Computer Science and Application 计算机科学与应用, 2025, 15(11), 33-41
Published Online November 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.1511280

Hans汉斯

基于CNN和视觉Transformer的哈希图像检索算法综述

任　欢[1]，赵虹阳[1]，刘小华[1,2*]

[1]新疆理工职业大学人工智能学院，新疆　图木舒克
[2]深圳职业技术大学人工智能学院，广东　深圳

收稿日期：2025年9月29日；录用日期：2025年10月26日；发布日期：2025年11月3日

摘　要

图像检索的核心目标是从预设的图像数据库中，精准定位并提取出与给定查询图像属于同一类别的所有相关图像。然而，由于传统算法通常采用简单的线性变换来构建哈希函数，并且参数优化中需要人为手动操作。因此，传统的检索方法往往存在着较大的提升空间。近年来，深度学习和哈希技术融合在拥有高检索效率同时拥有较高的检索准确度为图像检索领域提供了新思路。本文综述了各种深度哈希方法，评估了不同类别方法的原理及特性进行介绍，对各种方法的优缺点进行分析，实验结果表明，基于深度学习的哈希图像检索方法取得了较高的检索准确性。最后展望了深度学习在优化算法和计算能力方面的潜力，预测其将在图像检索中起到越来越关键的作用，为实际应用提供更精准的技术支持。

关键词

图像检索，深度学习，哈希

Survey on Hash Image Retrieval Algorithms Based on CNN and Vision Transformer

Huan Ren[1], Hongyang Zhao[1], Xiaohua Liu[1,2*]

[1]College of Artificial Intelligence, Xinjiang Vocational University of Technology, Tumushuke Xinjiang
[2]School of Artificial Intelligence, Shenzhen Polytechnic University, Shenzhen Guangdong

Received: September 29, 2025; accepted: October 26, 2025; published: November 3, 2025

Abstract

The core objective of image retrieval is to precisely locate and extract all relevant images belonging

*通讯作者。

文章引用: 任欢, 赵虹阳, 刘小华. 基于 CNN 和视觉 Transformer 的哈希图像检索算法综述[J]. 计算机科学与应用, 2025, 15(11): 33-41. DOI: 10.12677/csa.2025.1511280

to the same category as a given query image from a predefined image database. However, traditional algorithms typically employ simple linear transformations to construct hash functions, requiring manual parameter optimization. Consequently, conventional retrieval methods often exhibit significant room for improvement. In recent years, the integration of deep learning and hashing techniques has provided new insights for image retrieval, offering both high retrieval efficiency and accuracy. This paper reviews various deep hashing methods, evaluates the principles and characteristics of different categories of approaches, analyzes the advantages and disadvantages of each method, and presents experimental results demonstrating that deep learning-based hashing image retrieval methods achieve high retrieval accuracy. Finally, it explores the potential of deep learning in optimizing algorithms and computational capabilities, predicting that it will play an increasingly critical role in image retrieval, providing more precise technical support for practical applications.

Keywords

Image Retrieval, Deep Learning, Hash

1. 引言

图像检索，作为计算机视觉领域的重要研究课题，旨在实现在庞大的数据集中准确地找寻相似的图像。图像检索技术涉及广泛的应用场景，如各大购物软件、搜索引擎、医疗辅助诊断、智能监控等等。对图像检索技术的研究，不仅推动了深度学习技术的发展，更助力整个计算机视觉领域不断前进。当前，卷积神经网络、视觉 Transformer 等深度学习模型的研究热度居高不下，其在图像检索领域的应用探索也在持续深化。

