曹坤. 基于语义增强和动态关系感知的图文匹配方法[J]. 智能计算机与应用,2025,15(12):131-136. DOI:10.20169/j.issn.2095-2163.25070102

基于语义增强和动态关系感知的图文匹配方法

曹 坤

(青岛大学 计算机科学技术学院, 山东 青岛 266071)

摘 要: 在当前图文匹配领域中,主流的跨模态匹配方法主要分为基于全局特征的匹配方法与基于局部特征的细粒度匹配方法,前者侧重于对图像和文本整体语义进行对齐,能够捕捉宏观语义信息,但往往在关键细节的表达上存在不足;后者则注重细粒度特征的对齐,能够较好地挖掘图像和文本中的局部细节,然而其全局语义理解能力相对有限,导致在处理语义相似的困难负样本时表现欠佳。尽管上述方法在目前的图文匹配任务中已取得一定成效,但仍面临图文特征提取效果不佳、细节感知能力不足、噪声干扰等诸多挑战,进而本文提出一种基于语义增强与动态关系感知的图文匹配方法,通过构建模态内及模态间的关系感知机制,对全局特征和局部特征进行特征融合,实现更精准的语义对齐。在 Flickr30K 数据集上的实验结果表明,该方法显著提升了图文匹配的准确性。

关键词: 图文匹配;特征融合;语义增强;动态关系感知;跨模态匹配

中图分类号: TP391.41 **文献标志码**: A **文章编号**: 2095-2163(2025)12-0131-06

Image-text matching method based on semantic enhancement and dynamic relation awareness

CAO Kun

(College of Computer Science & Technology, Qingdao University, Qingdao 266071, Shandong, China)

Abstract: In the current field of image-text matching, mainstream cross-modal approaches can be broadly categorized into global feature-based methods and fine-grained local feature-based methods. The former focuses on aligning the overall semantics of images and texts, effectively capturing macro-level information, but often lacks precision in representing key details. In contrast, the latter emphasizes fine-grained feature alignment, which better captures local details within both modalities, yet tends to have limited global semantic understanding, resulting in suboptimal performance when dealing with semantically similar hard negatives. Although these methods have achieved notable progress in image-text matching tasks, they still face several challenges, including subpar feature extraction, insufficient fine-grained perception, and noise interference. To address these issues, this paper proposes an image-text matching method based on semantic enhancement and dynamic relation awareness. By constructing intra- and inter-modal relationship perception mechanisms, it integrates global and local features to achieve more precise semantic alignment. Experimental results demonstrate that the method significantly improves matching accuracy on the Flickr30K dataset.

Key words: image-text matching; feature fusion; semantic enhancement; dynamic relation perception; cross-modal matching

0 引 言

随着互联网和移动技术的发展,图文匹配技术已广泛应用于广告推荐、医疗图像检索、社交媒体分析等领域。图文匹配对于信息检索、自动标注以及跨模态内容理解至关重要。然而,传统的图文匹配方法往往依赖单一模态特征,导致匹配的准确性和鲁棒性有限[1-3]。面对海量的图像和文本数据,如何实现高效且精确的图文匹配,成为亟待解决的问题。

目前,图文匹配研究主要集中在全局特征匹配和局部特征匹配两类方法。全局特征匹配侧重于捕捉图像或文本的整体语义[4]。Kiros 等[5]通过结合卷积神经网络(Cnvolutional Neural Network, CNN)和长短期记忆网络(Long Short-Term Memory, LSTM)来提取图像和文本的全局特征;Faghri 等[6]则提出了改进的视觉语义嵌入模型(Visual Senmantic Embedding++, VSE++),并引入困难负样本以增强模型的约束能力。但该类方法忽视了图像细粒度语义信息,易导致匹配

作者简介: 曹 坤(1995—),男,硕士研究生,主要研究方向:计算机视觉。Email:838149033@qq.com。

收稿日期: 2025-07-01

失真。为了弥补这一缺陷,基于局部特征的匹配方法应运而生,如堆叠交叉注意力网络(Stacked Cross Attention Network, SCAN)和上下文感知注意网络(Context-Aware Attention Network, CAAN),通过细粒度语义的精确映射提升了匹配精度[7-9]。但这类方法仍难以实现模态间的全局语义对齐,尤其在处理困难负样本时存在局限性[10-11]。

