基于语义感知对比学习与Transformer的基于关键词的多样化图像检索

赵敏怡*
zhaomy20@fudan.edu.cn 计算机科学学院 复旦大学 上海，中国

王金鹏
wjp20@mails.tsinghua.edu.cn 清华大学深圳国际研究生院，清华大学 中国深圳

廖栋梁†
b
rightliao
@tencent.com 微信事业群 腾讯公司 中国广州

王怡如
d
orisyr
wang@tencent.com 微信事业群 腾讯公司 北京，中国

段焕中
b
oosterduan
@tencent.com 微信事业群 腾讯公司 北京，中国

周水庚†
sgzhou@fudan.edu.cn 计算机科学学院 复旦大学 中国上海

arXiv:2305.04072v1 [cs.IR] 6 May 2023

摘要

除了相关性，多样性也是跨模态图像检索系统一项重要但研究较少的性能指标，对用户体验至关重要。现有的多样性感知图像检索解决方案，要么对标准检索系统返回的原始检索结果进行显式后处理，要么尝试学习图像的多向量表示，以表征其多样化语义。然而，这两种方法都不足以平衡相关性与多样性。一方面，标准检索系统通常偏向共同语义，训练时很少采用多样性感知正则化，因此难以通过后处理提升多样性。另一方面，多向量表示方法无法保证学到稳健的多个投影。因此，无关图像以及具有罕见或独特语义的图像可能会被错误投影，从而降低某些典型算法（如 top-k）所生成结果的相关性和多样性。为解决这些问题，本文提出了一种名为 CoLT 的新方法，旨在生成更具代表性且稳健的表示，以准确分类图像。具体而言，CoLT 首先通过语义感知对比学习增强现有一对一跨模态系统的初步表示，从而提取语义感知图像特征。随后，开发了一种基于 Transformer 的词元分类器，将所有特征归入相应类别。最后，设计了一种后处理算法，从各个类别中检索图像，构成最终检索结果。在 Div400 和 Div150Cred 两个真实世界数据集上进行的大量实验表明，CoLT 能有效提升多样性，整体表现优于现有方法（$F1$ 分数更高）。

CCS 概念

• **信息系统 →信息检索多样性**。

关键词

跨模态检索、基于关键词的图像检索、多样化检索、Transformer

ACM 引用格式：

赵敏怡、王金鹏、廖栋梁、王怡如、段焕中和周水庚。2023年。基于关键词的多样化图像检索：语义感知对比学习与Transformer。在第*46*届国际*ACM*信息检索研究特别兴趣小组年会信息检索研究与发展会议（*SIGIR '23*），*2023*年*7*月*23*日至*27*日，中国台湾台北。ACM，美国纽约州纽约，11页。https://doi.org/10.1145/3539618.3591705

1 引言

随着万维网及其应用的普及，越来越多的数据被创建并发布到万维网上，这推动了网页搜索和信息检索技术的快速发展 [41, 57]。其中，跨模态数据检索 [2, 3, 46, 46, 48, 62, 63, 71]使用户能够便捷地获取不同形式的所需信息。一个典型示例是跨模态图像检索（简称CMIR） [5, 37, 49, 50, 59, 66]。CMIR以文本查询作为输入，检索语义匹配的图像，已被部署于Instagram和Flickr等众多网络应用中，并受到越来越多的研究关注 [12, 21, 27, 29, 54]。

近年来，得益于从万维网上抓取的大量图像-文本对，相关研究 [28, 30, 50, 60]和大规模一对一预训练编码器 [36, 53]显著推动了跨模态图像检索系统的发展。然而，现有模型往往会返回一组语义相似的检索图像。普通用户提交的查询，尤其是以缺乏具体上下文 [35, 47],的基于关键词的短文本形式提交的查询，通常很可能具有宽泛的语义，因此语义模糊且不确定。例如，给定粗粒度查询“狗”，用户可能期望获得具有多样化语义的狗的图像（例如 不同的品种、颜色和体型）。

*这项工作的主要部分是在作者于腾讯实习期间完成的。
†通讯作者。

特此授予许可：无需支付费用，即可为个人或课堂使用而制作本作品的全部或部分数字或纸质副本，但前提是不得为营利或商业利益而制作或分发副本，并且副本首页须附有本声明及完整引文。对于本作品中由作者以外的其他人拥有的组成部分，必须尊重其版权。注明出处后可进行摘要。若要以其他方式复制、再版、发布到服务器或重新分发至列表，必须事先获得具体许可和/或支付费用。请通过 permissions@acm.org 申请许可。*SIGIR '23，2023*年*7*月*23*日至*27*日，中国台湾台北。© 2023年版权归所有者/作者所有。出版权授权予 ACM。ACM ISBN 978-1-4503-9408-6/23/07...$15.00 https://doi.org/10.1145/3539618.3591705

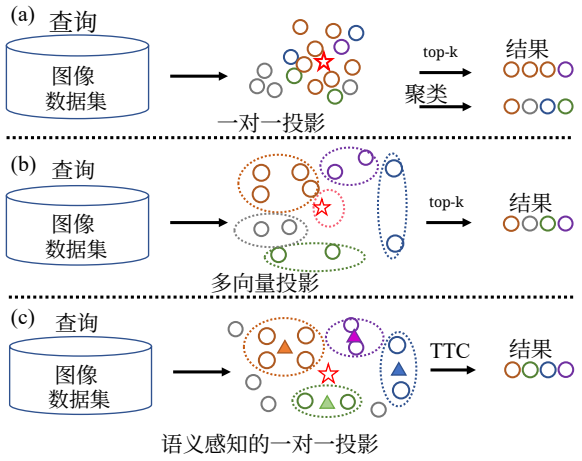

图1：（a）典型跨模态图像检索系统；（b）基于学习的多向量检索系统；（c）我们的CoLT方法的示意图。红色星形表示查询。不同颜色的点表示具有不同语义的图像。灰色点表示无关图像，三角形表示相应语义的原型。虚线圆圈表示投影区域。

显然，基于关键词的查询容易匹配到各种检索结果，但语义相似的图像列表无法满足不同用户的多样化需求，从而降低他们的检索体验 [43, 64]。

为了解决上述缺陷，本文提出了基于关键词的多样化图像检索 [15, 17, 18, 20], 任务，该任务以一段简短的基于关键词的文本作为输入，检索出相关性高且语义多样性丰富的图像列表。近期方法大致可分为两类。第一类是基于后处理的方法 [10, 25, 33, 38–40, 40, 61]。这类方法通常使用现有的跨模态编码器提取特征。随后，采用各种算法（例如重排序 [61] 和聚类 [33]）来提升多样性。然而，由于一对一投影的局限性，这些方法往往无法获得兼顾相关性与多样性的优质检索列表。例如，如图1（a）所示，一方面，在典型的一对一投影中，**(W1)** 查询特征（红色星形）可能会被常见语义的图像（棕色点）包围，这是由训练数据的长尾分布造成的，进而使前k个结果集充斥着语义相似的图像。另一方面，**(W2)** 由于忽略了对多样性的建模，不同语义的图像特征较难区分 [64], ，这会损害聚类等算法的性能。

第二类是一组基于学习的方法 [1, 42, 43, 52, 64, 70]，尝试利用各种技术（例如图度量学习 [5, 46] 和多实例学习 [55, 70]）来建模多样性。与将每幅图像投影到潜在空间中的一个向量的一对一投影相比，这些方法 [42, 52, 64]会将每幅图像（或文本查询）嵌入到相关特征周围的多个向量中，以获得其多样化表示，从而进行 top-k 搜索，即多向量投影。遗憾的是，这种投影不够稳健，无法处理具有罕见或独特语义的图像。如图 1(b) 所示，（**W3**）一些无关的离群点（灰色点）会被错误地投影为多样性的代表。此外，（**W4**）一些具有罕见或独特语义的图像（蓝色点）很可能会被投影到某些 top-k 算法无法到达的遥远区域。

