中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-47

论文引用格式:Han Junwei, Qian Xuelin, Xu Chang, Wang Haoyan, Zhang Dingwen. Green Visual AI: A Survey of Energy-efficient Techniques for Data and Models in Visual Intelligence[J/OL]. Journal of Image and Graphics, XXXX:1-47. DOI: 10.11834/jig.260181.(韩军伟, 钱学林, 许畅, 王浩研, 张鼎文. 绿色视觉AI:面向数据与模型的视觉智能节能化综述[J/OL]. 中国图象图形学报,XXXX:1-47. DOI: 10.11834/jig.260181.)
[DOI:10.11834/jig.260181]

绿色视觉AI:面向数据与模型的视觉智能节能化综述

韩军伟[1,2],钱学林[1],许畅[1],王浩研[1],张鼎文[1]

1. 脑与人工智能实验室,西北工业大学自动化学院,西安　710129; 2. 重庆邮电大学人工智能学院,重庆　400065

摘　要:在"双碳"战略深入实施与视觉智能(artificial intelligence,AI)产业快速发展的双重背景下,推动视觉智能技术的绿色化发展已成为实现经济社会可持续发展的重要路径。我国《"十四五"数字经济发展规划》明确提出推动算力基础设施绿色低碳发展。近年来,以深度学习为代表的视觉智能技术性能的飞跃很大程度上得益于模型规模的持续扩张与训练数据的海量增长,但由此引发的数据采集标注成本高、模型训练推理能耗大等问题,也对智能产业低碳转型构成现实挑战。在此背景下,绿色视觉AI作为兼顾技术性能与可持续发展的研究范式受到广泛关注,其核心目标是在保障模型任务性能的前提下,降低视觉智能部署前后的数据、算力、人力等各类成本,实现技术性能与能耗效益的协同。针对这一挑战,本文深入探讨面向绿色视觉AI的视觉智能技术节能化技术,从数据采集、数据标注、模型推理与模型迭代四个核心环节出发,介绍各环节中的节能策略与优化思路,梳理技术方案与发展现状,探究当前面临的主要挑战与未来研究方向,为视觉智能技术的绿色化、可持续化发展提供理论支撑与实践框架。

关键词:绿色视觉AI;数据节能;标注节能;推理节能;迭代节能

Green Visual AI: A Survey of Energy-efficient Techniques for Data and Models in Visual Intelligence

Han Junwei[1,2], Qian Xuelin[1], Xu Chang[1], Wang Haoyan[1], Zhang Dingwen[1]

1. Brain and Artificial Intelligence Laboratory, School of Automation, Northwestern Polytechnical University, Xi'an 710129; 2. School of Artificial Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065

Abstract: Against the profound backdrop of the advancing global "Dual Carbon" strategy and the rapid proliferation of the visual Artificial Intelligence (AI) industry, promoting the green transition of visual AI technologies has emerged as a crucial pathway toward sustainable socioeconomic development. National directives, such as China's "14th Five-Year Plan for Digital Economy Development" explicitly emphasize the necessity of advancing green and low-carbon computing infrastructures. In recent years, the revolutionary leaps in visual AI performance have been fundamentally driven by the "Scaling Law," which dictates that model capability scales in direct proportion to continuous expansions in model parameters and the exponential growth of training data. While this brute-force trajectory has unlocked unprecedented capabilities in complex perception tasks, it has simultaneously triggered severe resource bottlenecks. The prohibitive financial and tempo-

收稿日期:2026-04-08;**修回日期**:2026-05-24

*通信作者:钱学林,男,副教授,主要研究方向为计算机视觉。 E-mail: xlqian@nwpu.edu.cn;张鼎文,男,教授,主要研究方向为计算机视觉、三维视觉、医疗影像分析。 E-mail: zdw2006yyy@nwpu.edu.cn

基金项目:国家重点研发计划(2025YFF0514700);国家自然科学基金项目(62406252,62293543,62322605)

Supported by: National Key R&D Program of China (2025YFF0514700); National Natural Science Foundation of China (62406252, 62293543, 62322605)

ral costs associated with massive physical data collection and fine-grained manual annotation, coupled with the exorbitant energy consumption required for training and deploying massive architectures, pose stark challenges to the low-carbon transformation of the AI industry. In response to this impending crisis, "Green Visual AI" has garnered widespread attention as a pivotal research paradigm. The core objective of this study is to systematically review energy-efficient methodologies that harmonize technical performance with long-term ecological sustainability. This paper aims to provide a comprehensive theoretical foundation and practical framework by exploring optimization strategies that minimize data, computational, and human resource requirements across the entire lifecycle of visual AI systems. This review adopts a comprehensive survey methodology to systematically deconstruct the entire lifecycle of visual intelligence models. We construct a structured taxonomy organized around four core, interrelated stages: Data Collection, Data Annotation, Model Inference, and Model Iteration. To provide a holistic view of the field, we categorize and analyze the overarching strategies within each stage: In the Data Collection phase, the survey reviews literature aimed at circumventing costly physical data acquisition. We analyze data synthesis paradigms and data transfer mechanisms that reuse existing knowledge for new environments. In the Data Annotation phase, we evaluate methodologies designed to eliminate the reliance on exhaustive human-in-the-loop labeling. This encompasses weakly supervised learning and self-supervised learning frameworks. For Model Inference, the methodology categorizes hardware and algorithmic interventions into model lightweighting and inference acceleration. The review synthesizes approaches like knowledge distillation, efficient attention mechanisms, linear sequence architectures, and dynamic inference strategies. Finally, in the Model Iteration phase, we examine frameworks that mitigate the massive carbon footprint of repetitive retraining. The literature is organized into continual learning strategies that prevent catastrophic forgetting and parameter-efficient fine-tuning methods that adapt models with minimal parameter updates. The synthesis of current technological frameworks reveals a profound paradigm shift across all evaluated dimensions of the visual AI lifecycle, transitioning from resource-heavy processes to highly optimized, sustainable operations. Energy-Efficient Data Collection: The literature indicates a decisive shift from physical data collection toward virtual synthesis and cross-domain transfer. Generative Adversarial Networks, Diffusion Models, and Large Language Models are now highly capable of synthesizing high-fidelity, varied data. Furthermore, domain adaptation and open-vocabulary learning techniques enable models to effectively reuse source-domain knowledge, facilitating robust deployment in unknown target environments with minimal to zero new data acquisition. Annotation-Efficient Paradigms: To alleviate the immense human capital required for pixel-level annotations, the field is rapidly adopting self-driven learning signals. Weakly supervised methods successfully derive supervisory cues from coarse, image-level labels or pseudo-labels. More prominently, self-supervised strategies, particularly contrastive learning, Masked Image Modeling, and cross-modal alignment, extract intrinsic structural representations directly from massive unlabeled datasets. These methods effectively bypass the manual annotation bottleneck while yielding powerful, generalized foundation models. Low-Carbon Model Inference: Addressing the high-frequency energy costs of model deployment, current network lightweighting techniques exhibit remarkable efficacy. Knowledge distillation successfully transfers representational power from massive teacher models to compact student networks, shifting the computational burden away from the deployment phase. To break the quadratic complexity bottleneck of standard Transformers, researchers have developed linear attention mechanisms and linear sequence architectures. On the acceleration front, techniques like single-step sampling for diffusion models, Key-Value cache optimization, and dynamic routing architectures ensure models execute complex tasks with minimal computational overhead. Sustainable Model Iteration: In dynamic environments, retraining massive models from scratch for every concept drift is computationally prohibitive. The survey highlights continual learning strategies—including parameter regularization, data replay, and architectural increments—which successfully prevent catastrophic forgetting, allowing models to accumulate new knowledge continuously. Concurrently, PEFT methods, notably Adapters, Low-Rank Adaptation, and Prompt Tuning, permit adaptation to novel downstream tasks by updating only a minuscule fraction of parameters while freezing the pre-trained backbone, drastically lowering the barrier for sustained model evolution. The transition toward Green Visual AI represents a fundamental evolution of artificial intelligence from a resource-intensive discipline to a sustainable, eco-friendly ecosystem. This review demonstrates that achieving a synergistic balance between technological performance and energy efficiency is feasible through targeted, full-lifecycle optimizations. However, significant challenges remain. Future research trajectories may emphasize the integra-

tion of explicit physical constraints into generative models to ensure real-world consistency, the development of unified self-supervised frameworks for heterogeneous multi-modal data, and the deepening of hardware-software co-design. Ultimately, embedding energy efficiency into the foundational design principles of computer vision is imperative to ensure that the next generation of visual AI sustainably empowers industrial applications while actively advancing global carbon neutrality objectives.

Key words: Green visual AI; energy-efficient data collection; energy-efficient data annotation; energy-efficient model inference; energy-efficient model training

论文引用格式：Han Junwei, Qian Xuelin, Xu Chang, Wang Haoyan, Zhang Dingwen. Green Visual AI: A Survey of Energy-efficient Techniques for Data and Models in Visual Intelligence — SCID[J/OL]. Journal of Image and Graphics. DOI: 10.11834/jig.260181.（韩军伟，钱学林，许畅，王浩研，张鼎文. 绿色视觉AI:面向数据与模型的视觉智能节能化综述—SCID[J/OL]. 中国图象图形学报. DOI: 10.11834/jig.260181.）

0 引 言

在“双碳”战略深入实施与视觉智能产业快速发展的双重背景下，推动视觉智能技术的绿色化发展，是实现经济社会可持续发展的重要途径。我国《“十四五”数字经济发展规划》明确提出推动算力基础设施绿色低碳发展，优化算力资源调度；《“十五五”规划建议》进一步将智能化、绿色化、融合化确立为现代化产业体系建设方向，落实促进绿色低碳发展的科技政策。《“人工智能+制造”专项行动实施意见》则聚焦产业应用落地，要求推动模型轻量化部署，研发推广智能化绿色化协同解决方案。

近年来，以深度学习为代表的人工智能（artificial intelligence，AI）技术取得快速发展，其性能提升在很大程度上受益于模型规模的持续扩大与训练数据的指数级增长。从图像分类到自然语言处理等领域的研究进展表明，模型性能与参数量、数据量之间存在正相关性，这一现象也被概括为“规模化定律”（scaling law）。这一发展路径虽然推动人工智能在多模态交互、世界模型构建、具身智能等前沿方向取得突破性进展，但与此同时，其带来的资源消耗问题也日益凸显。在数据层面，高质量数据的采集往往受限于传感器部署、采集环境及隐私合规要求；而大规模数据标注则需要投入大量人力与时间成本，尤其在医疗、工业等专业领域，标注成本更为显著。例如，计算机视觉领域最著名的ImageNet数据包含1 400万张类别标注的图像；Google构建的Open Images数据集收录了超过900万张经过人工标注的图像，标注内容涵盖图像级标签、目标边界框以及视觉关系等多种不同细粒度的信息。据行业统计，2025年全球数据收集和标记市场规模约达327亿元。在模型层面，千亿乃至万亿参数规模的模型训练需要消耗海量计算资源与能源，带来显著的碳排放压力，同时高频次、长周期的推理任务进一步加剧了计算资源的持续占用。例如，OpenAI早期大模型训练成本可达数百至千万美元，国产DeepSeek-R1模型的增量训练成本虽优化至29.4万美元，但仍需512张H800显卡训练80小时。上述问题集中在视觉智能的数据与模型两大要素，贯穿视觉智能完整的生命周期，其资源消耗规模与绿色低碳发展的政策导向之间存在一定张力，也对视觉智能产业的可持续发展形成现实约束。

在此背景下，绿色视觉AI（green visual AI）作为兼顾技术性能与可持续发展的研究范式受到广泛关注。其核心目标是在保障视觉模型任务性能的前提下，围绕视觉智能全生命周期开展资源优化，以降低数据、算力、人力等各类成本。一个视觉智能的完整生命周期可划分为数据采集、数据标注、模型训练与模型推理四个核心环节，各环节之间相互关联，共同构成了技术落地的成本体系。针对数据采集与标注环节，绿色视觉AI重点关注如何以尽可能少的采集与标注成本获取满足任务需求的数据资源；针对模型训练与推理环节，则聚焦于模型在训练与推理阶段的轻量化设计以及在多场景下的高效复用与迭代，减少重复训练带来的计算开销。上述方向的协同推进，旨在实现数据的高效学习与模型的可持续演进，在推动技术性能持续提升的同时，兼顾资源消耗的合理控制。这不仅是对算力瓶颈与“双碳”政策

导向的积极回应，也为视觉智能技术从实验室走向大规模工业应用提供了可行路径。

综上，数据采集与标注成本的降低，为模型训练提供了高效、低成本的数据基础，减少了训练过程中因数据不足导致的性能问题；模型训练成本的优化，推动了可复用、可持续模型的设计，为模型推理环节的效率提升奠定基础；而模型推理成本的降低，又能推动轻量化视觉AI技术在边缘场景的落地，反哺数据采集环节的场景适配性。本综述将以绿色视觉AI为核心主旨，紧扣我国视觉智能产业绿色低碳发展的政策导向，针对当前视觉AI技术发展中的数据和模型资源消耗的挑战，围绕视觉智能生命周期中数据与模型两大要素所涉及的四个核心环节，从数据采集、数据标注、模型迭代与模型推理四个维度成本框架为主线展开系统性梳理。图1展示了本综述面向数据与模型的视觉智能节能化的技术与分类总结。最后，本综述总结绿色视觉AI的研究趋势，并对未来技术发展方向进行展望。

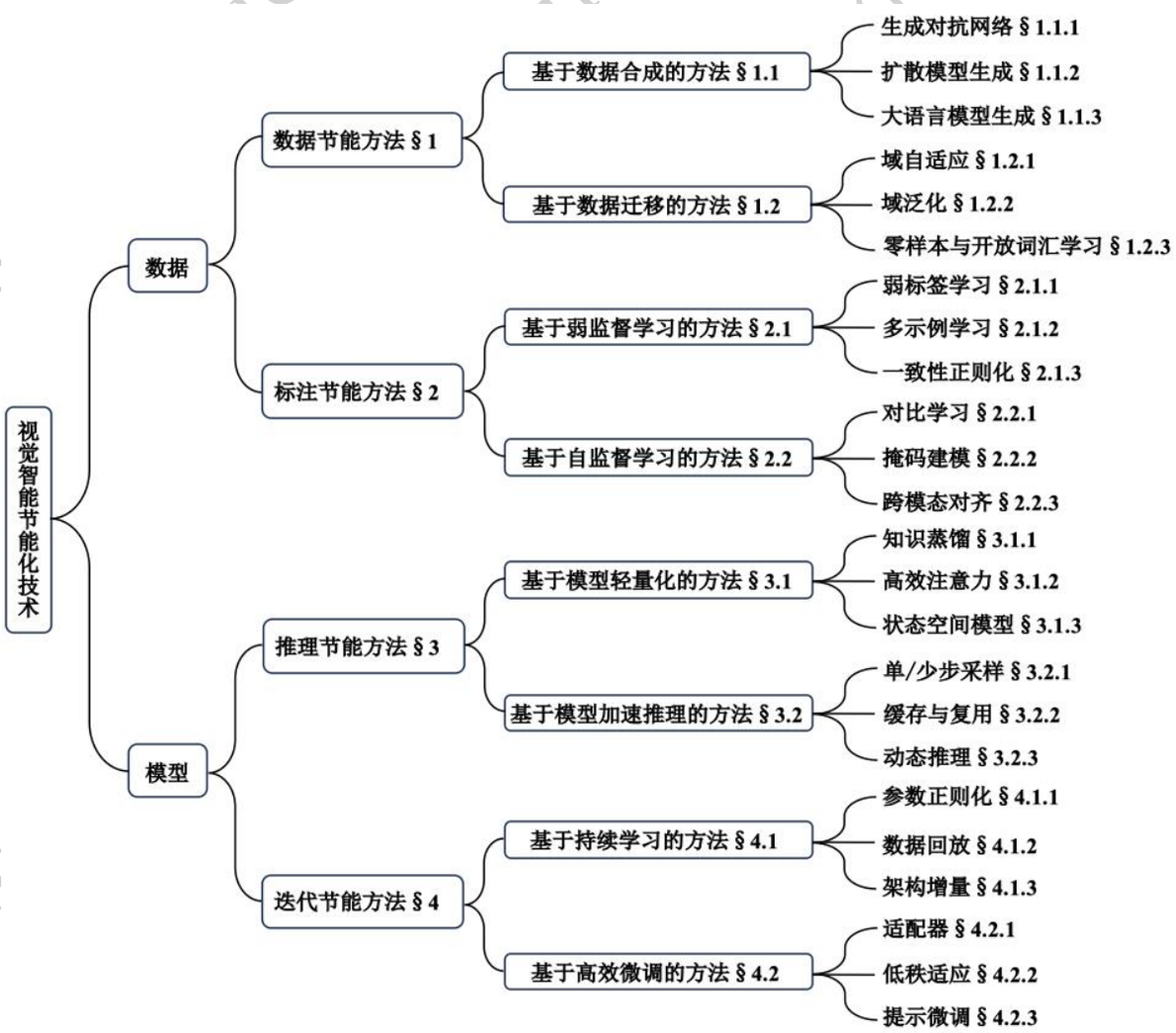

图1　面向数据与模型的视觉智能节能化技术总结

Fig. 1　Taxonomy of energy-efficient visual intelligence techniques from data and model perspectives

1　数据节能型AI方法

数据是视觉智能系统的基石。在当前大模型与多模态快速发展的时代，视觉智能系统对数据规模的需求达到了前所未有的高度。然而，特定领域（如医疗影像、工业缺陷检测）的高质量原始数据往往受限于隐私法规、采集设备或稀缺场景，获取难度极大且单价昂贵；通用领域数据的采集过程又伴随着巨大的存储消耗和清洗成本以及潜在的隐私风险。这

种数据需求与成本之间的矛盾，已成为制约AI技术普惠化与绿色化发展的首要瓶颈。因此，本章节围绕节约数据采集成本，通过数据合成与数据迁移两大类策略，利用技术手段从源头上减少对原始数据的依赖，规避资源浪费与合规风险。

1.1 基于数据合成的方法

数据合成技术通过在虚拟空间中生成仿真图像，替代物理世界中成本高昂的数据采集与人工测绘流程。该方法可基于少量真实样本或预训练大模型所蕴含的先验知识，利用生成式框架对目标数据的特征分布进行建模与泛化，进而批量生成符合特定物理规律与语义要求的数据样本。如图2所示，本节将围绕生成数据质量的发展历程，从生成对抗网络、扩散模型及大语言模型生成三个方向的研究进展进行阐述。

1.1.1 生成对抗网络

生成对抗网络（generative adversarial network，GAN）自提出以来，一直是视觉智能领域通过虚拟生成来扩充数据规模的重要工具。生成对抗网络的核心机制是基于生成器和鉴别器的对抗博弈：生成器通过映射随机噪声来拟合真实数据的潜在分布，其优化目标是使合成样本的分布尽可能逼近真实数据，从而能被鉴别器判定为真；而鉴别器的优化目标则是准确区分输入样本是来自真实数据集还是由生成器合成。在演变早期，GAN面临着对抗训练不稳定、模式崩溃以及高度依赖大规模真实数据等瓶颈。然而，随着网络架构的现代化演进，上述局限性得到了有效缓解。Huang等（2024）对GAN架构进行了系统评估，研究表明，GAN仍可在多个图像生成基准上取得与扩散模型相当甚至更优的生成质量；同时，GAN单步前向生成的特点使其在推理效率和计算开销方面相较多步采样的扩散模型更具优势。这使得GAN成为低成本、高效率获取大规模生成数据的重要技术框架之一。

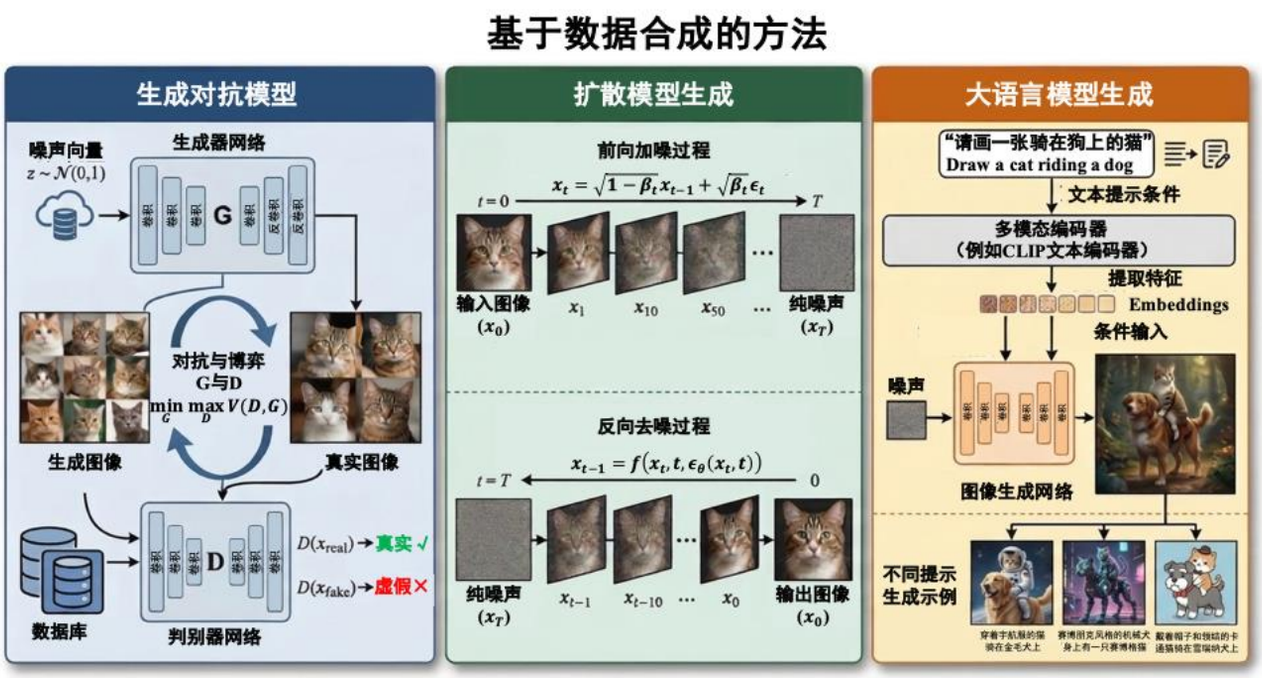

图2 （左）生成对抗网络的工作原理；（中）扩散模型前向与逆向过程示意图；（右）大语言模型生成框架示意图

Fig. 2 （Left）Working principle of GAN；（middle）Schematic of forward and reverse processes in diffusion models；（right）Overview of image generation with large language models

面向真实世界中物理采集受限的场景，数据高效与少样本生成成为GAN近些年的核心研究方向之一。当可获取的真实图像数量有限时，鉴别器容易仅依赖少量样本的局部特征即可完成分类，从而发生过拟合现象。为解决这一难题，Zhou等（2024）提出通过特征分布匹配的方式，在少样本冷启动条件下直接合成质量较高且具有多样性的扩增图像。Ni等（2024）则专门针对数据极度受限的合成场景，提出了基于李普希茨连续性约束的归一化方法，通过缓解判别器过拟合与训练不稳定问题，提升了有

限数据条件下GAN的生成质量与泛化能力。为了进一步丰富极少基础样本衍生出的数据集的多样性，Wang等(2025)引入了风格空间量化技术，缓解了有限数据条件下GAN潜在空间利用不足和生成质量下降的问题。

在丰富生成样本多样性的基础上，无监督的属性解耦与交互式可控生成技术，使GAN能够更具针对性地缓解数据集中长尾样本稀缺的问题。在工业检测或自然场景分析中，深度学习模型往往需要包含特定光照、视角或罕见姿态的训练数据以应对复杂的现实环境。借助可控生成技术，研究人员无需进行高昂的定向物理采集，仅通过调整潜在特征空间的对应参数，即可合成具有特定属性组合的图像，从而有效弥补真实数据集中长尾分布的缺陷，提升下游感知模型的泛化能力与鲁棒性。Pan等(2023)提出的DragGAN为交互式图像合成提供了一种有效方案，该方法允许用户通过在特征图上设定控制点与目标点，来精确调整目标对象的空间布局与姿态，从而降低了在物理世界中定向采集特定动作样本的时间与经济成本。此外，Lee等(2024)提出了一种线性可控的GAN架构，能够在无监督条件下将图像的多种属性进行解耦表征，这种机制支持对特定场景属性进行相对独立的线性编辑与组合重构，使得数据扩充过程从随机采样向按需定制转化，提高了生成样本对下游任务的实用价值。

随着GAN技术的不断发展与完善，其逐渐成为解决因专业门槛、数据隐私、高昂造价等特定领域数据获取难的有效途径之一。例如，在行人重识别和目标跟踪等任务中(Qian等，2019；Wang等，2025；Niu等，2025)，跨摄像头的数据采集与标注面临较高的经济成本与严格的隐私法规限制。Nguyen等(2024)利用GAN框架生成具有目标身份结构信息和不同服饰风格的合成图像，有效缓解了换装行人重识别场景中真实样本采集与服饰标注成本高的问题；Khaldi等(2024)通过合成多姿态的高空视角图像，为缓解航拍ReID中标注样本不足和视角变化问题提供了有效方案；此外，在工业异常检测与医疗影像分析等特定领域，Zhao(2025)提出的AnomalyHybrid框架通过跨域生成异常样本，降低了罕见缺陷样本的收集难度；Wang等(2025)提出的ODA-GAN实现了病理切片的虚拟染色，简化了传统化学染色流程，减少重复切片、实体染色和专家标注依赖。

除了上述特定垂直领域，GAN在数据合成上的泛化优势同样广泛应用至通用视觉场景。在遥感影像分析中，受限于卫星重访周期与云层遮挡，高质量多时相数据的获取具有挑战性。Yang等(2025)提出了一种结构表示引导的GAN框架，能够稳定合成去云的高分辨率遥感图像并保留地物结构，显著扩充了地表覆盖任务的可用数据池。

综上，基于GAN的数据合成技术正朝着高生成效率、少样本冷启动与精细化属性控制的方向持续演进。其生成与推理成本相较于传统采集方法具有显著优势，能够有效契合节约数据采集成本的核心诉求。然而，在面向复杂开放场景、缺乏语义对齐约束的大规模图像生成任务中，该技术在保真度与可控性方面仍面临一定局限。

1.1.2 扩散模型生成

扩散模型(diffusion model，DM)凭借其对复杂高维数据分布的优异拟合能力，已成为当前高保真数据合成领域的核心范式之一。与GAN的对抗训练机制不同，扩散模型的工作原理基于马尔可夫过程的渐进式演化：其前向阶段会不断向真实数据加入高斯噪声，直到数据逐渐退化为近似纯噪声；反向阶段则利用神经网络学习相应的去噪过程，从随机噪声出发，逐步恢复出符合目标分布的数据。这种基于似然估计的去噪机制，有效规避了GAN训练过程中常见的梯度不稳定与模式崩溃缺陷。在生成样本的多样性以及对真实数据分布的覆盖率上，扩散模型展现出了显著的优势，从而确立了其在复杂场景数据生成中的主流地位。

在应用层面，尽管以潜在扩散模型为代表的架构已初步实现了基于文本提示的条件生成，但在特定工业或垂直领域中，仅依靠文本的粗粒度引导往往难以满足对目标姿态、空间布局和物理约束的严格要求。因此，近期扩散模型的研究重点，已从基础的文本条件生成，演进为涵盖几何边缘、深度信息及多模态特征的细粒度空间精准控制。这种从粗粒度引导向高精度定向合成的跨越，使得扩散模型能够更加准确地重构和模拟真实物理世界的复杂分布，为缓解高质量数据采集的难题提供了一种极具价值的技术途径。