尽管视觉语言 BERT(Vision-and-Language BERT, ViLBERT)等模型在通用数据集上表现优异,但在实际应用中仍受限于以下因素:图文特征提取效果不佳,导致关键目标区域的特征难以突出[12];跨模态特征对齐困难,文本描述和图像内容的细节可能存在偏差;噪声干扰,背景元素可能影响匹配的准确性[13]。为解决这些问题,本文提出一种基于语义增强和动态关系感知的图文匹配方法,结合全局与局部特征,利用关系感知模块进行模态内外语义关联,通过关系融合模块强化全局特征的语义表达,并引入知识蒸馏,最终实现精确的跨模态匹配。

1 基于语义增强和动态关系感知的图文匹配方法

本文提出的基于语义增强和动态关系感知的图文匹配方法包含 3 个主要模块:特征提取和增强模块、关系感知融合模块和知识蒸馏模块如图 1 所示。

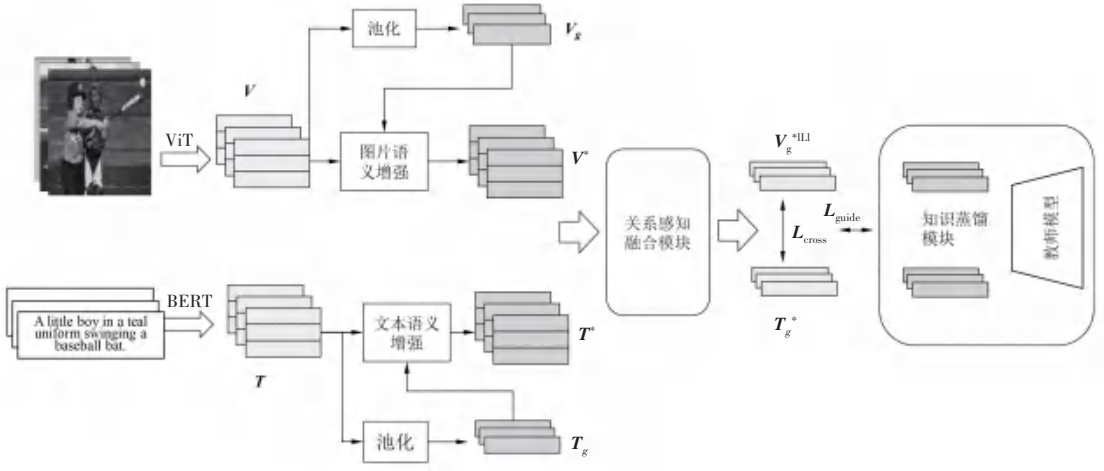

图 1 基于语义增强和动态关系感知的图文匹配方法

Fig. 1 Image-text matching method based on semantic enhancement and dynamic relation awareness

1.1 特征提取

1.1.1 图片特征提取

为充分表达图像信息并建立图像内的语义关联,本文采用预训练的 Vision Transformer 模型将输入图像划分为 n 个区域块,生成每个图片区域的特征表示 $q_i \in \mathbb{R}^{768}, i \in [1,n]$;通过全连接层的非线性变换,将特征映射到 d 维的嵌入空间,以便后续与文本特征进行计算和对比,变化公式如下式:

$$v_i = q_i W_v + b \tag{1}$$

其中,$W_v \in \mathbb{R}^{768\times d}$,$b$ 为可学习的参数。

此时每张图片的局部特征表示为 $V = \{v_i\}_{i=1}^{n} \in \mathbb{R}^{n\times d}$;将图片局部特征进行平均池化,得到其对应的全局特征:

$$v_g = \frac{1}{n}\sum_{i=1}^{n} v_i \tag{2}$$

1.1.2 文本特征提取

本文使用预训练的 BERT(Bidirectional Encoder Representations from Transformers, BERT)模型对文本特征进行提取。在 BERT 的输出层后引入一个全连接层,将提取的文本特征映射到与图片特征相同的 d 维特征空间中;文本局部特征经过平均池化,得到其对应的全局特征:

$$t_g = \frac{1}{m}\sum_{j=1}^{\mathrm{m}} t_j \tag{3}$$

其中,m 表示划分的文本区域数量,t 表示文本区域的特征向量。

1.2 语义增强模块

本文在图像的特征处理中,期望局部特征能够融入全局语义信息,以便细节特征也能体现出整体图像的语义。为此,本文采用一种基于自注意力机制的模态内语义增强方法,该方法将图像的全局特征作为查询(Query),与局部特征进行交互,进而生成融合了全局上下文信息的增强局部特征。具体来说,将图文的全局特征作为 Query,将局部特征分别作为键(Key)和值(Value),通过 Transformer 的自注意力机制,对 Query、Key 和 Value 进行线性变换,得到图片和文本的局部增强特征:

$$V^* = \mathrm{Softmax}\left(\frac{(\{v_g\}_{i=1}^{n}\cdot \boldsymbol{W}^Q)\cdot(\{v_i\}_{i=1}^{n}\cdot \boldsymbol{W}^K)^{\mathrm{T}}}{\sqrt{d}}\right)(\{v_i\}_{i=1}^{n}\cdot \boldsymbol{W}^V) \tag{4}$$

$$T^* = \mathrm{Softmax}\left(\frac{(\{t_g\}_{j=1}^{m}\cdot \boldsymbol{W}^Q)\cdot(\{t_j\}_{j=1}^{m}\cdot \boldsymbol{W}^K)^{\mathrm{T}}}{\sqrt{d}}\right)(\{t_j\}_{j=1}^{m}\cdot \boldsymbol{W}^V) \tag{5}$$

其中，$\boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V \in \mathbb{R}^{d\times d}$ 为可通过反向传播学习的参数矩阵。

1.3　关系感知融合模块

一般的图文匹配方法主要关注图像和文本对之间的一对一相似性，而忽略了不同图像和文本对之间潜在的语义关联。而通过显式学习图文对间的语义关系，有助于更全面地捕捉图文匹配中的细粒度信息。

本文首先构建了一个 $2N\times 2N$ 的关系感知矩阵 $\boldsymbol{H}$，N 表示同一批次图像和文本对的数量，用以表示在一个批次数据内图像和文本实例之间的语义关联程度。该矩阵划分为 4 个 $N\times N$ 的子块，分别表示图像 - 图像、图像 - 文本、文本 - 图像和文本 - 文本之间的语义关联[14]，如下式：

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_{\mathrm{I-I}} & \boldsymbol{H}_{\mathrm{I-T}} \\ \boldsymbol{H}_{\mathrm{T-I}} & \boldsymbol{H}_{\mathrm{T-T}} \end{bmatrix} \tag{6}$$

其中，$\boldsymbol{H}_{\mathrm{I-I}}$ 为图像 - 图像的模态内关系矩阵；$\boldsymbol{H}_{\mathrm{I-T}}$ 为图像 - 文本模态间的矩阵；$\boldsymbol{H}_{\mathrm{T-I}}$ 为文本 - 图像的模态间关系矩阵；$\boldsymbol{H}_{\mathrm{T-I}}$ 为文本的模态内关系矩阵。

通过建立图文数据模态内和模态间的关系感知，能够有效的补捉同批次内图像和文本之间的多层次语义关系，而每一个子矩阵受连接矩阵 $\boldsymbol{C}$ 和联系矩阵 $\boldsymbol{R}$ 的共同影响。连接矩阵 $\boldsymbol{C}$ 用于保留高于基准的连接而过滤掉关联度较弱的图文对；而联接矩阵 $\boldsymbol{R}$ 则用来表达图文之间的相似关联度[14]。

1.3.1　模态内关系感知

以图片模态内关系感知为例，为了衡量批次数据中各图像之间的内在关系，定义一个模态内的连接矩阵 $\boldsymbol{C}_{\mathrm{I-I}}$。元素 c_{xy} 表示第 x 张图片与该批次中第 y 张图片之间是否有关联关系，其定义如下式：

$$c_{xy} = \begin{cases} 1 & \text{if } s_{xy} \geqslant K_{\mathrm{th}}(\{s_{xy}\}_{y=1}^{N}) \\ 0 & \text{if } s_{xy} < K_{\mathrm{th}}(\{s_{xy}\}_{y=1}^{N}) \end{cases},\ x,y\in[1,N] \tag{7}$$

