

基于多级适配器增强与自适应特征 融合的零样本草图检索*

王伟光^{1,2}, 魏宏喜^{1,2}, 贺超^{1,2}

(1. 内蒙古大学计算机学院, 呼和浩特 010021; 2. 内蒙古大学蒙古文智能信息处理技术国家地方联合工程研究中心, 呼和浩特 010021)

摘要:针对现有零样本草图检索方法依赖单一深层特征导致特征提取不充分以及特征融合不足的问题,提出一种基于多级适配器增强与自适应特征融合的方法。该方法在不同网络层嵌入适配器,有效增强了边缘轮廓、局部结构和全局语义的多层次特征表示能力。进一步提出了自适应融合模块,借助可学习权重动态融合多级增强特征与原始视觉特征,实现局部和全局信息的融合。在3个数据集上进行对比实验,与当前最优方法相比,所提方法性能均有明显提升。

关键词:零样本草图检索; 零样本学习; 对比语言-图像预训练; 视觉 Transformer; 提示学习

中图分类号:TP391.41 **文献标志码:**A

随着移动智能设备和多点触控技术的广泛使用,基于草图的图像检索任务^[1]已成为跨模态检索的一个关键分支。手绘草图的生成已经从传统的以纸和笔为媒介发展到数字交互环境,极大提高了其作为视觉表示工具的价值。在现实世界的开放场景中,用户检索需求可能涉及大量未知类别,由于缺乏零样本学习能力,现有模型的性能显著下降。此外,获得大量高质量的草图具有挑战性,导致草图数据集稀缺^[2]。为了解决上述问题,零样本草图检索任务(Zero-shot sketch-based image retrieval, ZS-SBIR)^[3-5]引入了一种零样本学习机制,将数据集分为可见类别(训练集)和不可见类别(测试集),这需要模型在跨模态语义鸿沟和类别语义不一致的双重挑战下执行检索任务。

零样本草图检索的早期方法通常利用卷积神经网络同时从草图和自然图像中提取视觉特征。然而,由于卷积神经网络依赖于局部感受野,容易丢失草图的关键细节信息^[6]。为解决这一问题,SAKE方法引入了基于生成模型的特征合成路径,并借助知识蒸馏保留了在大规模数据集上预训练获得的判别性特征,从而减少模型在适应新类别时出现的灾难性遗忘现象^[7]。但该方法依赖复杂的师生网络结构,同时预训练教师模型若存在偏差,将进一步影响生成特征的准确性。为进一步强化语义传递,后续研究引入类别词嵌入技术,以建立跨模态和跨类别的语义关联^[8-10]。然而,提取的粗粒度语义信息并不能准确表示类别特征,这限制了模型在复杂草图场景下的判别能力。最近的研究转向对比语言-图像预训练(Contrastive language-image pre-training, CLIP)模型^[11],该类方法充分利用CLIP在零样本推理任务上的优势,结合文本语义信息,通过提示学习调整输入表示。尽管如此,标准CLIP模型仍存在两方面局限:一是其文本通常依赖于固定的手工模板,难以动态适配多样化的语义内容;二是视觉分支在微调过程中未能充分融合文本模态的引导信息^[12]。针对上述问题,DRCPL

* 收稿日期:2025-10-07; 修回日期:2025-12-31

基金项目:内蒙古自然科学基金项目(2024MS06029)

作者简介:王伟光(2001—),男,内蒙古呼和浩特人,2023级硕士研究生。E-mail:wangwg1019@163.com

通信作者:魏宏喜(1980—),男,辽宁鞍山人,教授,博士。主要从事智能信息处理研究。E-mail:cswxh@imu.edu.cn

将文本提示映射为视觉提示,实现了视觉-语言间的协作学习机制,显著提升了模型对未知类别的泛化性能^[13]。

尽管上述工作显著推动了零样本草图检索任务的发展,但它们仍然面临以下问题:(1)现有方法对视觉 Transformer 层的特征利用不足,导致对复杂草图的适应性差;(2)现有方法主要依赖视觉编码器最后一层的类别 token([CLS] token)作为图像表示,忽视了 Transformer 架构中不同层次捕获的多样化特征表示,无法有效整合多级语义信息,导致泛化能力受限。因此,本文提出了一种基于多级适配器增强与自适应特征融合的零样本草图检索方法(Multi-level adapter enhancement and adaptive feature fusion for zero-shot sketch-based image retrieval, MAEAF)。具体来说,首先提出了一个多级适配器增强模块(Multi-level adapter enhancement module, MAEM),该模块在视觉编码器的关键层插入轻量级的适配器。在视觉 Transformer 中,不同层捕获不同抽象级别的特征,在浅层,适配器可以增强草图轮廓和边缘特征的表示;在中层,适配器可以强化草图局部形状的表示;在深层,适配器可以调整语义特征,使草图的语义与自然图像的语义更好地对齐。此外,提出了一个自适应特征融合模块(Adaptive feature fusion module, AFFM),该模块通过自适应加权机制整合多层次优化特征,解决了单一特征表示的局限性。这种设计赋予了模型根据输入特征动态调整特征依赖的能力,即对于轮廓清晰、结构简单的草图,模型可赋予中层特征更高权重以充分利用其几何信息;对于语义复杂、需要上下文理解的图像,则可倾向于深层特征的全局语义表示。在 Sketchy^[14]、TU-Berlin^[15]和 QuickDraw^[16]3个基准数据集上进行了全面的实验。结果表明,与现有的最先进的零样本草图检索方法相比,提出的基于多级适配器增强与自适应特征融合的零样本草图检索方法具有更优越的性能。

