面向患者跨模态检索的图增强哈希网络模型

顾一凡[1,2)]，杨雪冰[1)]，朱承璋[3)]，张晨阳[1)]，张文生[4)*]

[1)] (中国科学院自动化研究所多模态人工智能系统全国重点实验室　北京　100190)
[2)] (中国科学院大学人工智能学院　北京　100049)
[3)] (中南大学文学与新闻传播学院　长沙　410083)
[4)] (广州大学计算机科学与网络工程学院　广州　510006)
(wensheng.zhang@gzhu.edu.cn)

摘　要：通过对图像、文本等多种模态的患者数据进行检索，医生能够更全面、深入地分析患者的病程状态，做出更精准的临床诊断和治疗决策。随着医疗数据规模的迅猛增长、低存储需求与高效率检索的挑战，哈希方法逐渐在患者数据跨模态检索领域成为流行范式。针对现有跨模态哈希方法在利用患者诊断类别的多标签信息以及实现模态间协同建模方面存在的不足，提出引入跨模态注意力的图增强哈希网络模型。首先设计图信息增强的标签编码器，利用诊断类别标签的共现关系构建标签图，并通过图卷积网络有效地提取患者的多标签嵌入表征；为了缓解多标签矩阵的稀疏性问题并丰富融合表征的语义信息，引入图像、文本与多标签嵌入表征之间的跨模态注意力融合机制；最后将诊断类别的多标签共现关系与语义信息有效地融入患者图像、文本数据的哈希编码中，进一步提升哈希编码的判别能力，并设计了模态内与模态间关联互补的相似性损失函数。在胸部 X 光成像与诊断报告数据集 MIMIC-CXR 上进行实验的结果表明，与 3 种经典跨模态哈希方法和 4 种先进的深度跨模态哈希方法相比，所提模型在多个编码位数下均展现出了性能优势。在 X 光成像检索诊断报告任务上，所提模型在哈希编码长度为 16、32、64 和 128 位时，平均精度均值相比次优方法分别提高了 2.98%、1.21%、0.63%和 0.53%；在诊断报告检索 X 光成像任务上，所提模型在哈希编码为上述 4 个长度时，平均精度均值相比次优方法分别提高了 0.91%、0.75%、1.03%和 0.57%。

关键词：患者跨模态检索；深度哈希；图神经网络；患者表征学习
中图法分类号：TP391.41　　**DOI：**10.3724/SP.J.1089.2024-00351

Graph Enhanced Hashing Networks for Cross-Modal Patient Retrieval

Gu Yifan[1,2)], Yang Xuebing[1)], Zhu Chengzhang[3)], Zhang Chenyang[1)], and Zhang Wensheng[4)*]

[1)] (*State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing*　100190)
[2)] (*School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing*　100049)
[3)] (*School of Literature and Journalism, Central South University, Changsha*　410083)
[4)] (*School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou*　510006)

Abstract: By retrieving patient data from multiple modalities such as images and text, physicians can analyze the patient's disease progression more comprehensively and deeply, thereby making more precise clinical diagnoses and treatments. Given the rapid growth of medical data and the challenges of low storage requirements and efficient retrieval, hashing methods have gradually become a popular paradigm in the field of cross-modal patient retrieval. However, existing cross-modal hashing methods still have deficiencies in utilizing the multi-label infor-

收稿日期：2024-07-03；修回日期：2024-10-22。**基金项目：**国家自然科学基金（62206293，62376287）；湖南省机器视觉与医学图像处理国际科技创新合作基地项目（2021CB1013）；湖南省自然科学基金（2022JJ30762，2023JJ70016）。**顾一凡**（1994—），男，博士生，主要研究方向为患者表征学习、机器学习、图神经网络；**杨雪冰**（1991—），男，博士，副研究员，硕士生导师，主要研究方向为机器学习、大数据知识挖掘；**朱承璋**（1978—），女，博士，副教授，博士生导师，CCF 会员，主要研究方向为机器学习、医学图像处理；**张晨阳**（1995—），男，博士，助理研究员，主要研究方向为机器学习、人工智能、计算机视觉；**张文生**（1965—），男，博士，研究员，博士生导师，CCF 会员，论文通信作者，主要研究方向为人工智能、机器学习、大数据知识挖掘。

mation of patient diagnostic categories and achieving collaborative modeling between modalities. Therefore, we propose a method named graph enhanced hashing with cross-modal attention (GEHCA), which is used for cross-modal patient retrieval. Firstly, a graph information enhanced label encoder is designed, which constructs a label graph using the co-occurrence relationships of diagnostic category labels and effectively extracts multi-label embedding representations of patients through a graph convolutional network. Secondly, to alleviate the sparsity of the multi-label matrix and enrich the semantic information of the fused representations, the model introduces a cross-modal attention fusion mechanism between image, text, and multi-label embeddings. Finally, to effectively integrate the multi-label co-occurrence relationships and semantic information of diagnostic categories into the hash codes of patient image and text data, and further enhance the discriminative power of the hash codes, a similarity loss function that utilizes intra-modal and inter-modal associations is designed. Experiments on the MIMIC-CXR dataset of chest X-ray images and diagnostic reports demonstrate that, compared with three classical cross-modal hashing methods and four advanced deep cross-modal hashing methods, GEHCA exhibits performance advantages under various hash code lengths. On the task of retrieving diagnostic reports using X-ray images, the proposed model achieves mean average precision improvements of 2.98%, 1.21%, 0.63%, and 0.53% compared to the second-best method, at hash code lengths of 16, 32, 64, and 128 bits, respectively. On the task of retrieving X-ray images using diagnostic reports, the proposed model achieves mean average precision improvements of 0.91%, 0.75%, 1.03%, and 0.57% compared to the second-best method at the above 4 hash code lengths.

Key words: cross-modal patient retrieval; deep hashing; graph neural networks; patient representation learning

随着医学检测手段的多样化和医学信息化技术的进步，多种模态的患者数据（包括诊断报告、X 光成像、CT 图像、磁共振成像、电子病历等）积累量急剧增加，使得面向患者数据的跨模态检索成为学术界和工业界的研究热点[1-2]。然而，不同模态的数据往往具有不同的分布和特征空间，如何高效且有效地缩小这些大规模异构数据之间的语义差距，并精确地计算不同模态数据之间的语义相似性，对于患者跨模态检索依然是一个重大挑战[3]。现有的跨模态检索方法，如主题模型[4]、子空间学习[5]、深度学习[6]等，通常将不同模态的原始特征映射到一个共同的实值子空间中，计算各实例不同模态间的语义相似性。但是，随着多模态患者数据规模的爆炸式扩增，基于实值的跨模态检索方法越来越难以满足低存储、高效率的检索需求。因此，跨模态哈希方法逐渐崭露头角，成为跨模态检索领域的一种流行范式。

跨模态哈希方法旨在将多模态实例的高维原始特征，借助哈希函数映射至低维的汉明空间中，以二值哈希编码的形式进行表征；随后，跨模态成对实例间的相关性通过简单的异或操作即可高效计算，提升了检索效率，降低了存储成本。跨模态哈希方法分为无监督与监督式 2 类：无监督方法不依赖明确的语义标签，通过保持原始多模态数据的结构信息学习哈希函数，以此捕捉潜在的语义信息[7]；监督式方法则在进行多模态实例的二值哈希编码时，利用类别标签的监督信息以确保跨模态间的语义相似性得以保持[8]。

近年来，深度学习模型在多个领域中表现出强大的特征提取能力，为跨模态哈希研究引入了新的思路[9]。已有工作将深度跨模态哈希（deep cross-modal hashing，DCMH）方法应用于患者医疗数据的跨模态检索任务[2,10-11]，它们主要聚焦于医学影像与诊断报告之间的跨模态检索，并利用医生标注的诊断类别作为语义多标签[12]，依据不同模态的 2 个患者实例是否共享相同的诊断来确定其成对相似性，进而监督跨模态哈希的训练过程；然而，这种多标签的利用方式未能有效地挖掘诊断类别的语义信息。为改进这些不足，部分方法设计了标签编码器，通过多层感知机（multi-layer perceptron，MLP）从多标签向量中提取特征，以指导图像、文本等模态的哈希编码过程[13-14]。但是，该方式存在 2 点局限：（1）未能充分利用多标签的共现关系。对于患者而言，不同诊断标签的共现关系反映了不同的患者病情。例如，如果一名患者的 X 光成像显示他有肺水肿和心脏肥大症状，而另一名患者被诊断有肺部阴影和肺水肿，尽管二者都表现出肺水肿症状，但前者更可能是心脏疾病，后者更可能是

肺部疾病[15]。因此，诊断的共现关系具有重要的医学价值，忽视这些标签间的关联会削弱哈希编码的相似性判别能力[6,16]。（2）医学跨模态数据的多标签矩阵普遍存在稀疏性问题[3,12]，使多标签提供的语义信息量受到局限，导致在协同建模过程中难以有效地缩减各模态数据之间的语义鸿沟。