其中，s_{xy} 为第 x 张图片与第 y 张图片之间的连接度评估值，该值通过计算图片全局特征之间的点积得到；K_{th} 是基于集合 $\{S_{xy}\}_{y=1}^{N}$ 计算的阈值，引入该阈值是为了筛选出与图像 x 最相关的图像对，从而过滤掉那些连接度较低的关系。

引入图像间的联系矩阵 $\boldsymbol{R}_{\mathrm{I-I}}$，中元素 r_{xy} 用于表示第 x 张图片与第 y 张图片之间的关联程度，其定义如下式：

$$r_{xy} = e^{-\frac{\|\boldsymbol{v}_{x-g}-\boldsymbol{v}_{y-g}\|^2}{\sigma^2}} \tag{8}$$

其中，$\boldsymbol{v}_{x-g}$ 和 $\boldsymbol{v}_{y-g}$ 分别表示第 x 和第 y 张图片在嵌入空间中的全局特征向量；$\|\boldsymbol{v}_{x-g}-\boldsymbol{v}_{y-g}\|^2$ 表示欧氏距离；σ 为控制相似度衰减速率的超参数。

最终基于矩阵 $\boldsymbol{C}_{\mathrm{I-I}}$ 和矩阵 $\boldsymbol{R}_{\mathrm{I-I}}$ 构建如图 2 所示的图片模态内关系感知矩阵 $\boldsymbol{H}_{\mathrm{I-I}}$，其定义如下式：

$$(\boldsymbol{H}_{\mathrm{I-I}})_{xy} = \begin{cases} (\boldsymbol{R}_{\mathrm{I-I}})_{xy} & \text{if } (\boldsymbol{C}_{\mathrm{I-I}})_{xy} = 1 \\ -\infty & \text{if } (\boldsymbol{C}_{\mathrm{I-I}})_{xy} = 0 \end{cases} \tag{9}$$

同理可得文本模态内关系感知矩阵 $\boldsymbol{H}_{\mathrm{T-T}}$。

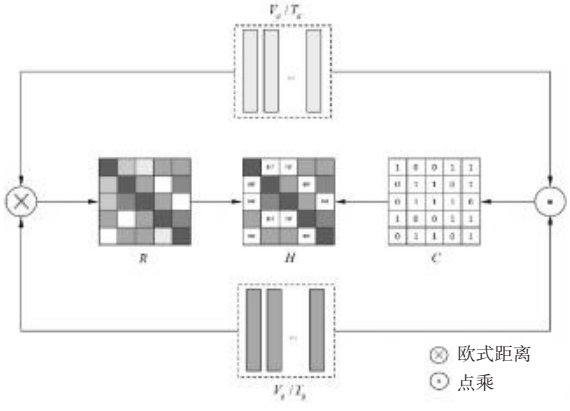

图 2　模态内关系感知

Fig. 2　Intra-modal relation awareness

1.3.2　模态间关系感知

与模态内关系感知不同，模态间关系的建模需要更细粒度的交互，例如图像与文本，旨在捕捉图片与文本之间的语义关联。

本文采用了基于局部区域匹配的方式进行模态间关系的量化。以图片到文本的模态间感知为例，将该批次的图片与文本之间具有较强相关性的连接关系进行保留，而较弱的连接则被舍弃，其连接矩阵 $\boldsymbol{C}_{\mathrm{I-T}}$ 内元素定义如下式：

$$c_{xy} = \begin{cases} 1 & \text{if } s_{xy} \geqslant K_{\mathrm{th}}(\{s_{xy}\}_{y=1}^{N}) \\ 0 & \text{if } s_{xy} < K_{\mathrm{th}}(\{s_{xy}\}_{y=1}^{N}) \end{cases},\ x,y\in[1,N] \tag{10}$$

其中，s_{xy} 代表图像中每个区域对文本中最相关特征的关注程度，即相似度的平均值，用以衡量该图像与该文本整体的关联程度，具体定义如下式：

$$s_{xy} = \frac{1}{n}\sum_{a=1}^{n}\max_{b\in[1,m]}[(\boldsymbol{v}_a^*)^{\mathrm{T}}\boldsymbol{t}_b^*] \tag{11}$$

