

中图法分类号: TP391.41; TP18 文献标识码: A 文章编号: 1006-8961(2026)08-3030-15

论文引用格式: Jiang K, Qian W H and Liu P. 2026. Image-text retrieval for Han portrait stones through the synergy of semantic enhancement and text reorganization. Journal of Image and Graphics, 31(8):3030-3044(姜坤, 钱文华, 刘朋. 2026. 语义增强与文本重组协同的汉画像石图文检索. 中国图象图形学报, 31(8):3030-3044)[DOI:10.11834/jig.250393]

语义增强与文本重组协同的汉画像石图文检索

姜坤, 钱文华*, 刘朋

云南大学信息学院, 昆明 650504

摘要: 目的 汉画像石是解读汉代历史文化的珍贵载体。图文检索技术能推动汉代文化研究向数字化转型, 为文化保护提供智能化解决方案。然而, 现有图文检索方法应用于汉画像石时, 存在领域偏移、图像与文本难以对齐等问题。本文提出一种语义增强与文本重组协同的图文检索方法以提高检索的准确性和全面性。方法 本文通过编码器分别完成汉画像石图像与文本特征的提取。在图像处理阶段, 基于汉画像石图像对应掩膜图像, 计算各语义对象权重, 进而推导出注意力权重。通过残差连接获取注意力特征, 并将所得注意力权重直接作为语义增强特征。通过多阶段特征融合将原始特征、注意力特征和增强特征融合得到组合特征以丰富特征表示。在文本处理阶段, 将文本与提取的图像语义对象嵌入随机提示模板, 生成结构更清晰的增强文本, 增强文本的语义表达能力。同时, 为适配本文数据集结构, 提出动态温度调整以及非对称相似度计算的方法, 基于特征相似度自适应调整温度。对图像到文本和文本到图像双向的相似度矩阵计算采用不同温度系数以提升跨模态匹配的准确性与稳定性。结果 通过本文所提方法验证, 文本到图像的检索指标 Mean_Recall、Recall@1、Recall@5 和 Recall@10 相较于排名第2的 BLIP (bootstrapping language-image pre-training) 模型分别提高了 15.20%、21.84%、14.29% 和 9.48%。图像到文本的检索指标 Mean_Recall、Recall@1、Recall@5 和 Recall@10 相较于排名第2的 X2-VLM (all-in-one pre-trained model for vision-language tasks) 模型分别提高了 17.47%、23.22%、18.45% 和 10.72%, 汉画像石图文可视化检索也得到了更好的效果。结论 本文方法显著增强了模型对汉画像石图像与其描述文本的对齐能力, 有效提升了图文检索在准确性、全面性等方面的表现。

关键词: 汉画像石; 图文检索; 语义增强; 文本重组; 动态温度

Image-text retrieval for Han portrait stones through the synergy of semantic enhancement and text reorganization

Jiang Kun, Qian Wenhua*, Liu Peng

School of Information Science and Engineering, Yunnan University, Kunming 650504, China

Abstract: Objective Han portrait stones are decorative stone carvings on funerary architectural structures, such as tombs and ancestral halls, dating back to the Han Dynasty. The contents of these stones depict mythological legends, institutional

收稿日期: 2025-10-28; 修回日期: 2026-01-15; 预印本日期: 2026-01-22

* 通信作者: 钱文华 whqian@ynu.edu.cn

基金项目: 国家自然科学基金项目 (62162065); 云南大学“双一流”建设联合专项 (202401BF070001-023); 云南省“视觉与文化计算创新团队”项目 (202505AS350009); 云南省科技厅应用基础研究计划项目 (202201AT070167)

Supported by: National Natural Science Foundation of China (62162065); Joint Special Project Research Foundation of Yunnan Province (202401BF070001-023); Yunnan Province Visual and Cultural Innovation Team (202505AS350009); Yunnan Fundamental Research Projects (202201AT070167)

systems, and everyday life scenes, including chariot processions, kitchen banquets, textile farming, and other scenes that directly reflect the social practices of the Han period. Iconic figures, such as the Queen Mother of the West (Xiwangmu), Fuxi and Nüwa, and feathered immortals (Yuren), represent the Han people's cosmology of "immortal souls". In particular, Xiwangmu images are prevalent in the tombs of the middle and lower classes but seldom appear in noble tombs, reflecting hierarchical differences in beliefs. Themes, such as ancient sages and war scenes, which combine architectural practicality and artistic decoration, serve educational and commemorative functions, embodying realism and romanticism. This study proposes an image-text retrieval method that leverages semantic information from mask images of Han portrait stones to enhance images and texts, addressing the domain shift problem present in existing image-text retrieval methods. The method introduces a semantic enhancement module and integrates semantic information with randomly generated prompt templates for text reorganization. This design enables the model to learn additional discriminative features, understand different objects in images, align text with images, and improve the accuracy and comprehensiveness of image-text retrieval.

Method This study utilizes the architecture of Chinese contrastive language-image pre-training (Chinese CLIP) using a vision transformer (ViT) as the image encoder and robustly optimized bidirectional encoder representations from Transformers (RoBERTa) model as the text encoder to extract image and text features. Semantic information from mask images and original text is used to calculate object-specific weights. Attention weights are derived by combining these weights with features from mask and original images. After a residual connection, these attention weights are used to generate attention features, while the weights themselves serve as semantic enhancement features. The original image features, attention features, and enhancement features are combined through a multi-stage feature fusion strategy at a ratio of 0.4:0.5:0.1 to form composite features. This design enables the model to focus on the primary subjects and important objects in images while enriching feature representation without information loss. Effective semantic objects are extracted from the mask images, and the original texts are embedded into randomly generated prompt templates incorporating these objects to produce enhanced texts. This process structures text descriptions, strengthens semantic expression, and improves matching with image features. Furthermore, dynamic temperature adjustment and asymmetric similarity calculation are introduced: the contrastive learning temperature is adaptively adjusted based on feature similarity (low temperatures applied to high-similarity samples to enhance discriminative learning and high temperatures applied to low-similarity samples to increase tolerance). The different temperature coefficients (0.9 for image-to-text and 1.1 for text-to-image) balance the matching difficulty in both directions, accounting for the distribution differences between image and text features to improve the accuracy and stability of cross-modal matching. **Result** In this study, a series of comparative experiments, ablation experiments, and visual comparison experiments is conducted on the proposed model to evaluate the model's performance and the contributions of each module. Results show that, on the Han portrait stone test set, the proposed method significantly improves the text-to-image retrieval metrics Mean_Recall (%), Recall@1 (%), Recall@5 (%), and Recall@10 (%) by 15.20%, 21.84%, 14.29%, and 9.48%, respectively, compared with the second-ranked bootstrapping language-image pre-training model in terms of precision and comprehensiveness. During the image-to-text retrieval, the metrics Mean_Recall (%), Recall@1 (%), Recall@5 (%), and Recall@10 (%) are improved by 17.47%, 23.22%, 18.45%, and 10.72%, respectively, compared with the second-ranked cross-modal vision-language model (X2-VLM), demonstrating significant improvements in image-text retrieval performance. Visualization experiments can effectively assist us in observing the improvement of the model's performance and the enhanced effects brought about by each added module, making the experimental results highly intuitive. Additionally, the proposed method maintains good retrieval performance in data robustness experiments. The average recall rate of the model consistently exceeds 70%, and it maintains excellent retrieval performance. This outcome indicates that the proposed model exhibits remarkable data robustness and is applicable to various scenarios, such as fuzzy retrieval. **Conclusion** The proposed method enhances the model's ability to align Han portrait stone images with their descriptive texts. The model can learn effective and crucial features by simultaneously enhancing text and images. The proposed dynamic temperature mechanism adapts to the structural characteristics of the Han portrait stone dataset and enhances the model's ability to learn from difficult samples, effectively improving the performance of image-text retrieval in terms of accuracy and comprehensiveness. However, when objects in images are similar but differ in positional information, the image-text retrieval results may suffer from confusion and false positives. Given the

limited scale and diversity of the dataset for Han portrait stone image-text retrieval, the model's cross-modal retrieval performance may be constrained. In the future, we will optimize the model's capability to semantically analyze the different positions and actions of objects in images and expand the Han portrait stone dataset.