在自动驾驶与复杂城市大尺度场景的视觉感知任务中(Li等，2025)，实地数据采集与三维测绘往往面临高昂的经济与时间开销。尤其是在雨雪等极端

天气或长尾分布的恶劣路况下，高质量连续数据的获取具有显著的挑战性与风险。为缓解上述数据稀缺问题，Li等(2024)提出了DrivingDiffusion框架。该方法利用三维空间布局作为先验引导，合成了具备多摄像头视角时空一致性的高保真驾驶视频，有效降低了自动驾驶场景中对真实多视角视频采集与人工标注的依赖。Gao等(2024)提出的MagicDrive框架引入了三维几何层面的细粒度约束机制，该模型允许研究人员通过输入目标边界框或鸟瞰图等条件信息，定向生成具备多视角一致性和真实感的不同天气与路况的街景图像，为大规模城市场景的数据扩充提供了一种低成本的虚拟合成路径。

在场景多样化生成的基础上，目标的跨域定制与属性精准解耦进一步提升了合成数据的多样性与实用性。Zhang等(2023)提出的ControlNet框架为扩散模型的空间约束提供了重要基础，该方法通过设计包含零卷积与可训练副本的独立网络结构，有效解析并融合了Canny边缘、深度图或人体骨架等外部输入信息，这种机制在不破坏预训练模型原有表征能力的前提下，实现了对生成图像几何与空间结构的精确控制。在自然语言指令引导方面，Brooks等(2023)提出的InstructPix2Pix探索了基于文本指令的图像编辑范式，该模型允许研究人员通过输入编辑指令，直接对源图像的属性(如天气、光照条件等)进行定向转换，从而在无需进行全天候重复物理采集的条件下，高效扩充具有同源特征的多场景数据。此外，针对特定目标的跨场景合成需求，Chen等(2024)提出了AnyDoor模型。该方法能够在零样本条件下，保持特定目标实体特征的高度一致性，将其自然地融合至全新背景中。这种免微调的定制化生成机制，为缓解感知模型在跨域场景下的泛化局限性提供了一种有效途径。

此外，扩散模型在缓解数据长尾分布与罕见异常样本合成方面展现出了显著的应用价值。在自然场景与工业缺陷检测等特定领域中，罕见目标或高危缺陷样本的物理收集往往面临极高的成本与客观困难。Zhao等(2023)提出的X-Paste框架利用Stable Diffusion合成稀缺的长尾实例并将其融入背景进行数据增强，有效扩充了低频类别的数据规模。在医疗影像分析场景中，Nazir等(2025)提出了DiffAug框架，其通过结合文本引导的扩散生成与自动分割验证机制，在正常医学图像上局部合成逼真的罕见异常病灶，有效提升了医疗分割模型在小样本条件下的准确率。

扩散模型的数据增强潜力也同样被广泛应用于通用视觉任务。在遥感影像分析中，传统的几何变换往往难以提供像素级的语义多样性。为此，Xie等(2025)提出了一种基于可控扩散模型的数据增强框架，通过结合像素级语义分割掩码与整体嵌套边缘图等结构先验，合成了具备高语义一致性的遥感图像，在保留原始数据集结构特征的同时极大提升遥感语义分割任务的数据丰富度。在目标检测领域，Vu等(2024)针对少样本目标检测提出了多视角数据增强策略，利用可控扩散模型合成兼具基础类低级特征与新颖类高级特征的困难样本，有效增强了检测器的泛化能力。而在视觉-语言目标跟踪任务中，为避免传统空间变换破坏场景构图与文本描述的对齐，Ge等(2025)提出的Gen4Track框架采用免微调的自纠正扩散模型动态生成带有高质量标注的连续视频帧，为跟踪任务提供了一种基于合成视频序列的新型增强范式。

值得注意的是，现代扩散模型不仅具备合成高质量图像的能力，还能在生成过程中同步提取相应的密集标注信号，从而在数据采集与人工标定两个维度上有效降低开销。针对扩散模型在空间位置精确控制上的局限性，Chen等(2023)提出了GeoDiffusion框架，该方法通过将几何坐标转化为空间约束条件引入生成过程，使得模型在合成复杂场景的同时，使合成图像能够与预设检测框保持较好对齐，为目标检测任务提供了一种高效的数据扩充方案。与此同时，Wu等人提出的DiffuMask(Wu等，2023)机制与DatasetDM(Wu等，2023a)框架，通过解析扩散模型内部交叉注意力的特征映射，在图像合成阶段同步提取出具有较高边界贴合度的像素级语义分割掩码与深度信息。这种“伴随生成(generation-with-annotation)”的联合合成范式，有效缓解了密集预测任务中对高昂像素级人工标注的依赖。

综上，基于扩散模型的数据合成技术在复杂场景建模与多模态条件控制方面取得了显著进展。该方法能够有效替代或部分替代成本高昂的物理采集流程，并在生成图像的同时提供语义标签，从而同时降低数据采集与标注成本。然而，其基于马尔可夫链的逆向采样过程计算开销较大，如何在保证生成质量的前提下提升采样效率，仍是其在绿色视觉AI

背景下需要持续优化的关键问题。

1.1.3 大语言模型生成

随着大语言模型(large language model, LLM)及视觉语义模型(vision-language model, VLM)的快速发展,其蕴含的丰富知识表征与逻辑推理能力正逐步应用于视觉数据合成任务。语言是人类对物理世界进行认知与描述的高度抽象符号系统,能够以简洁的形式表达复杂的物体属性、空间关系与语义概念。将大语言模型引入视觉数据生成流程,为数据生成提供了结构化的先验指导,使生成过程得以从单一的像素级拟合拓展为由语言指令全面驱动的语义精准定制,从而在抽象层面实现了复杂场景的灵活构建与多模态数据的高效生成。

在复杂场景布局与空间规划方面,大语言模型展现出了强大的应用潜力。在合成包含多个交互目标的场景时,传统视觉生成模型常出现实体属性错位或空间关系混乱的局限性。为解决此问题,Lian等(2023)提出了一种由大语言模型先验引导的扩散生成框架。该方法利用文本模型的上下文理解能力,将简短的自然语言指令扩展为包含详细空间坐标的结构化表述,进而引导底层视觉基座进行渲染,有效改善了复杂场景下多目标属性绑定的准确性。Feng等(2023)提出的LayoutGPT进一步将大语言模型作为独立的视觉规划模块,仅需自然语言输入,即可基于其内化的常识先验输出合理的目标边界框,为结构化数据的生成提供了一种无须人工干预的自动化路径。此外,Wang等(2025)提出的SKE-Layout通过结合真实物理空间数据与语言模型合成的虚拟数据,增强了感知模型在零样本条件下的空间布局泛化能力;Ran等(2025)则利用大语言模型的思维链(chain-of-thought)推理机制,直接从文本生成三维室内场景布局并进行自纠错,显著降低了传统计算机图形学建模过程中的人工构建成本。

在大规模多模态指令数据的自动化构建方面,大语言模型与视觉语言模型逐渐成为缓解人工标注与数据清洗高昂成本的有效途径。多模态模型的训练高度依赖高质量的“图像-文本”对或指令问答数据集。Chen等(2024)提出的ShareGPT4V工作利用具备较强视觉理解能力的闭源多模态模型,为大规模图像自动生成细粒度的高质量描述与指令数据,验证了利用强泛化能力基座模型构建高质量微调数据集的可行性。针对互联网抓取数据中普遍存在的噪声与低质量问题,Zhou等(2025)提出的MegaPairs框架则通过引入多种底层视觉相似度模型在开放域语料中挖掘异构关联图像对,并协同视觉语言模型进行渐进式指令生成。这种流水线式的数据合成机制,有效缓解了跨模态配对数据集中规模扩展与指令多样性难以兼顾的矛盾。

沿着这一技术脉络,基于大语言模型的数据生成策略已在多个垂直领域与通用视觉场景中得到了广泛应用。在工业缺陷检测领域,由于真实缺陷样本获取困难且分布极不均衡,Heo(2025)提出通过结构化提示引导模型输出缺陷类别、位置与置信度等信息,在零微调或少样本提示条件下实现缺陷识别,从而减少对大规模人工标注与模型重新训练的依赖。在医疗影像分析中,跨中心数据的分布偏移对模型的泛化能力提出了严峻挑战。为此,Duru等(2025)设计了一种基于大语言模型反馈的自适应数据扩充优化策略。该方法通过动态评估下游医疗模型的性能,利用文本模型的逻辑推理能力迭代搜索最优的增强参数序列,在无须额外人工干预的前提下有效提升模型的域间泛化能力。

此外,在遥感解译任务中,传统增强方法往往缺乏语义级别的多样性。Boussaid等(2025)引入大语言模型对遥感视觉问答数据中的问题文本进行语义保持的多样化改写,构建了具备高语义丰富度的多模态数据集,有效提升了视觉-语言模型在复杂地物环境下的理解能力。在更广泛的通用视觉任务中,针对目标检测场景下精确标注成本高昂的限制,Ge等(2025)利用大语言模型合成包含复杂场景布局与空间约束的训练数据,为目标检测器提供了关键的结构辅助线索;而在视觉-语言目标跟踪任务中,Ge等(2025)探讨了利用语言基座进行动态自我纠错并生成时序对齐标签的机制,为跟踪算法注入了更为连续且鲁棒的高阶语义特征,有效改善了目标在形态剧烈变化下的轨迹漂移问题。

总体而言,基于大语言模型指令驱动的数据合成方法,将生成式AI的应用范畴从基础的图像合成拓展至结构化场景规划与语义级自动化标注。该范式通过对物理关系建模与语义理解能力的引入,降低了复杂视觉任务中的人工干预成本,为构建高质量、大规模视觉数据集提供了新的技术路径。

1.2 基于数据迁移的方法

在现实开放环境中部署视觉模型时,常面临训

练数据与测试数据之间的分布差异问题。光照变化、天气条件、传感器类型或地理位置的差异，均可能导致源域训练集与目标域场景之间存在显著的视觉分布偏移。传统应对策略往往需要针对每一个新场景重新采集数据并进行人工标注，这一过程成本高昂且难以规模化推广。基于数据迁移的方法旨在借助特征对齐、域不变表征学习等技术，通过复用源域中已有的数据和知识，降低模型在目标域下对大规模标注数据的依赖。如图3所示，本节将围绕数据迁移技术中降低数据采集成本的三条路径：域自适应、域泛化以及零样本与开放词汇学习的研究进展进行阐述。

域自适应的主要目的，是在目标域样本规模有限或采集成本较低的条件下，通过缩小源域与目标域之间的特征分布差异，将模型在源域中学习到的知识迁移到目标域中。该方法能够增强模型对目标域数据的适应性，从而减少在新应用场景下重新收集和标注训练数据所需的成本。

1.2.1　域自适应技术

无监督域自适应（unsupervised domain adaptation，UDA）的目标是通过缩小源域与目标域在特征分布上的差异，将模型从有标签源域中获得的知识迁移到无标签目标域中。近期的UDA研究聚焦于更为严苛的现实物理与数据约束，例如源域标注数据极度受限，或跨域场景中存在由恶劣环境导致的严重视觉退化（Xu等，2025）。针对源域标注数据获取困难的限制，Zhang等（2025）探讨了单样本无监督域自适应问题，在源域每个类别仅提供单一标注样本的条件下，传统的全局分布对齐机制往往因统计量估计不足而失效。为此，该工作提出了一种跨域链接对比学习框架，通过在极其有限的源域锚点与未标记目标域样本之间构建可靠的语义关联，有效缓解了源域样本匮乏导致的特征坍塌现象，提升了模型在极端少样本条件下的域泛化能力。另一方面，在面对高度复杂的精细化场景时，由于目标间存在严重的视觉相似性与物理遮挡，传统的仅依赖RGB表观特征的域对齐机制容易产生语义混淆与边界模糊。为了突破单一视觉模态在纹理歧义场景中的局限性，Nadeem等（2025）提出了一种几何感知的多模态无监督域自适应框架。该方法创新性地从深度图中提取深度梯度以捕获微小的空间几何过渡，并通过深度梯度引导的交叉注意力机制对RGB特征进行精细化重构。这种跨模态几何约束有效克服了由于光照、土壤成分或生长阶段变化带来的域偏移，显著提升了密集预测任务在复杂环境下的边界锐度与目标区分度。

为了进一步降低大规模源域数据的传输与存储开销，并规避敏感数据的隐私风险，无源域自适应（source-free domain adaptation，SFDA）逐渐成为领域自适应的重要研究方向。SFDA不再直接依赖源域数据，而是仅借助预训练好的源域模型和无标签目标域数据，对模型进行微调并实现特征分布对齐。然而，由于缺乏真实标签的指导，该方法在目标域上生成的伪标签往往包含较多噪声，容易导致模型在自适应过程中产生误差累积。针对这一局限性，Tang等（2025）提出了一种代理去噪机制，在无法访问源域数据的条件下，通过引入额外的结构约束缓解了伪标签噪声的负面影响，实现了目标域特征的自我校准。另一方面，为了提升伪标签的初始质量，Tarashima等（2025）引入了多模态视觉-语言大模型，利用其丰富的通用先验知识为目标域样本生成更可靠的初始伪标签。这种跨模态先验的引入，有效缓解了传统SFDA方法易对源模型初始权重产生过拟合的倾向，在不增加额外人工标注成本的前提下，增强了模型应对较大域偏移的能力。

面向现实世界中动态且不可预知的环境变化，测试时自适应（test-time adaptation，TTA）为模型的高效跨域部署提供了一种轻量级方案。该范式允许模型在推理阶段，直接利用持续接收的无标注测试数据流进行在线参数更新或特征校准，从而在无须离线重新训练的前提下实现对新分布的动态适应。针对复杂场景下的多模态任务，Dong等（2025）提出了一种基于自适应熵感知优化的开放集TTA框架。该方法不仅能够动态调整模型以适应新的数据分布，还能有效识别并处理测试阶段出现的未知新类别。此外，在跨模态检索任务中，为缓解用户在新环境下产生的查询分布偏移，Li等（2025）引入了基于预测优化与联合目标函数的实时动态自适应机制，该策略能够在测试阶段即时更新检索模型，有效降低了对特定新环境重新收集与标定跨模态图文对的依赖与成本。

尽管无监督与无源域自适应方法在诸多场景中取得了显著进展，但在面临复杂且对精度要求较高的任务时，引入少量的目标域专家标注仍是进一步

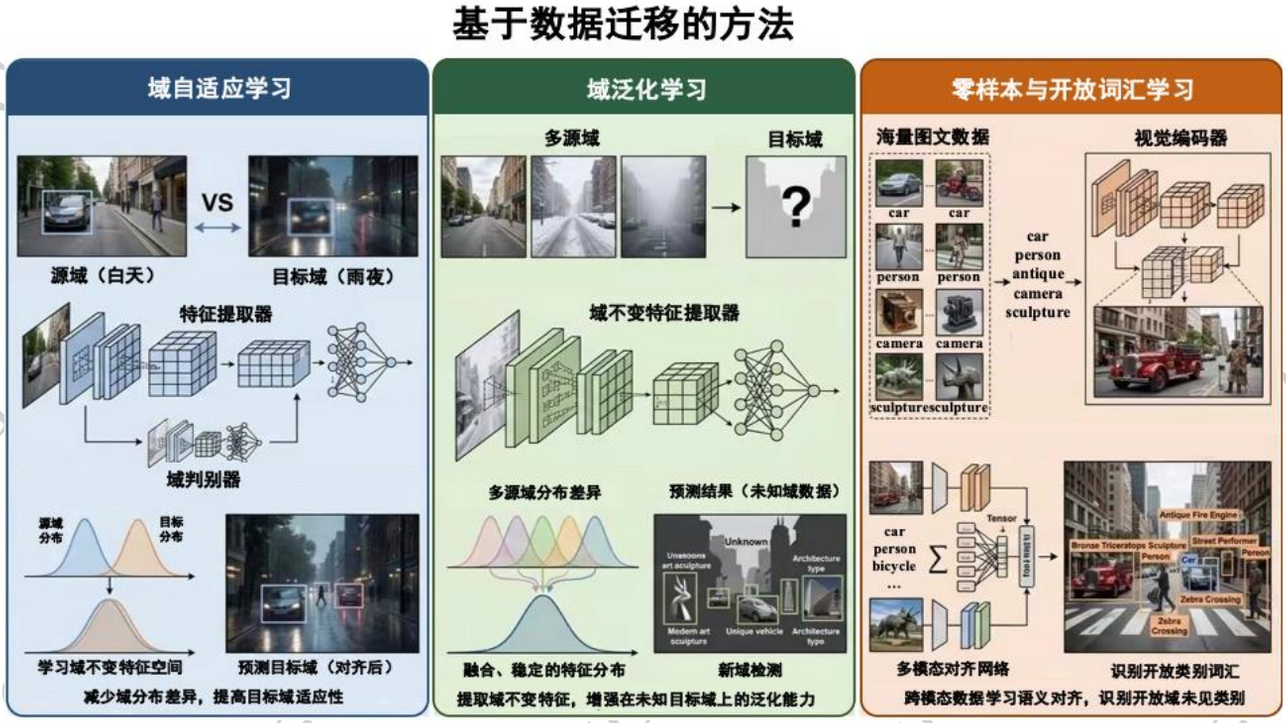

图3 （左）域自适应学习示意图；（中）域泛化示意图；（右）零样本学习与开放词汇学习示意图

Fig. 3 （Left） Illustration of domain-adaptive learning；（middle） Illustration of domain generalization；（right） Illustration of zero-shot learning and open-vocabulary learning

提升模型性能的有效途径。主动域自适应（active domain adaptation，ADA）旨在通过策略性采样，在有限的标注预算内最大化模型的跨域适应能力。针对无源设定下的密集预测任务，Li等（2026）探讨了无源主动域自适应在医疗视频分割中的应用。该方法利用时空相关性评估机制，综合评估视频帧中的空间相关性、时间运动密度与预测不确定性，筛选出少量高价值关键帧进行专家标注。这种策略在避免访问敏感医疗源数据的前提下，有效缓解了视频级像素标注的高昂人工开销。此外，Safaei等（2025）提出了一种协作式主动域自适应框架，该框架结合了主动学习与自训练机制，一方面筛选出模型预测不确定性较高的样本交由人工标注，另一方面将高置信度的预测结果作为伪标签直接纳入训练集。这种联合优化策略在严格控制标注预算的条件下，显著改善了模型在目标域的特征对齐效果。此外，Yang等（2026）针对跨域场景下目标检测模型性能下降的问题，提出了一种融合负教学和负学习的域自适应目标检测方法。该方法围绕平均教师框架下伪标签质量与类别判别偏差等问题进行改进，有助于提升模型在不同数据域之间迁移时的检测鲁棒性与特征对齐效果。

总体而言，当前域自适应技术的研究正围绕降低对目标域数据依赖这一核心目标持续深化。无源域自适应、在线自适应及主动域自适应等方向的探索，旨在进一步放宽对源域数据可访问性、目标域数据规模及成本的约束条件。这些进展有效提升了跨场景视觉模型在部署阶段的适应效率，为降低模型在新环境下的迭代成本、实现绿色视觉AI下的低碳高效部署提供了可行的技术路径。

1.2.2 域泛化技术

与域自适应方法可利用低成本采集的数据实现知识迁移不同，域泛化要求模型仅从一个或多个源域中学习，并在完全不接触任何目标域数据的情况下，直接泛化至未知的新场景。由于无需为目标域采集任何原始图像，域泛化从源头上规避了新场景数据采集的需求，实现了真正意义上的“零采集成本”模型部署，对于降低开放环境下模型迭代与维护成本具有重要意义（Chen等，2025；Son等，2025）。

为了提取具有强泛化能力的跨域不变特征，近期的研究深入探讨了模型优化的底层逻辑与损失空间。在域泛化任务中，寻找损失空间的平坦极小值

已被证明能有效降低分布外的泛化误差。然而，针对传统锐度感知最小化容易陷入“伪平坦”区域的局限性，Song等(2025)重新审视了多源域的优化目标，指出仅最小化全局损失的锐度无法保证特征在单个源域上的平坦性，并提出了一种基于各个域独立锐度的自适应微调策略。在梯度对齐层面，Wang等(2025)提出了一种算术元学习策略，通过精确估计各源域最优参数的质心位置，在确保各域梯度方向一致性的同时，实现了更为均衡的跨域参数聚合。此外，针对训练初期不同源域间梯度冲突导致的局部极小值问题，Ballas等(2025)提出了一种梯度引导退火算法，该方法在训练早期通过动态噪声注入与参数退火，引导模型寻找各分布梯度方向一致的平坦区域，从底层优化机制上提升了模型对分布偏移的鲁棒性。

在表征学习层面，随着视觉-语言模型的发展，引入多模态先验常识进行特征解耦成为了提升泛化能力的重要途径。预训练大模型虽然具备丰富的通用表征，但直接微调往往容易受限于源域的特定偏见。为此，Wen等(2025)提出了一种基于多样化文本提示引导的域泛化框架，该方法通过构建包含差异化上下文的提示集合来模拟未知的目标域分布，并明确抑制视觉编码器中对特定域敏感的冗余特征。针对传统单模态泛化难以处理跨模态间分布差异的瓶颈，Huang等(2025)进一步提出了一种基于统一表征的多模态域泛化框架，通过将不同模态对齐至同一特征空间，实现了特征在面对未知分布偏移时的同步优化。

针对源域数据多样性受限的严苛设定(如单域泛化)，利用生成式模型探索潜在空间并生成伪分布外数据成为了近期的研究热点。Xu等(2025)提出了一种渐进式对抗提示微调框架，通过在扩散模型的生成条件中显式剥离域无关的类别特征与域特定的风格特征，在无须目标域先验的条件下大幅扩充了训练集的数据流形。Thomas等(2025)深入剖析了扩散模型的潜在特征空间，提出了一种无监督的伪域发现机制。该研究证实扩散模型内部的特征能够有效分离出由于光照、视角等造成的隐藏域结构，并可直接用于增强下游分类器的跨域适应性。沿着免微调的生成路径，Noori等(2025)提出的FDS框架，利用多源条件扩散模型进行域插值与反馈引导的伪域合成，生成具有较高多样性的虚拟域样本，并筛选对分类器更具挑战性的合成样本参与训练，从而缓解单一训练分布导致的过拟合倾向。

针对专业领域或隐私保护场景下，域泛化技术也为降低数据采集门槛发挥了重要作用。例如，Tiwary等(2025)提出了一种基于朗之万动力学的数据增强策略，该方法利用基于能量的模型合成不同医院成像风格的中间过渡样本，在无须目标域数据的条件下，使得医学分割网络能够学习到更为鲁棒的解剖学形状先验而非浅层纹理特征，显著提升了模型在跨中心部署时的诊断可靠性。为了克服不同传感器物理分辨率与地理纬度带来的风格变异，Gong等(2025)提出了面向遥感域泛化的视觉基础模型，该框架通过引入地球物理风格模块与多任务预训练机制，提取了具备高度类间可分性的域无关地理空间特征，为跨区域、跨传感器的地表覆盖语义分割提供了一种通用的大尺度模型底座。在强调数据隐私保护的分布式场景中，Liao等(2025)探讨了联邦域泛化问题，利用特征与分类器权重相乘派生的“决策洞察矩阵”作为正则化约束，在完全去中心化的设定下实现了强隐私保护的跨域特征对齐。

总体而言，域泛化技术通过特征解耦、分布对齐等理论方法，结合生成式数据增强策略，持续提升视觉模型在未知场景下的泛化能力。该技术能够在完全不依赖目标域数据的前提下，使模型具备对陌生环境的适应能力，从而有效规避因场景变化带来的数据采集需求，为模型在开放环境下的低成本部署提供重要支撑。

1.2.3 零样本与开放词汇学习

域自适应与域泛化方法通常基于闭集假设，即模型在迁移至目标域时，仅能识别源域中预先定义的固定类别。然而，真实世界中的环境变化往往伴随着新类别的出现。零样本与开放词汇学习在数据迁移的语境下，使模型能够在无需为目标域采集任何图像的情况下，同时适应未知的视觉环境与未知的对象类别。从数据成本的角度来看，该范式在域自适应与域泛化的基础上，进一步节约了新类别数据采集的成本。

零样本学习(zero-shot learning，ZSL)的目标是通过建立已见类别与未见类别之间的语义联系，使模型能够在缺少训练样本的目标域中实现跨域泛化。在学习优化与特征对齐层面，广义零样本学习的核心挑战在于保持视觉与语义空间的一致性

(Cheng等,2023;Huang等,2024;Qorbani等,2025;Xu等,2025;Zhang等,2025a)。近期,Jiang等(2025)的一项研究提出了一种视觉与语义提示协同框架,利用提示微调技术实现了高效的跨模态特征适应,有效缓解了传统方法中视觉特征向语义空间映射时产生的特征混淆。随着视觉-语言大模型的普及,ZSL的研究重心已向如何无损引入大规模预训练语义知识偏移。针对大模型在下游特定领域微调时极易丧失零样本泛化能力的局限性,Deng等(2024)的研究探讨了零样本可泛化的增量学习任务,该工作提出了一种零干扰重参数化自适应机制,通过引入零干扰损失,在不显著增加显存开销的前提下,成功保留了基座模型对未知类别的零样本识别能力。

与侧重于预定义不可见类别的ZSL不同,开放词汇学习(open-vocabulary learning, OVL)要求模型在推理阶段能够处理无边界的任意自然语言概念。这一范式高度依赖于视觉-语言大模型底层先验知识的深度融合。在架构改进方面,Cheng等(2024)提出的YOLO-World在架构中融合了重参数化视觉—语言路径聚合网络,并引入区域—文本对比损失,从而有效提升了视觉信息与语言信息在底层特征层面的交互效率。此外,为应对真实物理世界中动态环境的演变,Xi等(2025)提出了一种统一的OW-OVD框架。该架构不仅继承了检测器的开放词汇泛化能力,还结合了开放世界目标检测的增量学习机制,使得模型能够主动发现未知目标并持续优化特征边界,为OVL在动态场景下的演进提供了新的理论支撑。

在垂直领域的应用场景中,Barsellotti等(2024)探索了免训练的开放词汇分割路径,利用离线扩散模型生成的局部概念辅助视觉匹配,在无须额外像素级训练开销的条件下实现了类别无关区域与语义类的精准对齐。Yuan等(2024)提出了一种统一的Open-Vocabulary SAM架构,该方法通过设计SAM2CLIP与CLIP2SAM双向知识传递模块,在单一网络内完美兼容了基于提示的医疗影像分析交互式分割与开放词汇识别任务。此外,针对遥感解译任务中地物尺度剧烈变化与边界模糊的难题,Li等(2025)提出的SegEarth-OV框架引入了空间信息恢复的上采样器,有效修正了开放词汇特征直接应用于遥感图像时产生的形态畸变;而在三维视觉领域,Jiang等(2024)进一步将视觉语言模型的开放语义常识注入三维点云网络,建立了3D几何结构与文本实体描述的细粒度对应关系,拓宽了开放词汇学习在三物理空间中的应用。

总体而言,零样本学习与开放词汇学习在引入域对齐机制后,突破了传统域迁移方法对可见类别的依赖以及源域与目标域同分布的假设。这类方法借助语言模型对类别语义的描述能力,帮助模型在未见类别或开放场景中实现跨域识别,从而降低因类别扩展或场景变化带来的数据采集需求。该范式对于构建适应性强、标注依赖低的通用视觉系统具有重要意义。

2 标注节能型AI方法

相比于数据采集,数据标注往往是成本更高、耗时更长的环节。随着传感器与互联网技术的普及,大规模原始数据的获取已相对便捷,但针对特定视觉任务(如像素级语义分割、高精度目标检测以及复杂多模态逻辑对)的精细化标注仍需投入大量人力资源;在医疗影像分析、工业缺陷检测等垂直领域,标注工作更是高度依赖专家知识,进一步提高了成本。此外,高强度的人工标注不可避免地会引入主观认知偏差与长期疲劳导致的标签噪声。这种对高质量数据标签的强依赖与高昂的人工标注成本之间的矛盾,已成为制约感知与多模态大模型向更广泛物理场景拓展的瓶颈。因此,本章节围绕节约数据标注成本,通过弱监督学习与自监督学习两大类策略,减少甚至消除对细粒度人工标注的依赖,降低高昂的人力开销与标签噪声风险。

