计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(11): 3002−3016, 2026

基于共性-特性协同建模与多专家融合的多特征图像描述模型

杨建锋[1] 朱东坡[2] 曹 瑞[1] 崔选皓[1] 王 平[1] 郑 杰[1]
[1](西北大学计算机学院 西安 710127)
[2](北京三快科技有限公司 北京 100190)
(yjf@nwu.edu.cn)

A Multi-Feature Image Captioning Model with Common-Specific Collaborative Modeling and Mixture-of-Experts Fusion

Yang Jianfeng[1], Zhu Dongpo[2], Cao Rui[1], Cui Xuanhao[1], Wang Ping[1], and Zheng Jie[1]
[1](*School of Computer Science*, *Northwest University*, *Xi'an* 710127)
[2](*Beijing Sankuai Technology Co.*, *Ltd.*, *Beijing* 100190)

Abstract Existing image captioning methods often struggle to simultaneously capture global semantics and fine-grained details, leading to overly generic descriptions or missing key information. To address this issue, we propose EAFCC, short for expert-aware fusion with common-specific collaboration, a multi-feature image captioning model with mixture-of-experts fusion. The model takes grid features, region features, and segmentation features as parallel inputs. In the encoding stage, a pre-fusion module first performs implicit alignment of heterogeneous visual features and models cross-feature common information within a shared semantic space, while preserving the basis for feature-specific representations. A shared Transformer encoder is then employed to learn the common representations, whose consistency and discriminability are further enhanced by a contrastive learning constraint. Meanwhile, three independent encoding branches are used to model feature-specific representations, so as to retain fine-grained complementary information associated with different feature types. A gated routing mechanism inspired by mixture-of-experts is further introduced to dynamically weight and fuse the specific information. Finally, the enhanced common features and the fused specific features are collaboratively integrated, processed by a skip-connected encoder, and fed into a Transformer decoder to generate image captions. Experimental results on MS-COCO dataset show that EAFCC achieves competitive performance on the CIDEr-D metric and outperforms existing single-feature and dual-feature methods, demonstrating the effectiveness of common-specific collaborative modeling and dynamic multi-feature fusion for image captioning.

Key words image captioning; multi-feature fusion; common-specific collaborative modeling; contrastive learning; Transformer

摘 要 针对现有图像描述方法难以同时兼顾全局语义与细粒度细节、生成结果易偏泛化或关键信息遗漏的问题,提出一种基于共性-特性协同建模与多专家融合的多特征图像描述生成模型 EAFCC(expert-aware fusion with common-specific collaboration)。该模型并行接收网格特征、区域特征与分割特征,在编码阶段首先通过预融合模块对多源异构特征进行隐式对齐。在共享语义空间中提取跨特征一致的共性信息,同时为各特征流保留差异化表达基础。随后利用共享 Transformer 编码器建模共性表征,并通过对比

收稿日期: 2025-12-24; **修回日期**: 2026-07-23
基金项目: 国家自然科学基金项目(62406247)
This work was supported by the National Natural Science Foundation of China (62406247).
通信作者: 曹瑞(cr@nwu.edu.cn)

学习约束进一步增强其一致性与判别性。同时，采用 3 条独立编码分支对不同特征流进行特性建模，以保留与特征类型相关的细粒度补充信息，并借助多专家门控路由机制实现特性信息的动态加权融合。最后，将增强后的共性特征与融合后的特性特征进行协同整合，经跳跃连接编码器后送入 Transformer 解码器生成图像描述。在 MS-COCO 数据集上的实验表明，EAFCC 在 CIDEr-D 指标上取得了较优性能，并优于现有单特征或双特征方法，验证了共性与特性联合建模的有效性。

关键词 图像描述；多特征融合；共性-特性协同建模；对比学习；Transformer

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202550936 **CSTR:** 32373.14.issn1000-1239.202550936

图像描述生成是计算机视觉与自然语言处理的交叉任务，其目标为根据给定图像自动生成准确且符合人类认知的自然语言描述，涵盖图像特征提取、图像理解、特征融合等关键环节[1]。随着深度学习技术的快速发展，虽然模型结构不断更新，但生成质量的瓶颈往往来自视觉信息供给不足与融合不充分，因此输入特征的选择及其融合策略逐渐成为决定描述准确性与细节丰富度的核心因素。

在早期的主流范式中，主要依赖目标检测网络提取区域特征作为图像信息[2]。区域特征以对象级表示刻画图像中的具体事物，显著降低视觉嵌入复杂度，有效提升描述准确性。然而，区域特征的对象级特性导致其缺乏全局视觉理解与细粒度信息，限制了描述的丰富性：生成结果常仅涵盖图像主体而忽视背景，且描述长度偏短。

为弥补区域特征的全局信息缺失，部分学者引入网格特征，利用其对整幅图像的全局概括能力补充细粒度与背景信息[3-4]。但网格特征与区域特征在特征粒度与语义空间上难以对齐，且网格特征的碎片化表达可能与区域特征产生信息冲突，导致模型语义混淆。

为进一步缓解上述 2 类特征的互补缺陷，本文在区域特征与网格特征基础上引入语义分割特征[5]。该特征以粗粒度语义边界刻画场景布局，可同时弥补区域特征背景信息不足与网格特征信息碎片化问题，使视觉信息更为全面、准确，为后续描述生成提供坚实基础。

基于上述分析，本文借鉴因式分解思想，将多源视觉表征拆分为跨特征共享的共性与各特征特有的特性，并据此提出多专家多特征共性与特性提取融合的图像描述模型 EAFCC（expert-aware fusion with common-specific collaboration）。针对多特征输入下的场景语义理解不足、特征融合冲突与冗余以及细节刻画不足等问题，本文分别引入预融合（pre-fusion, PF）模块、共性-特性协同建模与对比学习、多专家融合机制以及跳跃连接编码与解码策略，从而提升多特征建模能力与描述生成性能。为直观说明本文方法的设计动机及各模块之间的逻辑关系，图 1 给出了“问题-方法-效果”的整体对应框架。模型首先并行输入网格特征、区域特征与分割特征；随后通过预融合模块隐式对齐 3 类特征的语义空间，并在拉近特征距离的同时保留各特征独有信息；接着利用共享 Transformer[6] 编码器提炼跨特征的共性表征，并引入困难负样本对比学习策略进一步增强共性特征之间的一致性；针对各特征独有的特性信息，采用 3 条参数独立的 Transformer 编码器分别提取，并借鉴多专家系统的门控路由机制动态加权融合特性信息；最后，将融合后的特性特征与增强后的共性特征拼接，经跳跃连接编码器后送入 Transformer 解码器完成描述生成，以减少编码过程中的信息丢失。

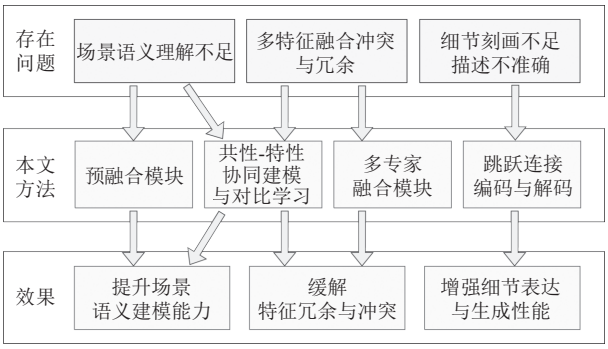


Fig. 1 The solutions for key issues in multi-feature image captioning

图 1 针对多特征图像描述关键问题的解决思路

本文的主要贡献包括 4 个方面。

1）提出共性-特性协同建模的多特征融合框架。针对单特征或拼接式融合难以兼顾一致性与互补性的问题，提出共性-特性协同建模的多特征融合框架：共性建模跨特征语义一致性，特性建模细粒度互补细节，并通过门控机制实现自适应融合。

2)设计基于对比学习的共性增强机制。引入困难负样本对比学习，在共享 Transformer 编码器输出的共性特征空间中，强化不同特征间语义一致性，抑制跨模态差异，提升共性表征的判别力与鲁棒性。

3)提出多专家门控路由的特性融合策略。借鉴混合专家网络(mixture-of-experts, MoE)思想，将各特征的特性表示视为“专家”，通过可学习的门控网络动态分配权重，实现特性信息的自适应加权融合，避免冗余，突出差异。

4)构建参数量上更具优势的高性能图像描述模型 EAFCC。在无需大规模预训练数据的前提下，于 MS-COCO 数据集上取得 135.2 的 CIDEr-D 分数，超越现有单特征、双特征及多数预训练模型，验证“小数据+细粒度架构”在资源受限场景下的有效性。

1 相关工作

国内外图像描述研究已由早期 CNN+RNN 模板或检索范式全面转向端到端 Transformer 架构，并在特征融合、预训练与轻量化等方向持续深入。现有工作聚焦单特征增强与大规模视觉-语言预训练[7]，国内近年来则提出跨层动态聚合、双重路径视觉-语言、多级语义对比、知识提示及蒸馏-压缩框架等创新，在 MS-COCO 数据集上 CIDEr-D 指标达到了 135.2，并兼顾中文场景与边缘部署需求。然而，现有方法仍普遍采用网格-区域双特征，对共性-特性协同建模与分割特征利用不足，本文所研究的 EAFCC 模型针对此空白展开系统研究。

1.1 基于 Transformer 编解码器结构的图像描述算法

图像描述研究已经由早期基于模板、检索和 CNN-RNN 的方法，逐步发展到以 Transformer 为核心的编解码框架[8-9]。在该方向上，研究人员分别从注意力分配优化[10]、层间特征交互[11]、记忆增强[12]以及视觉-语言上下文联合建模[13]等方面改进了 Transformer 图像描述框架，从而提升了模型对复杂视觉语义和语言依赖关系的表征能力。DLCT[14]，GRIT[15]和 DIFNet[16]进一步说明，Transformer 框架下的信息组织方式已成为影响性能的关键因素。与此同时，视觉语言预训练模型如 BLIP[17]，OFA[18]，OFASys[19]，SimVLM[20]，UNIMO[21]，Unified VLP[22]推动了图像描述与更广义视觉语言生成的融合。近 2 年，该方向又进一步延伸到检索增强和可控描述等新范式，例如 EVCap[23]利用外部视觉名称记忆(visual-name memory)增强开放世界描述能力，FineCaption[24]则支持任意粒度区域与掩码控制下的组合式描述生成。