Key words: Han portrait stones; image-text retrieval; semantic enhancement; text reorganization; dynamic temperature

0 引言

图文检索,又称跨模态检索,是一种通过文本描述检索相关图像或通过图像检索相关文本描述的技术。其核心目标是消除图像与文本之间的语义鸿沟,实现跨模态数据的精准对齐(Smeulders等, 2000)。图文检索的技术路径通常包含视觉特征提取、文本语义编码和跨模态对齐,通过残差网络(residual network, ResNet)(He等, 2016)或视觉变换器(vision Transformer, ViT)(Dosovitskiy等, 2021)提取图像的全局与局部特征,利用循环神经网络(recurrent neural network, RNN)(Lipton等, 2015)或Transformer完成文本语义编码,再借助联合嵌入空间或注意力机制等方式实现跨模态特征匹配(尹奇跃等, 2021)。例如,在网购场景中,用户输入商品名称可以检索出相关图像;医学领域中,医院通过影像数据可以检索出相关诊断报告等(姜定和叶茫, 2023)。

汉画像石是中国汉代特有的一种石刻艺术,是研究汉代社会、经济与文化的图像史书。如图1所示,图中两个人在敲鼓,两个人在奏乐,一个人在喝酒,描述的是一幅汉代的宴饮聚会图,阴线刻细腻流畅,浅浮雕立体感强,构图饱满且动感强烈,充满力量感与生命力。



图1 汉画像石实例

Fig. 1 Examples of Han portrait stones

汉画像石通过其多样丰富的内容展现了汉代的社会生产与风俗习惯,每一幅画作都蕴含了汉人的精巧构思和美好的愿景寄托。目前,针对汉画像石的研究主要集中在艺术价值(甘胜楠, 2019)和历史意义(渠超和翟京襄, 2009)等方面,缺乏对汉画像石图像和文本的数字化分类及检索(Chen等, 2023)。

因此,研究汉画像石图文检索技术,可以将分散的图文数据,如人物服饰、器物形制和场景布局等进行数字化分类,有利于用户通过关键词、图像特征(如车马出行、宴请奏乐)等快速定位相关材料,为汉代社会史、艺术史和科技史等研究提供技术支持(宋维涛等, 2025)。

图文检索需要提取出图像和文本的特征并计算出两个方向的相似度矩阵,然后按照相似度排序得出检索结果(Radford等, 2021)。早期的图文检索方法基于手工设计的视觉特征,如颜色、纹理等与文本关键词进行匹配,该方法依赖人工标注和简单统计。Smeulders等人(2000)总结了早期的基于内容的图像检索(content-based image retrieval, CBIR)技术方法,该方法通过颜色直方图、Gabor纹理等全局特征或尺度不变特征变换(scale-invariant feature transform, SIFT)等局部特征描述图像,结合文本标签进行检索。该方法直观、准确度高,但是依赖人工检索,时间成本高、步骤繁杂,不满足实时性要求。

随着神经网络的发展,双塔模型在跨模态匹配中经历了从全局特征对齐到细粒度交互的演进,其技术路径可划分为3个阶段。在全局特征对齐阶段, Faghri等人(2018)提出的VSE(visual semantic embedding)模型具有代表性,该模型采用卷积神经网络(convolutional neural network, CNN)提取图像全局特征、RNN编码文本序列,通过三元组损失优化特征对齐。但受限于模态本质差异及模型架构局限, VSE无法实现图像与文本的细粒度对齐。之后,局部特征交互被提出, Lee等人(2018)提出的SCAN(stacked cross attention network)模型成为重要突破,其结合快速区域卷积神经网络(faster region-based convolutional neural network, Faster R-CNN)(Ren等, 2015)提取图像区域特征,通过交叉注意力实现图像区域与文本单词的双向匹配,首次达成区域—单词级细粒度对齐,打破了全局特征的限制。不过, SCAN的双重注意力机制可能导致对齐不一致,且由于未引入大规模预训练,其对数据稀疏场景适应性较差。

此后,基于大规模图文数据的预训练模型成为主流。在对比学习与统一语义空间构建方面,Radford等人(2021)提出的CLIP(contrastive language-image pre-training)模型利用4亿图文对训练,通过对比学习构建统一语义空间,支持零样本检索,但其细粒度对齐能力有限,对中文等低资源语言支持不足。同期,Jia等人(2021)提出的ALIGN(automatic learning of inter-modality grounding)模型进一步验证了数据规模优先于数据质量,使用18亿噪声图文对训练,采用简化双塔结构提升训练效率,不过仍缺乏细粒度交互机制。Kim等人(2021)提出的ViLT(vision-and-language Transformer)模型则摒弃卷积与区域监督,将图像patch与文本token直接输入Transformer,首次实现图像与文本的“原生统一输入”,计算效率较SCAN提升10倍以上,但图像patch未结合语义信息,对复杂场景适应性不足。Wang等人(2022)提出的OmniVL(one foundation model for image-language and video-language tasks)模型采用了统一的Transformer编码器,先粗粒度对比全局特征再细粒度重排,通过“粗—细”两阶段对齐策略平衡效率与精度,显著提升检索任务的召回率与排序精度。Qin等人(2022)提出的Deep Evidential Learning模型,通过证据损失函数增强了模型对复杂数据的鲁棒性,首次将噪声建模融入跨模态对齐,解决了实际场景中数据标注质量参差不齐的问题。Zeng等人(2024)提出的X2-VLM(all-in-one pre-trained model for vision-language tasks)模型通过patch离散化提升细粒度检索能力,解决了ViLT仅依赖patch的局限。Li等人(2023)提出的BLIP(bootstrapping language-image pre-training)模型采用混合编解码器架构与CapFilt数据自举,通过ITC(image-text contrastive loss)和ITM(image-text matching loss)强化细粒度对齐,显著提升了细粒度对齐能力。Wang等人(2023)提出的双向图像文本编码器(bidirectional encoder representation from image and text, BEiT-3)模型以统一掩码建模替代对比学习,通过掩码图像patch和文本token实现语义对齐,提升了特征泛化性。Yang等人(2023)提出的Chinese-CLIP(Chinese contrastive language-image pre-training)模型则针对性解决中文场景语义鸿沟,构建2亿中文图文数据集,填补了跨语言泛化空白。Ge等人(2024)提出的视觉语义空间自强化网络3SHNet(visual semantic

spatial self-highlighting network)模型进一步突破,融合物体语义与空间位置信息,突破了孤立区域对齐的局限,显著提升对场景化描述的匹配精度。

然而,上述方法应用于汉画像石图文检索时存在以下问题:1)小样本图文特征提取不足。当前的预训练大模型用数量较大的图文对进行训练,以确保模型的有效训练和优化。由于汉画像石样本的稀缺性和抽象性以及相似性,模型在训练过程难以提取到关键特征,导致图像与文本无法正确对齐,最终导致图文检索的效果较差。2)领域偏移问题。当前的多模态预训练大模型虽然在大规模公开数据集上表现出色,但描述未见领域的图像时,表现出明显的领域偏移问题。模型的先验知识会主导检索过程,导致模型无法将图像信息与文本正确对齐,进而检索到无关的图像或文本,出现幻觉现象。3)图像文本难以对齐。汉画像石图像中通常包含许多场景与元素,模型难以将图像的小场景与描述文本语义相对应,从而导致错检和漏检。

为此,本文提出语义增强与文本重组协同的汉画像石图文检索方法,主要贡献在于:1)建立了汉画像石图文检索数据集,包括神仙传说、车马出行以及市井生活等题材;2)提出语义增强模块,通过汉画像石掩膜图像计算对象权重,生成注意力与增强特征,经多阶段融合形成组合特征,提升图像特征表达力与鲁棒性;3)提出文本重组模块,依据掩膜图像提取语义对象,将其与文本嵌入随机模板生成增强文本,结构化描述文本,提高图文匹配度;4)引入动态温度机制,对高相似度样本设低温增强区分,低相似度样本设高温提升容错,对图像到文本、文本到图像双向的相似度矩阵计算采用不同温度系数,平衡跨模态匹配难度,增强匹配准确性与稳定性。

1 本文方法

本文提出的图文检索模型总体结构如图2所示。图文检索模型由图像编码器、文本编码器、语义增强模块和文本重组模块构成。图像端,将汉画像石图像与其掩膜图像统一尺寸后输入图像编码器,对输出的语义和图像特征归一化;依据文本和掩膜图像计算语义权重并归一化,结合语义权重与图像特征得出注意力权重,经残差连接生成注意力特征,同时将注意力权重作为增强特征,最终融合图

