计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(9): 2183−2193, 2025

面向复杂噪声数据的鲁棒文本-图像行人检索方法

胡冰玉　徐艺心　余　珊　赵巨峰　杨宇翔
（杭州电子科技大学电子信息学院　杭州　310000）
（huby@hdu.edu.cn）

Robust Text-Based Person Retrieval Under Complex Noisy Data

Hu Bingyu, Xu Yixin, Yu Shan, Zhao Jufeng, and Yang Yuxiang
（*School of Electronics and Information*, *Hangzhou Dianzi University*, *Hangzhou* 310000）

Abstract　Text-based person retrieval, a pivotal task in multimodal intelligent surveillance systems, seeks to identify target pedestrian images from large-scale databases using free-form textual descriptions, with critical applications in public security and video forensics such as suspect tracking in criminal investigations and cross-camera forensic analysis. Conventional approaches predominantly operate under the idealized assumption of perfectly aligned image-text pairs, thereby overlooking the pervasive presence of complex noisy data—incorrect or ambiguous associations between visual instances and their textual annotations—that inevitably arise from human annotation biases, web-crawled data impurities, or semantic granularity mismatches between localized visual attributes and holistic textual contexts. To bridge this critical gap, we propose semantic-informed noisy correspondence learning, a unified framework that systematically addresses both noise identification and robust learning through dual innovative mechanisms. First, the semantic-aware noise identification criterion holistically integrates intra-modal semantic consistency with cross-modal interaction signals, enabling precise discrimination of false-positive and false-negative correspondences via adaptive thresholding. Second, the noise-robust complementary learning paradigm implements differentiated optimization strategies: positive learning with momentum contrastive alignment enforces discriminative feature learning on purified reliable pairs, while negative learning via entropy minimization constrains overfitting to noisy subsets. Extensive experiments on three public benchmarks validate the effectiveness of the proposed approach, achieving superior performance under both synthetic and real-world noise scenarios.

Key words　text-based person retrieval; noisy data; multimodal learning; feature representation; model robustness

摘　要　文本-图像行人检索（text-based person retrieval）作为多模态智能监控系统的核心任务，旨在通过自由形式的文本描述从大规模数据库中识别目标行人图像，在公共安全与视频取证领域具有关键应用价值，如刑事侦查中的嫌疑人追踪及跨摄像头取证分析. 传统方法通常基于图像-文本对完美对齐的理想化假设，忽视了实际场景中普遍存在的复杂噪声数据问题，即视觉实例与其文本标注间因人工标注偏差、网络爬取噪声，或局部视觉属性与全局文本语境间的语义粒度失配而产生的错误或歧义性关联. 为弥补这一缺陷，提出了一种语义感知噪声关联学习框架，通过双重创新机制系统性地实现噪声辨识与鲁棒学习. 首先，语义感知噪声辨识准则融合模态内语义一致性与跨模态交互信号，基于自适应阈值判定精准区分噪声关联；其次，噪声鲁棒互补学习范式实施差异化优化策略：对于可靠子集采用对比损失进行正向学习以增强特征判别性，而对噪声子集则通过反向学习以抑制过拟合. 在 3 个公开基准数据集上的大量实验

收稿日期：2025-06-04；**修回日期**：2025-07-09
基金项目：国家自然科学基金项目（62376080）
This work was supported by the National Natural Science Foundation of China (62376080).
通信作者：杨宇翔（yyx@hdu.edu.cn）

表明,该方法在合成噪声数据与真实噪声数据场景中均展现出优越性能.

关键词 文本-图像行人检索;噪声数据;多模态学习;特征表示;模型鲁棒性

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202550394 **CSTR:** 32373.14.issn1000-1239.202550394

文本-图像行人检索(text-based person retrieval, TBPR)是一项典型的跨模态检索任务,旨在通过自然语言描述,从大规模监控图像库中准确检索出目标行人图像[1]. 相比于传统的行人图像-图像检索任务,即行人重识别(person re-identification, Person ReID), TBPR具备更强的交互性与实用性,在很多现实场景中,尤其是公共安全与刑事侦查领域,用户往往无法提供查询图像,但可以基于目击信息提供相对详尽的文本描述,例如"穿着红色连帽衫、黑色裤子的中年男性". 这使得TBPR在跨摄像头追踪、视频证据检索等任务中具有重要应用价值[2].

