阿1

文本到图像人物检索的跨模态自适应双重关联

林迪轩，彭一星，京科孟\*，郑伟士

*摘要:text -to-image* person reidentification (ReID)旨在基于给定的文本描述检索人的图像。关键的挑战是学习来自视觉和文本模态的详细信息之间的关系。现有的作品侧重于学习一个潜在空间，以缩小模态差距，进一步构建两种情态之间的局部对应关系。然而，这些方法假设图像到文本和文本到图像的关联是模态不可知的，从而导致次优关联。在这项工作中，我们展示了图像到文本关联和文本到图像关联之间的差异，并提出了CADA:跨模态自适应双重关联，该关联精细地构建了双向图像到文本的详细关联。我们的方法具有基于解码器的自适应双关联模块，该模块可实现视觉和文本模式之间的完全交互，从而允许双向和自适应跨模态通信关联。具体而言，本文提出了一种双向关联机制:文本标记与图像块的关联(ATP)和图像区域与文本属性的关联(ARA)。基于错误关联聚合跨模态特征会导致特征失真的事实，我们对ATP进行了自适应建模。对于ARA建模，由于属性通常是一个人的第一个区分线索，我们建议通过使用相关图像区域预测被屏蔽的文本短语来探索属性级关联。最后，我们学习了文本和图像之间的双重关联，实验结果证明了我们的对偶公式的优越性。代码将会公开。

*索引术语-人员*重新识别*，*文本到图像

I. 介绍

文本到图像的人物检索是一项任务，涉及从最匹配给定文本描述查询[1]的大型图片库中重新检索感兴趣的人物。文本描述提供了一种自然而全面的方式来描述一个人的属性，并且比图像更容易获取。因此，近年来，文本到图像的人物检索受到越来越多的关注，受益于从个人相册搜索到公共安全的各种应用。

与基于图像的人物检索方法[2]-[7]相比，由于视觉和语言之间固有的表征差异所导致的显著的情态异质性，文本到图像的人物检索更具挑战性。解决这一问题的关键在于建立两种模态之间的对应关系。早期的方法采用全局匹配技术[8]，[9]在特征空间中将图像与其对应文本拉近。

Dixuan Lin, Yixing Peng, Jingke Meng and Wei-Shi Zheng are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China.

pengyx23@mail2.sysu.edu.cn

邮件：

asxlwsl@gmail.com，

，

mengjke@gmail.com and wszheng@ieee.org \* Corresponding author.

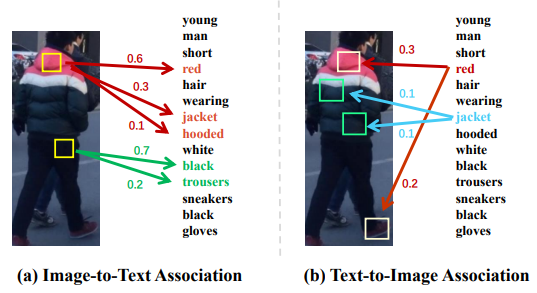


图1。跨模态双重关联示意图。(a)以图像块作为锚点的关联。(b)以文本标记作为锚点的关联。值得注意的是，这些关联依赖于锚。例如，红色夹克的图像片段与文本中的“红色”一词高度相关，而“红色”一词可以同时与图像中的红色鞋子和红色夹克相关联。

响应文本在特征空间中紧密相连。然而，由于全局匹配倾向于关注最显著的特征，细粒度的信息匹配对于Re-ID至关重要。最近的研究主要集中在探索两种模式之间更详细的对应关系，比如图像补丁和单词[10]、[11]，以及学习局部特征[12]、[13]。此外，还利用了来自 CLIP [14] 等有前景的视觉语言预训练模型的知识，并将其转移到文本到图像的人员检索任务 [15] [17]。。

尽管该领域最近取得了进展，但目前用于在视觉和文本之间建立跨模态关联的方法仍然不完全令人满意。造成这种情况的主要原因是建立与锚相关的详细对应关系的复杂性。具体来说，如果将一个图像块作为锚点，则多个单词可以在语义上关联，但这些关联的大小应该主要取决于图像块。例如，如图1所示，红色夹克的图像块可以与文本中的“红色”和“夹克”相关联，而“红色”这个词可以同时与图像中的红色鞋子和红色夹克相关联。此外，“夹克”一词可以与夹克覆盖的所有图像块相关联。这种差异意味着从图像到文本的关联并不一定意味着从文本到图像的关联。遗憾的是，以往的工作忽略了这一特点，仅仅采用一种模态作为学习的锚点，导致跨模态理解不足。

在这项工作中，我们提出了一种名为Cross-的新颖方法

arXiv:2312.01745v1 [cs。CV] 2023年12月4日

2

模态自适应双重关联(CADA)。我们的方法具有一个解码器，可以实现视觉和文本模态之间的完全交互，允许双向和自适应跨模态通信关联。双向关联意味着文本与图像的关联以及图像与文本的关联，这有利于模型理解来自两种模态的相应信息。为了有效地学习这些双重关联，我们的解码器与文本编码器共享除了交叉注意层之外的所有参数，这些参数只存在于解码器中。这是因为自注意层和前馈层在编码和解码任务中工作相似，因此共享参数可以提高训练效率，同时实现双跨模态交互和关联学习。具体来说，跨模态自适应双关联框架利用流行的编码器-解码器框架进行特征提取和跨模态交互。作为全面和详细关联的基础，来自特定模态编码器的特征被对齐，以实现大致全局级的关联(从句子到图像)。

作为我们方法的核心，我们增强了一个基于解码器的自适应双关联模块，该模块可以实现局部级的双关联。具体来说，该模块工作于两个任务:文本令牌到图像块(ATP)的关联和图像区域到文本属性(ARA)的关联。在ATP中，我们建议将信息从文本标记传递到图像块以构建标记级关联。我们施加了一个匹配约束来规范已经构建的关联应该汇总给与文本标记相关的图片块。此外，属性通常是区分一个人的第一印象。因此，我们提出ARA，通过掩码属性建模(mask Attribute Modeling, MAM)来探索文本短语和图像块之间的关联。该方法受视觉语言与训练模型中的掩码语言建模的成功启发，MAM通过自适应定位相关图像区域生成掩码文本短语。

