Computer Science and Application 计算机科学与应用, 2025, 15(6), 206-219
Published Online June 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.156171

Hans汉斯

基于改进Swin Transformer的深度哈希检索模型

李一昊，王直杰

东华大学信息学院，上海

收稿日期：2025年5月24日；录用日期：2025年6月23日；发布日期：2025年6月30日

摘　要

随着互联网和多媒体技术的飞速发展，数字图像已经成为现代社会中信息传播和交流的主要载体之一。人们每天都在生成和消费海量的图像数据，从社交媒体的图片分享到专业领域的图像分析，图像信息的规模和复杂性都在不断增长。与此同时，针对这些数据的检索需求也在快速增加，尤其是在需要快速定位和提取特定内容的场景中。然而，现实世界中的图像数据往往呈现出一种长尾分布的特性，即某些类别的数据非常丰富，而另一些类别的数据却极其稀缺。这种不平衡的数据分布为图像检索技术带来了巨大的挑战，尤其是在基于深度哈希技术的检索方法中，如何有效处理长尾分布成为研究的关键问题。针对这个问题，本文从模型层面构建了基于改进Swin Transformer哈希检索模型，以校验本文所设计长尾哈希检索模型在现实场景下的性能表现。详细内容如下：在面对长尾分布图像检索任务中对图像的局部的特征提取能力不足时，提出一种创新的解决方案。该方法核心在于利用双流网络架构将CNN的局部特征与Transformer的全局特征进行融合。同时，基于哈希层的输出数据设计了多目标损失函数。通过以上策略能够实现卷积的局部细节特征与自注意力的全局上下文特征的融合。实验结果表明，本模型能够实现高性能的哈希图像检索且优于当前主流模型，对各类数据集均取得最好或者次好的性能指标。

关键词

深度哈希检索，长尾分布，双流网络

Deep Hash Retrieval Model Based on Improved Swin Transformer

Yihao Li, Zhijie Wang

College of Information Science and Technology, Donghua University, Shanghai

Received: May 24th, 2025; accepted: Jun. 23rd, 2025; published: Jun. 30th, 2025

文章引用: 李一昊, 王直杰. 基于改进 Swin Transformer 的深度哈希检索模型[J]. 计算机科学与应用, 2025, 15(6): 206-219. DOI: 10.12677/csa.2025.156171

Abstract

With the rapid development of the Internet and multimedia technology, digital images have become one of the primary carriers for information dissemination and communication in modern society. People generate and consume vast amounts of image data every day—from picture sharing on social media to image analysis in professional fields, the scale and complexity of image information are continuously growing. At the same time, the demand for retrieving this data is also increasing rapidly, especially in scenarios where specific content needs to be quickly located and extracted. However, image data in the real world often exhibits a long-tail distribution characteristic—certain categories of data are highly abundant, while others are extremely scarce. This unbalanced data distribution poses significant challenges to image retrieval technologies, especially for retrieval methods based on deep hashing. Effectively addressing long-tail distribution has become a key research issue. To tackle this problem, this paper constructs a hash retrieval model based on an improved Swin Transformer at the model level, to evaluate the performance of the proposed long-tail hash retrieval model in real-world scenarios. Details are as follows: when the local feature extraction capability of images is insufficient in long-tail image retrieval tasks, an innovative solution is proposed. The core of this method lies in employing a two-stream network architecture that fuses the local features of CNNs with the global features of Transformers. Meanwhile, a multi-objective loss function is designed based on the output of the hash layer. This strategy enables the fusion of convolutional local detail features with the global contextual features from self-attention mechanisms. Experimental results demonstrate that this model can achieve high-performance hash-based image retrieval and outperforms current mainstream models, achieving the best or second-best performance indicators across various datasets.

Keywords

Deep Hash Retrieval, Long-Tail Distribution, Two-Stream Network

1. 引言

近年来，基于自注意力机制的 Transformer [1]模型突破了自然语言处理领域的边界，在计算机视觉领域展现出强大的适应能力。相较于传统卷积神经网络(CNN)依赖局部感受野的特性，Transformer 通过自注意力机制建立了长程依赖关系，既能捕捉图像全局特征，又具备并行计算优势，这一突破性进展在目标检测、图像分类等任务中得到了验证。其中，Vision Transformer (ViT)作为首个纯 Transformer 架构的视觉模型，通过将图像切分为序列化块并引入位置编码，在 ImageNet 图像分类基准上超越了传统 CNN [2]模型，证明了无卷积架构的可行性。

在深度哈希检索领域，现有方法普遍采用 CNN 架构进行特征提取与哈希编码。然而，这类方法因卷积操作的固有局限，难以有效建模图像全局语义关联，导致特征表达能力受限。与此形成对比的是，ViT 的自注意力机制能够动态建立图像块间交互，通过加权聚合全局上下文信息，既保留了局部细节特征，又强化了整体语义理解。这种双重视觉表征能力恰好弥补了 CNN 的结构缺陷，为哈希码学习提供了更丰富的特征支撑。基于此，本研究选择 Swin Transformer [3]作为基础架构，旨在通过其全局特征建模优势提升哈希检索性能。

另一方面，传统视觉 Transformer (ViT)及其轻量化变体(如 DeiT [4]、TinyViT [5])在浅层特征处理上存在显著效率瓶颈。ViT 直接将输入图像分割为序列化的图像块(Patch)，通过自注意力机制处理原始像素，这在浅层会带来两个问题：其一，自注意力在低层次特征(如边缘、纹理)上的计算成本高昂，而同等条件下，卷积操作可通过权重共享和局部感受野更好地提取此类特征；其二，图像块的线性投影层(Patch Embedding)缺乏空间层次性，导致浅层特征分辨率下降过快，丢失细节信息。尽管轻量化方法(如 DeiT)通过减少注意力头数、降低嵌入维度或裁剪层数来压缩模型，但这些策略本质上是对模型的“宽度”或“深度”进行静态裁剪，在降低计算量的同时牺牲了模型对多尺度特征的适应能力。更严重的是，轻量化模型往往依赖知识蒸馏或数据增强弥补性能损失，却未从根本上解决浅层处理效率与表达能力之间的矛盾。

为了解决上述问题，本文设计了一种融合 CNN 与 Transformer 的双流提取架构。同时，本文基于哈希层的输出数据设计了多目标损失函数。本文工作的主要贡献总结如下：

(1) 本文提出了一种改进 Swin Transformer 的哈希图像检索方法，该方法采用双流并行架构，使用两个独立的网络流，分别提取不同类型的特征，然后进行融合，提高模型性能。

