

图像数据集知识效能评估方法

丁昕苗^{1,2} 李好男^{1,2} 刘雨帆^{2,3} 刘彤飞^{2,3} 张克卿^{2,3}
李兵^{2,4} 原春锋^{2,3} 王坚^{2,4,5} 胡卫明^{2,3,6}

摘 要 当前数据交易面临价值评估难题, AI 模型训练缺乏数据集选择标准, 其原因在于不同数据集在信息密度、质量结构与任务适配性上差异显著. 现有方法多聚焦于数据质量或模型性能的单一维度, 缺乏面向信息密度与多维价值的系统性量化框架, 限制了任务导向场景下数据集的科学选择与高效利用. 为此, 提出数据集知识效能评估框架, 其是综合衡量数据集的有效知识含量及其在特定任务中适应能力的量化方法, 包含三个维度: 1) 知识提炼: 将无损数据集蒸馏算法引入效能评估, 通过知识密度比量化数据集有效知识承载能力; 2) 知识质量分析: 构建多维质量指标体系, 深入分析和评估数据集的知识质量; 3) 知识任务适配: 创新性提出任务驱动的评估适配机制, 通过契合度分析与权重优化实现对特定任务适配性的评估. 在 CIFAR-10/100、VGGFace2、ImageNet-1K 和 NABirds 等典型数据集上的实验结果表明, 所提方法能够较为准确地评估知识密度与任务适配性, 为数据集选择与应用提供理论依据和技术支撑.

关键词 知识效能评估; 多维质量分析; 任务适配; 数据集蒸馏

引用格式 丁昕苗, 李好男, 刘雨帆, 刘彤飞, 张克卿, 李兵, 原春锋, 王坚, 胡卫明. 图像数据集知识效能评估方法. 自动化学报, 2026, 52(9): 2130–2148

DOI 10.16383/j.aas.c250457 **CSTR** 32138.14.j.aas.c250457

Knowledge Efficiency Evaluation Method for Image Datasets

DING Xin-Miao^{1,2} LI Hao-Nan^{1,2} LIU Yu-Fan^{2,3} LIU Tong-Fei^{2,3} ZHANG Ke-Qing^{2,3}
LI Bing^{2,4} YUAN Chun-Feng^{2,3} WANG Jian^{2,4,5} HU Wei-Ming^{2,3,6}

Abstract Data trading faces significant challenges in value evaluation, and AI model training lacks criteria for dataset selection, because datasets differ markedly in information density, quality structure, and task adaptability. Existing approaches largely focus on a single dimension, either data quality or model performance, and provide no systematic framework for quantifying information density and multi-dimensional value, which hampers the scientific selection and efficient utilization of datasets in task-oriented scenarios. Therefore, we propose a dataset knowledge efficiency evaluation framework that comprehensively quantifies the effective knowledge content of a dataset and its adaptability to specific tasks, comprising three dimensions: 1) Knowledge distillation: Introducing a lossless dataset distillation algorithm into efficiency evaluation to quantify the effective knowledge capacity of datasets through the knowledge density ratio; 2) Knowledge quality analysis: Constructing a multi-dimensional quality indicator system to thoroughly analyze and evaluate dataset knowledge quality; 3) Knowledge-task alignment: Proposing an innovative task-driven evaluation-adaptation mechanism that leverages fit analysis and weight optimization to evaluate task-specific adaptability. Experiments on typical datasets, including CIFAR-10/100, VGGFace2, ImageNet-1K, and NABirds, show that the proposed method accurately evaluates knowledge density and task adaptability, providing theoretical basis and technical support for dataset selection and application.

Keywords knowledge efficiency evaluation; multi-dimensional quality analysis; task adaptation; dataset distillation

Citation Ding Xin-Miao, Li Hao-Nan, Liu Yu-Fan, Liu Tong-Fei, Zhang Ke-Qing, Li Bing, Yuan Chun-Feng, Wang Jian, Hu Wei-Ming. Knowledge efficiency evaluation method for image datasets. *Acta Automatica Sinica*, 2026, 52(9): 2130–2148

收稿日期 2025-09-18 录用日期 2026-03-04

Manuscript received September 18, 2025; accepted March 4, 2026

北京市自然科学基金 (JQ24022), 国家自然科学基金 (61876100, 62072286, 62372451, 62192785, 62372082, 62192782, 62532015), 中国人工智能学会-蚂蚁集团联合研究基金 (CAAI-MYJJ2024-02), 中国科协青年人才托举工程 (2024QNRC001) 资助

Supported by Beijing Natural Science Foundation (JQ24022), National Natural Science Foundation of China (61876100, 62072286, 62372451, 62192785, 62372082, 62192782, 62532015), CAAI-Ant Group Research Fund (CAAI-MYJJ2024-02), and Young Elite Scientists Sponsorship Program by CAST (2024QNRC001)

本文责任编辑 金连文

Recommended by Associate Editor JIN Lian-Wen

1. 山东工商学院信息与电子工程学院 烟台 264005 2. 中国科学院自动化研究所多模态人工智能系统全国重点实验室 北京 100190 3. 中国科学院大学人工智能学院 北京 100049 4. 人民中科(北京)智能技术有限公司 北京 102604 5. 北京大学软件与微电子学院 北京 102600 6. 上海科技大学信息科学与技术学院 上海 201210

1. School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai 264005 2. State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 3. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049 4. PeopleAI Inc., Beijing 102604 5. School of Software and Microelectronics, Peking University, Beijing 102600 6. School of Information Science and Technology, ShanghaiTech University, Shanghai 201210

近年来,数据交易过程与人工智能训练实践普遍面临一个关键问题:缺乏统一、科学的数据集价值评估标准.企业在数据交易中难以衡量数据资源价值,研究者在模型训练中也缺乏系统化的数据选择依据.这主要源于不同数据集在信息密度、质量结构与任务适配性方面差异显著,而现有方法多依赖单一指标,难以全面揭示数据集的知识承载能力与实际价值.与此同时,随着数据作为核心生产要素的战略地位不断提升,国家陆续出台了《关于促进数据产业高质量发展的指导意见》^[1]、《关于完善数据流通安全治理的实施方案》^[2]等政策文件,强调以技术创新驱动数据价值释放,构建安全高效的数据要素流通体系.在此背景下,如何科学、准确地评估数据资源的内在价值,特别是量化数据集的知识承载能力及其对不同任务的适应性,已成为实现数据要素高效配置与合理利用的重要技术问题.

由表 1 可见,现有方法通常分别聚焦数据规范性与可用性(数据质量)、经济或业务收益(数据价值)、单样本边际贡献(样本贡献度)或迁移适配评估(迁移适配预测),因而难以在同一框架下同时刻画数据集层面的知识承载、质量结构与任务适配.为此,本文提出数据集知识效能评估(dataset knowledge efficiency evaluation, DKEE)方法,作为衡量数据集有效知识含量及其在特定任务中适应能力的重要手段.与传统主要依赖单一维度指标或模型性能进行间接评价的方法不同,DKEE 从数据集的自身结构与内容特征出发,构建涵盖知识承载能力、质量结构特征以及任务契合度的多维度评估体系,以实现数据集价值的系统性量化,从而在概念边界上与传统数据质量、数据价值等评估范式形成明确区分.

然而,当前研究仍面临若干核心挑战:1)“知识效能”概念缺乏统一的理论定义,评估框架的理论基础有待完善;2)现有评估方法多局限于单一维度分析,缺乏系统性的多维度综合评估体系;3)缺少能够针对不同任务特性进行调整的评估机制,难以准确反映数据集在特定应用场景下的实际价值.这

些技术局限性在一定程度上制约了数据集评估方法的发展与应用.

针对上述问题,本文提出一种基于知识提炼、知识质量分析与知识任务适配的 DKEE 方法.该方法以知识提炼为基础、知识质量分析为支撑、知识任务适配为导向,构建了面向实际应用场景的评估框架.本文的主要贡献包括:

- 1) 理论提出:首先将数据集知识效能定义为“知识提炼-知识质量分析-知识任务适配”三要素协同作用的结果,为数据集评估研究提供新的视角;
- 2) 方法设计:构建涵盖代表性样本提取、多维质量建模与任务适配反馈的评估框架,具体包括基于数据集无损蒸馏知识核心子集提取、基于多样性指标的质量评估模块以及任务驱动的契合度建模与反馈机制;
- 3) 实验验证:在 CIFAR-10^[3]、ImageNet^[4] 等多个公开图像分类数据集上进行实验验证,结果表明该方法能够有效识别数据集在不同任务中的适配价值与质量特征.

本文所构建的 DKEE 方法在理论提出、方法设计与实验验证等方面进行了探索,为数据资源的评估与应用提供了技术参考.

1 相关工作

1.1 数据集蒸馏相关研究

本文的知识提炼模块采用数据集蒸馏方法,从大规模数据集中提取核心知识和代表性信息,旨在保留数据集的关键特征和学习价值.蒸馏思想最初应用于模型压缩与性能提升,即知识蒸馏(knowledge distillation, KD).从软标签蒸馏^[5]、中间层对齐^[6-7]到结构关系迁移^[8-10],相关方法不断发展.为缓解异构模型间的结构差异,Heo 等^[11]提出多层特征投影机制,Liu 等^[12]探索了跨域映射策略,Yang 等^[13]进一步发展了跨模态蒸馏方法.此外,知识蒸馏也被拓展至模型压缩与稀疏优化^[14-17],在 CNN^[18]与 Transformer^[19]等架构中均展现出良好的适用性.

表 1 不同数据评估范式的对比与本文方法定位
Table 1 Comparison of different data evaluation paradigms and positioning of our method

评估范式	评估对象	核心问题与边界	知识承载	多维质量	任务适配
数据质量评估	数据本身	关注数据是否规范及单一维度是否达标,通常难以刻画数据集整体知识结构	否	否	否
数据价值评估	经济层	关注数据的经济或业务收益,但强依赖具体场景与收益假设,可复用性弱	否	否	否
样本贡献度	单样本	刻画单样本对训练目标的边际贡献,计算代价高且难以形成数据集层整体结论	否	否	否
迁移适配预测	表征层	评估预训练表征的可迁移性,主要反映模型与任务匹配,难以解释数据本体与质量结构	否	否	是
DKEE (本文)	数据集层	统一刻画数据集中可被提炼的有效知识含量,并在不同任务场景下进行多维质量分析,从而支持训练前的数据选择与治理决策	是	是	是

传统知识蒸馏通常依赖完整的原始数据集进行监督学习,在数据访问受限的环境下存在局限性.为解决此问题, Wang 等^[20]引入数据集蒸馏 (dataset distillation, DD) 概念,旨在极大压缩数据量的同时保留原始数据集的核心知识,使模型在小规模子集上仍能达到与原数据集近似的性能. Zhao 等^[21]基于梯度匹配策略,从训练过程一致性角度提升了蒸馏样本的任务适应性. 后续研究提出基于核岭回归^[22-23]、分布匹配^[24-25]以及轨迹匹配^[26-30]等方法,从不同角度扩展了数据集蒸馏技术. 尽管上述方法在蒸馏效果上取得进展,但仍面临信息损失、泛化性能不足和计算开销大的挑战. 为进一步改善蒸馏效果, Guo 等^[31]提出 DATM (difficulty-aligned trajectory matching), 通过结合特征梯度与类别边界约束,达到了无损蒸馏的效果,实现对蒸馏样本结构的精准控制,在降低信息损失和提升任务适配能力方面表现出一定优势. 近年来也有工作从设计空间与机制层面对数据集蒸馏进行系统梳理,并提出更高效的数据归因近似方法,为蒸馏过程理解与效率分析提供补充^[32-35].

1.2 知识质量分析相关研究

知识质量分析是指从准确性、可靠性和适用性等维度系统评估数据集质量的过程,为数据集的有效利用提供基础保障. 传统数据质量评估方法主要关注数据集的基础质量和结构规范性,强调对数据在格式一致性、字段完整性、缺失率和异常值方面的检测. 这类方法在保障数据基本可用性和支持预处理流程方面发挥了重要作用,通过统一格式与排查损坏样本,有效提升了模型训练的稳定性与效率.