像特征、注意力特征和增强特征得到组合特征,以丰富特征的表示。文本端,从掩膜图像提取不重复的有效语义对象,嵌入随机提示模板重组为增强文本,以规范文本表达并丰富文本语义信息,经编码器提取并归一化特征。然后计算组合特征与增强文本的特征相似度,推导温度参数生成相似度矩阵以适应本文数据集的结构,最后通过对比损失优化模型。

1.1 图像编码器

图像编码器使用 ViT 提取图像的特征, ViT 通过

自注意力机制直接建模图像分块之间的全局依赖关系,自注意力机制(Vaswani 等,2017)允许模型动态聚焦于图像中的关键区域,避免了 CNN 因局部卷积操作导致的长距离语义断裂问题(Dosovitskiy 等,2021)。此外,为了与文本编码器协同优化,本文采用 ViT 作为模型的图像编码器。

为了确保输入模型的图像大小一致,图像在送入图像编码器之前会被统一成 384×384 像素的大小,然后进行特征提取。

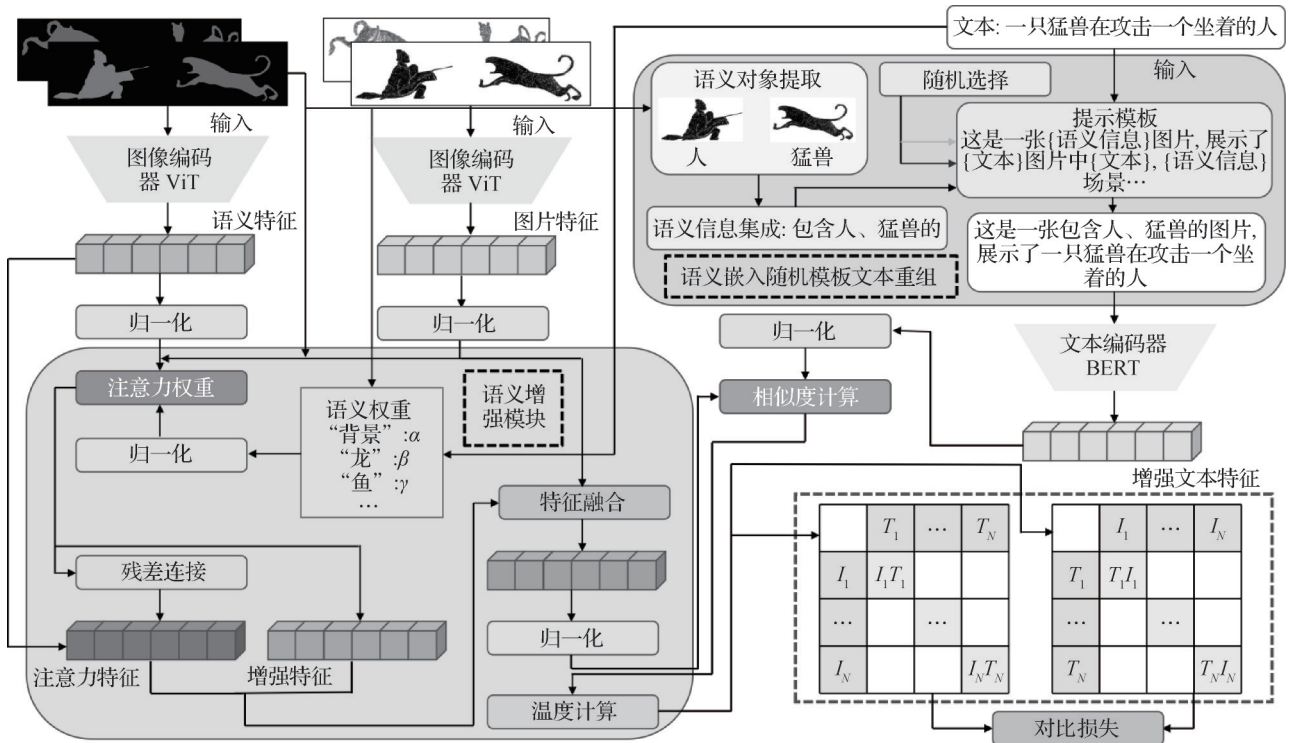


图2 模型总体结构图
Fig. 2 Overall structure diagram of the model

图3显示了图像编码器结构。图像进入编码器后首先被分块,将汉画像石图像转换为序列数据,使用 16×16 像素的 patch,总共被分成了 576 个 patch。而后对得到的 patch 进行向量化处理,每个 patch 通过线性投影层转换为向量,在降低复杂度的同时保留汉画像石的关键视觉信息,如线条、局部结构等。而后为每个 patch 添加位置编码,帮助模型理解汉画像石图像的空间结构,如图4车马出行图中人物、动物与马车之间的位置关系。将处理好的 patch 送入 Transformer 编码器层,编码器共有 12 层,每层包含多头自注意力、前馈神经网络、层归一化和残差连接模块,通过多个注意力头从线条方向、图案对称性等

不同角度提取特征,增强特征的多样性。最后进行全局平均池化,对 576 个向量 patch 求平均值,得到一个 1×512 维的特征向量。

1.2 文本编码器

文本编码器使用 RoBERTa(a robustly optimized

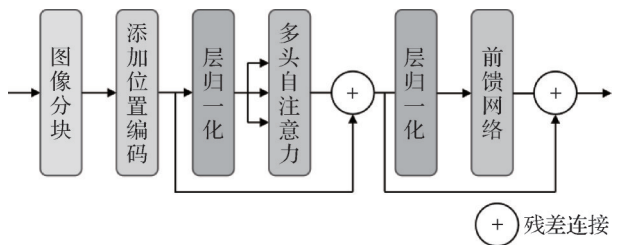


图3 图像编码器结构图
Fig. 3 Structure diagram of the image encoder



图4 车马出行图

Fig. 4 Picture of chariots and horses on the move

bert pretraining approach)(Liu等,2019),可以捕捉文本中的关键语义信息,具有较高的特征提取效率,由于其使用的Transformer-Encoder层可以更好地捕捉词元之间的语义关联,可以从文本中提取更多有效的特征,且该编码器对中文数据集的适应性较好,因此,本文选用RoBERTa模型作为文本编码器。

如图5所示,文本输入进到文本编码器后,编码器首先会将文本按照词或字的粒度分割为词元序列。而后文本编码器会为每个词元分配一个向量表示,这个向量蕴含了词元的语义信息。与图像编码器类似,为每个词元在文本序列中的位置添加位置编码,这样可以让模型理解词元的先后顺序,最终生成768维的向量序列。该向量序列会被送至Transformer编码器层,该层包含了多头自注意力机制和前馈网络以及层归一化,通过自注意力捕捉每个词元之间的语义关联,理解文本的整体语义。前馈神经网络对自注意力输出的特征进行非线性变换,进一步增强特征的表达能力,区分出细粒度特征。最后通过跨模态投影层将768维向量线性变换为512维,与图像特征维度对齐,实现汉画像石跨模态检索。

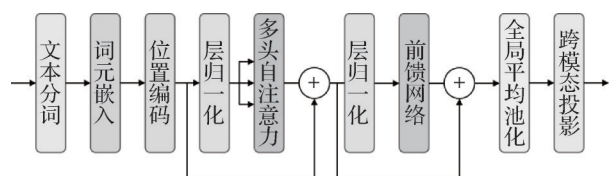


图5 文本编码器结构图

Fig. 5 Structure diagram of the text encoder

1.3 语义增强模块

为了让提取到的汉画像石图像特征中人物、动物等主体对象的特征更为突出,本文引入了语义增强模块,其结构如图6所示,根据掩膜图像所提供的语义信息对图像的特征进行增强,使那些关键主体的特征更加突出,并与动态温度机制结合,显著提升了模型的检索性能。

本文首先根据掩膜图像信息和文本信息计算语义权重,语义权重包括文本权重和图像权重两部分。文本权重的计算首先将该批次内的所有文本以空格隔开转换成字符串,然后统计总文本中该语义对象的出现次数,计算为

$$T_w = \frac{\sum_{i=1}^{|U_{t \in T} str(t)|} I(U_{t \in T} str(t)_i = n)}{|U_{t \in T} str(t)|} \quad (1)$$

