E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2025)11-3438-13

论文引用格式:Wang Y B, Hong X P and Huang Z W. 2025. Benchmark dataset and framework for continual AI-generated image detection. Journal of Image and Graphics, 30(11):3438-3450(王亚斌, 洪晓鹏, 黄智武. 2025. 面向AI生成图像持续检测基准数据集与框架研究. 中国图象图形学报, 30(11):3438-3450)[DOI:10.11834/jig.250167]

面向AI生成图像持续检测基准数据集与框架研究

王亚斌[1],洪晓鹏[2*],黄智武[3]

1. 西安交通大学,西安　710049; 2. 哈尔滨工业大学,哈尔滨　150001; 3. 南安普顿大学,南安普顿　SO171BJ,英国

摘　要:**目的**　针对人工智能生成内容(artificial intelligence generated content,AIGC)技术快速发展带来的高度逼真图像滥用风险,以及现有检测方法难以适应持续涌现的新型生成模型且缺乏持续学习能力的挑战,构建了面向该挑战的AI生成图像持续检测基准数据集,并提出一套针对该问题的持续检测框架。**方法**　首先,构建了面向持续学习的AI生成图像检测基准数据集,包含5种主流生成模型样本及真实图像,并按持续学习任务流组织其结构;其次,系统定义并研究了持续学习在AI生成图像检测任务下面临的问题,特别关注了现实场景约束下新颖的"混合二类与单类"的增量学习场景,并设计了3种基于不同程度样本回放约束的基准;最后,针对不同基准场景,改进现有持续学习方法进行任务适配,并为最严苛的无回放场景提出了一种通用转换框架,以修复在此场景下失效的方法。**结果**　在所提数据集上的实验验证了所提基准和方法的有效性。在允许回放的场景下,适配后的方法能够实现增量检测。在最严格的无回放场景下,传统无回放方法性能严重下降甚至失效,而应用本文提出的通用转换框架后,这些方法的性能显著提升,有效提升了检测准确率和版权识别能力,并大幅抑制了灾难性遗忘。**结论**　本文成功构建了面向持续学习的AI生成图像检测基准,深入分析了其中的关键挑战,并提出了有效的持续检测策略和解决方案,特别是为无回放场景下的持续学习提供了创新框架。研究成果为开发能够应对不断演进的AI生成技术的稳健、适应性强的检测系统提供了重要的方法论支撑和实证依据。本文相关数据和代码开源在https://huggingface.co/datasets/nebula/CAID和https://doi.org/10.57760/sciencodb.29781。

关键词:人工智能生成内容(AIGC)检测;持续学习(CL);深度伪造;基准数据集;数据集;灾难性遗忘;知识蒸馏(KD);提示调整

Benchmark dataset and framework for continual AI-generated image detection

Wang Yabin[1], Hong Xiaopeng[2*], Huang Zhiwu[3]

1. Xi'an Jiaotong University, Xi'an　710049, China; 2. Harbin Institute of Technology, Harbin　150001, China; 3. University of Southampton, Southampton　SO171BJ, United Kingdom

Abstract: **Objective** The rapid evolution of artificial intelligence-generated content (AIGC), particularly text-to-image models such as Stable Diffusion, Midjourney, and DALL-E, presents a dual-use dilemma. While enabling unprecedented creative applications, the proliferation of hyper-realistic AI-generated images poses severe societal risks, including the spread of disinformation, sophisticated fraud, and intellectual property infringement. This dynamic environment underscores the urgent need for robust and reliable detection technologies. However, conventional detection methods, which rely

收稿日期:2025-04-10;**修回日期**:2025-07-03;**预印本日期**:2025-07-10

*通信作者:洪晓鹏　hongxiaopeng@ieee.org

基金项目:国家自然科学基金项目(62376070, 62076195);中央高校基本科研业务费专项资金资助(AUGA-5710011522)

Supported by: National Natural Science Foundation of China (62376070, 62076195); Fundamental Research Funds for the Central Universities (AUGA-5710011522)

on offline training, are fundamentally ill-equipped for this dynamic environment. The "train-once, deploy-forever" paradigm causes them to fail when faced with the ceaseless emergence of novel generative models, leading to a swift decay in performance. This critical limitation highlights the need for a more adaptive approach. Continual learning (CL), a paradigm designed to enable models to learn sequentially from a continuous data stream while mitigating "catastrophic forgetting," offers a promising solution. Yet, its application to the diverse and fast-paced domain of modern AIGC detection is hindered by two significant, unaddressed gaps. First, there is a conspicuous absence of a specialized, large-scale benchmark for systematically evaluating continual AIGC detection methods. Second, and more critically, a unique real-world data constraint creates a novel and formidable learning challenge. This constraint arises because new generative models (positive samples) are often released and become accessible, while their corresponding, high-quality real training images (negative samples) remain proprietary and unavailable. This condition leads to a novel "mixed dual- and single-class" incremental learning problem: Initial learning tasks may possess both positive and negative samples, but subsequent tasks are often restricted to positive samples only. This scenario fundamentally violates the core assumptions of most existing CL algorithms, rendering them ineffective. To address these profound challenges, this study establishes a foundational methodology for the continual detection of AI-generated images. Our primary objective is to construct a comprehensive benchmark and a robust framework capable of dynamically adapting to an ever-expanding stream of generative models, particularly under these realistic and challenging data constraints. **Method** First, we introduce and release a continual AI-generated image detection (CAID) benchmark, the first large-scale dataset specifically tailored for this task. It contains high-quality images from five state-of-the-art generative models (Stable Diffusion v1.5, DALL-E 2, Imagen, Midjourney, and Parti) and corresponding real images, organized into a sequential task stream to simulate the real-world emergence of new AIGC technologies. Upon this benchmark, we formally define the CAID problem, which requires a model not only to perform binary classification (real vs. fake) but also to achieve multiclass source attribution (i.e., identifying the specific generator model), a task crucial for "copyright identification". To standardize evaluation, we design three benchmarks with increasing difficulty based on data replay constraints. Scenario 1 (full replay): a lenient setting where a small buffer of historical positive (fake) and negative (real) samples can be replayed. Scenario 2 (negative-only replay): a more practical setting where only historical negative samples can be replayed, reflecting IP or privacy restrictions. Scenario 3 (no replay): the most stringent setting, completely forbidding access to any past samples and forcing the model to learn incrementally from single-class data. Then, we propose tailored solutions for these scenarios. For Scenarios 1 and 2, we adapt existing CL methods using a "negative sample sharing" mechanism. This solution ensures that a small set of real images from the initial task is consistently available, providing the necessary negative class information to stabilize training and prevent the failure of standard loss functions. For the most severe no-replay scenario (Scenario 3), in which our experiments find that existing methods catastrophically fail, we propose a novel universal conversion framework. This framework is engineered to rescue failing methods by systematically addressing the core breakdown points of loss function invalidation, severe classifier output bias, and feature representation drift. We integrate three synergistic components to achieve this objective. First, we use knowledge distillation (KD), in which the model from the previous task acts as a "teacher" to guide the current model, preserving knowledge of past classes by matching its output logits. Second, we use a cosine-normalized classifier to replace the standard linear layer through cosine normalization (CN), which calibrates output logits across all tasks to mitigate bias. Finally, the framework utilizes prompt tuning (PT) by freezing the pretrained vision transformer backbone and training only a small set of learnable "prompt" parameters. This parameter-efficient approach drastically reduces overfitting on new single-class data and preserves the integrity of the learned feature space. **Result** Our extensive experiments on the CAID benchmark validate the efficacy of our methodology. In Scenarios 1 and 2, adapted CL methods like FOSTER and S-Prompts perform well, confirming the value of replaying historical data. The most compelling results emerge from Scenario 3. Standard replay-free methods (LwF, EWC, and S-Prompts) completely collapse, with their average accuracy (AA) dropping to random-guess levels (~50%), and suffer from extreme catastrophic forgetting. The application of our universal conversion framework resurrects these methods. Their AA surges dramatically to 65%, and, critically, catastrophic forgetting is largely eliminated, with AF scores dropping to near-zero levels. This finding provides unequivocal evidence that our framework successfully navigates the extreme challenges of no-replay, single-class incremental learning. An in-depth abla-

