| **论文标题（review+article）** | **发表来源** | **核心技术** |
| --- | --- | --- |
| 《An Overview Recent Trends and Challenges in Multi-Modal Image Retrieval Using Deep Learning》 | 2024 International Conference on Communication, Computing and Energy Efficient Technologies | 深度学习驱动的跨模态图像检索：特征提取与融合、多模态嵌入空间学习、注意力机制、对抗训练、评估指标优化 |
| 《Multi-modal REtrieval Augmented Generative Commonsense Reasoning (MORE)》 | NeurIPS 2024  多模态生成与推理领域顶会 | 跨注意力机制 + 软提示技术，文本 - 图像联合增强常识推理 |
| 《Unlocking the Multi-Modal Capability of Dense Retrieval via Visual Module Plugin (MARVEL)》 | ACL 2024  自然语言处理与多模态检索顶会 | 视觉模块插件 + 图像 - 标题对比预训练，统一文本 - 图像编码 |
| 《Visualized Text Embedding For Universal Multi-Modal Retrieval (VISTA)》 | ACL 2024 | 灵活架构 + 两阶段训练，文本编码器与图像理解融合 |
| 《Effective deep learning-based multi-modal retrieval》 | VLDB 2016  DB领域顶会 | 提出深度学习方法实现多模态检索，通过映射函数将不同媒体投影到公共空间，显著提升搜索准确率。 |
| 《Multi-modal image retrieval with random walk on multi-layer graphs》 | IEEE ISM 2016  多媒体领域检索顶会 | 多层图随机游走算法，融合多视图特征（颜色、边缘、文本） |

**技术总结/技术发展（依据 综述：An Overview Recent Trends and Challenges in Multi-Modal Image Retrieval Using Deep Learning）**

传统图像检索依赖单一模态（视觉特征或文本标注），无法捕捉多源数据的丰富语义，存在 “语义鸿沟” 问题。随着深度学习发展，多模态图像检索（融合视觉、文本、音频等模态）成为研究热点。本文核心目标是：系统梳理深度学习驱动的多模态图像检索技术趋势、架构设计、数据集与评估指标，剖析当前挑战并指出未来方向，为该领域研究提供全局参考。

（一）多模态特征提取技术（按模态分类）

1.1视觉模态技术

卷积神经网络（CNNs）：2012 年 AlexNet 后成为视觉特征提取核心，可自动学习颜色、纹理、空间布局等高层特征，解决传统手工特征（如 SIFT）泛化性差的问题，广泛用于图像分类、目标检测等基础任务，是多模态检索的视觉特征基础。

注意力机制：2017 年 Transformer 提出后快速普及，包括空间注意力（聚焦图像关键区域）、通道注意力（强化有效特征通道）、自注意力（捕捉像素间关联），可提升视觉特征的语义针对性，减少冗余信息干扰。

1.2文本模态技术

循环神经网络（RNNs）：2010-2015 年主流文本处理模型，擅长捕捉文本序列关系，用于处理图像标注、场景文本识别等，解决了传统词袋模型无法建模上下文的问题，但存在梯度消失缺陷（后被 LSTM/GRU 优化）。

联合嵌入技术：2016 年后兴起，将视觉与文本特征映射到共享向量空间，实现跨模态语义对齐，代表模型包括视觉语义嵌入（VSE）、多模态紧凑双线性池化（MCB），2021 年 CLIP 模型进一步通过大规模对比学习强化了联合嵌入的泛化性。

词嵌入与文本表示：2013 年 Word2Vec 提出后成为基础，将词语转化为低维向量；后续 BERT 等预训练模型（2018 年）实现句子级文本表示，为文本 - 视觉语义关联提供更精准的特征支持。

1.3音频模态技术

音频特征提取：2010 年前以传统信号处理为主，如梅尔频率倒谱系数（MFCC）、 spectrogram；2015 年后结合深度学习，将音频转化为数值特征，与视觉、文本特征融合用于多模态检索（如视频场景检索）。

跨模态融合：2018 年后主流方法为 “表示提取 - 模态对齐 - 融合建模”，通过注意力机制或双线性池化，整合多模态互补信息，减少模态异质性带来的融合难度。

（二）深度学习核心架构

孪生网络与三元组网络：2015 年后用于相似度度量，通过共享权重的子网络提取输入特征，计算相似度（孪生网络）或优化 “锚点 - 正例 - 负例” 距离（三元组网络），适用于少样本多模态检索（如医学图像检索）。

