引用格式：LI Haoran, XIONG Wei, CUI Yaqi, et al. Enhancing Remote Sensing Image Unsupervised Hashing Cross-modal Correlation with Similarity Matrix[J]. Acta Photonica Sinica, 2023, 52(1):0110003
李浩然,熊伟,崔亚奇,等. 相似度矩阵辅助遥感图像无监督哈希跨模态关联[J].光子学报,2023,52(1):0110003

相似度矩阵辅助遥感图像无监督哈希跨模态关联

李浩然,熊伟,崔亚奇,顾祥岐,徐平亮
(海军航空大学 信息融合研究所,烟台 264001)

摘　要:提出了一种使用相似度矩阵辅助遥感图像无监督哈希跨模态关联的方法,解决哈希码转化过程中造成的部分语义信息的损失问题。利用构建的原始特征以及哈希特征的相似度矩阵整合不同模态间的语义相关信息,以尽可能地保留模态内以及不同模态间语义的相关性,通过相似度矩阵间的语义对齐减小原始特征转换为哈希编码的特征信息损失,并结合对比学习的方法有效提高了遥感图像文本间无监督哈希跨模态关联效果。在两个公开数据集上的实验验证表明,所提方法优于现有基准方法,具有较好的性能。

关键词:遥感;无监督学习;跨模态检索;相似度矩阵;对比学习
中图分类号:TP751.1　　**文献标识码**:A　　**doi**:10.3788/gzxb20235201.0110003

0　引言

近年来,随着星载和机载等遥感探测手段的不断丰富,获取到的遥感数据类型更加多样、数据规模不断扩大,有力带动了遥感领域跨模态关联方法的发展。跨模态检索任务是指根据给定模态中的查询样本,然后从其他模态中检索出与其相关的数据,多模态数据通常包括图像、文本、视频、音频等[1]。而遥感图像和文本是情报信息的重要组成部分,遥感图像与文本信息之间关联关系的建立对多源情报数据的有效利用有着重要的意义,两者的相互印证有助于进一步提高获取情报信息的可靠性。

随着深度学习的不断发展,深度神经网络越来越广泛应用于获取不同模态的特征表示,实现将跨模态信息映射到同一特征空间中,有助于解决不同模态间的“异构鸿沟”问题。然而,现有的遥感图像文本跨模态关联方法多为基于实值表示学习,虽然关联效果较好但由于这类方法存储计算消耗大、效率低,故无法适应大规模快速检索任务的需求。

哈希方法实现检索速度快、效率高,随着当今遥感数据类型和规模的增长,在遥感跨模态领域受到越来越多的关注。近年来深度学习与哈希跨模态方法的结合,使得获取不同模态的特征表示更加高效,而且无监督的哈希跨模态关联方法无须标签信息就可实现关联,在解决跨模态关联问题上展现出更大的优势。但现有无监督深度哈希跨模态方法仍存在一些问题,通常分别对跨模态的相似度信息进行学习,而且没有标签信息辅助,会造成模型无法正确有效地获取不同模态之间的语义关联关系。此外,深度哈希方法大多通过深度神经网络获取的原始特征直接生成哈希码,生成的哈希特征难以获得令人满意的辨别性信息。

针对上述问题,本文提出了一种相似度矩阵辅助遥感图像无监督哈希跨模态关联方法,利用遥感图像和文本的原始特征和哈希特征分别构造对应的相似度矩阵,再通过矩阵损失函数的约束用原始特征构造的相似度矩阵来指导哈希特征相似度矩阵的生成,以捕获潜在的语义相关性尽可能保留与原始特征的语义相

基金项目:国家自然科学基金(Nos. 61790554, 62001499)
第一作者:李浩然,rizhaolihaoran@163.com
通讯作者:熊伟,xiongwei@csif.org.cn
收稿日期:2022-07-06;**录用日期**:2022-08-18

似性，减小生成哈希特征后的语义信息损失，并与无监督对比学习的方法相结合，增强学习特征表示的判别性，进一步了提高无监督哈希跨模态关联检索的性能。

1 相关工作

1.1 遥感图像跨模态关联

随着大规模多模态数据的不断丰富，跨模态关联检索方法在遥感领域的发展得到了更多关注[2]。文献[3]提出了一种用于遥感数据的新的跨模态信息关联检索模型，模型重于从不同的输入模式中学习统一的、有区别的嵌入空间，并可以用于单模态和跨模态信息检索场景。MAO Guo等[4]提出了一种视觉-语音关联学习网络，并构建了用于图像和语音关联的数据集，验证了遥感图像与语音数据之间关联关系构建的可能性。文献[5]基于不同模态信息间潜在的语义一致性，提出了一种通用的跨模态遥感信息关联学习方法，通过共同空间的构建实现了多种模态数据的相互检索。也有很多学者开始关注遥感图像与文本之间的关联问题，文献[6]设计了语义对齐模块，通过利用注意机制以增强遥感图像和文本之间的对应关系，然后通过设计的门函数过滤尽可能多的不必要信息。针对遥感多模态检索任务中的多尺度性和目标冗余问题，文献[7]设计了一种非对称多模态特征匹配网络实现跨模态遥感图像检索。文献[8]一种基于融合的关联学习模型用于遥感图像与文本间的跨模态检索，通过跨模态融合网络的设计提高跨模态信息之间的语义相关度。现有的遥感图像跨模态检索方法可以根据特征表示学习的方法不同分为基于实值表示和二进制表示两大类[9]，而上述模型大多为基于实值表示学习的跨模态关联方法。

1.2 哈希跨模态关联方法

哈希跨模态关联方法又可分为有监督和无监督两种，在计算机视觉领域对跨模态哈希关联方法的研究更为广泛，文献[10]提出了一种端到端深度跨模态哈希方法，将特征学习和哈希码学习集成到同一框架中。文献[11]提出了一种用于跨模态检索的多任务一致性保持对抗性哈希方法，通过利用标签信息学习一致的不同模态特征表示，用对抗性学习策略强化跨模态信息的语义一致性。文献[12]将矩阵分解和拉普拉斯约束结合到网络训练中，显式约束哈希码以保持原始数据的邻域结构，优化特征学习和二值化过程。文献[13]基于不同模态的邻域信息构建联合语义相似度矩阵，该矩阵同时集成了多模态相似信息，提出了跨模态检索的深度联合语义重构哈希法。文献[14]通过构造联合模态相似矩阵和基于分布的相似性决策和加权方法，充分保留了实例间的跨模态语义关联信息。文献[15]提出了一种用于无监督跨模态检索的多路径生成对抗哈希方法，该方法充分利用生成对抗网络GAN的无监督表示学习能力捕获跨模态数据的底层流形结构，实现多种模态信息关联。由于不需要人工标注节省了大量人力物力，无监督方法越来越受到更多地关注，上述除了文献[10]和文献[11]为有监督哈希方法外，其他方法都为无监督方法。跨模态哈希方法在遥感领域的研究相对较少，文献[16]研究了基于哈希网络的SAR与光学图像之间的遥感跨模态检索，通过引入图像转换的策略丰富了图像信息的多样性。文献[17]提出了一种新的无监督对比哈希算法用于解决遥感图像与文本间的跨模态关联问题，算法主要通过利用设计的一个多目标损失函数来进行无监督的跨模态表示学习。但无监督哈希方法在实值的二值化过程中会导致部分语义信息的损失以及原有结构被破坏[18]，而且没有充分考虑模态内数据结构和模态间邻域结构的匹配关联，因此对不同模态间语义一致性的挖掘及优化计算等仍是目前研究的一个重要方向。