(2) 本文设计了多目标损失函数来优化整个模型，通过自蒸馏、哈希代理损失和基于二元交叉熵的量化损失来优化模型。

(3) 在三个广泛使用的公共数据集上的综合实验结果证明了本文方法在长尾图像检索任务上的有效性和优越性。

2. 改进的长尾哈希检索模型

本文提出一种基于 CNN 与 Transformer 的长尾哈希检索模型，能兼二者之长，实现卷积的局部细节特征与自注意力的全局上下文特征的融合。首先，采用 EfficientNet-B3 [6]模型和 Swin Transformer [3]模型构成双流骨干网络，分别用于提取图像的局部特征和全局特征；其次，设计了多目标损失函数。整体的模型结构架构如图 1 所示。

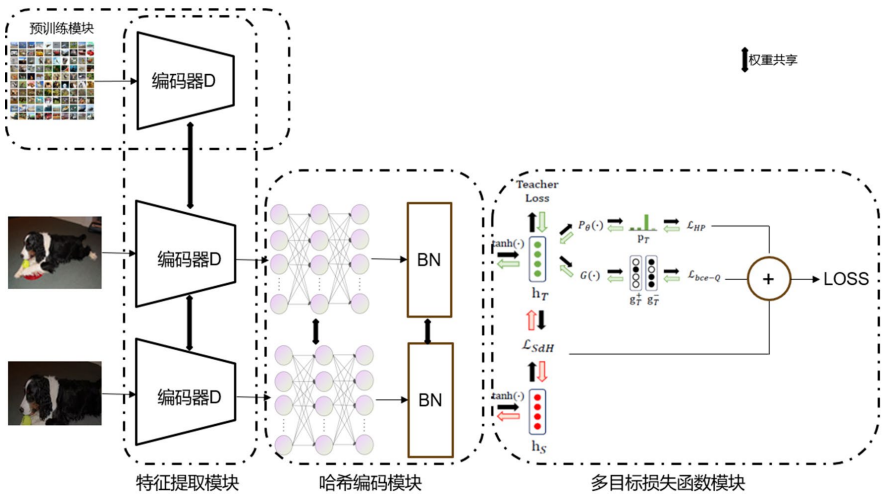


Figure 1. Improved hash retrieval model
图 1. 改进的哈希检索模型

整体的模型结构由预训练模块，特征提取模块 D，哈希编码模块以及损失函数模块组成，其中预训练模块在 ImageNet 数据集上对网络模型进行监督预训练，以学习中级图像特性。特征提取模块目标是提取图像特性，其中包括 class-token 特征、patch 特征。哈希编码模块对主干网络进行微调，以提高主干网络在提取输入图像特征方面的能力，改进的哈希检索模型的特征提取模块有两个主要的组成分支：局部

特征提取分支和全局特征提取分支，其中局部特征提取分支采用一种轻量级的 CNN 模型 EfficientNet-B3，EfficientNet 网络由多个堆叠的倒残差模块(MBConv)组成。

全局特征分支采用 Swin Transformer 模型。哈希编码模块对主干网络进行微调，以提高主干网络在提取输入图像特征方面的能力，并将特征提取模块提取到的连续码转化为二进制码。损失函数模块用于使模型学习类间距离更远、类内距离更近的特征信息，以增强模型分类能力。

模型运行流程如下：首先在 ImageNet 数据集上对进行监督预训练，以学习高质量的中级图像特性。定义输入的样本图像为 x，将输入图片 $x=\{x_n\}_{n=1}^{N}\ \mathbb{R}^{N\times d}$ (d 为被检索图片维度，N 为训练样本数量)送入模型中，并基于自蒸馏的哈希方案采用权重共享 Siamese Structure (暹罗结构、孪生网络)来同时对比一个图像的不同视图(增强结果)的哈希编码。配置了两个独立的增强组，分别生成弱变换视图和强变换视图，以构建简单教师和困难学生的训练框架。以随机抽样的方式控制难度：对组中的所有变换使用相同的超参数，并通过缩放它们自身的发生概率使它们发生的更少或更多。随后图像经过特征提取模块 D 得到连续码 $v=\{v_n\}_{n=1}^{N}\ \mathbb{R}^{N\times d}$，通过批处理归一化哈希编码层得到零平均连续码，送入多目标损失函数模块，基于样本两个不同的哈希编码，通过余弦相似度计算自蒸馏损失函数值，基于哈希编码输入分类层得到的输出和样本的语义标签，计算哈希代理损失函数值；通过哈希编码和二进制交叉熵的量化损失，计算量化损失函数值。最后整合多目标损失函数作为模型输出，进行前向传播与反向推导。最终将识别模型的置信度同检索模型检索出图片的汉明距离进行计算并作为最终用于排序的汉明距离。

2.1. 基于改进 Swin Transformer 的哈希特征提取模块

为保证学习到的特征具有足够的区分度和判别力，通过双流特征提取架构将视觉 Transformer 模型融合 CNN 模型，解决 Transformer 在捕捉空间局部细节特征上存在不足的问题。采用局部－全局特征提取模块，实现了局部细节特征与全局上下文特征的高效融合，提高主干网络提取输入图像特征方面能力。

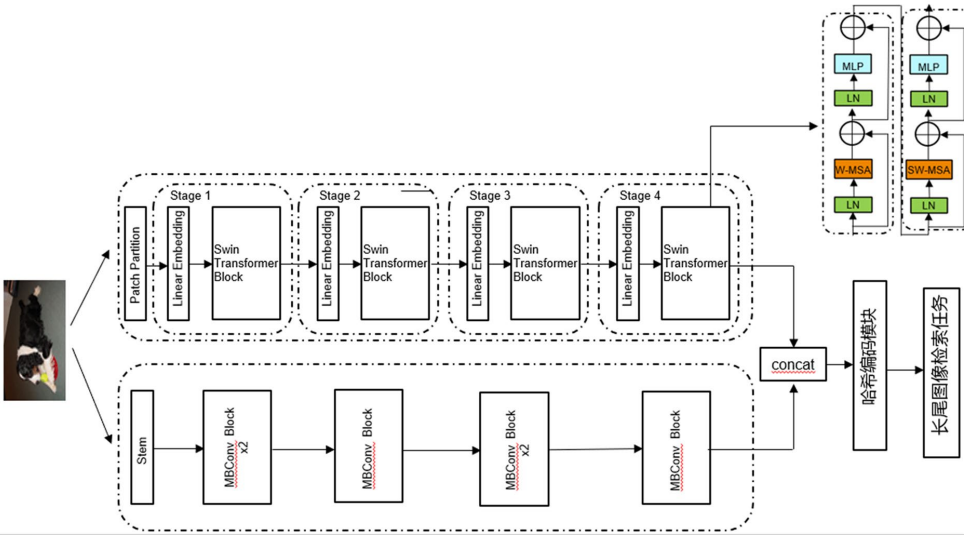


Figure 2. The overall of ILGS-Net
图 2. 局部－全局特征双流融合模型总体结构

卷积神经网络通过构建多层卷积层、池化层以及非线性激活函数，具备提取图像局部空间结构和纹理细节的能力。然而，当处理高分辨率遥感图像时，提取的信息越丰富，所需网络参数和计算量也会随之增加。为在计算效率和特征提取性能之间取得平衡，本文采用了一种轻量级卷积神经网络模型——