早期的图像检索问题，研究人员主要使用基于文本的图像检索(Text-based Image Retrieval, TBIR) [1]，通过手工文本来标注图像。TBIR 借助标注后的文本信息表征图像语义，最终通过图像与文本的匹配完成图像检索任务。但该方式通常需投入大量人力，用于图像信息的标注与管理，成本较高。此外，由于人类认知的主观差异和不同国家语言表述的差异，手工文本标注的歧义很大程度上影响了检索的准确性。基于此，研究人员开始探索基于内容的图像检索(Content-based Image Retrieval, CBIR) [2]。CBIR 主要依赖图像中颜色、形状、纹理等可视化信息作为图像检索的依据。然而，单独使用 CBIR 代价过于高昂和低效，现有的图像检索方法将 CBIR 和哈希技术结合起来进行特征学习。基于哈希的图像检索不仅在存储消耗存在很大优势，并且利用哈希函数进行学习极大的提高了检索效率。目前主流的哈希方法是使用 CNN (Convolutional Neural Network, CNN)和视觉 Transformer [3] (Vision Transformer, ViT)来挖掘特征信息，并使用哈希函数将其映射到二进制哈希码中。CNN 提取从低级到高级的图像特征，并通过权值共享和局部连接性减少模型参数。ViT 作为 Transformer 的扩展应用，允许通过自注意力捕获不同位置之间的依赖关系，进行全局关系建模。作为深度学习领域的突破性技术，CNN 和 ViT 始终占据关键地位，在此基础上也涌现出了诸多衍生模型。

本文首先对传统的哈希技术进行了简单的回顾，然后介绍了基于深度学习的哈希方法，主要针对 CNN 和 ViT 作为特征提取骨干网络的图像检索方法，以及常用数据集和算法对比分析。最后，对各种哈希方法的检索性能进行总结分析，并对深度哈希的未来进行展望。

2. 基于传统方法的图像检索

依据训练时对数据库数据的依赖与否，哈希方法可划分为两种类型。其中，数据无关哈希方法的哈希函数生成过程独立于数据集，通常由人工设计或随机生成，另一类则是数据相关哈希方法。Datar 等人[4]等人提出的局部敏感哈希(Locality Sensitive Hashing, LSH)是数据无关的经典代表之一，LSH 将元素数据空间中的两个相邻数据点通过相同的映射或投影变换后在新的数据空间中仍然相邻的概率很大，而不相邻的数据点概率很小。但 LSH 方法未考量原始数据特征，往往需依赖较长的哈希码才能实现理想检索效果。与之不同，数据相关方法在检索时会纳入原始数据信息；而依据构造函数过程中是否借助数据标签信息约束训练，数据相关哈希方法又可主要划分为有监督与无监督两类。

图像检索的无监督哈希技术的核心特征在于其无需依赖图像的标签信息，而是借助深度特征构建相似度矩阵，并在此基础上通过哈希函数完成哈希码的学习过程。回溯哈希算法在图像检索领域的早期应用，传统哈希方法是重要起点，其中具有代表性的包括 Weiss 等人[5]中提出的谱哈希(Spectral Hashing, SH)，该方法的核心思路是将哈希编码任务转化为图像分割问题。具体而言，就是把待处理图像划分为若干个不同的区域或部分。随后，为求解这一问题，SH 算法通过放松相关约束条件，将原本的图像分割任务转化为拉普拉斯特征图的降维问题，最终通过对降维问题的求解，得到图像数据对应的哈希编码。不过，这一方法存在明显缺陷：在计算辅助矩阵的过程中，会显著增加整体计算耗时，影响效率。另一项经典传统哈希方法是 Gong 等人[6]中提出的迭代量化(Iterative Quantization, ITQ)算法。该算法实现关键突破的核心，是聚焦于如何求解最优旋转矩阵这一问题，通过对这一核心问题的解决，有效优化了量化哈希码的生成流程，进而大幅降低了量化阶段产生的误差，在效率与精度平衡上实现了提升。从整体价值来看，无监督哈希方法为图像检索技术的发展提供了切实有效的技术支撑，对该领域的进步与实际应用推广起到了推动作用。但不可忽视的是，与监督哈希方法相比，无监督哈希方法的核心局限在于未利用数据集的标签信息参与哈希编码学习。这种信息缺失可能导致生成的哈希编码难以有效区分不同类别数据的特征差异，进而使得在后续的图像检索过程中，该方法存在一定难以规避的固有局限性。

