

推进可见的边红色人员重新识别：协同视觉文本推理和跨模式特征对齐

Yuxuan Qiu,¹ Liyang Wang,¹ Wei Song,¹ Jiawei Liu,¹ Zhiping Shi² and Na Jiang,¹ Member, IEEE

Abstract - 可见的国际人士重新认同 (VI-REID) 是一项关键的跨模式的分类任务，对公共安全和安全应用具有重要意义。现有的VI-Reid方法主要集中于为人检索提取模态不变特征。但是，由于红外图像中固有缺乏纹理信息，这些模态不变的特征倾向于强调全局文本。因此，经常错误地识别具有相似轮廓的个体，对安全系统和预测调查构成潜在风险。为了解决这个问题，本文创新了自然语言描述，以了解Vi-Reid的全球本地环境。特别是我们设计了一个框架，该框架可以共同优化可见的信号对准加上 (VIAP) 和视觉文本推理 (VTR)，并介绍了本地 - 全球联合措施 (LJM)，以增强度量标准，同时提出一种人类LLM的方法，以将文本描述合并到现有的交叉机构中的人类协作中，并将其融合到现有的交叉机构中。VIAP实现RGB和IR之间的跨模式对齐。它可以明确利用设计的频率感知方式对齐方式和关系增强的融合来探索在全球特征和模态不变信息中本地提示的潜力。VTR提出了汇总选择和双级推理机制，以迫使图像编码器根据文本描述注意重要区域。LJM建议将局部特征距离引入度量阶段度量标准，以增强使用细粒度信息的相关性。对流行的SYSU-MM01和REGDB数据集的广泛实验结果表明，所提出的方法显着超过了最先进的方法。该数据集可在<https://github.com/qyx596/vireid-caption>上公开获取。

I. 简介
Person重新识别 (REID) 是视频监视和取证中的关键技术，可以精确地识别个人跨多个相机视图的个体，以增强安全性监视功能。早期方法[2], [3]利用RGB图像作为输入已在封闭的数据集上实现了出色的性能，在该数据集中，画廊和查询集共享相似的环境。然而，在弱光或夜间条件下捕获的RGB图像无法充分显示行人的外观或轮廓，这对实用现实世界的应用构成了挑战。

为了解决不同照明环境下的识别，并入了红外 (IR) 图像，将REID扩展到可见的 - 红外REID (VI-REID) [4], [5]的跨模式任务。VI-REID的主要挑战在于，红外图像和可见图像之间的固有模态差异，这些图像表现出不同的视觉特征和特征表示。同时，在忽略当地信息的同时，经常将全球构造被捕获为判别性表示，导致了类似的轮廓的错误认同。基于文本的人检索可以通过文本描述启用匹配的人图像，与基于图像的人REID更好地与人类直觉的一致性。它还有效地处理语义上模棱两可的情况。但是，该方法仅支持文本到RGB检索，并且在弱光条件下的表现较差。

Index Terms - 重新认同，方式一致性，关系推理。

2024年7月23日收到；2025年1月5日和2025年1月26日修订；2025年1月27日接受。出版日期，2025年2月11日；当前版本的日期2025年2月24日。这项工作得到了北京自然科学基金会的部分支持。西安·贾蓬大学 (Xi'an Jiaotong University) 的人机混合增强情报的国家主要实验室，根据Grant Hmhai-202407；部分由中国国家纳特科学基金会根据62376166和62476260的赠款；纳瑟尔·达默 (Naser Dameer) 博士是2023年2023年2023年工程研究中心创新研究中心，教育部的数字学习技术综合和应用程序中心。*(Corresponding author: Na Jiang.)*

Yuxuan Qiu, Liyang Wang, Wei Song, Zhiping Shi和Na Jiang在中国首都师范大学的信息工程学院，中国（电子邮件：1201004027@cnu.edu.edu.edu.edu.edu.edu.cn; wly@cnu.edu.edu.edu.cn; jiangna@cnu.edu.cn）。

Jiawei Liu与中国科学技术大学自动化系，中国Hefei 230026（电子邮件：jwliu6@ustc.edu.cn）。数字对象标识10.1109/tifs.2025.3539946

为了弥补这些定义，在本文中，我们建议为VI-REID任务介绍自然语言规范。它的动机来自我们的日常生活。正如图1所述的那样，在实际情况下搜索独特的某些人时，我们不仅提供参考照片，而且还使用语言来描述性别和外观。这些自然的语言描述将帮助我们观察行人的不同地方区域，以确定他们是否是检索人。

自然语言的引入将里德 (Reid) 和vi-reid转化为多模式任务[6], [7]。但是，尽管其他方式在检索信息方面带来了多样性，但它们在模式一致性方面也构成了挑战。为了解决这个问题，我们建议使用VI-REID的恒星固定，并使用联合可见的 - 信号对准加 (VIAP)，视觉文本推理 (VTR) 和局部 - 全球关节措施 (LJM)。VIAP负责RGB和IR之间的跨模式学习

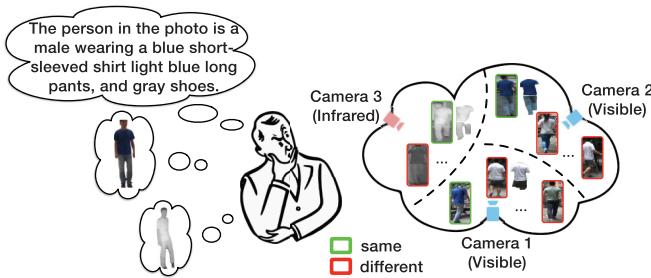


图1。动机的视觉解释。以图片为例，以SYSU-MM01数据集[1]中的个人（ID：0002）以人类的重点进行比较VI-ReID中整个身体的相似性，这使其很难找到相同的身份。但是，在考虑了自然语言描述之后，人类将根据关键字执行本地匹配，这使Vi-Reid更容易。此先验表明，自然语言描述对建立可见光和红外形态之间的全局 - 本地关系是有益的。

没有空间对齐。它利用频感知的模态对齐来同时学习模式间全局特征和模式内全部本地特征，然后通过关系增强融合（ R^2 融合）来实现这些特征的可区分摘要。VTR利用文本语义来参与图像中当地的显着区域。它设计了一种汇总选择机制，以获得可以弥补视觉令牌中缺乏本地信息的辅助特征。然后，它采用双重级别的推理和硬样品挖掘，以隐式指导VIAP通过提供全局局部指导来增强细粒度特征的学习。在测试阶段，LJM进一步优化了初始检索结果。与现有的相似性策略不同[8]，LJM利用局部相似性约束来通过合并细粒度信息来增强匹配项的相关性。

广泛的实验结果表明，在流行的SYSU-MM01和REGDB数据集上，提出的方法显着优于大多数最先进的方法。

总而言之，主要贡献如下：

- 1) 提出了一种新颖的星形架构，该结构将自然语言规格引入VI-ReID，并通过双重级别的推理来增强细粒度的视觉表示。
- 2) 设计使用频率感知的模态对准和增强关系融合的VIAP，从而明确地改善了低频本地信息加强全球线索的模式不变特征。
- 3) 设计具有局部全球关节相似性度量的LJM，利用本地特征进行群体检索，以减少空间和语义不连续性引起的干扰，从而优化排名列表。

与初步版本[9]相比，本文介绍了以下改进。通过合并了来自本地特征的低频信息以增强特征歧视，将VIA模块重新设计为VIAP。同时，采用了局部全球关节措施，并升级了文本编码器，这进一步增强了细粒度的语义提取和利用率，从而在复杂的情况下提高了检索性能。在

此外，添加了建议的数据集的构建方法，并公开提供了访问的链接。这些增强功能可在R1和MAP指标中提高5%。进行了LJM中的不同K设置和加权因子的实验，以验证所提出的组件的值。我们认为，通过将自然语言规范引入Vi-Reid中，可以提高含量的特征提取，以进行模态对齐，扩大检索范式并使与大语言模型或AI代理的集成。

ii. 相关工作

A. Visible-Infrared Person Re-Identification

Reid人最初是一项单模式任务[10]，主要集中在有效地描述RGB图像中的人。诸如AMR [11]之类的著名作品通过利用外观属性来生成更全面的人的描述性特征，而APN [12]动态段人行人图像在不同的捕获条件下促进了本地特征，从而实现了这一点。但是，基于RGB的REID方法在不同的照明条件下面临特征可靠性的挑战，这激发了整合RGB和IR的VI-REID的提议。

VI-REID进一步旨在克服模态差异并学习模态不变的特征，这些特征在可见和红外光谱中具有强大的功能[13], [14], [15]。已经提出了各种方法来弥合RGB和IR模式之间的差距。例如，XIV [16]引入了辅助X模式，以促进跨模式学习。WRIM-NET [17]被指导通过增强模态之间的局部区域相互作用来提取模态不变信息。Deen [5]采用跨模式嵌入模块来提取跨模态共享的犯罪性特征。RLE [18]模拟了在不同模态下材料表面的局部线性变换，以缩小模态间差异的数据。PMWGCN [19]利用小波变换来抑制特征中非平稳的高频噪声，从而改善了图像降解场景中的性能。Sgiel [20]利用人类姿势估计作为提取模态不变特征的指导提示，利用RGB和IR图像中人类轮廓的结构一致性。

