摘要

深度学习技术极大地促进了计算机视觉领域的发展。视觉感知算法在自动驾驶、智慧安防等场景实现了广泛的应用。与基于手工设计特征的传统计算机视觉算法相比，深度学习技术使视觉感知模型不仅能够从大规模数据中学习到更具有判别力的特征而且对各类视觉感知任务表现出强大的鲁棒性。然而，视觉感知模型的训练严重依赖采集自目标部署场景的高质量标注数据，受限于数据的收集和标注难度，训练集常常难以获取目标部署场景的数据标注、数据或任务定义。目标域数据标注的缺失会导致模型训练缺少监督信号，目标域数据的部分缺失会导致训练域和目标域的标签或者数据分布产生差异，目标域数据的完全缺失或者任务定义的缺失会使模型在训练过程中无法获取目标域信息。上述问题使训练场景和目标部署场景产生差异，严重影响模型性能。可泛化学习旨在提升模型处理未曾见过情景或任务的能力，从而解决上述场景中训练数据和目标域存在差异时的模型性能下降问题。为此，本文针对目标域标注缺失、目标域数据部分缺失、目标域数据完全缺失和目标域任务定义缺失四类可泛化学习场景展开研究，以提升模型在目标域的性能。本文的主要贡献包括：

针对**目标域标注缺失**问题，本文提出了基于域内域间相似度的无监督学习方法，该方法通过两阶段训练框架分别优化模型对图像域内和域间外观变化的鲁棒性。更进一步，方法提出了实例与摄像头风格归一化模块，该模块通过自适应实例批归一化减少了特征的域内外观变化，并利用变换归一化通过摄像头间的风格迁移增强了样本多样性提升了模型对域间外观变化的鲁棒性。上述方法直接提升了伪标签质量，仅利用无标注数据实现了对目标域高判别力模型的训练。实验结果显示，方法在 MSMT17 上实现了 64.4% 的第一查准率，性能领先同时期最先进方法 SpCL 超过 20%。

针对**目标域数据部分缺失**问题，本文面向部分类别数据缺失导致的训练集标签长尾分布和训练数据仅能覆盖部分目标域数据分布的问题，提出了解耦对比学习方法和增量模型增强设定。解耦对比学习方法有效避免了标签长尾分布时头部类别样本过多导致的模型偏差和尾部类别样本不足导致的特征欠表达问题，提升了模型在标签均匀分布目标域的性能。增量模型增强设定通过多次收集目标域数据实现训练域和目标域的分布对齐。为了降低训练数据多次获取导致的高训练复杂度，本文提出了记忆对比学习框架。该框架通过存储的旧样本特征和当前训练样本对模型的联合优化，实现模型在目标域性能的提升。实验结果显示，所提出方法在常用的长尾识别数据集 ImageNet-LT 上与对比学习方法相比正确率提升 6.5% 达到 57.7%。在增量模型增强设定下，方法在 ImageNet1K 上与直接微调基准相比将正确率从 55.5% 提升到 61.1%。

针对**目标域数据完全缺失**问题，本文面向决策平面偏移和特征表征能力差，提出了域适应提示学习方法和知识注入框架。域适应提示学习方法通过引入文本模态构建域适应分类器，实现分类器构成的决策平面对未知目标域分布的适应。知识注入框架利用可学习提示语句从文本编码器中提取任务相关知识，并通过知识注入模块完成对特征的知识注入，实现文本编码器中的丰富知识对特征的知识增强。得益于文本模态的加入，可泛化分类器在领域泛化基准 PACS 上与最先进单模态方法 PCL 相比，将正确率提升 5.0%。知识注入框架进一步提升了模型对未知目标分布、类别等的泛化能力，在 11 个数据集上与 CoCoOp 相比将平均正确率提升 4.4%。

针对**目标域任务定义缺失**问题，本文利用大语言模型将各类视觉感知任务统一为视觉问答形式，并提出了指代感知指令微调方法，实现了对未知视觉任务的可泛化学习。该方法利用现有数据集的标注，通过设计多样化的视觉指代感知任务，构建了高质量指令微调数据集。为了进一步增加指令微调的数据量，方法提出了自洽自举数据生成流程将任意目标检测密集标注拓展为目标框描述对。多样化的指令微调任务和丰富的数据使大语言模型获得对各类视觉感知任务的泛化能力。实验结果显示，模型在 12 个评测基准的零样本泛化性能超过同时期方法，与 Kosmos-2 相比，性能在 RefCOCO 上提升 24.7%。

本文围绕计算机视觉感知任务中的几类典型可泛化学习场景，提出了一组跨特征分布、跨类别、跨任务的可泛化模型训练方法。这些方法逐步消除了模型训练对目标域标注、数据和任务定义的依赖，最终实现了具有多模态和多任务能力的可泛化视觉感知模型。本文在各类基准测试集和各类可泛化设定上对所提出方法进行了验证，实验结果证明了本文方法相较于当前先进方法的性能优势。