tion study confirms that all three components (KD, CN, and PT) are indispensable and work synergistically to achieve this remarkable performance recovery. Furthermore, t-SNE visualizations verify that our framework significantly reduces feature drift, while frequency spectrum analysis highlights the inherent difficulty of the CAID dataset in comparison with traditional deepfakes, justifying the need for our approach. **Conclusion** This study makes a foundational contribution to the critical and burgeoning field of AIGC detection. We have constructed and released the first large-scale benchmark for CAID, formally defined the problem, and identified the novel and practical "mixed dual- and single-class" learning challenge. Our proposed solutions, particularly the innovative universal conversion framework, provide a robust and effective strategy for developing adaptive detection systems, demonstrating how to maintain performance even under the most stringent real-world data constraints. By open-sourcing our dataset and code, we aim to provide a solid foundation and catalyze future research in this vital area. Ultimately, our findings offer significant methodological support for building the next generation of future-proof detection systems capable of keeping pace with the relentless evolution of AI generation technologies.

Key words: artificial intelligence generated content (AIGC) detection; continual learning (CL); deepfake; benchmark dataset; dataset; catastrophic forgetting; knowledge distillation (KD); prompt tuning

0 引 言

近年来，人工智能生成内容（artificial intelligence generated content，AIGC）技术发展迅猛，在图像生成领域尤为突出。以 Stable Diffusion、Midjourney、DALL-E 2等为代表的模型已能生成高度逼真且富有多样性的图像，其质量甚至可与人类创作无异。这类技术在创意设计、艺术创作及娱乐产业中展现出广阔应用前景，但与此同时，生成内容的滥用也带来了严峻挑战。这些高度仿真的图像可被用于散布虚假信息、侵犯知识产权、实施网络欺诈等不良行为，对社会诚信体系、信息安全及版权保护构成重大威胁（Cetinic 和 She，2022）。因此，研发 AI 生成图像检测技术已成为当务之急。

学术界对早期 AI 生成模型的伪造检测已有充分研究。相关研究者提出了 FaceForensics++（Rössler 等，2019）、Celeb-DF（Li 等，2020）等广泛使用的评测数据集，并提出了基于频域特征分析或视觉伪影识别等技术的检测方法（Wang 等，2020；Li 等，2018；卓文琦 等，2023；冯才博 等，2024；王诗雨 等，2025；薛子育 等，2025；张晶 等，2025）。然而，这些方法主要针对早期特定类型的图像伪造（如人脸替换）或传统生成对抗网络（generative adversarial networks，GAN）模型生成的内容。面对新一代生成模型（尤其是扩散模型）所产出的高真实度且多样化的图像时，现有检测手段的有效性受到了严峻考验。

更为棘手的是，AI 图像生成技术正处于快速迭代与持续革新阶段，各类新型模型、算法及架构不断涌现。在此动态环境下，传统"静态训练"的检测模型（即在特定数据集上完成训练后离线部署使用）表现出明显泛化能力不足（Wang 等，2020）。这类检测器往往难以有效识别训练阶段未曾接触的新型生成模型产出的图像，致使其检测性能随时间推移迅速衰减，无法跟进 AI 生成技术的演进步伐。针对上述根本性挑战，持续学习（continual learning，CL）理念提供了一条颇具潜力的研究路径。持续学习旨在使模型能从持续数据流中不断吸收新知识，同时有效抑制"灾难性遗忘"现象（Lopez-Paz 和 Ranzato，2017），这一特性恰好契合 AI 生成图像检测领域对技术持续更新的迫切需求。

近年来，学界已在传统 Deepfake 检测中初步探索了持续学习的应用价值（Marra 等，2019；Kim 等，2021；Li 等，2023；Yang 等，2022），但是将其拓展至当前多元化、快速发展的 AI 生成图像检测领域时，会面临独特的难题。诸多前沿 AI 生成模型（如 Midjourney、DALL-E 2等）并未公开其核心技术细节或训练数据，导致持续学习过程中，研究者常能获取这些新模型生成的图像（正样本），却难以收集与之匹配的高质量真实图像。这形成了一种新型且复杂的"增量阶段负样本缺失"的持续学习问题，即模型在基础阶段进行标准的二分类训练，而在后续增量阶段则退化为仅有少量单类正样本可用的单类学习。

这一独特单类持续学习问题，对当前所有主流的持续学习方法构成了根本性挑战。首先，虽然基于样本回放（Rebuffi 等，2017；Lopez-Paz 和 Ranzato，2017；Wu 等，2019；Zhou 等，2021；Wang 等，2022a；Marra 等，2019）方法是目前的主流，这类方法在模型学习新任务的过程中，从一个预设的样本库中调取

并复现少量旧数据进行同步训练，以此巩固模型对旧知识的记忆。然而，在AI生成图像检测领域，样本回放方法因其巨大的存储开销和潜在的隐私版权风险而存在明显局限。其次，基于正则化的方法（Li和Hoiem，2018；Kirkpatrick等，2017；Wang等，2022b，2023；Wang等，2022c，d）虽然摒弃了样本存储，转而通过在学习新任务时对模型参数施加约束，来保护对旧知识至关重要的网络权重。例如Memory Aware Synapses（Aljundi等，2018）通过测量模型输出对参数变化的敏感度来进行约束。但是这些方法的核心机制均依赖于多类别数据提供的监督信号来有效更新模型或估算参数重要性，在单类学习场景下它们会性能骤降甚至完全失效（Hu等，2021）。

针对这些挑战，本文致力于构建一个具备持续适应能力的AI生成图像检测框架。为实现这一目标，首先设计并构建了面向持续学习的现代AI生成图像检测基准数据库，该库涵盖5种主流生成模型的高质量样本与真实图像，部分样例如图1所示。其次，系统性地定义并研究了持续AI生成图像检测（continual AI image detection，CAID）问题，深入分析了现实数据获取限制下的挑战及其对模型学习和灾难性遗忘的影响。最后，为解决现有无回放方法在“负样本缺失”场景下的失效问题，本文提出了一个通用的转换框架。该框架通过整合知识蒸馏、余弦归一化和提示调整技术，对现有方法进行系统性改造，使其能够在最严苛的单类增量学习条件下恢复性能，并有效抑制灾难性遗忘。

本文的主要贡献如下：1）定义并深入研究了持续AI生成图像检测（CAID）这一新问题，并为此构建了面向持续学习的大型基准数据库。2）特别关注并分析了CAID中由于现实数据限制而产生的新颖的“混合两类与单类”持续学习场景。3）针对所提出的问题和特定场景，设计了系统的基准评测方案，以标准化评估流程。最后，提出了适配现有方法及创新的解决方案。本文相关数据和代码开源在https://huggingface.co/datasets/nebula/CAID和https://doi.org/10.57760/sciencodb.29781。