为克服现有方法的不足（即，**W1~W4**），本文提出一种名为 CoLT 的新方法（即语义感知**Co**对比**L**学习与Transformer 的缩写）用于基于关键词的图像检索。具体而言，为了克服**W1**、**W2**和**W3**，CoLT 借助一种新颖的语义感知对比学习（SCL）损失，提取稳定、具有代表性且易于区分的图像特征。如图 1(c) 所示，SCL 的核心思想是将语义相似的图像（例如，同一品种的狗）投影到与其匹配的语义原型周围的向量空间中，同时使该原型与其他原型（例如，不同品种的狗和无关图像）以及查询特征保持适当距离，从而更好地建模多样性。为处理包含稀有语义的图像（**W4**），CoLT 不采用现有工作中的 top-k 算法，而是使用强大的基于 *Transformer* 的词元分类器（TTC）生成最终检索结果。具体而言，在 TTC 中，图像特征与查询特征被拼接为输入令牌序列。随后，TTC 将每个令牌分类到相关语义类别，以区分包含不同语义的图像。最后，我们设计了一种灵活的后处理算法，从不同语义类别（包括常见和稀有语义）中选择图像，组成最终检索结果。这种设计为我们的方法带来四方面优势：（i）语义相关性高。CoLT 提升了预训练跨模态编码器中稳健的一对一投影，其稳定性远高于近期基于多向量投影的方法。（ii）语义多样性高。CoLT 不仅通过语义感知对比学习使图像特征更易区分，还利用基于 Transformer 的词元分类器挖掘稀有语义。（iii）通用且易于使用。CoLT 可直接叠加在各种现有跨模态编码器的末端，无需修改其结构和参数，并以即插即用方式提升性能。（iv）易于控制。我们可以修改 CoLT 中的后处理算法，灵活平衡语义相关性和语义多样性，而无需重新实现模型。

本文的贡献总结如下：(1) 我们指出了现有方法的局限性，并提出了一种名为 CoLT 的基于关键词的多样化图像检索新方法。CoLT 首先提取高质量且具有区分性的语义感知特征，然后对这些特征进行分类，以生成最终检索列表。(2) 我们在 CoLT 中设计了语义感知对比损失，以提取更稳健、更具代表性的特征。(3) 为了更好地挖掘语义多样性，我们设计了一种基于 Transformer 的词元分类器来生成检索结果。(4) 我们在两个真实世界数据集 Div400 和 Div150Cred 上进行了大量实验，结果表明，我们的方法能有效提升多样性，整体表现优于现有方法，并取得更高的 $F1$ 分数。

表1：从特征投影、检索结果生成和性能三个维度，对CoLT与现有主要方法进行定性比较。

方法	投影	生成	性能
CLIP [36]	一对一	top-k	低多样性
MMR [38]	一对一	重排序	低多样性
UMONS [40]	一对一	聚类	低相关性
VMIG [64]	多向量	top-k	中等相关性与多样性
CoLT（我们的方法）	SCL	TTC	高相关性与多样性

2 相关工作

2.1 跨模态图像检索

典型的跨模态图像检索方法 [34] 大致可分为两类：基于跨模态相似度测量的方法 [44, 58, 65, 73]，直接计算跨模态距离；以及基于公共空间学习的方法 [4, 31,46, 56, 71]，通过注意力机制、生成对抗网络等多种技术，将查询和图像映射到共享空间。如今，得益于 Transformer 结构和预训练技术，大规模预训练编码器（例如 CLIP [36], ALIGN [22], GroupViT [53], 和 U-BERT [60]）已在基于相关性的检索任务中展现出优势。尽管这些方法显著提升了检索相关性，但它们忽视了对语义多样性的建模，从而损害了检索列表的语义多样性。

2.2 多样化检索

现有的多样化检索方法大致分为两类。第一类是基于后处理的方法 [6, 10, 25, 33, 38–40, 40, 61], ，通常使用现有的特征编码器 [7, 8, 14,26, 36] 生成特征，然后通过后处理算法挖掘多样性。其中， [40] 首先过滤掉不相关的图像，再通过 DBSCAN [9] 对其余图像进行聚类，以提升多样性。MMR [38] 被提出用于对检索列表重新排序，以平衡多样性与相关性。Bo 和 Gao [1] 提取关键词来控制结果的多样性。第二类包括近期提出的基于学习的方法，旨在潜在空间中表示语义多样性 [1, 42, 43, 52, 64, 70]。具体而言，Su等人 [43] 提出了一种动态意图图（GRAPH4DIV），以平衡文档中的内容与意图。Song 和 Soleymani [42] 利用多个 Transformer 提取视觉特征。Wu 和 Ngo [52]设计了一种非活跃词损失，以扩展语义概念，从而表示多种视频内容。VMIG [64] 通过多实例学习将每个图像和文本查询嵌入为多个向量。尽管这些方法成功提升了检索结果的语义多样性，但在保证语义相关性以及挖掘具有稀有语义的图像方面表现不佳。

2.3 我们的方法与现有工作的区别

为了阐述 CoLT 与典型现有方法之间的差异，我们在表 1 中从三个维度进行了定性比较：图像和查询如何投影？如何处理这……最终生成的检索结果如何？在语义相关性和多样性方面的性能又如何？如表 1 所示，近期的预训练跨模态编码器（例如 CLIP [36]）由于一对一投影的局限性，无法很好地建模语义多样性。两种典型的基于后处理的方法 MMR 和 UMONS，要么由于缺乏准确的多样性度量机制而难以建模多样性 [38]，要么由于将不相关的特征聚类在一起而无法保证相关性。最近提出的 VMIG 由于多向量投影的不确定性而存在鲁棒性问题，并且由于 top-k 搜索算法而导致罕见语义处理问题，从而带来不理想的性能。我们的方法 CoLT 是唯一同时实现高语义相关性和丰富语义多样性的检索方法，这得益于所提出的 语义感知对比学习 （SCL）和强大的 基于 *Transformer* 的词元分类器（TTC）。实验和可视化研究证明了我们方法的优势。

3 方法

3.1 概述

给定一个文本查询 Q 和一个图像数据集 D，我们的目标是生成一个检索列表 R，其中包含 K 张语义相关性高且具有多样性的图像。图 2 展示了我们的方法 CoLT 的架构，该方法由六个组件组成：一个固定特征编码器f，以 Q 和 D 为输入，生成初始查询特征 h_q 和视觉特征 $\{h_v^i\}$，即 $\{h_q,\ \{h_v^i\}\} = f(Q,\ \mathrm{D})$；一个视觉特征重编码器 g，借助语义感知对比学习（SCL）模块对视觉特征进行重编码；以及基于 *Transformer* 的词元分类器（TTC） ϕ，它将查询特征 h_q 和重编码后的图像特征 $\hat{}$ h_v^i 作为输入令牌序列，根据各词元的表示将其归入适当的语义类别。TTC 模块由两个子模块组成：令牌分类 Transformer，由 L 个 Transformer 编码器层构成；以及作为分类器的全连接层。最后，采用后处理模块，从这些类别中选出典型图像作为最终结果。在训练期间，使用逐词元数据增强模块充分利用训练数据，该模块置于 SCL 模块与 TTC 模块之间。

3.2 语义感知对比学习

在 CoLT 中，我们首先使用一个固定的预训练特征编码器 f提取初步的高质量且鲁棒的视觉特征和查询特征。然而，如上所述，这些一对一投影特征的区分度不足，无法支持有效的多样化检索。因此，我们采用由多层感知器实现、并由一种新颖的语义感知对比学习驱动的视觉特征重编码器 g，来优化图像特征，从而提升语义多样性。具体而言，对于每个视觉特征h_v^i，我们按如下方式对其表示进行重编码：