1.2 多视觉特征表示在图像描述中的应用

图像描述模型的性能高度依赖输入视觉特征的质量与互补性。现有方法常使用区域特征、网格特征以及分割特征等不同表征。区域特征通常由目标检测器提取，典型如文献 [25] 所依赖的检测式区域表示，其优势在于突出显著目标及局部属性；网格特征更强调全局空间布局与上下文关系；分割特征则能够补充像素级语义区域与边界信息，相关提取网络如 Faster R-CNN[26]和 UPSNet[27]分别对应区域级和全景/分割级表征来源。此外，SGAE[28]，GCN-LSTM[29]等方法还通过显式关系建模增强结构化语义表达。总体而言，不同视觉特征在感受野、语义粒度和关注重点上各有优势，单一特征往往难以同时兼顾全局场景、局部实体和区域边界，因此多视觉特征联合建模已成为提升图像描述准确性与细粒度表达能力的重要思路。

1.3 多特征交互与融合方法

围绕多特征利用，已有研究主要从多路输入和融合机制 2 个方面展开。文献 [1] 较早探索了多视角视觉表示在 Transformer 图像描述中的作用。代表性工作[30-33]分别从交叉注意力建模、循环融合、层次注意力以及双层协同交互等方面改进多特征信息的组织方式，从而增强了图像描述模型对异构视觉特征的联合表征能力。GRIT 则进一步表明双视觉特征能够在效率与性能之间取得较好平衡。除视觉特征间融合外，近年来还出现了检索增强图像描述方向，例如 GFF[33]，SmallCap[34]，M-RAT[35]，EVCap[23]通过引入检索文本、对象名称或外部记忆，为描述生成提供额外语言先验。与此同时，ConCap[36]和 FineCaption[24]代表了可控与组合式图像描述的发展趋势。尽管这些方法在融合策略上取得了显著进展，但多数工作更关注“如何融合”，而较少显式区分不同特征间的共享语义与特异信息。

1.4 共性-特性表示学习与对比增强方法

在多源信息建模中，仅通过拼接、加权或注意力交互进行融合，容易忽略不同特征之间同时存在的“共性”与“特性”。从更一般的视觉语言学习视角看，UNIMO[21]通过跨模态对比学习增强统一表征，NAAF[37]强调困难样本与判别性关系建模，EVE[38]则从模态感知与专家路由角度提升多模态表示能力。这些研究表明，显式建模共享信息与差异信息，并借助对比约束提升语义一致性，是增强多源特征表达能力的

重要方向。对于图像描述任务而言，现有多特征方法虽然已经开始关注跨特征交互，但对共享语义如何增强、特异信息如何保留以及两者如何协同融合的讨论仍不充分。基于此，本文在 Transformer 编解码框架下引入多视觉特征的共性-特性协同建模，通过预融合、共性增强和动态融合机制，提升标准图像描述任务中的多源视觉信息组织能力与描述生成质量。

2 方　　法

本文提出的 EAFCC 模型整体结构如图 2 所示。模型首先对输入图像进行多源视觉表征提取，利用 UPSNet 提取分割特征 $\boldsymbol{F}_{\mathrm{S}}$，利用 Faster R-CNN 提取区域特征 $\boldsymbol{F}_{\mathrm{R}}$ 和网格特征 $\boldsymbol{F}_{\mathrm{G}}$。3 类特征分别从场景语义布局、目标级局部信息和全局视觉上下文 3 个层面刻画图像内容，具有显著的互补性。然而，由于三者在表示粒度、语义侧重点和分布空间上存在差异，若直接拼接或粗粒度融合，容易引入语义错位与信息干扰，进而影响描述生成的准确性与细节完整性。

为此，本文围绕共性提取—特性建模—协同融合的思路构建统一的编码与解码框架。在编码阶段，模型首先将分割特征、区域特征和网格特征送入预融合共享编码器（PF）模块。该模块一方面对多源异构特征进行初步对齐与交互，缩小不同特征之间的表示差异；另一方面提取不同特征流之间共享的高层语义信息，为后续共性增强提供统一的表示基础。与传统仅关注特征拼接或简单注意力加权的方法不同，PF 模块不仅承担输入对齐作用，同时也是后续共性分支的语义起点。

在此基础上，模型进一步将经 PF 处理后的多路特征送入 3 条参数独立的特征编码分支，即 Encoder-S, Encoder-R, Encoder-G, 分别针对分割特征流、区域特征流和网格特征流提取各自保留的特性信息。这样可以在共享语义空间对齐的前提下，最大限度保留不同特征在目标边界、区域实体和场景全局结构上的差异化表达。随后，3 路特性特征被输入多专家路由融合网络（multi-feature fusion with mixture-of-experts, MMF），通过门控路由机制自适应建模各特征分支对当前图像描述任务的贡献程度，实现多特征特性信息的动态加权融合，从而得到融合后的特性表征。

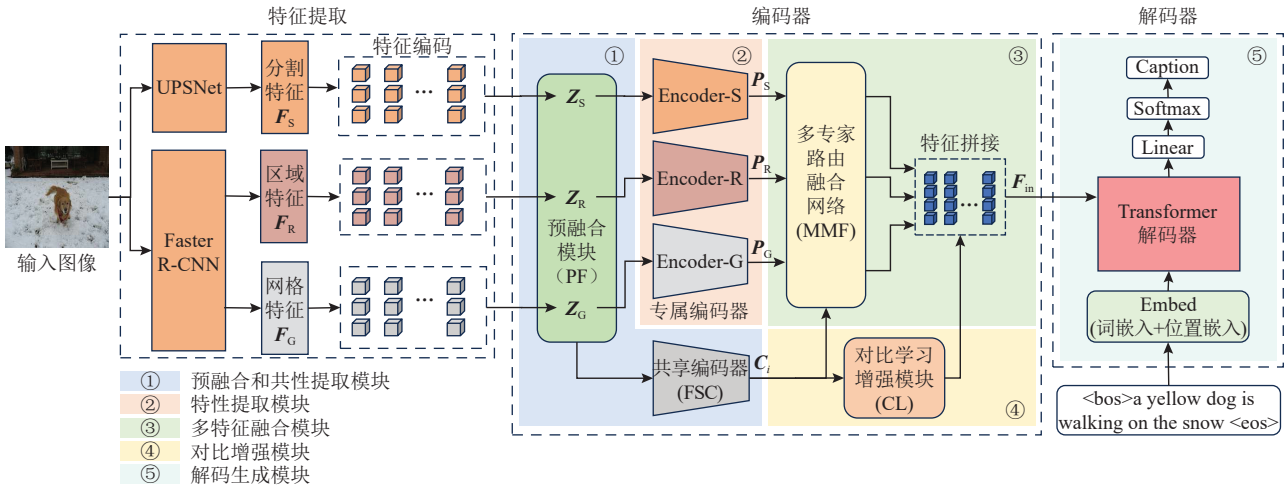


Fig. 2　Structure diagram of EAFCC model

图 2　EAFCC 模型结构图

与此同时，为进一步提升不同特征流在共享语义层面的表达一致性，本文在共性分支中引入对比学习增强模块（contrastive learning, CL）。该模块以 PF 输出的共享表征为基础，通过对比约束拉近语义相关的共性表示、区分潜在干扰表示，从而增强共性信息的判别性与鲁棒性。经过 CL 模块处理后，模型获得语义更加一致、表达更加稳定的共性特征表示。

最后，增强后的共性特征与 MMF 输出的融合特性特征在编码器末端进行协同汇聚，并共同送入 Transformer 解码器。解码阶段，模型结合词嵌入与位置嵌入，利用多层 Transformer 解码结构逐步生成自然语言描述，最终经线性映射与 Softmax 输出完整句子。通过上述设计，EAFCC 模型能够同时利用多源视觉特征中的共享语义与差异化细节，从而提升图像描述的语义准确性、信息完整性和表达丰富性。后文将依次介绍各模块的具体结构与实现方式。

2.1　预融合共享编码器（PF）

为缓解分割特征、区域特征和网格特征在表示

粒度与语义分布上的差异,并在初步对齐的同时尽可能保留各自的特异性信息,本文在编码器前端引入预融合模块,其结构如图 3 所示。PF 模块由共享参数的多头自注意力(multi-head attention)子层、共享层归一化子层、共享位置逐元素前馈子层以及特征流独立的层归一化子层组成,整体采用残差连接形式堆叠。

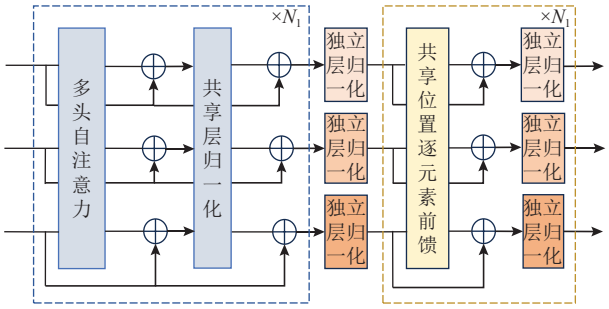


Fig. 3 Structure diagram of pre-fusion module network

图 3 预融合模块网络结构图

具体而言,给定 3 类输入特征$\boldsymbol{F}_{\mathrm{G}},\boldsymbol{F}_{\mathrm{R}},\boldsymbol{F}_{\mathrm{S}}$,首先通过共享参数的多头自注意力机制对 3 路特征进行统一建模,并结合共享层归一化操作对其进行初步对齐,从而缩小不同特征流在共享语义空间中的表示偏移。由于该阶段采用共享参数,不同特征流能够在一致的映射约束下学习更稳定的高层共性表示。随后,在对齐后的表示基础上,引入共享位置逐元素前馈网络进一步增强特征表达,并通过各特征流独立的层归一化操作来保留分割特征、区域特征和网格特征各自的特异性,从而避免过度共享导致的个性信息损失。通过这一设计,PF 模块能够在共享对齐与特性保持之间取得平衡,为后续共性增强与特性建模提供更合适的输入表示。

