

Domain Adaptation via Prompt Learning

通过提示学习进行领域适应

Chunjiang Ge¹ Rui Huang¹ Mixue Xie² Zihang Lai³

Chunjiang Ge¹ Rui Huang¹ Mixue Xie² Zihang Lai³

Shiji Song¹ Shuang Li² Gao Huang^{1,4}

Shiji Song¹ Shuang Li² Gao Huang^{1,4}

¹ Department of Automation, BNRist, Tsinghua University ² Beijing Institute of Technology

¹ 清华大学自动化系, BNRist ² 北京理工大学

³ Carnegie Mellon University ⁴ Beijing Academy of Artificial Intelligence

³ 卡内基梅隆大学 [latex1] 北京人工智能研究院

Abstract

摘要

Unsupervised domain adaption (UDA) aims to adapt models learned from a well-annotated source domain to a target domain, where only unlabeled samples are given. Current UDA approaches learn domain-invariant features by aligning source and target feature spaces. Such alignments are imposed by constraints such as statistical discrepancy minimization or adversarial training. However, these constraints could lead to the distortion of semantic feature structures and loss of class discriminability. In this paper, we introduce a novel prompt learning paradigm for UDA, named Domain Adaptation via Prompt Learning (DAPL). In contrast to prior works, our approach makes use of pre-trained vision-language models and optimizes only very few parameters. The main idea is to embed domain information into prompts, a form of representations generated from natural language, which is then used to perform classification. This domain information is shared only by images from the same domain, thereby dynamically adapting the classifier according to each domain. By adopting this paradigm, we show that our model not only outperforms previous methods on several cross-domain benchmarks but also is very efficient to train and easy to implement.

无监督领域适应 (UDA) 旨在将良好注释的源领域学习到的模型适应到目标领域, 在目标领域中仅提供未标记的样本。目前的 UDA 方法通过对齐源特征空间和目标特征空间来学习领域不变特征。这种对齐是通过统计差异最小化或对抗训练等约束来施加的。然而, 这些约束可能导致语义特征结构的扭曲和类别可区分性的丧失。在本文中, 我们提出了一种新的 UDA 提示学习范式, 称为通过提示学习进行领域适应 (DAPL)。与之前的工作相比, 我们的方法利用预训练的视觉-语言模型, 仅优化极少的参数。其主要思想是将领域信息嵌入到提示中, 提示是一种从自然语言生成的表示形式, 然后用于执行分类。该领域信息仅由来自同一领域的图像共享, 从而根据每个领域动态调整分类器。通过采用这一范式, 我们展示了我们的模型不仅在多个跨领域基准上优于以前的方法, 而且训练效率高且易于实现。

1. Introduction

1. 引言

Deep Learning has achieved great success in recent years [13, 17] with the help of large-scale annotated datasets [7]. Since annotating large-scale datasets is costly and time-consuming, researchers propose to train a model for an unlabeled domain by leveraging a related domain which is well-annotated. However, a model (e.g., a neural network) trained on an annotated domain may not generalize well to an unlabeled domain due to distribution shift [1, 2, 48]. The problem of Unsupervised Domain Adaptation (UDA) [10, 32, 37] has been proposed to study the transferring of knowledge under such domain shift.

深度学习在近年来取得了巨大的成功 [13, 17], 这得益于大规模标注数据集 [7]。由于标注大规模数据集既昂贵又耗时, 研究人员提出通过利用一个已标注的相关领域来训练一个无标注领域的模型。然而, 在标注领域训练的模型 (例如神经网络) 可能无法很好地推广到无标注领域, 因为存在分布偏移 [1, 2, 48]。无监督领域适应 (UDA) [10, 32, 37] 的问题被提出以研究在这种领域偏移下知识的转移。

Conventional UDA methods mainly resort to learning domain-invariant representations by aligning source and target domains. With similar features distribution led by domain alignment, the classifier trained on the source domain can be directly applied to the target data (Fig. 1, top). One typical

line of such methods is based on statistical discrepancy minimization [32, 34, 49, 55], Maximum Mean Discrepancy (MMD) [32] and Central Moment Discrepancy (CMD) [55]. Another typical line learns domain-invariant features via adversarial training by applying domain discriminators [11, 25, 33, 36]. Such methods confuse domain discriminators to reduce the difference between source and target domains in the feature space. However, reducing the discrepancy by aligning domains could lead to a loss of semantic information [47, 54]. Such loss comes from the entangled nature of semantic and domain information when the manifold structures of the data distributions are complex [3]. To remedy this, some recent UDA methods [4, 26, 47, 53] advocate preserving the semantic information to maintain the class discriminability. However, these methods suffer from a subtle trade-off between domain alignment and preserving semantic features [3, 45, 54] as two objectives could be adversarial. Learning disentangled semantic and domain representation could be an alternative since domain alignment could be discarded.

传统的 UDA 方法主要依赖于通过对齐源领域和目标领域来学习领域不变的表示。通过领域对齐导致的相似特征分布，训练于源领域的分类器可以直接应用于目标数据 (图 1, 顶部)。这类方法的一个典型方向是基于统计差异最小化 [32, 34, 49, 55]、最大均值差异 (MMD) [32] 和中心矩差异 (CMD) [55]。另一个典型方向是通过应用领域判别器 [11, 25, 33, 36] 进行对抗训练来学习领域不变特征。这类方法混淆领域判别器，以减少特征空间中源领域和目标领域之间的差异。然而，通过对齐领域来减少差异可能会导致语义信息的丧失 [47, 54]。这种损失源于当数据分布的流形结构复杂时，语义信息和领域信息的纠缠特性 [3]。为了解决这个问题，一些近期的 UDA 方法 [4, 26, 47, 53] 倡导保留语义信息以维持类别可区分性。然而，这些方法在领域对齐和保留语义特征之间存在微妙的权衡 [3, 45, 54]，因为这两个目标可能是对立的。学习解耦的语义和领域表示可能是一个替代方案，因为可以放弃领域对齐。

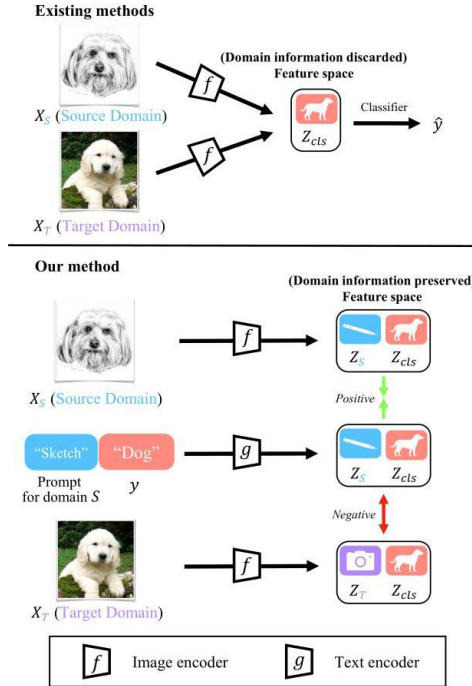


Figure 1. Overview of DAPL. We introduce the prompt tuning framework for domain adaptation. Top: conventional domain adaptation methods aim to remove domain-specific information via domain alignment or adversarial loss. This could lead to distorted feature representation when the manifold structures underlying the data distributions are complex [3]. Bottom: Our method preserves domain information and tunes a prompt for each domain. Our model learns with a contrastive objective.

图 1. DAPL 概述。我们引入了用于领域适应的提示调优框架。顶部：传统的领域适应方法旨在通过领域对齐或对抗损失来消除领域特定信息。当数据分布的流形结构复杂时，这可能导致特征表示失真 [3]。底部：我们的方法保留领域信息，并为每个领域调整一个提示。我们的模型通过对比目标进行学习。

To learn disentangled semantic and domain representation, we introduce the prompt learning method [16, 29, 31] to UDA, by learning a representation in a continuous label space. Fig. 2 illustrates our prompt design. The prompt consists of three parts: domain-agnostic context, domain-specific context, and class label (token). Each image corresponds to a ground truth class through the class label of prompt. For example, an image that shows "an art work of a dog" could correspond to the prompt "An image of a painting Dog". The domain-agnostic context represents general task information and is shared among all

images. The domain-specific context represents domain information and is shared in each domain. The class label distinguishes different categories.

为了学习解耦的语义和领域表示，我们将提示学习方法 [16, 29, 31] 引入到无监督领域适应 (UDA) 中，通过在连续标签空间中学习表示。图 2 说明了我们的提示设计。提示由三个部分组成：领域无关上下文、领域特定上下文和类别标签 (标记)。每个图像通过提示的类别标签对应于一个真实类别。例如，显示“狗的艺术作品”的图像可以对应于提示“狗的画作图像”。领域无关上下文表示一般任务信息，并在所有图像中共享。领域特定上下文表示领域信息，并在每个领域中共享。类别标签区分不同类别。

Such prompt learning method allows us to learn domain and category disentangled representation and avoids a loss of semantic information [47]. We apply a contrastive objective for training (Fig. 1, bottom). An image and a text form a pair of positive examples only when the domain and category of them are matched respectively, while any other cases are negative examples. By contrasting the representation of X_S and y , the image and text representation of the “sketch” and “dogs” are aligned in the feature space, respectively. Further, the text representation of “sketch” is pushed away from the “photo” domain by contrasting X_T and y . More details are discussed in Sec. 3.3. Hence, the representation of domain and category are aligned respectively. We adopt Contrastive Language Image Pretraining (CLIP) [42] as our backbone to facilitate prompt learning and contrastive learning.