$$\hat{h}_v^i = h_v^i + \beta g(h_v^i), \tag{1}$$

其中，β是一个超参数，用于控制学习到的重新编码特征$g(h_v^i)$。接下来，我们将详细介绍所提出的语义感知对比损失。

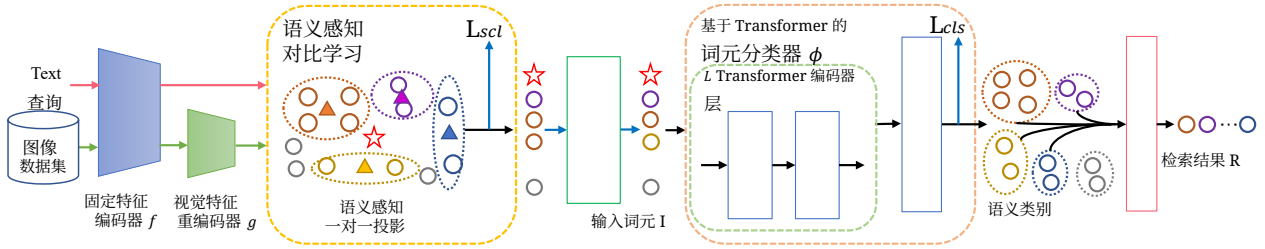

图 2：CoLT 的架构。蓝色连线仅在训练期间有效。

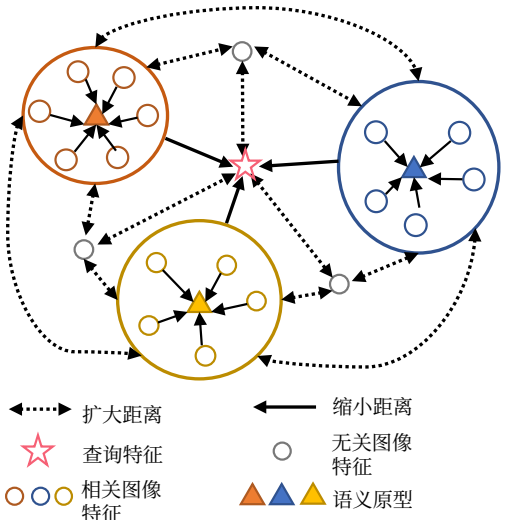

图 3：语义感知对比学习示意图。不同颜色表示不同类别。

如图 3 所示，语义感知对比学习（SCL）的目标是：（1）增大查询特征与无关特征之间的距离；（2）增大相关图像特征与无关图像特征之间的距离；（3）增大不同语义的图像特征之间的距离，使这些特征更易区分，同时提升多样性；（4）缩小查询特征与相关图像特征之间的距离，这与（1）和（2）一样有助于提高准确率；（5）缩小相似语义的图像特征之间的距离。在 SCL 中，我们使用存储在存储库 B 中的语义类别原型，高效地计算（3）和（5），从而避免使用过大的批次大小。因此，每个图像查询都会被投影到一个位置，使查询特征与其匹配的语义原型、不匹配的语义原型以及无关图像特征之间的距离恰当。

此处讨论所提出的语义感知对比学习的实现。在 SCL 中，正样本对包括：(1) 相关图像-查询特征对，以及 (2) 相关图像-类别原型对；负样本对包括：(3) 无关图像-查询特征对，以及 (4) 无关图像-类别原型对。对于查询 h_q，其相关图像特征集合为 $\{\hat{h}^{r,}{}_{v}{}^{i}\}$，无关图像特征集合为 $\{\hat{h}^{ir,}{}_{v}{}^{i}\}$。令 B($i$) 表示 i–个存储在存储库中的语义类别原型，$G(\cdot)$ 是一个将图像特征映射到相匹配的语义类别原型对应索引的函数，则 SCL 的损失可表示如下：

$$\mathcal{L}_{scl} = -log\frac{\overbrace{\Sigma_i exp(h_q \cdot \hat{h}_v^{r,i}/\tau)}^{(1)} + \overbrace{\Sigma_i exp(\mathcal{B}(G(\hat{h}_v^{r,i})) \cdot \hat{h}_v^{r,i}/\tau)}^{(2)}}{\underbrace{\Sigma_i exp(h_q \cdot \hat{h}_v^{ir,i}/\tau)}_{(3)} + \underbrace{\Sigma_{i,j} exp(\mathcal{B}(j) \cdot \hat{h}_v^{ir,i}/\tau)}_{(4)} + (1) + (2)}, \tag{2}$$

其中 τ 是用于控制温度的超参数。

存储库 B 中存储的类别原型在所提出的 SCL 中发挥着重要作用。因此，需要在训练期间对其进行初始化和更新，以获得准确且最新的表示。因此，我们使用固定特征编码器提取的细粒度文本描述特征来初始化存储库。至于更新，则使用指数移动平均（EMA）[11, 24, 68, 72] 来更新类别原型：

$$\mathcal{B}(G(\hat{h}_v^{r,i})) = \alpha\mathcal{B}(G(\hat{h}_v^{r,i})) + (1-\alpha)\hat{h}_v^{r,i}, \tag{3}$$

其中，α 是用于更新存储库的动量系数。

3.3 基于 Transformer 的词元分类

获得一组表征特征后，接下来的问题是如何生成兼具高相关性和多样性的最终结果。为此，我们开发了一种强大的基于 *Transformer* 的词元分类器 (TTC)，用于进行特征融合和词元分类。具体而言，我们将每个特征视为一个词元，并将查询特征h_q和 N图像特征 $\{\hat{h}_v^i\}_{i=1}^N$ 拼接起来，形成输入令牌序列，即 $\mathrm{I} = [h_q, \{\hat{h}_v^i\}_{i=1}^N]$。此处，为避免出现无关词元，我们仅使用与查询特征在语义上最相似的 N图像特征。我们根据图像特征与查询特征之间的余弦相似度对图像特征进行排序，并根据其细粒度语义标注生成相应的真实标注 $\mathrm{Y} = \{y_i\}_{i=1}^{N+1}$。值得一提的是，在 TTC 中，所有无关图像特征和查询特征都标记为特殊索引，以便进一步区分。如图 2 所示，L Transformer 编码器层 [45] 采用多头自注意力和前馈层来融合这些词元。随后，堆叠一个全连接层作为分类器进行预测。形式上，TTC 的预测可写为：

$$\{p_i\}_{i=1}^{N+1} = \phi(\mathcal{I}), \tag{4}$$

其中， p_i 表示第 i–个词元的预测分布概率。采用交叉熵损失作为分类损失来训练 TTC:

$$\mathcal{L}_{cls} = -\Sigma_i y_i log(p_i). \quad (5)$$

将每个图像词元分类到相应类别后，应用后处理算法 t 生成最终检索列表。也就是说，从每个语义类别中选出与查询特征相似度最高的 X 张图像:

$$\mathcal{R} = t(\{p_i\}_{i=1}^{N+1}, X). \quad (6)$$

最后，从 $\lfloor k/X \rfloor$ 个语义类别中选出图像后，即可得到长度为 k 的检索列表 R。

3.4 逐词元数据增强

由于缺乏特定的细粒度标注，直接在 D 上训练模型容易导致过拟合。因此，为了更好地发挥基于 Transformer 的词元分类模块的潜力，我们采用逐词元数据增强来丰富输入词元 I。具体而言，我们引入了四种不同的核操作:

查询扰动: 我们按如下方式扰动查询特征 h_q : 以Pq的概率，将查询特征与相关图像特征 $\hat{h}^{r,\ v^i}$进行 MIXUP [67] 。形式上，令 $\lambda \sim Beta$(1。0，1。0)，我们按如下方式生成扰动后的查询特征:

$$h_q = max(\lambda, 1.0 - \lambda)h_q + min(\lambda, 1.0 - \lambda)\hat{h}_v^{r,i}. \quad (7)$$

图像扰动: 我们对图像特征 $\hat{h}^{r,\ v^i}$进行如下处理：以 p_v 的概率将图像特征与相关查询特征 h_q 混合。通过从 $Beta$(1.0，1.0) 中采样 λ，我们得到

$$\hat{h}_v^{r,i} = max(\lambda, 1.0 - \lambda)\hat{h}_v^{r,i} + min(\lambda, 1.0 - \lambda)\hat{h}_q. \quad (8)$$

删除: 我们以 p_d 的概率删除图像特征。**复制:** 我们以 p_c 的概率复制图像特征。

在这 4 种操作中，查询扰动和图像扰动直接增强特征，而不修改语义感知表示，这有利于提升模型的鲁棒性；删除和复制操作则分别增强模型区分罕见标记和相似标记的能力。遵循 [51]，中的经验，我们按如下方式对输入标记 I 进行数据增强:依次采样以下数据增强操作：（1）查询扰动；（2）删除；（3）复制；（4）图像扰动，然后对每个标记分别执行选中的操作。

3.5 训练与评估算法

CoLT 的训练过程如算法1所示，可分为三个步骤。首先，通过 f(L2) 提取初始查询特征和图像特征。然后，使用所提出的语义感知对比学习（L3-L9）训练视觉特征重编码器，将初步特征重新编码为语义感知特征。最后，以查询特征和重新编码后的图像特征为输入，借助逐词元数据增强（L10-L16）训练基于 Transformer 的词元分类器。

评估过程如算法2所示，具体如下：我们首先生成初始特征（L2），然后对图像特征重新编码（L3）。随后，将这些特征作为标记，以生成预测分布概率（L4-L5）。最后，使用后处理算法 t 生成最终检索列表R。