针对上述局限性，本文提出引入跨模态注意力的图增强哈希网络（graph enhanced hashing with cross-modal attention，GEHCA）模型，用于患者医学影像与诊断报告的跨模态检索任务。该模型由图像编码器、文本编码器和标签编码器 3 个组件构成。图像编码器利用卷积神经网络（convolutional neural networks，CNN）提取医学影像特征；文本编码器先通过 MLP 处理诊断报告并提取特征，再对医学影像与诊断报告的嵌入表征进行二值化处理，生成哈希编码；为了充分利用患者诊断多标签间的共现关系，标签编码器依据诊断类别标签的点互信息（pointwise mutual information，PMI）构建标签图，并采用图卷积网络（graph convolutional networks，GCN）进行特征提取，获得每个患者实例的多标签嵌入表征，实现标签编码器的图结构信息增强。为了缓解多标签矩阵稀疏导致的语义信息不足，本文引入跨模态注意力机制，融合多标签嵌入表征与图像、文本模态的嵌入表征，丰富多标签表征的语义信息量。最后设计一个综合考虑模态内与模态间关联互补的相似性损失函数，对 3 个编码器进行联合端到端训练，将多标签的共现关联和语义信息在跨模态哈希过程中传递，提升哈希编码的判别力。在胸部 X 光成像与诊断报告数据集 MIMIC-CXR[12]上的实验结果表明，与多种跨模态哈希方法相比，GEHCA 模型在不同哈希编码位数下，在 X 光成像检索诊断报告和诊断报告检索 X 光成像这 2 项跨模态检索任务上，均展现出性能优势。

1 相关工作

本文提出的 GEHCA 模型是一种应用于患者医疗数据跨模态检索任务的哈希方法。本节首先概述跨模态哈希检索的相关工作，涵盖基于传统机器学习的方法和基于深度学习的方法，然后简要介绍跨模态哈希检索在医学领域的应用情况。

1.1 基于传统机器学习的跨模态哈希方法

跨模态哈希的核心目标是将不同模态的数据转换为紧凑的二值编码，并将其映射到一个公共的汉明空间中，利用快速的相似性计算方式提高大规模跨模态数据的检索速度，并降低存储成本[7]。根据在哈希过程中是否需要语义标签的监督，传统跨模态哈希方法分为无监督和监督式 2 类。

无监督跨模态哈希无需依赖明确的语义标签，而是通过保持原始多模态数据的内在结构信息来学习哈希函数[17]。其中，典型相关分析（canonical correlation analysis，CCA）[18]是一种标志性方法，其利用隐狄利克雷分布提取文本特征，采用尺度不变特征变换提取图像特征，并通过分析这些特征间的相关性实现 2 个模态数据的语义匹配。Ding 等[19]提出协同矩阵分解哈希（collective matrix factorization hashing，CMFH），利用潜在因子模型对同一实例的不同模态数据进行协同矩阵分解，学习得到统一的哈希编码，并通过整合多视角信息源来显著地提高检索精度；Wang 等[20]提出语义主题多模态哈希（semantic topic multi-modal hashing，STMH），分别采用主题模型和矩阵分解技术提取文本的语义主题和图像的语义概念，并基于特征间的关联性将它们映射到公共子空间中，通过判断特定主题或概念是否存在于文本或图像中直接生成哈希编码。尽管这类方法能够有效地利用数据自身的结构和分布特性，但因为缺少语义标签信息的监督，无监督跨模态哈希方法的检索性能仍有待提升[13]。

监督式跨模态哈希方法借助语义标签的监督作用，在保持多模态实例间相似性的前提下实现哈希编码的学习[21]。Zhang 等[22]提出一种语义相关性最大化策略（semantic correlation maximization，SCM），利用语义标签构建相似度矩阵，将语义信息整合到大规模数据哈希编码的学习过程中；Lin 等[23]提出语义保持哈希（semantics-preserving hashing，SePH），以训练数据的语义相似度作为监督信息，并将其转换为概率分布，通过在汉明空间内最小化 Kullback-Leibler 散度，促使跨模态哈希编码逼近此概率分布，之后采用基于抽样策略的核逻辑回归方法，将各模态数据特征映射为哈希编码；Meng 等[24]提出非对称监督一致特定哈希（asymmetric supervised consistency and specific hashing，ASCSH）方法，将映射矩阵分解为一致性和模态特定性 2 部分，有效地增强了哈希编码的判别能力。这些监督式跨模态哈希方法虽然提升了检索性能，但由于其二值编码的获取过程通常与特征提取过程分离，不可避免地导致哈希编码信息损失。

1.2 基于深度学习的跨模态哈希方法

鉴于深度学习模型在特征提取与表示方面的卓越能力，大量研究开始将深度学习融入跨模态哈

希方法中。其中最具标志性的方法是 DCMH[9]，其利用语义标签判定实例间的成对相似性，并以其作为监督信息，将深度特征提取与哈希编码生成整合至统一框架，实现端到端的哈希编码训练，并通过优化哈希编码与各模态实值特征间的量化正则项降低二值化过程的信息损失。这一核心思路也为后续方法所沿袭，催生出诸如基于离散潜在因子模型的跨模态哈希（discrete latent factor model-based cross-modal hashing，DLFH）[25]、一致性保持对抗哈希（consistency-preserving adversarial hashing，CPAH）[26]、深度多尺度融合哈希（deep multi-scale fusion hashing，DMFH）[27]等一系列具有影响力的研究工作。然而，这些方法仅将跨模态实例的多标签视为确定成对相似性的基准，没有充分发掘多标签本身蕴含的丰富语义信息。针对这一局限，部分研究开始将多标签也视作一种模态，通过提取多标签特征来引导其他模态数据的哈希编码过程。Li 等[14]提出自监督对抗哈希（self-supervised adversarial hashing，SSAH），利用自监督语义网络获取多标签标注的高级语义信息，以此指导特征学习过程，并在公共语义空间与汉明空间中维持模态间的关联性；Duan 等[13]提出多标签语义监督图注意力哈希（multi-label semantic supervised graph attention hashing，MS^{2}GAH），通过 MLP 从语义多标签中提取特征，为采用图注意力网络的图像与文本模态的哈希编码器建立细粒度语义关联。尽管上述方法利用多标签语义信息缩小了模态间语义鸿沟，但仍缺乏对标签间共现关联的有效探索，且难以应对多标签矩阵稀疏带来的语义信息量不足的问题。

1.3　患者数据的跨模态哈希检索

随着医学数据规模高速增长，低存储与高效率逐渐成为患者跨模态检索的核心要求，促使更多研究者将哈希方法应用于该任务。Yang 等[28]构建一个深度贝叶斯哈希框架，能够将跨模态数据映射至公共汉明空间，并从不平衡的跨模态神经影像中学习哈希编码，以服务于检索任务；Xu 等[10]提出多流形深度判别跨模态哈希（multi-manifold deep discriminative cross-modal hashing，MDDCH）方法，利用多模态流形相似度保持实例间的相关性，并结合高斯二元受限玻尔兹曼机直接生成哈希编码，实现患者医学影像的跨模态检索；Shi 等[11]提出深度监督信息瓶颈哈希（deep supervised information bottleneck hashing，DSIBH），将深层确定信息瓶颈机制扩展至跨模态场景，增强哈希编码判别力，提升其在 X 光成像与诊断报告的跨模态检索任务中的表现；Zhang 等[2]提出深度医疗跨模态注意力哈希（deep medical cross-modal attention hashing，DMCAH），运用注意力机制融合各模态数据的粗、细粒度特征，并将特征映射到公共空间来获取具有判别力的哈希编码。上述方法在患者数据跨模态检索任务中展现了较好的性能，但如何利用诊断类别标签间的共现信息来进一步强化各模态数据间的语义关联，仍是一个值得深入探究的问题。

2　GEHCA 模型

GEHCA 模型框架如图 1 所示。下面首先阐述本文中使用的数学符号并说明问题定义，然后分别介绍图像编码器、文本编码器、标签编码器和跨模态注意力模块，最后介绍整体目标函数和模型的学习策略。

2.1　符号和问题定义

本文中，图像数据表示为 $\boldsymbol{X}=\{\boldsymbol{x}_i\}_{i=1}^{N}\in\mathbb{R}^{N\times M_{\mathrm{X}}}$，文本数据则表示为 $\boldsymbol{T}=\{\boldsymbol{t}_i\}_{i=1}^{N}\in\mathbb{R}^{N\times M_{\mathrm{T}}}$，其中，$N$ 表示训练集中的患者实例总数，M_{X} 和 M_{T} 分别表示图像和文本数据的原始特征维度。训练集的多标签矩阵以 $\boldsymbol{L}=\{\boldsymbol{l}_i\}_{i=1}^{N}\in\{0,1\}^{N\times C}$ 表示，其中，C 表示类别标签总数。患者实例 i 包含 $\boldsymbol{x}_i$、$\boldsymbol{t}_i$ 和 $\boldsymbol{l}_i$ 这 3 种模态的数据。相似性集合用 S 表示，$s_{ij}\in S$ 为实例 i 和 j 之间的成对相似性。不失一般性，本文采用跨模态哈希方法中最常见的相似性定义方式[8]，即实例 i 和 j 如果共享至少一个类别标签，则 $s_{ij}=1$；否则，$s_{ij}=0$。明确了训练数据及实例间的成对相似性后，核心问题就是学习能保留语义相似性的各模态数据的哈希函数，得到哈希编码 $\boldsymbol{B}^{*}=\{\boldsymbol{b}_i^{*}\}_{i=1}^{N}\in\{-1,1\}^{N\times K}$；其中，$*=\{\mathrm{X},\mathrm{T}\}$ 表示图像和文本 2 种模态，$\boldsymbol{b}_i^{*}$ 表示实例 i 在对应模态下的 K 位哈希编码。s_{ij} 的条件概率可表示为

$$p\left(s_{ij}\middle|\boldsymbol{b}_i^{*},\boldsymbol{b}_j^{*}\right)=\begin{cases}\varphi\left(\Gamma_{ij}\right), & s_{ij}=1\\ 1-\varphi\left(\Gamma_{ij}\right), & s_{ij}=0\end{cases}=\varphi\left(\Gamma_{ij}\right)^{s_{ij}}\left(1-\varphi\left(\Gamma_{ij}\right)\right)^{1-s_{ij}}\tag{1}$$