与这些侧重于找到合并空间的方法相反，一些研究人员认为，应通过增强跨模式约束来改善特征不同的方法。为此，Maum [21]提出了一个基于内存增强的度量，以通过维护动态特征内存库来抑制模态差异。DART [22]设计了一个置信限制，以减轻模态噪声，从而根据每种模式的可靠性来调整每种模式的贡献。PartMix [4]进一步结合了模态熵约束和数据增强技术，以改善学习特征的可区分性和泛化。IDKL [23]通过从模态特征中蒸馏出隐式信息来增强可区分性。

生成对抗网络（GAN）还显示了VI-REID任务的显着改进[24]，

[25], [26], [27]。这些方法通常利用可见图像和红外图像之间的传输学习来生成跨模式数据。例如，FMCNET [28]使用GAN生成的虚拟模式来补偿原始模态信息并实现特征融合。Aligngan [29]通过对抗性学习使像素级别的模态对齐，而CMGAN [30]学习了模态转换的模式以进行跨模式图像训练，从而有效地减少了域间隙。此外，GML [31]设计了一种生成的度量学习方法，该方法通过生成对抗样本并将其标准化以产生强大的距离指标来改善性能。

尽管上述方法在绩效方面取得了巨大的成功，但它们仍然受到红外图像的固有限制的限制：隐藏局部细粒度线索。这种局限性导致学习模态不变的特征，这些特征主要强调全局内容，例如整体轮廓或粗糙的结构，同时忽略了对诸如人类重新识别的精细分类任务至关重要的局部区域。例如，在红外图像中通常会遮盖或不存在歧视性属性，例如配件，服装图案或步态特征。尽管甘斯（Gans）在有限的先验方面挣扎，但通常会产生不切实际和不可靠的产生的图像，这些图像严重胜任算法的发电，以使其无法看到。

为了解决这些问题，本文介绍了由人-GPT-4合作[32]生成的自然语言规范，以指导全球和本地注意力模式。这种方法背后的理由是，自然语言描述可以提供有关明显的本地属性（例如“戴红帽”或“携带背包”）的丰富人性化信息，这些信息可能在红外图像中不明显。

B. Text-Based Person Retrieval

目前，尚无现有的研究重点是使用自然语言来增强可见的红外重新识别（VI-REID）。但是，随着大型模型和多模式研究的兴起，基于文本的人检索（TPR）已成为研究热点。TPR，最初由Li等人于2017年提出。[33]是一个检索任务，可以根据文本描述搜索行人，可以视为一种交叉文本和视觉内容挖掘[34]。TPR中的一个显着挑战是有效地对齐两个不一致的编码空间。早期的效果主要利用了来自各自域的建立模型，例如Visual Realm中的VGG [35]和Resnet50 [36]，以及NLP域中的LSTM [37]和BERT [38]，以学习可靠的代码代码。在随后的模态对齐中，这些效果通常从全球环境的角度开始[39]，演变成各个不同粒度的自适应局部语义学习[40], [41]。有效的语义信息包括颜色[42]，性别，外观[43]和其他属性。

为了增强语义信息，视觉方式预处理模型已引入TPR [44], [45], [46]。例如，著名的对比语言图像预训练（剪辑）模型[47]，在广泛的文本图像对库中进行了培训，以强调潜在的潜力

模态对准。Yan等。[46]设计了一个框架，该框架利用剪辑模型提取细微的信息，推进基于文本的人检索的任务。该框架最佳地利用了许多图像文本对中嵌入的庞大知识，从而增强了跨模式转移学习。Jiang and Ye [7]引入了IRRA模型，这是原始剪辑模型的演变，从而可以获取更多敏感的文本图像嵌入。随后，Bai等。[48]提出了一种关系和敏感度感知的表示方法，该方法使用蒙版语言建模（MLM）来预测掩盖的令牌，获得了更具歧视性和信息性的特征。

上述方法中提到的各种跨模式一致性或推理策略在TPR中取得了巨大的成功。但是，这些方法仅学习可见图像和自然语言之间的模态不变特征。两种方式都包括细粒度的信息，通过语义提示促进了全球本地对齐。在这项工作中，我们将自然语言介绍给Vi-Reid，这意味着语言，可见和红外模式共存。红外图像中的细胞信息是潜在的，这使得与红外图像直接使自然语言保持一致。在先前的经验的基础上，我们通过两个跨模式学习过程（可见的红外和视觉语言）共同训练这三种方式。这种方法旨在通过利用每种方式的优势来有效地弥合方式之间的差距，并提高VI-REID的性能。

iii. 方法

在本节中，我们介绍了拟议的星级架构，并详细介绍了其组件，并描述了使用Human-LLM协作来增强具有自然语言描述的VI-REID数据集的方法。如图2所示，可见的图像 I_{RGB}^i ，红外图像 I_{IR}^i 和自然语言描述 T^i ，同一个人ID形成三重输入。给定培训集 \mathcal{X} 包含 M 三重样本 $(I_{RGB}^i, I_{IR}^i, T^i)$ 及其相应的IDS GT i ，Star-Reid的目的是利用自然语言描述 T^i 来指导Vi-Reid的区分特征提取。 I_{RGB}^i 和 I_{IR}^i 通过ViT [49]的共享变压器层提取功能。文本分支使用预读的BERT [38]作为骨干网络提取特征。但是，红外功能缺乏明显的细粒线索，因此很难直接与自然语言或执行关系推理。因此，Star Reid设计VIAP可以直接对齐可见的RGB和红外红外模式，然后通过VTR引入文本，以间接实现多模式学习。最后，LJM可以选择测量精细特征的相似性来优化排名列表。

与仅学习模态特征的现有方法不同，VIAP采用新颖的频率感知模态对准，以同步学习模式间全局效果和模式内部全局本地特征。同时， R^2 融合可以自适应地实现这些特征的可区分摘要。此外，VTR使用CLS令牌