2.1 基于弱监督学习的方法

弱监督学习通过利用低成本、易获取但往往带有误差或粒度较粗的标注信息来驱动模型训练。具体而言,弱监督学习通过设计鲁棒的算法机制、隐式的包-实例推理框架以及基于数据扰动的平滑性约束,引导模型从不完美的标签或无标签数据的内在规律中主动挖掘并提纯有效的监督信号。该方法的核心优势在于能够以较低的人力成本投入换取高质量的训练效果,在显著缓解精细化标注压力的同时,推动视觉感知系统向资源受限或专业门槛较高的场景延伸。本章节将围绕弱标签学习、多示例学习及一致性正则化三类主流方法进行梳理。

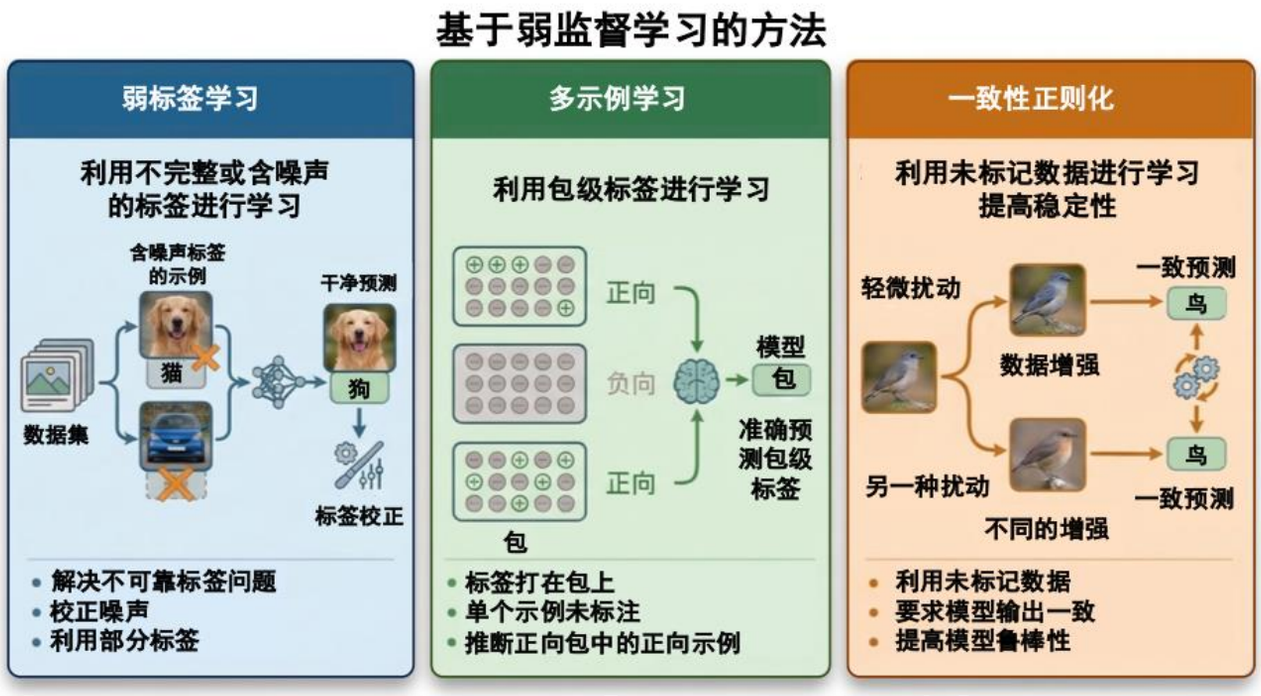

图4 （左）弱标签学习示意图；（中）多示例学习示意图；（右）一致性正则化示意图

Fig. 4 （Left）Illustration of weak label learning；（middle）Illustration of multi-instance learning；（right）Illustration of consistency regularization

2.1.1 弱标签学习

如图4左所示，为降低对高精度人工标注的依赖，研究者常利用质量有限但易于获取的弱标签来驱动模型训练。一种情况是标签本身可能存在一定的错误或偏差，要求设计鲁棒的训练机制以抑制噪声对模型性能的影响，即噪声标签学习。另一种情况是直接利用模型对无标注数据进行预测，将高置信度预测结果作为伪标签参与训练，实现自我迭代与性能提升，即伪标签学习。上述两类方法均旨在以较低的标注成本获取训练信号，但同时也需应对因标签质量参差不齐所带来的学习挑战（Kweon等，2025；Xue等，2023；Yu等，2025；Zhang等，2025b；Zhu等，2025）。

在噪声标签学习的研究中，核心挑战在于深度神经网络对训练数据中随机错误或偏差的过度拟合倾向，易导致模型性能下降。为从包含错误标注的数据中提取有效监督信号，近期研究致力于在训练层面阻断噪声对模型的影响。例如，Kim等（2024）提出一种基于训练动态分析的噪声检测方法，通过在潜在表示空间中对比干净样本与噪声样本的演化差异，实现对数据集中错误标注的识别。为进一步抑制过拟合，Kim等（2024）引入结构化标签机制，对特征空间中高度相似但标签错误的样本施加软标签约束，从而缓解随机噪声的干扰。针对长尾分布下的噪声标签问题，Zhou等（2024）设计了一种模型自举框架，利用预测置信度实现动态自我校正。针对精力疲劳引入的标注相关错误，Nagaraj等（2025）推导了一种鲁棒损失函数，能够在线估计并消除时序噪声的累积影响。此外，针对多视图感知中传感器损坏导致的错误标注，Xu等（2025）利用跨视图邻域信息实现无偏预测校准。

与被动容忍标签噪声的思路不同，伪标签学习借助深度特征、先验知识等信息主动挖掘无标注数据中的监督信号。由于信息本身的不确定性，如何提高伪标签的质量与可信度成为伪标签学习研究的核心关注点。针对3D目标检测中空间标定成本高昂的问题，Zhang等（2024）将伪标签生成过程解耦为二维属性与三维深度属性，缓解了深度估计误差带来的梯度冲突。在跨数据集的3D目标检测任务中，Zhang等（2024）通过互补增强策略筛选高纯度伪边界框，实现自动化标定替代人工标定。针对图像语义匹配等像素级对齐任务，Dünkel等（2025）提出利用3D感知的伪标签链自动生成匹配标签，使模型能够自发学习视觉对应关系。为解决伪标签生成

中的类别失衡问题，Lü等（2024）引入类别感知的置信度估计，增强了伪标签分布的多样性。在交叉学科领域，伪标签技术也取得了一定的进展。例如Juan等（2024）将该理念引入分子科学领域，结合启发式规则自动化生成分子属性伪标签，验证了弱监督范式在跨学科研究中的适用性。

综上，弱监督学习的核心在于通过改进算法机制设计，降低模型对标签质量与数量的需求。这类方法以算法能力替代部分人工标注投入，在压缩标注成本的同时，也挖掘了大规模无标签数据在模型预训练中的潜在价值。尽管在医疗、自动驾驶等高度敏感场景中，弱监督方法在标签噪声的敏感性和预测置信度的可靠性方面仍面临挑战，但其在构建大规模视觉感知系统、推动模型向低资源场景延伸方面，已展现出重要的工程实践意义。

2.1.2　多示例学习

如图4中所示，噪声标签学习与伪标签学习主要从训练层面出发解决标注质量不高（如标签错误或噪声）带来的问题；而多示例学习则从另一角度切入，通过降低标注的粒度来缓解大规模数据标注的压力。例如，在高分辨率图像分析或长时序视频处理等任务中，要求专家对每一个局部区域或时间帧进行精细标注，成本往往难以承受。多示例学习（multi-instance learning，MIL）则为此提供了一种弱监督学习范式：它将数据组织为包含多个实例的“包”，仅需提供包级别的粗粒度标签，引导模型学习包中触发该标签的关键特征（Ye等，2026）。这一机制在降低标注精度的同时，有效压缩了人工标注的投入。Jang等（2024）基于概率近似正确学习理论，推导了MIL算法实现实例级可学习性必须满足的理论边界与充分条件。该工作为粗粒度标签替代实例级精细标注提供了数学依据。

在长时序视频分析任务中，多示例学习能够在仅提供视频级粗粒度标签的条件下，挖掘触发异常标签的关键片段。Zhu等（2024）提出了一种长短期时间序列关联的视频异常事件检测方法，该方法将Transformer引入时间序列的多示例学习框架，并结合长短期注意力机制突出局部异常事件与正常事件之间的差异，从而提升了弱监督视频异常检测性能。在多模态视觉任务领域，Chen等（2024）提出了提示增强MIL框架，仅需针对长时序视频提供全局异常标签，即可自动识别出高辨别性的异常动作片段，并精确划定时空边界。针对小样本设定下MIL模型易出现的注意力偏差问题，Zhang等（2026）引入了注意力熵最大化正则化技术，显著提升了弱监督模型在数据匮乏环境下的泛化性能。为了进一步降低对大规模包标签的需求，Qu等（2024）将提示学习引入MIL框架，通过融合视觉与文本的先验知识，使得模型在极少样本下即可实现高效收敛。针对音视频解析中事件重叠与时间定位标注成本较高的问题，Zhou等（2024）提出了标签语义引导的事件解耦框架，利用视频级粗粒度标签文本作为语义先验，将音频与视觉片段特征投影至语义独立的标签嵌入空间，从而在时间维度上实现视听事件的解耦与定位。

在医疗图像分析等专业领域，多示例学习也展现出了优异的弱监督学习性能。例如，Castro-Macías等（2024）提出了一种多示例学习平滑算子，通过建模相邻实例之间的局部依赖关系，使模型在仅依赖切片级弱标签的条件下提升病灶区域的局部定位能力，从而缓解对医生手动像素级标注的依赖。Gou等（2025）通过可查询原型池对实例特征进行原型引导聚合，生成包级特征表示，并结合类别文本特征增强机制支持新数据到来时的动态原型学习与更新，从而缓解增量学习中的灾难性遗忘问题。

综上，多示例学习通过构建从“数据包”到“局部实例”的隐式推理机制，将人工标注的粒度从高成本的像素级或帧级降低至图像级或视频级。这一范式在一定程度上突破了精细化标注环节中人力高需求瓶颈，特别是为医疗影像、缺陷检测、目标识别、文档分类等多模态视觉任务的应用场景提供了切实可行的技术路径。

2.1.3　一致性正则化

如图4右所示，一致性正则化（consistency regularization）通常通过对无标签样本加入不改变其语义的扰动，并要求模型在不同变换视图下产生一致的预测结果，从而在损失函数中加入平滑约束。这一机制使模型能够在无标签数据上学习到对扰动鲁棒的表示，在一定程度上降低了对人工标注的依赖（Ding等，2025；Kashiani等，2025）。本小节将围绕一致性正则化在表征优化、跨模态学习以及复杂场景应用等方面的研究进展进行梳理。

早期的一致性正则化方法主要约束模型在输出层的预测分布，但在高维视觉任务中，这种末端约束往往难以充分传递样本间的语义关联。近期研究开

始将一致性约束从输出层向特征空间延伸。Lu等(2025)指出，仅在解码器端施加一致性约束难以影响骨干网络的特征提取过程，为此他们在编码器的中间层引入Wasserstein距离，通过优化无标签样本在特征空间中的对齐程度与分布均匀性，减少了语义分割任务对像素级标注的需求。在理论层面，Ni等(2024b)分析了一致性正则化的作用机制，发现其通过隐式调节权重梯度来增强模型的泛化能力，并在低标签率条件下保持了与预训练模型之间的知识一致性。针对特征空间中样本分布稀疏的问题，Wang等(2024)提出了一种基于密度的特征扰动策略，使模型在低密度区域保持预测一致性，提升了在困难无标签样本上的表征鲁棒性。此外，为缓解知识蒸馏中由噪声扰动和类别不平衡引发的预测偏差，Wang等(2025)设计了一种类内知识蒸馏方法，通过在同一类别内部共享教师模型的输出知识，促使学生模型形成更一致的类别预测，从而提升一致性正则化下的泛化稳定性。

除了图像空间维度，数据在连续时间维度以及多视点几何关系上的一致性正则化也为模型学习提供了有效的约束。例如，Vincent等(2025)提出语义相似性传播机制，通过跨帧传播分割预测并引入时间一致性约束，使模型能够利用稀疏标注视频中的连续帧相关性，在自主飞行场景下实现高时间一致性的视频语义分割。Luo等(2025)针对端到端半监督文本检测与识别任务设计了融合空间一致性与内容一致性的框架，通过在位置与转录信息的双向流动中生成可靠的层级伪标签，缓解检测与识别任务之间伪标签不一致的问题，并在少量标注数据条件下取得了接近甚至超过强监督文本检测识别器的性能。在时间序列分析领域，Chen等(2025)提出上下文感知的一致性学习框架，利用时序数据中的局部连续性动态校正相邻片段间的标签冲突，提升了模型在分段时间序列分类任务中的鲁棒性。

一致性正则化的应用范围也同样拓展到了视觉基础模型领域。针对视觉-语言大模型在少样本微调中容易过拟合的问题，Roy等(2024)提出一致性引导的提示学习方法，通过约束可训练模型与预训练模型在预测结果上的一致性，并结合扰动输入下的一致性正则化，在低标注依赖条件下保持模型的零样本泛化能力。在减少对人工反馈依赖方面，Wang等(2024)提出CREAM框架，利用多轮迭代间奖励分布的一致性对大语言模型的自我奖励机制进行正则化，从而降低了对专家偏好标注的需求。此外，在扩散模型的主体驱动生成任务中，Ni等(2025)引入噪声一致性正则化，通过约束微调模型在先验样本上的噪声预测与预训练基础模型保持一致，并增强主体潜变量在噪声扰动下的预测稳定性，在仅使用少量图像样本进行微调时提升主体身份保真度与生成多样性。

总体而言，一致性正则化通过挖掘数据内在的结构规律与时空连续性，使模型能够在无标签数据上构建自洽的监督信号。该方法借助特征空间中的语义对齐或时序维度上的逻辑一致性，引导模型从数据自身中学习鲁棒的表示，从而减少对外部人工标注的依赖。这一机制可为构建低标注成本的视觉感知模型提供可行的技术路径。

2.2 基于自监督学习的方法

与依赖外部人工标注的弱监督学习不同，自监督学习通过设计代理任务，从数据本身的结构中构造监督信号，使模型能够利用未标注数据直接进行训练。具体而言，自监督学习通过挖掘视觉数据中的空间结构、上下文关系及多模态信息的一致性，引导模型在特征空间中形成具有通用性与判别力的视觉表征。该方法的核心优势在于能够充分利用海量未标注数据，在显著降低标注成本的同时，为下游任务提供高质量的预训练模型。本章节将围绕对比学习、掩码建模及跨模态对齐三类主流方法进行梳理。

2.2.1 对比学习策略

经典对比学习通常会对同一张未标注图像进行不同形式的数据增强处理，例如随机裁剪、颜色扰动等，以此构建正样本对；同时，将同一批次中的其他图像作为负样本进行对比学习(如图5左所示)。通过在特征空间中拉近正样本间的距离，拉远负样本间的距离来学习对扰动鲁棒性且具有判别力的视觉特征。在近期的研究工作中，Tan等(2024)通过理论推导证明了标准对比学习在本质上等价于在样本相似度图上执行谱聚类。沿着这一理论脉络，Luthra等(2025)和Lee等(2025)从不同角度揭示了自监督对比学习与监督表征学习目标之间的近似关系，为InfoNCE等对比损失作为人工标签替代信号提供了理论解释。

然而，经典对比学习框架在实际应用中仍面临

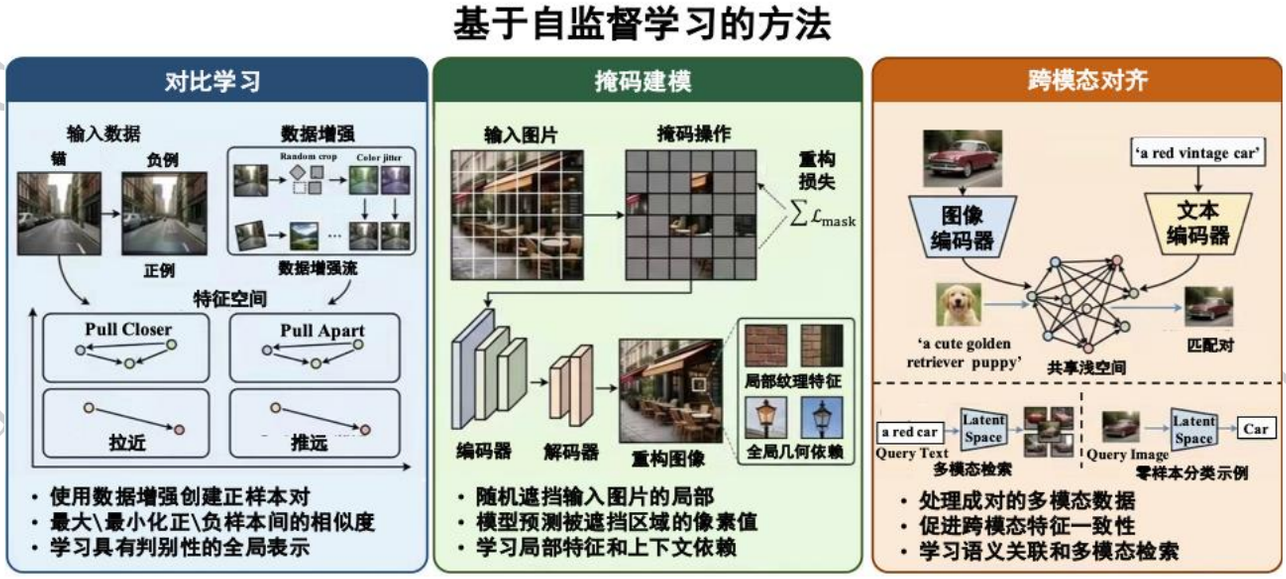

图5 （左）对比学习示意图；（中）掩码建模示意图；（右）跨模态对齐示意图

Fig. 5 （Left）Illustration of contrastive learning；（middle）Illustration of masked modeling；（right）Illustration of cross-modal alignment

计算开销与语义冲突两方面的挑战。以Chen等（2020）提出的SimCLR为代表的对比学习方法通常依赖大规模批次数据来保证负样本的多样性，这对计算资源提出了较高要求。针对这一问题，Sharma等（2024）将对比学习与最大化目标相结合，提出一种对批大小不敏感的对比损失函数，降低了模型训练对高性能计算资源的依赖。另一方面，在MoCo（He等，2020）、SimCLR（Chen等，2020）等框架中，由于缺乏标签信息，同一类别的不同图像可能被误作负样本进行区分，产生“假负样本”问题。为此，Zhou等（2024）通过引入零均值正则化，在一定程度上缓解了这种错误连接对特征表示的影响。此外，Sobal等（2025）突破了对比学习中正负样本二元对立的设定，提出基于连续相似度图的学习方式，使模型能够在无监督条件下捕捉样本间的语义相似性。

针对传统方法偏重全局特征、在密集预测任务中表现受限的问题，Hu等（2024）将多视图掩码机制与对比学习相结合，提出了一种无需标注即可同时学习全局语义与细粒度空间特征的预训练方法。Zhao等（2025）指出，大容量网络在无监督对比学习中容易偏向纹理等表层特征，他们通过在对比学习框架中引入人类视觉系统的感知偏置，在加速模型收敛的同时，改善了底层特征的语义纯度。此外，在低标注率与小样本场景分类任务中，Zhang等（2022）提出了一种面向小样本遥感图像场景分类任务的自监督学习方法。此方法采用教师网络与双学生网络协同预测，并在分类器前引入自监督对比学习，通过度量同类样本的类中心距离增强类间边界，从而提升了模型在小样本条件下的泛化能力。随着对比学习特征被广泛用作视觉基础模型，其安全性问题也逐步受到关注。Zhang等（2024）围绕自监督对抗训练展开研究，在无标签条件下提升预训练表征对对抗攻击的鲁棒性。

总体而言，对比学习作为自监督视觉预训练的重要范式之一，通过实例判别机制在无标注数据上学习具有判别力的视觉特征。然而，对比学习的核心仍依赖于样本间的显式对比，其在全局特征建模上的优势难以直接迁移至需像素级理解的密集预测任务，且对负样本质量与数据增强方式的依赖也限制了其在某些场景下的适用性。针对上述局限，掩码图像建模与跨模态语义对齐作为自监督学习的另外两条主流路径，从不同角度探索了更细粒度或更高效的特征学习方式。

2.2.2 掩码建模策略

掩码建模（masked image modeling，MIM）提供了自监督学习的另一类重要范式。如图5中所示，这类方法通过重构被遮挡的信号或图像内容，引导模型学习数据的内在结构与分布规律。该思想在自

然语言处理领域得到广泛验证，典型工作包括BERT(Devlin等，2019)和T5(Raffel等，2023)。

受此启发，计算机视觉领域也引入掩码建模策略学习图像表征先验。例如，Bao等(2022)提出的BEiT将图像离散化为视觉词元(Token)，并通过掩码预测任务进行预训练。He等(2022)提出的MAE采用非对称编码器-解码器架构与高比例随机掩蔽策略，在降低计算开销的同时引导模型学习图像的全局结构信息。随后，Yuan等(2024)提出的SemanticMIM框架引入了对比学习的语义约束，提升了掩码模型全局特征的可分性。Przewięźlikowski等(2025)系统分析了MIM表征的形成机制，改善了模型在免微调场景下的表示能力。在效率优化方面，Xiang等(2025)将掩码预测转换至小波频域，通过在频域中构建更紧凑且保留空间层级信息的重建目标，加速了自监督预训练过程并降低了计算开销。此外，Lee等(2025)将掩码机制与隐式神经网络表示相结合，增强了模型在处理分布外数据时的重建与特征鲁棒性；Wang等(2025)则利用掩码自编码器实现了红外与可见光图像的无监督融合，拓展了MIM在多模态感知中的应用。

随着掩码建模在视觉任务中展现出良好的泛化能力，其基于无标注数据的预训练优势正逐步拓展至三维空间感知、多模态融合及医疗影像等前沿领域。为应对大规模3D激光雷达点云高昂的标注成本，Cheng等(2025)提出LSV-MAE，通过对点云隐式特征的重建，引导模型理解复杂的物理空间几何结构。在3D人体重建任务中，Fiche等(2025)将人体姿态与形状离散化为Token序列，并训练掩码生成式自编码器根据输入图像和部分网格Token预测完整人体网格，从而以生成式掩码建模方式提升人体网格恢复能力。在医疗影像领域，Jang等(2026)提出将掩码图像重建为标准色彩空间，在利用无标签数据提取语义特征的同时，缓解了不同医疗设备引入的领域偏移问题。

掩码建模通过重建被遮挡的图像区域，引导模型从大量无标注数据中学习视觉特征的分布规律。该方法将特征提取的目标从拟合语义标签转向对原始像素结构的自监督学习，有效降低了对人工标注数据的依赖。这一范式为构建通用、高效的视觉基础模型提供了一条可行的技术路径，也在一定程度上支撑了当前视觉智能体系中大规模预训练的实践。

2.2.3　跨模态对齐策略

与单模态内的对比学习不同，跨模态对齐构成了一种基于语义关联的对比学习范式。如图5右所示，该方法利用大规模弱对齐的多模态数据，如图像与文本对，在共享特征空间中建立不同模态之间的映射关系，借助多模态相关性引导视觉特征表征学习(Zhang等，2021)。

在跨模态对齐的早期探索中，Radford等(2021)提出的CLIP与Jia等(2021)提出的ALIGN奠定了基于双流结构的对比学习基础。这些方法通过最大化匹配图文对在共享空间中的特征相似度，使视觉模型在大规模预训练后具备一定的零样本泛化能力。在此基础上，Yu等(2022)提出的CoCa进一步融合了对比学习与生成式建模。该模型采用编码器-解码器架构，在特征提取阶段进行跨模态对比对齐，同时在解码阶段引入图像描述生成任务。这种将判别式目标与生成式目标相结合的策略，在增强模型对细粒度视觉元素感知能力的同时，也提升了其语义表达能力。

随着跨模态双流架构的逐步成熟，近期研究开始关注模态间特征对齐的细粒度问题。Yamaguchi等(2025)的工作分析了视觉-语言基础模型中图像特征与文本特征分布分离的问题，并提出一种基于后预训练的对齐策略，在不破坏原有知识的前提下提升了模型在下游任务上的迁移性能。针对复杂时空场景中的对齐问题，Chen等(2025)进一步引入因果推断机制，通过剔除视频特征与问答文本之间的伪相关性，改善了视频时序定位任务中的特征一致性。

除了2D视觉任务，跨模态对齐在3D感知任务同样是研究热点之一，旨在降低3D数据的高昂标注成本。Huang等(2025)提出的3D CoCa框架将对比学习与生成式建模的统一思路拓展至点云领域，通过联合优化3D-文本对比学习与3D场景描述生成，在无需外部目标检测器或候选区域提取的情况下，实现了3D几何结构与语言描述在共享空间中的对齐。针对3D场景中分布外泛化能力不足的问题，Tang等(2026)进一步提出3D CoCa v2，引入测试时搜索机制，在不更新模型参数的前提下提升了跨模态特征的鲁棒性。

在多模态大模型热潮中，跨模态对齐策略也从

单一目标描述向场景级理解延伸。Sarkar等(2025)提出的CrossOver框架设计了一种场景级的灵活对齐策略,将RGB图像、点云、CAD模型及文本映射至统一的模态无关嵌入空间中,增强了模型在部分模态缺失情况下的检索与定位能力。在三维场景推理方面,Huang等(2025a)提出3D-R1,通过构建包含思维链的Scene-30K数据集,并在强化学习阶段引入感知奖励、语义相似性奖励与格式奖励,将三维空间定位精度与语言推理质量进行联合优化,推动跨模态对齐从特征匹配向几何推理层面拓展。

综上所述,跨模态对齐范式通过挖掘多模态数据间的语义关联,降低了对人工精细标注的依赖,为构建具备开放词汇与泛化能力的多模态基础模型提供了有效的技术支撑。

3 推理节能型AI方法

随着视觉智能与深度学习技术的快速发展,模型正从理论研究走向实际部署。在此过程中,推理效率成为制约模型落地应用的关键因素之一。与训练阶段成本投入不同的是,推理过程具有高频次、长周期持续运行的特点,其计算负载的累积不可避免地带来了能耗与资源占用。近年来,多模态大模型的参数规模已突破千亿级别,性能提升的同时也加剧了推理阶段的算力需求与能源消耗。本章节将从模型轻量化与推理加速两大技术路线出发,对算力与能耗高效的绿色视觉AI推理方法进行梳理。