近年来,得益于多模态深度表示学习的发展, TBPR取得了显著进展. 主流方法多围绕跨模态特征对齐展开,重点在于如何设计有效的图像与文本嵌入策略,使2种模态在统一的特征空间中实现语义对齐. 为提升匹配精度,研究者提出了多粒度特征匹配机制,如局部区域与文本短语的显式对应[3-4]、全局语义信息融合[5]等. 这些方法在多个公开数据集上均取得了良好的检索性能,推动了该领域的快速发展.

然而,现有方法普遍依赖于训练集中图像-文本样本对完全对齐的理想化假设,即每一对图像与文本描述都被视为真实且准确的匹配关系. 这一假设在实际场景中往往难以成立. 由于训练数据通常来自人工标注或网络爬取,图像-文本对之间不可避免地存在噪声关联(noisy correspondences, NC)问题,表现为配对错误、语义歧义或语义不完整等. 例如,如图1所示,一个描述为"穿着黑色长外套"的文本可能对应多张外观极为相似但身份不同的行人图像;也可能出现描述中未提及的关键属性,而模型却基于表面特征进行错误关联的情况. 这类噪声不仅导致特征学习出现偏差,更在测试阶段引发严重的匹配误判,从而降低了系统的鲁棒性和实用性[6].

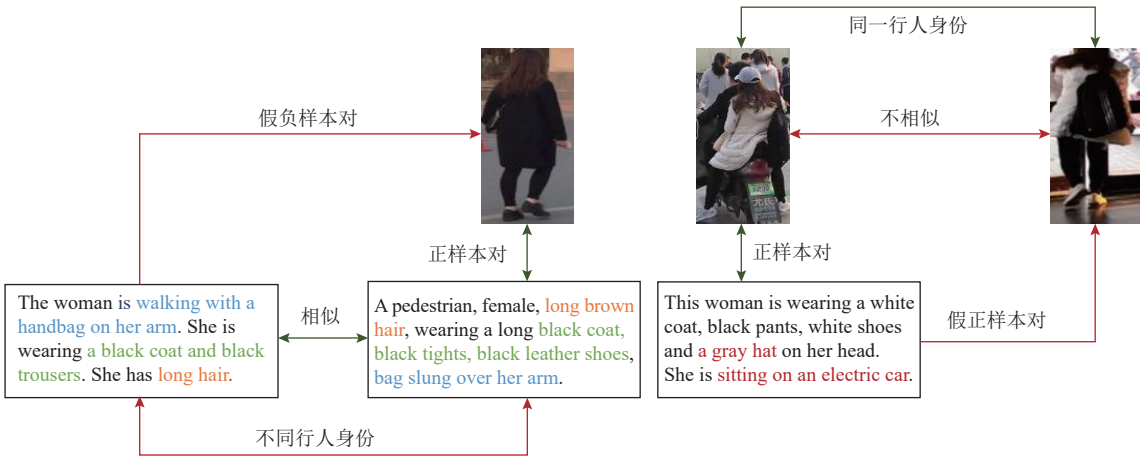


Fig. 1 False positive and false negative pairs from RSTPReid dataset

图1 RSTPReid数据集中的假正样本对和假负样本对

事实上,噪声关联问题已在跨模态检索、视频理解等多种多模态任务中被广泛关注[7-9]. 代表性方法多基于深度神经网络的记忆效应[10]:模型在训练早期倾向优先学习容易样本,使得真实配对样本具有更高的相似性. 具体来说,这些方法常构建样本对相似度的排序机制,并结合对比损失进行噪声样本筛选与建模. 然而, TBPR任务面临更为复杂的挑战. 行人图像与描述文本之间往往不是一一映射,而是多对多语义对应关系. 1个文本描述可能与多张语义相似但身份不同的图像关联,反之亦然. 因此,仅简单借助行人身份标签构建正负样本对来计算对比损失,无法充分表达模态间的语义关系,极易导致假正样本对(false positive)与假负样本对(false negative)的混淆,从而进一步加剧特征对齐误差的累积.