其中，$\varphi(\Gamma_{ij})=1/\left(1+\mathrm{e}^{-\Gamma_{ij}}\right)$ 表示 Sigmoid 函数，$\Gamma_{ij}=\left\langle\boldsymbol{b}_i^{*},\boldsymbol{b}_j^{*}\right\rangle$，$\langle\cdot,\cdot\rangle$ 表示内积运算。哈希编码 $\boldsymbol{b}_i^{*}$ 和 $\boldsymbol{b}_j^{*}$ 间的内积越大，则 $s_{ij}=1$ 的概率越高；否则，$s_{ij}=0$ 的概率越高。求解各模态哈希编码 $\boldsymbol{B}^{*}$，等同于求

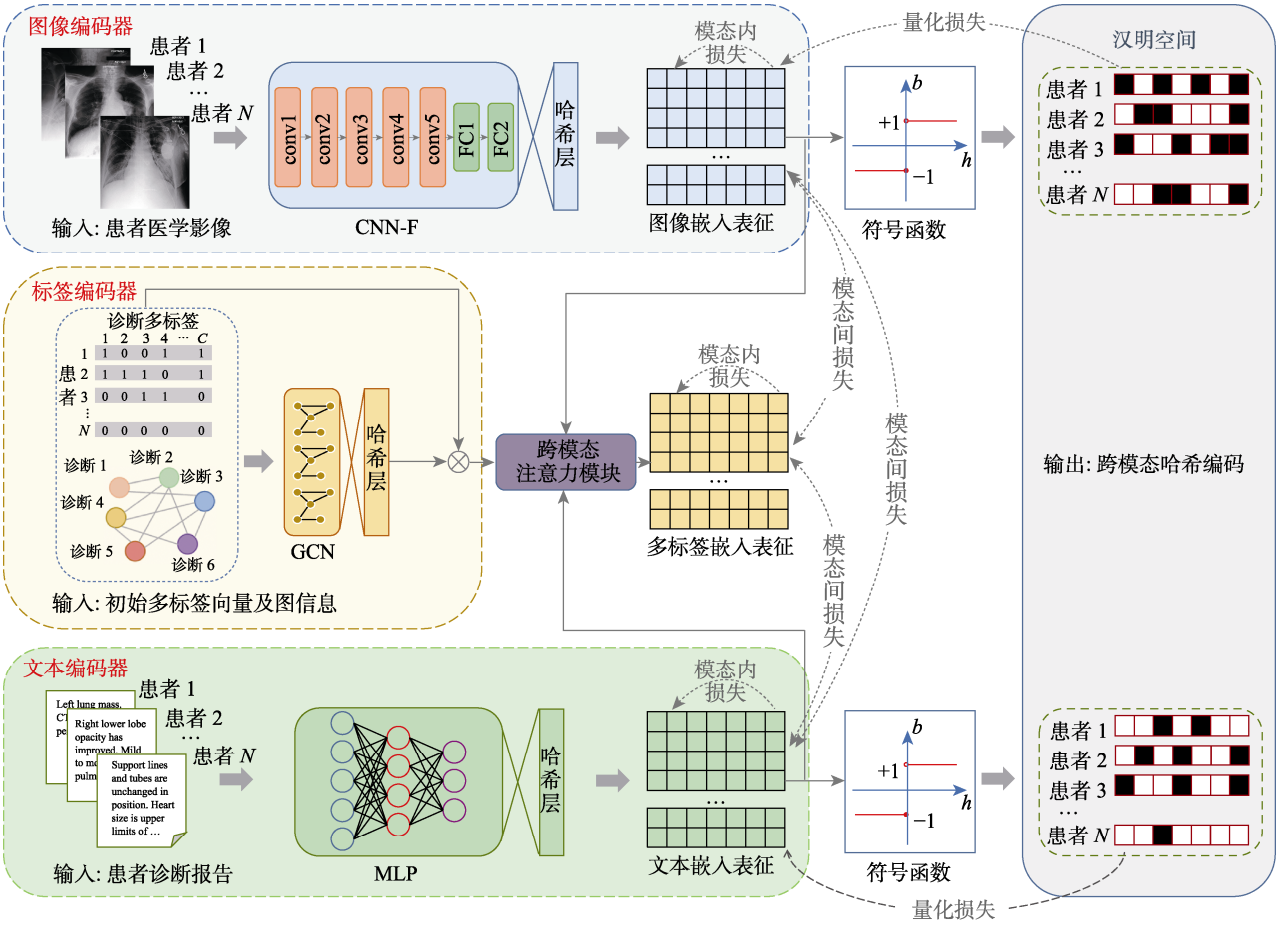


图 1 GEHCA 模型框架

解式（1）的最大似然估计。

2.2 图像编码器

参考文献[2-3,13]，本文采用 7 层 CNN-F[29]模型从原始图像数据 $\boldsymbol{X}$ 中提取特征，并以一个具有 K 个隐藏神经元的全连接层（fully connected layer，FC）作为哈希层，将 CNN-F 提取的特征输入该层中，输出图像数据的类哈希实值嵌入表征。该过程可以表示为 $\boldsymbol{H}^{\mathrm{X}}=E_{\mathrm{X}}\left(\boldsymbol{X};\Theta_{\mathrm{X}}\right)$。其中，$E_{\mathrm{X}}$ 表示图像编码器；Θ_{X} 表示 E_{X} 中的可训练参数；$\boldsymbol{H}^{\mathrm{X}}=\left\{\boldsymbol{h}_i^{\mathrm{X}}\right\}_{i=1}^{N}\in\mathbb{R}^{N\times K}$ 表示输出的嵌入表征。图像编码器的配置如表 1 所示，各符号含义如下：

（1）conv 表示卷积层；

（2）kernel.a×b×c 表示使用 a 个卷积核，每个卷积核尺寸为 b×c；

（3）st.a×b 表示在宽和高上的卷积核移动步长分别是 a 和 b；

（4）pad.a×b 表示在宽和高上的填充像素数分别为 a 和 b；

（5）pool.a×b 表示最大池化窗口尺寸为 a×b；

（6）FC 的数字（如 K）表示该层的隐藏神经元数，等于该层的输出维度。

表 1 图像编码器配置

序号	层类型	配置
1	conv1	kernel.64×11×11；st.4×4；pool.3×3
2	conv2	kernel.256×5×5；st.1×1；pad.2×2；pool.3×3
3	conv3	kernel.256×3×3；st.1×1；pad.1×1
4	conv4	kernel.256×3×3；st.1×1；pad.1×1
5	conv5	kernel.256×3×3；st.1×1；pad.1×1；pool.3×3
6	FC1	4 096
7	FC2	4 096
8	FC3	K

图像编码器网络中，前 7 层激活函数采用渗漏型整流线性单元（leaky rectified linear unit，Leaky ReLU）。为了使网络的实值输出尽量接近哈希编码的二值形式，减少后续二值化操作带来的信息损失，最后一层的激活函数选择双曲正切函数（tanh）；然后对图像编码器输出的实值嵌入表征 $\boldsymbol{H}^{\mathrm{X}}$ 进行二值化处理，获取图像模态对应的哈希编

码 $\boldsymbol{B}^{\mathrm{X}}$。该过程使用元素级符号函数实现，表示为

$$b_{ij}^{\mathrm{X}}=\operatorname{sgn}\left(h_{ij}^{\mathrm{X}}\right)=\begin{cases}1, & h_{ij}^{\mathrm{X}}>0\\ -1, & h_{ij}^{\mathrm{X}}\leqslant 0\end{cases} \tag{2}$$

其中，h_{ij}^{X} 表示 $\boldsymbol{h}_i^{\mathrm{X}}$ 的第 j 个元素，b_{ij}^{X} 表示 $\boldsymbol{b}_i^{\mathrm{X}}$ 的第 j 个元素。

2.3　文本编码器

为了从文本数据中提取特征，首先需要将诊断报告转换为向量形式。参考患者跨模态检索的相关研究[2,11]，统计训练数据集的诊断报告中 617 个出现频率最高的词汇，并根据是否包含这些词汇，将每份报告转换为一个 617 维的多热向量，作为原始文本的特征 $\boldsymbol{t}_i$；随后采用一个双层 MLP 来对 $\boldsymbol{T}$ 进行特征提取。将提取到的特征再输入到一个具有 K 个隐藏神经元的哈希层中，最终得到文本数据的类哈希实值嵌入表征。该过程可以表示为 $\boldsymbol{H}^{\mathrm{T}}=E_{\mathrm{T}}\left(\boldsymbol{T};\Theta_{\mathrm{T}}\right)$。其中，$E_{\mathrm{T}}$ 表示文本编码器；Θ_{T} 表示 E_{T} 中的可训练参数；$\boldsymbol{H}^{\mathrm{T}}=\left\{\boldsymbol{h}_i^{\mathrm{T}}\right\}_{i=1}^{N}\in\mathbb{R}^{N\times K}$ 表示输出的嵌入表征。文本编码器的配置如表 2 所示。