2 模型方法

本文所提方法的总体框架如图1所示，分为处理遥感图像和文本数据的两个结构相类似的网络模型分支。对于各模态信息的处理，首先通过对应模态的特征提取深度神经网络模型（ImgNet、TxtNet）获得遥感图像和文本的原始特征表示（Original feature），再通过哈希模块（Hash module）学习得到相应的哈希特征表示（Hash feature）。为了在哈希方法学习过程中尽可能保持不同模态信息间的语义相关性，分别构造了原始特征和哈希特征的相似度矩阵，通过原始特征相似度矩阵指导哈希特征相似度矩阵的学习，减小哈希学习过程的语义信息损失，并与对比学习相结合以进一步挖掘不同模态信息间潜在的语义相关性。模型整体主要通过相似度矩阵损失（Similarity matrix loss）以及对比损失（Contrastive loss）的约束进行关联关系的学习，

以进一步提高最终哈希特征模态间的语义相关性，实现对遥感图像文本跨模态信息之间关联关系的无监督学习。

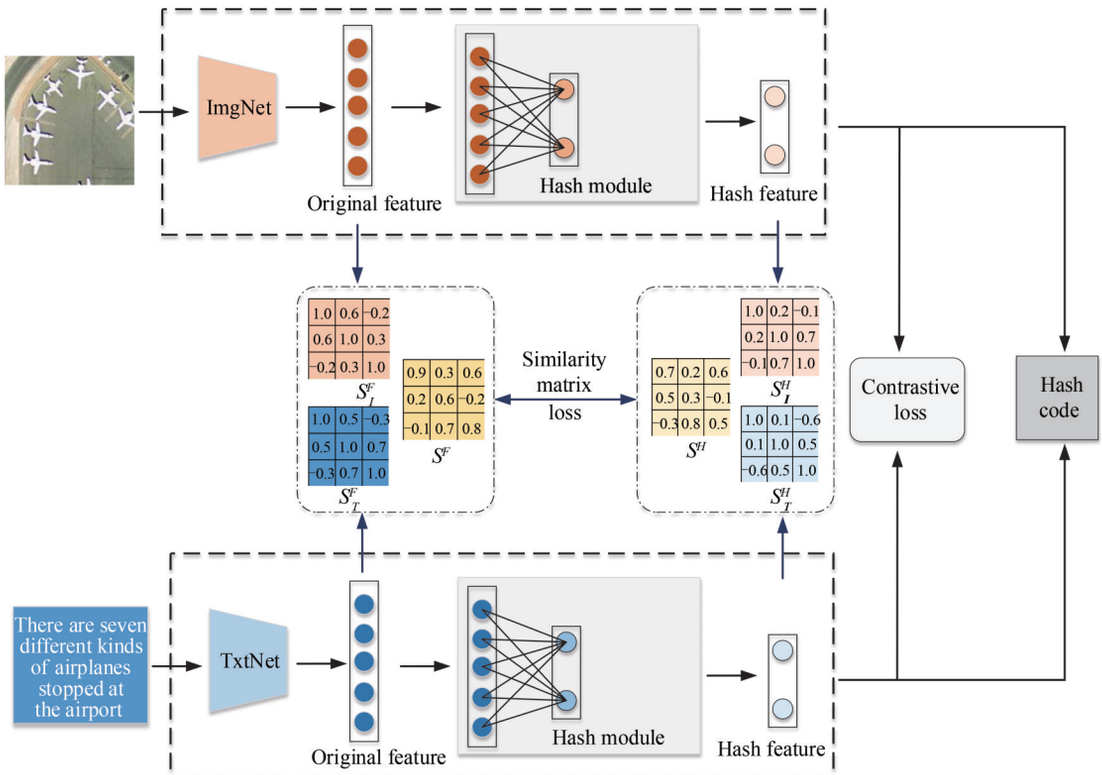


图 1　本文方法框架

Fig. 1　The structure of the proposed model

2.1　特征提取

各模态特征提取过程首先通过深度神经网络分别获取遥感图像和文本信息的原始特征表示，然后再进一步输入到哈希模块中学习相应的哈希特征表示。在本文提出的基准方法中，遥感图像的特征表示通过使用卷积神经网络模型 Resnet-18[19]提取，最后经全连接层得到输出维度与文本特征表示相同的遥感图像特征表示向量，文本的特征表示采用自然语言处理领域的 BERT[20]模型，将输出的最后四层隐藏状态求和得到最终的文本特征表示，特征向量的维度为 768 维。在哈希网络模型的学习训练阶段，遥感图像和文本特征提取网络模型的权重保持冻结，而且用于这两种模态的特征提取模型可以灵活替换。得到跨模态原始特征表示进一步输入到哈希模块中进行哈希特征的学习，通常方法的转化过程会使用符号函数将特征转换为“−1”和“+1”的形式，但这会造成神经网络模型在反向传播过程出现梯度消失的问题，导致模型无法训练。因此在训练过程中，本文方法在哈希模块由两层全连接层构成并采用反正切函数作为最后一层的激活函数[21]，使得最后输出类哈希编码的形式的哈希特征表示。哈希模块的目的是通过哈希函数的学习从遥感图像和文本原始语义特征信息生成准确的哈希特征表示，在这过程中不同模态中语义信息相似的实例能够表示成相似的哈希码。