综上所述，这项工作的贡献有三个方面。

1)我们提出了一种跨模态自适应双重关联(Cross-modal Adaptive Dual Association, CADA)方法，该方法将视觉和文本模态双向关联，而目前的方法将两个非等效关联视为一个关联。

2)引入文本标记到图像块的关联(Association of text tokens to image patches, ATP)，它允许将信息从文本模态传递到视觉模态，并自适应地聚合与文本标记锚点相关的图像块。

3)我们提出了一种图像区域与文本属性关联(Association of image regions to text attributes, ARA)方法来探索文本短语与图像patch之间的关系，该方法通过掩码属性建模(masking Attribute Modeling, MAM)实现，通过自适应定位相关图像区域来生成掩码文本短语。

我们进行了广泛的评估，以验证我们提出的方法在三个公共数据集(即CUHK [1]， ICFG-PEDES[21]和RSTPReid[22])上的有效性。结果表明，所提出的CADA方法在很大程度上优于最先进的方法，证明了我们所提出方案的有效性

II. 相关工作

*A. 文本到图像的人物检索*

文本到图像的人检索最早由[1]提出，发布了第一个基准数据集CUHK-PEDES。随后，提出了更具挑战性的数据集ICFG-PEDES[21]和RSTPReid[22]用于文本到图像的人物检索。与一般的文本-图像检索基准数据集(如COCO[23]和methods[9])相比，[19]的人物检索更具挑战性，因为它对跨模态细节对齐的要求很高。根据对齐中的策略，目前的工作可以分为两个分支:无跨模态交互方法和基于跨模态交互的方法。前一个方法分支[9],[24],[25]侧重于开发一个更好的特征提取模型在一个共享的潜在空间中去学习判别图像文本嵌入 ，提出CMPM损失[8]来对齐跨模态表示;IVT[20]则引入了一个额外的双向掩码建模块，来督促全局特征包含多层次信息。尽管这些方法效率很高，但它们的性能并不令人满意，因为它们只对齐全局表征，浪费了局部特征中丰富的细节信息。

由于全局对齐在包含整体细节方面存在不足，基于跨模态交互的方法旨在充分利用局部信息。最近的工作集中在局部级别的对应。部分工作专注于突出的图像块[1]、[26]-[28],旨在图像块-单词匹配框架，丁[21]利用注意机制捕捉身体部位之间的关系。更近期的作品放弃了分割和身体地标等外部线索，转而采用隐式对齐。AXM[29]在特征提取中使用多尺度卷积层，TIPCB[30]使用不同的残差块捕获不同尺度的语义信息，然后在同一尺度下对齐跨模态特征。CAIBC[31]分别对齐颜色相关特征和颜色不相关特征。CFine[17]采用Vision-Transformer[32]和BERT[18]作为主干进行特征提取，取得了更好的性能。

然而，目前的方法在构建局域级跨模态关联方面仍然不能令人满意。详细的对应关系复杂且与锚点相关。从图像到文本的关联并不意味着从文本到图像的关联。遗憾的是，以往的作品忽视了这一特点，仅仅采用一种模态作为学习的锚点，导致跨模态理解不足。

不同的是，我们引入了一个基于解码器的自适应双关联模块，使视觉和文本模态之间能够全面交互。这允许双向和自适应跨模态通信关联，这对于以详细和准确的方式学习两种模态之间的关系至关重要。

*B. 视觉语言预训练(VLP)*

视觉语言预训练(VLP)的过程是在大规模图像文本数据集上使用预训练好的参数初始化视觉语言模型，然后在下游任务上

3

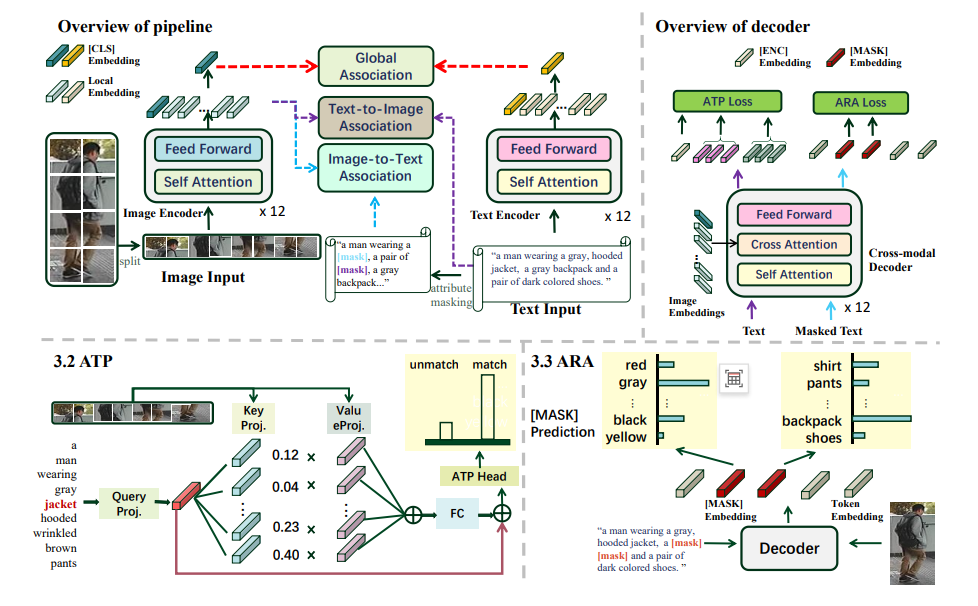


图2.。我们的CADA框架概述，其中包括两个特定于模态的编码器和一个跨模态的解码器(相同的颜色表示共享参数)。编码器在section . iii - a中的归一化分布拟合损失下提取全局特征并学习全局关联。解码器通过文本标记到图像补丁的关联(ATP)模块和文本标记到图像补丁的关联(ARA)模块学习跨模态双重关联。在ATP模块中，我们利用交叉注意机制，通过学习文本-图像匹配概率，将每个标记与相关图像补丁聚合在一起，这从本质上指导模型捕获文本标记与图像补丁的关联。在ARA模块中，强制模型通过来自图像补丁的提示来恢复被屏蔽的属性短语，这需要将图像补丁与文本令牌相关联。ATP和ARA的详细信息分别在章节iii - b和章节iii - c中阐述。