1 相关工作

1.1 零样本草图检索

零样本草图检索任务要求模型在仅学习可见类别草图和图像映射关系的基础上,通过类别间的语义关联,完成对不可见类别草图的有效检索。在跨域对齐方面,早期方法主要依赖生成对抗网络实现草图与图像间的特征映射^[6],后续方法进一步引入了关系感知度量学习和孪生网络等技术,以减小域间差异。随着视觉 Transformer(Vision transformer, ViT)^[17]的兴起,基于自注意力与交叉注意力机制的跨模态网络被用于建模图像与草图在图像块层面的语义对应^[3]。在知识迁移方面,SAKE^[7]利用知识蒸馏保留预训练模型的语义信息,结合词嵌入模型构建语义关联空间,并探索多模态预训练模型的潜力。随着 CLIP 模型域泛化能力的深入研究,其文本与视觉编码器之间的协同效应以及提示学习的共享策略增强了 CLIP 视觉与文本编码器间的知识交互,有效弥补了草图与自然图像嵌入之间的语义鸿沟,为跨域检索提供了有效途径^[10]。

1.2 视觉-语言预训练模型和提示学习

近年来,视觉-语言预训练模型已成为多模态学习中的重要研究方向。其中,CLIP 模型作为代表性方法,展现出卓越的跨模态理解与泛化能力。CLIP 采用双分支架构,分别使用视觉编码器与文本编码器提取图像与文本的嵌入表示。通过对比学习,CLIP 将匹配的图文对在嵌入空间中拉近,而非匹配对则推远,从而具备强大的零样本识别能力^[11]。

多模态提示学习是优化 CLIP 及视觉-语言预训练模型的关键策略,对增强模型适应特定任务的能力具有重要作用。CoOp 采用可学习的文本提示替代手工设计的提示模板,提升了零样本分类性能^[18]。然而,该方法主要侧重于文本提示的优化,对视觉模态的影响有限,所以,后续逐渐开始探索从文本提示到视觉提示的流动机制,旨在建立更强的模态间交互作用。DRCPL^[13]提出一种视觉与文本提示的联合优化方法,不仅影响文本编码器,还将提示通过语义映射传递至视觉编码器,从而增强图像特征的任务适应能力。

