DOI: 10.13718/j.cnki.xdzk.2024.05.002
刘传升，丁卫平，程纯，等. ViTH：面向医学图像检索的视觉 Transformer 哈希改进算法[J]. 西南大学学报(自然科学版)，2024，46(5)：11-26.

ViTH：面向医学图像检索的视觉 Transformer 哈希改进算法

刘传升，　丁卫平，　程纯，　黄嘉爽，　王海鹏

南通大学 信息科学技术学院，江苏 南通 226019

摘要：对海量的医学图像进行有效检索会给医学诊断和治疗带来极其重要的意义. 哈希方法是图像检索领域中的一种主流方法，但在医学图像领域的应用相对较少. 针对此，提出一种面向医学图像检索的视觉 Transformer 哈希改进算法. 首先使用视觉 Transformer 模型作为基础的特征提取模块，其次在 Transformer 编码器的前、后端分别加入幂均值变换(Power-Mean Transformation，PMT)，进一步增强模型的非线性性能，接着在 Transformer 编码器内部的多头注意力(Multi-Head Attention，MHA)层引入空间金字塔池化(Spatial Pyramid Pooling，SPP)形成多头空间金字塔池化注意力(Multi-Head Spatial Pyramid Pooling Attention，MHSPA)模块，该模块不仅可以提取全局的上下文特征，而且可以提取多尺度的局部上下文特征，并将不同尺度的特征进行融合. 最后在输出幂均值变换层之后将提取到的特征分别通过两个多层感知机(Multi-Layer Perceptrons，MLPs)，上分支的 MLP 用来预测图像的类别，下分支的 MLP 用来学习图像的哈希码. 在损失函数部分，充分考虑了成对损失、量化损失、平衡损失以及分类损失来优化整个模型. 在医学图像数据集 ChestX-ray14 和 ISIC 2018 上的实验结果表明，该研究所提出的算法相比于经典的哈希算法具有更好的检索效果.

关　键　词：医学图像检索；视觉 Transformer；哈希；幂均值变换；空间金字塔池化

中图分类号：TP391　　**文献标志码**：A

开放科学(资源服务)标识码(OSID)：

文 章 编 号：1673-9868(2024)05-0011-16

ViTH：Improved Vision Transformer Hashing Algorithm for Medical Image Retrieval

LIU Chuansheng，　DING Weiping，　CHENG Chun，
HUANG Jiashuang，　WANG Haipeng

School of Information Science and Technology，Nantong University，Nantong Jiangsu 226019，China

收稿日期：2023-05-27
基金项目：国家自然科学基金项目(61976120，62102199)；教育部人文社会科学研究青年基金项目(21YJCZH013)；江苏省自然科学基金项目(BK20231337)；江苏省高等学校自然科学研究重大项目(21KJA510004)；江苏省研究生科研与实践创新计划项目(SJCX22_1615).
作者简介：刘传升，硕士研究生，主要从事深度学习、多媒体信息检索研究.
通信作者：丁卫平，教授，博士研究生导师.

Abstract: Effective retrieval of huge number of medical images will bring extremely important significance to medical diagnosis and treatment. Hashing method is a mainstream method in the field of image retrieval, but the application in the field of medical images is relatively small. For this, an improved Vision Transformer Hashing algorithm for medical image retrieval is proposed. Firstly, the Vision Transformer model is used as the base feature extraction module; secondly, the Power-Mean Transform (PMT) is added to the front and back ends of the Transformer encoder respectively to further enhance the nonlinear performance of the model; and then the Spatial Pyramid Pooling (SPP) is introduced into the Multi-Head Attention (MHA) layer inside the Transformer encoder to form the Multi-Head Spatial Pyramid Pooling Attention (MHSPA) module, which not only extracts global contextual features, but also extracts multi-scale local contextual features and fuses features of different scales; finally, after outputting the Power-Mean Transformation layer, the extracted features are passed through two Multi-Layer Perceptrons (MLPs) respectively, and the MLP in the upper branch is used to predict the category of the image and the MLP in the lower branch is used to learn the hashing codes of the images. In the loss function part, pairwise loss, quantization loss, balanced loss, and classification loss are fully considered to optimize the whole model. Experimental results on the medical image dataset ChestX-ray14 and ISIC 2018 show that the proposed algorithm in this paper has better retrieval results compared to the classical hashing algorithm.

Key words: medical image retrieval; vision transformer; hashing; power-mean transformation; spatial pyramid pooling

近年来，X射线、核磁共振、计算机断层扫描、多普勒彩超等技术不断发展，由此所产生的医学影像也逐渐增加[1]. 不同的医学影像技术涵括了对人类不同身体部位的视觉解读，为了能够做出更合理的诊断，医生大多需要借助以往相关病例的影像资料来判断当前患者所患的病症[2]. 然而，针对如此海量的医学图像，如何快速且高效地检索到相关图像是一项重大挑战.

早期，基于文本的图像检索(Text-Based Image Retrieval, TBIR)是一种主流的检索技术，该技术一般通过图像文本形式的启发式信息(如标签、图像描述符等)进行检索，因此需要工作人员对每一张图像进行手工标注[3]. 然而，针对数以百万计的图像数据，手工标注每一幅图像显然是不现实的. 为了克服这一弊端，基于内容的图像检索(Content-Based Image Retrieval, CBIR)技术迅速兴起. CBIR是一种计算机视觉技术，它提供了一种在大型数据库中搜索相关图像的方法，这种搜索方法通过视觉特征(如颜色、形状和纹理等)来描述图像，而检索精度主要取决于这些选定的特征[4]. 在CBIR中，对于一张给定的待查询图像，系统会从数据库中检索出一些在颜色、形状和纹理等方面与其相似的图像返回给用户. 假设数据库中的图像和待查询图像都是由实值特征表示，搜索相关图像的最简单方法就是根据它们在特征空间中的距离进行排序，并返回距离最近的图像. 然而，对于大规模图像检索而言，CBIR同样面临着存储空间大、检索精度低且速度慢的缺点[5].

为了解决内存成本高、检索速度慢且精度低等一系列问题，基于哈希的图像检索方法被提出并逐渐得到应用. 哈希方法主要是将高维图像特征映射到低维汉明空间并生成紧凑的二进制哈希码，同时还能保持原始图像数据的相似性. 该方法极大地降低了特征维度，避免了维度过高问题，在检索精度和检索速度方面得到了极大改善[6-7]. 哈希方法具体分为两类：数据独立和数据依赖. 其中，在数据独立的哈希算法领域中，最著名的就是局部敏感哈希[8](Locality-Sensitive Hashing, LSH)及其变形算法. 该类算法采用随机映射的方式来获得哈希函数，并且一般需要足够长的哈希码位数才能够达到较高的精度. 相比之下，数据依赖的哈希算法只需要极短的哈希码就可以达到较为理想的精度，该类算法从训练集中学习哈希函数，故又称为学习哈希[9]. 因此，在实际应用中，数据依赖的哈希算法比数据独立的哈希算法更流行.

近年来，受益于深度学习在图像处理方面所表现出的强大性能，人们开始将哈希方法和深度学习相结合，提出了深度哈希算法. 深度哈希算法主要利用卷积神经网络(Convolutional Neural Network，CNN)来提取图像特征，然后利用提取到的特征进行哈希函数学习，这不仅有效避免了语义鸿沟问题，还极大地提高了检索性能[10]. 根据对标签信息的利用，深度哈希算法又分为无监督、半监督和监督 3 种方式. 一般来说，监督深度哈希算法的精度要高于其他两种方式，代表性的有基于成对标签的深度监督哈希[11]、深度成对监督哈希[12]、深度柯西哈希[13]等，以及基于三元组标签的深度三元标签监督哈希[14]、深度三元组量化[15]、基于注意力的三元哈希[16]等.

2017 年，Vaswani 等[17]提出了 Transformer 模型，并在自然语言处理领域取得了巨大成功. 2020 年，Carion 等[18]提出了 DETR 模型，并引入 Transformer 做目标检测任务. 2021 年，Dosovitskiy 等[19]在 Transformer 的基础上提出了视觉 Transformer 模型，并将其应用于计算机视觉领域；Han 等[20]在视觉 Transformer 的基础上提出了 TNT 模型，进一步提升了模型在数据上的学习能力和泛化性；Wang 等[21]提出了 PVT 模型，并将 Pyramid CNN 的思路引入 Transformer，大幅提高了输出结果的分辨率. 大量实验表明，Transformer 在各种计算机视觉任务中(如图像分类[19]、目标识别[22]等)优于许多基于 CNN 的方法. 近期，在哈希图像检索领域也出现了许多基于 Transformer 的模型，如 VTS[23]、TransHash[24]、HashFormer[25]、ViT2Hash[26]等，这些模型也取得了不错的效果.