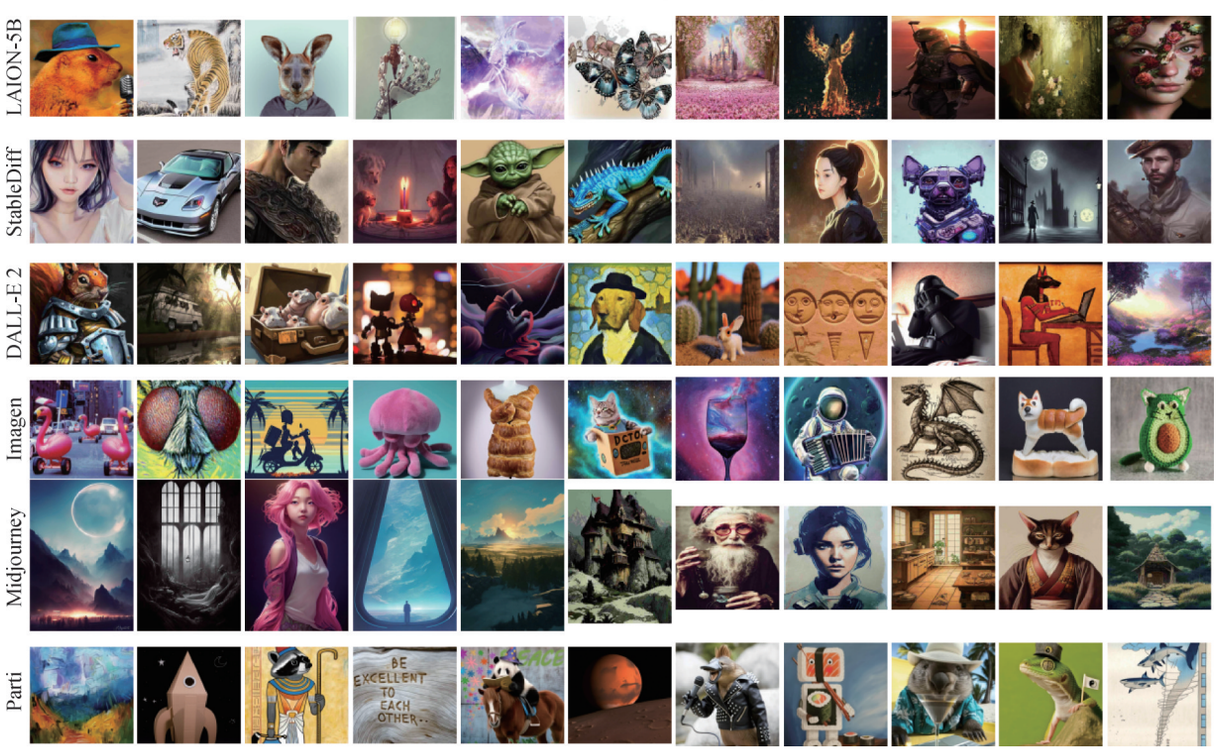

图1　面向持续学习的AI生成图像检测基准数据集示例

Fig. 1　Examples for continual learning AI-generated image detection benchmark dataset

1　连续AI伪造检测数据集

1.1　数据集构建

针对图像生成技术的迅猛发展及新型伪造作品层出不穷的现实挑战，为系统评估检测算法在动态环境中的适应能力，本文构建了一个评测数据集，并在此基础上明确定义了持续学习AI生成图像检测问题及其评价体系。

本文数据集构建遵循公开获取性、图像质量高

及来源可靠三大原则，主要包含两类核心样本：AI生成图像和真实图像。由于当前AI技术发展，其中样本多为美学价值高的艺术作品。作为检测任务的正样本，AI生成图像来自5种目前备受关注的前沿文本到图像生成模型：Stable Diffusion v1.5（SD）（Rombach等，2022）、DALL-E 2（Ramesh等，2022）、Imagen（Saharia等，2022）、Midjourney（Holz，2022）及Parti（Yu等，2022）。这些模型均能根据文本描述生成具有高度创造性和逼真度的图像。而作为负样本的真实图像则精选自LAION-5B数据集（Schuhmann等，2022）。图1直观展示了所建立数据集各类图像的代表性样本。所有伪造数据均从公开互联网可信数据源获得。

最终构建数据集的详细统计信息见表1。表1列出了每个模型对应的训练集和测试集的图像数量。数据集的一个关键特征，即除Stable Diffusion外，其他4个AI模型目前均未公开其训练所用的对应真实数据，直接导致了后续持续学习研究所面临的特殊挑战。

表1　所提出数据集统计信息

Table 1　Statistics of the proposed dataset

/幅

模型	训练集（真）	训练集（伪）	测试集（真）	测试集（伪）
SD	62 154	65 556	1 030	1 086
DALL-E 2	无	775	186	196
Imagen	无	178	44	46
Midjourney	无	4 306	1 027	1 077
Parti	无	155	38	40

1.2　研究问题定义

持续学习AI检测要求检测模型能够从不断出现的数据流（记为$\boldsymbol{D}_0, \boldsymbol{D}_1, \cdots, \boldsymbol{D}_S$，其中，$t=0$代表基础学习阶段，$t=1, \cdots, S$表示后续增量阶段）中持续获取新知识，同时有效抑制对已掌握知识的灾难性遗忘现象。该任务的最终目标是构建一个全能检测器，不仅能准确识别所有已学习过的生成模型类型，还能确定其具体生成模型来源（即实现版权溯源）。

在本研究的实验设计中，模型首先在基础学习阶段$t=0$接触$\boldsymbol{D}_0$，该阶段包含Stable Diffusion生成的AI图像（正样本，标记为$y=1$）以及从LAION-5B筛选的真实图像（负样本，标记为$y=0$）。随后，模型按时序逐步学习后续S个阶段的数据$\boldsymbol{D}_s$（$s=1, \cdots, S$），每个阶段分别对应一种新型生成模型及其生成作品。

本文数据集的一个显著特点——即大多数AI模型生成图像缺乏对应的公开真实数据。具体而言，当模型处理后续学习阶段$s \geqslant 1$的数据时，若严格遵循持续学习约束（即禁止访问或回放早期阶段数据，尤其是$t=0$阶段的负样本），则模型在这些阶段将仅能接触单一类别数据（即当前阶段的正样本，$y_i^s=1$）。这一约束条件导致CAID问题演变为一种混合了二类学习（基础阶段）与单类学习（后续阶段）的持续学习新范式。这种混合学习模式对传统持续学习方法提出了前所未有的挑战，因为现有大多数方法都依赖于每个学习阶段均具备多类别数据的假设前提。

1.3　评测基准与指标

为规范化CAID问题并全面评估不同方法应对该挑战的能力，本文设计了3种具有代表性的评测基准场景。这些场景主要基于对数据回放，即在学习新任务时访问历史任务样本的约束条件差异，体现了由浅入深的持续学习难度梯度：

1）基准场景1：完全回放。此为约束最宽松的设定。模型在学习新任务数据时，允许同时利用存储缓冲区内缓存的部分历史正样本，以及源自初始训练阶段的LAION-5B负样本进行回放训练。

2）基准场景2：仅负样本回放。此场景模拟了数据访问受限的实际约束，如涉及隐私或版权限制的环境。模型在学习新任务时，仅允许回放来自初始训练阶段的负样本，而严禁回放任何历史正样本。

3）基准场景3：无回放。此为最严格的持续学习条件，代表了对数据存储和隐私保护具有极高要求的应用场景。模型在学习新任务时，完全禁止访问和回放任何来自过去训练阶段的样本。该场景挑战的关键在于这种混合学习模式：模型初始阶段进行二类学习，而增量阶段则因无法回放历史样本，退化为仅能进行单类学习。

为全面评估模型在上述基准场景中的性能表现，本文采用了以下4项综合评价指标：

1）平均检测准确率（average accuracy，AA）。该指标衡量模型对全部5种伪造类型进行二分类判别（即区分真伪）的平均准确度。具体计算方法为：模型完成所有任务学习后，分别测算其在各任务测试

集上的准确率，然后取算术平均值。此指标直观反映了模型的整体检测能力。

2）平均精度均值（mean average precision，mAP）。通过计算模型在各任务测试集上的平均精度（average precision，AP，即 PR 曲线下面积），再对所有任务 AP 值取平均得到。与平均检测准确率相比，mAP 能够更全面地评估模型在不同决策阈值下的检测性能，特别适用于存在类别不平衡的测试场景。