针对上述问题,本文提出一种新颖的语义感知噪声关联学习(semantic-informed noisy correspondence

learning, SNCL)框架，通过双重创新机制系统性地实现噪声辨识与鲁棒学习，提升文本-图像行人检索模型在噪声数据环境下的鲁棒性. 首先，提出语义感知的噪声识别效用函数(semantic-informed noise identification criteria, SNI)，通过融合模态内的语义一致性与跨模态的交互信息，精准识别噪声样本. 具体而言，考虑到假阳性样本常因身份不同但语义相似而被误判为正样本，SNI 利用图像/文本模态内部的相似度结构，对训练对进行语义分析，从而缓解按照身份标签划分所导致的判别偏差. 其次，构建噪声鲁棒的互补学习(noise-robust complementary learning, NCL)范式，以差异化的优化策略分别处理可靠子集与噪声子集. 对于可靠的样本对，采用正向学习策略，借助对比损失强化语义对齐能力；对于噪声样本对，则实施反向学习，通过排除潜在干扰信息，防止噪声样本主导梯度方向. 该策略有效避免了现有方法对伪标签准确性的过度依赖，并在高噪声场景下展现出更优的鲁棒性与泛化能力.

综上所述，本文主要贡献包括：

1)提出了一种面向文本-图像跨模态行人检索的语义感知噪声关联学习框架，在合成噪声数据与真实噪声数据场景中均取得最优性能；

2)设计了一种语义感知的噪声识别效用函数，通过挖掘模态内语义关系修正正负样本对的构建，有效缓解了假正样本对/假负样本对问题；

3)构建了一种噪声鲁棒的互补学习范式，通过正/负协同优化策略提升训练效率与抗噪性能，突破现有方法在噪声数据环境下的性能瓶颈.

1 相关工作

1.1 文本-图像行人检索

文本-图像行人检索任务旨在利用自然语言描述从大规模图像库中精确定位目标行人. 现有方法主要分为显式特征对齐与隐式特征对齐 2 类. 显式特征对齐方法通过在特征空间中直接对齐图像与文本信息实现模态间语义一致[11,12]. 早期工作采用全局特征匹配策略，通过联合嵌入空间将整张图像与完整文本描述进行对齐[13]. 随着研究的深入，逐步发展为更精细的多粒度全局-局部匹配机制[14]. 隐式特征对齐方法则借助注意力机制，通过构建图像区域与文本元素间的细粒度对应关系实现语义对齐[15]. 例如，Yan 等人[16] 提出将图像和文本特征自适应地聚合到一组模态共享的语义中心，并在没有额外监督信息和复杂跨模态交互的情况下隐式学习图像与文本之间的局部细粒度对应关系.

尽管上述方法在标准数据集上取得了良好效果，但普遍建立在图像-文本对严格对齐的假设之上，忽略了现实中广泛存在的噪声关联问题. 由于人工标注误差、网络爬取误配等因素，在大规模跨模态数据集中，图像-文本对中的噪声关联广泛存在，严重干扰了跨模态表示学习，导致检索性能下降. 因此，如何提升 TBPR 模型在噪声关联条件下的鲁棒性，成为亟需解决的核心问题.

1.2 噪声关联学习

噪声关联学习 (noisy correspondence learning)旨在解决配对数据中的对齐错误问题，近年来在多模态推荐[17]、视频问答[18-19] 等多模态任务中受到了广泛关注. 与传统的噪声标签学习不同，噪声关联学习更侧重于在高层语义空间中捕捉数据的内在结构和语义信息，设计高效的跨模态相似度度量来识别噪声关联. 例如，Han 等人[20] 构建了一种元相似协同网络，利用元过程优化三元组排序损失，从正负元数据中学习区分性. 在可见光-红外跨模态行人重识别任务中，Yang 等人[21] 针对标签噪声问题提出了双重鲁棒训练策略，借助单样本置信度估计机制有效推导模态间的关联可信度. 2024 年，Qin 等人[6] 聚焦于 TBPR 任务中的噪声关联问题，提出结合双重嵌入结构与置信共识划分策略的鲁棒学习方法，能够在图像质量低或文本标注不准确时识别并利用可靠配对样本.

然而，现有研究多集中于假正样本对的识别与建模，忽视了 TBPR 任务中更为复杂的假负样本对现象. 由于行人外观具有普遍的类内差异与类间相似性，同时文本描述存在高度主观性与模糊性，不同身份的行人往往会被相同文本描述所覆盖，导致本应关联的图文对被误判为负样本，从而引发语义对齐偏差与特征退化. 当前对此问题的研究仍显不足，亟待进一步探索更具语义感知能力的噪声关联学习机制.