随着机器学习的发展,数据质量评估逐渐从基础结构检测扩展到更广泛的数据特性分析. Northcutt 等^[36]开发了置信学习方法来处理训练数据中的错误标注问题, Kang 等^[37]提出长尾分布数据的重平衡策略. 此外,一些研究关注数据集的统计特性^[38]和学习复杂性^[39]等方面. 然而,现有研究多聚焦于特定的单一质量维度,难以提供对数据集质量的全面评估. 面对多任务、多领域应用场景时,传统方法在系统性质量评估方面仍存在不足.

1.3 知识任务适配相关研究

知识任务适配是指评估和优化数据集与特定下游任务之间匹配程度的过程,旨在提高数据集在目标任务上的有效性和适用性. 在迁移学习和元学习等领域,知识任务适配性评估具有重要意义.

Nguyen 等^[40]提出 LEEP (log expected empirical prediction) 方法来评估预训练模型与目标数据

集的匹配度,通过分析预训练特征的条件概率分布来预测迁移性能. You 等^[41]开发了 LogME (log marginal evidence) 指标来预测迁移学习的性能,该方法基于最大证据原理评估特征质量. 在数据集选择方面, Swayamdipta 等^[42]从学习难度角度分析了数据集特性对任务性能的影响,提出数据图谱概念来刻画样本的学习特征. 此外,一些研究从任务间语义相似性和数据分布匹配的角度探索适配性评估方法. 尽管这些研究在特定场景下取得了一定成效,但现有方法多依赖预训练模型或实际训练过程来评估适配性,计算成本较高,缺乏通用性的知识任务适配评估框架,难以处理多样化的任务需求和数据特征. 此外,近年亦有研究从分布匹配与核方法角度对任务可迁移性进行评估,为刻画数据与任务之间的适配关系提供了新的量化视角^[43],同时国内学者围绕复杂任务规划、数据协同推断以及数据要素流通与交易机制等方向开展了研究与综述,从应用系统层面反映了数据质量与数据效能评估在实际任务决策中的重要性,为本文讨论数据集效能提供了补充参考^[44-46].

总体而言,现有方法在特定场景下提供了有益启发,但尚缺乏一个能够系统覆盖知识提炼、质量分析与任务适配的统一评估框架,因此难以直接与本文提出的 DKEE 方法进行对比.

2 数据集知识效能评估框架

为系统化评估数据集的知识承载能力,本文提出如图 1 所示的 DKEE 框架. 该框架综合考虑知识提炼、知识质量分析与知识任务适配三大要素,涵盖基于数据集蒸馏的知识密度测量、多维质量分析及任务驱动适配分析,旨在实现对数据集核心价值的量化评估.

2.1 数据集知识效能定义

数据集知识效能用于衡量单位数据量所蕴含的有效知识及其在特定任务下的适应能力. 设 $D = \{(x_i, y_i)\}_{i=1}^N$ 为包含 N 个样本-标签对的数据集,其中 $x_i \in \mathcal{X}$ 表示第 i 个样本, $y_i \in \mathcal{Y}$ 为对应标签, $T \in \mathcal{T}$ 为目标任务,则其知识效能定义为:

$$V(D, T) = \{g(D, T), q(D, T)\} \quad (1)$$

其中, $g(D, T)$ 为知识提炼得到的知识密度; $q(D, T)$ 为质量分析得到的多维质量指标,包含完整性、一致性、噪声比、多样性等.

2.2 基于数据集蒸馏的知识提炼

知识提炼维度通过无损数据集蒸馏算法,对

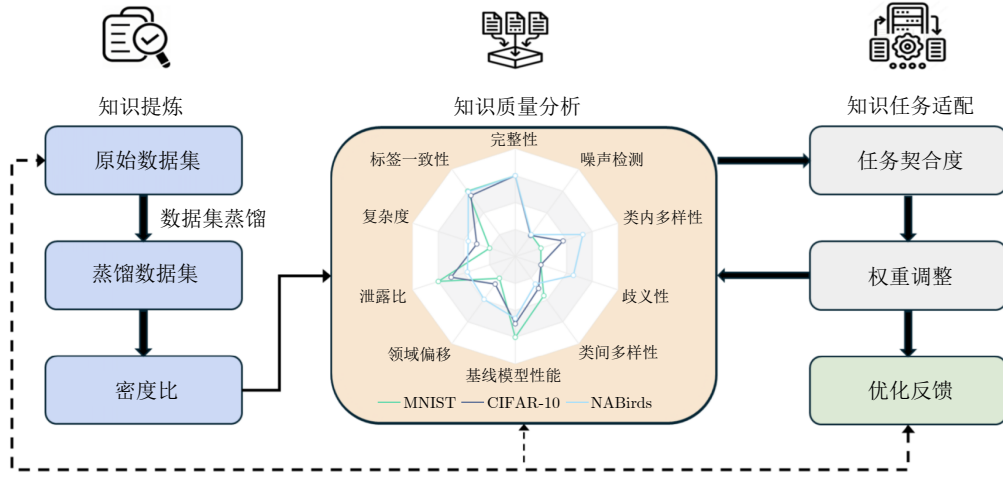


图 1 知识效能评估框架

Fig.1 Knowledge efficiency evaluation framework

比蒸馏前后各类别的保留样本数量, 量化数据集的核心知识承载能力. 该方法旨在从原始数据中提取具有代表性的关键信息, 生成在模型性能上与原数据集相当的合成子集, 从而衡量数据集的知识密度.

具体实现中, 首先对原始数据集 D 提取特征, 并通过主成分分析 (PCA)^[47] 降维以提升计算效率. 随后, 结合 K-means 聚类、梯度影响函数或模型不确定性等筛选策略, 从原始数据中提取代表性样本构建候选子集. $D' \subseteq D$ 表示从原始数据集 D 中选取的候选子集; D_{distill}^* 表示寻找到的最优蒸馏子集, 使得基于该子集训练的模型在原始数据分布上保持近似的预测性能, 从而在压缩数据规模的同时尽可能保留原始知识. 优化目标为:

$$D_{\text{distill}}^* = \arg \min_{D' \subseteq D} E_{(x, y) \sim P_D} [L(f_{D'}(x), y)] \quad (2)$$

其中, P_D 表示原始数据集 D 的数据分布; $E_{(x, y) \sim P_D}[\cdot]$ 表示在该分布下的期望; $L(\cdot, \cdot)$ 为特定任务的损失函数; $f_{D'}: \mathcal{X} \rightarrow \mathcal{Y}$ 表示在子集 D' 上训练得到的模型, $f_{D'}(x)$ 为其对输入样本 x 的预测输出.

为量化数据集的知识承载能力, 本文引入每类图像保留数量指标 IPC (images per class) 作为基础度量. 对于任意数据集 D , IPC 定义为:

$$\text{IPC}(D) = \frac{|D|}{K} \quad (3)$$

其中, $|D| = N$ 为数据集的样本总数; K 为类别总数. 为评估蒸馏效果, 分别计算原始数据集的 IPC 值 IPC_{orig} 与不同压缩比 r 下蒸馏数据集的 IPC 值 $\text{IPC}_{\text{distill}}(r)$. 通过调整 r , 寻找模型性能保持不变时的最小压缩比 r^* . 据此, 定义知识密度为:

$$g(D, T) = \frac{\text{IPC}_{\text{distill}}(r^*)}{\text{IPC}_{\text{orig}}} \quad (4)$$

$g(D, T)$ 表示在保证模型性能不变的情况下所需的必要样本比例, 间接反映了数据集的知识冗余程度. 若 g 越小, 说明在保持模型性能不变的情况下, 所需样本比例更少, 表明数据集中存在较多冗余信息 (更易压缩); 反之, 则表示数据集中冗余较少, 信息分布更均匀.

2.3 基于多维度的知识质量分析

为全面评估数据集的质量特征, 本文构建了多维度的知识质量分析体系. 该体系涵盖完整性、一致性、噪声检测、多样性等多个维度, 每个维度均从不同角度刻画数据集的内在特性, 为 DKEE 的综合评估提供支撑.

2.3.1 数据集完整性

数据集完整性 (data completeness) 衡量数据集中有效样本的占比, 反映数据集的可用程度. 该指标直接影响模型训练的稳定性与预测性能, 缺失或损坏的样本会导致训练过程不稳定, 可能引发过拟合或欠拟合问题. 完整性计算公式为:

$$C_{\text{comp}} = \frac{N_{\text{valid}}}{N} \times 100\% \quad (5)$$

其中, N_{valid} 为有效样本数. C_{comp} 值越高, 表明数据集质量越好, 为后续的知识提炼与任务适配提供可靠基础.

2.3.2 标签一致性

标签一致性 (label consistency) 指数据集中标签格式有效且可被正确解析的比例. 若标签存在格式缺失、越界、类型不符或无法解析等问题, 将直接

影响数据加载与训练的稳定性. 定义“格式错误/不可解析”指示变量

$$\mathcal{E}_i = \begin{cases} 1, & y_i \text{ 的格式无效或不可解析} \\ 0, & y_i \text{ 的格式有效且可解析} \end{cases}$$

则标签一致性计算公式为:

$$C_{\text{label}} = \left(1 - \frac{1}{N} \sum_{i=1}^N \mathcal{E}_i\right) \times 100\% \quad (6)$$

当所有样本的标签格式均有效且可解析时, $C_{\text{label}} = 100\%$; 随格式问题样本比例上升, 该值相应降低.

2.3.3 噪声检测

噪声检测 (noise detection) 识别数据集中偏离正常分布的异常样本, 这些样本通常源于标注错误、数据采集问题或样本自身异常. 噪声样本不仅影响训练稳定性, 还可能误导模型决策, 降低预测准确性. 本文结合 t-SNE (t-distributed stochastic neighbor embedding)^[48] 降维与 DBSCAN (density-based spatial clustering of applications with noise)^[49] 聚类方法构建噪声检测机制.

首先, 使用 t-SNE 将高维特征表示 $\{\phi(x_i)\}_{i=1}^N \subset \mathbf{R}^d$ (d 为特征维度) 降维至二维空间 $\{z_i\}_{i=1}^N \subset \mathbf{R}^2$, 保留样本间的邻接关系并增强类别可分性. 随后, 在降维空间中应用 DBSCAN 进行密度聚类分析, 它基于密度识别异常样本: 若某个样本在其邻域内未包含足够数量的近邻, 则将其视为噪声, 聚类标签标记为 -1. 设降维后被判定为噪声的样本集合为 Z_{noise} , 则噪声比例为:

$$R_{\text{noise}} = \frac{|Z_{\text{noise}}|}{N} \times 100\% \quad (7)$$

其中, $|Z_{\text{noise}}|$ 表示 DBSCAN 输出中聚类标签为 -1 的样本数. 该方法可有效识别低密度区域中的异常值, 尤其适用于非均匀分布或复杂结构的数据集. 通过将降维与聚类相结合, 能够在保持特征结构可视性的同时, 有效识别数据集中的潜在噪声样本.

2.3.4 类内多样性

类内多样性 (intra-class diversity) 衡量同一类别内部样本在特征空间中的差异程度. 适度的类内差异有助于模型捕捉类别内的细微变化, 增强泛化能力; 过低则可能导致欠拟合, 过高则可能引入噪声. 类内多样性 D_{intra} 通过类内方差量化:

$$D_{\text{intra}} = \frac{1}{K} \sum_{k=1}^K \frac{1}{n_k} \sum_{i=1}^{n_k} \left\| \phi(x_i^{(k)}) - \mu_k \right\|^2 \quad (8)$$

其中, n_k 为类别总数 K 中第 k 类的样本数量; $\phi(x_i^{(k)}) \in \mathbf{R}^d$ 为第 k 类中第 i 个样本的特征表示; $\mu_k = \sum_{i=1}^{n_k} \phi(x_i^{(k)}) / n_k$ 为该类的特征中心; $\|\cdot\|$ 表示欧几里得范数. D_{intra} 值较低表明类内样本较为集中, 有利于形成稳定的类别表示.