微调该模型。之前的许多工作[18]，[32]，[33]都证明了以这种方式转移知识可以提高下游任务的性能。一些作品在共享空间中对齐图像和文本的全局表征。CLIP[14]和ALIGN[34] 采用对比学习损失来学习强连贯的跨模态表征以实现跨模态对齐，用于跨模态对齐。其他一些作品则利用了局部特征的相互作用。ALBEF[19]将对比学习和掩码语言建模(MLM)与双编码器设计相结合来学习图像-文本关系。Coca[35]引入了一个解码器来聚合跨模态表示，BLIP[36]同时利用多阶段学习来提高图像-文本理解和基于图像的字幕能力。

文本-图像人检索的最新发展通过利用预训练的视觉语言模型，而不是依赖仅在单模态数据集上预训练的主干，取得了显着的改进。最近的研究[15]，[17]提出了CLIP驱动的框架，使用预训练的CLIP作为主干，并在文本到图像的人再识别数据集上进行微调

同样，这项工作也从VLP模型的成功中获得灵感，并利用他们的知识来增强文本到图像人物检索任务中的全局对齐和局部交互。

III. 跨模态自适应双重关联

*A.结构体系概述*

在本节中，我们将介绍拟议的跨模态自适应双关联(CADA)模型的总体架构。CADA框架是基于一种流行的编码器-解码器网络架构进行的。编码器用于从不同模态的数据中提取高级语义信息，解码器实现详细的双向跨模态交互，这是有效跨模态Re-ID的核心。对于编码器，我们提出了一个基于编码器的全局关联模块来大致对齐两模态的全局特征。对于解码器，我们提出了基于解码器的局部自适应双关联模块，以双向的方式增强视觉和文本模态的交互。我们的方法成功地将图像和文本的局部区域关联到标记级和属性级。

具体操作如下。

*1)基于编码器的全局关联:*

***图像编码器***：我们利用预训练的ViT作为图像编码器来提取视觉特征。对于给定的图像，其中H, W, C分别表示其高度，宽度和通道，我们将其分成N = H ×W/P 2个补丁的序列，其中P表示一个图块的长度。在patch序列的开头附加一个额外的可学习令牌作为全局视觉表示。添加了可学习的位置嵌入以合并空间信息。然后将patch序列馈送到视觉转换器中。图像编码器的输出表示为，其中vcls是令牌的dv-dim嵌入。

***文本编码器***:我们使用12层的BERT作为我们的文本编码器。对于给定的文本T，我们相应地将T中的每个词映射到它的词嵌入。使用特殊令牌表示整个句子，使用[PAD]将嵌入填充到固定长度。与位置嵌入聚合的词嵌入序列被馈送到文本编码器中,以获得文本表示，其中M为句子长度，tcls为令牌的dt-dim嵌入。

***全局关联***:全局关联是通过在共享潜在空间中对齐图像和文本的全局特征来学习的。先前的研究[17]，[30]，[31]证明了图像-文本对比学习可以通过跨模态投影匹配(cross-modal projection matching, CMPM)损失[25]有效地对齐两模态的全局特征。我们注意到CMPM损失以投影长度作为相似度，其中向量的大小可能会有很大的影响。为了缓解CMPM的缺点，我们提出了一种新的归一化分布拟合(NDF)损失，该损失采用余弦相似度代替投影，并使用正向KL散度和反向KL散度来加速分布拟合。

给定图像I和文本T，我们获得全局表示 和 ，具有可学习的线性变换 and 。我们定义作为相似函数。在由N对图像-文本组成的mini-batch中，I到T的图像-文本匹配概率计算为:

FORMULA

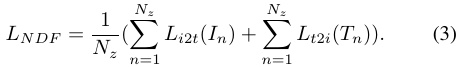
其中τ是一个温度参数，用于控制概率分布的清晰度。

假设在一个mini-batch中有Nz图像和文本。将ground-truth - one-hot标签记为qi2t∈RNz，则图像到文本的对比损失公式为:

FORMULA

KL(p||q)表示从分布p到分布q的KL散度，集中于p概率高而q概率低的样本。我们将式(2)中的前一部分称为正向KL散度，后一部分称为反向KL散度。在最小化正向KL散度侧重于拉低负对的匹配概率，而后向KL散度侧重于拉高正对的匹配概率。这样，就加速了对齐过程。

同样，文本到图像的对比损失可以通过交换式(1)、(2)中的I和T来表示。我们提出的归一化分布拟合(NDF)损失可以表示为:



*2) 基于解码器的局部自适应双关联:*编码器为基础的全局关联工作在整体图像和文本的表示上，这是不足以关联两种模态中全面和详细信息的。因此，我们提出了一种基于解码器的局部自适应双重关联方案，以增强T中单词与其单词嵌入的交互作用。特殊标记以双向方式处理视觉和文本模态，并自适应地关联标记级别和属性级别中的图像和文本的局部区域。具体而言，解码器为基础的局部自适应双重关联是通过将文本令牌与图像块（ATP）关联和将图像区域与文本属性（ARA）关联来实现的。我们在第III-B节和第III-C节详细介绍了ATP和ARA。

*B. 文本令牌与图像补丁的关联*

对于文本标记与图像补丁的关联(ATP)，我们基于文本标记锚点自适应聚合相关图像块，将信息从文本模态传递到视觉模态。由于单个单词可以与图像中的不同部分相关联，因此我们需要根据单词在句子中的位置来确定该单词所描述的特定部分。所以我们进行自注意机制和交叉注意机制来获得文本-图像聚合特征。直观地说，如果图像和文本真正匹配，它们之间的关联应该是很强的，因此相应的特征可以自然地聚合。相比之下，聚合来自负对(图像和文本指的是不同的人)的特征会导致特征失真。鉴于此，我们通过将文本嵌入和图像嵌入反馈到具有交叉注意层的解码器中，来交互单词和图像补丁之间的关联。通过评估聚合特征上的匹配损失，交叉注意层学习如何匹配和聚合来自不同模态的标记级细粒度特征。