EfficientNet-B3 作为局部特征流的基础。

EfficientNet-B3 的网络结构由层级化设计的 9 个核心阶段构成，采用模块化思路实现多尺度特征提取。初始阶段(Stage 1)为普通 3 × 3 卷积层，执行输入图像(300 × 300 分辨率)的基础空间特征映射，输出通道数经过复合缩放策略调整至 40。随后 Stage 2 至 Stage 8 为核心处理层，通过堆叠 18 个改进型倒残差模块(MBConv)实现深度特征抽象——对比 B0 模型，MBConv 总数增加了 40%。

Transformer 模型在处理图像数据时，会先将图片切割为固定大小的区块(例如 16 × 16 像素)，每个区块视为一个特征向量输入序列。通过内部的多头自注意力机制，模型能够自动计算不同区块之间的关联权重，例如识别出猫图像中耳朵与胡须的远距离位置关系，即使这两个特征在图像中相隔较远。这种全局关联能力对长距离依赖建模(如物体整体结构)尤为重要。但当面对长尾分布数据集时(例如某些罕见类别仅有少量样本)，为了更好地捕捉全图的细化特征，需要采用更精细的图像分块策略。每张图像的序列长度因此从 256 块增长到 1024 块，这直接导致自注意力计算量增加，使计算复杂度增长 16 倍。为了在同等提取能力下尽可能减少模型的计算开销，本文使用 Transformer 系列模型 Swin Transformer 作为全局特征流。如图 2 所示，输入图像 $X \in \mathbb{R}^{H\times W\times 3}$，其中 H 和 W 分别为输入图像的高度和宽度，3 表示 RGB 三通道)首先通过图像块划分(Patch Partition)操作被切割为 $N=\frac{H}{4}\times\frac{W}{4}$ 个非重叠像素块，每个块展平后为向量 $p_i \in \mathbb{R}^{H\times W\times 3}$ ($i=1,2,\cdots,N$)。这些块通过块嵌入(Patch Embedding)层线性映射到高维空间，最终得到初始嵌入序列 z_0 如公式(1)所示

$$z_0=\left[p_1W_e;p_2W_e;\cdots;p_NW_e\right]+E_{pos} \tag{1}$$

其中 $W_e \in \mathbb{R}^{48\times D}$ 是块嵌入的投影矩阵(D = 128 为嵌入维度)，$E_{pos} \in \mathbb{R}^{N\times D}$ 是学习得到的位置编码矩阵，用于保留图像的空间位置信息。

该序列随后通过四个层级处理阶段(Hierarchical Stages)逐步提炼全局特征，每个阶段由块合并(Patch Merging)和 Swin Transformer Block 堆叠组成。

以第一阶段为例，输入特征 z_0 首先经过块合并操作：将特征图划分为 2 × 2 的局部区域(每个区域包含 4 个相邻块 $z_{2i,2j}$、$z_{2i+1,2j}$、$z_{2i,2j+1}$、$z_{2i+1,2j+1} \in \mathbb{R}^{128}$)，在通道维度拼接为 $\mathbb{R}^{4\times 128}$ 后通过线性投影矩阵 $W_m \in \mathbb{R}^{4\times 128\times 256}$ 压缩至 256 维，得到降采样后的特征 $z_{merged} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 256}$。接下来，降采样后的特征通过若干交替的窗口自注意力(W-MSA)和移位窗口自注意力(SW-MSA)模块处理：在 W-MSA 中，特征图被划分为 $\frac{H}{8M}\times\frac{W}{8M}$ 个 $M\times M$ 窗口($M=7$)，每个窗口内的特征 $z_{win} \in \mathbb{R}^{M^2\times 256}$ 通过线性投影矩阵 $W_q, W_k, W_v \in \mathbb{R}^{256\times 64}$ 生成查询矩阵 $Q=z_{win}W_q$、键矩阵 $K=z_{win}W_k$ 和值矩阵 $V=z_{win}W_v$，并引入相对位置偏置矩阵 $B \in \mathbb{R}^{M^2\times M^2}$ 编码窗口内块间的位置差异，计算多头自注意力，公式如下：

$$\text{Attention}\left(Q,K,V\right)=\text{Softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}+B\right)V \tag{2}$$

其中 $d_k=64$ 为单头注意力维度(总头数 $h=4$，满足 $h\times d_k=256$)，多头输出经拼接和投影矩阵 $W_0 \in \mathbb{R}^{256\times 256}$ 聚合后与输入残差连接，再通过两层 MLP(包含权重矩阵 $W_1 \in \mathbb{R}^{256\times 1024}$ 和 $W_2 \in \mathbb{R}^{1024\times 256}$)，激活函数为 GELU)进一步增强非线性，最终输出特征为公式(3)：

$$z_{out}=\text{MLP}\left(\text{LayerNorm}\left(z_{attn}\right)\right)+z_{attn} \tag{3}$$

在后续模块中，窗口通过循环位移偏移 $\left[M/2\right]=3$ 个块，生成交叠的移位窗口以强制跨窗口信息交互，SW-MSA 的计算流程与 W-MSA 相同，但注意力计算时需考虑位移后窗口的索引变化，并通过掩码机制

避免无效区域参与计算。经过四个阶段的层级处理，每阶段依次将空间分辨率降低至 $\frac{H}{8}\times\frac{W}{8}$、$\frac{H}{16}\times\frac{W}{16}$ 和 $\frac{H}{32}\times\frac{W}{32}$ 通道维度扩展至 256、512 和 1024，最终在第四阶段输出特征图 $z_{final}\in\mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 256}$，对其沿空间维度执行全局平均池化(GAP)得到全局特征向量

$$F_{global}=\frac{1}{\frac{H}{32}\times\frac{W}{32}}\sum_{i=1}^{\frac{H}{32}}\sum_{j=1}^{\frac{W}{32}}z_{final}(i,j) \tag{4}$$

公式(4)中每个元素整合了图像的多尺度语义信息(例如物体整体形状、部件间拓扑关系及背景上下文)，并以紧凑形式支持哈希编码生成。整个过程中，Swin Transformer 通过局部窗口自注意力与层级降采样的结合，将计算复杂度从标准 Transformer 的 $O(N^2)$ 降低至 $O(M^2N)$ $(M\ll N)$，同时利用移位窗口实现跨窗口长程依赖建模，确保了全局特征对复杂视觉模式的鲁棒表征能力。

