Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(2), 455-463
Published Online March 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.152044

Hans汉斯

图像识别中人工智能模型的性能评估与改进

包婉莹

呼和浩特职业技术大学计算机与信息工程学院，内蒙古 呼和浩特

收稿日期：2026年1月8日；录用日期：2026年2月26日；发布日期：2026年3月9日

摘　要

明确图像识别领域人工智能模型性能评估的核心价值，梳理现有评估体系不足，可为模型改进提供理论与实践指引。采用文献梳理与逻辑分析相结合的方法，系统剖析主流评估指标、常用评估方法及典型数据集，归纳复杂场景下模型的性能短板并提出改进策略。结果表明，准确率、精确率等核心指标及单一指标评估、跨数据集验证等方法各有优劣；模型在光照变化、姿态差异、遮挡干扰、小样本数据等场景中存在明显性能瓶颈；数据增强、模型结构优化、迁移学习与多模型融合可有效提升模型性能。结论指出，需构建多维度综合评估体系，从数据、结构、算法多方面协同推进模型改进，以增强其复杂场景下的泛化与鲁棒性，助力图像识别技术实用化发展。

关键词

图像识别，人工智能模型，性能评估，模型改进，泛化能力

Performance Evaluation and Improvement of Artificial Intelligence Models in Image Recognition

Wanying Bao

Department of Computer and Information Engineering, Hohhot Vocational and Technical University, Hohhot Inner Mongolia

Received: January 8, 2026; accepted: February 26, 2026; published: March 9, 2026

Abstract

Clarifying the core value of performance evaluation of artificial intelligence (AI) models in the field of image recognition and sorting out the deficiencies of existing evaluation systems can provide theoretical and practical guidance for model improvement. By adopting a combination of literature

文章引用: 包婉莹. 图像识别中人工智能模型的性能评估与改进[J]. 人工智能与机器人研究, 2026, 15(2): 455-463.
DOI: 10.12677/airr.2026.152044

review and logical analysis, this study systematically analyzes mainstream evaluation metrics, common evaluation methods and typical datasets, summarizes the performance shortcomings of models in complex scenarios, and proposes improvement strategies. The results show that core metrics such as accuracy and precision, as well as methods including single-metric evaluation and cross-dataset validation, each have their own advantages and disadvantages; models have obvious performance bottlenecks in scenarios such as illumination changes, pose differences, occlusion interference, and small-sample data; data augmentation, model structure optimization, transfer learning, and multi-model fusion can effectively improve model performance. The conclusion points out that it is necessary to construct a multi-dimensional comprehensive evaluation system and promote model improvement collaboratively from multiple aspects of data, structure, and algorithms to enhance their generalization and robustness in complex scenarios, thereby facilitating the practical development of image recognition technology.

Keywords

Image Recognition, Artificial Intelligence Model, Performance Evaluation, Model Improvement, Generalization Ability

1. 引言

随着人工智能的发展，图像识别作为计算机视觉的重要研究方向，已在自动驾驶、生物识别、医学影像等领域广泛推广，深刻改变了社会生产与生活方式[1] [2]。算法模型是人工智能实现的核心，其特征提取与模式匹配设计直接决定系统性能优劣。在实际部署中，算法选择需兼顾高精度、强鲁棒性、快速响应能力及环境适应性等核心要素，以克服复杂场景中的各类干扰，保障系统稳定运行[3]。

性能评估是模型研发与应用的关键环节，核心目的是客观评判模型性能、定位不足，为优化提供依据[4]。当前评估体系存在诸多短板，如单一指标难以全面客观评估、数据集场景覆盖不全导致结果脱离实际、评估方法不统一阻碍横向对比，严重制约技术落地。同时，现有模型技术局限性显著，易受光照变化、姿态变形、物体遮挡等复杂情况影响导致识别精度下降，小样本场景下易过拟合、泛化能力不足，高分辨率处理及实时运行场景中计算效率偏低。因此，构建系统的评估框架、提出针对性优化策略以提升模型环境适应能力至关重要。相关研究聚焦模型性能评定与改进，通过梳理主流评估指标、考查框架及数据集搭建策略，构建综合完备的评估体系，深入剖析算法技术桎梏并提出改良手段，助力模型迭代升级，实现多应用场景下的便捷可靠运行[5]。

