arXiv:2303.00448v2 [cs.CV] 3 Apr 2023

结合共享知识优化的风格Transformer网络，用于图文检索

Wenrui Li、Zhengyu Ma、Jinqiao Shi 和 Xiaopeng Fan，*IEEE* 高级会员

摘要——**由于图文检索能够关联不同模态，且具有卓越的研究价值和广泛的现实应用，因此受到了广泛关注。然而，现有大多数方法未能充分考虑多模态中的高层语义关系（“风格嵌入”）和常识。为此，我们提出了一种新颖的结合公共知识优化的风格变换器网络（CKSTN），用于图文检索。其主要模块是公共知识适配器（CKA），由风格嵌入提取器（SEE）和公共知识优化（CKO）模块组成。具体而言，SEE 使用顺序更新策略，有效连接 SEE 不同阶段的特征。引入 CKO 模块，以动态捕获来自不同模态的常识潜在概念。此外，为了获取泛化的时序共识，我们提出了一种顺序更新策略，将 SEE 不同层的特征与先前的公共特征单元进行有效整合。在 MSCOCO 和 Flickr30K 数据集上的图文检索任务中，CKSTN 展现出优于当前先进方法的性能。此外，CKSTN 基于轻量级 Transformer 构建，由于其性能更优且参数更少，因此更适用于真实场景中的应用，也更加便捷实用。**

索引词—**图文检索，Transformer**

I. 引言

图Mage-text 检索 [1]–[3] 是指从包含两种模态的数据库中检索相关图像和文本的过程。这是一个融合了计算机视觉 [4] 和自然语言处理 [5], [6]的跨学科领域，已得到广泛研究。以往的大多数方法更关注细粒度对齐 [7]–[9]，而非高层语义关系。这些方法不仅没有充分利用全局信息，而且还存在参数数量庞大的问题。此外，目前的大多数方法 [10]–[16] 忽略了不同模态之间的常识，而常识对于理解语境中的词义以及网络的特征表示能力至关重要。

我们提出了一种新颖的结合公共知识优化的风格变换器网络（CKSTN），以解决上述问题。为提升模型在实际应用中的实用性，我们在 Transformer 中采用轻量级前馈网络替代注意力层，以减少模型参数和计算开销。我们提出了风格嵌入提取器（SEE），通过多层感知机（MLP）逐行提取特征图的“风格嵌入”，从而更好地利用高层语义关系。我们采用拼接和打乱操作 [17]，确保“风格嵌入”包含更多全局信息，并在特征图之间共享不同的行特征。风格嵌入提取器采用循环架构，通过有效融合低层和高层语义特征来提取多层级特征。“风格嵌入”不仅包含高层语义概念，还融入了低层特征中的像素级对齐关系。对于高层语义信息，我们还利用跨模态共享知识，这也能有效提升特征表示。最后，我们提出公共知识优化（CKO）模块，根据不同模态的“风格嵌入”动态更新公共知识中的潜在概念，从而促进特征调整模块的工作。我们提出了 CKO 中的顺序更新策略，以动态调整公共知识，并基于不同项目的风格嵌入优化不同模态之间的特征表示。本文的主要贡献总结如下：

1) 我们提出了一种新颖的结合共享知识优化的风格Transformer网络（CKSTN），该网络利用基于“风格嵌入”动态调整的常识，来优化两种模态的特征表示。

2) SEE通过拼接和随机打乱操作，在特征图之间共享不同的行特征，并利用多层感知机获得紧凑且高效的表示。我们提出了CKA，以捕获不同模态之间常识的潜在概念。

3) 我们引入顺序更新策略，将SEE中不同阶段的特征与先前的公共特征单元相结合，从而有助于获得泛化的时序共识。

II. 方法

A. 架构

图像和文本的特征提取器RAVLN 使用 Faster-RCNN [18]提取视觉特征，并使用预训练的 BERT [5]分别提取文本特征 [19] 。与几何边界框相关的信息被整合到