两个子网络(EfficientNet-B3 与 Swin Transformer)所提取的特征首先分别与对应的可学习权重相乘。这样的机制允许模型在训练过程中根据具体任务的需求动态调整各自特征的重要性。权重参数通过反向传播自动更新，从而使模型能够自适应地学习出在不同数据分布与任务条件下的最优特征融合策略。完成加权后，来自两个网络的特征被拼接为一个统一的特征向量，不仅增加了整体特征的维度，也使模型能够在后续处理过程中兼顾局部与全局的语义信息。该特征融合操作可由公式(5)表示。

$$F_{\text{combined}}=[\alpha\times F_{\text{EfficientNet}},\beta\times F_{\text{Swin}}] \tag{5}$$

式中：F_{combined} 为融合后的特征；$F_{\text{EfficientNet}}$ 和 F_{Swin} 为 EfficientNet-B3 和 Swin Transformer 的输出特征；α 和 β 是可学习的权重。

EfficientNet 与 Swin Transformer 在特征提取方面各具优势，二者的互补性使得双分支结构能够从不同角度解析图像内容，从而获得更全面的特征表达。通过融合两种结构的输出，模型不仅能够捕捉局部与全局信息，还增强了应对噪声与干扰的能力。由于两个网络对不同类型的扰动具有不同的敏感程度，组合使用有助于提升整体模型在复杂环境下的稳定性与鲁棒性。

2.2. 多目标哈希损失模块

由于基于哈希的检索系统需要计算图像与二值码之间的距离，因此对应的码需要通过 *sgin* 运算进行量化，从连续实空间到$\{-1,1\}$的离散汉明空间。在这个过程中，不断优化的图像表示被改变，并产生量化误差，从而降低了哈希码的判别能力。当输入图像被变换并偏离原始分布时，这个问题就更大了。为了避免由于变换而导致的性能下降，最常见的解决方案是通过使用具有各种变换的增强数据进行训练来泛化深度模型。然而，将这种增强策略应用于深度哈希是具有挑战性的，因为可能会出现表示上的差异。

为了解决这些问题，本文提出了一种多目标损失模块，该模块结合了自蒸馏哈希损失、哈希代理损失和基于二值交叉熵的量化损失，旨在优化哈希码的质量，提高图像检索的性能。

基于对余弦距离与汉明一致性之间关系的理解，该方法的核心理念是将难以直接优化的离散二值编码问题，转化为连续空间中可导的余弦距离优化。当图像经过不同变换(如裁剪、调色)生成两个视图时，通过神经网络生成对应的哈希码，并最小化这些连续哈希码间的余弦距离。由于在二值化后，哈希码的汉明距离与训练时的余弦距离存在严格线性关系，这种优化本质上迫使相似图片的二进制编码在汉明空间自动对齐。也就是说，可以利用哈希码 h_i 和 h_j 之间的余弦相似度来近似二值码 b_i 和 b_j 之间的汉明距离如公式(1)~(7)所示。

$$D_H\left(b_i,b_j\right)\simeq\frac{K}{2}\left(1-S\left(h_i,h_j\right)\right) \tag{6}$$

其中，$b_i=sign\left(h_i\right)$，$b_j=sign\left(h_j\right)$，$D_H\left(.,.\right)$表示汉明距离，公式中 S 为二进制编码的余弦相似度，其相似度越大，他们两者的距离便越小，结合余弦相似度和自蒸馏方案所产生的两个不同的哈希编码，可以得出自蒸馏损失函数。定义自蒸馏损失函数函数为 L_{SdH}，计算公式(7)为：

$$L_{SdH}\left(h_T,h_S\right)=1-S\left(h_T,h_S\right) \tag{7}$$

哈希代理损失的目的是学习带有温度标度的汉明一致性，以提高哈希码的判别能力。具体来说，对于每个类别，本文引入一个可训练的哈希代理 p_θ，并使用哈希码 h_T 计算类别的预测相似度 p_T：

$$p_T=\left[S\left(p_{\theta1},h_T\right),S\left(p_{\theta2},h_T\right),\cdots,S\left(p_{\theta N_{cls}},h_T\right)\right] \tag{8}$$

其中，$p_{\theta i}$ 是分配给第 i 个类的哈希代理，N_{cls} 表示类别数量，然后，通过计算哈希代理损失来学习类别标签 y 与预测相似度 p_T 之间的相似度：

$$L_{HP}\left(y,p_T,\tau\right)=H\left(y,\mathrm{softmax}\left(p_T/\tau\right)\right) \tag{9}$$

其中，τ 是温度标度超参数，$H\left(u,v\right)=-\sum_k u_k\log v_k$ 是交叉熵，softmax 运算沿着 p_T 的维度应用。

深度哈希方法的核心优化策略，是在模型训练过程中利用回归策略，将连续值的哈希编码逐步逼近实际的离散二值目标(如+1 和−1)。其本质是通过不断压缩每个编码维度与其对应二值端点之间的数值间距(例如计算欧式距离)，迫使中间态的浮点数值最终坍缩为符合二值特性的哈希位，从而显著降低由连续性编码向离散态转换过程中的信息精度损失。然而，由于哈希的目标是对每个比特的符号进行分类，因此一个更自然的选择是将其视为二进制分类，使用预定义的均值为 m 、标准差为 σ 的高斯分布估计器 $g\left(h\right)$ 来估计哈希码元素 h 的二值似然，公式为：

$$g\left(h\right)=\exp\left(-\frac{\left(h-m\right)^2}{2\sigma^2}\right) \tag{10}$$

所述的量化损失函数值采用下式计算：

$$L_{bce-Q}\left(h_T\right)=\frac{1}{K}\sum_{k=1}^{K}\left(H_b\left(b_k^+,g_k^+\right),\left(b_k^-,g_k^-\right)\right) \tag{11}$$

其中，$H_b=-u\log v+\left(1-u\right)\log\left(1-v\right)$ 是二值交叉熵，g_k^+,g_k^- 表示第 k 个哈希码元素的估计似然值，$g_k^+=g^+\left(h_k\right)$，$g_k^-=g^-\left(h_k\right)$；$b_k^+,b_k^-$ 表示二值似然标签，$b_k^+=\frac{1}{2}\left(sign\left(hk\right)+1\right)$，$b_k^-=1-b_k^+$，量化误差通过给定估计器的二进制分类损失来减少。

为了综合优化哈希码的质量，我们将这三种损失函数结合在一起，构建了一个多目标损失模块。具体来说，总损失函数定义为：

$$L_T\left(X_B\right)=\frac{1}{N_B}\sum_{n=1}^{N_B}L_{HP}+\lambda_1L_{SdH}+\lambda_2L_{bce-Q} \tag{12}$$

其中，λ_1 和 λ_2 是超参数，用于平衡不同训练目标的影响。通过这种方式，多目标损失模块能够综合优化哈希码的质量，提高图像检索的性能。

综上所述，本文提出的多目标损失模块通过结合自蒸馏哈希损失、哈希代理损失和基于二值交叉熵的量化损失，有效地解决了深度哈希学习中的关键问题，包括数据增强导致的表示不一致、预定义哈希目标的限制以及量化误差。通过这种方式，模型能够学习到更高质量的哈希码，从而显著提高图像检索