2. 图像识别中人工智能模型性能评估体系

2.1. 核心评估指标

评估指标是量化模型辨识能力的核心参数，图像识别领域主流指标包括准确率、精确率、召回率、F1 值、ROC 曲线及 AUC 数值，各指标有特定应用场景与局限性(见图 1)，需结合任务需求合理选择并综合考量[6]。

准确率是基础指标，即正确识别样本占总样本比例，计算直观、适用于样本均衡场景，但样本不均衡时存在缺陷，如医疗影像诊断中患病样本占比低，模型全判正常仍可能获得高准确率，无法反映对患

病样本的识别能力。精确率(判定正类中实际正类比例，反映可靠性)与召回率(实际正类中正确判定比例，反映捕获能力)侧重特定类别评估，二者常此消彼长，需依任务权衡，如安防危险目标识别优先保证召回率，商品分类需兼顾二者。

Figure 1. Performance evaluation metrics of artificial intelligence models in image recognition
图 1. 图像识别中人工智能模型性能评估指标

F1 分数作为二者综合指标，可平衡权衡关系，避免单一评价的信息偏差，广泛应用于人脸识别、车牌识别等场景。ROC 曲线以假正例率为横轴、真正例率为纵轴，全面体现分类器不同阈值下的性能变化，其面积 AUC 取值 0~1，数值越接近 1 预测能力越强，适用于二元或多类别分类任务，擅长区分目标与非目标变量，且对样本不均衡场景适应性良好。

2.2. 常用评估方法

评估方法通过量化指标实现客观测评，典型形式包括单一指标评估、综合指标评估、交叉验证及跨数据集验证。单一指标评估依赖准确率、F1 值等核心参数，操作简单、计算快速，适用于模型初筛，但难以反映复杂场景综合性能，不可作为性能验证和系统改进的独立工具。

综合指标评估选取多指标构建体系，多维度全面评估性能，如复杂图像分类同步采用准确率、精确率等指标，结果客观全面、应用广泛，但需合理筛选指标避免冗余。交叉验证可减少训练集与测试集划分的评估误差，提升结果稳健性，典型方式为 k 折交叉验证和留一法：k 折交叉验证将数据随机分若干子集，轮流作为测试集迭代训练测试并取平均值，能高效利用样本、降低过拟合风险，适用于小型数据集；留一法为其极端简化形式，可信度高但计算复杂度高，适用范围较窄。

跨数据集验证通过多源样本测试模型泛化能力，可筛选高适应能力模型、规避单数据集过拟合，在自动驾驶交通标志识别等领域优势显著，但需获取丰富优质的多类别数据集，既增加研发成本，还需解决标注标准不统一问题。

2.3. 典型评估数据集

数据集的质量、规模与适用范围决定其可信度，当前图像识别领域已形成多个知名公开数据集，覆盖分类、目标检测等场景，为模型验证提供可靠支撑。图像分类领域的 ImageNet 数据集类别多、标注精细，提供了统一训练标准，推动了 AlexNet、ResNet 等经典网络架构发展。

目标检测领域的 VOC 与 COCO 数据集是核心评估标准：VOC 数据集类别丰富、边界框标注精准，

是基础算法验证的理想平台；COCO 数据集通过扩充样本量、细化标签信息，提升了实例分割精度与复杂场景适应能力，二者均为计算机视觉技术发展提供关键支撑。语义分割领域的 Cityscapes 与 ADE20K 极具代表性：Cityscapes 聚焦城市道路场景，含车道线、车辆等标注，广泛应用于自动驾驶目标检测评价；ADE20K 整合多种环境特征与大量类别标签，为通用语义分割任务提供基准测试支撑，推动算法优化。此外，特定场景数据集应用广泛，如医学影像领域的 ChestX-ray 数据集聚焦肺部疾病诊断，生物特征识别领域的 LFW 数据库含多视角、多光照人脸验证样本，为精准评价特定任务模型提供重要实践支撑。