深度跨模态检索架构：2016 年后快速发展，核心是学习 “跨模态共享空间”，解决传统单模态检索的模态局限，支持 “图像查文本”“文本查图像” 等场景，典型应用包括图像 caption 检索、跨域图像匹配。

生成对抗网络（GANs）：2014 年 GANs 提出后，2018 年开始用于多模态检索，通过生成器生成跨模态样本（如文本生成图像特征）、判别器优化特征对齐，可缓解模态数据稀缺问题，提升检索鲁棒性，但存在训练不稳定缺陷。

四、技术发展时间线（按时间先后）

2010 年前：传统技术阶段，视觉依赖手工特征（SIFT、HOG），文本依赖词袋模型，音频依赖 MFCC，多模态融合仅为简单特征拼接，语义鸿沟显著。

2010-2015 年：单模态深度学习兴起，CNN（2012）主导视觉特征，RNN/LSTM（2013）优化文本处理，多模态检索开始尝试 “单模态特征 + 简单融合”，数据集以 Flickr30K（2013）、CUB-200-2011（2011）为主。

2016-2020 年：多模态融合深化，联合嵌入技术（VSE、MCB，2016）、孪生网络（2015）、跨模态检索架构（2016）成为核心，GANs（2018）引入多模态生成，MSCOCO（2014）成为主流数据集，评估指标 MAP、NDCG 广泛应用。

2021-2024 年：大模型驱动阶段，CLIP（2021）通过大规模对比学习革新联合嵌入，Transformer 注意力机制（2017 后）成为融合核心，细粒度检索、可扩展性优化成为重点，真实场景应用（如遥感图像检索）增多。

五、关键数据集与评估指标

核心数据集：MSCOCO（2014，12.3 万图像 + 5 句 / 图标注）、Flickr30K（2013，3.1 万图像 + 5 句 / 图标注）、CUB-200-2011（2011，1.1 万鸟类图像 + 细粒度标注），覆盖通用场景与细粒度场景。

核心评估指标：Precision/Recall（基础准确率）、MAP（平均精度均值，衡量排序质量）、NDCG（归一化折损累积增益，考虑结果排名权重），其中 MAP 和 NDCG 是多模态检索的主流指标。

六、当前挑战与未来方向

核心挑战：异质数据融合（模态分布差异大）、跨模态学习（域适应难）、可扩展性（大规模数据集检索效率低）、细粒度检索（ subtle 特征捕捉难）、模型可解释性（深度学习 “黑箱” 问题）。

未来方向：研发新型融合技术（如自适应模态权重）、强化迁移学习（跨域知识复用）、探索分布式检索架构（提升大规模数据效率）、结合多任务学习（缓解数据稀缺）、推进真实场景落地（如医疗、遥感）。

**多模态图像检索，multi-modal image retrieval**

**科普入门：**

**https://blog.csdn.net/m0\_59164520/article/details/142731035**

**https://www.bilibili.com/opus/985894512877371401**

**一、CSDN 文章《一文彻底搞懂多模态 - 多模态检索》核心总结（基础概念 + 学习路径）​**

1. 先搞懂 “单模态检索”：多模态的基础​

指只用一种数据类型（文本 / 图像 / 音频）查同类数据，是多模态的 “前身”：​

文本检索（最常见）：比如搜 “论文答辩技巧” 找相关文档，核心靠 3 步：​

分词建索引（把文本拆成词 / 短语，方便快速定位）；​

优化查询（比如识别 “答辩” 和 “ defence” 是一个意思）；​

按相似度排序（优先显示最相关的文档），用到倒排索引、向量空间模型等工具。​

图像检索：比如搜 “猫咪图片” 找相似图，核心靠 CNN（卷积神经网络）：​

先提特征（用 CNN 提取图片的颜色、纹理、形状等信息）；​

再比相似（算查询图和数据库图的特征相似度）；​

最后优化速度（用哈希、量化技术让检索更快）。​

音频检索：比如搜 “游戏音效” 找相似片段，靠分析音频的频谱（声音的 “频率画面”）、节奏、音色，再用算法比相似度。​

2. 核心：多模态检索怎么工作？​

指用多种数据（比如文本 + 图像）查信息，核心是 “让不同数据‘说同一种语言’”，4 步走：​

第一步：数据转向量（Embedding）：用 CNN（处理图像）、Transformer（处理文本）等模型，把文本、图像、音频都转换成机器能懂的 “向量”（类似一串数字）。​