的性能。

3. 实验准备与训练步骤

为了全面评估本文所提出的基于改进 Swin Transformer 的长尾哈希检索模型的性能，本研究设计并实施了一系列实验。首先，通过与当前主流的哈希检索模型进行对比实验，本文旨在量化分析所提模型在检索精度和效率方面的优势。其次，为了深入探究双流网络架构在识别图像目标区域方面的效能，本文设计了可视化实验，以直观呈现模型对图像关键区域的检测与识别能力。最后，为了剖析特征提取模块和多目标损失函数模块对长尾数据集检索性能的影响机制，本文进一步开展了消融实验，旨在通过逐一移除或替换关键模块，明确各模块对整体性能的贡献度。这些实验设计不仅为验证本文模型的有效性提供了坚实的实证基础，也为后续的模型优化与改进提供了明确的方向。

3.1. 数据集

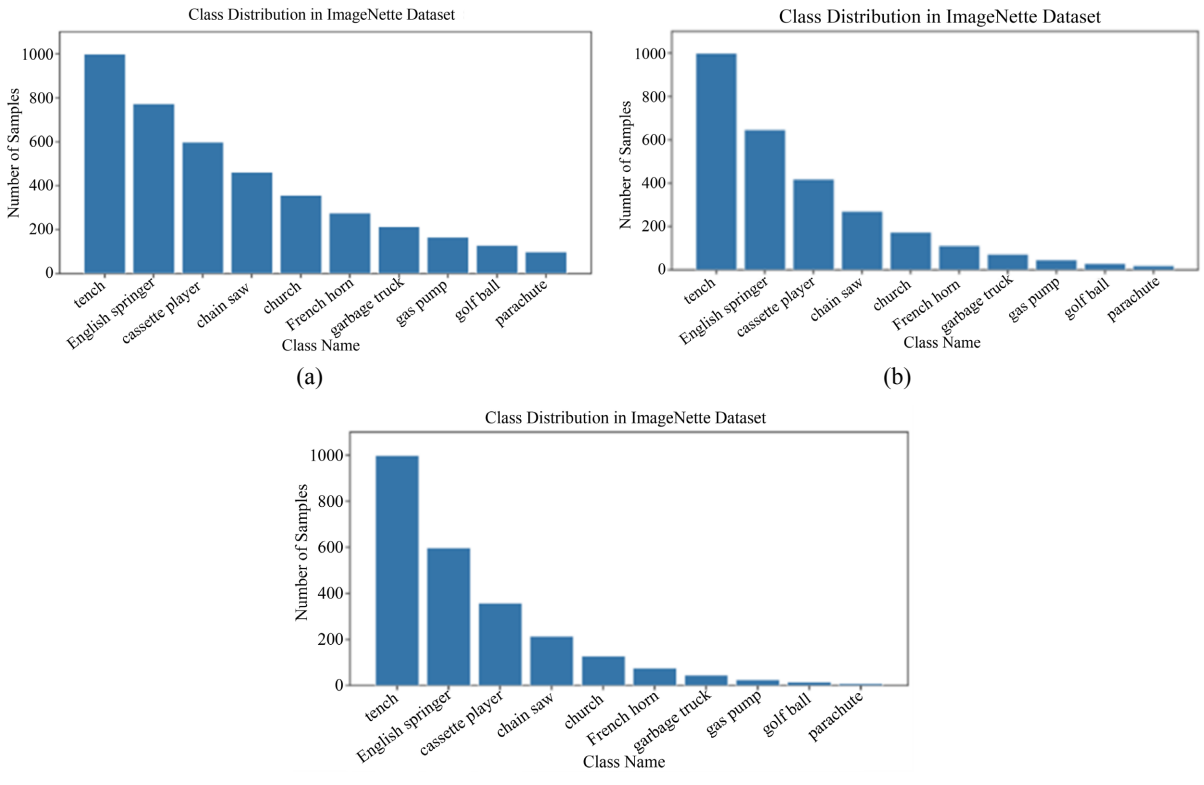


Figure 3. Distribution of ImageNette dataset at different imbalance rates. (a) The imbalance rate is 10; (b) The imbalance rate is 50; (c) The imbalance rate is 100

图 3. ImageNette 数据集在不同不平衡率上的分布情况。(a) 不平衡率为 10；(b) 不平衡率为 50；(c) 不平衡率为 100

为了评估本方法的检索性能，本文使用四个广泛适用于图像检索的数据集，对实验数据进行广泛评估。

NUS-WIDE [7]数据集是一个典型的多标签图像数据集。在本次实验中，专注于该数据集中的 21 个最频繁出现的类别，并从中选取了相应的图像用于性能评估。具体而言，查询集由 2040 张图像组成，训练集包含 10,000 张图像，而检索集则由 149,685 张图像构成。

MS-COCO [8]数据集由 80 个类别组成，其中查询集、训练集和检索集分别包含 5000 张、10,000 张和 117,218 张图像。

在本次实验中，我们使用了 ImageNette 和 CIFAR-10 数据集作为长尾数据集，采用了相同的长尾版本[9]-[11]，以模拟现实世界中数据分布的不平衡性。具体而言，ImageNette 数据集被划分为训练集和验证集，其中训练集包含 9469 张图像，验证集包含 3925 张图像。CIFAR-10 数据集被划分为训练集和验证集，图像数据集包括 50,000 张，其中有 10,000 张用于训练，10,000 张用于验证。