式中, T_w 表示文本权重, n 表示语义对象的名称, T 表示所有文本, t 表示单个文本。 I 为指示函数,括号内条件成立时取值为1,不成立则为0。 U 表示把批次内所有文本 t 拼接合并成一条完整长字符串。 str 表示拼接后长字符串的第 i 个词元/字符。

计算图像权重时,首先统计该批次内所有图像的颜色张量与当前输入语义对象的颜色张量相等的次数,然后得出该次数在该批次图像总像素数的占比,获得图像权重,计算为

$$I_w = \frac{\sum_{b=1}^B \sum_{h=1}^H \sum_{w=1}^W I(\bigcap_{p=1}^3 (S_{b,p,h,w} = O_{1,p,1,1}))}{B \cdot H \cdot W} \quad (2)$$

式中, I_w 表示图像权重, B 表示总批次大小, p 表示通道信息, H 、 W 表示图像高和宽, O 表示原图像, S 表示掩膜图像, b 表示具体批次, h 、 w 分别表示像素点的具体高和宽。

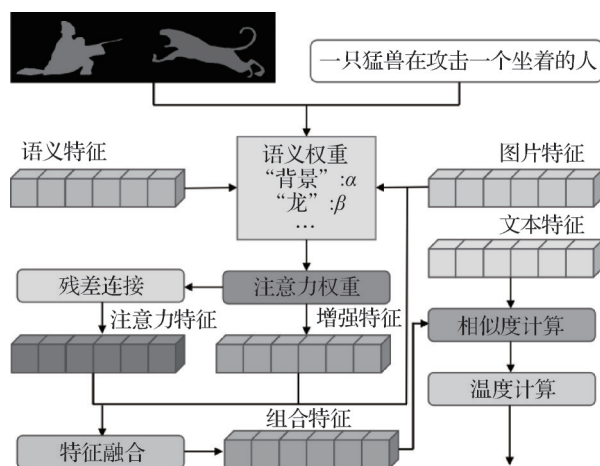


图6 语义增强模块结构图

Fig. 6 Structure diagram of the semantic enhancement module

语义权重的计算基于每个对象的图像权重和文本权重,由于掩膜图像中黑色背景的占比较高,并且黑色背景属于无效的语义信息,所以在权重加和时,图像权重的占比仅设置为0.1,文本权重的占比设

置为0.9,以确保计算出合理的语义权重,计算为

$$w[i] = 0.9 \times T_{\text{w}} + 0.1 \times I_{\text{w}} \quad (3)$$

式中, w 表示语义权重。

通过计算得到各语义对象的权重,过滤掉了一些无关信息,让模型关注权重更高的对象。在此基础上,根据语义权重计算出注意力特征和增强特征,首先根据原始图像特征初始化注意力权重,进而根据掩膜图像张量和目标语义对象颜色张量计算出语义掩码,然后根据语义掩码在所有像素的占比计算出该掩码的重要性,计算为

$$a^n = \frac{\sum_{h=1}^H \sum_{w=1}^W I\left(\bigcap_{p=1}^3 (S_{b,p,h,w} = O_{1,p,1,1}^n)\right)}{H \cdot W} \quad (4)$$

式中, a^n 表示掩码的重要性。

根据计算所得掩码重要性和语义权重计算出注意力权重,对注意力权重残差连接再乘以掩膜图像的特征则得到了最终的注意力特征。计算式为

$$F_{\text{att}} = \text{softmax}\left(\sum_{n \in N} w^n \cdot (F_{\text{seg}} \odot a^n)\right) \odot F_{\text{seg}} \quad (5)$$

式中, F_{att} 表示注意力特征, N 表示语义对象集合, F_{seg} 表示掩膜图像的特征, $\odot$ 是矩阵乘法, softmax 是激活函数(Jang等,2017)。

基于注意力权重计算增强特征,去掉了残差连接的步骤,直接对特征进行增强,丰富特征的表示。将注意力权重直接作为增强特征,计算为

$$F_{\text{enh}} = \sum_{n \in N} w^n \cdot (F_{\text{seg}} \odot a^n) \quad (6)$$

式中, F_{enh} 表示增强特征。

该模块设计了多阶段特征融合,将原始图像特征、注意力特征和增强特征融合得到最终的组合特征,计算为

$$F_{\text{comb}} = \alpha F_{\text{img}} + \beta F_{\text{att}} + \gamma F_{\text{enh}} \quad (7)$$

式中, F_{comb} 表示组合特征, F_{img} 表示汉画像石图像特征, α, β, γ 表示比例系数。

在得到组合特征后,计算其与增强文本特征的相似度。为了适应本文的特殊数据集结构,提出了动态温度调整和非对称相似度计算,模型会根据相似度高低自动调整对比学习的温度。当相似度较高时则使用较低的温度以增强区分度;相似度较低时则使用较高温度增加容错性,这样可以提升模型在不同难度的汉画像石样本上的学习效果,计算为

$$T = T_{\text{base}} \cdot [1 + a \cdot (1 - \frac{1}{n} \sum_{i=1}^n (F_{\text{img}} \cdot F_{\text{text}}^T))] \quad (8)$$

式中, T 为温度, $T_{\text{base}} = 0.07$,参考了CLIP的设置(Radford等,2021), a 为激进系数,表示相似度对温度的影响强度,本文设置为0.3, F_{text} 表示文本特征。

由于汉画像石图像多数场景表现出相同的语义信息,因此本文数据集设计文本与图像是一对多的关系,图像与文本是一对一的关系。为了应对此结构,本模型为双向的检索设置了不同的温度系数以平衡两种方向的匹配难度,充分考虑图像和文本特征的分布差异,提高跨模态匹配的准确性和稳定性,计算为

$$\begin{cases} L_{\text{img2text}} = \frac{F_{\text{img}} \cdot F_{\text{text}}^T}{0.9T} \\ L_{\text{text2img}} = \frac{F_{\text{text}} \cdot F_{\text{img}}^T}{1.1T} \end{cases} \quad (9)$$

式中, L 表示相似度矩阵,图像到文本和文本到图像检索的温度系数分别设置为0.9和1.1。

1.4 语义嵌入随机模板文本重组

本文提出了文本重组模块,它可以充分利用掩膜图像中的有效语义信息并嵌入规范提示模板,使得汉画像石描述文本更加规范化并且包含更多语义信息,与图像特征增强协同,提升汉画像石图文检索的性能。其结构如图7所示。

该模块首先从输入图像对应的掩膜图像中提取出有效且不重复的语义对象,然后将这些对象作为Context,语义对象获取的过程计算为

$$V = \{n \in N | n \notin \{g, \text{pad}, \text{cls}, \text{sep}\} \wedge \neg \exists n' \in N, n' = n\} \quad (10)$$

式中, V 表示有效语义对象, g 表示背景, n' 表示单个语义对象, pad (填充)、 cls (分类标记)、 sep (分隔符)等是特殊字符。

汉画像石描述文本进入该模块后,将为该文本选择一个随机提示模板,并将文本内容 Text 与包装后的语义对象信息Context嵌入到该模板中,计算式为 $\text{format}(t, s, n) = t.\text{replace}\{\text{text} \rightarrow s, \text{context} \rightarrow c\}$ (11)式中, t 是随机提示模板, s 表示文本,“ $\rightarrow$ ”表示嵌入模板。

2 实验结果与分析

2.1 实验数据

汉画像石的题材包罗万象、异彩纷呈,既生动揣摩了古人的现实生活图景,又深刻承载了先民的神

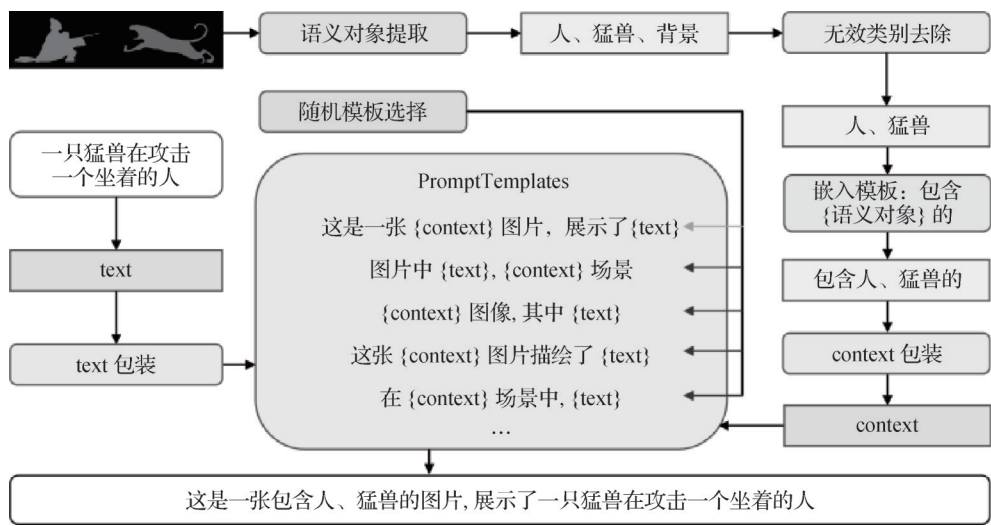


图 7 语义嵌入随机模板文本重组结构图

Fig. 7 Structure diagram of semantic embedding random template text reorganization

