学校代码：10284
分 类 号：TP181
密　　级：公开
U D C：004.8
学　　号：DG21330029

南京大学

博 士 学 位 论 文

论 文 题 目　深度学习中的数据增强与选择方法研究

作 者 姓 名　杨锁荣

专 业 名 称　计算机科学与技术

研 究 方 向　人工智能

导 师 姓 名　申富饶教授

2025 年 11 月 25 日

答辩委员会主席 武港山 教授

评　　阅　　人 武港山 教授

路通 教授

戴新宇 教授

张道强 教授

魏秀参 教授

论文答辩日期　2025 年 11 月 19 日

研究生签名：

导师签名：

Research on Data Augmentation and Selection Methods in Deep Learning

by

Suorong Yang

Supervised by

Professor Furao Shen

A dissertation submitted to

the graduate school of Nanjing University

in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Computer Science and Technology

Department of Computer Science and Technology

Nanjing University

November 25, 2025

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：深度学习中的数据增强与选择方法研究

计算机科学与技术 专业 2021 级博士生姓名： 杨锁荣

指导教师（姓名、职称）： 申富饶教授

摘 要

近年来，深度学习在计算机视觉、自然语言处理、多模态理解等领域取得了突破性进展，其成功得益于更大规模的模型架构、更强的算力以及更丰富的训练数据。然而，随着模型参数量和计算需求的急剧扩张，数据正在逐渐成为限制深度学习进一步发展的关键瓶颈：一方面，高质量数据的获取和标注成本高昂，且由于受到隐私与合规性的限制，数据集规模的提升进一步受限；另一方面，真实世界数据普遍包含噪声与冗余，降低了训练效率并影响模型的泛化能力。因此，如何科学地理解、度量与优化训练数据，以更高效地挖掘数据价值，成为当前人工智能发展中亟需解决的关键问题。本文以数据为核心研究对象，从“理解—增强—选择—协同”的角度系统展开研究，旨在构建统一的数据分析与优化框架，以提升深度模型的训练效率与泛化性。总结来说，本文涉及的主要工作包括：

1. **基于相似性与多样性的数据分析框架：**针对现有方法依赖模型性能间接度量数据效果的问题，提出了一套与训练无关的数据分析框架，从相似性与多样性两个维度刻画训练数据特性，相似性度量刻画数据和目标分布的偏离，而多样性则度量了数据在特征空间中是否具备足够的覆盖和差异性。该框架揭示了数据特性与模型性能之间的内在联系，为理解不同数据影响机制提供了可解释性工具，并为后续方法设计奠定理论基础。
2. **基于多样性提升的自适应数据增强研究：**在统一理论框架指导下，针对现有数据增强方法依赖固定或随机增强幅度、难以适配训练中模型与数据在高维空间中的动态变化的局限，提出了基于强化学习的自适应数据增强方法 AdaAugment，通过策略网络和目标网络的联合优化，该方法能够利用目标网络的实时反馈动态调整增强强度，使训练数据的相似性与多样性分

布自适应匹配模型状态，从而有效缓解欠拟合与过拟合风险，并显著提升模型的泛化性。

3. **基于相似性驱动的高质量数据选择研究：**真实世界数据中不可避免地伴随噪声、冗余和域偏移，如果不能正确区分高价值样本与有害样本，即使增强策略再完善，也难以发挥预期效果。针对这一问题，本章提出了一种基于多模态表征的数据选择方法。该方法通过轻量化适配器缓解域偏移，并结合语义对齐分数与多样性分数双重指标全面评估样本价值，再利用多目标优化与比例约束减轻群体效应。实验表明该方法在噪声场景下依然保持稳定表现，显著提升了数据选择机制的鲁棒性和训练效率。更重要的是，这一章的研究从一定程度上解决了“如何保证被利用的数据本身可靠”的问题，为下一章进一步探讨数据增强与选择的协同优化奠定了基础。
4. **面向相似性-多样性联合优化的增强与选择协同研究：**在前两章的研究中，本文分别针对泛化性问题提出了自适应数据增强方法，针对数据质量问题提出了基于多模态信息的数据选择方法。然而，当增强与选择被独立使用时，仍可能面临信息损失、多样性不足以及在噪声场景下增强策略失效等局限。为此，本文进一步提出了一种增强与选择的协同优化范式，将二者有机结合：一方面，数据选择为增强过程动态筛选高质量、信息量丰富的样本，保证增强作用集中在最有价值的训练实例上；另一方面，数据增强有效补偿了选择所带来的信息损失与多样性不足，从而在效率与泛化之间形成互补与平衡。使得可以在大幅压缩训练数据规模与计算开销的同时，依然保持甚至提升模型性能，并在含有噪声与损坏的复杂场景中展现出优异的鲁棒性与跨场景适应能力。本章作为前两章方法的综合与提升，标志着统一数据优化框架的完整落地。

总体而言，本文围绕“理解—增强—选择—协同”的主线，系统提出了从数据度量到自适应增强、再到多模态选择及其协同优化的一系列方法，构建了统一而完整的数据优化研究框架。研究成果不仅深化了对数据特性与模型性能内在关系的理解，也为实现高效、鲁棒且可扩展的深度学习提供了坚实的理论支撑与切实的实践路径。

关键词：数据智能，深度学习，数据增强，数据选择，多模态学习

南京大学研究生毕业论文英文摘要首页用纸

THESIS: Research on Data Augmentation and Selection Methods in Deep Learning

SPECIALIZATION: Computer Science and Technology

POSTGRADUATE: Suorong Yang

MENTOR: Professor Furao Shen

ABSTRACT

In recent years, deep learning has achieved breakthrough progress in computer vision, natural language processing, and multimodal understanding, driven by increasingly large-scale architectures, stronger computational power, and richer training data. However, as model parameters and computational demands continue to grow, data is gradually becoming the critical bottleneck for further advances in deep learning. On the one hand, the collection and annotation of high-quality data are costly, and further scaling is constrained by privacy and compliance requirements. On the other hand, real-world data inevitably contain noise and redundancy, reducing training efficiency and impairing model generalization. Therefore, a pressing challenge in modern artificial intelligence is how to scientifically understand, measure, and optimize training data to fully unlock its value for efficient and robust learning.

This dissertation takes data as the central research object and systematically investigates it from the perspectives of understanding, augmentation, selection, and synergy, with the goal of building a unified framework for data analysis and optimization to enhance both the efficiency and generalization of deep models. The main contributions are as follows:

1. **Similarity- and diversity-based data analysis framework:** To address the limitation that existing methods rely on model performance to indirectly evaluate data effectiveness, we propose a training-independent analysis framework that characterizes training data from the two dimensions of similarity and diversity.

Similarity measures the deviation between the data and the target distribution, while diversity evaluates whether the data provides sufficient coverage and heterogeneity in the feature space. This framework reveals the intrinsic relationship between data properties and model performance, offering interpretable tools for understanding different data impact mechanisms and laying the theoretical foundation for subsequent method design.

2. **Adaptive data augmentation with diversity enhancement:** Under the guidance of the unified theoretical framework, we propose AdaAugment, a reinforcement learning‑based adaptive data augmentation method, to overcome the limitations of existing augmentation techniques that rely on fixed or random strengths and fail to adapt to the dynamic changes of models and data in high-dimensional spaces. Through joint optimization of a policy network and a target network, AdaAugment leverages real-time feedback from the target network to dynamically adjust augmentation intensity. This allows the similarity and diversity distribution of the training data to adaptively match model states, thereby mitigating underfitting and overfitting risks and significantly improving generalization.
3. **High-quality data selection based on similarity modeling:** Real-world data inevitably contains noise, redundancy, and domain shifts. Without distinguishing high-value samples from harmful ones, even the best augmentation strategies cannot achieve their intended effect. To address this, we propose a multimodal representation‑based data selection method. This method employs lightweight adapters to alleviate domain shift, and evaluates sample value comprehensively through dual metrics of semantic alignment scores and diversity scores. Multi-objective optimization with ratio constraints is further applied to reduce group effects. Experiments demonstrate that the method maintains stable performance under noisy conditions, significantly improving the robustness and efficiency of data selection. More importantly, this study partially resolves the problem of "ensuring the reliability of the utilized data," paving the way for the following chapter on the coordination between data augmentation and selection.
4. **Co-optimization of augmentation and selection via similarity‑diversity joint**

modeling: Building upon the previous two chapters, this work proposes AdaAugment for the generalization problem and multimodal data selection for the data quality problem. However, when augmentation and selection are applied independently, issues such as information loss, insufficient diversity, and failure of augmentation under noisy scenarios may still arise. To overcome these limitations, we introduce a co-optimization paradigm that organically integrates augmentation and selection. On one hand, data selection dynamically filters high-quality, informative samples for augmentation, ensuring augmentation is concentrated on the most valuable training instances. On the other hand, data augmentation compensates for the information loss and lack of diversity brought by selection. This synergy achieves a balance between efficiency and generalization, enabling substantial reductions in data scale and computational cost while maintaining or even improving model performance, and exhibiting strong robustness and cross-domain adaptability under noisy and corrupted scenarios. As the integration and extension of the previous two chapters, this study marks the full realization of the proposed unified data optimization framework.

In summary, this dissertation follows the mainline of understanding–augmentation–selection–synergy, and systematically proposes a series of methods spanning data measurement, adaptive augmentation, multimodal selection, and their synergistic integration. Together, these contributions form a unified and complete framework for data optimization, which not only deepens the understanding of the intrinsic relationship between data characteristics and model performance, but also provides solid theoretical foundations and practical methodologies for achieving efficient, robust, and scalable deep learning.

KEYWORDS: Data Intelligence; Deep Learning; Data Augmentation; Data Selection; Multimodal Learning

目　录

插图目录

表格目录

第一章　绪论

1.1　引言

近年来，深度学习在计算机视觉、自然语言处理、多模态理解等领域取得了突破性进展，并迅速渗透至医疗、金融、交通等广泛应用场景。这一波浪潮的核心动力来自三个方面：更复杂高效的网络架构设计、更强大的计算资源、以及更大规模、更高质量的数据支撑。从最初的 AlexNet[1]和 VGG[2]，到后来的 ResNet[3-4]，再到近几年的 Transformer[5-6]和 GPT[7]等超大规模语言模型，研究者持续推动模型规模与深度的上限，深度学习模型的参数量和计算复杂度呈指数级增长，不断刷新图像识别、语言生成和跨模态推理等任务的性能上限。与此同时，GPU、TPU 等专用硬件架构的出现，以及分布式训练、并行计算和高效算子优化的发展，在很大程度上推动了深度学习的快速演进，使得更大规模模型的训练成为可能[8]。

然而，随着模型规模的不断扩张，模型、算力与数据三者的发展出现了显著的不均衡性[9-11]。计算需求的增长远超摩尔定律：自 2010 年以来，用于机器学习训练的计算量每 4-9 个月翻一倍；相比之下，硬件 FLOPs/美元的提升周期约为 2.5 年[12-13]。再模型参数方面，从视觉模型 Inception-v4 的 4260 万参数，到 GPT-4 超过 1 万亿的参数量，规模攀升前所未有。而在数据层面，虽然有 ImageNet、MS-COCO、Pascal 等经典数据集[14-16]，但近年来真正突破性的公共数据集却相对有限。主要原因是，高质量数据的采集与标注成本高昂，数据的隐私和合规性限制严格，且真实场景数据不可避免地存在噪声和冗余。这使得可用的高质量数据扩展日益受限。更为严峻的是，随着模型继续扩展到超大规模，算力和模型规模逐渐逼近数据资源上限，人工智能正面临“数据枯竭”的挑战。

在这一背景下，逐渐形成共识：人工智能的发展重心需要从“以模型为中心”转向“以数据为中心”的人工智能（Data-Centric AI）。所谓数据中心化，并非单纯增加数据规模，而是强调通过可解释的度量、系统化的治理与有针对性

的优化来提升数据的质量与利用效率，从而在资源不对称的格局下，为模型性能增长开辟新的可持续路径。数据的价值挖掘与智能管理正在成为新的研究前沿[17-18]。

本文正是沿着这一思路展开：以数据为核心对象，围绕“理解—增强—选择—协同”构建统一的数据分析与优化框架。首先，从理论上提出相似性与多样性度量的解释性框架，揭示数据特性与模型性能之间的内在联系；其次，面向训练动态性，设计自适应增强策略，在相似性与多样性之间实现动态平衡；再者，针对真实场景中的噪声与冗余，提出鲁棒的数据选择机制，以确保数据集本身的代表性与可靠性；最后，将增强与选择有机结合，通过协同优化在保证泛化的同时实现训练高效性。

1.2 国内外研究现状

近年来，随着深度学习的快速发展，国内外学术界和工业界逐渐认识到数据在人工智能系统中的核心地位，研究重心正从“以模型为中心”转向“以数据为中心”。国外以 Andrew Ng 提出的 Data-Centric AI 为代表，强调通过数据质量提升与治理来驱动模型性能改进；国内也涌现出一系列面向大规模任务的数据治理与优化研究[19-21]。概括而言，数据智能研究的起点在于理解与刻画数据特性，即以合理的指标体系揭示数据分布与模型性能之间的关系，从而为后续方法提供可解释的理论支撑。

然而，现阶段仍缺乏统一的理论体系与量化标准，相关研究大多依赖经验性方法。绝大多数数据评价体系仍建立在模型训练的最终性能之上，即通过验证集或测试集上的准确率、损失值等指标间接衡量数据价值。这类方法虽然直观，但存在两方面不足：其一，过度依赖具体模型结构与训练过程，难以分离数据特性本身的贡献，因而缺乏普适性和可解释性；其二，只能在训练完成后进行“事后评估”，无法为数据处理或方法设计阶段提供前置指导。受制于这些局限，数据度量与优化仍然倾向于经验驱动，亟需发展一种模型无关、可解释且可前置应用的系统性框架，以建立数据特性与模型性能之间的映射关系。

在理论体系尚未完善的背景下，研究者们从实践角度提出了两条最具代表性的技术路径：数据增强（Data Augmentation）和数据选择（Data Selection）。前

者通过对现有样本施加各种变换扩展数据空间，以提升样本多样性与覆盖度来缓解过拟合风险并增强泛化；后者则在大规模原始数据中“提纯”，通过评价与优化构造更具代表性的训练子集，以降低冗余与噪声、提升效率与稳定性。两条路径的共同目标，是在资源受限与数据质量参差的现实条件下，以更低成本释放数据价值。

具体到数据增强方面，国内外研究经历了从几何变换、颜色扰动到 Cutout[22]、Mixup[23]、CutMix[24] 等区域擦除和样本混合方法，再到 AutoAugment[25]、RandAugment[26] 等自动化搜索策略，近年亦出现基于模型反馈的增强优化机制，如 TeachAugment[27]、DADA[28]。然而，多数方法仍依赖随机或固定强度的增强，自动策略虽减轻了人工调参负担，却往往需要高额搜索成本，难以在大规模任务中推广[29]。为此，增强机制亟需具备面向训练动态的自适应能力：在早期提供更高相似性以稳健学习基本模式，在后期引入更高多样性以抑制过拟合并提升泛化，实现在“相似性—多样性”之间的动态平衡。

与之相对应，数据选择旨在构建尽可能小而性能无损的训练子集，代表性技术包括基于样本重要性分数的方法、基于分布覆盖的代表性选择方法以及基于优化的子集构造方法。典型代表如 Forgetting[30]、MoSo[31]、RL-Selector[32]、InfoBatch[33]等在不同场景展现出潜力。数据选择的价值在于，它能够有效去除冗余和噪声样本，显著降低存储与计算开销，从而提升训练效率和模型稳定性。在实际应用中，即使在在计算与存储资源有限的条件下，数据选择仍能保证模型达到与全量数据相近甚至更优的性能。然而，数据选择仍面临挑战：其一，业界普遍“偏好难样本”，但在单模态线索下难以区分“真正困难样本”与“噪声/错标样本”，在真实噪声场景中易产生偏差；其二，数据之间存在显著的交互作用，仅依赖独立评价每个数据点的重要性，容易忽视样本间的覆盖性与互补性，从而导致所选子集并非全局最优；其三，子集规模的缩减往往伴随着数据多样性的下降，这虽然在一定程度上提升了训练效率，却可能限制模型的泛化能力。由此可见，选择机制需要同时提升抗噪性、分布覆盖与多目标平衡能力，方能在真实应用中稳定奏效。

总体而言，数据增强关注“增加有效多样性”，而数据选择强调“减少无效冗余”，两者从不同方向提升数据利用效率，在效率与泛化之间形成天然张力。如何在统一视角下打通增强与选择，建立可解释的度量指导，并在训练动态与噪

声现实中实现协同优化，已成为当前数据中心化人工智能的关键挑战与重要发展方向，即“以更少代价发挥数据更大价值”的核心理念。

1.3 以数据为中心研究的关键问题与挑战

近年来，随着模型规模的持续扩大，人工智能的关注点正从“堆叠更大的模型”转向“如何更好地利用数据”。在这一趋势下，以数据为中心的人工智能（Data-Centric AI）逐渐成为推动智能系统发展的一条关键路径。与传统关注模型和训练的研究不同，以数据为中心围绕数据质量、数据价值和数据使用策略进行系统性的建模、分析与优化。以数据为中心的研究关注以下问题：数据度量、数据优化、理论与统一框架。其中数据度量构建可解释、可泛化的数据价值指标，数据优化通过增强、过滤、修剪等方式改善数据质量与任务匹配度；理论与统一框架整合度量方式与具体的优化技术。这些内容不是关注单一技术，而是一套自洽的研究范式，贯穿了模型训练的全流程，构成了以数据为中心的理论基础和方法体系，旨在突破当前深度学习对大规模数据和计算资源的高度依赖。本文将当前亟待解决的核心挑战总结为以下四点，它们共同构成“理解—增强—选择—协同”的主线，并分别对应后文的理论与方法路径。

缺乏与训练过程解耦的数据度量理论：现有数据价值的评估多依赖完整训练后的验证精度、损失或迁移表现等“事后度量”。这类方法虽然直观，但存在两方面局限：其一，计算开销高、对具体模型结构和优化过程高度敏感，难以跨任务、跨规模推广；其二，缺乏可解释性，难以揭示数据特性与模型性能之间的因果关系，也使得增强、选择方法缺少统一依据。亟需建立一种训练无关且具备可解释性的度量理论，将数据特性与模型性能建立起可操作的映射关系，为后续优化方法提供统一的理论基座。

缺乏对训练动态的适应性：数据增强是提升模型泛化能力的重要手段，但现有方法普遍依赖固定策略或随机扰动，长期停留在“训练前设计一次，训练期间不再变化”的静态范式。然而，深度模型的学习过程高度动态：一方面，不同样本本身的难度差异显著；另一方面，模型表征能力在训练过程中不断演化，比如，早期可能需要更强的扰动来提升鲁棒性，而后期则需要更精细的调整以避免过拟合。缺乏对训练动态的适配，使得增强效果与模型状态脱节，导致欠拟合或过

拟合。如何实现增强策略与训练进程的深度耦合，是提升泛化能力的关键挑战。

缺乏对噪声场景的鲁棒性：在真实世界的数据中，训练数据往往伴随不可避免的噪声、冗余和域偏移等问题，这些问题不仅增加了计算负担，有可能对模型学习产生误导，严重降低效率与泛化能力。现有的数据选择技术往往建立在一个较强的假设之上，即假设数据集中噪声样本极少甚至不存在。从而选择方法倾向于把所有“高损失”样本都视为难例而非噪声，这在复杂现实场景中容易误判，导致大量有害样本被保留，甚至使选择机制在噪声占比较高时完全失效。缺乏鲁棒性不仅削弱了选择方法的实用性，还可能使模型训练偏离正确的优化方向。因此，要实现真正的数据智能，必须具备一种能跨模态、跨分布识别语义一致性的机制，能在不依赖强假设的前提下区分高价值与有害样本的选择机制，同时兼顾分布覆盖与样本互补性。

增强和选择缺乏协同优化：增强用于扩充信息与多样性，选择用于压缩冗余与噪声。然而，现有研究往往将二者作为独立问题处理，忽视了它们之间潜在的互补性。单纯减少样本数量虽然能降低计算成本，却不可避免地损失部分信息和多样性，从而提升模型的过拟合风险；而单纯依赖增强手段增加样本多样性，又可能引入冗余或不必要的扰动，无法保证训练效率。增强与选择各自的局限性，使其难以在效率与性能之间取得平衡，若在统一框架下实现协同优化，用选择为增强提供高价值载体，用增强补偿选择导致的信息损失与多样性下降，从而实现高效性与泛化性的统一。

综上所述，上述四个问题相互联动、层层递进：度量理论为方法提供统一依据，动态增强直指泛化性提升，鲁棒选择聚焦效率与数据可靠性，而协同优化则综合二者以达成全局最优。围绕这一主线，本文在第二章提出统一理论框架，第三章研究自适应增强以提升泛化，第四章研究鲁棒选择以提升高效，第五章进一步实现增强与选择的协同优化，并通过系统实验验证其有效性与普适性。

1.4 本文的主要工作

依据内容的逻辑顺序，本文的组织结构如图1-1所示。

在第三章中，本章提出了一个基于相似性与多样性的训练数据分析框架。通过引入一套与模型训练无关的量化指标体系，从“相似性—多样性”两个基本维

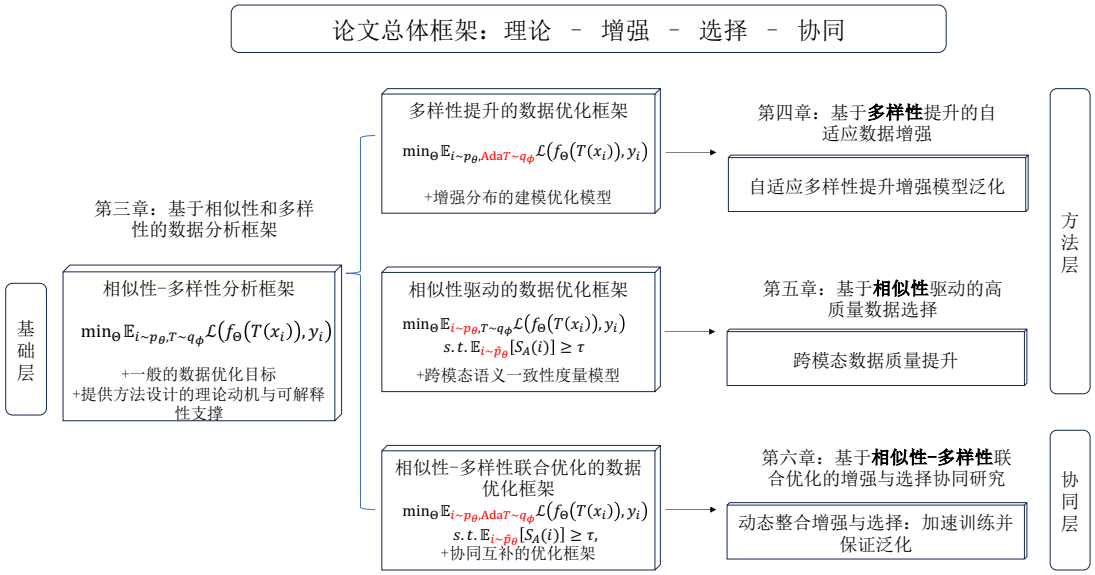


图 1-1　本文主要工作组织结构

度系统性地刻画了训练数据分布特性与模型性能之间的关系。该框架不仅揭示了不同数据集和模型在相似性与多样性上的差异化需求，还为理解数据增强和数据选择的作用机制提供了统一视角。实验结果表明，相似性与多样性指标能够有效解释不同训练数据特性的优劣，为后续方法设计和参数调优提供了理论依据与可解释性支撑。

在第四章中，在理论框架的指导下，针对传统数据增强方法自适应性不足、难以匹配训练动态的问题，本章提出基于强化学习的自适应数据增强方法 AdaAugment 来提升数据多样性。方法采用策略网络-目标网络的双模型结构，利用目标网络的实时反馈动态调控增强强度，使数据在训练不同阶段于“相似性—多样性”之间自适应平衡：早期避免欠拟合，后期缓解过拟合。与自动搜索类方法相比，AdaAugment 无需额外搜索开销即可实现分布自适应，显著提升了模型的泛化性能与训练效率。

在第五章中，考虑到若数据本身含噪或失配，则会严重影响增强的效力和模型的泛化。为此本章提出了一种基于相似性驱动的数据选择方法。与仅依赖图像模态的传统方法不同，该方法利用图像-文本双模态信息，通过语义对齐与数据多样性的联合建模，更加鲁棒地识别代表性样本和去除噪声样本。进一步地，本文在选择过程中引入多目标优化机制，有效缓解了样本间的群体效应问题。该方法在多种任务与噪声场景下稳定降低训练成本，同时保持或提升性能，为第六章

的协同优化奠定数据可靠性基础。

在第六章中，鉴于增强强调“增加有效多样性”、选择强调“减少无效冗余”，二者独立使用会在效率与泛化之间产生张力。本章提出统一的协同优化范式：由选择动态供给高质量、高信息量样本作为增强载体，增强则补偿选择带来的信息损失与多样性下降；并结合语义一致性与结构稀疏性进行联合建模，动态识别最适合增强的样本。在显著压缩训练数据规模与计算开销的同时，该框架保持甚至提升模型性能，并在含噪与跨域场景中展现出良好的鲁棒性与适应性。

综上所述，本文从理论分析、方法设计到框架整合，围绕数据中心化人工智能中的核心问题展开系统研究，形成了一条完整的研究链条：相似性与多样性的数据分析框架提供理论依据和可解释性基础，自适应增强指向多样性提升，相似性驱动的选择又聚焦效率与数据可靠性，最后进行协同优化实现二者统一。所提出的框架与方法丰富了数据优化的理论与实践，为实际任务中的高效、稳定、可扩展训练提供了可行路径。

第二章　数据增强与选择研究进展

在以数据为中心的视角下，提升数据价值大体有两条路径：其一是扩展有效多样性，即在不增加采集与标注成本的前提下，通过构造合理的变换来扩大可学习分布、注入先验不变性与正则化能力，这就是数据增强。其二是减少无效冗余与噪声，即通过筛选获得更小而更优的训练子集——即数据选择。本节将先概述数据增强的相关研究进展及其方法谱系，之后介绍数据选择方向的发展脉络与关键方法，最后简要讨论数据集蒸馏方向。

2.1　数据增强

数据增强（Data Augmentation）是一种广泛应用于机器学习和深度学习中的技术[11,34-36]，其核心是通过对原始数据进行各种变换和扩充，从而提高模型的性能和鲁棒性。其本质在于通过生成新的数据样本，增加训练集的规模和多样性，从而降低模型的过拟合风险。对于图像任务，常见的数据增强方式包括裁剪、翻转、颜色域变换等操作，这些方法可以在不增加数据采集成本的前提下，将原有训练集扩展至数倍以上。如果将多种操作进行组合（如裁剪 + 翻转），数据集规模的扩充效果则更加显著。对于文本数据，增强方式通常包括随机替换、插入、删除或同义词替换等，以生成新的文本变体。增强后的样本可被视为从接近原始分布的近似分布中采样而来，从而使扩充后的训练集覆盖更广泛的特征空间，帮助模型学习到更加稳健的判别边界并具备更强的泛化性能。从另一个角度来看，随着深度模型的规模越来越大，当模型复杂度超过数据集复杂度时，模型会遭遇过拟合问题。因此，为了获得更好的性能表现，可以在不改变模型规模大小的前提下，通过数据增强技术来平衡模型和数据集之间的复杂度差异，提高训练数据集的复杂度，缓解模型的过拟合风险。数据增强的目的就是增加训练数据的数量和多样性，以改善模型的泛化能力，减少过拟合现象的发生，并提高模型的鲁棒性，使其在面对不同的输入数据时表现更好。

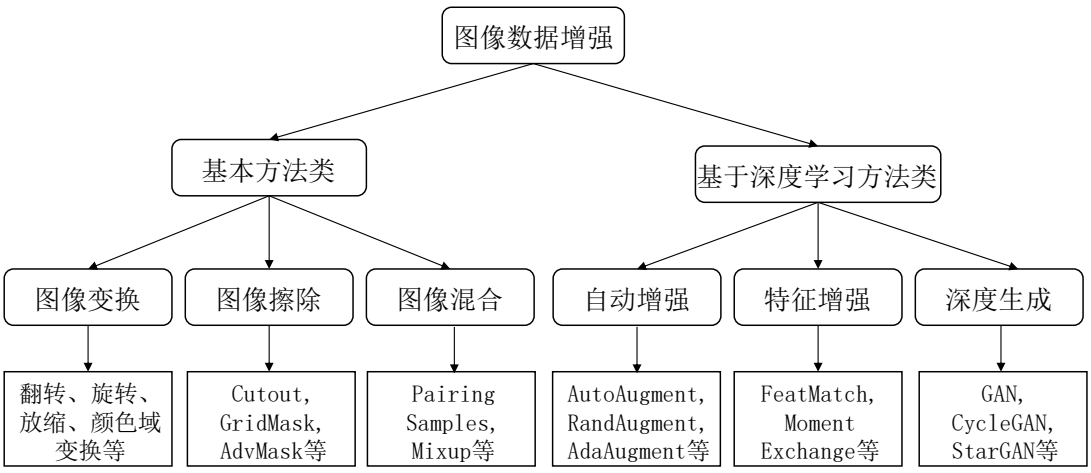


图 2-1　新型图像数据增强的分类方法

近年来，数据增强研究已成为深度学习领域的重要组成部分，研究者们提出了大量的数据增强方法，为了更系统地对其进行归纳，本章首先提出了一个分类方法，如图 2-1所示。该分类方法首先根据是否引入深度学习技术，将数据增强划分为两大类：基本方法和基于深度学习的方法。

- 基本方法：指直接在图像层面，通过一些基本的图像操作来生成新的数据。这类方法计算代价小，可以直接内嵌于模型训练过程中，用确定性的方式为模型训练提供更加多样化的训练数据。基本类方法又可以进一步细分为基础图像操作类、图像擦除类和图像混合类。
- 基于深度学习的方法：这类方法利用深度学习技术和模型来生成新数据，往往需要提前训练好生成式模型以保证数据的质量。该类别可进一步细分为自动增强类和特征增强类。

在这一新的分类体系下，基本方法强调低成本和实用性，而基于深度学习的方法则突出灵活性和适应性，二者共同构成了当下数据增强研究的全景格局。

2.1.1　基本方法类

基本方法类数据增强方法不依赖深度学习模型，而是直接在图像空间对原始数据进行变换。这类方法的核心思想是，用尽可能低计算代价的图像空间操作生成新数据。由于在数据生成阶段不涉及深度模型的训练，因此其计算开销远小于模型前向传播和反向传播的梯度计算过程，几乎可以忽略不计。正因如此，这类方法具有实现简单、计算复杂度低等优点，目前仍是最广泛使用的数据增

强手段之一。其主要目标是通过对图像的几何、纹理或内容进行操作，模拟现实场景中可能出现的多样模式，从而扩展训练数据的分布范围，提升模型的泛化能力。根据数据变换的形式，基本方法可进一步细分为三类[11]：图像变换、图像擦除[22,37-40]和图像混合[23-24,41-44]。

图像变换类方法，如旋转、翻转、裁剪等操作。这类技术的动机在于模拟现实场景中目标因视角、姿态、光照和拍摄环境变化而产生的多样性。例如，调整图像亮度可以模拟昼夜或光照强弱变化；旋转或翻转可以增强模型对目标方向和空间位置的鲁棒性。这类方法直接在像素层面操作图像，简单易行且能有效提升模型对不同输入场景的适应能力。

基于图像擦除的图像增强方法通常在图像中删除一个或多个子区域，其主要思想是将这些子区域的像素值替换为常量值或随机值。这类方法的动机是通过遮盖图中的一部分区域来增强模型对遮挡场景的鲁棒性，使得增强模型在失去一部分信息后，仍然具备在剩余信息中找到用于识别目标的显著性特征的能力。例如 Cutout[22]在训练卷积神经网络时随机掩盖掉输入的方形区域，去除输入图像的连续部分，有效地增加了现有样本的部分遮挡版本。Cutout 可以迫使模型更多地考虑完整的图像上下文，而不是依赖于少量特定视觉特征的存在。该方法不仅非常容易实现，而且它可以与现有的数据增强方法和其他正则化器结合使用，以进一步提高模型性能。Hide-and-Seek (HaS)[39]随机隐藏训练图像中的块，迫使网络在最有判别性的内容被隐藏时寻找其他相关的视觉内容。RandomErasing[38]在图像中随机选择一个矩形区域，用随机值或数据集的平均像素值替换其像素。训练不同遮挡程度的图像将有助于降低过拟合的风险，最终使模型更加鲁棒。随机翻转和随机裁剪也作用于图像层面，并与随机擦除密切相关，这两种技术都证明了具备提高图像识别精度的能力。避免连续区域的过度删除和保留是图像擦除方法的核心要求，一个成功的图像擦除方法应该在删除和保留图像区域信息之间取得合理的平衡。GridMask[37]既不删除连续区域，也不随机选择方块，删除的区域是一组空间上均匀分布的方块，其密度和大小可以控制。因此 GridMask 很好地平衡了连续区域的删除和保留。FenceMask[45]通过增强遮挡块的稀疏性和规则性，克服了小目标增强的困难，显著提高了基线性能。FenceMask 的设计思路是进一步稀疏化和正则化遮挡块。除此之外，AdvMask[40]不是随机遮盖图中的部分子区域来获得增强数据，而是利用对抗攻击方法首先学习并获得图像

的关键像素，在数据增强阶段，通过遮盖一部分关键像素来迫使模型学习其它非关键像素并用其进行分类，这有效增强了模型的鲁棒性和泛化性能，使得模型可以在目标被遮挡或者部分遮挡的场景下，仍然可以进行正确分类。这些方法通过增强模型在“缺失信息”场景下的判别能力，有效提升了模型在遮挡、模糊及部分损坏情况下的鲁棒性。

近年来，图像混合数据增强受到越来越多的关注。这些方法主要通过将两幅或多幅图像或图像的子区域混合为一幅来完成。通过混合多个图像，或者混合多个图像中的感兴趣目标到一张图像中，可以模拟现实中可能出现的多目标场景，从而可以增强模型对这一场景的泛化性能。其中，配对样本方法[46]通过将训练集中随机选择的两幅图像合成一幅新图像来扩大数据集。利用从训练集中随机选取的两幅图像，可以从 N 个训练样本中生成 N^2 个新样本，即配对样本。Mixup[23]不只是对两幅图像的强度进行平均，而是对样本对和它们的标签进行凸组合。因此，Mixup 在数据增强和监督信号之间建立了一种线性关系，并且能够正则化神经网络，使其在训练样本之间具有简单的线性行为。CutMix[24]用另一幅图像的块替换掉被移除的区域，而不是简单地从训练集中移除像素或混合图像，CutMix 可以生成更自然的图像。图像的真实标签也根据组合图像的像素数量成比例混合。FMix (Feature Mixup)[47]是 Mixup 的一种扩展，其通过混合特征而不是样本来实现这一点。具体而言，FMix 通过在输入图像中加入随机的掩码，并将其与另一幅图像的相应区域混合，从而生成一个新的图像。AugMix[48]是一种混合随机生成的增强操作和使用 Jensen-Shannon 损失来增强一致性的数据处理技术，它首先将多个增强操作混合为多个增强链，每个增强链由随机选取的 1 3 个增强操作组成，然后使用 Jensen-Shannon 散度作为一致性损失，在同一输入图像的不同增强链中执行分类器的一致性嵌入，最后将多个增强链的结果以凸组合的方式混合在一起。因此，整个过程通常是将同一图像在不同增强链中产生的结果进行混合，然后再将增强链的结果和原始图像结合起来。You Only Cut Once (YOCO)[49]对一幅图像在高度或宽度维度上切割成两幅相等的图像，在每一块内独立进行相同的数据增强，然后将增强块拼接在一起，形成一幅增强图像。Manifold Mixup[50]提出了一个新的方法。通过在随机线性组合的隐藏层特征上进行 Mixup 来实现数据增强。具体来说，它在神经网络的隐藏层特征上插入一组 Mixup 层，这些层将来自不同输入样本的隐藏层特征进行线性混合，从而

创建新的隐藏层特征。这样做的目的是让模型在更广泛的特征空间上进行学习，从而提高其泛化能力。

2.1.2 基于深度学习方法类

与人工设计的数据增强方法不同，近年来研究者们开始探索自动化的数据增强策略搜索，以期在更少人工干预的前提下获得更优的性能。自动增强一直是深度学习研究的前沿领域，并得到了广泛的研究。AutoAugment[25]自动地搜索可以达到最佳分类精度的数据增强策略。具体来说，AutoAugment 由搜索算法和搜索空间两部分组成。搜索算法使用强化学习技术，包含控制器和训练算法两部分，控制器在每一步使用 softmax 来预测一个决策，并将预测作为输入嵌入到下一步的决策生成过程中，以寻找验证精度最高的最佳策略。搜索空间包含许多子策略，详细说明了各种增强操作和应用这些操作的幅度，例如旋转和平移以及这两种操作的概率和参数，而搜索算法就是要找到这些子策略的最优组合。然而 AutoAugment 需要耗费大量的时间搜索对应数据集上的最优策略，Fast AutoAugment[51]通过基于密度匹配的更有效的搜索策略来寻找最优的增强策略。该方法通过学习将增强数据作为训练数据的缺失数据点的增强策略来提高给定网络的泛化性能，在策略搜索阶段通过贝叶斯优化来利用和探索一系列推理时间增强来恢复那些缺失的数据点。与此同时，Population Based Augmentation (PBA) 方法进一步减少 AutoAugment 的时间成本，该方法通过生成非平稳的增强策略计划而不是固定的增强策略来实现，PBA 学习增强策略的时间表，而不是固定的策略，这种选择是 PBA 的效率提高的主要原因。RandAugment[26]通过在一个小的代理任务上对增强策略进行单独搜索，然后将搜索到的结果迁移到更大的目标任务上。Tied-Augment[52]是近年来提出的一个通用框架，该框架前向传播过程中产生同一图像的两个增强视图，除了分类损失之外，他们还增加了一个相似度项来增强两个增强视图的特征之间的不变性，通过特征相似性来增加数据增强的有效性。传统的自动增强通常应用多个图像变换操作来生成数据，相比于之前的方法强调数据的多样性，在 TrivialAugment[53]中，每张图片只使用一次数据增强方式。因此相比于 AutoAugment、PBA 乃至 RandAugment，它的搜索成本几乎是可以忽略的。除此之外，EntAugment[54]提出了一种自适应数据增强的框架，通过关注数据增强的强度，而不是具体的操作，来调整训练过程中数