在本文实现中,PF 模块的堆叠层数设为$N_1=2$,设共享多头自注意力$MHSA_{\mathrm{sh}}(\cdot)$和共享位置逐元素前馈网络$LN_{\mathrm{sh}}(\cdot)$的参数分别为$\boldsymbol{\theta}_{\mathrm{att}}$和$\boldsymbol{\theta}_{\mathrm{ff}}$,共享层归一化参数为$(\boldsymbol{\alpha}_{\mathrm{sh}},\boldsymbol{\beta}_{\mathrm{sh}})$,3 条特征流独立层归一化参数分别为$(\boldsymbol{\alpha}_{\mathrm{G}},\boldsymbol{\beta}_{\mathrm{G}}),(\boldsymbol{\alpha}_{\mathrm{R}},\boldsymbol{\beta}_{\mathrm{R}}),(\boldsymbol{\alpha}_{\mathrm{S}},\boldsymbol{\beta}_{\mathrm{S}})$。则单层 PF 模块可表示为

$$\boldsymbol{M}_{\mathrm{G}}=LN_{\mathrm{sh}}(MHSA_{\mathrm{sh}}(\boldsymbol{F}_{\mathrm{G}};\boldsymbol{\theta}_{\mathrm{att}})+\boldsymbol{F}_{\mathrm{G}};\boldsymbol{\alpha}_{\mathrm{sh}},\boldsymbol{\beta}_{\mathrm{sh}}),\tag{1}$$

$$\boldsymbol{M}_{\mathrm{R}}=LN_{\mathrm{sh}}(MHSA_{\mathrm{sh}}(\boldsymbol{F}_{\mathrm{R}};\boldsymbol{\theta}_{\mathrm{att}})+\boldsymbol{F}_{\mathrm{R}};\boldsymbol{\alpha}_{\mathrm{sh}},\boldsymbol{\beta}_{\mathrm{sh}}),\tag{2}$$

$$\boldsymbol{M}_{\mathrm{S}}=LN_{\mathrm{sh}}(MHSA_{\mathrm{sh}}(\boldsymbol{F}_{\mathrm{S}};\boldsymbol{\theta}_{\mathrm{att}})+\boldsymbol{F}_{\mathrm{S}};\boldsymbol{\alpha}_{\mathrm{sh}},\boldsymbol{\beta}_{\mathrm{sh}}),\tag{3}$$

其中$\boldsymbol{M}_{\mathrm{G}},\boldsymbol{M}_{\mathrm{R}},\boldsymbol{M}_{\mathrm{S}}$表示经过共享对齐后的特征。在此基础上,进一步通过共享前馈网络$PWFF_{\mathrm{sh}}(\cdot)$和特征独立的层归一化$LN_i(\cdot),i\in\{\mathrm{G},\mathrm{R},\mathrm{S}\}$得到最终输出为

$$\boldsymbol{Z}_{\mathrm{G}}=LN_{\mathrm{G}}(PWFF_{\mathrm{sh}}(\boldsymbol{M}_{\mathrm{G}};\boldsymbol{\theta}_{\mathrm{ff}})+\boldsymbol{M}_{\mathrm{G}};\boldsymbol{\alpha}_{\mathrm{G}},\boldsymbol{\beta}_{\mathrm{G}}),\tag{4}$$

$$\boldsymbol{Z}_{\mathrm{R}}=LN_{\mathrm{R}}(PWFF_{\mathrm{sh}}(\boldsymbol{M}_{\mathrm{R}};\boldsymbol{\theta}_{\mathrm{ff}})+\boldsymbol{M}_{\mathrm{R}};\boldsymbol{\alpha}_{\mathrm{R}},\boldsymbol{\beta}_{\mathrm{R}}),\tag{5}$$

$$\boldsymbol{Z}_{\mathrm{S}}=LN_{\mathrm{S}}(PWFF_{\mathrm{sh}}(\boldsymbol{M}_{\mathrm{S}};\boldsymbol{\theta}_{\mathrm{ff}})+\boldsymbol{M}_{\mathrm{S}};\boldsymbol{\alpha}_{\mathrm{S}},\boldsymbol{\beta}_{\mathrm{S}}),\tag{6}$$

其中$\boldsymbol{Z}_{\mathrm{G}},\boldsymbol{Z}_{\mathrm{R}},\boldsymbol{Z}_{\mathrm{S}}$分别表示 PF 模块输出的 3 路特征表示。需要说明的是,式(1)~(6)描述了单层 PF,实际重复堆叠$N_1=2$,上层输出作为下层输入。

2.2 特征共性和特性提取建模模块

为了捕捉不同特征之间的共性信息,模型采用共享编码器机制。共享 Transformer 编码器采用标准结构,由多层堆叠,每层含多头自注意力机制、前馈神经网络,使用残差连接和层归一化技术,经多层堆叠对网格特征$\boldsymbol{F}_{\mathrm{G}}$、区域特征$\boldsymbol{F}_{\mathrm{R}}$和分割特征$\boldsymbol{F}_{\mathrm{S}}$统一编码,将不同输入特征映射到共享语义空间,提取共有语义信息,最终提炼出共性特征表示$\boldsymbol{C}_{\mathrm{G}},\boldsymbol{C}_{\mathrm{R}},\boldsymbol{C}_{\mathrm{S}}$。对于每个输入的特征信息流$\boldsymbol{Z}_{\mathrm{G}},\boldsymbol{Z}_{\mathrm{R}},\boldsymbol{Z}_{\mathrm{S}}$,共享编码器计算过程可以表示为

$$\boldsymbol{C}_i=ShareEncoder(\boldsymbol{Z}_i),\tag{7}$$

其中$ShareEncoder(\cdot)$是共享编码器操作,$i\in\{\mathrm{G},\mathrm{R},\mathrm{S}\}$时,$\boldsymbol{C}_i$分别为对应的共性特征表示$\boldsymbol{C}_{\mathrm{G}},\boldsymbol{C}_{\mathrm{R}},\boldsymbol{C}_{\mathrm{S}}$。

为保留不同特征流的差异化信息,本文在共享语义基础上进一步建模各特征流的特性表征。对于 PF 模块输出的分支特征表示$\boldsymbol{Z}_i$,采用参数独立的 Transformer 编码器进一步提取特性特征,计算过程可以表示为

$$\boldsymbol{P}_i=PrivateEncoder_i(\boldsymbol{Z}_i),\tag{8}$$

其中$PrivateEncoder_i(\cdot)$为特征编码器编码操作,$i\in\{\mathrm{G},\mathrm{R},\mathrm{S}\}$时,$\boldsymbol{P}_i$分别为对应的特性表征$\boldsymbol{P}_{\mathrm{G}},\boldsymbol{P}_{\mathrm{R}},\boldsymbol{P}_{\mathrm{S}}$。

2.3 基于对比学习的共性特征增强模块(CL)

在共享 Transformer 编码器获得初步共性表征后,不同特征流之间虽然已具备一定的共享语义基础,但由于分割特征、区域特征和网格特征在表示粒度与语义侧重点上存在天然差异,其共性表示仍可能存在分布偏移,进而影响后续特征协同与描述生成质量。为进一步提升共享语义表示的一致性与鲁棒性,本文在共性分支中引入基于对比学习的共性特征增强模块(CL)。

该模块以共享编码器输出的 3 路共性特征作为输入,通过对共享语义空间中的表征关系进行约束,增强语义相关共性表示之间的聚合性,并抑制无关干扰信息带来的偏移,从而获得更加紧致、稳定且具有判别性的共性表征。经过 CL 模块增强后,模型输出 3 路增强的共性特征,并将其与后续融合得到的特性特征共同用于描述生成,其具体损失函数形式与优化过程将在 2.7 节中详细介绍。

2.4 基于多专家的多特征融合模块（MMF）

为了有效地融合特性特征$\boldsymbol{P}_i$，模块采用基于多专家的图像多特征融合机制。该机制借鉴了多专家系统（mixture of expert, MoE）的思想[29]，将每个特征信息流的特性特征视为一个“专家”，利用门控路由网络学习每个特性特征的重要性权重，并根据权重对不同的特性特征进行加权融合。模块使用一个门控网络（gating network）来计算每个专家特性流的权重[10]。门控网络采用MLP结构，输入是拼接的特性特征$\boldsymbol{P}_i$，其公式为

$$\boldsymbol{w} \coloneqq \begin{pmatrix} \omega_{\mathrm{G}} \\ \omega_{\mathrm{R}} \\ \omega_{\mathrm{S}} \end{pmatrix} = Softmax(MLP(Concat(\boldsymbol{P}_{\mathrm{G}}, \boldsymbol{P}_{\mathrm{R}}, \boldsymbol{P}_{\mathrm{S}}))) , \tag{9}$$

其中ω_{G}，ω_{R}，ω_{S}分别表示3个分支的权重，经*Softmax*归一化后满足$\omega_{\mathrm{G}}+\omega_{\mathrm{R}}+\omega_{\mathrm{S}}=1$。得到门控权重$\boldsymbol{w}$之后，融合后的特性特征$\boldsymbol{F}_{\mathrm{fused}}$可以通过门控加权拼接的方式计算为

$$\boldsymbol{F}_{\mathrm{fused}} = Concat(\omega_{\mathrm{G}}\boldsymbol{P}_{\mathrm{G}}, \omega_{\mathrm{R}}\boldsymbol{P}_{\mathrm{R}}, \omega_{\mathrm{S}}\boldsymbol{P}_{\mathrm{S}}) 。 \tag{10}$$

2.5 跳跃连接编码器模块（FSC）

为了减少图像信息在编码过程中丢失的问题，提出了跳跃连接编码器模块，如图4所示。

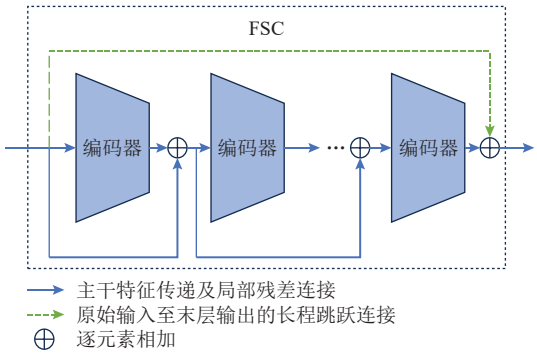


Fig. 4 Skip connection encoder module

图4 跳跃连接编码器模块

与常规编码方式相比，在对特征进行提取增强时，输入到下一层编码器的特征不仅包含上一层编码器的输出，还包含上一层的输入信息。在输入到下一个编码器之前，为降低计算资源的消耗，采用2个输入直接相加的方式，并且为最大程度降低特征在增强提取过程中的丢失，采用将初始输入的信息流直接与最高层编码器的输出流相加的方式得到最终输出。

设FSC模块采用的是N层Transformer编码器进行堆叠，$\boldsymbol{F}_0$是原始输入的特征，那么第1层到第$N-1$层的输出可以表示为

$$\boldsymbol{F}_n = Encoder_n(\boldsymbol{F}_{n-1}) + \boldsymbol{F}_{n-1} , \tag{11}$$