话信仰世界,更忠实记录了波澜壮阔的历史与声势浩荡的车马出行场景。

为了提高汉画像石图像的质量,本文从汉画像石博物馆和网络搜集整理素材。通过对原始质量不佳的汉画像石图像进行裁剪、清晰化和模糊部分消除等处理保留清晰可识别的部分,如图 8 所示。

此外,为了应对汉画像石图像数据集不足的问题,本文通过提取出汉画像石中的不同对象,如图 9 所示,将不同的对象组合拼凑成新的汉画像石图像,以此对汉画像石数据集进行扩充。

经过优化和扩充后的汉画像石图文检索数据集

题材包含市井生活、神仙传说和车马出行,图像数量分别约是 300、270 和 260 幅,根据素材自建文本描述语句共 479 句。数据集部分示例如图 10 所示,其中一条文本可以对应多个图像场景的语义表达,如图中文本“四个女娲”可以对应两幅图像,单幅图像只能对应单条文本,如图中四人骑马出行图对应唯一文本“四个人骑着马出行”。

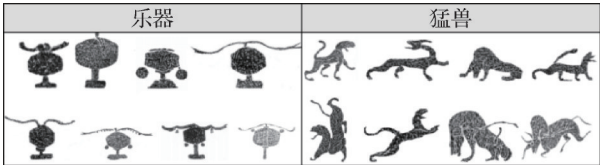


图 9 汉画像石对象提取示例图

Fig. 9 Example diagram of object extraction for Han portrait stones

2.2 实验设置

实验环境基于 Pytorch 的深度学习框架,显卡型号为 Geforce NVIDIA RTX 4090,模型的图像编码器为 ViT,文本编码器为 RoBERTa,为加速汉画像石图文检索任务的收敛,各模型均采用了与其相对应的预训练模型进行微调,使用单卡训练,训练过程中权重衰减为 0.01/10 epoch, batch_size 设置为 24,初始学习率设置为 2×10^{-5} ,优化器为 AdamW (Loshchilov 和 Hutter, 2019),并使用交叉熵损失函数 (Mao 等, 2023)对模型进行微调,通过对比损失 (van den Oord 等, 2019)优化模型,实验数据集按照 6:2:2 的比例进行分配,其中,训练集包含 280 个文本和 500 幅图

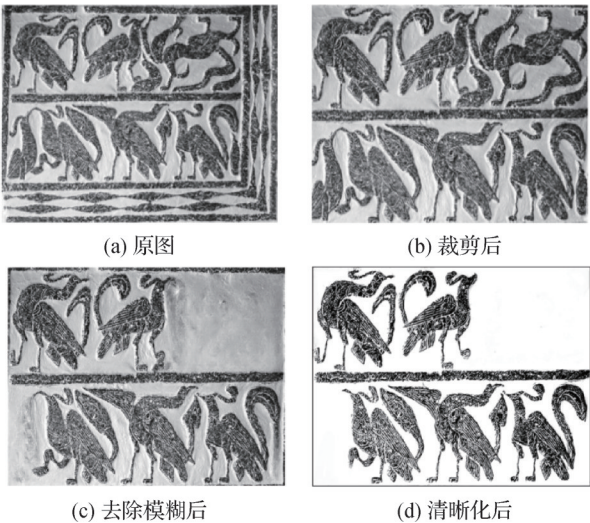


图 8 汉画像石数据集处理示例

Fig. 8 Examples of Han portrait stones dataset processing
((a)origin image; (b)after cutting; (c)after removing the blur;
(d)after clarification)

题材	文到图			图到文	
	文本	示例图像 1	示例图像 2	图像	唯一文本
市井生活	两个人敲鼓两个人在演奏				两个人敲鼓一个人坐着一个人跳舞
神仙传说	四个女娲				五个人在参拜女娲
车马出行	两辆马车和两个骑马的人在赶路				四个人骑着马出行

图 10 汉画像石图文检索数据集示例
Fig. 10 Examples of image-text retrieval dataset for Han portrait stones

像,验证集包含 95 个文本和 164 幅图像,测试集包含 104 个文本和 168 幅图像。

2.3 评价指标

为了验证模型的有效性,本文采用领域内常用的评估策略召回率 Recall@k (Hawking 等, 1999), 本文将 k 的值设定为 1、5 和 10, 通过计算图像到文本和文本到图像的 Recall@1、Recall@5、Recall@10, 进而反映检索结果的精确性、全面性和鲁棒性。Recall@k 通过计算前 k 个结果中真正相关的数量占比来评估模型的检索性能, 计算式为

$$Recall@k = \frac{TP@k}{TP@k + FN@k} \tag{12}$$

式中, FN@k 表示未被检索到的相关实例数量 (即漏检), TP@k 表示前 k 个结果检索到的数量。

2.4 特征融合参数实验

本文对特征融合参数进行讨论, 如表 1 所示, 表中对不同特征比例组合下的图文检索指标进行了对比。为了保留原始特征占有一定的比重, 本实验设置原始特征最低为 0.4、最高为 0.8, 根据不同的特征比例计算比较评价指标, 最终得出原始特征、注意力特征和增强特征按照 0.4、0.5 和 0.1 的比例是最佳的特征融合方案, 在图像到文本检索和文本到图像检索的实验中都获得最高的指标。通过实验可知, 以注意力特征和原始特征为主导、增强特征为辅助的搭配检索指标最高, 原始特征保留较多则会导致模型性能下降, 注意力特征的加入让模型可以更好地识别汉画像石图像中的关键部分。

2.5 与其他算法比较

为了验证本文模型的有效性, 选取 BLIP、ViLT、

VSE、X2-VLM、BEiT3、Chinese-CLIP、DSMD 和 3SHNet 等模型进行比较。本文对不同模型的图文检索性能进行了评估, 对比结果如表 2 所示, 从表中可以看出, 本文模型的图像检索评价指标 Mean_Recall、Recall@1、Recall@5 和 Recall@10 在文本到图像的检索中分别达到 79.48%、55.76%、87.5% 和 95.19%, 在图像到文本的检索中分别达到 78.37%、52.38%、85.71% 和 97.02%, 在精确性和全面性上, 文本到图像的检索指标 Mean_Recall、Recall@1、Recall@5 和 Recall@10 相较于排名第 2 的 BLIP 模型分别提高 15.20%、21.84%、14.29% 和 9.48%。图像到文本的检索指标 Mean_Recall、Recall@1、Recall@5 和 Recall@10 相较于排名第 2 的 X2-VLM 模型分别提高 17.47%、23.22%、18.45% 和 10.72%。可以看出, 本文所提模型在文本到图像检索和图像到文本检索的精确性和全面性都得到了提升, 解决了领域偏移问题, 图文特征能更好地进行对齐。

2.6 消融实验

为了验证本文提出的语义增强模块、动态温度模块以及文本重组模块的有效性, 本文设置了相应的消融模型, 实验结果见表 3。

模型首先通过添加动态温度模块提高了模型跨模态检索的稳定性和准确性。在此基础上, 引入多阶段特征融合模块, 使模型关注图像中的重要区域, 提取更多关键的特征。最后, 引入文本重组模块与多阶段特征融合模块协同以增强模型对齐汉画像石图像与文本特征的能力。实验结果显示, 在图文检索的精确性和全面性方面, 本文模型相较于其他模型都有一定程度的提升, 相较于不添加任何模块的模

表 1 不同特征比例融合下的指标对比
Table 1 Comparison of indicators under fusion of different feature proportions