据的变换幅度。AutoDA[55]使用一个识别模型来学习每个数据实例的底层增强策略，并通过一个交替的利用和探索过程来使用可微的工作流更新增强策略。在数据增强阶段，AutoDA 训练一个分类器，经过若干个步骤，然后通过探索阶段对分类器进行验证，并更新策略以最小化验证损失。另一类思路是对抗数据增强优化方法，例如 TeachAugment[27]通过利用教师模型，可以在不需要复杂参数调整的情况下生成信息丰富的转换图像。具体来说，对增强策略进行搜索，使增强图像对目标模型具有对抗性以及对教师模型具有可识别性。除了通过自动化的方式直接来搜索增强策略，自动增强方法类中的方法还被用于改善其它类方法中存在的潜在问题，例如 KeepAugment[56]利用显著性特征图来检测原始图像上的重要区域，然后在增强过程中保留这些信息区域不受干扰，从而保持语义稳定性。OHL-Auto-Aug[57]将增强策略建模为参数化的概率分布，分布的参数被视为超参数，进一步提出了一个双层框架，允许在网络训练的同时优化分布参数。SelectAugment[58]根据样本内容和网络训练状态，以确定的和在线的方式选择待增强的样本。具体来说，在每个批次中，首先确定增强比例，然后在该比例下决定是否对每个训练样本进行增广。他们将这一过程建模为两步马尔可夫决策过程，并采用分层强化学习来学习增强策略。通过这种方式，可以有效地缓解随机选择样本进行增广所带来的负面影响，提高数据增强的有效性。

为确保文献综述的全面性，本章不仅涵盖基于图像空间的数据增强方法，还系统性地纳入了特征空间增强技术的研究进展。这类方法不是在原始像素空间中进行操作，而是直接在神经网络学习得到的特征空间中生成或变换数据。例如，FeatMatch[59]设计了一个模块，该模块通过对从数据集中其他图像的特征中提取的一小组代表性原型进行软注意力来学习改进和增强输入图像特征，由于所提出的模块是在特征空间中学习和执行的，因此可以使用十分多样的数据转换方法。MoEx[60]，矩交换方法，通过鼓励模型利用潜在特征的矩信息，将一幅训练图像的学习特征的矩替换为另一幅训练图像的可学习特征的矩，并对目标标签进行插值迫使模型从这些矩中提取训练信号和归一化特征。但是总的来说，特征空间数据增强的计算成本也可能很高，因为它需要对输入数据进行复杂的变换和扩充。如果数据集很大，这可能会带来较大的训练代价，所以特征增强方法较少应用在实际的训练过程中。

综上，基于深度学习的方法通过策略搜索、自适应调节和特征空间建模等手

段，显著扩展了数据增强的能力谱系。它们在提高模型泛化性能和鲁棒性方面展现出巨大潜力，同时也暴露出计算成本较高、跨任务迁移性不足等问题，为后续研究提供了新的方向。

2.2 数据选择

数据选择（Data Selection）是近年来机器学习与深度学习领域中快速发展的重要研究方向[31,33,61-63]，其核心思想是在训练开始之前，预先从大规模原始数据集中挑选出一个更小但更具代表性的核心数据子集，用于模型训练。通过去除冗余和低质量样本，数据选择不仅能够显著降低存储与计算开销，还能减少噪声对模型训练的不良影响，从而提升模型的训练效率与稳定性。已有研究表明，即使只使用经过合理挑选的数据子集，模型在许多基准任务中的性能也可以接近甚至超过使用完整数据集的效果。与数据增强研究通过对现有样本进行变换来“扩充”数据规模不同，数据选择的思路是通过“压缩”来提升数据效率。这两类方法的出发点截然不同：数据增强强调在样本空间引入更多的多样性，从而提升模型对不同输入的泛化能力；而数据选择则关注如何在保证代表性和信息量的前提下，去除冗余和噪声样本，减少无效计算。前者在一定程度上缓解了模型过拟合风险，后者则更直接地面向训练效率优化。二者从不同角度提升了数据利用效率，在实际应用中具有天然的互补性。

作为提升数据高效训练的核心技术之一，在过去几年中，大量工作围绕数据选择展开研究，并逐渐形成了若干方向：静态数据选择[31,62,64]、动态数据剪枝[33,65-66]、和数据集蒸馏[67-70]。数据选择研究主要包含前两部分，本章为了保证全面性，系统性纳入了数据集蒸馏研究方向的讨论。

- 静态数据选择：在训练开始之前，根据预定义的重要性度量、分布覆盖指标或优化目标，从完整数据集中筛选出一个固定的高价值子集用于训练。这类方法的优势在于显著降低存储与训练成本，并在多种场景中实现接近甚至超越全数据集的性能。然而，由于子集在训练前已固定，其在跨任务或跨架构的泛化性可能受到一定限制。
- 动态数据选择：不同于静态方法在训练前一次性完成样本选择，动态方法在训练过程中实时更新选择结果[33,65-66]。其核心思想是根据训练状态或梯

度信息，动态识别和保留最具价值的样本，从而在保持模型性能的同时进一步减少无效计算。这类方法往往能够自适应不同的训练阶段，但需要额外的在线计算开销。

- 数据集蒸馏：与前两类方法通过“选择”子集不同，数据集蒸馏试图直接“合成”一个小规模的代理数据集，使得在该合成数据集上训练的模型能近似于全量数据的训练效果[70-72]。这种方法通常依赖生成式建模或优化技术，能够在极端压缩条件下保持性能，因此在超大规模数据场景中展现出巨大潜力。

2.2.1　静态数据选择

静态数据选择在训练开始之前，通过预定义的评价指标或优化目标，从完整数据集中筛选出一个固定的子集用于训练。其核心思想是：在不改变模型架构和训练流程的情况下，通过减少训练样本规模来降低计算与存储开销，同时尽可能保持或提升模型的泛化性能。由于选出的子集在训练前即已固定，这类方法通常计算成本较低，易于应用到大规模训练任务中，因此成为数据高效训练最早得到广泛研究的方向之一。静态数据选择主要可分为三类：基于样本重要性度量的方法、基于数据分布的方法以及基于优化的方法。

基于样本重要性度量的方法会为每个样本计算一个“重要性分数”，并选取分数较高或者较低的样本构成核心集。例如，Forgetting[30]方法会在训练中追踪样本预测从正确到错误的迁移次数，并将这种迁移定义为“遗忘事件”，最终会选择那些被频繁遗忘的样本来训练。EL2N 和 GraNd[64]分别基于样本预测误差向量的 ℓ_2 范数和梯度范数来度量样本的重要性，具体地，给定一个样本 (x, y)，在 t 时刻下，GraNd 分数定义为

$$\chi_t(x, y) = \mathbb{E}_{\mathbf{w}_t} \left\| g_t(x, y) \right\|_2, \tag{2-1}$$

其中 $g_t(x, y) = \nabla_{\mathbf{w}_t} \ell \left(p\left(\mathbf{w}_t, x\right), y \right)$ 是梯度范数,$\mathbf{w}_t$ 是。同时，给定一个样本的 EL2N 分数定义为误差向量的 ℓ_2 范数：$\mathbb{E} \left\| p\left(\mathbf{w}_t, x\right) - y \right\|_2$。Glister[73]是一种基于泛化能力的数据子集选择框架。该框架通过一个混合离散-连续的双层级优化问题，将数据选择过程形式化，其目标是选取能够最大化留出验证集上对数似然值的训练

数据子集。近几年来，CG-Score[74]提出了一种复杂度间隙评分（Complexity-Gap Score）来评估单个样本的影响力，该指标通过量化样本的不规则性，衡量每个数据实例对网络训练整体提升的贡献程度。Memorization[75]进一步衡量了删除或者保留每个样本对模型最终预测能力的影响，从而识别关键样本。MoSo[31]通过评估每个样本对最优经验风险的影响程度来确定其重要性，这种影响程度是通过衡量特定训练样本被排除出训练集时，经验风险的变化幅度来实现的。针对网络规模数据训练可能耗时数月的问题，Rho-Loss[76]提出了可减少验证集损失选择（Reducible Holdout Loss Selection, Rho-Loss）方法。传统优化方法通常选择” 困难”（如高损失）样本，但这些样本往往包含噪声或与任务相关性较低；而课程学习方法优先选择” 简单” 样本，但这些样本一旦学会后继续训练的价值有限。相比之下，Rho-loss 能够选择兼具可学习性、学习价值且尚未被充分掌握的样本。TDDS[77]提出了一种新颖的时序双重深度评分（Temporal Dual-Depth Scoring, TDDS）方法。该方法采用双重深度策略，在整合广泛训练动态与识别代表性样本之间实现平衡。在第一重深度中，其估计每个样本在整个训练过程中个体贡献的时间序列，确保训练动态的全面整合；在第二重深度中，聚焦于第一重深度中识别出的样本贡献的变异性，从而凸显具有良好泛化能力的样本。

基于数据分布的数据选择方法强调从整体上覆盖原始数据的分布特性，以保证子集的代表性和多样性。其核心是假设是：如果子集能较好地近似原始分布，那么在该子集上训练的模型也能保持较好的泛化性能。例如，Herding[78] 根据样本与类别中心的距离选择距离所属类别中心更近的样本作为最具代表性的样本。Greedy k-Center[79]使用贪心算法寻找能最大化覆盖率的核心集。D2 剪枝算法[80]将数据集表示为无向图，该算法通过在数据集图上进行前向与反向消息传递来实现核心集选择：首先通过融合邻域样本的难度信息来更新每个样本的难度评分，继而基于这些更新后的难度评分指导图采样方法，最终选出能够覆盖数据空间中多样且困难区域的核心集。Moderate-DS[62]提出一种” 适度核心集” 的概念。具体而言，给定任何数据选择的评分准则，不同场景倾向于选择不同得分区间的数据点。由于评分中位数是统计学中评分分布的代理指标，那些得分接近评分中位数的数据点可视为完整数据的代理样本，能够泛化到不同场景，因此被用于构建适度核心集。与此不同，Coverage-centric Coreset Selection (CCS)[81]从理论和实证的角度深入分析了当前核心集选择方法在高剪枝率下性能骤降的问

题。传统的一次性核心集选择方法往往依赖于重要性度量，在低剪枝率时表现良好，但在高剪枝率（例如 90%）下，往往会遭遇比随机选择还差的灾难性退化，其根本原因在于这些方法忽视了数据分布的整体覆盖性。为解决这一问题，CCS 将经典的几何集覆盖问题扩展为分布覆盖问题，提出了一种新的覆盖度量指标来衡量子集对整体分布的覆盖程度。在此基础上，CCS 设计了一种新的选择策略，同时考虑样本的重要性和整体分布覆盖性，从而在高剪枝率下依然保持较强的鲁棒性和准确性。

基于优化的方法将数据选择问题显式建模为一个优化过程，往往能在理论上保证选择的有效性与稳定性。Dataset Pruning[61] 是一种典型的优化驱动样本选择方法，其出发点是回答深度学习中的一个基本问题：并非所有训练样本对模型的泛化能力贡献相同。通过量化“移除某一部分训练样本对模型泛化性能的影响”，从理论上保证所构建的子集能够维持严格受限的泛化差距。换句话说，这个工作不仅评估单个样本的重要性，更强调在全局优化意义下寻找最优子集。Beyond[82] 的工作关注深度学习中广泛存在的神经网络缩放律（Neural Scaling Laws）。其核心思想是：如果能够找到一个高质量的数据剪枝度量指标，用于训练样本进行排序并在任意数据比例下选择最优子集，就有可能突破幂律缩放，甚至接近指数级缩放。在理论层面，作者证明了在存在理想剪枝度量的情况下，误差缩放律可以显著优于传统幂律。PRISM[83] 提出了一个参数化次模信息度量（PaRameterIzed Submodular Information Measures，PRISM）的通用框架。与传统的核心集选择方法仅关注代表性或多样性不同，该框架通过设计一类可参数化的次模函数，能够灵活建模不同任务场景下的目标需求，从而更好地支持引导式数据子集选择的实际应用。FASS[84]研究了在大规模数据场景下，如何在保持分类器性能的同时最小化训练数据子集规模的问题。作者首先建立了次模函数与数据似然函数（例如朴素贝叶斯和最近邻分类器）之间的联系，并将数据子集选择问题形式化为约束下的次模最大化问题。在此基础上，FASS 将不确定性采样与次模选择框架结合起来，提出了一种过滤式的主动次模选择方法。该方法在子集选择时既考虑了次模覆盖性，也利用了主动学习中的不确定性信息，以增强所选样本对模型训练的价值。Selective-Backprop[85]提出了一种基于损失驱动的动态样本选择策略，其核心思想是在每一次迭代中优先选择高损失样本进行反向传播。具体而言，Selective-Backprop 利用前向传播的结果来判断当前

样本是否足够重要，如果样本的损失较低，则跳过其反向传播，从而直接进入下一个样本。这样做的直接好处是显著减少了反向传播这一计算代价最高步骤的执行次数。实验证明，该方法在 CIFAR-10、CIFAR-100 和 SVHN 等多个数据集和不同的现代深度模型上均可实现加速效果，证明了其降低训练成本的潜力。RL-Selector[32]提出了一种基于强化学习的动态数据选择方法，旨在克服现有方法只依赖静态评分或预训练模型的局限。其核心创新在于引入 ϵ-sample cover 的概念，用于量化样本间的冗余关系，从而更好地刻画数据集的内在结构。与传统单一评分指标不同，ϵ-sample cover 通过建模样本之间的相互覆盖关系来衡量样本的代表性和多样性，使得选择过程能够避免重复选择大量冗余样本。基于这一指标，RL-Selector 将数据选择问题重新表述为一个强化学习过程。具体而言，它设计了一个轻量化的强化学习代理，通过不断与动态演化的数据分布交互，利用 ϵ-sample cover 作为奖励信号来优化选择策略。这样，样本选择不再是一次性的静态操作，而是一个与训练同步更新的动态过程，能够更好地适应数据分布和模型学习状态的变化。

2.2.2 动态数据选择

与静态数据选择在训练前一次性确定子集不同，动态数据选择方法强调在训练过程中根据模型的学习状态持续更新所选样本。其核心思想是：模型在不同阶段对样本的需求并不相同，早期可能需要更多代表性和多样性样本以快速建立全局认知，而在后期则更倾向于关注难例或边界样本以精细化决策。因此，动态方法通过在线样本选择或剪枝机制，在每一训练阶段动态调整数据子集，从而在不显著降低性能的前提下，有效减少训练计算开销。动态数据选择通常具备两个关键特征：1）与模型状态的联动性：不同于静态方法依赖于固定评分或先验度量，动态方法往往结合训练过程中的梯度、损失或预测变化，来刻画样本的重要性和学习价值。2）动态子集更新：样本的选择结果会随着训练迭代不断演化，避免了静态方法在容易出现的多样性损失或过拟合问题。两种动态数据剪枝方法 UCB 和 ϵ-greedy 选择得分最高的特定比例样本[65]，并在该周期内仅基于这些样本进行训练。InfoBatch[33]通过无偏动态数据剪枝实现无损训练加速。具体而言，InfoBatch 基于损失分布随机剪除部分低信息量样本，并重新缩放剩余样本的梯度以逼近原始梯度。作为即插即用且架构无关的框架，InfoBatch 在分类、

语义分割、视觉预训练和大模型指令微调任务中持续获得无损训练结果。He 等人[86]利用预测不确定性与训练动态信息来衡量样本的重要性。具体而言，研究者通过跟踪样本在整个训练过程中预测结果的变化情况，来评估样本对模型学习过程的贡献度。变化幅度较大的样本往往包含更高的信息量，因此在子集构建中被优先保留。与传统依赖静态打分指标或单次训练信号的方法不同，这种方式能够充分利用训练过程中积累的动态信息，因而更具鲁棒性与代表性。

2.2.3　数据集蒸馏

与数据选择不同，数据集蒸馏并非直接筛选原始数据样本，而是通过合成高度压缩的代理数据集来捕捉原始数据的核心特征。这种方法旨在使用极少量的人工合成样本替代海量原始数据，使得模型在合成数据上训练后能达到与原始数据相当的泛化性能。其核心挑战在于如何通过优化算法将原始数据集的信息密度最大化地编码到有限数量的合成样本中，从而突破传统数据选择方法在信息保留上的理论极限。该研究的代表性方法主要包括：

- 梯度匹配（Gradient Matching）：以[67]为例，通过最小化真实数据和蒸馏数据在梯度上的差异来生成合成样本，使模型在小数据集上训练的更新轨迹尽可能接近在完整数据集上的轨迹。
- 元学习（Meta-Learning）：如 MTT[69] 将蒸馏问题建模为一个双层优化问题，在“元层面”优化合成数据以提升模型在真实数据上的性能。
- 生成模型驱动（Generative Models）：如使用 GAN 或 Diffusion 模型生成可替代真实数据的代理数据，从而提升蒸馏样本的多样性和逼真度[70,72]。
- 数据凝练（Dataset Condensation）：进一步扩展了蒸馏思路，利用核函数、对比学习或隐空间建模来生成更加结构化和任务相关的合成数据[68,87-88]。

2.3　本章小结

本章系统回顾了数据增强与数据选择两条主要技术路线，梳理了其发展脉络与代表性方法。对于数据增强，我们从基本方法和深度学习方法两大类展开，分别总结了图像变换、擦除、混合、自动搜索增强以及特征增强等不同策略，并分析了它们在提升数据多样性、缓解过拟合和增强鲁棒性方面的作用。对于数据

选择，我们从静态选择、动态选择和数据集蒸馏三个角度进行了综述，讨论了基于重要性度量、分布覆盖和优化建模的不同方法，并进一步延伸到训练过程中动态剪枝及小规模代理集的构建。总体来看，数据增强和数据选择分别从“扩展”与“压缩”两个方向提升了数据价值，二者既有互补性也存在潜在的耦合关系。

第三章　基于相似性与多样性的数据分析框架研究

3.1　引言

深度学习的发展依赖于大规模训练数据，然而，不同数据集在规模、信息密度、和分布特征等方面存在显著差异。直接将统一的训练策略应用于不同数据集，往往会导致训练效率不佳，甚至削弱模型的泛化能力。已有研究表明，不同数据集、模型结构以及训练阶段对于相似性和多样性的偏好并不一致[89-91]。因此，从以数据为中心的人工智能角度出发，理解训练数据的特性如何影响模型性能，并建立这种关系的系统性分析框架，是实现高效与鲁棒学习的关键。

然而，虽然 Inception Score（IS）[92]和 Frechet Inception Distance (FID)[93]等评价指标已被广泛用于评估生成模型所生成图像的质量[94]，但这些指标主要关注生成图像的视觉质量和真实感。然而，对于训练数据而言，关注点并非图像的视觉效果，而是用于训练时的性能。例如，GridMask[37]通过删除图像中均匀分布的方形区域来生成增强图像用于训练；Mixup[23] 则通过对输入与标签成对进行凸组合来生成虚拟训练样本用于训练。尽管这些方法在提升模型性能上效果显著，但生成的样本并非自然场景的模拟，视觉上也往往缺乏意义，因此用 IS 或 FID 之类的指标来评估其质量是不合理的。由此可见，深度生成模型与数据增强方法在评估目标上存在根本差异。近期,[95]首次提出通过亲和性与多样性的实证研究来分析训练数据，但是这两项指标计算依赖于模型训练。由于它们基于训练结果进行度量，不仅可解释性不足，而且在计算开销上几乎与获得最终测试精度等价。因此，亟需建立一种新的量化标准，用于系统性地分析和利用数据特性，以获取更高效、更稳定的模型性能。

为解决上述问题，本章的研究目标是构建一个基于相似性（Similarity）与多样性（Diversity）的训练数据表征与分析框架，该框架通过两个与模型训练过程无关的量化指标，对训练数据的分布属性进行刻画，并揭示数据特性与模型性能之间的内在联系。具体而言，特别是，训练数据的相似性刻画训练数据和原始

分布之间的一致性，有助于避免因分布漂移导致的性能退化；多样性反映样本在特征空间的覆盖和差异性，保证模型能够学习到丰富的知识以提升泛化能力。与传统只针对特定方法的评价不同，本框架具有普适性：它不仅能够为训练数据的特性提供解释，也能够推广到核心数据集构建与动态数据优化场景中，从而为后续方法研究提供统一的分析空间。

本章的主要贡献如下：

- 提出了一套与模型训练无关的训练数据相似性-多样性度量体系，从相似性和多样性两个角度系统性地分析训练数据的有效性，为训练数据的分析和优化提供了统一框架。
- 通过系统对比量化结果与实际任务中的表现，本节揭示了相似性与多样性的作用因数据集而异，说明数据的有效性并非由单一维度决定，而是取决于二者的平衡。
- 本章的研究不仅加深了对数据机制的理解，也为增强方法的设计与参数调优提供了参考。同时，该框架能够作为高效的前置验证工具，降低大规模模型训练中的计算与时间成本。

更重要的是，本章提出的相似性-多样性框架为后续研究奠定了统一的理论基础和可解释的依据。接下来的章节将分别在此框架下展开：第四章利用强化学习机制实现自适应的数据增强来提升数据多样性；第五章将框架扩展到数据相似性优化场景，提出基于多模态学习的数据选择方法来保证数据质量；第六章则进一步探讨选择与增强的协同关系，构建面向高效训练的新型数据训练范式。

3.2　基于相似性与多样性的训练数据优化框架

3.2.1　数据优化的一般目标函数

在开始讨论之前，首先澄清两个概念，训练数据集和训练数据。在本文中，训练数据集（Training Dataset）或训练集指的是研究者在实验开始时就已经获取到的完整数据集合。它通常由多个样本及其对应的标签构成。例如，CIFAR-10[96]、ImageNet-1k[14]等数据集均可视为训练数据集。训练数据集在整个学习过程中保持不变，起到“数据源”的作用。在本文中，训练数据（Training Data）指的是在

模型在线训练的某一时刻实际被采样且施加变换后输入模型的那部分数据。因此，它是动态的、随训练进程而不断变化的。例如，在使用数据增强方法时，模型在某一迭代中接收到的训练数据可能与原始训练数据的分布有所不同[95]。这一区分在研究数据优化问题时尤为重要：在传统机器学习中，研究者常常将“训练数据集”和“训练数据”混用，或者定义训练数据构成训练数据集。而在数据优化问题中，本节更强调如何在训练过程中动态地调控训练数据的构成，从而影响模型的收敛速度与泛化性能。

为了更加系统地刻画训练数据对模型学习过程的影响，本文引入从数据本身出发的优化目标，其核心思想是：通过合理地调控训练数据的组成与使用方式，提高模型的学习效率与泛化能力。形式化地，可以将这一问题建模为以下期望风险最小化过程：

$$\min_{\Theta} \mathbb{E}_{i\sim p_\theta, T\sim q_\phi} \mathcal{L}\left(f_\Theta\left(T\left(x_i\right)\right), y_i\right), \tag{3-1}$$

其中，Θ 表示待学习的模型参数，p_θ 表示训练数据的原始分布，因此可以通过调整 p_θ 来决定哪些样本进入训练，q_ϕ 表示数据变换或生成分布，T 是从 q_ϕ 中采样的一种数据变换方式，用于控制实际用于训练的样本的增强方式，$\mathcal{L}(\cdot)$ 为任务相关的损失函数。这一目标函数提供了一种统一建模训练数据与模型学习过程关系的形式化表达，同时涵盖了原始训练数据的分布和在线训练过程中训练数据的变换。无论是通过调节已有数据的使用方式，还是结合不同的数据调控机制，这一框架都能加以涵盖。其核心思想在于：将训练数据的分布性因素显式引入到优化目标之中，从而系统地分析和解释数据特性对模型性能的影响。

3.2.2 相似性-多样性分析框架

为了深入探究不同训练数据特性对模型性能的影响，本节提出以相似性和多样性作为训练数据的核心度量。这两个维度不仅能够为已有方法的性能差异提供解释，还能为不同数据场景提供一个统一的分析空间，从而系统地刻画训练数据对模型学习行为的影响机制。由于原始训练数据集本身通常是固定的，需要借助数据增强（Data Augmentation，DA）来人为地调节训练数据的相似性和多样性，从而构造出覆盖全谱系的数据配置，以支持后续分析。下面本节首先介绍

相似性的定义与实现。

相似性度量

（1）背景知识　最优运输（Optimal Transport, OT）是一种比较概率分布的经典方法[97], 用于研究如何以最小代价将一个分布“搬运”到另一个分布。由于其在概率论、最优化以及机器学习中的广泛应用，OT 已成为衡量分布差异的重要工具。

在形式化定义中，OT 考虑两个概率分布 α 和 β。Kantorovich 形式的最优传输问题[98]可描述为：在所有同时满足边际分布为 α 和 β 的联合分布中，寻找一个传输方案，使得整体的搬运代价最小:

$$\mathcal{L}_c(\alpha,\beta) \triangleq \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} c(x,y)\mathrm{d}\pi(x,y), \tag{3-2}$$

其中，$c(x,y)$ 是一个代价函数，用来衡量 x 和 y 之间的距离；$\mathcal{U}(\alpha,\beta)$ 表示所有边际分布分别为 α 和 β 的联合分布集合：

$$\mathcal{U}(\alpha,\beta) \triangleq \left\{\pi \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y}) : P_{\mathcal{X}\sharp}\pi = \alpha, P_{\mathcal{Y}\sharp}\pi = \beta\right\}, \tag{3-3}$$

其中，$P_{\mathcal{X}\sharp}\pi = \alpha$ 和 $P_{\mathcal{Y}\sharp}\pi = \beta$ 分别是联合分布 π 在 $P_{\mathcal{X}}(x,y) = x$ 和 $P_{\mathcal{Y}}(x,y) = y$ 投影下的推前测度[98]。该问题的最优解被称为最优传输计划。在实际应用中，概率测度通常未知，而图像数据集由有限个离散样本构成。因此，可以将 α 和 β 定义为离散分布：$\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{\mathbf{x}^{(i)}}$ 和 $\beta = \sum_{i=1}^{m} \mathbf{b}_i \delta_{\mathbf{y}^{(j)}}$，其中，$\mathbf{a}$ 和 $\mathbf{b}$ 为概率单纯形中的向量，$\mathbf{x}^{(i)} \in \mathcal{X}$，$\mathbf{y}^{(j)} \in \mathcal{Y}$，$\delta_{\mathbf{x}^{(i)}}$ 和 $\delta_{\mathbf{y}^{(j)}}$ 分别是以 $\mathbf{x}^{(i)}$ 与 $\mathbf{y}^{(j)}$ 为中心的 Dirac 测度[99]。为了提升计算效率和稳定性，经典 OT 问题常通过加入熵正则化项进行平滑化，得到熵正则化最优传输问题：

$$\mathrm{OT}_\epsilon(\alpha,\beta) \triangleq \min_{\pi\in\mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{H}(\pi \mid \alpha\otimes\beta), \tag{3-4}$$

其中，$c(x,y)$ 表示样本对之间的代价，ϵ 是正则化系数，$\mathrm{H}(\pi \mid \alpha\otimes\beta)$ 表示相对熵，即 $\mathrm{H}(\pi \mid \alpha\otimes\beta) = \int \log(\mathrm{d}\pi/\mathrm{d}\alpha\mathrm{d}\beta)\mathrm{d}\pi$。该问题可通过 Sinkhorn 算法[100] 高效求解，从而使 OT 能够应用于大规模数据集之间的比较。

（2）基本定义 深度学习的一个基本假设之一是，训练集和测试集服从相同或相近的分布，即二者中的所有样本为独立同分布（i.i.d.）样本，且测试集是符合真实世界分布的。在这一假设下，若训练数据与真实分布存在过大偏离，模型的泛化能力将受到显著削弱。因此，度量不同训练数据集合之间的分布接近程度成为训练数据优化的首要问题。为了量化这一分布差异，本节采用最优运输（Optimal Transport）方法[97,101]。最优运输理论提供了一种较为严格且具备良好理论性质的分布距离度量框架，通过建立数据点之间的配对成本来衡量两个分布之间的差异。聚焦到有监督学习情境下，假设数据集是由一个联合分布中采样得到，记作 $\boldsymbol{D}=\{(\boldsymbol{x},y)\}\sim P(\mathcal{X},\mathcal{Y})$，其中 $\boldsymbol{x}$ 和 y 分别是原始图像特征及其标签。为了比较不同数据集的分布接近性，本节将相似性定义为不同数据集在特征-标签联合分布上的接近程度：

$$d((x,y),(x',y'))=\left(d_{\mathcal{X}}\left(x,x'\right)^{p}+d_{\mathcal{Y}}\left(y,y'\right)^{p}\right)^{1/p}, \tag{3-5}$$

其中定义 $d_{\mathcal{X}}\left(x,x'\right)^{p}$ 为特征空间中的欧式距离，而 $d_{\mathcal{Y}}$ 则度量标签之间的差异。与仅在输入特征空间度量不同，本节进一步显式引入标签差异度量 $d_{\mathcal{Y}}(y,y')$，以更准确反映监督学习中的判别信息。

（3）标签距离的建模 传统方法仅考虑输入图像特征分布的相似性，本章显式地引入了标签层面的差异度量 $d_{\mathcal{Y}}(y,y')$ 来构建联合分布。这是因为在有监督学习任务中，标签蕴含着更为直接的判别信息，而忽略标签差异将导致对分布间距离的估计有偏。举例来说，类别之间往往存在不同尺度的差异。对于细粒度类别，如金毛犬和哈士奇，它们在特征空间中的差异可以视为类内差异；而对于完全不相干的类别，如热狗和狗，则表现为明显的类间差异。这两类差异在特征空间中具有显著的尺度区别。基于此，在式(3-5)中同时建模了类别间的分布距离，以更全面地捕捉类间复杂关系。

基于类条件分布来建模标签距离。具体而言，对于标签 y，考虑其在特征空间上的条件分布：$C_y(X)=P(X|Y=y)$. 令 $X_y=\{x\mid(a,b)\in(\mathcal{X}\times\mathcal{Y}),b=y\}$，其中 X_y 为属于类别 y 的特征集合。通过这种方式，可以将标签间的距离转化为分布间的距离，进而可通过 p-Wasserstein 距离计算：$\mathrm{W}p^{p}\left(C_y,C'_{y'}\right)$。该度量不仅考

虑了特征的均值差异，还能捕捉类内分布和尺度差异，从而更准确地反映标签间的真实距离。因此，式(3-5)可以进一步表示为：

$$d((x,y),(x',y')) = \left(d_{\mathcal{X}}\left(x,x'\right)^{p} + \mathrm{W}_{p}^{p}\left(C_{y},C_{y'}'\right)\right)^{1/p}, \tag{3-6}$$

其中，$C_{y'}'$ 是标签 y' 的增强数据集上的条件概率分布。

（4）最优运输相似性度量　设原始训练数据集与实际变换后的用于训练的数据分别为 $\mathcal{D}$ 与 $\mathcal{D}_t$，则最优运输距离可定义为：

$$d_{\mathrm{OT}}\left(\mathcal{D}_t,\mathcal{D}\right) = \min_{\pi\in\mathcal{U}(C,C')}\int_{\mathcal{Z}\times\mathcal{Z}} d_{\mathcal{Z}}\left(z,z'\right)\pi\left(z,z'\right), \tag{3-7}$$

其中 $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$ 表示特征-标签联合空间，π 为耦合分布，$d_{\mathcal{Z}}$ 度量联合样本间的差异，$\mathcal{U}$ 由式(3-3)中定义。为了提高计算效率，引入熵正则化的 Sinkhorn 算法可快速求解[100,102]。据此，定义相似性度量为最优运输距离的负值：

$$similarity(\mathcal{D}_t,\mathcal{D}) = -d_{\mathrm{OT}}\left(\mathcal{D}_t,\mathcal{D}\right), \tag{3-8}$$

基于上述定义，较小的 OT 距离意味着两者具有更高的相似性，从而使得相应的数据增强方法的相似性度量值更接近于 0。在这种情况下，增强数据的分布与训练集的分布更加接近。基于训练集与测试集服从相同分布的假设，因为相似性度量值为负数，则该值更高的增强数据集更有可能逼近测试集，从而提升深度模型的性能。因此，本文提出的相似性度量在理论上与深度模型的最终泛化性能存在紧密关联。

3.2.3　多样性度量

当 $\mathcal{D}$ 和 $\mathcal{D}_t$ 完全一致，最优运输距离为零，此时相似性指标达到最大值。然而，在训练数据优化的目标下，最高的相似性往往不是最理想的。数据虽然完全贴合原始分布，但缺乏多样性，容易导致模型在训练过程中发生过拟合。这一点在许多数据增强的研究中也有体现：例如，若候选训练数据仅是对原始训练集的简单重复，虽然相似性指标最高，但这种“无差异”的数据集合并不能为模型提

供新的学习信号，反而可能削弱泛化性能。因此仅依赖相似性不足以全面刻画训练数据的有效性。为弥补这一不足，本章提出多样性度量，用于衡量训练数据在类别覆盖性和结构差异性上的表现。相似性度量了数据偏离目标分布的程度，而多样性度量了数据是否具备足够的覆盖与差异，两者互为补充，从而避免过拟合并提升模型的泛化能力。

在生物多样性和统计学研究中，生物多样性常被定义为“物种层面的差异性”的度量指标[103]，受此启发，类比到图像数据集，我们认为不同类别所包含的信息具有独特性，因此将多样性表征为类别间特征分布的非相关性。其次，多样性的计算仅基于图像中的有效成分[104]。从统计学视角来看，随机变量的协方差与其相关性直接相关[105]。在主成分分析（PCA）中，协方差矩阵的特征向量 $\mathbf{u}$ 代表数据的主方向，而对应的特征值 λ 则表示数据点沿该方向的差异程度。协方差矩阵的特征值和特征向量能够有效表征成分的非相关性[105]。

（1）基本定义 设某类别 k 的样本特征为 $\mathbf{A}_k \in \mathbb{R}^{m\times n}$，其中每列 $\boldsymbol{a}_i$ 表示一个样本的特征向量，m 为特征维度，n 为该类别的样本数。通过这种方式，可以将同类样本视为随机变量，并通过协方差矩阵刻画该类别分布的扩展性与相关性。本章将归一化特征矩阵的经验协方差矩阵 $\mathbf{S}$ 构造如下：

$$a_{ij}^{*} = \frac{a_{ij} - \overline{a_i}}{\sqrt{s_{ii}}}, i = 1, 2, ..., m; j = 1, 2, ..., n, \tag{3-9}$$

$$\mathbf{A}^{*} = [a_{ij}^{*}]_{m\times n}, \tag{3-10}$$

$$\mathbf{S} = \frac{1}{n-1}\mathbf{A}^{*}\mathbf{A}^{*T}, \tag{3-11}$$

其中，

$$\overline{a_i} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}, i = 1, 2, ..., m, \tag{3-12}$$

$$s_{ii} = \frac{1}{n-1}\sum_{j=1}^{n}(a_{ij} - \overline{a_i})^2, i = 1, 2, ..., m. \tag{3-13}$$

对协方差矩阵进行特征分解，有 $\mathbf{S} = \sum_{i=1}^{m} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$，其中 λ_i 为特征值，$\boldsymbol{u}_i$ 为对应的特征向量。特征向量 $\boldsymbol{u}_i$ 表示有效样本的一个方向，其对应的特征值 λ_i 表征该

方向上样本的显著程度[104]。

（2）多样性度量公式　由于训练数据本身的多样性取决于原始数据集的多样性，将类别 k 的原始训练数据 $\mathbf{A}_k$ 与实际用于训练的变换后的数据合并为 $\mathbf{A'}_k$，在多样性度量中，重点关注原始数据集和变换后用于训练的数据集之间的有效样本的非相关性，可以理解为 $d(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{train+aug}})$。分别对 $\mathbf{A}_k$ 和 $\mathbf{A'}_k$ 进行特征分解，得到 $\mathbf{u}_k$、λ_k 和 $\mathbf{u'}_k$、λ'_k。为避免噪声干扰，将特征值及对应特征向量按降序排列，仅保留前 t 个最大特征值及相关特征向量进行多样性分析，其中 t 根据以下方式确定:

$$t = \arg\min_{m'} \frac{\sum_{i=1}^{m'} \lambda_{ki}}{\sum_{i=1}^{m} \lambda_{ki}} \geqslant \theta. \tag{3-14}$$

这里 θ 是一个超参，m 表示 $\mathbf{A}_k$ 的特征值总数。数据集的主要信息包含在最大的 t 个特征值及其对应的特征向量中，而较小特征值对应的特征向量则与噪声相关[106-107]。最终，定义跨所有类别的平均多样性度量为：

$$diversity = \frac{1}{k} \sum_{i=1}^{k} ||\lambda_i^* \cdot \boldsymbol{u}_i^* - \lambda'^*_i \cdot \boldsymbol{u}'^*_i||_2^2, \tag{3-15}$$

其中 k 为类别数，λ_i^* 和 $\boldsymbol{u}_i^*$ 为前 t 个主成分及其特征向量。

多样性与模型性能之间也并非简单的单调关系，当候选数据与原始数据完全一致时，多样性达到最小，此时模型缺乏新的学习信号，容易陷入过拟合。而当多样性值过高时，也可能引入过多噪声或不合理的样本分布，例如将原本属于“猫”标签下的图像错误标注为“狗”，虽然会显著提高多样性指标，但实际会损害模型性能。因此，合理的多样性应当在“足够差异”与“不过度偏离”之间取得平衡，使训练数据既能提供新的判别信息，又不破坏整体分布的一致性。由此可见，相似性与多样性在刻画数据有效性时具有互补性：前者度量数据是否保持在合理的分布范围内，后者度量数据是否具有足够的覆盖与差异，二者结合才能全面反映数据优化对模型训练的真实影响。

3.3 实验验证

3.3.1 实验设置

为了系统性地研究训练数据的相似性与多样性对模型性能的影响，本节设计了一系列实验。核心思路是通过数据增强方法在相似性-多样性空间上人为调节训练数据的分布，从而构造不同的数据情景，并在多个基准数据集上评估模型性能。具体来说，在 MNIST[108]、CIFAR-10、CIFAR-100 和 ImageNet 等不同规模与复杂度的数据集上，选用 ResNet-18、ResNet-50[3]、WideResNet-28-10[109] 等代表性模型进行训练，以验证该框架的普适性。本实验的优化器选用 SGD，在在 CIFAR-10/100 上，动量 0.9，初始学习率 0.1，在 60/120/160 epoch 衰减 20%），权重衰减 0.0005；在 ImageNet 上，初始学习率 0.2（在第 30/60/80 个 epoch 衰减 10%），权重衰减 $1e^{-4}$。与此同时，在 CIFAR-10/100 上，批量大小为 128，训练 200 个 epochs。在 ImageNet 上，训练 112.6k 步，批量大小 1024。为了保证结果可比性，所有图像均经过预处理：将像素值除以 255 后进行数据集统计标准化，并移除了默认的随机裁剪与水平翻转。对于 ImageNet 数据集，由于原始图像尺寸不一致，所有图像均被统一缩放至 $(224, 224)$ 像素。部分 ImageNet 分类准确率结果引用自[95]。此外，本节采用 ImageNet 训练集的缩减子集进行相似性与多样性计算。因为是在特征空间中进行的计算，在 MNIST 数据集上，使用 ResNet-18 作为特征嵌入模型；对于 CIFAR-10、CIFAR-100 及 ImageNet 数据集，则采用 ResNet-50[3]作为特征嵌入模型。