2.3.5 类间多样性

类间多样性 (inter-class diversity) 度量各类别中心在特征空间中的离散程度, 反映类别之间的区分能力. 类间多样性越高, 表明类别分布越分散, 有助于构建明确的分类边界. 本文通过计算类别中心集合的方差度量类间多样性:

$$D_{\text{inter}} = \frac{1}{K} \sum_{k=1}^K \left\| \mu_k - \bar{\mu} \right\|^2 \quad (9)$$

其中, $\bar{\mu} = \sum_{k=1}^K \mu_k / K$ 为所有类别中心的均值. D_{inter} 值越大, 表明类别间的可分性越好, 有利于分类任务的执行, 数据集对分类任务的支撑能力越强, 有利于训练出具有良好泛化性能的分类模型.

2.3.6 复杂度

数据集复杂度 (dataset complexity) 衡量数据集的信息多样性和分布丰富程度, 反映其潜在特征容量. 复杂度越高的数据集包含更多结构变化和细节特征, 有助于提升模型对多样性信息的表征能力.

本文从三个维度评估图像复杂度: 灰度均值、标准差和信息熵. 对于数据集中的每张图像, 经灰度化后形成灰度向量 $[g_1, g_2, \dots, g_{n_p}]$, 其中 n_p 为单张图像的像素总数, 则该图像的灰度均值 μ_g 与标准差 σ_g 分别定义为:

$$\mu_g = \frac{1}{n_p} \sum_{i=1}^{n_p} g_i, \quad \sigma_g = \sqrt{\frac{1}{n_p} \sum_{i=1}^{n_p} (g_i - \mu_g)^2} \quad (10)$$

进一步, 将灰度值离散化为 256 个等级, 统计归一化后的灰度分布概率 $\{p_j\}_{j=1}^{256}$, 计算图像信息熵:

$$H_g = - \sum_{j=1}^{256} p_j \log_2 p_j \quad (11)$$

其中, p_j 表示灰度等级 j 的像素概率, 满足 $\sum_{j=1}^{256} p_j = 1$. 数据集层面的复杂度通过在所有图像上分别计算上述三个指标的均值得到, 形成三维复杂度向量 $[\bar{\mu}_g, \bar{\sigma}_g, \bar{H}_g]$, 其中 $\bar{\mu}_g$ 、 $\bar{\sigma}_g$ 、 $\bar{H}_g$ 分别为数据集中图像灰度均值、标准差与信息熵的平均值. 灰度均值反映整体亮度水平, 标准差衡量对比度变化, 信息熵刻画灰度分布的随机性, 三者共同构成数据集复杂性的综合评估.

2.3.7 歧义性

歧义性 (ambiguity) 衡量样本在类别判别边界

附近的不确定程度, 反映其在模型分类决策中的模糊性. 判别性强的样本预测概率集中于某一类别, 而边界样本的预测概率在多个类别间分布均衡. 本文基于多个教师模型的预测概率分布, 利用信息熵 $A_{\text{ambig}}(x_i)$ 度量样本的不确定性:

$$A_{\text{ambig}}(x_i) = \frac{-\sum_{c=1}^K \bar{p}(c|x_i) \log \bar{p}(c|x_i)}{\log K} \quad (12)$$

其中, $\bar{p}(c|x_i)$ 表示多个教师模型对样本 x_i 预测为第 c 类的概率的平均值. 为缓解单一模型可能带来的预测偏差, 本文对多个教师模型的输出概率进行平均处理后再计算熵值. $A_{\text{ambig}}(x_i)$ 越高, 说明样本在多模型判断下呈现出较强的不确定性, 通常位于类别决策边界区域.

2.3.8 领域偏移

领域偏移 (domain shift) 衡量训练集与测试集在特征分布上的差异程度. 分布差异会导致模型在实际部署场景中性能退化, 尤其在迁移学习与跨领域任务中, 是影响模型泛化能力的关键因素. 本文采用 Jensen-Shannon 散度 (JSD)^[50] 量化领域偏移 S_{shift} 程度:

$$S_{\text{shift}} = \text{JSD}(P, Q) = \frac{1}{2}(D_{\text{KL}}(P \parallel M) + D_{\text{KL}}(Q \parallel M)) \quad (13)$$

其中, P 和 Q 分别表示通过特征提取器如 CLIP (contrastive language-image pretraining)^[51] 或 VAE (variational autoencoder)^[52] 提取图像特征并归一化后的训练集和测试集概率分布; M 为两者的平均概率分布, 即 $M(b) = (P(b) + Q(b))/2$; KL (Kullback-Leibler) 散度^[53] 分别定义为:

$$D_{\text{KL}}(P \parallel M) = \sum_b P(b) \log \frac{P(b)}{M(b)} \quad (14)$$

$$D_{\text{KL}}(Q \parallel M) = \sum_b Q(b) \log \frac{Q(b)}{M(b)} \quad (15)$$

其中, b 表示特征空间离散化后的区间; $P(b)$ 、 $Q(b)$ 和 $M(b)$ 分别为对应区间的概率分布值. S_{shift} 值越大, 表明训练集与测试集的分布偏移越严重, 模型的域适应难度越高. 该指标对数据集分布稳定性评估和跨领域任务性能预测具有重要参考价值.

2.3.9 泄露比

泄露比 (leakage ratio) 衡量测试集与训练集样本之间的重复或高度相似程度. 泄露比越高, 意味着测试集中包含大量与训练集雷同的样本, 导致模型评估结果的客观性和泛化性降低. 本文通过计算

测试集样本与训练集样本的最大余弦相似度来量化泄露比:

$$L_{\text{leak}} = \frac{\sum_{i=1}^{N_{\text{test}}} \mathbb{I}\left(\max_{j=1, \dots, N_{\text{train}}} \text{sim}(u_i^{\text{test}}, u_j^{\text{train}}) \geq \varepsilon\right)}{N_{\text{test}}} \quad (16)$$

其中, N_{test} 和 N_{train} 分别为测试集和训练集的样本总数; $u_i^{\text{test}} \in \mathbf{R}^d$ 和 $u_j^{\text{train}} \in \mathbf{R}^d$ 分别表示测试样本和训练样本的特征向量; $\text{sim}(\cdot, \cdot)$ 表示余弦相似度; ε 为相似度阈值 (默认为 0.3); $\mathbb{I}(\cdot)$ 为指示函数. 当最大相似度超过阈值时, 判定为信息泄露. L_{leak} 值越高, 表明数据集的训练-测试划分独立性越差.

2.3.10 基线模型性能

基线模型性能 (baseline model performance) 通过多个经典网络架构 (如 ResNet、VGG 等) 对数据集进行基准测试, 评估数据集的特征可分性和学习难度. 若数据集在多个经典模型上均取得高准确率, 通常表明其具有良好的特征可分性和标注质量. 数据集的平均基线性能定义为:

$$P_{\text{base}} = \frac{1}{n_{\text{model}}} \sum_{m=1}^{n_{\text{model}}} \text{Acc}_m \quad (17)$$

其中, n_{model} 为基线模型数量; Acc_m 为第 m 个模型的测试准确率. 高 P_{base} 值表明数据集特征清晰、噪声较少、标签准确, 有利于模型学习; 低值则可能反映数据集具有更高的任务难度或更复杂的特征分布, 需结合其他质量指标进行综合分析.

2.4 基于知识任务适配的评估

在前文中, 本文将知识效能 $V(D, T)$ 定义为由知识密度 $g(D, T)$ 与质量指标集合 $q(D, T)$ 共同构成. 然而, $V(D, T)$ 本身是集合式的定义, 尚不能直接反映一个具体数值. 为实现可操作化的量化评估, 本文进一步提出基于知识任务适配的效能计算机制. 该机制通过识别任务类型并调整 $q(D, T)$ 中各指标的权重, 将 $g(D, T)$ 与 $q(D, T)$ 融合为最终的任务相关效能值 $V_{\text{task}}(D, T)$.

具体实现分为三个步骤: 首先, 利用多个预训练模型 (ResNet50 架构, 分别在 MNIST、CIFAR-10/100、ImageNet 等数据集上训练) 在目标数据集上进行迁移测试, 通过微调分类层计算准确率. 根据细粒度模型 (如 NABirds、VGGFace2) 与通用模型的相对性能判定任务类型. 其次, 基于任务类型调整各质量指标权重: 对于细粒度任务, 增大类内/类间多样性、歧义性等指标的权重; 对于通用任务, 则保持各指标的均衡权重. 此外, $g(D, T)$ 会作为独

立维度补充进入效能评估. 最后, 计算加权得分并进行任务相关的非线性映射. 设 $q(D, T) = \{q_1, \dots, q_{|q(D, T)|}\}$ 表示质量指标集合, 则其加权平均得分为:

$$V_{\text{task}}(D, T) = \frac{\sum_{i=1}^{|q(D, T)|} q_i w_i}{\sum_{i=1}^{|q(D, T)|} w_i} \quad (18)$$

其中, q_i 为第 i 个质量指标的归一化得分; w_i 为对应的任务权重. 不同任务类型通过权重配置的差异来体现: 通用任务强调各维度的整体均衡, 细粒度任务则突出类内/类间多样性与歧义性等关键因素. 此外, 在各质量指标归一化过程中引入轻量级 Sigmoid 平滑, 以避免分数贴边并增强区分度. 该机制实现了任务识别、权重调整和效能计算的统一流程, 提升了评估方法对不同任务场景的适应性.

3 实验验证与结果分析

3.1 实验设置与资源配置

本文在多个典型图像分类数据集上验证所提出的知识效能评估框架. 实验在两类计算平台上进行: 知识提炼实验部署于配备 4 块 NVIDIA A800 GPU (80 GB 显存) 的高性能计算平台, 知识质量分析与知识任务适配实验在配备 8 块 NVIDIA 2080Ti GPU (12 GB 显存) 的服务器上执行. 知识提炼部分采用无损蒸馏算法, 设置专家轨迹 buffer 为 100, 合成步长为 1000, 并训练至无损标准. 在知识质量分析阶段, 噪声检测使用 DBSCAN 聚类算法, 邻域半径设定为 0.5, 最小邻域样本数设定为 5.

本文模型采用 ResNet-18/50^[54]、VGG-16^[55]、DenseNet-121^[56]、EfficientNet-B0^[57]、MobileNetV2^[58] 和 PyramidNet^[59]. 选取的评估数据集涵盖不同复杂度层次, 包括 MNIST^[60] (手写数字识别)、CIFAR-10 和 CIFAR-100^[3] (低分辨率图像分类)、TinyImageNet^[61] (小型图像识别)、Places365^[62] (场景识别)、VGGFace2^[63] 和 LFW^[64] (人脸识别)、ImageNet^[4] (物体识别) 以及 NABirds^[65] 和 CUB^[66] (鸟类细粒度分类).

3.2 基于数据集蒸馏的知识提炼评估

在知识提炼维度, 本文采用无损数据集蒸馏算法量化数据集的知识密度. 通过对比蒸馏前后 IPC 的变化, 衡量数据集的知识承载能力. 图 2 展示了不同数据集每类蒸馏前后 IPC 变化对比, 整体规模上各数据集的样本量变化及知识密度汇总于表 2.