2 提出方法

2.1 问题定义

文本-图像行人检索任务旨在根据用户提供的自然语言描述，从大规模图像库中检索与该描述语义匹配的目标人物图像. 给定图像集合 $\mathcal{I}=\{I_i,y_i\}_{i=1}^{N_I}$，其中 N_I 代表图像的数量，$y_i\in Y=\{1,2,\cdots,C\}$ 表示行人身份标签，C 代表身份总数. 对于每幅图像，均配有若干

对应的文本描述, 记作 $\mathcal{T}=\{T_j\}_{j=1}^{N_T}=\{\{T_k^i\}_{k=1}^{K_i}\}_{i=1}^{N_I}$, 其中 $\{T_k^i\}_{k=1}^{K_i}$ 代表与图像 I_i 对应的文本集合, K_i 代表其数量, N_T 代表文本总数, 即 $N_T=\sum_{i=1}^{N_I}K_i$. 在 TBPR 任务中, 模型训练通常基于图像-文本对进行监督学习, 训练数据可以表示为 $\Omega=\{(I_i,T_i),y_i\}_{i=1}^{N}$, 其中 $I_i\in\mathcal{I},T_i\in\mathcal{T}$, N 为图像-文本对总数. 理想情况下, 每一对 (I_i,T_i) 应为语义一致的正确配对, 即文本 T_i 准确描述图像 I_i 所对应的行人. 然而, 在实际数据收集过程中, 由于标注误差、爬虫抓取偏差等因素, 训练数据中不可避免地混入语义不一致的噪声配对, 即图文对之间存在错误关联关系. 为建模这一现象, 本文引入二元标签 m_i 来表示 (I_i,T_i) 对的真实匹配关系, $m_i=1$ 表示匹配正确, $m_i=0$ 表示匹配错误, 最终带噪训练集可以表示为 $\Omega=\{(I_i,T_i),y_i,m_i\}_{i=1}^{N}$. 综上所述, TBPR 的总体目标是在含噪声训练数据的条件下, 学习一个鲁棒的跨模态检索模型, 使得对于任意查询文本 T_q, 模型能够准确计算其与图像集合中各图像之间的匹配相似度, 并根据相似度完成检索排序.

2.2 整体框架

为解决上述问题, 如图 2 所示, 本文提出了一种语义感知噪声关联学习框架. 首先, 设计了一种语义感知的噪声识别效用函数, 通过提取样本模态内和模态间的全局语义相似度信息, 有效提升标签噪声识别的准确率. 其次, 提出了一种噪声鲁棒的互补学习范式, 以差异化的优化策略分别处理可靠子集与噪声子集. 通过准确识别并充分利用噪声样本, 最终可以应对数据标签带噪条件下模型语义对齐失真的问题, 提升模型应对带噪监督信号的特征学习能力.

2.3 语义感知的噪声识别效用函数

为缓解噪声关联对模型训练带来的干扰, 首要任务是从训练数据中准确识别出潜在的噪声图文对, 从而避免错误监督对模型学习的误导. 已有关于噪声标签学习的研究[7-9]表明, 深度神经网络在训练早期更容易优先拟合容易样本, 该记忆效应[10]为噪声识别提供了理论基础. 具体而言, 相较于噪声配对, 真实匹配的图文样本对往往具有更高的特征相似度. 基于此, 主流方法通常采用对比损失作为噪声识别效用函数, 通过逐对计算训练样本的损失值分布 $\{\mathcal{L}(I_i,T_i)|(I_i,T_i)\in\Omega\}$, 将数据集划分为可靠子集与噪声子集, 其中, 损失函数 $\mathcal{L}(\cdot)$ 常采用基于对比学习的 InfoNCE 损失[22], 其形式如下:

$$\mathcal{L}_{\mathrm{NCE}}(I_i,T_i)=-\frac{1}{2}\log\frac{\exp(A(I_i,T_i)/\tau)}{\sum_{j=1}^{N}\exp\left(A\left(I_i,T_j\right)/\tau\right)}-\frac{1}{2}\log\frac{\exp(A(I_i,T_i)/\tau)}{\sum_{j=1}^{N}\exp\left(A\left(I_j,T_i\right)/\tau\right)},\tag{1}$$