这种提示学习方法使我们能够学习领域和类别解耦的表示，并避免语义信息的丢失 [47]。我们应用对比目标进行训练 (图 1, 底部)。只有当图像和文本的领域和类别分别匹配时，它们才形成一对正例，而其他情况则为负例。通过对比 X_S 和 y 的表示，“素描”和“狗”的图像和文本表示在特征空间中分别对齐。此外，通过对比 X_T 和 y ，“素描”的文本表示被推离“照片”领域。更多细节在第 3.3 节中讨论。因此，领域和类别的表示分别对齐。我们采用对比语言图像预训练 (CLIP) [42] 作为我们的骨干，以促进提示学习和对比学习。

Extensive experiments on two classic cross-domain benchmarks demonstrate that our method consistently yields promising performance, e.g., we achieve a sota performance of 74.5%/86.9% on Office-Home [51] and VisDA-2017 [39]. To summarize, the contributions of our work are three-fold:

在两个经典的跨领域基准上进行的大量实验表明，我们的方法始终产生令人满意的性能，例如，我们在 Office-Home [51] 和 VisDA-2017 [39] 上达到了最先进的性能 74.5%/86.9%。总而言之，我们工作的贡献有三方面：

- We propose Domain Adaptation via Prompt Learning (DAPL) for unsupervised domain adaptation. To the best of our knowledge, we are the first to apply prompt learning in unsupervised domain adaptation.
- 我们提出了通过提示学习进行领域适应 (DAPL) 的方法，用于无监督领域适应。据我们所知，我们是第一个在无监督领域适应中应用提示学习的研究。
- We propose to use domain-specific context in the prompt. Hence, we do not have to align domains at the cost of losing semantic information. Our method could learn continuous semantic representations for each category and domain.
- 我们建议在提示中使用特定领域的上下文。因此，我们不必以失去语义信息为代价来对齐领域。我们的方法可以为每个类别和领域学习连续的语义表示。
- The proposed DAPL has achieved state-of-the-art performance on Office-Home and VisDA-2017 dataset, improving the accuracy by 2.5%/2.5% over the strong baseline CLIP.
- 提出的 DAPL 在 Office-Home 和 VisDA-2017 数据集上达到了最先进的性能，相比强基线 CLIP 提高了 2.5%/2.5% 的准确性。

2. Related Work

2. 相关工作

Unsupervised Domain Adaptation. Unsupervised Domain Adaptation (UDA) adapts a model trained on a labeled source domain to an unlabeled target domain. Quite a few UDA methods learn domain-invariant features via minimizing the discrepancy between domains [32,34,46]. For example, Tzeng et al. [49] introduce an adaptation layer and a domain confusion loss to learn semantically meaningful and domain-invariant representations. DAN [32] aligns source and target domains by minimizing the

maximum mean discrepancy (MMD) on task-specific layers. Sun et al. [46] propose CORAL that aligns the second-order statistics of the source and target domain with a linear projection.

无监督领域适应。无监督领域适应 (UDA) 将一个在标记源领域上训练的模型适应到一个未标记的目标领域。相当多的 UDA 方法通过最小化领域之间的差异来学习领域不变特征 [32,34,46]。例如, Tzeng 等人 [49] 引入了一个适应层和一个领域混淆损失, 以学习语义上有意义且领域不变的表示。DAN [32] 通过最小化任务特定层上的最大均值差异 (MMD) 来对齐源领域和目标领域。Sun 等人 [46] 提出了 CORAL, 通过线性投影对齐源领域和目标领域的二阶统计量。

Inspired by generative adversarial networks (GANs) [11], another family of UDA methods apply adversarial learning to obtain domain-invariant representations [10, 25, 33]. For example, DANN [10] and CDAN [33] introduce a domain discriminator to distinguish source samples from target ones, while the feature extractor tries to generate domain-invariant features in order to fool the domain discriminator. Differently, MCD [43] plays the minimax game between a feature encoder and two classifiers, where two classifiers try to maximize their prediction discrepancy and the feature extractor aims to minimize that discrepancy.

受到生成对抗网络 (GANs) [11] 的启发, 另一类 UDA 方法应用对抗学习来获得领域不变的表示 [10, 25, 33]。例如, DANN [10] 和 CDAN [33] 引入了一个领域鉴别器, 以区分源样本和目标样本, 而特征提取器则试图生成领域不变的特征, 以欺骗领域鉴别器。不同的是, MCD [43] 在特征编码器和两个分类器之间进行极大极小博弈, 其中两个分类器试图最大化它们的预测差异, 而特征提取器则旨在最小化该差异。

Despite the success achieved by domain alignment, class discrimination also loses due to the distorted structure of semantic features [3, 47]. How to maintain class discriminability has also been considered by recent UDA works [5, 22, 24, 38, 47, 54]. To name a few, Li et al. [22] build attention-aware transport distance to learn discriminant features, along with an entropy-based regularization. Cui et al. [5] propose to enforce the prediction discriminability and diversity via batch nuclear-norm maximization (BNM). However, these methods have to make trade-offs between aligning domains and preserving class discriminability.

尽管领域对齐取得了成功, 但由于语义特征的扭曲结构, 类别区分能力也有所下降 [3, 47]。如何保持类别可区分性也被近期的 UDA 研究所考虑 [5, 22, 24, 38, 47, 54]。举几个例子, Li 等人 [22] 构建了注意力感知的传输距离来学习判别特征, 并结合基于熵的正则化。Cui 等人 [5] 提出了通过批量核范数最大化 (BNM) 来增强预测的可区分性和多样性。然而, 这些方法必须在对齐领域和保持类别可区分性之间进行权衡。


Domains	Prompt		
	Domain-agnostic	Domain-specific	Class label
Art	<div>  </div>	<div> "Painting" "Creation" ... </div>	<div> Dog Cat Cup ... </div>
Clipart		<div> "Icon" "Illustration" ... </div>	
Photo		<div> "Photo" "Real world" ... </div>	
Product		<div> "Product" "Manufactured" ... </div>	

Figure 2. Example prompt structure. Our proposed prompt consists of three parts: (a) Domain-specific prompt; (b) Domain-agnostic prompt; (c) Class label. The first two parts are continuous and learned from data. The words shown here are for illustrative purposes.

图 2. 示例提示结构。我们提出的提示由三个部分组成:(a) 特定领域的提示; (b) 与领域无关的提示; (c) 类别标签。前两部分是连续的, 并从数据中学习而来。这里展示的词语仅用于说明。

Compared with these methods, our method applies prompt learning to learn domain-specific visual concepts (i.e., the transparent background for "product" domain) for each domain.

与这些方法相比, 我们的方法应用提示学习来学习每个领域的特定领域视觉概念 (即 "产品" 领域的透明背景)。

Prompt Learning. Prompt learning, which is first introduced by Petroni et al. [40], has been widely studied in NLP during these years [19, 21, 27, 30, 40, 44]. Prompting means prepending instructions to the input and pre-training the language model so that the downstream tasks can be promoted. Petroni et al. [40] and Pörner et al. [41] use manually defined prompts to improve the performance of language models. However, manually created prompts may be sub-optimal or even inappropriate, which might fail to provide accurate instruction. To obtain more accurate estimation of the knowledge contained in language models, several methods have been proposed to automatically explore optimal prompts [19, 44, 57]. More recently, prompts have been integrated into vision-language models to learn generic visual representations [18, 42, 58]. Among them, ALIGN [18] and CLIP [42] are most pioneering ones. CLIP [42] learns state-of-the-art visual representations from natural language supervision by pre-training a vision language model on 400 million image-text pairs. Furthermore, Zhou et al. [58] use continuous representations to model prompts so that the task-relevant prompts can be automatically learned, namely CoOp. However, CoOp only develops a domain-agnostic prompt for visual recognition tasks while our work proposes to learn both domain-agnostic and domain-specific prompts to deal with distribution shift in UDA.

提示学习. 提示学习最早由 Petroni 等人提出 [40], 近年来在自然语言处理 (NLP) 领域得到了广泛研究 [19, 21, 27, 30, 40, 44]. 提示意味着在输入前添加指令, 并对语言模型进行预训练, 以便促进下游任务的执行. Petroni 等人 [40] 和 Pörner 等人 [41] 使用手动定义的提示来提高语言模型的性能. 然而, 手动创建的提示可能是次优的, 甚至不合适, 可能无法提供准确的指令. 为了更准确地估计语言模型中包含的知识, 提出了几种方法来自动探索最佳提示 [19, 44, 57]. 最近, 提示已被整合到视觉-语言模型中, 以学习通用视觉表示 [18, 42, 58]. 其中, ALIGN [18] 和 CLIP [42] 是最具开创性的. CLIP [42] 通过在 4 亿对图像-文本数据上对视觉语言模型进行预训练, 从自然语言监督中学习最先进的视觉表示. 此外, Zhou 等人 [58] 使用连续表示来建模提示, 以便任务相关的提示能够被自动学习, 即 CoOp. 然而, CoOp 仅为视觉识别任务开发了一个与领域无关的提示, 而我们的工作提出学习领域无关和特定领域的提示, 以应对 UDA 中的分布转移.

3. Method

3. 方法

Given a set of labeled source images $\mathcal{D}_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{N_s}$ and a set of unlabeled target images $\mathcal{D}_u = \{(\mathbf{x}_i^u)\}_{i=1}^{N_u}$, we adopt a model trained from a source domain to a target domain. Here, N_s and N_u denote the scale of source domain dataset \mathcal{D}_s and target domain dataset \mathcal{D}_u respectively. These two domains share the same K categories.

给定一组标记的源图像 $\mathcal{D}_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{N_s}$ 和一组未标记的目标图像 $\mathcal{D}_u = \{(\mathbf{x}_i^u)\}_{i=1}^{N_u}$, 我们采用从源领域训练的模型到目标领域进行迁移. 在这里, N_s 和 N_u 分别表示源领域数据集 \mathcal{D}_s 和目标领域数据集 \mathcal{D}_u 的规模. 这两个领域共享相同的 K 类别.