2.2　相似度矩阵构造

无监督跨模态哈希方法无法通过标签获得不同模态间的关联信息，但从深度神经网络中提取的特征中包含着丰富的语义信息。文献[12,22]已经证明，学习保留原始特征数据邻域结构的二进制码能够有效改进深度哈希网络的无监督训练，哈希特征相似度矩阵反映了哈希码在汉明空间中的邻域结构。所以本文分别利用原始特征表示和哈希特征表示构造相似度矩阵，通过原始特征之间的相似度与哈希特征之间的相似度的语义对齐来辅助增强不同模态语义信息间的相关性，提高无监督哈希方法的关联效果。

对于每个批次样本，遥感图像和文本对分别经 ImgNet、TxtNet 获得相应的原始特征表示，进一步经正

则化处理后表示为$\boldsymbol{I},\boldsymbol{T}$,其中遥感图像特征表示$\boldsymbol{I}=\{\boldsymbol{v}_i\}_{i=1}^{N},\boldsymbol{v}_i\in\mathbb{R}^{d_v}$文本特征表示$\boldsymbol{T}=\{\boldsymbol{t}_i\}_{i=1}^{N},\boldsymbol{t}_i\in\mathbb{R}^{d_t}$。然后采用计算余弦相似度的方法来构造这两种模态特征表示模态内以及模态间的相似度矩阵,用于描述遥感图像与文本的原始邻域结构信息及跨模态间的语义关系。遥感图像模态内的相似度矩阵表示为$\boldsymbol{S}_I^F=\left\{s_{I_{ij}}^F\right\}_{i,j=1}^{N}$,文本模态的相似度矩阵为$\boldsymbol{S}_T^F=\left\{s_{T_{ij}}^F\right\}_{i,j=1}^{N}$,两种模态间的相似度矩阵为$\boldsymbol{S}^F=\left\{s_{ij}^F\right\}_{i,j=1}^{N}$,其中$s_{I_{ij}}^F$,$s_{T_{ij}}^F$,$s_{ij}^F$的定义分别为

$$s_{I_{ij}}^F=\cos\left(\boldsymbol{I}_i,\boldsymbol{I}_j\right)=\frac{\boldsymbol{I}_i\boldsymbol{I}_j^{\mathrm{T}}}{\left\|\boldsymbol{I}_i\right\|_2\left\|\boldsymbol{I}_j\right\|_2},\ s_{T_{ij}}^F=\cos\left(\boldsymbol{T}_i,\boldsymbol{T}_j\right)=\frac{\boldsymbol{T}_i\boldsymbol{T}_j^{\mathrm{T}}}{\left\|\boldsymbol{T}_i\right\|_2\left\|\boldsymbol{T}_j\right\|_2},\ s_{ij}^F=\cos\left(\boldsymbol{I}_i,\boldsymbol{T}_j\right)=\frac{\boldsymbol{I}_i\boldsymbol{T}_j^{\mathrm{T}}}{\left\|\boldsymbol{I}_i\right\|_2\left\|\boldsymbol{T}_j\right\|_2}\tag{1}$$

不同模态的相似度矩阵通常互为补充,通过将遥感图像相似矩阵$\boldsymbol{S}_I^F$、文本相似矩阵$\boldsymbol{S}_T^F$以及两模态间的相似度矩阵$\boldsymbol{S}^F$融合成一个模态间联合相似矩阵,以获得对不同模态实例之间语义关系的准确描述。表示为

$$\boldsymbol{S}=\alpha\boldsymbol{S}_I^F+\beta\boldsymbol{S}_T^F+\gamma\boldsymbol{S}^F\tag{2}$$

式中,α,β,γ为权衡参数,$\alpha+\beta+\gamma=1$,$\alpha,\beta,\gamma\geqslant0$,用于调节不同模态邻域关系的重要性。

同样,对于遥感图像与文本经哈希模块得到的哈希特征,采用与原始特征表示构建相似度矩阵同样的计算方式,可以获得对应模态内及模态间哈希特征相似度矩阵$\boldsymbol{S}_I^H,\boldsymbol{S}_T^H,\boldsymbol{S}^H$,能够描述不同模态信息哈希特征间的相关关系邻域结构。如模态间的哈希特征相似度矩阵表示为$\boldsymbol{S}^H=\cos(\boldsymbol{H}_I,\boldsymbol{H}_T)$,其中的元素由遥感图像$\boldsymbol{i}$和文本$\boldsymbol{j}$对应的哈希特征计算

$$s_{ij}^H=\cos(\boldsymbol{H}_I,\boldsymbol{H}_T)_{ij}=\frac{\boldsymbol{H}_{Ii}\boldsymbol{H}_{Tj}^{\mathrm{T}}}{\left\|\boldsymbol{H}_{Ii}\right\|_2\left\|\boldsymbol{H}_{Tj}\right\|_2}\tag{3}$$

通常二进制哈希码对应的语义描述往往会偏离特征的语义描述,导致模型效果的下降。通过构造的联合模态相似矩阵来指导哈希相似度矩阵的生成,学习原始数据邻域结构,使得到哈希特征具有较高的实例间原始语义相关度,最终的二值哈希码能够尽可能保留更多地语义信息,可有效改进深度哈希网络模型的无监督训练[12]。

2.3 损失函数

为了有效地进行无监督跨模态哈希学习,本文设计了一个新的损失函数组合,通过对比损失与相似度矩阵损失的相互结合,增强哈希表示的判别性,提高了模型关联检索的准确性。

对比损失可以有效提高无监督表示学习能力,在本文模型中,采用归一化温度尺度交叉熵目标函数[23-24]进行对比损失的计算。在对比损失中主要考虑遥感图像文本对模态间的对比损失,匹配图文对的特征表示在共同特征空间的距离尽可能近,不匹配的图文对距离尽可能远离,通过使用对比损失来使匹配的遥感图像和文本之间的语义信息对齐。模态间对比学习能够使遥感图像和文本之间的互信息最大化,进一步发掘跨模态信息潜在的关联关系,使模型学习到的特征更具判别性增强模型表征学习能力。模态间的对比损失可定义为

$$L_{\mathrm{c}}=-\log\frac{S\left(f\left(x_j\right),g\left(y_j\right)\right)}{\sum_{k=1}^{M}S\left(f\left(x_j\right),g\left(y_k\right)\right)}\tag{4}$$

式中,$S(\boldsymbol{u},\boldsymbol{v})=\exp\left(\cos(\boldsymbol{u},\boldsymbol{v})/\tau\right)$,$\cos(\boldsymbol{u},\boldsymbol{v})=\frac{\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}$为余弦相似度,$\tau$为温度系数,$M$为批次大小。