有监督哈希方法的核心优势在于，其通过引入数据集的标签信息参与哈希码的学习过程，使得生成的哈希码能够更充分地保留数据间的语义信息。这一特性意味着，在有监督学习框架下，哈希编码的生成不仅会考量数据在特征层面的相似关系，还会主动纳入数据间的语义关联——例如不同样本在类别归属、场景属性等维度的内在联系。由此生成的哈希码，既能精准反映数据间的相似程度，又能更高效地捕捉它们背后的语义关联，最终为图像检索任务的准确性与效率提升提供关键支撑。在有监督哈希方法的早期发展阶段，涌现出多款经典算法。其中，Liu 等人[7]中提出的基于核函数的哈希方法(Supervised Hashing with Kernels, KSH)颇具代表性：该算法巧妙利用汉明距离与编码内积的等价特性，构建出兼具高效性与易优化性的目标函数；同时，依托内积的可分性优势，设计了一款贪婪算法，实现对哈希函数的逐比特求解；此外，通过引入基于核的哈希函数，有效解决了原始数据在特征空间中线性不可分的问题。另一项经典成果是 Shen 等人[8]中提出的监督离散哈希(Supervised Discrete Hashing, SDH)方法。该检索网络的关键技术突破在于：无需通过松弛策略简化问题，而是直接对离散优化问题进行求解，从而生成质量更高的哈希码。尽管相较于无监督哈希方法，有监督哈希在性能指标上实现了显著提升，但传统有监督哈希方法仍存在明显短板。一方面，这类方法多以简单线性变换构建哈希函数，且模型训练时需人工手动优化参数，这不仅提升了操作复杂度，也难以确保参数达到最优；另一方面，其性能严重依赖数据分布与算法构建假设的匹配程度，一旦实际数据分布偏离预设假设，所构建的哈希函数便难以实现预期检索效果。受这些不足影响，传统有监督哈希方法在性能优化与实用性拓展方面，仍有较大探索空间。

3. 基于深度学习的图像检索

深度哈希方法是图像检索领域备受关注的前沿方法，其通过深度神经网络将图像的特征映射到二维空间中，使得属于同一类别的图像具有相似的二进制编码，从而实现准确的图像检索。该方法能够极大的节省存储空间和计算资源。具体来说，通过 CNN 或视觉 Transformer 作为特征学习网络提取图像特征；然后，通过训练哈希函数将特征映射到相同长度的二进制编码中。训练过程可根据是否使用相似性标签的成对图像指导网络学习，使得具有相似的图像在编码空间中具有相似的编码；最后，通过学习到的哈希函数对需要检索的图像进行哈希编码，并通过比较哈希码之间的距离来确定图像之间是否相似。深度哈希图像检索将提取到的特征信息通过哈希函数转换为对应的二进制哈希码，相似图像的哈希码距离更近，不相似的图像的哈希码距离更远，从而获得最佳的检索效果。

卷积神经网络在计算机视觉领域表现突出，其核心通过卷积层、池化层与全连接层的协同作用，完成特征提取与分类任务。在卷积层中，多组卷积核会对输入图像执行卷积运算这一过程不仅能精准捕捉图像的局部空间特征，还借助权重参数共享机制，大幅降低了模型整体的参数规模，避免了传统模型参数冗余的问题。池化层的核心功能聚焦于卷积层输出特征图的降维处理：既能压缩特征图维度、降低后续计算负荷，又能最大程度保留分类所需的关键特征，进而显著优化模型的运算效率与泛化表现。全连接层则一般处于网络结构的末端，其作用是将前面卷积层与池化层提取到的高维特征，转化并映射到最终的输出类别或标签空间中；该层中每个神经元均与上一层所有神经元建立连接，以此实现对各类特征信息的全面整合，为最终分类决策提供支撑。在 CNN 的发展历程中，如 ResNet [9]是由何凯明等人在 2015 年提出的一种深度卷积神经网络，设计上采用模块化的思路，不同版本差异主要在深度和内部残差块的不同。ResNet-18 与 ResNet-34 设计上采用包含两层 3 × 3 的基本残差块。而 ResNet-50、ResNet-101 和 ResNet-512 等更深层版本则采用瓶颈结构的 1 × 1、3 × 3、1 × 1 三个卷积层组成。1 × 1 用于升维和降维操作，而 3 × 3 卷积用于特征提取，从而有效的降低了参数量与计算量，使得网络在深层结构下具备高效性。尤其是在缺少面对参数与计算资源的双重约束，瓶颈结构为模型深度延伸开辟了可行道路，最终推动数百层、乃至千层网络完成实际训练。大多数哈希图像检索方法采用 CNN 作为骨干网络进行特征提取，生成更准确的二进制哈希码。下面介绍一些使用 CNN 作为图像检索网络的方法。