3）平均遗忘度（average forgetting，AF）。作为持续学习领域的特有指标，AF 用于量化模型在学习新知识过程中对已掌握知识的遗忘程度。计算方式为：对每个先前任务，测量其刚学习完成时的准确率与整个学习流程结束后准确率之间的差值，然后对所有历史任务的遗忘程度取平均。AF 值越低（理想情况下接近零或为正值），表明模型遗忘的越少，抵抗灾难性遗忘的能力越强。

4）版权识别准确率（copyright identification accuracy，CA）。这是 CAID 问题下的关键应用指标，评估模型不仅能检测图像是否为 AI 生成，还能精确识别其生成模型来源的能力。计算时将每种生成模型视为独立类别，所有真实图像则归为统一的类别。CA 指标衡量了模型在区分不同 AI 生成源方面的辨识能力，这对于版权追溯、内容监管等实际应用至关重要。

2 解决方案

针对持续 AI 生成图像检测面临的特殊挑战，尤其是增量学习阶段可能出现的负样本（即真实图像）匮乏问题，本节对现有持续学习方法进行了适应性改造，并针对最为严苛的无回放场景提出一种通用改进框架。

2.1 面向允许样本回放场景的方法适配

传统持续学习方法大多针对多类别增量学习设计。为将其有效应用于 CAID 任务，对分类器结构进行了针对性调整。具体而言，保留其原有多类别输出层用于版权溯源（即识别不同生成模型来源），同时根据预测类别归属（属于真实类别还是任何 AI 生成模型）进行二分类判断（是否为生成内容）。给定输入样本 $\boldsymbol{x}$，模型输出各类别的逻辑预测值 $\varphi(\boldsymbol{x})$，最终预测类别为 $\hat{y} = \text{argmax}(\varphi(\boldsymbol{x}))$。若 $\hat{y}$ 对应真实类别，则判定为真实图像；若对应任一 AI 模型类别，则判定为伪造类别，并同时识别其具体生成模型。

CAID 任务在持续学习设置下面临一个独特的挑战。如表 1 所示的任务序列，其增量学习阶段往往只引入单一的新类别，即当前阶段需要学习的一种特定类型的伪造图像（正样本）。这种逐次引入单个正类别的特性，与许多标准持续学习基准不同，并对常用训练方法造成了障碍。具体而言，广泛使用的损失函数，例如交叉熵损失，通常需要训练批次中至少包含两个不同的类别才能有效计算。因此，在 CAID 的增量阶段，如果一个训练批次恰好只包含来自单一新类别的样本，这些损失函数便无法直接应用或可能导致训练不稳定（Hu 等，2021）。

为了克服这一由任务特性带来的固有困难，并确保在不同的数据回放约束下持续学习过程的有效性，本文提出并实施了负样本共享机制。该机制的核心思想非常直接：在所有增量学习阶段进行模型训练时，系统会强制性地、持续地从基础学习阶段（Task 0）存储的真实图像库（即负样本）中进行采样并回放。通过确保这些共享的负样本始终存在于训练数据流中，保证了无论当前增量任务引入何种正样本，每个训练批次都能够包含必要的负类别信息，从而满足标准损失函数的计算前提。

这一机制在两个场景中都发挥了关键作用：在场景 1 中，确保了依赖样本回放的主流方法（如 iCaRL（Rebuffi 等，2017）、Foster（Wang 等，2022a）等）即使在面对单一新类别时也能获得必要的负样本进行稳定训练；而在场景 2 中，由于禁止回放历史正样本，转而适配无（正）样本回放的方法（如 LwF（Li 和 Hoiem，2018）、S-Prompts（Wang 等，2022b）等），此时负样本共享机制同样必不可少，它为这些方法提供了计算损失和有效学习所需的负类别参照。因此，通过负样本共享机制，使得不同类型的持续学习方法均能有效应对这两个具有不同回放约束的基准场景。

2.2 面向无回放设定下方法通用转换框架

对于基准场景 3（无回放）代表了持续学习中最具挑战性的设定，它完全禁止任何形式的历史样本回放。这一极端约束不仅使得所有依赖回放的持续学习方法失效，更严峻的是，研究发现，由于增量学习阶段完全缺乏负样本（真实图像），多数现有的无回放（replay-free）方法亦会遭遇性能显著退化乃至完全失效的困境。

以经典的LwF(learning without forgetting)(Li和Hoiem,2018)和较新的S-Prompts(Wang等,2022b)等代表性无回放方法为例,它们通常为每个新任务增量式地引入新的分类头,并采用独立的损失函数进行优化。然而,在场景3这种仅有单类正样本可用于增量训练的极端条件下,该设计范式会暴露并放大以下3个关键问题。

1)损失函数失效。标准的分类损失函数(如交叉熵)在处理仅包含单一类别数据的训练批次时无法计算或提供有效的更新。虽然存在适用于单类学习的损失函数(Hu等,2021),但这些损失往往不适用于包含正负两类样本的基础学习阶段,导致难以在整个持续学习流程中采用统一有效的训练策略。

2)分类器输出偏置。由于各任务的分类头是独立优化且训练数据分布(尤其是类别数量)差异巨大,不同分类头的输出逻辑值(logits)尺度可能存在显著差异。这会导致模型在进行联合预测时,其决策边界严重偏向于近期学习过的、可能具有更大输出幅度的类别,尤其是在增量阶段仅用单类数据训练时,这种偏差(bias)现象尤为突出(Zhou等,2021)。

3)特征表示偏置现象。当仅使用单类别的增量数据对模型(或其部分参数)进行微调时,模型极易对当前任务数据产生过拟合。这会导致学习到的特征表示过度偏向当前类别,损害其对其他类别(包括已学习类别和未见类别)的泛化能力,并加速灾难性遗忘。传统方法中,如何精细地确定冻结哪些层、微调哪些层,或施加恰当的权重正则化,往往过程复杂且效果难以保证。

为系统性地解决上述在无回放场景下困扰现有无回放方法的共性问题,本文提出了一种通用转换框架,如图2所示。该框架旨在通过整合3种关键技术组件,对现有的无回放方法进行适应性改造,以修复其性能并提升其在极端约束下的鲁棒性。

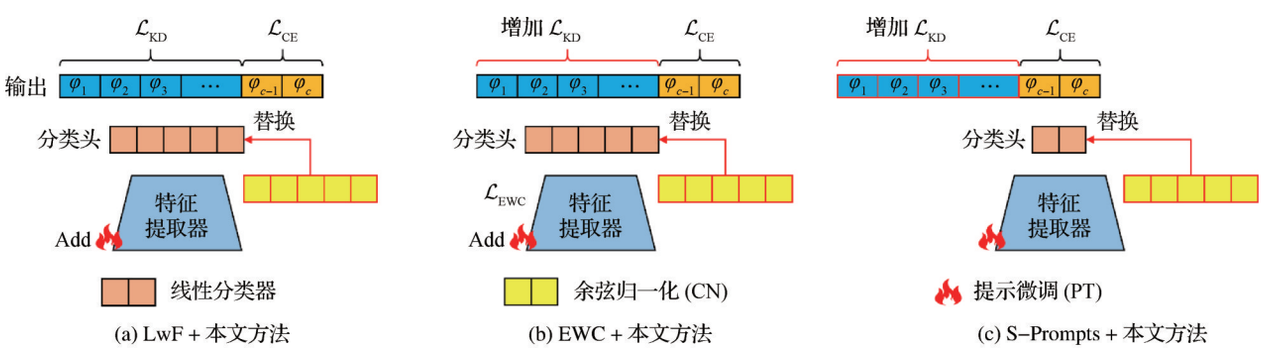

图2 本文通用转换框架示意图

Fig. 2 Overview of the proposed method((a)LwF+ours;(b)EWC+ours;(c)S-Prompts+ours)