基于上述分析，本研究提出面向医学图像检索的视觉 Transformer 哈希(Vision Transformer Hashing，ViTH)改进算法，同时这也是一种完全不采用 CNN 作为主架构的深度哈希算法. 本研究使用视觉 Transformer 作为基础特征提取模块来提取医学图像的视觉特征. 首先在 Transformer 编码器的前、后端分别加入了幂均值变换(Power-Mean Transformation，PMT)[27]来进一步增强模型的非线性性能，然后在 Transformer 编码器内部的多头注意力(Multi-Head Attention，MHA)层引入空间金字塔池化(Spatial Pyramid Pooling，SPP)形成多头空间金字塔池化注意力(Multi-Head Spatial Pyramid Pooling Attention，MHSPA)[28]模块，接着在输出幂均值变换之后将提取到的特征分别通过两个多层感知机(Multi-Layer Perceptrons，MLPs)，上分支的 MLP 用来预测图像的类别，下分支的 MLP 用来学习图像的哈希码，最后通过成对损失、量化损失、平衡损失以及分类损失来优化整个模型.

本研究主要贡献如下：

1) 提出了一种面向医学图像检索的视觉 Transformer 哈希改进算法.

2) 为了进一步提取具有细微差异的医学图像特征，本研究在多头注意力层中引入空间金字塔池化，形成多头空间金字塔池化注意力模块. 在损失函数中，除了成对损失、量化损失和平衡损失，该算法还设计了分类损失来进一步优化模型所学习的哈希码.

3) 本研究提出的算法不仅仅适用于 ChestX-ray14 和 ISIC 2018，还可以扩展到其他医学图像数据集. 另外，在 ChestX-ray14 和 ISIC 2018 医学图像数据集上验证了算法的有效性. 相比目前的算法，本研究取得了较好的检索效果.

1　相关工作

1.1　医学图像检索

医学图像具有相似性大、类别多等特性，在大量医学影像中高效准确地检索到所需图像一直是一项挑战. 近年来，很多学者逐渐将哈希算法应用到医学图像检索领域并取得了显著的效果.

Lu 等[29]结合模糊逻辑技术和深度神经网络提出深度模糊哈希，利用模糊规则来模拟数据背后的不确定性. Wang 等[30]提出基于细粒度相关分析的医学图像检索，有效减少了医学图像中的冗余信息. Xu 等[31]针对医学图像提出多流形深度判别跨模态哈希，多模态流形相似性集成了异构数据上的多个子流形以保持实例之间的相关性. Yang 等[32]提出一个名为 CenterHash 的深度贝叶斯哈希学习框架，它可以将多模态数据映射到共享的 Hamming 空间，并从不平衡的多模态神经图像中学习哈希码，解决了类间差异小

和模态间差异大所造成的难题.

1.2 Patch 编码器

假设输入图像 $\boldsymbol{I} \in \boldsymbol{R}^{H\times W\times C}$(其中 H, W 分别代表图像的高度和宽度, C 代表通道数),首先将 $\boldsymbol{I}$ 分为 N 个互不重叠的 patch,然后将 N 个 patch 展平成二维的 patches 向量 $\boldsymbol{X}_P \in \boldsymbol{R}^{N\times(P^2\times C)}$(其中 $N=HW/P^2$ 代表 patch 的总数),最后将 $\boldsymbol{X}_P$ 通过线性映射层映射到 D 维空间中,形成序列 $X_P^k \in \boldsymbol{R}^D$, $k=1,2,\cdots,N$. 位置嵌入被添加到 patch 编码器之后保留位置信息. 与文献[19]不同的是,本研究不使用 0 号 class token,而是将所形成的 N 个 patch 进行编码. 具体过程如下:

$$z_0=[x_P^1\boldsymbol{E};\ x_P^2\boldsymbol{E};\ \cdots;\ x_P^N\boldsymbol{E}]+\boldsymbol{E}_{\text{pos}} \tag{1}$$

式中: $\boldsymbol{E} \in \boldsymbol{R}^{P^2\times C\times D}$ 代表线性映射矩阵; $\boldsymbol{E}_{\text{pos}} \in \boldsymbol{R}^{N\times D}$ 代表位置嵌入矩阵.

1.3 Transformer 编码器

Transformer 编码器[17-19] 由 L 个 Transformer 块组成,每个 Transformer 块包含层归一化[33](Layer Normalization, LN)、多头注意力机制(Multi-Head Attention, MHA)块和多层感知机(Multi-Layer Perceptron, MLP)块,残差连接[34] 分别位于每个块之后. 因此,每个 Transformer 块的计算公式如下:

$$z_l'=\text{MHA}(\text{LN}(z_{l-1}))+z_{l-1} \qquad l=1,2,\cdots,L \tag{2}$$

$$z_l=\text{MLP}(\text{LN}(z_l'))+z_{l-1} \qquad l=1,2,\cdots,L \tag{3}$$

式中: MHA, LN 和 MLP 分别代表多头注意力机制块、层归一化以及多层感知机块.

2 ViTH 算法框架

本节详细介绍本研究所提出的算法(ViTH)框架. 首先介绍本研究所使用到的一些符号及含义,然后在此基础上给出 ViTH 的整体框架(图 1),并详细阐述各模块的具体作用,最后提出模型的损失函数并进行优化.

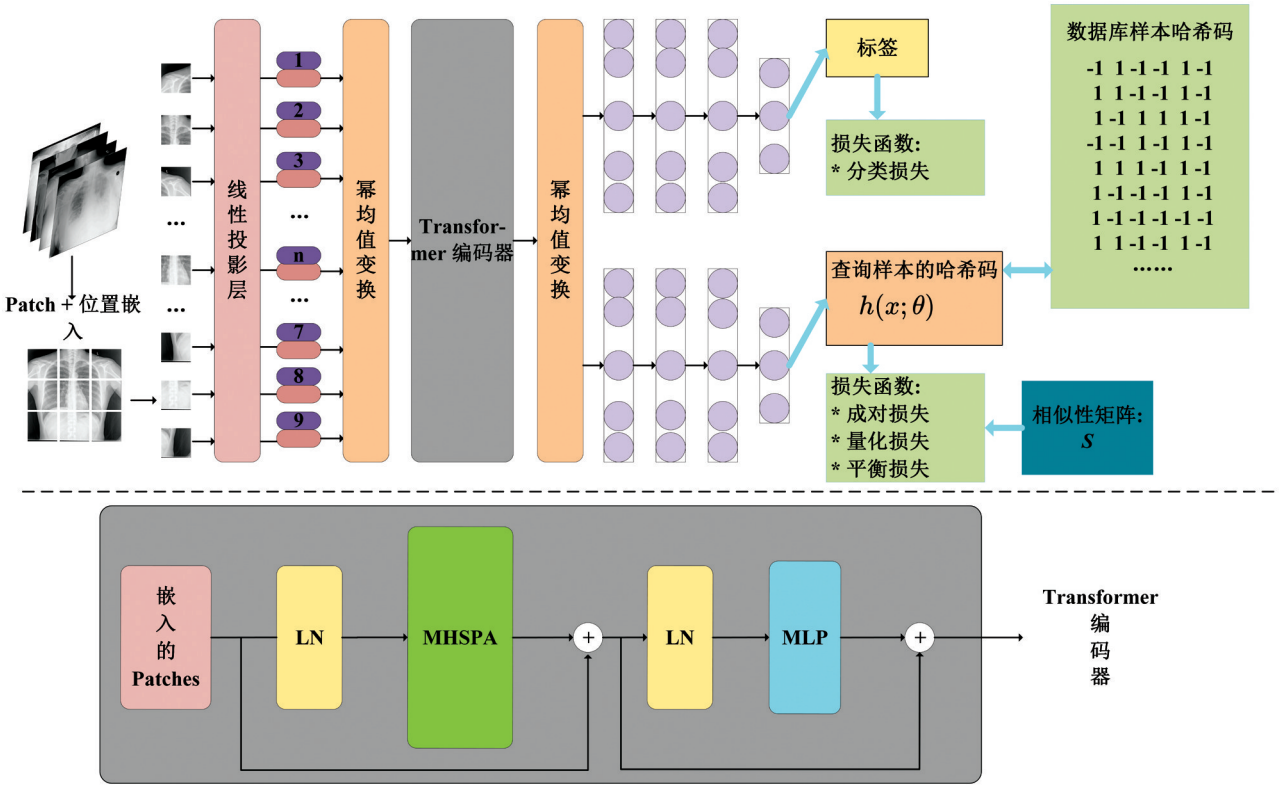

图 1 ViTH 整体框架

2.1　符号定义

本研究使用 $\boldsymbol{v}$ 等小写字母表示向量，$\boldsymbol{V}$ 等大写字母表示矩阵. $\boldsymbol{V}$ 的第 i 行第 j 列元素记为 V_{ij}；$\boldsymbol{V}$ 的第 i 行记为 V_{i*}；$\boldsymbol{V}$ 的第 j 列记为 V_{*j}；$\boldsymbol{V}$ 的转置记为 V^{T}. $\|\cdot\|$ 代表矩阵的 Frobenius 范数；$\odot$ 代表矩阵间的哈达玛积；$tr(\cdot)$ 代表矩阵的迹运算. 此外，符号函数和双曲正切函数的定义如下：

$$\operatorname{sign}(x)=\begin{cases}1 & x\geqslant 0,\\ -1 & x<0\end{cases} \tag{4}$$

$$\tanh(x)=\frac{\mathrm{e}^{x}-\mathrm{e}^{-x}}{\mathrm{e}^{x}+\mathrm{e}^{-x}} \tag{5}$$

给定医学图像数据库 $X=\{x_i\}_{i=1}^{n}$，其中 n 代表数据库样本的总数. $L=\{l_i\}_{i=1}^{n}\in\{0,1\}^{n\times c}$ 是数据库样本集 X 的标签信息，其中 c 代表 X 的类别数. 样本集 X 的成对相似性矩阵 $\boldsymbol{S}$ 一般定义为：如果 x_i 和 x_j 至少共享一个类别，则 $S_{ij}=1$；否则 $S_{ij}=0$. 这显然是有局限性的，因为两张图像共享的标签数越多，则它们之间的相似性越大；反之，如果没有共享的标签，则它们是不相似的. 本研究重新定义成对相似性矩阵[35] $\boldsymbol{S}=\dfrac{2\|l_i\cap l_j\|}{\|l_i\|+\|l_j\|-\|l_i\cap l_j\|}\in[0,2]^{n\times n}$. 如果 $S_{ij}=0$，则代表 x_i 与 x_j 不相似；如果 $S_{ij}=2$，则代表 x_i 与 x_j 完全相似；如果 S_{ij} 介于 0 到 2 之间，则代表 x_i 与 x_j 具有一定的相似性. 训练过程中，本研究从数据库样本集 X 中随机抽取 m 个样本作为查询集(训练样本)$Q=\{q_t\}_{t=1}^{m}$. 因此，查询集 Q 和数据库样本集 X 之间的成对相似性矩阵为 $\boldsymbol{S}'\in[0,2]^{m\times n}$.

