硕士专业学位论文

基于多模态特征融合与对比学习的条件图像检索方法研究

Research on Composed Image Retrieval Methods Based on Multimodal Feature Fusion and Contrastive Learning

作者：梁明秋

导师：卢思洋

北京交通大学

2025 年 10 月

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通大学可以将学位论文的全部或部分内容编入有关数据库进行检索，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

学校代码：10004 密级：公开

北京交通大学

硕士专业学位论文

基于多模态特征融合与对比学习的条件图像检索方法研究

Research on Composed Image Retrieval Methods Based on Multimodal Feature Fusion and Contrastive Learning

作者姓名：梁明秋 学　　号：25125289

导师姓名：卢思洋 职　　称：副教授

专业学位类别（领域）：计算机技术 学位级别：硕士

北京交通大学

2025 年 10 月

致谢

摘要

随着互联网多媒体数据的爆炸式增长，图像检索技术在电子商务、智能推荐、视觉问答等场景中发挥着日益重要的作用。传统图像检索主要依赖文本关键词或视觉内容进行独立查询，难以满足用户对“以图搜图并按描述修改”的精细化需求。为此，条件图像检索（Composed Image Retrieval, CIR）应运而生，其目标是根据给定的参考图像和自然语言描述（如“将红色连衣裙改为蓝色”），从大规模图像库中准确检索出符合语义修改的目标图像。该任务要求模型同时理解视觉与语言模态，并实现跨模态语义对齐与细粒度特征融合，具有极高的挑战性。

本文围绕条件图像检索中的多模态表征学习问题展开深入研究，针对现有方法在跨模态交互不充分、细粒度差异捕捉能力弱以及模型泛化性能不足等问题，提出了一系列创新性的解决方案。首先，为提升图文联合表征能力，本文设计了一种基于注意力机制的跨模态融合网络，通过引入视觉-语言交叉注意力模块，实现了图像区域与文本词元之间的动态对齐，增强了关键语义信息的提取能力。其次，针对真实场景中图像长宽比多样导致的特征失真问题，本文提出一种自适应图像预处理策略，在保持原始比例的基础上进行定向填充，有效提升了 CLIP 等预训练模型在非规则图像上的迁移性能。进一步地，为增强模型对相似负样本的判别力，本文构建了一个基于对比学习的组合网络（Combiner Network），将参考图像与文本描述编码后的特征进行非线性融合，并通过对比损失函数优化检索空间，显著提高了细粒度修改的识别精度。

本研究在多个公开基准数据集上进行了系统实验，包括 FashionIQ、CIRR 等具有代表性的条件图像检索数据集。实验结果表明，所提方法在 Recall@K 和 RecallSubset@K 等多项指标上均优于当前主流方法，尤其在低秩召回率（如 R@1）方面表现出显著优势，验证了模型在复杂语义理解和高相似度干扰下的鲁棒性。此外，本文还开发了交互式演示系统，支持用户输入任意图文对进行实时检索，展示了方法的实际应用潜力。

本论文的研究不仅推动了多模态学习与细粒度图像检索领域的发展，也为智能电商、个性化推荐等实际应用场景提供了高效的技术支撑，具有重要的理论价值与现实意义。

关键词：条件图像检索；多模态融合；对比学习；跨模态注意力；CLIP

ABSTRACT

序言

1

目录

1 引言

1.1 研究背景与意义

在当今数字化时代，图像已成为人们表达意图、传递信息的重要媒介。随着社交媒体、电商平台和智能设备的普及，海量图像数据被持续产生与共享。如何高效、精准地从这些数据中获取所需内容，成为计算机视觉领域的核心课题之一。传统的图像检索技术主要依赖于文本标签或视觉相似性进行匹配，然而这类方法在面对复杂语义查询时往往力不从心。例如，用户希望“将这件红色连衣裙换成蓝色款式”，仅靠关键字“蓝色连衣裙”可能无法准确反映其真实需求，而单纯基于视觉相似性的搜索也无法体现颜色替换这一语义操作。

为解决上述问题，条件图像检索（Composed Image Retrieval, CIR）作为一种新兴的多模态任务逐渐受到关注。该任务旨在结合一张参考图像和一段自然语言描述，生成一个“合成查询”，并据此检索出最符合语义修改的目标图像。它不仅要求模型具备强大的视觉理解能力，还需能够解析自然语言指令，并将其与图像内容深度融合，实现跨模态的语义推理。这一任务在时尚推荐、虚拟试衣、智能家居等领域具有广泛的应用前景。

近年来，尽管已有诸多工作尝试提升 CIR 系统的性能，如 TIRG、FiLM、MAAF 等模型通过门控机制、特征调制等方式实现图文融合，但仍存在若干关键挑战：一是现有方法在跨模态交互过程中往往采用简单的拼接或加权方式，缺乏深层次的语义对齐；二是面对图像尺寸不一、背景复杂的真实场景，预训练模型（如 CLIP）的迁移效果受限；三是多数方法忽视了对高相似负样本的判别能力，导致在细粒度修改（如颜色、纹理、部件增减）上表现不佳。