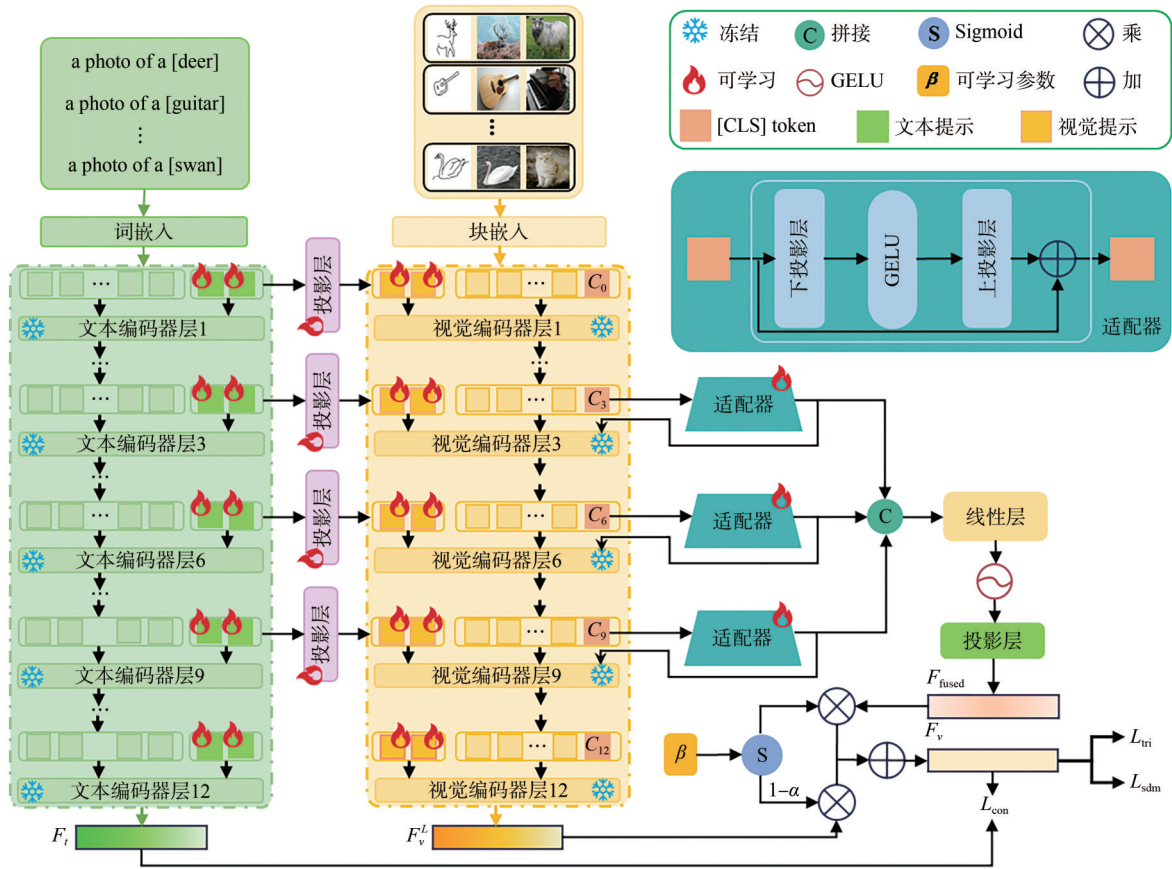
2 方法

2.1 问题定义

零样本草图检索的数据集由可见类别 C_{seen} 的训练集 $D_{\text{tr}} = \{S_{\text{tr}}, P_{\text{tr}}\}$ 和不可见类别 C_{unseen} 的测试集 $D_{\text{te}} = \{S_{\text{te}}, P_{\text{te}}\}$ 构成, 确保 $C_{\text{seen}} \cap C_{\text{unseen}} = \emptyset$ 。在这个数据集中, $S_{\text{tr}} = \{(s_i, y_i) | y_i \in C_{\text{seen}}\}_{i=1}^{n_s}$ 表示训练集中的草图集, $P_{\text{tr}} = \{(p_i, y_i) | y_i \in C_{\text{seen}}\}_{i=1}^{n_p}$ 表示训练集中的图像集, 其中 s_i 是草图样本, p_i 是图像样本, y_i 是类别标签, n_s 是草图数量, n_p 是图像数量。测试集中草图和图像的定义与训练集中的定义相似, 同时确保 $D_{\text{tr}} \cap D_{\text{te}} = \emptyset$ 。

2.2 模型结构

基于多级适配器增强与自适应特征融合的零样本草图检索方法(MAEAF)的框架如图1所示, 主要由CLIP、多级适配器增强模块和自适应特征融合模块组成。



在视觉编码器的第3、6、9层部署适配器, 并通过可学习的权重机制动态整合多级特征。

图1 MAEAF框架

Fig. 1 Framework of MAEAF

2.2.1 基于CLIP的提示学习

文本编码器采用基于CLIP的Transformer结构, 将一个带有类别名称的手工模板“a photo of a [类别]”输入文本编码器中, 对文本描述中的单词进行分词, 然后将它们投影到单词嵌入空间 $T_0 = [t_0^1, \dots, t_0^N] \in \mathbb{R}^{N \times d_t}$, 其中 N 表示嵌入 tokens 的数量, d_t 是文本特征的维度。为了灵活地学习特征表示, 模型冻结了文本和视觉编码器参数, 同时在文本编码器的浅层引入了可学习的 tokens。具体来说, 在文本编码器的前 H 层 Transformer 中引入 n 个可学习的文本提示 $P_i = [P_i^1, \dots, P_i^n] \in \mathbb{R}^{n \times d_t}$, 文本

提示一方面通过具有残差连接的浅层网络(ResMLP)^[13]生成相应的视觉提示 P'_i , 表示为 $P'_i = R_{\text{esMLP}}(P_i)$, $i = 0, 1, \dots, H-1$ 。另一方面, 文本提示与其他文本 tokens 进行拼接, 作为输入 $T_i = [P_i^1, \dots, P_i^n, t_i^{n+1}, \dots, t_i^N]$, 由文本编码器 E_t 的 L 层 Transformer 按顺序进行处理, 得到文本编码器的第 l 层输出结果, 表示为 $T_l = E_t^l(T_{l-1})$, $l = 1, 2, \dots, L$ 。文本编码器最后一层的输出 T_L 中的最后一个 token T_L^N 通过文本投影层 P_t 投影到公共视觉-语言嵌入空间, 以获得最终的文本表示 $F_t = P_t(T_L^N)$ 。