关键词：可泛化学习，长尾识别，持续学习，多模态学习

Learning Methods for Generalizable Visual Perception Models

Shiyu Xuan (Computer Applied Technology)
Directed by: Prof. Shiliang Zhang

ABSTRACT

Deep learning brings impressive development to computer vision, giving rise to many applications such as autonomous driving and smart security. Compared with hand-crafted features-based methods, deep learning enables visual perception models to not only learn discriminative features from large-scale data but also show strong robustness to various visual tasks. However, the training of visual perception models relies heavily on high-quality labeled data collected from target deployment scenarios. Due to the difficulty of data collection and labeling, obtaining labeled data, data, or task definitions for target deployment scenarios in the training set can be challenging. The lack of target domain labeled data will lead to the lack of supervision signals in model training. The partial lack of target domain data will lead to differences in labels or data distribution between the training domain and the target domain. The complete lack of target domain data or task definitions will make the model unable to obtain target domain information during training. The above problems cause differences between training scenarios and target deployment scenarios, seriously affecting model performance. Generalization learning aims to improve the model's ability to handle unseen scenarios or tasks, thereby solving the problem of model performance degradation when there are differences between the training data and the target domain. To this end, this thesis conducts research on four types of generalization learning scenarios: lack of labeled data from the target domain, partial lack of data from the target domain, complete lack of data from the target domain, and lack of target domain task definitions. The main contributions of this thesis include:

In terms of **lack of labeled data from the target domain**, this thesis proposes an unsupervised learning method based on intra-inter similarity learning. This method uses a two-stage training framework to enhance the robustness of the model to intra-inter domain appearance variations. Furthermore, the method proposes an instance and camera style normalization. The intra-domain appearance variation can be reduced through adaptive instance and batch normalization. The transform normalization increases sample diversity through style transfer

between cameras, improving the robustness of the model to inter-domain appearance variation. The above method directly improves the quality of pseudo-labels and achieves the training of high-discrimination models in the target domain using only unlabeled data. With only unlabeled data, the method achieves a rank-1 accuracy of 64.4% on MSMT17, outperforming the recent method SpCL by 20+%.

In terms of **partial lack of data from the target domain**, this thesis first proposes decoupled contrastive learning and patch-based self-distillation methods. This method avoids the model bias caused by head category samples and the under-representation of the tail classes, which significantly improves the model performance on the test set with uniform label distribution. Secondly, this thesis proposes an incremental model enhancement setting to achieve distribution alignment between the training set and the target domain by collecting data multiple times. To reduce the high training complexity caused by sequentially arrived training data, this thesis proposes a memory contrastive learning framework, achieving continuous enhancement of model performance. Experimental results show that the proposed method improves the accuracy by 6.5% to 57.7% compared with the contrastive learning method on the commonly used long-tail recognition data set ImageNet-LT. Under the incremental model enhancement setting, the method improves the accuracy from 55.5% to 61.1% on ImageNet1K compared with the fine-tuning baseline.

In terms of **complete lack of data from the target domain**, to learn generalizable models for unknown target domains, this thesis proposes the domain-adaptive prompt learning method. This method leverages text modality to construct domain-adaptive classifiers to adapt decision boundaries to the unknown target domains. By leveraging the rich knowledge of the text encoder in the vision-language model, this thesis enhances the visual features and classifier features. Specifically, the knowledge features are generated by several learnable prompt sentences. A knowledge injection framework is proposed to refine classifier features and visual features with knowledge features. The proposed method effectively improves the model's generalization ability to unknown target distributions and categories. Benefiting from the introduction of the text modality, the proposed method outperforms the recent single-modal DG method PCL by 5.0% on PACS. The knowledge injection framework further improves the model's generalization ability to unknown target distributions and categories. It outperforms CoCoOp by 4.4% under the base-to-new classes generalization setting.

In terms of **lack of target domain task definitions**, this thesis leverages the large language model to unify various visual tasks into the form of visual question answering and pro-

poses a referential comprehension instruction tuning method. This method constructs a high-quality instruction tuning data set by designing diverse referential comprehension tasks. To further involve more training data, the method proposes a self-consistent bootstrapping process to extend bounding box annotations into referring-expression-bounding-box pairs. Diversified instruction tuning tasks and data enable large language models to acquire various visual perception capabilities and generalize well to various unknown tasks. It also achieves the best zero-shot performance on 12 benchmarks and surpasses the accuracy of Kosmos-2 by 24.7% on RefCOCO.

The thesis conducts research on learning methods for generalizable vision models and proposes a set of cross-distribution, cross-category, and cross-task model training methods. These methods gradually eliminate the dependence of model training on target domain labeled data, data, and task definition. It achieves a generalizable model with multi-modal and multi-task capabilities. This thesis validates the proposed method on various benchmarks and various generalization settings. The experimental results demonstrate the superior performance of the proposed methods compared with recent advanced methods.

KEY WORDS: Generalizable Model Learning, Long-tailed Recognition, Incremental Learning, Multi-modal Learning