其中，$\boldsymbol{v}_a^*$ 和 $\boldsymbol{t}_b^*$ 分别表示图像第 a 个区域和文

本第 b 个单词进行过语义增强后的局部特征向量，n 和 m 分别为图像区域和文本单词的数量，max 操作用于捕捉局部区域之间的最强匹配关系。

在跨模态关系建模中，本文从图像和文本中提取最相关区域的前 k 个相似度评分，并将其组成向量 $\boldsymbol{o}$，将该向量定义为图文间的关注向量，用于表示图像与文本之间联系最紧密的区域，如下式：

$$\boldsymbol{o}_{xy} = \mathrm{Top}K(\{\max_{b\in[1,m]}[(\boldsymbol{v}_a^*)^{\mathrm{T}}\boldsymbol{t}_b^*]\}_{a=1}^{n}) \quad (12)$$

其次，将关注向量 $\boldsymbol{o}$ 输入多层感知机（MLP），以学习生成标量 r，该标量用于衡量图像与文本之间的关联程度。对该批次所有图文对的交互计算可构建出图像 - 文本之间的联系矩阵 $\boldsymbol{R}_{\mathrm{I-T}}$。

最后，基于矩阵 $\boldsymbol{C}_{\mathrm{I-I}}$ 和矩阵 $\boldsymbol{R}_{\mathrm{I-I}}$ 构建如图 3 所示的模态内关系感知矩阵 $\boldsymbol{H}_{\mathrm{I-T}}$：

$$(\boldsymbol{H}_{\mathrm{I-T}})_{xy} = \begin{cases} (\boldsymbol{R}_{\mathrm{I-T}})_{xy} & \text{if } (\boldsymbol{C}_{\mathrm{I-T}})_{xy} = 1 \\ -\infty & \text{if } (\boldsymbol{C}_{\mathrm{I-T}})_{xy} = 0 \end{cases} \quad (13)$$

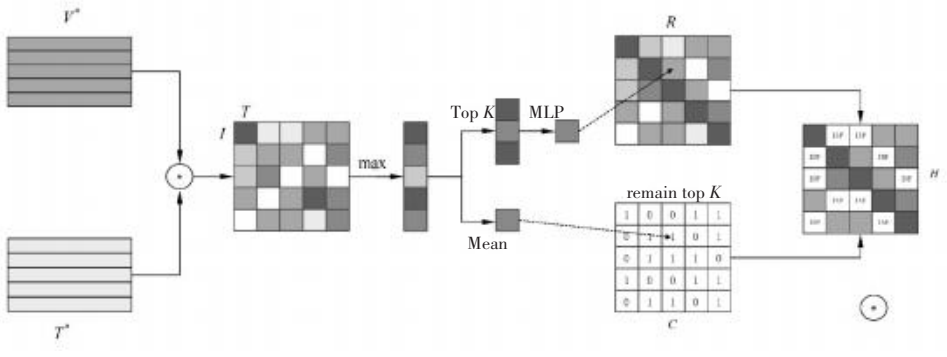

图 3 模态间关系感知

Fig. 3 Inter-modal relation awareness

同理可得文本模态间的关系感知矩阵 $\boldsymbol{H}_{\mathrm{T-I}}$。结合模态内关系感知矩阵和模态间关系感知矩阵，最终可得聚合后的关系感知矩阵 $\boldsymbol{H}$。此时，该矩阵不仅有效融合了图像与文本模态内部的特征关联信息，还同时捕获了跨模态语义交互以及不同实例间的潜在关联。

1.3.3 关系融合

为确保模态内和模态间的交互能够同时进行，将图文的全局特征进行拼接作为整体的输入特征。基于矩阵 $\boldsymbol{H}$ 引入多头注意力机制，实现对图文嵌入特征的自注意力感知[15]。通过堆叠若干层上述神经网络，得到了增强的图文全局嵌入特征 $\boldsymbol{v}_g^*$ 和 $\boldsymbol{t}_g^*$，如图 4 所示。