其中$n \in \{1,2,3,\cdots,N-1\}$，$\boldsymbol{F}_n$表示经过第$n$层编码器的输出，$Encoder_n(\cdot)$表示对输入编码器层的输入进行一系列操作，其中主要包括多头自注意力机制、前馈神经网络以及残差连接和层归一化处理。在第N层时，由于使用跳跃连接，其公式为

$$\boldsymbol{F}_N = Encoder_N(\boldsymbol{F}_{N-1}) + \boldsymbol{F}_0 。 \tag{12}$$

经过多层编码以及跳跃连接的特征将不同层次的特征进行汇总，使得最终特征既包含低层语义信息，又包含高层语义信息，实现了多尺度信息的保留。需要说明的是，FSC作为共享编码器和分支专属编码器的基本实现结构，当输入为$\boldsymbol{Z}_i$时，共享FSC编码器输出的$\boldsymbol{F}_N$为$\boldsymbol{C}_i$，分支专属FSC编码器输出的$\boldsymbol{F}_N$为$\boldsymbol{P}_i$。

2.6 解码与生成描述

共享编码器输出的共性特征$\boldsymbol{C}_{\mathrm{G}}, \boldsymbol{C}_{\mathrm{R}}, \boldsymbol{C}_{\mathrm{S}}$在训练过程中受到对比学习损失的约束，从而获得更强的一致性与判别性。在解码生成阶段，融合后的特性特征$\boldsymbol{F}_{\mathrm{fused}}$以及增强后的共性特征$\boldsymbol{C}_{\mathrm{G}}, \boldsymbol{C}_{\mathrm{R}}, \boldsymbol{C}_{\mathrm{S}}$将被拼接后作为最终Transformer解码器的图像特征输入为

$$\boldsymbol{F}_{\mathrm{in}} = Concat(\boldsymbol{F}_{\mathrm{fused}}, \boldsymbol{C}_{\mathrm{G}}, \boldsymbol{C}_{\mathrm{R}}, \boldsymbol{C}_{\mathrm{S}}) , \tag{13}$$

其中$\boldsymbol{F}_{\mathrm{in}}$表示解码器的输入特征。最后，通过多层堆叠的Transformer解码器的掩码多头自注意力机制和交叉注意力机制处理，生成完整的图像描述。

2.7 损失函数与训练策略

为同时增强多特征共性表征的判别能力，并提升图像描述生成的词级预测精度与序列级生成质量，本文采用对比学习约束与监督训练及强化学习微调的2阶段优化策略。整体而言，对比学习损失作用于编码阶段的共性增强模块，用于约束不同特征流在共享语义空间中的表示一致性；交叉熵损失用于第1阶段监督训练，以学习图像特征与目标描述之间的基本对齐关系；在此基础上，进一步引入基于自我批评序列训练（self-critical sequence training, SCST）[39]的强化学习方法，以CIDEr-D作为奖励函数对模型进行微调，从而直接优化序列级评价目标。

1）对比学习损失

为增强不同特征流共性信息的一致性并学习细粒度差异，本文在共性增强模块中引入基于InfoNCE（noise contrastive estimation）[11]的对比学习损失函数。以网格特征$\boldsymbol{F}_{\mathrm{G}}$和区域特征$\boldsymbol{F}_{\mathrm{R}}$为例，设同一训练批次内的每个图像样本$j$的正样本对为$(\boldsymbol{C}_j^{\mathrm{G}}, \boldsymbol{C}_j^{\mathrm{R}})$，完整的负样本池可表示为$N_j^{(\mathrm{G,R})} = \{\boldsymbol{C}_k^{\mathrm{R}} \mid k \neq j\} \cup \{\boldsymbol{C}_l^{\mathrm{S}} \mid l \neq j\}$，其对比学习损失函数如式（14）所示。

$$\mathcal{L}_{\mathrm{CL}}^{(j)}=-\ln\frac{\exp\left(sim(\boldsymbol{C}_j^{\mathrm{G}},\boldsymbol{C}_j^{\mathrm{R}})/\tau\right)}{\exp\left(sim(\boldsymbol{C}_j^{\mathrm{G}},\boldsymbol{C}_j^{\mathrm{R}})/\tau\right)+\sum_{k\neq j}\exp\left(sim(\boldsymbol{C}_j^{\mathrm{G}},\boldsymbol{C}_k^{\mathrm{R}})/\tau\right)+\sum_{l\neq j}\exp\left(sim(\boldsymbol{C}_j^{\mathrm{G}},\boldsymbol{C}_l^{\mathrm{S}})/\tau\right)},\tag{14}$$

其中j表示当前图像样本的索引，k和l表示同一批次内其他图像样本的索引，$\boldsymbol{C}_j^{\mathrm{G}},\boldsymbol{C}_j^{\mathrm{R}},\boldsymbol{C}_j^{\mathrm{S}}$分别表示第$j$个图像样本的网格特征、区域特征和分割特征的共性特征，$sim(\cdot,\cdot)$表示相似度函数，τ表示温度系数，B表示批次大小。

$\boldsymbol{F}_{\mathrm{G}}$与$\boldsymbol{F}_{\mathrm{S}}$以及$\boldsymbol{F}_{\mathrm{S}}$与$\boldsymbol{F}_{\mathrm{R}}$之间的计算方式与式(14)类似，最终通过3组对比损失的相加，确保所有特征信息流的两两之间的共性均被增强，从而增强3种特征信息之间的共性信息。

为了提升对比学习的难度和效率，模型进一步采用了困难负样本采样(hard negative sampling)策略。困难负样本指的是那些与正样本相似度较高的负样本，与随机负采样相比，困难负样本采样策略能够更有效地挖掘和区分不同特征信息之间的细微差异，从而提升对比学习的效果。

以$\boldsymbol{F}_{\mathrm{G}}$和$\boldsymbol{F}_{\mathrm{R}}$为例，困难负样本的选择策略为：对于每个正样本对$(\boldsymbol{C}_j^{\mathrm{G}},\boldsymbol{C}_j^{\mathrm{R}})$，模型首先计算其与批次内所有其他负样本的相似度，然后选择相似度最高的K个负样本作为困难负样本，构建困难负样本集合H_j，H_j可以表示为

$$H_j=TopK\left(\left\{sim(\boldsymbol{C}_j^{\mathrm{G}},\boldsymbol{z})|\boldsymbol{z}\in N_j^{(\mathrm{G,R})}\right\}\right),\tag{15}$$

其中$\boldsymbol{z}$表示从负样本池$N_j^{(\mathrm{G,R})}$中选取的任意候选负样本共性表征。将该策略融入对比学习损失函数之后，计算如式(16)所示。

$$\mathcal{L}_{\mathrm{CL_{HNS}}}^{(j)}=-\ln\frac{\exp\left(sim(\boldsymbol{C}_j^{\mathrm{G}},\boldsymbol{C}_j^{\mathrm{R}})/\tau\right)}{\exp\left(sim(\boldsymbol{C}_j^{\mathrm{G}},\boldsymbol{C}_j^{\mathrm{R}})/\tau\right)+\sum_{h\in H_j}\exp(h/\tau)},\tag{16}$$

其中h表示困难负样本集合H_j中的一个相似度值。根据式(15)，H_j由候选负样本与锚点特征$\boldsymbol{C}_j^{\mathrm{G}}$之间相似度最高的$K$个数值组成。

通过最小化基于困难负样本采样的对比学习损失函数$\mathcal{L}_{\mathrm{CL_{HNS}}}^{(j)}$，模型能够有效地增强不同特征共性的一致性，并学习细粒度的差异，最终得到增强后的共性特征$\boldsymbol{C}_{\mathrm{G}},\boldsymbol{C}_{\mathrm{R}},\boldsymbol{C}_{\mathrm{S}}$。

2)交叉熵监督训练

在第1阶段训练中，本文采用单词级交叉熵(CE)损失对模型进行监督训练，以提高解码器对目标词的预测能力。设目标描述序列为$Y^*=\left(y_1^*,y_2^*,\cdots,y_T^*\right)$，融合后的视觉表示为$\boldsymbol{F}_{\mathrm{in}}$，则交叉熵损失定义为

$$\mathcal{L}_{\mathrm{CE}}=-\sum_{t=1}^{T}\ln p\left(y_t^*\mid y_1^*,\cdots,y_{t-1}^*,\boldsymbol{F}_{\mathrm{in}}\right),\tag{17}$$

其中T表示目标描述序列的长度，t表示词语生成的时间步，y_t^*表示第t个真实目标词，上标“$*$”表示真实标注，$p(\cdot)$表示模型在时刻t对目标词的条件概率。该阶段主要用于建立稳定的图像-文本对齐关系，并为后续序列级优化提供良好的初始化参数。考虑到对比学习损失同样在第1阶段参与优化，因此第1阶段的总体目标函数可写为

$$\mathcal{L}_{\mathrm{stage1}}=\mathcal{L}_{\mathrm{CE}}+\frac{1}{B}\sum_{j=1}^{B}\mathcal{L}_{\mathrm{CL_{HNS}}}^{(j)},\tag{18}$$

其中B表示批次大小，j表示批次内的样本索引。

3)基于SCST的强化学习微调

在第1阶段训练收敛后，本文进一步采用基于自我批评序列训练(self-critical sequence training, SCST)的强化学习方法对模型进行微调，以直接优化图像描述任务中的序列级评价指标。具体而言，模型首先通过采样方式生成描述序列Y^{D}，并通过贪心解码生成基线序列$\hat{Y}$。以CIDEr-D作为奖励函数，强化学习损失定义为

$$\mathcal{L}_{\mathrm{RL}}=-\left(r(Y^{\mathrm{D}})-r(\hat{Y})\right)\sum_{t=1}^{T}\ln p\left(y_t^{\mathrm{D}}\mid y_1^{\mathrm{D}},\cdots,y_{t-1}^{\mathrm{D}},\boldsymbol{F}_{\mathrm{in}}\right),\tag{19}$$

其中，y_t^{D}表示采样描述序列中的第t个词，T表示生成序列的长度，$r(\cdot)$表示基于CIDEr-D计算得到的奖励值，$\hat{Y}$作为基线序列用于降低策略梯度估计的方差。当采样序列的奖励高于基线序列时，模型将增加该序列的生成概率；反之，则降低其概率。与仅使用交叉熵损失相比，该策略能够在一定程度上缓解暴露偏差问题，并使模型优化目标与最终评价指标更加一致。