算法1 CoLT 的训练。

输入： 固定特征编码器 f、视觉特征重编码器 g、基于 Transformer 的词元分类器 ϕ、查询 Q和图像数据集 D
$h_q, \{h_v^i\} = f(Q, \mathrm{D})$
使用细粒度描述初始化 B
当 g 尚未收敛**时，执行**
    $\hat{h}_v^i = h_v^i + \beta g(h_v^i)$
    $\hat{h}^{r,v^i}, \hat{h}^{ir,v^i} \sim g(h_v^i)$
    通过公式（2）计算 L_{scl}
    根据 L_{scl} 优化 g
    根据式（3）更新 B
当 ϕ 未收敛时 **执行**
    $\mathrm{I} = [h_q, \{\hat{h}_v^i\}_{i=1}^N]$
    根据第 3.4 节对 I 进行数据增强
    根据第 3.3 节获得最终的 I
    $\{p_i\}_{i=1}^{N+1} = \phi(\mathrm{I})$
    通过公式 (5) 计算 L_{cls}
    根据 L_{cls} 优化 ϕ
返回 g 和 ϕ

算法 2 CoLT 的评估。

输入： 固定特征编码器 f、视觉特征重编码器 g、基于 Transformer 的词元分类器 ϕ、查询 Q、图像数据集 D，以及后处理算法 t 及其超参数X2: $h_q, \{h_v^i\} = f(Q, \mathrm{D})$ 3: $\hat{h}_v^i = h_v^i + \beta g(h_v^i)$ 4: $\mathrm{I} = [h_q, \{\hat{h}_v^i\}_{i=1}^N]$5: $\{p_i\}_{i=1}^{N+1} = \phi(\mathrm{I})$ 6: $\mathrm{R} = t(\{p_i\}_{i=1}^{N+1}, X)$ 7: **返回**R

4 性能评估

4.1 研究问题

本节通过开展大量实验来评估所提出的方法，以回答以下研究问题:
RQ1: 与当前最先进的跨模态图像检索模型相比，CoLT 在相关性和多样性方面表现如何?

RQ2: 所提出的语义感知对比学习和基于 Transformer 的词元分类器能否有效提升相关性和多样性? **RQ3:** 不同组件/参数对 CoLT 的效果有何影响?

4.2 数据集与评估指标

这里简要总结了本文所使用的数据集和评估指标。更多详细信息请参阅 [19]。本文使用了两个数据集:

表2：在 Div400 上与当前最先进方法的性能比较。P@k 和 CR@k 分别是用于评估相关性和多样性的指标。F1@k 评估同时考虑相关性和多样性时的总体性能。

方法	P@10	P@20	CR@10	CR@20	F1@10	F1@20
IMRAM [5]	79.22%	76.98%	28.93%	33.79%	42.38%	46.96%
FCA-Net [13]	79.98%	78.42%	29.91%	34.90%	43.54%	48.30%
CLIP [36]	**90.17%**	**87.92%**	35.68%	52.97%	51.10%	65.60%
MMR [38]	86.79%	84.52%	36.96%	53.88%	51.85%	65.81%
UMONS [40]	79.28%	73.37%	44.71%	63.24%	57.17%	67.93%
DESA [35]	76.50%	73.82%	39.47%	52.82%	52.07%	61.58%
Su 等人 [43]	77.55%	74.50%	35.66%	45.77%	48.85%	56.70%
VMIG [64]	82.31%	83.01%	40.37%	59.46%	54.17%	69.28%
CoLT（我们的方法）	84.45%	85.48%	**47.06%**	**64.16%**	**60.44**%	**73.30**%

Div400： Div4001 由 MediaEval Workshop [18] 收集。它包含 396 个查询和 43,418 张图像。所有查询主要与旅游景点相关，查询的平均长度为 3.7 个词。平均而言，一个查询的真实标注涵盖数据集中图像的 11.8 个语义类别。每张图像都有一段粗粒度的文本描述（例如 "大本钟"）和一段细粒度的文本描述（例如 "局部视图"）。

Div150Cred： Div150Cred[2] 是从 2014 [16]中用于多样化社交图像检索的竞赛数据集衍生而来。该数据集共有153个查询和45，375张图像。一个查询的真实标注平均涵盖该数据集中22.6个图像语义类别。

使用三个指标来评估性能，包括用于衡量语义相关性的精确率 (*P*)、用于衡量语义多样性的聚类召回率 (*CR*)，以及用于衡量 *P*和 *CR*整体平衡性能的 *F*1 得分 。具体而言，我们计算前*k* 个结果的评估指标，其中 *k* 按照 [64]设置为 10 和 20。在本文的其余部分，我们分别使用 P@k、CR@k 和 F1@k 表示前*k* 个结果的 *P*、 *CR*和 *F*1 值。较高的 P@k 表示更好的相关性，而较高的 CR@k 则意味着更丰富的语义多样性。

4.3 实现细节

CoLT 使用 PyTorch-1.10 实现。所有实验均在 4 块显存为 24GB 的 NVIDIA 3090 GPU 上进行。模型使用 Adam [23] 优化器训练，视觉特征重编码器 g的学习率为 10^{-5}，Transformer 基础令牌分类器 ϕ 的学习率为 10^{-4}。批量大小设为 32。 τ、 α、 β 和 ϵ均设为较小值：遵循 [11, 24, 69, 72]，分别设为 $\tau = 0.2$、 $\alpha = 0.01$、 $\beta = 0.02$ 和 ϵ=0.01。通过消融实验，将 X、 N 和 L分别设为 1、200 和 8。数据增强所用的概率遵循 [51]进行设置。具体而言，我们设置了 $p_q = 0.5$、 $p_v = 0.2$、$p_d = 0.2$ 和 $p_c = 0.2$。每个数据集中的所有不同语义类别（简称语义）均存储为原型。因此，Div400 数据集存储了 629 个原型，Div150Cred 数据集则存储了 725 个原型。

4.4 与最先进方法的比较（RQ1）

为展示我们的方法 CoLT 的有效性，我们将其与若干最先进方法进行比较，包括三种典型的跨模态图像检索方法：IMRAM [5], FCA-Net [13]

[1]http://multimediaeval.org/mediaeval20142014 2 /多样化图像 http://campus.pub.ro/lab7/bionescu/Div150Cred.html

表 3：与 Div150Cred 上最先进方法的性能比较。

方法	P@10	P@20	CR@10	CR@20	F1@10	F1@20
CLIP [36]	**96.02%**	**95.04%**	23.48%	35.32%	37.73%	51.51%
MMR [38]	95.37%	94.23%	23.50%	35.49%	37.71%	51.56%
UMONS [40]	77.40%	84.15%	26.69%	**40.10%**	39.69%	54.32%
VMIG [64]	90.81%	89.96%	23.83%	37.97%	37.75%	53.40%
CoLT（我们的方法）	93.41%	94.39%	**27.53%**	39.30%	**42.52**%	**55.49**%

和 CLIP [36], 两种基于后处理的多样化检索方法：MMR [38] 和 UMONS [40], , 以及三种基于学习的多样化检索方法：DESA [35],GRAPH4DIV [43]和 VMIG [64]。由于 MMR、UMONS 和 VMIG 需要特征编码器，为了公平起见，我们使用相同的特征编码器 CLIP [36] 来实现它们以及我们的方法 CoLT。对于 MMR 和 UMONS，我们采用网格搜索获得其最佳结果。总体而言，得益于强大的特征编码器，我们的结果高于原论文中的结果。例如，VMIG 在 DIV400 数据集上的 P@20、CR@20 和 F1@20 值分别从 78.27%、59.01% 和 67.29% 提升至 83.01%、59.46% 和 69.28%。Div400 和 Div150Cred 上的实验结果分别见表2和表3。最佳结果以粗体标出，次佳结果则加下划线。