1)知识蒸馏技术(knowledge distillation,KD)。借鉴LwF(Li和Hoiem,2018)的核心思想,引入知识蒸馏损失函数,如式(1),强制当前学习模型模拟上一阶段模型的输出响应特征。这为模型提供了有效的更新约束。

$$\mathcal{L}_{KD}=-\sum_{i=1}^{C}\frac{\exp(v_i/\tau)}{\sum_{j=1}^{C}\exp(v_j/\tau)}\log\left(\frac{\exp(z_i/\tau)}{\sum_{j=1}^{C}\exp(z_j/\tau)}\right) \tag{1}$$

式中,v_i是图像$\boldsymbol{x}$的旧阶段模型预测,z_i是当前模型预测,τ是温度系数,C是类别数。

2)余弦归一化机制(cosine normalization,CN)。采用余弦归一化分类器(Hou等,2019)替代标准线性分类器。通过对特征向量和分类器权重进行L_2范数归一化处理,将输出逻辑值有效约束在[-1, 1]区间内,从而校准不同分类头的输出幅度,有效缓解分类器输出偏置问题,计算式为

$$\varphi(\boldsymbol{x})=\left(\frac{\boldsymbol{W}}{\|\boldsymbol{W}\|_2}\right)^{\mathrm{T}}\left(\frac{f(\boldsymbol{x})}{\|f(\boldsymbol{x})\|_2}\right) \tag{2}$$

式中,$f(\boldsymbol{x})$是由模型骨干网络(即冻结的ViT)从输入图像$\boldsymbol{x}$中提取的高维特征向量(或称特征嵌入);$\boldsymbol{W}$是分类层的权重矩阵,由每个类别i对应的权重向量$\boldsymbol{W}_i$拼接而成;$\varphi(\boldsymbol{x})$是模型最终输出的逻辑值(logits)向量。

在持续学习框架下,权重矩阵$\boldsymbol{W}$是动态管理的:当学习新任务时,会为新类别增量式地添加并初始化新的权重向量到矩阵$\boldsymbol{W}$中。

3)提示调整策略(prompt tuning,PT)。借鉴最

新研究成果(Wang等,2022c,d),采用提示调整技术替代传统全模型微调方法。具体实现为冻结预训练模型(如ViT)的主体参数,仅在输入端添加少量可学习的"提示"参数进行训练。对于一个输入图像$\boldsymbol{x}$,首先被ViT的嵌入层处理成一个补丁(patch)嵌入序列$\boldsymbol{E}_{\text{patch}}=[\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_N]$。提示词即额外定义一组可学习的、与任务相关的参数$\boldsymbol{P}_{\text{prompt}}=[\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_L]$,其中$L$是提示的长度,且每个提示词参数$\boldsymbol{p}_i$的维度与嵌入$\boldsymbol{e}_j$的维度相同。在将图像特征送入Transformer编码器之前,这组提示参数会与图像的补丁嵌入序列在序列维度上进行拼接,形成一个扩展后的新输入序列$\boldsymbol{Z}_0$,即

$$\boldsymbol{Z}_0=[\boldsymbol{P}_{\text{prompt}};\boldsymbol{E}_{\text{patch}}] \tag{3}$$

式中,[;]表示序列拼接操作。

在后续的持续学习阶段,模型的主干网络参数保持完全冻结,只有这部分新增的、数量极少的提示参数$\boldsymbol{P}_{\text{prompt}}$会通过梯度反向传播进行更新。这种微调方式限制了模型在学习新知识时对原有参数空间的改动,从而有效抑制了对旧任务特征表示的破坏,显著缓解了特征表示偏置和灾难性遗忘问题。

如图2所示,将该通用框架应用于多种典型无回放方法,包括LwF、EWC(elastic weight consolidation)和S-Prompts等。具体转换流程包括:首先,将模型主干网络替换为基于ViT的提示微调结构;其次,将原有分类头更换为余弦归一化分类器;最后,在总体损失函数中引入知识蒸馏损失项$\mathcal{L}_{\text{KD}}$。

3 实验与结果分析

本节旨在通过一系列实验,评估目前增量学习方法在所构建的数据集上的性能。详细介绍实验设置,并对3种不同持续学习场景下的结果进行深入分析,同时通过消融实验验证所提关键技术的有效性,并通过可视化手段进一步揭示模型行为与数据特性。

3.1 对比方法

本文实验评估了多种先进的持续学习方法在持续伪造检测任务上的性能。对于基准场景1,选择经典基于回放的持续学习方法,如iCaRL(Rebuffi等,2017)、BiC(Wu等,2019)、Gradient Episodic Memory(GEM)(Lopez-Paz和Ranzato,2017)、Coil(Zhou等,2021)以及Foster(Wang等,2022a)。对于基准场景2和场景3,选择经典无回放的方法,如LwF(Li和Hoiem,2018)、EWC(Kirkpatrick等,2017)和S-Prompts(Wang等,2022b)。上述所有方法均针对本任务使用所提出的策略改造。

3.2 实现细节

本文所有实验基于PyTorch框架,使用NVIDIA RTX 3090 GPU进行。除非特别说明,模型骨干网络采用在ImageNet上预训练的ViT-B/16。各对比方法遵循其原始实现的核心超参数设定。对于ViT骨干网络,统一使用学习率0.001,训练10个epoch。对于应用了转换框架的方法采用学习率0.1,训练30个epoch,以适应提示微调。优化器为带动量的SGD(stochastic gradient descent)(momentum=0.9),配合余弦退火学习率调度策略。知识蒸馏损失同LwF,批处理大小设为128。数据增强仅使用标准操作:缩放至224×224像素、随机水平翻转和随机裁剪。

3.3 结果分析

表2展示了所提出的3个基准实验场景及联合训练上界的结果,清晰地揭示了不同持续学习方法在应对CAID挑战时的性能差异。表2不仅包含了各项方法的mAP(平均精度均值)和CA(概念遗忘率)作为综合指标,还详细列出了它们在检测Stable Diffusion、DALL-E 2、Imagen、Midjourney和Parti这5种不同图像生成模型时各自的AA(平均准确率)表现。

在约束最为宽松的基准场景1,即允许回放历史伪造样本和真实样本时,基于ViT骨干网络的Foster方法展现出最佳的综合性能。无论是在1 000样本还是500样本的缓冲区设置下,Foster均在AA和CA上领先,例如在1 000样本缓冲时分别达到75.57%和60.27%。这显示其特征增强和压缩策略与样本回放结合能有效缓解遗忘。相较之下,Coil(Zhou等,2021)等其他方法表现略逊一筹。缓冲区大小的减少会导致所有方法性能下降,但Foster的相对优势得以保持。

对于基准场景2,挑战随之增大,因为模型仅能回放基础阶段的真实负样本。在此条件下,基于ViT的S-Prompts方法脱颖而出,在1 000样本缓冲下取得了最高的AA(67.90%)。尤为突出的是其极低的平均遗忘度(AF:-2.77%),远优于LwF(-18.08%)和EWC(-16.00%),这归功于其提示学

习机制能有效隔离参数，减少对旧知识的干扰。缓冲区减至500时，S-Prompts性能虽然下降但仍保持领先。对比基准场景1，基准场景2整体性能的显著降低清晰地揭示了回放历史正样本对于维持模型性能的关键作用。ViT骨干和负样本共享策略在此更困难的场景下依然是取得较好结果的重要因素。

基准场景3构成了最严苛的考验，完全禁止任何样本回放。在此极端条件下，未经改进的原始无

表2　增量学习方法在本数据集上的性能

Table 2　Performance comparison of incremental learning methods on the proposed dataset