在第2阶段中，模型以强化学习损失为主进行微调，因此总体目标可表示为

$$\mathcal{L}_{\mathrm{stage2}}=\mathcal{L}_{\mathrm{RL}}。\tag{20}$$

4)总体训练流程

综上，本文采用2阶段训练策略：第1阶段通过交叉熵损失与对比学习损失的联合优化，学习稳定的视觉语义表示和基础语言生成能力；第2阶段在此基础上利用基于CIDEr-D奖励的SCST方法进行序列级微调，以进一步提升生成描述的整体语义相关性与评价指标表现。该训练方式使编码端的共性增强、解码端的词级监督以及最终的序列级优化形

成互补，从而共同提升图像描述生成性能。

3 实　　验

3.1 数据集与实现细节

本文使用 COCO 2014 数据集进行实验，其中包含 123 287 张图片，每张图片对应 5 句英文描述。为便于将实验结果与其他模型性能进行对比，遵循主流的数据集划分模式，将训练集图片设置为 113 287 张，验证集和测试集图片均设置为 5 000 张。对于文本描述的处理，采用通用处理方式，将所有图片参考描述中出现次数少于 5 的单词筛除，使用经过处理后的 10 201 个词汇作为候选词库。此外，对于数据集中的每一句文本描述，删除其中所有标点符号，并将其中的大写格式转换为小写格式，在每句描述的开头插入<bos>表示句子的开始，在每句描述的结尾插入<eos>表示句子的结束，并且使用<pad>对长度不足 20 个单词的图片描述句子进行填充。

在训练实现上，参考 M^2Transformer[12] 的常用训练范式，并结合本文模型进行调整。第 1 阶段以词级交叉熵损失为主，对模型进行监督训练，同时结合对比学习损失对共享语义表征进行约束，以提升模型的基础生成能力和共性特征表示质量；第 2 阶段在第 1 阶段参数初始化的基础上，采用基于 SCST 的强化学习方法进行微调，并以 CIDEr-D 作为奖励信号，直接优化图像描述任务中的序列级评价指标。模型训练过程中采用 Adam 优化器，并使用分段学习率策略进行参数更新。所有实验均在 NVIDIA RTX3090 上完成，训练轮数设为 60，批次大小设为 50。

3.2 参数配置

为便于复现与对比，本文配置汇总如表 1 所示，所有实验均重复 5 次并取平均值，以确保结果的可重复性与公平性。

Table 1　Experimental Environment Settings

表 1　实验环境设置

类别	参数项	参数
实验环境	PyTorch	1.11.0
	CUDA	11.3
	Python	3.8
	GPU	RTX 3090（24 GB）
输入	预处理尺寸	短边 600px（保持长宽比） 再随机水平翻转、标准化
	网格特征图尺寸	7×7=49 个网格，每个网格 2 048 维 （ResNet-101conv5_x）
	分割特征图尺寸	7×7=49 个空间位置，每个位置 133 维 （COCO133 类，UPSNet 语义头）
	区域特征数量	36 个 RoI/图（Faster R-CNN， *IoU*>0.7 取前 36 个）
	文本词汇量	10 201 词，最大句子长度 20
优化器	类型	Adam
	批次大小	50

3.3 指标说明

为验证本文方法的有效性，采用以下评价指标进行性能比较。

1）BLEU-*n*[5]。基于生成描述与参考描述在 1~4 元词组上的重叠情况进行评价，其中 *n* 越大越关注生成句在长序列层面的连贯性与流畅性。具体而言，BLEU-1 主要反映词汇层面的匹配准确性，BLEU-4 则更强调短语组合和整体表达的自然性。

2）ROUGE[40]。基于生成描述与参考描述之间的最长公共子序列进行评价，从召回率角度衡量生成句对参考句关键信息的覆盖程度。该指标更关注整体内容的完整性与句子结构相似性，ROUGE-L 值越高，说明生成描述越能保留人工参考句中的主要语义信息。

3）METEOR[7]。引入同义词、词干还原和词序惩罚，用加权 *F*-measure 调和精度与召回，并惩罚片段乱序，缓解 BLEU 对同义表达过于苛刻的问题，更贴近人工判断。

4）CIDEr-D[41]。先用 TF-IDF 给 *n*-gram 赋权，突出稀有却关键的视觉词，再计算余弦相似并叠加长度惩罚，抑制重复刷分，使高分描述必须与多数人编写的参考句达成共识。

5）SPICE[42]。句子解析成（对象，属性，关系）三元组场景图，按 *F*1 计算语义元组匹配率，专评内容准确性：即使词面不同，只要语义关系对得上，就能获得高分。

本文方法在 COCO 2014 数据集上进行实验，并与公开文献中报告的代表性方法结果进行对比。

3.4 对比实验

在相同测试集上，将本文方法与 Up-Down，AoANet，X-Transformer 等代表性方法和训练策略 SCST[39] 进行对比。模型的网格特征和分割特征提取的骨干网络均为 ResNet-101，分割特征大小为 7×7，并且在共享编码器部分添加了 PF 模块，所有编码器均采用提出的 FSC 模块。实验结果如表 2 所示。

表 2 结果显示，EAFCC 在 5 类指标上均优于单特征输入的 M^2Transformer[12]，CIDEr-D 相对提升 3.05%，其余指标平均提升 2.65%。相较于双特征融合的代表模型 DLCT，BLEU-4 提升 2.01%，CIDEr-D 提升 1.05%，

Table 2 Comparative Experimental Results of EAFCC Model on COCO2014 Dataset

表 2 EAFCC 模型在 COCO2014 数据集上对比实验结果

模型	B4↑	M↑	R↑	C↑	S↑
SCST[39]	34.3	26.7	55.7	114.0	
Up-Down[25]	36.3	27.7	56.9	120.1	21.4
GCN-LSTM[29]	38.2	28.5	58.5	128.3	22.0
AoANet[10]	38.9	29.2	58.8	129.8	22.4
EAT[11]	39.9	28.9	59.0	127.6	22.6
M^2Transformer[12]	39.1	29.2	58.6	131.2	22.6
X-Transformer[13]	39.7	29.5	59.1	132.8	23.4
DLCT[14]	39.8	29.5	59.1	133.8	23.0
CIIC[43]	40.2	29.5	59.4	133.1	23.2
RSTNet[44]	39.3	29.4	58.8	133.3	23.0
ConCap[36]	40.5	**30.9**		133.7	**23.8**
EAFCC（本文）	**40.6**	29.8	**59.8**	**135.2**	23.2

注：黑体数值表示最优值。B4，M，R，C，S 依次表示 BLEU-4，METEOR，ROUGE，CIDEr-D，SPICE。↑表示数值越高性能越优。各指标均按原始评测得分乘以 100 的尺度展示。

验证了引入分割特征并在共享语义基础上建模分支依赖补充信息的有效性。与 2022 年 CIIC 及 2023 年 ConCap 相比，CIDEr-D 分别提高 1.58% 与 1.12%。上述增益可归因于 3 个特征输入以及共性-特性 2 阶段融合策略：PF 模块先对齐语义空间并保留特征个性，MMF 模块再动态加权特性向量，使解码器同时获得一致的高层语义与细粒度线索。然而，EAFCC 在 METEOR 与 SPICE 上仍低于 ConCap，差异分别为 1.1 个百分点与 0.6 个百分点。分析表明 ConCap 通过文本提示显式注入语义重心，而这 2 个指标对同义词与语义命题更敏感，因此提示机制会带来额外增益。

表 3 给出了各模型在第 1 阶段交叉熵训练下的对比结果，主要用于考察模型在不引入序列级强化学习优化时的基础生成能力，而表 2 中的最终结果则对应在第 1 阶段训练基础上进一步经过强化学习微调后的性能表现。通过分别报告 2 阶段结果，可以更清晰地分析模型结构设计本身的有效性，以及强化学习微调对最终描述生成质量的增益作用。可以看出，在不引入强化学习微调的情况下，EAFCC 已表现出较强的基础生成能力，在多数评价指标上优于对比模型，并在 SPICE 指标上与 X-Transformer 持平。这说明，本文所提出的多特征共性-特性协同建模框架在交叉熵训练阶段即可学习到较为有效的视觉语义表示与语言生成能力。

在此基础上，进一步引入基于 CIDEr-D 奖励的 SCST 强化学习微调后，模型最终性能继续提升，说明 2 阶段训练策略具有必要性。其中，第 1 阶段交叉熵训练主要用于建立稳定的词级预测能力；第 2 阶段则通过直接优化序列级评价指标，进一步提升生成描述的整体语义相关性与表达质量。由此可见，本文模型的性能提升既来源于网络结构设计本身，也得益于 2 阶段训练策略所带来的进一步优化。

Table 3 Comparative Results of Cross-Entropy Loss Training Across Different Models in Stage 1

表 3 不同模型在阶段 1 的交叉熵损失训练对比结果

模型	B1↑	B4↑	M↑	R↑	C↑	S↑
Up-Down[25]	77.2	36.2	27.0	56.4	113.5	20.3
GCN-LSTM[29]	77.3	36.8	27.9	57.0	116.3	20.9
SGAE[28]	**77.6**	36.9	27.7	57.2	116.7	20.9
AoANet[10]	77.4	37.2	28.4	57.5	119.8	21.3
X-Transformer[13]	77.3	37.0	28.7	57.5	120.0	**21.8**
SMALLCAP[34]		37.0	27.9		119.7	21.3
EAFCC（本文）	**77.6**	**37.5**	**28.8**	**57.6**	**120.2**	**21.8**

注：黑体数值表示最优值。B1，B4，M，R，C，S 依次表示 BLEU-1，BLEU-4，METEOR，ROUGE，CIDEr-D，SPICE。↑表示数值越高性能越优。各指标均按原始评测得分乘以 100 的尺度展示。

随着视觉-语言预训练模型规模的扩大，模型通过亿级图文对预训练获得强泛化能力，但在 COCO 描述任务上通常仅采用交叉熵微调，且推理开销大。相比于这些预训练多模态多任务模型，本文研究的优势在于能够在计算资源受限的情况下针对特定领域进行细粒度的图像描述任务。表 4 给出了与近年预训练多模态多任务模型在 COCO 数据集上的图像描述生成任务的性能对比结果。