相似度矩阵损失通过利用构造的相似度矩阵,同时考虑模态内以及模态间哈希特征的相似度信息与原始特征的相似度信息语义对齐。通过最小化原始特征与哈希特征相似度矩阵之间的重构误差来学习原始特征的邻域结构,发掘模态内及模态间潜在的语义相关性,以弥补转化为哈希特征后语义信息的不足。本文所提方法的相似度矩阵损失设计为

$$L_{\mathrm{m}}=L_{\mathrm{inter}}+L_{\mathrm{intra}}\tag{5}$$

$$L_{\mathrm{inter}}=\left\|\eta\boldsymbol{S}^F-\boldsymbol{S}^H\right\|_F^2+\sum_{i=1}^{N}\left\|\eta\cdot1-\mathrm{diag}\left(s_{ii}^H\right)\right\|^2\tag{6}$$

$$L_{\text{intra}}=\frac{1}{2}\left(\left\|\boldsymbol{S}_I^F-\boldsymbol{S}_I^H\right\|_F^2+\left\|\boldsymbol{S}_T^F-\boldsymbol{S}_T^H\right\|_F^2\right)+\frac{1}{2}\left(\left\|\eta\boldsymbol{S}^F-\boldsymbol{S}_I^H\right\|_F^2+\left\|\eta\boldsymbol{S}^F-\boldsymbol{S}_T^H\right\|_F^2\right) \tag{7}$$

式中，η为权衡参数，其设置能够使相似度矩阵间的语义对齐更加灵活。通过模态内及模态间相似度矩阵的语义对齐，尽可能学习保留遥感图像文本原始特征的邻域结构关系，充分挖掘潜在语义相关信息，更好地建立不同模态间的关联关系。

最终的总体损失函数是对比损失函数和相似度矩阵损失的加权和为

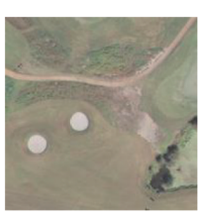
$$L=\lambda L_c+\mu L_m \tag{8}$$

式中，λ，μ为平衡两损失之间关系的超参数。

3 实验结果与分析

3.1 数据集与实验设置

在本文实验中，使用UCM数据集[25]和RSICD数据集[26]，RSICD数据集包含31类共10 921幅遥感影像，每幅影像包含5句对应的文本描述，UCM数据集包含2 100张21类遥感图像，每张图像同样有5句相关的文本描述。两个数据集中遥感图像及对应文本描述的部分样例如图2中所示。在训练过程中，每张图片只使用一个随机选择的文本描述，输入图像的大小为224×224，遥感图像的特征提取采用预训练的Resnet-18，文本描述的特征提取使用预训练模型‘bert-base-uncased’提取。批大小设置为256，学习率设为0.000 3，共训练100次，训练时采用Adam优化器。损失函数中超参数分别设置为$\alpha=0.25$，$\beta=0.25$，$\gamma=0.5$，$\eta=1.5$，$\lambda=0.001$，$\mu=0.1$。实验在PyTorch框架下进行编译，模型在搭载一块GeForce RTX 2080Ti GPU的工作站上运行。



· This is a part of a golf course with green turfs and some bunkers and trees.
· A part of a golf course with some bunkers and trees while a trail goes through the turfs.
· A part of a golf course with a trail goes through the turfs and some bunkers and trees.
· Some bunkers and trees with a trail goes through the turfs in the golf course.
· Some green bunkers and trees with a trail goes through the turfs in the golf course.

(a) UCM

· A football field with several buildings surrouded.
· A rectangular playground and many tall buildings surrounded.
· Many buildings and green trees are around a playground.
· Many buildings are in different blocks with many green trees and a playground.
· A playground is surrounded by many trees and buildings.

(b) RSICD

图2 数据集中遥感图像及对应的文本描述

Fig. 2 Remote sensing image and corresponding captions from the datasets

为有效评估所提出的方法，在实验中进行了遥感图像检索文本以及文本检索的关联检索任务，采用均值平均准确率mAP作为模型的评价指标，mAP是评价模型关联检索性能的常用指标，其定义为

$$\text{mAP}=\frac{1}{N}\sum_{i=1}^{\text{N}}\frac{1}{r_i}\sum_{j=1}^{k}P_i(j)\times\text{rel}_i(j) \tag{9}$$

式中，N是查询集的大小，r_i是与查询样本i相关的项数，k是数据库中的样本数。

在本文实验中，采用前20个检索到的样本的均值平均准确率（表示为mAP@20）作为模型的评价指标，Top-k的准确率$P(k)$是通过按与查询样本的汉明距离排序的返回前k个样本的准确率计算，定义为

$$P(k)=\frac{\sum_{i=1}^{k}\text{rel}(i)}{k} \tag{10}$$

式中，rel(i)是一个样本相关性指示符，如果查询和检索到的样本相匹配，则该指示符等于1，否则为0。汉明距离是由两个二进制码中不同位的个数定义的，汉明排序用于哈希值的排序，它根据查询和检索样本之间

的汉明距离对检索到的样本进行排序。这里$P(k)$较大的值对应更好的检索结果。

3.2 对比实验结果

为了验证所提方法的有效性，在数据集RSICD和UCM上将其与文献[17,28]中的方法DUCH以及几种不同的跨模态哈希基准方法进行了对比，分别为CPAH[11]、DJSRH[13]、JDSH[14]。为了保证对比的公平，对比实验数据划分与文献[17]保持一致，数据集通过随机选择分为训练集、查询集和检索集（分别为50%、10%和40%），在相同的实验设置下训练模型。表1和表2分别展示了本文方法在RSICD和UCMerced数据集上使用mAP@20评估图像到文本和文本到图像检索两种任务与上述基准方法的对比结果，实验对比了四种不同哈希码长度B=16,32,64,128。

表1 RSICD数据集上不同方法的mAP@20比较

Table 1 The mAP@20 comparison of different methods on the RSICD dataset

Method	Image to Text (I→T)				Text to Image (T→I)			
	B=16	B=32	B=64	B=128	B=16	B=32	B=64	B=128
CPAH	0.428	0.587	0.636	0.696	0.452	0.598	0.667	0.706
DJSRH	0.411	0.665	0.688	0.722	0.422	0.685	0.705	0.733
JDSH	0.385	0.720	0.796	0.815	0.418	0.751	0.799	0.815
DUCH	0.684	0.791	0.836	0.829	0.697	0.780	0.824	0.826
Proposed	0.708	0.802	0.823	0.832	0.736	0.818	0.845	0.850