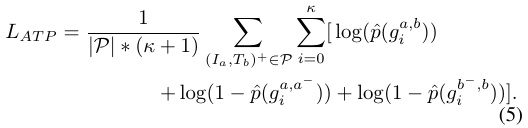
ATP中的解码器架构和交叉注意机制如图2所示。给定一个图像-文本对(Ia，Tb)，我们首先对Tb进行标记和嵌入以获得文本嵌入。在序列的开头附加特殊的令牌[ENC]。然后令牌嵌入被馈送到交叉注意层。图像嵌入 = 也被馈送到交叉注意层。我们对交叉注意层中的关联进行建模。在操作交叉注意时，将作为查询，将作为键和值。我们表示作为解码器的最终输出嵌入。将ei与加权组合进行聚合后得到嵌入的输出。

5

其中相关图像patch获得更高的权重。如果一个令牌能与某些图像patch对应，那么它的聚合嵌入应该是和谐的，否则会混淆。为了保证相关的图像特征能够自然地与文本标记相对应地聚合，最直接的方法是用二值分类损失约束中的每个元素，以预测文本标记及其对应的聚合特征是正(匹配)还是负(不匹配)。考虑到训练效率和无意义单词(例如“a”、“of”等)产生的噪声，我们将分成组，并计算每组的匹配损失。具体来说，我们将分组大小为p的聚合特征和切割步幅r分组: ，其中。第i组表示为。我们应用对每组进行平均池化，得到作为组表示。另外，我们取作为额外的组。对于每个组表示 (0<i<k)，我们将其馈送到分类器中，以计算“匹配”和“不匹配”的概率，公式为:

FORMULA

其中为Ia和Tb基于的匹配概率。我们期望对正对得到高匹配概率，对负对得到低匹配概率。对于每个，我们选择一个硬否定文本Ta来自对不同人具有最高全局相似度的小批文本并且表示不同的人，对于每个Tb，我们以相同的方式选择。总体上，ATP损失计算如下:



(5)

其中P是一个小批量中的正对的集合。(Ia，Tb)+表示正的图像-文本对，其中和表示同一个人。通过最小化LATP，该模型仅从细节上区分有差异的负面图像-文本对，迫使其正确构建文本标记和图像块之间的关联。

*C. 图像区域与文本属性的关联*

在ATP模块中，我们的模型学习理解从文本中的单词到图像块的关联。在本节中，我们将讨论图像到文本的关联。为了学习图像到文本的关联，我们提出了一个ARA模块，该模块主要关注图像区域与文本属性的关联。由于属性通常是识别一个人时首先想到的东西，因此对于属性级跨模态关联来说，这是一个值得探索的重要特征。为了实现这一目标，我们提出MaskAttributeModeling,（MAM）技术，该技术的灵感来自掩码语言建模在视觉语言预训练[15]，[18]，[19]，[37]中的成功。在MAM中，我们随机掩盖文本描述中的属性短语，并训练解码器通过定位相关图像区域来预测正确的短语。

与MLM相比，MAM更注重与属性相关的词，而不是随机屏蔽所有词，因为图像中的属性对人物图像的检索很重要。为了实现这一点，我们使用自然语言工具包NLTK1进行词性标注，然后以“[形容词][名词]”的模式表示短语作为属性，例如“黑鞋”和“长直发”。

屏蔽一个属性短语意味着该短语中的每个单词都被[MASK]取代。解码器学习预测整个被屏蔽的属性短语。具体来说，给定一个正的图像-文本对，我们通过随机屏蔽具有屏蔽率α的属性短语来获得屏蔽文本Tb′。然后我们将输入解码器以获得最终嵌入 a,b ' 作为as Sec[III-B]中的子集。'，由中每个[MASK]令牌的标记嵌入组成。表示中屏蔽令牌的数量。预测过程可以公式化为一个分类问题。在BERT[18]之后，我们持有一个大小为的词汇表，包括所有可能的单词。我们将输入一个分类器来计算词汇表中每个单词的概率:

FORMULA

其中是对每个掩码令牌的预测。我们取被屏蔽词的标签′来自作为真值标签，ARA损失计算为:

FORMULA

其中P是mini-batch中正图像-文本对的集合。我们应用KL散度来对齐预测分布和ground-truth分布。ARA模块通过使用相关图像区域预测正确的属性来实现图像到文本的关联。

*D. 训练与推理*

**训练。**综上所述，我们的跨模态自适应双关联(CADA)框架的整体损失可以计算为:

FORMULA

其中λ为权衡参数

**推理**，在测试阶段，我们在先前工作[15]中定义的全局匹配和局部匹配推理协议下进行评估。

**全局匹配推理:**我们分别通过图像编码器和文本编码器提取全局特征，然后利用全局特征之间的余弦相似度SG(Eq.(1))作为最终匹配分数。

**局部匹配推理:**将图像和文本特征序列送入解码器进行交互并计算匹配分数SL(即Eq.(4)中的i.e., )。为了提高效率，我们基于SG对图库中的所有图像进行排序，并选择最上面的

1 https://www.nltk.org/

6

η候选图像来计算SL。我们取SG+SL作为前η候选人的最终匹配分数，SG作为其余候选人的最终匹配分数。在所有基准上进行测试时，我们设置η= 32。在不同η值下的性能讨论见第iv - e节。

IV. 实验

*A. 数据集和评估指标*

**数据集**。我们在CUHK-PEDES[1]、ICFG-PEDES[21]和RST-PReid[22]三个公共数据集上评估了我们的方法。

CUHK-PEDES是首个可供使用的文字-图像人物检索数据集。它包含40,206张图像和80,412个文本描述，涉及13,003个身份，每个图像都有2个文本描述注释。按照官方协议，数据集在训练集中分为34,054张图像和68,108个描述，用于11,003个身份;在验证集中分为3,078张图像，用于1,000个身份，包含6,156个描述;在测试集中分为3,074张图像，包含1,000个人的6,148个描述。