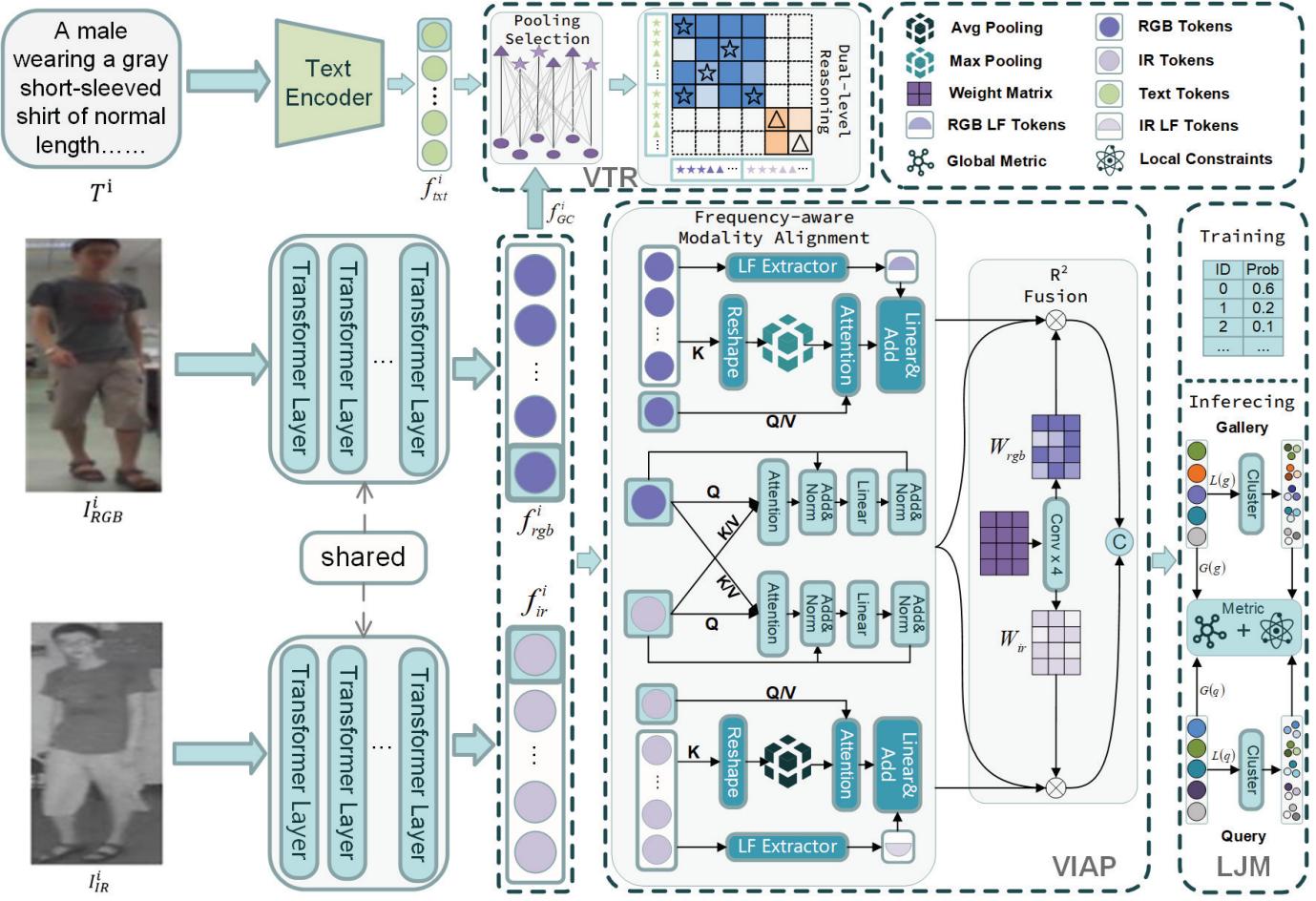


图2。拟议的恒星里德的整体体系结构。它是带有VIAP和VTR的三个分支结构。RGB和IR输入被输入重量共享的变压器层，以学习模态不变特征，然后通过监督的VIAP实现视觉对齐。在此基础上，VTR利用了汇总选择和双级推理，以增强视觉特征对局部细粒线索的关注。在推理阶段，可以选择使用LJM来通过使用局部细粒信息来扩展测量维度，从而提高结果的准确性。最好的颜色观看。

和双层推理的辅助功能。通过合并选择提取了三种模式的贴片令牌的辅助特征。通过这种方式，Star-Reid可以通过Textual Guidance在视觉方式中增强细粒度信息。在以下小节中说明了更多详细信息。

A. Visible-Infrared Alignment Plus

由于RGB和IR图像之间缺乏空间一致性，因此现有方法提取的模态不变特征通常包含全局信息。这导致许多局部细粒度的特征被忽略。同时，在红外模式中，低频信息包含有助于识别的关键提示。为了应对这些问题，可以在VIAP中设计频率感知的模态对准和 R^2 融合可以挖掘本地显着性线索。

以三重输入 (I_{RGB}^i, I_{IR}^i, T^i) 为例，基本可见特征 f_{rgb}^i 和红外功能 f_{ir}^i 是由带有共享权重的骨架提取的。然后将每个基本功能分解为CLS令牌 $G(f_{rgb/ir}^i)$ 和一个补丁令牌 $L(f_{rgb/ir}^i)$ 。使用CLS代币作为输入，频率吸引的方式对齐首先学习模式间全局特征

$f_{Grbg/ir}^i$ 由 (1) 和 (2)。

$$f_{Grbg}^i = \text{Softmax} \left(\frac{(G(f_{rgb}^i)G(f_{ir}^i)^\top)}{\sqrt{d_{ir}}} \right) G(f_{ir}^i) \quad (1)$$

$$f_{Gir}^i = \text{Softmax} \left(\frac{(G(f_{ir}^i)G(f_{rgb}^i)^\top)}{\sqrt{d_{rgb}}} \right) G(f_{rgb}^i) \quad (2)$$

在 (1) 和 (2) 中， $d_{rgb/ir}$ 表示 $G(f_{rgb/ir}^i)$ 的维度。通过此计算，获得的全局功能 f_{Grbg}^i 和 f_{Gir}^i 可以提高模式间相似性。在模态对齐中，这些与原始CLS令牌同样至关重要。

此外，考虑到A训练样本中的 I_{RGB}^i 和 I_{IR}^i 都不能通过同一设备同时捕获 I_{RGB}^i 和 I_{IR}^i ，也应学习模式内的功能。频率感知的模态对准将基本功能 $f_{rgb/ir}^i$ 作为输入，计算模式内全部本地特征 f_{GLrgb}^i 和 f_{GLir}^i 根据 (3) 和 (4)。

$$f_{GLrgb}^i = \text{Softmax} \left(\frac{(G(f_{rgb}^i)RMax(L(f_{rgb}^i))^\top)}{\sqrt{d_{Lrgb}}} \right) G(f_{rgb}^i) \quad (3)$$

$$f_{GLir}^i = \text{Softmax} \left(\frac{(G(f_{ir}^i)RAvg(L(f_{ir}^i))^\top)}{\sqrt{d_{Lir}}} \right) G(f_{ir}^i) \quad (4)$$

在这一部分中，CLS代币充当查询和价值，而补丁令牌则充当关键。 $RMax()$ 表示重塑后的最大池， $RAvg()$ 表示重塑后的平均池， $d_{Lrgb/Lir}$ 表示池池后 $L(f_{rgb/ir}^i)$ 的维度。由于本地向量 $L(f_{rgb/ir}^i)$ 和全局令牌 $G(f_{rgb/ir}^i)$ 之间的尺寸差异，因此需要一个重塑操作才能将 $L(f_{rgb/ir}^i)$ 从 $D \times HW$ 更改为 $D \times H \times W$ 。然后在合并操作中进行重塑向量。对于可见图像 I_{RGB}^i ，采用最大池来突出显示歧视性颜色或纹理细节。对于红外图像 I_{IR}^i ，平均池用于合并健壮的形状上下文。通过这种方式，所达到的 f_{GLrgb}^i 和 f_{GLir}^i 可以反映全球 - 本地关系，从而促进模态对齐。

为了进一步增强特征歧视，尤其是在红外模式中，提取局部低频信息以促进细粒度的跨模式比对，如(5)和(6)，(6)，(6)，(6)，(6)

$$c_{rgb}^i = H_l(L(f_{rgb}^i)) \quad (5)$$

$$c_{ir}^i = H_l(L(f_{ir}^i)) \quad (6)$$

其中 H_l 表示Haar小波低频提取器，而 $c_{rgb/ir}^i$ 表示相应的低频令牌。

随后，低频代币被重塑并分配到平均或最大池，然后是单独的投影仪，以增强(7)和(8)，(8)，(8)，(8)，(8)，(8)，(8)，(8)，(7)和(8)，

$$f_{GLFrgb}^i = Proj_{rgb}(RMax(c_{rgb}^i)) + f_{GLrgb}^i \quad (7)$$

$$f_{GLFir}^i = Proj_{ir}(RAvg(c_{ir}^i)) + f_{GLir}^i \quad (8)$$

其中 $Proj_{rgb/ir}$ 表示低频功能的线性投影仪， f_{GLFrgb}^i 和 f_{GLFir}^i 表示通过低频信息增强的功能。

目前，VIAP可以获得原始的全局特征 $G(f_{rgb/ir}^i)$ ，模式间特征 f_{Grgb}^i 和 f_{Gir}^i ，内模态特征 f_{GLFrgb}^i 和 f_{GLFir}^i 。它们由 R^2 融合自适应地融合，而权重 $W_{rgb/ir}$ 则是从星形修复中学到的。如图2的子图中所示，第一次初始化 $2 \times 3 \times 768$ 一个矩阵，然后执行4层 1×1 卷积的层以获得适合RGB和IR模态的权重。Fusion特征 f_{gl}^i 的计算被定义为

$$\begin{aligned} f_{gl}^i &= Cat(Sum(W_{rgb}^0 \cdot f_{Grgb}^i, W_{rgb}^1 \cdot G(f_{rgb}^i), W_{rgb}^2 \cdot f_{GLFrgb}^i), \\ &\quad Sum(W_{ir}^0 \cdot f_{Gir}^i, W_{ir}^1 \cdot G(f_{ir}^i), W_{ir}^2 \cdot f_{GLFir}^i)) \end{aligned} \quad (9)$$

其中 $Cat(a, b)$ 表示 a 与 b ， $Sum(a, b, c)$ 代表matrix添加。集成的 f_{gl}^i 反映了跨模态和全局 - 本地对齐，从而有效地改善了模态不变特征的歧视。表I总结了上面使用的符号。

B. Visual-Textual Reasoning

在检索任务中，包括性别或服装在内的自然语言描述可以通过本地显着线索来快速找到对象范围。因此，本文介绍了自然语言规范以改善Vi-Reid，这也是

表I符号摘要

Notation	Description
I_{RGB}^i	visible image
I_{IR}^i	infrared image
T^i	natural language description
f_{rgb}^i	visible feature
f_{ir}^i	infrared feature
$f_{Grgb/Gir}^i$	inter-modality global feature
$d_{rgb/ir}$	dimension of inter-modality global feature
$f_{GLrgb/GLir}^i$	intra-modality global-local feature
$f_{GLFrgb/GLFir}^i$	low-frequency strengthened global-local feature
$d_{Lrgb/Lir}$	dimension of intra-modality global-local feature
$c_{rgb/ir}^i$	low-frequency component
$W_{rgb/ir}$	fusion weights
$G(f)$	extract CLS token from f
$L(f)$	extract patch tokens from f
$RAvg(f)$	reshape and apply average pooling to the feature f
$RMax(f)$	reshape and apply max pooling to the feature f
$H_l(f)$	extract low-frequency token from f
$Proj_{rgb/ir}(f)$	extract projected features from f