其中 $A(\cdot,\cdot)$ 为模态间特征对齐度, 通常采用余弦相似度计算, τ 是温度参数, 用于调节相似度分布的平滑程度. 然而, 在 TBPR 任务中, 图文匹配天然存在假负样本问题, 即不同行人图像之间可能具有高度语义相似性, 从而使其对应的文本描述无法被简单视为负样本. 例如, 2 个身穿相似服饰的行人, 其描述“穿白色 T 恤的男子”可能同时适用于多个样本对, 因此简单地使用对比损失函数难以避免假负样本的干扰. 为了解决这一挑战, 本文提出一种语义感知的噪声识别效用函数, 通过引入模态内部语义相似性, 修正

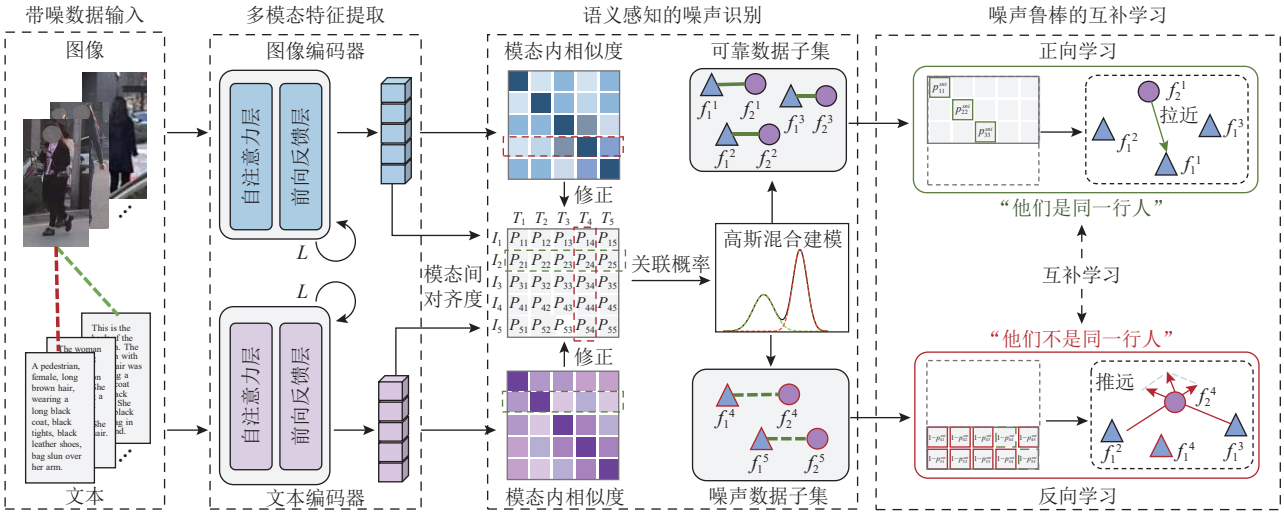


Fig. 2 The proposed semantic-informed noisy correspondence learning framework

图 2 所提出的语义感知噪声关联学习框架

行人数据多重关联特性引起的假负样本的影响. 首先, 计算图像-文本对关联概率:

$$P_{\mathrm{SNI}}(I_i,T_i)=\frac{1}{2}(P_{\mathrm{SNI}}^{\mathrm{i2t}}(I_i,T_i)+P_{\mathrm{SNI}}^{\mathrm{t2i}}(I_i,T_i)), \tag{2}$$

其中:

$$P_{\mathrm{SNI}}^{\mathrm{i2t}}(I_i,T_i)=\frac{\exp(A(I_i,T_i)/\tau)}{\sum_{j=1}^{N}(1-S(I_i,I_j))^{\eta}\exp(A(I_i,T_j)/\tau)}, \tag{3}$$

$$P_{\mathrm{SNI}}^{\mathrm{t2i}}(I_i,T_i)=\frac{\exp(A(I_i,T_i)/\tau)}{\sum_{j=1}^{N}(1-S(T_i,T_j))^{\eta}\exp(A(I_j,T_i)/\tau)}. \tag{4}$$