从表2和表3可以看出：1）典型的跨模态图像检索方法（包括大规模预训练编码器 CLIP）在基于相关性的检索中表现出色，但无法很好地实现多样化检索。例如，尽管 CLIP 的相关性性能最佳，但其多样性得分不如其他方法。2）基于后处理的方法只能在准确率和多样性之间实现有限的权衡。例如，如表2所示，MMR 带来的多样性提升非常有限（在 Div400 上，CR@10 从 35.68% 增至 36.96%）。至于 UMONS，尽管其多样性有较大提升，但准确率得分大幅下降（在 Div400 上，P@10 从 90.17% 降至 79.28%）。因此，它们的 $F1$ 得分并不理想。3）近期提出的基于学习的方法在相关性和多样性得分之间实现了平衡。例如，VMIG 在 CR@10 和 CR@20 上优于大多数现有方法，且相关性得分优于 UMONS。然而，由于多向量投影存在不足，其相关性和多样性都受到限制。4）我们的方法 CoLT 获得了最佳多样性得分、较高的精确率，并且在 Div400 和 Div150Cred 上的整体 $F1$ 得分明显最高。具体而言，在 Div400 上，CoLT 的 F1@20 分别显著高于 CLIP 和 VMIG 7.70% 和 4.02%。这表明 CoLT 能够获得相关性高且语义多样性丰富的检索结果。此外，我们还展示了 CoLT 的一个变体，它在相关性和多样性两方面均优于 CLIP；我们将在第4.6节讨论详情。

4.5 可视化（RQ2）

为了更好地展示所提技术的优势并研究其如何提升性能，我们通过统一流形近似与投影 [32] 可视化了 DIV400 数据集测试集中的一些案例。图4展示了我们的语义感知表示与现有方法表示之间的可视化对比。

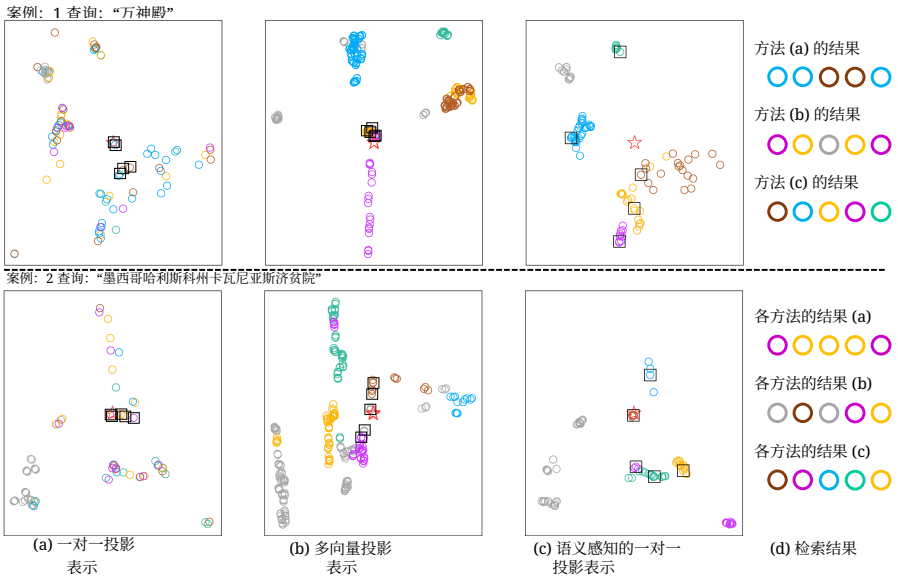

图 4：不同表示的可视化比较。(a) 由跨模态编码器 f 生成的一对一投影（OOP）表示。(b) 多向量投影（MVP）表示。(c) 由 g 重新编码的语义感知的一对一投影（ SA-OOP ）表示。检索到的图像以黑色方块标记。(d) 不同方法生成的最终检索结果。

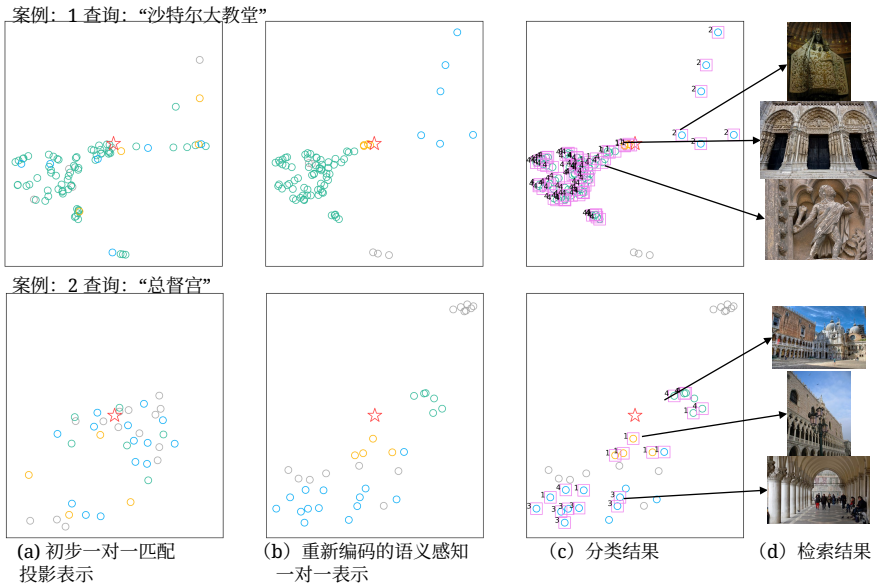

图5：CoLT 结果的可视化。(a) 由跨模态编码器 f 生成的一对一投影（OOP）表示。(b) 由重编码器 g生成的语义感知一对一投影（SA-OOP）表示。(c) TTC ϕ 的分类结果。为使图示清晰，未标注无关图像。框周围的数字是 TTC 预测的类别 ID。(d) 通过我们的后处理算法获得的最终检索结果。

从图4(a)可以看出，CLIP 提取的初步 OOP 表示能够区分一些无关图像。然而，其缺点也显而易见：（1）查询与一些具有共同语义的图像特征（蓝色和棕色第一种情况中的点）；（2）各种语义的图像相互混合。因此，这类表示不适合挖掘多样性。接下来，让我们关注多向量投影（MVP）。如图 4(b) 所示，每个图像和查询都被投影

表4：CoLT 在 Div400 上的消融实验。

ID	变体	P@20	CR@20	F1@20
1	不使用 SCL&TTC	87.92%	52.97%	65.60%
2	SCL + TTC	85.48%	64.16%	73.30%
3	不使用 SCL	84.26%	62.94%	72.06%
4	UMONS	73.37%	63.24%	67.93%
5	SCL + DBSCAN	75.94%	63.90%	69.40%
6	SCL + 排名前-k	89.06%	89.93%	67.79%
7	不使用 DA	86.34%	62.19%	72.30%
8	未修正 f	76.52%	51.25%	61.38%

表 5：SCL 中配对构建的效果。

变体	P@20	CR@20	F1@20
全部 4 对	85.48%	**64.16**%	**73.30**%
不含配对（2）	85.04%	63.78%	72.89%
不含配对（4）	**85.72%**	62.17%	72.07%

分成多个点以丰富多样性。然而，一方面，一些离群点被错误地投影到查询特征的邻域中，以表示多样性（第1种情况下的一个灰色点，而第2种情况下有两个）。另一方面，一些稀有语义的图像特征被投影到远处区域（第1种情况下的绿色点），而top-k算法无法到达这些区域。因此，如图4(d)所示，一些无关图像被选中，而一些稀有语义的图像未被检索到。最后，我们检查了我们的SCL的表示以及由TTC检索到的图像。从图4(c)中我们可以看到：(1) 与图4(a)中的典型OOP表示相比，相同语义的图像的表示聚集在一起，并且具有更强的可区分性；(2) 一些无关图像也被推离。例如，在第2种情况下，一些灰色点被推到了左下角。这证明了我们的SCL的优势和有效性。随后，采用TTC和后处理，从每个类别中对图像进行分类和选择，包括第1种情况下绿色点所代表的稀有语义类别，以形成最终结果。

我们还对TTC的分类结果进行了可视化，以进一步展示其效果。可视化结果如图5所示，由此可见：(1) TTC能够区分不同语义和无关图像。以第1个案例为例，黄色点被归为第1类，大多数绿色点被归为第4类，蓝色点则被归为第2类。无关图像也得到了正确分类。这证明了所提出的TTC的有效性。(2) TTC的分类性能还有进一步提升的空间。例如，如第2个案例所示，TTC错误地将一个绿色点归为第1类。总之，可视化结果充分展示了TTC的强大能力。

4.6 消融实验（RQ3）

我们在Div400上进行了消融实验，以展示不同模块的贡献以及我们方法中某些参数的影响。采用P@20、CR@20和F1@20作为评估指标。结果见表4至表10。

整体性能提升。如表 4 第 1 行和第 2 行所示，我们的方法显著提高了多样性得分和 $F1$ 得分，分别从 52.97% 提升至 64.16%，以及从 65.60% 提升至 73.30%，