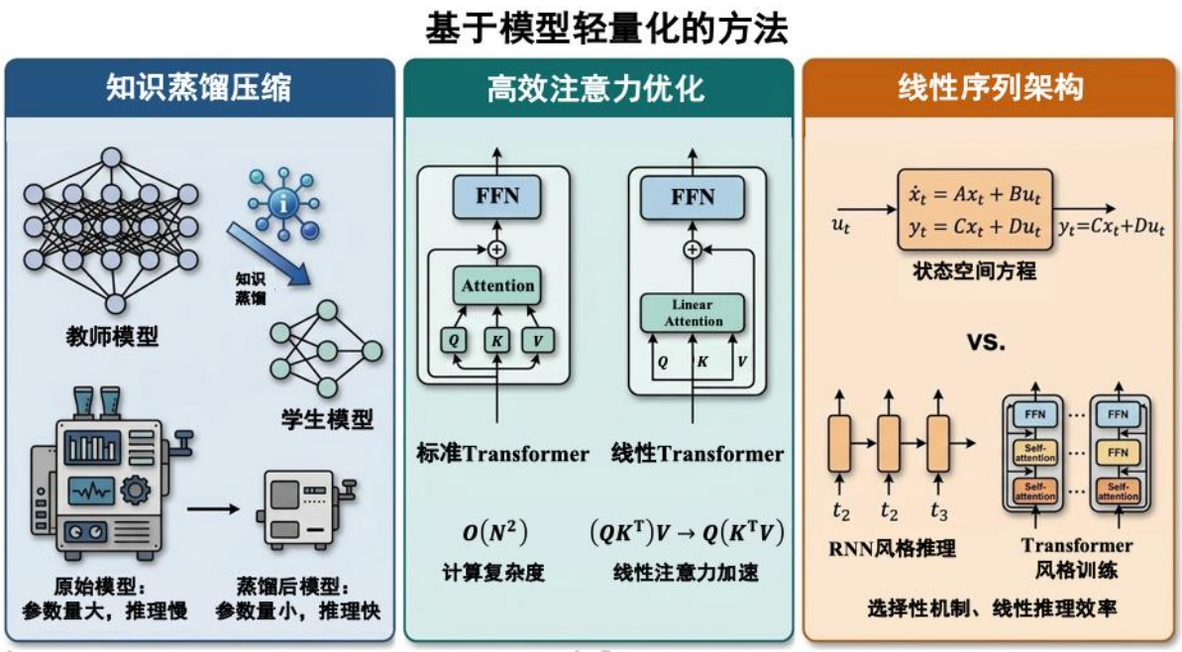

图6 (左)知识蒸馏压缩示意图;(中)高效注意力优化示意图;(右)线性序列架构示意图

Fig. 6 (Left) Illustration of knowledge distillation; (middle) Illustration of efficient attention mechanism; (right) Illustration of linear-time sequence models

3.1 基于模型轻量化的方法

模型轻量化技术的核心目标是在尽可能保持模型原有性能的前提下,通过减少参数量、降低计算复杂度及优化内存占用等途径,实现模型推理成本的压缩。该技术旨在从模型结构设计与参数压缩的角度入手,在性能与效率之间寻求平衡。

3.1.1 知识蒸馏压缩

如图6左所示,知识蒸馏(knowledge distillation, KD)是一种通过使用学生小模型学习教师大模型以实现模型轻量化的一种模型压缩技术。自Hinton等(2015)确立“教师-学生”范式以来,KD已从简单的分类任务扩展至大语言模型、多模态感知等复杂场景,形成了涵盖知识提取、知识对齐、模型优化的完整技术体系。通过让学生模型学习教师模型的显性输出与隐式知识,打破模型性能与计算成本的强绑定关系。在训练中,它引导学生模型学习

教师模型的显性输出与隐式特征，从而将高质量表示压缩至参数精简的架构中。这使得学生模型在显著降低推理成本的同时，能够保持与原模型相近的预测性能（Gao等，2026；Fei等，2025）。

知识提取阶段的核心不仅仅在于缩减参数规模，也在于压缩单次前向传播的计算强度。以DistilBERT为例，Sanh等（2019）通过三重损失函数对齐软标签与隐藏层状态，在参数规模减少约40%的情况下保留了97%的性能表现。在多模态场景中，Jang等（2025）通过引入知识浓缩层显式提取跨模态语义关系，使学生模型无需重复构建完整的高维对齐空间。该层将大视觉语言模型的高维表征压缩适配至轻量模型维度，在蒸馏阶段完成跨模态知识迁移，有效降低轻量化模型推理阶段的特征计算开销，适配资源受限设备的部署需求。此外，针对图像领域的蒸馏研究也不断丰富，例如Guo等（2023）提出的语义感知蒸馏与能够跨越异构网络及不同图像尺寸的像素级蒸馏（2024）方法，进一步拓展了知识提取的应用场景与灵活性。

知识对齐阶段的核心在于修正学生模型的中间推理轨迹，防止误差累积。如果仅将蒸馏聚焦于输出层，学生模型通常难以学习复杂任务中的深层推理过程。Wang等（2025）通过追踪内部知识流指出，视觉-语言模型的多模态知识在跨越关键层后，中间网络的表征分布会迅速收敛且演化趋于稳定，层间变化幅度极小。针对蒸馏忽视中间层表征对齐的局限，Hao等（2025）提出了Low-Rank Clone（LRC）框架，通过低秩投影直接对齐前馈网络的激活值。在跨模态感知方面，Liu等（2025）提出的MonoTAKD框架设计了残差蒸馏策略，仅提取点云教师模型中独有的空间几何特征残差，并将其注入到视觉学生模型中，避免了无差别地模仿全量特征带来的噪声干扰。上述研究表明，细粒度的过程对齐能在不增加计算强度的前提下，有效提升模型在复杂场景下的推理鲁棒性。

模型优化阶段则从数据规模、系统架构及迭代策略三个方面降低全链路成本。在数据规模层面，Wang等（2025）提出的基于神经特征函数的数据集蒸馏方法，将大量训练样本压缩为少量合成数据原型，大幅降低了显存开销与计算复杂度。在系统架构方面，Cao等（2025）针对多视觉编码器并行计算所带来的算力冗余问题，采用专家机制将多个教师编码器的知识高效蒸馏并整合至单一学生编码器中。这不仅解决了多教师间的知识冲突，还将推理成本从多编码器并行降低到单编码器的水平，实现了结构层面的优化。此外，在迭代与训练策略方面，分步优化的思想被广泛应用。例如，Zhang等（2025）针对低分辨率病理图像提出了多步混合知识蒸馏策略，通过分阶段训练有效克服了特定领域的目标检测瓶颈；而面向大模型领域，Shu等（2025）提出的LLaVA-MoD框架设计了一种渐进式蒸馏策略，该策略通过模仿蒸馏与偏好蒸馏两阶段迭代，在仅使用0.3%的训练数据和激活23%的参数的条件下，使2B学生模型在多项基准上超越了7B教师模型，降低知识迁移过程的计算与数据成本。

通过对教师模型输出分布与中间表示的深度对齐，知识蒸馏成功将高算力消耗压缩至离线训练阶段，使在线推理能够由轻量化的学生模型承担。这种“重训练、轻推理”的结构不仅显著降低了单次推理的边际成本，也为大模型在资源受限环境下的部署提供了可行的折中路径。

3.1.2 高效注意力优化

Transformer架构在自然语言处理、计算机视觉及多模态任务中应用愈发广泛，现已成为现代模型的核心骨干结构。无论是语言模型、视觉模型，还是跨模态融合框架，Transformer中的自注意力机制承担着全局信息建模的关键功能。然而，在模型规模与输入序列长度不断扩展的背景下，自注意力的计算复杂度逐渐成为影响实际部署效率的重要问题。如图6中所示，在标准自注意力机制中，计算与内存开销随序列长度呈平方级增长，即$O(N^2)$复杂度。当模型处理长文本、高清视频帧序列或高分辨率图像时，注意力矩阵的构建与存储将带来显著的推理时延与显存占用增长。这种复杂度增长不仅提高了单次推理成本，也限制了模型在边缘设备或资源受限环境中的可部署性。

针对这一瓶颈，高效注意力机制研究成为模型优化的一个重要方向。其中，线性注意力（linear attention）作为代表性分支，通过核函数映射、矩阵分解或递归聚合等方式对注意力计算进行线性化改造，将复杂度降低至$O(N)$或$O(NlogN)$。其基本思想在于避免显式构建$N \times N$的注意力矩阵，而通过可分离核函数或特征映射实现全局信息的线性聚

表1　高效注意力机制技术原理。

Table 1　Technical Principles of Efficient Attention Mechanisms.

方法	技术原理
Linear Attention	非负特征映射 + 递归转化
RFA	随机傅里叶特征近似
PolaFormer	查询 - 键符号分解
xLSTM	矩阵记忆单元 + 指数门控
ToST	词元二阶统计量算子
DuoAttention	检索头与流式头并行
Tiled Flash Linear Attention	块内切片 + 分层并行
QuickLLaMA	查询感知动态稀疏筛选

合，从而在保持长程依赖建模能力的同时显著降低计算与内存开销。为便于梳理不同方法之间的差异，表1汇总了本节讨论的典型高效注意力机制及其核心实现原理。

线性注意力的发展可追溯至对Transformer与循环神经网络（RNN）等价性的理论探索。Katharopoulos等（2020）证明，在采用如elu(x) + 1非负特征映射后，自注意力机制可转化为具有恒定内存占用的递归形式，从而将时间和空间复杂度降至线性级别。尽管该方法在复杂语言任务上的表征能力尚有不足，但在图像生成等场景中展现出的推理加速效果验证了线性化路径的工程可行性。为缓解线性化带来的近似误差，Peng等（2021）提出随机特征注意力（RFA），利用随机傅里叶特征近似Softmax核，并引入时间衰减机制以增强局部建模能力，在效率与精度之间取得了更稳定的平衡。为解决核函数近似“静态映射”的局限，Meng等（2025）提出的线性注意力（PolaFormer）通过将查询-键对显式分解为同号和异号两个独立流，解决了符号信息丢失的问题，使模型聚焦于关键区域，在保持线性复杂度的同时显著提升了表达的灵活性。近期，Beck等（xLSTM，2026）通过引入矩阵记忆单元与指数门控机制，从记忆容量扩展和动态调节两个维度对线性循环模型的性能边界进行了系统分析。与依赖核函数近似的方法不同，该工作聚焦于记忆结构与信息流控制机制的改进，提升了线性模型在大规模任务中的表达能力与训练稳定性。

除了从核函数或递归结构内部进行改进外，部分研究开始从更宏观的序列建模视角出发，探索新的基础算子或新型计算范式以减少对标准自注意力的依赖。Wu等（2025）提出的Token Statistics Transformer（ToST）通过变分最大编码率约简（MCR^2）理论，将注意力机制重新推导为基于词元二阶统计量的线性算子，无需计算词元间的两两相似度，在长序列建模中实现了线性复杂度与强可解释性的统一。Xiao等（2025）提出的DuoAttention则从功能解耦的视角出发，设计了检索头与流式头并行的双路径架构。该方法通过仅对检索头保留全量上下文缓存，而对流式头采用轻量化的固定长度缓存，在维持长上下文能力的同时，显著降低了预填充与解码阶段的内存与延迟开销。

Beck等（2025）指出算法层面的线性化并一定带来系统级的效率优势。若缺乏针对硬件的并行优化，线性模型在实际训练中的速度可能不及经过工程优化的标准注意力实现。为此，他们提出Tiled Flash Linear Attention，通过块内切片与分层并行策略提升GPU上的算术强度。另一方面，当上下文长度扩展至百万级时，即便复杂度为线性，也可能产生大量冗余计算，这使得动态稀疏机制成为重要的补充方向。Li等（QuickLLaMA，2025）采用查询感知的上下文筛选策略，在推理阶段动态选择与当前查询高度相关的记忆块，从而减少无效计算。这一机制将效率优化从公式层面的线性化拓展至信息选择层面，为超长文档理解与多轮对话场景提供了可扩展的实现路径。

通过核函数近似、算子创新和硬件适配的动态稀疏化，高效注意力策略通过多种技术手段有效地将模型在长序列及常规序列上的计算复杂度从二次降至线性或次线性。这些创新不仅减少了推理过程中的计算成本和显存占用，还为Transformer架构的规模化应用解决了性能瓶颈，是实现高效、绿色视觉AI推理的重要技术基础。

3.1.3　线性序列架构

前述方法通过优化注意力机制或压缩模型规模，在一定程度上缓解了计算复杂度问题，但尚未突破序列建模范式本身的效率边界。从更宏观的视角来看，序列建模长期面临训练效率与推理效率之间的根本性矛盾：以Transformer为代表的注意力机制凭借并行训练能力和全局建模优势占据主导地位，但其自注意力的平方级复杂度导致长序列推理时延

高、显存占用大，限制了在资源受限场景下的部署；而以RNN、LSTM为代表的循环神经网络虽具备线性推理效率和恒定显存占用，适用于流式处理，却因时间步间的强依赖而无法并行训练，难以支撑大规模模型的构建。

状态空间模型（state space models，SSM）的引入为序列建模提供了一种新的技术路径（Gu等，2021）。该方法将序列建模任务重构为连续动态系统的离散化过程，在理论上兼顾了循环神经网络在线性推理阶段的效率优势与Transformer在并行训练方面的能力，从而在一定程度上有助于缓解训练效率与推理效率之间的矛盾。SSM在保持较低计算复杂度的同时，实现了对长序列的有效建模，成为当前长序列建模研究中的重要方向之一。

在长序列建模中，如何在保持线性复杂度的同时避免信息随时间衰减，是状态空间模型的一个核心问题。Gu等（2022）通过引入HiPPO框架对连续时间状态方程进行结构化参数化，构建了结构化状态空间序列模型（structured state space sequence model，S4）。该模型通过将状态转移矩阵分解为正规矩阵与低秩矩阵之和，实现了对状态空间模型的高效计算，首次在长序列基准上取得了与Transformer相当的性能。在此基础上，Gu和Dao（2024）提出的Mamba架构引入了选择性扫描机制将SSM参数设计为输入的函数（如图6右所示），使系统能根据当前Token动态决定信息的保留或遗忘，从数学上揭示了Transformer注意力与SSM的对偶性，并配合硬件感知的并行算法与I/O优化，在训练速度和推理显存上实现了对同规模Transformer的超越。

为突破单卡显存对长序列建模的限制，Dutt等（2026）针对状态空间模型设计了高效的张量并行策略，通过优化SSM状态块的分片机制与通信协议，采用量化All-Reduce技术降低同步开销，解决了选择性扫描在多GPU集群中的通信瓶颈问题。此外，Li等（2025）提出了MaIR模型，针对原有Mamba扫描方式容易破坏图像局部特征的问题，他们设计了一种嵌套S形扫描策略。这种新方法让Mamba在遍历图像时，既能看清局部细节，又能保持整体连续性，改进了Mamba在图像恢复中的表现。在底层视觉图像处理的类似延伸中，Chen等（2026）提出了Retinex-rawmamba，将Mamba应用于低光照Raw图像增强，有效协同了去马赛克与去噪过程。此外，针对红外与可见光图像融合，Yang等（2026）利用Mamba架构的线性复杂度优势显著降低了推理能耗，证明了该范式在高效多模态感知任务中的潜力。

随着SSM在空间结构感知建模方面的局限性被逐步认识，混合架构设计成为提升其表达能力的主流路径。Lenz等（2025）通过交替堆叠Mamba层与稀疏注意力层，在保持线性效率的同时增强了模型对全局信息的聚焦能力。Dong等（2025）针对人体姿态序列设计了一种时空图Mamba模块PS-Mamba，利用双向扫描机制捕捉关节间的空间依赖与时序动态，缓解了长视频中动作连贯性丢失的问题。Jin等（2025）面向LiDAR点云提出UniMamba框架，采用组高效扫描机制统一建模空间与通道维度的表示，避免了传统序列化处理对三维结构信息的破坏，实现了高精度、低延迟的目标检测。此外，在复杂视觉场景的感知挖掘方面，Zhang等（2025）将Mamba与胶囊路由机制相结合，有效建模了伪装目标检测任务中的局部与整体关系。

通过构建结构化参数空间并结合选择性扫描策略，状态空间模型在兼顾推理阶段线性复杂度的同时，有效地解决了训练过程中的并行化瓶颈。随着混合架构的兴起及其在垂直领域的广泛应用，这一建模范式正逐步克服在全局依赖建模方面的不足，在长时序预测、三维感知等任务中展现出良好的适应性。

3.2 基于模型加速推理的方法

模型轻量化技术主要从模型结构设计的角度出发，通过压缩参数量与计算复杂度实现推理效率的提升。然而，对于已部署的大规模模型（如扩散模型、大语言模型），重新设计架构或调整参数的成本较高，难以在实际应用中灵活采纳。针对这一问题，推理加速技术则期望在不改变模型权重或仅进行少量微调的前提下，通过优化采样算法、改进计算调度等方式，提升模型的运行效率。

3.2.1 单/少步采样策略

扩散模型自诞生起，凭借着卓越的生成质量和多样性超越了生成对抗网络和变分自编码器等方法成为图像、视频和音频等模态数据生成的主流范式。受限于马尔可夫链的步进式去噪架构，该类模型在推理阶段往往需执行数百乃至上千次的迭代采样，方能完成单次样本生成。这种多步递进式采样机制虽然有助于提升生成质量，但也带来了较高的推理

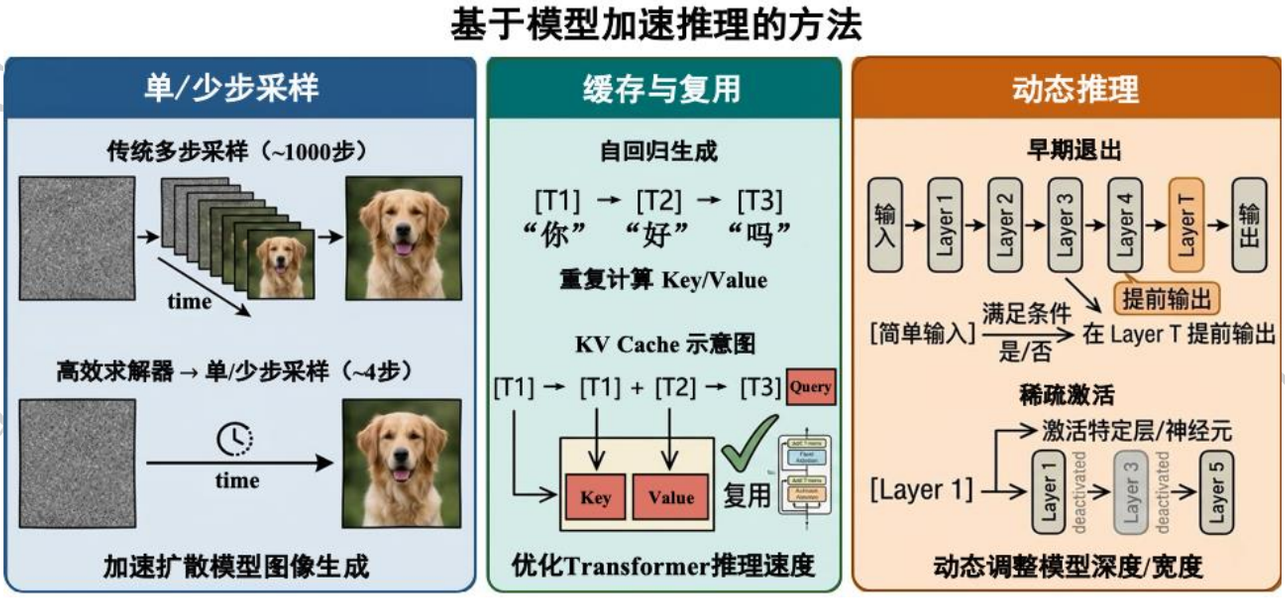

图7 （左）单/少步采样策略示意图；（中）缓存与复用策略示意图；（右）动态推理模型策略示意图

Fig. 7 （Left）Illustration of one/few-step denoising in diffusion models；（middle）Illustration of cache and reuse strategies in transformers；（right）Illustration of dynamic inference in neural networks.

表2 扩散模型采样加速方法与所需步数。

Table 2 Sampling Acceleration Methods for Diffusion Models and Sampling Steps.

方法	步数
DDIM	20～50步
DPM-Solver	10～20步
递归蒸馏方法	4～8步
LCM	4～8步
对抗扩散蒸馏方法	2～4步
CM	1步
分布匹配蒸馏 (DMD)	1步
Rectified Flow+ ODE Solver	1～4步

时延，限制了模型在实时交互与低延迟应用场景中的部署。因此，许多研究学者关注采样过程的加速，旨在保证生成质量的前提下压缩迭代步数以实现推理过程加速。表2分析了本节讨论的扩散模型采样方法及所需采样步数。

早期加速工作致力于在不重新训练模型的前提下，通过改进微分方程（ODE/SDE）求解器来减少截断误差。Song等（2021）提出的DDIM中，构建了非马尔可夫的前向过程，推导出了确定性的采样轨迹，允许用户在时间步上进行“跳跃”采样，首次实现了在较少步数（20～50步）下的高质量生成。随后，Lu等（2022）提出了DPM-Solver，这是一种专为扩散概率模型设计的高阶常微分方程求解器。通过利用扩散模型得分函数的解析性质，DPM-Solver能够以二阶甚至三阶精度逼近去噪轨迹，将高质量采样所需的步数进一步稳定在10～20步，且无需额外训练。尽管求解器优化显著提升了效率，但其性能下限仍受限于预训练模型的轨迹曲率，难以突破至单步生成。

为了突破求解器的理论极限，研究者转向直接学习从噪声到数据的短路径映射，将多步去噪过程压缩为极少的函数调用。Salimans等（2022）构建递归训练框架，将教师模型的N步去噪过程蒸馏为学生模型的N/2步，并通过多轮迭代逐步压缩采样轨迹，使模型在4～8步采样条件下仍能保持较好的生成质量。Sauer等（2024）提出的对抗扩散蒸馏引入了对抗训练机制，利用定制的判别器网络强制单步或少步生成器的输出分布与真实数据分布对齐。与传统知识蒸馏不同，对抗扩散蒸馏不关注中间特征的匹配，而是直接优化生成结果的感知质量，成功在2～4步内实现了与百步迭代相当的生成效果。然而，这类基于对抗的方法往往面临训练不稳定及模式坍塌的挑战，限制了其进一步压缩至单步的能力。

近期的许多研究通过重构生成轨迹的几何结构或引入严格的一致性约束，从根本上解决了多步迭

代问题，实现了无需对抗训练的稳定单步生成。

在轨迹重构方面，Lipman等(2023)提出的流匹配(flow matching)框架，以及Liu等(2023)提出的Rectified Flow，通过构建从噪声分布到数据分布的最优传输路径，利用“重整(rectification)”技术将原本弯曲的随机微分方程轨迹拉直为线性路径。这种直线化特性使得简单的欧拉求解器仅需极少步即可精确逼近生成结果。Esser等(2024)进一步在Stable Diffusion 3中规模化应用了Rectified Flow Transformers，实现了质量与速度的双重突破。与此同时，Song等(2023)提出的一致性模型(consistency models，CM)开辟了另一条路径。一致性模型通过施加“自一致性”约束(即任意时间步t的输出应直接指向最终结果x_0)，训练模型直接从噪声映射到数据，无需模拟整个扩散过程。Luo等(2023)进一步推出了潜在一致性模型(LCM)，将一致性约束应用于潜在空间，实现了4～8步的高清图像生成，并被广泛集成于开源社区。针对一致性模型训练偏差问题，Song等(2024)通过移除教师网络的指数移动平均，并辅以Pseudo-Huber损失函数、对数正态噪声调度及课程学习策略，解决了原有方法的训练偏差问题，大幅提升了单步生成的保真度。此外，Yin等(2024)提出的分布匹配蒸馏(DMD)，通过直接匹配生成分布与目标分布，解决了单步生成中的多样性缺失问题。近期，So等(2025)提出了Picard Consistency Models通过引入相位一致性约束与对抗损失，显著加速了Picard迭代的收敛速度，减少了生成步骤数量，同时保持了与原始模型一致的输出质量。

扩散模型的采样算法经历了从数值求解器改进到生成轨迹重构的演进过程。基于流匹配、一致性模型等技术的发展，当前采样步数已可压缩至单步或极少数步，显著提升了生成速度。这一进展在降低推理延迟的同时，也使生成式模型逐步从追求单一生成质量，向兼顾生成效率与保真度的方向演进。

3.2.2　缓存与复用策略

如图7中所示，在基于Transformer架构的自回归生成中，在每次解码迭代中，模型均需度量当前查询向量(Query)与既有历史词元键向量(Key)的相似度得分，并以此为权重对相应的值向量(Value)执行聚合操作。为了避免对历史词元的Key和Value进行重复计算，通常会在每一步推理时缓存对应的Key-Value对(即KV Cache)，从而将时间复杂度从二次降为线性，显著提升推理速度。在此基础上，相关研究通过更优的显存分配机制、更高效的缓存复用策略，以及按需进行的动态剪枝与卸载技术，在尽量不影响生成质量的前提下缓解显存压力。

为缓解大模型生成式推理中的显存利用率低下问题，Yu等(2022)提出的Orca系统引入了细粒度迭代式调度机制，通过在Token级别进行动态批处理，有效打破了传统静态Batching的等待瓶颈。在此基础上，针对推理过程中因KV Cache动态增长导致的显存碎片问题，Kwon等(2023)提出的vLLM系统进一步引入了分页注意力机制。该方法借鉴操作系统的虚拟内存管理思想，将KV Cache划分为非连续的物理块并通过页表将其映射为逻辑上连续的序列，大幅提升了显存利用率与并发吞吐能力。而针对超长序列所面临的单卡显存墙，Li等(2024)提出的DistFlashAttn则从分布式训练的维度展开了显存高效协同优化，通过改进跨节点通信与显存布局，显著提升了长上下文场景下的系统并行训练效率。

当序列长度超出显存容量限制时，必须对KV缓存进行选择性丢弃。早期的滑动窗口策略往往采用简单的历史丢弃机制，容易造成长程信息缺失。Xiao等(2024)发现Transformer仅需保留初始词元与局部滑动窗口即可维持生成稳定性，这一发现为StreamingLLM提供了理论支撑，实现了无限长文本的低成本流式生成。为进一步在有限显存中保留全局关键信息，Zhang等(2023)通过动态评分识别并保留对当前生成贡献最大的关键词元，同时淘汰冗余信息。在此基础上，Li等(2024)提出的SnapKV通过观察窗口提前识别后续生成中可能被关注的关键KV，显著提升了模型在长上下文检索任务中的准确率。Tang等(2024)提出的Quest算法则引入查询感知的动态稀疏机制，根据当前查询实时选择参与计算的关键KV，在降低显存占用的同时减少了计算开销。此外，这类缓存稀疏与检索机制正逐步向多模态领域延伸。例如，Man等(2025)针对超长视频理解提出了自适应跨模态显存缩减方法(AdaCM^2)，而Di等(2025)则通过上下文视频KV缓存检索机制，有效支撑了流式视频的问答推理。

在极端资源受限场景下，即使经过模型剪枝等优化，KV缓存仍可能超出单块GPU的显存容量。此时，显存卸载技术通过将KV缓存的部分数据迁移至CPU内存或NVMe存储，以扩展可用存储空间，

从而突破硬件限制。Sheng等(2023)提出的FlexGen设计了一套CPU-GPU协同交换策略,通过流水线机制掩盖I/O延迟,使得在单张消费级显卡上即可运行数十亿参数规模的模型推理。针对卸载过程中I/O访问成为性能瓶颈的问题,Lee等(2024)提出的InfiniGen引入了动态预测与智能预取机制,通过推测未来生成步骤所需的关键KV缓存,实现按需加载与预取,提升了生成式推理的整体效率。

综上,KV缓存的优化研究通过引入分页式管理、前缀共享机制以及动态剪枝策略,有效缓解了自回归生成过程中的显存压力。在此基础上,显存卸载技术进一步拓展了存储层次,使得在有限硬件资源下支持百万级词元的超长序列推理成为可能,提升了生成式推理系统的吞吐上限与资源利用效率。

3.2.3 动态推理模型

常规的神经网络多基于静态计算图构建,即模型的计算路径对所有的输入样本在推理时是固定不变的。如图7右所示,对于简单样本,完整的计算路径可能导致不必要的资源开销;而对于复杂样本,固定的网络深度又可能限制模型的表达能力。动态推理(dynamic inference)旨在打破这一僵化模式,通过引入条件计算机制,使模型能够根据输入样本的复杂度、内容特征或上下文状态,自适应地调整计算深度、宽度、粒度或网络结构,从而实现"简单样本快算、复杂样本精算"的绿色推理目标。本节将从动态深度、稀疏宽度、动态粒度及结构重参数化四个维度,系统梳理该领域的最新进展。