3. 图像识别中人工智能模型的性能短板分析

3.1. 复杂环境下的鲁棒性不足

鲁棒性是系统在外部干扰下稳定运行的核心属性，当前多数图像识别模型在复杂场景中抗扰动能力薄弱，易受光照变化、姿态差异、遮挡现象及背景干扰影响，导致识别精度显著下降。

光照动态变化会改变图像亮度、对比度等关键属性，而主流模型多基于恒定光照数据集训练，局限性明显。如人脸识别在室内稳定光线中表现良好，室外强光或低照度环境下效果骤降；交通标志识别也会因阴雨天、夜间照明不足出现视觉模糊、色彩失真，影响检测精度。目标姿态的多样性大幅提升识别难度，现有训练数据难以覆盖所有典型姿态，对少见姿态分类性能较差，例如行人检测中，正面图像准确率较高，但弯腰、蹲坐等非标准姿势的检测精度和召回率均不理想。

遮挡干扰同样关键，目标部分或完全遮挡会导致模型无法获取完整特征，易出现误判、漏判，如监控中行人被树木、建筑遮挡，医疗影像中病灶被其他组织遮挡等场景均存在此类问题。此外，复杂环境中的背景噪声与目标特征相似度高，对模型精准定位目标、排除干扰形成考验，如自然场景下动物分类中，若目标颜色、纹理与环境过于接近，易发生误判或漏检。

3.2. 小样本数据下的泛化能力欠缺

泛化能力直接决定模型迁移应用效果，深度学习模型在大规模标注数据支撑下可实现优异性能，但在罕见疾病影像诊断、工业缺陷检测等小样本场景中，因数据获取难、标注成本高，现有技术难以满足需求，模型易出现过拟合，适应性不足。

小样本条件下，训练数据量不足导致模型无法充分提取目标特征，对输入高度敏感、易受噪声干扰，泛化能力薄弱，新测试样本分类精度低。以罕见病医学影像识别为例，可用标注样本稀少使模型易过拟合现有数据集，面对未知病例诊断准确性差，无法满足临床应用要求。同时，小样本数据集类别分布失衡，低频标签对应的目标特征难以被全面提取，导致识别精度下降、误判风险提升。如工业零部件缺陷检测中，异常样本远少于正常样本，训练过程被主流模式主导，罕见缺陷分类性能大幅下滑，无法达到生产质量控制标准。

3.3. 实时性与准确率的平衡难题

自动驾驶、实时监测等场景对识别精度和响应速度均有严苛要求，现有模型普遍面临两者难以兼顾的困境：提升准确率需采用复杂架构和多维度参数设置，必然增加计算负担、降低处理效率；简化模型结构以提升实时性，则会削弱特征提取能力，导致整体性能下降。

深度卷积神经网络虽精度优异，但复杂结构、海量参数带来的高计算成本对硬件资源提出挑战，普通终端设备难以实现高效实时运行，如自动驾驶中毫秒级延迟可能引发重大安全风险。轻量化模型虽能在一定程度上提升运算效率，但需警惕准确率下降导致的功能失灵风险。此外，图像分辨率选择也存在权衡：高分辨率图像细节丰富、检测精度高，但像素量大、计算复杂度高，削弱系统实时处理能力；低分

辨率图像运算负荷小、响应速度快，但关键特征提取受限，可能降低目标识别准确率。如何平衡两者矛盾，是当前图像识别领域的研究重点。

4. 图像识别中人工智能模型的改进策略

针对上述性能短板，本文提出系统化的改进策略(见图 2)，从数据增强、模型结构优化、迁移学习和多模型融合四个维度展开。

4.1. 基于数据增强的模型改进

数据增强技术会把原始训练样本经由很多途径来加以扩展和变形处理，进而达成数据集容量的增长，特征维度的拓展，以此提升模型的稳健性与适应能力，该方法特别适合小样本学习场景和高复杂度任务环境中的应用需求[7] [8]。