ICFG-PEDES来自MSMT-17数据集[38]，它具有更多的身份和文本描述。ICFG-PEDES包含54,522个文本描述，54,522个图像，4,102人。在训练集中分为3102个身份的34,674个图像-文本对，在测试集中分为1,000个身份的19,848个图像-文本对。

RSTPReid与MSMT-17相同，包括20,505张图像，41,010个文本描述，4,101人。每个人包含5张由15个不同相机拍摄的图像，每张图像有2个对应的文本描述。训练包含3701个身份，验证集和测试集分别包含200个和200个身份。

评估指标。对于所有实验，我们报告了Rank-1、Rank-5和Rank-10的精度和平均平均精度(mAP)，以评估我们模型的性能。

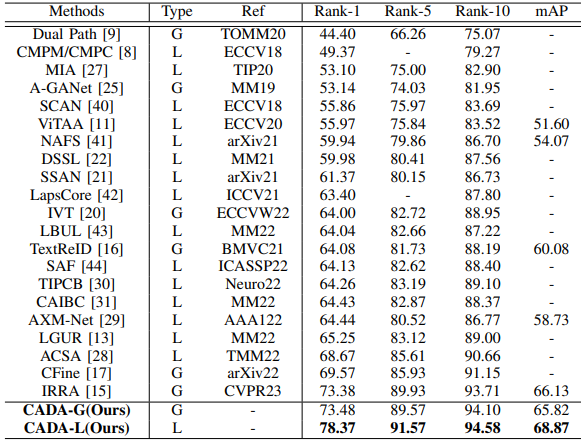
*B. 实现细节*

我们使用12层的ViT-B/16[32]作为图像编码器，使用12层的BERT-base[18]作为文本编码器。我们使用BLIP[36]参数初始化我们的模型，该参数在129M的图像-文本对上进行了预训练。在训练过程中，我们采用随机水平翻转、随机擦除和随机裁剪来进行数据增强。将每张图像的大小调整为224 ×224。输入句子的最大文本长度设置为72。嵌入维数dv，dt= 768，共享潜在维数d = 256。MAM的掩蔽率α为0.8，每个掩蔽字被替换为一个特殊的令牌[MASK]。Eq.(1)中NDF损耗的温度参数τ设为0.02。Eq.(8)中的参数λNDF设为0.1。我们使用权重衰减为0.05的AdamW优化器[39]。我们的模型训练的批大小为96，进行了40次迭代。学习率初始化为1e−5，使用余弦学习率衰减调度器。

表一

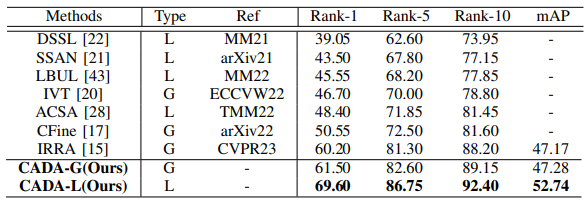
CUHK-PEDES数据集上的性能比较。

“类型”中“G”和“L”表示全局匹配和局部匹配推理协议



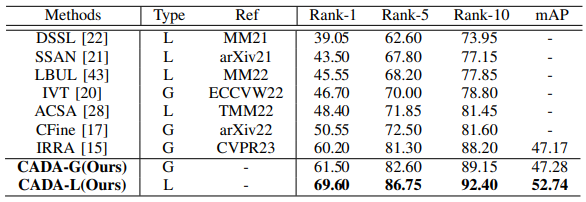
表二

ICFG-PEDES上的性能比较。



表三

RST-PReid上的性能比较。



*C. 与其他方法的比较*

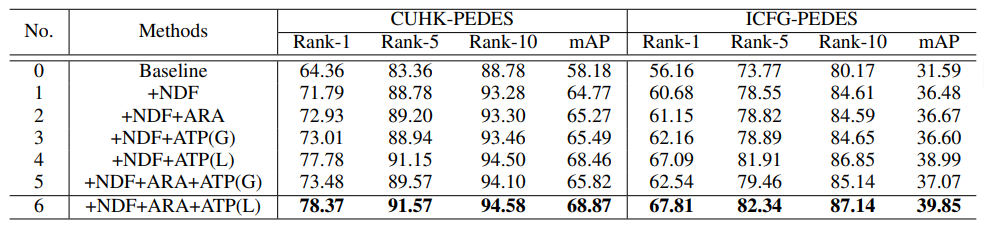
我们将我们的方法与最先进的方法在三个基准数据集上进行比较，如表1、表2和表3所示。

**CUHK-PEDES:**在CUHK-PEDES数据集上与相关方法的性能比较见表1。可以观察到，我们提出的CADA在两个推理协议上都达到了最先进的结果。在全局匹配推理协议下，CADA达到73.48%

7

表四

消融研究 “G”和“l”分别代表全局匹配和局部匹配推理协议。



Rank-1精度和65.82%的mAP，与最强的竞争对手(IRRA[15])相当接近，并略有提高，即Rank-1精度提高了0.10%。在局部匹配推理协议下，CADA Rank-1准确率达到78.37%(+4.99%)，mAP准确率达到68.87%(+2.74%)，显著优于IRRA[15]。结果表明，我们提出的基于双向公式的CADA可以更好地学习详细的跨模态关联。此外，全局匹配结果表明，建立双向局部关联有助于缩小全局表征的模态差距。

**ICFG-PEDES:**如表2所示，我们提出的CADA在全局匹配推理协议下获得的结果与最先进的IRRA[15]方法相当。在局部匹配推理协议下，我们的方法在Rank-1上的准确率达到67.81%，比最新的IRRA[15]和CFine[17]分别高出+4.35%和+6.98%。值得注意的是，由于复杂背景和不稳定光照的负面影响，ICFG-PEDES的全局匹配不足。结果表明，CADA可以在局部水平上双向学习跨模态关联，克服背景和光照的差异。

RSTPReid:我们还对新发布的基准数据集RSTPReid进行了实验，结果如表III所示。可以观察到，我们提出的CADA在很大程度上优于其他现有方法，在Rank-1和mAP上分别超过IRRA[15] +9.40%和+5.57%。此外，与其他局部匹配方法相比，CADA也优于所有现有方法，在Rank-1精度上超过最新的局部匹配方法CFine[17] +19.05%。