表2 UCM数据集上不同方法的mAP@20比较

Table 2 The mAP@20 comparison of different methods on the UCM dataset

Method	Image to Text (I→T)				Text to Image (T→I)			
	B=16	B=32	B=64	B=128	B=16	B=32	B=64	B=128
CPAH	0.706	0.802	0.891	0.914	0.782	0.891	0.987	0.982
DJSRH	0.686	0.711	0.735	0.754	0.738	0.755	0.776	0.800
JDSH	0.462	0.751	0.82	0.829	0.509	0.794	0.884	0.904
DUCH	0.760	0.794	0.844	0.870	0.799	0.851	0.916	0.927
Proposed	0.789	0.816	0.841	0.860	0.848	0.894	0.926	0.951

通常哈希码位数越多模型效果越好，因为其可以存储的语义信息更丰富。在RSICD数据集上，除在B=64时图像检索文本任务的mAP@20低于基准方法的DUCH方法，在其他情况下均优于其他对比算法，而且在哈希位数少的情况下模型的优势更明显，例如当B=16时与DUCH方法相比，在两种检索任务上mAP@20分别提升了2.4%、3.9%，而当B=128时对应提升分别为0.3%、2.4%。在UCM数据集上，除了当B=64,128时图像检索文本任务的mAP@20低于DUCH方法外，在其他情况下本文方法在不同哈希码长度下的两种检索任务上与无监督的方法DJSRH，JDRH，DUCH相比均取得了最好的mAP@20分数，而且优势相对比较明显。而与有监督方法CPAH相比，在哈希吗位数为16和32时本文方法优于该方法，尤其在哈希码长度为16时优势更为明显，但在哈希码位数为64和128时与方法CPAH还有一定差距。分析两个表中数据可以得出，在哈希码位数较少时本文方法的优势相对更明显，这更好地说明的本文方法的有效性，哈希码位数较少时其能够存储的信息有限，因此在哈希码转化过程中位数较少时的语义信息损失可能会更严重，哈希码位数较多时其本身可存储的语义信息更加丰富，所以模型对其提升效果相对有限。而本文模型的设计能够使得到的哈希特征中尽可能保留更多的语义信息，减少原始特征转化为哈希特征时的语义损失，使得生成的哈希码中可保留更具辨别性的语义信息，且在哈希码位数少时更有效，使得关联结果更加准确，更适于大规模遥感图像文本间的准确关联检索任务。

对比算法文献中仅把数据集的50%用作模型训练学习可能不够充分，为充分学习遥感图像文本间的关联关系并进一步检验本文模型的有效性，我们对数据集划分比例进行了优化。进行优化后数据集训练集与测试集的比例为7:3，并在相同的实验条件及划分下与遥感领域最新提出无监督跨模态哈希关联检索方法DUCH进行了对比，实验结果对比如表3所示。此外，我们将遥感图像原始特征提取采用的预训练Resnet-18

替换为预训练的vit[27]后进行模型效果的对比，一方面能够验证模型的灵活性及有效性，另一方面也可进一步探索遥感图像特征提取网络模型对最终关联检索效果的影响。

表3　方法优化后实验结果对比（mAP@20）

Table 3　The mAP@20 comparison of different methods after optimization

Dataset	Method	Image to Text (I→T)				Text to Image (T→I)			
		B=16	B=32	B=64	B=128	B=16	B=32	B=64	B=128
RSICD	DUCH	0.698	0.795	0.838	0.838	0.732	0.825	0.85	0.856
	Proposed	0.748	0.815	0.839	0.85	0.796	0.845	0.868	0.872
	Proposed(vit)	0.805	0.875	0.893	0.907	0.819	0.882	0.892	0.891
UCM	DUCH	0.768	0.824	0.896	0.910	0.802	0.862	0.938	0.953
	Proposed	0.831	0.871	0.899	0.912	0.873	0.922	0.942	0.957
	Proposed(vit)	0.905	0.915	0.933	0.939	0.923	0.947	0.961	0.969

从表3可以看出，当仅对数据集划分比例优化后模型对各模态特征表示的学习更充分，而本文方法与DUCH方法的效果对比提升也更加明显，且在两个数据集上不同任务下的mAP@20指标均优于DUCH方法，说明了改变后模型学习到的语义信息更加丰富，也进一步验证了本文方法的有效性。本文方法同样在B=16和32时的效果提升最为显著，再次说明本文所提方法能更好地保留原始语义相关性，能够在哈希学习过程中减小语义损失，实现更准确的关联检索。此外，当遥感图像特征提取模型采用预训练的vit时，所提方法在两种任务的各个指标上都会有更进一步的提高，模型整体性能都有进一步提升，说明了原始特征提取模块对模型效果同样有重要影响，同时也验证了本文模型的灵活性及可扩展性，可根据需要设计替换相关模块，使模型在应用时能够灵活调整。

3.3　消融实验

为了验证本文模型中各损失函数设置发挥的作用，进行消融实验对不同损失函数的有效性进行检验。我们在不同哈希码长度下都进行了模型简化实验，以充分检验所提模型的整体性能，在本文中两种数据集划分比例上的实验结果如表4和表5中所示，其中，“L_m+L_c”表示本文的完整方法，采用对比损失与相似度矩阵损失的加权组合（$\lambda=0.001$，$\mu=0.1$）；“L_m”表示模型仅使用相似度矩阵损失（$\lambda=0$，$\mu=1$）；“L_c”模型只采用对比损失（$\lambda=1$，$\mu=0$）。

表4　消融实验结果（50% 训练）

Table 4　Ablation experiment results (50% training)

Dataset	Method	Image to Text (I→T)				Text to Image (T→I)			
		B=16	B=32	B=64	B=128	B=16	B=32	B=64	B=128
RSICD	L_m+L_c	0.708	0.802	0.823	0.832	0.736	0.818	0.845	0.850
	L_m	0.632	0.770	0.817	0.834	0.656	0.785	0.824	0.848
	L_c	0.703	0.787	0.818	0.828	0.710	0.796	0.820	0.833
UCM	L_m+L_c	0.789	0.816	0.841	0.86	0.848	0.894	0.926	0.951
	L_m	0.747	0.817	0.851	0.859	0.813	0.877	0.934	0.943
	L_c	0.760	0.808	0.842	0.863	0.823	0.890	0.935	0.948

分析模型的消融实验结果可以看出，在仅使用一种损失的情况下，虽然在部分任务中的指标可以达到较好的效果，但由于部分语义信息的损失造成模型的整体性能并不理想。同样这在哈希码长度较小时表现得更为突出，而本文模型中通过相似度矩阵损失与对比损失的加权结合，使得模型在哈希学习过程中能够保留更多模态内及模态间的语义相关信息，且能够使学习的哈希特征更判别性，提高了模型整体关联检索的性能，实验验证表明本文所提方法是有效的。为更直观的对比分析模型简化实验，将表4和表5中的实验结果分别绘制在图3和图4中。

