深度学习在组合图像检索中的应用综述

王治浩　周新 *
大连海事大学　辽宁省大连市　116000

摘　要：图像检索作为计算机视觉与信息检索交叉的重要研究领域，近年来在深度学习技术的驱动下迎来了显著的发展。传统方法依赖手工设计的特征提取手段，往往在面对大规模数据时暴露出适应性不足与表达能力受限的短板。而深度学习的出现，通过自动特征学习和端到端的训练方式，为这一领域带来了深刻的变革。其中，卷积神经网络凭借其强大的特征提取能力，成为技术革新的核心支柱。与此同时，近年来兴起的 Transformer 架构、哈希编码技术以及注意力机制等新方法，进一步推动了检索精度与效率的提升。本文特别关注“组合图像检索”这一新兴研究方向，探讨了 CNN 与 Transformer 在处理局部细节和全局语义特征时的独特优势。此外，通过多模态特征融合以及深度学习与传统方法的结合，检索性能得以显著增强。然而，当前应用仍面临若干挑战：对大规模标注数据的依赖性、计算资源的高消耗，以及模型解释性不足等问题不容忽视。展望未来，随着轻量级模型设计和自监督学习等技术的不断成熟，这些难题有望逐步破解，从而推动图像检索技术迈上新的台阶。这一领域的持续探索，不仅体现了学术研究的深度，也预示着广阔的应用前景。

关键词：图像检索；深度学习；组合图像检索；特征提取；CNN；Transformer；多模型融合；自监督学习

1. 引言

1.1 图像检索的背景与意义

图像检索作为信息检索领域的重要分支，旨在从大规模图像数据库中快速、准确地找出与查询图像内容相似的图像。随着互联网和多媒体技术的迅猛发展，图像数据呈现出爆炸式增长的趋势。例如，社交媒体平台每天产生数以亿计的图像，电子商务网站依赖图像展示商品，医学领域需要处理大量的诊断影像。这些海量图像数据的管理和高效检索已成为亟待解决的关键问题。图像检索技术的意义不仅体现在学术研究中，如推动计算机视觉和模式识别领域的发展，还在实际应用中发挥着不可替代的作用，包括数字图书馆的资源管理、电子商务的商品推荐、医学影像的辅助诊断以及安全监控的目标识别等。

1.2　深度学习在图像检索中的兴起

深度学习作为一种基于多层神经网络的机器学习方法，自 2012 年 AlexNet 在 ImageNet 图像分类竞赛中取得里程碑式成果以来，迅速改变了计算机视觉的研究格局。其核心优势在于能够自动学习数据的特征表示，这一特性在图像检索领域尤为重要。传统图像检索依赖手工设计的特征，如颜色直方图或 SIFT，这些特征在复杂场景下的表达能力有限。而深度学习，尤其是卷积神经网络，通过多层卷积和池化操作，能够提取图像的高层语义特征，显著提升检索的准确性。

1.3“组合图像检索”的概念

在深度学习时代，组合图像检索的具体实现形式多样。一方面，可以通过多模型特征融合，利用多个深度学习模型提取的特征进行集成；另一方面，可以将深度学习特征与传统手工特征相结合，发挥两者的互补优势。此外，跨模态特征组合也成为重要方向，例如在图像 – 文本检索任务中融合视觉特征和语义特征。刘萌等对基于深度学习的图像 – 文本匹配研究进行了综述，指出深度学习在跨模态检索中的广泛应用及其在提升检索准确性方面的显著效果 [4]。然而，组合图像检索也面临挑战，如特征融合策略的选择和计算复杂度的增加，如何设计高效的组合方法成为当前研究的重点。

2. 传统图像检索方法回顾

传统 CBIR 方法的局限性在面对大规模和多样化的图像数据集时尤为明显。首先，手工设计的特征通常局限于低级语义，无法理解图像的深层含义。例如，颜色直方图可能无法区分具有相似颜色分布但语义完全不同的图像。其次，特征提取和表示的效率较低，难以满足海量图像实时检索的需求。此外，相似性度量的选择也对检索结果影响显著，常用的欧氏距离或余弦相似度在高维特征空间中容易遭遇“维度

灾难”，从而降低检索的准确性。因此，研究者提出了更高级的局部特征描述符。例如，Lowe 提出的 SIFT[5] 通过检测图像中的关键点并提取其局部不变特征，在图像匹配和物体识别中取得了广泛应用；Dalal 和 Triggs 提出的 HOG[6] 则通过计算局部区域的梯度方向直方图，有效描述了物体的形状和边缘信息。这些方法的引入显著提升了特征的鲁棒性和判别性，但在高层语义理解和大规模检索效率方面仍存在明显短板。