*D. 消融研究*

为了充分展示CADA框架中每个模块的有效性，我们在CUHK-PEDES和ICFG-PEDES两个基准上进行了消融研究。我们采用双编码器(即vitb /16和基于bert)，以CMPM[25]损耗为基线。实验结果如表4所示。

**归一化分布拟合损失的有效性**。对比No.0和No.1的结果，我们可以观察到，用提出的LNDF替代常用的交叉模态投影匹配(CMPM)损

失，Rank-1和mAP在CUHK-PEDES上分别提高了7.43%和6.59%，在ICFG-PEDES上分别提高了4.52%和4.89%。结果表明，LNDF在对齐跨模态全局表示方面更有效，因为它消除了矢量幅度变化的影响。

**文本标记与图像补丁(ATP)关联的有效性**。ATP模块通过利用文本标记和图像块的完整交互以及构建关联显著提高了性能。比较No.1和No.4的结果，很明显，构建标记级关联极大地提高了精度和召回率的性能，Rank-1精度提高了+4.85%，+5.94%，mAP精度提高了+3.69%，+2.51%。此外，由于ATP学习了细粒度局部特征的交叉模态关系，从而缩小了共享潜在空间中全局特征的距离。如No.1与No.3所示，即使解码器在推理中不使用，ATP也可以提高Rank-1的精度+1.22%，+1.48%。结果表明，在构建标记级关联过程中，CADA解码器通过检查局部细节的一致性来学习推断匹配概率，同时编码器学习将全局特征更精确地映射到潜在空间中。

**图像区域与文本属性关联(ARA)的有效性。**ARA是通过掩码属性建模实现的，该建模用于学习图像区域与文本属性之间的关联。在No.1和No.2中可以看到，直接将ARA模块添加到全局匹配框架中，而不进行任何标记级交互，模型在Rank-1精度上分别提高了+1.14%和+0.47%，证明通过建立区域-属性关联，增强了单模态编码器的性能，缩小了模态差距。通过比较No.3与No.5、No.4与No.6，我们可以观察到ARA模块对全局匹配推理和局部匹配推理增益的提高。

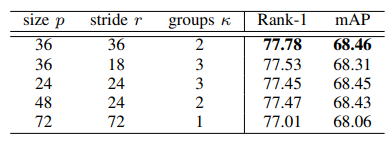
*E. 超参数分析*

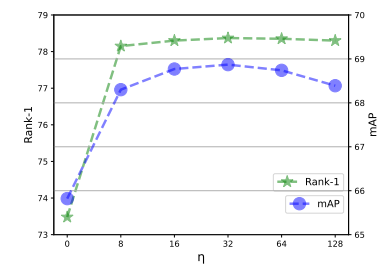
**局部特征分组分析。**考虑到训练效率和无意义单词(例如，“a”，“of”等)的噪声，我们将局部聚合特征分成组，以挖掘Sec.III-B中的局部关联。我们评估分组如何

8

表五

在CUHK-PEDES数据集的III-B部分，使用不同的分割步幅和组大小进行评估。





(a)CUHK-PEDES数据集在不同η下的表现。

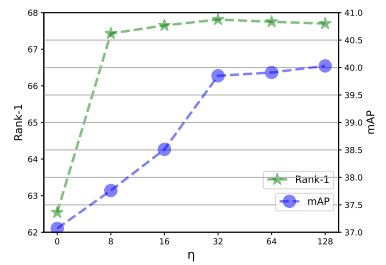
(b)ICFG-PEDES数据集在不同η下的表现。

图3 不同η下的性能。

Size p和拆分步幅r会影响文本到图块关联。我们在不同的p和r组合下评估了我们的模型，并在表v中报告了结果。可以观察到，当将组大小设置为32且没有重叠时，性能达到最佳。在这样的分割方法中，我们保证每个组包含足够的局部聚合特征来判断匹配。

**遮罩属性建模的掩码率分析。**我们分析了MAM掩蔽率的影响，见第iii - c节。掩码率设置的实验结果如图4所示。我们可以观察到，将屏蔽率设置为0.8可以在Rank1和mAP指标上获得更好的性能。

**交互作用对选择的分析。**在局部匹配推理协议中，我们混合了全局匹配分数SG

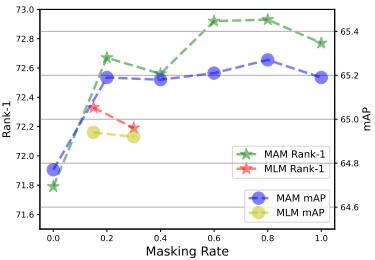
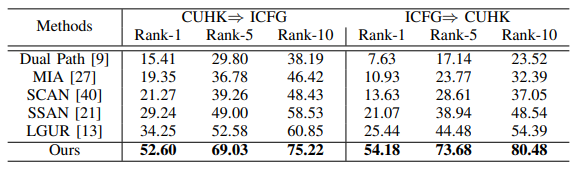


图4. 不同MAM掩蔽率下CUHK-PEDES的评价。本文还比较了掩蔽率为0.15和0.3的MLM的结果。

表六

领域泛化任务的性能比较。

A⇒B表示我们将A数据集作为训练集，B数据集作为测试集。



Eq.(1)和Eq.(4)中的局部匹配分数SL，以提高匹配精度。显然，简单地计算所有图像-文本对的交互分数会带来巨大的计算成本，对于M个图像和N个文本，其复杂度为O(MN)。因此，对于每个文本，我们只选择具有最高η全局相似性的η图像，然后进行η次局部交互以获得它们的匹配分数，将计算复杂度缩小到O(M + ηN)。我们的实验证明，与全局匹配方法相比，η可以设置得很低，以避免高的计算成本，并大大提高性能。我们在不同η值下的CUHK-PEDES和ICFG-PEDES上对我们的方法进行了评估，实验结果如图3所示。我们还设置η= 0作为比较，这意味着没有进行相互作用。可以观察到，当我们设置η= 32时，我们已经可以得到很大的精度提升，这在很大程度上节省了计算成本。值得注意的是，即使当我们限制η< 10时，CADA仍然优于所有现有的最先进的方法，证明了我们模型的效率。