式(3)和式(4)分别用于衡量图像到文本(i2t)和文本到图像(t2i)的正确匹配概率, 其中$A(\cdot,\cdot)$为模态间特征对齐度, 通常采用余弦相似度计算. $S(\cdot,\cdot)$为模态内特征相似度, 用于刻画样本在同一模态下的语义关系. 以式(3)为例, 当一个图像在模态内存在多个语义接近的邻居样本时, 相似度$S(\cdot,\cdot)$会较高, 使得分母中该负样本对应的加权项衰减, 从而降低噪声对损失的影响. η是超参数, 用于控制衰减速度. 基于上述设计, 给定跨模态特征对齐网络f_θ, 最终噪声样本对识别效用函数定义为:

$$l(\theta)=\{l_i\}_{i=1}^{N}=\{-\log(P_{\mathrm{SNI}}(I_i,T_i))\}_{i=1}^{N}. \tag{5}$$

在计算所有样本对的关联概率后, 本文使用双峰高斯混合模型(Gaussian mixture model, GMM)[23]对其分布进行建模:

$$p(l|\theta)=\sum_{k=1}^{K}\beta_k\phi(l|k), \tag{6}$$

其中β_k和$\phi(l|k)$分别为混合系数和第k个分量的概率密度. 基于神经网络的记忆效应, 本文将均值较低的分量, 即损失较低的分量视为可靠子集, 其余视为噪声子集. 通过EM算法对GMM进行优化后, 可得到第i对样本属于可靠子集的概率, 计算公式为:

$$p_i=p(k=0|l_i)=\frac{p(k=0)\,p(l_i|k=0)}{p(l_i)}, \tag{7}$$

其中k是均值较低的高斯分量. 此外, 为了将关联标签的估计过程与模型学习的优化过程协同统一, 我们引入了时序集成技术[24], 以实现更稳定可靠的训练效果:

$$\hat{p}_i[t]=\beta p_i[t]+(1-\beta)\,p_i[t-1], \tag{8}$$

其中β为动量系数, t为迭代次数. 根据置信阈值将数据分为可靠子集Ω^{c}和噪声子集Ω^{n}. 默认阈值设定为0.5.

$$\hat{m}_i=\begin{cases}1, & \hat{p}_i>0.5,\\ 0, & \hat{p}_i\leqslant 0.5.\end{cases} \tag{9}$$

最终得到的可靠子集Ω^{c}和噪声子集Ω^{n}为:

$$\begin{aligned}\Omega^{\mathrm{c}}&=\{(I_i,T_i)|\hat{m}_i=1,\forall(I_i,T_i)\in\Omega\},\\ \Omega^{\mathrm{n}}&=\{(I_i,T_i)|\hat{m}_i=0,\forall(I_i,T_i)\in\Omega\}.\end{aligned} \tag{10}$$

2.4 噪声鲁棒的互补学习范式

在完成训练样本对的可靠性划分后, 下一步的关键在于设计一种对噪声具备鲁棒性的目标函数, 以实现对模型训练过程的有效引导. 当前大多数方法主要关注于对噪声数据对进行标签修正, 常通过为噪声样本分配软标签的方式来缓解错误监督. 然而, 这类方法在实际应用中面临2个关键挑战: 一方面, 模型性能高度依赖于标签校正的准确性; 另一方面, 在高噪声比例的场景中, 错误的软标签容易产生误导性梯度, 甚至导致训练过程陷入平凡解的风险[25]. 为此, 本文从互补对比学习(complementary contrastive learning)的思想出发[26], 提出一种噪声鲁棒的互补学习范式. 该范式的核心思想是在不同的样本子集上实施差异化学习策略: 在可靠子集上执行正向学习(positive learning)以强化正确监督, 在噪声子集上实施反向学习(negative learning)以规避错误信息, 从而提升整体学习的鲁棒性.

1)正向学习策略. 对于被判定为可靠的图文样本对集合Ω^{c}, 其关联标签具有高度可信度. 因此, 在这些样本上进行标准的对比学习有助于稳定模型收敛, 强化图文模态间的表示对齐. 对应的正学习损失定义为:

$$\mathcal{L}_{\mathrm{p}}=-\sum_{i\in\Omega^{\mathrm{c}}}\log(P_{\mathrm{SNI}}(I_i,T_i)). \tag{11}$$