动态深度策略允许模型在中间层提前输出结果,避免不必要的深层计算。早期工作如Kaya等(2018)提出的Shallow-Deep Networks奠定了基于置信度的早退理论基础,随后Schuster等(2022)的CALM将其成功应用于大语言模型,通过设定置信度阈值实现自适应退出。近期研究进一步针对长上下文与多模态场景进行了深度优化,He等(2025)提出了自适应子层跳过机制,专门解决长上下文LLM推理中的冗余计算问题;Lu等(2025)深入分析了具身智能体的早退行为特征,为更精准的行动决策提供了理论依据;Bajpai等(2025)则将早退扩展至视觉-语言模型(VLM),显著提升了多模态任务的推理鲁棒性。此外,Xu等(2025)从体系结构角度出发,将早退机制与投机采样(speculative decoding)相结合,提出了Speculative Early Exiting,在系统层面进一步挖掘了动态深度的加速潜力。

混合专家模型(mixture of experts, MoE)通过稀疏激活机制,实现了"大参数量、低计算量"的突破。其核心思想是将网络宽度动态化,每步仅激活部分专家(experts)参与计算。Fedus等(2022)提出的Switch Transformer确立了Top-1路由的稀疏MoE范式,证明了其在万亿参数模型扩展上的有效性;Clark等(2022)进一步建立了路由语言模型的统一扩展律(Scaling Laws),为MoE的设计提供了理论指导。当前,工业界实践已走向极致高效,Liu等(2024)推出的DeepSeek-V2通过精细化的多头潜在注意力(MLA)与高稀疏度MoE架构,在保持强劲性能的同时大幅降低了训练与推理成本,证明了稀疏宽度策略在大模型时代的核心价值。

除了调整网络层级,动态推理还可细化至词元粒度,通过丢弃、合并或路由冗余Token来减少计算量。在视觉领域,Rao等(2021)的DynamicViT首创了基于重要性评分的动态词元剪枝,而Bolya等(2023)提出的Token Merging(ToMe)则通过合并相似词元实现了极速推理。这一思想近期被成功迁移至长文本与多模态场景:Fu等(2024)提出的LazyLLM专门针对长上下文LLM推理,动态识别并剪枝无关词元,显著降低了显存与计算开销;Zeng等(2025)则展示了自适应Token跳过(Adaptive Token Skipping)机制在视觉-语言模型中的可扩展加速能力,证明了细粒度动态调整在复杂架构中的通用性与高效性。

与前三种运行时动态调整不同,结构重参数化通过"训练时多分支、推理时单路"的策略,实现了无损加速。Ding等(2021)的RepVGG开创了这一范式,通过结构重参数化将多分支CNN等价合并为单路VGG风格网络;Ding等(2022)的RepLKNet将其扩展至大卷积核设计,进一步提升了性能。Termöhlen等(2023)提出的ReVT将重参数化技术迁移至Vision Transformer,在不增加推理开销的前提下引入了局部归纳偏置,有效提升了模型在语义分割等任务中的领域泛化能力。最新的前沿探索Huo等(2026)提出了REPSPEC,将结构重参数化与投机采样结合,设计了重参数化模型,探索了结构优化在生成式加速中的新边界。

作为一种自适应的感知范式,动态推理依托条件执行机制,赋予了模型依据输入数据难易程度灵

活配置前向计算链路的能力。无论是通过早退机制提前结束简单样本的前向传播，还是通过稀疏激活机制动态选择参与计算的参数，这种“按需计算”的设计思路有望为构建资源高效的视觉AI系统提供了更灵活的实现方式。

4 迭代节能型AI方法

随着模型规模与应用场景的持续扩展，模型的生命周期成本逐步从单次训练转向持续的更新与维护。传统的“全量数据重新训练”或“面向多任务建立独立模型副本”的迭代策略，不仅会导致算力、存储及维护负担的剧增，更会在连续更新时触发灾难性遗忘。该现象表现为模型在拟合新任务特征时，对历史数据分布的泛化能力发生严重退化，进而必然产生重新训练与知识对齐的额外损耗。在资源受限与可持续部署的背景下，如何在控制计算与存储开销的同时实现模型的稳定迭代，已成为模型工程化落地的重要议题。围绕这一目标，本章节重点从持续学习与参数高效微调两条关键技术路径展开讨论。

4.1 基于持续学习的方法

持续学习（continual learning，CL）旨在研究模型如何在获取新知识的同时保持对已有任务的性能稳定，在连续任务或数据流场景下实现知识的有效积累与可持续适应。Lyu等（2025）指出，持续学习通过平衡模型的“可塑性”与“稳定性”，使系统能够在不经历灾难性遗忘的前提下处理非平稳数据流。从模型迭代成本的角度看，该方法使模型能够在一次部署后持续学习新数据，减少重复训练带来的计算与时间开销。如图8所示，本节将从基于正则化策略、基于回放策略与基于架构增量策略三种方法展开。

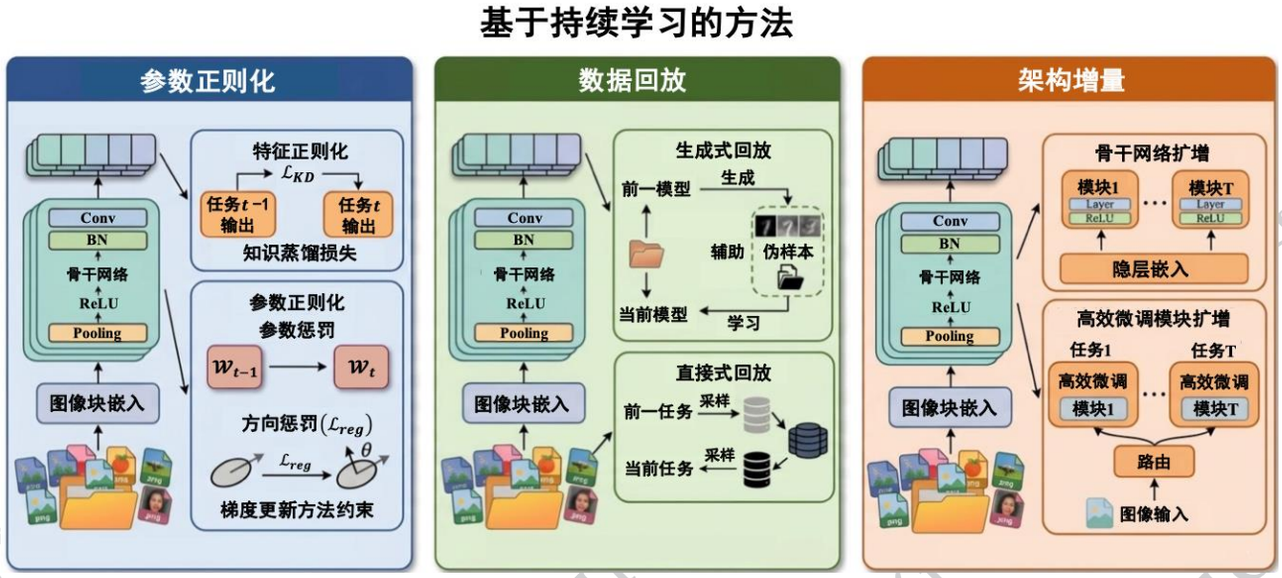

图8 持续学习三种核心方法的技术路线对比

Fig. 8 Comparison of technical routes for three continual learning methods

4.1.1 参数正则化策略

在持续学习框架下，基于正则化的方法（optimization-based）通过修改优化目标来缓解灾难性遗忘。该类方法不依赖历史数据回放，也不改变模型结构，而是在损失函数中引入约束项，限制新任务训练过程中对旧任务重要参数的偏移。其基本思想是：旧任务知识编码于参数空间的特定解附近，通过对关键参数施加约束，可以在一定程度上维持历史性能。

前期的研究工作注重刻画参数的重要性。例如，Kirkpatrick等（2017）提出的Elastic Weight Consolidation（EWC）利用Fisher信息矩阵的对角近似估计参数对旧任务的贡献，并通过二次正则项加以约束。Aljundi等（2018）提出的Memory Aware Synapses（MAS）则依据参数扰动对模型输出变化的敏感度定义重要性，使方法能够适用于无监督场景。近期，

Zhu等(2025)在参数空间中为不同任务构建了高度解耦、低干扰的子空间约束,从而在参数层面实现了有效的任务隔离。沿着参数空间约束的思路,Cheng等(2025)提出通过自适应正交投影来约束模型参数的更新方向,在数学层面有效实现了持续学习中可塑性与稳定性的平衡。与此同时,Wang等(2025)面向测试时自适应目标检测场景,提出敏感度引导的剪枝策略,通过正则化约束筛选并精简冗余参数,在适配域外数据分布、缓解灾难性遗忘的同时,显著提升了模型的自适应推理效率。

除参数空间约束外,函数空间正则化提供了另一条路径。不同于参数层面约束,Li等(2017)提出的LwF算法,是知识蒸馏技术落地于持续学习的典型方案。该方法借助最小化新旧模型预测分布的KL差异,约束参数迭代过程,保障模型对历史任务的预测结果基本不变。近期,这一思路被拓展至结构化数据场景。Jhajj等(2025)针对知识图谱持续演化问题,对传统EWC进行改造,提出结合实体关系拓扑结构的正则化策略,在估计参数重要性的同时引入结构信息,以缓解动态知识注入过程中对既有推理路径的干扰。近期的研究进一步关注模型压缩与表示分解。Yang等(2025)提出基于贝叶斯压缩的持续学习框架,将潜在表示划分为共享与私有成分,并通过变分推断压缩冗余参数,以降低存储与计算开销,同时减弱任务间特征干扰。函数空间正则化策略展现了更强的任务适应性。例如,Lee等(2025)通过插值基础模型、历史模型和当前模型的分类头权重,在权重空间进行集成,并结合增强数据的知识蒸馏正则化损失项,在函数空间层面维持了模型的行为一致性。Gong等(2024)针对持续分割任务,提出了一种解耦物体性学习与类别识别的方法,通过特征蒸馏强化模型对物体边界的感知能力,在特征空间层面施加约束,确保模型在更新时保持对物体形状的识别能力。此外,针对任务无关的类增量学习场景,Wu等(2025)探讨了如何通过正则化手段约束特征表达,从而有效引导模型克服新知识学习过程中的语义漂移现象。

基于正则化的方法通过在优化目标中引入参数重要性约束或函数空间稳定性惩罚,在无需依赖历史数据的情况下有效缓解了灾难性遗忘问题。这类参数层面的调控机制,为模型在多域适应及持续迭代等任务中维持既有能力提供了可行的技术路径。

4.1.2 数据回放策略

基于回放策略的方法(rehearsal-based)通过在新任务训练过程中显式混合少量历史数据(真实样本或生成样本),缓解灾难性遗忘。其基本思路是通过重现旧任务的数据分布,使模型在参数更新时同时受到新旧样本的约束,从而减缓性能退化。与正则化方法通过修改优化目标不同,回放策略保持原有损失函数形式不变,而是在数据层面构造联合分布。如图8中所示,该框架主要可分为两类,一是直接回放存储的真实历史样本;二是通过生成模型或历史特征合成伪样本进行隐式回放,两者均通过与当前样本混合实现联合学习。

真实样本回放技术的核心机制在于:筛选并保存以往任务里的部分原始图像与标注资料,待到开展新任务时,再将这些保留的旧数据与新获取的数据合并在一起进行联合优化。例如,Yin等(2025)提出了增强实例回放方法,在持续语义分割任务中,通过存储和回放旧类别的实例块,确保旧类别的特征能够稳定学习。Kang等(2025)面向持续学习任务,创新性引入学习与休整协同调控思路,通过动态调度回放间隔与自监督特征挖掘,合理规划模型训练节奏,缓解长期迭代产生的性能衰减,提升持续学习整体稳定性。该方法通过自监督学习获取任务视图,从而智能规划模型在学习和回放状态之间的转换,以提升持续学习的整体效率。在三维视觉任务中,Thengane等(2025)提出了CLIMB-3D框架,该框架针对不平衡的3D实例分割任务,在经典回放策略的基础上进行了扩展,解决了点云数据中类别不平衡带来的问题。此外,Cheng等(2024)将回放策略应用于图像恢复任务,实现了连续多种恶劣天气的有效去除。

存储真实样本虽然对解决知识遗忘问题直接有效,但不可避免会带来隐私泄露、存储开销和数据版权等隐患。因此,基于隐式样本或生成样本的回放方法逐渐得到了广泛关注。这类方法不直接保存原始数据,而是通过生成模型模拟历史任务的数据分布,合成具有代表性的伪样本,或在特征空间中进行知识回放。Ye等(2025)提出了一种基于动态可扩展记忆分布的持续学习方法,利用生成模型建模和回放历史任务的数据分布,从而实现无任务边界的持续学习。Zhang等(2025)通过扩散模型合成高质量的历史任务图像,成功缓解了新旧任务之间的领

域差异，有效减轻了遗忘。与此同时，Zhu等(2025)针对持续图像分割任务，提出了基于视觉查询的重放机制。该方法通过存储和重放高度抽象的视觉查询特征，实现了对旧任务知识的隐式保留，显著降低了存储成本并提升了模型的可塑性。

综上，数据回放策略是实现高效稳健持续学习的重要技术。通过在训练中借助新老数据的联合优化，模型得以在吸收新信息的同时锚定旧有记忆，从而从根本上抑制了灾难性遗忘现象的发生。该策略已广泛应用于目标检测、语义分割和图像识别等持续迭代任务中。

4.1.3 架构增量策略

基于架构增量的方法(architecture-based)通过在结构层面减少任务间参数共享以缓解灾难性遗忘。其核心假设在于，当多个任务共用同一组参数时，梯度更新可能产生冲突；为此，该类方法通过扩展网络结构或划分相对独立的参数子空间，为新任务分配专用容量，从而降低任务间的相互干扰。与正则化方法通过对参数更新施加约束、回放方法通过重现旧任务数据的方式不同，架构增量策略主要在模型结构层面实现任务隔离。

Rusu等(2016)提出渐进式神经网络(progressive neural networks，PNN)，该方法用于缓解持续学习中的灾难性遗忘问题。该方法在引入新任务时新增一列网络结构，并冻结已有任务对应的参数。不同任务之间通过横向连接共享中间层表示，使新任务能够利用先前学习到的特征。由于各任务参数彼此独立，干扰现象得到缓解，但模型规模会随着任务数量增加而持续增长，从而带来额外的存储与计算成本。为控制模型容量，Mallya等(2018)提出在固定网络中通过“训练—剪枝—掩码”流程为不同任务分配稀疏子网络，并用二值掩码锁定已分配参数。Serra等(2018)提出HAT进一步引入可学习的任务掩码，通过硬注意力机制动态选择神经元。这类方法在参数层面实现了任务隔离，但在大规模模型中直接剪枝或复制网络的代价较高。此外，Yang等(2024)将渐进式适应与剪枝策略应用于显著性预测，有效提升了模型在多图像域中的增量学习能力。

在大语言模型场景下，整体复制或剪枝数十亿参数并不现实，因此研究重心转向提示空间的扩展。Wang等(2022)提出提示池(prompt pool)机制，其维护一组可学习提示向量，在新输入到来时基于特征相似度检索相关提示拼接至输入端，而主干模型保持冻结。若现有提示无法覆盖新任务分布，可向提示池中添加新条目。该方法通过限制更新范围在提示参数内，实现了较为轻量的任务隔离。Jayasuriya等(2025)进一步提出子空间感知提示适配，通过控制提示向量在特征空间中的分布，使新增提示在语义子空间上与已有提示保持相对分离，从而减少提示间干扰。

随着混合专家模型(mixture-of-experts，MoE)的发展，动态路由成为架构增量的重要实现方式。一些研究，如Wang等(2025)、Zhang等(2025)探索在检测到新任务分布时动态实例化专家模块，或通过强化学习优化路由策略，使不同任务流量分配至不同专家，从而降低共享参数带来的冲突。此外，Li等(2025)提出Dynamic Integration of Task-Specific Adapters，通过维护适配器库，并在推理阶段通过动态网关聚合多个适配器输出，实现对不同任务模块的组合调用。这种方式避免为每个任务保存完整模型，同时保留结构层面的相对隔离。

除显式结构扩展外，也有工作通过参数空间中的正交约束实现隐式隔离。例如，Hu等(2024)针对长时持续学习中严格参数隔离易导致模型欠拟合的局限，构建了任务感知正交稀疏网络。该方法打破了传统的“旧知识保留”与“新知识学习”的二元网络划分，在稀疏子网络中划分出第三个参数区域，专门用于探索新旧任务间的语义共性。通过结合锐度感知的正交学习机制，模型能够在这个专属区域内动态搜索并优化跨任务的共享参数。这一机制使得参数高效迁移在减少任务干扰、避免灾难性遗忘的同时，有效挖掘了持续学习中的共享知识。

综上，架构增量策略通过扩展网络分支或引入动态路由机制，在结构层面实现任务间干扰的隔离。在大模型背景下，这一思路逐步向轻量化方向发展，通过提示池扩展与正交子空间投影等方式，在控制参数规模的前提下实现多任务知识的持续积累。

4.2 基于高效微调的方法

预训练大模型在海量数据上习得的泛化能力，使其成为众多下游任务的基础架构。然而，当面向具体场景或任务进行适配时，若从零训练或对全参数进行更新，将带来显著的计算与存储开销，难以支持多任务场景下的高效迭代。参数高效微调技术通过冻结主干网络、仅优化少量额外参数，在保持模型

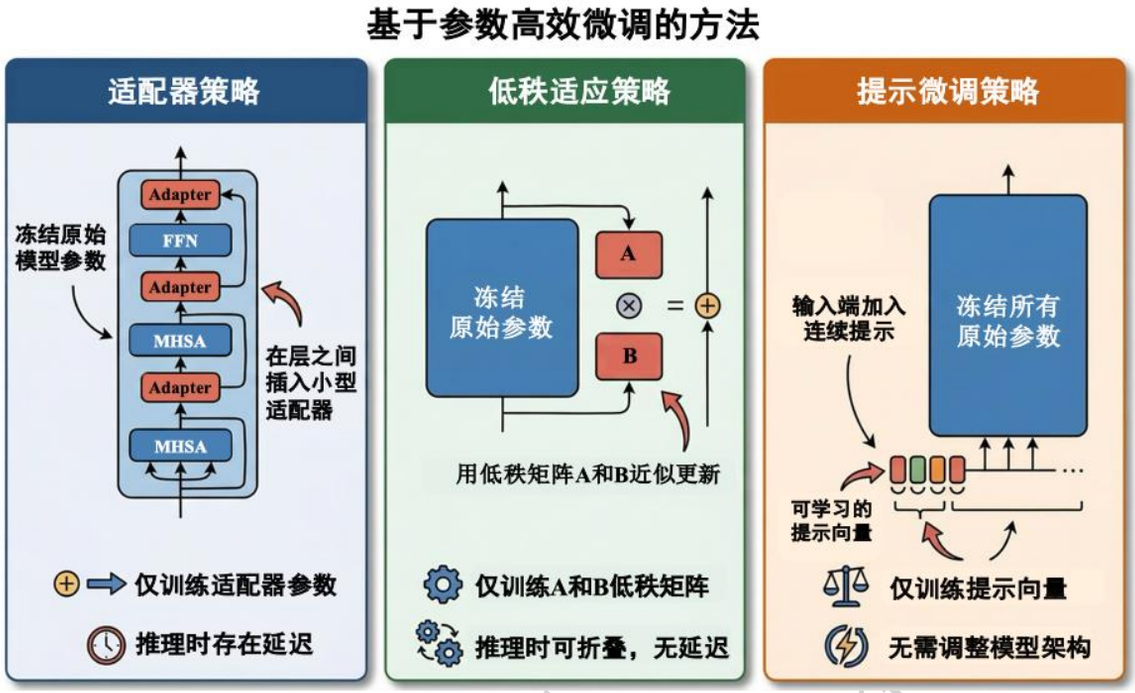

图9　三种主流参数高效微调方法的结构对比

Fig. 9　Structural comparison of three mainstream parameter-efficient fine-tuning (PEFT) methods

原有泛化能力的同时，大幅降低任务适配过程中的计算成本与存储开销。如图9所示，当前主流的参数高效微调技术主要围绕三类架构设计展开：适配器插入式微调（adapter）、低秩分解式微调（LoRA）与提示学习式微调（prompt tuning），本节将围绕这三类主流方法进行梳理介绍。

4.2.1　适配器策略

适配器方法是参数高效微调领域较早形成体系的代表性思路之一，其核心思想是在预训练模型的现有层之间（通常在多头注意力层与前馈神经网络层之间）插入轻量级的可训练模块（adapter），同时冻结主干网络的所有参数。这种设计使得模型在面对新任务时，无需更新庞大的预训练权重，仅需优化极少数的新增参数即可完成适配。

适配器方法的奠基性工作由Houlsby等（2019）提出。他们设计了标准的Bottleneck Adapter结构：包含一个下投影层（将高维特征映射到低维空间）、一个非线性激活函数（如ReLU或GELU）以及一个上投影层（恢复至原始维度）。这种编码器-解码器结构有效减小了adapter模块的参数量，并且能有效捕捉任务特定的特征表示。由于主干参数被冻结，优化器仅需维护极少量的梯度状态，显著降低了微调过程中的显存峰值，使得在资源受限设备上对模型进行快速定制或迭代新任务成为可能。

为进一步压缩适配器的参数量并提升推理效率，后续研究在结构设计上进行了深度优化。Yin等（2025）提出的Mona适配器引入了多认知视觉滤波器与特征分布优化技术，并辅以可学习的尺度归一化层，在仅需调整约5%参数的条件下，在目标检测、语义分割等视觉任务上实现超越全参数微调的优异性能。Xie等（2025）提出的Mamba-Adapter将状态空间模型引入适配器内部，该设计巧妙地利用状态空间模型在长距离依赖建模上的优势，弥补了适配器感受野受限的短板。针对模型在复杂真实场景下的泛化效率，Bi等（2025）的工作NightAdapter提出了频域适配策略，在推理阶段通过重点适配敏感频带来实现对夜间等复杂光照条件的高效泛化。在参数选择策略上，Zhang等（2025）的CLIP-AST方法进一步优化了适配器的微调效率，采用自适应学习率机制动态识别模型中最关键的参数进行调整，避免了冗余更新。Tian等（2025）则围绕动态结构化适配器概念提出了MetaPEFT框架。该方法通过元学习策略使模型能够根据具体任务和数据分布，自动搜索并生成最优的适配器拓扑结构。

近期的研究也在推动适配器理念与结构重参数化的深度融合，旨在构建“训练时动态增强、推理时

零开销融合”的隐式适配机制。这类方法在微调阶段引入额外的动态参数或复杂算子作为“临时适配器”以增强表征能力，而在推理时通过数学等价变换将其无损内化至主干权重中，从而减小推理阶段的架构差异与额外开销。Lou等(2025)提出的Over-LoCK将上下文混合动态卷积作为一种轻量级动态适配器插入微调流程。该方法仅在训练时激活以增强复杂场景的表征能力，推理时则将其数学等价合并至主干静态卷积中，实现了零额外计算开销的高效部署。类似地，Hu等(2025)提出的TEAFormer模型引入可重参数化的滑动窗口组件作为隐式适配器，在冻结主干的前提下高效捕捉局部细节，并在部署前通过结构重参数化将其等价为标准线性单元。

适配器策略的灵活性、可组合性和可扩展性优点，使其能够有效应用于多模态和多任务场景。例如，在开放词汇目标检测任务中，Fu等(2025)提出的LLMDet方法通过微调轻量适配器，使得模型与大语言模型的语义空间对齐，有效提升了开放词汇检测中的泛化能力。在医学影像分析任务中，Tian等(2025)提出的连续模态参数适配器，根据输入图像的模态动态生成专属参数，使得同一视觉网络能够自适应处理不同模态的医学图像，有效解决了跨模态检索中的领域鸿沟问题。Ye等(2025)则在医学多模态学习中，提出了时间序列和文本主导的并行适配器架构。该架构允许任意模态作为主导，并增强另一模态的特征，从而实现模态间的知识融合与交互，显著提升了任务性能。

综上，适配器方法通过在预训练模型中插入轻量级模块，实现对原始模型能力的快速迭代更新。这种模块化、灵活化的结构设计为多任务共享与跨任务迁移提供了高效的微调技术。

4.2.2　低秩适应策略

低秩适应(low-rank adaptation，LoRA)基于“内在秩缺陷”假设，即预训练大模型在适配下游任务时，权重更新矩阵ΔW具有极低的内在秩。

如图9中所示，Hu等(2022)提出的该方案冻结预训练权重W_0，将权重更新参数化为两个低秩矩阵A和B的乘积($\Delta W = BA$)，矩阵A和B的参数量可以仅占原模型的1%，甚至更少。由于更新量在推理阶段可通过数学等价变换合并回原权重($W = W_0 + BA$)，该方法在不引入额外推理延迟的前提下，实现了与全量微调相当的性能表现。

原始LoRA对所有层采用固定的秩r和统一的学习率，忽略了不同层对任务适配需求的差异性。针对秩分配僵化问题，Zhang等(2023)提出的AdaLoRA引入了奇异值分解(SVD)机制，将低秩矩阵参数化为奇异值三元组，并根据参数重要性评分动态剪枝不重要的奇异值，从而在总参数量预算不变的前提下，自适应地将秩资源分配给关键层。针对收敛速度缓慢问题，Hayou等(2024)提出的LoRA+通过理论分析发现，矩阵A(投影到低秩空间)与矩阵B(投影回高维空间)的梯度范数存在显著差异，因此提出了差异化学习率策略(B的学习率远大于A)，在不增加任何计算成本的情况下显著加速收敛过程并提升模型性能。此外，Wang等(2024)提出的LoRA-GA进一步通过梯度近似技术优化初始化策略，使LoRA的初始更新方向更接近全量微调的全局最优解，从源头上减少了训练迭代次数。

在结构解耦方面，Liu等(2024)提出的DoRA指出，全量微调过程中权重更新通常同时包含幅度与方向两个维度，而标准LoRA难以对二者进行显式解耦。为此，DoRA将预训练权重分解为幅度向量与方向矩阵，并仅对方向部分引入低秩适应，从而在保持参数高效的同时，使模型更新更加接近全量微调的学习行为。在张量化参数微调方面，Veeramacheneni等(2025)提出的CaRA方法针对视觉Transformer中多头注意力权重的高维结构，提出了一种基于张量分解的低秩适应策略，在保持模型表达能力的同时显著减少可训练参数数量。

面对连续任务流或多任务场景时，直接复用单一LoRA更新容易产生任务干扰并导致灾难性遗忘。近期研究开始探索将LoRA模块化，结合参数隔离与共享机制以支持高效迭代。例如，He等(2025)提出CL-LoRA，通过构建任务共享适配器与任务特定适配器的双适配器结构，在无回放的类增量学习场景中同时建模跨任务共享知识与任务特有表示。Liu等(2025)提出的LoRA Subtraction机制通过分离任务漂移分量，提高长任务序列下的稳定性，降低了频繁回退或重新训练的需求。Zhang等(2025)在多模态场景中结合混合专家与增量LoRA模块，实现任务间参数分配与动态路由，进一步减轻迭代负担。这些进展标志着LoRA已从单一的微调工具，演变为支持持续进化、多任务共存的动态架构核心组件。

作为参数高效微调的经典方法之一，LoRA将全量参数更新近似为低秩矩阵的乘积形式，在不引入额外推理延迟的前提下取得了与全量微调相近的性能。后续研究进一步引入动态秩分配、量化协同与参数隔离等机制，使LoRA成为降低大模型迭代门槛、支持在线持续训练的主流方法之一。

4.2.3　提示微调策略

如图9右所示，提示微调（prompt tuning）的核心思想是不修改预训练模型的任何内部权重，也不插入额外的网络层，而是通过构建并优化连续可学习的向量序列（soft prompts），将其作为虚拟词元拼接到输入端或注入到注意力机制中，从而引导模型生成符合下游任务要求的输出。与适配器和LoRA相比，该方法参数量极少，且完全保留了主干模型的原始推理逻辑，特别适用于存储资源极度受限或需严格保持模型行为一致性的场景。