表 2　文本编码器配置

序号	层类型	配置
1	FC1	M_{T}
2	FC2	M_{T}
3	FC3	K

在文本编码器网络中，前 2 层激活函数采用 ReLU，最后一层的激活函数同样选择 tanh；然后采用式（2）中的元素级符号函数 $\operatorname{sgn}(\cdot)$ 对 $\boldsymbol{H}^{\mathrm{T}}$ 进行二值化，得到文本模态对应的哈希编码 $\boldsymbol{B}^{\mathrm{T}}$。

2.4　标签编码器

为了有效挖掘并利用诊断类别多标签间的语义关联，本文设计一个图信息增强的标签编码器获取多标签嵌入表征。编码器结构如图 2 所示。

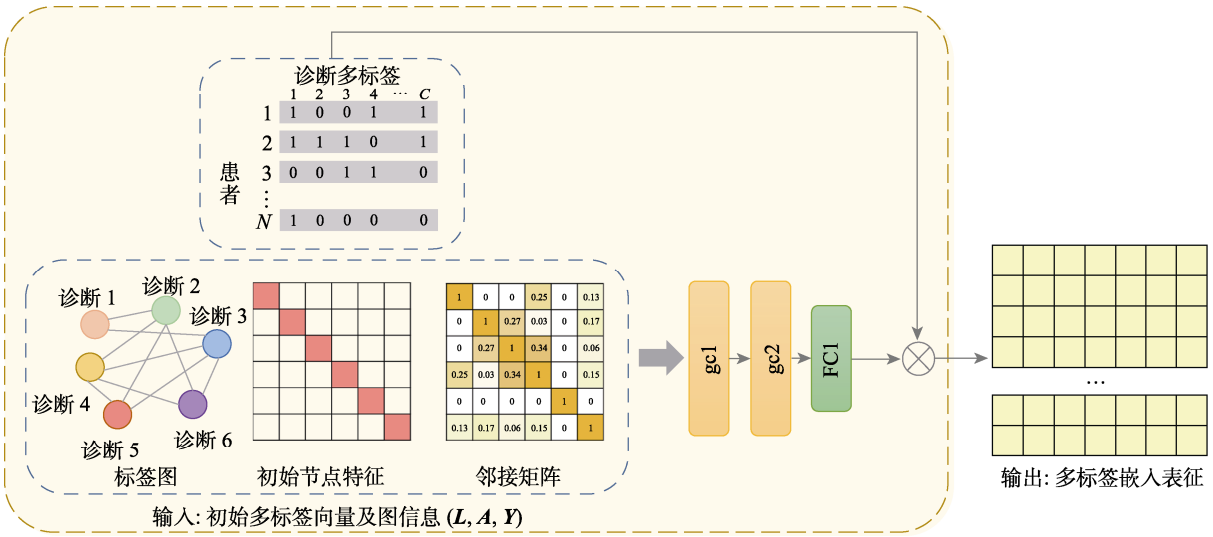


图 2　标签编码器结构

首先将多标签组织为一个图 $G=(V,E)$。节点集 V 表示诊断类别标签集，每个节点对应一项特定的诊断，节点总数与诊断类别总数 C 相等。节点的初始特征表示为 $\boldsymbol{Y}=\left\{\boldsymbol{y}_i\right\}_{i=1}^{C}\in\mathbb{R}^{C\times M_{\mathrm{L}}}$，其中，$\boldsymbol{y}_i\in\mathbb{R}^{M_{\mathrm{L}}}$ 表示节点 i 的 M_{L} 维特征向量。本文中，$\boldsymbol{Y}$ 被设置为一个 $C\times C$ 的单位矩阵，即每个节点的特征都被初始化为一个独热向量。边集为 E，其中，节点 i 和节点 j 间的边连接记为 $e_{ij}\in E$。图 G 的拓扑结构可以通过邻接矩阵 $\boldsymbol{A}\in\mathbb{R}^{C\times C}$ 表示，其中，每个元素 $\boldsymbol{A}_{ij}$ 代表边 e_{ij} 的权重。为了有效地捕捉诊断之间的共现关联，采用 PMI 确定 $\boldsymbol{A}_{ij}$ 的值。诊断类别 i 和 j 之间 PMI 值的计算公式为

$$\mathrm{PMI}(i,j)=\mathrm{lb}\left(\frac{n(i,j)}{n(i)n(j)}\times N\right),$$

其中，$n(i,j)$ 表示同时标记 i 和 j 这 2 项诊断的患者实例总数，$n(i)$ 和 $n(j)$ 分别表示标记了诊断 i 和诊断 j 的患者实例数。得到 PMI 值后，$\boldsymbol{A}_{ij}$ 可以通过如下方式确定

$$\boldsymbol{A}_{ij}=\begin{cases}\mathrm{PMI}(i,j), & \mathrm{PMI}(i,j)>0\\ 0, & \mathrm{PMI}(i,j)\leqslant 0\end{cases}。$$

此外，每个节点与其自身也是高度关联的。为了避

免损失节点自身信息，构建邻接矩阵 $\boldsymbol{A}$ 时，需将其与一个单位矩阵相加，补充节点自连接，得到 $\tilde{\boldsymbol{A}}$。相应地，度矩阵可表示为 $\tilde{\boldsymbol{D}}$，其中，$\tilde{\boldsymbol{D}}_{ii}=\sum_j \tilde{\boldsymbol{A}}_{ij}$；然后采用多层 GCN 对标签图进行特征提取。图卷积层的表达式为

$$\boldsymbol{Z}^{(q)}=\sigma\left(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{Z}^{(q-1)}\boldsymbol{W}^{(q-1)}\right)。$$

其中，q 表示层数；$\sigma(\cdot)$ 表示激活函数；$\boldsymbol{Z}^{(0)}$ 表示节点的初始特征矩阵 $\boldsymbol{Y}$；$\boldsymbol{Z}^{(q)}$ 表示第 q 层输出的节点嵌入矩阵；$\boldsymbol{W}^{(q-1)}\in\mathbb{R}^{F\times F'}$ 表示该图卷积层的权重参数矩阵，F 和 F' 分别代表输入和输出维度。随着图卷积层数增多，节点可以聚合其更高阶邻域的特征，进一步挖掘图结构信息，为了避免图卷积层数过多引发的过平滑问题[30]，采用 2 层图卷积层。GCN 的输出 $\boldsymbol{Z}^{(2)}$ 被输入到一个具有 K 个隐藏单元的哈希层中，得到类哈希表征 $\boldsymbol{Z}^{(3)}\in\mathbb{R}^{C\times K}$，此时，$\boldsymbol{Z}^{(3)}$ 中的每行是每个诊断类别标签自身的嵌入向量。为了得到患者实例 i 的多标签嵌入表征，需结合多标签标注 $\boldsymbol{l}_i$ 对各诊断类别的嵌入向量进行融合，计算公式为

$$\boldsymbol{z}_i=\sigma\left(\boldsymbol{l}_i\boldsymbol{Z}^{(3)}\Big/\sum_{j=1}^{C}l_{ij}\right) \tag{3}$$

其中，$\boldsymbol{z}_i\in\mathbb{R}^k$ 表示患者实例 i 的多标签类哈希实值嵌入表征；$\sigma(\cdot)$ 表示一个非线性激活函数。

最终，以上多标签嵌入表征的编码过程可以表示为 $\boldsymbol{Z}=E_{\mathrm{L}}(\boldsymbol{L},\boldsymbol{A},\boldsymbol{Y};\Theta_{\mathrm{L}})$。其中，$E_{\mathrm{L}}$ 表示标签编码器；Θ_{L} 表示需要训练的网络参数；$\boldsymbol{Z}=\{\boldsymbol{z}_i\}_{i=1}^{N}\in\mathbb{R}^{N\times K}$ 表示输出的多标签嵌入表征矩阵。标签编码器的配置如表 3 所示，其中，gc 代表图卷积层，其后配置列的数字表示该层的输出维度；2 层图卷积层均采用 ReLU 作为激活函数；由于哈希层的输出并非标签编码器的最终输出，因此未设置激活函数，而是将式（3）中的激活函数设置为 tanh，减少多标签与图像、文本模态嵌入表征的异质性，便于引导图像和文本编码器的训练过程。

表 3　标签编码器配置

序号	层类型	配置
1	gc1	C
2	gc2	C
3	FC1	K

2.5　跨模态注意力模块

设计标签编码器的核心是，利用生成的多标签嵌入表征指导图像与文本模态数据的哈希函数学习过程。然而，由于多标签矩阵普遍稀疏，多标签嵌入表征往往面临语义信息量不足的问题。因此，本文引入跨模态注意力模块[3]，旨在通过注意力机制，将图像和文本模态的丰富语义信息融入多标签嵌入表征中，增强其信息量。为清晰起见，本节后续内容中将多标签也表述为患者实例的另一种模态。跨模态注意力模块的结构如图 3 所示，其中，a^{XL} 和 a^{TL} 分别代表图像与多标签嵌入表征间、文本与多标签嵌入表征间的注意力权重系数，$\otimes$ 表示元素级相乘，$\oplus$ 表示对位相加。