2)反向学习策略. 相反地, 对于被判定为噪声的样本对集合Ω^{n}, 直接用于正向学习可能引入错误监督. 为避免这种风险, 本文采取反向学习策略, 鼓励模型在特征空间中将不匹配的样本对进行区分. 具体地, 定义反向学习损失为:

$$\begin{aligned}\mathcal{L}_{\mathrm{n}}=&-\frac{1}{N}\sum_{i\in\Omega^{\mathrm{n}}}\sum_{j\neq i}\log\left(1-P_{\mathrm{SNI}}^{\mathrm{t2i}}\left(I_i,T_j\right)\right)-\\ &\frac{1}{N}\sum_{i\in\Omega^{\mathrm{n}}}\sum_{j\neq i}\log\left(1-P_{\mathrm{SNI}}^{\mathrm{i2t}}\left(I_j,T_i\right)\right).\end{aligned} \tag{12}$$

将正/反学习策略进行联合, 构建最终的整体鲁棒学习目标函数如下:

$$\mathcal{L}=\mathcal{L}_{\mathrm{p}}+\gamma\mathcal{L}_{\mathrm{n}}, \tag{13}$$

其中γ为正/反学习损失之间的权重平衡系数, 用于控制2个样本子集在训练过程中的相对影响力. 在实际训练中, 可根据噪声比例自适应设置该参数, 以兼顾收敛速度与模型鲁棒性.

综上所述, 本文所提出的噪声鲁棒互补学习范

式针对复杂噪声数据场景,采用差异化策略分别处理可靠样本子集与噪声样本子集,有效强化模型对判别性特征的学习能力,显著抑制图文对噪声对模型性能的干扰与削弱,为鲁棒跨模态检索模型的构建提供了有效的优化机制.

3 实 验

3.1 实验设置

1)数据集.为充分验证所提方法的有效性,本文在3个常用基准数据集上进行了实验,分别为CUHK-PEDES[1]、ICFG-PEDES[27]和RSTPReid[28].这些数据集涵盖了不同规模和多样性的行人图像-文本对.

CUHK-PEDES数据集包含40 206张图像,涵盖13 003个独立身份,并配有80 412条文本描述.数据集划分为11 003个身份用于训练、1 000个身份用于验证、1 000个身份用于测试.

ICFG-PEDES数据集包含54 522张图像,涉及4 102个身份,其中3 102个身份用于训练、1 000个身份用于测试.每张图片均有对应的文本描述.

RSTPReid数据集包含20 505张图像和4 101个身份,每张图像配有2个文本描述.数据集划分为3 701个身份用于训练、200个身份用于验证、200个身份用于测试.

2)评价指标.为了全面评估检索性能,我们采用了Rank-K指标(K=1, 5, 10),该指标广泛应用于基于图像和文本的检索任务中.此外,我们还结合了平均精度均值(mAP)和平均逆负惩罚均值(mINP),以便对比现有技术并提供更为全面的性能评估[2,6].

3)实现细节.本文使用来自CLIP-ViT-B/16和CLIP-Xformer[29]的权重初始化图像和文本编码器.所有图像大小调整为384×128.在训练过程中,应用随机水平翻转、带填充的裁剪和擦除来进行图像增强.温度系数τ、超参数β、γ和η默认设置分别为0.02、0.8、1.0和1.5,并在后续实验中进行了超参数分析.本文方法使用Adam优化器进行训练,训练周期为60轮,初始学习率为1E−5,批次大小为64,并使用余弦学习率衰减.所有实验均在一张Tesla NVIDIA V100 GPU上进行.

3.2 性能比较

为验证所提出方法的有效性,本文在3个主流数据集上分别与文本-图像行人检索方法IRRA[2]、SSAN[3]、IVT[4]、CFine[5]、RDE[6]以及通用的噪声关联学习方法DECL[25]、CLIP-C[29]进行了系统对比实验.为模拟实际应用中图像-文本对存在未对齐问题的情况,我们合成了20%、50%和80%三种不同噪声比例的训练集,以更真实地评估模型在高噪声环境下的鲁棒性.实验结果如表1~3所示,全面展示了各方法在不同噪声水平下的检索性能表现.