本研究部分得到了中国国家重点研发计划（2021YFF0900500）和国家自然科学基金（NSFC）项目（U22B2035和62206141）的资助。
（通讯作者：Jinqiao Shi。）Wenrui Li和Xiaopeng Fan就职于中国哈尔滨150001哈尔滨工业大学计算机科学与技术系（电子邮箱：liwr@stu.hit.edu.cn；fxp@hit.edu.cn）。Zhengyu Ma就职于中国深圳518055鹏城实验室（电子邮箱：mazhy@pcl.ac.cn）。Jinqiao Shi就职于中国北京100876北京邮电大学网络空间安全学院（电子邮箱：shijinqiao@bupt.edu.cn）。

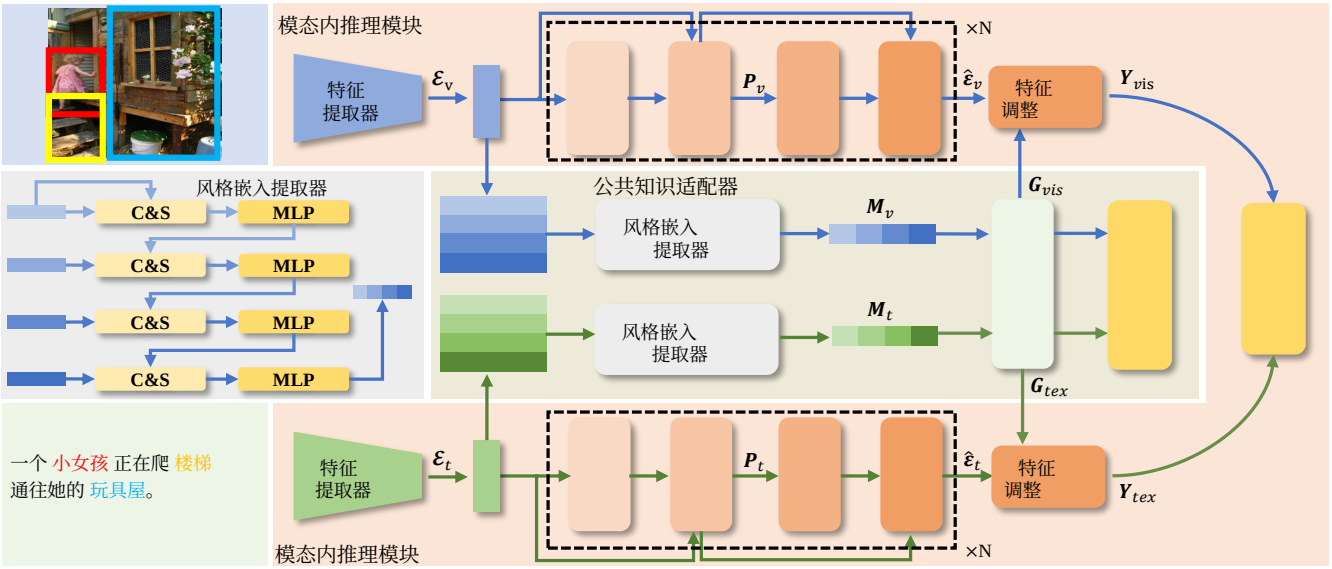


图1：CKSTN 的架构。“C&S”表示拼接和随机打乱操作。SEE 能够有效提取代表全局信息的高层语义关系。CKA 能够动态捕获不同模态的共同潜在概念。最后，利用学得的常识进一步辅助形成物品特征表示。