第二步：放进 “共同向量空间”：把不同模态的向量放到同一个 “空间” 里（比如把文本向量和图像向量都放进坐标系），这步叫 “跨模态对齐”，确保它们能互相比较。​

第三步：模态融合：把多种向量 “合并”，提取共同特征（比如文本 “黑色猫咪” 和猫咪图片的向量融合后，能更精准对应 “黑色猫咪” 这个语义）。​

第四步：比相似排序：用 “余弦相似度”“欧氏距离” 等简单方法，算查询向量（比如你输的 “黑色猫咪” 文本）和数据库向量的相似度，按高低返回结果。​

3. 进阶：跨模态检索（“跨类型” 查数据）​

指用一种数据查另一种数据，比如用文本 “黑色猫咪趴在地上” 找对应的图片，核心流程和多模态类似，但更强调 “语义对齐”：​

关键：让不同模态的向量 “语义一致”（比如文本 “猫咪” 和猫咪图片的向量，在空间里的位置要靠近，代表同一个意思）；​

举例子：用文本查图像时，先把文本转向量，再在图像向量库中找最像的，返回对应的图片。​

4. 实用：多模态大模型怎么学？​

文章给了 “从入门到就业” 的 4 阶段学习路径（约 2 个月能入门）：​

初阶（10 天）：搞懂大模型基础，比如 “怎么用代码连 GPT-3.5”“怎么写 Prompt 让 AI 更听话”；​

高阶（30 天）：实战做工具，比如搭 “ChatPDF”（上传 PDF 就能对话）、学向量数据库（存向量的数据库，多模态检索必备）；​

训练（30 天）：学微调模型，比如训练自己的 “宠物图片检索模型”，懂 Transformer 结构、轻量化微调；​

商业（20 天）：学部署和落地，比如在阿里云部署模型、处理内容安全、算法备案（企业落地必备）。​

**二、B 站文章《性能赶超 GPT-4! 多模态检索最新成果刷爆 SOTA!》核心总结（2024 年最新技术）​ 技术点对应上面几篇会议文章**

1. MORE 框架：让 AI 更有 “常识” 的多模态检索​

解决的问题：以前的语言模型（比如 GPT-3.5）查信息时，不会结合图像补全常识（比如查 “滑雪”，只看文本可能漏 “雪山环境” 的信息）。​

怎么做：结合文本 + 图像一起检索，用 “跨注意力机制” 挑有用信息（比如忽略无关的风景图，只留滑雪者 + 雪山的图），还会故意用 “无关结果” 训练（让模型学会 “没用的信息就忽略”）。​

效果：在 “常识生成任务”（比如按 “滑雪” 生成描述）中，比 GPT-4 还强 ——GPT-4 的 Bleu4（衡量生成质量的指标）是 30.0，MORE 能到 32.53。​

2. MARVEL 模型：让 “文本检索器” 能懂图像​

解决的问题：以前的检索模型要么只懂文本，要么只懂图像，很难兼顾。​

怎么做：给成熟的 “文本检索器”（比如 T5-ANCE）装个 “视觉插件”—— 用 CLIP（图像模型）提图像特征，再用 “图像 - 标题对比训练”（比如让模型学 “图书馆图片” 和 “Library Reading Room” 文本是一对），微调时只改文本模型参数，成本低。​

效果：在所有多模态检索基准测试中都拿了 SOTA（当前最好成绩），还建了大规模数据集 ClueWeb22 方便大家研究。​

3. VISTA 嵌入模型：让文本和图像 “融合更深”​

解决的问题：以前的模型很难让文本和图像的向量 “完全对齐”（比如文本 “红色苹果” 和红色苹果图的向量可能不够近）。​

怎么做：​

架构灵活：把强文本编码器和 ViT（图像编码器）结合，让文本和图像的 “标记”（比如文本的 “苹果”、图像的 “苹果区域”）直接交互；​

数据够多：自动生成大量高质量图文对（比如给每张图配精准文本）；​

分两阶段训：先用水印数据对齐文本和图像，再用高质量数据优化。​

效果：在多模态检索任务中表现炸裂，比如 R@5（前 5 个结果命中的概率）达 96.3%，远超之前的模型（比如 GeMKR 的 90.3%）。​

4. SciMMIR 基准：给 “科学领域” 多模态检索定标准​

解决的问题：以前的多模态数据集多是日常场景（比如猫咪、风景），没有专门给科学领域（比如论文、实验数据）用的。​