哈希学习的目的是针对每一个数据库样本 x_i 学习一个哈希函数 $h(x_i)\in\{-1,+1\}^{k}$，以便将各图像映射为紧凑的二进制码 $B=\{b_i\}_{i=1}^{n}\in\{-1,+1\}^{n\times k}$，并且所学习的二进制码 B 可以保持图像在原始空间中的相似性.

2.2　框架

2.2.1　特征学习模块

本研究使用视觉 Transformer 来提取图像特征. 具体来说，首先将输入图像裁剪成 9 个大小相同且互不重叠的 patch，然后将各 patch 展平成二维的 patches 向量. 类似地，本研究使用一个可学习的线性投影层将各向量映射到 $D(D=2\,048)$ 维的空间中，得到序列 $X_P^k\in R^D$，$k=1,2,\cdots,9$. 位置嵌入则被添加到 patch 编码器之后，其作用是给各 patch 添加相对位置，以防丢失位置信息. 与研究[19]不同的是，本研究不使用 0 号 class token，而是将所形成的 9 个 patch 进行编码，如公式(1)所示.

此外，本研究在 Transformer 编码器的前后端分别加入 PMT[27] 操作来增强模型的非线性. 假设 PMT 的输入为 x、输出为 l，PMT 则将 x 变换为 $[\ln(x+\beta),\ln^2(x+\beta)]$，其中 β 是一个常数(本研究取 $\beta=1$). 在前向传播过程中，假设 $\dfrac{\partial l}{\partial y}$ 是 PMT 输出的梯度(此处 $y=[\ln(x+\beta),\ln^2(x+\beta)]$)，由链式法则得：

$$\frac{\partial l}{\partial y}=\frac{\partial l}{\partial[\ln(x+\beta),\ln^2(x+\beta)]} \tag{6}$$

$$\frac{\partial l}{\partial x^{\mathrm{T}}}=\frac{\partial l}{\partial y^{\mathrm{T}}}\frac{\partial y}{\partial x^{\mathrm{T}}}=\frac{\partial l}{\partial[\ln(x+\beta)]^{\mathrm{T}}}\frac{\ln(x+\beta)}{\partial x^{\mathrm{T}}}+\frac{\partial l}{\partial[\ln^2(x+\beta)]^{\mathrm{T}}}\frac{\ln^2(x+\beta)}{\partial x^{\mathrm{T}}} \tag{7}$$

通过前向和反向传播，PMT 被集成到整个模型中. 在训练过程中可以学习到更复杂的信息，增强模型的非线性.

本研究所使用的 Transformer 编码器的深度是 6，并且每个 Transformer 编码器都是由 LN，MHSPA，MLP 以及残差连接组成. 图 2 为 SPP 模块，该模块将输入特征图经过自适应平均池化层形成 1×1，2×2，4×4 和 16×16 的特征子图，并将形成的特征子图展平拼接. SPP 的主要作用是对输入特征图进行不同尺度的特征提取，并生成融合多个区域信息的多尺度特征. 在模型中，本研究将 SPP 模块嵌入在 MHA 中形成

MHSPA模块，如图3所示.

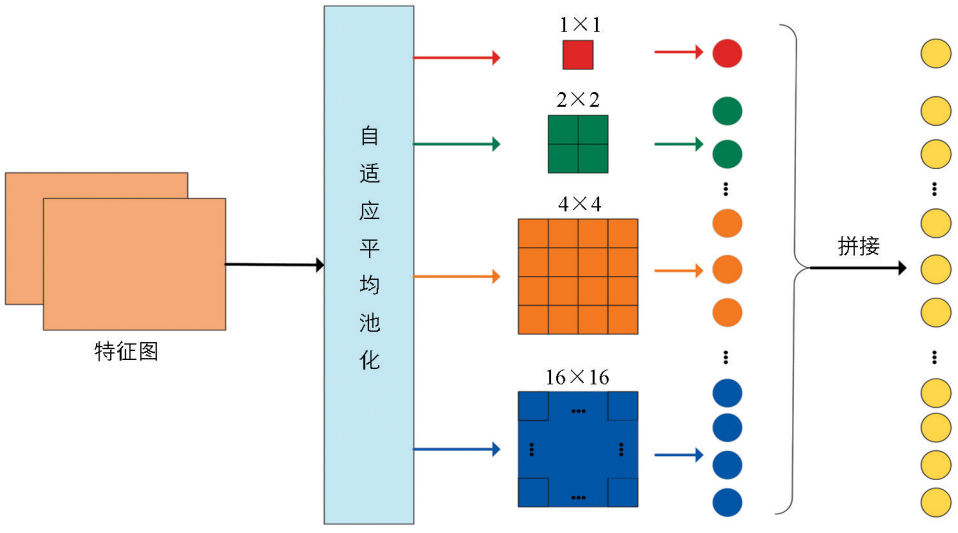

图2 空间金字塔池化(Spatial Pyramid Pooling, SPP)

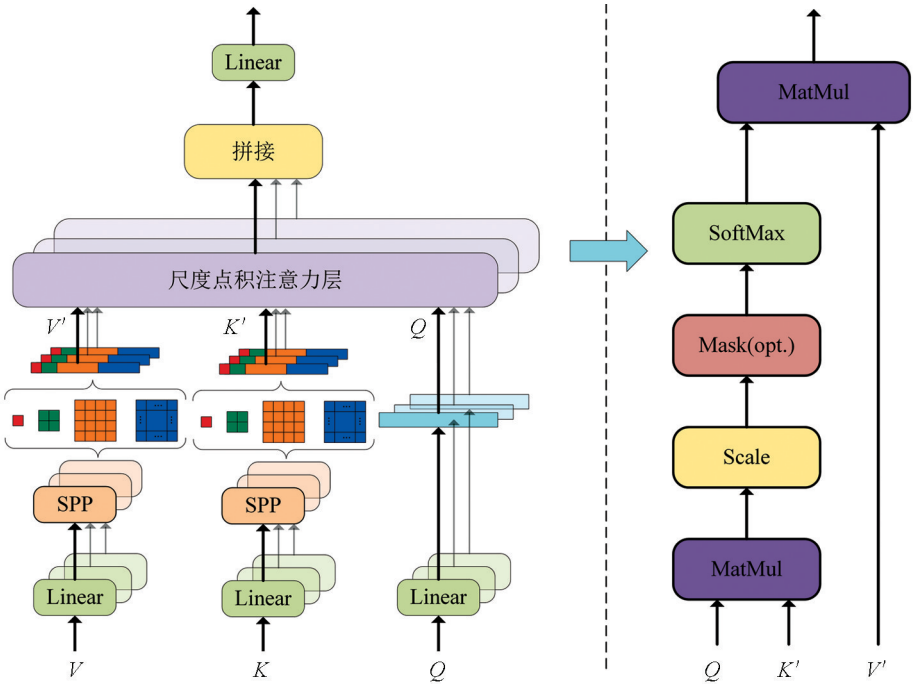

图3 多头空间金字塔池化注意力机制(Multi-Head Spatial Pyramid Pooling Attention, MHSPA)

注意力机制是将查询$\boldsymbol{Q}$和一组键值对$\boldsymbol{K}$，$\boldsymbol{V}$映射到输出，其中$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$以及输出都是矩阵形式. 输出是通过对$\boldsymbol{V}$的加权来计算的，分配给每个$\boldsymbol{V}$的权重则是通过查询$\boldsymbol{Q}$与相应键$\boldsymbol{K}$的兼容性函数来计算. 在MHSPA模块中，$\boldsymbol{V}$和$\boldsymbol{K}$先通过SPP模块，然后再将查询$\boldsymbol{Q}$和经过池化的键值对$\boldsymbol{K}'$和$\boldsymbol{V}'$通过尺度点积注意力层(图3)，该层计算查询$\boldsymbol{Q}$和键值对$\boldsymbol{K}'$和$\boldsymbol{V}'$之间的注意力分数. 假设尺度点积注意力层的查询$\boldsymbol{Q}$和键$\boldsymbol{K}'$的维度为d_k，则MHSPA模块的输出为：

$$\mathrm{MHSPA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}'^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}' \tag{8}$$