如何提升提示向量与视觉特征之间的语义对齐是一个关键问题。Ren等（2025）提出的DA-VPT通过引入度量学习机制，引导提示向量的分布，使其能够在图像补丁词元与类别词元之间传递语义信息，从而提升视觉Transformer在下游视觉识别与分割任务中的迁移性能。针对生成任务中提示优化难以收敛的问题，Wang等（2025）将提示学习视为双线性映射，通过去相关和方差均衡化技术加速了提示调优的收敛。针对固定层级提示难以适应不同任务需求的问题，Shang等（2025）提出的PRO-VPT通过提示重定位机制动态调整提示向量在Transformer各层中的分布，并结合提示剪枝删除冗余提示，从而实现更为灵活有效的提示微调。此外，Wang等（2024）提出利用下游数据的原型特征来初始化提示向量，显著提升了模型的微调性能。

为了克服浅层提示在复杂任务和中小模型上表现不佳的局限，研究开始关注深度提示与结构化优化。Li等（2025）提出的DPC通过构建双提示协作机制，将提示解耦为基类提示和新类提示，成功解决了基类性能与新类泛化的冲突，提供了一个通用的深度提示微调框架。Deng等（2025）提出的Seq2Time将提示学习与时间感知结合，提升对视频时序结构的感知与建模能力。此外，针对提示初始化敏感性问题，Guo等（2025）通过多模态表示学习策略，利用语义信息引导提示学习，提高了方法在不同视觉架构间的泛化能力。

随着研究的深入，提示微调已经发展成为一种前沿技术，广泛应用到计算机视觉、自然语言处理等任务中。例如，Dai等（2025）提出的PropVG框架，通过多粒度提示在视觉定位任务中实现精确定位；Jiang等（2025）通过视觉与语义提示的协作，提升了广义零样本分类的性能。在通用微调方法层面，Xiao等（2025）提出的ViaPT方法，通过生成实例感知提示，为小样本下的参数高效微调提供了新思路。在更广泛的视觉场景中，提示微调也展现了极强的跨域适应性。例如在复杂目标感知方面，Luo等（2024，2026）提出的VSCode及其改进版VSCode-V2，先后通过2D提示学习与动态双阶段优化，实现了通用的显著性与伪装目标检测；在底层视觉领域，Sun等（2025）结合频率感知学习，提出渐进式提示驱动的低光照图像增强方法；而在医学图像领域，Yao等（2025）通过对视觉语言模型进行提示微调，高效完成了细胞核的实例分割与分类任务。

综上，提示微调通过引入少量可学习提示向量，将模型适配的粒度由“完整网络”转变为“少量提示向量”，使得在同一预训练主干上维护多个轻量级任务适配模块成为可能。这种“主干冻结、提示更新”的参数高效范式，从而帮助大型模型突破硬件条件的壁垒，使得在算力匮乏的场景中依然能够稳健地执行多任务协同与长周期演进。

5　总结与展望

5.1　数据节能型AI方法

尽管基于数据合成与迁移的技术在降低数据采集成本方面取得了显著进展，但面向复杂开放的真实世界，现有技术仍面临诸多挑战。当前主流的数据生成模型主要依赖于对海量真实数据的统计分布进行学习，能够合成视觉逼真度较高的图像。然而，这种基于纯数据驱动的生成机制往往缺乏对现实世界物理规律的内在建模能力，使得生成的图像或视频序列在物理一致性方面仍存在局限。例如，在自动驾驶或具身智能等动态场景中，生成模型合成的视频可能呈现出车辆运动不符合动力学规律、物体遮挡违反光学成像原理、刚体接触部位发生非物理形变等现象。因此，未来研究可重点关注物理规律引导的数据生成技术，探索将偏微分方程、运动学约束及三维渲染引擎等物理先验显式地引入生成模型

的潜在表示与损失函数设计之中。这种融合物理知识与数据驱动的方法，有望使合成数据不仅具备视觉逼真度，同时满足三维几何与物理规律的一致性，从而为在高危垂直领域中以虚拟合成替代真实数据采集、降低数据获取成本与安全风险。

在生成模型的多模态扩展方面，当前基础模型虽已实现输入端的多模态统一，能够通过文本或空间布局引导视觉内容生成，但其生成端仍以单一模态为主。为此，未来研究可面向构建端到端统一的多模态生成基础模型，探索模态统一的特征编码和解码过程机制。通过建立统一的潜在空间，实现对文本、图像、视频及三维结构数据的同步生成与语义对齐。同时，这一发展方向有助于打破多模态数据生成之间的独立设计，为构建时空与语义一致的复杂场景训练数据提供基础支撑。

此外，对于以“零采集成本”为目标的数据迁移技术而言，如何在未知场景下实现有效评估，仍是当前研究尚未充分解决的问题。现有域泛化与开放词汇方法在公开基准上展现了良好的跨域迁移能力，但这些评估通常依赖于已完备采集且精确标注的静态数据集，属于后验式验证。在不接触目标域真实数据的前提下，如何对模型在全新场景中的实际表现进行可信评估，目前仍缺乏有效的理论支撑与度量手段。针对这一挑战，未来研究可关注无目标域先验的跨域评估体系构建。一方面，可探索基于无监督分布差异的度量指标，用于评估生成数据或迁移模型在目标域上的适用性；另一方面，可借助模型内在的不确定性估计机制，对模型在未知场景下的泛化边界进行量化预测。建立科学的无数据依赖的跨域评估基准，将有助于为模型在陌生环境中的部署提供更可靠的理论保障。

5.2 标注节能型AI方法

近年来，自监督与弱监督算法在摆脱高质量标签依赖、从无标注样本及含噪标签中提取有效监督特征方面，展现出了突破性的成效，有效降低了对大规模精细人工标注的依赖。然而，当前主流方法大多针对单一类型的弱监督信号进行设计，例如仅利用噪声类别标签，或仅依靠对比学习中的实例判别任务。这种单一的表征学习机制难以对多源异构的弱监督信息进行有效融合，导致数据中潜在的有效信息未被充分利用。以医疗影像分析或工业质检为例，同一数据集可能同时包含粗粒度标签、带有噪声的历史诊断文本，以及大量无标注的影像数据。现有单一框架难以同时处理并整合这些来自不同来源的监督线索。针对上述问题，未来研究可探索能够统一建模多种弱监督与自监督信息的联合学习框架。通过设计多任务损失函数或基于因果推断的信息融合机制，模型可根据不同监督信号的置信度动态调整其贡献权重。这种多范式统一的表征学习方法，有助于更充分地利用低成本获取的异构数据，在提升模型性能的同时，进一步推动弱监督学习向“零人工标注”的目标迈进。

以视觉-语言大模型为代表的跨模态对比学习，已成为实现零样本泛化与降低标注成本的主流技术路径之一。通过对海量互联网图文对进行对比对齐，这类模型展现出较强的开放词汇识别能力。然而，当前主流的跨模态对齐机制主要停留在图像级与文本描述级的整体语义匹配，缺乏对局部视觉实体及复杂空间关系的精细建模。当将这类基于全局对齐的模型直接应用于目标检测、实例分割或视觉定位等密集预测任务时，模型往往难以将文本中的具体实体准确映射至图像中对应的像素区域。针对这一局限，未来研究可关注提升跨模态对齐的语义细粒度。在无需额外像素级标注的自监督条件下，可探索层级化与结构化的语义对齐方法，借助无监督区域发现、密集视觉提示或细粒度对比损失等机制，推动模型从全局相似度匹配向局部实体特征精准映射演进。这一方向的发展，有望增强多模态模型在像素级感知任务中的适用性，进一步降低密集预测任务对精细人工标注的依赖。

此外，从绿色视觉AI的全生命周期视角来看，自监督预训练虽然为基础模型提供了强大的泛化能力，理论上能够降低下游任务对标注数据的依赖，但其工程化部署仍面临显著的算力成本挑战。因此，未来研究可推动数据标注成本节约与模型迭代成本节约的协同优化，将自监督预训练模型的基础表征能力与参数高效微调技术深度结合。通过发展适配于视觉任务的低秩自适应、提示微调等轻量化机制，使模型在仅更新少量参数的前提下实现对新任务的快速适配。这种数据效率与算力效率的统一，将为可持续、低碳化的视觉AI部署提供更为可行的技术路径。

5.3 推理节能型AI方法

尽管现有研究在降低模型推理成本方面取得了

积极进展，但随着模型规模持续增长与应用场景日趋复杂，当前方法在知识融合效率、结构优化自动化及推理过程自适应能力等方面仍存在进一步探索的空间。在模型轻量化过程中，多教师蒸馏面临知识冲突与信息冗余的挑战。当不同教师模型在特征表达或决策逻辑上存在差异时，简单的知识平均或特征对齐策略可能引入干扰，影响学生模型的学习稳定性。未来研究可探索更具适应性的知识融合机制，例如根据任务相关性对教师知识进行动态选择与加权融合，在蒸馏过程中识别各教师模型对当前任务的贡献程度，从而实现有选择的知识迁移。通过上述方式，有望缓解多教师知识间的冲突，提升学生模型在多任务或多模态场景下的学习效率与稳定性，从而在减少模型推理计算量的同时，有效保持模型推理性能。

在动态推理与模型结构优化方面，结构重参数化通过将训练阶段的复杂结构等效转换为推理阶段的简洁形式，在保持模型功能不变的前提下有效降低了计算开销。这种思路已在多种视觉模型中取得良好效果。然而，现有方法大多围绕特定网络结构进行设计，例如针对卷积网络中的多分支模块进行参数折叠，缺乏统一的理论分析框架与通用化工具，因此难以直接迁移到不同类型的模型架构之中。针对这一问题，未来研究可探索更加通用的结构重参数化机制，通过建立统一的等价变换理论或自动化结构搜索方法，支持在Transformer、状态空间模型等不同架构中的参数合并与计算重构，并系统分析多分支、多尺度结构中的计算冗余与可合并规律。在保证模型功能等价的前提下，这类通用重参数化框架有望进一步降低推理成本，为复杂模型在资源受限环境中的高效部署提供新的技术路径。

此外，神经架构搜索与推理效率优化的结合也是一个可探索的研究方向。近年来，神经架构搜索在自动化模型设计方面取得了显著进展，通过在大规模结构空间中进行系统搜索，能够自动发现性能优良的网络结构。然而，现有多数研究主要关注搜索算法本身的效率以及最终模型的预测性能，往往以准确率作为主要优化目标，而较少在搜索阶段系统性地考虑模型的计算复杂度与推理成本。因此，未来工作可在架构搜索过程中同时优化准确率与推理成本，例如将计算量、参数规模或推理时延作为多目标优化约束，引导搜索算法自动发现兼顾效率与性能的网络结构。这一方向的深入，有助于为绿色视觉AI在边缘端与云端的协同部署提供更灵活的模型生成工具。

5.4 迭代节能型AI方法

近年来，多模态大模型在统一处理文本、图像等多模态信息方面取得了显著进展，通常依赖大规模跨模态数据进行集中式训练，从而学习到具有较强泛化能力的统一表示。然而，大规模数据的集中式训练不仅需要消耗大量计算资源，带来较高的时间与经济成本，并且也难以支持模型在不断引入新任务或新模态数据时的高效迭代。针对这一问题，未来研究可探索将持续学习引入多模态大模型的训练与更新过程，通过构建基于持续学习的迭代训练框架，使模型在接收新任务或新模态数据时能够进行渐进式更新，而无需依赖大规模数据的集中式训练。在这一过程中，需要重点研究多任务与多模态之间的知识冲突问题，例如通过参数隔离、动态模块化结构或知识蒸馏等机制，实现新知识的有效吸收与旧知识的稳定保持。该研究方向有望在降低训练成本的同时，提高多模态模型持续扩展能力。

另一方面，基于提示微调的参数高效微调方法虽然利用少量可学习的提示向量实现模型对下游任务的快速适配，但性能在一定程度上仍依赖于提示向量的初始化方式、长度设置以及插入位置等人工经验，且训练过程中缺乏对提示结构的自适应优化机制，这在一定程度上限制了模型的收敛效率和跨任务泛化能力。针对这一问题，未来研究可探索引入自动化与自适应优化策略，对提示向量的生成与更新过程进行动态调节，例如通过强化学习或元学习方法学习更优的提示生成策略，使模型能够在训练过程中自动调整提示向量的结构、位置或权重分布。通过这种自适应优化机制，有望提升提示微调的收敛效率，并使提示表示能够更加准确地对齐跨场景多任务的需求。

最后，持续学习主要关注缓解灾难性遗忘并维持模型迭代过程中的知识稳定，而参数高效微调则针对单任务或少量任务的模型快速适配。现有研究通常将这两类方法分别进行设计，但是二者之间的协同机制应是一个有意义的研究方向。例如，当模型需要在长任务序列中不断扩展能力时或需要逐步拓宽应用领域和范围时，持续学习技术可以有效管理跨任务的知识共享与累积，而高效微调策略则可

带来更低的参数更新成本。同时，更少的参数更新成本有望减少对原参数调动带来的知识冲突。这种融合策略有望在控制模型规模增长的同时实现知识的持续积累，为构建可长期迭代演化且计算开销可控的智能化系统提供新的技术途径。

6 结束语

本文围绕绿色视觉AI的发展目标，从数据节能型、标注节能型、推理节能型和迭代节能型视觉AI方法出发，全面梳理了面向视觉智能全生命周期的资源高效技术研究进展。在数据采集层面，本文综述了基于数据合成与域迁移的方法，展示了生成模型与知识复用技术在降低真实数据依赖方面的潜力；在标注成本层面，本文分析了弱监督学习与自监督学习的主流路径，揭示了利用未标注数据与内在监督信号替代人工标注的有效机制；在推理成本层面，本文探讨了模型轻量化与推理加速两类技术路线，呈现了从算法优化到系统协同的多种效率提升手段；在迭代成本层面，本文总结了持续学习与参数高效微调的研究现状，阐述了模型在多任务场景下可持续演进的技术路径。

总体而言，绿色视觉AI作为应对人工智能资源消耗挑战的重要研究方向，正经历从单一环节优化向全生命周期系统协同的演进。随着数据高效学习、模型轻量化设计、可持续迭代机制等技术的持续突破，绿色视觉AI有望在保障性能的前提下，显著降低视觉智能系统的资源开销与环境影响。面向未来，随着“双碳”战略的深入推进与视觉智能产业的高质量发展需求，绿色视觉AI将在构建可持续、可扩展、可落地的现代化、产业化体系建设中发挥日益重要的作用。

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]

Bajpai D J and Hanawal M K. 2025. FREE: fast and robust vision language models with early exits//Findings of the Association for Computational Linguistics: ACL 2025. [s. l.]: ACL: 23599-23615 [DOI: 10.18653/v1/2025.findings-acl.1209]

Ballas A and Diou C. 2025. Gradient-guided annealing for domain generalization//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 20558-20568[DOI: 10.1109/CVPR52734.2025.01914]

Bao H B, Dong L, Piao S H and Wei F R. 2022. BEiT: BERT pre-training of image transformers[EB/OL]. [2026-03-13]. http://arxiv.org/abs/2106.08254.pdf

Barsellotti L, Amoroso R, Cornia M, Baraldi L and Cucchiara R. 2024. training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3689-3698[DOI: 10.1109/CVPR52733.2024.00354]

Beck M, Pöppel K, Lippe P and Hochreiter S. 2025. Tiled flash linear attention: more efficient linear RNN and xLSTM kernels//Proceedings of the 39th Annual Conference on Neural Information Processing Systems. San Diego, USA: Curran Associates

Beck M, Schweighofer K, Böck S, Lehner S and Hochreiter S. 2026. xLSTM scaling laws: competitive performance with linear time-complexity//Proceedings of the 14th International Conference on Learning Representations. Rio de Janeiro, Brazil: ICLR

Bi Q, Yi J J, Huang H M, Zheng H, Zhan H L, Huang Y W, et al. 2025. NightAdapter: learning a frequency adapter for generalizable night-time scene segmentation//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 23838-23849[DOI: 10.1109/CVPR52734.2025.02220]

Bolya D, Fu C Y, Dai X, Zhang P, Feichtenhofer C and Hoffman J. 2023. Token merging: your ViT but faster//Proceedings of the 11th International Conference on Learning Representations. [s.l.]: ICLR

Boussaid H, Kwon N, Kurtz C, Wendling L and Lobry S. 2025. LLM-driven data augmentation for visual question answering//Proceedings of 2025 Joint Urban Remote Sensing Event. 1-4[DOI: 10.1109/JURSE60372.2025.11076083]

Brooks T, Holynski A and Efros A A. 2023. InstructPix2Pix: learning to follow image editing instructions//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 18392-18402 [DOI: 10.1109/CVPR52729.2023.01764]

Cao J J, Zhang Y, Huang T, Lu M, Zhang Q Z, An R, et al.. 2025. MoVE-KD: knowledge distillation for VLMs with mixture of visual encoders//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 19846-19856[DOI: 10.1109/CVPR52734.2025.01848]

Castro-Macías F M, Morales-Álvarez P, Wu Y, Molina R and Katsaggelos A K. 2024. Sm: enhanced localization in multiple instance learning for medical imaging classification//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 77494-77524

Chaudhary A. 2020. Semi-supervised learning in computer vision[EB/OL]. [2026-03-13]. https://amitness.com/posts/semi-supervised-learning

Chen J R, Cao T Y, Xu J, Li J H, Chen Z L, Xiao T, et al. 2025. Con4m: context-aware consistency learning framework for segmented time series classification[EB/OL]. [2026-03-12]. http://arxiv.org/abs/2408.00041.pdf

Chen J X, Li L, Su L, Zha Z J and Huang Q M. 2024. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 18319-18329 [DOI: 10.1109/CVPR52733.2024.01734]

Chen K, Xie E, Chen Z, Wang Y B, Hong L Q, Li Z G, et al. 2023. GeoDiffusion: text-prompted geometric control for object detection data generation//Proceedings of International Conference on Learning Representations 2024. Vienna, Austria: OpenReview: 8979--9001

Chen L, Li J S, Dong X Y, Zhang P, He C H, Wang J Q, et al. 2024. ShareGPT4V: improving large multi-modal models with better captions//Proceedings of Computer Vision – ECCV 2024. Milan, Italy: Springer Nature Switzerland: 370-387 [DOI: 10.1007/978-3-031-72643-9_22]

Chen T, Kornblith S, Norouzi M and Hinton G. 2020. A simple framework for contrastive learning of visual representations [EB/OL]. [2026-03-13]. http://arxiv.org/abs/2002.05709.pdf

Chen W X, Liu Y, Chen B L, Su J D, Zheng Y S and Lin L. 2025. Cross-modal causal relation alignment for video question grounding//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 24087-24096 [DOI: 10.1109/CVPR52734.2025.02243]

Chen X M, Han L F, Huang P L, Feng X X, Zhang D W and Han J W. 2026. Retinex-rawmamba: bridging demosaicing and denoising for low-light RAW image enhancement. IEEE Transactions on Circuits and Systems for Video Technology, 36(1): 406-420 [DOI: 10.1109/TCSVT.2025.3589476]

Chen X, Huang L H, Liu Y, Shen Y J, Zhao D L and Zhao H S. 2024. AnyDoor: zero-shot object-level image customization//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6593-6602 [DOI: 10.1109/CVPR52733.2024.00630]

Chen X N, Huo S, Jiang B R, Hu H L and Chen X H. 2025. Single domain generalization for few-shot counting via universal representation matching//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4639-4649 [DOI: 10.1109/CVPR52734.2025.00437]

Cheng D, Hu Y S, Wang N S, Zhang D W and Gao X B. 2025. Achieving plasticity-stability trade-off in continual learning through adaptive orthogonal projection. IEEE Transactions on Circuits and Systems for Video Technology, 35(8): 7485-7498 [DOI: 10.1109/TCSVT.2025.3547916]

Cheng D, Wang G R, Wang B, Zhang Q, Han J G and Zhang D W. 2023. Hybrid routing transformer for zero-shot learning. Pattern Recognition, 137: 109270 [DOI: 10.1016/j.patcog.2022.109270]

Cheng D, Ji Y L, Gong D, Li Y, Wang N N, Han W J, et al. 2024. Continual all-in-one adverse weather removal with knowledge replay on a unified network structure. IEEE Transactions on Multimedia, 26: 8184-8196 [DOI: 10.1109/TMM.2024.3377136]

Cheng N, Luo C Y, Li H, Ma S K, Lei S G and Li P. 2025. LSV-MAE: a masked-autoencoder pre-training approach for large-scale 3D point cloud data. IEEE Access, 13: 135708-135721 [DOI: 10.1109/ACCESS.2025.3594614]

Cheng T H, Song L, Ge Y X, Liu W Y, Wang X G and Shan Y. 2024. YOLO-World: real-time open-vocabulary object detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 16901-16911 [DOI: 10.1109/CVPR52733.2024.01599]

Clark A, de Las Casas D, Guy A, Freeman J, Rae J, Hernandez D, et al. 2022. Unified scaling laws for routed language models//Proceedings of the 39th International Conference on Machine Learning. [s.l.]: PMLR: 4057-4086

Dai M, Cheng W X, Zhuang J D, Liu J J, Zhao H S, Feng Z H, et al. 2025. PropVG: end-to-end proposal-driven visual grounding with multi-granularity discrimination//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 7058-7068

Deng A D, Gao Z P, Choudhuri A, Planche B, Zheng M, Wang B, et al. 2025. Seq2Time: sequential knowledge transfer for video LLM temporal grounding//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13766-13775 [DOI: 10.1109/CVPR52734.2025.01285]

Deng J R, Zhang H J, Ding K , Hu J H, Zhang X X and Wang Y K. 2024. Zero-shot generalizable incremental learning for vision-language object detection//Proceedings of Advances in Neural Information Processing Systems. 136679-136700 [DOI: 10.52202/079017-4342]

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2026-03-13]. http://arxiv.org/abs/1810.04805.pdf

Di S Z, Yu Z L, Zhang G H, Li H Y, Zhong T, Cheng H, et al. 2025. Streaming video question-answering with in-context video KV-cache retrieval//Proceedings of the 13th International Conference on Learning Representations. Singapore: ICLR

Ding X, Wang L, Koniusz P and Gao Y S. 2025. Graph your own prompt [EB/OL]. [2026-03-13]. https://arxiv.org/abs/2509.23373v2.pdf

Ding X H, Zhang X Y, Han J G and Ding G G. 2022. Scaling up your

kernels to 31×31: revisiting large kernel design in CNNs//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11953-11965[DOI: 10.1109/CVPR52688.2022.01166]

Ding X H, Zhang X Y, Ma N N, Han J G, Ding G G and Sun J. 2021. RepVGG: making VGG-style ConvNets great again//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13728-13737[DOI: 10.1109/CVPR46437.2021.01352]

Dong H Y and Lee G H. 2025. PS-Mamba: spatial-temporal graph mamba for pose sequence refinement//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 8568-8578

Dong H, Chatzi E and Fink O. 2025. Towards robust multimodal open-set test-time adaptation via adaptive entropy-aware optimization//Proceedings of International Conference on Learning Representations 2025. Singapore: OpenReview: 77426-77451

Dünkel O, Wimmer T, Theobalt C, Rupprecht C and Kortylewski A. 2025. Do it yourself: learning semantic correspondence from pseudo-labels//Proceedings of the IEEE/CVF International Conference on Computer Vision: 5834-5844

Duru A and Temizel A. 2025. Adaptive augmentation policy optimization with LLM feedback[EB/OL]. [2026-03-12]. http://arxiv.org/abs/2410.13453.pdf

Dutt A, Shah N, Masarani H and Gandhi A. 2026. Scaling state-space models on multiple GPUs with tensor parallelism [EB/OL]. [2026-03-13]. https://arxiv.org/abs/2602.21144.pdf

Esser P, Kulal S, Blattmann A, Entezari S, Müller J, Saini H, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR

Fedus W, Zoph B and Shazeer N. 2022. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research, 23(120): 1-39

Feng W X, Zhu W R, Fu T, Jampani V, Akula A, He X H, et al. 2023. LayoutGPT: compositional visual planning and generation with large language models//Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.: 18225-18250

Fiche G, Leglaive S, Alameda-Pineda X and Moreno-Noguer F. 2025. MEGA: masked generative autoencoder for human mesh recovery//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5366-5378[DOI: 10.1109/CVPR52734.2025.00505]

Fu Q C, Cho M, Merth T, Mehta S, Rastegari M and Najibi M. 2024. LazyLLM: dynamic token pruning for efficient long context LLM inference[EB/OL].[2026-03-13]. https://arxiv.org/abs/2407.14057.pdf

Fei S C, Gao X H, Hu J W, Hou X L, Li L and Ren J, 2025. Knowledge distillation-based distributed dynamic 3d gaussian splatting for large scale scene reconstruction. Expert Systems with Applications, 305: 130758[DOI: 10.1016/j.eswa.2025.130758]

Fu S H, Yang Q Z, Mo Q J, Yan J K, Wei X H, Meng J K, et al. 2025. LLMDet: learning strong open-vocabulary object detectors under the supervision of large language models//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14987-14997[DOI: 10.1109/CVPR52734.2025.01396]

Gao R Y, Chen K, Xie E Z, Hong L Q, Li Z G, Yeung D Y, et al. 2024. MagicDrive: street view generation with diverse 3D geometry control//Proceedings of International Conference on Learning Representations 2024. Vienna, Austria: OpenReview: 22841-22860

Ge J W, Zhang X Y, Cao J X, Zhu X L, Liu W J, Gao Q Q, et al. 2025. Gen4Track: a tuning-free data augmentation framework via self-correcting diffusion model for vision-language tracking//Proceedings of the 33rd ACM International Conference on Multimedia. Tucson, USA: Association for Computing Machinery: 3037-3046[DOI: 10.1145/3746027.3754956]

Ge S R, Cao J and He R. 2025. Improving object detection models via LLM-based training data synthesis. International Journal of Computer Vision, 133(12): 8436-8451[DOI: 10.1007/s11263-025-02560-x]

Ghojogh B and Ghodsi A. 2024. Diffusion models: tutorial and survey [EB/OL]. [2026-03-13]. https://osf.io/w7jcm_v1/

Gao Y Y, Dai Y L, Li H, Ye W C, Chen J Y, Chen D P, et al. 2026. Cosurfgs: 3D surface gaussian splatting with collaborative distributed learning for large-scale scene reconstruction. International Journal of Computer Vision, 134(5): 195[DOI: 10.1007/s11263-025-02627-9]

Gong Y Z, Yu S Y, Wang X Y and Xiao J M. 2024. Continual segmentation with disentangled objectness learning and class recognition//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3848-3857[DOI: 10.1109/CVPR52733.2024.00369]

Gong Z Y, Wei Z X, Wang D, Hu X X, Ma X Z, Chen H R X, et al. 2025. CrossEarth: geospatial vision foundation model for domain generalizable remote sensing semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025: 1-18[DOI: 10.1109/TPAMI.2025.3649001]

Gou J X, Ji L P, Liu P and Ye M. 2025. Queryable prototype multiple instance learning with vision-language models for incremental whole slide image classification. Proceedings of the AAAI Conference on Artificial Intelligence, 39(3): 3158-3166[DOI: 10.1609/aaai.v39i3.32325]

Gu A and Dao T. 2024. Mamba: linear-time sequence modeling with selective state spaces//Proceedings of the 1st Conference on Lan-

guage Modeling. [s.l.]: COLM

Gu A, Goel K and Re C. 2022. Efficiently modeling long sequences with structured state spaces//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: ICLR

Gu A, Johnson I, Goel K, Saab K, Dao T, Rudra A, et al. 2021. Combining recurrent, convolutional, and continuous-time models with linear state space layers//Proceedings of the 35th Annual Conference on Neural Information Processing Systems. [s.l.]: Curran Associates: #44

Guo G Y, Han L F, Wang L, Zhang D W and Han J W. 2023. Semantic-aware knowledge distillation with parameter-free feature uniformization. Vis. Intell.1(1):6[DOI:10.1007/s44267-023-00003-0]