同时处理RGB、IR和文本的第一个重新认同工作。但是，现有的VI-Reid数据集缺乏文本描述。为了解决这个问题，我们将统一属性（例如，包，衣服）的图像进行粗略注释，并以JSON格式上传到GPT-4中，以生成标准化的语言描述。

使用上述自然语言规范指导视觉图像以不可避免地提取细粒度的特征，需要在视觉和语言之间进行跨模式学习。考虑到这两种方式之间的显着差异以及红外模式中缺乏明显的细粒度信息。建议VTR在文本 f_{txt}^i 和 f_{rgb}^i 和 f_{ir}^i 的视觉融合之间进行关系推理。VTR包含汇总选择和双级推理。池选择从patch tokens提取辅助功能ptokens $P(f_{rgb/ir/txt}^i)$ 从通道max boming中提取，该功能提供了用于关系推理的本地特征。与仅取决于CLS令牌的现有推理方法不同，双级推理利用CLS令牌和PTOKENS来对全局和本地视觉监督进行监督。这可以隐式地帮助VIAP增强纤维性特征的学习。

在视觉监督中，全局视觉到文本的关系推理采用了(10)，(10)中的成功对比损失 L_{V2T}^G

$$L_{V2T}^G = -\log \frac{\exp(S(f_{GC}^i, G(f_{txt}^i)))}{\sum_{k=1}^N \exp(S(f_{GC}^i, G(f_{txt}^k)))} \quad (10)$$

其中 f_{GC}^i 请参阅 $Cat(G(f_{rgb}^i), G(f_{ir}^i))$ ， $S(a, b)$ 计算 a 和 b 之间的相似性， N ， N 表示批次。本地视觉到文本损失 L_{V2T}^L 只需要用Ptokens $P(f_{rgb/ir/txt}^i)$ 替换(10)中的CLS令牌。

在计算文本到视觉损失时，我们发现批次中的任何三重输入都可能具有多个阳性。这意味着相同的描述 T^i 可能对应于不同的 (I_{RGB}^i, I_{IR}^i) 。面对这些积极因素，视觉和语言之间的推理越好，对比度损失越小。VTR应该更加注意具有很大错误的训练样本，因为它们很难在文本和视觉方式之间提示。因此，推理能力 w_p^i 旨在将对比损失修改为 L_{T2V} 。以全局约束为例，相关计算在(11)和(12)中定义

$$w_p^i = \frac{L_C(f_{GC}^i, G(f_{txt}^i))}{\sum_{p \in P(T)} L_C(f_{GC}^p, G(f_{txt}^i))} \quad (11)$$

$$L_{T2V}^G = -w_p^i \log \frac{\exp(S(f_{GC}^i, G(f_{txt}^i)))}{\sum_{k=1}^N \exp(S(f_{GC}^k, G(f_{txt}^i)))} \quad (12)$$

其中 L_C 是指标准对比损失， $P(T)$ 是批处理中 T^i 的所有阳性索引集， w_p^i 代表了标准化的重量，它反映了对VTR更新的影响。使用 $P(f_{rgb/ir/txt}) L_{T2V}^L$ 也适用相同的计算。他们共同努力帮助VTR实施艰难的示例采矿。视觉监督的总损失 L_{VLS} 被定义为(13)，

$$L_{VLS} = L_{V2T}^G + \alpha L_{V2T}^L + L_{T2V}^G + \alpha L_{T2V}^L \quad (13)$$

其中 α 根据经验设置为0.3。然后将对象函数 L_{OB} 总结为，

$$L_{OB} = L_{id} + L_{tri} + L_{VLS} \quad (14)$$

其中 L_{tri} 代表带有 f_{gl}^i 的三重态损耗函数， L_{id} 是使用 $G(f_{rgb/ir}^i)$ 的跨熵损失函数的总和。他们共同负责更新Star-Reid。

C. Local-Global Joint Measure

除了对特征部分中的本地信息充分关注外，在检索阶段还使用了全球和本地特征的联合度量。

在此阶段，给出了一个人图像 q ，并给出了包含大量图像的图库 g ，此阶段的目的是将大量图像与类似于所选的 q 的人映像匹配。

常见的检索策略是将图像输入训练的模型中，以提取功能并根据 q 和 g 的特征相似性获取检索列表。在单峰重新认可任务中，有必要解决由整体相似性引起的识别错误问题，而不是局部特征相似性。在跨模式检索中，这种缺乏对当地细节的关注的影响可能更大。

因此，在跨模式的重新识别任务中，有必要进一步跨越模态差异带来的特征相似性度量的影响，并进一步优化排名策略。图3显示了排名列表。

我们利用联合全球-本地优化思想。基于全局特征和汇总本地特征的相邻样本相似性分别获得，然后以A的形式将两个样本相似性汇总并融合



图3。具有和没有LJM的视觉比较。第一列是查询图像。第2-11列是前十名结果。绿色框表示与地面图的一致性，而红色框表示错误。黄色标记表明本地提示在LJM中发挥作用。最好的颜色观看。

距离矩阵以获得一个关节约束距离矩阵进行排名检索。

距离矩阵是通过相似性在全球范围内获得的，并且总体姿势被用作确保识别精度的约束。对于局部特征，红外图像和粘性图像之间的模态差异导致模态不平衡问题。由于可见特征的过多语义信息，即使在执行特征指标时相同的身份，红外功能也可能无法准确匹配。为了解决由模态差异引起的这种检索失衡，我们为两种模态的本地特征执行了合并操作，以减少可见特征的复杂背景噪声的效果，同时增强红外功能中包含的人的前景信息。

因此，特定描述的本地特征的约束操作如下：我们将平均汇总应用于从查询和画廊中提取的本地特征，并使用k-重视方法来获取图库的K-Nearest邻居，并根据汇总功能获得基于汇总的图像。最后，将这两个矩阵与特定的权重相结合并融合在一起，以实现我们的联合全部本地测量策略，这仅在推理阶段使用。该策略在内，包括本地约束，可以定义为：

$$D_{Joint} = D_{Global} + \beta R_{Local} \quad (15)$$

$$R_{Local} = 0.3 \cdot \Delta_{Ja}(q_l, g_l, K) + 0.7 \cdot \Theta(q_l, g_l) \quad (16)$$

其中 D_{Joint} 表示关节度量约束， D_{Global} 表示全局度量， R_{Local} 表示局部约束，用于代表全局和局部特征的权重因子的参数 β 。 Δ_{Ja} 表示使用本地功能来计算jaccard距离， Θ 是余弦距离， q_l, g_l 表示局部特征的查询和画廊。 K 表示基于距离选择的样本数量，经验设置为10。通过 Δ_{Ja} ，使用本地特征启用了样本扩展，将单一对单一的相似性测量转换为多to-multi组检索。

在VI-REID中使用本地特征有助于解决由遮挡或观察点变化引起的空间和语义不一致之处，作为代表性的局部特征，可以更好地判别能力，并提供准确的参考信息。此外，基于组的度量优化在复杂方案中增强了性能。

如图3所示，第一行和第三行是初始的VI-REID结果，而第二行和第四行是通过LJM的结果。从这些结果来看，很明显，诸如背包，帽子和水瓶之类的局部特征以及全局特征都被考虑到LJM中。LJM通过全球本地相似性改善了初始结果。

Algorithm 1 STAR-ReID Processing Procedure

```

Input: Triple input ( $I_{RGB}^i, I_{IR}^i, T^i$ ) with the same ID
Parameter: Textual branch  $\mathcal{T}_{txt}$ ,  $I_{RGB}^i$  and  $I_{IR}^i$  use shared visual
branch  $\mathcal{T}_{img}$ 
1: Initialization part of  $\mathcal{T}_{img}$  from ViT
Initialization part of  $\mathcal{T}_{txt}$  from BERT
2: while in train stage do
3: Extract  $f_{rgb/ir}^i = \mathcal{T}_{img}(I_{RGB/IR}^i)$ ,  $f_{txt}^i = \mathcal{T}_{txt}(T^i)$ 
4: Calc. inter-modality global features  $f_{Grb}^i$  and  $f_{Gir}^i$ 
according to (1)–(2)
5: Calc. intra-modality global-local features  $f_{GLrb}^i$  and
 $f_{GLir}^i$  according to (3)–(4)
6: Calc. low-frequency strengthened intra-modality fea-
tures  $f_{GLFrb}^i$  and  $f_{GLFir}^i$  according to (5)–(8)
7: Achieve fusion feature  $f_{gl}^i$  and realize frequency-aware
modality alignment according to (9)
8: Achieve PTokens  $P(f_{rgb/ir/txt}^i)$  and conduct visual-
textual dual-level reasoning according to (10)–(13)
9: Optimize  $\mathcal{T}_{txt}, \mathcal{T}_{img}$  according to (14)
10: end while
11: while in test stage do
12:   if mix-retrieve then
13:     Calc. the inner product between  $G(f_{txt}^i)$  of probe
and  $f_{gl}^i$  of gallery
14:     Sort by inner product and retain the Top- $N$  of the
gallery
15:     Achieve results with  $f_{gl}^i$  of retained gallery and
probe
16:   end if
17:   if local-global joint measure then
18:     Achieve distance with  $RMax(L(f_{rgb}^i))$  and
 $RAvg(L(f_{ir}^i))$  according to (15)–(16)
19:   end if
20:   Achieve results with  $f_{gl}^i$  of gallery and probe
21: end while
```