实验结果显示不同数据集存在显著的知识密度差异. MNIST 表现出极高的冗余度: 原始 IPC 为

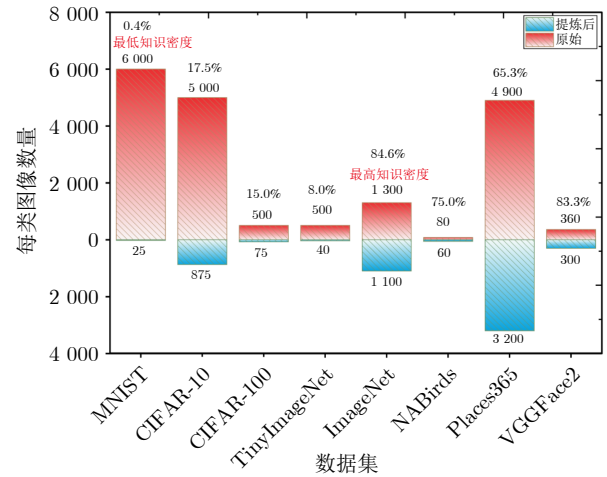


图 2 蒸馏前后数据集每类 IPC 变化

Fig. 2 Per-class IPC changes of the datasets before and after distillation

表 2 蒸馏前后数据集整体规模知识密度对比

Table 2 Overall-scale knowledge density comparison of datasets before and after distillation

数据集	类别数 K	原始样本	蒸馏后样本	$g(D, T)$
MNIST ^[60]	10	60 000	250	0.004
CIFAR-10 ^[3]	10	50 000	8 750	0.175
CIFAR-100 ^[3]	100	50 000	7 500	0.150
TinyImageNet ^[61]	200	100 000	8 000	0.080
ImageNet ^[4]	1 000	1 300 000	1 100 000	0.846
NABirds ^[65]	555	44 400	33 300	0.750
Places365 ^[62]	365	1 788 500	1 168 000	0.653
VGGFace2 ^[63]	9 131	3 287 160	2 739 300	0.833

6000, 蒸馏后仅保留 25 张 (信息保留率约为 0.004), 表明其样本高度同质, 极少量样本即可表达完整知识. 原因在于手写数字的拓扑特征高度稳定: 数字“0”的封闭环形、“1”的单笔直线、“8”的上下双环结构等核心特征在各种书写变化中保持不变, 使得极少样本即可代表完整知识.

CIFAR-10 的原始 IPC 为 5000, 蒸馏后需保留 875 张 (信息保留率为 0.175); CIFAR-100 的原始 IPC 为 500, 蒸馏后需保留 75 张 (信息保留率为 0.150). 32×32 的低分辨率是决定其知识密度的一大因素, CIFAR-10 虽然类别少, 但每类需要覆盖更大的类内变化, 如“汽车”类涵盖轿车、SUV、卡车等车型; 而 CIFAR-100 将这些粗类进一步细分, 如“轿车”独立为一类, 每个细分类的样本分布更集中, 部分抵消了类别数量增加带来的复杂度提升.

TinyImageNet 原始 IPC 为 500, 蒸馏后需保留 40 张 (信息保留率为 0.080), 呈现中等冗余性.

其 64×64 的分辨率相比 CIFAR 提供更多细节 (像素数提升 4 倍), 使得单张图像包含更丰富的信息, 但 200 个类别使每类仅有 500 样本, 限制了类内变化的充分采样. 以“蘑菇”类为例, 能展现菌盖纹理和菌柄形态, 但 500 张样本难以覆盖所有种类. 相比之下, ImageNet 展现出极高的知识密度. 其原始 IPC 为 1300, 蒸馏后仍需保留 1100 张图像 (信息保留率高达约 0.846), 以“狗”类为例, 包含 120 余个品种 (哈士奇、柯基、德牧等), 每个品种又有年龄、毛色、季节等多维变化. 这表明 ImageNet 样本在类别内部具有高度异质性, 任务信息分布广泛, 难以进一步压缩, 展现出了极高的知识承载能力.

对于 NABirds, 其原始 IPC 约为 80, 蒸馏后仍需保留约 60 张样本 (信息保留率为 0.750). 作为典型的细粒度识别数据集, NABirds 的类别区分高度依赖局部形态与微小差异, 例如喙部形状、羽毛纹理分布与颜色过渡等. 此类判别线索往往分散在不同样本及不同局部区域中, 难以被少量样本充分覆盖, 因而呈现出较强的不可压缩性.

Places365 则反映了场景理解任务的复杂性. 其原始 IPC 接近 4900, 蒸馏后仍需保留约 3200 张样本 (信息保留率约为 0.653). 不同于以单一物体为中心的分类任务, 场景识别需要同时刻画空间布局、上下文关系与多目标共现模式. 例如“机场候机厅”与“购物中心”在局部物体层面可能高度相似, 但在全局结构与语义组合上存在显著差异. 由于关键知识更多体现在全局统计与语义共现结构中, 场景数据集在蒸馏过程中难以被大幅压缩.

在 VGGFace2 中, 其原始 IPC 约为 360, 蒸馏后仍需保留约 300 张样本 (信息保留率高达约 0.833). 尽管单个身份的样本数量相对有限, 但人脸识别仍具有显著的类内变化, 包括姿态、光照、表情、年龄与遮挡等因素. 模型需要在多种变化条件下学习身份不变的判别表征, 使得有效信息分散在大量不同成像条件中, 进而提高了蒸馏压缩的难度.

综合来看, 这些结果表明: 数据集的知识可蒸馏性不仅取决于样本规模或类别数量, 更与任务属性、类内异质性以及判别线索的空间分布密切相关.

3.3 多维度知识质量评估

3.3.1 数据集完整性与标签一致性

数据集的基础质量直接影响模型训练效果和知识提取能力. 本文首先从完整性和标签一致性两个维度进行评估, 以确保数据集的基本可用性.

完整性反映数据集中有效样本的占比, 体现数据采集和预处理的质量水平. 从表 3 可见, 除 CIFAR-100 检测到 26 张重复图像 (有效率为 99.96%) 外, 其余各数据集均达到 100% 的样本有效率, 未发现无效或重复图像. 这表明主流数据集在构建过程中普遍采用严格的去重和验证机制, 确保了数据的非冗余性. 即使是规模超过百万的 ImageNet, 也保持了完美的样本有效率, 体现了良好的数据管理水平.

标签一致性衡量数据集中标签格式是否有效且可被正确解析的比例. 本文从类别索引越界、词表外类别以及非法编码等维度对各数据集进行格式一致性检测, 本次评测的各数据集均未检出格式错误, 因此 $C_{\text{label}} = 100\%$. 这一结果说明这些公开基准在发布环节具有较高的格式规范性, 有利于数据加载与模型训练的稳定性.

需要说明的是, 标签不仅涉及“格式是否有效、可解析”, 还包含其语义正确性. 已有研究表明, 主流公开数据集中普遍存在语义层面的标注错误, 如图 3 所示. Northcutt 等^[36] 基于置信学习的可疑样本检索与人工复核, 发现验证集中的错误样本比例分别为: MNIST (10000 张) 中有 15 张、CIFAR-10 (10000 张) 中有 54 张、CIFAR-100 (10000 张) 中有 584 张、ImageNet (50000 张) 中有 2916 张. 这些结果表明, 尽管在格式层面不存在问题, 但语义层面的标注错误在公开数据集中仍然普遍存在, 且

表 3 数据集完整性与标签一致性检测结果
Table 3 Dataset completeness and label consistency detection results

数据集	总图像数	重复图像数	有效率 (%)	正确标签数	标签一致性 (%)
MNIST	70000	0	100.00	70000	100.00
CIFAR-10	60000	0	100.00	60000	100.00
CIFAR-100	60000	26	99.96	60000	100.00
TinyImageNet	130000	0	100.00	130000	100.00
Places365	2204960	0	100.00	2204960	100.00
VGGFace2	3311286	0	100.00	3311286	100.00
ImageNet	1431167	0	100.00	1431167	100.00
NABirds	48562	0	100.00	48562	100.00

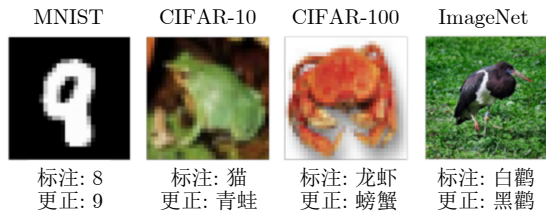


图 3 语义层面标注错误

Fig.3 Semantic-level annotation errors

其核实往往依赖人工逐一检查与确认, 耗费大量人力与时间. 因此, 本节仅报告格式一致性结果, 不将其等同于语义无误.

3.3.2 噪声检测

基于 DBSCAN 聚类的噪声检测揭示了各数据集在样本分布上的差异特征. 表 4 展示了检测结果, 各数据集的噪声比例呈现显著差异.

表 4 数据集噪声检测结果
Table 4 Dataset noise detection results

数据集	总样本数	正常样本数	噪声样本数	噪声比例 (%)
MNIST	60 000	42 238	17 762	29.60
CIFAR-10	50 000	33 078	16 922	33.84
CIFAR-100	50 000	33 269	16 731	33.46
TinyImageNet	100 000	89 462	10 538	10.54
Places365	1 803 460	1 801 968	1 492	0.08
VGGFace2	3 141 890	3 140 174	1 716	0.05
ImageNet	1 281 167	1 279 738	1 429	0.12
NABirds	23 912	6 021	17 891	74.82

从噪声比例由大到小的顺序来看, NABirds 的最高噪声比例 (74.82%) 反映了细粒度识别中跨类视觉相似性的固有挑战, 如“黄莺” (yellow warbler) 与“威尔逊莺” (Wilson’s warbler) 仅在头部羽冠颜色上有细微差别, 在特征空间中几乎重叠, 导致大量此类样本被判定为噪声.

值得注意的是, NABirds 的高噪声比例 (74.82%) 与其分类准确率 (74.93%) 并未呈现负相关关系. 这表明, 在细粒度识别任务中, 基于密度的噪声检测所识别的大量“噪声样本”并不等同于标签错误样本, 而更可能是位于类别判别边界附近的统计离群样本. 这类样本通常表现为跨类别外观高度相似, 仅在局部细节上存在差异的困难样本. 进一步的对照实验显示 (见附录 A), 在剔除部分此类样本后, 模型在 NABirds 上的分类准确率出现明显下降, 说明这些样本在训练过程中有助于模型关注更具判别性的局部特征 (如羽毛纹理、颜色分布与结构细节), 从而促进细粒度特征表示的学习并提升判别边界的

精细化程度.

深入分析噪声来源发现: CIFAR-10 (33.84%) 与 CIFAR-100 (33.46%) 受限于 32×32 的低分辨率, 导致类别边界模糊, 如“卡车”与“汽车”在低分辨率下轮廓相似, “猫”与“狗”的侧面剪影难以区分, 这些边界样本自然被判定为噪声; MNIST (29.60%) 的噪声比例主要源于手写风格的个体差异, 例如数字“7”的书写存在欧式风格 (带横线) 和美式风格 (无横线) 两种变体, 这些变体在特征空间中形成独立簇, 导致部分样本被识别为离群点.

VGGFace2 (0.05%) 的噪声主要来自光照、姿态、表情等人脸自然变化, 特别是极端侧脸 (大于 75°) 和强逆光条件下的人脸图像, 其特征表示与正面光照充足的标准人脸差异巨大; 相比之下, ImageNet (0.12%) 和 Places365 (0.08%) 同样保持较低噪声水平, 更多得益于其多阶段数据清洗与人工审核流程对异常样本的有效控制.

综上, 噪声检测从特征空间角度刻画了数据集的分布结构, 噪声比例反映样本分布的紧致性及判别边界附近样本的相对占比. 该结果可作为对传统质量指标的补充, 用于识别分布更分散或边界更易混叠的数据集, 并刻画其困难样本聚集程度的差异.

3.3.3 类内多样性

类内多样性衡量同一类别内部样本的特征差异程度. 如图 4 所示, 各数据集呈现明显的层次差异.

结果显示, VGGFace2 (0.048 12) 具有最高的类内多样性, 反映了人脸图像在光照、表情、姿态、年龄等多维度上的显著变化, 构成了高度复杂的特征空间. ImageNet (0.034 85) 和 NABirds (0.032 01) 次之: 前者包含的自然物体类别内部涵盖了从不同姿态、视角到背景环境的广泛变化, 如“狗”类别包含数百个品种的多形态; 后者作为细粒度数据集,

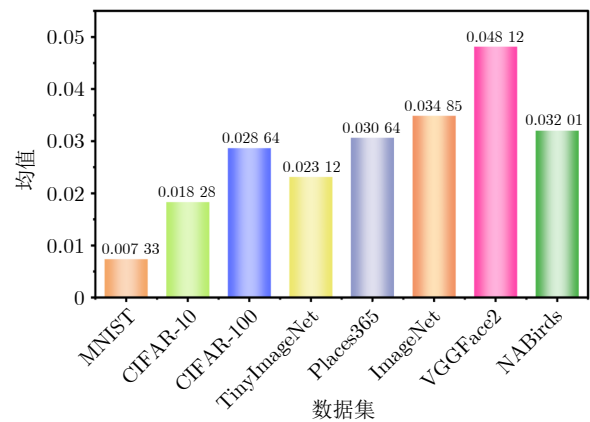


图 4 数据集类内多样性统计

Fig.4 Intra-class diversity statistics of the datasets

虽然聚焦于鸟类, 但同一鸟种的不同个体在羽毛纹理、体态姿势、季节性变化等方面仍保持较高差异.