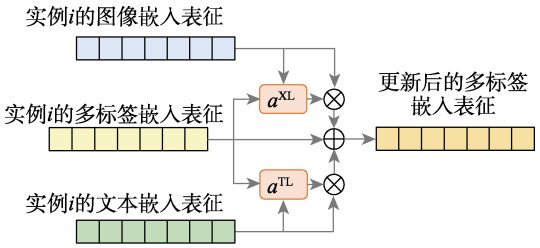


图 3　跨模态注意力模块结构

实例 i 的图像与多标签、文本与多标签之间的跨模态注意力权重计算过程表示为

$$\begin{cases} a_i^{\mathrm{XL}}=\dfrac{\left\langle \boldsymbol{h}_i^{\mathrm{X}},\boldsymbol{z}_i\right\rangle}{\left\|\boldsymbol{h}_i^{\mathrm{X}}\right\|_2\left\|\boldsymbol{z}_i\right\|_2} \\ a_i^{\mathrm{TL}}=\dfrac{\left\langle \boldsymbol{h}_i^{\mathrm{T}},\boldsymbol{z}_i\right\rangle}{\left\|\boldsymbol{h}_i^{\mathrm{T}}\right\|_2\left\|\boldsymbol{z}_i\right\|_2} \end{cases} \tag{4}$$

其中，$\|\cdot\|_2$ 表示向量的 2-范数。融合图像与文本的嵌入表征，将实例 i 的多标签嵌入表征更新为

$$\boldsymbol{h}_i^{\mathrm{L}}=\boldsymbol{z}_i+a_i^{\mathrm{XL}}\boldsymbol{h}_i^{\mathrm{X}}+a_i^{\mathrm{TL}}\boldsymbol{h}_i^{\mathrm{T}} \tag{5}$$

最终，获得的多标签嵌入表征矩阵可表示为 $\boldsymbol{H}^L=\{\boldsymbol{h}_i^L\}_{i=1}^N\in\mathbb{R}^{N\times K}$。利用式（4）（5），能够有效地补充多标签嵌入的语义信息，缓解多标签矩阵稀疏带来的不利影响。

2.6　目标函数和学习策略

如第 2.1 节所述，对各模态哈希编码 $\boldsymbol{B}^*$ 的求解过程等同于对式（1）的最大似然估计。因此，将其等价转换为求解负对数似然最小值的形式，即可得到跨模态哈希过程的损失函数，公式为

$$\begin{aligned} &L_{\mathrm{b}}\left(\boldsymbol{B}^*,\boldsymbol{B}^*\right)=-\ln p\left(S\mid\boldsymbol{B}^*,\boldsymbol{B}^*\right)= \\ &-\ln\sum_{s_{ij}\in S}p\left(s_{ij}\mid\boldsymbol{b}_i^*,\boldsymbol{b}_j^*\right)= \\ &\sum_{s_{ij}\in S}p\left(\ln\left(1+\exp\left(\left\langle\boldsymbol{b}_i^*,\boldsymbol{b}_j^*\right\rangle\right)\right)-s_{ij}\left\langle\boldsymbol{b}_i^*,\boldsymbol{b}_j^*\right\rangle\right) \end{aligned} \tag{6}$$

最小化该损失函数，即可获得保持实例间相似性的跨模态精确哈希编码。然而，由于生成哈希编码过程中采用的符号函数式（2）在非 0 值处的梯度始终为 0，会引发训练过程中的梯度消失问题，使反向传播机制失效，因此该优化问题难以直接求解。针对这一问题，采用一种最为常用的解决策略——连续松弛[31]，即用实值的类哈希嵌入表征 $\boldsymbol{H}^*$ 代替二值哈希编码 $\boldsymbol{B}^*$，将式（6）改写为

$$L_{\mathrm{h}}\left(\boldsymbol{H}^*,\boldsymbol{H}^*\right)=\sum_{s_{ij}\in S}\omega_{ij}\left(\ln\left(1+\exp\left(\left\langle \boldsymbol{h}_i^*,\boldsymbol{h}_j^*\right\rangle\right)\right)-s_{ij}\left\langle \boldsymbol{h}_i^*,\boldsymbol{h}_j^*\right\rangle\right)。$$

其中，$\omega_{ij}=\left(|S|/|S_1|\right)^{s_{ij}}\left(|S|/|S_0|\right)^{1-s_{ij}}$ 表示用于缓解相似和不相似实例对数量不平衡的权重参数；$S_1=\left\{s_{ij}\in S:s_{ij}=1\right\}$ 和 $S_0=\left\{s_{ij}\in S:s_{ij}=0\right\}$ 分别表示相似实例对集合和不相似实例对集合。

具有相同诊断标签的患者实例中，即使数据模态不同，也应当具有语义相似性。为了实现各模态协同建模，设计一个模态间相似性损失函数，用于将诊断多标签间的共现关联和语义信息补充到图像、文本模态的哈希编码中。其计算公式为

$$L_{\text{inter-modal}}=L_{\mathrm{h}}\left(\boldsymbol{H}^{\mathrm{X}},\boldsymbol{H}^{\mathrm{T}}\right)+L_{\mathrm{h}}\left(\boldsymbol{H}^{\mathrm{X}},\boldsymbol{H}^{\mathrm{L}}\right)+L_{\mathrm{h}}\left(\boldsymbol{H}^{\mathrm{T}},\boldsymbol{H}^{\mathrm{L}}\right)。$$

通过该损失函数，多标签嵌入表征中蕴含的标签图结构信息可以被传递到图像与文本的哈希编码训练过程中，实现对跨模态哈希的图增强效果，丰富哈希编码的信息量。借助跨模态注意力机制，实现多标签嵌入语义信息融合，同时结合图像与文本模态之间的成对损失函数，能够进一步减少各模态表征之间的异质性，增强哈希编码的判别能力。此外，由于同模态数据的各实例间的相似性也应得到保持，因此本文还引入了一个模态内相似性损失函数，表示为

$$L_{\text{intra-modal}}=L_{\mathrm{h}}\left(\boldsymbol{H}^{\mathrm{X}},\boldsymbol{H}^{\mathrm{X}}\right)+L_{\mathrm{h}}\left(\boldsymbol{H}^{\mathrm{T}},\boldsymbol{H}^{\mathrm{T}}\right)+L_{\mathrm{h}}\left(\boldsymbol{H}^{\mathrm{L}},\boldsymbol{H}^{\mathrm{L}}\right)。$$

为了减少实值嵌入表征 $\boldsymbol{H}^*$ 二值化为哈希编码时产生的信息损失，在损失函数中补充量化正则项 $L_{\text{quantization}}=\sum_{i=1}^{N}\left\|\boldsymbol{h}_i^{\mathrm{X}}-\boldsymbol{b}_i^{\mathrm{X}}\right\|_2^2+\left\|\boldsymbol{h}_i^{\mathrm{T}}-\boldsymbol{b}_i^{\mathrm{T}}\right\|_2^2$。最终，GEHCA 模型的整体目标函数可以表示为

$$\min_{\Theta_{\mathrm{X}},\Theta_{\mathrm{T}},\Theta_{\mathrm{L}}} L=L_{\text{inter-modal}}+\alpha L_{\text{intra-modal}}+\beta L_{\text{quantization}}\quad(7)$$

其中，α 和 β 表示平衡各损失项影响的超参数。由于式（7）是可微的，因此模型的全部参数可以通过标准反向传播机制和小批量梯度下降算法进行优化，图像、文本模态数据的哈希编码过程能够以端到端的方式联合训练。

3　实验及结果分析

3.1　数据集描述和预处理

本文实验采用 MIMIC-CXR 数据集[12]用于评估算法性能。该数据集是一个大型公开的胸部 X 光成像数据库，收录了由贝斯以色列女执事医疗中心收集的 2011—2016 年 227 827 名患者的放射学研究影像资料，共包含 377 110 张胸部 X 光成像图片，此外，每名患者都有一个自由文本形式的放射学诊断报告。该数据集共包含 14 个诊断标签，分别是肺不张（Atelectasis）、心脏肥大（Cardiomegaly）、肺实变（Consolidation）、肺水肿（Edema）、心纵隔扩大（Enlarged Cardiomediastinum）、骨折（Fracture）、肺部病变（Lung Lesion）、肺部阴影（Lung Opacity）、健康（No Finding）、胸腔积液（Pleural Effusion）、其他胸膜症状（Pleural Other）、肺炎（Pneumonia）、气胸（Pneumothorax）和辅助装置（Support Devices）。对于每个患者实例，各标签对应的诊断如果为阳性，则标记为 1，阴性则标记为 0，无法确定则标记为−1。表 4 所示为 MIMIC-CXR 数据集的数据样例。