利用全连接神经网络从图像中提取的视觉特征中。在预训练的 BERT 中，标准 Transformer 架构可以使用自注意力机制对句子中词语之间的记忆连接进行编码。给定输入的视觉和文本特征 $\boldsymbol{X}_c \in \mathbb{R}^{n\times d_o}$，$n$ 和 d_o分别表示$\boldsymbol{X}_c$的序列长度和维度，以及 $c \in (\boldsymbol{v}, \boldsymbol{t})$。视觉和文本特征提取器的输出可表示为：$\boldsymbol{\mathcal{E}}_v = E_{vis}(\boldsymbol{X}_v)$ 和 $\boldsymbol{\mathcal{E}}_t = E_{tex}(\boldsymbol{X}_t)$，其中$\boldsymbol{\mathcal{E}}_v \in \mathbb{R}^{n\times d_{in}}$ 和 $\boldsymbol{\mathcal{E}}_t \in \mathbb{R}^{n\times d_{in}}$。

模态内推理模块。在特征提取器之后，使用一系列轻量级 Transformer 进一步探索模态内语义关系。视觉流水线中的一个完整轻量级 Transformer 层可表示为：

$$\begin{aligned}\boldsymbol{\mathcal{E}}_v^a &= \frac{(\boldsymbol{\mathcal{E}}_v\boldsymbol{T}_Q \cdot (\boldsymbol{\mathcal{E}}_v\boldsymbol{T}_K)^T)}{\sqrt{d_e}}\boldsymbol{\mathcal{E}}_v\boldsymbol{T}_V,\\ \boldsymbol{P}_v &= \text{Add\&Norm}(\boldsymbol{\mathcal{E}}_v^a, \boldsymbol{\mathcal{E}}_v),\\ \hat{\boldsymbol{\mathcal{E}}_v} &= \text{Add\&Norm}(\text{FFN}(\boldsymbol{P}_v), \boldsymbol{P}_v),\end{aligned} \tag{1}$$

其中，$\boldsymbol{T}_{(\cdot)} \in \mathbb{R}^{d_{in}\times d_e}$ 表示查询、键和值的可学习线性投影，d_e 表示公共空间的维度，Add&Norm(·)包含层归一化操作和残差连接，FFN(·)表示轻量级前馈网络，由两个采用 GLUE激活函数的非线性层组成。

常识适配器。合理利用高层语义信息能够优化特征表示。本文提出 SEE 来编码高层特征（“风格嵌入”）之间的语义关系。SEE 的第 i 层输出定义如下：

$$\boldsymbol{R}_i = \text{C\&S}(\text{CLIP}_i(\boldsymbol{\mathcal{E}}_v), \boldsymbol{M}_{i-1}), \boldsymbol{M}_i = \text{MLP}(\boldsymbol{R}_i), \tag{2}$$

其中，$\text{CLIP}_i(\cdot): \mathbb{R}^{n\times d_e} \rightarrow \mathbb{R}^{n\times(d_e/m)}$ 表示用于裁剪并返回特征图第 i行的裁剪函数，并且，m 是 SEE 中的层数。$\text{C\&S}(\cdot,\cdot): (\mathbb{R}^{n\times d_m}, \mathbb{R}^{n\times d_m}) \rightarrow \mathbb{R}^{n\times(2*d_m)}$，$d_m = d_e/m$表示拼接和随机打乱操作，该操作能够包含更多全局信息，并在特征图与多层感知机之间共享不同的行特征。$\text{MLP}(\cdot): \mathbb{R}^{n\times(2*d_m)} \rightarrow \mathbb{R}^{n\times d_m}$ 是一个由 3 层线性层组成的多层感知机。

公共知识优化（CKO）由一组公共特征单元 $\boldsymbol{S} = \{\boldsymbol{S}_1, \boldsymbol{S}_2, ..., \boldsymbol{S}_k\}$ 构成，用于表示两种模态之间的常识。CKO 的示意图如图 2 所示。对于来自 SEE 最后一层的视觉和文本输入 $\boldsymbol{M}_c \in \mathbb{R}^{n\times d_m}$ 和 $c \in (\boldsymbol{v}, \boldsymbol{t})$，CKO 可以根据输入向量计算各公共特征单元的重要性，并将它们有效地组合起来，可表示为：

$$\boldsymbol{S}_o = \sum_{i=1}^{k} \text{Softmax}(\boldsymbol{S}_i\boldsymbol{M}_c)\boldsymbol{M}_c, \tag{3}$$