Places365 (0.030 64) 与 CIFAR-100 (0.028 64) 处于中等水平: 前者的场景类别内部存在布局、装饰风格的变化, 但受限于场景的相对固定性; 后者虽有 100 个类别, 但 32×32 的低分辨率限制了细节表达. TinyImageNet (0.023 12) 虽然具有更高分辨率 (64×64), 但其精简的类别设计导致样本分布相对集中. CIFAR-10 (0.018 28) 仅有 10 个基础类别, 类间差异明显使得类内变化空间有限. MNIST (0.007 33) 最低, 作为二值化的手写数字数据集, 其结构高度规范, 变化主要局限于笔画粗细和倾斜角度.

高类内多样性有助于模型学习更鲁棒的类别表征, 提升对新样本的适应能力; 但过高可能会引入噪声, 增加学习难度. 低类内多样性虽利于快速收敛, 但可能限制模型的泛化能力.

3.3.4 类间多样性

类间多样性度量不同类别中心在特征空间中的离散程度, 直接影响分类边界的清晰度. 表 5 展示了各数据集的测量结果.

表 5 数据集类间多样性测量结果
Table 5 Dataset inter-class diversity measurement results

数据集	类间多样性均值
MNIST	0.004 162
CIFAR-10	0.002 321
CIFAR-100	0.009 125
TinyImageNet	0.004 166
Places365	0.011 322
ImageNet	0.012 894
VGGFace2	0.029 938
NABirds	0.009 781

VGGFace2 (0.029 938) 和 ImageNet (0.012 894) 类间多样性最高, 体现了优秀的类别设计. VGG-Face2 中不同身份的人脸在骨骼结构、五官特征等方面具有显著差异, 形成了清晰的身份边界. ImageNet 包含 1 000 个类别, 从动物、植物到人造物品, 跨类别样本分布在完全不同的语义区域, 如“汽车”与“猫”在视觉特征上几乎没有重叠. Places365 (0.011 322) 的场景类别具有明确区分, “办公室”与“山地”、“厨房”与“海滩”等在空间布局和视觉元素上差异显著. NABirds (0.009 781) 虽然聚焦于鸟类, 但是不同鸟种在体型、羽色、喙形等特征上仍保持良好的可区分性.

CIFAR-100 (0.009 125) 处于中等水平, 其 100

个类别中存在语义相近的组, 如“虎”与“豹”同属猫科, “橡树”与“枫树”同为树木, 这种层次化的类别结构导致部分类间边界相对模糊. TinyImageNet (0.004 166) 和 MNIST (0.004 162) 类间多样性明显降低: 前者可能因为在构建缩减版时选择了视觉特征相对接近的类别子集; 后者的手写数字中, “1”与“7”、“3”与“8”等在笔画结构上存在相似性. CIFAR-10 (0.002 321) 最低, 虽然只有 10 个基础类别, 但这些类别在 32×32 的低分辨率下特征表达受限, 导致类间距离缩小. 类间多样性与分类准确率呈正相关, 但 CIFAR-10 是例外, 其类间多样性最低 (0.002 321), 却有 94.67% 的高准确率, 原因在于任务简单 (仅 10 类), 抵消了低分辨率的负面影响, 这也说明单一指标不足以反映整体难度, 需要多维度联合评估.

类间多样性高的数据集有利于构建稳定的决策边界、提升分类准确率, 是评估数据集知识效能的重要指标. 类间多样性不足会增加分类难度, 特别是训练样本有限时, 模型难以学习到有效的判别特征.

3.3.5 图像复杂度

图像复杂度衡量数据集的信息量与结构多样性. 本文从灰度均值、标准差和信息熵三个维度进行综合评估, 分别反映图像的亮度水平、对比度变化和信息分布的随机性.

如图 5 和图 6 所示, 各数据集的复杂度呈现显著差异. 可将数据集分为三个层次.

1) 低复杂度数据集. MNIST (熵 0.669 7, 均值 0.001 36, 方差 0.003 28) 由于二值化特性, 像素分布集中于黑白极端, 仅能反映有限的书写差异, 整体结构简单. VGGFace2 (熵 0.022 1, 均值 0.000 71, 方差 0.000 88) 在人脸图像预处理过程中进行了严

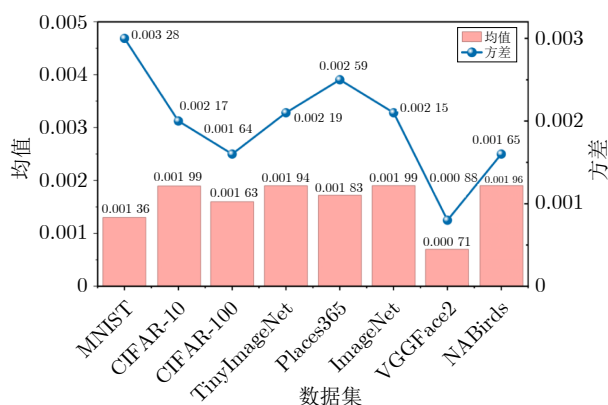


图 5 数据集灰度均值与标准差分布

Fig. 5 Distribution of grayscale mean and standard deviation of the datasets

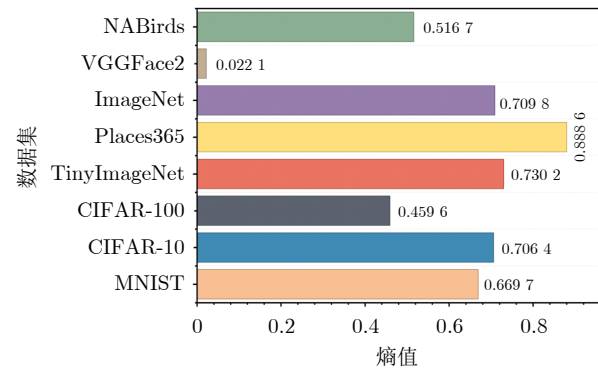


图 6 数据集信息熵分布

Fig.6 Information entropy distribution of the datasets

格的对齐与归一化, 像素分布高度一致, 导致统计复杂度较低.

2) 中等复杂度数据集. CIFAR-100 (熵 0.4596, 均值 0.001 63, 方差 0.001 64) 尽管类别数扩展至 100, 但分辨率仅为 32×32 且单类样本有限, 特征差异不足, 复杂度下降. NABirds (熵 0.5167, 均值 0.00196, 方差 0.00165) 作为细粒度鸟类数据集, 虽然类别间差异细微, 但是羽毛纹理与色彩变化仍提供了一定信息量, 整体复杂度处于中等水平.

3) 高复杂度数据集. CIFAR-10 (熵 0.7064, 均值 0.001 99, 方差 0.002 17) 在低分辨率下覆盖 10 个通用类别, 信息量较 MNIST 更丰富. TinyImageNet (熵 0.7302, 均值 0.00194, 方差 0.00219) 在分辨率提升至 64×64 且包含 200 个类别时, 细节表现更丰富. ImageNet (熵 0.7098, 均值 0.001 99, 方差 0.002 15) 类别数达到 1 000, 样本规模庞大且清洗严格, 均值与方差均衡, 熵值体现其多样性. Places365 (熵 0.8886, 均值 0.001 83, 方差 0.002 59) 覆盖 365 个场景类别, 熵值最高, 反映出空间布局、光照条件和语义元素的高度差异性, 整体复杂度最为显著.

图像复杂度直接影响模型的学习难度和所需的网络容量. 高复杂度数据集需要更深的网络架构来捕获丰富的特征层次, 而低复杂度数据集则可以用相对简单的模型达到良好效果.

3.3.6 基线模型性能

为进一步揭示不同数据集在可学习性上的差异, 本文在多个经典模型上进行基准测试, 结果如表 6 所示.

在 MNIST 上, 所有模型的准确率均超过 99.00%, 最高达 99.75%, 表明其特征高度可分、结构清晰. CIFAR-10 (平均准确率为 94.67%) 仅有 10 个基础类别, 虽受限于 32×32 分辨率, 但类间差异明显, 保持了良好的可学习性. VGGFace2 (平均准确率为 92.39%) 虽然类别数达 8 631 个, 但人脸识别的类内一致性强、判别特征明确, 使其保持较高准确率. CIFAR-100 (平均准确率为 78.29%) 在与 CIFAR-10 相同的低分辨率下需要区分 100 个类别, 类间相似性显著增加.

ImageNet (平均准确率为 74.75%) 包含 1 000 个类别, 尽管图像质量高、标注规范, 但庞大的类别规模和细粒度的语义差异 (如不同品种的狗) 仍带来较大挑战. NABirds (平均准确率为 74.93%) 作为细粒度鸟类识别数据集, 需要区分极其细微的羽毛纹理和形态差异, 不同模型的性能差异较大 (63.70% ~ 80.30%), 反映了网络架构对细粒度特征捕获能力的显著差异. TinyImageNet (平均准确率为 62.04%) 在 64×64 的分辨率下包含 200 个类别, 信息密度与类别复杂度的矛盾尤为突出. Places365 (平均准确率为 55.83%) 的场景类别存在大量类内差异 (如不同风格的卧室) 和类间相似性 (如咖啡厅与餐厅), 是最具挑战性的任务.

平均准确率清晰地展示了数据集的难度梯度. 这种差异不仅反映了任务本身的复杂度, 也体现了

表 6 经典模型在不同数据集上的测试准确率 (%)
Table 6 Test accuracy of classic models on different datasets (%)

模型	MNIST	CIFAR-10	CIFAR-100	TinyImageNet	Places365	VGGFace2	ImageNet	NABirds
VGG16 ^[55]	99.20	93.25	73.50	56.80	55.00	89.65	71.50	74.90
ResNet18 ^[54]	99.50	93.02	77.10	58.70	54.70	91.20	69.80	76.21
ResNet50 ^[54]	99.60	93.62	79.20	61.50	55.20	92.75	76.15	79.55
DenseNet121 ^[56]	99.70	95.04	79.30	64.20	56.20	93.10	74.90	80.30
MobileNetV2 ^[58]	99.40	94.50	75.30	59.10	54.00	90.50	71.80	—
EfficientNet-B0 ^[57]	99.70	96.00	80.10	66.00	57.00	94.00	77.10	63.70
PyramidNet ^[59]	99.75	97.25	83.54	68.00	58.70	95.50	82.00	—
平均准确率	99.55	94.67	78.29	62.04	55.83	92.39	74.75	74.93

注: “—”表示该数据集未公开基准结果.