表 5 消融实验结果(70% 训练)

Table 5 Ablation experiment results (70% training)

Dataset	Method	Image to Text (I→T)				Text to Image (T→I)			
		B=16	B=32	B=64	B=128	B=16	B=32	B=64	B=128
RSICD	L_m+L_c	0.748	0.815	0.839	0.85	0.796	0.845	0.868	0.872
	L_m	0.682	0.802	0.842	0.851	0.692	0.820	0.862	0.869
	L_c	0.741	0.812	0.836	0.849	0.770	0.828	0.851	0.867
UCM	L_m+L_c	0.831	0.871	0.899	0.912	0.873	0.922	0.942	0.957
	L_m	0.817	0.859	0.869	0.879	0.841	0.898	0.913	0.930
	L_c	0.835	0.867	0.885	0.898	0.869	0.917	0.940	0.952

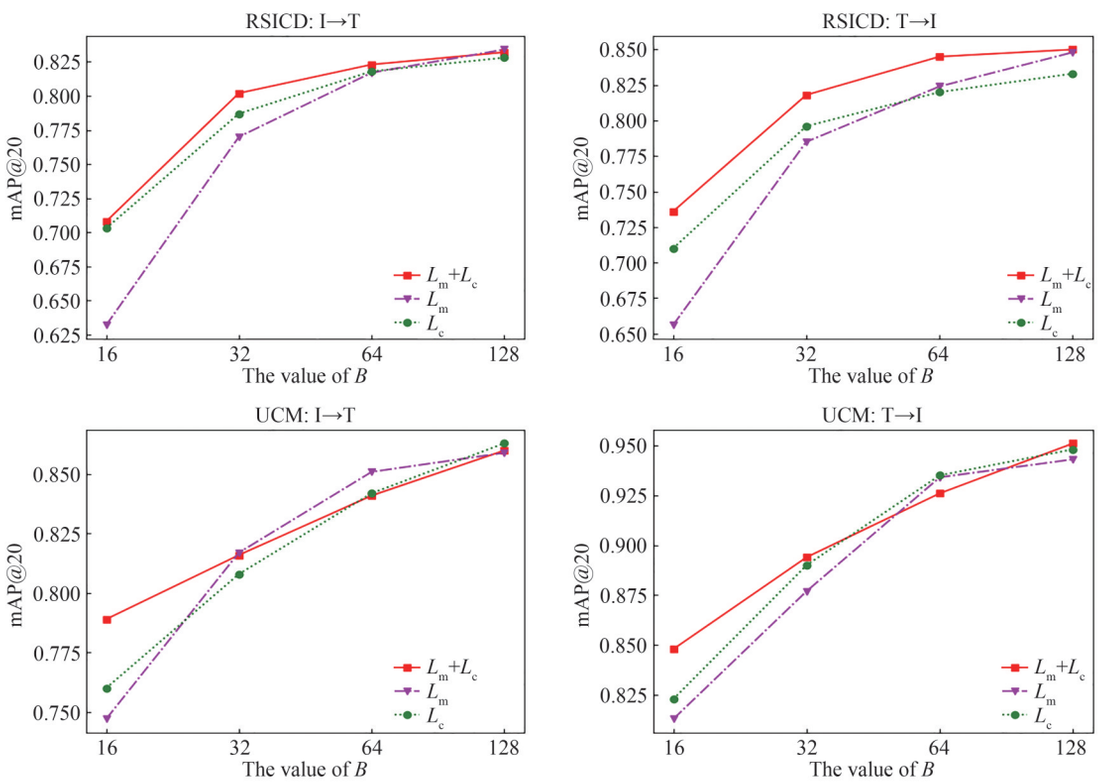
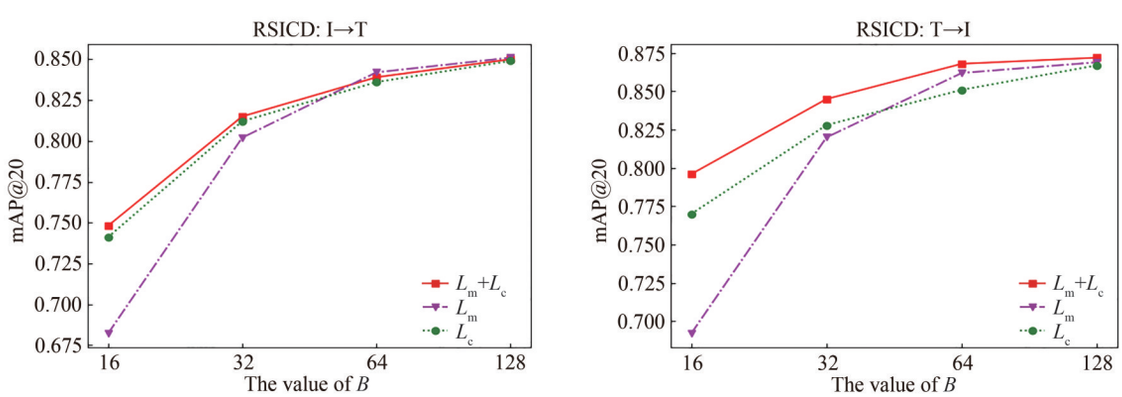


图 3 消融实验结果(50% 训练)

Fig. 3 Ablation experiment results (50% training)



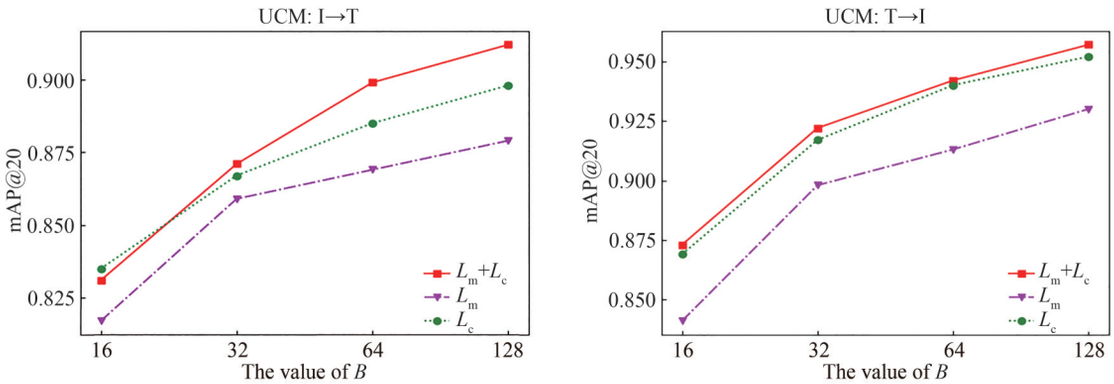


图4　消融实验结果(70%训练)