为了调整不同的数据特性，本节使用的数据增强方法涵盖从高相似性/低多样性到低相似性/高多样性的全谱系增强。本节几乎纳入了现有的主流增强策略，力求构建一个尽可能全面的实验设置，使用的方法包括：

- 几何类：旋转、平移、裁剪等；
- 擦除类：Cutout[22]、AdvMask[110]、GridMask[37]、RandomErasing[38]、Hide-and-Seek[39]；
- 混合类：Mixup[23]、CutMix[24]；
- 自动增强类：AutoAugment[25]、RandAugment[26]、Fast-AutoAugment[51]、Keep-Augment[56]；
- 属性变换：Color jitter、Posterize、Contrast、PatchGaussian。

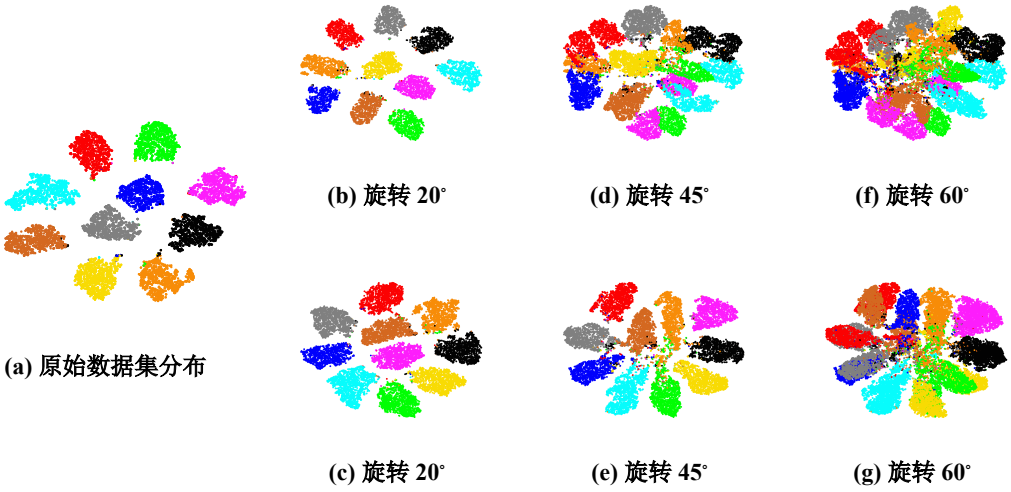


图 3-1　使用 MNIST[108]数据集对图像嵌入空间进行可视化分析，该嵌入采用预训练的 ResNet-18[3]模型得到深度特征。利用 t-SNE 方法在不同数据变换水平下可视化图像嵌入空间的分布变化。图（b）（d）（f）无原始数据，图（c）（e）（g）含原始数据

通过这种设计，可以更系统地揭示数据特性（相似性与多样性）与模型性能之间的潜在关系，从而为后续训练数据优化算法的设计与评估提供坚实的实证基础。

3.3.2　数据集分布可视化实验

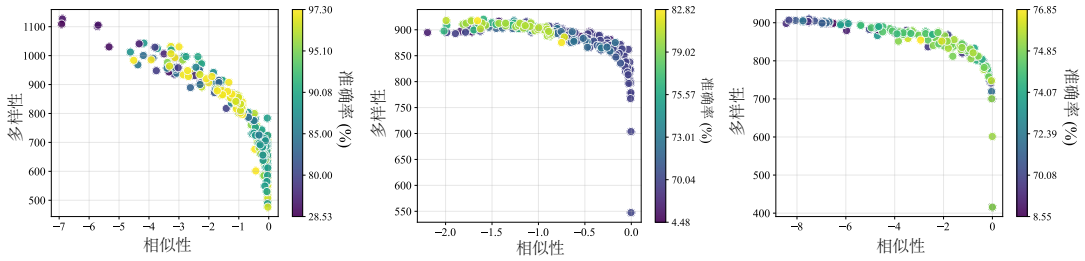
为了直观验证所提出的相似性与多样性度量是否能够合理刻画训练数据的分布特性，本节首先在 MNIST 数据集上进行可视化实验。该实验通过人为控制数据变换幅度，逐步改变数据分布，从而考察相似性与多样性指标的动态变化趋势。

基于 MNIST 数据集，采用旋转变换作为调控手段，分别设置旋转角度为 20°、45° 和 60° 三种情况。为了比较不同分布情境下的表现，实验分别考虑两种场景：1）仅包含旋转后的数据集（无原始数据），即完全由增强数据构成；2）原始数据与旋转数据结合（含原始数据），即增强数据作为原始数据的补充。为便于可视化分析，本节使用在 MNIST 训练集上预训练的 ResNet-18 作为嵌入模型，将图像映射到特征空间，再利用 t-SNE 方法对嵌入结果进行二维可视化。

如图 3-1所示，当对第一行中 20°、45° 和 60° 的旋转增强进行分析时，可以观察到随着旋转角度的增大，增强数据集的分布与原始数据集逐渐产生差异。在 20° 旋转角度下，增强数据集基本保持原有分布特性，表明原始数据集与增强数据集之间具有高度相似性。然而当旋转角度增大至 45° 和 60° 时，嵌入空间中数

表 3-1 不同旋转角度增强下 MNIST 测试集的相似性与多样性度量。

旋转角度	相似性	多样性
20°	-1.49	719.02
45°	-6.77	861.85
60°	-12.46	1269.45



(a) CIFAR-10 数据集 (b) CIFAR-100 数据集 (c) ImageNet 数据集

图 3-2 在 CIFAR-10[96]、CIFAR-100[96]和 ImageNet[111]数据集上，数据增强方法在相似性-多样性平面上的有效性研究

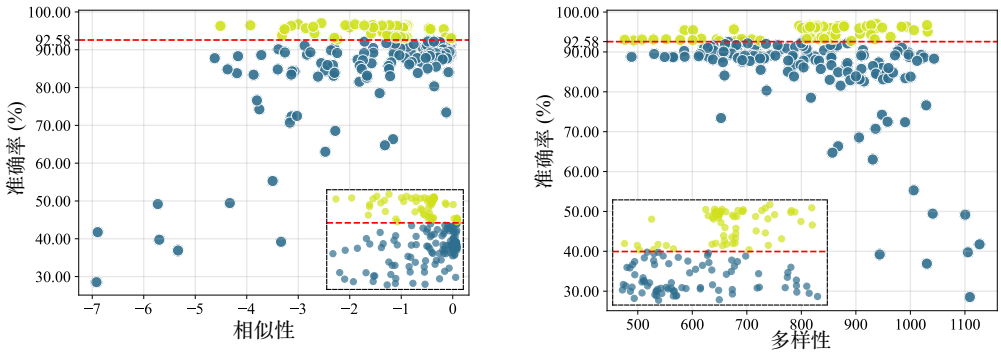
据集的整体分布发生显著变化，导致相似性度量值降低。特别是当旋转角度达到 60° 时，数据分布发生剧烈变化，不同类别间的分类边界变得模糊不清。表 3-1 给出了不同旋转角度增强下相似性与多样性度量的对应统计结果。

为了进一步探究多样性和数据集分布之间的关系，在图 3-1 第二行展示了旋转 20°、45° 和 60° 时原始数据集与增强数据集的联合嵌入结果。嵌入结果表明，随着旋转角度的增大，增强数据集的复杂度显著增加。具体而言，在 20° 和 45° 旋转角度下，各类别的聚类分布与原始数据集基本保持一致。但当旋转角度达到 60° 时，数据分布发生根本性改变，部分聚类甚至出现重叠现象。这种分布上的显著偏离会对模型训练产生负面影响。

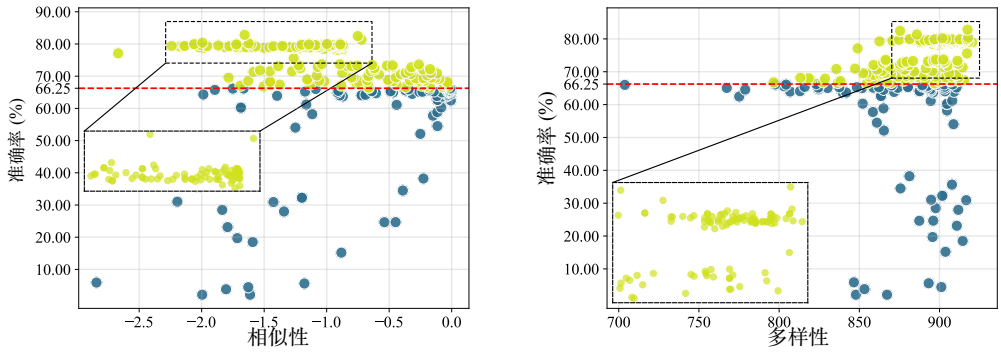
该实验清楚地表明：相似性与多样性度量能够合理反映数据分布的变化趋势，并为分析不同训练数据配置下的性能表现提供了定量工具。

3.3.3 性能评估实验

为了系统性验证所提出的相似性与多样性度量在不同数据条件下的有效性，本节在 CIFAR-10、CIFAR-100 和 ImageNet 三个基准数据集上开展了大规模实验。实验共覆盖 220 种不同的数据变换策略在 CIFAR-10 和 CIFAR-100 上的表现，以及 143 种方法在 ImageNet 上的表现。



(a) CIFAR-10: 相似性　　(b) CIFAR-10: 多样性



(c) CIFAR-100: 相似性　　(d) CIFAR-100: 多样性

图 3-3　相似性与准确性、以及多样性与准确性之间的关系。红色虚线表示完全未使用数据增强的基线模型准确率

如图3-2所示，相似性和多样性度量呈反比关系，这与理论预期一致，较高的相似性往往对应较低的多样性，反之亦然。尽管二者存在负相关关系，但仅凭相似性或多样性任一指标都无法全面解释增强方法的有效性。如图3-3所示，在不同数据集上获得最优性能的方法并不一定对应于最高或最低的相似性或多样性值。事实上，那些相似性或多样性极端的方法往往效果较差，这说明任何单一指标都不足以全面评估增强方法的效果。例如，将图像旋转 60° 和垂直翻转会产生较高的多样性值，但这两种增强方法在 CIFAR-10 上的最终准确率分别仅为 36.9% 和 39.2%。相比之下，常用的裁剪和水平翻转增强方法能达到 96.2% 的更高准确率。这表明，过高的相似性可能导致过拟合，因为增强数据与训练集几乎相同；但过高的多样性又可能导致增强数据与训练集分布差异过大，使得深度模型难以学习潜在分布，从而导致性能下降。因此，仅凭相似性或多样性都无法全面解释增强方法的有效性，两者在研究中具有互补性。

尽管在三个数据集上的最佳方法表现出不同的相似性－多样性特征，但我

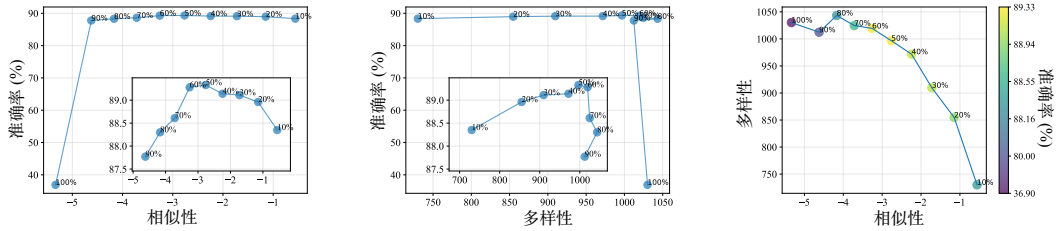
们观察到它们集中分布在相似性–多样性平面的特定区域，即所谓的“候选区间”。以 ResNet-50 作为嵌入模型时，该区域的特征是相似性约在-2.5 至-1 之间，多样性约在 800 至 900 之间。通过进一步分析这些数据集上最佳增强方法的相似性-多样性特征发现这些方法对度量的偏好源于数据集拟合难度的差异。如图 3-2(a) 所示，CIFAR-10 上表现最佳的增强方法主要集中于具有较高相似性和多样性的区域（这一现象在图 3-3(a) 中更为明显）。然而从图 3-2(b) 可见，对 CIFAR-100 而言，高多样性的增强方法更为有益（这一结论也可通过图 3-3(c) 和图 3-3(d) 验证）。这是因为 CIFAR-100 相比 CIFAR-10 具有更多类别且每类训练样本更少，我们认为，拟合 100 个类别的复杂度显著高于 10 个类别，因此需要更多样化的样本来达到最佳性能，使得高多样性的数据增强对 CIFAR-100 至关重要。如图 3-2(c) 所示，ImageNet 上最佳增强方法同时具有较高的相似性和多样性度量值，其中更高多样性更为有益。这可能是因为 ImageNet 训练集与测试集分布相似，为避免过拟合，在相同分布下提供更多样化的样本更为有利。

因此，影响模型性能的一个关键因素是平衡训练数据的相似性与多样性偏好，不同数据集和任务对这一平衡点的需求有所不同。这一发现也解释了为什么某些增强技术（如 GridMask 和 AutoAugment）在不同数据集上采用不同的参数和策略。

3.3.4 基于相似性与多样性的准确率分析

为了进一步验证所提出相似性与多样性度量的合理性，本节在 CIFAR-10 数据集上开展了一个旋转增强比例调控实验，展示所提出的度量方法与数据增强所引入的变化程度保持一致性。该度量方法能够定量反映数据增强方法所带来的变化幅度，并可进一步用于指导增强方法的参数调优。

如图 3-4所示，本节在 CIFAR-10 数据集上评估了旋转增强 *rotate* (60°) 在不同增强比例（10% 至 100%）下的相似性与多样性表现。通过调节增强数据所占的比例，可以人为地控制增强技术引入的变化程度，例如，当比例为 10% 时，表示仅有 10% 的数据被旋转 60°，其余 90% 保持原始状态。从图 3-4(a)和图 3-4(b)可以观察到，随着增强比例的增加，相似性度量逐渐降低而多样性度量持续上升。这一现象符合直观预期：引入更多的旋转样本会使数据分布逐渐远离原始训练集，从而增加差异性。与此同时，图 3-4(c)显示，模型准确率在比例增加的过程



(a) 相似性　　(b) 多样性　　(c) 相似性和多样性

图 3-4　通过将 CIFAR-10 数据集上 *Rotate*(60°) 增强比例从 10% 调整至 90%，这里展示了相似性与多样性度量的变化趋势

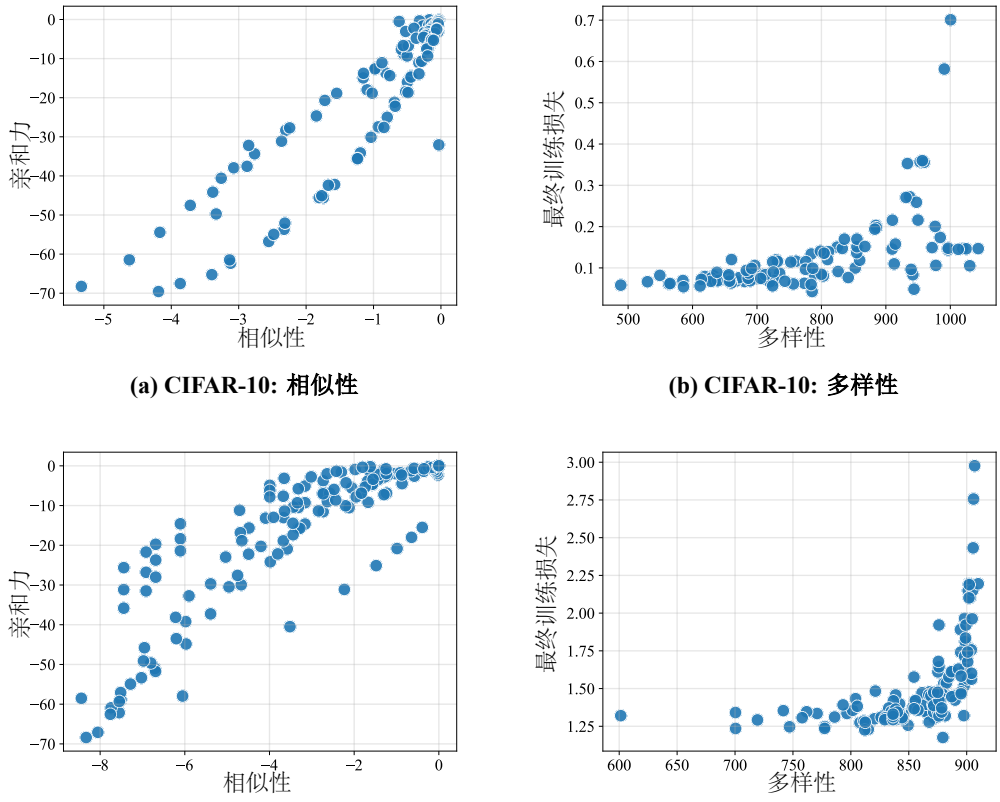
中并非单调变化，而是呈现“先升后降”的趋势。当比例在 50% 至 60% 的比例区间中，模型准确率达到峰值。此时，相似性和多样性均保持在相对较高水平，模型获得了最佳的泛化性能。这一发现与图 3-2(a) 中的结果相一致，进一步验证了度量的有效性。

值得注意的是，当增强比例过低时，训练数据与原始分布高度相似，模型容易陷入过拟合，导致泛化性能不足。随着增强比例的增加，多样性逐步增加，模型的泛化性能得到改善。但若增强比例继续升高，导致增强数据与原始训练集和测试集的分布差距过大时，模型性能会急剧下降。例如，当增强比例超过 60% 时，测试准确率迅速下滑；当比例达到 100% 时，模型性能最差。这表明，适度的相似性与多样性平衡对于模型性能至关重要：过低会导致过拟合，过高则可能引入噪声或分布偏移。

3.3.5　与其他度量方法的比较

在本节中，将所提出的度量方法与文献[95] 中提出的亲和度（affinity）和最终训练损失（Final Training Loss, FTL）在 CIFAR-10 和 ImageNet 数据集上进行了对比。实验结果表明，尽管相似性和多样性是基于数据集分布距离的度量方法，但它们与数据增强方法在模型训练中的表现密切相关，且在实际应用价值上优于现有工作。

具体而言，如图 3-5(a) 所示，相似性度量与亲和度之间存在较高的正相关性。同时，从图 3-5(b) 可以观察到，高多样性值的数据增强方法往往对应较高的 FTL。为了进一步分析这种相关性，本节在表 3-2 中给出了所提出度量与其他指标之间的 Pearson 相关系数[112] 和 Spearman 相关系数[113]。从表中可以看出，相



(a) CIFAR-10: 相似性　(b) CIFAR-10: 多样性

(c) ImageNet: 相似性　(d) ImageNet: 多样性

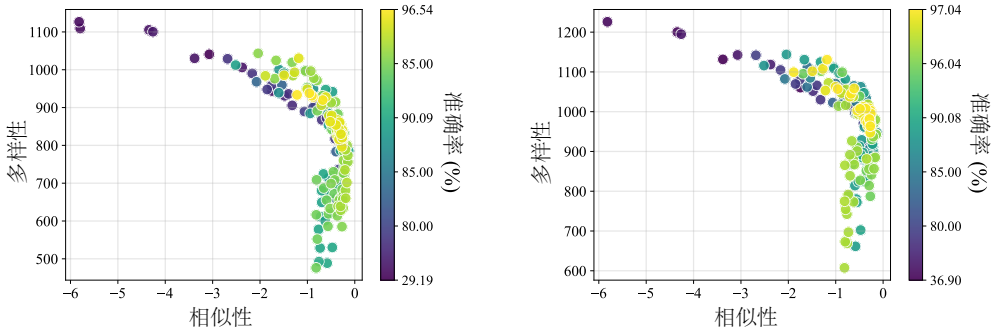
图 3-5　本章提出的度量指标与 CIFAR-10 及 ImageNet 数据集上的 affinity、FTL 之间的关系如下图所示

似性与亲和度、多样性与 FTL 之间均呈显著正相关关系，且所有相关系数的 p 值均低于 1×10^{-9}，表明该结论具有高度统计置信度。由此可见，本章提出的度量不仅能够反映数据增强方法在模型训练中的有效性，还具有较强的实际应用价值。

然而，亲和度和 FTL 需要完整的训练过程才能计算，这带来了高昂的计算成本和较低的效率，使其难以在模型训练前被应用于数据增强方法的设计。同时，最优增强策略往往需要在不同数据集上联合优化亲和度和 FTL[95]，而不考虑数据集之间的差异。因此，亲和度和 FTL 本质上缺乏一种透明的方法论来指导增强策略在实际中的设计与应用。相比之下，本文提出的基于距离的度量方法清晰地揭示了数据增强技术的有效性来源：即通过生成在特定数据集上具有适当相似性和多样性的数据。由于每个数据集仅需训练一个嵌入模型即可计算该度量，因此效率远高于前者。更重要的是，相似性和多样性能够直接用于指导增强方法的设计、参数调优以及增强效果的初步验证。例如，通过调整数据增强方

表 3-2　在 CIFAR-10 和 ImageNet 数据集上，本文提出的度量指标与 affinity 之间的皮尔逊相关系数（PCC）和斯皮尔曼相关系数（SCC）如下所示。FTL：最终训练损失；PCC：皮尔逊相关系数；SCC：斯皮尔曼相关系数

数据集	指标	PCC	SCC
CIFAR-10	相似性 & 亲和力	0.91	0.89
	多样性 &FTL	0.75	0.58
ImageNet	相似性 & 亲和力	0.86	0.87
	多样性 &FTL	0.70	0.78



(a) ResNet-50　　**(b) Wide-ResNet-28-10**

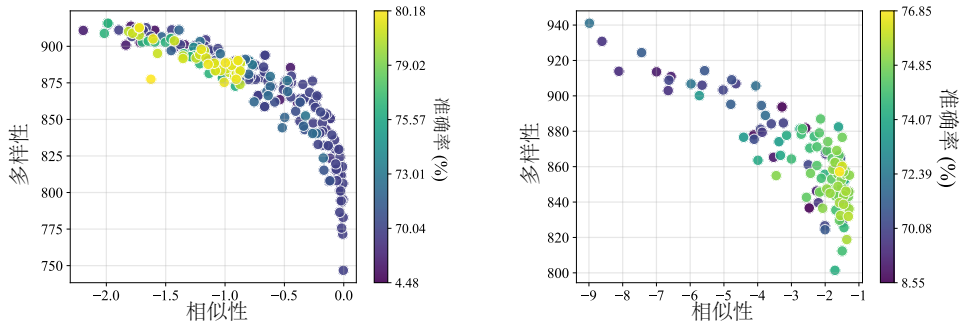
图 3-6　相似性度量是在 CIFAR-10 数据集上，分别使用 ResNet-50 和 Wide-ResNet-28-10 模型，通过计算增强数据集与测试集之间的距离得到的。实验结果表明，性能最优的数据增强方法往往能够获得较高的相似性度量值

法的参数，可以在相似性或多样性维度上对齐至“候选区间”，从而实现最佳性能。

3.3.6　训练数据与测试集分布关系的分析

为了进一步验证相似性与多样性度量在解释模型性能方面的有效性，在 CIFAR-10 数据集上考察了训练数据与测试数据之间的分布关系。尽管测试集在数据优化过程中始终不可见，但深度学习的基本假设指出：如果训练数据能够更好地贴近测试数据的分布，那么在该训练集上学习到的模型也更有可能在测试集上获得优异的泛化性能。因此，训练数据与测试数据之间的相似性在一定程度上可以作为模型泛化能力的前置指标。

实验结果如图 3-6 所示，本节采用 ResNet-50 与 Wide-ResNet-28-10 两种模型进行分析，结果表明：当训练数据与测试数据的分布更为接近时，模型在测试集上的表现也显著更优。这一观察与理论预期一致，进一步验证了相似性度量的合理性和解释力。然而，实验结果同时也揭示出，仅靠相似性并不足以全面解释



(a) DenseNet121[114]用于提取 CIFAR-100 上的嵌入特征 **(b) Vision Transformer[115]用于提取 ImageNet 上的嵌入特征**

图 3-7 不同嵌入模型对度量框架的影响

模型性能。结合前文对训练数据内部特性的分析可以看到，最佳的性能往往出现在既保持较高相似性，又维持一定多样性的区域。这说明过高的相似性虽然能够减少分布偏移，但如果缺乏足够的差异性，模型容易在训练中过拟合，从而限制泛化性能。

另一方面，单纯依赖多样性也同样具有局限性。例如，过度追求数据的差异性可能使训练分布与测试分布产生显著偏移，从而削弱模型的学习效果。因此，真正有效的训练数据往往需要在相似性与多样性之间找到平衡：一方面保证分布贴近目标任务，另一方面通过适度差异提供新的学习信号，避免模型陷入过拟合。

3.4 消融实验

为了研究不同嵌入模型的影响，本节在 CIFAR-100 上训练了 DenseNet121[114] 模型，在 ImageNet 上训练了 Vision Transformer[115] 模型，以此获得特征表示。使用 ResNet-50 作为嵌入模型的结果见图 3-2。可以观察到，无论采用哪种嵌入模型，“候选区间”始终存在。

具体而言，与图 3-2 中的结果类似，相似性与多样性度量总体上呈反比关系，大多数表现最优的方法都集中在相似性-多样性平面的特定区域。图 3-7(a) 显示，CIFAR-100 同样更偏好高多样性的增强方法；其在图 3-2(b) 与图 3-7(a) 中的度量结果之间的 Spearman 相关系数高达 0.91。与图 3-2(c) 的结果类似，在图 3-7(b) 中，ImageNet 上的最佳增强方法也集中在相似性和多样性均较高的区域。

因此，即使采用不同的嵌入模型，本章仍然可以得到一致的结论：最佳方法

依然集中在“候选区间”内。区别仅在于，不同模型下该“候选区间”的具体数值范围可能略有差异。

3.5　本章小结

本章围绕训练数据的相似性与多样性两个核心维度，构建了一个系统化的分析框架，并通过理论建模与实证实验对其有效性进行了全面探讨。与传统仅依赖模型性能来评价数据质量的方法不同，本章提出的度量指标完全独立于模型训练，因而具有高效性和可解释性。首先，本章基于最优传输（Optimal Transport）方法定义了相似性度量，用于刻画不同数据集在特征-标签联合分布上的接近程度，从而反映训练数据与目标分布保持一致性的程度。其次，借鉴生物多样性和统计学方法，提出了基于主成分协方差结构的多样性度量，用于刻画类别间的覆盖性和样本的结构性差异。这两种度量互为补充：相似性衡量数据不过度偏离目标分布，而多样性则衡量了数据在特征空间中具备足够的覆盖和差异。通过在 MNIST、CIFAR-10、CIFAR-100 和 ImageNet 等数据集上的大规模实证实验，验证了该框架的合理性和有效性。实验结果表明：（1）相似性与多样性在不同数据集和模型上的偏好存在差异；（2）最佳的训练性能往往并不依赖于单一维度的极值，而是集中在相似性和多样性之间的一个“候选区间”，我们提出了一个候选区间的近似估计方法，可以用来快速判断当前数据集的相似性多样性程度；（3）所提出的度量能够有效揭示训练数据的内在特性，并为已有方法的解释和新方法的设计提供统一视角。

综上，本章的研究不仅深化了对训练数据在模型性能中作用机制的理解，也为后续章节的自适应数据增强、跨模态数据选择以及动态数据优化方法的提出奠定了理论和实践基础。

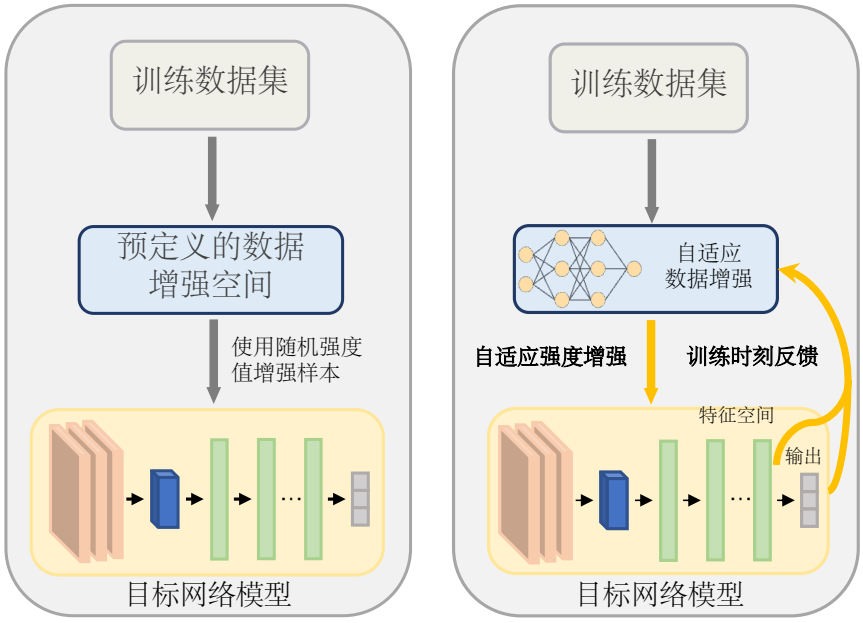
第四章　基于多样性提升的自适应数据增强研究

4.1　引言

第三章提出了一个统一的数据优化目标函数，并进一步构建了一个训练数据分析框架，系统性地刻画了数据相似性与多样性对模型性能的影响机制。实验结果表明，不同数据集及不同训练阶段的模型在这两个维度上的偏好并不相同。然而，这一框架本质上仍是分析与解释的工具，尚未解决一个更为关键的实践问题：在实际训练过程中，如何动态调节训练数据的分布，使其能够自适应地匹配目标模型的学习状态。正如式(3-1)所定义的那样，如何在训练中对 q_ϕ 进行动态优化，从而借助数据增强技术对数据的多样性进行有效提升，而这是从“事后分析”走向“训练中调控”的关键挑战。

近年来，随着深度神经网络（DNN）的快速发展，模型规模持续扩大，对高质量训练数据的依赖愈发显著[11,34-35,116-117]。然而，标注数据往往稀缺且昂贵，使得如何在有限数据条件下通过提高数据多样性来缓解模型过拟合问题成为深度学习的核心难题之一[118-120]。在这一背景下，数据增强（Data Augmentation, DA）成为最常用且行之有效的技术之一。通过人为构造的变换，DA 能够在一定程度上缓解过拟合、提升数据多样性，从而增强模型的泛化性能[66,121-124]。然而，现有的主流方法仍存在两个局限性：其一，增强幅度通常是固定或随机的[22-24,37,40-41,125]，无法适配训练过程的动态变化；其二，自动化增强策略虽然一定程度上减轻了人工调参负担，但需要消耗巨大的计算资源针对每个特定的数据集进行搜索优化，难以在大规模实际任务中推广[25-27,53,58]。这些问题导致增强数据的多样性往往与模型的最优需求不匹配，从而影响模型性能。

更具体地说，在训练早期，如果过强的增强幅度引入过多数据变化性，往往会带来噪声和分布漂移，使模型难以有效收敛，导致欠拟合；而在训练后期，如果增强幅度过弱，训练数据缺乏足够的多样性，又可能加剧模型对训练集的过拟合风险。换言之，传统增强方法在不同训练阶段与模型需求之间存在显著的“错



(a) 传统数据增强　　**(b) 自适应数据增强**

图 4-1　传统数据增强与自适应数据增强的对比。(a) 传统方法通常采用固定或随机的增强强度，无法适应深度模型的训练状态。(b) 本文提出的 AdaAugment 能够基于模型的实时反馈，为每个训练样本动态确定增强强度，从而实现模型状态感知的数据增强

位”。第三章的实验框架已经揭示：高效的训练数据应当既保持分布与目标任务的贴近，又具备适度差异性。因此，一个理想的增强机制应当能够面向模型的学习状态，动态调控训练数据的多样性水平。

解决上述挑战的关键在于基于目标模型的实时反馈，自适应地调节增强数据的变换幅度。如图 4-1 所示，本章对比了传统的数据增强（DA）机制与自适应 DA 机制。与依赖随机或预设增强强度的传统方法不同，自适应 DA 机制能够根据目标网络在训练过程中的实时反馈，动态调整数据增强操作的强度。此外，与大多数现有的自适应 DA 方法不同，该自适应 DA 机制更强调优化增强强度，而非增强操作本身的具体形式。这种自适应策略显著减少了搜索空间，能够直接调节增强数据的变异性，而无需依赖具体的增强策略设计。实现自适应 DA 的一个直接途径是构建一种能够反映每个训练样本实时学习状态的度量指标。然后，利用该度量来决定最优的增强强度水平，从而与模型的训练状态保持一致。然而，理论分析表明，判定模型的学习状态（例如欠拟合或过拟合风险）是不可判定问题[126-128]。

为应对这些局限与挑战，本章提出了一种新颖且无需人工调参的自适应数据增强方法 AdaAugment。与依赖人工经验或固定策略的数据增强方法不同，

AdaAugment 采用强化学习算法，自适应地为每个训练样本确定具体的增强强度。AdaAugment 的核心是一个双模型架构：策略网络与目标网络。策略网络在训练过程中，根据目标网络的实时反馈学习增强策略，从而为每个样本动态确定增强强度；同时，目标网络利用这些自适应增强的数据进行训练。两个网络被联合优化，无需单独对目标网络进行重复训练，显著提升了该方法在不同数据集上的实用性。

通过这种机制，所学习到的策略可以自适应地调节训练数据的相似性与多样性分布，以契合目标模型的学习状态，从而优化增强数据中引入的有效信息。具体而言，在策略网络训练过程中，通过对比完全增强数据与未增强数据的损失，分别估计欠拟合与过拟合风险。然后，将这些损失与由 AdaAugment 自适应增强数据所得的损失进行对比，从而构建奖励信号。通过这种方式，AdaAugment 能够根据模型训练状态进行动态调整，有效实现模型与数据双重自适应的数据增强，而无需任何关于特定数据集的先验知识。本章在多个基准数据集上的实验结果表明，AdaAugment 不仅显著提升了模型的泛化性能，还在计算效率上保持了竞争力，展示了其在实际应用中的广泛适用性。本章的主要贡献如下：

- 设计了一种全新的自适应数据增强方法，结合强化学习，学习了无需人工干预的数据增强策略自动优化，为后续自适应数据增强研究提供了新范式。
- 设计了双网络结构，包括策略网络和目标网络，策略网络根据目标网络的实时反馈学习增强策略，两者联合优化协同学习，避免了额外的重训练步骤，提升了方法的实用性。
- 实验验证了 AdaAugment 方法的有效性，在保证计算效率的前提下，体现出强大的泛化能力。
- AdaAugment 通过动态调控训练数据的多样性来平衡欠拟合与过拟合风险，从而为后续基于数据特性的训练优化提供了新的思路与实践路径。

4.2 本章工作

4.2.1 方法概述

AdaAugment 的核心目标是在训练过程中动态调节数据增强的强度来同时缓解欠拟合和过拟合风险，以更好地适配深度模型在不同阶段的学习状态。具体来

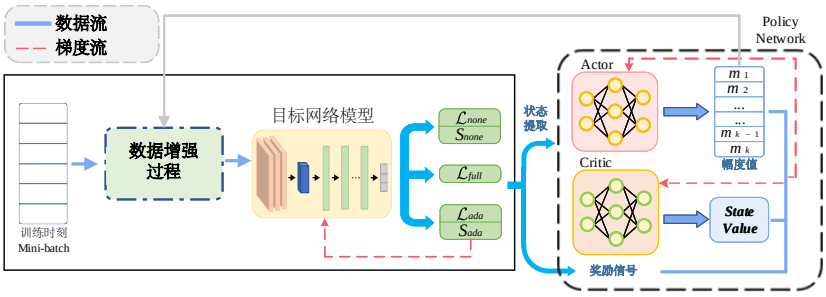


图 4-2　AdaAugment 方法的双模型框架示意图

说，它在训练初期阶段施加较弱的增强来加速模型收敛，并在后期逐渐提升增强幅度以减轻过拟合风险。这一切通过模型自发的行为完成，实现了更平衡的泛化能力。这种自适应调节可以形式化为一个逐样本的决策问题，从而无需依赖人工设计的度量指标。为实现这一目标，图4-2展示了 AdaAugment 的双模型架构：在训练目标网络的同时，引入一个策略网络来动态优化数据增强操作的强度。该双模型框架支持两个网络的联合优化，无需对目标网络进行额外的重复训练，在每一轮训练中根据策略网络的输出增强样本。经过自适应增强的数据样本再用于优化目标网络的参数。两者通过联合优化协同学习，这样的设计既避免了额外的独立重训练过程，又确保了增强策略能够随着模型训练状态的演化而实时调整，从而保证了方法的实用性和高效性。

4.2.2　强化学习建模

预备知识　为了将数据增强强度的动态调节形式化为一个决策问题，本章引入强化学习（Reinforcement Learning, RL）框架。RL 通常通过马尔可夫决策过程（Markov Decision Process，MDP）进行形式化描述，其包含以下组成部分：状态空间 $\mathcal{S}$、动作空间 $\mathcal{A}$、状态转移函数 $\mathcal{P}：\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$（表示通过执行动作实现状态转移的概率）、奖励函数 $\mathcal{R}$、折扣因子 $\gamma\in[0,1]$ 以及时间步长 T。在给定状态 $\boldsymbol{s}\in\mathcal{S}$，强化学习智能体根据策略 $\pi(a\mid\boldsymbol{s})$ 确定动作 $a\in\mathcal{A}$。在此框架下，强化学习的目标是寻找最优策略 π^*，以最大化期望累积奖励。

假设训练数据集 $\mathcal{D}$ 包含 N 个训练样本，每个样本的形式为 $(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}$。其中 $\boldsymbol{x}$ 表示原始数据，$\boldsymbol{y}$ 是一个 k 维的独热编码向量（k 为类别总数），用于指示 $\boldsymbol{x}$ 的真实标签。增强操作定义为 $e(m,\boldsymbol{x})$，其中 e 从增强空间 $\mathcal{E}$ 中随机选取，m 表示 e 对应的强度值。虽然 e 是从 $\mathcal{E}$ 中随机选择的，但所应用的强度值由 AdaAugment

表 4-1 增强空间 $\mathcal{E}$ 的定义

操作类型	S_{Max}	对称性
identity	-	-
auto contrast	-	-
equalize	-	-
color	1.9	-
contrast	1.9	-
brightness	1.9	-
sharpness	1.9	-
rotation	30°	±
translate_x	10	±
translate_y	10	±
shear_x	0.3	±
shear_y	0.3	±
solarize	256	-
posterize	4	-