式中：$\boldsymbol{K}'=\mathrm{SPP}(\boldsymbol{K})$，$\boldsymbol{V}'=\mathrm{SPP}(\boldsymbol{V})$. $\boldsymbol{Q}$和$\boldsymbol{K}'^{\mathrm{T}}$之间的点积计算每个查询与所有键之间的相似度，softmax(•)

函数对相似度分数进行归一化，以获得每个查询总和为 1 的注意力权重，比例因子$\frac{1}{\sqrt{d_k}}$用于降低点积幅度的影响. 通过使用尺度点积注意力层，模型可以更好地处理输入中的长距离依赖关系，同时避免了点积的数量级过大导致的数值不稳定问题. 若 Transformer 编码器的输入为 z_l，$l=1, 2, \cdots, L$，则输出 Z_o 为：

$$Z_o = \mathrm{MLP}(\mathrm{LN}(\mathrm{MHSPA}(\mathrm{LN}(z_l)) + z_l)) + \mathrm{MHSPA}(\mathrm{LN}(z_l)) + z_l \tag{9}$$

将 Transformer 编码器的输出再次经过 PMT 操作，并将输出分别通过两个 MLP，其中上分支的 MLP 用来预测输入图像的类别，下分支的 MLP 用来生成哈希码.

2.2.2　损失函数

本研究主要考虑了成对损失 L_P、量化损失 L_Q 以及平衡损失 L_B[36-38]. 此外，为了能更好地提高精度，还添加了分类损失 L_C 这一约束项.

成对损失 L_P：训练过程中，本研究通过最小化相似度矩阵 $\boldsymbol{S}'$ 和查询 — 数据库样本之间的哈希码内积 $u_i v_j^T$ 的 L_2 损失来保持查询样本和数据库样本之间的相似性. 成对损失具体定义如下：

$$L_P = \sum_{t=1}^{m}\sum_{j=1}^{n} \| u_t v_j^{\mathrm{T}} - kS'_{tj} \|^2$$
$$\text{s. t. } U = [u_1, u_2, \cdots, u_m]^{\mathrm{T}} \in \{-1, +1\}^{m\times k}, V = [v_1, v_2, \cdots, v_n]^{\mathrm{T}} \in \{-1, +1\}^{n\times k} \tag{10}$$

式中：u_i 和 v_j 分别代表查询样本和数据库样本的哈希码；$u_t = h(q_t) = \mathrm{sign}(\Phi(q_t; \theta_h))$；$\Phi(q_t; \theta_h) \in R^k$ 代表哈希编码分支的输出；θ_h 代表哈希编码分支的网络参数.

由于公式(10)的求解是一种离散优化问题，这是极其难以求解的，因此本研究使用双曲正切函数 $\tanh(\cdot)$ 来近似符号函数 $\mathrm{sign}(\cdot)$，即：

$$\begin{aligned} L_P &= \sum_{t=1}^{m}\sum_{j=1}^{n} \| u_t v_j^{\mathrm{T}} - kS'_{tj} \|^2 = \\ &\sum_{t=1}^{m}\sum_{j=1}^{n} \| h(q_t) v_j^{\mathrm{T}} - kS'_{tj} \|^2 = \\ &\sum_{t=1}^{m}\sum_{j=1}^{n} \| \mathrm{sign}[\Phi(q_t; \theta_h)] v_j^{\mathrm{T}} - kS'_{tj} \|^2 \approx \\ &\sum_{t=1}^{m}\sum_{j=1}^{n} \| \tanh[\Phi(q_t; \theta_h)] v_j^{\mathrm{T}} - kS'_{tj} \|^2 \end{aligned} \tag{11}$$

量化损失 L_Q：由于在成对损失中使用了双曲正切函数 $\tanh(\cdot)$ 来近似符号函数 $\mathrm{sign}(\cdot)$，因此本研究在网络的实值输出和哈希码之间添加一个正则项，即量化损失 L_Q：

$$L_Q = \sum_{t=1}^{m} \| \tanh(\Phi(q_t; \theta_h)) - u_t \|^2 \tag{12}$$

平衡损失 L_B：为了使哈希码尽可能地充满整个 2^k 的汉明空间，并保证每一比特的平衡性，本研究提出了平衡损失 L_B. 该损失函数会确保每一比特上 -1 和 $+1$ 出现的概率尽可能地相等，定义如下：

$$L_B = \sum_k \left(\left| \sum_{t=1}^{m} \mathrm{mean}(u_t) \right|^2 + \left| \sum_{j=1}^{n} \mathrm{mean}(v_j) \right|^2 \right) \tag{13}$$

分类损失 L_C：由于不同图像的标签个数有所差别，有的图像只包含一种疾病，有的却包含多种疾病，本研究在训练过程中针对单标签和多标签图像使用不同的损失函数.

1）多标签分类损失. 针对多标签图像，分类损失定义如下：

$$L_{\mathrm{multi}} = -\frac{1}{c} * \sum_{t=1}^{m} \left\{ l_t * \log[(1+\exp(-\mathrm{pred}(q_t; \theta_c)))^{-1}] + (1-l_t) * \log\left[\frac{\exp(-\mathrm{pred}(q_t; \theta_c))}{1+\exp(-\mathrm{pred}(q_t; \theta_c))}\right] \right\} \tag{14}$$

2）单标签分类损失. 针对单标签图像，分类损失定义如下：

$$L_{\text{single}}=-\frac{1}{m}*\sum_{t=1}^{m}\{l_t*\log[\text{pred}(q_t;\theta_c)]+(1-l_t)*\log[1-\text{pred}(q_t;\theta_c)]\} \tag{15}$$

其中，c 代表数据库样本集 X 的类别数；$\text{pred}(q_t;\theta_c)$ 代表图像 q_t 的预测标签；l_t 代表图像的真实标签；θ_c 代表预测图像类别分支的参数.

因此，总的分类损失函数如下：

$$L_C=L_{\text{multi}}+L_{\text{single}} \tag{16}$$

2.3 优化策略

由 2.2.2 节得，ViTH 模型总的损失函数如下：

$$\min_{V,\theta_h,\theta_c} L=L_P+\alpha L_Q+\beta L_B+\gamma L_C$$
$$\text{s. t. } V\in\{-1,+1\}^{n\times k} \tag{17}$$

其中，α，β 和 γ 都是不敏感的超参数. 本研究使用交替优化算法对式(17)进行求解. 也就是说，在其他参数固定的情况下优化一个参数.

2.3.1 固定 V 和 θ_c，更新 θ_h

当 V 和 θ_c 固定时，使用反向传播算法来更新参数 θ_h. 因此，针对每个查询图像 q_t，梯度计算如下：

$$\frac{\partial L}{\partial z_t^h}=\frac{\partial L_P}{\partial z_t^h}+\alpha\frac{\partial L_Q}{\partial z_t^h}+\beta\frac{\partial L_B}{\partial z_t^h}+\gamma\frac{\partial L_C}{\partial z_t^h}=$$
$$2\sum_{t=1}^{m}\{[(\tanh(z_t^h)v_j^T-kS_{tj}^{'})v_j^T]+\alpha[\tanh(z_t^h)-u_t]\}\odot[1-\tanh(z_t^h)^2] \tag{18}$$

其中 $z_t^h=\Phi(q_t;\theta_h)$. 有了 $\frac{\partial L}{\partial z_t^h}$ 之后，我们可以根据链式法则计算 $\frac{\partial L}{\partial \theta_h}$. 另外，由于 L_B 和 L_C 均独立于 z_t^h，所以 $\frac{\partial L_B}{\partial z_t^h}=0$，$\frac{\partial L_C}{\partial z_t^h}=0$.

2.3.2 固定 V 和 θ_h，更新 θ_c

当 V 和 θ_h 固定时，使用反向传播算法来更新参数 θ_c. 因此，针对每个查询图像 q_t，梯度计算如下：

$$\frac{\partial L}{\partial z_t^c}=\frac{\partial L_P}{\partial z_t^c}+\alpha\frac{\partial L_Q}{\partial z_t^c}+\beta\frac{\partial L_B}{\partial z_t^c}+\gamma\frac{\partial L_C}{\partial z_t^c}=$$
$$\gamma\left(\frac{\partial L_{\text{multi}}}{\partial z_t^c}+\frac{\partial L_{\text{single}}}{\partial z_t^c}\right)=$$
$$\gamma\left\{-\frac{1}{c}*\sum_{t=1}^{m}\frac{\partial}{\partial z_t^c}\left[l_t*\log(\exp(z_t^c))+\log\frac{\exp(-z_t^c)}{1+\exp(-z_t^c)}\right]\right\}+$$
$$\gamma\left\{-\frac{1}{m}*\sum_{t=1}^{m}\frac{\partial}{\partial z_t^c}\left[l_t*\log\frac{z_t^c}{1-z_t^c}+\log(1-z_t^c)\right]\right\}=-$$
$$\frac{\gamma\varepsilon}{c}\sum_{t=1}^{m}\left[l_t-\frac{1}{1+\exp(-z_t^c)}\right]-\frac{\gamma\varepsilon}{m}\sum_{t=1}^{m}\left[l_t*\frac{1}{z_t^c(1-z_t^c)}-\frac{1}{1-z_t^c}\right] \tag{19}$$

式中：$z_t^c=\text{pred}(q_t;\theta_c)$；$\varepsilon$ 是一个常数. 有了 $\frac{\partial L}{\partial z_t^c}$ 之后，我们可以根据链式法则计算 $\frac{\partial L}{\partial \theta_c}$. 另外，由于 L_P，L_Q 和 L_B 均独立于 z_t^c，所以 $\frac{\partial L_P}{\partial z_t^c}=0$，$\frac{\partial L_Q}{\partial z_t^c}=0$，$\frac{\partial L_B}{\partial z_t^c}=0$.