表 6：使用不同特征编码器时的性能。

变体	P@20	CR@20	F1@20
ViT+BERT	87.75%	52.39%	65.60%
+CoLT	85.48%	64.16%	73.30%
R50+BERT	87.01%	52.10%	65.17%
+CoLT	85.62%	60.83%	72.94%
GroupViT	83.68%	52.02%	64.21%
+CoLT	82.62%	70.91%	70.91%

分别而言，相关性得分仅略有下降。这证明了我们方法的优越性。

SCL 的效果。这里我们考察所提出的 SCL 的效果。具体而言，我们首先设计了一个移除 SCL、直接应用 TTC 的变体。显然，如表 4 的第 2 行和第 3 行所示，包括相关性、多样性和 $F1$ 分数在内的所有指标均有所下降。此外，我们还设计了一个将 SCL 与 UMONS 的思路相结合、通过 DBSCAN 生成最终检索结果的变体。比较第 4 行和第 5 行的结果可知，使用 SCL 时的性能优于原始 UMONS。原因在于，SCL 能够使图像特征更具可区分性，而这类表示更适用于现有的后处理方案。

接下来，我们考察 SCL 中构造的配对的效果。如第 3.2 节所述，SCL 使用 4 种配对。其中，(1) 和 (3) 是对比学习中的常见配对，用于对齐图像和查询，而 (2) 和 (4) 则在区分不同语义的图像方面发挥着重要作用。因此，我们分别移除 (2) 和 (4)，以考察它们的影响。如表 5 所示，不含配对 (2) 和配对 (4) 时，多样性得分和 F1 分数均有所下降。相反，它们对相关性得分的影响较小。这证明了 SCL 的有效性——使表示更具可区分性，从而提升多样性。

TTC 的效果。为检验所提出的基于 Transformer 的令牌分类器的效果，我们设计了两种变体，分别用 DSCAN（表 4 第 5 行）或 top-k （表 4 第 6 行）替换 TTC，以生成检索结果。显然，这些变体的表现不如我们的方法（表 4 第 2 行）。这证明了 TTC 的优势。

逐令牌数据增强的效果。这里我们实现了一种移除逐令牌数据增强模块的变体。该变体的结果见表 4 第 7 行。显然，与我们的方法相比（表 4 第 2 行），该变体的性能较差。**为何固定跨模态特征编码器?** 在 CoLT 中，我们固定跨模态特征编码器 f，以更好地保留预训练知识。为支持这一设计，我们实现了一种对特征编码器 f进行微调的变体。实验结果见表 4 第 8 行。显然，与我们的方法相比（表 4 第 2 行），包括相关性、多样性和 $F1$ 在内的所有指标均显著下降。这可能是因为微调过多参数容易导致过拟合。

CoLT 能否支持多种特征编码器? 如上所述，CoLT 具有通用性，即可以与多种跨模态编码器配合，实现多样化图像检索。为验证这一点，我们尝试了三种不同的编码器配置，包括 ViT [8] 和 BERT [7]（由 [36], R50 [14] 开发），以及 BERT [7]（由.....实现

表7：从每个类别中选取的图像数量与性能的关系。

X	P@20	CR@20	F1@20
1	85.48%	64.16%	73.30%
2	88.09%	58.54%	70.34%
3	88.45%	57.28%	69.54%

表 8：耗时对比。我们报告了在 3090 GPU 上进行单次检索的结果。

	CLIP	MMR	UMONS	VMIG	CoLT（我们的方法）
时间（毫秒）	18.06	24.10	22.65	86.77	30.23

表 9：主要组件的时间开销对比。我们报告了在 3090 GPU 上进行一次检索的结果。f 和 g 采用并行方式进行测试。

	f	g	ϕ
时间（毫秒）	18.06	0.37	11.80

通过 [36], 以及 GroupViT [53] 中提出的编码器。实验结果如表 6 所示，从中可以看出：(1) 所有这些预训练跨模态编码器在基于相关性的检索方面表现良好，但在 CR@20 方面表现较差；(2) 应用我们的方法 CoLT 后，多样性得分显著提升，而精确率仅略有下降。因此，最终取得了更优的 F1 分数。这验证了 CoLT 能够与各种特征编码器良好协作，从而提升性能。

CoLT 能否灵活地平衡准确率与多样性? 如上所述，无需修改网络参数，我们便可以灵活地权衡检索结果的相关性与多样性。这是通过控制超参数 X 实现的。如第 3.3 节所述，后处理算法将从每个语义类别中选取 X 张图像，以构建长度为 k 的检索列表 R。因此，较小的 X 会从每个类别中选取较少的图像，但能够涵盖更多不同的类别（通过 $\lfloor k/X \rfloor$ 估算），这有利于提高检索列表 R 的多样性，但可能会损害相关性，因为模型在稀有语义类别上的分类准确率较低。相反，较大的 X，即从每个共同语义类别中选取更多图像，有利于提高准确率，但由于利用的类别较少，会限制语义多样性。我们在表 7 中展示了 X 对性能影响的实验结果。可以看出，当 $X = 1$ 时，多样性达到最佳；而当 X=3 时，准确率最高。这表明 CoLT 能够适应各种检索设置，体现了我们方法的灵活性。本文默认设置 X=1，以获得最佳多样性和 $F1$ 得分。

时间开销。我们首先将本方法的时间开销与各种 SOTA 方法进行比较。实验结果见表 8。一方面，我们的方法 CoLT 比最先进的基于学习的方法 VMIG 快 2.87× 倍。另一方面，与基于后处理的方法相比，我们的方法耗时略多。例如，CoLT 比 MMR 多耗时 6.23 毫秒。这证明了我们方法的高效性。

接下来，我们进一步考察 CoLT 各主要模块的耗时：固定特征编码器 f、视觉特征重编码器 g，

表10：参数 L 对TTC的影响。

L	P@20	CR@20	F1@20	时间（毫秒）
6	85.82%	61.34%	71.54%	8.93
8	85.48%	64.16%	73.30%	11.80
10	85.29%	61.67%	71.58%	18.56

以及 TTC ϕ。实验结果见表 9。可以看出，g 和 ϕ 的耗时远少于特征编码器 f。原因在于，g 是一个简单的多层感知机，而 ϕ 由多个可并行运行的 Transformer 编码器层组成。值得一提的是，在实际应用中，f 和 g 生成的图像特征可以离线缓存。因此，主要耗时来自 TTC ϕ，而这一耗时非常有限（根据表 9，为 11.80 毫秒）。这也验证了我们方法的高效性。

参数 L的影响。本文研究 Transformer 层数 L的影响。一方面，由于训练数据有限，较大的 L 更有可能导致过拟合。另一方面，较小的 L 无法充分发挥 TTC 的潜力。因此，我们通过网格搜索确定 L的取值。如表 10 所示，当 L=8时，性能最佳。

参数 N的影响。本文考察输入基于 Transformer 的词元分类器 ϕ 的图像数量 N 对性能的影响。直观而言，较大的 N 会包含语义更多的图像。另一方面，较大的 N 会引入更多无关图像，从而可能使词元分类更加困难。相反，较小的 N 包含的无关图像较少，但语义也较少。因此，较小的 N 和较大的 N 都不适用于 TTC。我们在两个数据集上通过网格搜索确定 N。根据实验结果，对于 DIV400 数据集，我们将 N 设为 200，因为 $N = 150$ 和 $N = 250$ 的 F1@20 分数分别为 72.10% 和 72.13%，低于 $N = 200$ 对应的 73.30%。而在 Div150Cred 数据集上，当 $N = 200$ 时性能最佳（$F1$ 为 55.49%），其次是 250（$F1$ 为 55.32%）。因此，我们将该超参数设为 200。

5 结论

本文针对基于关键词的多样化图像检索问题，提出了一种名为语义感知分类 Transformer（CoLT）的新方法来完成这一任务。与现有方法不同，CoLT 首先通过语义感知对比学习提取具有高度代表性的图像和查询特征，然后利用基于 Transformer 的词元分类器融合这些特征，并将其归入相应的类别。最后，应用一种后处理算法，从各个类别中灵活地选择图像，以形成检索结果。CoLT 具有四个方面的优势：高语义相关性、高语义多样性、通用性以及易于使用和易于控制。在两个数据集 Div400 和 Div150Cred 上进行的大量实验验证了我们方法的优越性。