特征比例			文本到图像检索			图像到文本检索		
原始特征	注意力特征	增强特征	r@1/%	r@5/%	r@10/%	r@1/%	r@5/%	r@10/%
0.4	0.1	0.5	45.19	81.73	92.30	46.42	86.30	94.64
0.4	0.2	0.4	49.03	81.73	92.30	44.64	80.95	95.83
0.4	0.3	0.3	48.07	81.73	95.19	39.28	84.52	92.26
0.4	0.4	0.2	48.07	85.57	95.19	47.12	84.52	93.45
0.5	0.1	0.4	48.07	85.57	95.19	52.38	84.52	93.45
0.5	0.2	0.3	49.03	85.57	95.19	45.23	83.33	94.64
0.5	0.3	0.2	46.15	79.80	88.46	35.71	75.59	91.07
0.5	0.4	0.1	42.30	81.73	91.34	41.66	79.16	89.88
0.6	0.1	0.3	47.11	77.88	89.42	47.02	79.16	95.23
0.6	0.2	0.2	44.23	75.96	88.46	39.28	81.54	94.04
0.6	0.3	0.1	43.26	79.80	92.30	42.26	85.11	94.64
0.7	0.1	0.2	49.03	81.73	94.23	42.85	84.52	94.04
0.7	0.2	0.1	44.23	76.92	90.38	36.30	73.21	89.28
0.8	0.1	0.1	40.38	78.84	89.42	43.45	79.16	91.07
0.4	0.5	0.1	55.76	87.57	95.19	52.38	85.71	97.02

注:加粗字体表示各列最优结果。r 是 Recall 的缩写。

表 2 本文方法与其他方法指标对比
Table 2 Comparison of metrics between the method in this paper and other methods

方法	文本到图像检索				图像到文本检索			
	Mean_r	r@1	r@5	r@10	Mean_r	r@1	r@5	r@10
VSE(Faghri 等,2018)	38.13	22.11	42.30	50.00	38.68	19.64	41.66	54.76
ViLT(Kim 等,2021)	39.67	11.90	45.23	61.90	38.77	10.57	45.19	60.57
X2-VLM(Zeng 等,2024)	42.94	16.34	47.11	65.38	60.90	29.16	67.26	86.30
BEiT3(Wang 等,2023)	59.61	28.84	67.30	82.69	59.29	26.92	64.42	86.53
BLIP(Li 等,2023)	64.28	33.92	73.21	85.71	48.39	21.15	55.76	68.26
Chinese-CLIP(Yang 等,2023)	46.79	21.15	49.03	70.19	35.74	16.07	37.50	53.57
DSMD(Liang 等,2025)	34.52	10.71	39.88	52.97	33.52	13.09	36.90	50.59
3SHNet(Ge 等,2024)	55.12	27.88	61.53	75.96	60.31	29.16	68.45	83.33
本文	79.48	55.76	87.50	95.19	78.37	52.38	85.71	97.02

注:加粗字体表示各列最优结果。r 是 Recall 的缩写。

型,文本到图像检索中,本文模型在图文检索指标 Recall@1、Recall@5 和 Recall@10 分别提升 41.34%、47.12% 和 42.31%;图像到文本检索中,本文模型在图文检索指标 Recall@1、Recall@5 和 Recall@10 分别

提升 41.08%、55.36% 和 50.00%。结合消融实验结果看,本文添加的每个模块对模型图文检索的效果具有不同程度的增益作用,共同促进了图文检索性能。
此外,为了可视化模块添加对于模型效果的增

益,针对此进行了正负样本分离度可视化实验,如图 11 所示。从图中可以看出,初始模型的正负样本存在较多重叠,峰值距离较近,区分度较低,仅添加

相较于初始模型正负样本重叠区域有所减小,并且峰值距离拉远。集成了两个模块的模型正负样本的重叠区域最小,并且峰值距离最远,模型的区分度效果最好,达到0.93。

表 3 消融实验
Table 3 Ablation experiment

模块			文本到图像检索			图像到文本检索		
多阶段特征融合	文本重组	动态温度	r@1	r@5	r@10	r@1	r@5	r@10
×	×	×	14.42	40.38	52.88	11.30	30.35	47.02
×	×	√	24.03	44.23	56.73	12.50	34.52	52.38
×	√	×	37.50	70.19	83.63	32.73	68.45	91.07
×	√	√	36.53	72.11	89.42	38.09	74.40	89.28
√	×	×	29.80	60.57	73.07	21.42	57.14	74.40
√	×	√	38.46	62.50	75.96	23.21	50.00	67.85
√	√	×	42.30	77.88	90.38	42.26	80.35	94.04
√	√	√	55.76	87.50	95.19	52.38	85.71	97.02

注:加粗字体表示各列最优结果。r 是 Recall 的缩写,×表示删除该模块,√表示添加该模块。

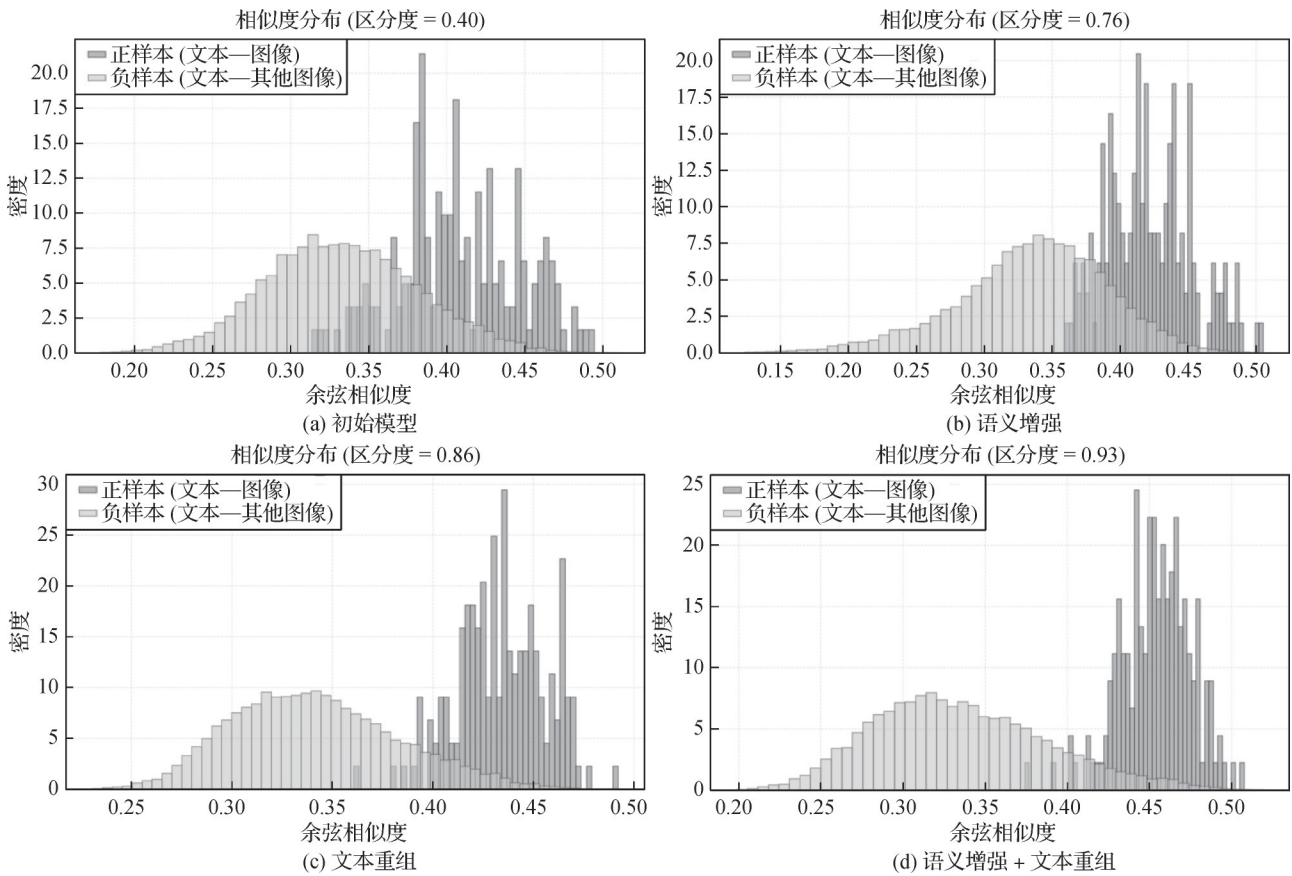


图 11 正负样本分离度可视化对比图

Fig. 11 Visualization comparison of positive and negative sample separation((a)initial model;(b)semantic enhancement model;(c)text reorganization model;(d)semantic enhancement and text reorganization)