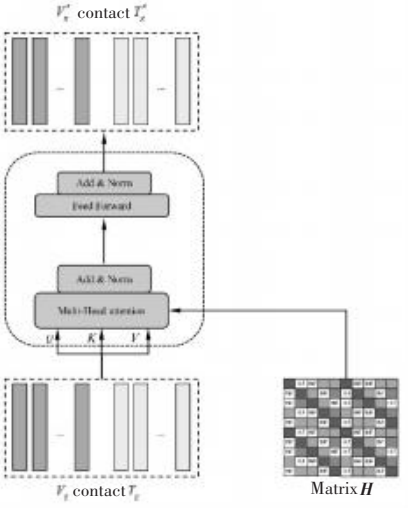

图 4 关系融合增强

Fig. 4 Relation fusion enhancement

1.4 知识蒸馏和损失函数

得到增强后的图文全局特征后，将经模型增强后的图文全局特征之间的相似度定义为得分 S：

$$S_g^{\mathrm{I-T}} = \frac{(\boldsymbol{v}_g^*)^{\mathrm{T}}\boldsymbol{t}_g^*}{\|\boldsymbol{v}_g^*\|\,\|\boldsymbol{t}_g^*\|}, S_g^{\mathrm{T-I}} = \frac{(\boldsymbol{t}_g^*)^{\mathrm{T}}\boldsymbol{v}_g^*}{\|\boldsymbol{v}_g^*\|\,\|\boldsymbol{t}_g^*\|} \quad (14)$$

为有效区分批量图文对中的困难负样本对，使用基于铰链的三元组损失函数，其定义如下式：

$$L_{\mathrm{cross}} = \max[0, \alpha - S(\mathrm{I},\mathrm{T}) + S(\mathrm{I},\hat{\mathrm{T}})] + \max[0, \alpha - S(\mathrm{I},\mathrm{T}) + S(\hat{\mathrm{I}},\mathrm{T})] \quad (15)$$

其中，超参数 α 用于控制模型对正负样本对区分的严格程度。

该损失函数从图文两个方向共同约束模型学习，从而更全面地捕捉图文之间的双向语义关联，进一步提升语义对齐的准确性。

为了进一步提升模型的学习能力，本文引入了知识蒸馏机制，利用基于对比文本-图像对的预训练模型（Contrastive Language-Image Pre-training，CLIP）跨模态对齐能力来指导下游任务的学习。另外，本文使用一种基于 KL（Kullback-Leibler）离散度的蒸馏损失函数，用于指导增强特征的学习。蒸馏损失函数定义，如下式：

$$L_{guide} = [\mathrm{KL}(\boldsymbol{R}_{I-T}^{teacher}, \boldsymbol{R}_{I-T}^{student}) + \mathrm{KL}(\boldsymbol{R}_{T-I}^{teacher}, \boldsymbol{R}_{T-I}^{student})]/2 \quad (16)$$

结合三元组损失和蒸馏损失，定义模型的融合损失函数，如下式：

$$L_{total} = (1-\lambda)L_{cross} + \lambda L_{guide} \quad (17)$$

其中，λ 是权重超参数，用于平衡三元组损失和蒸馏损失对总损失的贡献。

2　实验结果及分析

2.1　数据集及实验环境

Flickr30K 数据集包含 31 783 幅图像，每幅图像均配有 5 条文本描述[16]。本实验将其中 29 000 幅图像用于训练，1 000 幅作为验证，剩余 1 000 幅用于测试。

本文实验在 NVIDIA RTX4090 上运行，代码使用 PyTorch 框架，使用 Adam 优化器，批次大小为 128，训练周期为 20，初始学习率为 0.000 2，每 10 个周期缩减 0.1 倍。ViT 模型采用 12 层 Transformer Encoder 结构，每张图片被划分为 16×16 的 patch。图文特征维度 d 为 1 024，控制相似度衰减速率参数 σ 值为 1，损失函数参数 α 为 0.3，权重参数 λ 为 0.5。

2.2　实验结果分析

将本文提出的基于语义感知的图文匹配方法与全局匹配方法 VSE++[6] 以及局部匹配方法 SCAN[8]、CLIP[9] 在相同数据集上进行实验，结果见表 1。通过对比正确结果召回率（R_{sum}）指标可以发现，本文方法表现出了较好的图文匹配效果，相较于当前较为先进的 CLIP 方法提升了 3.6 个百分点。

表 1　实验比对结果

Table 1　Experimental comparison results

方法	图片-文本/%			文本-图片/%			R_{sum}/%
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++	51.6	79.1	84.3	41.3	70.2	78.9	405.4
SCAN	66.1	88.3	95.5	49.7	79.1	83.1	461.8
CLIP	76.3	**95.6**	97.5	56.9	83.3	88.2	497.8
本文	**77.4**	95.4	**97.6**	**57.1**	**84.1**	**89.8**	**501.4**