3.1. Preliminaries

3.1. 基础知识

We adopt CLIP [42] as our backbone. Our model is comprised of an image encoder $f(\cdot)$ and a text encoder $g(\cdot)$. The image encoder can be a ResNet [13] or Vision Transformer (ViT) [8], and the text encoder is a Transformer [50]. The image and text input can be directly transformed from high dimensional space into a low dimensional feature space by the encoders.

我们采用 CLIP [42] 作为我们的主干网络. 我们的模型由一个图像编码器 $f(\cdot)$ 和一个文本编码器 $g(\cdot)$ 组成. 图像编码器可以是 ResNet [13] 或视觉变换器 (ViT) [8], 而文本编码器是一个变换器 [50]. 图像和文本输入可以通过编码器直接从高维空间转换为低维特征空间.

CLIP [42] is trained with image-text pairs in a contrastive manner. Each input text describes a category in the format of "a photo of a [CLASS]" ([CLASS] is the class token). A positive pair is an image \mathbf{x}_i with its corresponding text \mathbf{t}_i describing the category of \mathbf{x}_i . A negative pair is an image \mathbf{x}_i with an irrelevant description $\mathbf{t}_j, j \neq i$ in the mini-batch. The training objective is to maximize the cosine similarity of positive pairs and minimize the cosine similarity of negative pairs. The contrastive learning objective aligns the image and text representation in the same feature space.

CLIP [42] 以对比方式使用图像-文本对进行训练。每个输入文本以“一个 [CLASS] 的照片”的格式描述一个类别 ([CLASS] 是类标记)。正对是指一张图像 \mathbf{x}_i 及其对应的文本 \mathbf{t}_i ，描述 \mathbf{x}_i 的类别。负对是指一张图像 \mathbf{x}_i 及其在小批量中的无关描述 $\mathbf{t}_j, j \neq i$ 。训练目标是最大化正对的余弦相似度，并最小化负对的余弦相似度。对比学习目标将图像和文本表示对齐到同一特征空间。

With the aligned features, the model is capable of performing zero-shot inference. By forwarding K category descriptions, an image \mathbf{x} would belong to the category \hat{y}_i with the largest similarity:

通过对齐的特征，模型能够执行零-shot 推理。通过前向传递 K 类别描述，一张图像 \mathbf{x} 将属于与其相似度最大的类别 \hat{y}_i ：

$$P(\hat{y} = i | \mathbf{x}) = \frac{\exp(\langle g(\mathbf{t}_i), f(\mathbf{x}) \rangle / T)}{\sum_{k=1}^K \exp(\langle g(\mathbf{t}_k), f(\mathbf{x}) \rangle / T)}, \quad (1)$$

$$\hat{y}_i = \arg \max_k P(\hat{y}_i = k) \quad (2)$$

where T is a user-defined hyper-parameter (temperature) and $\langle \cdot, \cdot \rangle$ denotes the cosine similarity. 其中 T 是用户定义的超参数 (温度)，而 $\langle \cdot, \cdot \rangle$ 表示余弦相似度。

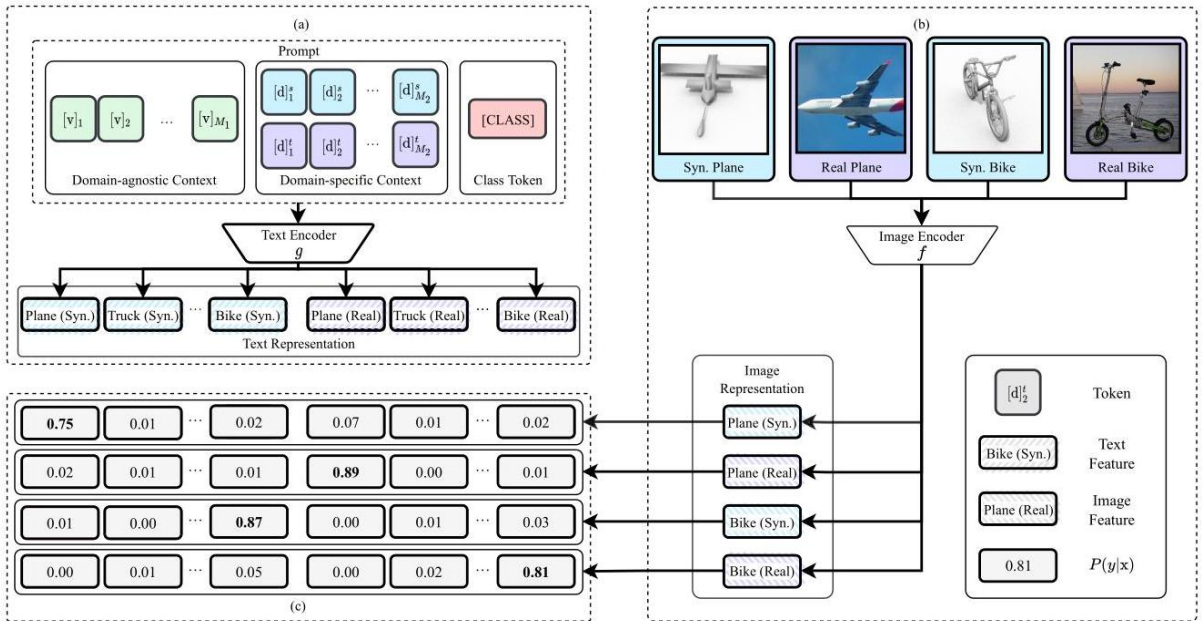


Figure 3. Domain Adaptation via Prompt Learning (DAPL): (a) DAPL trains the learnable context variables: domain-agnostic context variables and domain-specific context variables, and [CLASS] token which are combined and encoded by a text encoder. (b) An image encoder encodes images from different domains. (c) Next, cosine similarity between text and image features is computed and the positive pairs (with matched domain and class) are encouraged to align. The classification probability are defined in Eq. (6) and a cross-entropy loss is applied between the image feature and the ground truth class to train the networks.

图 3. 通过提示学习进行领域适应 (DAPL): (a) DAPL 训练可学习的上下文变量: 领域无关的上下文变量和领域特定的上下文变量, 以及 [CLASS] 令牌, 这些变量通过文本编码器进行组合和编码。 (b) 图像编码器对来自不同领域的图像进行编码。 (c) 接下来, 计算文本和图像特征之间的余弦相似度, 并鼓励正对 (具有匹配的领域和类别) 对齐。分类概率在公式 (6) 中定义, 并在图像特征与真实类别之间应用交叉熵损失以训练网络。

The input text described above is a manually designed prompt comprised of a sequence of discrete tokens. The manually designed prompts are transformed into fixed vectors in the word embedding space. Since these vectors could be sub-optimal for the representation of categories, we could optimize the continuous embedding of the tokens. The continuous representation \mathbf{t}_k allows for a more precise description of semantic features which are important to the context variable learning.

上述输入文本是一个手动设计的提示, 由一系列离散令牌组成。手动设计的提示被转换为词嵌入空间中的固定向量。由于这些向量可能对类别的表示不是最优的, 我们可以优化令牌的连续嵌入。连续表示 \mathbf{t}_k 允许对上下文变量学习重要的语义特征进行更精确的描述。

Existing prompt learning methods adopt a domain-agnostic style that context is shared across all domains and all categories. It follows a unified style:

现有的提示学习方法采用领域无关的风格，即上下文在所有领域和所有类别之间共享。它遵循统一的风格：

$$\mathbf{t}_k = [\mathbf{v}]_1 [\mathbf{v}]_2 \dots [\mathbf{v}]_{M_1} [\text{CLASS}]_k, \quad (3)$$

where $[\mathbf{v}]_{m_1}, m_1 \in \{1, 2, \dots, M_1\}$ is a vector with the same dimension as the word embedding, and M_1 is the number of context tokens applied in the prompt.

其中 $[\mathbf{v}]_{m_1}, m_1 \in \{1, 2, \dots, M_1\}$ 是与词嵌入相同维度的向量，而 M_1 是在提示中应用的上下文令牌的数量。

3.2. Domain Adaptation via Prompt Learning

3.2. 通过提示学习进行领域适应

Since the domain-agnostic context alone cannot deal with the distribution shift between domains, we propose to use Domain-Specific Context (DSC) to capture unique features of each domain. To be specific, our proposed prompt contains two counterparts, a domain-agnostic context and a domain-specific context. We use $[\mathbf{d}]_{m_2}^d, m_2 \in \{1, 2, \dots, M_2\}$ to denote domain-specific tokens, which have the same dimension as word embeddings. The domain-specific context is shared among all categories but specially designed for each domain $[\mathbf{d}]_i^s \neq [\mathbf{d}]_j^u, i, j \in \{1, 2, \dots, M_2\}$. The number of domain-specific tokens is denoted by M_2 . Domain indicator denotes the source and target domains $d \in \{s, u\}$. The overall prompt is defined in the following format:

由于领域无关的上下文无法处理领域之间的分布变化，我们建议使用领域特定上下文 (DSC) 来捕捉每个领域的独特特征。具体来说，我们提出的提示包含两个部分，一个领域无关的上下文和一个领域特定的上下文。我们用 $[\mathbf{d}]_{m_2}^d, m_2 \in \{1, 2, \dots, M_2\}$ 来表示领域特定的标记，它们与词嵌入具有相同的维度。领域特定上下文在所有类别之间共享，但为每个领域特别设计 $[\mathbf{d}]_i^s \neq [\mathbf{d}]_j^u, i, j \in \{1, 2, \dots, M_2\}$ 。领域特定标记的数量用 M_2 表示。领域指示符表示源领域和目标领域 $d \in \{s, u\}$ 。整体提示的定义格式如下：

$$\mathbf{t}_k^d = [\mathbf{v}]_1 [\mathbf{v}]_2 \dots [\mathbf{v}]_{M_1} [\mathbf{d}]_1^d [\mathbf{d}]_2^d \dots [\mathbf{d}]_{M_2}^d [\text{CLASS}]_k. \quad (4)$$

When [CLASS] token in the text feature space could not fully model the difference among each class, the domain-agnostic context could follow a class-specific style [42] denoted by class-specific context. Each class could be initialized with different tokens:

当文本特征空间中的 [CLASS] 标记无法充分建模每个类别之间的差异时，领域无关的上下文可以遵循由类别特定上下文表示的类别特定风格 [42]。每个类别可以用不同的标记进行初始化：

$$\mathbf{t}_k^d = [\mathbf{v}]_1^k [\mathbf{v}]_2^k \dots [\mathbf{v}]_{M_1}^k [\mathbf{d}]_1^d [\mathbf{d}]_2^d \dots [\mathbf{d}]_{M_2}^d [\text{CLASS}]_k. \quad (5)$$

The trainable class-specific context could learn a more fine-grained representation than only [CLASS] token [58]. Our main results are based on class-specific context and domain-specific context as Eq. (5).