自适应确定。这种方法能有效调节增强数据的变异性，在训练数据的多样性和一致性之间实现更优平衡，从而提升模型泛化能力。

在表 4-1中提供了 $\mathcal{E}$ 的具体细节，其中包含广泛使用的图像操作及其对应的最大强度值。值得注意的是，该增强空间 $\mathcal{E}$ 的设置与[25-26,51,53]中的设定基本一致。与先前工作不同，本章的方法自适应地确定增强空间 $\mathcal{E}$ 中的强度值，而非在训练前预设固定值。具体而言，实际应用的强度值通过 $S_{Max} \times m$ 计算确定，其中 S_{Max} 是相应变换允许的最大强度。此外，对于对称变换，其对称方向将进行随机选择。

状态设计 在 RL 框架下，状态向量 $\boldsymbol{s}$ 需要准确反映样本难度、模型训练状态以及增强操作强度。为此，AdaAugment 引入了特征映射（Feature Mapping）的方式获取这些信息：

- $\boldsymbol{s}_{none}$：目标网络在未增强样本 $\boldsymbol{x}$ 上的特征表示，用于衡量样本在当前训练状态下的固有难度；
- $\boldsymbol{s}_{ada}$：目标网络在自适应增强样本 $\tilde{\boldsymbol{x}}$ 上的特征表示，反映增强强度与模型状态的交互结果。

这两个特征向量共同构成状态输入，为策略网络提供决策依据，通过这种设计，策略网络能够学习到在不同训练阶段应采用怎样的增强强度，从而使训练数据的多样性水平更好地与模型收敛过程相匹配。

动作设计　策略网络负责确定增强数据的增强强度 $\boldsymbol{m}$。尽管训练过程中每个小批量的组成具有随机性，但强度 $\boldsymbol{m}$ 以样本为单位进行操作，与每个训练样本一一对应。为简化表述，将当前小批量数据的增强强度记为 $\boldsymbol{m}$，其维度等于批量大小，且每个 $m \in \boldsymbol{m}$ 被严格限制在区间 $[0,1]$ 内。

当 $m=0$ 时，不施加任何数据增强；而当 $m=1$ 时，则表示采用对应增强操作的最大强度。因此，越接近 0 的强度值会产生与原始数据更相似的样本，而越接近 1 的强度值则会产生多样性更高的数据。自适应增强样本 $\tilde{\boldsymbol{x}}$ 可通过以下方式获得：

$$\tilde{\boldsymbol{x}} = e(m, \boldsymbol{x}), \tag{4-1}$$

其中 e 为随机选取的增强操作，m 则是根据动作策略确定的强度值。公式 (4-1) 使我们能够优化增强数据在相似性与多样性之间的权衡偏好。

因此，策略网络无需知晓具体的增强类型和方向，仅需优化强度值即可。更重要的是，这种对相似性-多样性偏好的调整体现了 AdaAugment 在训练早期加速模型收敛、在训练后期缓解过拟合风险的有效机制。

奖励函数设计　考虑一个由参数 θ 表征的目标分类模型 f_θ，输入样本 $\boldsymbol{x} \in \mathbb{R}^n$，$f_\theta(\boldsymbol{x})$ 表示网络输出。设 $\mathcal{L}$ 为交叉熵损失函数，$\mathcal{L}(f_\theta(\boldsymbol{x}), \boldsymbol{y})$ 表示原始样本 $(\boldsymbol{x}, \boldsymbol{y})$ 的损失项。

本文提出的方法旨在通过基于目标模型反馈控制数据增强操作的强度，在训练早期促进模型拟合，在训练后期缓解过拟合。为此，本节基于增强策略定义了三个损失项：首先，$\mathcal{L}_{full}(f_\theta(\boldsymbol{x}^+), \boldsymbol{y})$ 表示采用最大增强强度样本的损失，即 $\boldsymbol{x}^+ = e(m=1, \boldsymbol{x})$；其次，$\mathcal{L}_{none}(f_\theta(\boldsymbol{x}^-), \boldsymbol{y})$ 表示未增强样本的损失，即 $\boldsymbol{x}^- = e(m=0, \boldsymbol{x}) = \boldsymbol{x}$；最后，基于公式 (4-1)，自适应增强数据的损失记为 $\mathcal{L}_{ada}(f_\theta(\tilde{\boldsymbol{x}}, \boldsymbol{y}))$，其中增强强度由策略网络确定。受课程学习[89]的启发，将奖励函数设计如下：

$$r = \lambda(\mathcal{L}_{full} - \mathcal{L}_{ada}) + (1-\lambda)(\mathcal{L}_{ada} - \mathcal{L}_{none}), \tag{4-2}$$

其中 $\lambda \in [0,1]$ 为调节因子，在训练过程中从 1 逐渐衰减至 0。这种自适应机制激励强化学习模块在训练初期施加较弱的数据增强以加速模型学习，并在后期逐步增加增强强度以提高样本多样性并降低过拟合风险。

表 4-2 A2C 网络结构

	层序	层类型	维度
Actor	1	线性层	(512, 512)
	2	线性层	(512, 256)
	3	线性层	(256, 1)
Critic	1	线性层	(512, 512)
	2	线性层	(512, 256)
	3	线性层	(256, 1)

策略学习 策略的目标是确定实例级别的数据增强操作强度。为了实现策略学习，本节采用了轻量化且广泛使用的 A2C 算法[129]。其优势在于结构简单，适合在本任务场景中高效应用。A2C 的网络结构如表 4-2 所示，由 Actor 网络 θ_a 和 Critic 网络 θ_c 组成。Actor 网络负责学习策略，即在给定状态下的动作概率分布 $\pi(a \mid s)$；Critic 网络则用于估计特定状态下的价值函数，记作 $V_{\theta_c}(s)$。值得注意的是，Critic 网络仅包含若干线性层，保证了整体的计算开销较低。

在更新 Actor 和 Critic 网络时，本节针对本研究的具体问题场景重新形式化了损失函数：

- Actor 网络的损失函数定义为:

$$\mathcal{L}_{\text{actor}} = -\log \pi_{\theta_a}(a \mid s_{none})(r + \gamma V_{\theta_c}(s_{ada}) - V_{\theta_c}(s_{none})). \tag{4-3}$$

  该损失函数鼓励策略网络在给定状态下采取能够带来更高奖励的动作。

- Critic 网络的损失函数定义为：

$$\mathcal{L}_{\text{critic}} = \mathbb{E}[(r + \gamma V_{\theta_c}(s_{ada}) - V_{\theta_c}(s_{none}))^2]. \tag{4-4}$$

  用于最小化价值估计与实际奖励之间的误差，从而提升 Critic 网络对状态价值的估计精度。

为了更直观地展示 AdaAugment 的整体流程，算法 1 给出了详细的算法步骤，涵盖了策略学习与目标网络训练的完整交互过程。

算法 1 AdaAugment 的整体工作流程

Require: 训练数据集 $\mathcal{D}$，批量大小 B，增强操作空间 $\mathcal{E}$，交叉熵损失函数 $\mathcal{L}$，总训练轮数 T

随机初始化目标网络 f_θ，Actor 网络参数 θ_a 和 Critic 网络参数 θ_c
初始化增强幅度向量 $\boldsymbol{m} \leftarrow \mathbf{0}$
for $t = 0: T-1$ **do**
    从数据集 $\mathcal{D}$ 中采样一个 mini-batch $\{\boldsymbol{x}_i, y_i\}_{i=1}^{B}$
    从增强操作空间 $\mathcal{E}$ 中随机采样一个增强操作 e
    根据式 (4-1) 生成自适应增强样本 $\tilde{\boldsymbol{x}}_i$（对应幅度 $m \in \boldsymbol{m}$）以及完全增强样本 $\boldsymbol{x}_i^+$
    分别计算无增强样本损失 $\mathcal{L}_{none}(f_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i)$、完全增强样本损失 $\mathcal{L}_{full}(f_\theta(\boldsymbol{x}_i^+), \boldsymbol{y}_i)$ 以及自适应增强样本损失 $\mathcal{L}_{ada}(f_\theta(\tilde{\boldsymbol{x}}_i, \boldsymbol{y}_i))$
    基于 Actor 网络 θ_a 确定当前 batch 的自适应增强幅度 $\boldsymbol{m}$
    根据式 (4-2) 计算奖励值 r
    使用式 (4-3) 与式 (4-4) 更新 Actor 网络参数 θ_a 与 Critic 网络参数 θ_c
    基于 $\mathcal{L}_{ada}(f_\theta(\tilde{\boldsymbol{x}}_i, \boldsymbol{y}_i))$ 更新目标网络 f_θ
end for

Ensure: 训练好的目标网络 f_θ

4.2.3　理论分析

在数据增强中，增强幅度 m 与增强样本的损失值 $\mathcal{L}$ 近似成正比，即 $m \propto \mathcal{L}$。因此，通常可以近似得到如下不等式：

$$\mathcal{L}_{none} \leqslant \mathcal{L}_{ada} \leqslant \mathcal{L}_{full},$$

这一关系表明，$\mathcal{L}_{none}$ 与 $\mathcal{L}_{full}$ 可以分别作为模型在训练过程中潜在的过拟合与欠拟合状态的实时指标。与此同时，这些损失值会随着训练的推进不断动态演化。

在奖励函数式 (4-2) 中，项 $\mathcal{L}_{full} - \mathcal{L}_{ada}$ 在训练初期具有重要意义。此时策略网络倾向于采用较小的增强幅度，以最大化 $\mathcal{L}_{full}$ 与 $\mathcal{L}_{ada}$ 的差异。换言之，初期训练阶段更相似的增强样本有助于模型快速捕捉全局模式，从而加速收敛[130-131]。随着训练的推进，策略网络逐渐转向优化 $\mathcal{L}_{ada} - \mathcal{L}_{none}$，进而倾向于施加更大的增强幅度，以提高 $\mathcal{L}_{ada}$ 与 $\mathcal{L}_{none}$ 之间的差异。此时更具多样性的增强样本能够有效缓解过拟合风险。从整体机制来看，AdaAugment 的动态调控过程与课程学习[89,91] 的思想高度契合：在初期，模型先从相似样本中学习稳定的全局模式；在后期，再逐渐引入更多差异性强的样本，以增强模型的鲁棒性和泛化性能。

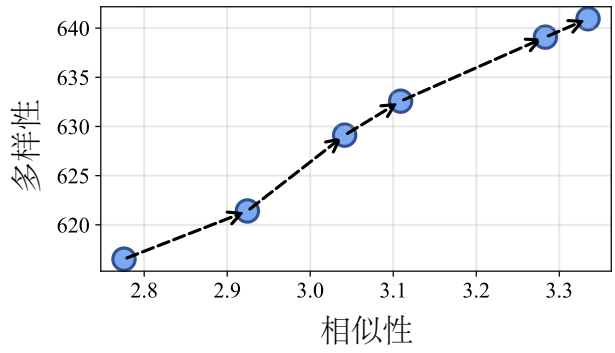


图 4-3 使用 ResNet-50 在 CIFAR-10 数据集上对整个训练过程中自适应增强数据集的评估

为了进一步验证 AdaAugment 在训练过程中对相似性-多样性权衡的自适应调控能力，在不同训练阶段计算了增强数据的相似性与多样性指标。值得强调的是，这里所采用的度量方法直接沿用了第三章中提出的分析框架[121]，即基于特征-标签联合分布的相似性度量与类别覆盖驱动的多样性指标。这不仅保证了评估方法的一致性，也凸显了第三章方法在实际优化场景中的解释力与适用性。具体而言，我们在训练过程中每 50 个 epoch 统计一次全体样本的平均增强幅度，并由此计算对应的相似性与多样性指标。如图 4-3 所示，随着训练的进行，相似性度量逐渐降低，而多样性度量逐渐升高，这与前文的理论分析保持一致。这一结果不仅验证了 AdaAugment 的内在适应性，还进一步证明了第三章所提出的相似性-多样性度量在刻画训练数据与模型性能关系上的普适性。换句话说，AdaAugment 通过动态调控增强强度，使训练数据的分布特性在训练过程中自动迁移到最优的相似性-多样性平衡区间，并与模型不断演化的学习状态保持协调，从而实现更高效、更鲁棒的训练。

4.3 实验验证

4.3.1 实验设置

基准数据集 为了全面验证 AdaAugment 的有效性，参考已有研究[28,40,53]，本节在多个不同类型的基准数据集上开展实验，涵盖从粗粒度分类到长尾分布和细粒度识别等多种任务场景。

- 粗粒度分类任务：包括 CIFAR-10/100[96]、Tiny-ImageNet[132] 和 ImageNet-

1k[14] 等经典数据集。这些数据集规模和复杂度各不相同，有助于检验 AdaAugment 在小规模到大规模数据条件下的普适性与稳健性。

- 长尾分类任务：在 ImageNet-LT 和 Places-LT[133] 等长尾数据集上进行实验，重点验证 AdaAugment 在类别分布极度不均衡条件下的普适性和稳健性。
- 细粒度分类任务：在 Oxford Flowers[134]、Oxford-IIIT Pets[135]、FGVC-Aircrafts[136] 和 Stanford Cars[137] 等数据集上进行评估，这些任务更依赖模型在区分细微特征上的能力，可检验 AdaAugment 对细粒度识别的适应性和优势。
- 迁移学习与可视化分析：在迁移学习场景中，测试 AdaAugment 是否能提升模型的跨数据集泛化性和可迁移性。此外，还进行了特征可视化实验与收敛性分析，以进一步揭示 AdaAugment 在训练动态与特征学习方面的作用机理。
- 复杂度与消融实验：系统分析了 AdaAugment 在计算效率与资源开销方面的影响，并通过消融实验验证了不同模块和参数设置对整体性能的贡献。

对比方法　本节将 AdaAugment 与当前最具代表性和广泛应用的 13 种数据增强方法进行比较，这些方法涵盖了从经典手工设计到自动化与对抗式增强的不同思路，包括 HaS[39]、Fast-AutoAugment (FAA)[51], DADA[28]、Cutout[22]、CutMix[24]、MADAug[131]、AdvMask[40]、GridMask[37]、AutoAugment[25]、RandAugment[26]、TeachAugment[27]、TrivialAugment[53] 和 RandomErasing[38]。这些方法几乎覆盖了现有主流增强策略的代表性范式，能够为 AdaAugment 提供充分且公平的对比基准。

其中 AutoAugment 与 Fast-AutoAugment 使用了 16 种操作，包括 Shear X/Y、Rotate、AutoContrast、Invert、Equalize、Solarize、Posterize、Contrast、Color、Brightness、Sharpness、Cutout 和 Sample pairing。它们在 CIFAR-10/100 上随机选择 25 种增强策略，在 ImageNet-1k 上选择 24 种，每种策略包含 2 个增强操作。RandAugment 包含 14 种操作，与表 4-1 相同，并在增强空间中随机选择 N 个操作，使用固定的强度 M。TeachAugment 通过训练增强模型来表示几何增强与颜色增强，从而覆盖了 AutoAugment 中的大部分操作。TrivialAugment 使用 21 种增强操作，除了表 4-1 中列出的操作外，还包括 Cutout、Invert、Flip-X/Y、Sample pairing、Blur 和 Smooth。此外，在 CIFAR-10/100 上，CutMix 的概率为 0.5，在

表 4-3 CIFAR-10/100 数据集上的测试准确率 (%)（平均值 ± 标准差）。* 表示先前文献中报道的结果

方法	CIFAR-10				CIFAR-100			
	ResNet-18	ResNet-50	WRN-28-10	ShakeShake	ResNet-18	ResNet-50	WRN-28-10	ShakeShake
baseline	95.28±0.14*	95.66±0.08*	95.52±0.11*	94.90±0.07*	77.54±0.19*	77.41±0.27*	78.96±0.25*	76.65±0.14*
HaS[39]	96.10±0.14*	95.60±0.15	96.94±0.08	96.89±0.10*	78.19±0.23	78.76±0.24	80.22±0.16	76.89±0.33
FAA[51]	95.99±0.13	96.69±0.16	97.30±0.24	96.42±0.12	79.11±0.09	79.08±0.12	79.95±0.12	81.39±0.16
DADA[28]	95.58±0.06	95.61±0.14	97.30±0.13*	97.30±0.14*	78.28±0.22	80.25±0.28	82.50±0.26*	80.98±0.15
Cutout[22]	96.01±0.18*	95.81±0.17	96.92±0.09	96.96±0.09*	78.04±0.10*	78.62±0.25	79.84±0.14	77.37±0.28
CutMix[24]	96.64±0.62*	96.81±0.10*	96.93±0.10*	96.47±0.07	79.45±0.17	81.24±0.14	82.67±0.22	79.57±0.10
MADAug[131]	96.49±0.10	97.12±0.11	97.48±0.12	97.37±0.11	79.39±0.19	81.40±0.10	83.01±0.11	81.67±0.18
AdvMask[40]	96.44±0.15*	96.69±0.10*	97.02±0.05*	97.03±0.12*	78.43±0.18*	78.99±0.31*	80.70±0.25*	79.96±0.27*
GridMask[37]	96.38±0.17	96.15±0.19	97.23±0.09	96.91±0.12	75.23±0.21	78.38±0.22	80.40±0.20	77.28±0.38
AutoAugment[25]	96.51±0.10*	96.59±0.04*	96.99±0.06	97.30±0.11	79.38±0.20	81.34±0.29	82.21±0.17	82.19±0.19
RandAugment[26]	96.47±0.32	96.25±0.06	96.94±0.13*	97.05±0.15	78.30±0.15	80.95±0.22	82.90±0.29*	80.00±0.29
TeachAugment[27]	96.47±0.09	96.40±0.10	97.50±0.12	97.29±0.14	79.27±0.17	80.54±0.23	82.81±0.20	81.30±0.20
TrivialAugment[53]	96.28±0.10	97.07±0.08	97.18±0.11	97.30±0.10	78.67±0.19	81.34±0.18	82.75±0.26	82.14±0.16
RandomErasing[38]	95.69±0.10	95.82±0.17	96.92±0.09	96.46±0.13*	75.97±0.11*	77.79±0.32	80.57±0.15	77.30±0.18
AdaAugment	**96.75**±0.06	**97.34**±0.13	**97.66**±0.07	**97.41**±0.06	**79.84**±0.27	**81.46**±0.12	**83.23**±0.23	**82.82**±0.25

ImageNet-1k 上为 1.0；Cutout 与 AdvMask 的概率均为 1.0。

实验配置 公式 (4-3) 和公式 (4-4) 中的折扣因子 γ 设置为 0.99，该设置遵循了已有研究[138-140] 的设置建议。所有图像均通过将像素值除以 255 并进行数据集统计量归一化进行预处理。参照先前工作[40,53]的设置，对于 CIFAR-10 和 CIFAR-100 数据集，采用 ResNet-18/50[3]、Wide-ResNet-28-10（WRN-28-10）[109] 和 Shake-Shake-26x32（ShakeShake）[141] 架构。Shake-Shake 模型训练 1800 个周期，使用带动量的 SGD 优化器（Nesterov 动量），学习率 0.01，批量大小 256，权重衰减 $1e^{-3}$，并采用余弦学习率衰减策略。其余所有网络训练 300 个周期，使用带动量的 SGD 优化器（Nesterov 动量），学习率 0.1，批量大小 128，权重衰减 $5e^{-4}$，同样采用余弦学习率衰减。对于 Tiny-ImageNet 数据集，将图像尺寸调整为 64×64，使用 ImageNet 预训练权重初始化模型，随后采用各种增强方法进行微调。每个实验均进行三次独立随机试验。对于 ImageNet-1k 数据集，遵循[25,53] 的设置，采用 ResNet-50 作为目标网络。将图像尺寸调整为 224×224，学习率 0.1，批量大小 256，Nesterov 动量参数 0.9，权重衰减 0.4。需要注意的是，由于计算成本巨大，每种情况下的实验仅执行一次。Baseline 方法仅采用填充与水平翻转作为默认增强策略。为保障公平比较，所有方法均采用相同的训练配置。如无特殊说明，所有实验均在配备 8 块 NVIDIA-2080TI-GPU 的服务器上完成。

表 4-4　Tiny-ImageNet 数据集上的图像分类准确率 (%)（平均值 ± 标准差）

方法	ResNet-18	ResNet-50	WRN-50-2	ResNext-50
baseline	61.38±0.99	73.61±0.43	81.55±1.24	79.76±1.89
HaS[39]	63.51±0.58	75.32±0.59	81.77±1.16	80.52±1.88
FAA[51]	68.15±0.70	75.11±2.70	82.90±0.92	81.04±1.92
DADA[28]	70.03±0.10	78.61±0.34	83.03±0.18	81.15±0.34
Cutout[22]	68.67±1.06	77.45±0.42	82.27±1.55	81.16±0.78
CutMix[24]	64.09±0.30	76.41±0.27	82.32±0.46	81.31±1.00
MADAug[131]	70.16±0.76	78.62±0.32	82.38±0.42	81.41±1.26
AdvMask[40]	65.29±0.20	78.84±0.28	82.87±0.55	81.38±1.54
GridMask[37]	62.72±0.91	77.88±2.50	82.25±1.47	81.05±1.33
AutoAugment[25]	67.28±1.40	75.29±2.40	79.99±2.20	81.28±0.33
RandAugment[26]	65.67±1.10	75.87±1.76	82.25±1.02	80.36±0.62
TeachAugment[27]	70.05±0.57	70.56±0.44	82.95±0.13	81.39±0.97
TrivialAugment[53]	69.97±0.96	78.41±0.39	82.16±0.32	80.91±2.26
RandomErasing[38]	64.00±0.37	75.33±1.58	81.89±1.40	81.52±1.68
AdaAugment	**71.25**±0.64	**79.11**±1.51	**83.07**±0.78	**81.92**±0.29

4.3.2　基准数据集上的对比实验结果

CIFAR-10/100 数据集实验结果　表 4-3 展示了 AdaAugment 在 CIFAR-10 和 CIFAR-100 上的实验结果，选择了当前广泛使用的多种深度网络架构作为目标模型，包括 ResNet-18/50[3]、Wide-ResNet-28-10[109] 以及 ShakeShake-26-32[141]。通过在不同模型和不同数据集上的验证，本节能够系统性地评估 AdaAugment 的普适性与稳健性。实验结果表明，AdaAugment 在所有模型和两个数据集上都实现了稳定且显著的性能提升，相较于已有的最先进数据增强方法，具有明显优势。以 ResNet 系列为例，AdaAugment 分别在 ResNet-18 和 ResNet-50 上带来了 1.47% 和 1.66% 的提升，在 Wide-ResNet-28-10 和 ShakeShake-26-32 上，提升幅度更大，分别达到了 2.14% 和 2.51%。这说明 AdaAugment 的有效性并不依赖于特定架构，而是能够在不同规模和复杂度的模型上普遍奏效。尽管在较小规模的模型上不同方法间的性能差异相对有限，因为这些模型在表达能力上受到限制；然而在更大规模的模型上，AdaAugment 的优势被进一步放大。例如，在 ShakeShake 模型上，AdaAugment 在 CIFAR-10 上比 CutMix 高出近 1%，在 CIFAR-100 上甚至超过了 3.3%。这表明 AdaAugment 在更复杂模型上能够更好地释放潜力，有效缓解因模型容量增加而带来的过拟合风险。

这一结果进一步验证了 AdaAugment 的核心机制：通过利用目标网络的实时

表 4-5 在 Tiny-ImageNet 数据集上的对比实验结果，这里报告的是 ResNet-50 模型的分类准确率

baseline	HaS	FAA	DADA	Cutout	CutMix	GridMask
49.73±.37	52.05±.12	52.83±.02	53.07±.28	50.89±.52	52.31±.20	50.37±.11
MADAug	AA	RA	TeachAug	TA	RE	AdaAugment
54.95±.24	53.37±.13	50.63±.53	53.97±.23	53.88±.03	50.74±.27	**55.98**±.12

反馈，自适应地调节增强操作的强度。不同于传统方法依赖固定或随机的增强幅度，AdaAugment 能够动态匹配训练过程中模型的状态，使生成的数据分布与模型的学习需求保持一致。这种自适应调节有效提升了训练样本的利用效率，避免了过强变异导致的欠拟合和过弱变异导致的过拟合，从而显著提升了泛化性能。

另一方面，实验结果也揭示了一个有趣的结论：模型规模更大并不一定能带来更优的性能。例如，在仅采用基线增强操作的情况下，ResNet-50 的基线性能甚至低于 ResNet-18，这很可能是由于较大模型面临更容易陷入过拟合风险。然而，当引入更先进的数据增强方法时，ResNet-50 的性能显著超过了 ResNet-18，凸显出数据增强在大模型训练中的关键作用。由此可见，AdaAugment 不仅能提升整体性能，还能够显著改善大模型在小数据场景下的适应性，进一步证明其方法论上的价值与实用性。

Tiny-ImageNet 数据集实验结果 首先在中等规模且难度较高的 Tiny-ImageNet 数据集上评估了 AdaAugment 的有效性。实验分别在预训练的 ResNet-18、ResNet-50、Wide-ResNet-50-2 (WRN-50-2) 和 ResNeXt-50[142] 模型上进行，结果如表 4-4 所示。可以看到，AdaAugment 在所有架构上均取得了显著提升，展现出方法在不同规模与复杂度网络下的普适性。例如，在 ResNet-18 上提升幅度高达 9.87%，这一大幅度的性能改进尤其凸显了 AdaAugment 在小模型和中等数据规模下的潜力。在 ResNet-50 上，准确率提升 5.5%，显示出其在较大模型上的稳定适应性；在 WRN-50-2 和 ResNeXt-50 上也分别取得了 1.52% 和 1.16% 的提升。

进一步地，在 Tiny-ImageNet 上从零开始训练 ResNet-50，结果见表 4-5。即便在这种更具挑战性的设定下，AdaAugment 依然展现出稳定的改进，准确率提升超过 1%。这充分说明，AdaAugment 的优势不仅体现在迁移学习场景中，也能在从头训练的场景下保持有效性。

这些结果表明，AdaAugment 能够根据目标模型的学习状态自适应地调整增

表 4-6　ImageNet-1k 数据集上使用 ResNet-50 的 Top-1 准确率 (%)。部分结果引用自[27-28,53]

baseline	HaS	FAA	DADA	Cutout	CutMix	MADAug	GridMask	AA	RA	TeachAug	TA	RE	AdaAugment
77.1	77.2	77.6	77.5	77.1	77.2	**78.3**	77.9	77.6	77.8	78.0	77.9	77.3	**78.3**

表 4-7　使用 4-V100-GPU 服务器在 ImageNet-1k 数据集上采用 ViT-Base、ViT-Large 和 Swin-Transformer 架构的性能表现 (%)

模型	基线	AdaAugment
ViT-B	81.46	**82.55**
ViT-L	83.50	**84.67**
Swin-Transformer	85.00	**85.65**

强幅度，从而在不同规模和不同初始化方式的模型中均显著提升性能。这凸显了本章方法的泛化能力与实用潜力。

ImageNet-1k 数据集实验结果　为了进一步验证 AdaAugment 在大规模场景下的有效性，在更具挑战性的 ImageNet-1k 数据集[14] 上进行了实验。需要指出的是，由于 AdvMask 的稀疏对抗攻击模块计算开销过大[40]，因此未在 ImageNet-1k 上纳入比较。

实验结果如表 4-6 和表 4-7 所示。在该数据集上，所有数据增强方法在 ResNet-50 及更先进的基线模型上均显著优于仅使用简单增强的基线模型。其中，AdaAugment 在 ResNet-50 上取得了最优性能，相较于基线方法提升了 1.1%，且无需引入额外的模型结构改动，也未显著增加训练开销。这表明自适应增强机制在实际大规模训练任务中同样具有很高的性价比。

在与其他方法的对比中，值得注意的是，虽然 MADAug 在 ImageNet-1k 上的结果与 AdaAugment 接近，但其采用了双层优化框架（bi-level optimization），需要投入远高于 AdaAugment 的计算成本。相比之下，AdaAugment 在保持轻量化设计的同时，依然实现了与其相当甚至更优的性能，这使得 AdaAugment 在效率与性能的权衡上具有明显优势。此外，表 4-7 展示了在 ViT-B、ViT-L 和 Swin-Transformer 等更先进的视觉 Transformer 架构上的实验结果。可以看到，AdaAugment 依然带来了一致性的提升。这些结果表明，AdaAugment 不仅适用于 CNN 架构，同样能够在基于 Transformer 的视觉模型中展现优势，凸显了方法的通用性与可扩展性。

综上，与传统方法对整个数据集采用固定增强操作不同，AdaAugment 针对

表 4-8 不同数据增强方法在 CIFAR-10 上的迁移测试准确率 (%)。预训练的 ResNet-50 模型分别在 CIFAR-100（上行）和 Tiny-ImageNet（下行）数据集上训练

baseline	HaS[39]	FAA[51]	DADA[28]	Cutout[22]	CutMix[24]	MADAug[131]	GridMask[37]	AA[25]	RA[26]	TeachAug[27]	TA[53]	RE[38]	AdaAugment
91.53±.03	92.51±.24	92.28±.13	92.58±.09	92.42±.20	92.81±.47	92.84±.10	91.49±.10	92.82±.04	92.78±.23	92.83±.18	92.80±.16	92.55±.05	**93.06**±.25
64.02±.05	66.84±.06	70.32±.63	69.04±.43	65.54±.75	69.29±.09	72.82±.32	64.88±.43	69.53±.53	64.68±.97	69.98±.17	71.53±.35	64.56±.27	**76.86**±.12

每张图像在训练过程中动态确定增强幅度，实现了模型与样本适配。这种灵活性极大地提升了模型的泛化能力，并进一步验证了自适应增强的必要性与优势。

4.3.3 迁移学习实验结果

迁移学习（Transfer Learning）是评估模型可迁移性和泛化能力的重要手段[143]，在数据增强方法的研究中也被广泛采用[40,56,121]。相比单一数据集上的性能提升，迁移学习能够更真实地反映模型所学特征的普适性与稳健性，即模型是否能在不同数据分布之间迁移时依旧保持较高的准确率。在实验中，首先在 CIFAR-100 和 Tiny-ImageNet 数据集上，使用不同的数据增强方法对 ResNet-50 进行预训练。随后，将这些预训练好的模型迁移至 CIFAR-10 数据集上进行微调，并比较在测试集上的准确率表现。该实验的核心思想在于：更优的数据增强方法能够在预训练阶段提供更具代表性和多样性的特征学习，从而在迁移到新的目标任务时展现更强的适应性和泛化性。

表 4-8 给出了不同数据增强方法在 CIFAR-10 上的迁移学习测试结果。整体来看，尽管不同方法之间的迁移精度差异相对有限，但 AdaAugment 在所有实验设定中均取得了稳定的领先。相比于其他主流方法，AdaAugment 在预训练阶段通过自适应增强机制有效调节了数据的相似性与多样性，使得模型学到的特征更具普适性，从而在迁移任务中表现更佳。这一结果凸显了 AdaAugment 的两个关键优势：

1. 提升特征可迁移性：通过动态调控增强幅度，AdaAugment 能够使预训练模型学到的表示更具普适性，而不是过度依赖特定数据集的分布特征。
2. 增强泛化性能：在微调阶段，迁移后的模型在新任务上能够更快收敛，并表现出更好的泛化能力。这说明 AdaAugment 有效缓解了源任务中的过拟合风险，使得模型在面对新的数据分布时具备更强的适应性。

综上，迁移学习实验进一步验证了 AdaAugment 不仅在原始任务上提升模型精度，而且在跨任务场景中同样展现了出色的性能。这说明 AdaAugment 作为一

表 4-9　ImageNet-LT 和 Places-LT 数据集的 Top-1 分类准确率 (%)。* 表示原始论文中报道的结果

数据集	方法	闭集设置				开集设置			
		Many-shot	**Medium-shot**	**Few-shot**	**Overall**	**Many-shot**	**Medium-shot**	**Few-shot**	**F-measure**
ImageNet-LT	OLTR	43.2±0.1*	35.1±0.2*	18.5±0.1*	35.6±0.1*	41.9±0.1*	33.9±0.1*	17.4±0.2*	44.6±0.2*
	OLTR+**AdaAugment**	**45.9**±0.1	**38.3**±0.1	**22.0**±0.2	**39.0**±0.1	**44.1**±0.1	**36.8**±0.1	**20.8**±0.2	**45.8**±0.1
Places-LT	OLTR	**44.7**±0.1*	37.0±0.2*	25.3±0.1*	35.9±0.1*	**44.6**±0.1*	36.8±0.1*	25.2±0.2*	46.4±0.1*
	OLTR+**AdaAugment**	43.7±0.1	**41.1**±0.1	**29.5**±0.2	**39.6**±0.1	43.9±0.1	**40.8**±0.1	**28.9**±0.1	**50.4**±0.1

表 4-10　采用 ResNet-50 架构的 AdaAugment 方法在细粒度数据集上的测试准确率 (%)（平均值 ± 标准差）

数据集	基线	AdaAugment
Oxford Flowers[134]	89.47±0.08	**97.17**±0.14
Oxford-IIIT Pets[135]	89.73±0.18	**91.95**±0.24
FGVC-Aircraft[136]	77.25±0.09	**90.92**±0.05
Stanford Cars[137]	82.13±0.03	**84.76**±0.20

种自适应数据增强方法，能够切实增强深度模型的泛化能力与可迁移性。

4.3.4　长尾数据集实验结果

尽管多数数据增强方法尚未在大规模长尾偏置数据集上进行系统验证，为了进一步验证 AdaAugment 的通用性和有效性，将其应用于 ImageNet-LT 和 Places-LT 两个典型的长尾分类数据集[133]。实验严格基于 OLTR[133] 提供的代码框架，完全遵循其超参数设定与训练策略，以确保公平对比。表 4-9 展示了实验结果，可以看到，AdaAugment 在 many-shot、medium-shot、few-shot 以及 open-set 四个场景下均带来了显著的性能提升。尤其值得注意的是，在 closed-set 设定下，AdaAugment 的准确率提升超过 3%，这充分表明了其在缓解长尾分布带来的不平衡问题上具有强大能力。

这种提升主要得益于 AdaAugment 的自适应增强机制：通过在训练过程中动态调控增强强度，它能够为 head classes 提供足够的多样性以避免过拟合，同时为 tail classes 提供相对适度的相似性以增强判别能力。与固定或随机增强不同，AdaAugment 能够更好地平衡相似性与多样性，从而在长尾数据场景中实现更加鲁棒的表示学习。

4.3.5 细粒度数据集上的实验结果

为了更全面地验证 AdaAugment 的有效性，本小节将其应用于多个细粒度（fine-grained）图像分类任务，包括 Oxford Flowers[134]、Oxford-IIIT Pets[135]、FGVC-Aircraft[136] 以及 Stanford Cars[137]。这些数据集的共同特点是类别间差异细微、判别难度较高，因此对增强方法的质量和适应性要求更高。在实验中，本节统一采用在 ImageNet 上预训练的 ResNet-50[3] 作为主干网络，并在目标细粒度数据集上进行微调。为确保公平性，所有方法在相同的实验设置下进行比较。

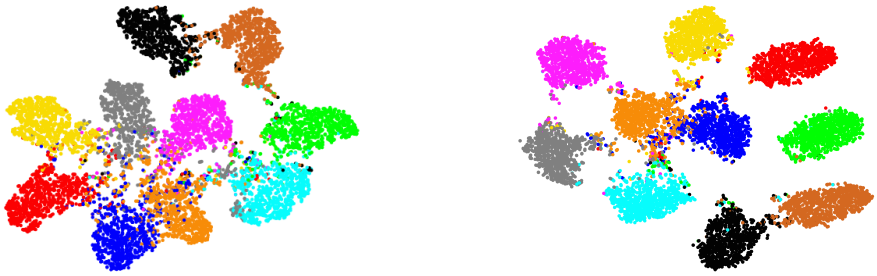
表 4-10 给出了各数据集的测试准确率，可以清晰地看到 AdaAugment 在所有细粒度数据集上均带来了显著的性能提升，在 Oxford Flowers 上，准确率提升达到 7.70%，从 89.47% 显著提高至 97.17%，在 Oxford-IIIT Pets 上，提升 2.22%, 在 FGVC-Aircraft 上，提升最为突出，达到 13.67%，在 Stanford Cars 上，尽管数据集本身难度较大，AdaAugment 仍实现了 2.63% 的提升。这些结果表明，AdaAugment 能够在类别边界模糊、判别难度极高的场景下，依然有效地缓解过拟合风险并增强模型的判别性。其自适应增强机制通过动态调节增强强度，使得训练数据在相似性与多样性之间保持合理平衡，从而帮助模型更好地捕捉细粒度类别的局部差异。这不仅提高了模型在细粒度分类上的表现，也再次验证了 AdaAugment 在复杂任务下的普适性与有效性。

4.3.6 分析实验结果

可视化实验效果分析

数据增强的核心目标在于提升模型的泛化能力，使其不仅能在训练数据上拟合良好，还能在未知测试数据上表现稳健。为更直观地展示 AdaAugment 的有效性，本节在 CIFAR-10 数据集上设计了可视化实验，对比分析采用 AdaAugment 与未采用该方法训练的模型在特征空间中的表现差异。具体而言，利用 t-SNE 算法[144] 将 CIFAR-10 测试集的特征嵌入进行降维可视化，这些特征嵌入均来自不同训练模型的特征图。理论上，具备更强泛化能力的模型应当展现出更清晰的特征分布，其表现为类间分离度更高、类内紧凑性更强，从而反映出模型对不同类别的判别性。

图 4-4 展示了基线模型与采用 AdaAugment 训练模型的 t-SNE 可视化结果。



(a) 基线
DI=5.03×10^{-5}

(b) AdaAugment
DI=1.1×10^{-4}

图 4-4　基于 CIFAR-10 数据集使用 t-SNE 算法的可视化效果分析。嵌入模型为 ResNet-50。DI：Dunn 指数

与图 4-4(a) 相比，图 4-4(b) 的分布结构显著优化，表现为类间分离度更高、类内紧凑性更强。这种几何结构的优化表明 AdaAugment 有效提升了模型在特征层面的判别能力。

为了对可视化结果进行定量评估，本节进一步采用 Dunn Index (DI)[145] 对聚类质量进行定量评估。DI 指标通过类间最小距离与类内最大距离的比值来衡量特征嵌入的判别性，值越大代表聚类结构越清晰、泛化能力越强，其定义如下：

$$DI = \frac{\min_{1 \leqslant i \neq j \leqslant m} \delta\left(C_i, C_j\right)}{\max_{1 \leqslant j \leqslant m} \Delta_j}, \tag{4-5}$$

