, L. Niu, and D. Xu. Exploiting low-rank structure from latent domains for domain generalization. In ECCV, pages 628-643, 2014.

[66] J. Yang, R. Yan, and A. G. Hauptmann. Adapting svm classifiers to data with shifted distributions. In Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on, pages 69-76. IEEE, 2007.

[67] T. Yao, Y. Pan, C.-W. Ngo, H. Li, and T. Mei. Semi-supervised domain adaptation with subspace learning for visual recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.