经典的基于 CNN 深度哈希方法如 CNNH [10] (Convolutional Neural Network Hashing, CNNH)，通过将卷积神经网络和哈希映射结合起来，学习更精确的特征映射。Cao 等人[11]提出一种具有收敛性保证的深度学习哈希新架构，采用连续方法从不平衡相似性数据中精确学习二进制哈希码。Fan 等人[12]设计出深度极化网络(Deep Polarized Network, DPN)，用于监督学习精确的二进制哈希码。该网络能让不同类别间的哈希码具备高度可分离性，增大不同类之间的哈希码距离，减少同一类别的哈希码距离。Yuan 等人[13]提出了基于中心相似度的图像视频检索技术(Central Similarity Quantization, CSQ)。该网络核心是利用全局相似度度量方法推动相似数据对的哈希码向同一公共中心靠拢，同时促使不同数据对的哈希码分别聚集到各自不同的中心，从而提高哈希学习的效率以及检索的准确性。Xu 等人[14]提出哈希引导铰链函数(Hashing-guided Hinge Function, HHF)，应用于深度哈希检索。该函数采用一种新的铰链损失函数，对网络生成哈希码进行监督。它能避免哈希学习过程中陷入局部度量的最优最小值，解决网络学习里的语义丢失问题，进而促进更多特征的提取。虽然 CNN 作为网络架构在图像检索领域的性能上有优势，但仍存在潜在局限性。CNN 依靠卷积核计算小区域像素关系来提取图像特征，这导致 CNN 更侧重于局部信息的提取。近年来，研究人员发现，专注于全局特征提取的 ViT 在图像检索领域展现出不错的性能。

作为最近在计算机视觉领域出现的 CNN 的替代架构，首先更深入研究 Transformer 应用程序扩展 ViT 及其后续出现的变体是很重要的。Transformer 首次被提出是用来解决自然语言处理(Natural Language

Processing, NLP)数据问题。作为 Transformer 的一种扩展应用，ViT 展现出了强大的能力，能够有效地建模全局信息，为图像处理任务带来了新的可能性。Chen 等人[15]设计了骨干网络 ViT，以结合双流多粒度来学习特征，然后使用了贝叶斯学习方案。Li 等人[16]在使用 ViT 作为骨干网络以提高检索准确率的同时，提出了一种新的平均准确率损失。Ren 等人[17]设计了一种检索模型，该模型结合了对比学习和视觉转换器以及哈希技术。总的来说，ViT 在图像检索领域的应用处于蓬勃发展阶段，各种方法不断涌现[18]-[20]。

4. 实验

4.1 常用数据集

本文采用 CIFAR-10 [21]、NUS-WIDE [22]两个常用的公开图像数据集展开分析。CIFAR-10 数据集涵盖 60,000 张 RGB 图像，作为单标签数据集，它共划分 10 个类别且每个类别均包含 6000 张图片；在算法对比实验中，会从每个类别里随机选取 500 张图像用作训练集，同时从每个类别再选取 100 张图像作为测试集，不过该数据集的图像不仅像素大小固定，整体尺寸还相对较小，这使得图像所包含的细节信息较为匮乏，进而给检索任务带来了一定挑战，CIFAR-10 数据集图片如图 1 所示。