采用 ViT 作为视觉编码器, 输入图像被划分为固定大小的图像块, 这些图像块被映射到块嵌入, 以获得初始图像块嵌入 $V_0 \in \mathbb{R}^{M \times d_v}$, 其中 M 是图像块的数量, d_v 是视觉特征的维度。将可学习的 [CLS] token c_v 、视觉 tokens 和从文本提示投影得到的视觉提示 P'_i 拼接在一起, 作为视觉编码器的输入 $I_l = [c_v, V_l, P'_i]$, 最后得到视觉编码 E_v 的第 l 层输出, 表示为 $I_l = E_v^l(I_{l-1})$, $l = 1, 2, \dots, L$ 。

2.2.2 多级适配器增强模块

尽管视觉 Transformer 在图像理解任务中表现出色, 但在零样本草图检索场景中, 预训练的 ViT 模型往往难以充分捕捉草图的抽象特性和结构信息。传统设计通常仅在网络末端插入单一适配器, 这限制了模型对不同层次特征的针对性增强。为此, 提出了多级适配器增强模块(MAEM), 具体结构如图 1 所示, 通过在 ViT 的关键层次部署适配器, 实现对不同级别特征的层次化增强。

适配器的结构设计为轻量级的瓶颈结构, 可以在保持参数效率的同时实现有效的特征变换。给定输入特征 $x \in \mathbb{R}^d$, 适配器 F_{adapter} 结构的表达式为 $h = \text{GELU}(W_{\text{down}} \cdot x)$, $x_{\text{out}} = x + W_{\text{up}} \cdot h$, 其中, $W_{\text{down}} \in \mathbb{R}^{d \times (d/r)}$ 是降维投影矩阵, $W_{\text{up}} \in \mathbb{R}^{(d/r) \times d}$ 是升维投影矩阵, $r = 4$ 是固定的缩减率, x_{out} 是通过残差连接得到的, 以确保训练的稳定性。

在视觉编码器的第 $l \in L = \{3, 6, 9\}$ 层部署适配器, 这些层分别对应 ViT 中不同抽象级别的特征: 浅层特征主要编码边缘和轮廓信息, 中层特征整合局部特征的结构信息, 深层特征形成高级语义概念。对于每一层 $l_i \in L$, 适配器仅应用于代表全局图像的 [CLS] token c_{l_i} , 表示为 $\tilde{c}_{l_i} = F_{\text{adapter}}^{l_i}(c_{l_i})$, 其中, $\tilde{c}_{l_i}$ 是第 l_i 层适配器优化后的 [CLS] token。该 token 随后被放回序列中替换原始 token, 表示为 $[c_{l_i}, V_{l_i}, P'_{l_i}] \leftarrow [\tilde{c}_{l_i}, V_{l_i}, P'_{l_i}]$ 。最终, 获得适配器增强的特征集合 $A = \{\tilde{c}_3, \tilde{c}_6, \tilde{c}_9\}$ 。

2.2.3 自适应特征融合模块

单一的特征表示难以充分适应草图与自然图像之间的领域差异, 并且传统的特征融合方法通常采用简单的拼接或平均池化, 未能充分考虑不同层次特征的贡献差异。为此, 提出了自适应特征融合模块(AFFM), 通过可学习的权重机制动态整合多级特征, 实现从局部到全局的全面表征。

自适应特征融合模块将来自不同层次的适配器增强特征进行融合。首先进行特征拼接, 表示为 $f_{\text{concat}} = [\tilde{c}_3 \oplus \tilde{c}_6 \oplus \tilde{c}_9]$, 其中, $\oplus$ 表示特征维度的拼接操作, $\tilde{c}_3, \tilde{c}_6, \tilde{c}_9$ 分别表示来自第 3、6、9 层的适配器增强后的 [CLS] token。拼接后的特征通过一个非线性融合网络进行维度变换和特征交互, 表示为 $f_{\text{fused}} = \text{GELU}(W_{\text{fusion}} \cdot f_{\text{concat}} + b_{\text{fusion}})$, 其中, $W_{\text{fusion}} \in \mathbb{R}^{d \times 3d}$ 是一个融合权重矩阵, $b_{\text{fusion}} \in \mathbb{R}^d$ 是偏置项, 通过引入 GELU 激活函数进行非线性变换。视觉特征表示经过一个线性投影层 W_{proj} 映射到公共嵌入空间中, 表示为 $F_{\text{fused}} = W_{\text{proj}}(f_{\text{fused}})$ 。

视觉编码器的最后一层输出中的 [CLS] token c_L 通过视觉投影层 P_v 投影到公共视觉-语言嵌入空间上, 得到原始视觉特征, 表示为 $F_v^L = P_v(c_L)$ 。同时, 引入可学习的融合权重参数来平衡多级融合特征和原始特征的重要性, 表示为 $\alpha = \sigma(\beta) = 1/(1 + e^{-\beta})$, $F_v = \alpha \cdot F_{\text{fused}} + (1 - \alpha) \cdot F_v^L$, 其中, β 是可学习的标量参数, $\sigma(\cdot)$ 是 Sigmoid 函数, 以确保权重在 $(0, 1)$ 范围内。