可训练的类别特定上下文能够学习比仅使用 [CLASS] 标记更细粒度的表示 [58]。我们的主要结果基于类别特定上下文和领域特定上下文，如公式 (5) 所示。

We have $2K$ categories since we apply different prompts $\mathbf{t}_k^s, \mathbf{t}_k^u$ for the source and the target domain respectively. Given a set of training samples $\{\mathbf{x}_i^s, y_i^s\}_{i=1}^{N_s}$ of the source domain, we could obtain the probability that a training sam-

我们有 $2K$ 个类别，因为我们分别为源领域和目标领域应用不同的提示 $\mathbf{t}_k^s, \mathbf{t}_k^u$ 。给定一组源领域的训练样本 $\{\mathbf{x}_i^s, y_i^s\}_{i=1}^{N_s}$ ，我们可以获得一个训练样本的概率，

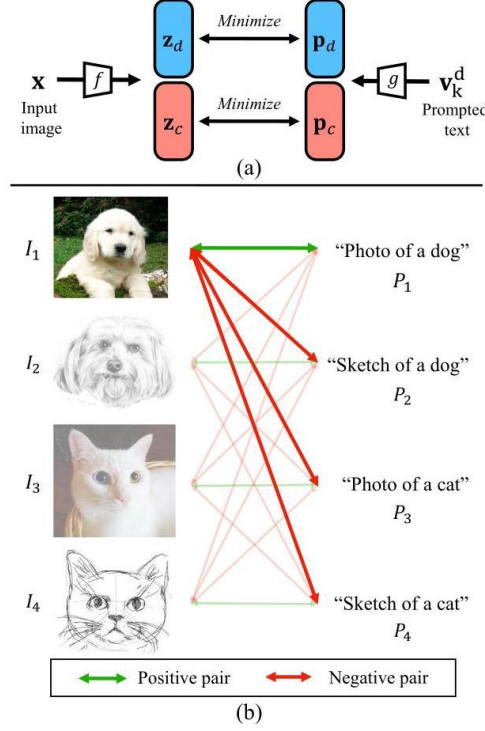


Figure 4. Contrastive learning helps transfer learning. (a) We assume that visual representation implicitly contains two parts: domain information (\mathbf{z}_d) and class information (\mathbf{z}_c). Similarly, the language feature contains two parts: domain information (\mathbf{p}_d) and class information (\mathbf{p}_c). By minimizing the distance between positive pairs (shown in green) and maximizing the distance between negative pairs (shown in red), we show that the domain information and class information can be disentangled. Such disentangled representations can be applied for transfer learning. See Sec. 3.3 for details.

图 4. 对比学习有助于迁移学习。(a) 我们假设视觉表示隐含地包含两个部分: 领域信息 (\mathbf{z}_d) 和类别信息 (\mathbf{z}_c)。类似地, 语言特征也包含两个部分: 领域信息 (\mathbf{p}_d) 和类别信息 (\mathbf{p}_c)。通过最小化正样本对 (以绿色显示) 之间的距离, 并最大化负样本对 (以红色显示) 之间的距离, 我们表明领域信息和类别信息可以被解耦。这种解耦的表示可以应用于迁移学习。有关详细信息, 请参见第 3.3 节。

ple belongs to the k -th category:

样本属于 k 类别:

$$P(\hat{y}_i^s = k | \mathbf{x}_i^s) = \frac{\exp(\langle g(\mathbf{t}_k^s), f(\mathbf{x}_i^s) \rangle / T)}{\sum_{d \in \{s, u\}} \sum_{j=1}^K \exp(\langle g(\mathbf{t}_j^d), f(\mathbf{x}_i^s) \rangle / T)}.$$

(6)

With the probability of the image \mathbf{x}_i belonging to class k , we minimize the standard cross-entropy loss given ground truth label y_i^s . The loss is computed as follow:

通过最小化给定真实标签 y_i^s 的标准交叉熵损失, 我们计算图像 \mathbf{x}_i 属于类别 k 的概率。损失计算如下:

$$\mathcal{L}_s = -\frac{1}{N_s} \sum_{i=1}^{N_s} \log P(\hat{y}_i^s = y_i^s). \quad (7)$$

To further exploit the unlabeled data, We generate pseudo labels on the target domain. We choose from K classes with maximum predicted probability as the pseudo label y^u of the training data \mathbf{x}^u :

为了进一步利用未标记的数据, 我们在目标领域生成伪标签。我们从 K 类别中选择最大预测概率的类别作为训练数据 \mathbf{x}^u 的伪标签 y^u :

$$y^u = \arg \max_k P(\hat{y}^u = k | \mathbf{x}^u), k = \{1, 2, \dots, K\}. \quad (8)$$

We only generate pseudo labels for unlabeled data whose maximum prediction probability is larger than a fixed threshold τ for the quality of pseudo labels. We make use of the zero-shot inference ability

of CLIP to generate pseudo labels as described in Sec. 3.1. We train the prompt of target domain \mathbf{t}_k^u with these unlabeled images and their pseudo labels with the contrastive objective Eq. (6):

我们仅为最大预测概率大于固定阈值 τ 的未标记数据生成伪标签，以确保伪标签的质量。我们利用 CLIP 的 zero-shot 推理能力生成伪标签，如第 3.1 节所述。我们使用这些未标记图像及其伪标签，以对比目标领域的提示 \mathbf{t}_k^u ，并使用对比目标 Eq. (6) 进行训练：

$$\mathcal{L}_u = -\frac{1}{N_u} \sum_{i=1}^{N_u} \mathbb{I}\{P(\hat{y}_i^u = y_i^u | \mathbf{x}_i^u) \geq \tau\} \log P(\hat{y}_i^u = y_i^u | \mathbf{x}_i^u) \quad (9)$$

where $\mathbb{I}\{\cdot\}$ is an indicator function. Overall, our proposed Domain Adaptation via Prompt Learning (DAPL) method could be trained in an end-to-end manner with a total contrastive loss:

其中 $\mathbb{I}\{\cdot\}$ 是一个指示函数。总体而言，我们提出的通过提示学习的领域适应方法 (DAPL) 可以以端到端的方式进行训练，具有总对比损失：

$$\mathcal{L} = \mathcal{L}_s(\mathcal{D}^s) + \mathcal{L}_u(\mathcal{D}^u) \quad (10)$$

Existing domain adaptation methods train their classifier on the source domain to learn a conditional probability distribution $P(y | \mathbf{x}^s)$. By aligning the marginal distribution of $P(f(\mathbf{x}^s))$ and $P(f(\mathbf{x}^u))$ they could directly make use of the conditional probability for inference on the target domain. When the conditional probability distribution varies $P(y | \mathbf{x}^s) \neq P(y | \mathbf{x}^u)$, these methods could suffer the risk of performance drop [52]. Our method does not align marginal distributions but learns two conditional probability distributions $P(y | \mathbf{x}^s)$ and $P(y | \mathbf{x}^u)$ by learning two sets of prompts $\mathbf{t}_k^s, \mathbf{t}_k^u, k \in \{1, 2, \dots, K\}$. Hence, our method could deal with both conditional distribution shift and marginal distribution shift. The overview of DAPL is shown in Fig. 3.

现有的领域适应方法在源域上训练其分类器，以学习条件概率分布 $P(y | \mathbf{x}^s)$ 。通过对齐 $P(f(\mathbf{x}^s))$ 和 $P(f(\mathbf{x}^u))$ 的边际分布，它们可以直接利用条件概率在目标领域进行推断。当条件概率分布发生变化 $P(y | \mathbf{x}^s) \neq P(y | \mathbf{x}^u)$ 时，这些方法可能面临性能下降的风险 [52]。我们的方法并不对齐边际分布，而是通过学习两组提示 $\mathbf{t}_k^s, \mathbf{t}_k^u, k \in \{1, 2, \dots, K\}$ 来学习两个条件概率分布 $P(y | \mathbf{x}^s)$ 和 $P(y | \mathbf{x}^u)$ 。因此，我们的方法可以处理条件分布转移和边际分布转移。DAPL 的概述如图 3 所示。

3.3. Disentanglement by Contrastive Learning

3.3. 通过对比学习进行解耦

We adopt a contrastive loss \mathcal{L} as the optimization objective. Here, we provide an intuitive explanation for why this objective achieves the desired goal: the visual encoder and text encoder each encodes the input into two disentangled latent representations, separating domain information from the intrinsic class information. Only when both the class and the domain information are aligned, the distance between the textual feature and the image feature is minimized. By minimizing the distance between such positive pairs (maximizing the similarity), the probability of the correct label is maximized (see Eq. (6)).