其中 $\text{Softmax}(x) = 1/(1+e^{-x})$。本文提出记忆门，以选择性地保留或排除融合特征中的信息。视觉流水线中的记忆门定义如下：

$$\boldsymbol{G}_{vis} = \text{ReLU}(\boldsymbol{W}_o\boldsymbol{S}_{ovis} + b_o), \tag{4}$$

其中 $\boldsymbol{W}_o \in \mathbb{R}^{n\times n}$ 是可学习权重度量，b_o 是偏置项。经过记忆门后，风格嵌入与常识进行合理整合。

为了提取有效的常识，我们提出一种顺序更新策略，将 SEE 不同阶段的特征与先前的公共特征单元 $\boldsymbol{S}_{t-1} \in \mathbb{R}^{n\times d_m}$连接起来。公共特征单元的更新方式如下：

$$\boldsymbol{S}_t = (\boldsymbol{G}_{vis}\boldsymbol{G}_{tex})\boldsymbol{S}_{t-1} + (1 - \boldsymbol{G}_{vis}\boldsymbol{G}_{tex})\boldsymbol{S}_{ovis}\boldsymbol{S}_{otex}. \tag{5}$$

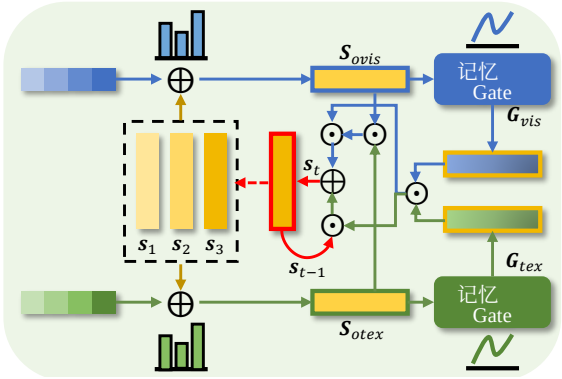


图 2：公共知识优化的架构。中心的红色线条表示 s_t 的顺序更新策略。更新细节可参见公式（3-5）。提取的s_t 与视觉和文本风格嵌入融合，以更新和优化下一个视觉和文本特征输入。

最后，轻量级 Transformer 推理得到的特征将根据风格嵌入进行全局调整。视觉流水线中的特征调整模块定义如下：

$$\boldsymbol{Y}_{vis} = \text{Softmax}(\boldsymbol{W}_g \boldsymbol{G}_{vis})\hat{\boldsymbol{\mathcal{E}}}_v, \tag{6}$$

其中，$\boldsymbol{W}_g \in \mathbb{R}^{n\times n}$ 是用于微调记忆门输出的可学习权重度量。

B. 训练策略

我们在单张 NVIDIA V100s GPU 上训练模型 30 个 epoch，选择 Adam 作为优化器。我们按照 [4], 中提出的方法提取自底向上视觉特征，并将两种模态的特征映射到 1024 维的嵌入空间，再使用轻量级 Transformer 和 SEE 模块探索它们之间的关系。为避免过拟合，我们采用预热策略，在前 10 个 epoch 中将学习率设为 $1e^{-5}$ 至 $1e^{-4}$，随后线性衰减至 $1e^{-5}$。我们将风格嵌入层的层数设为 4；对于三元组排序损失，困难负样本间隔设为 0.2。为促使模型学习更准确的特征表示，并增大嵌入空间中不匹配对之间的距离，我们根据视觉特征和文本特征计算对比损失和匹配损失。

对比损失。一个批次包含 N 对图像和句子，图像与句子之间的批内相似度可定义为：

$$\mathcal{L}_{i2t} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp((\boldsymbol{G}_{vis})_i^T(\boldsymbol{G}_{tex})_i/\tau)}{\sum_N^{j=1} exp((\boldsymbol{G}_{vis})_i^T(\boldsymbol{G}_{tex})_i/\tau)}. \tag{7}$$