在一些复杂应用环境下，改进系统适应能力，最重要的是采用多种优化手段：利用亮度，对比度等参数随时模仿光照变化模式，提升设备应对光强度波段变化的处理能力；通过旋转，翻转之类的几何变换方法建立多种姿态数据库，提升模型的空间解析准确性；随机裁剪，遮挡技术用于设计典型干扰情形，增强模型对干扰的抵抗能力，从而扩大训练样本的覆盖范围，强调特征提取的主体部分，明显提升目标检测算法在恶劣环境下的鲁棒性和准确性。

在处理小样本数据集的时候，GAN 作为一种生成式模型是相当突出的，它存在两个主要组成部分，生成器与判别器，生成器的任务是生成与真实样例相媲美的合成样本，判别器则是执行真假的归类工作，在反馈循环中使生成的样本质量不断进步，经过检验合格的合成样本数据，可以加入原样本集，不仅增加原始数据的容量达到扩充效果，还可以同时减轻模型拟合现象，对整体泛化性产生明显的优化作用，尤其是在罕见病医学图像疾病诊断环境中，利用 GAN 技术产生大类别量不同形式的模拟影像资料后，找出典型病症实例的检验精确度比较明显提升了。

跨域数据增强技术利用多源异构数据集来扩充样本量，采用合适算法把原始数据映射到目标领域特征空间，以此优化模型泛化能力的稳定性，在自动驾驶交通标志识别场景中，融合仿真平台生成包含不同气象状况和路面状态等多种数据集合，再结合实际采集样本一起训练网络模型，可明显提升系统应对复杂环境的能力及鲁棒性[9]-[11]。

4.2. 基于模型结构优化的改进

优化模型结构是提升整体性能的必要途径，我们可以在模型结构上做出改进，在保持准确率的情况下大幅度降低计算复杂性，提升运行效率，同时以实时性和精确性的动态折中达到目标，进一步强化模型的特征表征能力、鲁棒性和跨场景兼容性。

为解决实时性不够的问题，可采取轻量化模型设计方法，这类办法简化网络架构，压缩参数量，改善计算步骤，在大幅提高运算速度的情况下，维持性能稳定，常见的轻量化方案包括 MobileNet 和 ShuffleNet 等，这些方法利用深度可分离卷积等新技术，既减少冗余参数又保持高效的特征提取能力，MobileNet 采用深度可分离卷积代替普通卷积，极大提升运行速度，并被广泛应用在移动终端及其他资源受限场景中的部署工作上[12] [13]。

为加强特征表示能力，改善模型在复杂情形下的适应性，可以创建融合多尺度信息提取与注意机制的崭新架构，多尺度整合技术依靠综合各个层级的空间属性，既满足局部细节分析的需求，又契合全局语义把握的要求，以此来改善目标检测的准确度，注意模块采用动态权重分配方法，强调关键区域，抑制次要背景干扰，进一步提升分类精确性，在目标检测领域当中，特征金字塔网络融合通道级和空间级

主要组件，在应对微小目标及遮挡情形时，表现出更好的性能表现。

针对深度神经网络体系中常常会出现的梯度消失或者爆炸现象，可以通过残差链接和稠密链接两种机制来予以解决。残差链接经过建立 shortcut 路径形式，简化误差逆推程序，遏制梯度衰退，并增进特征提取水平及其易发展的特性[14] [15]，而稠密连接则借助累积跨层链接提升特征传送的有效性，如此便改善信息流通质量并增进训练的稳定性程度，于是两种手段可促使整体目标检测精确性提升的趋向，在分类阶段反映得更为明朗。

4.3. 基于迁移学习的改进

迁移学习利用源域知识向目标域的快速传播，很好地解决了目标领域数据稀少和标注成本高的问题，显著提高了模型对未知环境的泛化能力，尤其是对小样本图像分类问题。

迁移学习在图像识别方面的主导应用模式可以概括为：通过大规模通用数据集，例如 ImageNet 来训练迁移学习的模型参数，在针对新的目标任务训练初始权重后，这些初始权重可以使模型提前学习到边缘，纹理等通用特征，从而为未来学习的小样本量带来极大的微调优势，不易产生过拟合现象，泛化能力更强。