最终，所提出的Star-Reid的整个过程与上述方程相结合，在算法1中逐步证明。在训练阶段，使用VIAP和VTR来实现可区分的模态不变特征。VIAP进行模态对准，这在第3-6行中进行了描述。VTR结合了自然语言规范，以增强对局部细粒度特征的关注。它的过程在第7-8行中显示。（14）中定义的 L_{OB} 负责更新参数。

在测试阶段，这项工作提供了两种检索模式。一个是VI-REID中的标准相似性测量值，另一个是带有自然语言规范的混合式测量。对于第二种模式，我们首先计算探针的文本令牌 $G(f_{txt})$ 的内部产物，并将融合功能 f_{gl} 特征 f_{gl} ，然后保留顶部 - N 画廊图像。在此基础上，使用探针的 f_{gl} 和保留画廊实现余弦距离。文本描述可以有效地缩小搜索范围，这对于在实际场景中促进星座至关重要。

D. Human-LLM Collaborative Natural Language Description Generation

由于在现有的可见红外重新标识数据集（例如SYSU-M M01和REGDB）中没有自然语言规格，因此我们在行人ID级别注释了这些数据集。

仅依靠手动注释需要不同的注释人员来确保一致的描述规则，这是耗时且艰苦的。很难通过直接使用现有的图像到文本生成算法来获得REID数据集中裁剪图像的准确描述。背后的原因是这些算法缺乏客观的指导，无法执行有利于REID任务的解析和语言组织。

为此，我们执行两阶段的注释。通过手动注释为基于算法的自动注释提供语义取向，同时通过自动发电算法降低人工和时间成本。

在手动注释中，粗糙标签包括11个尺寸：性别，顶部色，顶部类型，顶长，裤子颜色，裤子类型，袋子类型，帽子类型，眼镜，鞋类，鞋子颜色和音符。“注释”维度用于描述特殊模式，配件和项目，以增强标题描述的能力。此标签信息将使用相应的维数名称作为生成GPT-4的详细说明的密钥以结构化的JSON格式存储。

在GPT-4一代中，先前注释的粗标签被用作GPT-4提示的一部分，该提示旨在产生更光滑的自然语言描述。这回合促进了随后的文本编码在语义上监督模型的情况下。提示如下：

A photo taken by a surveillance camera shows an individual, whose attire characteristics will be described using a structured JSON data as follows: [JSON]. Please succinctly describe the person in the photo in English. In the JSON, the ‘top color’ field represents the color of this person’s clothing; a field value of ‘none’ indicates that he is not carrying corresponding items; The “top type” field could be either ‘short sleeve’ or ‘long sleeve’; The “top length” field refers to its clothing length, which can be either ‘normal’ or ‘long’; The ‘pants color’ field is his pants’ color. The description should start with ‘The person in the photo is’.

[JSON]的位置，它将被特定的JSON数据替换。重要的是要注意，GPT-4的温度参数需要调整到0.6左右，以防止其自动填写图像中不存在的细节。同时，由于有限

表II标题的分布

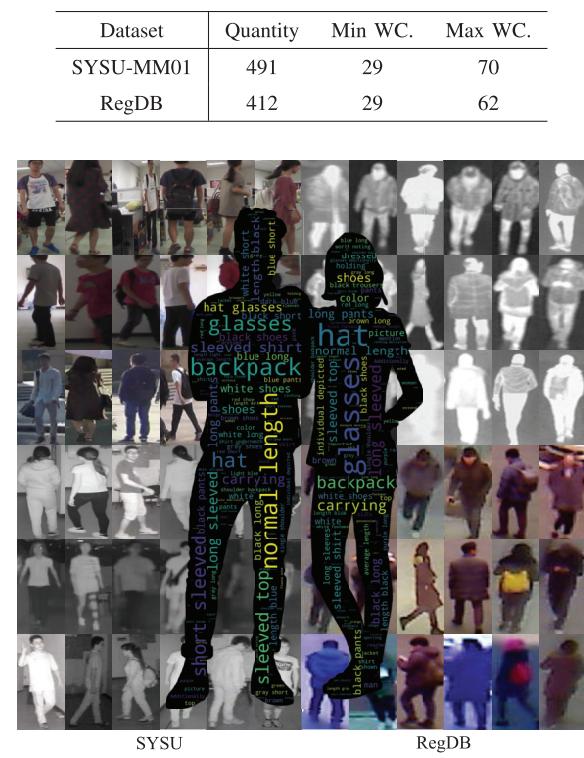


图4。SYSU-MM01和REGDB数据集上自然语言描述的单词云分布。最好的颜色观看。

详细的维度，在某些情况下，不同的ID与一代结束后相同的字幕对应。在这种情况下，需要手动修改字幕才能进行区分。这些措施有助于避免通过文本语义信息对网络进行错误的监督。

iv. 实验

A. Dataset and Implementation Details

为了验证Star-Reid的效率，我们在两个基准数据集上进行了实验：SYSU-MM01 [1]和REGDB [50]。SYSU-MM01数据集包含30,071个可见的和15,792个红外图像，其中包括491个独特的行人。相比之下，REGDB数据集包含412个身份，每个身份以10个可见和10个红外图像表示。图5提供了这些数据集的代表性示例。

表II阐明了两个数据集的身份级别上自然语言描述的分布。值得注意的是，单词计数存在很大的差异，这是归因于某些场景（例如条纹衣服，携带的物体）的复杂性，需要更详细的描述。图4可视化字幕的云分布，分别代表sysu-mm01和regdb的左右半部分。

值得注意的是，与REGDB相比，SYSU-MM01提出了更具挑战性和权威的基准。所有报告的结果在后续表中代表平均



图5。来自SYSU-MM01数据集的样本。最左侧的部分是描述文本，ID人的相应轮廓图像是从cam1到CAM6从左到右。最好的颜色观看。

十个独立的实验运行。我们采用等级-1准确性 (R1) 和平均平均精度 (MAP) 作为我们的评估指标。

所提出的Star-Reid框架是使用Pytorch实施的，并在单个NVIDIA RTX3090TI GPU上进行了训练。我们利用了64个批量的大小，并训练了150个时期的型号。优化过程与余弦退火学习率调度程序一起采用了ADAMW优化器。初始学习率设置为 3×10^{-4} ，重量衰减为 1×10^{-4} 。

B. Comparison With State-of-the-Art Methods

表III和表IV分别显示了SYSU-MM01和REGDB数据集上的未出色算法的比较。SYSU-MM01数据集涉及全搜索和室内搜索设置，而REGDB则在红外评估为可见 (i to v)，并可见红外线 (v to i)。最好的和第二好的结果以粗体和下划线突出显示。

在表III中，我们提出的Star-Reid方法在两个搜索设置中都表明了卓越的性能。如果没有LJM模块，Star-Reid在全搜索设置中的R1为79.92%，地图为77.23%，R1的R1为86.74%，室内搜索模式中的R1和地图为88.12%，已经超过了先前的先前尚未达到的现状方法。另外[17], [18]通过以局部特征丰富特征表示来实现竞争结果，而所提出的方法通过全球本地对准进一步增强了性能。随着LJM模块的添加（表示为星条期（我们的）），全面搜索的R1进一步提高到R1的82.93%，地图为80.47%，R1的R1为88.04%，Indoor-Search的地图为89.58%。这种对以前的方法和我们的显着改善

了比较

表III与SYSU-MM01数据集上的最新方法进行

Methods	Venue	SYSU-MM01 [1]			
		All Search		Indoor Search	
		R1	mAP	R1	mAP
FMCNet [28]	CVPR'22	66.34	62.51	68.15	74.09
DART [22]	CVPR'22	68.72	66.29	72.52	74.94
MAUM [21]	CVPR'22	71.68	68.79	76.97	81.94
CMT [51]	ECCV'22	71.88	68.57	76.90	79.91
CIIFT [52]	ECCV'22	74.08	74.79	81.82	85.61
MSCLNet [53]	ECCV'22	76.99	71.64	78.49	81.17
CAJ+ [54]	TPAMI'23	71.48	68.15	78.36	81.98
ProtoHPE [55]	MM'23	71.92	70.59	77.81	81.31
DEEN [5]	CVPR'23	74.70	71.80	80.30	83.30
PartMix [4]	CVPR'23	77.78	74.62	81.52	84.38
PMWGCN [19]	TIFS'24	66.82	64.88	72.64	76.19
DARD [27]	TIFS'24	69.33	65.65	77.21	81.91
MFCS [56]	TMM'24	70.59	67.49	75.98	80.24
RLE [18]	NIPS'24	75.40	72.40	84.70	87.70
WRIM-Net [17]	ECCV'24	77.40	75.40	86.20	88.10
J-ReID [9]	ICME'24	77.83	75.68	82.52	85.72
STAR-ReID (w/o LJM)		79.92	77.23	86.74	88.12
STAR-ReID (Ours)		82.93	80.47	88.04	89.58