分辨率、类别数量、类内变异等多维因素的综合影响, 为数据集选择和模型设计提供了重要的定量参考。

3.3.7 歧义性

歧义性衡量数据集中类别边界的模糊程度, 直接影响模型的判别难度。表 7 展示了基于多教师模型预测熵的歧义性测量结果。

表 7 歧义性、领域偏移与泄露比测量结果 (%)
Table 7 Measurement results of ambiguity domain shift and leakage ratio (%)

数据集	歧义性	领域偏移	泄露比
MNIST	99.88	0.120	4.99
CIFAR-10	97.75	0.170	5.15
CIFAR-100	99.04	0.210	4.49
TinyImageNet	95.36	0.010	8.74
Places365	99.69	0.040	73.35
ImageNet	99.64	0.020	62.82
VGGFace2	99.85	0.036	78.43
NABirds	99.67	0.370	2.39

TinyImageNet (95.36%) 歧义性最低, 说明其类间边界相对清晰。这一结果可能得益于子集构建和数据清洗过程, 但受限于 64×64 的分辨率, 细粒度特征表达不足。CIFAR-10 (97.75%) 紧随其后, 仅包含 10 个通用类别, 边界清晰, 如飞机与汽车差异显著。CIFAR-100 (99.04%) 虽扩展至 100 类, 但出现“虎”与“豹”、“橡树”与“枫树”等语义相近类别, 使歧义性略有上升。

ImageNet (99.64%)、Places365 (99.69%)、NABirds (99.67%) 的歧义性处于较高水平, 体现出任务的复杂性。ImageNet 包含 1000 个类别, 仅犬类就覆盖 120 多个品种, 如“哈士奇”与“阿拉斯加雪橇犬”之间的细微差别使得边界判别极具挑战。Places365 的“咖啡厅”与“餐厅”、“办公室”与“会议室”等场景在视觉与语义层面高度重叠。NABirds 则因细粒度特征 (如羽毛纹理与色彩差异) 带来边界模糊, 凸显细粒度识别的挑战性。

VGGFace2 (99.85%) 与 MNIST (99.88%) 歧义性最高: 前者的人脸图像因姿态、光照、表情变化及高度相似的面部特征 (如双胞胎) 增加了边界不确定性; 后者虽然仅包含十个数字, 但书写风格差异显著, 例如数字“7”的欧式横线与美式无横线之间的差异, 使其在特征空间中形成大量模糊区域, 从而推高整体歧义性。

综上所述, 歧义性反映数据集在类别边界处的模糊性差异: 低歧义性数据集适合用于验证算法的基准性能, 而高歧义性数据集则更能揭示模型在细

粒度判别和复杂场景下的能力差距。因此, 歧义性不仅是衡量数据集内在难度的重要指标, 也为数据集知识效能的全面评价和任务适配分析提供关键支撑。

3.3.8 领域偏移

领域偏移衡量训练集与测试集在特征分布上的差异, 是影响模型泛化的关键因素。表 7 展示了采用 JSD 散度量化的测量结果, 各数据集呈现不同程度的分布一致性。

令人意外的是, TinyImageNet (0.010%) 展现出最佳的分布一致性, 甚至优于其母数据集 ImageNet (0.020%)。这可能得益于 TinyImageNet 作为精心挑选的子集, 在构建时特别注重训练/测试的均衡性。Places365 (0.040%) 虽然涵盖 365 个场景类别, 但通过标准化的采集流程保持了良好的一致性。

VGGFace2 的领域偏移仅为 0.036%, 处于较低水平, 表明其训练集与测试集在成像条件及整体视觉统计分布上具有较强一致性。需要指出的是, 尽管分布差异有限, VGGFace2 仍采用典型的开集评测设定, 即训练与测试身份集合完全不重叠。在该设定下, 模型难以通过记忆已见身份获得优势, 而必须学习具备跨身份可迁移性的表征与判别机制, 从而在保持低分布偏移的同时对身份泛化能力提出更高要求。

MNIST (0.120%)、CIFAR-10 (0.170%) 和 CIFAR-100 (0.210%) 这些经典数据集的偏移略高, 但仍处于低水平。MNIST 的手写数字在风格上相对稳定, CIFAR 系列虽然来源多样, 但在创建时就考虑了分布均衡。这些成熟数据集的低偏移保证了实验的可重复性, 也解释了它们为何成为算法评测的基准。

NABirds (0.370%) 的偏移明显上升, 反映了野外数据采集的现实挑战。同一种鸟类在不同季节的羽毛色彩差异显著——繁殖期的雄鸟色彩鲜艳, 而越冬期可能暗淡许多。地理位置的差异也带来背景环境的变化, 从森林到湿地, 这些自然变异增加了模型泛化的难度。

领域偏移的差异揭示了不同的数据集定位: 低偏移数据集适合算法开发和公平比较, 高偏移数据集则推动了领域自适应、元学习等前沿方向的研究。这一指标提醒我们, 在选择数据集时不仅要考虑准确率, 更要关注其泛化场景。

3.3.9 泄露比

泄露比通过检测测试集与训练集之间的高相似样本 (余弦相似度大于 0.3) 来评估数据集的隔离程

度. 这一指标对于准确评估模型的真实泛化能力具有重要意义, 因为过高的泄露比可能导致模型通过记忆而非学习获得虚高的测试性能.

实验结果如表 7 所示, 各数据集的泄露比呈现出差异化的分布. VGGFace2 (78.43%)、Places365 (73.35%)、ImageNet (62.82%) 的泄露比远高于其他数据集, 这主要源于自然场景与大规模类别构成中的高度冗余. 例如, VGGFace2 由于姿态、光照和表情差异, 导致部分跨划分的人脸样本被判定为相似; Places365 的“教室-会议室”或“餐厅-咖啡厅”等场景难以完全区分; ImageNet 的跨类别图像在采集过程中也存在较强的视觉重叠.

TinyImageNet (8.74%) 处于中等水平. 它受限于 64×64 的分辨率和有限样本量, 特征空间中不同样本间可能表现出较强重叠. CIFAR-10 (5.15%)、CIFAR-100 (4.49%)、MNIST (4.99%) 以及 NABirds (2.39%) 的泄露比相对较低, 说明这些数据集在构建过程中较好地保持了训练与测试划分的独立性. MNIST 的手写数字虽风格各异, 但整体结构相对简单; CIFAR 系列在构建时充分考虑了类别均衡; 而 NABirds 在划分时有效避免了相同个体或近似场景的重复, 从而保证了较高的样本异质性.

泄露比的高低直接影响模型评估的可靠性. 高泄露比可能使模型在测试阶段表现优于实际应用, 而低泄露比则更能真实反映模型的泛化性能. 因此, 在选择基准数据集时, 泄露比应作为一个重要的考量因素.

3.4 知识任务适配评估结果分析

本节通过任务适配机制对各数据集进行综合评估. 评估过程遵循任务类型识别、权重调整和适配评分计算三个关键步骤. 在知识任务适配评估阶段, 本文通过分析多源预训练模型在目标数据集上的迁移性能差异模式来判别其任务属性. 为使该判别主

要反映预训练知识来源差异, 而尽量减少网络结构与容量因素带来的干扰, 本文统一采用 ResNet50 作为基础架构. ResNet50 结构成熟、表征能力稳定, 且在多类视觉任务中具有良好的可迁移性, 适合作为控制变量下的对比基准.

此外, 本文补充了以 ResNet18 为统一基架的对照实验 (见附录 B). 在保持预训练来源设置、微调策略与任务类型判别规则一致的前提下, ResNet18 所得到的任务类型判定结果与 ResNet50 保持一致, 说明本文基于迁移性能模式的任务识别结论具有良好的稳定性与可复现性.

3.4.1 基于迁移性能的任务类型识别

为识别数据集的任务类型, 本文测试了 8 个预训练模型在 10 个目标数据集上的迁移性能, 结果如表 8 所示. 这些模型均采用 ResNet50 架构, 分别在不同源数据集上预训练, 通过微调分类层适配到目标任务. 其中 LFW 和 CUB 作为陌生数据集, 用于验证方法的泛化性.

通过分析表 8 中的迁移性能模式, 系统识别数据集的任务类型, 将 VGGFace2、NABirds、LFW 和 CUB 识别为特定 (细粒度) 任务, 其余数据集 (MNIST、CIFAR 系列、TinyImageNet、Places365、ImageNet) 则为通用分类任务, 因为它们在通用模型上表现更佳.

需要指出的是, 本文在任务类型识别阶段以图像分类任务作为验证载体, 主要出于实验可控性与迁移性能模式可解释性的考虑, 而非方法本身对任务形式的限制. 基于迁移性能的任务识别机制关注的是“不同预训练知识来源在目标数据集上的相对适配关系”, 这一判别依据并不依赖于分类任务特有的损失函数或预测形式. 在目标检测与语义分割等更复杂视觉任务中, 不同数据集同样会在跨模型迁移表现上呈现出与任务判别难度、判别粒度及结

表 8 预训练模型在不同数据集上的迁移准确率 (%)
Table 8 Transfer accuracy of pre-trained models on different datasets (%)

源模型	目标数据集									
	MNIST	CIFAR-10	CIFAR-100	TinyImageNet	Places365	ImageNet	VGGFace2	NABirds	LFW ^[64]	CUB ^[66]
MNIST	83.68	7.42	10.98	0	0	0	0	0	0.81	0.78
CIFAR-10	25.19	91.90	72.81	25.82	61.65	79.41	52.47	65.00	1.59	0.67
CIFAR-100	5.30	8.35	76.92	10.17	31.55	56.00	23.96	42.66	2.66	0.86
TinyImageNet	1.39	3.91	20.52	44.14	26.26	62.34	20.83	36.87	4.01	1.85
Places365	0.45	0.68	2.51	8.81	45.60	39.30	8.78	26.19	5.18	2.59
ImageNet	0.13	0.29	1.50	7.85	27.61	69.15	5.56	38.71	8.80	28.39
VGGFace2	0.22	0.26	0.33	1.28	9.64	23.59	98.00	17.95	98.46	1.69
NABirds	0.24	0.33	0.42	1.35	2.60	26.85	1.59	73.63	18.29	80.84

构复杂性相关的差异模式. 因此, 该基于迁移性能的任务类型识别思想在方法层面具备向其他视觉任务推广的潜力. 本文在附录 C 中进一步给出了目标检测数据集上的指标统计结果, 用以验证核心评估维度在非分类任务中的可迁移性.

3.4.2 任务适配评分结果

根据任务适配机制 (式 (18)), 系统首先将各个指标归一化后计算加权得分, 在不同任务类型中通过权重配置的差异来体现最终效能值 $V_{\text{task}}(D, T)$.

表 9 展示了各数据集在通用任务与特定任务场景下的 $V_{\text{task}}(D, T)$ 结果. 在通用任务模式下, ImageNet (0.546) 得分最高, 体现了其在多样性、标签质量与复杂度方面的综合优势; Places365 (0.537) 和 TinyImageNet (0.519) 紧随其后, 表现出良好的平衡性; NABirds (0.510) 的得分同样接近这一水平, 表明该细粒度鸟类数据集具有一定的可迁移性. CIFAR-100 (0.456)、CIFAR-10 (0.439) 与 MNIST (0.420) 得分相对较低, 主要由于分辨率受限导致特征表达能力不足. VGGFace2 (0.420) 的得分处于最低水平, 反映出其在通用任务下的适配性受限于身份识别场景的任务特性.

当系统识别 NABirds 为鸟类细粒度任务并相应调整权重后, 其 $V_{\text{task}}(D, T)$ 从 0.510 提升至 0.536, 识别 VGGFace2 为身份识别任务并相应调整权重后, 其 $V_{\text{task}}(D, T)$ 从 0.420 提升至 0.477. 这种差异表明, 同一数据集在不同任务语境下的价值评估结果可能出现显著差异, 只有结合任务特性进行调整, 才能获得科学合理的知识效能评价.

需要指出的是, DKEE 并非由多个相互独立、可线性叠加的评估模块构成, 而是一个具有层级依赖关系的整体评估框架. 其中, 知识提炼与知识质量分析模块共同刻画数据集所蕴含知识的基本状态. 若移除其中任一模块, 评估将只能保留其中之一

维的信息, 无法同时覆盖“知识提炼”与“知识质量分析”这两个基础方面, 从而使对数据集知识效能的刻画不够完整且不充分. 任务适配机制通过对比未启用时的通用任务评分与启用后的特定任务评分之间的差异, 形成了对任务适配模块的消融验证, 从而验证了任务适配模块的必要性与有效性.

综上, 任务适配实验的核心结论在于: 数据集的知识效能并非静态属性, 而是与具体任务需求密切相关. 通用分类数据集在不同任务设定下表现出较为稳定的效能评分, 而细粒度数据集在引入任务适配机制后, 其知识效能得到提升. 这一结果表明, 若忽略任务语境, 细粒度数据集的价值可能被系统性低估; 而通过任务驱动的权重调整, DKEE 能够更准确地反映数据集在特定应用场景中的实际效能, 从而为面向具体任务的数据集选择与训练决策提供更具针对性的评估依据.