我们采用对比损失 \mathcal{L} 作为优化目标。在这里，我们提供一个直观的解释，说明为什么这个目标能够实现预期的目标：视觉编码器和文本编码器各自将输入编码为两个解耦的潜在表示，将领域信息与内在类别信息分离。只有当类别信息和领域信息都对齐时，文本特征与图像特征之间的距离才会最小化。通过最小化这种正样本对之间的距离（最大化相似性），正确标签的概率被最大化（见公式 (6)）。

First, we assume that the visual representation $f(\mathbf{x}_i^d)$ contains two parts: domain information of domain d and the intrinsic class information of class c (Fig. 4 (a), \mathbf{z}_d and \mathbf{z}_c). Similarly, the language embedding $g(\mathbf{t}_k^d)$ contains the same two parts: domain information of domain d and the class information of class c (Fig. 4 (a), \mathbf{p}_d and \mathbf{p}_c). Next, we show that such domain information and class information can be disentangled by optimizing the contrastive objective.

首先，我们假设视觉表示 $f(\mathbf{x}_i^d)$ 包含两个部分：领域 d 的领域信息和类别 c 的内在类别信息（图 4(a), \mathbf{z}_d 和 \mathbf{z}_c ）。类似地，语言嵌入 $g(\mathbf{t}_k^d)$ 也包含相同的两个部分：领域 d 的领域信息和类别 c 的类别信息（图 4(a), \mathbf{p}_d 和 \mathbf{p}_c ）。接下来，我们展示了通过优化对比目标可以解耦这些领域信息和类别信息。

Figure 4 (b) provides an illustrative example. In this example, there are four image-text pairs with two classes (cat, dog) and two domains (photo, sketch). Take the image I_1 , prompts P_1 and P_2 as an example. The image can form a positive pair with prompt P_1 and a negative pair with prompt P_2 . By optimizing the contrastive objective, the distance between image feature $f(I_1)$ and the sentence embedding of $g(P_1)$ is minimized, whereas the distance between image feature $f(I_1)$ and the sentence embedding of $g(P_2)$ is maximized. We claim that this forces the class information of dog disentangled

from the domain representation of photo or sketch. Suppose on the contrary that the domain information and the class information are still entangled in the representation, i.e. the domain representation (\mathbf{p}_d^1 and \mathbf{p}_d^2) contains the class information of dog. In this case, I_1 and P_2 still matches and the distance between $f(I_1)$ and $g(P_2)$ could be further maximized by removing this class information. In other words, we reduce class information in domain representation by optimizing the contrastive loss. Similarly, taking (I_1, P_3) as negative pair, we remove domain information from class representation - otherwise $f(I_1)$ still matches $g(P_3)$ because of the entangled domain information of photo in class representation. Combining these two negative pairs, the domain representation and the intrinsic class information can be forced to disentangle with each other by minimizing the contrastive objective.

图 4 (b) 提供了一个说明性的例子。在这个例子中，有四个图像-文本对，包含两个类别（猫，狗）和两个领域（照片，素描）。以图像 I_1 、提示 P_1 和 P_2 为例。该图像可以与提示 P_1 形成正对，与提示 P_2 形成负对。通过优化对比目标，图像特征 $f(I_1)$ 和 $g(P_1)$ 的句子嵌入之间的距离被最小化，而图像特征 $f(I_1)$ 和 $g(P_2)$ 的句子嵌入之间的距离被最大化。我们声称这迫使狗的类别信息与照片或素描的领域表示解耦。假设相反，领域信息和类别信息仍然在表示中纠缠，即领域表示 (\mathbf{p}_d^1 和 \mathbf{p}_d^2) 包含狗的类别信息。在这种情况下， I_1 和 P_2 仍然匹配，并且通过去除该类别信息，可以进一步最大化 $f(I_1)$ 和 $g(P_2)$ 之间的距离。换句话说，我们通过优化对比损失来减少领域表示中的类别信息。类似地，以 (I_1, P_3) 作为负对，我们从类别表示中去除领域信息——否则由于类别表示中照片的纠缠领域信息， $f(I_1)$ 仍然与 $g(P_3)$ 匹配。结合这两个负对，通过最小化对比目标，可以迫使领域表示和内在类别信息相互解耦。

4. Experimental Results

4. 实验结果

We conduct extensive experiments on UDA benchmarks to verify the validity of our proposed method. We next present the datasets used in our experiments, comparisons with baseline methods, ablation studies of our method and visualization of results.

我们在 UDA 基准上进行了广泛的实验，以验证我们提出的方法的有效性。接下来，我们将介绍实验中使用的数据集、与基线方法的比较、我们方法的消融研究以及结果的可视化。

4.1. Datasets and Experimental Settings

4.1. 数据集和实验设置

Office-Home [51] is a large-scale benchmark for visual cross-domain recognition. It collects a total of 15,500 images from four distinct domains: Art(Ar), Clip Art(Cl), Product(Pr), and Real World(Rw). Besides, each domain contains the objects of 65 categories in the office and home environments. To evaluate our method, we conduct 12 UDA tasks, i.e., $\text{Ar} \rightarrow \text{Cl}, \dots, \text{Rw} \rightarrow \text{Pr}$.

Office-Home [51] 是一个用于视觉跨域识别的大规模基准。它从四个不同的领域收集了总共 15,500 张图像：艺术 (Ar)、剪贴画 (Cl)、产品 (Pr) 和现实世界 (Rw)。此外，每个领域包含办公室和家庭环境中 65 个类别的对象。为了评估我们的方法，我们进行 12 个 UDA 任务，即 $\text{Ar} \rightarrow \text{Cl}, \dots, \text{Rw} \rightarrow \text{Pr}$ 。

VisDA-2017 [39] is a more challenging dataset for synthetic-to-real domain adaptation with 12 categories. It contains 152,397 synthetic images, generated by rendering the 3D models with different angles and light conditions, and 55,388 real-world images, collected from MSCOCO [28]. Following [33] and [43], we use the synthetic images as source domain and real-world images as target domain.

VisDA-2017 [39] 是一个更具挑战性的数据集，用于合成到真实域的适应，共有 12 个类别。它包含 152,397 张合成图像，这些图像是通过以不同角度和光照条件渲染 3D 模型生成的，以及 55,388 张从 MSCOCO [28] 收集的真实世界图像。遵循 [33] 和 [43] 的方法，我们将合成图像作为源域，将真实世界图像作为目标域。

Implementation details. For Office-Home, we use pre-trained CLIP model and adopt ResNet-50 [14] as its image encoder. We fix the parameters in the encoders and the prompt is trained with the mini-batch SGD optimizer for 200 epochs, where the batch size is set to be 32. The initial learning rate is set to 0.003 and decayed with a cosine annealing rule [35]. For VisDA-2017 [39], the results are obtained by leveraging the pre-trained CLIP model with ResNet-101 [14] as the image encoder. The parameters of the image and text encoders are fixed and we train the prompt for 25 epochs using the mini-batch SGD optimizer with a batch of 32. The learning rate is set to 0.003 initially and decayed with a cosine annealing rule. As for the hyper-parameters, the length of context tokens M_1 and domain-specific tokens

M_2 are both set to 16 . Other choices of token numbers are discussed in Sec. 4.3. Our context vectors are randomly initialized using a zero-mean Gaussian distribution with a standard deviation of 0.02 . The pseudo labeling threshold τ is set to 0.6 for Office-Home and 0.5 for VisDA-2017 [39]. Further discussion about the value of τ is shown in Sec. 4.3.

实施细节。对于 Office-Home, 我们使用预训练的 CLIP 模型, 并采用 ResNet-50 [14] 作为其图像编码器。我们固定编码器中的参数, 并使用小批量 SGD 优化器对提示进行训练, 训练周期为 200 轮, 其中批量大小设置为 32。初始学习率设置为 0.003, 并采用余弦退火规则进行衰减 [35]。对于 VisDA-2017 [39], 结果是通过利用预训练的 CLIP 模型与 ResNet-101 [14] 作为图像编码器获得的。图像和文本编码器的参数固定, 我们使用小批量 SGD 优化器训练提示 25 轮, 批量大小为 32。学习率初始设置为 0.003, 并采用余弦退火规则进行衰减。至于超参数, 上下文令牌的长度 M_1 和特定领域令牌的长度 M_2 均设置为 16。其他令牌数量的选择在第 4.3 节中讨论。我们的上下文向量使用均值为零、标准差为 0.02 的高斯分布随机初始化。伪标签阈值 τ 在 Office-Home 中设置为 0.6, 在 VisDA-2017 [39] 中设置为 0.5。关于 τ 值的进一步讨论见第 4.3 节。

4.2. Comparison with State-of-the-Art DA Methods

4.2. 与最先进的领域适应方法的比较

4.2.1 Quantitative Evaluation

4.2.1 定量评估

Results on Office-Home are shown in Tab. 1, where our method obviously outperforms all other baselines w.r.t the average accuracy of 12 tasks. Note that there exists a large performance gap between the feature alignment-based methods (e.g., DANN [10] and CDAN+E [33]) and SRDC [47]. The possible reason may be that excessive feature alignment would hamper the discrimination of target data. While such potential risk will not happen in our method, since we do not force feature alignment across domains. Particularly, our method further surpasses the state-of-the-art method SRDC [47]) by a large margin of 3.2% in terms of the average accuracy. We owe the performance improvement to the more suitable visual concepts for the target domain that are generated from our learned prompts. And the superior performance of our method shows that simple prompt learning is effective for UDA problems.