致谢

赵敏怡的研究部分受到了 2022 年腾讯犀牛鸟研究精英培养计划的资助。周水庚受到了中国国家重点研发计划资助，项目编号：2021YFC3340302。

参考文献

[1] Yuan Bo 和 Xinbo Gao。2019。基于多样化文本特征的图像检索。《神经计算》357（2019），116–124。[2] Deng Cai、Yan Wang、Huayang Li、Wai Lam 和 Lemao Liu。2021。利用单语翻译记忆的神经机器翻译。载于 ACL。ACL，7307–7318。[3] Da Cao、Yawen Zeng、Xiaochi Wei、Liqiang Nie、Richang Hong 和 Zeng Qin。2020。通过联合建模排序和定位进行对抗式视频片段检索。载于 ACM多媒体会议论文集。ACM，898–906。[4] Hui Chen、Guiguang Ding、Zijia Lin、Sicheng Zhao 和 Jungong Han。2019。具有语义一致性的跨模态图文检索。载于 ACM多媒体会议论文集。ACM，1749–1757。[5] Hui Chen、Guiguang Ding、Xudong Liu、Zijia Lin、Ji Liu 和 Jungong Han。2020。IMRAM：基于循环注意力记忆的迭代匹配，用于跨模态图文检索。载于 CVPR 论文集。电气电子工程师学会，12652–12660。[6] Laming Chen、Guoxin Zhang 和 Eric Zhou。2018。用于行列式点过程的快速贪心最大后验推断，以提升推荐多样性。《神经信息处理系统进展》31（2018）。[7] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019。BERT：用于语言理解的深度双向变换器预训练。载于《2019 年计算语言学协会北美分会：人类语言技术会议论文集》，第 1 卷（长短篇论文）。4171–4186。[8] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xi- aohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。2020。一幅图像胜过 16×16 个词：用于大规模图像识别的变换器。arXiv 预印本 arXiv:2010.11929（2020）。[9] Martin Ester、Hans-Peter Kriegel、Jörg Sander、Xiaowei Xu 等。1996。一种用于在含噪大型空间数据库中发现簇的基于密度的算法。载于 KDD，第 96 卷。226–231。[10] Iftah Gamzu、Marina Haikin 和 Nissim Halabi。2020。语音购物中的空查询改写。载于信息检索研究特别兴趣小组年会论文集。ACM，1369–1378。[11] Yixiao Ge、Feng Zhu、Dapeng Chen、Rui Zhao 等。2020。利用混合记忆进行自步对比学习，以实现领域自适应目标重识别。《神经信息处理系统进展》33（2020），11309–11321。[12] Yun Gu、Khushi Vyas、Mali Shen、Jie Yang 和 Guang-Zhong Yang。2021。用于内镜图像检索的基于深度图的多模态特征嵌入。《IEEE 神经网络与学习系统汇刊》33，2（2021），481–492。[13] Ning Han、Jingjing Chen、Guangyi Xiao、Zhang Hao、Yawen Zeng 和 Hao Chen。2021。用于文本视频检索的细粒度跨模态对齐网络。载于 ACM多媒体会议论文集。ACM，3826–3834。[14] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。2016。用于图像识别的深度残差学习。载于 CVPR 论文集。电气电子工程师学会，770–778。[15] Bogdan Ionescu、Alexandru-Lucian Gînsca、Maia Zaharieva、Bogdan Boteanu、Mihai Lupu 和 Henning Müller。2016。在 MediaEval 2016 中检索多样化社交图像：挑战、数据集与评估。载于 MediaEval 2016 研讨会论文集工作札记，第 1739 卷。CEUR-WS.org。[16] Bogdan Ionescu、Adrian Popescu、Mihai Lupu、Alexandru Lucian Gînscă、Bogdan Boteanu 和 Henning Müller。2015。Div150cred：社交图像检索结果多样化及用户标签可信度数据集。载于第六届 ACM多媒体系统会议论文集。207–212。[17] Bogdan Ionescu、Adrian Popescu、Anca-Livia Radu 和 Henning Müller。2016。社交图像检索中的结果多样化：基准测试框架。《多媒体工具与应用》75，2（2016），1301–1331。[18] Bogdan Ionescu、Anca-Livia Radu、María Menéndez、Henning Müller、Adrian Popescu 和 Loni Babak。2014。Div400：社交图像检索结果多样化数据集。载于 2014 年多媒体系统会议。ACM，29–34。[19] Bogdan Ionescu、Maia Rohm、Bogdan Boteanu、Alexandru-Lucian Gînsca、Mihai Lupu 和 Henning Müller。2021。社交媒体图像检索多样化技术的基准测试。《IEEE多媒体汇刊》23（2021），677–691。[20] Bogdan Ionescu、Maia Rohm、Bogdan Boteanu、Alexandru Lucian Gînscă、Mihai Lupu 和 Henning Müller。2020。社交媒体图像检索多样化技术的基准测试。《IEEE多媒体汇刊》23（2020），677–691。[21] Zhong Ji、Yuxin Sun、Yunlong Yu、Yanwei Pang 和 Jungong Han。2020。用于跨模态零样本哈希的属性引导网络。《IEEE 神经网络与学习系统汇刊》31，1（2020），321–330。[22] Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。2021。利用噪声文本监督扩展视觉及视觉语言表征学习。载于国际机器学习会议。PMLR，4904–4916。[23] Diederik P Kingma 和 Jimmy Ba。2014。Adam：一种随机优化方法。arXiv 预印本 arXiv:1412.6980（2014）。[24] Frank Klinker。2011。指数移动平均与移动指数平均。《数学学期报告》58，1（2011），97–107。[25] Saar Kuzi、Abhishek Narwekar、Anusri Pampari 和 ChengXiang Zhai。2019。帮我搜索：利用用户-系统协作构建查询，以提高困难查询的准确率。载于信息检索研究特别兴趣小组年会论文集。ACM，1221–1224。[26] V. Quoc Le 和 Tomás Mikolov。2014。《句子和文档的分布式表示》。载于《国际机器学习大会论文集》。JMLR.org，1188–1196。[27] 廖栋梁、Jin Xu、Gongfu Li、Huang Weijie、Liu Weiqing 和 Li Jing。2019。《通过时间过程与内容特征的深度融合预测在线文章的流行度》。载于《AAAI论文集》。AAAI，200–207。[28] Haoliang Liu、Tan Yu 和 Ping Li。2021。《扩充与收缩：通过丰富和减少交互实现快速文本图像检索》。载于《2021年自然语言处理实证方法会议论文集》。9796–9809。[29] Shenglan Liu、Muxin Sun、Lin Feng、Hong Qiao、Shuyuan Chen 和 Yang Liu。2021。《用于多特征图像检索的社交邻域图与多图融合排序》。IEEE神经网络与学习系统汇刊 32, 3 (2021)，1389–1399。[30] Zheyuan Liu、Cristian Rodriguez-Opazo、Damien Teney 和 Stephen Gould。2021。《使用预训练视觉语言模型检索真实场景图像》。载于《IEEE/CVF国际计算机视觉会议论文集》。2125–2134。[31] Minnan Luo、Xiaojun Chang、Zhihui Li、Liqiang Nie、Alexander G. Hauptmann 和 Qinghua Zheng。2017。《从简单到复杂的跨模态学习排序》。计算机视觉与图像理解 (2017)，67–77。[32] Leland McInnes、John Healy 和 James Melville。2018。《UMAP：用于降维的统一流形近似与投影》。arXiv预印本 arXiv:1802.03426 (2018)。[33] Liang Peng、Yi Bin、Xiyao Fu、Jie Zhou、Yang Yang 和 Heng Tao Shen。2017。《CFM@MediaEval 2017：通过重排序与层次聚类完成多样化社交图像检索任务》。载于《MediaEval 2017研讨会工作笔记论文集》，第1984卷。[34] Yuxin Peng、Xin Huang 和 Yunzhen Zhao。2018。《跨媒体检索概述：概念、方法、基准与挑战》。IEEE视频技术电路与系统汇刊 28, 9 (2018)，2372–2385。[35] Xubo Qin、Zhicheng Dou 和 Ji-Rong Wen。2020。《使用自注意力网络实现搜索结果多样化》。载于《信息与知识管理国际会议论文集》。ACM，1265–1274。[36] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021。《通过自然语言监督学习可迁移视觉模型》。载于《国际机器学习会议》。PMLR，8748–8763。[37] Nikhil Rasiwasia、José Costa Pereira、Emanuele Coviello、Gabriel Doyle、Gert R. G. Lanckriet、Roger Levy 和 Nuno Vasconcelos。2010。《一种跨模态多媒体检索新方法》。载于《ACM多媒体会议论文集》。ACM，251–260。[38] Jean-Michel Renders 和 Gabriela Csurka。2017。《NLE@MediaEval'17：结合跨媒体相似度与嵌入来检索多样化社交图像》。载于《MediaEval 2017研讨会工作笔记论文集》，第1984卷。[39] Mustafa Ilker Sarac 和 Pinar Duygulu。2014。《Bilkent-RETINA参加MediaEval 2014多样化社交图像检索任务》。载于《MediaEval 2014研讨会工作笔记论文集》，第1263卷。[40] Omar Seddati、Nada Ben-Lhachemi、Stéphane Dupont 和 Saïd Mahmoudi。2017。《UMONS @ MediaEval 2017：多样化社交图像检索》。载于《MediaEval 2017研讨会工作笔记论文集》，第1984卷。[41] Amit Singhal 等。2001。《现代信息检索：简要概述》。IEEE数据工程通报 24, 4 (2001)，35–43。[42] Yale Song 和 Mohammad Soleymani。2019。《用于跨模态检索的多义视觉语义嵌入》。载于《CVPR论文集》。IEEE，1979–1988。[43] Zhan Su、Zhicheng Dou、Yutao Zhu、Xubo Qin 和 Ji-Rong Wen。2021。《为搜索结果多样化建模意图图》。载于《信息检索研究特别兴趣小组年会论文集》。ACM，736–746。[44] Hanghang Tong、Jingrui He、Mingjing Li、Changshui Zhang 和 Wei-Ying Ma。2005。《基于图的多模态学习》。载于《ACM多媒体会议论文集》。ACM，862–871。[45] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、N. Aidan Gomez、Lukasz Kaiser 和 Illia Polosukhin。2017。《Attention Is All You Need》。载于《神经信息处理系统大会论文集》。5998–6008。[46] Bokun Wang、Yang Yang、Xing Xu、Alan Hanjalic 和 Heng Tao Shen。2017。《对抗式跨模态检索》。载于《ACM多媒体会议论文集》。ACM，154–162。[47] Huanwen Wang、Yawen Zeng、Jianguo Chen、Zhouting Zhao 和 Hao Chen。2022。《用于基于会话的推荐的时空图神经网络》。《专家系统与应用》(2022)。[48] Kaiye Wang、Qiyue Yin、Wei Wang、Shu Wu 和 Liang Wang。2016。《跨模态检索综合综述》。arXiv预印本 arXiv:1607.06215 (2016)。[49] Liwei Wang、Yin Li 和 Svetlana Lazebnik。2016。《学习保持深层结构的图像文本嵌入》。载于《CVPR论文集》。IEEE，5005–5013。[50] Zihao Wang、Xihui Liu、Hongsheng Li、Lu Sheng、Junjie Yan、Xiaogang Wang 和 Jing Shao。2019。《跨模态自适应消息传递：用于文本图像检索》。载于《IEEE/CVF国际计算机视觉会议论文集》。5764–5773。[51] Jason Wei 和 Kai Zou。2019。《简易数据增强：提升文本分类任务性能的数据增强技术》。arXiv预印本 arXiv:1901.11196