2.3　消融实验

为了评估语义增强模块、语义关系感知模块和知识蒸馏模块对方法性能的影响，本文设计了在 Flickr30K 数据集上的消融实验，结果见表 2。可发现引入全局特征与自注意力机制可增强局部语义表达，提升匹配精度；加入模态内外语义关系感知模块，有助于识别困难负样本，提高准确率；引入 CLIP 教师模型后，模型跨模态理解与特征表达能力显著增强。综上，本文方法在图文匹配准确性方面取得了一定提升。

表 2　消融实验

Table 2　Ablation experiment

编号	调节项				图像-文本/%			文本-图像/%			R_{sum}/%
	全局匹配	语义增强	关系感知	知识蒸馏	R@1	R@5	R@10	R@1	R@5	R@10	
1	√				54.9	76.8	80.1	43.3	66.1	77.1	398.3
2	√	√			63.1	80.7	87.3	49.5	69.7	81.8	432.1
3	√	√	√		73.1	87.2	94.1	55.9	77.8	87.7	475.8
4	√	√	√	√	**76.1**	**95.5**	**97.9**	**57.3**	**84.5**	**89.6**	**500.9**

3　结束语

本文提出一种融合语义增强与动态关系感知的图文匹配方法，结合全局与局部特征，通过语义增强模块优化模态内表达，利用关系感知与融合模块建模并强化图文间深层关联；引入预训练 CLIP 模型进行知识蒸馏，提升匹配精度。实验结果表明：该方法在图文匹配任务中表现优越，但仍面临模型复杂、计算开销大等问题，后续将致力于结构优化与性能提升。

参考文献

[1] 刘萌，齐孟津，詹圳宇，等. 基于深度学习的图像-文本匹配研究综述[J]. 计算机学报，2023，46(11)：2370-2399.

[2] 刘长红，曾胜，张斌，等. 基于语义关系图的张量融合网络的

图像文本检索[J]. 计算机应用, 2022, 42(10): 3018-3024.
[3] 张振兴, 王亚雄. 图文跨模态检索研究综述[J]. 北京交通大学学报, 2024, 48(2): 23-36.
[4] CHEN Jiacheng, HU Hexiang, WU Hao, et al. Learning the best pooling strategy for visual semantic embedding [C]// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2021: 15784-15793.
[5] KIROS R, SALAKHUTDINOV R, ZEMEL R S. Unifying visual-semantic embeddings with multimodal neural language models[J]. arXiv preprint arXiv, 1411.2539, 2014.
[6] FAGHRI F, FLEET D J, KIROS J R, et al. Vse++: Improving visual-semantic embeddings with hard negatives [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018:6951-6959.
[7] WANG Y, SU Y T, LI W H, et al. Rare-aware attention network for image-text matching [J]. Information Processing and Management, 2023, 60(3): 103280.
[8] LEE K H, CHEN X, HUA G, et al. Stacked cross attention for image-text matching [C]// Proceedings of the 15th European Conference on Computer Vision. Cham: Springer, 2018: 212-228.
[9] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of International Conference on Machine Learning. IMLS, 2021: 8748-8763.
[10] OTTO C, SPRINGSTEIN M, ANAND A, et al. Characterization and classification of semantic image-text relations [J]. International Journal of Multimedia Information Retrieval, 2020, 9(1): 31-45.
[11] ZHAO G S, ZHANG C F, SHANG H, et al. Generative label fused network for image-text matching [J]. Knowledge-Based Systems, 2023, 263: 110280.
[12] LU J, BATRA D, PARIKH D, et al. ViLBERT: Pretraining task-agnostic visiolinguistic representations for Vision-and-Language Tasks [C]// Proceedings of the 32nd Conference on Neural Information Processing Systems. NeurIPS, 2020: 8-14.
[13] GAO D H, JIN L B, CHEN B, et al. FashionBERT: Text and image matching with adaptive loss for cross-modal retrieval[C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2020: 2251-2260.
[14] FU Zheren, MAO Zhendong, SONG Yan, et al. Learning semantic relationship among instances for image-text matching [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2023: 15159-15168.
[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of Neural Information Processing Systems. NeurIPS. 2017: 5998-6008.
[16] WANG L, CERVANTES C M, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models [C]// Proceedings of the 2015 IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2015: 2641-2649.