在 Office-Home 上的结果如表 1 所示, 我们的方法在 12 个任务的平均准确率方面明显优于所有其他基线。值得注意的是, 基于特征对齐的方法 (例如, DANN [10] 和 CDAN+E [33]) 与 SRDC [47] 之间存在较大的性能差距。可能的原因是过度的特征对齐会妨碍目标数据的区分性。而在我们的方法中, 这种潜在风险不会发生, 因为我们并不强制在不同领域之间进行特征对齐。特别是, 我们的方法在平均准确率方面大幅超越了最先进的方法 SRDC [47], 差距为 3.2%。我们将性能提升归功于从我们学习的提示中生成的更适合目标领域的视觉概念。我们方法的优越性能表明, 简单的提示学习对于无监督领域适应 (UDA) 问题是有效的。

Results on VisDA-2017 [39] are presented in Tab. 2. It can be observed that our method achieves the highest average accuracy of 86.9% over the 12 classes, outperforming the state-of-the-art method STAR [36] by a large margin of 4.2% . Note that CLIP in Tab. 2 means zero-shot CLIP which adopts "a photo of a [CLASS]" as the hand-crafted prompt. Even the hand-crafted prompt method already has an impressive performance, our DAPL still achieves a 2.5% absolute improvement over it. The reason why the accuracy of truck is significantly boosted may be that the concept of "truck" is more discriminative in the language model. Furthermore, with the help of prompt learning, DAPL outperforms CLIP by 7.5%, 14%, 9.2% on "knife", "person" and "plant". In general, despite the simplicity of ours method, the encouraging results validate the efficacy of our prompt learning method.

在 VisDA-2017 [39] 上的结果如表 2 所示。可以观察到, 我们的方法在 12 个类别上达到了最高的平均准确率 86.9%, 大幅超越了最先进的方法 STAR [36], 差距为 4.2%。值得注意的是, 表 2 中的 CLIP 指的是零样本 CLIP, 它采用 "一个 [CLASS] 的照片" 作为手工制作的提示。即使手工制作的提示方法已经表现出令人印象深刻的性能, 我们的 DAPL 仍然在其基础上实现了 2.5% 的绝对提升。卡车的准确率显著提高的原因可能是 "卡车" 这一概念在语言模型中更具区分性。此外, 在提示学习的帮助下, DAPL 在 "刀具"、"人" 和 "植物" 上超越了 CLIP, 提升幅度为 7.5%, 14%, 9.2%。总的来说, 尽管我们的方法相对简单, 但令人鼓舞的结果验证了我们提示学习方法的有效性。

4.2.2 Training Time Analysis

4.2.2 训练时间分析

We train all the models with 1 NVIDIA RTX 2080 Ti GPU. Our method is much more efficient than other methods. For example, DAPL, MCD [43] and DANN [10] take 5.3h, 13.4h, 38.3h to train on VisDA-2017, respectively. Because we only fine-tune the prompt with very few parameters, it is much easier and faster to optimize the model.

我们使用 1 个 NVIDIA RTX 2080 Ti GPU 训练所有模型。我们的方法比其他方法高效得多。例如，DAPL、MCD [43] 和 DANN [10] 在 VisDA-2017 上的训练时间分别为 5.3 小时、13.4 小时和 38.3 小时。由于我们仅对非常少量的参数进行微调，因此优化模型要容易得多和快得多。

Table 1. Accuracy (%) on Office-Home [51] for unsupervised domain adaptation (ResNet-50 [13]). The best accuracy is indicated in bold. Method | |Ar → ClAr → PrAr → RwCl → ArCl → PrCl → RwPr → ArPr → ClPr → RwRw → ArRw → ClRw → Pr | Avg

表 1. 在 Office-Home [51] 上进行无监督领域适应的准确率 (ResNet-50 [13])。最佳准确率用粗体表示。方法 | |Ar → ClAr → PrAr → RwCl → ArCl → PrCl → RwPr → ArPr → ClPr → RwRw → ArRw → ClRw → Pr | Avg

ResNet-50 [13]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DANN [10]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN [34]	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CDAN+E [33]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
BSP+CDAN [4]	52.0	68.6	76.1	58.0	70.3	70.2	58.6	50.2	77.6	72.2	59.3	81.9	66.3
SymNets [56]	47.7	72.9	78.5	64.2	71.3	74.2	63.6	47.6	79.4	73.8	50.8	82.6	67.2
ETD [22]	51.3	71.9	85.7	57.6	69.2	73.7	57.8	51.2	79.3	70.2	57.5	82.1	67.3
BNM [5]	52.3	73.9	80.0	63.3	72.9	74.9	61.7	49.5	79.7	70.5	53.6	82.2	67.9
GSDA [15]	61.3	76.1	79.4	65.4	73.3	74.3	65.0	53.2	80.0	72.2	60.6	83.1	70.3
GVB-GD [6]	57.0	74.7	79.8	64.6	74.1	74.6	65.2	55.1	81.0	74.6	59.7	84.3	70.4
RSDA-MSTN [12]	53.2	77.7	81.3	66.4	74.0	76.5	67.9	53.0	82.0	75.8	57.8	85.4	70.9
SPL [53]	54.5	77.8	81.9	65.1	78.0	81.1	66.0	53.1	82.8	69.9	55.3	86.0	71.0
SRDC [47]	52.3	76.3	81.0	69.5	76.2	78.0	68.7	53.8	81.7	76.3	57.1	85.0	71.3
CLIP [42]	51.6	81.9	82.6	71.9	81.9	82.6	71.9	51.6	82.6	71.9	51.6	81.9	72.0
DAPL	54.1	84.3	84.8	74.4	83.7	85.0	74.5	54.6	84.8	75.2	54.7	83.8	74.5

ResNet-50 [13]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DANN [10]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN [34]	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CDAN+E [33]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
BSP+CDAN [4]	52.0	68.6	76.1	58.0	70.3	70.2	58.6	50.2	77.6	72.2	59.3	81.9	66.3
SymNets [56]	47.7	72.9	78.5	64.2	71.3	74.2	63.6	47.6	79.4	73.8	50.8	82.6	67.2
ETD [22]	51.3	71.9	85.7	57.6	69.2	73.7	57.8	51.2	79.3	70.2	57.5	82.1	67.3
BNM [5]	52.3	73.9	80.0	63.3	72.9	74.9	61.7	49.5	79.7	70.5	53.6	82.2	67.9
GSDA [15]	61.3	76.1	79.4	65.4	73.3	74.3	65.0	53.2	80.0	72.2	60.6	83.1	70.3
GVB-GD [6]	57.0	74.7	79.8	64.6	74.1	74.6	65.2	55.1	81.0	74.6	59.7	84.3	70.4
RSDA-MSTN [12]	53.2	77.7	81.3	66.4	74.0	76.5	67.9	53.0	82.0	75.8	57.8	85.4	70.9
SPL [53]	54.5	77.8	81.9	65.1	78.0	81.1	66.0	53.1	82.8	69.9	55.3	86.0	71.0
SRDC [47]	52.3	76.3	81.0	69.5	76.2	78.0	68.7	53.8	81.7	76.3	57.1	85.0	71.3
CLIP [42]	51.6	81.9	82.6	71.9	81.9	82.6	71.9	51.6	82.6	71.9	51.6	81.9	72.0
DAPL	54.1	84.3	84.8	74.4	83.7	85.0	74.5	54.6	84.8	75.2	54.7	83.8	74.5

Table 2. Accuracy (%) on VisDA-2017 [39] for unsupervised domain adaptation (ResNet-101 [13]). The best accuracy is indicated in bold.

表 2. 在 VisDA-2017 [39] 上进行无监督领域适应的准确率 (ResNet-101 [13])。最佳准确率用粗体表示。

Method	plane	bicycle	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg
ResNet-101 [13]	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DANN [10]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
JAN [34]	75.7	18.7	82.3	86.3	70.2	56.9	80.5	53.8	92.5	32.2	84.5	54.5	65.7
MCD [43]	87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
CDAN+E [33]	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.9
BSP+CDAN [4]	92.4	61.0	81.0	57.5	89.0	80.6	90.1	77.0	84.2	77.9	82.1	38.4	75.9
SWD [20]	90.8	82.5	81.7	70.5	91.7	69.5	86.3	77.5	87.4	63.6	85.6	29.2	76.4
DWL [54]	90.7	80.2	86.1	67.6	92.4	81.5	86.8	78.0	90.6	57.1	85.6	28.7	77.1
MODEL [23]	94.8	73.4	68.8	74.8	93.1	95.4	88.6	84.7	89.1	84.7	83.5	48.1	81.6
CGDM [9]	93.4	82.7	73.2	68.4	92.9	94.5	88.7	82.1	93.4	82.5	86.8	49.2	82.3
STAR [36]	95.0	84.0	84.6	73.0	91.6	91.8	85.9	78.4	94.4	84.7	87.0	42.2	82.7
CLIP [42]	98.2	83.9	90.5	73.5	97.2	84.0	95.3	65.7	79.4	89.9	91.8	63.3	84.4
DAPL	97.8	83.1	88.8	77.9	97.4	91.5	94.2	79.7	88.6	89.3	92.5	62.0	86.9

方法	飞机	自行车	公交车	汽车	马	刀	摩托车	人	植物	滑板	火车	卡车	平均
ResNet-101 [13]	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DANN [10]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
JAN [34]	75.7	18.7	82.3	86.3	70.2	56.9	80.5	53.8	92.5	32.2	84.5	54.5	65.7
MCD [43]	87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
CDAN+E [33]	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.9
BSP+CDAN [4]	92.4	61.0	81.0	57.5	89.0	80.6	90.1	77.0	84.2	77.9	82.1	38.4	75.9
SWD [20]	90.8	82.5	81.7	70.5	91.7	69.5	86.3	77.5	87.4	63.6	85.6	29.2	76.4
DWL [54]	90.7	80.2	86.1	67.6	92.4	81.5	86.8	78.0	90.6	57.1	85.6	28.7	77.1
模型 [23]	94.8	73.4	68.8	74.8	93.1	95.4	88.6	84.7	89.1	84.7	83.5	48.1	81.6
CGDM [9]	93.4	82.7	73.2	68.4	92.9	94.5	88.7	82.1	93.4	82.5	86.8	49.2	82.3
STAR [36]	95.0	84.0	84.6	73.0	91.6	91.8	85.9	78.4	94.4	84.7	87.0	42.2	82.7
CLIP [42]	98.2	83.9	90.5	73.5	97.2	84.0	95.3	65.7	79.4	89.9	91.8	63.3	84.4
DAPL	97.8	83.1	88.8	77.9	97.4	91.5	94.2	79.7	88.6	89.3	92.5	62.0	86.9

Table 3. Ablation: the effectiveness of domain-specific context (DSC). Domain-specific context is crucial for achieving good performance. The numbers show classification accuracy (%) on VisDA-2017 [39] dataset. Higher values are better. The numbers in brackets show absolute improvement from baseline.