句子到图像的相似度定义如下：

$$\mathcal{L}_{t2i} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp((\boldsymbol{G}_{tex})_i^T(\boldsymbol{G}_{vis})_i/\tau)}{\sum_N^{j=1} exp((\boldsymbol{G}_{tex})_i^T(\boldsymbol{G}_{vis})_j/\tau)}. \tag{8}$$

对比损失的公式为 $\mathcal{L}_{con} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i}$

匹配损失。我们采用基于 hinge 的三元组排序损失 [5] 来计算匹配损失，该损失着重于困难负样本 l' 和 k'。区域-词语相似度矩阵$\boldsymbol{C} \in \mathbb{R}^{dk\times dl}$ 通过计算 $\boldsymbol{Y}vis$ 与 $\boldsymbol{Y}tex$之间的余弦相似度得到，其中 dk 表示第 k 张图像的区域特征索引集合，dl 表示第 l 个句子中句子词语的索引。匹配损失定义如下：

$$\mathcal{L}_{kl} = \max_{l'}[\gamma + \boldsymbol{C}_{kl'} - \boldsymbol{C}_{kl}]_+ + \max_{k'}[\gamma + \boldsymbol{C}_{k'l} - \boldsymbol{C}_{kl}]_+, \tag{9}$$

其中，边际参数是负样本对与真正匹配的视觉-文本嵌入之间的最小距离，$[s]_+ \equiv \max(s, 0)$。

总损失的公式为 $\mathrm{L}_{all} = \mathrm{L}_{con} + \mathrm{L}_{kl}$。

III. 实验

A. 结果比较

我们的模型与 Flick30K [20] 和 MSCOCO [21] 数据集上的最先进基线模型进行了比较。为统一评估标准，我们仅评估基础模型，不使用集成。我们采用排名 k 处的召回率（R@K）作为评估指标，以衡量各比较方法的检索性能。TERAN [28],ViLBERT [29], 和 LTAN [19] 使用 Transformer 获取注意力图。DIME [27] 使用“校正身份单元”和“全局-局部引导单元”提取通用知识，以促进不同模态之间的交互。然而，DIME 提取的通用知识无法动态调整。我们的模型使用 Transformer 提取不同模态之间的注意力关系，并提出 CKO 模块，以动态优化特征形成。

表 1 显示，我们的 CKSTN 性能优异，超过了大多数基线模型。在 MSCOCO 上，CKSTN 始终优于最先进的方法 DERN：在图像检索和句子检索的 R@1 上，相对提升分别为 11.8% 和 5.9%。此外，我们模型的 Rsum 提高了 22.1。在 Flickr30K 上，我们的方法优于最先进的模型 LTAN：在图像检索和句子检索的 R@1 上，分别提升了 3.1% 和 2.4%。表 2 将我们的模型与基于 Transformer 的模型进行了比较，结果表明，我们的模型以更少的参数实现了更好的性能。ViLBERT 拥有 275M 个参数，而我们的模型仅有 175M 个参数。考虑到现实世界应用对实用性和效率的要求，我们的方法更适合部署，也更加便捷。

B. 消融实验

CKO、SEE 和匹配损失的有效性。为了展示我们提出的 CKO 和 SEE 对图文检索的有效性，我们将未使用 CKO 和 SEE 的模型与图3左侧的模型进行比较，并将其分别记为“未使用公共知识优化”和在 SEE 中层数为 0 的“标准”模型。对于 CKO，我们发现，在 SEE 为 4 层时，“标准”模型在图像检索和句子检索的 R@1 上分别提升了 2.5% 和 3.1%。CKO 能够捕捉不同模态间共有的潜在概念，从而更好地帮助形成特征