其中，$\delta(C_i, C_j)$ 表示不同簇 C_i 与 C_j 之间的类间距离，Δ_j 表示簇 C_j 内部的平均类内距离。DI 值越高，说明聚类效果越优。在本实验中，基线模型（图 4-4(a)）的 DI 值为 5.03×10^{-5}，而 AdaAugment（图 4-4(b)）的 DI 值提升至 1.05×10^{-4}，相较于基线提升了 108.7%。这一结果表明，AdaAugment 在改善特征嵌入结构、提升泛化能力方面具有显著优势。

复杂度分析

大多数现有的数据增强方法通常依赖于训练过程中预定义的随机增强幅度，这种机制虽然几乎不会引入额外的计算开销，但其局限性同样明显：由于增强幅度在训练过程中保持固定或随机分布，它往往难以与深度模型动态演化的学习状态相匹配，从而可能导致训练早期欠拟合或训练后期过拟合的问题。与之相

表 4-11 在 CIFAR-10 数据集上辅助策略网络的额外架构参数与训练时间分析。实验设备为 2 块 NVIDIA RTX2080TI GPU 和 Intel(R) Xeon(R) CPU E5-2678 @ 2.50GHz 处理器

模型	FLOPs	参数量	GPU 耗时	准确率提升
ResNet-18	1.82G	+0.15M	+0.41h$_{\pm 0.03}$	+1.47%$_{\pm 0.06}$
ResNet-50	4.14G	+0.60M	+0.49h$_{\pm 0.03}$	+1.68%$_{\pm 0.13}$
WRN-28-10	5.25G	+0.19M	+0.43h$_{\pm 0.03}$	+2.04%$_{\pm 0.07}$

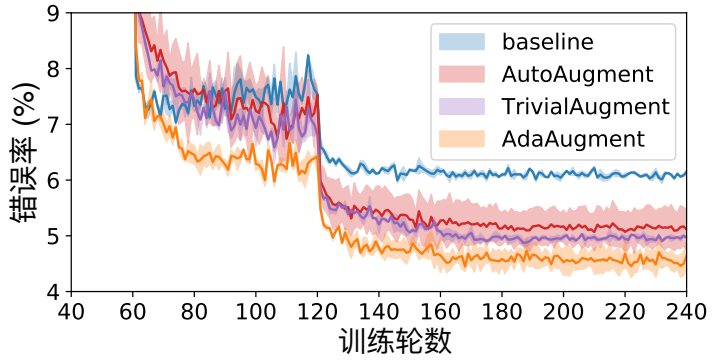


图 4-5 使用 ResNet-50 在 CIFAR-10 数据集上的训练过程收敛性分析

比，AdaAugment 在在线训练中引入了一个辅助策略网络（policy network），用于实时自适应地确定增强操作的幅度。虽然这种机制不可避免地会增加少量参数和计算开销，但这是为了实现更有效数据增强而进行的权衡策略。

本节遵循先前研究采用的方法论[25,51,58]，本节全面分析 AdaAugment 的参数和时间复杂度以评估其效率。由于策略网络的参数复杂度与目标网络的特征空间相关，在表 4-11中报告了在 ResNet-18/50 和 WRN-28-10 上的参数复杂度。值得注意的是，可以观察到**策略网络的参数复杂度仅使 ResNet-18 增加 1.3%**（原参数 11.7M），**ResNet-50 增加 2.4%**（原参数 25.5M），以及 **WRN-28-10 增加 0.52%**（原参数 36.5M）。此外，这些网络的总体训练成本仅有轻微增加。

尽管训练成本有边际增加，AdaAugment 实现了显著的准确率提升：ResNet-18、ResNet-50 和 WRN-28-10 分别获得 1.47%、1.68% 和 2.04% 的显著改进。这一改进凸显了 AdaAugment 在效果与效率之间的卓越平衡，证明了其在最小化训练成本的同时提升模型性能的能力。

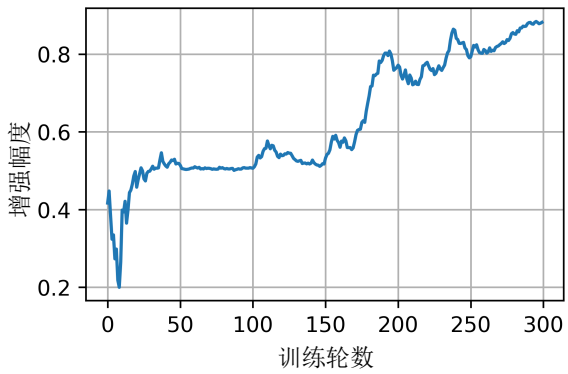


图 4-6　使用 ResNet-50 在 CIFAR-100 数据集上整个训练过程中增强幅度值的动态变化

收敛性分析

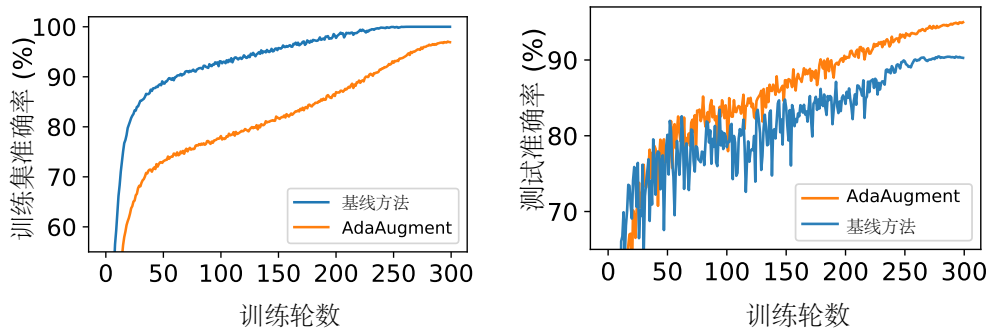
除了参数与计算复杂度外，训练过程中算法的收敛性也是评估数据增强方法有效性的重要指标。本节进一步分析了 AdaAugment 在整个训练过程中的收敛特性，并与两种代表性增强方法 AutoAugment 与 TrivialAugment 进行了对比。实验设计如下：在 CIFAR-10 数据集上，本节采用 ResNet-18 模型，并使用多步学习率衰减策略进行训练。初始学习率设为 0.1，在第 60、120、160、220 和 280 个 epoch 时按 0.2 倍进行衰减，同时在训练开始阶段引入 5 个 epoch 的 warm-up。

从图 4-5中可以观察到，AdaAugment 显著提升了模型性能，特别是在第二次学习率下降之后。此外，即使在首次学习率下降后，AdaAugment 仍始终保持比其他方法更低的错误率。这些实证发现不仅证明了 AdaAugment 在模型性能提升方面的卓越效能，更强调了其在训练过程中加速模型收敛的能力。因此，AdaAugment 有效降低了深度网络的训练难度。

自适应增强幅度的动态演化

为了展示自适应增强幅度在训练过程中的演化趋势，本节在 CIFAR-100 数据集上使用 ResNet-50 进行实验。学习率初始值设为 0.1，并在第 60、120、160、220 和 280 个 epoch 时依次乘以 0.2，同时在训练初期设置了 5 个 epoch 的 warmup 策略。

如图 4-6 所示，随着训练的进行，增强幅度整体呈现逐步增加的趋势，因为增强幅度正比于数据的变换程度，所以这与我们在理论分析中提出的“早期侧重相似性、后期提升多样性”的机制基本一致。这种动态调控不仅帮助模型在早



(a) 训练集准确率 **(b) 测试集准确率**

图 4-7 AdaAugmen 在缓解过拟合风险方面的有效性，使用更小的 CIFAR-10 数据训练 ResNet-50 模型

期快速学习全局模式，还能在后期有效缓解过拟合，提升泛化能力。

综上，收敛性分析表明，AdaAugment 不仅在最终精度上优于主流方法，更能够在整个训练过程中提供更加稳定和高效的收敛轨迹。这一特性有效降低了深度网络的训练难度，使得模型在不同学习阶段都能获得匹配的增强信号，从而实现更高质量的表示学习。

缓解过拟合的有效性

尽管现代深度模型能很好地拟合训练数据的分布，甚至在训练集上达到接近 100% 的精度，但这并不总意味着它们具有良好的泛化能力。尤其是在数据量有限的情况下，模型往往会过度依赖训练样本，从而陷入严重的过拟合，导致在测试集上的表现显著下降。数据增强的核心目标，正是通过人为地增加数据多样性，缓解这种风险并提升泛化能力。

为更清晰地评估 AdaAugment 在降低过拟合风险方面的有效性，本节构建了一个高过拟合风险的场景下并进行了实验。具体来说，在相对简单的 CIFAR-10 数据集上训练容量较大的 ResNet-50 模型。为了进一步放大过拟合风险，随机将训练集规模缩减 50%，而测试集保持不变。

如图 4-7 所示，在未使用 AdaAugment 的情况下，模型训练精度迅速接近 100%，但测试精度上升缓慢且波动明显，这种现象表明模型出现严重过拟合，泛化能力不足。相比之下，在引入 AdaAugment 后，模型的训练精度增长速度更慢，且整体训练精度保持在较低水平。然而，测试精度显著更高且更加稳定。这一结果表明，AdaAugment 能够有效缓解过拟合现象，并显著提升模型的泛化性能。

表 4-12　使用 ResNet-18 在 CIFAR-10 数据集上增强空间的性能影响分析

增强操作个数	2	4	6	8	10	12	14
Accuracy (%)	$95.99_{\pm 0.02}$	$96.18_{\pm 0.04}$	$96.38_{\pm 0.05}$	$96.54_{\pm 0.10}$	$96.61_{\pm 0.11}$	$96.67_{\pm 0.09}$	$\mathbf{96.75}_{\pm 0.06}$

表 4-13　自适应强度 m 的影响分析：基于 CIFAR-100 数据集和 ResNet-18/50 架构，AdaAugment 与不同 m 设置值的对比研究

模型	$m = 0.5$	随机 m	线性 m	正弦函数 m	AdaAugment
R-18	$78.62_{\pm 0.32}$	$77.08_{\pm 0.30}$	$78.38_{\pm 0.27}$	$78.58_{\pm 0.32}$	$\mathbf{79.84}_{\pm 0.27}$
R-50	$80.23_{\pm 0.29}$	$80.61_{\pm 0.19}$	$80.28_{\pm 0.31}$	$80.65_{\pm 0.15}$	$\mathbf{81.46}_{\pm 0.12}$

4.3.7　消融实验

增强空间的影响

尽管 AdaAugment 所使用的增强空间规模并未超过现有方法[25-26,51,53,131]，但其核心优势在于增强强度的自适应调整。为了进一步验证基础增强操作对性能的影响，本节设计了增强空间规模的对比实验。具体而言，通过逐步减少增强空间中的操作数量，分析其对模型性能的影响。实验结果如表 4-12 所示。

值得注意的是，该表中的增强类型数量始终少于其他基线方法。从结果可以观察到，随着增强操作数量的减少，模型精度有所下降，但下降幅度非常缓慢。例如，当 $N = 2$ 时，增强空间中仅包含两类增强操作，限制了增强数据的多样性，但准确率下降仍然不足 0.8%。这说明 AdaAugment 的性能提升并非依赖于庞大的增强操作集合，而是依靠其策略网络在训练过程中动态调节增强强度，从而在有限的增强空间中依然能够获得显著的泛化性能提升。这一结果强调了 AdaAugment 在资源受限或任务场景操作集有限时的适用性与鲁棒性。

强化学习模块的影响

通过引入强化学习（RL）模块，AdaAugment 可以在训练过程中自适应地调节数据增强操作的强度，从而避免了人工参数设计和调节。在这一部分实验中，本节评估了 RL 模块的具体贡献。在保持增强空间一致的前提下，我们测试了不同的 m 设置，包括固定值、随机值，以及线性或正弦函数递增的方式。结果如表 4-13 所示，可以看到，固定或预设的增强强度虽然在一定程度上提升了性能，但仍明显不及 RL 模块带来的效果。特别是在 CIFAR-100 上，AdaAugment 相较

表 4-14 不同强化学习模块在 CIFAR-10/100 数据集上对 ResNet-18 的性能影响分析

数据集	DDPG[146]	SAC[147]	Ours
CIFAR-10	96.73±0.06	**96.82**±0.04	96.75±0.06
CIFAR-100	79.27±0.19	79.80±0.29	**79.84**±0.27

表 4-15 折扣因子 γ 在 CIFAR-100 数据集上对 ResNet-18/50 的性能影响分析

γ	0.2	0.4	0.6	0.8	0.99
R-18	78.94±0.33	78.95±0.20	78.99±0.29	79.24±0.25	**79.84**±0.27
R-50	80.61±0.15	80.44±0.20	80.68±0.17	80.90±0.19	**81.46**±0.12

于最佳人工设定仍额外提升了约 1 个百分点。这一结果说明，AdaAugment 的性能优势主要来自于增强强度的自适应调节机制，而不是增强空间本身的设计。

此外，进一步探究了不同 RL 算法在 AdaAugment 中的适用性。除了采用轻量化的 A2C 架构外，还引入了 Soft Actor-Critic (SAC)[147] 和 Deep Deterministic Policy Gradient (DDPG)[146] 两种常见算法进行对比。实验结果如表 4-14 所示，可以看到三者在性能上的差异非常有限，这表明 AdaAugment 对 RL 模块的具体架构选择具有较强的鲁棒性。换言之，关键在于“是否引入自适应机制”，而非具体使用哪种 RL 算法。

折扣因子 γ 的影响

在实现中，本节按照已有研究[138-140] 将折扣因子 γ 设置为常数 0.99。为进一步分析其作用，测试了不同的 γ 值，结果如表 4-15 所示。可以看到，在设定的 $\gamma = 0.99$ 时，AdaAugment 的表现最佳。

参数 λ 的影响

最后，我们分析了奖励函数式(4-2) 中参数 λ 的影响。具体而言，将 λ 固定为 0 和 1，而不是动态调整。实验结果如表4-16 所示，验证了参数 λ 的稳定性，说明其在不同取值下均能保持较好的鲁棒性。

表 4-16　调节因子 λ 在 CIFAR-100 数据集上对 ResNet-18/50 的性能影响分析

λ	0	1	Ours
ResNet-18	78.53±0.36	78.70±0.30	**79.84**±0.27
ResNet-50	80.21±0.19	80.74±0.29	**81.46**±0.12

4.4　方法讨论与展望

本研究提出了 AdaAugment，一种无需调参的自适应数据增强方法。与传统依赖手工设定的增强策略不同，AdaAugment 引入了强化学习模块，根据每个训练样本的实时训练进度自适应地调整增强幅度。这种调整方式能够有效应对模型训练过程中的不同阶段，从而提高训练效率和最终的模型表现。实验结果表明，AdaAugment 在多个基准数据集上均表现出优越的模型性能，并且引入的训练开销可忽略不计。然而，尽管本章的方案在多个方面取得了良好的效果，但也存在一些局限性，值得进一步探讨。

首先，虽然 AdaAugment 通过从完全增强和未增强的数据中推导损失来估算欠拟合和过拟合的风险，但这一额外的损失计算引入了相较于传统训练方法的两次前向传播。这一操作在计算上是额外的，虽然在 GPU 小时数上增加不多，但这一额外的计算步骤仍可能对计算资源产生一定负担。为进一步提高整体效率，未来的工作可以探索采用其他替代性指标来估算欠拟合和过拟合风险，避免不必要的计算消耗。

其次，在本章设计的增强空间中，图像变换的应用强度可以通过一个实值参数（如 $m \in [0, 1]$）进行调节。然而，大多数现有的数据增强方法，如 Cutout[22]、AdvMask[40] 和 AutoAugment[25] 等，并不支持对增强强度进行明确的调整。因此，未来的研究可以尝试探索现有高级数据增强方法的可调版本，使得增强强度可以通过连续的实值进行显式调整。这样一来，本章的方法增强空间可能会得到进一步扩展，从而带来性能上的进一步提升。

最后，虽然 AdaAugment 在适应性增强上取得了较好效果，但由于其采用的是全局强化学习策略，未来的工作可以考虑根据特定任务或样本特性为每种数据类型动态分配不同的调整策略或权重，从而进一步提升方法的灵活性和性能。这也为模型训练中的动态数据增强机制提供了新的研究方向。

总之，AdaAugment 为数据增强领域提供了一种新的思路，通过强化学习动

态调整增强强度，提升了模型的训练效果和泛化能力。未来，我们期待能在增强方法的扩展性、计算效率和任务适应性等方面进行更多的探索和优化，以进一步提升模型在复杂任务和多模态数据场景中的表现。

4.5 本章小结

本章提出了一种创新性的自适应数据增强方法 AdaAugment，其核心思想是在训练过程中根据目标模型的实时反馈动态调整增强强度，从而突破传统数据增强方法固定或随机幅度的局限。与依赖人工启发式设计或高成本搜索策略的方法不同，AdaAugment 通过策略网络和目标网络的协同优化，实现了真正的免调参自适应增强框架。基于第三章提出的相似性-多样性分析框架，AdaAugment 不仅在理论上契合了“合理的训练数据应在相似性与多样性之间保持平衡”的原则，还在实践中通过动态调控增强强度，使模型能够在不同训练阶段自适应地提升数据多样性来改善模型泛化性能。通过大规模的实验结果，我们证明了其广泛的适用性与卓越的性能提升效果。AdaAugment 的提出为后续探索基于数据特性和模型状态的自适应训练优化提供了新思路，也为数据驱动的高效深度学习奠定了重要基础。

第五章　基于相似性驱动的高质量数据筛选研究

5.1　引言

经过前几章的介绍，数据在深度学习中起到非常重要的作用，尤其是数据的相似性和多样性特性，决定了训练数据本身性能的同时，又使得模型在多个领域都取得了最先进的性能表现[148-151]。然而，这种成功往往伴随着显著的代价：一方面，庞大的数据规模导致数据存储和模型训练成本急剧增加，严重依赖专用计算基础设施，从而限制了模型在实际应用中的可扩展性；另一方面，真实场景下采集的数据往往包含大量的冗余样本与噪声，这不仅降低了训练效率，还可能对模型的泛化性能造成负面影响。而这些问题从多样性的角度是很难被解决的，因此本章关注相似性来解决这个问题。

为缓解冗余数据带来的问题并提升训练效率，研究者当下主要关注两类主要方法：动态剪枝与数据选择。动态剪枝方法[33,65]通过在训练过程中动态挑选最具影响力的样本来降低训练成本，尽管在加速训练方面取得了一定效果，但由于这类方法是在在线训练的过程中进行，所以对于使用的算法和技术复杂度有着非常高的限制。与此不同，数据选择方法[31,61-62,64] 在训练前挑选一个高价值的固定子集进行训练，不仅能显著降低存储和训练成本，还能保证在多种训练场景下取得一致甚至更优的泛化性能，因此成为当前更具潜力的研究方向。现有的数据选择方法通常基于三种设计思路：样本重要性分数[31,64]、数据分布[62,81] 以及基于优化的方法[61,73]。这些方法往往通过计算单个样本的重要性或难度来挑选数据子集。

尽管这些思路取得了可观的效果，但也存在明显的局限性。一方面，多数方法仅依赖图像模态信息，在面对噪声样本时容易出现歧义，即无法有效独立样本的语义相似性。例如，难以区分“真正困难的样本”与“噪声或错误标注的样本”，这会导致选择结果偏差。另一方面，数据集存在群体效应（Group Effect）[61,152]，现有方法大多直接选取最高或最低分的样本，但忽略了样本间的交互作用。然

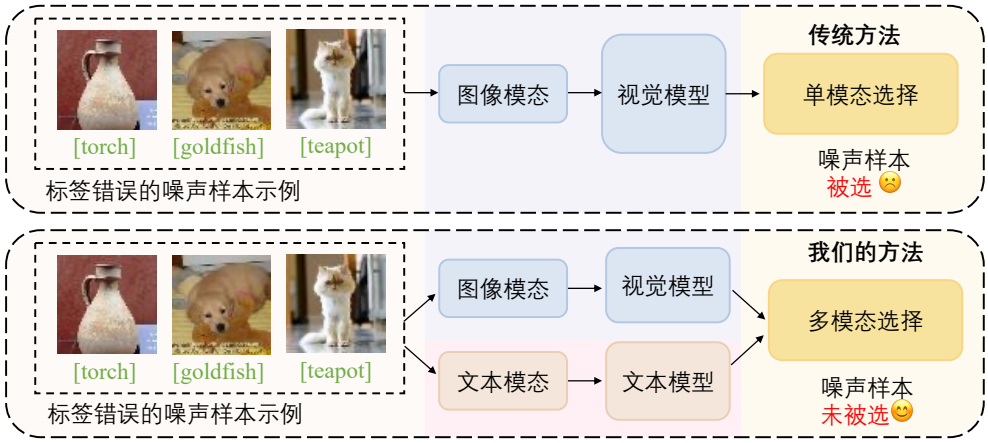


图 5-1　相似性驱动的数据筛选的优势。传统单模态方法（上图部分）在处理噪声和损坏数据时存在局限，而本章的方法（下图部分）在有效过滤噪声与损坏数据的同时，能够识别具有多样性的类别代表性样本

而在实际训练中，高分样本与低分样本之间的组合可能对整体性有显著影响，仅从单样本角度进行选择难以捕捉这种“群体效应”。基于第三章提出的优化视角，如式(3-1)中定义的那样，可以将数据选择问题理解为在固定增强分布 q_ϕ 的条件下，通过调整样本分布 p_θ 来优化训练数据的相似性与多样性平衡。

因此，本章提出了一种基于相似性驱动的高质量数据筛选方法。与传统仅依赖单一图像模态的方法不同，本章的方法利用图像-文本双模态信息，通过引入类别文本作为补充信号，使样本选择过程更加鲁棒与泛化。具体来说，该框架包括三个关键模块：（1）数据集适配模块（Dataset Adaptation）：通过引入图像和文本适配器，将大规模预训练模型中的跨模态知识有效迁移到目标数据集。（2）样本评分模块（Sample Scoring）：基于适配后的多模态特征，计算语义对齐分数（Semantic Alignment Score, SAS）与样本多样性分数（Sample Diversity Score, SDS），分别衡量样本的图文语义一致性与局部模式的多样性，从而在选择过程中兼顾代表性与多样性。（3）选择优化模块（Selection Optimization）：针对群体效应问题，引入多目标优化策略，在给定选择率下寻找最优子集，从而实现样本级与数据集级的双重优化。实验结果表明，本章提出的方法能够有效提升模型的泛化性能，特别是在噪声数据和复杂任务中的表现优于现有方法。同时，在数据选择效率方面，本章提出的方法展现了较高的计算效率，避免了昂贵的优化成本，并且能够保证在大规模任务中的可扩展性。本章的主要贡献如下：

- 本章深入分析了现有方法仅依赖图像模态的局限性，首次提出了一种基于

CLIP 的数据选择框架，该框架估计图像和文本特征之间的相似度关系，从而实现了更为鲁棒和泛化能力的数据选择方法。

- 本章的框架包含了数据集适配和样本评分两个模块，旨在促进多模态知识的有效迁移，并全面评估样本的重要性。该双模态设计能够有效地从数据集中去除噪声和异常样本。
- 实验结果表明，本章的方法在性能、跨架构泛化以及在噪声场景下的鲁棒性方面超越了现有的最先进方法。同时，本章的方法在性能与选择效率之间实现了最佳的平衡，为未来的数据选择研究奠定了坚实的基准。

5.2 本章工作

5.2.1 方法概述

本章提出了一种基于多模态信息驱动的数据选择框架，其核心目标是通过跨模态特征融合与优化，提升数据选择的鲁棒性与泛化能力。不同于仅依赖单一模态图像信息的传统方法，本章的方法借助预训练的视觉-语言基础模型 CLIP[153-154]，联合利用图像与类别文本描述，进行更准确地区分噪声样本、冗余样本和真正具代表性的样本。整个方法的框架如图 6-2 所示，主要包含三个关键模块：1）数据集适配（Dataset Adaptation）：通过轻量化的图像与文本适配器缓解预训练与目标数据集之间的域偏移问题，实现知识迁移。2）样本评分（Sample Scoring）：引入语义对齐分数（Semantic Alignment Score, SAS）与样本多样性分数（Sample Diversity Score, SDS），从语义一致性和局部分布两方面度量样本的重要性。3）选择优化（Selection Optimization）：通过多目标优化方法确定最终子集，既保证了样本的代表性和多样性，又克服了“群体效应”的问题。在这一框架下，所选数据子集不仅能够覆盖完整的类别语义空间，同时具备较强的判别能力与泛化能力。下面将对每个模块进行详细介绍。

5.2.2 数据集适配

由于 CLIP 等预训练模型通常在大规模异构数据集上训练，其特征分布与目标任务数据集之间可能存在领域漂移[155-157]，尤其是当预训练的数据集和目标数据集在分辨率、尺寸、分布上存在差距的时候。如果直接使用原始特征，可能

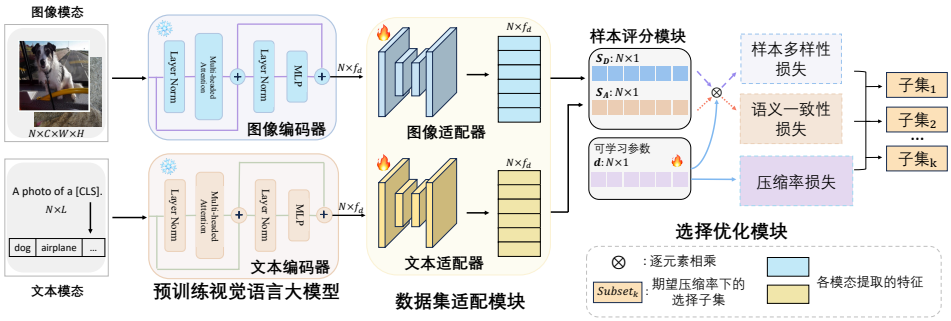


图 5-2　本章提出的方法包含数据集适配、样本评分和选择优化三个模块。数据集适配模块用于学习数据集特定知识。样本评分模块计算两个评分 S_A 和 S_D 以评估样本重要性，最后选择优化模块根据预期选择比例确定最优子集

导致样本评分不准确。因此本章在图像模态与文本模态分别引入了维度保持的适配器，记为 A_I 与 A_T。在冻结预训练 CLIP 权重的前提下，对两个适配器进行有监督微调以实现知识迁移。为保持高效性，两个适配器均采用简单的多层感知机结构。

具体而言，微调过程采用 InfoNCE 损失[158-159]，该损失函数通过最大化图像与文本表示之间的互信息来优化适配器的表现。文本表示采用提示语 “A photo of [CLS]” 描述类别信息，其中 [CLS] 代表对应类别标签。通过此方式，适配器可以有效对齐并捕获双模态的相关特征，同时增强模型区分正负样本对的能力，从而提升针对目标数据集深度表示的鲁棒性和准确性。

5.2.3　样本评分

对于分类数据集，训练样本的学习过程本质上与获取对应类别的知识密切相关。因此，一个样本是否 “代表性强”，可以通过其与类别语义的对齐程度来衡量；同时，子集是否 “覆盖性强”，则取决于样本的多样性。为此，本节设计了两个互补的指标：语义对齐评分和样本多样性分数。

（1）语义对齐分数　用来衡量训练样本与其对应类别之间的语义相似性。具体而言，由于图像和文本特征处于同一嵌入空间[153]，SAS 通过计算嵌入图像与对应文本深度描述之间的余弦相似度得到。第 i 个样本的 SAS 定义为：

$$S_{Ai} = \ell_{cos}(A_I(E_I(I_i)), A_T(E_T(T_i))), \tag{5-1}$$

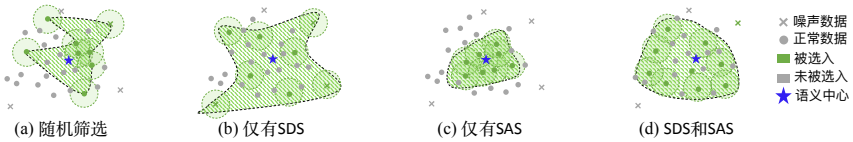


图 5-3 SAS 和 SDS 的有效性示意图。圆圈和叉号分别代表正常样本和噪声样本，不同颜色对应选择结果。SDS（b）选择多样样本但可能包含噪声。SAS（c）可避免噪声样本但可能丢失广泛类别信息。同时使用两个评分（d）能在保持高多样性的同时选择具有类别代表性的样本

其中 $\boldsymbol{I}_i$ 表示第 i 个样本，$\boldsymbol{T}_i$ 为 $\boldsymbol{I}_i$ 对应类别的文本描述，$\boldsymbol{E}_I$ 和 $\boldsymbol{E}_T$ 分别为预训练的图像和文本编码器。高 SAS 值的样本更能代表其所属类别，有助于减少噪声样本的影响。

（2）样本多样性分数 衡量样本在同类样本空间中的“局部稀疏性”。对于经过选择的数据集，数据量的减少可能限制所选数据的多样性。正如在第三章所探讨的那样，这对训练数据集至关重要[121]。为解决这个问题，本节引入多样性视角来全面评估训练数据的效果。SDS 定义为每个样本与其同类 k 个最近邻样本的平均距离：

$$\boldsymbol{S}_{Di} = \frac{1}{k} \sum_{\boldsymbol{I}_j \in \mathrm{KNN}(\boldsymbol{I}_i)} \| A_I(E_I(\boldsymbol{I}_i)) - A_I(E_I(\boldsymbol{I}_j)) \|, \tag{5-2}$$

其中采用 KNN 算法获取每个样本的邻居样本，距离度量基于 ℓ_2 范数，k 通常设置为每类样本数的 10%。SDS 可理解为特征空间中训练样本的局部密度：若某样本拥有更多距离相近的邻居（即较低 SDS），其训练效果更容易被其他邻居样本替代，因此选择具有较高 SDS 的样本通常更有利。

图 5-3展示了 SAS 和 SDS 的效果。SDS 有助于样本多样性，但所选数据点可能包含噪声（图 5-3(b)）；SAS 能选择语义合适的样本（这些样本基本围绕样本中心分布，图 5-3(c)），但这些样本可能过于集中而缺乏多样性。当同时使用 SDS 和 SAS 时（图 5-3(d)），能够以更少数据覆盖整个类别空间，选择既具有语义代表性又保持多样性的样本，从而提升数据选择的有效性。

5.2.4 选择优化模块

本方法通过 SGD 多目标优化确定选择结果，即确定给定选择比例下的最优样本组合是什么，该优化模块可以提高计算效率并快速收敛。具体而言，本节引

入样本级参数 $\boldsymbol{d}$ 表示选择决策，其中元素 1 表示选中，0 表示未选中。虽然严格二值参数因缺乏梯度而难以在神经网络中优化，但采用 sigmoid(·) 函数将 $\boldsymbol{d}$ 的连续值推向近似二值化，优化后，$\boldsymbol{d}$ 被严格二值化以明确指示最终样本选择。初始化时，$\boldsymbol{d}$ 所有元素设为 1。

为监督优化过程，引入三个损失项。第一项 $\mathcal{L}_{sa}$ 优先选择高 SAS 样本作为代表性样本，其定义为：

$$\mathcal{L}_{sa} = -\frac{1}{N}\sum_{i}^{N} \text{sigmoid}(\boldsymbol{d}) * \boldsymbol{S}_{Ai}, \tag{5-3}$$

其中 N 为样本总数。$\mathcal{L}_{sa}$ 惩罚低 SAS 样本，鼓励选择语义对齐更好的样本。

第二项损失 $\mathcal{L}_{sd}$ 鼓励选择具有更高 SDS 的多样性样本：

$$\mathcal{L}_{sd} = -\frac{1}{N}\sum \text{sigmoid}(\boldsymbol{d}) * \boldsymbol{S}_{Di}. \tag{5-4}$$

为缓解群体效应，针对特定选择比例优化所选数据集，以识别最优子集。引入选择损失项 $\mathcal{L}_{s}$ 确保选择过程符合目标比例。由于从连续实值参数优化中推导精确选择率较为困难，本节采用直通估计器（Straight Through Estimator，STE）[160]估计实际选择率并计算梯度。STE 允许梯度在反向传播过程中通过离散决策，有效结合连续和二值参数的优势实现高效优化和准确样本选择。$\mathcal{L}_{s}$ 定义为：

$$\mathcal{L}_{s} = \sqrt{\left[\frac{1}{N}\sum_{i} \text{STE}\left[\mathbb{I}\left(\text{sigmoid}(\boldsymbol{d}_i)_{\text{i}} > 0.5\right)\right] - s_r\right]^2}, \tag{5-5}$$

其中 $\mathbb{I}$ 为指示函数，s_r 表示预期选择比例。该损失项引导参数 $\boldsymbol{d}$ 趋向近二值化，确保 1 的数量与预期样本量一致。由于选择通过自适应优化指导，最终选择比例可能略微偏离目标值。为最小化偏差，使用阈值 θ（本工作中设为 5×10^{-4}）约束 $\mathcal{L}_{s}$，确保实际选择比例与期望值差异小于 $\pm 0.05\%$。最终整体损失函数为：

$$\mathcal{L} = \mathcal{L}_{sa} + \alpha\mathcal{L}_{sd} + \beta\mathcal{L}_{s}, \tag{5-6}$$

其中 α 和 β 为调节损失项间数值差异的系数，可根据不同任务场景或数据集方

算法 2 总体流程

Require: 数据集 $\mathcal{D}$，训练样本总数 N，训练总轮数 T，选择比例 s_r，阈值 θ，预训练图像与文本编码器 E_I 和 E_T，微调后的图像与文本适配器 A_I 和 A_T

初始化选择向量 $\boldsymbol{d} \leftarrow \mathbf{1}$
初始化得分向量 $\boldsymbol{s} \leftarrow \mathbf{0}$
for i=0:N-1 **do**
    根据公式 (5-1) 计算语义对齐分数 (SAS) $\boldsymbol{S}_A$
    使用 KNN 算法计算样本 $\boldsymbol{x}_i$ 的 K 个邻居 $\boldsymbol{x}'$
    根据公式 (5-2) 计算样本多样性分数 (SDS) $\boldsymbol{S}_D$
end for
for t=0:T-1 **do**
    根据公式 (5-3) 计算损失 $\mathcal{L}_{sa}$
    根据公式 (5-4) 计算损失 $\mathcal{L}_{sd}$
    根据公式 (5-5) 计算子集比例损失 $\mathcal{L}_s$
    根据公式 (5-6) 计算总损失 $\mathcal{L}$
    使用带动量的 SGD 优化器更新 $\boldsymbol{d}$
    if $\mathcal{L}_s \leqslant \theta$ **then**
        终止循环
    end if
end for

Ensure: 最终选择结果 $\boldsymbol{d}$

便设置。完整流程见算法 2。

复杂度分析 本方法包含三个主要组件：1）数据集适配涉及微调图像和文本适配器。由于适配器由简单线性层构成，参数量小，前向和反向传播计算高效；2）样本评分过程中，计算 SAS 和 SDS 的复杂度分别为 $O(N)$ 和 $O(Kf_d)$（K 为类别数，f_d 为特征维度，通常为 512）。KNN 算法的复杂度为 $O(N_k f_d)$（N_k 为每类样本数）。鉴于 K 和 f_d 为常数且 N_k 远小于 N，该过程总复杂度约为 $O(N)$；3）$\boldsymbol{d}$ 的选择优化是数值优化过程，不涉及深度模型，因此复杂度与参数量成正比，即 $O(|w|) = O(N)$。

5.3 实验验证

5.3.1 实验设置

对比基线方法 为了全面验证所提方法的有效性，本节选择了当前最具代表性且广泛使用的十种数据选择基线方法进行比较：(1) Random，即从数据集中随机采样子集作为基准；(2) MoSo[31]，通过计算样本对经验风险的边际影响来确定

其重要性；(3) Glister[73]，基于子集选择的贪心优化方法；(4) Herding[78]，通过均值匹配原则来挑选代表性样本；(5) Forgetting[30]，依据样本在训练过程中被遗忘的次数来度量其价值；(6) GraNd 和 (7) EL2N[64]，通过梯度范数与误差范数来衡量样本的贡献；(8) Self-supervised Selection (SSP)[82]，借助自监督信号评估样本代表性；(9) CG-Score[74]，结合对比学习和图结构的方法；以及 (10) Moderate-DS[62]，在样本重要性与多样性之间实现权衡。这些方法覆盖了现有数据选择的主要范式，确保实验对比的全面性和公正性。

参数设置　本节提出的方法在参数设置上具有高度的简洁性与可解释性。具体而言，损失函数中多样性约束的权重系数 α 与预期选择比例 s_r 成正比，用于平衡样本多样性的重要性。在所有数据集上，将 α 直接设置为 s_r，从而避免了额外的参数调优开销。另一系数 β 用于调节不同损失项之间的数值差异，其值在所有数据集上均统一设为 2。这种简洁的超参数设计充分体现了方法的泛化性和实用性。

基准数据集和网络模型　为了全面评估所提方法的有效性与鲁棒性，在多个主流公开基准数据集上进行了系统实验，这些数据集在计算机视觉研究中被广泛采用，能够充分检验数据选择方法的普适性与适应性。具体而言，本节选取了三个层次的任务场景：(1) **CIFAR-10/100**[161]：该数据集包含 60,000 张 32×32 尺寸的彩色图像，分别分为 10 类和 100 类，其中 CIFAR-100 的类别粒度更细，对方法的泛化能力要求更高。(2) **Tiny-ImageNet**[132]：该数据集是 ImageNet 的子集，包含 200 个类别，每个类别有 500 张训练样本和 50 张验证样本，图像分辨率为 64×64。其数据规模和复杂度介于 CIFAR 与 ImageNet 之间，非常适合验证方法在中等规模场景下的表现。(3) **ImageNet-1k**[14]：这是最具挑战性的大规模分类数据集，包含超过 120 万张训练图像和 50,000 张验证图像，覆盖 1000 个类别。该数据集对算法的可扩展性、计算效率以及鲁棒性提出了更高要求。

在模型架构方面，不仅关注传统的卷积神经网络（CNN），还扩展到近年来表现突出的视觉 Transformer。具体包括：ResNet-18/50[3]、VGG-16[2]、DenseNet-121[114] 等经典 CNN 网络，以及 Vision Transformer (ViT)[115] 和 Swin-Transformer[162] 等 Transformer 架构。通过在多种异构架构上进行验证，能够更客观地评估方法