表 3. 消融实验: 领域特定上下文 (DSC) 的有效性。领域特定上下文对于实现良好的性能至关重要。数字显示了在 VisDA-2017 [39] 数据集上的分类准确率 (%)。数值越高越好。括号中的数字显示了相对于基线的绝对提升。

Domain-agnostic	Domain-specific	Cls. Acc.
Manual	x	84.4
Unified	x	85.5 (+1.1)
Class-specific	x	86.2 (+1.8)
Unified	✓	86.9 (+2.5)
Class-specific	✓	86.9 (+2.5)

与领域无关	特定领域	分类准确率
手动	x	84.4
统一	x	85.5 (+1.1)
特定类别	x	86.2 (+1.8)
统一	✓	86.9 (+2.5)
特定类别	✓	86.9 (+2.5)

4.3. Ablation Study

4.3. 消融研究

To give a more detailed analysis of our method, we conduct several ablation studies on VisDA-2017 [39]. All of the variant models are trained with the same training hyper-parameters as described in Sec. 4.1.

为了对我们的方法进行更详细的分析，我们在 VisDA-2017 [39] 上进行了几项消融研究。所有变体模型均使用与第 4.1 节中描述的相同训练超参数进行训练。

Ablation: domain-specific context. To prove the effectiveness and necessity of domain-specific context, we compare the performances of these following prompt settings on VisDA-2017 [39] dataset: (1) the manually designed prompt "a photo of [CLASS]" as the baseline; (2) the domain-agnostic prompt in

the form of unified context (as shown in Eq. (3)); (3) the domain-agnostic prompt in the form of class-specific context; (4) the domain-agnostic prompt in the form of unified context with domain-specific context (as shown in Eq. (4)); and (5) the domain-agnostic prompt in the form of class-specific context with domain-specific context (as shown in Eq. (5)).

消融实验: 特定领域上下文。为了证明特定领域上下文的有效性和必要性, 我们比较了以下提示设置在 VisDA-2017 [39] 数据集上的表现: (1) 手动设计的提示“a photo of [CLASS]”作为基线; (2) 以统一上下文形式的领域无关提示 (如公式 (3) 所示); (3) 以类特定上下文形式的领域无关提示; (4) 以统一上下文形式结合特定领域上下文的领域无关提示 (如公式 (4) 所示); 以及 (5) 以类特定上下文结合特定领域上下文的领域无关提示 (如公式 (5) 所示)。

Manual prompt: "A photo of a [CLASS]" Prompts Domain agnostic prompt: " $[v]_1[v]_2 \dots [v]_M [CLASS]$ "
" Our prompt: " $[v]_1[v]_2 \dots [v]_{M_1}[d]_1^k[d]_2^k \dots [d]_{M_2}^k [CLASS]$ "

手动提示: "A photo of a [CLASS]" 提示领域无关提示: " $[v]_1[v]_2 \dots [v]_M [CLASS]$ " 我们的提示: " $[v]_1[v]_2 \dots [v]_{M_1}[d]_1^k[d]_2^k \dots [d]_{M_2}^k [CLASS]$ "

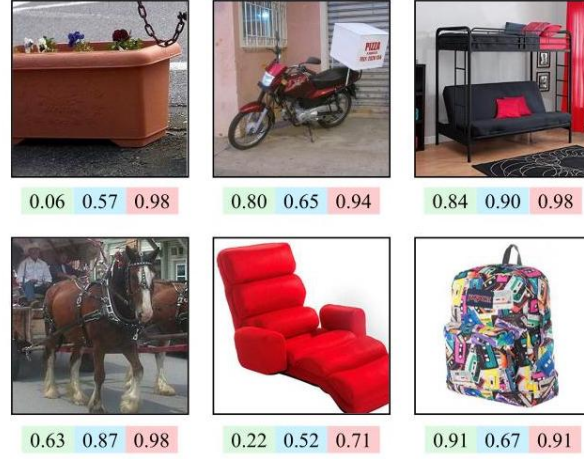


Figure 5. Prediction confidence from VisDA-2017 (top) and Office-Home dataset (bottom). Confidence of the ground-truth class predicted using different prompting methods. Blue: manually designed prompt. Green: domain-agnostic prompt. Pink: our proposed method. Predictions given by our method show the highest confidence.

图 5. 来自 VisDA-2017(上) 和 Office-Home 数据集 (下) 的预测置信度。使用不同提示方法预测的真实类别的置信度。蓝色: 手动设计的提示。绿色: 领域无关提示。粉色: 我们提出的方法。我们的方法给出的预测显示出最高的置信度。

The results of the above experiments are listed in Tab. 3. Even the manually design prompt is a strong baseline, our proposed DAPL (4) and (5) achieves 2.5% absolute improvement than the hand-crafted baseline (1). By comparing (2) with (3), we can observe that learning prompt with class-specific context can have a better performance than with unified context when domain-specific context is not used. Because the differences between classes can be better modeled by the class-specific context. Combining domain-specific context with the unified context (i.e., (4)) can further bring 1.4% performance improvement to (2). Besides, consistent performance improvement is also attained from (3) to (5). These improvements over the domain-agnostic context alone demonstrate the necessity of domain-specific context, which helps to capture the unique underlying domain information. Finally, by comparing (4) with (5), we know that tuning class-specific context with domain-specific context does not still yield improvement like (2) over (1). This is because distribution shift is the predominant factor in UDA, and modeling fine-grained discrepancy between classes may not further improve the performance. Thus, we choose the combination of unified context and domain-specific context in the paper.

上述实验的结果列在表 3 中。即使手动设计的提示是一个强基线, 我们提出的 DAPL (4) 和 (5) 相较于手工制作的基线 (1) 实现了 2.5% 的绝对改进。通过比较 (2) 和 (3), 我们可以观察到, 当不使用特定领域的上下文时, 学习具有类特定上下文的提示的性能优于统一上下文。这是因为类之间的差异可以通过类特定上下文更好地建模。将特定领域的上下文与统一上下文结合 (即 (4)) 可以进一步带来 1.4% 的性能提升。此外, 从 (3) 到 (5) 也获得了一致的性能提升。这些在仅使用领域无关上下文的基础上的改进证明了领域特定上下文的必要性, 它有助于捕捉独特的潜在领域信息。最后, 通过比较 (4) 和 (5), 我们知道, 使用领域特定上下文调整类特定上下文并没有像 (2) 相对于 (1) 那样带来改进。这是因为分布

转移是 UDA 中的主要因素，而建模类之间的细粒度差异可能不会进一步提高性能。因此，我们在本文中选择了统一上下文和领域特定上下文的组合。

Ablation: context token length. We conduct experiments in Tab. 4 to explore the influence of context token length. The lengths of domain-agnostic and domain-specific context tokens are denoted by M_1 and M_2 , respectively. From the results, we can see that the performance is a little lower when $M_1 < M_2$. Overall, the token length has little effect on the performance of our method. This implies the continuous representation could be learned with a small number of tokens.

消融实验: 上下文令牌长度。我们在表 4 中进行实验，以探索上下文令牌长度的影响。领域无关和领域特定上下文令牌的长度分别用 M_1 和 M_2 表示。从结果中可以看出，当 $M_1 < M_2$ 时，性能略低。总体而言，令牌长度对我们方法的性能影响很小。这意味着可以用少量令牌学习连续表示。

Table 4. Ablation: context token length. The accuracy (%) of different length combinations on VisDA-2017 [39] dataset (with ResNet-101 as image encoder). The values shown are (M_1, M_2) , i.e., context length of domain-agnostic prompt and domain-specific prompt. The best performance is denoted in bold.

表 4. 消融实验: 上下文令牌长度。不同长度组合在 VisDA-2017 [39] 数据集上的准确率 (%) (使用 ResNet-101 作为图像编码器)。所示值为 (M_1, M_2) ，即领域无关提示和领域特定提示的上下文长度。最佳性能以粗体表示。

Content token length	(4, 28)	(8, 24)	(28, 4)	(16, 16)	(24, 8)
Cls. Acc.	86.6	86.8	86.9	86.9	86.9

内容令牌长度	(4, 28)	(8, 24)	(28, 4)	(16, 16)	(24, 8)
分类准确率	86.6	86.8	86.9	86.9	86.9

Ablation: pseudo label threshold. In Tab. 5, we present the sensitivity of our method to the hyper-parameter τ by ranging it from 0.4 to 0.7. It seems that our method is not sensitive to τ because of the trade-off between quality and quantity of pseudo labels. For example, when τ is set to 0.7, the model is trained with fewer but more confident pseudo labels and the quality of pseudo labels may make up the performance drop brought by the reduced quantity.