表 I：两个基准数据集上的实验结果。

模型	MS-COCO 数据集							Flickr30K 数据集							
	图像检索			句子检索			*Rsum*	图像检索			句子检索			*Rsum*	场所
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10		
DREN [22]	61.8	88.2	96.0	76.6	95.7	98.3	516.6	53.6	81.5	87.0	74.3	93.5	96.2	483.8	TCSVT 22
SSAMT [23]	62.7	89.6	95.3	78.2	95.6	98.0	519.4	54.8	81.5	88.0	75.4	92.6	96.4	488.7	ICMR 22
CGMN [24]	63.8	90.7	95.7	76.8	95.4	98.3	520.7	59.9	85.1	90.6	77.9	93.8	96.8	504.1	TOMM 22
GraDual [25]	63.7	90.8	95.6	76.8	95.9	98.3	521.1	57.7	83.1	90.5	76.1	93.7	96.7	497.8	WACV 22
MACM [26]	63.1	90.5	95.6	77.2	96.7	98.5	521.6	58.8	84.1	90.0	77.6	94.5	97.1	502.1	IJCNN 22
DIME [27]	63.0	90.5	96.2	77.9	95.9	98.3	521.8	60.1	85.5	91.8	77.4	95.0	97.4	507.2	SIGIR 21
TERAN [28]	65.0	91.2	96.4	77.7	95.9	98.6	524.8	59.5	84.9	90.6	75.8	93.2	96.7	500.7	TOMM 21
LTAN [19]	67.5	92.2	97.4	79.6	97.9	**99.7**	534.3	60.5	85.1	91.1	76.5	93.5	96.5	503.2	ICASSP 22
CKSTN	**69.1**	**92.3**	**98.8**	**81.1**	**98.1**	99.3	**538.7**	**62.3**	**86.8**	**91.3**	**78.3**	**92.9**	**96.9**	**508.5**	

表 II：Flickr30K 数据集上的方法比较。

模型	R@1	R@5	R@10	*Rsum*	#参数	场所
ViLBERT [29]	58.2	84.9	**91.5**	234.6	275M	NeurIPS 19
TERAN [28]	59.5	84.9	90.6	235.0	216M	TOMM 21
LTAN [19]	60.5	85.1	91.1	236.7	181M	ICASSP 22
CKSTN	**62.3**	**86.8**	91.3	**240.4**	**175M**	

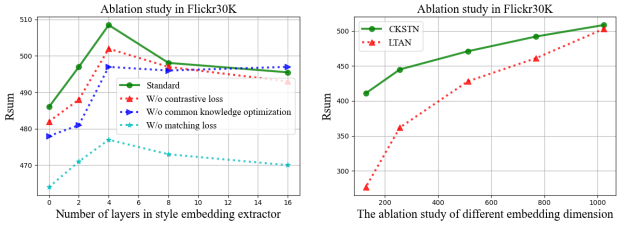


图3：Flickr30K 数据集上的消融实验。

表示。采用 4 层 SEE 的模型相比未使用 SEE 的模型，Rsum 相对提升了 5.2%。值得一提的是，在未使用 CKO 的模型中，增加 SEE 的层数反而会提升模型性能，这与使用 CKO 的模型呈现的变化不一致。使用 CKO 时，借助常识进行宏观调节，只需较少的 SEE 层便能提取项目的全局信息。未使用 CKO 时，更细致的风格嵌入则能更好地辅助特征表示。

样式嵌入提取器中不同层数的影响。我们设计了相关实验，通过逐步将层数增加到2层、4层、8层和16层，探究SEE中层数的影响。图3左侧将未经任何修改的CKSTN模型 обозначʏ为“标准”模型。我们的研究结果表明，在一定范围内，将层数从2层增加到4层可以提高检索准确率。然而，当层数进一步增加到8层时，模型的表征能力开始下降。

对我损失的影响。中的对比损失来衡量视觉特征与文本特征的风格嵌入之间的差异，从而对最终模型进行微调。如图3左侧所示，不使用对比损失时，结果略低于“标准”结果，这与预期一致。