为了更全面地评估模型在不同数据分布情况下的性能，我们设置了不同的不平衡因子，分别为 100、50 和 10，如图 3 和图 4 所示。

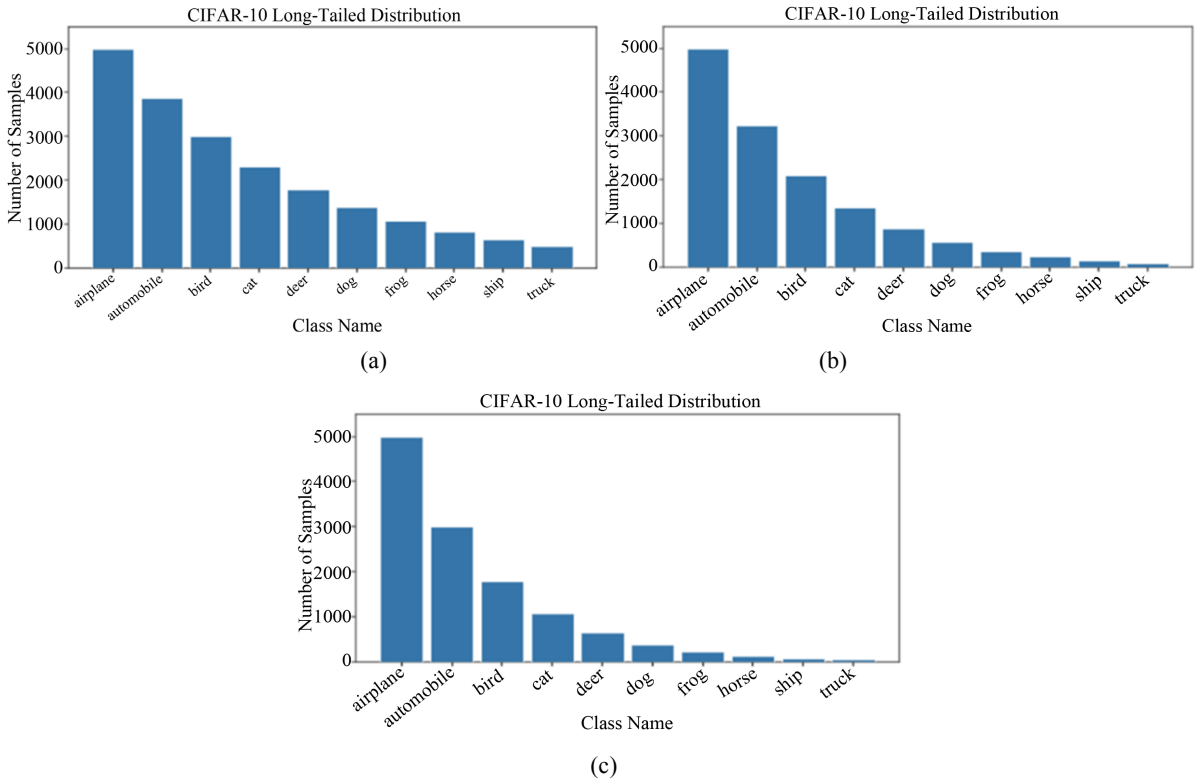


Figure 4. Distribution of the CIFAR-10 dataset at different imbalance rates. (a) The imbalance rate is 10; (b) The imbalance rate is 50; (c) The imbalance rate is 100

图 4. CIFAR-10 数据集在不同不平衡率上的分布情况。(a) 不平衡率为 10；(b) 不平衡率为 50；(c) 不平衡率为 100

3.2. 实验设置

本篇文章采用 Pytorch 深度学习框架，该框架的优势在于提供强大的 GPU 张量计算和深度神经网络的自动求导系统。模型 Batch 数设为 32，epoch 为 50，采取 Adam 优化器，学习率为 0.001，采用余弦退火算法对学习率进行优化，实验使用预训练的 Swin Transformer 模型，主干网络权重已提前在 ImageNet 上训练好。本文使用的哈希算法生成的哈希码长度分别为 32 和 64 位，每 10 个 epoch 进行一次测试，并报告最佳结果。此外，本节实验的检索评估指标与目前主流的深度哈希检索研究相一致，主要采用平均精度(mAP)来衡量检索效果。

4. 实验结果与分析

4.1. 模型对比实验

本文设计基于改进 Swin Transformer 的深度图像检索性能对比实验。本文将其与多种哈希学习方法

在不同编码模式(32、64 位)上进行比较，选择了几种经典的和目前最先进的哈希方法。它们包括无监督哈希方法 LSH [12]，三种传统有监督哈希方法 CCA-ITQ [13]、COSDISH [14]、FSDH [15]以及深度监督哈希方法 HashNet [16]、DCH [17]、GreedyHash [18]、CSQ [19]、DPN [20]与 Orthohash [21]所有对比算法的参数均参考原始文献进行设置。其中表 1 为本文模型与传统检索模型的对比实验结果，表 1 结果表明，本文提出的方法相较于浅层哈希方法 CCA-ITQ [13]、COSDISH [14]、LSH [12]、FSDH [15]在数据集 MS-COCO、NUS WIDE 上取得最优检索结果。相较于深度哈希方法(HashNet [9]、DCH [10]、GreedyHash [18]、CSQ [19]、DPN [20]、Orthohash [21])在 NUS WIDE 数据集检索中表现最佳，较 Orthohash 深度哈希检索模型性能提高了约 3%，较表现次好的 DPN 性能提高了约 2%。NUS WIDE 数据集检索中居次。在图像检索过程中，较长的哈希码能够提供更多的细节信息，从而使得模型能够更精准地捕捉到图像之间的相似性，进而显著提高检索精度。

Table 1. mAP results of comparative experiments on different data sets
表 1. 在不同数据集上的对比实验的 mAP 结果

方法	NUS WIDE		MS-COCO	
	32	64	32	64
LSH	51.34	52.48	36.48	37.13
CCA-ITQ	56.34	57.57	40.15	40.13
COSDISH	57.55	64.56	46.89	48.31
FSDH	54.16	54.63	45.62	45.86
HashNet	69.45	71.54	77.15	78.65
DCH	79.86	81.24	80.42	81.95
GreedyHash	79.58	80.94	72.65	74.51
CSQ	82.64	83.56	83.44	**86.48**
DPN	85.64	86.67	81.51	85.32
OrthoCos	83.54	85.48	78.76	79.81
Ours	**87.12**	**88.27**	**83.92**	85.63

Table 2. mAP results of comparative experiments on ImageNette and CIFAR-10
表 2. 在 ImageNette 与 CIFAR-10 上对比实验的 mAP 结果

方法	ImageNette			CIFAR-10		
	100	50	10	100	50	10
BBN	80.15	82.81	87.15	82.48	83.16	88.56
Hybrid-SC	83.61	87.29	89.73	83.49	88.48	90.15
BCL	85.56	88.64	91.96	86.31	89.48	92.48
Ours	**87.64**	**90.46**	**92.43**	**88.18**	**91.63**	**93.86**

为了更全面地评估本模型在长尾分布数据集上的展示效果，本文在长尾 ImageNette、长尾 CIFAR-10 数据集上进行实验，其中数据集不平衡率设置为 100，50 与 10，并同目前流行的长尾哈希检索算法 BBN [11]、Hybrid-SC [22]、BCL [23]进行比较，所有数据集的结果均记录在表 2 中，其中加粗的代表最高的 mAP 数值，下划线表示次高 mAP 数值。通过表 2 可知，本文提出的方法相较于传统的不平衡图像检索

算法，在长尾 CIFAR-10 数据集上，较先进的不平衡图像检索算法 BCL 提高了约 1%~2%，较 BBN 提高了约 7%，较 Hybrid-SC 提高了约 3%~4%。在长尾 ImageNette 数据集上，较先进的不平衡图像检索算法 BCL 提高了约 1%，较 BBN 提高了约 6%，较 Hybrid-SC 提高了约 3%。