*F. 参数数量的分析*

在我们的方法中，我们在文本编码器和跨模态解码器中共享参数，并且在我们的模型中总共使用了36个注意层:视觉编码器中有12个自注意层，文本编码器中有12个自注意层+ 12个交叉注意层。

为了评估增加参数的影响，我们使用不同数量的注意层进行了进一步的实验

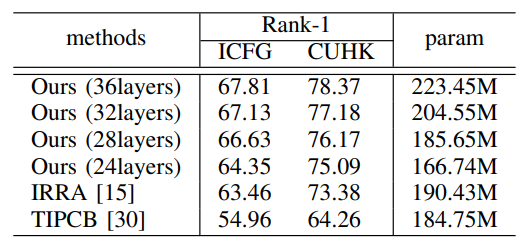
9



图5 文本标记到图像补丁关注的示例。对于每个人，我们在红色框架中显示图像，在蓝色框架中显示标记到补丁关注图像。注意权重用红点标记。

表七

模型在不同数量的注意力层下的表现和参数数量的比较结果



并报告参数的个数，如表七所示。结果表明，减少注意层的数量会导致相应的性能下降。然而，值得注意的是，当我们将层数减少到24层时，我们的模型包含的参数比IRRA[15]少，但仍然优于它。

综上所述，我们的实验结果表明(1)增加参数的数量确实提高了性能。(2)我们的改进不仅仅是因为参数的增加，因为即使在保持相同数量的参数时，我们的方法仍然优于IRRA[15]。

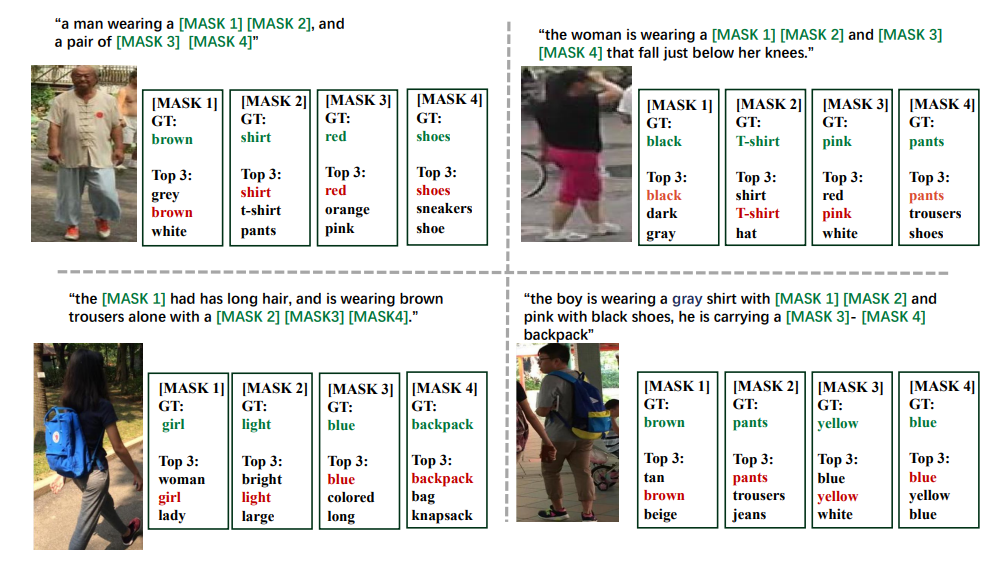
*G. 定性结果*

图5显示了给定文本单词的图像补丁的注意权重。它清楚地表明，对于颜色相关的单词(例如“红色”和“蓝色”)，我们的CADA精确地捕获了相应颜色的图像块。对于与衣服或身体部位相关的单词(如“pants”、“hair”)，CADA自适应地关注整个相关区域。我们还展示了与图像无关的单词(例如“the”)的注意图，并得到了一个不规则的注意图，揭示了这些单词在构建图像-文本关联时可能会引起混淆，这证明了文本令牌与图像块(ATP)关联的有效性。

在图6中，我们说明了我们提出的方法中图像区域与文本属性关联(ARA)模块的影响。对于每个被屏蔽的属性词，我们在Eq.(6)中显示了前3个预测。在来自图像块的线索下，我们的模型成功地在描述性文本中建立了从视觉信息到属性(如性别、颜色、服装和身体部位)的对应关系。

*H. 更多评价*

**领域泛化任务的比较。**我们提出的CADA有效地学习了细粒度细节的关联。由于它在细粒度细节之间建立关联的有效性，背景或光照变化的影响可能是微妙的。进一步验证泛化



我们在两个基准上进行了域泛化分析，其中我们直接使用在源基准上预训练的CADA对目标基准进行评估。结果如表7所示。我们提出的CADA在很大程度上优于所有方法。例如，我们的CADA在Rank-1准确率上分别比LGUR b[13]高出18.35%和28.47%。

V. 结论

在这项工作中，我们指出文本和图像之间的关联与锚相关，并开发了一个跨模态自适应双关联(CADA)框架。我们的框架由一个全局级关联和一个基于解码器的自适应双关联模块组成。该模块通过文本标记与块(ATP)的关联以及图像区域与文本属性(ARA)的关联，实现了视觉和文本模态之间的双向和自适应跨模态对应关联。具体来说，在ATP中，我们建议将信息从文本标记传递到图像块，以建立标记级关联。相比之下，在ARA中，我们提出通过自适应定位相关图像区域来生成屏蔽文本短语。通过构建双向跨模态关联，我们在所有三个流行基准上都取得了显着的性能，证明了我们方法的有效性。

REFERENCES

[1] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, “Person search with natural language description,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[2] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, “Pose-guided feature align- ment for occluded person re-identification,” in *IEEE/CVF international conference on computer vision (ICCV)*, 2019.

[3] H. Fan, H.-M. Hu, S. Liu, W. Lu, and S. Pu, “Correlation graph convolu- tional network for pedestrian attribute recognition,” *IEEE Transactions on Multimedia(TMM)*, vol. 24, pp. 49–60, 2020.