与 LTAN 的进一步分析。LTAN 也是近期采用轻量级 Transformer 压缩模型参数的最先进论文之一。如图3右侧所示，当中间层的维度缩小时，LTAN 的结构会遭遇“悬崖效应”。然而，加入我们提出的 CKO 模块后，不仅可以缓解“悬崖效应”，还可以在中间层维度相同的情况下优化特征表示。

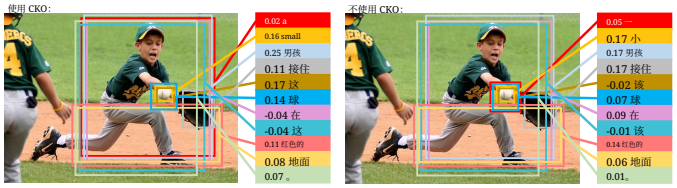


图4：区域-词语匹配的可视化对比。

C. 区域-词语匹配的可视化

图4展示了每个词与其关联度最高的对应图像区域之间的余弦相似度。视觉特征与文本特征之间的亲和度采用余弦相似度进行衡量，与训练阶段相同。若不使用 CKO，实例对象（如男孩、球）之间对齐关系的相似度就会下降。参见图4中的红色区域。冠词“a”修饰的是“boy”。然而，如果没有常识优化模块，冠词“a”就会错误地修饰“ball”。这也表明，常识包含全局信息，可用于优化特征表示。

IV. 结论

所提出的风格嵌入模块通过拼接和打乱操作，在特征图之间共享不同的行特征。为了捕捉不同模态的潜在概念，并根据风格嵌入动态调整这些概念，我们提出了公共知识优化操作。实验结果验证了 CKSTN 的有效性：在两个基准数据集上，其性能优于以往最先进的方法，同时参数量少于基于 Transformer 的方法。