2.1 深度学习在图像检索中的早期应用

深度学习技术的突破，尤其是卷积神经网络的兴起，为图像检索领域带来了革命性的变化。与传统方法依赖手工特征不同，CNN 通过多层卷积和池化操作，能够自动学习图像的特征表示，极大地提升了特征的表达能力。早期，研究者尝试将预训练的 CNN 模型直接应用于图像检索任务，开启了深度学习在该领域的初步探索。Krizhevsky 等在 2012 年提出的 AlexNet[7] 是一个标志性工作，其在 ImageNet 图像分类竞赛中的成功不仅推动了深度学习的发展，也为图像检索提供了新的思路。Razavian 等 [8] 在研究中发现，使用预训练的 AlexNet 模型提取的卷积层特征，在多个图像检索基准数据集上的表现显著优于传统手工特征，验证了深度学习特征的优越性。这一发现表明，深度学习模型在大规模数据集上学习到的通用特征具有很强的迁移能力，可以直接应用于检索任务。

2.2 深度学习特征提取方法

特征提取作为图像检索的核心环节，直接决定了检索系统的性能优劣。传统的手工特征设计依赖于研究者对图像特性的先验知识，而深度学习的出现彻底改变了这一局面。在组合图像检索中，深度学习特征提取方法主要包括卷积神经网络和 Transformer 两大类，二者共同推动了检索技术的前沿发展。

2.2.1 卷积神经网络

卷积神经网络自 2012 年 AlexNet 在 ImageNet 竞赛中取得突破性成果以来，已成为计算机视觉领域的支柱性技术，其在图像检索中的应用尤为广泛。CNN 通过多层卷积和池化操作，能够提取图像从边缘纹理到语义信息的特征，形成具有高度抽象性的全局表示。这种层次化的特征学习能力，使其在图像检索任务中表现出强大的适应性和准确性。然而，CNN 也存在一定局限性。由于其依赖局部感受野的特性，CNN 在处理全局上下文信息时往往力不从心。

2.2.2 Transformer 在图像检索中的应用

Transformer 核心的自注意力机制赋予了模型捕捉序列中长程依赖关系的能力。随着 Vision Transformer 的提出，Transformer 在图像检索领域展现出巨大潜力。与 CNN 的局部感受野不同，Transformer 通过全局自注意力机制，能够同时建模图像中所有像素或区域之间的关系，为图像检索提供了全新的特征提取范式。

在图像检索中，Vision Transformer 是最具代表性的应用之一。Dosovitskiy 等人首次提出 ViT，将图像分割为固定大小的图像块，并将其作为序列输入 Transformer 编码器，从而学习图像的全局表示 [15]。他们在图像分类任务上的成功验证了 ViT 的优越性，随后研究者开始将其应用于图像检索。例如，El-Nouby 等人基于 ViT 提出了一种图像检索方法，通过 Transformer 的全局建模能力提取鲁棒特征，在 Revisiting Oxford 和 Revisiting Paris 数据集上显著提高了检索准确性 [16]。这一成果表明，Transformer 在处理复杂图像场景时，能够有效捕捉全局依赖关系，从而提升特征的表达能力。

2.3 组合多种特征提取方法

尽管 CNN 和 Transformer 在图像检索中取得了显著进展，但单一特征提取方法往往难以全面捕捉图像的多样化信息。在组合图像检索的范式下，研究者开始探索通过整合多种特征提取方法，构建更全面、更鲁棒的图像表示，以进一步提升检索性能。

2.3.1 多模型特征融合

多模型特征融合的核心思想在于，通过集成多个深度学习模型提取的特征，构建一个综合的特征表示，从而提升检索的准确性。例如，Khayyat 和 Elrefaei 提出了一种多层次特征融合方法，结合多个 CNN 模型在不同层次提取的特征，通过加权融合生成最终的特征向量 [17]。他们在多个图像检索数据集上的实验结果表明该方法显著优于单一模型，特别是在处理复杂场景时表现尤为突出。Yang 等人通过综述系统回顾了基于深度学习的草图图像检索方法，指出多模型融合在跨模态任务中的重要性 。他们的研究表明，通过融合 CNN 在不同深度的特征，可以有效提升草图检索的鲁棒性。