Table 4 Comparative Results of Pre-Trained Vision-Language Models

表 4 预训练视觉语言模型对比结果

模型	预训练图像数量	参数量	B4↑	C↑
$BLIP_{base}$[17]	14×10^6	223×10^6	38.6	129.7
$BLIP_{CapFilt-L}$[17]	129×10^6		39.7	133.3
OFASys[19]	20×10^6	270×10^6		134.8
OFA_{Tiny}[18]	20×10^6	33×10^6	38.1	128.7
$SimVLM_{base}$[20]	1.8×10^9	86×10^6	39.0	134.8
UNIMO[21]			39.6	127.7
Unified VLP[22]	3×10^6		39.5	129.3
EAFCC（本文）		**31.9×10^6**	**40.6**	**135.2**

注：黑体数值表示最优值。↑表示数值越高性能越优。B4，C 依次表示 BLEU-4，CIDEr-D。B4 和 C 的值按原始评测得分乘以 100 的尺度展示。

如表 4 所示，BLIP 系列、OFA 系列、SimVLM 等模型依赖 14×10^6~1.8×10^9 级图文预训练，在 COCO 上仅做轻量交叉熵微调即可取得 133 以上的 CIDEr-D，

表现出强大的跨任务泛化能力。然而，其大规模预训练与推理资源需求在边缘或域特定场景中难以承受。

表 4 的结果表明，尽管 BLIP、OFA 和 SimVLM 等方法依赖大规模视觉语言预训练，并在模型参数量和预训练数据规模上具有明显优势，EAFCC 在不借助额外预训练图像的条件下，仍在 MS-COCO 测试集上取得了 135.2 的 CIDEr-D 和 40.6 的 BLEU-4，优于表中多数预训练模型。进一步从模型规模来看，EAFCC 的参数量为 31.9×10^6，明显低于 $\text{BLIP}_{\text{base}}$ 的 223×10^6 和 OFASys 的 270×10^6。该结果表明，在资源受限条件下，针对图像描述任务进行专门的多特征建模与融合设计，同样能够取得具有竞争力的结果。

3.5 消融实验

为系统验证本文所提出各组成模块的有效性，以不包含预融合共享编码器（PF）、跳跃连接编码器（FSC）、对比学习共性增强模块（CL）以及基于多专家的多特征融合模块（MMF）的基础模型构建 EAFCC-base，并在此基础上逐步引入不同模块开展消融实验。同时，为进一步说明多特征共性-特性协同建模框架的整体优势，还将 EAFCC-base 与单特征模型 M^2Transformer、双特征模型 DLCT[14] 以及 DIFNet[16] 进行对比，实验结果如表 5 所示。

Table 5 Ablation Experiment Results of EAFCC Model on COCO2014 Dataset

表 5 EAFCC 模型在 COCO2014 数据集上的消融实验结果

模型	增加模块	B1↑	B2↑	B3↑	B4↑	M↑	R↑	C↑	S↑
M^2Transformer[12]		80.8			39.1	29.2	58.6	131.2	22.6
DLCT[14]		81.4			39.8	29.5	59.1	133.8	23.0
DIFNet[16]		81.5			40.0	29.5	59.2	134.9	23.1
EAFCC-base		81.6	66.4	52.2	40.3	29.6	59.2	134.0	23.0
EAFCC-base	PF	81.4	66.4	52.0	39.9	29.6	59.1	133.4	22.9
EAFCC-base	FSC	81.5	66.5	52.0	40.0	29.7	59.3	133.3	23.0
EAFCC-base	PF+FSC	81.8	66.7	52.2	40.5	29.7	59.5	134.3	23.1
EAFCC-base	PF+FSC+CL	82.0	66.7	52.3	40.5	29.6	59.5	134.7	23.0
EAFCC-base	PF+FSC+MMF	82.1	66.8	52.4	40.4	**29.8**	59.4	134.9	23.1
EAFCC-base	全部模块	**82.4**	**67.3**	**52.6**	**40.6**	**29.8**	**59.8**	**135.2**	**23.2**

注：黑体数值表示最优值。B1，B2，B3，B4，M，R，C，S 依次表示 BLEU-1，BLEU-2，BLEU-3，BLEU-4，METEOR，ROUGE，CIDEr-D，SPICE。↑表示数值越高性能越优。各指标均按原始评测得分乘以 100 的尺度展示。

从整体结果来看，即使不引入 CL 和 MMF，EAFCC-base 仍表现出较强的竞争力，整体性能优于单特征建模的 M^2Transformer 和双特征融合的 DLCT，并在多数指标上超过 DIFNet。这说明，相较于传统单一视觉表征或简单双特征融合方式，本文构建的 3 特征共性-特性协同建模框架能够更充分地利用分割特征、区域特征和网格特征之间的互补信息，从而为图像描述生成提供更有效的视觉语义表示。

在单模块消融方面，PF 与 FSC 的作用呈现出不同特点。仅引入 PF 模块时，模型整体性能未出现明显改善，部分指标甚至略有波动。这表明，预融合操作虽然能够在一定程度上缩小多源特征之间的表示差异，但若缺乏后续特性保持与信息补偿机制，其优势尚难以充分转化为下游生成性能。相比之下，仅引入 FSC 模块后，模型在若干指标上表现出一定提升，说明跳跃连接机制有助于缓解深层编码过程中信息衰减的问题，并在一定程度上改善描述生成质量。但从整体上看，FSC 单独使用所带来的增益仍较为有限，说明仅依赖层间信息传递尚不足以充分释放多特征建模的潜力。

进一步地，当 PF 与 FSC 联合引入后，模型性能较基线模型出现更稳定的提升，且整体表现优于两者单独作用时的结果。这说明，预融合所带来的跨特征对齐能力需要借助跳跃连接机制进行信息保留与补偿，才能更有效地服务于后续特征建模与描述生成。换言之，PF 与 FSC 在模型中并非彼此独立地发挥作用，而是存在较明显的协同关系。

在此基础上，进一步分析高层语义增强与动态融合机制的贡献。实验结果表明，在 PF+FSC 结构上引入 CL 模块后，模型在核心生成指标上继续取得提升，说明对比学习约束能够增强多源特征在共享语义空间中的一致性，从而提升共性表征的判别能力。进一步引入 MMF 模块后，模型性能再次改善，表明多专家路由融合机制能够更灵活地建模不同特征分支对当前图像描述任务的贡献，并有效整合差异化特性信息。总体来看，CL 更侧重于提升共享语义表示的紧致性与一致性，而 MMF 更侧重于增强特性信息的动态互补能力，两者在功能上具有较强的互补性。

最终，当 PF，FSC，CL，MMF 全部引入后，模型取得了表 5 中的最优结果。该结果说明，本文所设计的各模块均能对性能提升产生积极作用，且不同模块之间并非简单叠加关系，而是通过预融合对齐、特性保持、共性增强与动态融合协同发挥作用，共同提升了图像描述生成的语义准确性、细节完整性和整体表达质量。

此外，为进一步分析 PF 的迭代次数 N_1 对模型性能的影响，本文在 EAFCC-base 上设置不同迭代次数开展消融实验，结果如表 6 所示。

Table 6 Ablation Experimental Results of Iteration Numbers in PF Module of EAFCC-base

表 6 EAFCC-base 的 PF 模块中迭代次数消融实验结果

迭代次数 N_1	B1↑	B2↑	B3↑	B4↑	M↑	R↑	C↑	S↑
1	81.5	66.5	52.0	40.2	**29.6**	59.2	133.6	22.9
2	81.6	**66.7**	**52.3**	**40.3**	**29.6**	**59.5**	**134.0**	**23.0**
3	**81.7**	66.4	51.9	40.0	29.5	59.3	**134.0**	**23.0**
4	81.4	66.3	51.8	39.8	29.5	59.1	133.9	22.9

注：黑体数值表示最优值。B1，B2，B3，B4，M，R，C，S 依次表示 BLEU-1，BLEU-2，BLEU-3，BLEU-4，METEOR，ROUGE，CIDEr-D，SPICE。↑表示数值越高性能越优。各指标均按原始评测得分乘以 100 的尺度展示。

通过表 6 可知，迭代次数为 2 时模型效果最好，当迭代次数再增加时，模型性能反而出现下降趋势。类似现象在文献 [16] 中也得到了验证，证明迭代次数过多会导致图像的低层语义特征占比增加，从而忽略相对重要的高层语义特征，导致模型性能下降。因此本文 EAFCC 模型最终采用 2 次迭代，达到了相对较好的性能。

在 EAFCC-base 模型上对 FSC 的跳连方式进行消融实验，由于编码器一共由 6 层 Transformer 编码器堆叠而成，所以分别采用跳跃间隔为“1”“3”“6”“1, 6”“3, 6”以及全连接进行实验，结果如表 7 所示。

Table 7 Ablation Experimental Results of Skip Intervals in the EAFCC-base Skip-Connection Encoder

表 7 EAFCC-base 跳跃连接编码器中跳跃间隔消融实验结果

跳跃间隔	B1↑	B2↑	B3↑	B4↑	M↑	R↑	C↑	S↑
1	81.4	66.4	51.8	40.1	29.3	58.7	133.7	22.8
3	**81.6**	66.4	51.5	40.3	29.1	58.5	133.5	22.7
6	81.2	66.2	51.5	40.0	29.0	58.1	133.1	22.7
1，6	**81.6**	**66.7**	**52.3**	**40.4**	**29.6**	**59.5**	**134.0**	**23.0**
3，6	81.6	66.5	51.7	**40.4**	29.4	59.1	133.9	23.0
全连接	81.3	66.2	51.3	40.0	29.0	58.0	133.1	22.6

注：黑体数值表示最优值。B1，B2，B3，B4，M，R，C，S 依次表示 BLEU-1，BLEU-2，BLEU-3，BLEU-4，METEOR，ROUGE，CIDEr-D，SPICE。↑表示数值越高性能越优。各指标均按原始评测得分乘以 100 的尺度展示。

表 7 的实验结果证实，本文对跳跃连接方式进行了系统性消融：在单独验证逐级跳连与首尾跳连的基础上，进一步考察了两者并联以及稠密全连接跳连的配置，结果一致表明逐级与首尾越级并联这一稀疏方案在 BLEU-4, CIDEr-D, SPICE 上均取得最优值，验证其可为梯度回传提供充足且低冗余的路径，从而强化模型对底层视觉特征的利用并生成更细粒度的描述。相比之下，稠密全连接跳跃连接在相同层数编码器下性能显著下降，提示当网络深度不足时，过度密集的连接易加剧参数耦合并诱发过拟合，导致泛化性能受损。