依据目标领域数据集规模以及源域特征相关性的强弱，可以灵活选用对应的迁移学习策略，若样本数量不多且两者较为相似，则冻结大部分预训练模型参数，仅对顶层模块实施微调，如此一来便能充分利用通用表征并避免过拟合现象的发生，若数据分布存在较大差异，则要采用端到端全网络微调的方法来适应特定应用场景，倘若处于极端情况，即数据极为匮乏，那么需把预训练模型当作特征提取工具来使用，配合简易分类算法迅速完成训练[16]，在简化架构的同时增强泛化能力。

领域自适应技术的核心目标是提高迁移学习的效果，其本质就是通过减小源域和目标域之间的分布差异来优化任务效果，常用的领域自适应方法有领域对抗训练和度量学习策略。前者通过构建专门的领域判别网络来引导模型提取跨域不变的特征；后者则通过调整距离函数来改变样本在特征空间中的相似性程度，以此提高泛化能力和适应小样本环境的能力。

4.4. 基于多模型融合的改进

多模型融合是通过融合多个模型的预测结果，充分利用多个模型的优势，弥补不足，从而提高整体识别性能。不同的模型在特征提取、决策判断上各具特点，对不同样本的识别优势不同，通过融合可以取长补短，提高鲁棒性、泛化能力和准确率。

集成学习主要构建一种系统性的策略设计框架，具体包括投票法、加权投票法和堆叠法等，投票法汇总各模型预测结果，以多数类别作为最终判定，在多算法性能接近时具备较强的适应能力，加权投票法则依据单个模型的表现赋予相应权重，并根据综合评分选出最优解，最大程度发挥优质子模型效果的贡献，而堆叠法则把基础模型输出转化为新特征供元模型使用，以此达到提取和优化深层次关联信息的目的。

实践应用当中，依照具体任务需求合理挑选基础模型十分关键，针对图像分类问题，ResNet，VGG 系列架构是首选对象，经由技术整合之后，分类准确性得以进一步改善，目标检测领域则建议采用 YOLO 或者 Faster R-CNN 这类主流方案，而且借助多模态融合策略，定位精确度和召回率都获得明显改善[17] [18]。

集成学习方法融合了 bagging 和 boosting 等经典算法，目的在于提升预测精度，bagging 方法构建出多个相互独立的基础模型，然后将各个基础模型的输出结果做加权平均处理，以此来削减方差并提升整体稳定性，boosting 方法按照迭代顺序逐步训练基础模型[19] [20]，每次更新都会动态调整权重分配策略，从而有效缩减偏差，提升准确性，大幅度改善综合性能表现。