Fig. 4　Ablation experiment results (70% training)

3.4　参数分析实验

本小节主要对损失函数(式(8))中的两个平衡参数λ和μ进行分析，以探究不同参数设置对算法性能的影响。我们在数据集RSICD上(50%训练)分析两个超参数对关联效果的影响，先将μ的值固定为0.1，λ的取值分别为0.1，0.01，0.001，0.000 1时，在不同哈希码长度得到的实验结果如图5所示，从结果可以看出当$\lambda=0.001$时模型的效果最好。同样，将λ的值固定为0.001，μ的取值分别为1，0.1，0.01，0.001时，在不同哈希码长度得到的实验结果如图中所示，可以看出当$\mu=0.1$时模型的性能达到最佳。因此，在$\mu=0.1$且$\lambda=0.001$时，两种损失能够更好地结合，使得本文方法达到最优的关联效果。

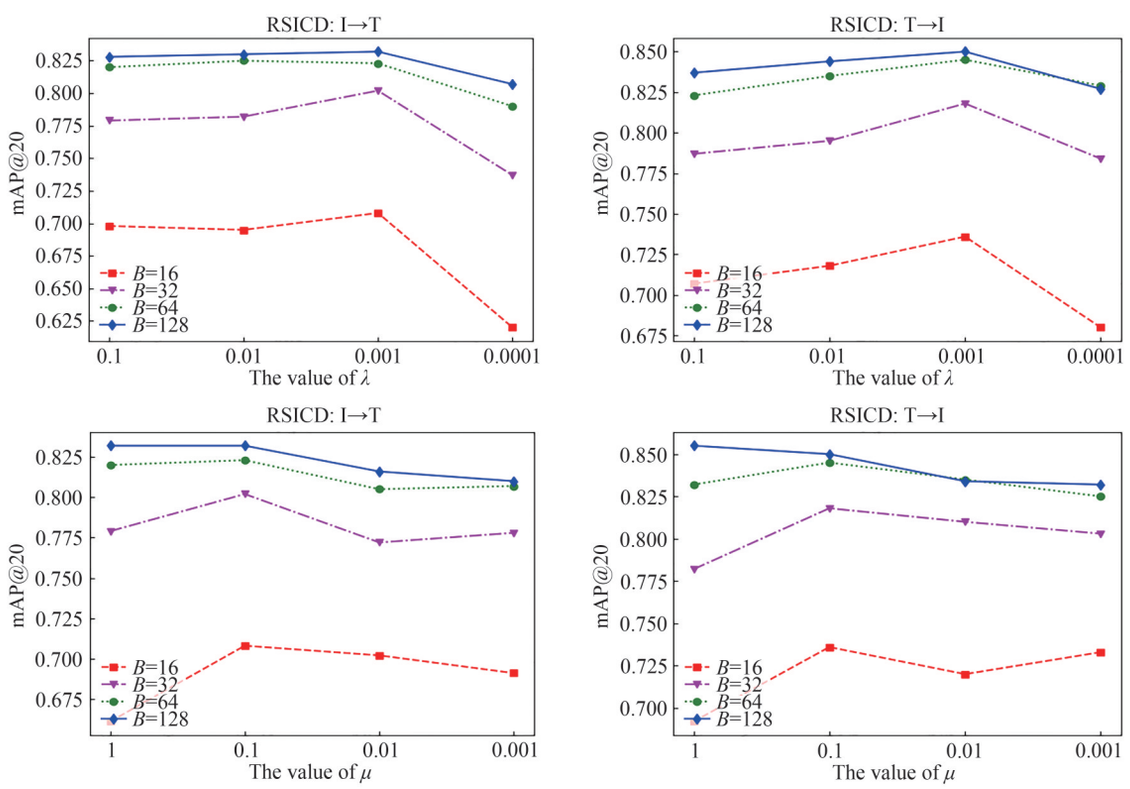


图5　不同参数设置下的模型性能(mAP@20)

Fig. 5　Model performance under different parameter settings (mAP@20)

4　结论

本文提出了一种相似度矩阵辅助遥感图像无监督哈希跨模态关联方法，通过构建相似度矩阵损失来对齐哈希特征与原始特征的语义信息，并与对比学习方法相结合，以增强最终哈希特征的语义相关性，减小哈

希码转化过程的语义信息损失。损失函数采用相似度矩阵损失与对比损失的加权和,两者的结合有效提高了无监督跨模态哈希关联的准确性,更适用于大规模跨模态遥感图像关联检索任务,而且在遥感领域的两个基准数据集上验证了所提方法的有效性。但模型原始特征提取模块的设计对各模态的语义信息丰富性考虑不够充分,相似度矩阵的计算方式相对较简单,将来工作可以作进一步改进以提高关联的准确性。

参考文献

[1] PENG Yuxin, HUANG Xin, ZHAO Yunzhen. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(9): 2372-2385.

[2] LI Yansheng, MA Jiayi, ZHANG Yongjun. Image retrieval from remote sensing big data: a survey[J]. Information Fusion, 2021, 67: 94-115.

[3] CHAUDHURI U, BANERJEE B, BHATTACHARYA A, et al. CMIR-NET: a deep learning based model for cross-modal retrieval in remote sensing[J]. Pattern Recognition Letters, 2020, 131(2): 456-462.

[4] MAO Guo, YUAN Yuan, LU Xiaoqiang. Deep cross-modal retrieval for remote sensing image and audio[C]. 2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS), IEEE, 2018.

[5] LV Yafei, XIONG Wei, ZHANG Xiaohan. A general cross-modal correlation learning method for remote sensing[J]. Geomatics and Information Science of Wuhan University, 2022, 47(11):1887-1895.
吕亚飞, 熊伟, 张筱晗. 一种通用的跨模态遥感信息关联学习方法[J]. 武汉大学学报(信息科学版), 2022, 47(11): 1887-1895.

[6] CHENG Qimin, ZHOU Yuzhuo, FU Peng, et al. A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 4284-4297.

[7] YUAN Zhiqiang, ZHANG Wenkai, FU Kun, et al. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60:4404119.

[8] LV Yafei, XIONG Wei, ZHANG Xiaohan, et al. Fusion-based correlation learning model for cross-modal remote sensing image retrieval[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 6503205.

[9] WANG Kaiye, YIN Qiyue, WANG Wei, et al. A comprehensive survey on cross-modal retrieval[J/OL].(2016-07-21). https://arxiv.org/abs/1607.06215.