2.3 损失函数

2.3.1 对比学习损失

对比学习损失通过计算视觉和文本表示之间的余弦相似度, 将视觉特征与相应的文本表示对

齐,损失函数 L_{con} 定义为 $L_{\text{con}} = \frac{1}{N} \sum_{i=1}^N -\ln P(y|F_v^i)$, 其中, $P(y|F_v^i) = \frac{\exp[S(F_v^i, F_y^i)/\rho]}{\sum_{c=1}^C \exp[S(F_v^i, F_c^i)/\rho]}$, ρ 表

示温度超参数, N 表示小批量大小, C 表示可见类别的数量, $S(\cdot)$ 表示余弦相似性。

2.3.2 三元组损失

三元组损失旨在最小化草图 F_s 与正样本图像 F_p^+ 间的距离,同时最大化与负样本图像 F_p^- 的距离,损失函数 L_{tri} 定义为 $L_{\text{tri}} = \max\{D(F_s, F_p^+) - D(F_s, F_p^-) + m, 0\}$, 其中, $D(\cdot, \cdot)$ 是余弦距离函数, m 是边界超参数。

2.3.3 相似性分布匹配损失

相似性分布匹配损失 L_{sdm} 旨在最小化草图图像相似性分布 P_i 和标签匹配分布 Q_i 之间的 Kullback-Leibler(KL)散度,其公式为

$$L_{\text{sdm}} = K_L(P_i \| Q_i) = \frac{1}{2N} \sum_{i=1}^{2N} \sum_{j=1}^{2N} p_{i,j} \ln \left(\frac{p_{i,j}}{q_{i,j} + \epsilon} \right),$$

其中, $\epsilon = 1\text{E}-10$, 用来避免零除法误差; $p_{i,j}$ 是匹配概率; $q_{i,j}$ 是真实匹配概率, $q_{i,j} = Y_{i,j} / \sum_{k=1}^{2N} Y_{i,k}$; $Y_{i,j} = 1$ 表示草图和图像属于同一类别, $Y_{i,j} = 0$ 表示草图和图像属于不同类别。

将上述3个损失函数结合起来进行训练,总损失函数 L 表示为 $L = L_{\text{con}} + L_{\text{tri}} + L_{\text{sdm}}$ 。

3 实验

3.1 数据集和评估指标

采用3个广泛使用的数据集,并遵循既定的数据集划分协议^[5]。Sketchy^[14]数据集包含75471张草图和73002张图像,共有125个类别,其中100个类别用于训练,25个类别用于测试,定义为S-I。另一种划分方法^[1]使用104个类别进行训练,21个类别进行测试,定义为S-II。TU-Berlin^[15]数据集由20000张草图和204489张图像组成,共有250个类别,其中220个类别用于训练,30个类别用于测试。QuickDraw^[16]数据集包含330000张草图和204000张图像,共有110个类别,其中80个类别用于训练,30个类别用于测试。

为进行公平比较,实验采用了与文献[7]相同的评估指标,包括平均精度均值(Mean average precision, mAP)和精确率(Precision, P)。所有类别和前200张图像的平均精度均值分别记为mAP@all和mAP@200,检索前100张图像和200张图像的精确率分别记为 $P@100$ 和 $P@200$ 。

3.2 实验配置

所有实验均在32GB V100 GPU上使用PyTorch框架进行。实验选择CLIP的ViT-B/16骨干网络作为视觉编码器,同时使用基于Transformer的骨干网络作为文本编码器,即 L 设置为12。所提方法采用与文献[13]相同的参数设置,其中输入图像大小设置为 224×224 ,提示学习的提示深度 H 设置为9,提示tokens的数量 n 设置为2,边界参数 m 设置为0.3,温度超参数 ρ 设置为0.07。实验使用初始学习率为 $5\text{E}-06$ 的Adam优化器,在训练过程中使用余弦衰减策略衰减到 $3\text{E}-06$,权重衰减率为 $5\text{E}-04$ 。此外,实验以32的批处理大小训练了40个轮次。

3.3 实验结果

为了验证所提方法的有效性,选择了13种最先进的零样本草图检索方法。表1提供了MAEAF与使用CNN、ViT和CLIP作为骨干模型在3个基准数据集上的比较分析。结果表明,MAEAF在大部分评估指标上均优于现有的最先进的方法。与DRCPL方法相比,所提方法在Sketchy S-I、Sketchy S-II、TU-Berlin和QuickDraw数据集上将平均精度均值分别提高了1.5%、1.3%、2.6%和0.7%,并将精确率分别提高了1.0%、1.2%、2.1%和1.0%。

实验结果表明,使用提示学习的CLIP模型可以更全面地理解零样本草图检索任务中的多域数

据。此外,多级适配器增强模块在浅层、中层、深层部署适配器可以针对性地增强视觉的细节、局部及全局特征,显著提升了模型对视觉多层次特征的建模能力。自适应特征融合模块动态调整多级融合特征与原始视觉编码器输出特征的平衡,解决了单一特征表示难以适应草图抽象特性和自然图像丰富细节的问题,实现了从局部细节到全局语义的全面特征增强。