本文进一步分析负样本数参数 K 对模型性能的影响，结果如表 8 所示。

Table 8 Sensitivity Analysis of Negative Sample Number K

表 8 负样本数 K 的敏感性分析

评价指标	K=2	K=3	K=5	K=16	K=32	K=50	K=64
BLEU-1	75.5	75.8	75.6	75.8	**76.0**	**76.0**	75.8
BLEU-4	35.8	35.9	35.9	35.7	36.0	**36.3**	36.1
METEOR	27.9	**28.0**	27.8	27.9	27.9	27.9	27.9
ROUGE	56.3	56.5	56.2	56.4	56.4	**56.6**	56.3
CIDEr-D	114.9	115.5	115.3	115.1	115.5	**116.3**	115.4

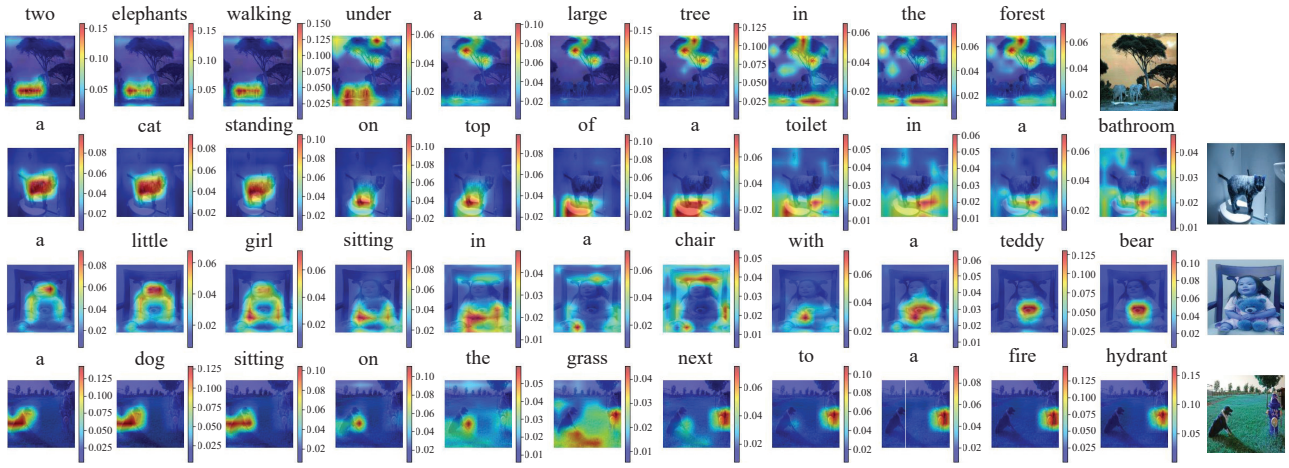
注：黑体数值表示最优值。

在其他训练设置保持不变的条件下，分别测试了 K=2, 3, 5, 16, 32, 50, 64 的情况。实验结果表明，不同评价指标下的最优 K 值并不完全一致，这主要是由于 BLEU, METEOR, ROUGE, CIDEr-D 对生成描述质量的关注重点不同。总体来看，模型在合理范围内对 K 的变化较为稳定，但当 K 取值过小时，性能会出现一定下降；随着 K 增大，模型表现逐渐提升并趋于平稳。其中，K=50 在核心指标 CIDEr-D 上取得最佳结果，并在其他指标上保持了较好的综合表现，因此本文最终将 K 固定为 50。

3.6 可视化分析

为验证模型特征融合的性能，从测试集中选取了不同场景的图片进行图像描述生成，并保留生成每个词时的注意力权重，以注意力图的方式呈现，如图 5 所示。为验证多特征输入融合模型与双特征输入模型（DLCT）以及采用大规模数据训练的 OFA_{Tiny} 模型在解码过程中的差异，采用这 3 个模型对相同图像进行描述并生成对应的注意力热图，如图 6 所示。

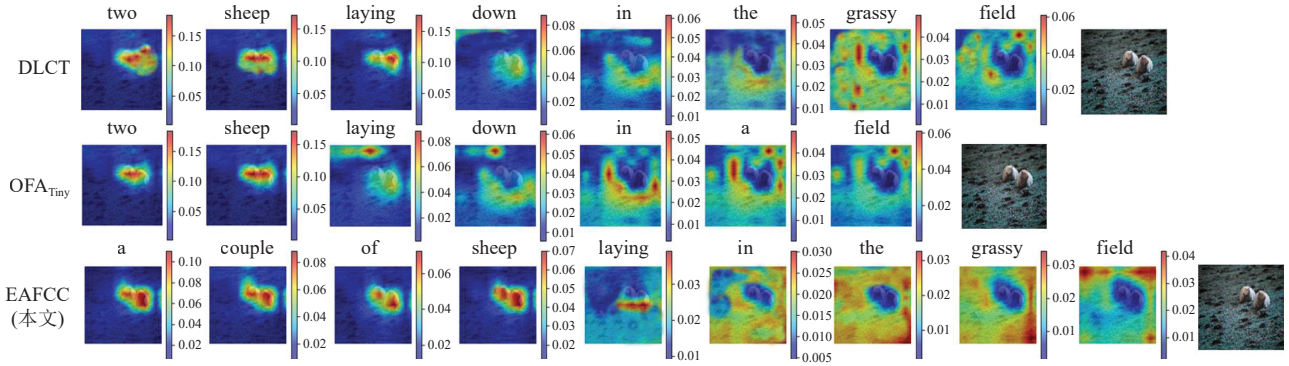
从图 5 和图 6 中可以看到，模型在生成背景信息的相关描述时，注意力关注更加准确，并且对相关动作的注意力体现得明显细致，如“Laying”一词对应的注意力，且生成的描述也更加具体准确。表 9 报告了 EAFCC 与 DLCT, DIFNet 及多模态预训练基线 OFA_{Tiny} 在测试集上的自动生成结果对比。其中，样本 1~2 具备高视觉相似度，用以评估对比学习模块的区分能力，样本 3~4 包含复杂背景与遮挡，用以检验多特征融合机制对细节与语境的保持能力。相比于使用大规模数据训练的 OFA_{Tiny} 模型，可视化结果显示，EAFCC 在生成每个词语时均能准确聚焦对应的图像语义区



注：标尺的值是模型在生成上方那个单词时，分配给不同图像区域的注意力权重。

Fig. 5 Attention maps for generating descriptions in different scenarios

图 5 不同场景下生成描述的注意力图



注：标尺的值是模型在生成上方那个单词时，分配给不同图像区域的注意力权重。

Fig. 6 Attention visualization comparative diagrams generated by different models

图 6 不同模型生成的注意力可视化对比图

域，并在多种场景下保持较高的语义对齐精度。例如对于包含“two elephants walking under a large tree in a forest”的图像，模型在生成“walking”时关注象腿区域，在生成“under”时聚焦于树下空间，在生成“forest”时则激活背景区域，体现出对空间关系的理解。

EAFCC 在生成的描述上更加具体，且由于采用多特征融合方式，对图像背景和细节信息的描述更加完整，体现了本文研究方法的有效性。这说明 EAFCC 在生成描述过程中，能够实现词语与图像区域之间的准确对齐，保证描述内容在主体、动作、背景及空间关系等方面的一致性与准确性。

为验证共性分支与特性分支之间的表示差异，本文基于已训练模型在验证集上提取各分支输出特征，并计算其平均余弦相似度，结果如图 7 所示。可以看到，共性分支内部表示的相似度整体较高，其中 $\boldsymbol{C}_{\mathrm{G}},\boldsymbol{C}_{\mathrm{R}},\boldsymbol{C}_{\mathrm{S}}$ 之间的平均余弦相似度约为 0.57；相比之下，特性分支 $\boldsymbol{P}_{\mathrm{G}},\boldsymbol{P}_{\mathrm{R}},\boldsymbol{P}_{\mathrm{S}}$ 之间的平均余弦相似度约为 0.24，明显低于共性分支。该结果说明，共享编码与共性增强机制能够在一定程度上拉近不同特征流在共享语义空间中的表示，而分支专属编码器则进一步保留了与特征类型相关的差异化信息。

进一步地，分割特性分支 $\boldsymbol{P}_{\mathrm{S}}$ 与其他分支之间的相似度整体较低，表明该分支更倾向于保留与语义区域和边界结构相关的特有信息。与此同时，部分共性分支与特性分支之间仍具有较高相关性，这说明 EAFCC 中的特性并非是与共性完全割裂的独立表示，而是在共享语义基础上进一步形成的分支依赖补充表示。

4 结 论

针对图像描述任务中视觉表征利用不足、跨特征对齐与互补融合不充分的问题，本文在 Transformer 编码器-解码器框架下提出共性表征与特性表征协同

Table 9 Comparison of Image Captioning Generative Results on the Test Set

表 9 测试集图像描述生成结果对比

图像	描述
	**GT1**：A giraffe is standing in the middle of the grass. **GT2**：A giraffe standing alone in a forest. **M²Transformer**：A giraffe standing in the forest. **DLCT**：A giraffe standing alone in the green field. **DIFNet**：A giraffe standing in forest. **OFA$_{Tiny}$**：A giraffe is brown and white. **EAFCC:** A giraffe is standing in the **tall green bushes**.
	**GT1**：A young giraffe leaning over a tall bush in a dry field. **GT2**：Giraffe eating leaves off of a tree in tall bushes. **M²Transformer**：A giraffe is eating leaves in the field. **DLCT**：A giraffe is eating green leaves in the field. **DIFNet**：A giraffe eating leaves in a field. **OFA$_{Tiny}$**：A giraffe is standing in the grass. **EAFCC**：A **young** giraffe is eating green leaves in the **dry bushes**.
	**GT1**：Well made be in clean and neat bedroom setting. **GT2**：A very luxurious bedroom filled with nice furniture. **M²Transformer**：A bedroom filled with a lot of furniture. **DLCT**：A grand bedroom includes a bed and a sofa. **DIFNet**：A bedroom filled with nice furniture. **OFA$_{Tiny}$**：A bedroom with a bed or beds in a hotel room. **EAFCC**：A luxurious and **clean** bedroom filled with nice furniture such as bed and sofa.
	**GT1**：A man in a vest and plaid pants rides a bicycle. **GT2**：A man rides a bicycle down the street. **M²Transformer**：A man rides a bicycle on the street. **DLCT**：A man with plaid and hat rides a is riding a bicycle. **DIFNet:** A man riding a bicycle on the street. **OFA$_{Tiny}$**：A man with a hat riding a bicycle. **EAFCC**：A man with leather hat riding a bicycle on road with **greenery in the middle**.