的跨架构泛化能力。

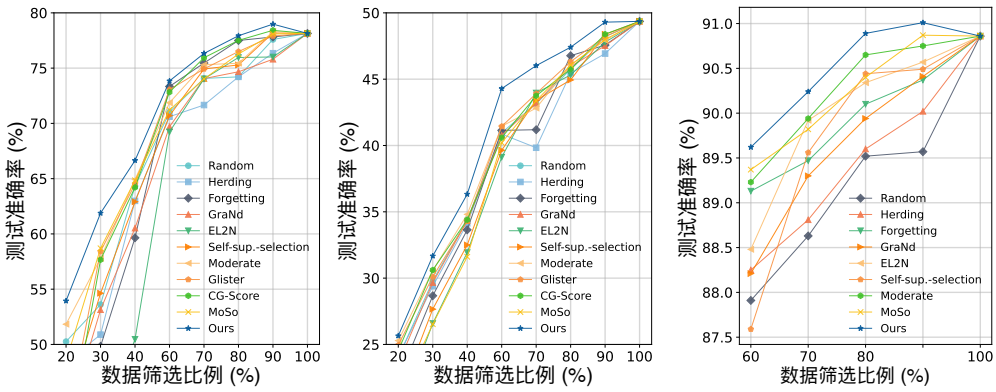
训练设置 在训练配置方面，严格遵循已有研究[62,82,163] 的实验设定，以确保比较结果的公平性和可复现性。

- **CIFAR-10/100**：采用批量大小为 128，优化器为带动量的 SGD（momentum=0.9），权重衰减设为 5×10^{-4}，初始学习率为 0.1，总训练轮数为 200。学习率在第 60、120、160 轮时分别衰减为原来的 1/5。
- **Tiny-ImageNet**：采用批量大小为 256，优化器为 SGD（momentum=0.9），权重衰减设为 1×10^{-4}，初始学习率为 0.1，总训练轮数为 90。学习率在第 30、60 轮时分别衰减为原来的 1/10。
- **ImageNet-1k**：在大规模实验中，参考[54,62,82]，使用 VISSL 库[164] 进行训练。具体配置为：基础学习率 0.01，批量大小 256，优化器为 SGD（momentum=0.9），权重衰减 1×10^{-3}，总训练轮数为 105。

5.3.2 基准数据集上的对比实验结果

本节对所提出的方法与现有最先进的数据选择方法进行了系统性的性能对比。按照已有研究[62,82] 的实验标准，在 CIFAR-100 和 Tiny-ImageNet 数据集上报告 Top-1 分类准确率，在更大规模的 ImageNet-1k 数据集上报告 Top-5 分类准确率。需要特别指出的是，由于计算开销过于庞大，Glister 与 CG-Score 两种方法未被纳入 ImageNet-1k 的对比实验：其中，Glister 依赖于双层优化问题的迭代求解[73]，而 CG-Score 则需要对大规模 Gram 矩阵进行求逆运算，两者在大规模数据集上均难以实际运行，计算代价极为昂贵。

从图 5-4 所示的结果中可以观察到，提出的方法在所有数据集上均取得了最佳性能，并在多个基准上展现出稳定且显著的优势。尤其是在 Tiny-ImageNet 和 ImageNet-1k 这类更具挑战性的中大规模数据集上，本章的方法相较于其他对比基线均取得了明显的性能提升，显示出其在复杂任务和大规模场景中的优越性。相比之下，现有方法在 CIFAR-100 这类小规模数据集上往往只能带来较为有限的增益，而本章的方法在此类任务中同样能够实现更为可观的提升，凸显了其跨规模的普适性与鲁棒性。



(a) CIFAR-100　　(b) Tiny-ImageNet　　(c) ImageNet-1k

图 5-4　在 CIFAR-100 (a)、Tiny-ImageNet (b) 和 ImageNet-1k (c) 数据集上，本方法与多种数据选择基线方法的对比效果示意图

此外，在高选择率（例如 90%）的实验设定下，本章的方法不仅能够在保持较低训练开销的同时接近无损地维持全数据集的性能，甚至在部分场景中超越完整数据集和所有现有基线的表现。这一结果说明，本章提出的方法不仅能够在削减训练成本方面发挥重要作用，同时也具备作为数据清洗工具的潜力，能够有效剔除噪声与冗余样本，从而在保证或提升模型性能的同时，优化数据使用效率。

5.3.3　不同模型架构下的泛化性实验

进一步评估了所选子集在不同选择阶段所使用的深度网络架构上的泛化能力。这一实验的核心目标是验证方法是否能够在跨架构的情境下依然保持稳定和优越的表现，而不仅仅局限于特定的网络结构。

具体而言，在 Tiny-ImageNet 数据集上选取了 VGG-16 与 DenseNet-121 两种经典且结构差异较大的模型进行训练，并在前一阶段通过数据选择得到的子集上进行训练。与 ResNet 系列的主流模型相比，VGG-16 缺乏残差结构，属于相对较浅但参数量庞大的卷积网络；DenseNet-121 则以其密集连接著称，强调特征复用与梯度高效传递。这两类架构的差异为检验数据选择方法的鲁棒性提供了理想的试验平台。

实验结果如表 5-1 所示。整体来看，提出的方法在两种网络架构上均优于所有基线方法，进一步证明了其良好的架构泛化性。值得注意的是，虽然在 DenseNet-121 上，不同方法间的性能差异相对较小，但模型依然保持了稳定的领先地位。

表 5-1 Tiny-ImageNet 数据集上的测试准确率 (%)。实验采用 VGG-16 与 DenseNet-121 架构

方法 / 选择比例	VGG-16				Densenet-121			
	70%	80%	90%	100%	70%	80%	90%	100%
Random	47.39±2.72	49.38±0.23	51.15±0.64	57.23±1.08	59.55±0.20	60.78±0.18	61.03±0.22	62.22±0.23
EL2N	48.30±2.95	48.75±1.65	49.01±1.31	57.23±1.08	59.61±0.00	60.38±0.04	61.16±0.47	62.22±0.23
GraNd	50.79±1.26	46.84±1.38	54.73±0.49	57.23±1.08	59.62±0.02	60.84±0.09	61.10±0.05	62.22±0.23
MoSo	50.47±1.01	50.12±0.83	50.07±0.43	57.23±1.08	59.27±0.33	59.86±0.07	60.00±0.37	62.22±0.23
Herding	48.59±0.07	45.77±0.12	50.77±1.24	57.23±1.08	59.00±0.28	60.03±0.35	61.15±0.12	62.22±0.23
Glister	48.74±2.29	50.05±0.02	49.42±1.81	57.23±1.08	59.98±0.01	60.62±0.34	61.28±0.18	62.22±0.23
CG-Score	48.73±2.70	48.49±1.88	49.62±1.08	57.23±1.08	59.74±0.15	60.55±0.20	61.14±0.11	62.22±0.23
Self-sup. prototypes	48.38±1.38	49.98±1.49	54.71±0.84	57.23±1.08	59.56±0.03	60.22±0.12	60.91±0.29	62.22±0.23
Forgetting	47.50±2.43	48.59±1.77	49.82±0.62	57.23±1.08	58.54±0.15	60.39±0.46	61.12±0.10	62.22±0.23
Moderate-DS	50.78±0.93	49.31±0.41	49.25±0.77	57.23±1.08	59.41±0.18	60.42±0.14	61.44±0.11	62.22±0.23
Ours	**53.40**±3.20	**52.25**±0.58	**56.34**±2.93	57.23±1.08	**60.12**±0.06	**60.93**±0.03	**61.59**±0.03	62.22±0.23

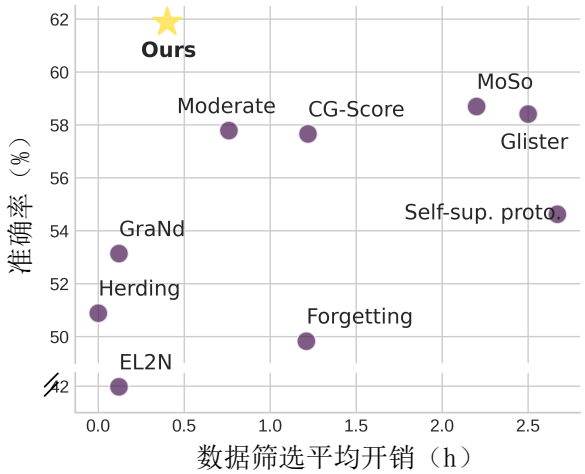


图 5-5 CIFAR-100 数据集上效果与效率的对比分析。结果基于 ResNet-50 架构在 30% 选择比例下测得，实验设备为 4 块 2080TI GPU

在 VGG-16 上，差异更加显著：随着选择比例从 70% 提升到 90%，本章的方法分别比当前最优基线高出 2.61%、2.20% 和 1.63%。这表明所选子集不仅在标准架构上表现优异，在面对结构设计差异较大的模型时，同样能够提供强有力的支持。

这一结果充分说明了本章方法所构建的子集具备高度的通用性与可迁移性，其有效性并不依赖于特定的模型结构，而是能够跨越不同的深度学习架构保持稳定的优势。这一特性在实际应用中尤为重要，因为在多样化的任务场景中，研究者与工程师可能会采用不同类型的模型进行训练。实验结果表明，该方法能够在不同架构下普遍适用，为广泛的深度学习任务提供了稳健的数据选择策略。

5.3.4　训练效率对比

为了进一步评估不同方法在数据选择过程中的效率表现，对比分析了各方法在效果与效率（Effectiveness vs. Efficiency）之间的平衡情况。结果如图 5-5 所示，可以清晰地观察到：首先，在整体趋势上，本章的方法在保持较低选择开销的同时，取得了最高的分类精度，展现出效果与效率兼具的特性。与基于简单启发式或早期训练指标的方法（如 Herding、EL2N 和 GraNd）相比，这些方法的选择成本最低，因为它们往往依赖于预定义指标或在训练早期就直接进行样本筛选。然而，这种方式的准确率往往受限，难以充分挖掘样本的重要性。而本章的方法虽然在选择阶段的开销略高于它们，但换来的是显著更高的精度提升，这种“轻微增加的成本—显著提升的性能”组合更具实用价值。

其次，相较于依赖复杂优化过程的方法（如 MoSo 与 Glister），本章的框架展现出了明显优势。这两类方法需要迭代式地解决复杂的双层优化问题或矩阵反演运算，不仅选择过程耗时，而且在大规模数据集上难以应用。而本章的方法通过结合多模态特征设计高效评分函数，并在选择优化阶段采用高效的梯度优化策略，使得其计算开销显著低于上述优化类方法，同时在性能表现上也取得了更大优势。

综上，实验结果验证了方法在效率与性能之间的卓越平衡能力：既能够避免启发式方法带来的性能不足，又克服了复杂优化方法在大规模场景下效率低下的问题。因此在实际应用中，能够在有限资源环境下快速完成高质量的数据选择，为后续模型训练提供坚实的数据基础。

5.3.5　噪声场景下的鲁棒性

对噪声标签的鲁棒性

在真实场景中，数据集往往不可避免地存在标签噪声，即部分样本的标签被错误翻转。然而，构建既干净又多样化的数据集需要大量时间与成本，因此评估数据选择方法在此类复杂条件下的表现具有重要意义。在本研究中，我们在 CIFAR-100 和 Tiny-ImageNet 数据集上引入对称噪声[165]，并设置 20% 的噪声比例，以模拟真实世界中的标签错误情况，并在这个数据设置下进行选择。

实验结果如表 5-2 所示，提出的方法在噪声标签环境下展现出极强的鲁棒

表 5-2 CIFAR-100 与 Tiny-ImageNet 噪声标签数据集上的实验结果（准确率，%，平均值 ± 标准差）。其中 20% 的标签受到干扰。同时报告了所选 CIFAR-100 数据集中噪声数据比例（%）的数值分析

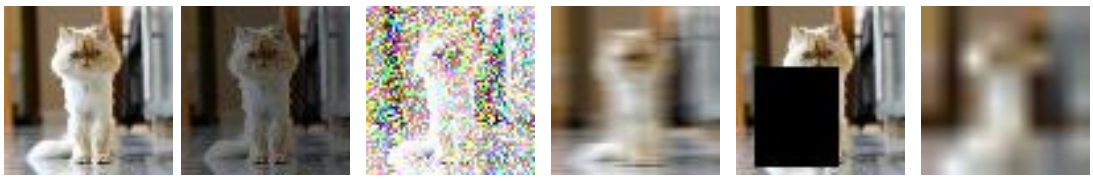
方法 / 选择比例	CIFAR-100 (标签噪声)		Tiny-ImageNet (标签噪声)		噪声比例	
	20%	30%	20%	30%	20%	30%
Random	34.47±0.64	43.26±1.21	17.78±0.44	23.88±0.42	20.80	19.83
MoSo	31.01±0.67	43.73±0.14	21.55±0.37	27.80±0.16	7.78	8.82
Moderate-DS	40.25±0.12	48.53±1.60	19.64±0.40	24.96±0.30	0.30	0.31
Glister	28.51±1.46	43.16±1.31	21.61±0.19	25.45±0.23	21.21	21.95
Herding	42.29±1.75	50.52±3.38	18.98±0.44	24.23±0.29	35.00	30.56
Forgetting	36.53±1.11	45.78±1.04	13.20±0.38	21.79±0.43	23.00	21.76
GraNd	31.72±0.67	42.80±0.30	18.28±0.32	23.72±0.18	5.00	5.14
EL2N	29.82±1.19	33.62±2.35	13.93±0.69	18.57±0.31	22.00	21.80
Self-sup. prototypes	31.08±0.78	41.87±0.63	15.10±0.73	21.01±0.36	21.70	20.21
CG-Score	6.82±1.60	20.07±0.45	8.35±0.65	15.31±0.90	45.09	39.69
Ours	**46.05±0.21**	**58.34±0.36**	**26.09±0.12**	**33.13±0.25**	**0.25**	**0.32**

性，显著优于现有主流方法。在 CIFAR-100 上，相比此前性能最优的基线方法提升超过 10.12%，在 Tiny-ImageNet 上提升 4.41%。进一步地，表中还展示了不同方法选择子集中噪声数据的占比，可以看到本章的方法仅选择了 0.24% 的噪声样本，远低于其他方法。这一结果不仅直接提升了选取数据集的整体质量，也进一步解释了模型泛化性能的显著提升。

这种鲁棒性的根本原因在于公式 (5-1) 中提出的语义对齐分数。在存在噪声标签的情况下，图像内容与标签之间的语义对齐关系被破坏，从而导致 SAS 显著降低。在优化过程中，这类低 SAS 样本被赋予较低的选择概率，从而大幅减少其进入最终训练集的可能性。相比之下，大多数基线方法仅依赖图像模态特征进行选择，缺乏对语义一致性的判别能力，因此在面对标签噪声时往往出现性能退化，甚至部分方法表现还不如随机选择。虽然 Moderate 方法同样能够选择相对较少比例的噪声样本，但其整体性能依旧低于本章的方法。这一差异进一步突显了本章的方法在噪声环境下能够做出更优的选择，有效规避噪声干扰，从而实现更高质量的数据选择与更鲁棒的模型训练。

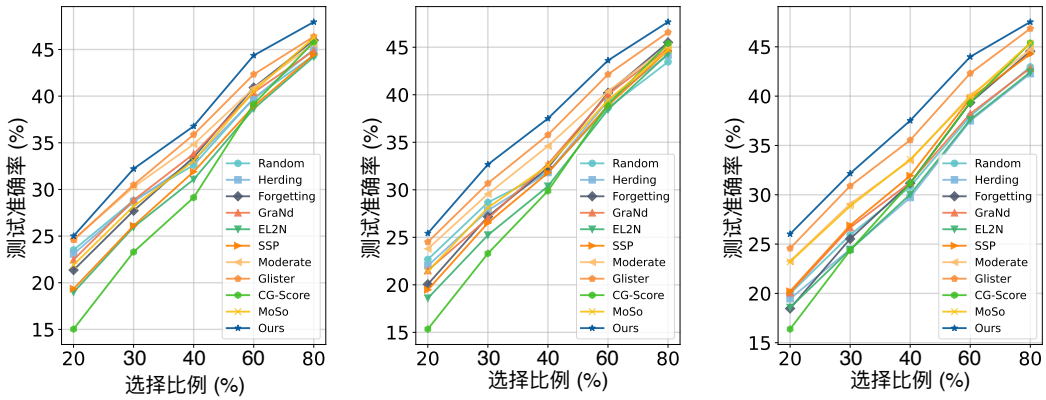
对图像受损的鲁棒性

除了标签噪声，真实世界中还普遍存在图像语义层面的受损，例如模糊、遮挡、分辨率下降、天气因素和噪声等，这些都会显著影响训练数据的质量和模型的鲁棒性。为了模拟这一场景，我们在实验中引入了五类常见的真实图像干扰[62]，包括高斯噪声、随机遮挡、分辨率下降、雾化和运动模糊，并分别设定



(a) 原图　(b) 雾化　(c) 高斯噪声　(d) 运动模糊　(e) 随机遮挡　(f) 分辨率降低

图 5-6　图像损坏类型示意图，包含雾化、高斯噪声、运动模糊、随机遮挡和分辨率降低



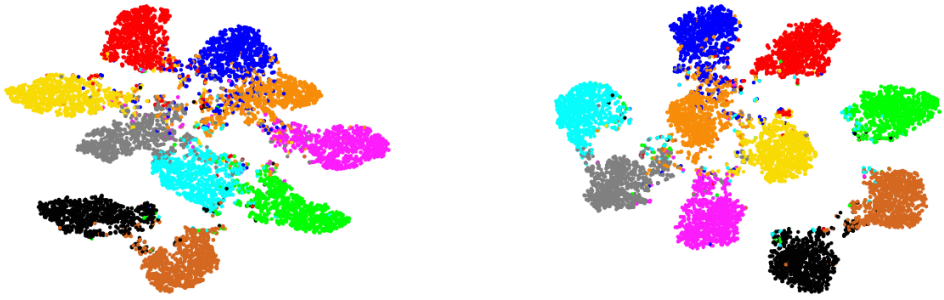
(a) 5% 受损率　(b) 10% 受损率　(c) 20% 受损率

图 5-7　对受损图像的鲁棒性对比实验

5%、10% 和 20% 的受损率进行测试，受损图像示例如图 5-6所示。

实验结果如图 5-7 所示，相较于其他基线方法，本章的方法在不同受损率下均保持了更强的鲁棒性，即便在受损率高达 20% 的情况下，仍能维持较高的泛化性能。其根本原因在于，在选择过程中引入了文本模态的辅助，使得 SAS 在判别图像与类别语义的一致性时能够更敏感地发现受损样本。一旦图像受到破坏，这种语义一致性会下降，从而降低 SAS 值，使这些样本被剔除的可能性大大增加。

相反，像 Forgetting 这类方法往往倾向于优先选择“困难样本”，而受损图像通常更难被正确分类，因此更容易被此类方法选中，导致泛化性能恶化。总而言之，在高噪声和受损环境下，本章的方法始终表现出色，这不仅验证了其鲁棒性，也凸显了多模态信息在数据选择过程中的价值。



(a) Baseline
DI = 3.0×10^{-5}

(b) 90% 数据子集
DI=4.3×10^{-5}

图 5-8 测试集分布可视化。DI：Dunn 指数

5.3.6 泛化性实验

可视化分析

为了更直观地展示所选数据集在提升模型泛化能力方面的有效性，在 CIFAR-10 测试集上进行了可视化实验。具体而言，分别使用完整数据集和经过本章方法筛选后的子集（选择比例为 90%）训练两个模型，并提取它们在测试集上的嵌入表示。随后，采用 t-SNE[144] 对这两组嵌入特征进行降维可视化，以直观展示数据分布和类别区分情况。

从图 5-8 可以清晰地观察到，使用所选子集训练的模型在特征层面展现出更加理想的分布结构。相比于使用完整数据集训练的模型，子集训练得到的模型在嵌入空间上表现为类间分离度更大、类内聚合性更强，这一几何结构上的优化充分表明本章方法在提升判别性表示学习方面的优势。为了进一步对可视化结果进行定量验证，引入 Dunn Index (DI)[166] 作为聚类质量的度量指标。DI 综合衡量了类间距离与类内紧凑性，其数值越大，说明聚类结果越理想。实验结果显示，在移除 10% 数据后，本章方法所选子集训练的模型的 DI 值相较于使用完整数据集训练的模型提升了 43%。这一显著提升不仅从数值层面验证了可视化结果，也进一步表明所选子集在有效过滤冗余和噪声样本后，反而增强了模型的特征判别能力和泛化性能。

表 5-3　ImageNet-1k 数据集上 Swin-T、ViT-B 和 ViT-L 模型在 4 卡 A100 服务器上的性能与成本节约对比 (%)

模型	80%	90%	100%
ViT-B	81.13	**81.46**	**81.46**
ViT-L	84.37	**84.74**	84.59
Swin-T	78.05	**78.63**	78.31
开销节省 (%)	20.62%	10.31%	-

ViT 架构下的泛化性实验

为了进一步验证所选子集在更复杂模型架构上的普适性与适应性，将其应用于基于 Transformer 的先进模型，包括 Swin Transformer、ViT-Base 和 ViT-Large。与前文在 CNN 架构上的实验一致，结果如表 5-3 所示。可以清晰地观察到，本章的方法在这些更复杂的模型架构上依旧能够保持几乎无损的性能，甚至在某些场景下实现超过全量数据训练的效果。这一结果与之前章节的实验结果相互印证，表明所选子集不仅能够在 ResNet、DenseNet 等传统 CNN 架构下实现高效且稳定的表现，同时也能够很好地迁移至 ViT 系列及其改进模型中。在实际应用中，这意味着本章的方法能够跨越不同的深度学习范式（CNN 与 Transformer）。更重要的是，相较于使用全量数据训练的方式，在保持精度几乎不变的前提下，显著降低了训练成本。这一特性对于大规模模型的训练尤为关键，因为 ViT-Large 和 Swin-Transformer 等模型通常对计算资源和存储有极高要求，而所选子集能够有效缓解这一问题，使得在资源受限环境下依然可以高效完成模型训练。

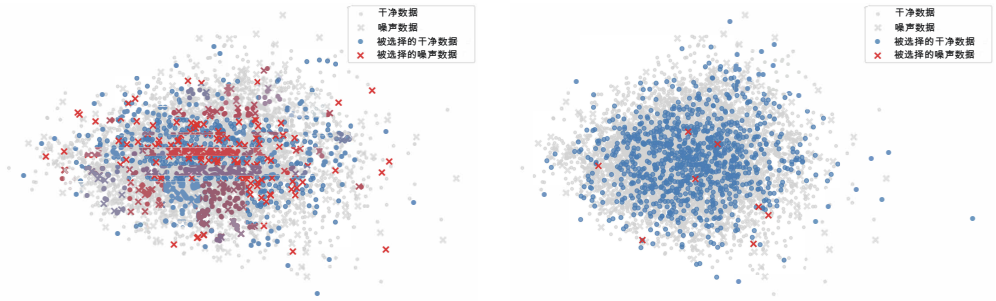
综上所述，本节实验进一步验证了所提出方法在不同类型、不同复杂度的深度网络架构上的广泛适用性和高度泛化能力，说明本章的数据选择方法能够在不同的网络架构下均保持一致的优势。

泛化至更具挑战性的基准数据集

为了进一步检验所选数据集在更复杂场景下的泛化性与鲁棒性，在 ResNet-18 与 ResNet-50 两种典型网络架构上进行了系统性实验。具体而言，模型分别在完整数据集与本章方法选择的子集上进行训练，然后在更具挑战性的 ImageNet 系列衍生测试集上进行评估，包括 ImageNet-Hard[167]、ImageNet-R[168] 和 ImageNet-A[169]。这些数据集专门用于衡量模型在真实复杂分布、分布外样本以

表 5-4 基于更具挑战性的 ImageNet-1k 基准数据集的泛化性能评估

数据集	模型	80%	90%	100%
ImageNet-Hard	R-18	10.89	**11.33**	10.85
	R-50	14.75	**14.98**	14.75
ImageNet-A	R-18	1.65	**2.04**	1.12
	R-50	3.17	**3.31**	3.09
ImageNet-R	R-18	32.99	**33.70**	33.03
	R-50	36.60	**37.11**	36.16



(a) 随机选择 **(b) 本章的方法**

图 5-9 噪声环境下数据选择效果示意图。噪声比例与选择比例均为 20%

及具有较强干扰因素下的表现，被广泛认为是验证深度模型泛化能力和鲁棒性的“硬核”基准。

实验结果如表 5-4 所示，可以观察到，使用本章方法所选子集训练得到的模型在三个挑战性基准上的表现均优于在完整数据集上训练的模型，并在多数情况下取得了显著提升。例如，在 ImageNet-Hard 上，ResNet-18 与 ResNet-50 的准确率均比全量数据训练更高；在 ImageNet-A 上，本章方法的提升更为明显，ResNet-18 提升近 0.9%，而 ResNet-50 的准确率也超过了使用全量数据训练的结果；在 ImageNet-R 上，本章的方法同样实现了稳定的增益。这一趋势表明，所选子集不仅未损失关键信息，反而在面对更具挑战性和分布外的数据时展现了更优的泛化能力。

更重要的是，这些改进是在降低训练成本的前提下实现的。相比于使用全量数据进行训练，本章的方法在减少数据规模和训练时间的同时，依然能够在高难度测试基准上取得更强的鲁棒性和更优的性能。这一结果充分验证了本章方法在真实复杂任务场景下的实用性与高性价比。

表 5-5　对本章的模块在 CIFAR-100（C-100）和 Tiny-ImageNet（T-IN）上的评估结果

	w/o adapter	w/o $\mathcal{L}_{sa}$	w/o $\mathcal{L}_{sd}$	w/o $\mathcal{L}_{s}$	w/o adp&$\mathcal{L}_{sa}$	w/o adp&$\mathcal{L}_{sd}$	w/o adp&$\mathcal{L}_{s}$	w/o adp&$\mathcal{L}_{sd}$&$\mathcal{L}_{sa}$	Ours
C-100	78.20±0.18	78.42±0.46	78.85±0.05	78.48±0.32	78.11±0.16	78.21±0.07	77.10±0.29	77.47±0.31	**78.98±0.09**
T-IN	46.68±0.12	46.79±0.39	49.14±0.09	46.01±0.38	47.23±0.06	46.70±0.33	45.79±0.11	45.69±0.10	**49.30±0.12**

噪声场景下的可视化分析

为了进一步验证所选数据集在复杂环境下的泛化能力，在带有噪声标签的数据条件下进行了可视化实验。具体而言，在 CIFAR-10 数据集上人为注入 20% 的标签噪声，并设定数据选择比例为 20%。随后，基于预训练模型使用 t-SNE 对选择的数据点进行可视化投影。

实验结果如图 5-9 所示：(a) 表示随机选择的结果，可以看到大量噪声样本（红色叉号）被误选进入子集，导致噪声样本在特征空间中分布广泛且难以区分，削弱了模型的判别能力；而 (b) 则展示了本章方法在相同设定下的结果，可以清晰观察到被选择的噪声样本数量极少，同时大部分被选择的样本均为干净样本（蓝色点）。这说明，本章方法在噪声条件下能够更有效地过滤掉错误样本，仅保留具有代表性和高质量的数据，从而获得更鲁棒的特征表示。与前文中在数值实验上的分析一致，该可视化结果进一步证明了所选数据集在复杂真实场景中的泛化性和稳定性优势。

5.3.7　消融实验

数据集适配模块的影响

为了验证数据集适配模块的影响，设计了一个对比实验：在不使用微调后的图像与文本适配器的情况下，直接采用预训练 CLIP 模型的特征来计算 $\boldsymbol{S}_A$ 和 $\boldsymbol{S}_D$。这种方式跳过了自适应阶段，从而能够直观体现数据集适配模块对性能的贡献。

实验结果如表 5-5 所示，在 90% 的选择率下，若不进行数据集适配模块，整体分类准确率均显著下降，尤其是在 Tiny-ImageNet 上，性能下降幅度超过了 2%。这一结果清楚地表明，简单地依赖 CLIP 的预训练特征不足以在新任务上实现良好的泛化能力，适配模块对于弥合预训练数据与目标数据集之间的语义和分布差异至关重要。进一步分析表明，这种差异在与预训练域差距较大的数据集上更加显著。例如，在 CIFAR 系列数据集上，图像尺寸小、分布域与 CLIP 预训练

表 5-6 本章的方法与平均图像特征在不同噪声和选择比例下的噪声抑制性能对比。噪声比例指所选数据集中引入的噪声比率

	噪声比例 (%)	20		50		70	
	选择比例 (%)	20	30	20	30	20	30
图像特征中心	选入噪声比例 (%)	16.39	25.35	20.00	29.95	20.22	30.16
	准确率 (%)	28.42	38.35	16.56	23.19	11.18	14.61
Ours	选入噪声比例 (%)	**0.24**	**0.32**	**0.43**	**0.68**	**0.80**	**4.30**
	准确率 (%)	**46.05**	**58.34**	**52.56**	**60.72**	**51.50**	**56.80**

所用的大规模图文对齐数据存在较大差异。如果缺乏适配，图像与文本模态之间的对齐关系会受到严重影响，导致 $\boldsymbol{S}_A$ 的判别能力下降，进而影响整体选择效果。相反，当引入轻量化的图像与文本适配器后，模型能够更好地迁移和吸收预训练知识，从而保证所计算的语义对齐分数和多样性分数更加可靠，最终带来更高质量的子集选择。

文本模态的影响

在已有方法中[62]，通常将类别的平均图像特征作为原型，并计算嵌入图像与对应原型之间的欧氏距离，用于筛除噪声标签[170]。为了验证文本模态在本章方法中的作用，在实验中也采用了这种基于距离的方式来评估噪声样本，并在公式 (5-1) 中用图像原型特征替代文本特征。

如表 5-6所示，在 20% 噪声比例的情况下，实验结果表明准确率显著下降：在 20% 的选择率下，准确率从 46.05% 降至 16.39%；在 30% 的选择率下，从 58.34% 降至 38.35%。这种大幅度的性能退化充分验证了在数据选择过程中引入文本模态的重要性。文本模态不仅提供了额外的语义信息，还在面对噪声样本时能有效提升判别能力，从而显著增强了数据选择的鲁棒性。

损失项的影响

在表 5-5 中，对公式 (5-6) 中各个损失项及其组合的效果进行了分析。整体来看，完整的损失函数能够实现最高的准确率。

当去除 $\mathcal{L}_{sa}$ 时，选择过程更倾向于保留多样化的样本，但会遗漏一些类别代表性样本，导致模型性能显著下降。这表明 $\mathcal{L}_{sa}$ 对于确保数据集在语义层面上的代表性至关重要。当去除 $\mathcal{L}_{sd}$ 时，选择结果更多地集中在类别中心的样本，即最具代表性的样本。虽然性能下降幅度略小，但由于缺乏足够的多样性，所选数据

集的覆盖性不足，降低了泛化能力。这说明 $\mathcal{L}_{sd}$ 的引入保证了数据集在代表性与多样性之间的平衡。当去除 $\mathcal{L}_{s}$ 时，由于无法在预期选择比例下获得明确的二值化选择结果，采用直接对分数 $\boldsymbol{d}$ 排序并选择高分样本的方式。然而，这种方式会退化为完全基于分数的选择方法，无法解决群体效应，从而导致性能显著下降。

综上所述，三个损失项在方法中各司其职：$\mathcal{L}_{sa}$ 保证了语义代表性，$\mathcal{L}_{sd}$ 保证了多样性，$\mathcal{L}_{s}$ 则保证了选择比例的约束和群体效应的缓解。三者协同作用，才能使所选数据集在有效性、鲁棒性和泛化性上同时达到最佳。

5.4　方法讨论与展望

尽管本章提出的自适应数据选择方法在多个基准任务上展现出了显著的优势，但从更高层次的视角来看，仍存在若干值得深入探讨的局限与潜在改进方向。以下从偏差传播、多模态不平衡、以及选择优化机制三个方面展开讨论。

预训练模型偏差的传导风险. 本方法高度依赖于预训练的 CLIP 模型来构建多模态嵌入空间。然而，已有研究表明 CLIP 在语义空间中存在潜在偏差与不均衡问题[155]，例如对某些类别或语义模式的过度强调。这种偏差可能通过样本选择过程被放大并传导至下游模型，从而影响模型的稳健性。未来工作可以考虑通过引入偏差感知的损失函数或再训练策略来缓解此类问题，例如在保持多模态对齐能力的同时引入偏差正则项，或在数据选择过程中动态约束偏差的传播。

模态不平衡数据场景的挑战. 本方法主要针对图像-文本平衡数据集展开研究，在此类场景下，文本模态能够提供有力的类别语义指导。然而在真实应用中，模态往往存在不平衡，例如文本描述简短或含噪，而视觉模态信息丰富，此时过度依赖某一模态会削弱整体选择效果。未来研究可探索模态权重的动态分配机制，根据具体数据分布自适应调整图像与文本模态的重要性。此外，还可结合跨模态数据增强策略，在模态不足时引入合成补充，以进一步提升对不平衡场景的鲁棒性。

选择优化与群体效应的进一步研究. 本章提出的多目标优化机制有效缓解了“群体效应”，即高分样本与低分样本之间的交互对整体性能的影响。然而，

该优化过程仍依赖于手工设定的损失加权参数（如 α 与 β），可能在不同数据集和任务中表现不一。未来可考虑通过元学习（meta-learning）或强化学习框架，自动学习适应不同任务的损失权重，从而实现真正意义上的自适应优化。此外，还可探索在更复杂的下游任务（如检测、分割、多模态生成）中的迁移效果，以进一步验证选择机制的普适性。

5.5 本章小结

本章围绕第三章提出的基于相似性与多样性分析的训练数据研究框架，提出了一种全新的致力于相似度提升的数据选择方法，实现了更加鲁棒且具备良好泛化性的样本筛选机制。该框架包含三个核心模块：数据集自适应、样本评分与选择优化。数据集自适应模块通过引入轻量化适配器有效缓解了预训练模型与目标数据之间的域偏移问题，确保多模态特征的有效迁移；样本评分模块利用语义对齐分数与样本多样性分数双重指标，从语义代表性与特征多样性两个维度全面衡量样本的重要性；选择优化模块则通过多目标优化与比例约束，有效避免了单一评分机制下的群体效应，从而获得兼具代表性与多样性的高质量数据子集。大量实验结果表明，所提出的方法不仅在经典基准数据集上实现了优于现有方法的表现，而且在更具挑战性的场景下，例如噪声场景下，展现出了显著的鲁棒性。特别是，相比依赖单一模态的传统方法，本方法利用多模态信息有效滤除了噪声和异常样本，使得模型在大规模与复杂任务中依然保持稳定的泛化能力。同时，该框架在效率上亦具备较大优势，避免了昂贵的双层优化或大规模矩阵运算，保证了实际应用中的可扩展性。

综上所述，本章提出的数据选择方法为数据高效学习提供了一种新的范式，不仅提升了训练效率与模型性能，还在鲁棒性、公平性以及跨架构泛化方面展现出潜力。然而，静态的数据选择在面对动态训练过程和复杂任务需求时仍存在一定局限。因此，下一章将进一步探讨如何将动态样本选择机制与自适应数据增强策略相结合，构建一种协同式的高效数据训练范式，以更好地支持大规模深度学习模型在复杂多模态场景下的训练与应用。

第六章　基于相似性-多样性联合优化的增强与选择协同研究

6.1　引言

第三章提出了一个统一的训练数据优化目标函数，并进一步基于相似性与多样性构建了分析框架，系统性地揭示了训练数据分布特性与模型性能之间的关系。该框架不仅为理解数据增强和数据选择的作用机制提供了统一视角，也为后续研究奠定了理论基础。第四章和第五章分别聚焦于优化增强分布 q_ϕ 与优化样本分布 p_θ，提出了自适应数据增强方法来提升多样性和基于多模态表征的数据选择方法来提升相似性，展示了两类技术在提升模型泛化能力与训练效率上的潜力。然而，正如第三章所揭示的，相似性与多样性之间存在内在张力：数据选择往往在减少冗余和噪声的同时降低了样本多样性，而数据增强虽然能增加多样性，却可能引入噪声和分布漂移，从而影响模型的稳定性。换言之，单纯依赖选择或增强都难以同时兼顾效率与泛化，这使得如何在统一框架下协同利用两者成为关键挑战。

尤其是，随着模型变得越来越复杂、参数越来越多，为了有效训练这些模型，所需的数据集规模也不断扩大。这一趋势虽然推动了模型性能的提升，但是也限制了模型的训练效率，尤其是面对海量数据时，如何有效处理和利用这些数据成为了亟待解决的问题。此外，真实世界中的数据集通常包含噪声样本，这些无效数据不仅增加了训练成本，还可能严重影响模型的训练效果和泛化能力。具体来说，上一章提出的基于多模态表征学习的数据压缩方法，通过有效考虑数据的语义一致性和局部分布特性，展示了数据压缩在提升训练效率方面的巨大潜力。然而，尽管数据选择方法通过筛选出最具代表性的样本，降低了训练成本且不损害性能，但训练样本量的减少也伴随着样本多样性的损失和信息量的减少，这潜在地会增加模型过拟合的风险，从而可能削弱模型的泛化能力。与此同时，数据增强技术正是为了解决数据多样性不足的问题。数据增强通过对原始样本进行多样化变换，能够有效增加训练数据的多样性，从而减少过拟合的风险并改

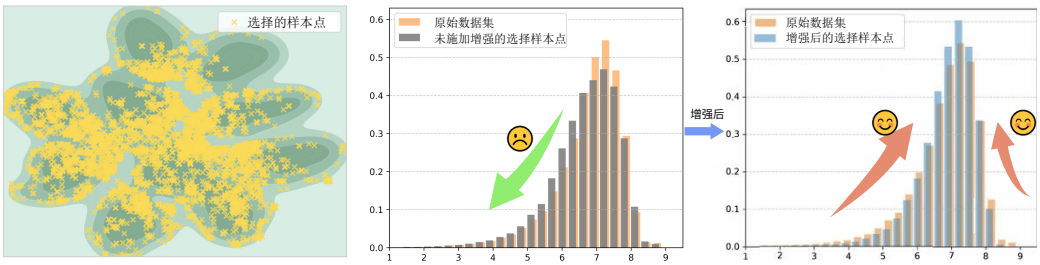


图 6-1　使用 t-SNE 算法对本方法所选数据点分布的可视化（左图），以及在 CIFAR-10 数据集上未增强（中图）与增强后（右图）所选数据的密度直方图对比。选择比例为 10%

善模型的泛化能力。然而，如何有效结合数据选择与增强仍是一个挑战，现有数据选择方法通常优先选择具有代表性和较难的样本，并未专门面向数据增强设计。虽然增强能提升所选数据的多样性并进一步增强模型鲁棒性，但如第三章所讨论的，对复杂样本应用增强可能引入歧义或噪声，增加训练难度[54,56]。因此，将数据选择与增强整合到统一框架中，成为平衡效率与泛化的有潜力的方向。