方法的比较

表IV与REGDB数据集上的最新

Methods	Venue	RegDB [50]			
		I to V		V to I	
		R1	mAP	R1	mAP
DART [22]	CVPR'22	81.97	73.78	83.60	75.67
MSCLNet [53]	ECCV'22	83.86	78.31	84.17	80.99
MAUM [21]	CVPR'22	86.95	84.34	87.87	85.09
FMCNet [28]	CVPR'22	88.38	83.86	89.12	84.43
CAJ+ [54]	TPAMI'23	84.88	78.55	85.69	79.70
PartMix [4]	CVPR'23	84.93	82.52	85.66	82.27
ProtoHPE [55]	MM'23	88.69	81.99	88.74	83.72
MFCS [56]	TMM'24	83.88	75.16	85.34	76.39
DARD [27]	TIFS'24	85.53	85.09	86.19	85.39
PMWGCN [19]	TIFS'24	88.77	81.61	90.61	84.53
J-ReID [9]	ICME'24	89.37	87.63	89.42	87.64
STAR-ReID (w/o LJM)		89.52	88.17	90.69	89.57
STAR-ReID (Ours)		90.48	91.77	91.89	93.31

自己没有LJM的基线证明了我们提出的方法的有效性。

在表IV中，我们提出的方法显着胜过地图上提到的所有作品。这些进一步表明，VIAP和VTR具有有效改进的模态 - 不变特征。

特别值得注意的是，在竞争性的PartMix [4]中也探讨了当地的暗示。这证明局部细粒度的对齐在Vi-Reid中起着重要作用。同时，这间接表明，使用文本注意本地明显区域比现有的本地信息探索更合理。

此外，PMWGCN [19]还展示了竞争结果，因为它类似地包含了类似于拟议的星条期的低频信息。但是，该提出的方法进一步探讨了局部特征内的低频信息，从而减轻了空间和语义不一致。最后，提出的VTR，VIAP和LJM的密切集成最佳结果。

C. Ablation Study

在本节中，进行一项消融研究以评估恒星中的每个组件的贡献，如表V所示。“基础”代表PMT [57]，这是我们提出的方法的基线。“FMA”是指频率吸引的方式对齐。

与基线相比，用局部低频信息提取器引入VIAP进一步增强了特征可区分性，R1和MAP分别增加了7.59%和9.27%。LJM的添加进一步提高了性能，R1达到76.92%，MAP 75.99%。

纳入文本特征提取的BERT还显示出显着的改进，将R1增加到72.09%，并将其映射到70.78%。当将VTR和VIAP结合时，显着提高了性能，在R1中获得79.92%和77.23%

SYSU-MM01数据集的表V 消融研究		
Settings	R1	mAP
Base [57]	67.53	64.98
Base+FMA	74.68	73.09
Base+VIAP	75.12	74.25
Base+VIAP+LJM	76.92	75.99
Base+BERT	72.09	70.78
Base+VTR	72.93	72.21
Base+VTR+VIAP	79.92	77.23
STAR-ReID (Ours)	82.93	80.47
STAR-ReID (Mix-retrieve)	83.11	80.93

SYSU-MM01数据集上LJM中不同K值的结果
表VI结果

K	mAP	mAP INC. (%)	Duration (ms)	Duration INC. (%)
0	77.23	0%	67445	0%
5	78.16	1.20%	69367	2.85%
10	80.47	4.20%	69610	3.21%
20	80.39	4.09%	70069	3.89%
30	79.34	2.73%	70702	4.83%

regdb数据集LJM中不同K值的表VII结果

K	mAP	mAP INC. (%)	Duration (ms)	Duration INC. (%)
0	88.17	0%	67923	0%
5	90.03	2.11%	69905	2.92%
10	91.77	4.08%	70118	3.23%
20	91.22	3.46%	70653	4.02%
30	90.15	2.25%	71349	5.04%

在地图中。这表明频率感知的模态对准和视觉文本推理对于我们的方法都是必不可少的。

我们的完整方法（星星雷德）以82.93%的R1和80.47%的地图实现了最先进的性能。这表明每个组件都是必不可少的，并有助于该方法的整体效果。

此外，在表IV的最后一行中显示了使用探针文本的图库的混合回答结果。在检索模式下的顶部N仅设置为画廊项目总数的50%。很难发现，使用自然语言规范可以将R1进一步提高到83.11%，并映射到80.93%。在带有大量画廊数据的实际情况下，混合检索将通过缩小搜索范围而变得更有效。

D. Hyperparameter Analysis

为了评估LJM的值，在所有搜索模式下以SYSU-MM01测试集进行了不同的K设置和时间消耗的实验，并以红外为单位模式进行了REGDB测试集，如表VI和表VII所示。

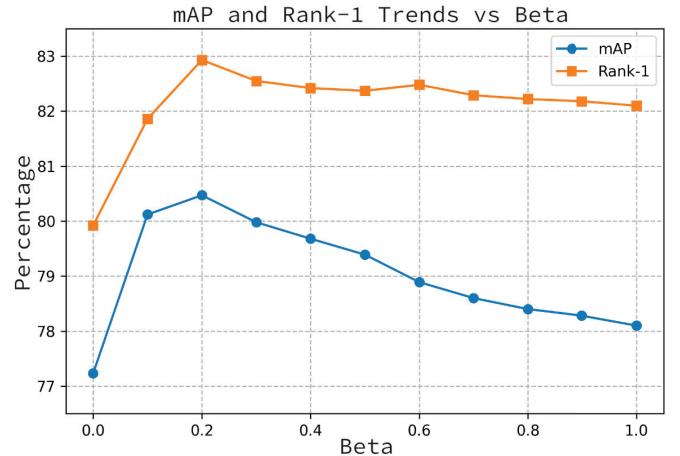


图6。全局和局部特征关节测量的加权因子。蓝线代表地图，橙色线表示等级-1。 $\beta \in [0.0, 1.0]$ 。最好的颜色观看。

如表中所述，具有不同K值的LJM通常会增强MAP，这是因为其在构造组检索中使用局部特征，该特征在测量过程中提供了参考信息。但是，随着K的增加，计算成本也上升，因此，基于应用程序方案和画廊的大小选择适当的K至关重要，以平衡性能和效率。

例如，如表VI中所示，当K = 10相对于原始结果，精度提高了4.20%，而推理时间则增加了3.21%。在这一点上，时间和准确性达到了有利的成本效果比。在大规模的场景中，结合文本描述以进行混合检索可以在应用LJM之前先缩小画廊的范围，从而在可管理的时间成本中实现准确性的改进，同时保持主要的工作流量。

分析加权因子对LJM的影响的实验也可以如图6所示进行。 β 从0增加到0.2，MAP和RANK-1度量指标都显着改善，这表明合并局部效果增强了模型性能。超过 $\beta = 0.2$ ，等级-1保持相对稳定，而地图逐渐下降。这种趋势表明，在保持高排名1的准确性时，过度强调本地特征可能会稍微降低总排名质量。地图和等级-1曲线之间的差异揭示了顶级准确性和整体检索性能之间的交易，从而在各种情况下提供了模型优化的见解。因此，选择 $\beta = 0.2$ 作为关节测量的重量。

E. Visualization Analysis

除定量分析外，本节还提供了四个可视化分析。

首先，为了更好地理解视觉文本推理的价值（VTR），我们使用Grad-CAM [58]在最后一个变压器层中可视化注意力特征，如图8所示。该图比较了文本描述（a），原始输入图像（B），基线注意图（C），以及与VTR Applied（d）的注意力图（c），以及对四个样本主题的注意力图。

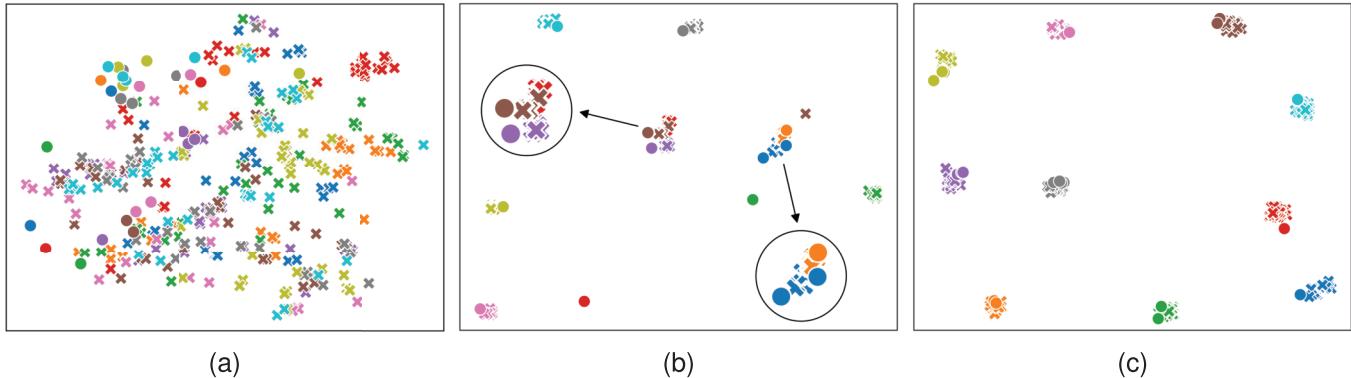


图7. 在SYSU-MM01中的十个随机测试ID上的T-SNE可视化。 (a) 原件。 (b) 基线。 (c) 我们的。相同的ID标记有统一的颜色。交叉和圆圈分别表示可见和红外功能。最好的颜色观看。

A female wearing a yellow long-sleeved top of normal length, black long pants, and white shoes. She is wearing a hat and glasses, and she is not carrying a backpack.