2.3.3 固定 θ_h 和 θ_c，更新 V

当 θ_h 和 θ_c 固定时，我们可以重写式(17). 具体如下：

$$\min_{V} L = L_P + \alpha L_Q + \beta L_B + \gamma L_C =$$

$$\| \widetilde{U} V^{\mathrm{T}} - k S' \|^2 + \alpha \| \widetilde{U} - \widetilde{V} \|^2 + \beta \sum_{k=1}^{k} | V_{*k} \cdot 1 | + \tau =$$

$$\| \widetilde{U} V^{\mathrm{T}} \|^2 - 2k \operatorname{tr}(V^{\mathrm{T}} S'^{\mathrm{T}} \widetilde{U}) - 2\alpha \operatorname{tr}(\widetilde{U} \widetilde{V}^{\mathrm{T}}) + \beta \sum_{k=1}^{k} | V_{*k} \cdot 1 | + \tau =$$

$$\| V \widetilde{U}^{\mathrm{T}} \|^2 - 2\operatorname{tr}(V[k \widetilde{U}^{\mathrm{T}} S' + \alpha \overline{V}^{\mathrm{T}}]) + \beta \sum_{k=1}^{k} | V_{*k} \cdot 1 | + \tau =$$

$$\operatorname{tr}(V[\widetilde{U}^{\mathrm{T}} V \widetilde{U}^{\mathrm{T}} - 2k \widetilde{U}^{\mathrm{T}} S' - 2\alpha \overline{V}^{\mathrm{T}}]) + \beta \sum_{k=1}^{k} | V_{*k} \cdot 1 | + \tau$$

$$\text{s. t. } V \in \{-1, +1\}^{n \times k} \tag{20}$$

式中：$\widetilde{U} = [\widetilde{u}_1, \widetilde{u}_2, \cdots, \widetilde{u}_t, \cdots, \widetilde{u}_m]^{\mathrm{T}}$；$\widetilde{u}_t = \tanh(\Phi(q_t; \theta_h))$，$1 = [1, 1, \cdots, 1]^{\mathrm{T}} \in R^{1 \times n}$；$\tau$ 是一常数. 另外，由于本研究的查询样本是从数据库样本中随机抽取的，所以 $\widetilde{V} \in R^{m \times k}$ 代表查询样本在数据库中所对应的哈希码，$\overline{V} \in R^{n \times k}$ 代表将 $\widetilde{V}$ 嵌入在 V 中的编码.

为了简便运算，本研究令

$$Z = (\widetilde{U}^{\mathrm{T}} V \widetilde{U}^{\mathrm{T}} - 2k \widetilde{U}^{\mathrm{T}} S' - 2\alpha \overline{V}^{\mathrm{T}})^{\mathrm{T}}$$

因此，式(20) 进一步转化为：

$$\min_{V} L = tr(V Z^{\mathrm{T}}) + \beta \sum_{k=1}^{k} | V_{*k} \cdot 1 | + \tau$$

$$\text{s. t. } V \in \{-1, +1\}^{n \times k} \tag{21}$$

针对式(21)，对数据库编码 V 采用逐比特更新的策略. 也就是说，在更新某一比特哈希码时，固定其他比特保持不变. 因此，等式(21) 可以重写为：

$$\min_{V_{*k}} L = tr(V_{*k} Z_{*k}^{T}) + \beta \sum_{k=1}^{k} | V_{*k} \cdot 1 | + \tau$$

$$\text{s. t. } V_{*k} \in \{-1, +1\}^{n \times k} \tag{22}$$

因此，式(22) 的最优解如下：

$$V_{*k} = \begin{cases} -\operatorname{sign}(Z_{*k} + \beta \cdot 1^{\mathrm{T}}) & V_{*k} \cdot 1 \geqslant 0 \\ -\operatorname{sign}(Z_{*k} - \beta \cdot 1^{\mathrm{T}}) & V_{*k} \cdot 1 < 0 \end{cases} \tag{23}$$

2.4　样本外扩展

在模型训练完成之后，对于查询集 $Q = \{q_t\}_{t=1}^{m}$ 之外的样本，本研究通过样本外扩展生成对应样本的哈希码. 具体如下：

$$b_i = h(x_i) = \operatorname{sign}[\Phi(x_i; \theta_h)] \tag{24}$$

2.5　算法实现步骤

本研究算法实现步骤如下：

算法 1　视觉 Transformer 哈希算法(ViTH)

输入　数据库样本集 $X = \{x_i\}_{i=1}^{n}$，标签矩阵 $\boldsymbol{L} = \{l_i\}_{i=1}^{n} \in \{0, 1\}^{n \times c}$，码长 k，批处理大小 b，迭代数目 *iters* 和 *epochs*.

输出　数据库样本哈希码 V，最优神经网络参数 θ_h 和 θ_c.

(1) 初始化神经网络参数 θ_h 和 θ_c，数据库样本哈希码 V；

(2) **for** *iter* = 1 *to iters* **do**

(3)　　随机生成查询集 $Q = \{q_t\}_{t=1}^{m}$，计算查询 — 数据库的相似度矩阵 $\boldsymbol{S}' \in [0, 2]^{m \times n}$；

(4) **for** $epoch=1$ *to* $epochs$ **do**

(5) **for** $batch=1$ to $\frac{m}{b}$ **do**

(6) 计算批数据的哈希码，根据式(18)和链式法则更新θ_h，根据式(19)和链式法则更新θ_c；

(7) **end**

(8) **end**

(9) **for** $bit=1$ *to* k **do**

(10) 根据式(23)更新V；

(11) **end**

(12) **end**

3 实验结果与分析

3.1 实验数据集

ChestX-ray14：该数据集共包含112 120幅正面的胸部X射线图像，每幅图像的尺寸为1 024×1 024. 除健康人群的图像外，每幅图像都带有一种或多种常见的胸部疾病，共14种. 本研究只关注带有疾病的X射线图像，由于各类疾病的数量是极其不平衡的，因此剔除了数量极少的种类，并对相对较少的疾病图像通过旋转一定角度、水平翻折、垂直翻折等操作进行扩充. 最终，本研究从该数据集获得了47 723张图像，并使其达到相对平衡的状态. 实验过程中，本研究从每种疾病中随机抽取100张图像作为测试样本，其余图像作为数据库样本，训练样本则是每次从数据库样本中随机抽取10 000张.

ISIC 2018：该数据集是一个皮肤镜像数据集，包含黑色素瘤、黑素细胞痣、基底细胞癌等7种疾病，每张图像的尺寸为600×450×3，共计10 208张图像. 由于数据集中各种疾病的数目相差过大，因此本研究剔除了数目极少的2种疾病，并将剩余的5种疾病通过旋转一定角度、水平翻折、垂直翻折等操作进行扩充. 最终，本研究从该数据集获得了16 345张图像，并使其达到相对平衡的状态. 实验过程中，本研究从每种疾病中随机抽取100张图像作为测试样本，其余图像作为数据库样本，训练样本则是每次从数据库样本中随机抽取5 000张.

3.2 实验评估指标

本节主要介绍实验中使用的评估标准.

平均精度(Mean Average Precision，MAP)被广泛用于衡量汉明距离排序的准确性. 为了得到MAP的值，本研究首先引入平均精度(Average Precision，AP). 对于第i个查询图像，AP(i)的定义如下：

$$\mathrm{AP}(i)=\frac{1}{M}\sum_{n=1}^{N}\frac{R_n}{n}\times rel_n \tag{25}$$

式中：M是从数据库中检索到相关图像的总数；N代表数据库中的总样本数；R_n代表前n个返回图像中与查询图像相关的图像总数. 如果返回图像中第n张图像与查询图像相关，则$rel_n=1$；否则，$rel_n=0$. MAP则是所有AP的平均值，即$\mathrm{MAP}=\mathrm{mean}[\mathrm{AP}(i)]$.

Precision@K代表返回样本中前K个与查询样本相似的平均准确率.

$$\mathrm{Precision}@K=\frac{\sum_{i=1}^{K}rel_i}{K} \tag{26}$$

式中：如果返回图像中第i张图像与查询图像相关，则$rel_i=1$；否则，$rel_i=0$.

3.3 实验设置

本研究针对ViTH模型的所有实验都是基于Windows系统的服务器完成的(其中CPU配置为：Intel

(R) Core(TM) i9-10940X CPU @ 3.30GHz，GPU 配置为：NVIDIA GeForce RTX 3090). 经过对超参数的实验分析，本研究设置 $\alpha=100$，$\beta=500$，$\gamma=10$. 另外 batch size 为 64，迭代次数为 50，总 epoch 为 20，初始学习率为 0.000 1.