本研究提出一种面向相似性-多样性联合优化的增强与选择协同框架，该框架以相似性驱动的选择机制保证所选样本的代表性与可靠性，同时结合数据增强补偿选择带来的信息损失与多样性不足，为深度学习模型的高效训练提供了一种新的范式。在模型训练过程中，低密度样本通常对应未充分学习或表征不足的数据点（如分类边界）。对这些样本应用增强变换可强化模型学习并提升鲁棒性。然而，噪声或异常值样本通常也表现为低密度点，这会增加引入噪声的风险。为解决该问题，本章引入了基于预训练多模态模型 CLIP[153]的语义相似性分布。通过优先选择具有高稀疏性和强语义对齐的样本，本章的方法利用密度与语义相似性的联合分布实现有效且鲁棒的样本选择。

如图 6-1左子图所示，所选数据点主要聚集在簇间边界区域。同时，图 6-1中的密度直方图显示，经过增强后数据分布更加均衡：低密度与高密度样本减少，更多数据点向中等密度区域集中。与未使用数据增强的分布相比，重新分布凸显了本章框架增强稀疏区域的能力，从而提升了模型在整个数据分布上的泛化性能。本章在多个基准数据集上的实验结果表明，在保持甚至提升泛化能力的同时，显著加速了训练。例如在 ImageNet-1k 上，本章的方法在取得与使用完整数据集训练相当的性能，将训练效率提升了一倍。此外，本章的框架展现出强大的跨架构和跨场景泛化能力，能有效缓解噪声数据影响，增强实际应用中的适应性。本章的主要贡献如下：

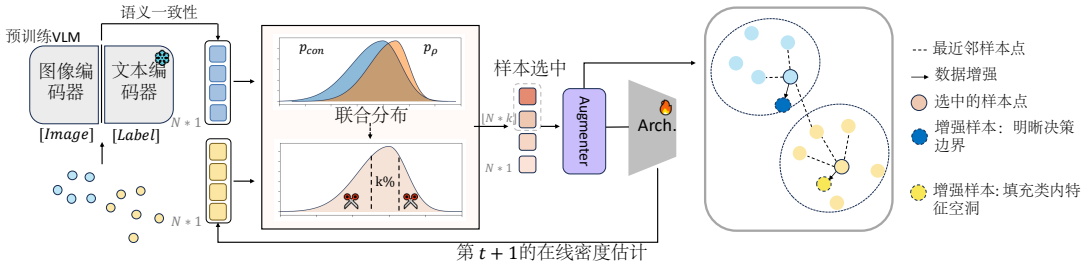


图 6-2 本研究所提出数据训练方法的框架示意图：本方法的核心思想是构建融合密度分布与语义一致性分布的联合分布，从而优先选择低密度且语义一致的样本。经过数据增强后，簇内区域的增强稀疏样本有助于填补表征不足的空间，而位于簇间决策边界附近的样本则能更清晰地区分分类决策，从而提升模型泛化能力

- 提出了一种创新的增强与选择协同优化框架，从相似性-多样性联合建模的角度统一两者。
- 引入了基于密度和语义相似性的联合分布，实现高质量样本选择与增强的互补协同，降低噪声与歧义风险。
- 在大规模基准数据集上的实验验证了方法在效率与泛化之间取得的优越平衡，并证明其在复杂噪声场景下的鲁棒性和跨域适应能力。

6.2 本章工作

6.2.1 方法概述

如图 6-2所示，本章提出了一种新的数据训练框架，结合了动态数据选择和增强技术，以同时提高训练效率和模型的泛化能力。与以往方法单独关注样本选择或增强不同，本章的方法在同一框架下引入两个互补的分布：其一是由任务特定模型在训练过程中动态估算得到的密度分布，用于识别欠表示或位于决策边界附近的样本；其二是通过冻结的预训练多模态模型计算得到的语义一致性分布，用于衡量图像与对应文本标签之间的对齐程度。低密度区域往往包含欠学习的样本，这类样本若能通过增强补充，将有助于模型更好地学习判别特征。然而，这些区域也可能混杂噪声或模糊实例，直接使用会显著增加训练复杂性。为此，语义一致性分布作为重要的补充，可以有效过滤掉语义对齐性较弱的样本。通过联合建模密度与语义一致性，本章提出的框架能够构建一个同时刻画样本信息量与语义正确性的联合分布。在训练中，具有较高联合分数的样本将被优先选择并进行增强，从而兼顾样本的代表性和鲁棒性。

6.2.2　基于密度分布的样本选择

设数据集为 D，服从分布 $P(D)$。在训练的第 t 次迭代中，动态数据选择模块旨在从 D 中选取一个大小不超过 k 的子集 $\hat{D}_t$，以最小化在分布 $P(D)$ 下的期望损失，即优化目标为：

$$\hat{D}_t = \underset{\hat{D}_t \subseteq D, |\hat{D}_t| \leqslant k}{\arg\min} \mathbb{E}_{z \sim P(D)} \left[\mathcal{L} \left(z, \hat{\theta}_{\mathcal{A}(\hat{D}_t)} \right) \right] \tag{6-1}$$

其中，z 表示测试样本，$\mathcal{L}$ 为损失函数，$\hat{\theta}_{\hat{D}_t}$ 为在 $\hat{D}_t$ 上最小化经验风险的参数，$\mathcal{A}$ 表示应用于选定子集的增强操作。

在实际数据分布中，样本在特征空间中往往呈现明显的非均匀性：高密度区域对应于冗余样本，低密度区域则常包含决策边界附近或长尾类的代表性样本。传统随机采样策略对两者一视同仁，容易导致模型在主流分布上过拟合而在边界区域欠学习。因此，本节提出的基于密度分布的样本选择策略通过优先挑选低密度区域样本来改善模型的判别能力。具体而言，低密度样本往往往往对应于欠学习或表示不足的数据点，例如决策边界附近的样本。这类样本经过选择并施加增强操作后，可以更好地帮助模型学习判别特征，从而提升跨区域的泛化能力。通过增强这些稀疏样本，弥补了其在原始训练数据中的不足，使模型在面对不同数据分布时表现更加稳健。因为数据选择是在训练过程中同步进行的，为了在在线训练过程中高效估算样本的密度分布，本节采用基于 HNSW（Hierarchical Navigable Small World）[171]的近似最近邻搜索结构以保证计算效率。HNSW 将样本组织为多层小世界图结构，在高层执行粗粒度搜索，在底层执行精细匹配，从而在保证较高检索精度的同时显著降低时间复杂度。与传统的搜索不同，HNSW 的检索复杂度接近对数级 $O(\log N)$，能在大规模数据集上高效完成密度估计，因此非常适合于在线动态选择任务。对于每个样本 x，找到其最近的 k 个邻居，记为 $NN(x)$，样本的密度通过计算 x 与其邻居的 ℓ_2 距离的均值来估算：

$$\rho_{x_i} = \frac{1}{k} \sum_{j \in NN(x_i)} ||x_i - x_j||, \tag{6-2}$$

较高的 ρ_{x_i} 值表示该样本处于低密度区域，即该样本与周围样本差异较大，代表潜在的决策边界或语义长尾区域。为了便于后续联合建模，密度得分经过 Min-

Max 归一化后转化为密度概率分布 $p_\rho(x)$。在密度驱动的选择阶段，增强策略采用局部邻域约束增强（locality-preserving augmentation）。即仅对低密度区域内的样本进行轻量级增强，如仿射扰动、随机裁剪或微小噪声注入，以生成邻近样本。这种设计兼顾了两点优势：1）局部结构保持：增强样本与原始样本保持一致的语义结构，避免破坏低密度区域的决策边界形态；2）有效填充稀疏区域：在特征空间稀疏区域生成更多邻域样本，减轻欠表示问题，从而提升模型在边界附近的泛化稳定性。

然而，低密度区域并非总是高价值样本的代表，它可能同时包含噪声、错误标注或异常数据。如果不加区分地持续选择这些异常点，将导致训练不稳定甚至性能退化。为此，本节进一步在密度选择的基础上引入语义一致性约束：通过预训练的多模态模型（如 CLIP）计算样本的语义一致性分布 $p_{con}(x)$，并仅在 $p_{con}(x)$ 超过阈值的情况下才执行增强操作。这种密度-语义双约束的选择机制确保了模型能够专注于语义合理但代表性不足的样本，从而在保证训练效率的同时显著提升鲁棒性。

6.2.3 跨模态一致性驱动的鲁棒选择

在真实世界数据中，噪声数据普遍存在，可能来源于错误标注、图像损坏或异常样本。这类噪声本质上表现为样本 x 与其标签 y 之间语义内容的不匹配，即视觉信息与其对应语义描述之间缺乏一致性。若仅依赖单一模态的特征分布（如图像空间密度或损失信号）进行样本筛选，模型往往难以识别这些语义不一致的噪声样本，从而导致模型在训练过程中引入偏差甚至过拟合。为提升样本筛选的鲁棒性，本节在密度驱动选择的基础上，引入了跨模态一致性约束，通过语义对齐机制提升数据选择的可靠性与稳定性。

密度分布反映了数据的结构性信息，而跨模态一致性反映了样本的语义合理性。二者结合可实现结构-语义双结构筛选机制：低密度区域有助于挖掘判别性样本，而跨模态一致性约束则能进一步提出语义异常的噪声样本，从而实现对高价值训练样本的更精细控制。为了度量样本的语义对齐程度，本节采用预训练的 CLIP 模型，将图像和文本嵌入到共享的多模态空间中，并在该空间中进行语义对齐评估。CLIP 模型在大规模图文配对数据上预训练，通过联合优化视觉编码器 $E_I(\cdot)$ 与文本编码器 $E_T(\cdot)$，使得语义一致的图像-文本对在特征空间中接

近，而不相关的对则远离。因此，样本 (x_i, y_i) 的跨模态一致性可通过余弦相似度量化：

$$con(x_i) = \ell_{cos}\big(E_I(x_i), E_T(y_i)\big), \tag{6-3}$$

其中 E_I 与 E_T 分别表示冻结的图像与文本编码器。较高的 $con(x_i)$ 值表示更强的语义一致性，即图像内容与标签语义更加匹配。

尽管 CLIP 在通用图文任务上具备较强的零样本能力，但直接应用于特定领域（如医疗影像）时，仍可能面临分布偏移问题。为此，本节在冻结 CLIP 主干的基础上，为视觉与文本编码器引入轻量化适配器（Adapter）结构[156]，通过小规模参数更新实现领域自适应。这一设计既保留了 CLIP 的通用语义对齐能力，又避免了全量微调带来的高昂计算开销。适配器仅在领域特征空间内调整跨模态特征的分布，使一致性度量在特定领域下更具鲁棒性。通过对所有训练样本计算跨模态相似度，可得到一致性得分集合 $con(x_i)$。然后，通过 Min-Max 缩放将一致性得分归一化为一致性分布 $p_{con}(x)$，其中较高的值表示更强的语义对齐。由于图像-标签对的对齐来自于预训练的视觉-语言模型，并且在训练过程中保持独立，因此可以预先计算一致性分布，在样本选择时直接使用，避免了在线训练中额外的计算开销。

6.2.4　数据增强器

为了结合结构稀疏性和语义一致性，定义了一个联合分布，将密度分布和一致性分布结合起来：

$$p_{sel}(x_i) = p_\rho(x_i) * p_{con}(x_i). \tag{6-4}$$

这里，p_ρ 随着模型训练的进行而动态变化，反映样本在特征空间中的稀疏程度。$p_{con}(x)$ 来自跨模态一致性估计，反映样本语义内容与标签之间的匹配程度。通过联合分布，本节能够同时捕捉样本的信息量与可靠性。在训练过程中，具有较高联合分布得分的样本被优先选择并施加数据增强操作，从而在保证语义合理性的前提下提升样本多样性。

在增强策略上，采用 TrivialAugment 作为增强器，它被广泛使用并提供了一种计算高效的增强策略。如表 6-1所示，在增强过程中，每个图像仅应用一个轻量级变换。这带来了两个关键优势：一是对在线训练过程的计算开销几乎可以忽

表 6-1 数据增强操作及其幅度范围

增强操作	取值范围	基于幅度
恒等变换（Identity）	-	×
X 方向剪切（ShearX）	[0.0, 0.99]	✓
Y 方向剪切（ShearY）	[0.0, 0.99]	✓
X 方向平移（TranslateX）	[0.0, 32.0]	✓
Y 方向平移（TranslateY）	[0.0, 32.0]	✓
旋转（Rotate）	[0.0, 135.0]	✓
亮度调整（Brightness）	[0.0, 0.99]	✓
色彩调整（Color）	[0.0, 0.99]	✓
对比度调整（Contrast）	[0.0, 0.99]	✓
锐度调整（Sharpness）	[0.0, 0.99]	✓
位深压缩（Posterize）	[2, 8]	✓
曝光反转（Solarize）	[255.0, 0.0]	✓
自动对比度（AutoContrast）	-	×
直方图均衡化（Equalize）	-	×

略不计，二是由于每个图像仅进行一次轻微的变换，增强后的样本保持在原始局部特征空间内，非常适合动态数据选择框架的目标：填补类内空隙或增强决策边界。因此，选定样本的连贯性得以保持，同时稀疏区域中的数据多样性得到了增强。最终，使用这些增强后的样本进行训练可以显著提高模型的表现。

6.2.5 复杂度分析

在本方法中，主要的计算开销来源于密度估计。本节采用基于 HNSW 图的近似最近邻搜索结构，其查询和更新的时间复杂度均为 $\mathcal{O}(\log(n))$，其中 n 为数据点的总数。若设训练总轮数为 T，则整体复杂度为 $\mathcal{O}(T \cdot \log(n))$。在常见的大规模训练场景中，通常满足 $T \ll n$，因此方法整体复杂度仍保持在 $\mathcal{O}(\log(n))$ 的量级，保证了良好的可扩展性。此外，数据增强作为模型训练中的标准步骤，本身属于轻量化操作，对整体训练效率的影响几乎可以忽略。因此，在集成数据选择与增强的情况下，本节的方法依然能够在大规模数据集上保持较高的计算效率，兼顾理论上的可扩展性与实践中的可落地性。

6.3　实验验证

6.3.1　实验设置

数据集与网络结构　为全面验证所提出方法的有效性，在多个经典基准数据集上开展实验，涵盖不同规模与难度。具体包括 CIFAR-10/100[161]、Tiny-ImageNet[132] 和 ImageNet-1k[14]。此外，为评估方法在复杂环境下的鲁棒性，进一步在带有噪声标签的数据集上进行实验。为了深入考察方法的泛化能力，还扩展到更具挑战性的测试集，包括 ImageNet-A/O[169]、ImageNet-Hard[167] 和 ImageNet-R[168]。在网络结构方面，不仅在常见的 ResNet 系列（ResNet-18、ResNet-50）上进行验证，还在更先进的 Transformer 架构（ViT-B/L、Swin-Transformer）上进行实验，以系统评估本方法在不同深度模型上的鲁棒性与可扩展性。

对比方法　为了体现所提出方法的优势，选取了当前具有代表性的静态与动态数据选择方法进行对比，涵盖了广泛的主流思路。静态选择方法包括 Random、EL2N[64]、GraNd[64]、Herding[78]、Forgetting[30]、Moderate-DS[62]、Self-sup. Prototypes[82]、MoSo[31] 和 DP[61]。动态选择方法则包括 UCB[65]、ϵ-Greedy[65]、Glister[73] 以及 InfoBatch[33]。这些方法基本覆盖了目前数据选择与动态数据剪枝领域的代表性基线。

实验配置　在实验配置方面，严格遵循已有工作的设置[33,62]，以确保对比的公平性。具体而言，模型训练均采用 SGD/LARS 优化器，动量设为 0.9，权重衰减为 $5e^{-4}$，并配合余弦退火调度策略与 OneCycle 学习率调度器。为了保持方法的一致性，在框架中采用 TrivialAugment[53] 作为数据增强策略。同时，参照 InfoBatch[33]，在动态数据剪枝过程中引入了 annealing 和 re-scaling 技术，以保证各方法在动态数据规模变化下的对比公平性。

此外，在所有数据集上，使用 InfoNCE 损失对适配器进行微调，微调轮数为 15，以保证多模态嵌入的有效性。对于 InfoBatch，由于其采用软剪枝策略，动态选样数量可能随训练过程而变化，因此在报告其结果时保持与所提方法相同的 forward pass 数量，以确保实验比较的公平与一致。

表 6-2 与最先进基线方法的准确率 (%) 对比。所有方法均在 CIFAR-10/100 数据集上使用 ResNet-18 架构，在 Tiny-ImageNet 数据集上使用 ResNet-50 架构进行训练。需要注意的是，由于部分方法缺乏开源代码和参数设置，无法重现实验结果。Random* 表示每轮训练周期中随机选择样本

Dataset	CIFAR-10			CIFAR-100			Tiny-ImageNet		
Whole Dataset	95.6			78.2			45.0		
Selection Ratio (%)	30	50	70	30	50	70	30	50	70
Random	90.2	92.3	93.9	69.7	72.1	73.8	29.8	37.2	42.2
Herding[78]	80.1	88.0	92.2	69.6	71.8	73.1	29.4	31.6	39.8
EL2N[64]	91.6	95.0	95.2	69.5	72.1	77.2	26.6	37.1	44.0
GraNd[64]	91.2	94.6	95.3	68.8	71.4	74.6	29.7	36.3	43.2
Glister[73]	90.9	94.0	95.2	70.4	73.2	76.6	30.1	39.5	43.9
Forgetting[30]	91.7	94.1	94.7	69.9	73.1	75.3	28.7	33.0	41.2
Moderate-DS[62]	91.5	94.1	95.2	70.2	73.4	77.3	30.6	38.2	42.8
Self-sup. prototypes[82]	91.0	94.0	95.2	70.0	71.7	76.8	27.7	37.9	43.4
MoSo[31]	91.1	94.2	95.3	70.9	73.6	77.5	31.2	38.5	43.4
DP[61]	90.8	93.8	94.9	-	73.1	77.2	-	-	-
Random*	93.0	94.5	94.8	74.4	75.3	77.3	41.5	42.8	43.1
UCB[65]	93.9	94.7	95.3	-	75.3	77.3	-	-	-
ϵ-Greedy[65]	94.1	94.9	95.2	-	74.8	76.4	-	-	-
InfoBatch[33]	94.7	95.1	95.6	76.5	78.1	78.2	42.2	43.2	43.8
Ours	**94.9**	**95.5**	**96.0**	**77.6**	**78.9**	**79.5**	**44.9**	**47.0**	**49.4**

表 6-3 使用 ResNet-50 架构、60% 选择比例时 ImageNet-1k 数据集的实验结果。需要注意的是，由于高昂的计算成本和设备内存成本[62]，Glister 和 CG-Score 方法未报告结果。部分结果引用自[33]。时间指标为壁钟时间；总计算量 (n*h) 表示 GPU 总时数，其中 n 为计算节点数量，实验设备为 8 卡 A100 服务器

Method	Herding	EL2N	GraNd	Forgetting	SSP	Moderate	MoSo	UCB	Infobatch	Glister	CG-Score	Ours	Whole Dataset
Acc. (%)	71.1	72.3	71.0	72.5	70.0	73.1	74.0	75.8	76.5	-	-	**76.9**	76.4
Time (h)	10.5	10.5	10.5	10.5	10.5	10.5	10.5	10.5	10.5	10.5	10.5	10.5	17.5
Overhead (h)	>17.5	>17.5	>17.5	>17.5	>24.0	>17.5	>17.5	0.03	0.0028	-	-	0.53	0.0
Overall (n*h)	>224.0	>224.0	>224.0	>224.0	>276.0	>224.0	>224.0	**84.0**	**84.0**	-	-	88.2	140.0

6.3.2 基准数据集上的对比实验

如表 6-2 所示，在 CIFAR-10/100 上使用 ResNet-18，在 Tiny-ImageNet 上使用 ResNet-50，对比不同选择率下的性能表现。实验结果表明，在仅使用部分训练数据的情况下，本章的方法依然能够保持与完整数据集训练相当的性能。具体而言，在 CIFAR-10/100 上，仅使用 50% 的数据即可达到与全量数据几乎无差别的准确率；在规模更大的 Tiny-ImageNet 上，仅使用 30% 的数据同样能够保持与全数据训练接近的性能水平。相比之下，现有方法通常需要在更高的选择率下才能达到无损训练效果。例如，在 CIFAR-10/100 上往往需要超过 60% 的选择率，在 Tiny-ImageNet 上则需要超过 70% 的选择率才能保证与完整数据集相当的性能。这一对比结果凸显了本章方法在低选择率下依然保持较强泛化能力的优势。

值得注意的是，在相同选择率下，本章的方法在各个数据集上均显著优于现有方法。尤其是在 Tiny-ImageNet 这样更大规模、更加复杂的数据集上，本章的

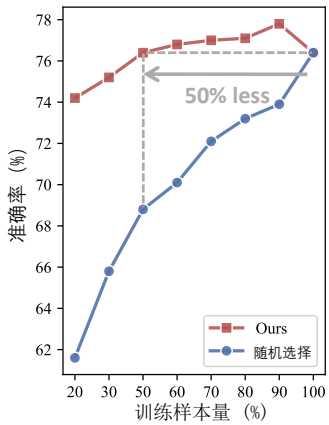


图 6-3　在 ImaegNet-1k 数据集上不同选择比例下的性能表现。实验设备为 4 卡 A100 服务器

方法在相同训练开销下带来了平均至少 2.7% 的性能提升。随着训练数据规模的增加，这种性能差距进一步扩大，充分证明了所提出框架在提升数据效率与模型性能方面的有效性和优越性。

6.3.3　ImageNet-1k 上的实验结果

表 6-3 展示了本章方法在 ImageNet-1k 数据集上的评估结果，选择比例为 60%。实验表明，本章的方法不仅实现了与全量数据训练相当甚至更优的性能，还显著降低了训练成本。具体而言，在减少约 40% 的训练开销的同时，最终准确率相比全量数据提升了 0.5%，对应训练时间减少约 56 个 GPU 小时。这一结果清晰地表明，本章的方法在大规模数据集上能够同时兼顾性能与效率。

值得注意的是，大多数静态数据选择方法通常需要训练额外的代理模型来评估样本在整个训练过程中的影响，因此其计算开销远高于本章的方法。而相比动态剪枝类方法，在效率相当的前提下，最终准确率更高。因此，该实验结果凸显了本章方法在大规模数据场景下的普适性和竞争力。

进一步的分析结果如图 6-3 所示，展示了在不同选择比例下的性能变化趋势。结果表明，本章的方法在仅使用 50% 的训练数据时即可达到无损性能。当数据量降低到 20% 时，虽然准确率略微下降约 2%，但训练开销几乎减少了 80%。相比之下，随机选择方法在低选择比例下会出现显著的准确率下降，而本章的框架依然能够保持稳定的性能。同时，大多数现有基线方法通常需要至少 60% 的训练数据才能达到无损效果，而本章的方法将这一比例进一步降低，证明了其在

表 6-4 使用 ResNet-50 在含噪声与损坏数据的 Tiny-ImageNet 数据集上的实验结果。噪声比例为 20%

方法 / 选择比例 (%)	Noisy		Corrupted	
	20	30	20	30
Random	17.8	23.9	20.0	25.9
Herding	19.0	24.2	35.0	30.6
Moderate-DS	19.6	25.0	23.3	29.1
EL2N	13.9	18.6	18.6	24.4
GraNd	18.3	23.7	20.0	26.7
Forgetting	13.2	21.8	18.5	25.5
Self-sup. prototypes	15.1	21.0	20.2	26.9
CG-Score	8.4	15.3	16.4	24.4
Glister	21.6	25.5	21.2	22.0
MoSo	7.4	11.3	23.1	28.8
Random*	33.8	36.5	35.1	36.9
InfoBatch	34.9	37.1	35.1	38.1
Ours	**35.9**	**39.6**	**39.1**	**42.0**

保证性能的同时，具备更强的数据压缩潜力。

6.3.4 噪声场景下的鲁棒性实验

在真实应用中，训练数据往往不可避免地受到噪声污染，其中既包括错误标注，也包括由环境或采集条件引起的图像损坏。这类噪声会显著降低深度模型的训练效果和泛化能力。为了系统评估本章方法在此类场景中的适用性与鲁棒性，在实验中分别构造了两种噪声环境：（1）标签噪声，通过对一部分样本进行对称性标签翻转，模拟错误标注情况；（2）引入五种典型的现实噪声，包括高斯噪声、随机遮挡、分辨率变化、雾化以及运动模糊。在此基础上，将所提出的框架与现有代表性方法进行对比，以验证其在噪声环境下的有效性和实用性。

如表 6-4 所示，本章的方法在两类噪声场景下均表现出显著优势，始终优于现有的对比方法。在 Tiny-ImageNet 数据集上，即便在 20% 标签噪声的高强度环境下，本章的方法仍能比现有最优方法提升约 4% 的准确率。更重要的是，实验进一步表明，在存在多种图像损坏的情况下，本章的方法依然能够保持稳定的性能优势。

这种优势主要得益于框架的双重机制：一方面，基于密度的样本选择机制能够优先关注低密度、代表性不足的区域；另一方面，多模态语义一致性约束有效过滤掉低语义对齐度的噪声样本。由于噪声样本往往表现为稀疏或低密度，单纯

表 6-5　使用数据增强方法的 Tiny-ImageNet 数据集实验结果。选择比例分别为 30%, 50%, 和 70%

全集	52.0		
选择比例	30%	50%	70%
Random	29.8	37.2	42.2
Herding	31.6	39.2	45.6
EL2N	32.0	40.1	45.9
GraNd	32.2	40.5	46.2
Glister	33.1	42.2	46.5
Forgetting	27.2	36.2	44.2
Moderate-DS	33.8	41.5	46.6
Self-sup. proto.	33.4	41.1	46.6
MoSo	32.6	41.5	45.9
Random*	42.1	43.9	45.2
InfoBatch	43.2	45.9	48.3
Ours	**44.9**	**47.0**	**49.4**

依赖密度选择容易引入不良样本，而通过跨模态一致性校正，有效避免了这一问题。这种协同策略确保增强操作仅作用于真正有价值的样本，从而兼顾了数据效率与鲁棒性。综上所述，通过同时利用低密度样本的潜在信息和多模态语义一致性的纠偏能力，本章的方法为复杂噪声环境下的深度学习提供了一种可靠且高效的解决方案。在大规模真实应用中，这种鲁棒性尤为关键，进一步证明了所提出方法的实践意义和应用潜力。

6.3.5　数据增强对模型性能的影响实验

为了进一步评估本章方法的有效性，在不同选择率下引入数据增强到不同的基线方法中，并与这些方法进行对比，实验结果如表 6-5 所示。结果显示，在所有选择率设定下，本章的方法始终优于其他方法。

值得注意的是，虽然数据增强在整体上提升了所有方法的性能，但本章的方法在不同选择率下依然取得了更为突出的表现。这说明本章的方法不仅仅是对数据选择和数据增强的简单叠加，而是能够有效识别最适合进行增强的样本，从而在保持效率的同时显著提升模型的最终性能。通过对这些关键样本的针对性增强，本章的方法在本质上放大了数据增强的收益，使增强操作更加有针对性和高效。这一机制不仅优化了模型在有限样本条件下的泛化能力，还进一步证明了本方法在数据效率与性能提升上的优势。

表 6-6 基于本方法训练的模型在 ImageNet-Hard、ImageNet-A、ImageNet-R 和 ImageNet-O 数据集上的泛化性能。ImageNet-O 报告 AUPR 指标 (%)，其他数据集报告准确率 (%)。所有模型均采用 ResNet-50 架构

选择比例 (%)	20	30	50	60	70	80	90	全集
ImageNet-A	1.9↓1.2	2.1↓1.0	2.9↓0.2	3.1↑0.0	3.4↑0.3	3.4↑0.3	**3.5↑0.4**	3.1
ImageNet-R	37.2↑1.0	38.5↑2.3	39.3↑3.1	39.8↑3.6	39.9↑3.7	40.6↑4.4	**41.0↑4.8**	36.2
ImageNet-O	15.4↑2.2	15.8↑2.6	16.1↑2.3	16.3↑2.5	16.3↑2.5	16.4↑2.6	**16.5↑2.7**	13.2
ImageNet-Hard	14.2↓0.5	15.3↑0.6	15.9↑1.2	16.5↑1.8	16.7↑2.0	17.2↑2.5	**17.5↑2.8**	14.7

6.3.6 在困难基准数据集上的泛化性实验

为了进一步验证所提出框架的泛化能力，在多个具有挑战性的基准数据集上开展实验，包括 ImageNet-Hard[167]、ImageNet-R[168] 以及 ImageNet-A/O[169]。这些数据集在真实场景中更接近模型部署时可能面临的复杂分布偏移与噪声扰动，因此能够更全面地评估所选数据子集在跨域和高难度任务中的表现。具体而言，在不同选择比例下，利用所提出方法筛选出的数据子集对 ResNet-50 模型进行预训练，并在上述挑战性数据集上进行测试。在评测指标方面，遵循标准设置：对 ImageNet-O 采用精确率-召回率曲线下面积作为度量指标，对其他数据集则采用分类精度作为评价标准。

实验结果如表 6-6 所示。可以观察到，即使在使用更少训练样本的情况下，本章的方法依然能够在这些困难基准上保持甚至提升模型的泛化性能。这表明，基于联合分布驱动的选择与增强机制在减少数据规模的同时，并不会牺牲数据的有效性与完整性。相反，当选择比例逐渐增加时，本章的方法在多个基准上的表现甚至超过了使用完整数据集进行训练的模型，进一步验证了所提出框架在提升数据质量和促进模型泛化方面的优势。总体来看，该结果凸显了本章的方法在复杂与非理想场景中的实用性：通过动态选择低密度且语义一致性高的样本，并结合轻量化增强策略，不仅有效压缩了训练数据规模，还在分布偏移和噪声干扰严重的场景下展现出更优的鲁棒性与泛化能力。这一发现进一步说明，数据压缩与增强的协同优化为高效且可泛化的数据训练范式提供了新的可能性。

6.3.7 在 ViT 架构下的泛化性实验

为了进一步验证所提出方法的可扩展性，在多种先进的深度神经网络架构上进行了实验，包括基于 Transformer 的 ViT-B、ViT-L[115] 和 Swin-T[162]。具体

表 6-7　基于 ViT-B、ViT-L 和 Swin-T 等先进架构在 ImageNet-1k 数据集上的实验结果（使用 4 卡 A100 服务器）。Overhead 表示 GPU 计算时数（h），S_r 指数据选择比例

S_r(%)	50	60	70	80	90	全集
ViT-B	82.6	82.9	83.2	83.2	**83.3**	82.5
ViT-L	85.2	85.3	85.6	**85.7**	**85.7**	84.6
Swin-T	84.1	84.1	84.2	84.2	**84.3**	84.2

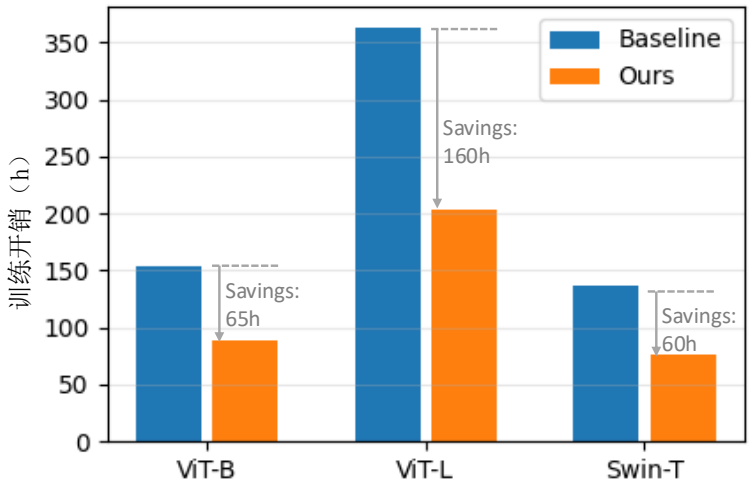


图 6-4　本方法在基于 ViT 的架构上实现无损性能所带来的总体成本节约。实验在 ImageNet-1k 数据集上使用 4 卡 A100 GPU 服务器进行

而言，在不同的选择比例下，利用所提出的动态选择与增强协同框架对这些架构进行训练，并系统性评估其在大规模图像分类任务中的表现。

实验结果如表 6-7 所示。可以观察到，即使在仅使用一半数据（50%）的情况下，本章的方法依然能够实现与完整数据集相当甚至更优的性能表现。这表明所提出的框架不仅适用于传统的 ResNet 系列模型，而且在 Transformer 系列架构上同样具有高度的泛化性和鲁棒性。该结果凸显了本章方法的模型无关性，能够在不同的网络结构和训练范式下保持一致的优势。

此外，在图. 6-4 中进一步展示了这些架构在实际训练中的计算成本，以及通过本框架实现的损失无损情况下的成本节省效果。可以明显看出，在大规模 Transformer 架构上，本方法能够显著减少数百小时的训练开销，而不牺牲模型精度。结合前文在困难基准上的实验结果，这一发现再次强调了方法的高效性与通用性：在保证性能的同时，显著提升了跨架构和跨场景的训练效率。

表 6-8 大规模数据集上模型训练前的微调与特征嵌入开销分析（基于单卡 V100 GPU 服务器）

数据集	Fine-tuning	Embedding	Overall training
Tiny-ImageNet	0.39h	0.03h	21.0h
ImageNet-1k	1.25h	0.17h	84.0h

6.3.8 训练加速表现的进一步分析

尽管本章的方法在在线训练过程中几乎不引入额外开销，但在正式开始目标模型训练之前，仍需要进行一次性的预计算过程，包括基于 CLIP 模型的特征嵌入与适配器的微调。为了全面评估这些潜在的前期开销，我们在表 6-8 中对比分析了适配器微调和 CLIP 特征嵌入所需的时间与资源消耗。

从结果可以观察到，这些预计算操作所带来的开销与目标模型的标准训练相比可以忽略不计。例如，在 CIFAR-100 和 Tiny-ImageNet 上，适配器微调仅需少量的训练轮次，且模型参数规模远小于目标网络，因此整体计算代价极低。同时，特征嵌入仅需执行一次前向传播过程，不涉及梯度反传，因此在大规模数据集上依然可以高效完成。更为重要的是，这些预计算步骤仅在训练开始前执行一次，与选择比例无关。在完成预计算后，所有的实验设置和选择比例均可直接复用生成的特征嵌入与一致性分布，因此不会在后续训练中引入任何额外负担。换句话说，一旦完成该步骤，在线训练的开销几乎与标准训练完全一致。

这一结果进一步说明，所提出的方法在保持性能与泛化优势的同时，其整体计算代价是可控且可扩展的。对于实际应用而言，即使在大规模数据集和复杂模型架构下，该方法的额外成本也完全在可接受范围内，从而保证了其在真实场景中的实用性与可扩展性。

6.3.9 消融实验

在表 6-9 中，我们系统性地评估了所提出框架中不同组件在 Tiny-ImageNet 数据集上、使用 ResNet-50 模型的效果，并在不同数据选择比例下（30%、50%、70%）进行了对比分析。为揭示各模块的作用机理与协同效果，下面将从三个核心模块的角度进行逐点讨论。

密度分布模块 p_ρ。当仅依赖密度分布 p_ρ 进行样本选择时，模型性能显著低于完整框架。这是因为低密度区域样本往往包含稀疏样本和潜在异常点，它们虽然

表 6-9　基于 ResNet-50 在 Tiny-ImageNet 数据集上分析密度分布、一致性分布与增强器的影响。表中为测试准确率（%）。数据选择比例为 30%, 50%, 和 70%

p_ρ	p_{con}	aug.	30%	50%	70%
✓			39.0	40.7	42.5
	✓		42.0	45.6	45.8
		✓	41.6	45.9	48.3
	✓	✓	42.5	46.3	49.3
✓	✓		41.5	43.1	44.3
✓		✓	41.1	45.1	48.5
✓	✓	✓	**43.5**	**47.5**	**50.2**

在特征空间中欠表示，但直接选择这些样本会引入语义歧义，从而影响训练稳定性与模型性能。因此，p_ρ 的引入主要贡献在于提供了结构层面的样本覆盖信息，但它需要额外的语义约束来避免引入噪声。

语义一致性分布 p_{con}。当仅使用跨模态一致性分布 p_{con} 进行筛选时，准确率普遍高于仅用密度的结果，说明语义一致性在过滤异常与噪声样本方面具有显著作用。p_{con} 通过预训练视觉-语言模型（CLIP）计算图像与标签语义的对齐度，有效识别语义不匹配的样本。在过滤异常样本、弱对齐样本的过程中发挥了关键作用，使得最终选择的子集不仅覆盖稀疏区域，还具备较强的语义代表性，从而提高了整体数据质量。尤其在低选择比例（30%）时，p_{con} 的优势更为突出，说明语义约束在数据稀缺时对于稳定训练尤为关键。

数据增强模块。单独引入数据增强模块时，模型性能略高于密度或一致性单独作用的情形。尤其是在低选择比例下，增强模块的作用更加明显。原因在于：通过联合分布选出的样本本身更适合于增强操作，增强能够进一步扩充其局部邻域的数据多样性，有效缓解样本减少带来的信息不足问题。这种“选择-增强协同”机制使得模型能够在有限数据条件下依然保持强泛化能力，并在多个比例下都取得了优于现有方法的结果。然而若缺乏合理的样本筛选（即缺乏 p_ρ 或 p_{con} 的指导），增强可能放大噪声样本的影响，导致模型学习方向偏移。

总体而言，实验结果清晰地表明，框架中的三个模块缺一不可：密度分布 p_ρ 提供了稀疏区域信息，语义一致性分布 p_{con} 保证了样本的语义正确性，而数据增强模块则进一步提升了训练的多样性与鲁棒性。去掉任一模块都会导致性能大幅下降。这一发现不仅验证了所提出方法的合理性，也凸显了不同模块之间的互补性与协同作用，为未来进一步优化提供了坚实基础。

6.4 方法讨论与展望

本章提出了一种新颖的在线数据训练框架，该框架将动态数据选择与数据增强结合起来，在保证高效性的同时，显著提升了模型的泛化能力。尽管方法展现了良好的潜力和实验效果，仍有若干值得进一步探讨和扩展的方向。

目前的框架主要在图像分类任务中进行了验证。尽管实验结果显示出色的效果，但其他任务（如目标检测、语义分割、视频理解、多模态检索）对样本的代表性和多样性要求更高。未来的研究可以将该框架推广到下游复杂任务中，探索在不同任务目标下动态选择与增强的最佳协同方式。

本研究使用了轻量级的 TrivialAugment 来保持增强开销可控。然而，不同任务和数据分布可能对增强策略敏感，简单增强可能无法充分挖掘稀疏区域样本的潜力。未来可以探索可学习的增强策略或任务自适应增强，通过强化学习或元学习等机制，使增强过程动态匹配当前的训练状态。

虽然本章的框架在在线训练中引入的开销可以忽略不计，但在预处理阶段仍需要进行适配器微调与特征嵌入计算。虽然这是一次性操作，但在超大规模数据集下仍可能成为瓶颈。未来可探索更高效的近似语义一致性估计方法，或通过增量更新策略避免每次任务都进行完整的特征重计算，从而进一步降低系统成本。

6.5 本章小结

本章在第三章提出的“基于相似性与多样性的训练数据分析框架”之上进一步展开研究。第三章从全局角度揭示了高效训练数据应当同时兼顾语义相似性与多样性，为后续方法设计提供了理论基础。在此框架下，第四章聚焦于如何通过自适应增强策略动态调节数据的多样性，第五章则提出了结合多模态表征学习的数据选择方法来提升数据的相似性，从而有效提升了样本选择的鲁棒性与泛化能力。本章进一步探索了数据高效训练的新范式，即动态选择与数据增强的协同框架。不同于以往独立地考虑样本选择或数据增强，本章提出的框架能够主动识别最适合增强的样本，并结合语义一致性与结构稀疏性进行联合建模。这样一来，在降低训练数据规模和开销的同时，能够有效弥补因样本压缩而带来的信息量损失和多样性不足，显著降低过拟合风险。通过在 CIFAR-10/100、

Tiny-ImageNet 以及 ImageNet-1k 等大规模基准数据集上的系统实验，我们验证了该框架在多种数据规模和网络架构下的有效性。实验结果表明：在数据量显著减少的情况下，所提出的方法依然能够实现无损的性能表现；在相同数据规模下，方法则能够显著增强模型的泛化性能。进一步地，在含有噪声与损坏的复杂场景中，本章的方法展现出强大的鲁棒性与跨场景适应能力，证明其在真实应用环境中的潜力。

第七章　结束语

7.1　全文总结

深度学习模型规模和算力不断扩张，尤其是大模型的广泛应用，高质量数据却逐渐成为发展的瓶颈，如何更高效、更智能地理解和利用数据，已成为推动人工智能持续发展的关键问题。本文以数据为核心研究对象，从数据智能的视角出发，围绕“理解—增强—选择—协同”的主线，构建了一个统一的数据分析与优化框架，并在此基础上开展了系统研究。主要工作与贡献总结如下：

1. 在第三章中，本文提出了统一的数据优化目标函数，并基于该目标构建了相似性多样性分析框架。该框架从相似性和多样性两个角度系统性地刻画训练数据分布特性，并揭示了其与模型泛化性能之间的内在联系。实验结果表明，不同数据集和训练阶段在相似性与多样性上的偏好差异显著，为理解数据影响机制提供了可解释性工具，也为后续方法提供了理论支撑。
2. 在第四章中，基于上述统一框架，本章针对现有数据增强方法固定或随机调整数据的多样性，导致无法有效提升多样性，甚至引入噪声数据的问题，提出了自适应数据增强方法 AdaAugment。该方法通过强化学习策略网络与目标网络的协同优化，有效提升增强数据的多样性并使其能够自适应匹配模型状态。实验结果表明，AdaAugment 在显著提升模型泛化性能的同时兼顾了计算效率，展现出较强的实用性与扩展性。
3. 在第五章中，基于第三章数据优化框架，我们聚焦于相似性驱动的数据选择研究。该方法引入语义相似性分数与多样性分数作为双重指标，并结合多目标优化与比例约束，有效缓解了群体效应问题。在含噪与域偏移场景下，该方法依然能够稳定筛选具有高度语义代表性的样本，显著提升了数据选择机制的可靠性与训练效率，为后续增强与选择的协同优化奠定了坚实基础。
4. 在第六章中，本章进一步探索在第三章统一框架的指导下，协同优化数据

增强和数据压缩，实现了在统一框架下对两者的有机结合。该方法通过样本稀疏性与语义一致性约束，在选择环节动态保留高价值样本的同时，利用增强机制补偿由于选择导致的信息损失与多样性不足。实验结果表明，该方法能够在效率与泛化之间取得更优平衡，并在复杂噪声环境下展现出优异的鲁棒性与跨域适应性，标志着统一数据优化框架的完整落地。

7.2 未来工作展望

尽管本文在数据相似性-多样性度量、自适应数据增强、多模态数据选择以及二者协同优化方面取得了一定进展，但仍存在诸多值得深入探索的方向：

1. 更通用的数据训练框架：本文在第三章中的数据优化框架分析目前主要聚焦于图像分类场景。未来可以将该框架扩展到更多任务和模态（如文本、语音、多模态推理），并探索跨任务的一致性度量方法；
2. 自适应机制的扩展版本：第四章提出的 AdaAugment 展示了自适应增强在动态调节数据分布方面的潜力，但其研究范围仍主要集中于图像分类等标准任务。未来工作可以面向更大规模的训练场景（如百亿级参数模型或万亿级数据），研究如何提升自适应机制的可扩展性与计算效率，例如通过分布式优化、近似推理或轻量化策略网络来降低开销，从而保证在大规模训练中仍具备实用性。
3. 多模态与复杂场景下的数据选择：第五章表明，多模态表征能够有效解决单模态数据选择在噪声识别与样本代表性方面的不足。但在大规模、多模态异构数据（如视频-文本、图-语言）场景中，如何高效建模样本间的语义交互与群体效应，仍然是一个开放性问题。未来可以结合知识图谱或因果推断方法，提升数据选择的解释性与稳定性。
4. 数据智能系统的工程化与应用落地：本文的研究主要集中在方法论与实验验证层面。未来可在更贴近实际的应用场景中推进数据智能系统的落地，例如智能医疗、自动驾驶、金融风控等领域。同时，在工程实现上需要关注可扩展性与稳定性，例如结合分布式计算平台与高效存储系统，以支撑万亿级别样本的数据优化与训练。

参考文献

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25.