[4] Y.-X. Peng, J. Jiao, X. Feng, and W.-S. Zheng, “Consistent discrepancy learning for intra-camera supervised person re-identification,” *IEEE Transactions on Multimedia(TMM)*, vol. 25, pp. 2393–2403, 2022.

[5] X. Ning, K. Gong, W. Li, L. Zhang, X. Bai, and S. Tian, “Feature refine- ment and filter network for person re-identification,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, pp. 3391–3402, 2020.

[6] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, “Diverse part dis- covery: Occluded person re-identification with part-aware transformer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[7] Y. Shi, Z. Wei, H. Ling, Z. Wang, J. Shen, and P. Li, “Person retrieval in surveillance videos via deep attribute mining and reasoning,” *IEEE Transactions on Multimedia(TMM)*, vol. 23, pp. 4376–4387, 2020.

[8] Y. Zhang and H. Lu, “Deep cross-modal projection learning for image- text matching,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[9] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, “Dual-path convolutional image-text embeddings with instance loss,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, pp. 1–23, 2020.

[10] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan, “Pose-guided multi-granularity attention network for text-based person search,” in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[11] Z. Wang, Z. Fang, J. Wang, and Y. Yang, “Vitaa: Visual-textual attributes alignment in person search by natural language,” in *Proceedings of the European conference on computer vision (ECCV)*, 2020.

[12] X. Gong, Z. Yao, X. Li, Y. Fan, B. Luo, J. Fan, and B. Lao, “Lag-net: Multi-granularity network for person re-identification via local attention system,” *IEEE transactions on multimedia(TMM)*, vol. 24, pp. 217–229, 2021.

11

[13] Z. Shao, X. Zhang, M. Fang, Z. Lin, J. Wang, and C. Ding, “Learn- ing granularity-unified representations for text-to-image person re-identification,” in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2022.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning (ICML)*, 2021.

[15] D. Jiang and M. Ye, “Cross-modal implicit relation reasoning and align- ing for text-to-image person retrieval,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[16] X. Han, S. He, L. Zhang, and T. Xiang, “Text-based person search with limited data,” *arXiv preprint arXiv:2110.10807*, 2021.

[17] S. Yan, N. Dong, L. Zhang, and J. Tang, “Clip-driven fine-grained text- image person re-identification,” *arXiv preprint arXiv:2210.10276*, 2022.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[19] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi,“Align before fuse: Vision and language representation learning with momentum distillation,” 2021.

[20] X. Shu, W. Wen, H. Wu, K. Chen, Y. Song, R. Qiao, B. Ren, and X. Wang, “See finer, see more: Implicit modality alignment for text- based person retrieval,” in *Proceedings of the European conference on computer vision (ECCV)*, 2023.

[21] Z. Ding, C. Ding, Z. Shao, and D. Tao, “Semantically self-aligned network for text-to-image part-aware person re-identification,” *arXiv preprint arXiv:2107.12666*, 2021.

[22] A. Zhu, Z. Wang, Y. Li, X. Wan, J. Jin, T. Wang, F. Hu, and G. Hua,“Dssl: deep surroundings-person separation learning for text-based per- son retrieval,” in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2021.

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, “Microsoft coco: Common objects incontext,” in *Proceedings of the European conference on computer vision (ECCV)*, 2014.

[24] Y. Wang, C. Bo, D. Wang, S. Wang, Y. Qi, and H. Lu, “Language person search with mutually connected classification loss,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[25] J. Liu, Z.-J. Zha, R. Hong, M. Wang, and Y. Zhang, “Deep adversarial graph attention convolution network for text-based person search,” in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2019.

[26] T. Chen, C. Xu, and J. Luo, “Improving text-based person search by spatial matching and adaptive threshold,” in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[27] K. Niu, Y. Huang, W. Ouyang, and L. Wang, “Improving description- based person re-identification by multi-granularity image-text align-ments,” *IEEE Transactions on Image Processing*, vol. 29, pp. 5542– 5556, 2020.

[28] Z. Ji, J. Hu, D. Liu, L. Y. Wu, and Y. Zhao, “Asymmetric cross- scale alignment for text-based person search,” *IEEE Transactions on Multimedia(TMM)*, pp. 1–11, 2022.

[29] A. Farooq, M. Awais, J. Kittler, and S. S. Khalid, “Axm-net: Implicit cross-modal feature alignment for person re-identification,” in *Proceed- ings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[30] Y. Chen, G. Zhang, Y. Lu, Z. Wang, and Y. Zheng, “Tipcb: A simple but effective part-based convolutional baseline for text-based person search,” vol. 494, 2022, pp. 171–181.

[31] Z. Wang, A. Zhu, J. Xue, X. Wan, C. Liu, T. Wang, and Y. Li,“Caibc: Capturing all-round information beyond color for text-based person retrieval,” in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2022.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*,“An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[33] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for imagerecognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[34] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, “Scaling up visual and vision-language representation learning with noisy text supervision,” in *International Conference on Machine Learning (ICML)*, 2021.

[35] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu,“Coca: Contrastive captioners are image-text foundation models,” *arXiv preprint arXiv:2205.01917*, 2022.

[36] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *International Conference on Machine Learning (ICML)*, 2022.

[37] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, and J. Huang, “Vision-language pre-training with triple contrastive learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[38] L. Wei, S. Zhang, W. Gao, and Q. Tian, “Person transfer gan to bridge domain gap for person re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[39] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[40] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[41] C. Gao, G. Cai, X. Jiang, F. Zheng, J. Zhang, Y. Gong, P. Peng, X. Guo, and X. Sun, “Contextual non-local alignment over full-scale representation for text-based person search,” in *arXiv preprint*, 2021.

[42] Y. Wu, Z. Yan, X. Han, G. Li, C. Zou, and S. Cui, “Lapscore: language-guided person search via color reasoning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*,2021.

[43] Z. Wang, A. Zhu, J. Xue, X. Wan, C. Liu, T. Wang, and Y. Li, “Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold,” in *Proceedings of the ACM International Conference on Multimedia(ACM MM)*, 2022.

[44] S. Li, M. Cao, and M. Zhang, “Learning semantic-aligned feature representation for text-based person search,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.