3.4　实验结果与分析

3.4.1　对比实验

本节将所提出的算法(ViTH)与其他几种经典算法进行对比，包括 DSH[11]，DPSH[12]，IDHN[37]，DBDH[38]，VTS[23]，SADH[39]. 简单介绍如下：

DSH[11]：该算法设计了一个 CNN 架构，将成对的图像(相似/不相似)作为训练输入，并鼓励每个图像的输出接近离散值(例如+1/−1). 另外，设计损失函数并通过对输入图像的监督信息进行编码，同时对实值输出进行正则化以逼近所需的离散值，从而最大限度地提高输出空间的可辨别性.

DPSH[12]：一种深度成对监督哈希，用于对具有成对标签的样本点联合执行特征学习和哈希码学习.

IDHN[37]：一种将语义标签进行归一化并计算其成对量化相似度的哈希方法. 该方法将成对相似度分为硬相似度和软相似度两种情况，并对这两种情况使用不同的损失函数.

DBDH[38]：一种深度平衡离散哈希方法，该方法没有使用传统的连续松弛策略，从而减少了连续松弛带来的量化误差. 在损失函数中，离散值是通过成对损失和平衡控制项来计算的. 学习到的二进制哈希码同时保持相似关系和标签一致性. 在保持成对相似性的同时，该方法保持哈希码的平衡以提高检索性能.

VTS[23]：一种利用预训练的视觉 Transformer 模型来进行图像检索的方法. 该方法以现有的经典哈希算法为基础，利用 ViT 模型作为通用的特征提取模块，并将 ViT 中的 MLP 替换为各经典算法的哈希模块.

SADH[39]：提出一种具有自监督非对称语义挖掘和边距可扩展约束的新型深度哈希方法. 该方法实现了一个自监督网络，在语义特征字典和语义代码字典中充分保留给定数据集语义的语义信息，高效准确地引导特征学习网络使用非对称学习来保留多标签语义信息策略，并通过进一步利用语义词典以及边距可缩放约束来生成哈希码.

表 1 和表 2 分别是目前经典的哈希算法在 ChestX-ray14 和 ISIC 2018 数据集上的检索结果. 另外为了节省时间，本研究中对比算法(DSH，DPSH，IDHN，DBDH)的实验使用预训练好的 AlexNet[40] 网络作为特征提取模块. 从表 1(图 4)和表 2(图 5)可以看出，ViTH 整体上要优于其他经典算法. 图 6 和图 7 是 ViTH 分别在 ChestX-ray14 和 ISIC 2018 数据集上随机检索到的前 10 张图像展示. 其中，蓝色框代表检索到的图像与查询图像至少共享一个类别，红色框代表检索到的图像与查询图像没有共享任一类别.

表 1　不同哈希算法在 ChestX-ray14 上的检索结果

算法	不同码长下的 Precision@10			
	8bit	12bit	24bit	36bit
DSH	0.434	0.444	0.411	0.441
DPSH	0.449	0.442	0.396	0.402
IDHN	0.407	0.413	0.453	0.481
DBDH	0.374	0.400	0.380	0.409
VTS	0.405	0.416	0.468	0.516
SADH	0.422	0.484	0.509	0.519
ViTH	**0.583**	**0.634**	**0.731**	**0.758**

表2 不同哈希算法在ISIC 2018上的检索结果

算法	不同码长下的MAP			
	8bit	12bit	24bit	36bit
DSH	0.549	0.625	0.662	0.687
DPSH	0.630	0.651	0.713	0.706
IDHN	0.569	0.623	0.641	0.627
DBDH	0.597	0.607	0.614	0.550
VTS	0.415	0.639	0.714	0.715
SADH	0.574	0.642	0.686	0.696
ViTH	**0.740**	**0.754**	**0.746**	**0.719**