2.3.2 深度学习与传统方法的结合

在实际研究中，这种结合策略已经展现出了显著的效果。例如，方潜生等人提出了一种将 HOG（方向梯度直方

图) 与深度特征相融合的方法，用于草图到图像的检索任务。HOG 特征以其捕捉图像边缘和形状信息的卓越能力著称，而深度学习特征则擅长提取更高层次的语义信息。两者结合后，检索的准确性得到显著提升。这种方法尤其适用于用户手绘草图的检索场景，因为它巧妙地利用了传统特征的局部描述能力和深度学习的全局理解能力。

3. 优势

3.1 自动特征学习

自动特征学习是深度学习在图像检索领域最核心的优势之一。传统图像检索方法依赖人工设计的特征，这些特征的提取需要大量的领域知识和手动调优，难以适应复杂多样的图像内容。而深度学习的神经网络的层次结构能从原始数据中自主学习特征表示，避免了繁琐的特征工程过程。这一特性在组合图像检索中尤为关键。

3.2 端到端训练

端到端训练作为深度学习的一项核心特性，在图像检索领域展现出显著的价值，尤其是在组合图像检索这一复杂任务中，其作用尤为突出。相较于传统方法，端到端训练将原本割裂的特征提取、特征表示和相似性度量等环节，整合进一个统一的优化框架，直接针对最终的检索目标进行调整。这种方式不仅简化了模型设计的过程，还大幅提升了系统的整体性能，避免了传统方法因各步骤独立优化而产生的误差传递问题。

传统的图像检索流程通常将任务分解为若干独立模块：首先通过设计的算法提取图像特征，然后构造特征表示，最后利用特定的相似性度量方法计算图像间的匹配程度。这种分步优化的方式虽然直观，但在实际应用中往往面临一个难题：每个模块的优化目标可能并不完全一致，导致误差在环节间逐步累积，最终影响检索的精度和效率。而深度学习的端到端训练则完全不同，它通过一个整体的神经网络结构，将所有步骤无缝衔接起来，直接以检索指标作为优化目标。这种一体化的训练方式使得模型能够更专注于捕捉与任务直接相关的特征，从而显著提升效果。

4. 局限性

4.1 对大规模数据集的依赖

以实例级图像检索为例，Babenko 等人指出，高质量的检索数据集（如 Landmarks 数据集）往往规模有限，远不及 ImageNet 这类分类数据集的丰富性，这直接限制了深度学习模型的训练效果 [9]。在组合图像检索中情况更为复杂。例如，在跨模态检索任务中，需要同时具备图像和文本的标注数据，而现有数据集的规模和多样性仍不足以满足复杂模型的需求。刘萌等人在其综述中提到，跨模态数据集的稀缺性是制约深度学习在该领域发展的主要瓶颈之一 [4]。在实际中，我们常通过数据增强或迁移学习缓解这一问题。未来，自监督学习和少样本学习等技术可能成为解决数据依赖问题的关键方向。

4.2 计算资源需求

以近年来兴起的 Transformer 架构为例，Dosovitskiy 等人提出的 Vision Transformer 在图像检索中表现出色，但其处理高分辨率图像时，内存和计算资源需求急剧增加，普通硬件难以承受 [15]。为解决这一问题，刘泽华等人提出了 Swin Transformer，通过分层设计和窗口注意力机制降低了计算复杂度，但在实际应用中仍需高性能 GPU 支持 [17]。在组合图像检索中，多种特征提取方法的融合进一步加剧了资源消耗。

5. 挑战与展望

5.1 当前研究的挑战

随着深度学习模型结构的日益复杂，其在图像检索中的应用虽然带来了性能提升，但也伴随着可解释性不足的问题。可解释性直接关乎模型的可靠性和用户信任。以 Vision Transformer 为例，Dosovitskiy 等人提出的这一架构在图像检索任务中表现出色，其自注意力机制能够有效捕捉全局特征，但其复杂的计算过程使得研究者难以追踪模型的决策依据 [15]。他们在实验中虽然验证了 ViT 的高性能，却也坦言模型的内部逻辑仍需进一步剖析。

在组合图像检索任务中，模型复杂性问题尤为突出。Khayyat 和 Elrefaei 提出了一种多模型特征融合方法，通过集成多种特征提取技术显著提升了检索精度 [17]。然而，这种方法在提升性能的同时，也使得模型的内部机制更加隐晦，调试和优化变得异常困难。