参考文献

[1] M. Zhao、H. Zhang 和 J. Sun，“一种基于多趋势结构描述符的新型图像检索方法”，视觉通信与图像表示期刊，2016年。[2] F. Shang、H. Zhang、L. Zhu 和 J. Sun，“基于字典学习的对抗式跨模态检索”，神经计算，2019年。[3] X. Xu、H. Lu、J. Song、Y. Yang、H. T. Shen 和 X. Li，“用于零样本跨模态检索的自监督三元对抗网络”，IEEE《控制论汇刊》，2020年。[4] P. Anderson、X. He、C. Buehler、D. Teney、M. Johnson、S. Gould 和 L. Zhang，“用于图像描述生成和视觉问答的自底向上与自顶向下注意力”，载于 IEEE计算机视觉与模式识别会议论文集，2018年。[5] J. D. M.-W. C. Kenton 和 L. K. Toutanova，“BERT：面向语言理解的深度双向变换器预训练”，载于 NAACL-HLT 论文集，2019年。[6] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，“注意力就是一切”，《神经信息处理系统进展》，第30卷，2017年。[7] J. Song、Y. Yang、Z. Huang、H. T. Shen 和 R. Hong，“用于实时大规模近重复视频检索的多特征哈希”，载于第19届ACM多媒体国际会议论文集，2011年。[8] J. Song、Y. Yang、Z. Huang、H. T. Shen 和 J. Luo，“用于大规模近重复视频检索的有效多特征哈希”，IEEE《多媒体汇刊》，2013年。[9] W. Li、Z. Ma、L.-J. Deng、X. Fan 和 Y. Tian，“用于鲁棒图文检索的基于神经元脉冲传输与推理网络”，IEEE《视频技术电路与系统汇刊》，2022年。[10] K.-H. Lee、X. Chen、G. Hua、H. Hu 和 X. He，“用于图文匹配的堆叠交叉注意力”，载于欧洲计算机视觉会议论文集，2018年。[11] H. Chen、G. Ding、X. Liu、Z. Lin、J. Liu 和 J. Han，“IMRAM：通过循环注意力记忆进行迭代匹配，以实现跨模态图文检索”，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2020年。[12] X. Xu、T. Wang、Y. Yang、L. Zuo、F. Shen 和 H. T. Shen，“用于图文匹配的语义一致性跨模态注意力”，IEEE《神经网络与学习系统汇刊》，2020年。[13] K.-H. Lee、X. Chen、G. Hua、H. Hu 和 X. He，“用于图文匹配的堆叠交叉注意力”，载于欧洲计算机视觉会议论文集，2018年。[14] H. Lan 和 P. Zhang，“学习并整合用于图文检索的多层级匹配特征”，IEEE《信号处理快报》，2022年。[15] Y. Liu、H. Liu、H. Wang 和 M. Liu，“通过对比学习正则化用于图文匹配的视觉语义嵌入”，IEEE《信号处理快报》，2022年。[16] M. Zheng、W. Wang 和 Q. Li，“通过判别式潜在空间学习实现细粒度图文检索”，IEEE《信号处理快报》，2021年。[17] S. Mehta、M. Ghazvininejad、S. Iyer、L. Zettlemoyer 和 H. Hajishirzi，“DELIGHT：深度轻量级变换器”，载于国际学习表征会议，2021年。[18] S. Ren、K. He、R. Girshick 和 J. Sun，“Faster R-CNN：利用区域提议网络实现实时目标检测”，载于《神经信息处理系统进展》，2015年。[19] W. Li 和 X. Fan，“使用轻量级变换器进行图文对齐与检索”，载于 IEEE声学、语音与信号处理国际会议，2022年。[20] P. Young、A. Lai、M. Hodosh 和 J. Hockenmaier，“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度量”，载于《计算语言学协会会刊》，2014年。[21] T.-Y. Lin、M. Maire、S. Belongie、J. Hays、P. Perona、D. Ramanan、P. Doll´ar 和 C. L. Zitnick，“Microsoft COCO：常见物体上下文”，载于欧洲计算机视觉会议论文集，2014年。[22] S. Yang、Q. Li、W. Li、X. Li 和 A.-A. Liu，“面向图文检索的特征与上下文双层表征增强”，IEEE《视频技术电路与系统汇刊》，2022年。[23] Z. Fan、Z. Wei、Z. Li、S. Wang、H. Shan、X. Huang 和 J. Fan，“构建短语级语义标签，形成用于图文检索的多粒度监督”，载于多媒体检索国际会议

，2022年。

[24] Y. Cheng、X. Zhu、J. Qian、F. Wen 和 P. Liu，《用于图文检索的跨模态图匹配网络》，*ACM* 多媒体计算、通信与应用汇刊，2022年。[25] S. Long、S. C. Han、X. Wan 和 J. Poon，《GraDual：用于图文匹配的基于图的双模态表征》，载于 *IEEE/CVF* 冬季计算机视觉应用会议论文集，2022年。[26] H. Zhang、W. Wu 和 M. Zhang，《基于多头结构的图文检索多注意力级联模型》，载于 *2022*年国际神经网络联合会议，2022年。[27] L. Qu、M. Liu、J. Wu、Z. Gao 和 L. Nie，《用于图文检索的动态模态交互建模》，载于 第*44*届国际 *ACM* 信息检索研究与发展会议论文集，2021年。[28] N. Messina、G. Amato、A. Esuli、F. Falchi、C. Gennaro 和 S. Marchand-Maillet，《使用 Transformer 编码器进行跨模态检索的细粒度视觉文本对齐》，*ACM* 多媒体计算、通信与应用汇刊，2021年。[29] J. Lu、D. Batra、D. Parikh 和 S. Lee，《ViLBERT：面向视觉与语言任务的任务无关视觉语言表征预训练》，载于 神经信息处理系统进展，H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alch´e-Buc、E. Fox 和 R. Garnett 编，2019年。