(2019). [52] Jiaxin Wu 和 Chong-Wah Ngo。2020。用于即席视频搜索的可解释嵌入。载于*ACM*多媒体会议论文集。ACM，3357–3366。[53] Jiarui Xu、Shalini De Mello、Sifei Liu、Wonmin Byeon、Thomas Breuel、Jan Kautz 和 Xiaolong Wang。2022。GroupViT：语义分割从文本监督中涌现。载于*IEEE/CVF*计算机视觉与模式识别会议论文集。18134–18144。[54] Jin Xu、Bo Tang、Haibo He 和 Hong Man。2017。基于相关性与冗余性准则的半监督特征选择。*IEEE*神经网络与学习系统汇刊28，9（2017），1974–1984。[55] Ruicong Xu、Li Niu、Jianfu Zhang 和 Liqing Zhang。2020。一种用于活动图像到视频检索的基于提议的方法。载于*AAAI*论文集。AAAI出版社，12524–12531。[56] Caixia Yan、Qinghua Zheng、Xiaojun Chang、Minnan Luo、Chung-Hsing Yeh 和 Alexander G. Hauptmann。2020。用于零样本目标检测的保语义图传播。*IEEE*图像处理汇刊（2020），8163–8176。[57] Xiaojun Yang、Lunjia Liao、Qin Yang、Bo Sun 和 Jianxiang Xi。2021。间歇性交互多智能体系统的有限能量输出编队。富兰克林研究所期刊（2021），6462–6489。https://doi.org/10.1016/j.jfranklin.2021 06 009. . [58] Yi Yang、Feiping Nie、Dong Xu、Jiebo Luo、Yueting Zhuang 和 Yunhe Pan。2011。一种基于半监督排序和相关性反馈的多媒体检索框架。*IEEE*模式分析与机器智能汇刊34，4（2011），723–742。[59] Peter Young、Alice Lai、Micah Hodosh 和 Julia Hockenmaier。2014。从图像描述到视觉指称：用于事件描述语义推理的新型相似度指标。计算语言学协会汇刊2（2014），67–78。[60] Tan Yu、Hongliang Fei 和 Ping Li。2022。用于快速、可扩展文本-图像检索的 U-BERT。载于*2022*年*ACM SIGIR*信息检索理论国际会议论文集。193–203。[61] Maia Zaharieva 和 Patrick Schwab。2014。用于检索多样化社交图像的统一框架。载于*MediaEval 2014*研讨会工作笔记论文集，第1263卷。[62] Yawen Zeng、Da Cao、Shaofei Lu、Hanling Zhang、Jiao Xu 和 Qin Zheng。2022。片段至关重要：通过对抗式基于语言的视频片段检索。学习。*ACM Trans. Multim. Comput. Commun. Appl.* 18 (2022), 56:1–56:21. [63] Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. 2021. 用于跨模态视频片段检索的多模态关系图。在*CVPR* 会议论文集。电气电子工程师学会，2215–2224。[64] Yawen Zeng, 王怡如，廖栋梁, Gongfu Li, Weijie Huang, Jin Xu, Da Cao, and Hong Man. 2022. 基于变分多实例图的基于关键词的多样化图像检索。*IEEE*神经网络与学习系统汇刊(2022)。[65] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. 2012. 用于跨媒体检索的基于近邻的有效异构相似性度量。在多媒体建模进展会议论文集。Springer, 312–322。[66] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. 2014. 通过形式逻辑对信息检索建模：综述。*IEEE*视频技术电路与系统汇刊24, 6 (2014), 965–978。[67] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup：超越经验风险最小化。*arXiv* 预印本 *arXiv:1710.09412* (2017)。[68] Lu Zhang, Yang Wang, Jiaogen Zhou, Chenbo Zhang, Yinglu Zhang, Jihong Guan, Yatao Bian, and 周水庚. 2022. 分层小样本目标检测：问题、基准与方法。在第*30*届*ACM*多媒体国际会议论文集。2002–2011。[69] 赵敏怡，Yi Xu, and 周水庚. 2021. 用于视频压缩伪影消除的递归融合与可变形时空注意力。在第*29*届*ACM*多媒体国际会议论文集。5646–5654。[70] Wanqing Zhao, Ziyu Guan, Hangzai Luo, Jinye Peng, and Jianping Fan. 2017. 面向对象的图像检索的深度多实例哈希。在*IJCAI* 会议论文集。3504–3510。[71] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. 深度监督跨模态检索。在*CVPR* 会议论文集。电气电子工程师学会，10394–10403。[72] Linchao Zhu and Yi Yang. 2020. 通过区域自注意力增强情景记忆，实现长尾视觉识别。在*IEEE/CVF*计算机视觉与模式识别会议论文集。4344–4353。[73] Yueting Zhuang, Yi Yang, and Fei Wu. 2008. 挖掘异构多媒体数据的语义关联以实现跨媒体检索。*IEEE*多媒体汇刊10, 2 (2008), 221–229。