A male wearing a black short-sleeved shirt of normal length, black pants, and black shoes. He is not carrying a backpack, wearing a hat, or glasses.

A male wearing a white short-sleeved shirt of normal length, blue shorts, a backpack, no hat, glasses, and black shoes.

A male wearing a blue short-sleeved shirt of normal length, gray pants, and gray shoes. He is carrying a backpack and is not wearing a hat or glasses.

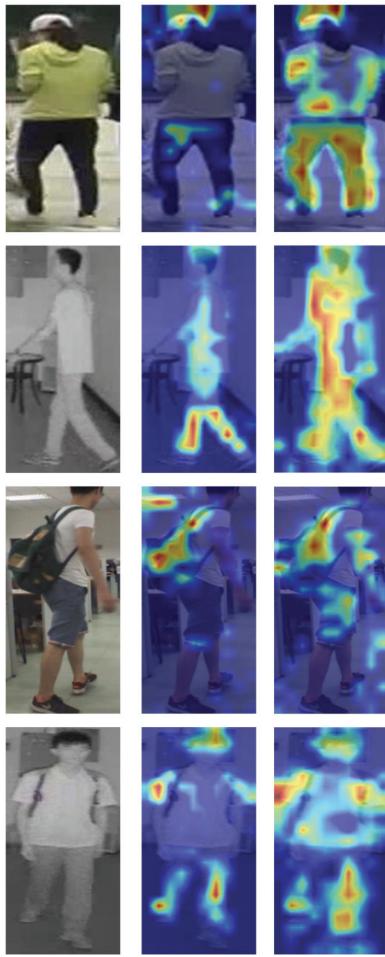


图8。通过Grad-CAM [58]的注意可视化[58]。(a) 标题 (b) 原始输入。
。(c) 基线。(d) 基线+ vtr。不同的颜色突出显示指导描述。黄色比
绿色更突出。最好的颜色观看。

比较图8中的列 (c) 和 (d)，我们观察到包含属性描述的文本通过VTR引起对局部显着区域的关注。与 (c) 列中更多使用的注意力相比，列 (d) 中更为集中和精确的热图显而易见。颜色编码在

文本描述（黄色比绿色更为突出）与热图中的注意力强度相对应，这表明该模型如何优先考虑某些属性而不是其他属性。

这种注意力优化改善了RGB和IR表示之间的模态 - 同时增强了模态不变特征的表达。审视表明，将自然语言引入VI-REID（可见的红外重新识别）和VTR的效果。通过利用文本描述，该模型侧重于针对区分各个不同成像方式的个体至关重要的特定属性，这可能会导致更加稳健，更准确的重新识别结果。

为了验证从星号中提取的模态不变特征的区别，在SYS U-MM01测试集中随机采样了十个ID，以进行定性可视化。基线[57]和Star-Reid的特征在图7中用T-SNE [59]绘制。

如图7所示，我们使用T-SNE可视化特征分布，以说明我们提出的方法的效果。子图(a)呈现原始的复杂数据分布。在(b)中，基线模型的许多身份簇紧密相互交织，表明在跨模态的不同身份方面很困难。圆圈突出了这种身份混乱的例子。

相反，子图 (c) 演示了我们的星级模型（基线+ vtr）实现的特征分布。在这里，每个特定身份的特征簇都以清晰的边界分离。这种分离表明，通过我们所提出的VIAP 和VTR模块，Star-Reid有效地改善了模态不变特征的歧视

(b) 和 (c) 之间的比较为学习歧视性和模态不变特征的出色表现提供了强有力的证据，从而解决了跨模式人员重新认同的关键挑战。

图9展示了LJM对VI-REID样品距离分布的影响。该方法有效地增加了正样品和负样品之间的分离，负样品的峰（红色曲线）从0.218转移到0.271。这种增强仅是通过优化相似性度量而不改变特征空间来实现的。这

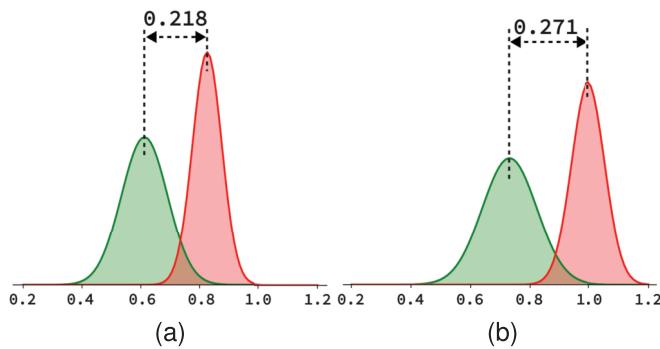


图9。添加局部特征之前/之后的样品之间距离的分布。绿色曲线代表正品，而红色曲线表示负样品。（a）常见测量。（b）拟议的LJM。最好的颜色观看。

宽大的差距表明，以自然语言为指导的局部细粒度特征是可以区分的，这突出了LJM对Vi-Reid的重要性。这种改进的分布表明，在实际情况下，可以更好地准确性，从而有可能降低误报并提高系统性能。该视觉证据强调了LJM通过重新确定的相似性测量技术来推进VI-REID的效果。

V. 结论

在本文中，我们提出了一个新颖的星级框架，以学习VI-ReID的歧视性表示。这是引入自然语言规范以挖掘可见和红外形态之间潜在细粒度线索的第一个。它可以通过VI-ALP和VTR有效地进行多模式学习。特定地，VIAP可以增强视觉界面的交互和全局本地信息融合，从而明确改善了模态不变的特征。VTR负责视觉学习，根据自然语言，这隐含地对视力中的显着区域表示了更多的关注。

此外，我们引入了LJM，以通过结合局部特征距离来增加强度量，从而提高了使用细粒度信息的相关性。我们还提出了一种人类合作方法，将文本描述纳入现有的跨模式重新认同数据集中，利用人类专业知识和大语言模型的力量。

在SYSU-MM01和REGDB数据集上进行的广泛实验证明，Star-Reid的表现优于大多数最先进的方法，并带有文本方法的混音重新训练，显示出对现实世界中多个美容场景的特殊希望。所提出的框架显示出在弱光条件下改善人识别的重要潜力，并且自然语言规范的引入为与大语言模型或AI代理的集成提供了可能性，并且可以更好地与现实世界中的互动场景 - 对目标人进行高度明显的互动式的互动，从而使视频范围进行了反复分析，并进行了证明是相互证明的。而且我们认为，多模式联合学习（结合视觉，红外和文本信息）将是人重新认同的未来趋势，开放