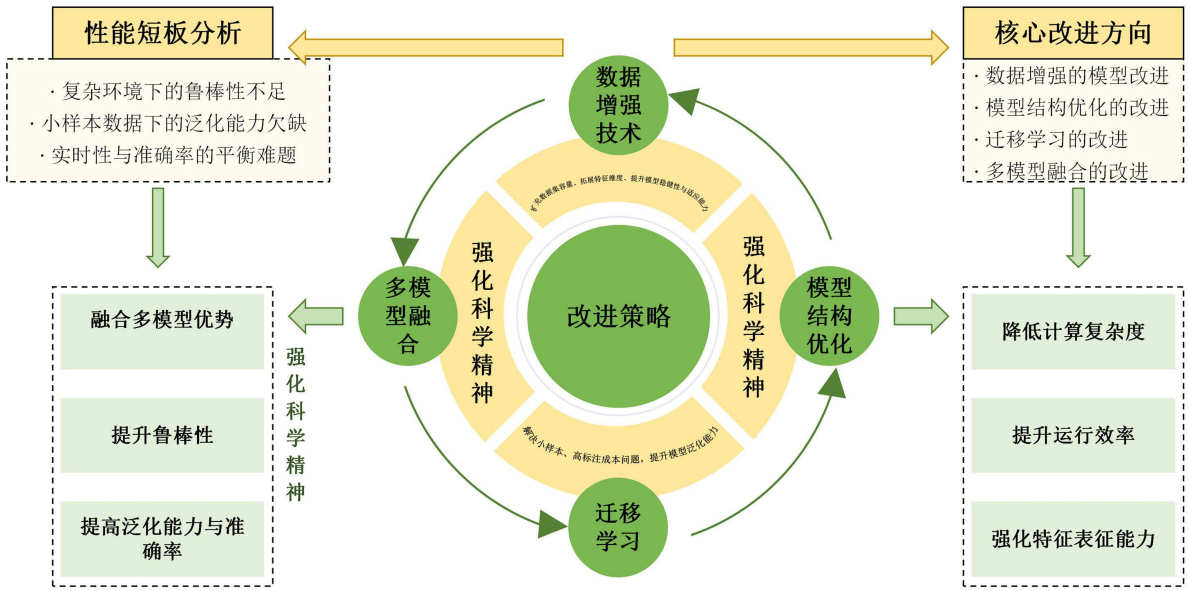

Figure 2. Improvement strategies of artificial intelligence models in image recognition
图 2. 图像识别中人工智能模型的改进策略

5. 结果与讨论

5.1. 评估体系构建结果

本文梳理评估指标、建立多样的综合评价体系，具有核心指标和多方面量化标准，对模型整体表现进行全方位考察，综合运用经典测评手段，结合具体任务特性及数据属性，选出最优算法，提升可信度；对特定任务类型选取典型数据集作为参照样本，为实验设计提供参考，打破单一参数或技术路线的局限性，对模型全方位特性加以描述，为后续改进工作提供坚实基础。

5.2. 改进策略效果分析

针对模型鲁棒性不足、泛化能力欠缺、实时性与准确率平衡难等短板，本文提出的四大改进策略可从不同角度提升综合性能。

数据增强技术通过样本变换和生成模型创建合成实例来实现数据集规模的扩展和特征维度的拓展，从而使算法在复杂场景中具有更好的适应性，同时也能有效缓解小样本训练带来的过拟合问题。GAN (Generative Adversarial Network)因为其生成高质量虚拟样本速度快的特点，在支持小样本学习任务时具有明显的优势，模型泛化能力和稳定性都得到了显著的提升。

模型结构优化轻量化设计与多尺度融合以及注意力机制、残差/稠密连接等技术结合，以保证精度的前提下降低运算量提高速度，保证较好的实时性与精度。轻量化模型针对性缺失资源环境，增加特征提取的多尺度融合与注意力机制模型，提升残差/稠密连接提升训练模型的稳定训练。

迁移学习策略利用大规模预训练知识，小样本问题，模型快速收敛、防止过拟合、泛化能力增强，迁移策略根据数据集场景选择，领域自适应技术减小域分布差异，增强适应性。

多模型融合方法把各种算法独有的优点融合起来，意图提高预测的准确性，提升抵抗干扰的能力以及改进泛化的能力，利用多种预测结果通过集成学习技术进行系统性的综合处理，一方面能够降低方差波动产生的负面影响，另一方面优化偏倚校正的效果，从而实现整体模型稳定性和精确度的全面

改善。

6. 结论

研究聚焦图像识别模型性能及优化路径，通过系统性文献梳理与推演构建评价体系，深入剖析技术缺陷并提出破解方案。研究发现，模型性能评估需依托系统化框架，单一维度与方法难以全面反映实际效果，本研究搭建的涵盖核心指标、常用技术及典型数据集的综合体系，可实现多维度客观评判，为后续优化提供科学依据。当前模型存在明显短板：环境适应性弱，对光照变化、姿态偏差等外部扰动容忍度低；泛化能力不足，小样本场景易出现过拟合；实时性与精度的平衡问题未有效解决，制约技术实用化进程。而系统化改进方案成效显著，数据增强可扩充样本特征维度，强化模型稳定性与泛化能力；优化网络架构兼顾实时处理效能与识别精度；迁移学习破解小样本训练难题，多模态融合策略则拓展技术应用边界。

参考文献

[1] 张丽英, 张永兴, 席云, 等. 基于人工智能的农机自动驾驶系统设计与优化[J]. 中国农机装备, 2025(12): 7-9.