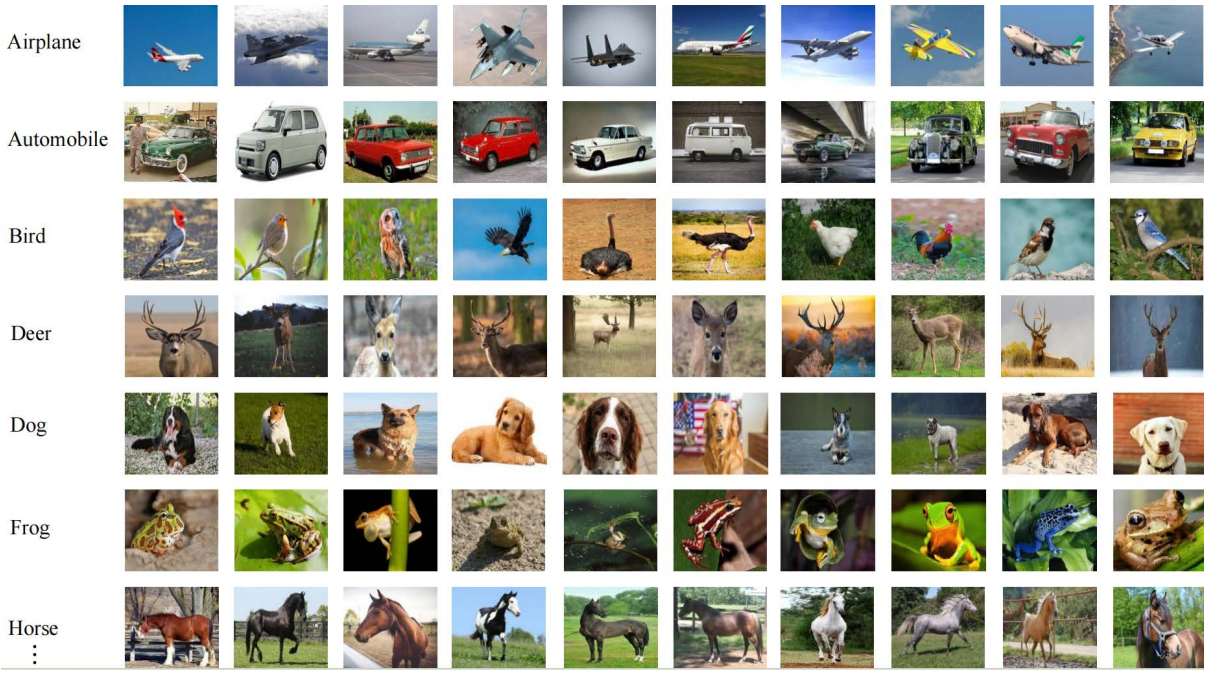

Figure 1. CIFAR-10 dataset images
图 1. CIFAR-10 数据集图片

NUS-WIDE 该数据集包含 269,648 张图像，是一个多标签数据集。对比实验共选取了 21 个类别的 195,834 张图像。然后，每个类别随机抽取 500 张图像构建训练集，每个类别随机抽取 100 张图像构建测试集。该数据集涵盖了多种不同的主题和内容，这使得它具有很好的代表性，适用于各种应用场景。但是，由于图像质量各异，有些图像拍摄质量较高，清晰度较高，而另一些可能存在模糊、光照不足等问题，这反映了真实世界中的图像多样性，使得检索任务具有挑战性。NUS-WIDE 数据集图片如图 2 所示。

4.2. 评价指标

本文选用平均精度均值(mAP)作为评估指标，该指标在图像检索系统性能评估中应用广泛。其中，平

均精度(AP)针对单次查询而言，指的是检索结果里所有真实相关项的准确率的平均值；而 mAP 则是对 N 次查询分别计算 AP 值后，再将这些 AP 值取平均，以此来综合衡量图像检索系统的整体表现。由于被检索的数据集往往包含数量庞大的图像，在实际计算 mAP 时，通常会限定仅采用前 k 个返回结果进行计算。以 NUS-WIDE 数据集为例，mAP@5000 就代表在检索过程中，仅依据前 5000 个返回结果来计算每次查询的 AP 值，最后对 N 次查询的 AP 值求平均得到 mAP。一般来说，mAP 数值越高，意味着图像检索系统的性能越优，能够更精准地返回与查询内容相关的图像。