需要进一步说明的是, 上述任务适配评估结论建立在数据集处于相同或相近数据生成机制的假设下. 对于分布理想化、冗余度较低的合成数据, 其在知识提炼阶段可能呈现出较高的可压缩性与知识密度, 但该现象并不必然对应其在真实任务中的实际效能. 在此情形下, 任务适配机制通过引入跨域迁移表现作为约束, 能够在一定程度上揭示合成数据在 sim-to-real 场景下面临的分布不匹配问题, 从而避免仅依据蒸馏或结构性指标高估其在真实应用中的价值的问题.

3.5 应用局限性与改进方向

尽管 DKEE 能够统一刻画知识提炼、质量结构与任务适配, 但在面向数据交易等实际部署场景时仍存在若干现实挑战.

DKEE 的设计目标定位于数据集引入、训练前审计与数据治理等决策阶段的离线评估框架. 在该离线评估设定下, DKEE 的整体耗时处于可接受范围内 (如附录 D 所示). 但是, 本文观察到评估耗时随数据集规模增大而增加, 因此在对时效要求极高的实时数据交易场景中, 可能面临计算与时延压力. 为此, 后续工作可考虑采用分阶段评估策略, 例如利用轻量代理模型与子集采样进行快速预筛, 再对候选数据集触发完整评估, 从而在保证评估结论可靠性的同时有效降低整体计算成本.

此外, 本文对 DKEE 的方法设计与实验验证主要建立在图像数据集上, 而在实际数据评估与数据交易过程中, 数据集往往以多模态形式存在. 为更好地适配此类场景, 后续工作将进一步拓展 DKEE 至多模态数据集 (如图文、视频与语音). 具体而言,

表 9 不同任务类型下的数据集适配效能
Table 9 Dataset adaptation efficiency under different task types

数据集	通用任务 $V_{\text{task}}(D, T)$	特定任务 $V_{\text{task}}(D, T)$
MNIST	0.420	0.420
CIFAR-10	0.439	0.439
CIFAR-100	0.456	0.456
TinyImageNet	0.519	0.519
Places365	0.537	0.537
ImageNet	0.546	0.546
VGGFace2	0.420	0.477
NABirds	0.510	0.536

在知识提炼中建模跨模态对齐与互信息,在质量分析中评估模态一致性与语义漂移,在任务适配中进行多模态任务驱动的权重协同调整,从而提升DKEE在多模态数据评估中的通用性与可迁移性.

4 结束语

本文提出一种融合知识提炼、质量分析和任务适配的数据集知识效能评估框架,并在MNIST、CIFAR、TinyImageNet、Places365、VGFace2、ImageNet和NABirds等典型数据集上进行了验证.

本研究的主要贡献包括: 1) 通过无损数据集蒸馏技术量化了数据集的知识密度,揭示了不同数据集在信息承载能力上的显著差异——从MNIST的0.004保留率到ImageNet的0.846保留率; 2) 构建了涵盖完整性、一致性、噪声水平、类内/类间多样性、复杂度、歧义性、领域偏移、泄露比和基线性能十个维度的知识质量评估体系,实现了数据集质量的全面量化; 3) 设计了基于迁移性能的任务类型识别机制,通过权重调整实现任务相关的适配评分,NABirds在细粒度任务评估下效能值从0.510提升至0.536,验证了本文方法的有效性.

本文方法的核心创新在于认识到数据集效能评估不应脱离具体任务语境. 通过将知识提炼的客观度量与任务适配的主观需求相结合,为数据集选择和应用提供了更准确的决策支持. 这对于数据资源的合理配置和高效利用具有重要意义. 未来工作将在以下方向展开: 探索跨模态数据集(文本、音频、视频)的统一评估框架; 研究自动化的权重优化策略,减少人工干预; 将评估框架扩展到实际的数据交易和定价场景中.

附录 A NABirds 数据集中困难样本剔除实验补充结果

在前文实验结果的基础上,为进一步验证困难样本在细粒度识别任务中的作用,附录A给出了NABirds数据集中困难样本剔除实验的补充分析结果. 具体而言,本文依据模型在训练过程中的预测不确定性对样本进行排序,并从训练集中剔除占比最高的20%困难样本. 在保持网络结构、训练策略与评估协议不变的条件下,对模型进行重新训练与测试. 表A1给出了困难样本剔除前后模型在NABirds数据集上的分类准确率对比结果.

可以观察到,剔除困难样本后模型分类准确率由56.0%降至44.0%,出现显著下降. 这表明困难样本虽训练难度更高,但往往包含关键的细粒度判别线索,对模型学习具有重要贡献.

表 A1 NABirds 数据集中困难样本剔除前后的准确率对比 (%)

Table A1 Accuracy comparison before and after removing hard samples from NABirds dataset (%)

实验设置	分类准确率
原始数据集	56.0
剔除 20% 困难样本	44.0

附录 B 基于 ResNet18 的跨数据集迁移性能补充结果

为验证任务类型识别结论不依赖于特定主干网络,本文补充开展ResNet18对照实验,用于检验在网络容量与结构变化条件下,各源模型在不同目标数据集上的迁移性能相对分布是否保持一致,从而评估任务类型判别的稳定性与可复现性.

具体而言,在与ResNet50完全一致的预训练来源设置与迁移评估协议下,采用ResNet18作为统一主干完成跨数据集迁移测试,结果见表B1. 结果显示,ResNet18条件下的迁移性能模式与ResNet50基本一致,并得出一致的任务类型判定结论. 这表明本文的任务识别主要刻画源数据集与目标任务之间的知识适配关系,而非由特定主干的容量或结构差异所驱动.

附录 C COCO 数据集在通用任务下的指标评估结果

为验证DKEE核心评估维度对非分类数据集的适用性与稳定性,本文以COCO数据集为代表,给出了其在通用评估设定下的多维指标统计结果(如表C1所示). 本文并未针对目标检测任务单独设计评估流程,而是将COCO视为一种具有检测属性的数据集,用以检验DKEE所采用的知识提炼与质量分析指标在不同任务数据形态下的可解释性. 从指标分布来看,COCO在知识密度、完整度、标签一致性等维度上呈现出稳定且合理的统计特征,说明这些评估维度并不依赖分类任务特有的假设. 在歧义性与复杂度等指标上,未出现明显偏置或失效现象. 上述结果表明,DKEE所采用的核心评估维度在非分类数据集上同样具有可解释性,为其向目标检测与其他复杂视觉任务扩展提供了初步实证依据.

附录 D 单次完整评估流程的计算开销统计

为评估本文方法在实际应用中的计算成本,本文对单次完整评估流程的典型耗时进行了统计分

表 B1 基于 ResNet18 的跨数据集迁移分类准确率 (%)
Table B1 Cross-dataset transfer classification accuracy based on ResNet18 (%)

源模型	目标数据集							
	MNIST	CIFAR-10	CIFAR-100	TinyImageNet	Places365	ImageNet	VGGFace2	NABirds
MNIST	99.36	17.37	3.31	1.10	0.27	0.10	0.29	0.26
CIFAR-10	44.11	81.13	14.64	4.57	0.73	0.29	0.32	0.48
CIFAR-100	62.41	67.56	79.26	17.80	2.34	1.29	0.30	0.42
TinyImageNet	31.38	55.80	33.59	99.36	17.97	16.04	2.38	1.52
Places365	91.60	68.84	39.47	41.81	52.17	32.92	16.50	3.69
ImageNet	94.64	77.74	52.80	58.05	38.80	63.61	25.24	22.14
VGGFace2	94.59	78.20	52.56	57.83	38.91	63.51	25.65	0.23
NABirds	54.28	27.35	8.23	3.50	4.47	1.81	1.23	22.41

表 C1 COCO 数据集在通用分类任务下的多维评估结果
Table C1 Multi-dimensional evaluation results of the COCO dataset for the general classification task

指标	结果
知识密度	95%
完整度	100%
标签一致性	100%
噪声比	93.08%
类内多样性	0.034485
类间多样性	0.01
均值	0.0018
方差	0.0022
熵	0.7269
基线模型性能	36.28%
歧义性	0.9984
领域偏移	0.059
泄露比	11.03%
知识效能	4.47

析. 所统计的时间包括知识提炼与数据质量分析两个核心阶段, 所有实验均在相同硬件配置与实现框架下完成, 以保证结果的可比性.

表 D1 给出了在不同规模与任务属性的数据集

表 D1 单次评估流程的耗时统计 (h)
Table D1 Time cost statistics of a single evaluation process (h)

数据集	知识提炼	质量分析
MNIST	1	0.5
CIFAR-10	2	0.5
TinyImageNet	8	1.0
Places365	15	15.0
VGGFace2	30	20.0
ImageNet	20	20.0
NABirds	5	1.0

上, 单次完整评估流程中各阶段的平均耗时情况. 可以观察到, 评估时间随数据集规模与任务复杂度增大而增加. 其中, 知识提炼阶段由于涉及数据集蒸馏与多轮优化, 其计算开销在大规模或高语义复杂度数据集 (如 Places365、VGGFace2) 上占据主导地位; 相比之下, 质量分析阶段整体耗时较为稳定, 主要受特征提取与统计分析过程影响.

参考文献

1 National Development and Reform Commission, National Data Administration, Ministry of Education, Ministry of Finance, State Administration for Financial Regulation, China Securities Regulatory Commission. Guiding opinions on promoting the high-quality development of the data industry [Online], available: https://www.gov.cn/zhengce/zhengceku/202412/content_6995430.htm, September 1, 2025
(国家发展改革委, 国家数据局, 教育部, 财政部, 金融监管总局, 中国证监会. 关于促进数据产业高质量发展的指导意见 [Online], available: https://www.gov.cn/zhengce/zhengceku/202412/content_6995430.htm, 2025-09-01)

2 National Development and Reform Commission, National Data Administration, Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security, State Administration for Market Regulation. Implementation plan for improving data circulation security governance and better promoting the marketization and valorization of data elements [Online], available: https://www.ndrc.gov.cn/xxgk/zcfb/tz/202501/t20250115_1395692.html, September 1, 2025
(国家发展改革委, 国家数据局, 中央网信办, 工业和信息化部, 公安部, 市场监管总局. 关于完善数据流通安全治理更好促进数据要素市场化价值化的实施方案 [Online], available: https://www.ndrc.gov.cn/xxgk/zcfb/tz/202501/t20250115_1395692.html, 2025-09-01)

3 Krizhevsky A, Hinton G. Learning Multiple Layers of Features From Tiny Images, Technical Report TR-2009-001, Department of Computer Science, University of Toronto, Canada, 2009.

4 Deng J, Dong W, Socher R, Li L J, Li K, Li F F. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE, 2009. 248–255

5 Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv: 1503.02531, 2015.

6 Romero A, Ballas N, Kahou S E, Chassang A, Gatta C, Bengio Y. FitNets: Hints for thin deep nets. arXiv preprint arXiv: 1412.6550, 2014.