[10] JIANG Qingyuan, LI Wujun. Deep cross-modal hashing[C]. IEEE Conference on Computer Vision & Pattern Recognition, 2017:3270-3278.

[11] XIE De, DENG Cheng, LI Chao, et al. Multi-task consistency-preserving adversarial hashing for cross-modal retrieval[J]. IEEE Transactions on Image Processing, 2020, 29:3626-3637.

[12] WU Gengshen, LIN Zijia, HAN Jungong, et al. Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval[C]. Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018.

[13] SU Shupeng, ZHONG Zhisheng, ZHANG Chao. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[14] LIU Song, QIAN Shengsheng, GUAN Yang, et al. Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval[C]. ACM, 2020.

[15] ZHANG Jian, PENG Yuxin. Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval[J]. IEEE Transactions on Multimedia, 2020, 22(1):174-187.

[16] XIONG Wei, XIONG Zhenyu, ZHANG Yang, et al. A deep cross-modality hashing network for SAR and optical remote sensing images retrieval[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 5284-5296.

[17] MIKRIUKOV G, RAVANBAKHSH M, DEMIR B. Unsupervised contrastive hashing for cross-modal retrieval in remote sensing[C]. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: 4463-4467.

[18] OU Wenhua, LIU Bin, ZHOU Yonghui, et al. Survey on the crossmodal retrieval research[J]. Journal of Guizhou Normal University (Natural Sciences), 2018, 36(2): 114-120.
欧卫华, 刘彬, 周永辉, 等. 跨模态检索研究综述[J]. 贵州师范大学学报(自然科学版), 2018, 36(2): 114-120.

[19] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[20] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J/OL].(2018-10-11). https://arxiv.org/abs/1810.04805.

[21] CAO Zhangjie, LONG Mingsheng, WANG Jianmin, et al. Hashnet: deep learning to hash by continuation[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 5608-5617.

[22] SHEN Fumin, XU Yan, LIU Li, et al. Unsupervised deep hashing with similarity-adaptive and discrete optimization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 3034-3044.
[23] CHEN Ting, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]. International Conference on Machine Learning, 2020: 1597-1607.
[24] ZHANG Han, JingyuKOH, BALDRIDGE J, et al. Cross-modal contrastive learning for text-to-image generation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition, 2021: 833-842.
[25] QU Bo, LI Xuelong, TAO Dacheng, et al. Deep semantic understanding of high resolution remote sensing image[C]. 2016 International Conference on Computer, Information and Telecommunication Systems (Cits), 2016: 1-5.
[26] LU Xiaoqiang, WANG Binqiang, ZHENG Xiangtao, et al. Exploring models and data for remote sensing image caption generation[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 56(4): 2183-2195.
[27] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J/OL].(2020-10-22). https://arxiv.org/abs/2010.11929.
[28] MIKRIUKOV G, RAVANBAKHSH M, DEMIR B. Deep unsupervised contrastive hashing for large-scale cross-modal text-image retrieval in remote sensing[J/OL].(2022-01-20). https://arxiv.org/abs/2201.08125v1.

Enhancing Remote Sensing Image Unsupervised Hashing Cross-modal Correlation with Similarity Matrix

LI Haoran, XIONG Wei, CUI Yaqi, GU Xiangqi, XU Pingliang
(*Research Institute of Information Fusion*, *Naval Aviation University*, *Yantai* 264001, *China*)

Abstract: With the continuous enrichment of satellite-borne and airborne remote sensing detection methods, the types of remote sensing data obtained are more diverse and the data scale is constantly expanding, which strongly drives the development of cross-modal correlation methods in the field of remote sensing. Cross-modal retrieval task refers to retrieving relevant data from other modes according to the query samples in a given mode. Multi-modal data usually includes images, text, video, audio, etc. Remote sensing image and text are important components of intelligence information, and the establishment of correlation between remote sensing image and text information is of great significance to the effective use of multi-source intelligence data. The mutual verification of the two is helpful to further improve the reliability of acquiring intelligence information. Remote sensing data usually contains rich information, but getting serviceable knowledge from the massive data effectively can be very challenging. With the continuous development of deep learning, deep neural networks are more and more widely used to obtain feature representations of different modes. Mapping cross-modal information into the same feature space is helpful to solve the "heterogeneous gap" problem among different modalities. Hashing method achieves fast retrieval speed and high efficiency. With the growth of remote sensing data type and scale, it has attracted more and more attention in the field of remote sensing cross modal. However, the existing unsupervised deep hashing cross-mode methods still have some problems. Usually, the similarity information across modes is learned separately, and without the assistance of label information, the model cannot obtain the semantic correlation between different modes correctly and effectively. In addition, most deep hash methods generate hash codes directly from the original features obtained by deep neural networks, and the generated hash features are difficult to obtain satisfactory discrimination information. In general, data from different modalities could give people a comprehensive description of the same object. As a result of the applicability and flexibility of cross modal retrieval, multiple methods of it have been widely explored in the computer vision community. In recent years, some studies have been conducted on cross-modal retrieval in the field of remote sensing. But most of the existing cross-modal correlation methods in remote sensing field are based on real value representation, which has problems of slow correlation retrieval speed and large memory consumption. However, the hash coding method can effectively improve the efficiency of association retrieval and is more suitable for large-scale and rapid association retrieval tasks. However, some semantic information will be lost during the transformation of the hash code. Therefore, this paper proposes an unsupervised hash-cross-modal association method for remote sensing images assisted by a similarity matrix. The constructed original feature and the similarity

matrix of the hash feature are used to integrate the semantic correlation information between different modes, so as to preserve the semantic correlation within modes and between different modes as much as possible, and reduce the loss of feature information when the original feature is converted to hash code through semantic alignment between similarity matrices. The loss function uses the weighted sum of the similarity matrix loss and contrast loss. The combination of the two effectively improves the accuracy of unsupervised cross-modal hash association, which is more suitable for large-scale cross-modal remote sensing image association retrieval tasks. Experimental results on benchmark datasets in the remote sensing field show that the proposed method performs better than the existing benchmark method. However, the design of the original feature extraction module of the model does not fully consider the semantic information richness of each mode, and the calculation method of the similarity matrix is relatively simple. Future work can be further improved on the accuracy of association.

Key words: Remote sensing; Unsupervised learning; Cross-modal retrieval; Dimilarity matrix; Contrastive learning

OCIS Codes: 100.2960; 280.4788; 200.3050

Foundation item: National Natural Science Foundation of China (Nos. 61790554, 62001499)