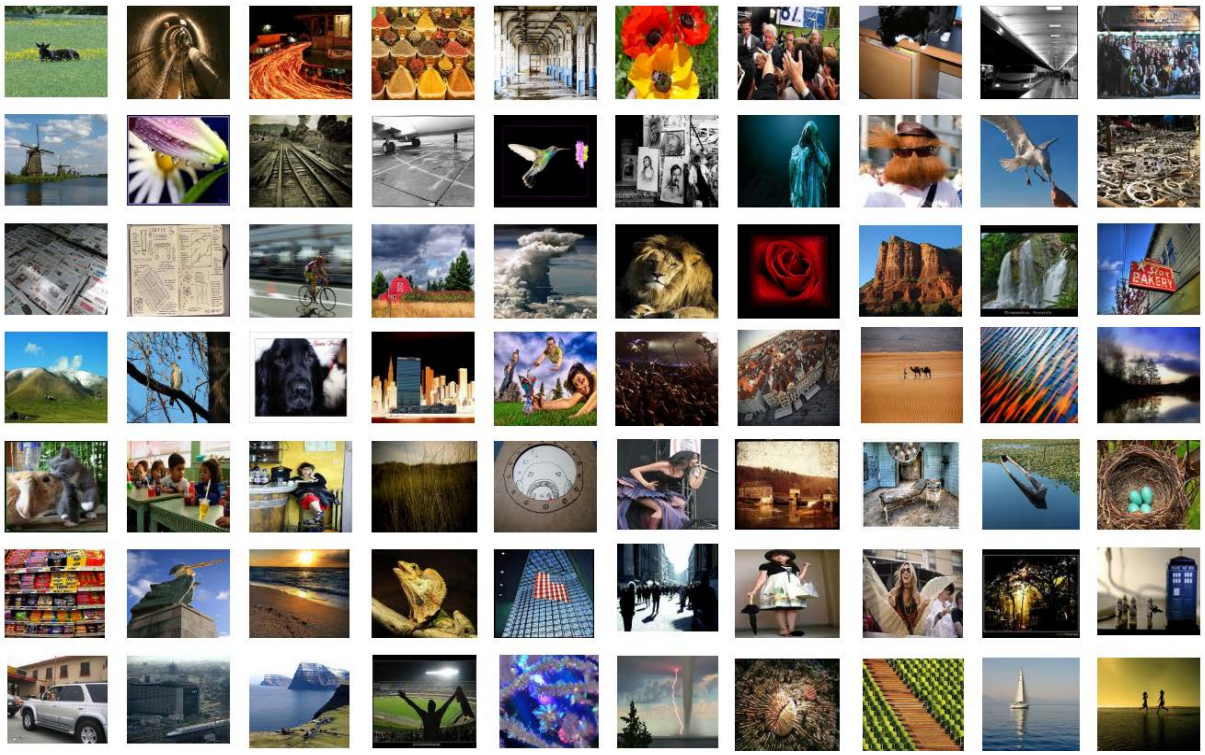

Figure 2. NUS-WIDE dataset images
图 2. NUS-WIDE 数据集图片

4.3. 对比试验

本文选择 DPN [12]，CSQ [13]，HashNet [11]，HHF [14]，TransHash [15]，Hashformer [16]，CVTH [17]，CMTH [18]以及 HPMPA [19]等算法进行性能比较。文中呈现的所有结果数据，均取自相关研究论文，为结论分析提供严谨的数据支撑。表 1 是 CIFAR-10 数据集上不同长度哈希码的 mAP 值。

Table 1. mAP values of Hash codes with different lengths on the CIFAR-10 dataset
表 1. CIFAR-10 数据集上不同长度哈希码的 mAP 值

method	CIFAR-10 (mAP@ALL)			
	16 bit	32 bit	48 bit	64 bit
DPN [12]	0.825	0.838	0.830	0.829
CSQ [13]	-	-	-	-
HashNet [11]	-	-	-	-

续表

HHF [14]	0.975	0.976	0.978	0.979
TransHash [15]	0.908	0.911	0.914	0.917
Hashformer [16]	0.912	0.917	0.921	0.924
CVTH [17]	0.910	0.929	0.937	0.946
CMTH [18]	0.952	0.951	0.955	0.961
HPMPA [19]	0.987	0.983	0.982	0.985