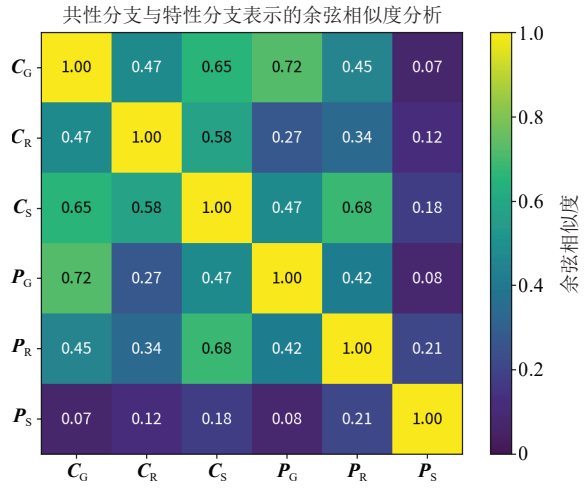


Fig. 7 Average cosine similarity heatmap of common branch and specific branch output representations

图 7 共性分支与特性分支输出表示的平均余弦相似度热力图

融合机制。首先对网格、区域和分割 3 类特征进行共享语义建模；继而以独立编码器在共性基底上强化模态专有细节，形成特性表征。其次，在共性分支中引入对比学习约束，拉近语义匹配对、推远异类对；特性分支通过门控路由实现动态加权，抑制冗余、突出差异。最后，增强后的共性与特性向量拼接输入解码器，使交叉注意力于每个时间步精准对齐对应视觉成分，生成细粒度且高保真的描述。消融实验与对比实验验证所提方法的有效性。

作者贡献声明：杨建锋提出研究构思并指导研究开展，审阅和修改论文；朱东坡负责方法实现、模型搭建、实验与数据分析，并撰写论文；曹瑞审阅和修改论文，统筹审稿意见回复并负责最终定稿；崔选皓参与论文审阅与修改；王平参与论文审阅与修改；郑杰提出改进思路，并提供方法优化建议与分析支持。

参 考 文 献

[1] Gurari D, Li Qing, Stangl A J, et al. VizWiz grand challenge: Answering visual questions from blind people[C]//Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 3608−3617

[2] Aneja J, Deshpande A, Schwing A G. Convolutional image captioning[C]//Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 5561−5570

[3] Yu Jun, Li Jing, Yu Zhou, et al. Multimodal Transformer with multi-view visual representation for image captioning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(12): 4467−4480

[4] Lou Liangshan, Lu Ke, Xue Jian. A novel cross-fusion method of different types of features for image captioning[C]//Proc of the 33rd Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2023: 1−8

[5] Mathur N, Baldwin T, Cohn T. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics[C]// Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 4984−4997

[6] Cui Yin, Yang Guandao, Veit A, et al. Learning to evaluate image captioning[C]//Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 5804−5812

[7] Bernardi R, Cakici R, Elliott D, et al. Automatic description generation from images: A survey of models, datasets, and evaluation measures[J]. Journal of Artificial Intelligence Research, 2016, 55: 409−442

[8] Al-Shamayleh A S, Adwan O, Alsharaiah M A, et al. A comprehensive literature review on image captioning methods and metrics based on deep learning technique[J]. Multimedia Tools and Applications, 2024, 83(12): 34219−34268

[9] Ghandi T, Pourreza H, Mahyar H. Deep learning approaches on image captioning: A review[J]. ACM Computing Surveys, 2023, 56(3): 1−39

[10] Huang Lun, Wang Wenmin, Chen Jie, et al. Attention on attention for image captioning[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4634−4643

[11] Li Guang, Zhu Linchao, Liu Ping, et al. Entangled Transformer for image captioning[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 8928−8937

[12] Cornia M, Stefanini M, Baraldi L, et al. Meshed-memory Transformer for image captioning[C]//Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10578−10587

[13] Pan Yingwei, Yao Ting, Li Yehao, et al. X-linear attention networks for image captioning[C]//Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10971−10980

[14] Luo Yunpeng, Ji Jiayi, Sun Xiaoshuai, et al. Dual-level collaborative Transformer for image captioning[C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 2286−2293

[15] Nguyen V Q, Suganuma M, Okatani T. GRIT: Faster and better image captioning transformer using dual visual features[G]//LNCS 13696: Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 167−184

[16] Wu Mingrui, Zhang Xuying, Sun Xiaoshuai, et al. DIFNet: Boosting visual information flow for image captioning[C]//Proc of the 35th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 18020−18029

[17] Li Junnan, Li Dongxu, Xiong Caiming, et al. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//Proc of the 39th Int Conf on Machine Learning. New York: PMLR, 2022: 12888−12900

[18] Wang Peng, Yang An, Men Rui, et al. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework[C]//Proc of the 39th Int Conf on Machine Learning. New York: PMLR, 2022: 23318−23340

[19] Bai Jinze, Men Rui, Yang Hao, et al. OFASys: A multi-modal multi-task learning system for building generalist models[J]. arXiv preprint, arXiv: 2212.04408, 2022

[20] Wang Zirui, Yu Jiahui, Yu Adams Wei, et al. SimVLM: Simple visual language model pretraining with weak supervision[J]. arXiv preprint, arXiv: 2108.10904, 2021

[21] Li Wei, Gao Can, Niu Guocheng, et al. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning[J]. arXiv preprint, arXiv: 2012.15409, 2020

[22] Zhou Luowei, Palangi H, Zhang Lei, et al. Unified vision-language pre-training for image captioning and VQA[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 13041−13049

[23] Li Jiaxuan, Vo D M, Sugimoto A, et al. EVCap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 13733−13742

[24] Hua Hang, Liu Qing, Zhang Lingzhi, et al. FineCaption: Compositional image captioning focusing on wherever you want at any granularity[C]//Proc of the 38th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2025: 24763−24773

[25] Anderson P, He Xiaodong, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]// Proc of the 31st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6077−6086

[26] Ren Shaoqing, He Kaiming, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137−1149

[27] Xiong Yuwen, Liao Renjie, Zhao Hengshuang, et al. UPSNet: A unified panoptic segmentation network[C]//Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 8818−8826

[28] Yang Xu, Tang Kaihua, Zhang Hanwang, et al. Auto-encoding scene graphs for image captioning[C]//Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 10685−10694

[29] Yao Ting, Pan Yingwei, Li Yehao, et al. Exploring visual relationship for image captioning[G]//LNCS 11218: Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 711−727

[30] Zhang Jing, Xie Yingshuai, Ding Weichao, et al. Cross on cross attention: Deep fusion transformer for image captioning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 4257−4268

[31] Jiang Wenhao, Ma Lin, Jiang Yugang, et al. Recurrent fusion network for image captioning[G]//LNCS 11206: Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 499−515

[32] Wang Weixuan, Chen Zhihong, Hu Haifeng. Hierarchical attention network for image captioning[C]//Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 8957−8964

[33] Zhao Shanshan, Li Lixiang, Peng Haipeng. Incorporating retrieval-based method for feature enhanced image captioning[J]. Applied Intelligence, 2023, 53(8): 9731−9743

[34] Ramos R, Martins B, Elliott D, et al. SmallCap: Lightweight image captioning prompted with retrieval augmentation[C]//Proc of the 36th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 2840−2849

[35] Song Jiayan, Pan Renjie, Zhou Jun, et al. M-RAT: A multi-grained retrieval augmentation Transformer for image captioning[C]//Proc of the 17th Asian Conf on Computer Vision. Berlin: Springer, 2024: 185−203

[36] Wang Ning, Xie Jiahao, Wu Jihao, et al. Controllable image captioning via prompting[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 2617−2625

[37] Zhang Kun, Mao Zhendong, Wang Quan, et al. Negative-aware attention framework for image-text matching[C]//Proc of the 35th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 15661−15670

[38] Chen Junyi, Guo Longteng, Sun Jia, et al. EVE: Efficient vision-language pre-training with masked prediction and modality-aware MoE[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 1110−1119

[39] Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning[C]//Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 7008−7024

[40] Lin C Y. ROUGE: A package for automatic evaluation of summaries[C]//Proc of the Text Summarization Branches Out. Stroudsburg, PA: ACL, 2004: 74−81

[41] Vedantam R, Lawrence Z C, Parikh D. CIDEr: Consensus-based image description evaluation[C]//Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 4566−4575

[42] Anderson P, Fernando B, Johnson M, et al. SPICE: Semantic propositional image caption evaluation[G]//LNCS 9909: Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 382−398

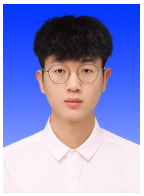
[43] Liu Bing, Wang Dong, Yang Xu, et al. Show, deconfound and tell: Image captioning with causal inference[C]//Proc of the 35th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 18041−18050

[44] Zhang Xuying, Sun Xiaoshuai, Luo Yunpeng, et al. RSTNet: Captioning with adaptive attention on visual and non-visual words[C]//Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 15465−15474



Yang Jianfeng, born in 1981. Master, senior engineer, master supervisor. His main research interests include software development methodologies and data science, with a focus on applications in higher education.

杨建锋, 1981 年生。硕士, 高级工程师, 硕士生导师。主要研究方向为软件开发方法与数据科学及其在高校中的应用。



Zhu Dongpo, born in 1999. Master, engineer. His main research interest includes multi-modal fusion and application.

朱东坡, 1999 年生。硕士, 工程师。主要研究方向为多模态融合应用。



Cao Rui, born in 1991. PhD, lecturer, master supervisor. Member of CCF. His main research interests include text representation learning, applications and optimization of generative large language models, and art generation using diffusion models.

曹 瑞, 1991 年生。博士, 讲师, 硕士生导师。CCF 会员。主要研究方向为文本表示学习、生成式大语言模型应用、优化及扩散模型艺术生成。



Cui Xuanhao, born in 2003. Master candidate. His main research interest includes large model development and application.

崔选皓, 2003 年生。硕士研究生。主要研究方向为大模型开发应用。



Wang Ping, born in 2002. Master candidate. His main research interest includes large model visual reasoning.

王 平, 2002 年生。硕士研究生。主要研究方向为大模型视觉推理。



Zheng Jie, born in 1985. PhD, professor, master supervisor. His main research interest includes edge general intelligence.

郑 杰, 1985 年生。博士, 教授, 硕士生导师。主要研究方向为边缘通用智能。