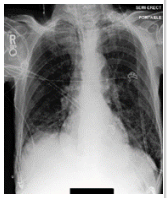
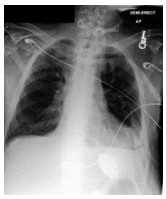
对 MIMIC-CXR 数据集进行预处理，包括去除包含不确定性诊断（即标签值为−1）的实例，以及存在诊断标签矛盾的实例（即存在其他阳性诊断，但仍被错误标记为健康的实例）；保留各实例的正向 X 光成像（如果某实例无正向 X 光成像，则保留其背向 X 光成像）。最终，选取 73 837 个实例的 X 光成像和诊断报告用于实验评估。每个诊断报告都依据最高频出现的 617 个词汇转换为多热向量，每个实例对应的多标签也被表示为一个 14 维多热向量。在测试阶段，参照文献[2,11]，随机选取 762 个实例作为查询集，其余 73 075 个实例用作检索库。在训练阶段，从检索库中随机选择 14 000 个实例作为训练集，并额外选取 1 000 个实例作为验证集。

3.2　实验设置

3.2.1　评价指标

采用平均精度均值（mean average precision，MAP）作为患者跨模态检索的评价指标。对于第 q 个查询，其检索结果的平均精度（average precision，AP）公式为

表 4 MIMIC-CXR 数据集样例

X 光成像	诊断报告	诊断标签	患者 ID
	FINAL REPORT HISTORY: Subarachnoid hemorrhage. FINDINGS: In comparison with the study of ___, the endotracheal tube remains in good position while the side port of the nasogastric tube remains above the esopha-gogastric junction and should be pushed forward several centimeters. Again there is upper lobe predominant emphysema with pulmonary fibrosis that predominantly affects the lower lungs. No evidence of acute focal pneumonia.	其他胸膜症状、辅助装置	s50001598
	FINAL REPORT INDICATION: Hypotension, altered mental status. COMPARISON: Chest radiograph, ___, ___. FINDINGS: Frontal chest radiograph was provided for review. The cardiomediastinal and hilar contours are stable. Again seen is a small left pleural effusion, much decreased compared to the most recent prior study. There is no right pleural effusion. There is no pneumothorax. There has been overall improvement in the bibasilar and right apical opacities since the most recent prior study. A tracheostomy tube is again noted in acceptable position. There has been interval removal of left internal jugular central venous catheter. Again seen is a venous central line terminating near the inferior cavoatrial junction. Contrast is seen in the stomach. IMPRESSION: Overall improvement with small residual left pleural effusion.	胸腔积液	s52105600
	FINAL REPORT EXAMINATION: CHEST (PORTABLE AP)CHEST (PORTABLE AP)i INDICATION: ___ year old woman with acute SOB // Acute SOB COMPARISON: Chest radiographs since ___ most recently ___. IMPRESSION: Mild interstitial edema it is worse, in the left lung. Moderate cardiomegaly chronic. Pleural effusions small if any. No pneumothorax.	心脏肥大、肺水肿、胸腔积液	s53999282

$$\mathrm{AP}_q=\frac{1}{N_q}\sum_{i=1}^{N_q}\frac{r_i}{i}\text{。}$$

其中，N_q 表示在检索库中与当前查询 q 相关的实例总数；r_i 表示在前 i 个检索结果中与查询 q 相关的实例数量。实验中，如果检索返回的实例与查询实例共享至少一个诊断类别标签，则认为二者相关；反之，则认为它们不相关。获得全部查询的 AP 值之后，对其求平均，即可得到

$$\mathrm{MAP}=\frac{1}{Q}\sum_{q=1}^{Q}\mathrm{AP}_q\text{。}$$

其中，Q 表示查询实例总数。MAP 值越高，表明检索性能越好。

3.2.2 对比方法和实现细节

实验中，选择 7 种具有代表性的跨模态哈希方法作为对比方法，包括 3 种经典的传统跨模态哈希方法和 4 种先进的深度跨模态哈希方法。

（1）CCA[18]：一种无监督跨模态哈希方法，通过分析图像与文本模态中提取的代表性特征之间的相关性，生成具有语义匹配特性的哈希编码。

（2）SCM[22]：利用标签信息，在保持 2 个模态间最大相关性的前提下学习模态特定的编码转换。

（3）SePH[23]：采用概率分布的方式构建相似度矩阵，同时最小化 Kullback-Leibler 散度，对各模态数据生成统一形式的二进制编码。

（4）DCMH[9]：将实例间的成对相似性作为监督信息，将特征学习与哈希编码生成过程整合到一个端到端的框架中进行训练。

（5）SSAH[14]：将自监督学习范式引入跨模态哈希中，利用多标签模态数据学习哈希函数，指导并监督其他模态数据的哈希编码过程。

（6）DSIBH[11]：将深层确定瓶颈机制扩展至跨模态中，减少与标签无关的冗余信息，增强各模态哈希码的判别力。

（7）DMCAH[2]：利用注意力机制，对各模态数据的粗、细粒度特征进行整合，并将这些特征映射至公共空间来获取相应的哈希编码。

除 CCA 外，以上方法均属于监督式跨模态哈希方法，所有方法依据其原始文献中推荐的参数进行复现。为公平比较，GEHCA 模型的图像编码器采用 CNN-F 在训练集的 X 光成像上提取的 4 096 维特征，同样作为 CCA、SCM 和 SePH 这 3 种传统方法图像模态的原始特征。在 GEHCA 的训练过程中，小批量大小设定为 128，采用 Adam 优化器进行优化，学习率和权重衰减均设置为 10^{-4}。式(7)中，设模态内相似损失项系数 α =1，量化正则项系数 β =0.01。为防止过拟合，实验中，GEHCA 模型训练的最大 epoch 数限制为 100，并采用早停策略，耐心值设为 10 个 epoch，即在验证集上，如果模型的损失函数值在连续 10 个 epoch 内没有下降，则停止训练。

实验中，传统跨模态哈希方法均通过 scikit-learn 工具包[32]实现，深度跨模态哈希方法均采用 PyTorch[33]实现；使用的服务器搭载 Intel(R)Xeon(R) Gold 5218R@ 2.10 GHz CPU，251 GB 内存和 NVIDIA GeForce RTX 3090 显卡。

3.3　对比实验

在 MIMIC-CXR 数据集上，针对 X 光成像检索诊断报告（简称为 X→T）与诊断报告检索 X 光成像（简称为 T→X）这 2 项任务，8 种方法的 MAP 指标结果如表 5 所示，其中，所有数值均为 5 次随机实验的平均值；参考文献[2,9,11,14]，哈希编码位数设定为 16、32、64 和 128 位。可以看出：（1）在 X→T 任务上，与次优方法相比，GEHCA 在编码长度为 16、32、64 和 128 位时，MAP 得分分别提高 2.98%、1.21%、0.63%和 0.53%；与其他方法相比，SSAH[14]、DSIBH[11]和 GEHCA 展现出更明显的性能优势，且在编码位数低时仍有较好的表现，原因是这些方法均设计了标签编码器来提取多标签中的语义特征并指导哈希过程，丰富了哈希编码信息量。（2）在 T→X 任务中，由于本实验对诊断报告文本的预处理方式相对简单，8 种方法的 MAP 得分在直观上差异较小，但 GEHCA 仍表现出最佳的检索性能；与次优方法相比，GEHCA 在 16、32、64 和 128 位编码长度下，MAP 得分分别提升 0.91%、0.75%、1.03%和 0.57%。

表 5　在 MIMIC-CXR 数据集上 8 种方法的 MAP 得分结果对比

方法	X→T				T→X			
	16 位	32 位	64 位	128 位	16 位	32 位	64 位	128 位
CCA[18]	0.519 4	0.517 5	0.516 4	0.516 1	0.523 2	0.520 4	0.517 8	0.515 5
SCM[22]	0.515 3	0.518 5	0.515 4	0.519 1	0.514 8	0.517 9	0.517 9	0.515 7
SePH[23]	0.525 6	0.518 2	0.529 7	0.532 8	0.513 7	0.525 5	0.523 7	0.521 5
DCMH[9]	0.525 8	0.549 5	0.559 0	0.581 3	0.517 1	0.517 4	0.523 7	0.524 1
SSAH[14]	0.588 0	0.578 7	0.579 9	0.590 5	0.518 0	0.526 6	0.526 6	0.529 7
DSIBH[11]	0.652 0	0.668 3	0.666 8	0.663 2	0.524 0	0.532 2	0.534 6	0.539 4
DMCAH[2]	0.555 7	0.557 6	0.559 0	0.578 3	0.526 7	0.528 8	0.531 3	0.529 7
GEHCA	**0.671 4**	**0.676 4**	**0.671 0**	**0.666 7**	**0.531 5**	**0.536 2**	**0.540 1**	**0.542 5**

注：粗体表示最优值，下划线表示次优值。

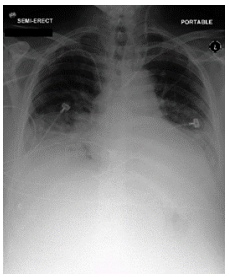
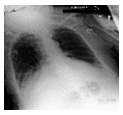
为了评估表 5 结果中 GEHCA 模型性能优势的统计显著性，采用 Wilcoxon 秩和检验[34]对该模型与其他 7 种对比方法的 5 次随机实验结果进行对比，如果 2 种方法秩和检验的 p 值小于 0.05，则认为具有显著性能差异。在 MIMIC-CXR 数据集上的 2 个检索任务中，4 种哈希编码长度下，以 GEHCA 作为对照方法的 p 值对比如表 6 所示。可以看出，X→T 任务中，当编码长度为 64 位和 128 位时，GEHCA 与 DSIBH[11]的性能差异并不显著；其他情况下，与其他方法相比，GEHCA 其他均展现出显著的性能优势。