/%

设定	存储数量	方法	StableDiff	DALL-E 2	Imagen	Midjourney	Parti	AA	AF	mAP	CA
场景1	1 000	iCaRL	70.32	73.02	78.41	70.29	68.42	72.09	−6.80	89.46	56.05
		BiC	72.54	78.04	79.55	72.96	70.32	74.68	**−5.48**	90.37	53.09
		GEM	66.35	70.11	84.09	82.75	67.11	74.08	−8.99	85.30	57.43
		Coil	70.57	78.57	86.36	75.95	64.47	75.18	−5.81	90.08	56.97
		Foster	72.40	76.46	81.82	73.48	73.68	**75.57**	−5.73	**91.15**	**60.27**
	500	iCaRL	66.30	69.05	76.14	68.77	65.79	69.21	−7.78	88.83	**55.02**
		BiC	69.75	68.78	73.86	70.72	71.05	70.83	−7.47	88.85	43.85
		GEM	58.79	60.05	72.73	62.74	60.53	62.97	−16.85	52.43	24.76
		Coil	72.73	75.40	75.23	65.43	65.79	70.92	**−5.80**	**89.69**	46.18
		Foster	68.05	68.52	81.82	68.16	69.74	**71.26**	−7.18	89.28	48.05
场景2	1 000	LwF	52.81	55.82	64.77	51.89	67.11	58.48	−18.08	87.14	62.94
		EWC	60.78	62.17	77.27	64.35	59.21	64.76	−16.00	78.18	38.09
		S-Prompts	57.61	67.72	76.14	70.91	67.11	67.90	**−2.77**	75.63	48.76
		LwF + 本文方法	54.58	50.79	61.36	73.67	57.89	59.66	−16.67	89.99	59.66
		EWC + 本文方法	67.67	60.05	79.55	66.44	76.32	**70.01**	−18.62	78.16	54.96
		S-Prompts + 本文方法	57.28	61.90	`77.27	71.01	69.74	67.46	−3.14	**91.02**	**68.99**
	500	LwF	50.52	55.03	54.55	59.65	61.84	56.32	−25.19	72.32	52.54
		EWC	57.14	65.87	69.32	67.06	53.95	62.67	−19.74	67.14	36.27
		S-Prompts	55.67	64.29	72.73	61.55	67.11	64.27	−5.80	73.67	37.53
		LwF + 本文方法	67.25	53.17	67.05	85.36	63.16	67.20	−8.74	**86.36**	58.83
		EWC + 本文方法	75.66	63.23	72.95	74.90	61.58	**69.66**	−10.96	84.90	54.28
		S-Prompts + 本文方法	58.22	58.73	75.00	67.40	69.74	65.82	**−4.33**	85.98	**66.75**
场景3	0	LwF	51.32	51.06	51.14	51.19	51.32	51.21	−11.70	51.24	28.19
		EWC	51.01	50.06	51.14	51.19	49.32	50.54	−11.65	63.77	23.54
		S-Prompts	49.72	48.94	48.86	48.81	48.68	49.01	−11.04	60.32	39.20
		LwF + 本文方法	84.74	50.79	60.23	73.91	56.58	65.25	−0.66	**87.70**	56.99
		EWC + 本文方法	87.57	49.21	70.45	63.83	61.84	**66.58**	−0.43	83.32	51.39
		S-Prompts + 本文方法	95.65	48.41	57.95	69.49	52.63	64.83	**−0.21**	87.18	**68.52**
上界	无	ViT-FT	96.08	88.36	89.77	95.44	86.84	91.30	无	99.20	94.35
		ViT-PT	93.48	81.22	77.27	93.39	86.84	86.44	无	98.92	90.15

注：加粗字体表示最佳平均性能。

回放方法，如 LwF、EWC 和 S-Prompts，几乎完全失效。它们的平均检测准确率（AA）均在 50% 左右（相当于随机猜测），mAP 和 CA 指标极低，且遗忘非常严重（AF 约 −11%）。这有力地印证了本文之前的判断：在增量阶段仅有单类正样本且无任何回放可用时，这些方法会因损失函数失效、分类头输出偏差和特征表示漂移等问题而崩溃。然而，应用了本文提出的通用转换框架（标记为+本文方法）后，局面发生了根本性转变。例如，改进后的 LwF 的 AA 从 51.21% 提升至 65.25%，EWC 从 50.54% 提升至 66.58%，S-Prompts 从 49.01% 提升至 64.83%，相对提升幅度均在 20%～30% 以上。更令人瞩目的是，这些改进后的方法在平均遗忘度（AF）上表现卓越，普遍接近于零（如，LwF 为 −0.66%，EWC 为 −0.43%，S-Prompts 为 −0.21%），抗遗忘能力远超上面两个场景下的方法。这充分证明，本文结合了提示微调（PT）、余弦归一化（CN）和知识蒸馏（KD）的框架能够有效稳定特征表示、校准分类器输出并利用历史知识，从而在零回放的极端约束下成功实现了高效的持续学习。甚至部分改进后的方法在 AA 等指标上反超了某些基准场景 2 方法，显示了该框架强大的性能恢复和抗遗忘能力。

最后，将所有持续学习结果与联合训练上界进行比较，后者代表了理想数据条件下的最优性能。基于 ViT 的完全微调（ViT-FT）达到了最高的 AA（91.30%）、mAP（99.20%）和 CA（94.35%），显著优于其他设置。ViT 的提示调优（ViT-PT）表现也相当出色，证明了其高效性。然而，持续学习方法与联合训练上界之间存在的显著性能差距，清晰地表明在数据流式到达且资源受限的持续学习场景下，实现高效学习并完全避免知识遗忘仍然是一项充满挑战的研究课题。

3.4 消融实验

为了系统性地评估提出的通用转换框架中各个核心组件：知识蒸馏（KD）、余弦归一化（CN）和提示微调（PT）的独立贡献及其组合效果，针对最严苛的无回放场景进行了一系列消融实验。实验结果如表 3 所示。

实验首先考察了基线情况，即不采用任何 KD、CN 或 PT 组件的配置，仅仅进行连续微调。其性能表现极其不佳，平均检测准确率（AA）仅为 50.01%，接近随机猜测水平，同时伴随着严重的灾难性遗忘（平均遗忘度 AF 为 −14.53%）。这一结果清晰地验证了本文之前的判断：在无回放的极端条件下，标准方法难以有效学习并会遭遇性能崩溃。

表 3 所提框架主要模块消融实验

Table 3 Ablation experiments on the main modules of the proposed framework

/%

KD	CN	PT	AA	AF
−	−	−	50.01	−14.53
√	−	−	51.21	−11.70
−	√	−	51.20	−12.20
−	−	√	51.74	−11.01
√	√	−	59.43	−5.17
−	√	√	58.20	−5.53
√	−	√	57.83	−9.12
√	√	√	65.25	−0.66

注：“√”表示采用，“−”表示不采用。

随后，评估了单独引入每个组件的效果。无论是单独引入知识蒸馏（KD）、余弦归一化（CN）还是提示调优（PT），相对于基线虽然带来了微小的性能提升（AA 分别提升至 51.21%、51.20% 和 51.74%），但改善幅度有限，且遗忘问题（AF 仍在 −11%～−12% 区间）依然非常严重。这表明，尽管每个组件可能针对性地缓解了无回放场景下的某个特定问题（如 KD 提供学习信号，CN 校准输出，PT 减少过拟合），但单一策略的力量不足以应对该场景面临的多重、交织的挑战。