5.2 未来研究方向

自监督学习作为一种新兴范式，通过利用未标注数据自动生成监督信号，为解决数据稀缺问题提供了新的可能性。在图像检索领域，自监督学习的应用正迅速升温。Chen 等人提出的 SimCLR 方法通过对比学习在未标注图像上训练模型，其性能已可媲美有监督学习 。他们在实验中展示了自

监督预训练模型在下游任务中的优越性，尤其是在数据量有限时表现尤为突出。He 等人则将自监督学习直接应用于图像检索任务，他们通过在大型未标注数据集上预训练模型，显著提升了特征提取的鲁棒性和检索精度 。

参考文献：

[1] 甄俊杰 , 应自炉 , 赵毅鸿 , 等 . 深度学习和迭代量化在图像检索中的应用研究 [J]. 信号处理 , 2019(5):7.DOI:CNKI:SUN:XXCN.0.2019-05-025.

[2] 贺周雨、冯旭鹏、刘利军、黄青松 . 面向大规模图像检索的深度强相关散列学习方法 [J]. 计算机研究与发展 , 2020, 57(11):14.DOI:10.7544/issn1000-1239.2020.20190498.

[3] 张凯 , 姚宇 , 伍岳庆 , 等 . 深度哈希算法在心脏超声图像检索中的应用 [J]. 计算机应用 , 2019, 39(S02):6.DOI:CNKI:SUN:JSJY.0.2019-S2-015.

[4] 刘萌 , 齐孟津 , 詹圳宇 , 等 . 基于深度学习的图像 – 文本匹配研究综述 [J]. 计算机学报 , 2023, 46(11):2370-2399. DOI:10.11897/SP.J.1016.2023.02370.

[5] 彭晏飞 , 宋晓男 , 武宏 , 等 . 结合深度学习与相关反馈的遥感图像检索 [J]. 中国图象图形学报 , 2019(3):15.DOI:CNKI:SUN:ZGTB.0.2019-03-010.

[6] 赖心瑜 , 陈思 , 严严 , 等 . 基于深度学习的人脸属性识别方法综述 [J]. 计算机研究与发展 , 2021(012):058.

[7] 方潜生 , 李惠 , 苏亮亮 , 等 . 基于 HOG 与深度特征融合的草图 – 图像检索 [J]. 计算机仿真 , 2023, 40(8):258-263. DOI:10.3969/j.issn.1006-9348.2023.08.050.

[8] 苗壮 , 赵昕昕 , 李阳 , 等 . 基于 Swin Transformer 的深度有监督哈希图像检索方法 [J]. 湖南大学学报 : 自然科学版 , 2023, 50(8):62-71.

[9] 李志义 , 许洪凯 , 段斌 . 基于深度学习 CNN 模型的图像情感特征抽取研究 [J]. 图书情报工作 , 2019, 63(11):12. DOI:10.13266/j.issn.0252-3116.2019.11.011.

[10] 万方 , 强浩鹏 , 雷光波 . 自监督深度离散哈希图像检索 [J]. 中国图象图形学报 , 2021, 026(011):2659-2669.

[11] 崔少国 , 熊舒羽 , 刘畅 , 等 . 基于深度哈希卷积神经网络的医学图像检索 [J]. 重庆理工大学学报（自然科学版）, 2020(008):034.

[12] DUAN Wenjing,CHEN Shaoping. 具备高层语义特征的离散哈希图像检索算法 [J]. 计算机工程与应用 , 2019, 55(13):6.DOI:10.3778/j.issn.1002-8331.1804-0057.

[13] 单连平 , 窦强 . 基于深度学习的海战场图像目标识别 [J]. 指挥控制与仿真 , 2019, 41(1):1-5.

[14] 彭金喜 , 苏远歧 , 薛笑荣 . 基于深度学习和同生矩阵的 SAR 图像纹理特征检索方法 [J]. 计算机科学 , 2019, 46(B06):5.DOI:CNKI:SUN:JSJA.0.2019-S1-040.

[15] Chen W, Liu Y, Wang W, et al. Deep learning for instance retrieval: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(6): 7270-7292.

[16] Liu P, Guo J M, Wu C Y, et al. Fusion of deep learning and compressed domain features for content-based image retrieval[J]. IEEE Transactions on Image Processing, 2017, 26(12): 5706-5717.

[17] Khayyat M M, Elrefaei L A. Manuscripts image retrieval using deep learning incorporating a variety of fusion levels[J]. IEEE Access, 2020, 8: 136460-136486.