在 MIMIC-CXR 数据集上，GEHCA 与经典传统跨模态哈希方法 CCA、代表性的 DCMH 方法前 5 个跨模态检索结果的对比样例如表 7 所示。可以看出，对于一个被诊断为肺不张和胸腔积液的查询实例，X→T 任务中，CCA 的前 5 个检索结果中有 3 个与查询实例不相关，DCMH[9]中有 2 个不相关，而 GEHCA 的前 5 个检索结果全部与查询实例相关。在 T→X 任务中，CCA 的前 5 个检索结果中有 2 个与查询实例不相关，DCMH[9]中有 1 个不相关，而 GEHCA 的检索结果均与查询实例相关。实验结果表明，与 CCA 和 DCMH[9]相比，GEHCA

表 6　在 MIMIC-CXR 数据集上 GEHCA 与 7 种方法 MAP 得分的 Wilcoxon 秩和检验 p 值对比

方法	X→T				T→X			
	16 位	32 位	64 位	128 位	16 位	32 位	64 位	128 位
CCA[18]	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9
SCM[22]	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9
SePH[23]	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9
DCMH[9]	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9
SSAH[14]	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9
DSIBH[11]	0.015 9	0.015 9	0.841 3	0.547 6	0.007 9	0.007 9	0.007 9	0.007 9
DMCAH[2]	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9	0.007 9

表 7 在 MIMIC-CXR 数据集上 3 种方法的前 5 个跨模态检索结果对比

任务	查询	方法	前 5 个检索结果				
X→T	 肺不张、胸腔积液	CCA[18]	A small left pleural effusion is present. Left pectoral pacemaker in situ No pneumonia...	Compared to prior study 1 day ago, there is slightly more pulmonary edema.	There is now pulmonary edema of moderate severity. A small left pleural…	The size of the cardiac silhouette has minimally increased. No evidence of pneumonia…	Borderline size of the cardiac silhouette. Moderate retro-cardiac atelectasis…
			肺部阴影、肺炎	肺水肿	肺不张、肺水肿、胸腔积液、…	心脏肥大	肺不张、心脏肥大、骨折
		DCMH[9]	There is a large right pleural effusion, which obscures the right heart border…	No change compared to the prior study with a moderate to large left hydropneumothorax…	Massive cardiomegaly with retrocardiac and right basilar atelectasis. No larger pleural...	They might potentially reflect improving pulmonary edema, but attention …	…which likely represents a combination of pleural fluid, consolidation, and…
			胸腔积液	气胸	肺不张、心脏肥大、肺部阴影	肺水肿、骨折、肺部阴影、…	肺不张、肺实变、胸腔积液、…
		GEHCA	Moderate right and small left pleural effusions unchanged. Previous …	Likely small left and right pleural effusions. Retrocardiac atelectasis is un-changed.	Mild pulmonary edema has improved. Bibasilar atelectasis and small effusions...	…pleural effusion has been evacuated and a small pneumothorax…	Stable right-sided pleural effusion. No new acute cardiopulmonary process…
			肺不张、心脏肥大、胸腔积液	肺不张、胸腔积液、辅助装置	肺不张、肺水肿、胸腔积液、…	肺不张、胸腔积液、气胸、…	肺部病变、肺部阴影、胸腔积液
T→X	PROVISIONAL FINDINGS IMPRESSION (PFI): __ __ __ 11:50 AM IMPRESSION: Small bilateral pleural effusions with associated atelectasis. FINAL REPORT HISTORY: __-year-old male with right-sided pleural effusion and status post right colectomy, liver transplant, and splenectomy. STUDY: Portable AP semi-upright chest radiograph. COMPARISON: Multiple chest radiographs from __ through __. FINDINGS: There has been interval removal of the endogastric tube. The left central line tip sits at the left brachiocephalic vein while the right central line tip sits in the upper SVC. The heart size is at the upper limits of normal but is unchanged. The mediastinal and hilar contours are normal. Additionally, bibasilar opacities are somewhat worsened with no clear delineation of the costophrenic angles. These findings are compatible with small bilateral pleural effusions with associated worsening atelectasis at both bases. There is no pneumothorax. IMPRESSION: Small bilateral pleural effusions with associated atelectasis. 肺不张、胸腔积液	CCA[18]					
			肺不张、心脏肥大、胸腔积液	肺部阴影、辅助装置	心脏肥大、胸腔积液、…	肺部阴影、肺炎	心脏肥大、胸腔积液、…
		DCMH[9]					
			肺不张、气胸、胸腔积液、…	心脏肥大、辅助装置	肺不张、肺水肿、辅助装置	肺不张、肺水肿、胸腔积液、…	胸腔积液
		GEHCA					
			肺不张、胸腔积液	肺不张、气胸、辅助装置	胸腔积液、气胸、辅助装置	肺水肿、胸腔积液、…	胸腔积液、辅助装置

注：下划线表示检索结果中与当前查询重合的诊断类别标签。

能检索到更精确的候选结果，有助于为医生和患者提供更有价值的参考。

3.4 消融实验

在 MIMIC-CXR 数据集上进行消融实验，验证 GEHCA 模型中的 3 个变体 GEHCA-G 和 GEHCA-A 和 GEHCA-B 对跨模态检索性能的影响。其中，GEHCA-G 表示去除图增强的变体，通过将标签编码器的 2 层图卷积层替换为包含相同神经元数的 FC 来实现；GEHCA-A 表示去除跨模态注意力模块的变体，将式（5）中的 $\boldsymbol{h}_i^{\mathrm{L}}$ 更换为标签编码器的输出 $\boldsymbol{z}_i$；GEHCA-B 指同时移除以上 2 个设计的变体。实验中，选取 16 位和 128 位作为哈希编码长度的典型代表，消融实验结果如图 4 所示。在 X→T 任务中，当编码长度为 16 位和 128 位时，与 GEHCA 相比，GEHCA-G 的 MAP 得分分别下降 1.50%和 0.45%，GEHCA-A 的 MAP 得分分别下降 2.60%和 0.58%，GEHCA-B 的 MAP 得分分别下降 3.64%和 2.16%。在 T→X 任务中，与 GEHCA 相比，不同编码长度时，GEHCA-G 的 MAP 得分分别下降 0.41%和 0.11%，GEHCA-A 的 MAP 得分分别下降 0.15%和 0.28%，GEHCA-B 的 MAP 得分分别下降 0.66%和 0.31%。消融实验结果表明，挖掘诊断标签的共现关联信息，并引入跨模态注意力机制对多标签嵌入表征的语义信息量进行补充，均有助于提升模型性能。特别是在 X→T 任务中，当编码位数较低时，GEHCA-A 和 GEHCA-B 能有效地增加哈希编码的信息量和判别力，在降低存储开销的同时提高检索准确率。

3.5 参数分析

在数据集 MIMIC-CXR 上，编码长度为 16 位

时，以验证集实例作为查询进行参数分析实验，探究式(7)中超参数 α 和 β 对模型检索性能的影响，结果如图 5 所示。可以看出，当 α 的取值从 0.01 增长到 1 时，GEHCA 在 2 项跨模态检索任务上的 MAP 得分有所提高，而当 α 值超过 10 时，模型性能开始迅速下降；在 β 值为 0.000 1~0.1 时，GEHCA 性能变化幅度较小，$\beta=0.01$ 时的 MAP 得分略高于其他取值点位，而 β 值超过 0.1 后，模型性能则明显下降。综上，如第 3.2 节所述，本文实验中设定 $\alpha=1$，$\beta=0.01$。