表 1 MAEAF 与现有方法的性能比较

Table 1 Performance comparison between MAEAF and existing methods

%

骨干网络	方法	会议/期刊	数据集							
			Sketchy S-I		Sketchy S-II		TU-Berlin		QuickDraw	
			mAP@all	P@100	mAP@200	P@200	mAP@all	P@100	mAP@all	P@200
CNN	SEM-PCYC ^[1]	CVPR'19	34.9	46.3	45.9	37.0	29.7	42.6	17.7	18.4
	SAKE ^[7]	ICCV'19	54.7	69.2	49.7	59.8	47.5	59.9	—	—
	Sketch3T ^[5]	CVPR'22	—	—	57.9	64.8	50.7	67.1	—	—
	DDFUS ^[19]	IPM'25	73.3	83.6	—	—	52.8	61.9	15.6	21.6
ViT	TVT ^[20]	AAAI'22	64.8	79.6	53.1	61.8	48.4	66.2	14.9	29.3
	PSKD ^[21]	MM'22	68.8	78.6	56.0	64.5	50.2	66.2	15.0	29.8
	ZSE-SBIR ^[3]	CVPR'23	69.8	79.7	52.5	62.4	54.2	65.7	14.5	21.6
	MFCPKD ^[22]	TCSVT'25	84.7	87.8	73.6	78.8	66.2	72.8	22.4	28.5
CLIP	CLIP-AT ^[12]	CVPR'23	—	—	72.3	72.5	65.1	73.2	20.2	38.8
	MARL ^[23]	WACV'24	—	—	69.1	75.5	70.5	77.7	32.7	42.5
	SpLIP ^[10]	ECCV'24	80.2	86.7	76.4	77.3	<u>73.1</u>	<u>78.2</u>	34.2	44.6
	DRCPL ^[13]	TBD'24	<u>85.9</u>	<u>88.5</u>	74.0	<u>79.4</u>	72.3	76.6	<u>37.0</u>	43.4
	SADLnet ^[24]	ICASSP'25	77.7	83.8	73.6	71.2	69.1	76.6	23.7	30.8
	MAEAF (本文方法)		87.4	89.5	<u>75.3</u>	80.6	74.9	78.7	37.7	<u>44.4</u>

注:粗体和下划线分别表示最优结果和次优结果。

3.4 消融实验

为评估所提模型中多级适配器增强模块(MAEM)与自适应特征融合模块(AFFM)的有效性,在所有基准数据集上进行了消融实验,实验结果如表 2 所示。其中实验 1 为基准模型(未添加 MAEM 和 AFFM),用于验证基础框架的性能;实验 2 仅添加 MAEM,用于单独评估层次化特征增强的效果;实验 3 同时添加 MAEM 和 AFFM,用于验证两个模块的协同作用。

表 2 每个组件的消融实验

Table 2 Ablation experiments of each component

%

实验	MAEM	AFFM	数据集							
			Sketchy S-I		Sketchy S-II		TU-Berlin		QuickDraw	
			mAP@all	P@100	mAP@200	P@200	mAP@all	P@100	mAP@all	P@200
1	×	×	85.9	88.5	74.0	79.4	72.3	76.6	37.0	43.4
2	✓	×	86.5	89.1	74.6	80.0	74.2	78.0	37.3	43.8
3	✓	✓	87.4	89.5	<u>75.3</u>	80.6	74.9	78.7	37.7	44.4

注:粗体表示最优结果。

由表 2 可知,仅添加 MAEM 的实验 2 相比基准实验 1,在所有数据集上的性能都有提升,表明 MAEM 通过对不同层次特征的针对性增强,有效提升了模型的特征表征能力。在此基础上添加 AFFM 的实验 3,模型达到最佳性能,表明 AFFM 通过动态平衡多级融合特征与原始特征的权重,解决了单一特征表示的局限性,与 MAEM 形成了良好的协同效应。

3.5 定性结果和计算量分析

图 2 展示了在 Sketchy 和 TU-Berlin 数据集上 DRCPL 和 MAEAF 方法返回的前 5 个图像检索结果。绿色边框表示返回正确的检索结果,红色边框表示返回错误的检索结果。从图 2 可以看出,所提方法获得了更准确的结果,检索结果再次验证了 MAEAF 方法的有效性。