4.2. 消融实验

为了证明本文所提出的提取模块的有效性，本节对其进行消融实验。如表 3 所示。本文通过将模型中的特征提取模块用 AlexNet、ResNet50 与 VGG 模块分别替换，在不平衡数据集 ImageNette 数据集(64 位 bit)与 CIFAR-10 数据集(64 位 bit)上利用 3 种不平衡因子和 mAP 作为性能指标进行实验验证。

Table 3. Ablation experiment results of replacing feature extraction module
表 3. 替换特征提取模块的消融实验结果

	ImageNette			CIFAR-10		
特征提取模块	100	50	10	100	50	10
Alexnet	67.34	70.16	72.57	73.04	76.57	78.16
Resnet50	83.28	86.49	88.15	85.73	88.17	90.83
VGG	79.91	82.65	84.64	79.93	83.48	85.29
Ours	**87.64**	**90.46**	**92.43**	**88.18**	**91.63**	**93.86**

对比表 3 中在数据集 ImageNette 与 CIFAR-10 的检索结果，可以看出，不平衡率越低，检索效果越好，这是因为类别分布越均衡时，模型能够从各类样本中学习到更加充分的判别特征，从而避免了对多数类的过拟合和对少数类的忽视。在不平衡数据集 ImageNette 上，本文提出的特征提取模块相较于传统的特征提取模块 AlexNet、ResNet50 与 VGG 分别提升约 0.20、0.04、0.08。

在非平衡数据集 CIFAR-10 上，本文提出的特征提取模块相较于传统的特征提取模块 AlexNet、ResNet50 与 VGG 分别提升约 0.17、0.05、0.11。经过实验验证，本文中所介绍的特征提取模块相对于传统特征提取模块具备明显的优势和合理性。

4.3. 可视化实验

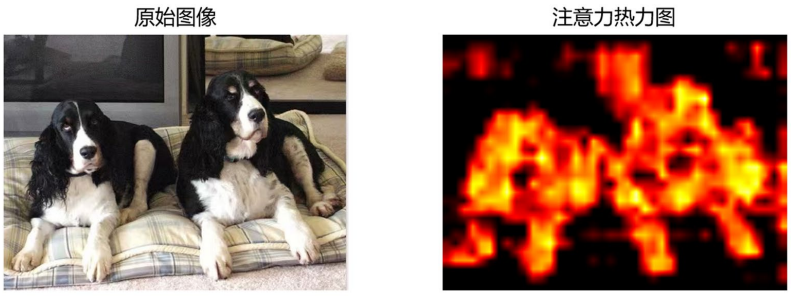
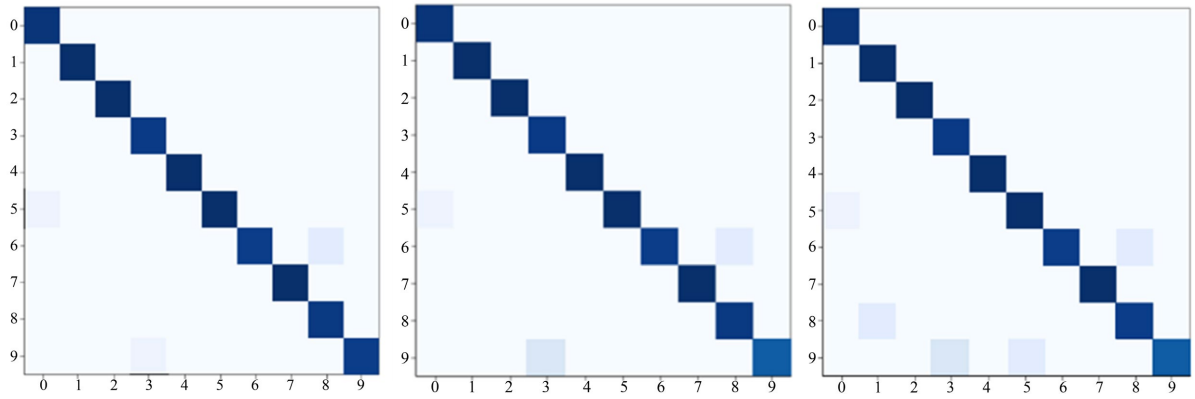


Figure 5. Visualized attention heat map
图 5. 可视化的注意力热力图

为了展示该特征提取模块的效果，本节对注意力识别进行可视化。可视化结果如图 5 所示。图 5 依次为原始图像、注意力热力图。注意力热力图(Attention Heatmap)是注意力热图的一种可视化方式，它将注意力热图应用到原始图像上，以显示出图像中关注的区域。透过叠加注意力热图和原始图像，可以更直观地了解模型对各个区域的关注程度。通过对图 5 中的注意力热力图进行分析可知，红色的部分表示

模型对 应该区域的关注程度较高，即认为该区域包含与任务相关的有用信息。相反，较黑色的部分表示模型对该区域的关注程度较低，即认为该区域可能包含与任务无关的不重要信息。

为了进一步探究不平衡率对模型中性能检测指标以及哈希码的紧凑性与判别性的影响，在ImageNette 和 CIFAR-10 上将模型输出的 64 bit 哈希编码的实验结果利用混淆矩阵方法进行可视化。图 6 的混淆矩阵中，用蓝色表示识别准确度，颜色的深浅与模型识别的准确率成正比。水平方向代表样本预测标签，垂直方向代表样本真实的标签，水平轴和垂直轴均代表图像类别。图中对角线均为高亮状态即说明同一类别的相关性最高。可以看出，大多数样本被正确分类，其对应位置集中分布于对角线上，显示出模型在主干类别上的识别能力较强。与此同时，部分非对角线区域存在轻微的误差分布，尤其集中在某些相似类别之间，这可能是由于这些类别在特征空间中存在较高的语义重叠或纹理结构接近所导致的。



(a) 不平衡率为 10　(b) 不平衡率为 50　(c)不平衡率为 100

Figure 6. Confusion matrix results for ImageNette
图 6. ImageNette 的混淆矩阵结果

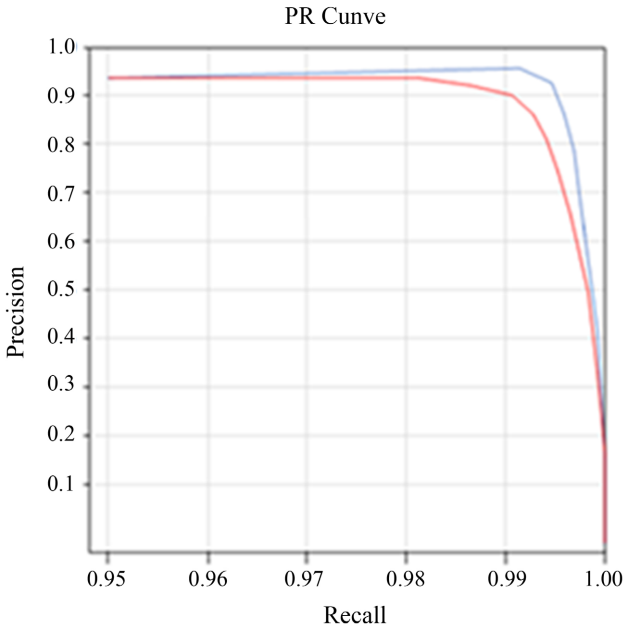


Figure 7. Comparison of PR curves between this method and existing methods
图 7. 本文方法和现有方法的 PR 曲线对比