[2] 韩永刚. 基于人工智能算法的图像识别技术分析[J]. 通讯世界, 2025, 32(11): 152-154.

[3] 雷郑波, 涂凯, 张永乐, 等. 一类分布鲁棒指数追踪模型及算法[J/OL]. 运筹学学报(中英文), 1-21[2025-12-28].

[4] 刘平献, 张明明, 王鹏, 等. 基于大模型的便民热线工单智能知识推荐系统的算法优化与性能评估[J]. 数字技术与应用, 2025, 43(3): 16-18.

[5] 孟彬, 杨帆. 基于深度强化学习的数据中心资源调度算法研究[J]. 软件, 2025, 46(11): 1-3.

[6] 阮春珠, 林旭怡, 张燕. 人工智能辅助学习系统技术架构优化与标准化性能评估[J]. 大众标准化, 2025(22): 164-166.

[7] Grill, J.-B., *et al*. (2020) Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. *Proceedings of the* 34*th International Conference on Neural Information Processing Systems*, Vancouver, 6-12 December 2020, 21271-21284.

[8] Chen, X. and He, K. (2021) Exploring Simple Siamese Representation Learning. 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 15745-15753. https://doi.org/10.1109/cvpr46437.2021.01549

[9] Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., *et al*. (2021) Emerging Properties in Self-Supervised Vision Transformers. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 9630-9640. https://doi.org/10.1109/iccv48922.2021.00951

[10] Li, J., *et al*. (2023) Uniform Masking: Enabling MAE Pre-Training for Pyramid-Based Vision Transformers. *IEEE*/*CVF International Conference on Computer Vision*, *ICCV* 2023, Paris, 1-6 October 2023, 1190-1199.

[11] Chen, X., Ding, M., Wang, X., *et al.* (2024). Context Autoencoder for Self-Supervised Representation Learning. *International Journal of Computer Vision*, **132**, 208-223. https://doi.org/10.1007/s11263-023-01852-4

[12] Li, J., *et al*. (2022) BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation. *Proceedings of the* 39*th International Conference on Machine Learning*, Baltimore, 17-23 July 2022, 12888-12900.

[13] Wang, P., *et al*. (2022) OFA: Unifying Architectures, Tasks, and Modalities via a Simple Sequence-to-Sequence Framework. *Proceedings of the* 39*th International Conference on Machine Learning*, Baltimore, 17-23 July 2022, 23318-23340.

[14] Alayrac, J.-B., *et al*. (2022) Flamingo: A Visual Language Model for Few-Shot Learning. *NeurIPS* 2022, New Orleans, 28 November-9 December 2022, 23716-23736.

[15] Driess, D., *et al*. (2023) PaLM-E: An Embodied Multimodal Language Model. *International Conference on Machine Learning*, *ICML* 2023, Honolulu, 23-29 July 2023, 8469-8488.

[16] Zhu, D., *et al*. (2023) MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models.

[17] Luo, Z., *et al*. (2024) Cheap and Quick: Efficient Vision-Language Instruction Tuning for Large Models. *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, *CVPR* 2024, Seattle, 16-22 June 2024, 42444-42457.

[18] Team Gemini (2025) Gemini 1. 5: Unlocking Multimodal Understanding Across Millions of Tokens of Context.

[19] OpenAI (2023) GPT-4V(ision) System Card & Benchmark Results.

[20] Liu, H., *et al*. (2023) Visual Instruction Tuning (LLaVA). *Proceedings of the* 37*th International Conference on Neural Information Processing Systems*, New Orleans, 10-16 December 2023, 34892-34916.