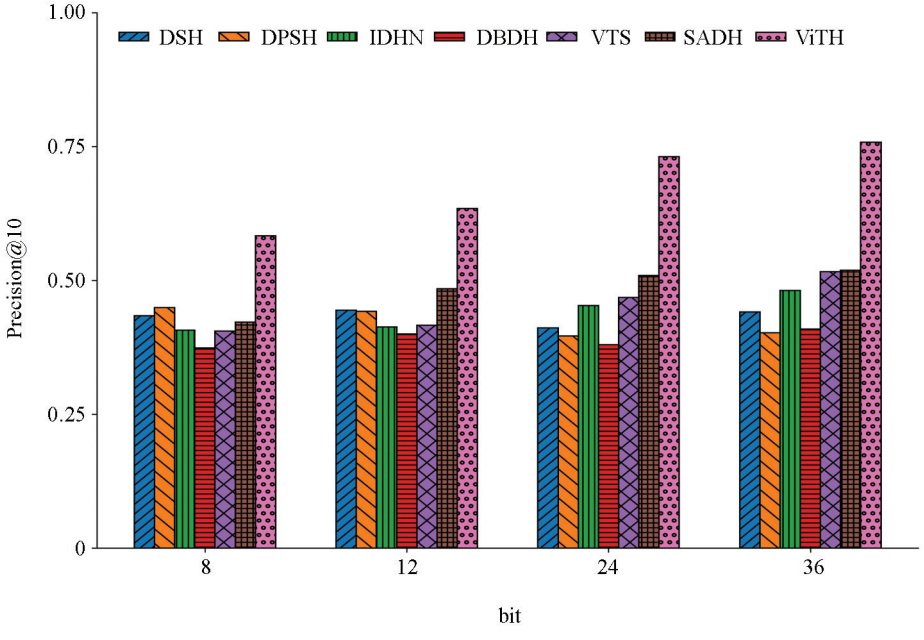

图4 不同哈希算法在ChestX-ray14数据集上的对比实验

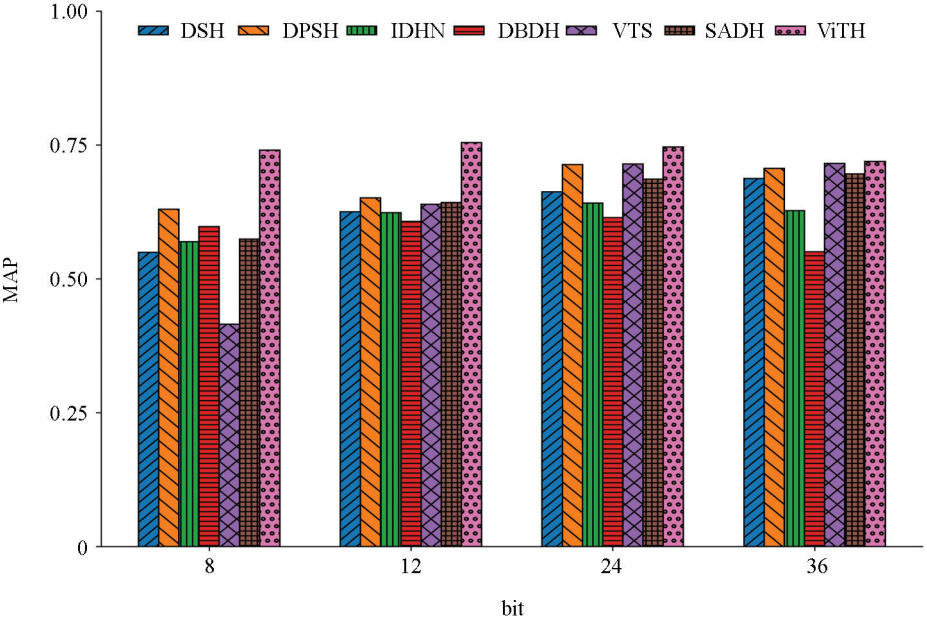

图5 不同哈希算法在ISIC 2018数据集上的对比实验

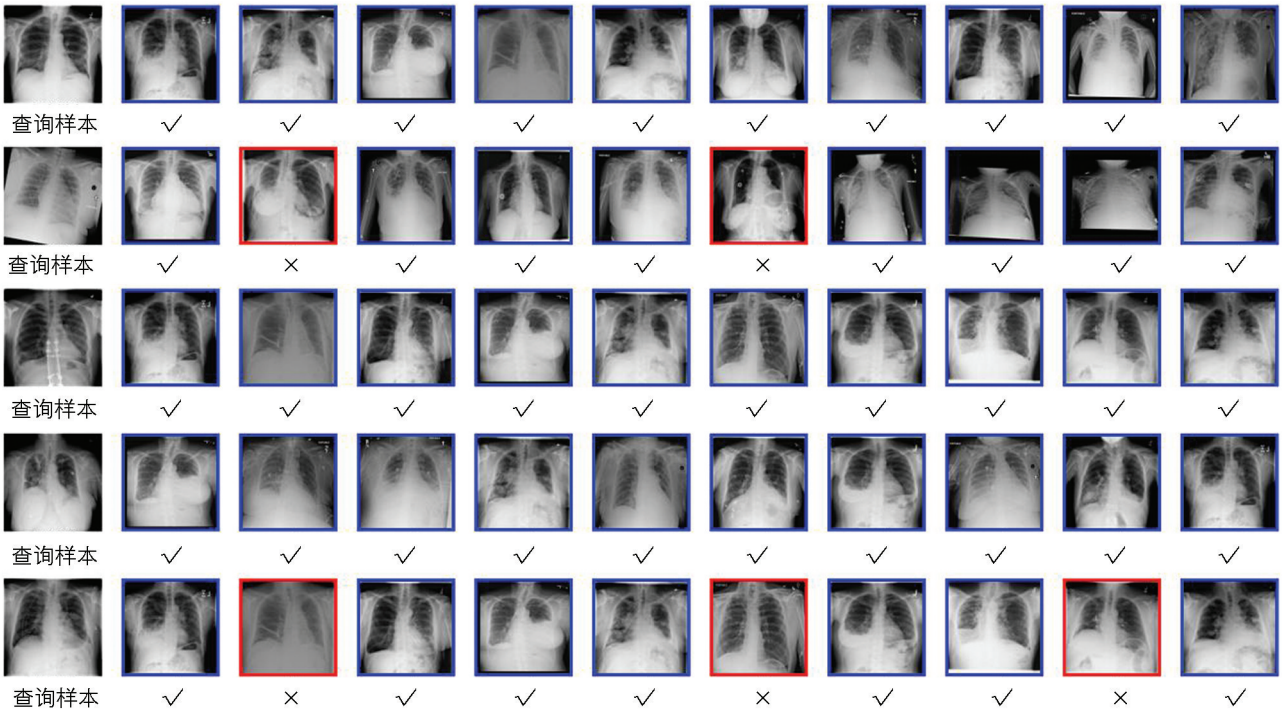

图 6　ViTH 在 36 位哈希编码下随机检索到的前 10 张图像(从左到右，返回的图像按汉明距离降序排列)

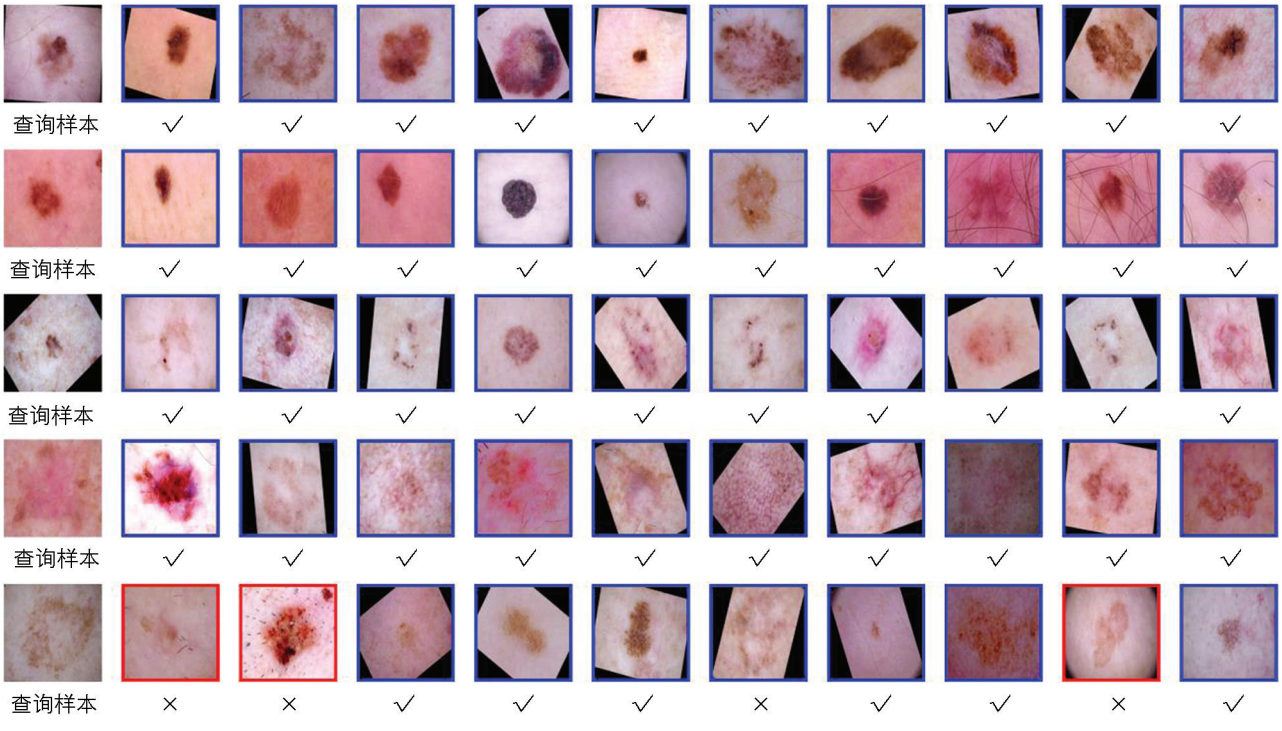

图 7　ViTH 在 12 位哈希编码下随机检索到的前 10 张图像(从左到右，返回的图像按汉明距离降序排列)

3.4.2　超参数分析

深度学习中，不同数值的超参数对模型的性能影响至关重要[41]. 本节主要针对 α，β 和 γ 这 3 个超参数，使用 Precision@10 在 ChestX-ray14 上对 36 位哈希编码进行超参数分析. 其中，α 代表量化损失 L_Q 的权重系数，β 代表平衡损失 L_B 的权重系数，γ 代表分类损失 L_C 的权重系数. 本研究分别设置 $\alpha\in\{100, 200, 300, 400, 500\}$，$\beta\in\{0.5, 5, 50, 500, 5\,000\}$，$\gamma\in\{0.1, 1, 10, 100, 1\,000\}$进行实验分析. 除超参

数的取值外，其他参数的取值仍使用3.3中的实验设置. 图8—图10分别是α，β和γ在ChestX-ray14上不同取值下的结果. 由图8—图10可知，当$\alpha=100$，$\beta=500$，$\gamma=10$时，检索性能最优.

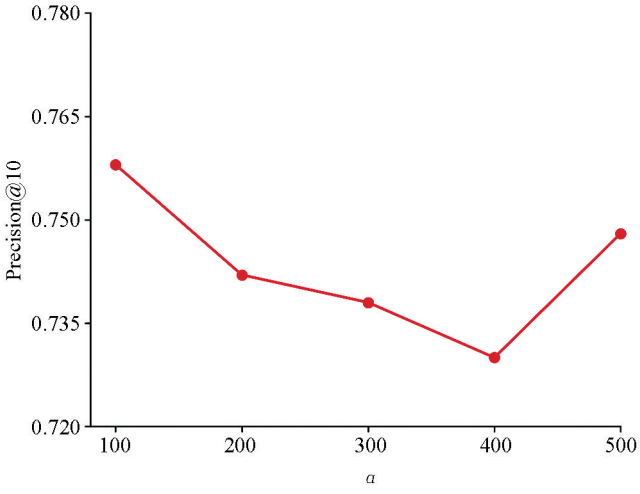

图8 36位哈希编码下超参数α不同时的MAP

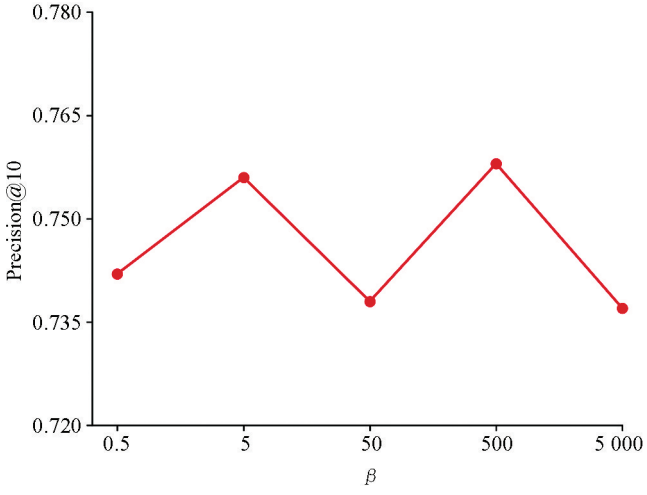

图9 36位哈希编码下超参数β不同时的MAP

4 结论与展望

本研究提出一种面向医学图像检索的视觉Transformer哈希(ViTH)算法. 在Transformer编码器的前后端分别加入PMT模块来进一步增强模型的非线性. 鉴于医学图像之间差异性小且难以分辨的问题，本研究在Transformer编码器内部引入MHSPA模块，该模块不仅可以提取图像的全局上下文特征，而且可以提取多尺度的局部上下文特征，并将不同尺度的特征进行融合. 在损失函数方面，本研究不仅考虑了传统的成对损失、量化损失，还添加了平衡损失和分类损失以对哈希码的映射进一步约束. 本研究在ChestX-ray14和ISIC 2018两个医学图像数据集上与其他多个先进的哈希算法进行实验比较，证明了本研究算法在检索性能方面具有较好的优越性，对关键超参数的变化具有鲁棒性.