为了展示本文方法在长尾数据集的优越性，将本文方法和 BCL 方法的检索结果进行可视化如图 7 所示，所使用的数据集为 ImageNette，不平衡率为 10，模型选择 64 bit 哈希编码。在图 7 的 PR 曲线中，P 表示 Precision，R 表示 Recall，该曲线反映了准确率与召回率之间的关系。曲线越靠近右上方，代表模型性能越好。从实验结果可以看出，所提出的方法较 BCL 方法有更佳的检索性能。本文方法的优越性可能主要得益于该方法采用了双流网络架构，关注图片的全局和局部特征，同时采用了自蒸馏损失等多目标损失，增强了同类数据之间的相似性保持，同时拉开了不同类数据之间的距离。

5. 本文小结

在本文中，我们提出了一种基于双流网络架构的深度哈希检索模型，旨在提高长尾数据集下的检索效率和准确性。该模型通过融合 EfficientNet-B3 和 Swin Transformer 的优势，实现了对图像全局和局部特征的高效提取。EfficientNet-B3 以其高效的计算能力和强大的局部特征提取能力而闻名，而 Swin Transformer 则在捕捉全局特征和注意力机制方面表现出色。通过将两者结合，我们的模型能够更全面地理解图像内容，从而为后续的哈希编码提供更丰富的特征表示。

在特征提取的基础上，我们进一步引入了多种损失函数来优化哈希码的质量。具体而言，我们采用了余弦相似度来计算自蒸馏损失，通过最小化同一样本视图之间的余弦距离，确保模型在数据增强后的表示一致性。此外，我们还引入了基于哈希代理的相似性学习损失，通过学习带有温度标度的汉明一致性，提高哈希码的判别能力。最后，基于二元交叉熵的量化损失被用于减少哈希码从连续空间到离散空间的量化误差，从而进一步提高哈希码的质量。

通过这种多目标优化策略，我们的模型在长尾数据集上取得了良好的检索效率和准确性。实验结果表明，这种结合双流网络架构和多目标损失函数的设计不仅能够有效提高模型对长尾数据的适应能力，还能够在大规模数据集中实现高效的检索性能。这种设计为长尾哈希检索领域提供了一种新的解决方案，具有重要的理论和实践意义。为展现本文所提模型的性能，分别设计了模型在在通用数据集 NUS-WIDE、MS-COCO、CIFAR-10 以及不平衡数据集 CIFAR-10、ImageNette 上的对比实验与消融实验，同时本文进行可视化实验验证。实验表明，本文所提供的网络架构在长尾图像检索任务中取得优秀性能。

参考文献

[1] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, **30**.

[2] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, **25**.

[3] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., *et al*. (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. 2021 *IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 9992-10002. https://doi.org/10.1109/iccv48922.2021.00986

[4] Touvron, H., Cord, M. and Jégou, H. (2022) DeiT III: Revenge of the Vit. In: *Lecture Notes in Computer Science*, Springer, 516-533. https://doi.org/10.1007/978-3-031-20053-3_30

[5] Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., *et al*. (2022) TinyViT: Fast Pretraining Distillation for Small Vision Transformers. In*: Lecture Notes in Computer Science*, Springer, 68-85. https://doi.org/10.1007/978-3-031-19803-8_5

[6] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L. (2018) MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 4510-4520. https://doi.org/10.1109/cvpr.2018.00474

[7] Chua, T., Tang, J., Hong, R., Li, H., Luo, Z. and Zheng, Y. (2009) NUS-WIDE. *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

[8] Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., *et al*. (2014) Microsoft COCO: Common Objects in Context. In: *Lecture Notes in Computer Science*, Springer, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

[9] Cao, K., Wei, C., Gaidon, A., *et al*. (2019) Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss.

Advances in Neural Information Processing Systems, **32**.

[10] Cui, Y., Jia, M., Lin, T., Song, Y. and Belongie, S. (2019) Class-Balanced Loss Based on Effective Number of Samples. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 9260-9269. https://doi.org/10.1109/cvpr.2019.00949

[11] Zhou, B., Cui, Q., Wei, X.S., *et al*. (2020) BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition. *Proceedings of the IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 9719-9728.

[12] Slaney, M. and Casey, M. (2008) Locality-Sensitive Hashing for Finding Nearest Neighbors [Lecture Notes]. *IEEE Signal Processing Magazine*, **25**, 128-131. https://doi.org/10.1109/msp.2007.914237

[13] 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

[14] Kang, W., Li, W. and Zhou, Z. (2016) Column Sampling Based Discrete Supervised Hashing. *Proceedings of the AAAI Conference on Artificial Intelligence*, **30**, 1230-1236. https://doi.org/10.1609/aaai.v30i1.10176

[15] Gui, J., Liu, T., Sun, Z., Tao, D. and Tan, T. (2018) Fast Supervised Discrete Hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 490-496. https://doi.org/10.1109/tpami.2017.2678475

[16] 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, 5609-5618. https://doi.org/10.1109/iccv.2017.598

[17] Cao, Y., Long, M., Liu, B., *et al*. (2018) Deep Cauchy Hashing for Hamming Space Retrieval. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 1229-1237.

[18] Su, S., Zhang, C., Han, K., *et al*. (2018) Greedy Hash: Towards Fast Optimization for Accurate Hash Coding in CNN. *Advances in Neural Information Processing Systems*, **31**, 1-10.

[19] 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. https://doi.org/10.1109/cvpr42600.2020.00315

[20] 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*, Seattle, 13-19 June 2020, 3080-3089. https://doi.org/10.24963/ijcai.2020/115

[21] Hoe, J.T., Ng, K.W., Zhang, T., *et al*. (2021) One Loss for All: Deep Hashing with a Single Cosine Similarity Based Learning Objective. *Advances in Neural Information Processing Systems*, **34**, 24286-24298.

[22] Wang, P., Han, K., Wei, X., Zhang, L. and Wang, L. (2021) Contrastive Learning Based Hybrid Networks for Long-Tailed Image Classification. 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 943-952. https://doi.org/10.1109/cvpr46437.2021.00100

[23] Li, T., Cao, P., Yuan, Y., Fan, L., Yang, Y., Feris, R., *et al*. (2022) Targeted Supervised Contrastive Learning for Long-Tailed Recognition. 2022 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 6908-6918. https://doi.org/10.1109/cvpr52688.2022.00679