因此，本文聚焦于多模态特征融合与对比学习机制的设计，致力于构建一个更加鲁棒、精细且可扩展的条件图像检索框架。通过引入先进的注意力机制、优化图像预处理流程，并设计高效的组合网络结构，本文旨在提升模型对图文联合语义的理解能力，特别是在处理细微视觉变化方面的表现。研究成果不仅有助于推动多模态学习理论的发展，也为构建智能化、个性化的视觉搜索引擎提供了坚实的技术基础。

1.2 国内外研究现状

1.3 研究内容与工作

1.4 论文组织结构

本论文围绕基于多模态特征融合与对比学习的条件图像检索方法展开研究，旨在提升模型在细粒度语义修改下的检索精度与鲁棒性。全文共分为五章，各章节内容安排如下：

第 1 章 引言。本章首先阐述条件图像检索的研究背景与现实意义，分析其在智能电商、个性化推荐等场景中的应用价值。随后，系统综述多模态图像检索、条件图像检索任务以及视觉-语言预训练模型的研究现状，指出现有方法在跨模态交互、特征对齐和泛化能力方面存在的挑战。在此基础上，明确本文的主要研究内容与创新贡献，并给出全文的组织结构。

第 2 章 相关理论与技术基础。本章为后续方法研究提供理论支撑。首先，形式化定义条件图像检索任务，并介绍 FashionIQ、CIRR 等常用数据集及 Recall@K 等评价指标。接着，系统介绍多模态表示学习的基本框架，包括视觉与文本特征提取方法。重点阐述 CLIP 等视觉-语言预训练模型的工作原理，以及注意力机制、Transformer 架构、对比学习等核心技术，为后续模型设计奠定基础。

第 3 章 基于多模态融合的条件图像检索方法。本章提出本文的核心方法。首先介绍模型的整体架构，然后依次阐述三个关键技术模块：针对图像比例失真的自适应预处理策略、实现图文深度交互的跨模态特征融合模块（含交叉注意力机制），以及用于生成联合表征的组合网络（Combiner Network）与对比学习优化策略，形成一套完整的条件图像检索解决方案。

第 4 章 实验结果与分析。本章通过实验验证所提方法的有效性。首先介绍实验设置，包括数据集、评价指标和实现细节。随后，在 FashionIQ 和 CIRR 数据集上进行消融实验和对比实验，分析各模块的贡献并评估模型性能。

第 5 章 总结与展望。本章对全文研究工作进行系统总结，凝练主要研究成果与创新点，并指出当前方法在复杂语义理解、多轮交互等方面存在的局限性，对未来可能的研究方向进行展望。

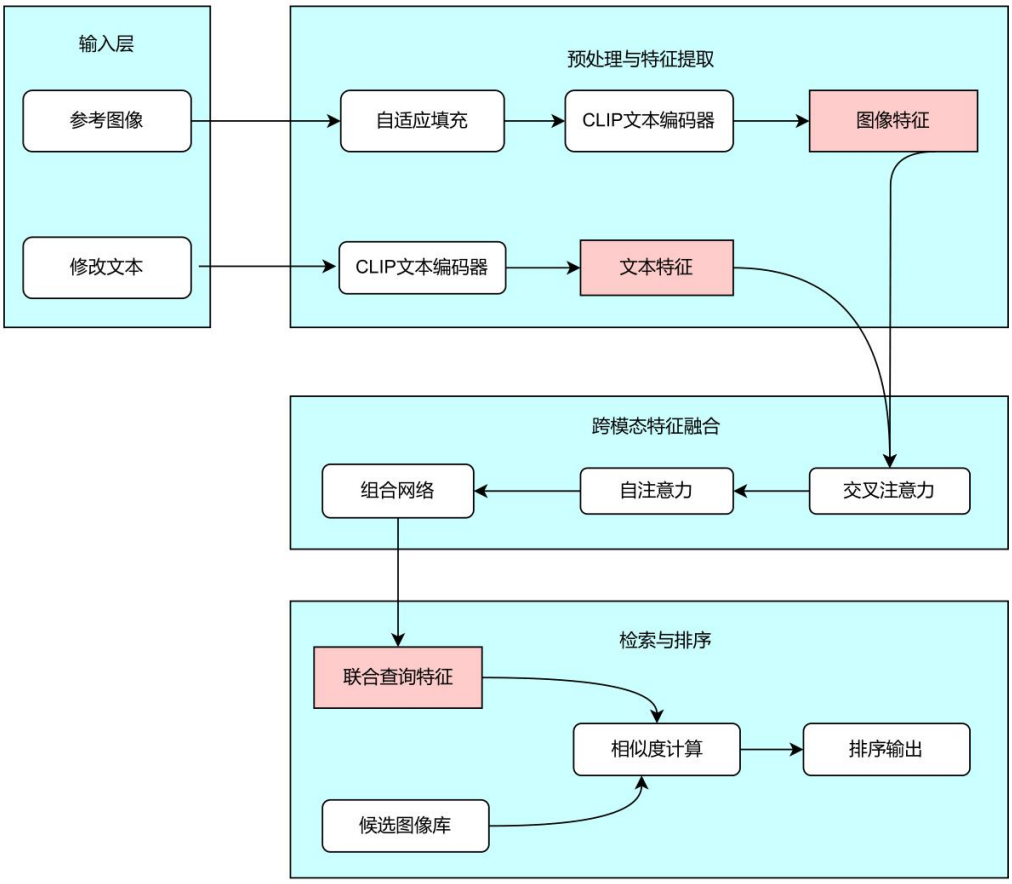
输入层
参考图像
修改文本
预处理与特征提取
自适应填充
CLIP文本编码器
图像特征
CLIP文本编码器
文本特征
跨模态特征融合
组合网络
自注意力
交叉注意力
检索与排序
联合查询特征
相似度计算
排序输出
候选图像库