在 CIFAR-10 数据集不同长度哈希码的 mAP@ALL 指标下，CNN 类算法里，DPN 表现相对一般，HHF 则表现出色，mAP 值随哈希码长度增加逐步提升；ViT 类算法整体更优，TransHash、Hashformer 的 mAP 值随哈希码长度增加逐步上升，CVTH 提升幅度明显，CMTH 整体表现较好，而 HPMPA 表现顶尖，四个不同哈希码长度下的 mAP 值超过 CNN 类里表现最佳的 HHF。表 2 是 NUS-WIDE 数据集上不同长度哈希码的 mAP 值。

Table 2. mAP values of Hash codes with different lengths on the NUS-WIDE dataset
表 2. NUS-WIDE 数据集上不同长度哈希码的 mAP 值

method	**NUS-WIDE (mAP@5000)**			
	16 bit	**32 bit**	**48 bit**	**64 bit**
DPN [12]	0.847	0.859	0.863	0.862
CSQ [13]	0.810	0.825	-	0.839
HashNet [11]	0.662	0.699	0.711	0.716
HHF [14]	-	-	-	-
TransHash [15]	0.726	0.739	0.753	0.749
Hashformer [16]	0.732	0.742	0.759	0.760
CVTH [17]	-	-	-	-
CMTH [18]	0.769	0.783	0.783	0.800
HPMPA [19]	0.907	0.920	0.930	0.930

在 NUS-WIDE 数据集不同长度哈希码的 mAP@5000 指标下，CNN 类算法里，DPN 的 mAP 值随哈希码长度增加先升后降，CSQ 有部分数据且整体表现尚可，HashNet 的 mAP 值则随哈希码长度增加逐步提升但整体偏低；ViT 类算法中，Hashformer 的 mAP 值随哈希码长度增加逐步上升，CMTH 表现较好，而 HPMPA 的 mAP 值在各哈希码长度下都处于较高水平，明显优于其他 ViT 类算法以及 CNN 类算法。

5. 结语

本文对深度哈希图像检索算法进行了系统综述，多维度梳理领域研究脉络与核心成果。在算法模型现状方面，重点总结了卷积神经网络(CNN)与视觉 Transformer (ViT)：CNN 以局部特征提取和层级融合为基础，早期是深度哈希核心框架，研究聚焦于通过损失函数优化提升哈希码区分度；ViT 依托自注意

力机制捕捉全局特征，弥补 CNN 长距离依赖建模局限，研究涵盖架构适配与轻量化设计，呈现两类模型并行发展、互补优化的态势。当前，面向图像检索的哈希学习仍是热门且具挑战性的方向。虽然各类算法已取得理论与应用成果，但技术环境与应用需求的迭代，对算法检索性能、能效比及场景针对性提出更高要求。因此，哈希学习需从学术与产业角度持续完善：理论上构建统一评价标准、完善离散优化框架；研究上探索场景化算法设计、哈希码动态自适应。随着研究深入与软硬件发展，哈希学习有望在基础理论突破、效能提升、应用扩展等方面取得进展，为大规模图像高效检索提供支撑。

参考文献

[1] Barrios, J.M., Diaz-Espinoza, D. and Bustos, B. (2009) Text-Based and Content-Based Image Retrieval on Flickr: Demo. 2009 *Second International Workshop on Similarity Search and Applications*, Prague, 29-30 August 2009, 156-157. https://doi.org/10.1109/sisap.2009.30

[2] Hörster, E., Lienhart, R. and Slaney, M. (2007) Image Retrieval on Large-Scale Image Databases. *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, Amsterdam, 9-11 July 2007, 17-24. https://doi.org/10.1145/1282280.1282283

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al.* (2020) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

[4] Datar, M., Immorlica, N., Indyk, P. and Mirrokni, V.S. (2004) Locality-Sensitive Hashing Scheme Based on P-Stable Distributions. *Proceedings of the Twentieth Annual Symposium on Computational Geometry*, Brooklyn, 8-11 June 2004, 253-262. https://doi.org/10.1145/997817.997857

[5] Weiss, Y., Torralba, A. and Fergus, R. (2008) Spectral Hashing. *Advances in Neural Information Processing Systems*, **21**, 1753-1760.