进一步地，考察了两两组合这些组件的影响。将 KD 与 CN 结合（AA：59.43%，AF：−5.17%）、CN 与 PT 结合（AA：58.20%，AF：−5.53%）或 KD 与 PT 结合（AA：57.83%，AF：−9.12%），均能观察到比单组件更显著的性能改善，尤其是在缓解遗忘方面（AF 显著降低至 −5%～−9% 的区间）。这有力地说明了这些技术组件之间存在积极的协同效应，它们的组合能够更有效地应对场景 3 的复杂困境。

最终，当同时集成所有 3 个核心组件（KD + CN + PT）时，得到了最佳的性能表现。该完整框架配置取得了最高的平均检测准确率（AA：65.25%），相较于基线提升超过 15%。更为关键的是，它几乎完全消除了灾难性遗忘，实现了接近于零的平均遗忘度（AF：−0.66%）。这一结果不仅在准确率上远超所

有其他配置，在抗遗忘能力上更是实现了质的飞跃。

综上所述，消融实验（表3）的结果表明了提出的通用转换框架中知识蒸馏、余弦归一化和提示调优3个组件的必要性和互补性。

3.5 可视化实验分析

为可视化验证本文所提转换框架对性能的提升，对模型输出的特征进行了可视化。如图3所示，应用所提方法后，各个AI模型类别的特征子空间在经历全部5个学习阶段后表现出明显更小的位移。这直观表明本文方法能够减少特征表示在持续学习过程中的漂移，从而有助于维持模型的稳定性和性能。

此外，为了更深入地理解数据集中包含的AI图像的特性及其检测难度，将本文方法与传统的Deepfake图像进行了对比分析。利用离散傅里叶变换（discrete Fourier transform，DFT）计算并可视化了各类模型生成图像的平均频率频谱，结果如图4所示。分析结果清晰地揭示了一个重要现象：数据集收集的AI图像和真实的图像在频率频谱上表现出高度的相似性。然而，作为对比参照的Deepfake图像则在频谱图中呈现出显著且易于观察的周期性模式。这一对比有力地表明，所提出的AI检测任务所面临的巨大挑战性。

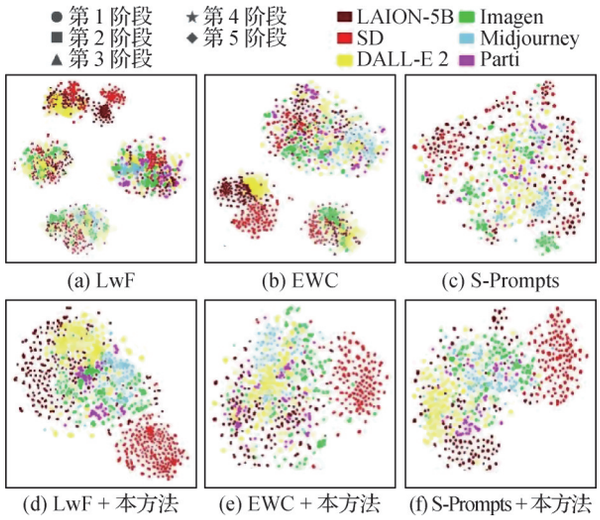

图3 不同方法在增量后特征空间t-SNE可视化

Fig. 3 t-SNE visualization of incremental feature space for different methods ((a) LwF; (b) EWC; (c) S-Prompts; (d) LwF+ours; (e) EWC+ours; (f) S-Prompts+ours)

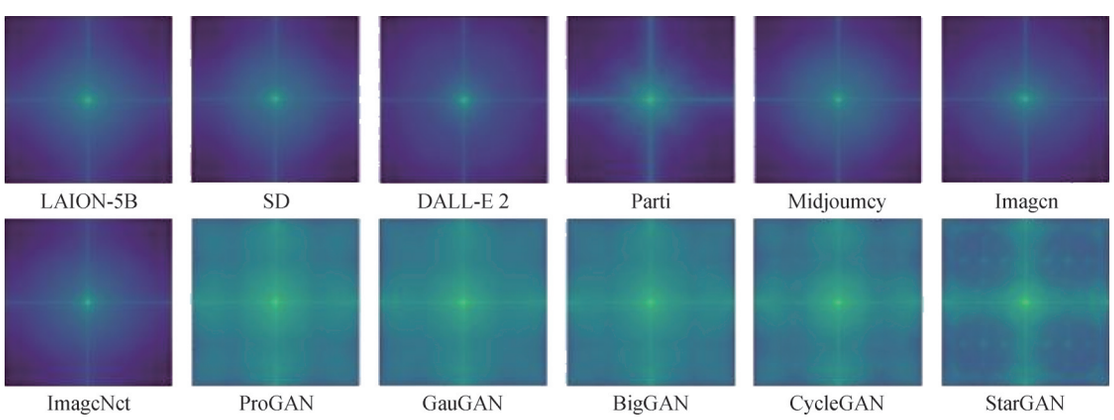

图4 不同生成模型的生成图像平均频率频谱可视化

Fig. 4 Visualization of average frequency spectra for different image generators

4 结 论

AI生成技术的迅速发展与广泛传播，使得对其进行有效检测与精确的来源识别（版权溯源）成为一项新兴且重要的研究议题。本文针对这一挑战，系统性地开展了研究工作，主要贡献可归纳为4个方面：1）构建并发布了专门面向此任务的检测数据库；2）定义持续AI检测任务，并设计了3个具有代表性的、难度递增的评测基准场景，这些场景依据对数据回放的不同约束进行划分；3）针对不同场景下的挑战，提出了相应的适配策略与改进框架，包括适配现有持续学习方法的关键技术——“负样本共享机制”，以及专为应对最严苛无回放场景而设计的“通用转换框架”；4）进行了全面的实验评估。

面向先进生成模型持续检测是一个充满挑战且具有广阔前景的研究领域。本文通过系统地构建、定义与评估，期望能为该领域的后续探索奠定坚实基础，并激发更多创新性的研究方向。展望未来，若干有价值的研究路径值得进一步探索，例如：可以扩

展至更多的生成模型进行持续学习，从公共资源持续收集和扩充更新的伪造模型，以及深入探索如何有效利用与生成过程相关联的文本提示信息，以增强检测与溯源的性能。

参考文献（References）

Aljundi R, Babiloni F, Elhoseiny M, Rohrbach M and Tuytelaars T. 2018. Memory aware synapses: learning what (not) to forget//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 144-161 [DOI: 10.1007/978-3-030-01219-9_9]

Cetinic E and She J. 2022. Understanding and creating art with AI: review and outlook. ACM Transactions on Multimedia Computing, Communications, and Applications, 18(2): #66 [DOI: 10.1145/3475799]

Feng C B, Liu C X, Wang Y Y and Zhou Q D. 2024. Face forgery detection with image patch comparison and residual map estimation. Journal of Image and Graphics, 29(2): 457-467 (冯才博, 刘春晓, 王昱烨, 周其当. 2024. 结合图像块比较与残差图估计的人脸伪造检测. 中国图象图形学报, 29(2): 457-467) [DOI: 10.11834/jig.230149]

Holz D. 2022. Midjourney: exploring new mediums of thought and expanding the imaginative powers of the human species [EB/OL]. [2025-04-10]. https://halotool.com/tool/midjourney

Hou S H, Pan X Y, Loy C C, Wang Z L and Lin D H. 2019. Learning a unified classifier incrementally via rebalancing//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 831-839 [DOI: 10.1109/CVPR.2019.00092]

Hu W P, Qin Q, Wang M Y, Ma J W and Liu B. 2021. Continual learning by using information of each class holistically//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual: AAAI Press: 7797-7805 [DOI: 10.1609/aaai.v35i9.16952]

Kim M, Tariq S and Woo S S. 2021. CoReD: generalizing fake media detection with continual representation using distillation//Proceedings of the 29th ACM International Conference on Multimedia. New York, USA: Association for Computing Machinery: 337-346 [DOI: 10.1145/3474085.3475535]

Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu A A, Milan K, Quan J, Ramalho T, Grabska-Barwinska A, Hassabis D, Clopath C, Kumaran D and Hadsell R. 2017. Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences of the United States of America, 114(13): 3521-3526 [DOI: 10.1073/pnas.1611835114]

Li C Q, Huang Z W, Paudel D P, Wang Y B, Shahbazi M, Hong X P and Van Gool L. 2023. A continual deepfake detection benchmark: dataset, methods, and essentials//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE: 1339-1349 [DOI: 10.1109/WACV56688.2023.00139]

Li Y Z, Chang M C and Lyu S W. 2018. In Ictu Oculi: exposing AI created fake videos by detecting eye blinking//Proceedings of 2018 IEEE International Workshop on Information Forensics and Security (WIFS). Hong Kong, China: IEEE: 1-7 [DOI: 10.1109/WIFS.2018.8630787]

Li Y Z, Yang X, Sun P, Qi H G and Lyu S W. 2020. Celeb-DF: a large-scale challenging dataset for DeepFake forensics//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 3204-3213 [DOI: 10.1109/CVPR42600.2020.00327]

Li Z Z and Hoiem D. 2018. Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(12): 2935-2947 [DOI: 10.1109/TPAMI.2017.2773081]

Lopez-Paz D and Ranzato M A. 2017. Gradient episodic memory for continual learning//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6470-6479

Marra F, Saltori C, Boato G and Verdoliva L. 2019. Incremental learning for the detection and classification of GAN-generated images//Proceedings of 2019 IEEE International Workshop on Information Forensics and Security (WIFS). Delft, Netherlands: IEEE: 1-6 [DOI: 10.1109/WIFS47025.2019.9035099]

Ramesh A, Dhariwal P, Nichol A, Chu C and Chen M. 2022. Hierarchical text-conditional image generation with CLIP latents [EB/OL]. [2025-04-10].https://arxiv.org/pdf/2204.06125.pdf

Rebuffi S A, Kolesnikov A, Sperl G and Lampert C H. 2017. iCaRL: incremental classifier and representation learning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 5533-5542 [DOI: 10.1109/CVPR.2017.587]

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10674-10685 [DOI: 10.1109/CVPR52688.2022.01042]

Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J and Niessner M. 2019. FaceForensics++ : learning to detect manipulated facial images//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 1-11 [DOI: 10.1109/ICCV.2019.00009]

Saharia C, Chan W, Saxena S, Lit L, Whang J, Denton E, Ghasemipour S K S, Karagol Ayan B, Mahdavi S S, Gontijo-Lopes R, Salimans T, Ho J, Fleet D J and Norouzi M. 2022. Photorealistic text-to-image diffusion models with deep language understanding//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates

Inc.: 2643

Schuhmann C, Beaumont R, Vencu R, Gordon C, Wightman R, Cherti M, Coombes T, Katta A, Mullis C, Wortsman M, Schramowski P, Kundurthy S, Crowson K, Schmidt L, Kaczmarczyk R and Jitsev J. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1833

Wang F Y, Zhou D W, Ye H J and Zhan D C. 2022a. FOSTER: feature boosting and compression for class-incremental learning//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 398-414 [DOI: 10.1007/978-3-031-19806-9_23]

Wang S Y, Feng C B, Liu C X and Jin Y S. 2025. Multivariate and soft blending sample-driven image-text alignment for deepfake detection. Journal of Image and Graphics, 30(5): 1334-1345 (王诗雨, 冯才博, 刘春晓, 金逸胜. 2025. 多元软混合样本驱动的图文对齐人脸伪造检测. 中国图象图形学报, 30(5): 1334-1345) [DOI: 10.11834/jig.240252]

Wang S Y, Wang O, Zhang R, Owens A and Efros A A. 2020. CNN-generated images are surprisingly easy to spot... for now//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 8692-8701 [DOI: 10.1109/CVPR42600.2020.00872]

Wang Y B, Huang Z W and Hong X P. 2022b. S-prompts learning with pre-trained transformers: an Occam's razor for domain incremental learning//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 411

Wang Y B, Ma Z H, Huang Z W, Wang Y W, Su Z and Hong X P. 2023. Isolation and impartial aggregation: a paradigm of incremental learning without interference//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press: 10209-10217 [DOI: 10.1609/aaai.v37i8.26216]

Wang Z F, Zhang Z Z, Ebrahimi S, Sun R X, Zhang H, Lee C Y, Ren X Q, Su G L, Perot V, Dy J and Pfister T. 2022c. DualPrompt: complementary prompting for rehearsal-free continual learning//Proceedings of the 17th European Conference on Computer Vision—ECCV 2022. Tel Aviv, Israel: Springer: 631-648 [DOI: 10.1007/978-3-031-19809-0_36]

Wang Z F, Zhang Z Z, Lee C Y, Zhang H, Sun R X, Ren X Q, Su G L, Perot V, Dy J and Pfister T. 2022d. Learning to prompt for continual learning//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 139-149 [DOI: 10.1109/CVPR52688.2022.00024]

Wu Y, Chen Y P, Wang L J, Ye Y C, Liu Z C, Guo Y D and Fu Y. 2019. Large scale incremental learning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 374-382 [DOI: 10.1109/CVPR.2019.00046]

Xue Z Y, Liu Q T, Zhou K N, Wang W and Jiang X H. 2025. Physical properties synthetic dataset and forgery detection method for shadow-consistency analysis. Journal of Image and Graphics: 1-20 (薛子育, 刘庆同, 周康能, 王伟, 姜秀华. 2025. 面向光影一致性分析的物理属性合成数据集及伪造检测方法. 中国图象图形学报, 2025: 1-20) [DOI: 10.11834/jig.240705]

Yang T Y, Huang Z Y, Cao J, Li L and Li X R. 2022. Deepfake network architecture attribution//Proceedings of the 36th AAAI Conference on Artificial Intelligence. AAAI Press: 4662-4670 [DOI: 10.1609/aaai.v36i4.20391]

Yu J H, Xu Y Z, Koh J Y, Luong T, Baid G, Wang Z R, Vasudevan V, Ku A, Yang Y F, Ayan B K, Hutchinson B, Han W, Parekh Z, Li X, Zhang H, Baldridge J and Wu Y H. 2022. Scaling autoregressive models for content-rich text-to-image generation. Transactions on Machine Learning Research

Zhang J, Xu P, Liu W J, Guo X X, and Sun F. 2025. Negative instance generation for cross-domain facial forgery detection. Journal of Image and Graphics, 30(2): 421-434 (张晶, 许盼, 刘文君, 郭晓萱, 孙芳. 2025. 多样性负实例生成的跨域人脸伪造检测. 中国图象图形学报, 30(2): 421-434) [DOI: 10.11834/jig.240160]

Zhou D W, Ye H J and Zhan D C. 2021. Co-transport for class-incremental learning//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event, China: Association for Computing Machinery: 1645-1654 [DOI: 10.1145/3474085.3475306]

Zhuo W Q, Li D Z, Wang W and Dong J. 2023. Data-free model compression for light-weight DeepFake detection. Journal of Image and Graphics, 28(3): 820-835 (卓文琦, 李东泽, 王伟, 董晶. 2023. 面向轻量级深度伪造检测的无数据模型压缩. 中国图象图形学报, 28(3): 820-835) [DOI: 10.11834/jig.220559]

作者简介

王亚斌,男,博士研究生,主要研究方向为增量学习和伪造检测。E-mail:iamwangyabin@stu.xjtu.edu.cn

洪晓鹏,通信作者,男,教授,主要研究方向为计算机视觉、模式识别和深度学习。E-mail:hongxiaopeng@ieee.org

黄智武,男,讲师,主要研究方向为模式识别和图像生成。E-mail:zhiwu.huang@soton.ac.uk