7 Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: Proceedings of the 5th International Con-

- ference on Learning Representations. Toulon, France: OpenReview.net, 2017.
- 8 Tung F, Mori G. Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019. 1365–1374
- 9 Liu Y F, Cao J J, Li B, Yuan C F, Hu W M, Li Y X. Knowledge distillation via instance relationship graph. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 7089–7097
- 10 Tian Y L, Krishnan D, Isola P. Contrastive representation distillation. arXiv preprint arXiv: 1910.10699, 2019.
- 11 Heo B, Kim J, Yun S, Park H, Kwak N, Choi J Y. A comprehensive overhaul of feature distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019. 1921–1930
- 12 Liu Y F, Cao J J, Li B, Hu W M, Ding J T, Li L, et al. Cross-architecture knowledge distillation. *International Journal of Computer Vision*, 2024, **132**(8): 2798–2824
- 13 Yang Z Y, Liu J G, Huang J, He X D, Mei T, Xu C L. Cross-modal contrastive distillation for instructional activity anticipation. In: Proceedings of the 26th International Conference on Pattern Recognition (ICPR). Montreal, Canada: IEEE, 2022. 5002–5009
- 14 Yu J H, Yang L J, Xu N, Yang J C, Huang T. Slimmable neural networks. In: Proceedings of the International Conference on Learning Representations. New Orleans, USA: OpenReview.net, 2019.
- 15 Liu Y F, Cao J J, Li B, Hu W M, Maybank S. Learning to explore distillability and sparsability: A joint framework for model compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(3): 3378–3395
- 16 Liu Y F, Cao J J, Bai W M, Li B, Hu W M. Learning from the raw domain: Cross modality distillation for compressed video action recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023. 1–5
- 17 Fang Z Y, Wang J F, Hu X W, Wang L J, Yang Y Z, Liu Z C. Compressing visual-linguistic model via knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 1408–1418
- 18 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada: Curran Associates Inc., 2012. 1097–1105
- 19 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 6000–6010
- 20 Wang T Z, Zhu J Y, Torralba A, Efros A A. Dataset distillation. In: Proceedings of the International Conference on Learning Representations. New Orleans, USA: OpenReview.net, 2019.
- 21 Zhao B, Mopuri K R, Bilen H. Dataset condensation with gradient matching. arXiv preprint arXiv: 2006.05929, 2020.
- 22 Nguyen T, Chen Z R, Lee J. Dataset meta-learning from kernel ridge-regression. In: Proceedings of the International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2021.
- 23 Nguyen T, Novak R, Xiao L C, Lee J. Dataset distillation with infinitely wide convolutional networks. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2021. Article No. 397
- 24 Zhao B, Bilen H. Dataset condensation with distribution matching. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2023. 6503–6512
- 25 Zhao G L, Li G B, Qin Y P, Yu Y Z. Improved distribution matching for dataset condensation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 7856–7865
- 26 Cazenavette G, Wang T Z, Torralba A, Efros A A, Zhu J Y. Dataset distillation by matching training trajectories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 10708–10717
- 27 Yin Z Y, Xing E, Shen Z Q. Squeeze, recover and relabel: Dataset condensation at ImageNet scale from a new perspective. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 3219
- 28 Zhang X, Du J W, Liu P, Zhou J T. Breaking class barriers: Efficient dataset distillation via inter-class feature compensator. In: Proceedings of the 13th International Conference on Learning Representations. Singapore: OpenReview.net, 2025.
- 29 Chen M Y, Du J W, Huang B, Wang Y, Zhang X B, Wang W. Influence-guided diffusion for dataset distillation. In: Proceedings of the 13th International Conference on Learning Representations. Singapore: OpenReview.net, 2025.
- 30 Qi D, Li J, Gao J Y, Dou S G, Tai Y, Hu J L. Towards universal dataset distillation via task-driven diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 10557–10566
- 31 Guo Z Y, Wang K, Cazenavette G, Li H, Zhang K P, You Y. Towards lossless dataset distillation via difficulty-aligned trajectory matching. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2024.
- 32 Shao S T, Zhou Z K, Chen H R, Shen Z Q. Elucidating the design space of dataset condensation. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 3146
- 33 Schioppa A. Efficient sketches for training data attribution and studying the loss landscape. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 1190
- 34 Liu T F, Liu Y F, Li B, Hu W M, Li Y M, Ma C G. Noise-optimized distribution distillation for dataset condensation. In: Proceedings of the 33rd ACM International Conference on Multimedia. Dublin, Ireland: ACM, 2025. 10352–10360
- 35 Liu P, Du J W. The evolution of dataset distillation: Toward scalable and generalizable solutions. arXiv preprint arXiv: 2502.05673, 2025.
- 36 Northcutt C G, Athalye A, Mueller J. Pervasive label errors in test sets destabilize machine learning benchmarks. In: Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1). Virtual Event: OpenReview.net, 2021.
- 37 Kang B Y, Xie S N, Rohrbach M, Yan Z C, Gordo A, Feng J S, et al. Decoupling representation and classifier for long-tailed recognition. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview.net, 2020.
- 38 Uddin S, Lu H H. Dataset meta-level and statistical features affect machine learning performance. *Scientific Reports*, 2024, **14**(1): Article No. 1670
- 39 Ho T K. A data complexity analysis of comparative advantages of decision forest constructors. *Pattern Analysis & Applications*, 2002, **5**(2): 102–112
- 40 Nguyen C, Hassner T, Seeger M, Archambeau C. LEEP: A new

- measure to evaluate transferability of learned representations. In: Proceedings of the 37th International Conference on Machine Learning. Virtual Event: PMLR, 2020. 7294–7305
- 41 You K C, Liu Y, Wang J M, Long M S. LogME: Practical assessment of pre-trained models for transfer learning. In: Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR, 2021. 12133–12143
 - 42 Swayamdipta S, Schwartz R, Lourie N, Wang Y Z, Hajishirzi H, Smith N A, et al. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Virtual Event: ACL, 2020. 9275–9293
 - 43 Dai M Y, Raffee A H, Jain A, Correa J. Evaluating transferability in retrieval tasks: An approach using MMD and kernel methods. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 22390–22400
 - 44 Li Juan-Juan, Qin Rui, Ding Wen-Wen, Wang Ge, Wang Tan, Wang Fei-Yue. A new framework for web3-powered decentralized autonomous organizations and operations. *Acta Automatica Sinica*, 2023, **49**(5): 985–998
(李娟娟, 秦蕊, 丁文文, 王戈, 王坛, 王飞跃. 基于 web3 的去中心化自治组织与运营新框架. 自动化学报, 2023, **49**(5): 985–998)
 - 45 Zhang Hui, Yan Xing-Yu, Mao Jian-Xu, Biekezhati Baheti, Du Rui, Wang Yao-Nan. A review of intelligent data collaborative inference techniques for source-grid-load systems. *Acta Automatica Sinica*, 2025, **51**(11): 2387–2411
(张辉, 颜星雨, 毛建旭, 别克扎提·巴合提, 杜瑞, 王耀南. 面向源网荷的智能化数据协同推断技术研究综述. 自动化学报, 2025, **51**(11): 2387–2411)
 - 46 Qiyang Li-Wei, Wang Shuai, Yuan Yong, Ni Xiao-Chun, Wang Fei-Yue. Smart contracts: Architecture and research progresses. *Acta Automatica Sinica*, 2019, **45**(3): 445–457
(欧阳丽炜, 王帅, 袁勇, 倪晓春, 王飞跃. 智能合约: 架构及进展. 自动化学报, 2019, **45**(3): 445–457)
 - 47 Hotelling H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 1933, **24**(6): 417–441
 - 48 van der Maaten L, Hinton G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008, **9**: 2579–2605
 - 49 Ester M, Kriegel H P, Sander J, Xu X W. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland, USA: AAAI Press, 1996. 226–231
 - 50 Lin J H. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 1991, **37**(1): 145–151
 - 51 Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR, 2021. 8748–8763
 - 52 Kingma D P, Welling M. Auto-encoding variational Bayes. In: Proceedings of the International Conference on Learning Representations. Banff, Canada: OpenReview.net, 2014.
 - 53 Kullback S, Leibler R A. On information and sufficiency. *The Annals of Mathematical Statistics*, 1951, **22**(1): 79–86
 - 54 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 770–778
 - 55 Simonyan S, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: OpenReview.net, 2015.
 - 56 Huang G, Liu Z, van der Maaten L, Weinberger K Q. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 2261–2269
 - 57 Tan M X, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR, 2019. 6105–6114
 - 58 Sandler M, Howard A, Zhu M L, Zhmoginov A, Chen L C. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 4510–4520
 - 59 Han D, Kim J, Kim J. Deep pyramidal residual networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 6307–6315
 - 60 Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278–2324
 - 61 Le Y, Yang X. Tiny ImageNet Visual Recognition Challenge, Technical Report, Stanford University, USA, 2015.
 - 62 López-Cifuentes A, Escudero-Viñolo M, Bescós J, García-Martín A. Semantic-aware scene recognition. *Pattern Recognition*, 2020, **102**: Article No. 107256
 - 63 Cao Q, Shen L, Xie W D, Parkhi O M, Zisserman A. VGGFace2: A dataset for recognising faces across pose and age. In: Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). Xi'an, China: IEEE, 2018. 67–74
 - 64 Huang G B, Ramesh M, Berg T, Learned-Miller E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In: Proceedings of the Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition. Marseille, France: ECCV 2008 Workshop, 2008. 1–11
 - 65 van Horn G, Branson S, Farrell R, Haber S, Barry J, Ipeirotis P. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 595–604
 - 66 Wah C, Branson S, Welinder P, Perona P, Belongie S. The Caltech-UCSD Birds-200-2011 Dataset, Technical Report CNS-TR-2011-001, California Institute of Technology, USA, 2011.



丁昕苗 山东工商学院信息与电子工程学院教授。2013 年获得中国矿业大学 (北京) 博士学位。主要研究方向为图像与视频理解, 机器学习和网络安全。

E-mail: dingxinmiao@126.com

(DING Xin-Miao Professor at the School of Information and Electronic Engineering, Shandong Technology and Business University. She received her Ph.D. degree from China University of Mining and Technology, Beijing in 2013. Her research interests include image and video understanding, machine learning, and internet security.)



李好男 山东工商学院信息与电子工程学院硕士研究生。主要研究方向为机器学习, 模式识别。

E-mail: 2023420059@sdtbu.edu.cn

(LI Hao-Nan Master student at the School of Information and Electronic Engineering, Shandong Tech-

nology and Business University. His research interests include machine learning and pattern recognition.)



刘雨帆 中国科学院自动化研究所副研究员. 2023 年获得中国科学院大学博士学位. 主要研究方向为计算机视觉, 模型压缩和视频理解. 本文通信作者. E-mail: liuyufan@ia.ac.cn
(**LIU Yu-Fan** Associate researcher at the Institute of Automation, Chinese Academy of Sciences. She received her Ph.D. degree from University of Chinese Academy of Sciences in 2023. Her research interests include computer vision, model compression, and video understanding. Corresponding author of this paper.)



刘彤飞 中国科学院自动化研究所博士研究生. 主要研究方向为数据集压缩. E-mail: liutongfei@ia.ac.cn
(**LIU Tong-Fei** Ph.D. candidate at the Institute of Automation, Chinese Academy of Sciences. His main research interest is dataset compression.)



张克卿 中国科学院自动化研究所博士研究生. 主要研究方向为大语言模型安全, 知识蒸馏. E-mail: zhangkeqing2025@ia.ac.cn
(**ZHANG Ke-Qing** Ph.D. candidate at the Institute of Automation, Chinese Academy of Sciences. His research interests include large language model safety and knowledge distillation.)

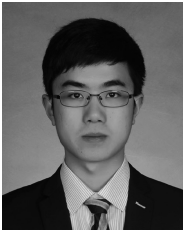


李 兵 中国科学院自动化研究所研究员. 2009 年获得北京交通大学博士学位. 主要研究方向为视频理解, 颜色恒常性, 视觉显著性, 多实例学习和网络内容安全. E-mail: libing@ia.ac.cn
(**LI Bing** Researcher at the Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree from Beijing Jiaotong University in 2009. His research interests include video understanding, color constancy, visual saliency, multi-instance learning, and web content security.)

tute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree from Beijing Jiaotong University in 2009. His research interests include video understanding, color constancy, visual saliency, multi-instance learning, and web content security.)



原春锋 中国科学院自动化研究所研究员. 2010 年获得中国科学院自动化研究所博士学位. 主要研究方向为模式识别, 计算机视觉. E-mail: yuan Chun-feng@ia.ac.cn
(**YUAN Chun-Feng** Researcher at the Institute of Automation, Chinese Academy of Sciences. She received her Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2010. Her research interests include pattern recognition and computer vision.)



王 坚 中国科学院自动化研究所工程师. 主要研究方向为机器学习, 计算机视觉. E-mail: wangjian@ia.ac.cn
(**WANG Jian** Engineer at the Institute of Automation, Chinese Academy of Sciences. His research interests include machine learning and computer vision.)



胡卫明 中国科学院自动化研究所研究员. 1998 年获得浙江大学博士学位. 主要研究方向为视觉运动分析, 网络不良信息识别和网络入侵检测. E-mail: huweiming@ia.ac.cn
(**HU Wei-Ming** Researcher at the Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree from Zhejiang University in 1998. His research interests include visual motion analysis, recognition of web objectionable information, and network intrusion detection.)