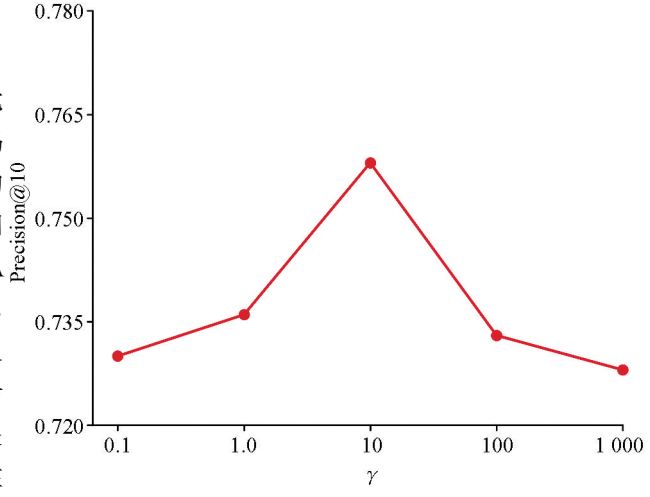

图10 36位哈希编码下超参数γ不同时的MAP

另外，本研究算法主要应用在两个领域：① 医学图像检索与快速诊断. ViTH算法可以支持医生和研究人员快速获取与特定病例相关的图像. 这有助于提高诊断效率，尤其是在紧急情况下迅速获取相关图像进行诊断. ② 医学图像相似性分析. ViTH算法可以量化医学图像之间的相似性，从而帮助医学研究人员进行更准确的图像分析.

最后，本研究虽然在ChestX-ray14和ISIC 2018上取得了良好的实验效果，但仍然存在一些局限性：① 现实中各疾病的发病概率是不同的，从而导致医学图像数据集中各类别之间存在不均衡现象，因此模型在检索过程中可能更倾向于占比较大的类别. ② 本研究仅关注单模态医学图像，对多模态数据并不适用. 这些局限性也将是本团队未来工作中的重要研究方向.

参考文献：

[1] RAHMAN M M, BHATTACHARYA P, DESAI B C. A Framework for Medical Image Retrieval Using Machine Learning and Statistical Similarity Matching Techniques with Relevance Feedback [J]. IEEE Transactions on Information Technology in Biomedicine: A Publication of the IEEE Engineering in Medicine and Biology Society, 2007, 11(1):

58-69.

[2] OWAIS M, ARSALAN M, CHOI J, et al. Effective Diagnosis and Treatment through Content-Based Medical Image Retrieval (CBMIR) by Using Artificial Intelligence [J]. Journal of Clinical Medicine, 2019, 8(4): 462.

[3] UNAR S, WANG X Y, ZHANG C, et al. Detected Text-Based Image Retrieval Approach for Textual Images [J]. IET Image Processing, 2019, 13(3): 515-521.

[4] QAYYUM A, ANWAR S M, AWAIS M, et al. Medical Image Retrieval Using Deep Convolutional Neural Network [J]. Neurocomputing, 2017, 266: 8-20.

[5] 曾宪华，袁知洪，王国胤，等. 基于多特征多核哈希学习的大规模图像检索 [J]. 中国科学：信息科学，2017，47(8)：1109-1126.

[6] 刘颖，程美，王富平，等. 深度哈希图像检索方法综述 [J]. 中国图象图形学报，2020，25(7)：1296-1317.

[7] SINGH A, GUPTA S. Learning to Hash: A Comprehensive Survey of Deep Learning-Based Hashing Methods [J]. Knowledge and Information Systems, 2022, 64(10): 2565-2597.

[8] ANDONI A, INDYK P. Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions [J]. Communications of the ACM, 2008, 51(1): 117-122.

[9] KONG W H, LI W J. Isotropic Hashing [C] //Proceedings of the 25th International Conference on Neural Information Processing Systems, Nevada, USA, 2012: 1646-1654.

[10] 陈昌红，彭腾飞，干宗良. 基于深度哈希算法的极光图像分类与检索方法 [J]. 电子与信息学报，2020，42(12)：3029-3036.

[11] LIU H M, WANG R P, SHAN S G, et al. Deep Supervised Hashing for Fast Image Retrieval [J]. International Journal of Computer Vision, 2019, 127(9): 1217-1234.

[12] LI W J, WANG S, KANG W C. Feature Learning Based Deep Supervised Hashing with Pairwise Labels [C] //Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, USA, 2016: 1711-1717.

[13] CAO Y, LONG M S, LIU B, et al. Deep Cauchy Hashing for Hamming Space Retrieval [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Utah, USA, 2018:1229-1237.

[14] WANG X F, SHI Y, KITANI K M. Deep Supervised Hashing with Triplet Labels [C] //Asian Conference on Computer Vision, Taipei, China, 2016: 70-84.

[15] LIU B, CAO Y, LONG M S, et al. Deep Triplet Quantization [C] //Proceedings of the 26th ACM international conference on Multimedia, Seoul, Republic of Korea, 2018: 755-763.

[16] FANG J S, FU H Z, LIU J. Deep Triplet Hashing Network for Case-Based Medical Image Retrieval [J]. Medical Image Analysis, 2021, 69: 101981.

[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you Need [C] //Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017: 5998-6008.

[18] CARION N, MASSA F, SYNNAEVE G, et al. End-to-End Object Detection with Transformers [C] //European Conference on Computer Vision, Glasgow, UK, 2020: 213-229.

[19] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale [C] //The Ninth International Conference on Learning Representations, Vienna, Austria, 2021: 1-21.

[20] HAN K, XIAO A, WU E H, et al. Transformer in Transformer [J]. Advances in Neural Information Processing Systems, 2021, 34: 15908-15919.

[21] WANG W H, XIE E Z, LI X, et al. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions [C] //Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, IEEE: 568-578.

[22] HE S T, LUO H, WANG P C, et al. TransReID: Transformer-Based Object Re-Identification [C] //Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, IEEE: 15013-15022.

[23] DUBEY S R, SINGH S K, CHU W T. Vision Transformer Hashing for Image Retrieval [C] //Proceedings-IEEE International Conference on Multimedia and Expo, Taipei, China, 2022 : 1-6.

[24] CHEN Y B, ZHANG S, LIU F X, et al. TransHash: Transformer-Based Hamming Hashing for Efficient Image Retrieval [C] //Proceedings of the 2022 International Conference on Multimedia Retrieval, Newark, NJ, USA, 2022: 127-136.

[25] LI T, ZHANG Z, PEI L S, et al. Hash Former: Vision Transformer Based Deep Hashing for Image Retrieval [J]. IEEE Signal Processing Letters, 2022, 29: 827-831.

[26] GONG Q K, WANG L D, LAI H J, et al. ViT2Hash: Unsupervised Information-Preserving Hashing [EB/OL]. 2022: arXiv: 05541. http://arxiv.org/abs/2201.05541.

[27] ZHANG C L, WU J X. Improving CNN Linear Layers with Power Mean Non-Linearity [J]. Pattern Recognition, 2019, 89: 12-21.

[28] HE X Z, TAN E L, BI H W, et al. Fully Transformer Network for Skin Lesion Analysis [J]. Medical Image Analysis, 2022, 77: 102357.

[29] LU H M, ZHANG M, XU X, et al. Deep Fuzzy Hashing Network for Efficient Image Retrieval [J]. IEEE Transactions on Fuzzy Systems, 2021, 29(1): 166-176.

[30] WANG X Q, LAN R S, WANG H D, et al. Fine-Grained Correlation Analysis for Medical Image Retrieval [J]. Computers and Electrical Engineering, 2021, 90: 106992.

[31] XU L M, ZENG X H, ZHENG B C, et al. Multi-Manifold Deep Discriminative Cross-Modal Hashing for Medical Image Retrieval [J]. IEEE Transactions on Image Processing, 2022, 31: 3371-3385.

[32] YANG E K, LIU M X, YAO D R, et al. Deep Bayesian Hashing with Center Prior for Multi-Modal Neuroimage Retrieval [J]. IEEE Transactions on Medical Imaging, 2021, 40(2): 503-513.

[33] BA J L, KIROS J R, HINTON G E. Layer Normalization [EB/OL]. arXiv: 1607.06450. http://arxiv.org/abs/1607.06450.

[34] HE K M, ZHANG X Y, REN S Q, et al. Deep Residual Learning for Image Recognition [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770-778.

[35] 邹细涛. 多标记跨模态语义哈希图文检索研究 [D]. 重庆: 西南大学, 2022.

[36] JIANG Q Y, LI W J. Asymmetric Deep Supervised Hashing [C] //Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, USA, 2018: 3342-3349.

[37] ZHANG Z, ZOU Q, LIN Y W, et al. Improved Deep Hashing with Soft Pairwise Similarity for Multi-Label Image Retrieval [J]. IEEE Transactions on Multimedia, 2020, 22(2): 540-553.

[38] ZHENG X T, ZHANG Y C, LU X Q. Deep Balanced Discrete Hashing for Image Retrieval [J]. Neurocomputing, 2020, 403: 224-236.

[39] YU Z Y, WU S, DOU Z H, et al. Deep Hashing with Self-Supervised Asymmetric Semantic Excavation and Margin-Scalable Constraint [J]. Neurocomputing, 2022, 483: 87-104.

[40] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Image Net Classification with Deep Convolutional Neural Networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[41] 曾超, 白琮, 马青, 等. 基于对抗投影学习的跨模态哈希检索 [J]. 计算机辅助设计与图形学学报, 2021, 33(6): 904-912.

责任编辑 包颖
崔玉洁