Guo G Y, Zhang D W, Han L F, Liu N, Cheng M M and Han J W. 2024. Pixel distillation: cost-flexible distillation across image sizes and heterogeneous networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12):9536-9550[DOI:10.1109/TPAMI.2024.3421277]

Guo H R, Zeng F H, Zhu F, Wang J Y, Wang X K, Zhou J G, et al. 2025. Continual learning for generative AI: from LLMs to MLLMs and beyond [EB/OL]. [2026-03-13]. https://arxiv.org/abs/2506.13045.pdf

Guo Y C and Gu X D. 2025. MMRL: multi-modal representation learning for vision-language models//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE:25015-25025[DOI:10.1109/CVPR52734.2025.02329]

Han Y Z, Huang G, Song S J, Yang L, Wang H H and Wang Y L. 2022. Dynamic neural networks: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(11): 7436-7456 [DOI:10.1109/TPAMI.2021.3117837]

Hao J T, Huang Q, Liu H, Xiao X Y, Ren Z C and Yu J. 2025. A token is worth over 1,000 tokens: efficient knowledge distillation through low-rank clone//Proceedings of the 39th Annual Conference on Neural Information Processing Systems. [s.l.]: Curran Associates: #8862

Hayou S, Ghosh N and Yu B. 2024. LoRA+: efficient low rank adaptation of large models//Proceedings of the 41st International Conference on Machine Learning. [s.l.]: PMLR:20798-20823

He J P, Duan Z H and Zhu F Q. 2025. CL-LoRA: continual low-rank adaptation for rehearsal-free class-incremental learning//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 30534-30544 [DOI: 10.1109/CVPR52734.2025.02843]

He K M, Chen X L, Xie S N, Li Y H, Dollár P and Girshick R. 2021. Masked autoencoders are scalable vision learners[EB/OL]. [2026-03-13]. http://arxiv.org/abs/2111.06377.pdf

He K M, Fan H Q, Wu Y X, Xie S N and Girshick R. 2020. Momentum contrast for unsupervised visual representation learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9726-9735[DOI: 10.1109/CVPR42600.2020.00975]

He Z M, Yao Y Z, Zuo P F, Gao B, Li Q Y, Zheng Z Z, et al. 2025. AdaSkip: adaptive sublayer skipping for accelerating long-context LLM inference//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI: 24050-24058[DOI: 10.1609/aaai.v39i22.34579]

Heo S. 2025. Prompt driven mmultimodal large language models for concrete defect identification. IEEE Access, 13: 160278-160287 [DOI:10.1109/ACCESS.2025.3605263]

Hinton G, Vinyals O and Dean J. 2015. Distilling the knowledge in a neural network [EB/OL]. [2026-03-13]. https://arxiv.org/pdf/1503.02531.pdf

Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, et al. 2019. Parameter-efficient transfer learning for NLP//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR:2790-2799

Hu E, Shen Y L, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. 2022. LoRA: low-rank adaptation of large language models//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: ICLR

Hu J K, Yao Z J, Jin L J, He H Z and Lu Y Y. 2025. Enhancing image restoration transformer via adaptive translation equivariance//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE:16047-16057

Hu K, Xiao Y, Zhang Y and Gao X P. 2024. Multi-view masked contrastive representation learning for endoscopic video analysis. Advances in Neural Information Processing Systems, 37: 47987-48014[DOI:10.52202/079017-1521]

Hu Y S, Cheng D, Zhang D W, Wang N N, Liu T L and Gao X B. 2024. Task-aware orthogonal sparse network for exploring shared knowledge in continual learning//Proceedings of the 41st International Conference on Machine Learning. [s.l.]: PMLR

Huang H, Xia Y, Zhou S S, Wang H T, Wang S L and Zhao Z. 2025. Bridging domain generalization to multimodal domain generalization via unified representations//Proceedings of the IEEE/CVF International Conference on Computer Vision:22488-22498

Huang P L, Zhang D W, Cheng D, Han L F, Zhu P F and Han J W. 2024. M-RRFS: a memory-based robust region feature synthesizer for zero-shot object detection. International Journal of Computer Vision, 132(10): 4651-4672 [DOI: 10.1007/s11263-024-02112-9]

Huang T, Zhang Z Y and Tang H. 2025a. 3D-R1: enhancing reasoning in 3D VLMs for unified scene understanding[EB/OL]. [2026-03-13]. http://arxiv.org/abs/2507.23478.pdf

Huang T, Zhang Z Y, Wang Y M and Tang H. 2025b. 3D CoCa: contrastive learners are 3D captioners[EB/OL]. [2026-03-13].

http://arxiv.org/abs/2504.09518.pdf

Huang Y W, Gokaslan A, Kuleshov V and Tompkin J. 2024. The GAN is dead; long live the GAN! a modern baseline GAN//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 44177-44215[DOI:10.52202/079017-1402]

Huo F Y, Tan J C, Liu J H, Jiang Z X, Li J C, Wang J G, et al. 2026. RepSpec: structural re-parameterized draft model training for speculative decoding//Proceedings of the 14th International Conference on Learning Representations. Rio de Janeiro, Brazil: ICLR

Jang E J, Kang M, Kim S, Sagong M and Park S H. 2026. Revisiting masked image modeling with standardized color space for domain generalized fundus photography classification//Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. Daejeon, the Republic of Korea: Springer Nature Switzerland:538-548[DOI:10.1007/978-3-032-04981-0_51]

Jang J and Kwon H Y. 2024. Are multiple instance learning algorithms learnable for instances? //Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.:10575-10612

Jang J, Ma C F and Lee B. 2025. VL2Lite: task-specific knowledge distillation from large vision-language models to lightweight networks//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 30073-30083 [DOI:10.1109/CVPR52734.2025.02799]

Jayasuriya D, Tayebati S, Ettori D, Krishnan R and Trivedi A R. 2025. SPARC: subspace-aware prompt adaptation for robust continual learning in LLMs [EB/OL].[2026-03-13]. https://arxiv.org/abs/2502.02909.pdf

Jhajj G and Lin F. 2025. Elastic weight consolidation for knowledge graph continual learning: an empirical evaluation//NORA: the 1st Workshop on Knowledge Graphs & Agentic Systems Interplay. [s.l.]

Jia C, Yang Y F, Xia Y, Chen Y T, Parekh Z, Pham H, et al. 2021. Scaling up visual and vision-language representation learning with noisy text supervision[EB/OL].[2026-03-13]. http://arxiv.org/abs/2102.05918.pdf

Jiang H J, Li Z X, Yu X H, Hu Y L, Yin B C, Yang J, et al. 2025. Visual and semantic prompt collaboration for generalized zero-shot learning//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 20275-20285[DOI:10.1109/CVPR52734.2025.01888]

Jiang L, Shi S S and Schiele B. 2024. Open-vocabulary 3D semantic segmentation with foundation models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 21284-21294 [DOI: 10.1109/CVPR52733.2024.02011]

Jiang X G, Yu L, Lin G J and Zhang Y C. 2025. PMD-transformer: a domain generalization approach for person re-identification. IEEE Access, 13:93178-93189[DOI:10.1109/ACCESS.2025.3573921]

Jin X, Su H S, Liu K, Ma C, Wu W, Hui F, et al. 2025. UniMamba: unified spatial-channel representation learning with group-efficient mamba for LiDAR-based 3D object detection//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1407-1417 [DOI: 10.1109/CVPR52734.2025.00139]

Juan X, Zhou K X, Liu N H, Chen T L and Wang X. 2024. Molecular data programming: towards molecule pseudo-labeling with systematic weak supervision//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 308-318[DOI:10.1109/CVPR52733.2024.00037]

Kang H, Seifer G, Lee D and Ryu J. 2025. Do your best and get enough rest for continual learning//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10077-10086 [DOI: 10.1109/CVPR52734.2025.00942]

Kashiani H, Talemi N A and Afghah F. 2025. FreqDebias: towards generalizable deepfake detection via consistency-driven frequency debiasing//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 8775-8785[DOI:10.1109/CVPR52734.2025.00820]

Katharopoulos A, Vyas A, Pappas N and Fleuret F. 2020. Transformers are RNNs: fast autoregressive transformers with linear attention//Proceedings of the 37th International Conference on Machine Learning. [s.l.]: PMLR:5156-5165

Kaya Y, Hong S and Dumitras T. 2018. Shallow-deep networks: understanding and mitigating network overthinking//Proceedings of the 35th International Conference on Machine Learning. [s.l.]: PMLR: 2529-2538

Khaldi K, Nguyen V D, Mantini P and Shah S. 2024. Unsupervised person re-identification in aerial imagery//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. Waikoloa, USA: IEEE: 260-269 [DOI: 10.1109/WACVW60836.2024.00034]

Kim N R, Lee J S and Lee J H. 2024. Learning with structural labels for learning with noisy labels//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE:27600-27610[DOI:10.1109/CVPR52733.2024.02607]

Kim S, Lee D, Kang S, Chae S, Jang S and Yu H. 2024. Learning discriminative dynamics with label corruption for noisy label detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE:22477-22487 [DOI:10.1109/CVPR52733.2024.02121]

Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu A A, et al. 2017. Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114 (13):3521-3526

Kumar J, Pillai J and Doermann D. 2011. Document image classification

and labeling using multiple instance learning//Proceedings of 2011 International Conference on Document Analysis and Recognition. Beijing, China: IEEE: 1059-1063 [DOI: 10.1109/ICDAR.2011.214]

Kweon H and Yoon K-J. 2025. WISH: weakly supervised instance segmentation using heterogeneous labels//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 25377-25387 [DOI: 10.1109/CVPR52734.2025.02363]

Kwon W, Li Z, Zhuang S, Sheng Y, Zheng L, Yu C H, et al. 2023. Efficient memory management for large language model serving with PagedAttention//Proceedings of the 29th Symposium on Operating Systems Principles. Koblenz, Germany: ACM: 611-626 [DOI: 10.1145/3600006.3613165]

Lee B. 2025. Understanding self-supervised contrastive learning through supervised objectives [EB/OL]. [2026-03-13]. http://arxiv.org/abs/2510.10572.pdf

Lee J, Hayat M and Yun S. 2025. Tripartite weight-space ensemble for few-shot class-incremental learning//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 15329-15338 [DOI: 10.1109/CVPR52734.2025.01428]

Lee S, Kim M, Chae Y and Stenger B. 2024. Linearly controllable GAN: unsupervised feature categorization and decomposition for image generation and manipulation//Proceedings of Computer Vision - ECCV 2024. Milan, Italy: Springer Nature Switzerland: 229-245 [DOI: 10.1007/978-3-031-73235-5_13]

Lee S, Lee J and Kang M. 2025. MINR: implicit neural representations with masked image modelling [EB/OL]. [2026-03-13]. http://arxiv.org/abs/2507.22404.pdf

Lee W, Lee J, Seo J, Moon H, Lee S and Kim G W. 2024. InfiniGen: efficient generative inference of large language models with dynamic KV cache management//Proceedings of the 18th USENIX Symposium on Operating Systems Design and Implementation. Santa Clara, USA: USENIX Association: 155-172

Lenz B, Lieber O, Arazi A, Shachaf G, Mitliagkas I and Shoham Y. 2025. Jamba: hybrid transformer-mamba language models//Proceedings of the 13th International Conference on Learning Representations. Singapore: ICLR

Li B Y, Zhao H Y, Wang W X, Hu P, Gou Y B and Peng X. 2025. MaIR: a locality- and continuity-preserving mamba for image restoration//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 7491-7501 [DOI: 10.1109/CVPR52734.2025.00702]

Li D C, Shao R L, Xie A Z, Xing E P, Ma X Z, Stoica I, et al. 2024. DistFlashAttn: distributed memory-efficient attention for long-context LLMs training//Proceedings of the 1st Conference on Language Modeling. [s.l.]: COLM

Li H, Li J F, Zhang D W, Wu C M, Shi J Q, Zhao C, et al, 2025. VDG: vision-only dynamic gaussian for driving simulation. IEEE Robotics and Automation Letters, 10(5): 5138-5145 [DOI: 10.1109/LRA.2025.3555938]

Li H Y, Li Y M, Tian A X, Tang T H, Xu Z C, Chen X J, et al. 2025. A survey on large language model acceleration based on KV cache management. Transactions on Machine Learning Research

Li H B, Hu P, Zhang Q J, Peng X, Liu X T and Yang M X. 2025. Test-time adaptation for cross-modal retrieval with query shift//Proceedings of International Conference on Learning Representations 2025. Singapore: OpenReview: 76517-76544

Li H Y, Wang L, Wang C, Jiang J, Peng Y and Long G D. 2025. DPC: dual-prompt collaboration for tuning vision-language models//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 25623-25632 [DOI: 10.1109/CVPR52734.2025.02386]

Li J Y, Shi H, Wu S, Zheng C, Li Z G, Jiang X, et al. 2025. QuickLLaMA: query-aware inference acceleration for large language models//Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE: ACL: 508-528

Li J L, Wang H Q, Wang W M, Qin J, Wang Q and Zhu L. 2026. Source-free active domain adaptation for efficient medical video polyp segmentation//Proceedings of Medical Image Computing and Computer Assisted Intervention - MICCAI 2025. Daejeon, the Republic of Korea: Springer Nature Switzerland: 499-509 [DOI: 10.1007/978-3-032-05127-1_48]

Li J S, Wang S K, Qian B, He Y H, Wei X, Wang Q, et al. 2025. Dynamic integration of task-specific adapters for class incremental learning//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 30545-30555 [DOI: 10.1109/CVPR52734.2025.02844]

Li K Y, Liu R X, Cao X Y, Bai X R, Zhou F, Meng D Y, et al. 2025. SegEarth-OV: towards training-free open-vocabulary segmentation for remote sensing images//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10545-10556 [DOI: 10.1109/CVPR52734.2025.00986]

Li X F, Zhang Y F and Ye X Q. 2024. DrivingDiffusion: layout-guided multi-view driving scenarios video generation with latent diffusion model//Proceedings of Computer Vision - ECCV 2024. Milan, Italy: Springer Nature Switzerland: 469-485 [DOI: 10.1007/978-3-031-73229-4_27]

Li Y H, Huang Y B, Yang B W, Venkitesh B, Locatelli A, Ye H C, et al. 2024. SnapKV: LLM Knows What You are Looking for Before Generation//Advances in Neural Information Processing Systems 37. [s.l.]: Curran Associates, Inc.: 22947-22970 [DOI: 10.52202/079017-0722]

Li Z Z and Hoiem D. 2017. Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(12): 2935-2947 [DOI: 10.1109/TPAMI.2017.2773081]

Lian L, Li B Y, Yala A and Darrell T. 2023. LLM-grounded Diffusion: enhancing prompt understanding of text-to-image diffusion models with large language models[EB/OL]. [2026-03-13]. https://arxiv.org/abs/2305.13655v3.pdf

Liao T C, Xie B H, Fu L L, Huang S, Deng B W, Chen C, et al. 2025. Federated domain generalization with decision insight matrix//Proceedings of Thirty-Fourth International Joint Conference on Artificial Intelligence: 5689-5697[DOI: 10.24963/ijcai.2025/633]

Lin X T, Ren P Z, Yeh C H, Yao L N, Song A and Chang X J. 2021. Unsupervised person re-identification: a systematic survey of challenges and solutions[EB/OL]. [2026-03-13]. http://arxiv.org/abs/2109.06057.pdf

Lipman Y, Chen R T Q and Ben-Hamu H. 2023. Flow matching for generative modeling//Proceedings of the 11th International Conference on Learning Representations. [s.l.]: ICLR

Liu A X, Feng B, Wang B, Wang B X, Liu B, Zhao C G, et al. 2024. DeepSeek-V2: a strong, economical, and efficient mixture-of-experts language model [EB/OL]. [2026-03-13]. https://arxiv.org/abs/2405.04434.pdf

Liu H, Wu C, Cheng J H, Chai W H, Wang S Y, Liu G W, et al. 2025. MonoTAKD: teaching assistant knowledge distillation for monocular 3D object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 22266-22275 [DOI: 10.1109/CVPR52734.2025.02074]

Liu S Y, Wang C Y, Yin H X, Molchanov P, Wang Y, Cheng K, et al. 2024. DoRA: weight-decomposed low-rank adaptation//Proceedings of the 41st International Conference on Machine Learning. [s.l.]: PMLR: 28092-28117

Liu X and Chang X B. 2025. LoRA subtraction for drift-resistant space in exemplar-free continual learning//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 15308-15318[DOI: 10.1109/CVPR52734.2025.01426]

Liu X C, Gong C Y and Liu Q. 2023. Flow straight and fast: learning to generate and transfer data with rectified flow//Proceedings of the 11th International Conference on Learning Representations. [s.l.]: ICLR

Lou M and Yu Y Z. 2025. Overlock: an overview-first-look-closely-next convnet with context-mixing dynamic kernels//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 128-138

Lu C, Zhou Y H, Bao F, Chen J F, Li C X and Zhu J. 2022. DPM-solver: a fast ODE solver for diffusion probabilistic model sampling in around 10 steps//Proceedings of the 36th Annual Conference on Neural Information Processing Systems. [s.l.]: Curran Associates: #418

Lu C Y, Derakhshandeh K and Chaterji S. 2025. Improving semi-supervised semantic segmentation with sliced-Wasserstein feature alignment and uniformity//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 20233-20243 [DOI: 10.1109/CVPR52734.2025.01884]

Lu Q Y, Ding L, Cao S Y, Zhang K J, Zhang J X and Tao D C. 2025. Runaway is ashamed, but helpful: on the early-exit behavior of large language model-based agents in embodied environments// Findings of the Association for Computational Linguistics: EMNLP 2025. [s.l.]: ACL: 24014-24027

Lü S, Kang M and Li X M. 2024. Alleviating imbalanced pseudo-label distribution: self-supervised multi-source domain adaptation with label-specific confidence//Proceedings of Thirty-Third International Joint Conference on Artificial Intelligence: 4669-4677 [DOI: 10.24963/ijcai.2024/516]

Luo D L, Zhu H S, Zhang Z Y, Liang D K, Xie X D, Liu Y L, et al. 2025. SemiETS: integrating spatial and content consistencies for semi-supervised end-to-end text spotting//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9329-9338 [DOI: 10.1109/CVPR52734.2025.00871]

Luo S M, Tan Y Q, Huang L B, Li J and Zhao H. 2023. Latent consistency models: synthesizing high-resolution images with few-step inference[EB/OL].[2026-3-13]. https://arxiv.org/abs/2310.04378.pdf

Luo Z Y, Liu N, Yang X G, Zhang D W, Fan D P, Khan F S, et al. 2026. Vscode-v2: dynamic prompt learning for general visual salient and camouflaged object detection with two-stage optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 48(3): 3137-3153[DOI: 10.1109/TPAMI.2025.3635136]

Luo Z Y, Liu N, Zhao W B, Yang X G, Zhang D W, Fan D P, et al. 2024. Vscode: general visual salient and camouflaged object detection with 2D prompt learning//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 17169-17180 [DOI: 10.1109/CVPR52733.2024.01625]

Luthra A, Yang T B and Galanti T. 2025. Self-supervised contrastive learning is approximately supervised contrastive learning[EB/OL]. [2026-03-13]. http://arxiv.org/abs/2506.04411.pdf

Lyu F, Wang L, Li X, Zheng W S, Zhang Z, Zhou T and Hu F Y. 2025. Comprehensive survey of continual learning. Journal of Image and Graphics, 30(8): 2599-2632 (吕凡, 王亮, 李玺, 郑伟诗, 张彰, 周涛, 胡伏原 . 2025. 持续学习研究进展 . 中国图象图形学报, 30(8): 2599-2632)[DOI: 10. 11834/jig. 240661]

Mallya A and Lazebnik S. 2018. PackNet: adding multiple tasks to a single network by iterative pruning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7765-7773[DOI: 10.1109/CVPR.2018.00810]

Man Y B, Huang Y, Zhang C M, Li B Z, Niu W and Yin M. 2025.

Adacm2: on understanding extremely long-term video with adaptive cross-modality memory reduction//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 8534-8544 [DOI: 10.1109/CVPR52734.2025.00798]

Mansourian A M, Ahmadi R, Ghafouri M, Babaei A M, Golezani E B, Ghamchi Z Y, et al. 2025. A comprehensive survey on knowledge distillation. Transactions on Machine Learning Research

Meng W K, Luo Y D, Li X, Jiang D M and Zhang Z. 2025. PolaFormer: polarity-aware linear attention for vision transformers//Proceedings of the 13th International Conference on Learning Representations. Singapore: ICLR

Nadeem N, Asad M H, Anwar S and Bais A. 2025. MaskAdapt: unsupervised geometry-aware domain adaptation using multimodal contextual learning and RGB-depth masking//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, USA: IEEE: 5422-5432 [DOI:10.1109/CVPRW67362.2025.00539]

Nagaraj S, Gerych W, Tonekaboni S, Goldenberg A, Ustun B and Hartvigsen T. 2025. Learning under temporal label noise [EB/OL]. [2026-03-12]. http://arxiv.org/abs/2402.04398.pdf

Nazir M, Aqeel M and Setti F. 2025. Diffusion-based Data Augmentation for Medical Image Segmentation//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision Workshops. Honolulu, USA: IEEE: 1341-1350 [DOI: 10.1109/ICCVW69036.2025.00143]

Nguyen V D, Mantini P and Shah S K. 2024. Contrastive clothing and pose generation for cloth-changing person re-identification//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE: 7541-7549 [DOI:10.1109/CVPRW63382.2024.00749]

Ni Y and Koniusz P. 2024a. CHAIN: enhancing generalization in data-efficient GANs via Lipschitz continuity constrained normalization//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6763-6774 [DOI: 10.1109/CVPR52733.2024.00646]

Ni Y, Wen S, Koniusz P and Cherian A. 2025. Noise consistency regularization for improved subject-driven image synthesis//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville, USA: IEEE: 3107-3117 [DOI: 10.1109/CVPRW67362.2025.00294]

Ni Y, Zhang S and Koniusz P. 2024b. PACE: marrying generalization in parameter-efficient fine-tuning with consistency regularization. Advances in Neural Information Processing Systems, 37: 61238-61266 [DOI:10.52202/079017-1958]

Niu K, Yu H Y, Zhao M Y, Fu T, Yi S Y, Lu W, et al. 2025. ChatReID: Open-ended Interactive Person Retrieval via Hierarchical Progressive Tuning for Vision Language Models//Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 24245-24254 [DOI: 10.1109/ICCV51701.2025.02247]

Noori M, Cheraghalikhani M, Bahri A, Vargas Hakim G A, Osowiechi D, Yazdanpanah M, et al. 2025. FDS: feedback-guided domain synthesis with multi-source conditional diffusion models for domain generalization//Proceedings of 2025 IEEE/CVF Winter Conference on Applications of Computer Vision. Tucson, USA: IEEE: 8504-8514 [DOI:10.1109/WACV61041.2025.00824]

Pan X G, Tewari A, Leimkühler T, Liu L, Meka A and Theobalt C. 2023. Drag your GAN: interactive point-based manipulation on the generative image manifold//Proceedings of ACM SIGGRAPH 2023 Conference Proceedings. Los Angeles, USA: Association for Computing Machinery: 1-11 [DOI:10.1145/3588432.3591500]

Peng H, Pappas N, Yogatama D, Schwartz R, Smith N and Kong L. 2021. Random feature attention//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: ICLR

Przewięźlikowski M, Balestriero R, Jasiński W, Śmieja M and Zieliński B. 2025. Beyond [cls]: exploring the true potential of masked image modeling representations//Proceedings of the IEEE/CVF International Conference on Computer Vision: 23442-23452

Qian X L, Fu Y W, Xiang T, Jiang Y G, Xue X Y. 2019. Leader-based multi-scale attention deep architecture for person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2): 371-385 [DOI:10.1109/TPAMI.2019.2928294]

Qorbani R, Villani G, Panagiotakopoulos T, Colomer M B, Härenstam-Nielsen L, Segu M, et al. 2025. Semantic library adaptation: LoRA retrieval and fusion for open-vocabulary semantic segmentation//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9804-9815 [DOI:10.1109/CVPR52734.2025.00916]

Qu L H, Yang D K, Huang D, Guo Q H, Luo R K, Zhang S T, et al. 2024. PPathology-knowledge enhanced multi-instance prompt learning for few-shot whole slide image classification//Proceedings of Computer Vision – ECCV 2024. Milan, Italy: Springer Nature Switzerland: 196-212 [DOI:10.1007/978-3-031-73247-8_12]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision [EB/OL]. [2026-03-13]. http://arxiv.org/abs/2103.00020.pdf

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer [EB/OL]. [2026-03-13]. http://arxiv.org/abs/1910.10683.pdf

Ran X J, Li Y X, Xu L N, Yu M L and Dai B. 2025. Direct numerical layout generation for 3D indoor scene synthesis via spatial reasoning [EB/OL]. [2026-03-12]. http://arxiv.org/abs/2506.05341.pdf

Rao Y M, Zhao W L, Liu B L, Lu J W, Zhou J and Hsieh C J. 2021.