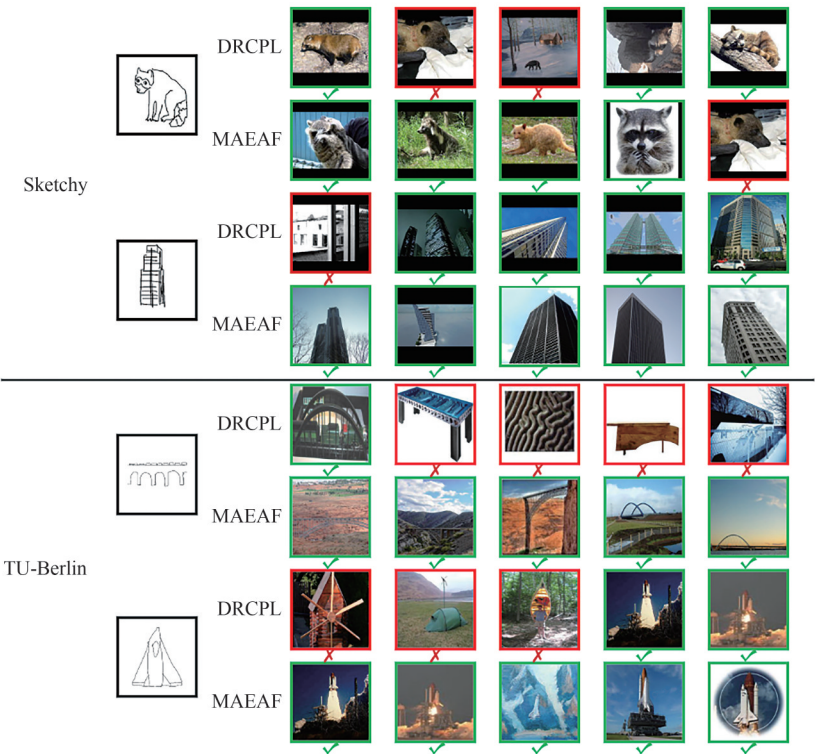


图 2 在 Sketchy 和 TU-Berlin 测试集上 DRCPL 和 MAEAF 的前 5 个检索结果
Fig. 2 Top 5 retrieval results for DRCPL and MAEAF on the Sketchy and TU-Berlin test sets

将所提方法与 4 种零样本草图检索方法(ZS-SBIR)在计算复杂度方面进行了比较。从表 3 中可以看出,所提方法具有较少的可学习参数量,是因为模型仅对视觉提示、文本提示以及多级适配器增强模块(MAEM)和自适应特征融合模块(AFFM)中的参数进行训练。尽管 CLIP-AT 实现了较低的浮点计算量,但所提方法在合理增加计算量的情况下实现了更好的检索性能。

表 3 MAEAF 与现有方法的计算复杂度比较

Table 3 Comparison of computational complexity between MAEAF and existing methods				
方法	骨干网络	总参数量/M	可学习参数量/M	FLOPs/G
SEM-PCYC ^[1]	GAN+VGG-16	137.9	137.9	15.5
ZSE-SBIR ^[3]	ViT-B/16	102.2	102.2	19.5
CLIP-AT ^[12]	ViT-B/32	151.3	31.5	3.1
DRCPL ^[13]	ViT-B/16	132.6	8.3	11.4
MAEAF(本文方法)	ViT-B/16	135.2	10.9	11.4

4 结论

本文提出了一种基于多级适配器增强与自适应特征融合的零样本草图检索方法(MAEAF),该方法设计了多级适配器增强模块,通过在不同深度(浅层、中层、深层)部署轻量级瓶颈结构适配器,分别增强边缘轮廓、局部结构和全局语义的特征表示,显著提升了模型对多层次视觉特征的建模能力。此外,提出了一个自适应特征融合模块,该模块动态调整多级融合特征与原始输出特征的平衡,实现了从局部到全局的全面特征增强。在3个基准数据集上进行了全面的实验,结果表明,所提方法能够有效地对齐草图和图像特征,并具有优异的零样本草图检索性能。

参考文献:

- [1] DUTTA A, AKATA Z. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 5084-5093.
- [2] 高志杰. 基于视觉关系的零样本草图检索研究[D]. 成都: 电子科技大学, 2024.
- [3] LIN F Y, LI M K, LI D, et al. Zero-shot everything sketch-based image retrieval, and in explainable style[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada: IEEE, 2023: 23349-23358.
- [4] 司佳宝, 赵翔宇, 刘和鑫, 等. STConVM: 基于多视图多模态的对比学习方法识别空间域[J]. 内蒙古大学学报(自然科学版), 2025, 56(5): 532-544.
- [5] SAIN A, BHUNIA A K, POTLAPALLI V, et al. Sketch3T: Test-time training for zero-shot SBIR[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE, 2022: 7452-7461.
- [6] YU Q, YANG Y X, LIU F, et al. Sketch-a-net: A deep neural network that beats humans[J]. International Journal of Computer Vision, 2017, 122(3): 411-425.
- [7] LIU Q, XIE L X, WANG H Y, et al. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea: IEEE, 2019: 3661-3670.
- [8] ZHOU Y H, LIU D W, MOK P Y. Zero-shot sketch based image retrieval via modality capacity guidance[C]//Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. Jeju, Republic of Korea: IJCAI, 2024: 1780-1787.
- [9] DONG S Y, ZHU M R, WANG N N, et al. Adapt and align to improve zero-shot sketch-based image retrieval [DB/OL]. (2023-05-09)[2025-09-13]. <https://arxiv.org/abs/2305.05144v1>.
- [10] SINGHA M, JHA A, GUPTA D, et al. Elevating all zero-shot sketch-based image retrieval through multimodal prompt learning[C]//Computer Vision-ECCV 2024. Cham: Springer, 2025: 1-19.
- [11] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the International Conference on Machine Learning. Online: PMLR, 2021: 8748-8763. <https://arxiv.org/abs/2103.00020>.
- [12] SAIN A, BHUNIA A K, CHOWDHURY P N, et al. CLIP for all things zero-shot sketch-based image retrieval, fine-grained or not[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada: IEEE, 2023: 2765-2775.
- [13] ZHUO G Y, SHU Z Q, YU Z T. Deep residual coupled prompt learning for zero-shot sketch-based image retrieval [J]. IEEE Transactions on Big Data, 2025, 11(3): 1493-1507.
- [14] SANGKLOY P, BURNELL N, HAM C, et al. The sketchy database: Learning to retrieve badly drawn bunnies[J]. ACM Transactions on Graphics, 2016, 35(4): 119.
- [15] EITZ M, HAYS J, ALEXA M. How do humans sketch objects?[J]. ACM Transactions on Graphics, 2012, 31(4):

- 1-10.
- [16] HA D, ECK D. A neural representation of sketch drawings[DB/OL]. (2017-04-16)[2025-09-13]. <https://arxiv.org/pdf/1704.03477v2>.
 - [17] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[DB/OL]. (2020-10-22)[2025-09-13]. <https://arxiv.org/abs/2010.11929v2>.
 - [18] ZHOU K Y, YANG J K, LOY C C, et al. Learning to prompt for vision-language models[DB/OL]. (2021-09-02)[2025-09-13]. <https://arxiv.org/abs/2109.01134>.
 - [19] ZHANG Q, ZHANG J, SU X D, et al. Domain disentanglement and fusion based on hyperbolic neural networks for zero-shot sketch-based image retrieval[J]. *Information Processing & Management*, 2025, 62(1): 103963.
 - [20] TIAN J L, XU X, SHEN F M, et al. TVT: Three-way vision transformer through multi-modal hypersphere learning for zero-shot sketch-based image retrieval[C]//*Proceedings of the AAAI Conference on Artificial Intelligence*. Vancouver, Canada: AAAI, 2022: 2370-2378.
 - [21] WANG K, WANG Y F, XU X, et al. Prototype-based selective knowledge distillation for zero-shot sketch based image retrieval[C]//*Proceedings of the 30th ACM International Conference on Multimedia*. New York, NY, USA: ACM, 2022: 601-609.
 - [22] LI C X, ZHANG D L, HU Z Z, et al. Modality fused class-proxy with knowledge distillation for zero-shot sketch-based image retrieval[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025, 35(6): 6158-6169.
 - [23] LYOU E, LEE D, KIM J, et al. Modality-aware representation learning for zero-shot sketch-based image retrieval [C]//*2024 IEEE/CVF Winter Conference on Applications of Computer Vision*. Waikoloa, HI, USA: IEEE, 2024: 5634-5643.
 - [24] ZHANG Q, ZHANG J, BAO F L, et al. Structural-aware disentangled learning with CLIP for hyperbolic zero-shot sketch-based image retrieval[C]//*2025 IEEE International Conference on Acoustics, Speech and Signal Processing*. Hyderabad, India: IEEE, 2025: 1-5.

(责任编辑 黄宝琦)

Zero-Shot Sketch-Based Image Retrieval Based on Multi-Level Adapter Enhancement and Adaptive Feature Fusion

WANG Weiguang^{1,2}, WEI Hongxi^{1,2}, HE Chao^{1,2}

(1. *College of Computer Science, Inner Mongolia University, Hohhot 010021, China;*

2. *National and Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, Inner Mongolia University, Hohhot 010021, China*)

Abstract: To address the limitations of existing zero-shot sketch-based image retrieval (ZS-SBIR) methods, which rely on a single deep feature leading to inadequate feature extraction and insufficient feature fusion, the approach of multi-level adapter enhancement and adaptive feature fusion (MAEAF) is proposed. Adapters are embedded in different network layers, effectively enhancing the multi-level feature representation capabilities of edge contours, local structures, and global semantics. Furthermore, an adaptive fusion module is proposed to dynamically fuse multi-level enhanced features with original visual features using learnable weights, achieving the integration of local and global information. Finally, extensive experiments are conducted on three benchmark datasets. The experimental results demonstrate that the proposed method MAEAF outperforms existing ZS-SBIR methods.

Key words: zero-shot sketch-based image retrieval; zero-shot learning; CLIP; vision transformer; prompt learning