提高VI-REID系统的准确性和鲁棒性的新途径。

参考

- [1] A. Wu, W. Zheng, H.-X. Yu, S. Gong and J. Lai, “RGB-Infrared Cross-Modality Re-Identification”，在*Proc. ICCV*, 2017年10月, 第5380-5389页。
- [2] W. Li等人，“DC形式：重新认同的人的多样和紧凑变压器”，*Proc. AAAI*, 第1卷。37, 2023年6月, 第1415–1423页。
- [3] H. Rao and C. Miao, “Transg: 基于变压器的骨架图原型对比度学习，结构 - 区域对比促使人们重新认同”，在*Proc. IEEE/CVF Conf. Comput. CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*
- [4] M. Kim, S. Kim, J. Park, S. Park and K. Sohn, “PartMix: 研究可见的infrared人重新认同的部分发现的调节策略”，载于*Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog-CVCF Conf. Comput. Vis. Pattern Recog-nit. (CVPR) nit. (CVPR)*, 6月2023年, 第2023页, 第22118-2218-2221218页。
- [5] Y. Zhang and H. Wang, “可见的infrared人重新识别的各种嵌入扩展网络和低光跨模式基准”，载于*Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. CVF Conf. Comput. Vis. Pattern Recog-nit. (CVPR) nit. (CVPR)*, 6月2023年, 第2023页, 第18621-18621-186322232页。
- [6] Y. Zhang and H. Wang, “可见的infrared人重新识别的各种嵌入扩展网络和低光跨模式基准”，载于*Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog-nit. (CVPR) nit. (CVPR)*, 6月2023年, 第2023页, 第2153-2162页。
- [7] S. Yan, N. Dong, L. Zhang and J. Tang, “剪贴式驱动的文本 - 图像人的重新认同”，*IEEE Trans. Image Process.*, 第1卷。32, 第6032-6046、2023页。
- [8] Z. Zhong, L. Zheng, D. Cao and S. Li, “重新列入k-Reciprocal编码的人”，在*Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017年7月, 2017年, 第1318-1327页。
- [9] N. Jiang, Y. Qiu, W. Song, J. Liu, Z. Shi and L. Wang, “对人重新识别的联合视觉 - 文本推理和可见的 - 信号对齐方式”，*Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2024, pp. 1-6。
- [10] N. Gheissari, T. B. Sebastian and R. Hartley, “使用时空外观的人重新识别”，*Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006年6月, 第1528-1535页。
- [11] Y. Shi, Z. Wei, H. Ling, Z. Wang, J. Shen and P. Li, “通过深度属性采矿和推理中的监视视频中的人检索”，*IEEE Trans. Multimedia*, 第1卷。23, pp. 4376-4387, 2021。
- [12] Y. Shi等人，“以政策梯度检索人的自适应和健壮的分区学习”，*IEEE Trans. Multimedia*, 第1卷。23, pp. 3264-3277, 2021。
- [13] H. Liu, S. Ma, D. Xia and S. Li, “Sfanet: Spectrum-Aware-Aware-Aware for fause from from for fausible-Inversion for for fausible-Invisible-Infrade Person Reidentification reidentification”，*IEEE Trans. Neural Netw. Learn. Syst.*, 第1卷。34, 不。4, 第1958-1971年, 2023年4月。
- [14] C. Chen, M. Ye, M. Lin, “可见的红外人重新识别的结构感知位置变压器”，*IEEE Trans. Image Process.*, 第1卷。31, pp. 2352-2364, 2022。
- [15] Z. Cui, J. Zhou and Y. Peng, “DMA：可见的Infrared人重新识别的双态度意识对齐”，*IEEE Trans. Inf. Forensics IEEE Trans. Inf. Forensics Security*, 第1卷。19, 第2696-2708、2024页。
- [16] D. Li, X. Wei, X. Hong and Y. Gong, “具有X模态的红外可视跨模式重新认同”，在*Proc. AAAI Conf. Artif. Intell. Intell.*, 第1卷。34, 2020年4月, 第4610-4617页。
- [17] 15111, 意大利米兰, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler and G. Varol, 编辑, 瑞士: Springer, 2024年, 第55-72页, doi: 10.1007/978-3-3-031-031-731-73668--1。跨频谱重新标识的数据增强的透视图, 在*Proc. 38th Annu. Conf. Neural Inf. Process. Syst.*, 2024年, 第1-12页中。[在线的]。可用: <https://openreview.net/> 论坛? 19, pp. 2800-2813, 2024。
- [18] J. Feng, A. Wu and W.-S. Zheng, “基于形状的特征学习，用于可见的红外人重新认同”，在*Proc. IEEE/CVF Conf. CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023年6月, 第22752-22761页。

- [21] J. Liu, Y. Sun, F. Zhu, H. Pei, Y. Yang和W. Li, “学习记忆的跨模式人重新识别的单向指标”，在*Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit./CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) (CVPR)*, pp. 193666-1935中。[22] M. Yang, Z. Huang, P. Hu, T. Li, J. LV和X. Peng, “在*Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*中学习双嘈杂的标签，用于可见的infrared人的重新认同”，载于2022年8月2022年，第14308-14317-14308-14317页。[23] K. Ren和L. Zhang, “对可见的红外人重新识别的隐性歧视知识学习”，载于*Proc. IEEE/CVF Conf./CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*
Comput. Vis. Pattern Recognit. (CVPR)
- Comput. Vis. Pattern Recognit. (CVPR)*, 美国华盛顿州西雅图，美国华盛顿州，2024年6月，2024年，第393-402页，doi: 10.1109/cvpr55224454545.2024.00045。[24] G. Wang, T. [25] 150, 第155-161页，2021年10月。[26] 38, 不。第1页，第279-294页，2022年1月。[27] Z. Wei, X. Yang, N. Wang和X. Gao, “可见的红外人员重新识别的双反向表示表示”，19，第2186-2200，2024年。[28] [29] [30] P. Dai, R. Ji, H. Wang, Q. Wu和Y. Huang, “带有生成对抗性训练的跨模式的人重新认同”，载于*Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018年7月，第677-683页。[31] D. Liu等人，“针对对抗性稳健的开放世界人的生成度量学习”，*ACM Trans. Multimedia Comput., Commun., Appl.*, 第1卷。19, 没有。1, 第1-19页，2023年1月，doi: 10.1145/3522714。[32] J. Achiam等人，“GPT-4技术报告”，2023, *arXiv:2303.08774*。[33] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue和X. Wang, “具有自然语言描述的人搜索”，载于*Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017年7月，第1970-1979年。[34] Z.-J. Zha, M. Wang, J. Shen和T.-S. Chua, *Text Mining in Multimedia*. 美国马萨诸塞州波士顿: Springer, 2012年, 第361-384页, DOI: 10.1007/978-1-4614-3223-4_11。[35] K. Simonyan和A. Zisserman, “非常深的卷积网络，用于大规模图像识别，”，2014年, *arXiv:1409.1556*。[36] K. He, X. Zhang, S. Ren和J. Sun, “图像识别的深度残留学习”，载于*Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016年6月，第70-778页。[37] S. Hochreiter和J. Schmidhuber, “长期记忆”，*Neural Comput.*, vol. 9, 不。8, 第1735-1780页，1997年。[38] J. Devlin, M.-W. Chang, K. Lee和K. Toutanova, “Bert: 深层双向变压器的预训练，以了解语言理解”，2018年, *arXiv:1810.04805*。[39] Shen, “具有实例丢失的双路径卷积图像文本嵌入”，*ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, 不。2, 第1-23页，2020年5月。
- [40] S. Li, M. Cao和M. [41] C. Gao等人，“基于文本的人搜索的全尺度代表上的上下文非本地对齐”，2021, *arXiv:2101.03036*。[42] [43] P. Zheng等人，“具有基于文本的人搜索的辅助指南的关系感知的聚合网络”，*World Wide Web*, 第1卷。25, 不。4, pp. 1565-1582, 7月2022年。[44] X. Han, S. He, L. Zhang和T. Xiang, “具有有限数据的基于文本的人搜索”，2021年, *arXiv:2110.10807*。[45] [46] [47] A. Radford等人，“从自然语言监督中学习可转移的视觉模型”，*Proc. Int. Conf. Mach. Learn.*, 第1卷。139, 2021, 第8748-8763页。[48] Y. Bai等人，“RASA: 基于文本的人搜索的关系和敏感性意识到的表示形式”，2023, *arXiv:2305.13653*。[49] A. Dosovitskiy等人，“图像价值16×16个单词：用于图像识别的变压器，以规模识别”，在*Proc. ICLR*, 2021年, 第1-21页中。[50] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao和S. C. H. Hoi, “人重新认同的深度学习：调查和前景”，*IEEE Trans. Pattern Anal. Mach. Intell.*, 第1卷。44, 不。6, pp. 2872-2893, Jun. 2022。[51] K. Jiang, T. Zhang, X. Liu, B. Qian, Y. Zhang, and F. Wu, “Cross-modality transformer for visible-infrared person re-identification,” in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 480-496。[52] [53] Y. Zhang, S. Zhao, Y. Kang和J. Shen, “模态协同效应的综合学习，伴随着可见的 - 信号人的重新认同，” *Proc. ECCV*, 2022年, 第462-479页。[54] 46, 不。4, pp. 2299-2315, Apr. 2024。[55] G. Zhang, Y. Zhang, and Z. Tan, “ProtoHPE: Prototype-guided high-frequency patch enhancement for visible-infrared person re-identification,” in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 944-954。[56] 26, 第8172-8183、2024页。[57] H. Lu, X. Zou和P. Zhang, “学习渐进式态度 - 可见的可见界人士的共享变压器重新识别”，*Proc. AAAI Conf. Artif. Intell.*中，第1卷，第1卷。37, 2023, pp. 1835-1843。[58] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh和D. Batra, “Grad-Cam: 通过基于梯度的本地化来自深网的视觉解释”，*Int. J. Comput. Vis.*, 第1卷。128, 不。2, 第336-359页, 2019年10月。[59] L. van der Maaten和G. Hinton, “使用T-SNE可视化数据”，*J. Mach. Learn. Res.*, 第1卷。9, 第2579-2605页, 2008年11月。