2.7 可视化实验分析

为了更直观地体现出本文模型在汉画像石图文检索任务中性能的提升,通过可视化实验反映模块添加对本文模型的增益效果,如图 12 所示。图 12(a)显示了文本到图像的检索结果,图中实线边框表示该文本对应的正确检索图像,从初始前 10 个结果检索不到的情况,到进行微调、单图像增强和单文本增强,模型在全面性和精确性方面都得到了提升,本文提出的模型在检索的结果中前 6 个都是正确的,相较于其他模型,本文提出的模型可以捕捉到不同图像之间的同一语义表达,使得检索结果更加精确和全面。图 12(b)显示了图像到文本的检索结果,加粗字体的文

本表示该图像对应的正确描述文本,从初始模型前 10 个结果检索不到正确文本的情况,到进行微调、单图像增强和单文本增强,模型的性能得到了不错的提升,可以在前 5 个检索结果中检索到正确文本。本文提出的模型将图像增强与文本增强结合,提取到更丰富的图像特征和文本特征,模型能更好地将图文特征进行对齐,在检索效果上得到了不错的提升,可以达到检索到的正确结果中排名第一。

2.8 鲁棒性实验

为了验证本文模型的鲁棒性,对于图像,采取了随机裁剪、添加高斯噪声、模糊化与亮度变化等随机操作,对于文本,采取了同义词替换、顺序变换等操

文本: 一只猛兽在攻击一只鸟文本到图像检索 Top10										
初始										
微调										
单图像增强										
单文本增加										
图像+文本增强										

(a) 文本到图像检索

图像:											图像到文本检索 Top10									
初始	一条龙和四条鱼	一条龙与两只马	两条龙 两个人在敲鼓 一个人在站着 一个人在喝酒 一个人在坐着 一个人在跳舞		六条龙	一只猛兽在攻击女娲	一条龙在追一只猛兽和一个人	两个骑马的人 一只猛兽 一条龙	四只猛兽在追三个骑马拿长矛的人	一条龙在追赶一只猛兽	三只猛兽 两个女婢 一只鸟 一个人									
微调	五只鸟 一只青蛙 七条鱼	五只鸟 四条鱼 两只青蛙 一匹马 四个人	六只鸟 四条鱼 一只青蛙	五只鸟 四条鱼 和一只青蛙	两只鸟在飞 两个人在敲鼓 一个人站着 有四条鱼和两只青蛙		五只猛兽 两只鸟 一条龙 两条鱼	四条鱼和两只青蛙	两个人在交谈 五条鱼 一只青蛙	六条鱼和两只青蛙	三条鱼 两个女婢 一只青蛙									
单图像增强	六只鸟 四条鱼 一只青蛙	五只鸟 一只青蛙 七条鱼	五只鸟 四条鱼 和一只青蛙	五只鸟 四条鱼 两只青蛙 一匹马 四个人	五只猛兽 两只鸟 一条龙 两条鱼		六条鱼和两只青蛙	两只鸟在飞 两个人在敲鼓 一个人在坐着 有四条鱼和两只青蛙	两个女婢 一只青蛙 两只鸟 一只猛兽	两只鸟 两只青蛙 一条鱼	七个人在交谈 池塘里有三条鱼和一只青蛙									
单文本增强	八只鸟	五只鸟 四条鱼 和一只青蛙	五只鸟 一只青蛙 七条鱼	六只鸟 四条鱼 一只青蛙	五只猛兽 两只鸟 一条龙 两条鱼		六只鸟	五只猛兽	五只鸟 四条鱼 两只青蛙 一匹马 四个人	五只鸟	两条龙 四只鸟 一只青蛙									
图像+文本增强	五只猛兽 两只鸟 一条龙 两条鱼	四只猛兽在狩猎 五只鹿	六只猛兽	七只猛兽	两只猛兽在追 五只鹿		三只猛兽在捕猎 三只鱼	两只猛兽在攻击 两只马和两只鹿 和一个骑马的人		五只鸟 四条鱼 两只青蛙 一匹马 四个人	三头猛兽攻击 四头鹿	两只猛兽 两只鸟 一只鹿								
排名	1	2	3	4	5	6	7	8	9	10										

(b) 图像到文本检索

图 12 可视化实验对比图

Fig. 12 Visualization experiment comparison diagram ((a) text2img retrieval; (b) img2text retrieval)

作,在此情况下测试模型的检索性能,得到的指标对比如表4所示。虽然指标相较于原始数据有所下降,但是平均召回率始终可以保持不低于70%,模型依旧保持较好的检索性能,具有较好的鲁棒性。

本文进行了鲁棒性可视化实验,以验证鲁棒性,

如图13所示,图像到文本的检索性能相较于初始数据有所衰减,但是模型依旧可以在前10个检索结果内检索出正确的结果。文本到图像的检索,模型依旧能够保持良好的精确性与全面性,在有所干扰的情况下依旧可以检索出多个准确的结果。通过该实

表4 鲁棒性实验对比
Table 4 Comparison of data robustness experiments

方法	文本到图像检索				图像到文本检索			
	Mean_r	r@1	r@5	r@10	Mean_r	r@1	r@5	r@10
图像衰减	72.11	42.30	82.69	91.34	73.01	41.66	82.73	94.64
文本衰减	75.96	50.96	86.53	90.38	77.97	51.19	87.50	95.23
图像和文本衰减	70.19	42.30	76.92	91.34	74.00	42.85	85.71	93.45
本文	79.48	55.76	87.50	95.19	78.37	52.38	85.71	97.02

注:加粗字体表示各列最优结果。r是 Recall 的缩写。

图像到文本检索 Top10											
高斯噪声		两条龙	两只猛兽	两只猛兽在攻击一条龙	一条龙在攻击两只猛兽	一条龙和一只猛兽	五只猛兽,两只鸟,一条龙,两条鱼	两条龙和两个女蜗	三条龙两个女蜗	一条龙在攻击一只猛兽	两只猛兽在打架还有一条龙在旁边
随机裁剪		四个女蜗	三只猛兽在攻击一个人	两个女蜗和一只猛兽	四条鱼	七个女蜗	四只猛兽在打架	两个女蜗一只青蛙两只鸟一只猛兽	六只猛兽	六个女蜗	四个人在交谈
亮度变化		六只鸟	八只鸟	一只猛兽在追击五只鸟	五只鸟	有三只鸟在飞两只猛兽在战斗	四只鸟两条龙	七只鸟和两个人	一个人在攻击正在打架的两只猛兽三只鸟在飞	两只鸟和两只鹿	三只鸟六个人在交谈
模糊化		两只鸟两只青蛙一条鱼	四条鱼和两只青蛙	三条鱼两个女蜗一只青蛙	六条鱼和两只青蛙	三条鱼一个人在驱赶三只猛兽	四只鸟在抓两条鱼	五只鸟四条鱼和一只青蛙	五只鸟一只青蛙七条鱼	三只猛兽在捕猎三只鱼	三条鱼
操作	排名	1	2	3	4	5	6	7	8	9	10

(a) 图像到文本检索

文本到图像检索 Top10 (对文本采取调换顺序、同义词替换等操作)											
同义词替换: 凶兽→四足兽人→雅士 马车→汉代交通工具 骑马射箭→骑射 龙→五爪金龙 鸟→神鸟											
一只四足兽在攻击一只神鸟											
两条五爪金龙											
一位雅士											
两个人骑马两个人在骑射一辆汉代交通工具一个人射箭一个人站着拿着长矛											

(b) 文本到图像检索

图13 鲁棒性可视化实验图

Fig. 13 Visualization experiment diagram of data robustness((a)img2text retrieval;(b)text2img retrieval)