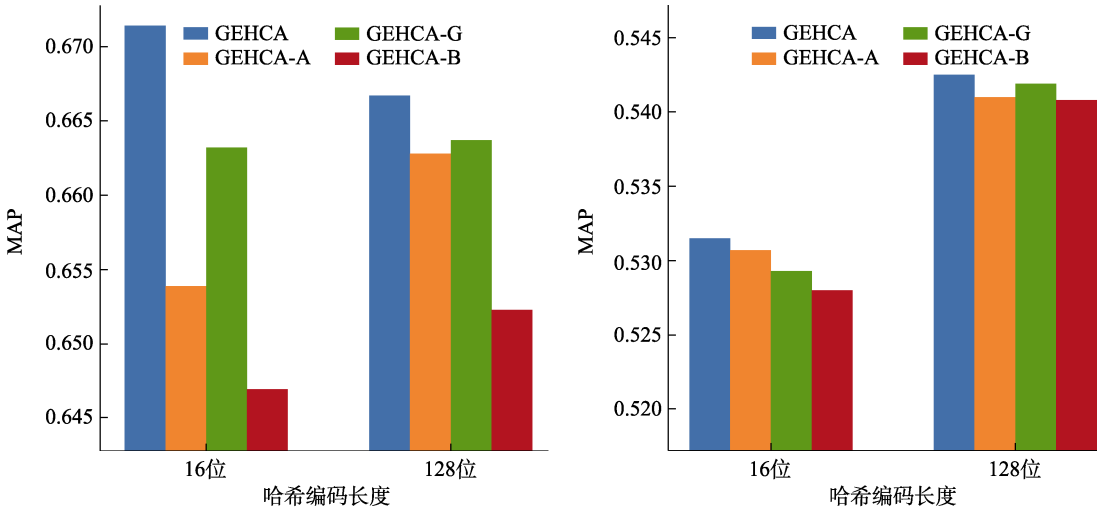


图 4　在 MIMIC-CXR 数据集上的消融实验结果

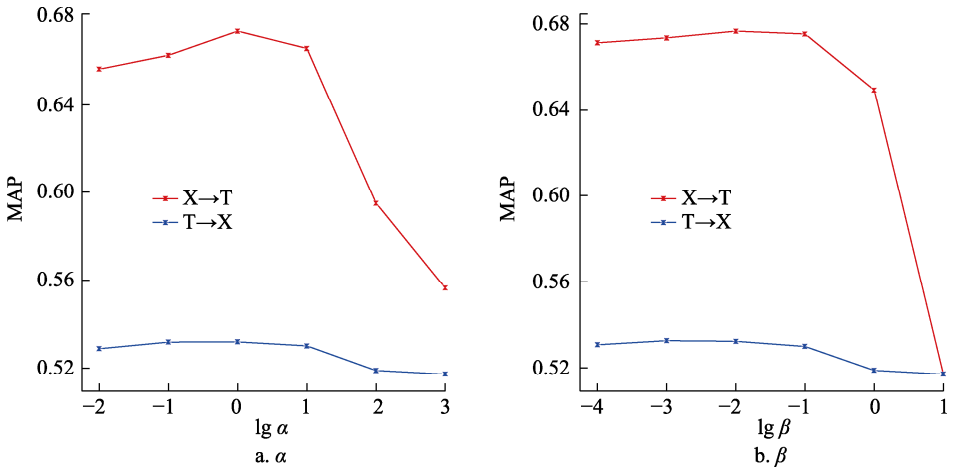


图 5　不同 α 值和 β 值对 MAP 得分的影响

4　结　语

本文提出一种 GEHCA 模型，用于患者医学影像与诊断报告文本间的跨模态检索。除了采用图像编码器及文本编码器提取相应模态数据特征并进行哈希编码外，还设计了图信息增强的标签编码器，用于将诊断类别标签的共现关联信息融入多标签嵌入表征中；为了缓解多标签矩阵的稀疏性问题，引入图像、文本与多标签嵌入表征间的跨模态注意力机制，通过整合其他模态的特征，丰富多标签嵌入的语义信息量；设计了一个模态内与模态间相结合的相似性损失函数，将多标签嵌入表征的共现关联和语义信息传递至医学影像和诊断报告的哈希编码学习过程中，增强哈希编码的判别力，提高患者跨模态检索的性能。在 MIMIC-CXR 上数据集进行实验的结果表明，在多个编码位数下，GEHCA 模型在跨模态检索任务上具备性能优势，并进一步验证了图增强和跨模态注意力 2 项关键设计对该模型性能的贡献。但是，GEHCA 模型也具有一定的局限性，对于模型的验证主要聚焦于医

学数据的跨模态检索任务，并未在通用的图像、文本跨模态数据集上对模型进行评估，因此该模型的广泛适用性需进一步验证。

未来的工作聚焦于 2 个方面：（1）拓展模型适用的模态，使其能灵活地应用于图像和文本模态之外更广泛的医学数据类型，包括但不限于表格型数据（如结构化电子病历）、基因组学数据以及蛋白组学数据等；（2）探索生成式大语言模型在医学跨模态数据研究的创新应用，将检索报告模式改进为生成报告模式，在临床实践中实现更便捷、高效的应用体验。

参考文献（References）：

[1] Ding Guohui, Zhang Qi, Fang Shichao, *et al*. Review of multi-modal retrieval in medicine[J]. Computer Engineering and Applications, 2023, 59(1): 26-36(in Chinese)
(丁国辉, 张琦, 房士超, 等. 多模态检索在医学领域的研究综述[J]. 计算机工程与应用, 2023, 59(1): 26-36)

[2] Zhang Y, Ou W H, Shi Y F, *et al*. Deep medical cross-modal attention hashing[J]. World Wide Web, 2022, 25(4): 1519-1536

[3] Zou X T, Wu S, Zhang N, *et al*. Multi-label modality enhanced attention based self-supervised deep cross-modal hashing[J]. Knowledge-Based Systems, 2022, 239: Article No.107927

[4] Jia Y Q, Salzmann M, Darrell T. Learning cross-modality similarity for multinomial data[C] //Proceedings of the International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2011: 2407-2414

[5] Wang K Y, He R, Wang L, *et al*. Joint feature selection and subspace learning for cross-modal retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(10): 2010-2023

[6] Qian S S, Xue D Z, Fang Q, *et al*. Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(4): 4794-4811

[7] Zhu L, Zheng C Q, Guan W L, *et al*. Multi-modal hashing for efficient multimedia retrieval: a survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(1): 239-260

[8] Qiang H P, Wan Y, Liu Z Y, *et al*. Discriminative deep asymmetric supervised hashing for cross-modal retrieval[J]. Knowledge-Based Systems, 2020, 204: Article No.106188

[9] Jiang Q Y, Li W J. Deep cross-modal hashing[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 3270-3278

[10] 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

[11] Shi Y F, Chen S H, You X G, *et al*. Deep supervised information bottleneck hashing for cross-modal retrieval based computer-aided diagnosis[OL]. [2024-07-03]. https://arxiv.org/abs/2205.08365

[12] Johnson A E W, Pollard T J, Berkowitz N R, *et al*. MIMIC-CXR: a large publicly available database of labeled chest radiographs[OL]. [2024-07-03]. https://arxiv.org/abs/1901.07042v2

[13] Duan Y X, Chen N, Zhang P Y, *et al*. MS^{2}GAH: multi-label semantic supervised graph attention hashing for robust cross-modal retrieval[J]. Pattern Recognition, 2022, 128: Article No.108676

[14] Li C, Deng C, Li N, *et al*. Self-supervised adversarial hashing networks for cross-modal retrieval[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 4242-4251

[15] Winegardner K, Scrivani P V, Gleed R D. Lung expansion in the diagnosis of lung disease[J]. Compendium, 2008, 30(9): 479-489

[16] Gu Y F, Yang X B, Sun M X, *et al*. Graph-guided deep hashing networks for similar patient retrieval[J]. Computers in Biology and Medicine, 2024, 169: Article No.107865

[17] Tu R C, Jiang J, Lin Q H, *et al*. Unsupervised cross-modal hashing with modality-interaction[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(9): 5296-5308

[18] Rasiwasia N, Pereira J C, Coviello E, *et al*. A new approach to cross-modal multimedia retrieval[C] //Proceedings of the 18th ACM International Conference on Multimedia. New York: ACM Press, 2010: 251-260

[19] Ding G G, Guo Y C, Zhou J L. Collective matrix factorization hashing for multimodal data[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 2083-2090

[20] Wang D, Gao X B, Wang X M, *et al*. Semantic topic multimodal hashing for cross-media retrieval[C] //Proceedings of the 24th International Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2015: 3890-3896

[21] Tang J, Wang K, Shao L. Supervised matrix factorization hashing for cross-modal retrieval[J]. IEEE Transactions on Image Processing, 2016, 25(7): 3157-3166

[22] Zhang D Q, Li W J. Large-scale supervised multimodal hashing with semantic correlation maximization[C] //Proceedings of the 28th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2014: 2177-2183

[23] Lin Z J, Ding G G, Hu M Q, *et al*. Semantics-preserving hashing for cross-view retrieval[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3864-3872

[24] Meng M, Wang H T, Yu J, *et al*. Asymmetric supervised consistent and specific hashing for cross-modal retrieval[J]. IEEE Transactions on Image Processing, 2021, 30: 986-1000

[25] Jiang Q Y, Li W J. Discrete latent factor model for cross-modal hashing[J]. IEEE Transactions on Image Processing, 2019, 28(7): 3490-3501

[26] Xie D, Deng C, Li C, *et al*. Multi-task consistency-preserving adversarial hashing for cross-modal retrieval[J]. IEEE Transactions on Image Processing, 2020, 29: 3626-3637

[27] Nie X S, Wang B W, Li J J, *et al*. Deep multiscale fusion hashing for cross-modal retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(1): 401-410

[28] 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

[29] Chatfield K, Simonyan K, Vedaldi A, *et al*. Return of the devil in the details: delving deep into convolutional nets[OL] [2024-07-03]. https://www.arxiv.org/abs/1405.3531

[30] Li Q M, Han Z C, Wu X M. Deeper insights into graph convolutional networks for semi-supervised learning[C] //Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 3538-3545

[31] Cao Z J, Long M S, Wang J M, *et al*. HashNet: deep learning to hash by continuation[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 5609-5618

[32] Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in Python[J]. The Journal of Machine Learning Research, 2011, 12: 2825-2830

[33] Paszke A, Gross S, Massa F, *et al*. PyTorch: an imperative style, high-performance deep learning library[C] //Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019: Article No.721

[34] Hollander M, Sethuraman J. Nonparametric statistics: rank-based methods[M] //Wright J D. International Encyclopedia of the Social & Behavioral Sciences. 2nd ed. Amsterdam: Elsevier, 2015: 891-897