图 1 整体结构图

2 相关理论与技术

2.1 条件图像检索任务定义

2.1.1 任务形式化描述

2.1.2 常用数据集与评价指标

2.2 多模态表示学习基础

2.2.1 视觉特征提取

2.2.2 文本特征编码

2.3 视觉-语言预训练模型（CLIP）

2.4 注意力机制与 Transformer 架构

2.5 对比学习与度量学习方法

2.6 本章小结

3 基于多模态融合的条件图像检索方法

3.1 模型整体架构设计

3.2 自适应图像预处理策略

3.2.1 图像长宽比问题分析

3.2.2 定向填充与 CLIP 适配优化

3.3 跨模态特征融合模块

3.3.1 视觉与文本编辑器

3.3.2 交叉注意力机制设计

3.4 组合网络与对比学习优化

3.4.1 Combiner Network 结构

3.4.2 损失函数设计：对比损失与训练策略

3.5 本章小结

4 实验结果与分析

4.1 实验环境及数据集

4.2 消融实验与结果分析

4.3 本章小结

5 总结与展望

5.1 总结

5.2 展望

参考文献

[内容为五号宋体。] 参考文献是文中引用的有具体文字来源的文献集合。按照 GB 7714《文后参考文献著录规则》的规定执行。

参考文献以文献在整个论文中出现的次序用[1]、[2]、[3]……形式统一排序、依次列出。

参考文献的表示格式为:

著作：[序号]作者.译者.书名[M].版本(第一版不著录).出版地:出版社,出版时间:引用部分起止页.

期刊：[序号]作者.译者.文章题目[J].期刊名,年份,卷号(期数):引用部分起止页.

会议论文集：[序号]作者.译者.文章名[C]. //编者.论文集名,会议地址，会议时间.出版地:出版者，出版年.引用部分起止页.

学位论文：[序号]作者.题名[D].保存地点:保存单位,年份.引用部分起止页.

专利：[序号]专利申请者.专利文献题名[P].国别,专利文献种类,专利号.发布日期:引用部分起止页.

技术标准：[序号]起草责任者.标准代号.标准顺序号——发布年.标准名称.出版地.出版者.出版年份:引用部分起止页.

报纸: [序号]作者.题名[N].报纸名，出版日期(版次)

附录 A

[附录标题]

[内容为五号宋体。] 附录是作为论文主体的补充项目，并不是必须的。

论文的附录依序用大写正体英文字母 A、B、C……编序号，如：附录 A。

索引

[内容为五号宋体。] 按照需要编排分类索引、著者索引、关键词索引等。

作者简历及攻读硕士/博士学位期间取得的研究成果

[内容采用五号宋体]　包括教育经历、工作经历、攻读学位期间发表的论文和完成的工作等。行距 16 磅，段前后各为 0 磅。

一、作者简历

二、发表论文

[1]

[2]

[3]

.

.

.

三、参与科研项目

[1]

[2]

[3]

.

.

.

四、专利

[1]

[2]

[3]

.

.

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名： 签字日期： 年 月 日

学位论文数据集

表 1.1： 数据集页

关键词*	密级*	中图分类号	UDC	论文资助
学位授予单位名称*		学位授予单位代码*	学位类别*	学位级别*
北京交通大学		10004		
论文题名*		并列题名*		论文语种*
作者姓名*			学号*	
培养单位名称*		培养单位代码*	培养单位地址	邮编
北京交通大学		10004	北京市海淀区西直门外上园村 3 号	100044
专业学位类别（领域）*		研究方向*	学制*	学位授予年*
论文提交日期*				
导师姓名*			职称*	
评阅人	答辩委员会主席*		答辩委员会成员	
电子版论文提交格式 文本（ ） 图像（ ） 视频（ ） 音频（ ） 多媒体（ ） 其他（ ） 推荐格式：application/msword；application/pdf				
电子版论文出版（发布）者		电子版论文出版（发布）地		权限声明
论文总页数*				
共 33 项，其中带*为必填数据，为 21 项。				