验结果分析可以得知本文模型具有良好的鲁棒性。

3 结 论

针对汉画像石样本较少、内容丰富多彩的特点, 本文提出了一种语义增强与文本重组协同的汉画像石图文检索模型。引入了语义增强模块, 通过掩膜图像提供的语义信息计算得出注意力特征和增强特征, 通过多阶段特征融合得到组合特征, 在保留原始特征的基础上又突出关键特征。此外, 基于本文数据集的特殊性设置了相应的温度模块, 通过图文相似度计算温度, 提高了跨模态检索的全面性与准确性。通过引入文本重组模块, 利用从掩膜图像中提取出的有效语义信息和随机规范提示模板对文本进行重组, 这样既丰富了文本的语义表达, 又增强了文本的规范性。实验表明, 本文模型较原始模型在精确性和全面性以及鲁棒性上有显著提升。然而, 当图像中的对象比较相似但方位信息不一致时, 图文检索的效果可能会出现混淆与误检。此外, 由于汉画像石图文检索的数据集的规模与多样性有限, 模型跨模态检索的性能可能会受到约束。未来, 将扩充汉画像石数据集, 并且优化模型对图像中对象的不同位置与动作进行语义分析的能力。

参考文献(References)

- Chen W, Liu Y, Wang W P, Bakker E M, Georgiou T, Fieguth P, et al. 2023. Deep learning for instance retrieval: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7270-7292 [DOI: 10.1109/TPAMI.2022.3218591]
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2021. An image is worth 16×16 words: transformers for image recognition at scale//*Proceedings of the 9th International Conference on Learning Representations*. [s.l.]: #ICLR
- Faghri F, Fleet D J, Kiros J R and Fidler S. 2018. VSE++: improving visual-semantic embeddings with hard negatives//*Proceedings of 2018 British Machine Vision Conference*. Newcastle, UK: BMVA Press: #12
- Gan S N. 2019. Study on the Dance Art in the Stone Relief of the Han Dynasty in Nanyang from the Perspective of Imagology. Zhengzhou: Zhengzhou University (甘胜楠. 2019. 图像学视域下南阳汉画像石中的舞蹈艺术研究. 郑州: 郑州大学)
- Ge X R, Xu S P, Chen F H, Wang J, Wang G X, An S, et al. 2024. 3SHNet: boosting image-sentence retrieval via visual semantic-spatial self-highlighting. *Information Processing and Management*, 61(4): #103716 [DOI: 10.1016/j.ipm.2024.103716]
- Hawking D, Voorhees E M, Craswell N and Bailey P. 1999. Overview of the TREC-8 web track//*Proceedings of 1999 Eighth Text Retrieval Conference*. Gaithersburg, USA: NIST
- He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]
- Jang E, Gu S X and Poole B. 2017. Categorical reparameterization with gumbel-softmax//*Proceedings of the 5th International Conference on Learning Representations*. Toulon, France: ICLR
- Jia C, Yang Y F, Xia Y, Chen Y T, Parekh Z, Pham H, et al. 2021. Scaling up visual and vision-language representation learning with noisy text supervision//*Proceedings of the 38th International Conference on Machine Learning*. [s.l.]: JMLR.org: 4904-4916
- Jiang D and Ye M. 2023. Transformer network for cross-modal text-to-image person re-identification. *Journal of Image and Graphics*, 28(5): 1384-1395 (姜定, 叶茫. 2023. 面向跨模态文本到图像行人重识别的Transformer网络. *中国图象图形学报*, 28(5): 1384-1395) [DOI: 10.11834/jig.220620]
- Kim W, Son B and Kim I. 2021. ViLT: vision-and-language transformer without convolution or region supervision//*Proceedings of the 38th International Conference on Machine Learning*. [s.l.]: JMLR.org: 5583-5594
- Lee K H, Chen X, Hua G, Hu H D and He X D. 2018. Stacked cross attention for image-text matching//*Proceedings of the 15th European Conference on Computer Vision*. Munich, Germany: Springer: 212-228 [DOI: 10.1007/978-3-030-01225-0_13]
- Li J N, Li D X, Savarese S and Hoi S. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models//*Proceedings of the 40th International Conference on Machine Learning*. Honolulu, USA: JMLR.org: #814
- Liang Z Y, Liang M Y, Huang W, Li Y W, Liu W, Shao Y X, et al. 2025. Dynamic self-adaptive multiscale distillation from pre-trained multimodal large model for efficient cross-modal retrieval//*Proceedings of the 33rd ACM International Conference on Multimedia*. Dublin, Ireland: ACM: 2851-2859 [DOI: 10.1145/3746027.3754844]
- Lipton Z C, Berkowitz J and Elkan C. 2015. A critical review of recurrent neural networks for sequence learning [EB/OL]. [2025-10-28]. <https://arxiv.org/pdf/1506.00019.pdf>
- Liu Y H, Ott M, Goyal N, Du J F, Joshi M, Chen D Q, et al. 2019. RoBERTa: a robustly optimized BERT pretraining approach [EB/OL]. [2025-10-28]. <https://arxiv.org/pdf/1907.11692.pdf>
- Loshchilov I and Hutter F. 2019. Decoupled weight decay regularization//*Proceedings of the 7th International Conference on Learning Representations*. New Orleans, USA: ICLR
- Mao A Q, Mohri M and Zhong Y T. 2023. Cross-entropy loss functions:

- theoretical analysis and applications//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 23803-23828
- Qin Y, Peng D Z, Peng X, Wang X and Hu P. 2022. Deep evidential learning with noisy correspondence for cross-modal retrieval//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 4948-4956 [DOI: 10.1145/3503161.3547922]
- Qu C and Zhai J X. 2009. Analysis of Han dynasty stone reliefs in Xuchang. Legal System and Society, (7): #311 (渠超, 翟京襄. 2009. 浅析许昌汉画像石. 法制与社会, 7: #311) [DOI: 10.19387/j.cnki.1009-0592.2009.07.201]
- Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: JMLR.org: 8748-8763
- Ren S Q, He K M, Girshick R and Sun J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks//Proceedings of the 29th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 91-99
- Smeulders A W M, Worring M, Santini S, Gupta A and Jain R. 2000. Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12): 1349-1380 [DOI: 10.1109/34.895972]
- Song W T, Liao L Y, Zhang H T, Li L, Yu T X, Zhao Y S, et al. 2025. Applications and prospects of artificial intelligence in the cultural heritage. Journal of Image and Graphics, 30(12): 3707-3739 (宋维涛, 廖聆宇, 张浩天, 李琳, 俞天秀, 赵永生, 等. 2025. 人工智能在文物行业的应用与展望. 中国图象图形学报, 30(12): 3707-3739) [DOI: 10.11834/jig.240765]
- van den Oord A, Li Y Z and Vinyals O. 2019. Representation learning with contrastive predictive coding [EB/OL]. [2025-10-28]. <https://arxiv.org/pdf/1807.03748.pdf>
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010
- Wang J K, Chen D D, Wu Z X, Luo C, Zhou L W, Zhao Y C, et al. 2022. OmniVL: one foundation model for image-language and video-language tasks//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #412
- Wang W H, Bao H B, Dong L, Bjorck J, Peng Z L, Liu Q, et al. 2023. Image as a foreign language: BEIT pretraining for vision and vision-language tasks//Proceedings of 2023 Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 19175-19186 [DOI: 10.1109/CVPR52729.2023.01838]
- Yang A, Pan J S, Lin J Y, Men R, Zhang Y C, Zhou J R, et al. 2023. Chinese CLIP: contrastive vision-language pretraining in Chinese [EB/OL]. [2025-10-28]. <https://arxiv.org/pdf/2211.01335.pdf>
- Yin Q Y, Huang Y, Zhang J G, Wu S and Wang L. 2021. Survey on deep learning based cross-modal retrieval. Journal of Image and Graphics, 26(6): 1368-1388 (尹奇跃, 黄岩, 张俊格, 吴书, 王亮. 2021. 基于深度学习的跨模态检索综述. 中国图象图形学报, 26(6): 1368-1388) [DOI: 10.11834/jig.200862]
- Zeng Y, Zhang X S, Li H, Wang J W, Zhang J P and Zhou W C S. 2024. X²-VLM: all-in-one pre-trained model for vision-language tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(5): 3156-3168 [DOI: 10.1109/TPAMI.2023.3339661]

作者简介

姜坤,男,硕士研究生,主要研究方向为深度学习和计算机视觉。E-mail: 2476270892@qq.com

钱文华,通信作者,男,教授,博士生导师,主要研究方向为计算机视觉和文化计算。E-mail: whqian@ynu.edu.cn

刘朋,男,博士研究生,主要研究方向为计算机视觉和图像处理。E-mail: pengliu0606@gmail.com