DynamicViT: efficient vision transformers with dynamic token sparsification//Proceedings of the 35th Annual Conference on Neural Information Processing Systems.[s.l.]: Curran Associates: #1068

Ren L, Chen C, Wang L and Hua K. 2025. DA-VPT: semantic-guided visual prompt tuning for vision transformers//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, IEEE: 4353-4363 [DOI: 10.1109/CVPR52734.2025.00411]

Roy S and Etemad A. 2024. Consistency-guided prompt learning for vision-language models//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR

Rusu A A, Rabinowitz N C, Desjardins G, Soyer H, Kirkpatrick J, Kavukcuoglu K, et al. 2016. Progressive neural networks [EB/OL]. [2026-03-13]. https://arxiv.org/abs/1606.04671.pdf

Safaei B, VS V and Patel V M. 2025. Certainty and uncertainty guided active domain adaptation//Proceedings of 2025 IEEE International Conference on Image Processing (ICIP). Anchorage, USA: IEEE: 2342-2347[DOI:10.1109/ICIP55913.2025.11084455]

Salimans T and Ho J. 2022. Progressive distillation for fast sampling of diffusion models// Proceedings of the 10th International Conference on Learning Representations.[s.l.]: ICLR

Sanh V, Debut L, Chaumond J and Wolf T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [EB/OL]. [2026-03-13]. https://arxiv.org/abs/1910.01108.pdf

Sarkar S D, Miksik O, Pollefeys M, Barath D and Armeni I. 2025. CrossOver: 3D scene cross-modal alignment//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 8985-8994 [DOI: 10.1109/CVPR52734.2025.00840]

Sauer A, Lorenz D, Blattmann A and Rombach R. 2024. Adversarial diffusion distillation//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer:87-103[DOI:10.1007/978-3-031-73016-0_6]

Schuster T, Fisch A, Gupta J, Dehghani M, Bahri D, Tran V Q, et al. 2022. Confident adaptive language modeling//Proceedings of the 36th Annual Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1269

Serra J, Suris D, Miron M and Karatzoglou A. 2018. Overcoming catastrophic forgetting with hard attention to the task//Proceedings of the 35th International Conference on Machine Learning. Stockholmsmässan, Stockholm, Sweden: PMLR:8024-8033

Shang C K, Li M K, Zhang Y Q, Chen Z, Wu J L, Gu F Q, et al. 2025. PRO-VPT: distribution-adaptive visual prompt tuning via prompt relocation//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 1558-1568[DOI:10.1109/ICCV51701.2025.00153]

Sharma R, Ji K Y, Xu Z Q and Chen C Y. 2024. AUC-CL: a batchsize-robust framework for self-supervised contrastive representation learning//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR

Sheng Y, Zheng L M, Yuan B H, Li Z H, Ryabinin M, Chen B D, et al. 2023. FlexGen: high-throughput generative inference of large language models with a single GPU//Proceedings of the 40th International Conference on Machine Learning. Honolulu, Hawaii, USA: PMLR:31094 - 31116

Shu F X, Liao Y, Zhang L, Zhuo L, Xu C N, Zhang G H, et al. 2025. LLaVA-MoD: making LLaVA tiny via MoE-knowledge distillation// Proceedings of the 13th International Conference on Learning Representations. Singapore: ICLR

So J, Shin J, Jang C and Park E. 2025. PCM: Picard consistency model for fast parallel sampling of diffusion models//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 23313-23322 [DOI: 10.1109/CVPR52734.2025.02171]

Sobal V, Ibrahim M, Balestriero R, Cabannes V, Bouchacourt D, Astolfi P, et al. 2025. X-sample contrastive loss: improving contrastive learning with sample similarity graphs//Proceedings of the 13th International Conference on Learning Representations. Singapore: ICLR

Son H M, Zhao Z, Rezaei S and Liu X. 2025. Domain generalization in-the-wild: disentangling classification from domain-aware representations[EB/OL]. [2026-03-13]. https://arxiv.org/abs/2508.21769.pdf

Song J M, Meng C L and Ermon S. 2021. Denoising diffusion implicit models//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: ICLR

Song Y and Dhariwal P. 2024. Improved techniques for training consistency models//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR

Song Y, Dhariwal P, Chen M and Ilya S. 2023. Consistency models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR:32211-32252

Song Y J, Hwang Y, Lee J H, Lee H C and Lim D-Y. 2025. DGSAM: domain generalization via individual sharpness-aware minimization [EB/OL]. [2026-03-12]. http://arxiv.org/abs/2503.23430.pdf

Sun X Y, Cheng D, Li Y, Wang N N, Zhang D W, Gao X B, et al. 2025. Progressive prompt-driven low-light image enhancement with frequency aware learning. IEEE Transactions on Multimedia, 27: 6620-6634[DOI:10.1109/TMM.2025.3586101]

Tan Z Q, Zhang Y F, Yang J Q and Yuan Y. 2024. Contrastive learning is spectral clustering on similarity graph//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR

Tang H, Huang T and Zhang Z Y. 2026. 3D CoCa v2: contrastive learners with test-time search for generalizable spatial intelligence[EB/

OL]. [2026-03-13]. http://arxiv.org/abs/2601.06496.pdf

Tang J M, Zhao Y L, Zhu K, Xiao G X, Kasikci B, Han S, et al. 2024. QUEST: query-aware sparsity for efficient long-context LLM inference//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR:47901-47911

Tang S, Su W X, Gan Y, Ye M, Zhang J W and Zhu X T. 2025. Proxy denoising for source-free domain adaptation//Proceedings of the 13th International Conference on Learning Representations. Singapore: ICLR

Tarashima S H, Shu X Q and Tagawa N. 2025. ViLAaD: enhancing "attracting and dispersing" source-free domain adaptation with vision-and-language model[EB/OL]. [2026-03-12]. http://arxiv.org/abs/2503.23529.pdf

Termöhlen J A, Bartels T and Fingscheidt T. 2023. A re-parameterized vision transformer (ReVT) for domain-generalized semantic segmentation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision Workshops. Paris, France: IEEE:4378-4387[DOI:10.1109/ICCVW60793.2023.00472]

Thengane V, Lahoud J, Cholakkal H, Anwer R M, Yin L, Zhu X T, et al. 2025. CLIMB-3D: continual learning for imbalanced 3D instance segmentation[EB/OL]. [2026-03-13]. https://arxiv.org/abs/2502.17429.pdf

Thomas X and Ghadiyaram D. 2025. What's in a latent? Leveraging diffusion latent space for domain generalization//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 2183-2194 [DOI: 10.1109/ICCV51701.2025.00211]

Tian Y, Ji K Y, Zhang R Z, Jiang Y K, Li C Y, Wang X S, et al. 2025. Towards all-in-one medical image re-identification//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 30774-30786 [DOI: 10.1109/CVPR52734.2025.02866]

Tian Z C, Liu Y Y and Sun Q R. 2025. Meta-learning hyperparameters for parameter efficient fine-tuning//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE:23037-23047[DOI:10.1109/CVPR52734.2025.02145]

Tiwary P, Bhattacharyya K P P A. 2025. LangDAug: Langevin data augmentation for multi-source domain generalization in medical image segmentation//Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need//Proceedings of the 31st Annual Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates:6000-6010

Veeramacheneni L, Wolter M, Kuehne H and Gall J. 2025. Canonical rank adaptation: an efficient fine-tuning strategy for vision transformers//Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR.

Vincent C, Kim T and Meeß H. 2025. High temporal consistency through semantic similarity propagation in semi-supervised video semantic segmentation for autonomous flight//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1461-1471 [DOI: 10.1109/CVPR52734.2025.00144]

Vu A K N, Truong T Q, Nguyen V-T, Ngo T D, Do T-T and Nguyen T V. 2025. Multi-perspective data augmentation for few-shot object detection//Proceedings of the 13th International Conference on Learning Representations. Singapore: ICLR

Wang J, Lan X, Zhou J Z, Tian Y X and Lv J C. 2025. Style quantization for data-efficient GAN training//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 7696-7706 [DOI: 10.1109/CVPR52734.2025.00721]

Wang J X, Li G H, Liu J Q, Xu Z Y, Chen X R and Wei J M. 2025. RelVid: relational learning with vision-language models for weakly video anomaly detection. Sensors, 25 (7) [DOI: 10.3390/s25072037]

Wang J S, Cao N Q, Ding Y, Xie M Y, Gu F Q and Chen C. 2025. SKE-Layout: spatial knowledge enhanced layout generation with LLMs//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 19414-19423[DOI:10.1109/CVPR52734.2025.01808]

Wang K Y, Fu X Y, Lu X, Ge C J, Cao C Z, Zhai W, et al. 2025. Efficient test-time adaptive object detection via sensitivity-guided pruning//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10577-10586[DOI:10.1109/CVPR52734.2025.00989]

Wang L P, Chen S, Jiang L N, Pan S, Cai R Z, Yang S, et al. 2025. Parameter-efficient fine-tuning in large language models: a survey of methodologies. Artificial Intelligence Review, 58:#227[DOI:10.1007/s10462-025-11236-4]

Wang L Y, Zhang X L, Jia C M and Ma S W. 2025. MAFS: masked autoencoder for infrared-visible image fusion and semantic segmentation. IEEE Transactions on Image Processing, 34: 6490-6505 [DOI:10.1109/TIP.2025.3611602]

Wang L, Xu L C, Yang X, Huang Z H and Cheng J. 2025. Debiased distillation for consistency regularization//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI:7799-7807[DOI: 10.1609/aaai.v39i8.32840]

Wang Q Z, Qian X L, Li B, Fu Y W, Xue X Y. 2025. Image-text-image knowledge transfer for lifelong person re-identification with hybrid clothing states. IEEE Transactions on Image Processing, 34:5584-5597[DOI:10.1109/TIP.2025.3602745]

Wang S B, Yang Y C, Liu Z Y, Sun C H, Hu X M, He C H, et al. 2025. Dataset distillation with neural characteristic function: a min-max perspective//Proceedings of 2025 IEEE/CVF Conference on

Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 25570-25580[DOI:10.1109/CVPR52734.2025.02381]

Wang S W, Yu L X and Li J. 2024. LoRA-GA: low-rank adaptation with gradient approximation//Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates:54905-54931[DOI:10.52202/079017-1741]

Wang S D, Zhang Y J, Zhu Y, Li J N, Wang Z Z, Liu Y W, et al. 2025. Towards understanding how knowledge evolves in large vision-language models//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 29858-29868 [DOI: 10.1109/CVPR52734.2025.02779]

Wang T, Wang M K, Wang Z Z, Wang H K, Xu Q, Cong F Y, et al. 2025. ODA-GAN: orthogonal decoupling alignment GAN assisted by weakly-supervised learning for virtual immunohistochemistry staining//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 25920-25929[DOI: 10.1109/CVPR52734.2025.02414]

Wang X Y, Bai H H, Yu L M, Zhao Y and Xiao J M. 2024. Towards the uncharted: density-descending feature perturbation for semi-supervised semantic segmentation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3303-3312[DOI:10.1109/CVPR52733.2024.00318]

Wang X Y, Liu Y C, Cheng W, Zhao X, Chen Z Z, Yu W C, et al. 2025. MixLLM: dynamic routing in mixed large language models// Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics. Albuquerque, USA: ACL:10912-10922

Wang X, Wang S A, Ding Y H, Li Y H, Wu W T, Rong Y, et al. 2024. State space model for new-generation network alternative to transformers: a survey [EB/OL]. [2026-03-13]. https://arxiv.org/abs/2404.09516.pdf

Wang X R, Zhang J, Qi L and Shi Y H. 2025. Balanced direction from multifarious choices: arithmetic meta-learning for domain generalization//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 30577-30587[DOI:10.1109/CVPR52734.2025.02847]

Wang Y Z, Cheng L C, Fang C W, Zhang D W, Duan M N and Wang M. 2024. Revisiting the power of prompt for visual tuning//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR:50233 - 50247

Wang Y Z, Duan M N and Kong S. 2025. Attention to the burstiness in visual prompt tuning! //Proceedings of 2025 IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 4253-4263[DOI: 10.1109/ICCV51701.2025.00405]

Wang Z, He W, Liang Z, Zhang X, Bansal C, Wei Y, et al. 2025. CREAM: consistency regularized self-rewarding language models// Proceedings of the 13th International Conference on Learning Representations. Singapore: ICLR

Wang Z F, Zhang Z Z, Lee C-Y, Zhang H, Sun R X, Ren X Q, et al. 2022. Learning to prompt for continual learning//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 139-149 [DOI: 10.1109/CVPR52688.2022.00024]

Wen C G, Peng Z L, Huang Y, Yang X K and Shen W. 2025. Domain generalization in CLIP via learning with diverse text prompts//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9559-9569[DOI: 10.1109/CVPR52734.2025.00893]

Wu F W, Cheng L C, Tang S G, Zhu X F, Fang C W, Zhang D W, et al. 2025. Navigating semantic drift in task-agnostic class-incremental learning//Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR.

Wu W J, Zhao Y Z, Chen H, Gu Y C, Zhao R, He Y F, et al. 2023a. DatasetDM: synthesizing data with perception annotations using diffusion models// Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.: 54683-54695

Wu W J, Zhao Y Z, Shou M Z, Zhou H and Shen C H. 2023b. DiffuMask: synthesizing images with pixel-level annotations for semantic segmentation using diffusion models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE:1206-1217[DOI:10.1109/ICCV51070.2023.00117]

Wu Z Y, Ding T J, Lu Y F, Pai D, Zhang J Y, Wang W D, et al. 2025. Token statistics transformer: linear-time attention via variational rate reduction//Proceedings of the 13th International Conference on Learning Representations. Singapore: ICLR

Xi X, Huang Y Y, Luo R H and Qiu Y. 2025. OW-OVD: unified open world and open vocabulary object detection//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 25454-25464 [DOI: 10.1109/CVPR52734.2025.02370]

Xiang W Z, Liu C, Yu H Y and Chen X L. 2025. Wavelet-driven masked image modeling: a path to efficient visual representation/// Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI: 8611-8619[DOI: 10.1609/aaai.v39i8.32930]

Xiao G X, Tang J M, Zuo J W, Guo J X, Yang S, Tang H T, et al. 2025. DuoAttention: efficient long-context LLM inference with retrieval and streaming heads//Proceedings of the 13th International Conference on Learning Representations. Singapore: ICLR

Xiao G X, Tian Y D, Chen B D, Han S and Lewis M. 2024. Efficient streaming language models with attention sinks//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR

Xiao X, Zhang Y B, Li X J, Wang T Y, Wang X, Wei Y X, et al. 2025. Visual instance-aware prompt tuning//Proceedings of the 33rd ACM International Conference on Multimedia. New York,

NY, USA: ACM:2880-2889[DOI:10.1145/3746027.3754858]

Xie F, Nie J H, Tan Y, Zhang W K and Zhao H S. 2025. Mamba-Adaptor: state space model adaptor for visual recognition//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 20124-20134[DOI: 10.1109/CVPR52734.2025.01874]

Xie M J, Gong J Y, Gao Z and Cao M. 2025. Data augmentation for remote sensing semantic segmentation via controllable diffusion models//Proceedings of IGARSS 2025 - 2025 IEEE International Geoscience and Remote Sensing Symposium. Brisbane, Australia: IEEE:6132-6136[DOI:10.1109/IGARSS55030.2025.11242875]

Xing J L, Liu J P, Wang J, Sun L L, Chen X, Gu X X, et al. 2024. A survey of efficient fine-tuning methods for vision-language models — prompt and adapter. Computers and Graphics, 119: #1012 [DOI:10.1016/j.cag.2024.01.012]

Xu G Z, Guo H, Yi L, Ling C, Wang B Y and Yi G. 2025. Revisiting source-free domain adaptation: a new perspective via uncertainty control//Proceedings of the 13th International Conference on Learning Representations. Singapore: ICLR

Xu J C, Lo S-Y, Safaei B, Patel V M and Dwivedi I. 2025. Towards zero-shot anomaly detection and reasoning with multimodal large language models//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 20370-20382[DOI:10.1109/CVPR52734.2025.01897]

Xu J, Pan J, Zhou Y, Zhang Y and Jiang J. 2025. Specee: accelerating large language model inference with speculative early exiting//Proceedings of the 52nd Annual International Symposium on Computer Architecture. Tokyo, Japan: IEEE: 467-481 [DOI: 10.1145/3695053.3730996]

Xu S L, Sun Y, Li X F, Duan S Y, Ren Z W, Liu Z, et al. 2025. Noisy label calibration for multi-view classification//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI:21797-21805[DOI:10.1609/aaai.v39i20.35485]

Xu Z P, Cheng D, Jiang X Y, Wang N N, Li D S and Gao X B. 2025. Adversarial domain prompt tuning and generation for single domain generalization//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 18584-18595[DOI:10.1109/CVPR52734.2025.01732]

Xue W H, Yang Y, Li L, Huang Z L, Wang X G, Han J W, et al. 2023. Weakly supervised point cloud segmentation via deep morphological semantic information embedding. CAAI Transactions on Intelligence Technology, 9 (3): 695-708 [DOI: 10.1049/cit2.12239]

Yamaguchi S, Feng D W, Kanai S, Adachi K and Chijiwa D. 2025. Post-pre-training for modality alignment in vision-language foundation models//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4256-4266[DOI:10.1109/CVPR52734.2025.00402]

Yang J J, Wang W J, Chen K Y, Liu L Q, Zou Z X and Shi Z W. 2025. Structural representation-guided GAN for remote sensing image cloud removal. IEEE Geoscience and Remote Sensing Letters, 22: 1-5[DOI:10.1109/LGRS.2024.3516078]

Yang K H, Han J W, Guo G Y, Fang C W, Fan Y Z, Cheng L C, et al. 2024. Progressive adapting and pruning: domain-incremental learning for saliency prediction. ACM Transactions on Multimedia Computing, Communications, and Applications, 20(8): #243[DOI:10.1145/3661312]

Yang L R, Wang J Q, Zhai Y J, Su P. 2026. Domain adaptive object detection via joint negative teaching and negative learning. Journal of Image and Graphics, 31(4): 1108-1124 (杨立然, 王佳琪, 翟永杰, 苏攀. 2026. 融合负教学和负学习的域自适应目标检测. 中国图象图形学报, 31(4): 1108-1124) [DOI: 10.11834/jig.250264]

Yang T Y, Huo H T, Guo B F, Zheng B W and Liu X W. 2026. Multi-level Mamba network for infrared and visible image fusion. Journal of Image and Graphics, 31(4): 1184-1200 (杨天宇, 霍宏涛, 郭宝峰, 郑博文, 刘晓文. 2026. 用于红外与可见光图像融合的多层级 Mamba 网络. 中国图象图形学报, 31(4): 1184-1200)[DOI: 10. 11834/jig. 250243]

Yang Y, Guo D D, Chen B and Hu D X. 2025. Continual learning with Bayesian compression for shared and private latent representations. Neural Networks, 185: 107167 [DOI: 10.1016/j. neunet. 2025. 107167]

Yao J R, Guo G Y, Zheng Z H, Xie Q, Han L F, Zhang D W, et al. 2025. Prompting vision-language model for nuclei instance segmentation and classification. IEEE Transactions on Medical Imaging, 44(11): 4567-4578[DOI:10.1109/TMI.2025.3579214]

Ye F and Bors A G. 2025. Online task-free continual learning via dynamic expansionable memory distribution//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 20512-20522 [DOI: 10.1109/CVPR52734.2025.01910]

Ye J X, Zhang W Q, Li Z Y, Li J, Zhao M and Tsung F. 2025. MedualTime: a dual-adapter language model for medical time series-text multimodal learning//Proceedings of the 34th International Joint Conference on Artificial Intelligence. Montreal, Canada: IJCAI: 7913 - 7921[DOI:10.24963/ijcai.2025/880]

Ye L F, Hamidi S M, Chi Z X, Li G, Pilanci M, Ogawa T, et al. 2026. ASMIL: attention-stabilized multiple instance learning for whole slide imaging[EB/OL]. [2026-03-13]. https://arxiv.org/abs/2603.06658.pdf

Yin D S, Hu L Y, Li B, Zhang Y Q and Yang X. 2025. 5%>100%: breaking performance shackles of full fine-tuning on visual recognition tasks//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 20071-20081[DOI:10.1109/CVPR52734.2025.01869]

Yin H M, Feng T L, Lyu F, Shang F H, Liu H Y, Feng W, et al. 2025. Beyond background shift: rethinking instance replay in continual

semantic segmentation//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9839-9848[DOI: 10.1109/CVPR52734.2025.00919]

Yin T W, Gharbi M, Zhang R, Shechtman E, Durand F, Freeman W T, et al. 2024. One-step diffusion with distribution matching distillation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6613-6623[DOI: 10.1109/CVPR52733.2024.00632]

Yu G I, Jeong J S, Kim G-W, Kim S and Chun B G. 2022. Orca: a distributed serving system for transformer-based generative models//Proceedings of the 16th USENIX Symposium on Operating Systems Design and Implementation. Carlsbad, USA: USENIX Association: 521-538

Yu J H, Wang Z R, Vasudevan V, Yeung L, Seyedhosseini M and Wu Y. 2022. CoCa: contrastive captioners are image-text foundation models[EB/OL]. [2026-03-13]. http://arxiv.org/abs/2205.01917.pdf

Yu Y, Ren B T, Zhang P Y, Liu M X, Luo J W, Zhang S F, et al. 2025. Point2RBox-v2: rethinking point-supervised oriented object detection with spatial layout among instances//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 19283-19293 [DOI: 10.1109/CVPR52734.2025.01796]

Yuan H B, Li X T, Zhou C, Li Y N, Chen K and Loy C C. 2024. Open-vocabulary SAM: segment and recognize twenty-thousand classes interactively//Proceedings of Computer Vision – ECCV 2024. Milan, Italy: Springer Nature Switzerland: 419-437[DOI: 10.1007/978-3-031-72775-7_24]

Yuan Y K, Dou H Z, Guo F J and Li X. 2024. SemanticMIM: marring masked image modeling with semantics compression for general visual representation[EB/OL]. [2026-03-13]. http://arxiv.org/abs/2406.10673.pdf

Zeng W L, Huang Z Y, Ji K X, Wang X and Yan Y C. 2025. Skip-vision: efficient and scalable acceleration of vision-language models via adaptive token skipping//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 21384-21397[DOI: 10.1109/ICCV51701.2025.01986]

Zhang D W, Cheng L B, Liu Y, Wang X G and Han J W. 2025. Mamba capsule routing towards part-whole relational camouflaged object detection. International Journal of Computer Vision, 133: 7201-7221[DOI: 10.1007/s11263-025-02530-3]

Zhang D W, Huang G H, Zhang Q, Han J G, Han J W and Yu Y Z. 2021. Cross-modality deep feature learning for brain tumor segmentation. Pattern Recognition, 110: 107562[DOI: 10.1016/j.patcog.2020.107562]

Zhang D W, Li H, He D Q, Liu N, Cheng L C, Wang J D, et al. 2025a. Unsupervised pre-training with language-vision prompts for low-data instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(10): 8642-8657[DOI: 10.1109/TPAMI.2025.3579469]

Zhang D W, Li H, Zeng W Y, Fang C W, Cheng L C, Cheng M-M, et al. 2025b. Weakly supervised semantic segmentation via alternate self-dual teaching. IEEE Transactions on Image Processing, 34: 3086-3095[DOI: 10.1109/TIP.2023.3343112]

Zhang D C, Zhang K, Chu S M, Wu L, Li X and Wei S. 2025. MORE: a mixture of low-rank experts for adaptive multi-task learning//Findings of the Association for Computational Linguistics: ACL 2025. [s.l.]: ACL: 1311-1324

Zhang H Z, Feng T and You J X. 2025. Router-R1: teaching LLMs multi-round routing and aggregation via reinforcement learning//Proceedings of the 39th Annual Conference on Neural Information Processing Systems. San Diego, USA: Curran Associates

Zhang J C, Li J M, Lin X R, Zhang W, Tan X, Han J Y, et al. 2024. Decoupled pseudo-labeling for semi-supervised monocular 3D object detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognitio. Seattle, USA: IEEE: 16923-16932[DOI: 10.1109/CVPR52733.2024.01601]

Zhang L M, Rao A Y and Agrawala M. 2023. Adding conditional control to text-to-image diffusion models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3813-3824[DOI: 10.1109/ICCV51070.2023.00355]

Zhang Q R, Chen M S, Bukharin A, He P C, Cheng Y, Chen W Z, et al. 2023. Adaptive budget allocation for parameter-efficient fine-tuning//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR

Zhang R Z, Tang S and Cao J. 2024. Self-supervised adversarial training via diverse augmented queries and self-supervised double perturbation//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 43788-43808

Zhang R, Yang Y X, Li Y, Wang J B, Miao Z, Li H, et al. 2022. Self-supervised learning based few-shot remote sensing scene image classification. Journal of Image and Graphics, 27(11): 3371-3381(张睿, 杨义鑫, 李阳, 王家宝, 苗壮, 李航, 等. 2022. 自监督学习下小样本遥感图像场景分类. 中国图象图形学报, 27(11): 3371-3381)[DOI: 10.11834/jig.210486]

Zhang S Z, Lv X Q, Xing Y H, Wu Q R, Xu D and Zhang Y N. 2025. Revisiting generative replay for class incremental object detection//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 20340-20349[DOI: 10.1109/CVPR52734.2025.01894]

Zhang X S, Han L F, Xu C C, Zheng Z H, Ding J, Fu X H. 2025. MHKD: multi-step hybrid knowledge distillation for low-resolution whole slide images glomerulus detection. IEEE Journal of Biomedical and Health Informatics, 29(2): 767-774[DOI: 10.1109/JBHI.2024.3513716]

Zhang X-W, Zhang D L, Peng Y-X, Ouyang Z, Meng J K and Zheng W-S. 2025. VIPerson: flexibly generating virtual identity for person

re-identification//Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 23374-23384

Zhang Y, Bin M Y, Zhang Y Y, Wang Z Y, Han Z and Liang C. 2025. Link-based contrastive learning for one-shot unsupervised domain adaptation//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4916-4926[DOI: 10.1109/CVPR52734.2025.00463]

Zhang Y, Deng Y-X, Guo M-H and Hu S-M. 2025. Adaptive parameter selection for tuning vision-language models//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4280-4290 [DOI: 10.1109/CVPR52734.2025.00404]

Zhang Y L, Li H L, Sun Y X, Shui Z Y, Li J X, Zhu C L, et al. 2026. AEM: attention entropy maximization for multiple instance learning based whole slide image classification//Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. Daejeon, the Republic of Korea: Springer Nature Switzerland: 45-55[DOI: 10.1007/978-3-032-04981-0_5]

Zhang Z W, Chen M H, Xiao S, Peng L, Li H J, Lin B B, et al. 2024. Pseudo label refinery for unsupervised domain adaptation on cross-dataset 3D object detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 15291-15300 [DOI: 10.1109/CVPR52733.2024.01448]

Zhang Z Y, Sheng Y, Zhou T Y, Chen T L, Zheng L M, Cai R S, et al. 2023. H2O: heavy-hitter oracle for efficient generative inference of large language models//Proceedings of the 37th Annual Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.: 34661-34710

Zhao H Q, Sheng D M, Bao J M, Chen D D, Chen D, Wen F, et al. 2023. X-Paste: revisiting scalable copy-paste for instance segmentation using CLIP and stablediffusion//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR: 42098-42109

Zhao J R, Li T Q, Jiang D H, Wu S H, Ramirez A and Lee T S. 2025. Perceptual inductive bias is what you need before contrastive learning//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9621-9630 [DOI: 10.1109/CVPR52734.2025.00899]

Zhao Y. 2025. AnomalyHybrid: a domain-agnostic generative framework for general anomaly detection//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville, USA: IEEE: 3118-3127[DOI: 10.1109/CVPRW67362.2025.00295]

Zhou J X, Guo D, Mao Y X, Zhong Y R, Chang X J and Wang M. 2024. Label-anticipated event disentanglement for audio-visual video parsing//Proceedings of Computer Vision – ECCV 2024. Milan, Italy: Springer Nature Switzerland: 35-51[DOI: 10.1007/978-3-031-72684-2_3]

Zhou J J, Xiong Y P, Liu Z, Liu Z, Xiao S T, Wang Y Z, et al. 2025. MegaPairs: massive data synthesis for universal multimodal retrieval//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. Vienna, Austria: Association for Computational Linguistics: 19076-19095 [DOI: 10.18653/v1/2025.acl-long.935]

Zhou X, Liu X M, Zhang F L, Wu G, Zhai D M, Jiang J J, et al. 2024. Zero-mean regularized spectral contrastive learning: implicitly mitigating wrong connections in positive-pair graphs//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR

Zhou Y Y, Li X H, Liu F Z, Wei Q Y, Chen X X, Yu L Q, et al. 2024. L2B: learning to bootstrap robust models for combating label noise//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 23523-23533[DOI: 10.1109/CVPR52733.2024.02220]

Zhou Y B, Ye Y T, Zhang P Y, Wei X and Chen M S. 2024. Exact fusion via feature distribution matching for few-shot image generation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8383-8392 [DOI: 10.1109/CVPR52733.2024.00801]

Zhou Z C, Hu C F and Wang Y X. 2024. A intelligent speech recognition method based on stable learning[EB/OL]. [2026-03-13]. https://www.researchsquare.com/article/rs-4019203/v1.pdf

Zhu H, Zhang Y F, Dong J H and Koniusz P. 2025. BiLoRA: almost-orthogonal parameter spaces for continual learning//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 25613-25622 [DOI: 10.1109/CVPR52734.2025.02385]

Zhu L H, Wang X G, Feng J P, Cheng T H, Li Y Y, Jiang B, et al. 2025. WeakCLIP: adapting CLIP for weakly-supervised semantic segmentation. International Journal of Computer Vision, 133(3): 1085-1105[DOI: 10.1007/s11263-024-02224-2]

Zhu X, Qian X, Shi Y, Tao X, Li Z. 2024. Video anomaly detection with long-and-short-term time series correlations. Journal of Image and Graphics, 29(7): 1998-2010 (朱新瑞, 钱小燕, 施俞洲, 陶旭东, 李智昱. 2024. 长短期时间序列关联的视频异常事件检测. 中国图象图形学报, 29(7): 1998-2010) [DOI: 10.11834/jig.230406]

Zhu Y C, Shi C, Wang D Y, Tang J J, Wei Z X, Wu Y, et al. 2025. Rethinking query-based transformer for continual image segmentation//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4595-4606 [DOI: 10.1109/CVPR52734.2025.00433]

作者简介

许畅, 男, 硕士研究生, 主要研究方向为多模态生成。 E-mail: changxu@mail.nwpu.edu.cn

王浩研，男，硕士研究生，主要研究方向为视觉目标识别跟踪。E-mail：15114996579@163.com