消融实验: 伪标签阈值。在表 5 中，我们展示了我们的方法对超参数 τ 的敏感性，范围从 0.4 到 0.7。我们的研究表明，由于伪标签的质量与数量之间的权衡，我们的方法对 τ 并不敏感。例如，当 τ 设置为 0.7 时，模型使用较少但更自信的伪标签进行训练，伪标签的质量可能弥补了由于数量减少而带来的性能下降。

Table 5. Ablation: pseudo label threshold. The accuracy (%) of different threshold τ on VisDA-2017 [39] dataset (with ResNet-101 image encoder). The best performance is denoted in bold.

表 5. 消融实验: 伪标签阈值。不同阈值 τ 在 VisDA-2017 [39] 数据集上的准确率 (%) (使用 ResNet-101 图像编码器)。最佳性能以粗体表示。

Threshold τ	0.4	0.5	0.6	0.7
Cls. Acc.	86.9	86.9	86.7	86.6

阈值 τ	0.4	0.5	0.6	0.7
分类准确率	86.9	86.9	86.7	86.6

4.4. Visualization

4.4. 可视化

In Fig. 5, we compare the prediction confidence of the ground truth category on the target domain when using three different prompts: (a) a hand-crafted prompt; (b) the prompt with only domain-agnostic context; and (c) the prompt with domain-agnostic context and domain-specific context.

在图 5 中，我们比较了使用三种不同提示时目标领域真实类别的预测置信度: (a) 手工制作的提示; (b) 仅包含领域无关上下文的提示; (c) 同时包含领域无关上下文和领域特定上下文的提示。

For the third example of the top row, the plant only takes up a small area of the image. Hence, the prompt "a photo of a plant" is inappropriate for the image, while "a photo of a plant with a pot" might be a better match. Therefore, the hand-crafted prompt performs poorly on this example. In contrast, the learnable prompt yields a more confident prediction than the manually designed prompt. For the last image of the bottom row, it is a good match for the prompt "a photo of a backpack". The learnable domain-agnostic context performs worse than the manually designed prompt. By learning

domain information of "product", the domain-specific context enables the model with more confidence to predict the image as a backpack. Overall, these comparison results with different prompts validate that learnable domain-agnostic and domain-specific contexts improve the performance of our model when combined.

对于顶部行的第三个示例，植物仅占图像的一小部分。因此，提示“植物的照片”对于该图像不合适，而“带花盆的植物照片”可能更为匹配。因此，手工制作的提示在这个示例中表现不佳。相比之下，学习型提示的预测信心高于手动设计的提示。对于底行的最后一张图像，它与提示“背包的照片”非常匹配。学习型的领域无关上下文的表现不如手动设计的提示。通过学习“产品”的领域信息，领域特定的上下文使模型更有信心将图像预测为背包。总体而言，这些不同提示的比较结果验证了学习型领域无关和领域特定上下文在结合时提高了我们模型的性能。

5. Conclusion

5. 结论

In this paper, we introduce a novel prompt learning method for unsupervised domain adaptation, which is free of aligning features between domains as conventional methods do [32]. Instead, we design domain-specific context for each domain to advocate learning distinct domain representations of the source and the target domain. By making use of the prompt learning, We build a bridge between multimodality methods and domain adaptation methods. Extensive results have demonstrated the advantage of our method. Prompt learning methods can be extended to other visual tasks in unsupervised domain adaptation in the future, e.g., semantic segmentation.

在本文中，我们提出了一种新颖的无监督领域适应的提示学习方法，该方法不需要像传统方法那样在领域之间对齐特征 [32]。相反，我们为每个领域设计了领域特定的上下文，以促进学习源领域和目标领域的不同领域表示。通过利用提示学习，我们在多模态方法和领域适应方法之间架起了一座桥梁。大量结果证明了我们方法的优势。提示学习方法可以扩展到未来无监督领域适应中的其他视觉任务，例如语义分割。

Acknowledgements

致谢

This work is supported in part by the National Science and Technology Major Project of the Ministry of Science and Technology of China under Grants 2018AAA0100701, the NSFC under Grant 62022048, the Guoqiang Institute of Tsinghua University.

本工作部分得到了中国科技部国家科技重大项目（项目编号:2018AAA0100701）、国家自然科学基金（项目编号:62022048）以及清华大学国强研究院的支持。

References

参考文献

- [1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151-175, 2010. 1
- [2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NeurIPS*, pages 137-144, 2006. 1
- [3] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. Learning disentangled semantic representation for domain adaptation. *IJCAI*, 2019:2060-2066, 2019. 1,2
- [4] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *ICML*, volume 97, pages 1081-1090, 2019. 2, 7
- [5] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qing-ming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *CVPR*, pages 3940-3949, 2020. 2, 7
- [6] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *CVPR*, pages 12455-12464, 2020.7

- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, pages 248-255. Ieee, 2009. 1
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl-vain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021. 3
- [9] Zhekai Du, Jingjing Li, Hongzu Su, Lei Zhu, and Ke Lu. Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In CVPR, 2021. 7
- [10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In ICML, pages 1180-1189, 2015. 1, 2, 6, 7
- [11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In NeurIPS, pages 2672-2680, 2014. 2
- [12] Xiang Gu, Jian Sun, and Zongben Xu. Spherical space domain adaptation with robust pseudo-label loss. In CVPR, pages 9101-9110, 2020. 7
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770-778, 2016. 1, 3, 7
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770-778, 2016. 6
- [15] Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Unsupervised domain adaptation with hierarchical gradient synchronization. In CVPR, pages 4043-4052, 2020. 7
- [16] Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Juanzi Li, and Maosong Sun. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. arXiv preprint arXiv:2108.02035, 2021. 2
- [17] Gao Huang, Zhuang Liu, Geoff Pleiss, Laurens Van Der Maaten, and Kilian Weinberger. Convolutional networks with dense connectivity. TPAMI, 2019. 1
- [18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In ICML, volume 139, pages 4904-4916, 2021. 3
- [19] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neu-big. How can we know what language models know. TACL, 8:423-438, 2020. 3
- [20] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In CVPR, pages 10285-10295, 2019. 7
- [21] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. arXiv: 2104.08691, 2021. 3
- [22] Mengxue Li, Yiming Zhai, You-Wei Luo, Pengfei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In CVPR, 2020. 2, 7
- [23] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In CVPR, pages 9641-9650, 2020. 7
- [24] Shuang Li, Chi Liu, Qiuxia Lin, Binhui Xie, Zhengming Ding, Gao Huang, and Jian Tang. Domain conditioned adaptation network. In AAAI, volume 34, pages 11386-11393, 2020. 2
- [25] Shuang Li, Chi Harold Liu, Binhui Xie, Limin Su, Zheng-ming Ding, and Gao Huang. Joint adversarial domain adaptation. In ACM MM, pages 729-737, 2019. 2
- [26] Shuang Li, Shiji Song, Gao Huang, Zhengming Ding, and Cheng Wu. Domain invariant and class discriminative feature learning for visual domain adaptation. TPAMI, 27(9):4260-4273, 2018. 2
- [27] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In ACL/IJCNLP, pages 4582-4597, 2021. 3
- [28] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In ECCV, volume 8693, pages 740-755, 2014. 6
- [29] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hi-roaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586, 2021. 2
- [30] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hi-roaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv: 2107.13586, 2021. 3
- [31] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yu-jie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. arXiv:2103.10385, 2021. 2

- [32] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97-105, 2015. 1, 2, 9
- [33] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, pages 1647-1657, 2018. 2, 6, 7
- [34] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, pages 2208-2217, 2017. 1, 2, 7
- [35] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*, 2017. 6
- [36] Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *CVPR*, pages 9111-9120, 2020. 2, 6,
- [37] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *TKDE*, 22(10):1345-1359, 2009. 1
- [38] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *CVPR*, pages 2239- 2247, 2019. 2
- [39] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv:1710.06924*, 2017. 2, 6, 7, 8
- [40] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. Language models as knowledge bases? In *EMNLP-IJCNLP*, pages 2463-2473, 2019. 3
- [41] Nina Pörner, Ulli Waltinger, and Hinrich Schütze. BERT is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised QA. *arXiv: 1911.03681*, 2019. 3
- [42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pages 8748-8763, 2021. 2, 3, 4,7
- [43] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tat-suya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, pages 3723-3732, 2018. 2, 6,7
- [44] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. Auto-prompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, pages 4222-4235, 2020. 3
- [45] Petar Stojanov, Zijian Li, Mingming Gong, Ruichu Cai, Jaime Carbonell, and Kun Zhang. Domain adaptation with invariant representation learning: What transformations to learn? *NeurIPS*, 34, 2021. 2
- [46] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, pages 443- 450. Springer, 2016. 2
- [47] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *CVPR*, pages 8725-8735, 2020. 2, 6, 7
- [48] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521-1528, 2011. 1
- [49] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 1,2
- [50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko-reit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998- 6008, 2017. 3
- [51] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5385- 5394, 2017. 2, 6, 7
- [52] Jindong Wang, Yiqiang Chen, Wenjie Feng, Han Yu, Meiyu Huang, and Qiang Yang. Transfer learning with dynamic distribution adaptation. *TIST*, 11(1):1-25, 2020. 5
- [53] Qian Wang and Toby P. Breckon. Unsupervised domain adaptation via structured prediction based selective pseudo-labeling. In *AAAI*, pages 6243-6250, 2020. 2, 7
- [54] Ni Xiao and Lei Zhang. Unsupervised domain adaptation via structurally regularized deep clustering. In *CVPR*, pages 15242-15251, 2021. 2, 7
- [55] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (CMD) for domain-invariant representation learning. In *ICLR*, 2017. 1, 2
- [56] Yabin Zhang, Hui Tang, Kui Jia, and Minghui Tan. Domain-symmetric networks for adversarial domain adaptation. In *CVPR*, pages 5031-5040, 2019. 7
- [57] Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [MASK]: learning vs. learning to recall. In *NAACL-HLT*, pages 5017-5033, 2021. 3

[58] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. arXiv: 2109.01134, 2021. 3, 4