[2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. ArXiv preprint arXiv:1409.1556, 2014.

[3] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[4] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(06): 1229-1251.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30.

[6] 刘文婷, 卢新明. 基于计算机视觉的 Transformer 研究进展[J]. 计算机工程与应用, 2022, 58(06): 1-16.

[7] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report[J]. ArXiv preprint arXiv:2303.08774, 2023.

[8] 仝卫国, 李敏霞, 张一可. 深度学习优化算法研究[J]. 计算机科学, 2018, 45(S2): 155-159.

[9] NICKOLLS J, DALLY W J. The GPU computing era[J]. IEEE micro, 2010, 30(2): 56-69.

[10] ALZUBAIDI L, ZHANG J, HUMAIDI A J, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions[J]. Journal of big Data, 2021, 8(1): 53.

[11] YANG S, XIAO W, ZHANG M, et al. Image Data Augmentation for Deep Learning: A Survey[J]. ArXiv preprint arXiv:2204.08610, 2022.

[12] EpochAI. Trends in GPU Price-Performance[Z]. https://www.lesswrong.com/posts/c6KFvQcZggQKZzxr9/trends-in-gpu-price-performance. 2021.

[13] 翁寿松. 摩尔定律与半导体设备[J]. 电子工业专用设备, 2002(04): 196-199.

[14] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[15] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. 2014: 740-755.

[16] EVERINGHAM M, ESLAMI S A, VAN GOOL L, et al. The pascal visual object classes challenge: A retrospective[J]. International journal of computer vision, 2015, 111(1): 98-136.

[17] 刘德荣, 李宏亮, 王鼎. 基于数据的自学习优化控制: 研究进展与展望[J]. 自动化学报, 2013, 39(11): 1858-1870.

[18] 何清, 李宁, 罗文娟, 等. 大数据下的机器学习算法综述[J]. 模式识别与人工智能, 2014, 27(04): 327-336. DOI: 10.16451/j.cnki.issn1003-6059.2014.04.009.

[19] 张引, 陈敏, 廖小飞. 大数据应用的现状与展望[J]. 计算机研究与发展, 2013, 50(S2): 216-233.

[20] 张欣. 生成式人工智能的数据风险与治理路径[J]. 法律科学 (西北政法大学学报), 2023, 41(05): 42-54. DOI: 10.16290/j.cnki.1674-5205.2023.05.006.

[21] 李继峰, 张成龙, 刘鑫, 等. 面向人工智能的数据治理框架[J]. 大数据, 2025, 11(01): 3-20.

[22] DEVRIES T, TAYLOR G W. Improved regularization of convolutional neural networks with cutout[J]. ArXiv preprint arXiv:1708.04552, 2017.

[23] ZHANG H, CISSE M, DAUPHIN Y N, et al. Mixup: Beyond Empirical Risk Minimization[C]//International Conference on Learning Representations. 2018.

[24] YUN S, HAN D, OH S J, et al. Cutmix: Regularization strategy to train strong classifiers with localizable features[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 6023-6032.

[25] CUBUK E D, ZOPH B, MANE D, et al. Autoaugment: Learning augmentation strategies from data[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 113-123.

[26] CUBUK E D, ZOPH B, SHLENS J, et al. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space[C]//LAROCHELLE H, RANZATO M, HADSELL R, et al. Proceedings of the Advances in Neural Information Processing Systems: vol. 33. 2020: 18613-18624.

[27] SUZUKI T. Teachaugment: Data augmentation optimization using teacher knowledge[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2022: 10904-10914.

[28] LI Y, HU G, WANG Y, et al. DADA: Differentiable Automatic Data Augmentation[J]., 2020.

[29] 高友文, 周本君, 胡晓飞. 基于数据增强的卷积神经网络图像识别研究[J]. 计算机技术与发展, 2018, 28(08): 62-65.

[30] TONEVA M, SORDONI A, COMBES R T D, et al. An empirical study of example forgetting during deep neural network learning[J]. ArXiv preprint arXiv:1812.05159, 2018.

[31] TAN H, WU S, DU F, et al. Data pruning via moving-one-sample-out[J]. Advances in Neural Information Processing Systems, 2024, 36.

[32] YANG S, LI P, SHEN F, et al. RL-Selector: Reinforcement Learning-Guided Data Selection via Redundancy Assessment[J]. ArXiv preprint arXiv:2506.21037, 2025.

[33] QIN Z, WANG K, ZHENG Z, et al. InfoBatch: Lossless Training Speed Up by Unbiased Dynamic Data Pruning[C]//The Twelfth International Conference on Learning Representations. 2024.

[34] XU M, YOON S, FUENTES A, et al. A Comprehensive Survey of Image Augmentation Techniques for Deep Learning[J/OL]. Pattern Recognition, 2023, 137: 109347. https://www.sciencedirect.com/science/article/pii/S0031320323000481. DOI: https://doi.org/10.1016/j.patcog.2023.109347.

[35] SHORTEN C, KHOSHGOFTAAR T M. A survey on image data augmentation for deep learning[J]. Journal of big data, 2019, 6(1): 1-48.

[36] CHEUNG T H, YEUNG D Y. A survey of automated data augmentation for image classification: Learning to compose, mix, and generate[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023.

[37] CHEN P, LIU S, ZHAO H, et al. Gridmask data augmentation[J]. ArXiv preprint arXiv:2001.04086, 2020.

[38] ZHONG Z, ZHENG L, KANG G, et al. Random erasing data augmentation[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 34: 07. 2020: 13001-13008.

[39] SINGH K K, LEE Y J. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3544-3553.

[40] YANG S, LI J, ZHANG T, et al. AdvMask: A sparse adversarial attack-based data augmentation method for image classification[J]. Pattern Recognition, 2023, 144: 109847.

[41] DORNAIKA F, SUN D. LGCOAMix: Local and Global Context-and-Object-Part-Aware Superpixel-Based Data Augmentation for Deep Visual Recognition[J]. IEEE Transactions on Image Processing, 2024, 33: 205-215. DOI: 10.1109/TIP.2023.3336532.

[42] KIM J H, CHOO W, JEONG H, et al. Co-mixup: Saliency guided joint mixup with supermodular diversity[J]. ArXiv preprint arXiv:2102.03065, 2021.

[43] KIM J H, CHOO W, SONG H O. Puzzle mix: Exploiting saliency and local statistics for optimal mixup[C]//International conference on machine learning. 2020: 5275-5285.

[44] KANG M, KIM S. GuidedMixup: an efficient mixup strategy guided by saliency maps[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 37: 1. 2023: 1096-1104.

[45] LI P, LI X, LONG X. Fencemask: a data augmentation approach for pre-extracted image features[J]. ArXiv preprint arXiv:2006.07877, 2020.

[46] INOUE H. Data augmentation by pairing samples for images classification[J]. ArXiv preprint arXiv:1801.02929, 2018.

[47] HARRIS E, MARCU A, PAINTER M, et al. Fmix: Enhancing mixed sample data augmentation[J]. ArXiv preprint arXiv:2002.12047, 2020.

[48] HENDRYCKS D, MU N, CUBUK E D, et al. Augmix: A simple data processing method to improve robustness and uncertainty[J]. ArXiv preprint arXiv:1912.02781, 2019.

[49] HE Y, XIAO L, ZHOU J T. You Only Condense Once: Two Rules for Pruning Condensed Datasets[J]. Advances in Neural Information Processing Systems, 2024, 36.

[50] VERMA V, LAMB A, BECKHAM C, et al. Manifold mixup: Better representations by interpolating hidden states[C]//International conference on machine learning. 2019: 6438-6447.

[51] LIM S, KIM I, KIM T, et al. Fast AutoAugment[C/OL]//WALLACH H, LAROCHELLE H, BEYGELZIMER A, et al. Proceedings of the Advances in Neural Information Processing Systems: vol. 32. Curran Associates, Inc., 2019. https://proceedings.neurips.cc/paper/2019/file/6add07cf50424b14fdf649da87843d01-Paper.pdf.

[52] KURTULUŞ E, LI Z, DAUPHIN Y, et al. Tied-augment: controlling representation similarity improves data augmentation[C]//International Conference on Machine Learning. 2023: 17994-18007.

[53] MÜLLER S G, HUTTER F. Trivialaugment: Tuning-free yet state-of-the-art data augmentation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2021: 774-782.

[54] YANG S, SHEN F, ZHAO J. EntAugment: Entropy-Driven Adaptive Data Augmentation Framework for Image Classification[C]//European Conference on Computer Vision. 2024: 197-214.

[55] CHEUNG T H, YEUNG D Y. Adaaug: Learning class-and instance-adaptive data augmentation policies[C]//International Conference on Learning Representations. 2021.

[56] GONG C, WANG D, LI M, et al. KeepAugment: A simple information-preserving data augmentation approach[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2021: 1055-1064.

[57] LIN C, GUO M, LI C, et al. Online hyper-parameter learning for auto-augmentation strategy[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 6579-6588.

[58] LIN S, ZHANG Z, LI X, et al. Selectaugment: Hierarchical deterministic sample selection for data augmentation[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 37: 2. 2023: 1604-1612.

[59] OH M H, PARK B G. Feature-based Augmentation for Semi-Supervised Learning[J].,

[60] LI B, WU F, LIM S N, et al. On feature normalization and data augmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2021: 12383-12392.

[61] YANG S, XIE Z, PENG H, et al. Dataset Pruning: Reducing Training Data by Examining Generalization Influence[C]//International Conference on Learning Representations. 2023.

[62] XIA X, LIU J, YU J, et al. Moderate Coreset: A Universal Method of Data Selection for Real-world Data-efficient Deep Learning[C]//The Eleventh International Conference on Learning Representations. 2023.

[63] 吴小坤, 邓可晴. 算法偏见背后的数据选择、信息过滤与协同治理[J]. 中国出版, 2024(06): 10-15.

[64] PAUL M, GANGULI S, DZIUGAITE G K. Deep learning on a data diet: Finding important examples early in training[J]. Advances in Neural Information Processing Systems, 2021, 34: 20596-20607.

[65] RAJU R S, DARUWALLA K, LIPASTI M. Accelerating deep learning with dynamic data pruning[J]. ArXiv preprint arXiv:2111.12621, 2021.

[66] YANG S, YE P, SHEN F, et al. When Dynamic Data Selection Meets Data Augmentation[J]. ArXiv preprint arXiv:2505.03809, 2025.

[67] LEI S, TAO D. A comprehensive survey to dataset distillation[J]. ArXiv preprint arXiv:2301.05603, 2023.

[68] LIU S, YE J, YU R, et al. Slimmable dataset condensation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2023: 3759-3768.

[69] DU J, JIANG Y, TAN V Y, et al. Minimizing the accumulated trajectory error to improve dataset distillation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2023: 3749-3758.

[70] ZHANG L, ZHANG J, LEI B, et al. Accelerating dataset distillation via model augmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2023: 11950-11959.

[71] CAZENAVETTE G, WANG T, TORRALBA A, et al. Dataset Distillation by Matching Training Trajectories[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 2022: 4750-4759.

[72] ZHOU Y, NEZHADARYA E, BA J. Dataset Distillation using Neural Feature Regression[J]. ArXiv preprint arXiv:2206.00719, 2022.

[73] KILLAMSETTY K, SIVASUBRAMANIAN D, RAMAKRISHNAN G, et al. Glister: Generalization based data subset selection for efficient and robust learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 35: 9. 2021: 8110-8118.

[74] NOHYUN K, CHOI H, CHUNG H W. Data Valuation Without Training of a Model[C]//The Eleventh International Conference on Learning Representations. 2023.

[75] FELDMAN V, ZHANG C. What neural networks memorize and why: Discovering the long tail via influence estimation[J]. Advances in Neural Information Processing Systems, 2020, 33: 2881-2891.

[76] MINDERMANN S, BRAUNER J M, RAZZAK M T, et al. Prioritized training on points that are learnable, worth learning, and not yet learnt[C]//International Conference on Machine Learning. 2022: 15630-15649.

[77] ZHANG X, DU J, LI Y, et al. Spanning training progress: Temporal dual-depth scoring (tdds) for enhanced dataset pruning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2024: 26223-26232.

[78] WELLING M. Herding dynamical weights to learn[C]//Proceedings of the 26th Annual International Conference on Machine Learning. 2009: 1121-1128.

[79] SENER O, SAVARESE S. Active learning for convolutional neural networks: A core-set approach[J]. ArXiv preprint arXiv:1708.00489, 2017.

[80] MAHARANA A, YADAV P, BANSAL M. D2 pruning: Message passing for balancing diversity and difficulty in data pruning[J]. ArXiv preprint arXiv:2310.07931, 2023.

[81] ZHENG H, LIU R, LAI F, et al. Coverage-centric Coreset Selection for High Pruning Rates[C]//The Eleventh International Conference on Learning Representations. 2023.

[82] SORSCHER B, GEIRHOS R, SHEKHAR S, et al. Beyond neural scaling laws: beating power law scaling via data pruning[C]//OH A H, AGARWAL A, BELGRAVE D, et al. Advances in Neural Information Processing Systems. 2022.

[83] KOTHAWADE S, KAUSHAL V, RAMAKRISHNAN G, et al. PRISM: A Unified Framework of Parameterized Submodular Information Measures for Targeted Data Subset Selection and Summarization[C]//Thirty-Sixth AAAI Conference on Artificial Intelligence. 2022.

[84] WEI K, IYER R, BILMES J. Submodularity in data subset selection and active learning[C]//International conference on machine learning. 2015: 1954-1963.

[85] JIANG A H, WONG D L, ZHOU G, et al. Accelerating Deep Learning by Focusing on the Biggest Losers[EB/OL]. 2019. https://arxiv.org/abs/1910.00762. arXiv: 1910.00762 [cs.LG].

[86] HE M, YANG S, HUANG T, et al. Large-scale dataset pruning with dynamic uncertainty[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2024: 7713-7722.

[87] YANG E, SHEN L, WANG Z, et al. An efficient dataset condensation plugin and its application to continual learning[J]. Advances in Neural Information Processing Systems, 2023, 36.

[88] ZHAO B, BILEN H. Dataset condensation with differentiable siamese augmentation[C]//International Conference on Machine Learning. 2021: 12674-12685.

[89] SOVIANY P, IONESCU R T, ROTA P, et al. Curriculum learning: A survey[J]. International Journal of Computer Vision, 2022, 130(6): 1526-1565.

[90] WANG X, CHEN Y, ZHU W. A survey on curriculum learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[91] AZAD A S, GUR I, EMHOFF J, et al. Clutr: Curriculum learning via unsupervised task representation learning[C]//International Conference on Machine Learning. 2023: 1361-1395.

[92] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training gans[J]. Advances in Neural Information Processing Systems, 2016, 29.

[93] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in Neural Information Processing Systems, 2017, 30.

[94] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[95] GONTIJO-LOPES R, SMULLIN S, CUBUK E D, et al. Tradeoffs in data augmentation: An empirical study[C]//Proc. Int. Conf. on Learning Representations. 2020.

[96] KRIZHEVSKY A, HINTON G. Convolutional deep belief networks on cifar-10[J]. Unpublished manuscript, 2010, 40(7): 1-9.

[97] ALVAREZ-MELIS D, FUSI N. Geometric Dataset Distances via Optimal Transport[C]//LAROCHELLE H, RANZATO M, HADSELL R, et al. Proceedings of the Advances in Neural Information Processing Systems: vol. 33. Curran Associates, Inc., 2020: 21428-21439.

[98] KANTOROVICH L V. On the translocation of masses[C]//Dokl. Akad. Nauk. USSR (NS): vol. 37. 1942: 199-201.

[99] CHIZAT L, PEYRÉ G, SCHMITZER B, et al. Unbalanced optimal transport: Dynamic and Kantorovich formulations[J]. Journal of Functional Analysis, 2018, 274(11): 3090-3123.

[100] CUTURI M. Sinkhorn Distances: Lightspeed Computation of Optimal Transport[C/OL]//BURGES C, BOTTOU L, WELLING M, et al. Proceedings of the Advances in Neural Information Processing Systems: vol. 26. Curran Associates, Inc., 2013: 2292-2300. https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

[101] COURTY N, FLAMARY R, HABRARD A, et al. Joint distribution optimal transportation for domain adaptation[C/OL]//GUYON I, LUXBURG U V, BENGIO S, et al. Proceedings of the Advances in Neural Information Processing Systems: vol. 30. Curran Associates, Inc., 2017: 3733-3742. https://proceedings.neurips.cc/paper/2017/file/0070d23b06b1486a538c0eaa45dd167a-Paper.pdf.

[102] PEYRE G, CUTURI M. Computational Optimal Transport: With Applications to Data Science[J/OL]. Foundations and Trends® in Machine Learning, 2019, 11(5-6): 355-607. http://dx.doi.org/10.1561/2200000073. DOI: 10.1561/2200000073.

[103] MAGURRAN A E. Measuring biological diversity[J]. Current Biology, 2021, 31(19): R1174-R1177.

[104] RAHANE A A, SUBRAMANIAN A. Measures of complexity for large scale image datasets[C]//2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). 2020: 282-287.

[105] XIE P, SINGH A, XING E P. Uncorrelation and evenness: a new diversity-promoting regularizer[C]//Proc. Int. Conf. Mach. Learn. 2017: 3811-3820.

[106] JOLLIFFE I T, CADIMA J. Principal component analysis: a review and recent developments[J]. Philos. Trans. Roy. Soc. A, Math. Phys. Eng. Sci., 2016, 374(2065): 20150202.

[107] BAILEY S. Principal component analysis with noisy and/or missing data[J]. Publications Astronomical Soc. Pacific, 2012, 124(919): 1015.

[108] LECUN Y, CORTES C, BURGES C J. The MNIST database of handwritten digits, 1998[J]. URL http://yann. lecun. com/exdb/mnist, 1998, 10(34): 14.

[109] ZAGORUYKO S, KOMODAKIS N. Wide Residual Networks[C/OL]//RICHARD C. WILSON E R H, SMITH W A P. Proc. Brit. Mach. Vis. Conf. (BMVC). BMVA Press, 2016: 87.1-87.12. https://dx.doi.org/10.5244/C.30.87. DOI: 10.5244/C.30.87.

[110] YANG S, LI J, ZHANG T, et al. AdvMask: A sparse adversarial attack-based data augmentation method for image classification[J]. Pattern Recognition, 2023, 144: 109847.

[111] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. Int. J. Comput. Vis., 2015, 115(3): 211-252.

[112] COHEN I, HUANG Y, CHEN J, et al. Pearson correlation coefficient[J]. Noise reduction in speech processing, 2009: 1-4.

[113] MYERS L, SIROIS M J. Spearman correlation coefficients, differences between[J]. Encyclopedia of statistical sciences, 2004, 12.

[114] HUANG G, LIU Z, van der MAATEN L, et al. Densely Connected Convolutional Networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[115] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. ArXiv preprint arXiv:2010.11929, 2020.

[116] 武川, 苏杭. 基于大语言模型的目标情感分析数据增强研究[J]. 情报理论与实践, 1-12.

[117] 葛轶洲, 许翔, 杨锁荣, 等. 序列数据的数据增强方法综述[J]. 计算机科学与探索, 2021, 15(07): 1207-1219.

[118] REN X, LIN W, YANG X, et al. Data Augmentation in Defect Detection of Sanitary Ceramics in Small and Non-i.i.d Datasets[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(11): 8669-8678. DOI: 10.1109/TNNLS.2022.3152245.

[119] TANG Y, LI B, LIU M, et al. AutoPedestrian: An Automatic Data Augmentation and Loss Function Search Scheme for Pedestrian Detection[J]. IEEE Transactions on Image Processing, 2021, 30: 8483-8496. DOI: 10.1109/TIP.2021.3115672.

[120] ZHANG H, XU Z, HAN X, et al. Data Augmentation Using Bitplane Information Recombination Model[J]. IEEE Transactions on Image Processing, 2022, 31: 3713-3725. DOI: 10.1109/TIP.2022.3175429.

[121] YANG S, GUO S, ZHAO J, et al. Investigating the effectiveness of data augmentation from similarity and diversity: An empirical study[J]. Pattern Recognition, 2024, 148: 110204.

[122] GAO J, HUA Y, HU G, et al. Discrepancy-Guided Domain-Adaptive Data Augmentation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(8): 5064-5075. DOI: 10.1109/TNNLS.2021.3128401.

[123] PU Y, HAN Y, WANG Y, et al. Fine-Grained Recognition With Learnable Semantic Data Augmentation[J]. IEEE Transactions on Image Processing, 2024, 33: 3130-3144. DOI: 10.1109/TIP.2024.3364500.

[124] 刘子扬, 王朝坤, 章衡. 图对比学习方法综述[J]. 软件学报, 1-20. DOI: 10.13328/j.cnki.jos.007417.

[125] MAI Z, HU G, CHEN D, et al. MetaMixUp: Learning Adaptive Interpolation Policy of MixUp With Metalearning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(7): 3050-3064. DOI: 10.1109/TNNLS.2020.3049011.

[126] BASHIR D, MONTAÑEZ G D, SEHRA S, et al. An information-theoretic perspective on overfitting and underfitting[C]//Advances in Artificial Intelligence: 33rd Australasian Joint Conference. 2020: 347-358.

[127] SEHRA S, FLORES D, MONTANEZ G D. Undecidability of underfitting in learning algorithms[C]//2021 2nd International Conference on Computing and Data Science (CDS). 2021: 591-594.

[128] LI H, RAJBAHADUR G K, LIN D, et al. Keeping Deep Learning Models in Check: A History-Based Approach to Mitigate Overfitting[Z]. 2024. arXiv: 2401.10359 [cs.SE].

[129] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning[C]//International conference on machine learning. 2016: 1928-1937.

[130] LIU T Y, MIRZASOLEIMAN B. Data-Efficient Augmentation for Training Neural Networks[J]. Advances in Neural Information Processing Systems, 2022, 35: 5124-5136.

[131] HOU C, ZHANG J, ZHOU T. When to learn what: Model-adaptive data augmentation curriculum[C]//Proceedings of the IEEE International Conference on Computer Vision. 2023: 1717-1728.

[132] CHRABASZCZ P, LOSHCHILOV I, HUTTER F. A downsampled variant of imagenet as an alternative to the cifar datasets[J]. ArXiv preprint arXiv:1707.08819, 2017.

[133] LIU Z, MIAO Z, ZHAN X, et al. Large-Scale Long-Tailed Recognition in an Open World[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019.

[134] NILSBACK M E, ZISSERMAN A. Automated flower classification over a large number of classes[C]//2008 Sixth Indian conference on computer vision, graphics & image processing. 2008: 722-729.

[135] PARKHI O M, VEDALDI A, ZISSERMAN A, et al. Cats and dogs[C]//2012 IEEE conference on computer vision and pattern recognition. 2012: 3498-3505.

[136] MAJI S, RAHTU E, KANNALA J, et al. Fine-grained visual classification of aircraft[J]. ArXiv preprint arXiv:1306.5151, 2013.

[137] KRAUSE J, DENG J, STARK M, et al. Collecting a large-scale dataset of fine-grained cars[J]., 2013.

[138] MNIH V. Asynchronous Methods for Deep Reinforcement Learning[J]. ArXiv preprint arXiv:1602.01783, 2016.

[139] SHEN M, HOW J P. Robust opponent modeling via adversarial ensemble reinforcement learning in asymmetric imperfect-information games[J]. ArXiv preprint arXiv:1909.08735, 2019.

[140] HUMPLIK J, GALASHOV A, HASENCLEVER L, et al. Meta reinforcement learning as task inference[J]. ArXiv preprint arXiv:1905.06424, 2019.

[141] GASTALDI X. Shake-Shake regularization[J]., 2017.

[142] XIE S, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1492-1500.

[143] ZHUANG F, QI Z, DUAN K, et al. A Comprehensive Survey on Transfer Learning[J]. Proc. IEEE, 2021, 109(1): 43-76. DOI: 10.1109/JPROC.2020.3004555.

[144] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. J. Machine Learning Research, 2008, 9(11).

[145] NCIR C E B, HAMZA A, BOUAGUEL W. Parallel and scalable Dunn Index for the validation of big data clusters[J]. Parallel Computing, 2021, 102: 102751.

[146] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL]. 2019. https://arxiv.org/abs/1509.02971. arXiv: 1509.02971 [cs.LG].

[147] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//International conference on machine learning. 2018: 1861-1870.

[148] LIU H, LI C, LI Y, et al. Improved baselines with visual instruction tuning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2024: 26296-26306.

[149] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models[J]. ArXiv preprint arXiv:2302.13971, 2023.

[150] 吴敖, 王海龙, 柳林, 等. 情感识别大模型研究综述[J]. 计算机科学与探索, 1-32.

[151] 王文晟, 谭宁, 黄凯, 等. 基于大模型的具身智能系统综述[J]. 自动化学报, 2025, 51(01): 1-19. DOI: 10.16383/j.aas.c240542.

[152] KOH P W W, ANG K S, TEO H, et al. On the Accuracy of Influence Functions for Measuring Group Effects[C/OL]//WALLACH H, LAROCHELLE H, BEYGELZIMER A, et al. Advances in Neural Information Processing Systems: vol. 32. Curran Associates, Inc., 2019. https://proceedings.neurips.cc/paper/2019/file/a78482ce76496fcf49085f2190e675b4-Paper.pdf.

[153] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. 2021: 8748-8763.

[154] 谷金晶, 覃天宝, 普园媛, 等. 基于 CLIP 语义特征增强的图像描述[J]. 计算机应用研究, 1-8. DOI: 10.19734/j.issn.1001-3695.2025.08.0270.

[155] ALABDULMOHSIN I, WANG X, STEINER A, et al. CLIP the Bias: How Useful is Balancing Data in Multimodal Learning?[J]. ArXiv preprint arXiv:2403.04547, 2024.

[156] POTH C, STERZ H, PAUL I, et al. Adapters: A unified library for parameter-efficient and modular transfer learning[J]. ArXiv preprint arXiv:2311.11077, 2023.

[157] LI G, HOU W, HU D. Progressive spatio-temporal perception for audio-visual question answering[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 7808-7816.

[158] PARULEKAR A, COLLINS L, SHANMUGAM K, et al. Infonce loss provably learns cluster-preserving representations[C]//The Thirty Sixth Annual Conference on Learning Theory. 2023: 1914-1961.

[159] OORD A V D, LI Y, VINYALS O. Representation learning with contrastive predictive coding[J]. ArXiv preprint arXiv:1807.03748, 2018.

[160] BENGIO Y, LÉONARD N, COURVILLE A. Estimating or propagating gradients through stochastic neurons for conditional computation[J]. ArXiv preprint arXiv:1308.3432, 2013.

[161] KRIZHEVSKY A, HINTON G, et al. Learning multiple layers of features from tiny images[J]., 2009.

[162] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE International Conference on Computer Vision. 2021: 10012-10022.

[163] YANG S, YANG H, GUO S, et al. Not All Data Matters: An End-to-End Adaptive Dataset Pruning Framework for Enhancing Model Performance and Efficiency[J]. ArXiv preprint arXiv:2312.05599, 2023.

[164] GOYAL P, DUVAL Q, REIZENSTEIN J, et al. VISSL[Z]. https://github.com/facebookresearch/vissl. 2021.

[165] LI S, XIA X, GE S, et al. Selective-supervised contrastive learning with noisy labels[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2022: 316-325.

[166] NCIR C E B, HAMZA A, BOUAGUEL W. Parallel and scalable Dunn Index for the validation of big data clusters[J]. Parallel Computing, 2021, 102: 102751.

[167] TAESIRI M R, NGUYEN G, HABCHI S, et al. Imagenet-hard: The hardest images remaining from a study of the power of zoom and spatial biases in image classification[J]. Advances in Neural Information Processing Systems, 2024, 36.

[168] HENDRYCKS D, BASART S, MU N, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2021: 8340-8349.

[169] HENDRYCKS D, ZHAO K, BASART S, et al. Natural adversarial examples[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2021: 15262-15271.

[170] WU Z F, WEI T, JIANG J, et al. Ngc: A unified framework for learning with open-world noisy data[C]//Proceedings of the IEEE International Conference on Computer Vision. 2021: 62-71.

[171] MALKOV Y A, YASHUNIN D A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 42(4): 824-836.

致　　谢

首先，衷心感谢我的父母。是他们始终如一的爱、理解与支持，让我有勇气追求学术理想、坚持到最后。无论是在我最顺利的时刻，还是最困难的阶段，父母都给予我无条件的鼓励与力量。我所取得的每一份成绩，都凝聚着他们的辛劳与信任。

其次，诚挚感谢我的导师申富饶教授。感谢您在学业上给予我的悉心指导与无私帮助。您温和而坚定的教诲，不仅引领我在科研的道路上不断前行，更让我学会了如何以严谨的态度面对学术问题。正是您始终如一的信任与鼓励，让我在遇到困难时能够重新振作、坚持不懈，最终顺利完成博士学业。

同时，我也要感谢在此期间给予我支持与帮助的所有合作者、同门与朋友们。感谢你们在研究、生活与成长道路上给予的真诚交流与鼓励。每一次讨论、每一次相伴，都是我宝贵的记忆与财富。

攻读博士学位期间的学术成果

部分已发表学术论文

1. **Yang, S.**, Yang, H., Guo, S., Shen, F., & Zhao, J. (2025). IPF-RDA: An Information-Preserving Framework for Robust Data Augmentation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**. (*CCF-A* 类期刊)
2. **Yang, S.**, Li, P., Xiong, X., Shen, F., & Zhao, J. (2025). AdaAugment: A tuning-free and adaptive approach to enhance data augmentation. **IEEE Transactions on Image Processing**. (*CCF-A* 类期刊)
3. **Yang, S.**, Li, P., Shen, F., & Zhao, J. (2025). RL-Selector: Reinforcement Learning-Guided Data Selection via Redundancy Assessment. In: **Proceedings of the IEEE International Conference on Computer Vision (ICCV)**. (*CCF-A* 类会议)
4. **Yang, S.**, Ye, P., Shen, F., & Zhou, D. (2025). When Dynamic Data Selection Meets Data Augmentation: Achieving Enhanced Training Acceleration. In: **Proceedings of the 42nd International Conference on Machine Learning (ICML)** (*CCF-A* 类会议)
5. **Yang, S.**, Zhang, T., Xu, Z., Li, P., Xu, B., Shen, F., & Zhao, J. (2025). Supervised Contrastive Learning with Prototype Distillation for Data Incremental Learning. **Neural Networks**, 107651. (*CCF-B* 类期刊)
6. **Yang, S.**, Ye, P., Ouyang, W., Zhou, D., & Shen, F. (2024). A Clip-Powered Framework for Robust and Generalizable Data Selection. In: **Proceedings of the 13th International Conference of Learning Representation (ICLR)** (Spotlight，录用率 5%)
7. **Yang, S.**, Shen, F., & Zhao, J. (2024). EntAugment: Entropy-driven adaptive data augmentation framework for image classification. In: **European Conference on**

Computer Vision (ECCV). (*CCF-B* 类会议)

8. **Yang, S.**, Guo, S., Zhao, J., & Shen, F. (2024). Investigating the effectiveness of data augmentation from similarity and diversity: An empirical study. **Pattern Recognition**. (*CCF-B* 类期刊)
9. **Yang, S.**, Li, J., Zhang, T., Zhao, J., & Shen, F. (2023). AdvMask: A sparse adversarial attack-based data augmentation method for image classification. **Pattern Recognition**. (*CCF-B* 类期刊)
10. Yang, H., **Yang, S.**, Zhang, L., Dou, H., Shen, F., & Zhao, J. (2024). CS-QCFS: Bridging the performance gap in ultra-low latency spiking neural networks. **Neural Networks**. (*CCF-B* 类期刊)
11. Xiong, X., Wang, X., **Yang, S.**, Shen, F., & Zhao, J. (2024). GMNI: Achieve good data augmentation in unsupervised graph contrastive learning. **Neural Networks**. (*CCF-B* 类期刊)
12. **杨锁荣**，杨洪朝，申富饶，赵健. 面向深度学习的图像数据增强综述. 软件学报. 2024. (*CCF-A* 类期刊)

申请专利

1. 申富饶，**杨锁荣**，李春华，秦辞海，杨洪朝，张天玥，陈昊，“一种基于图像配准的视频流变化检测方法和系统”. CN202211342122.X
2. 申富饶，**杨锁荣**，李俊，赵健，“一种基于实时定位轨迹数据进行动态滤波优化的方法”. CN201911005643.4
3. 申富饶，**杨锁荣**，李俊，赵健，“实时定位轨迹数据进行局部线性插值和预测的方法”. CN202010316990.5
4. 申富饶，陈昊，**杨锁荣**，杨洪朝，卢俣金，张凌茗，刘佩涵，李若彤，赵健，“一种基于对比检测的机器人巡检系统”. CN202310111632.4
5. 申富饶，张天玥，时晓峰，**杨锁荣**，赵健. “一种基于自组织增量图的半监督多媒体数据流分类方法”. CN202211315078.3

科研项目参与情况

1. 国家自然基金面上项目，面向增量式无监督学习的新型神经网络研究。
2. 科技部重大项目-科技创新 2030 项目，基于神经可塑性的脉冲网络高效学习机制与类脑智能系统。
3. 国家自然基金面上项目，基于深度感知增量式联想记忆神经网络的信息融合系统研究。
4. 上海市科技重大项目