[6] Gong, Y., Lazebnik, S., Gordo, A. and Perronnin, F. (2013) Iterative Quantization: A Procrustean Approach to Learning Binary Codes for Large-Scale Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 2916-2929. https://doi.org/10.1109/tpami.2012.193

[7] Liu, W., Wang, J., Ji, R., Jiang, Y. and Chang, S. (2012) Supervised Hashing with Kernels. 2012 *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, 16-21 June 2012, 2074-2081. https://doi.org/10.1109/cvpr.2012.6247912

[8] Shen, F., Shen, C., Liu, W. and Shen, H.T. (2015) Supervised Discrete Hashing. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 37-45. https://doi.org/10.1109/cvpr.2015.7298598

[9] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[10] Xia, R., Pan, Y., Lai, H., Liu, C. and Yan, S. (2014) Supervised Hashing for Image Retrieval via Image Representation Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **28**, 2156-2162. https://doi.org/10.1609/aaai.v28i1.8952

[11] Cao, Z., Long, M., Wang, J. and Yu, P.S. (2017) HashNet: Deep Learning to Hash by Continuation. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 5608-5617. https://doi.org/10.1109/iccv.2017.598

[12] Fan, L., Ng, K.W., Ju, C., Zhang, T. and Chan, C.S. (2020) Deep Polarized Network for Supervised Learning of Accurate Binary Hashing Codes. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 825-831. https://doi.org/10.24963/ijcai.2020/115

[13] Yuan, L., Wang, T., Zhang, X., Tay, F.E., Jie, Z., Liu, W., *et al.* (2020) Central Similarity Quantization for Efficient Image and Video Retrieval. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 3083-3092. https://doi.org/10.1109/cvpr42600.2020.00315

[14] Xu, C., Chai, Z., Xu, Z., Li, H., Zuo, Q., Yang, L., *et al.* (2023) HHF: Hashing-Guided Hinge Function for Deep Hashing Retrieval. *IEEE Transactions on Multimedia*, **25**, 7428-7440. https://doi.org/10.1109/tmm.2022.3222598

[15] Chen, Y., Zhang, S., Liu, F., Chang, Z., Ye, M. and Qi, Z. (2022) Transhash: Transformer-Based Hamming Hashing for Efficient Image Retrieval. *Proceedings of the* 2022 *International Conference on Multimedia Retrieval*, Newark, 27-30 June 2022, 127-136. https://doi.org/10.1145/3512527.3531405

[16] Li, T., Zhang, Z., Pei, L. and Gan, Y. (2022) HashFormer: Vision Transformer Based Deep Hashing for Image Retrieval. *IEEE Signal Processing Letters*, **29**, 827-831. https://doi.org/10.1109/lsp.2022.3157517

[17] Ren, X., Zheng, X., Zhou, H., Liu, W. and Dong, X. (2022) Contrastive Hashing with Vision Transformer for Image Retrieval. *International Journal of Intelligent Systems*, **37**, 12192-12211. https://doi.org/10.1002/int.23082

[18] 杨梦雅, 赵琰, 薛亮. 基于改进的 Vision Transformer 深度哈希图像检索[J]. 陕西科技大学学报, 2025, 43(4): 183-191.

[19] 刘华咏, 徐明慧. 基于混合注意力与偏振非对称损失的哈希图像检索[J]. 计算机科学, 2025, 52(8): 204-213.

[20] Song, C.H., Yoon, J., Choi, S. and Avrithis, Y. (2023) Boosting Vision Transformers for Image Retrieval. 2023 *IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 2-7 January 2023, 107-117. https://doi.org/10.1109/wacv56688.2023.00019

[21] Krizhevsky, A. and Hinton, G. (2009) Learning Multiple Layers of Features from Tiny Images. Technical Report TR-2009. University of Toronto.

[22] Chua, T., Tang, J., Hong, R., Li, H., Luo, Z. and Zheng, Y. (2009) NUS-WIDE: A Real-World Web Image Database from National University of Singapore. *Proceedings of the ACM International Conference on Image and Video Retrieval*, Santorini, 8-10 July 2009, 1-9. https://doi.org/10.1145/1646396.1646452