首先,实验结果表明,传统的文本-图像行人检索方法在噪声关联干扰下表现出明显的性能退化,揭示了训练数据中图文匹配错误对模型学习过程产生的不利影响.以CUHK-PEDES数据集为例,IRRA方法在无噪声条件下的Rank-1准确率和mAP分别为73.38%和66.13%.当噪声比例提升至50%时,性能显著下降,分别降至62.41%和55.52%.这一现象表明,

Table 1 Performance Comparison with State-of-the-art Methods on CUHK-PEDES Dataset

表1 在CUHK-PEDES数据集上与最先进方法的性能比较 %

噪声比例	方法	Rank-1	Rank-5	Rank-10	mAP	mINP
0%	SSAN	61.37	80.15	86.73		
	IVT	65.59	83.11	89.21		
	CFine	69.57	85.93	91.15		
	IRRA	73.38	89.93	93.71	66.13	50.24
	RDE	75.94	90.14	94.12	67.56	51.44
	本文方法	**76.87**	**91.11**	**94.66**	**68.58**	**52.45**
20%	SSAN	46.52	68.36	77.42	42.49	28.13
	IVT	58.59	78.51	85.61	57.19	45.78
	IRRA	69.74	87.09	92.20	62.28	45.84
	CLIP-C	66.41	85.15	90.89	59.36	43.02
	DECL	70.29	87.04	91.93	62.84	46.54
	RDE	74.46	89.42	93.63	66.13	49.66
	本文方法	**76.24**	**90.61**	**94.20**	**67.47**	**50.73**
50%	SSAN	13.43	31.74	41.89	14.12	6.91
	IVT	50.49	71.82	79.81	48.85	36.60
	IRRA	62.41	82.23	88.40	55.52	38.48
	CLIP-C	64.02	83.66	89.38	57.33	40.90
	DECL	65.22	83.72	89.28	57.94	41.39
	RDE	71.33	87.41	91.81	63.50	47.36
	本文方法	**74.50**	**89.77**	**93.93**	**66.00**	**49.10**
80%	SSAN	0.18	0.83	1.45	0.47	0.24
	IVT	34.03	55.49	66.16	33.90	23.29
	IRRA	38.63	56.69	64.18	34.60	21.84
	CLIP-C	57.38	78.05	84.97	51.08	34.83
	DECL	47.90	71.57	80.17	44.51	29.86
	RDE	64.99	83.15	89.52	57.84	41.07
	本文方法	**67.43**	**85.49**	**90.81**	**60.06**	**43.18**

注:黑体数值为最优性能.

Table 2 Performance Comparison with State-of-the-art Methods on ICFG-PEDES Dataset

表 2 在ICFG-PEDES数据集上与最先进方法的性能比较 %

噪声比例	方法	Rank-1	Rank-5	Rank-10	mAP	mINP
0%	SSAN	54.23	72.63	79.53		
	IVT	56.04	73.60	80.22		
	CFine	60.83	76.55	82.42		
	IRRA	63.46	80.24	85.82	38.05	7.93
	RDE	67.68	82.47	87.36	40.06	7.87
	本文方法	**68.85**	**83.08**	**88.96**	**41.96**	**9.97**
20%	SSAN	40.57	62.58	71.53	20.93	2.22
	IVT	50.21	69.14	76.18	34.72	8.77
	IRRA	60.76	78.26	84.01	35.87	6.80
	CLIP-C	55.25	74.76	81.32	31.09	4.94
	DECL	61.95	78.36	83.88	36.08	6.25
	RDE	66.54	81.70	86.70	39.08	7.55
	本文方法	**67.85**	**82.70**	**87.64**	**40.47**	**9.64**
50%	SSAN	18.83	37.70	47.43	9.83	1.01
	IVT	43.03	61.48	69.56	28.86	6.11
	IRRA	52.53	71.99	79.41	29.05	4.43
	CLIP-C	51.60	71.89	79.31	28.76	4.33
	DECL	57.50	75.09	81.24	32.64	5.27
	RDE	63.76	79.53	84.91	37.38	6.80
	本文方法	**65.89**	**81.13**	**84.41**	**41.45**	**9.50**
80%	SSAN	0.28	0.99	1.90	0.27	0.15
	IVT	21.10	37.10	45.64	13.68	2.32
	IRRA	28.19	44.14	51.27	14.36	1.41
	CLIP-C	44.84	65.24	73.27	24.27	3.42
	DECL	40.53	61.49	69.84	21.78	2.97
	RDE	56.02	74.00	80.62	30.67	4.60
	本文方法	**57.40**	**76.86**	**84.13**	**34.55**	**8.41**

注：黑体数值为最优性能.

Table 3 Performance Comparison with State-of-the-art Methods on RSTPReid Dataset

表 3 在RSTPReid数据集上与最先进方法的性能比较 %

噪声比例	方法	Rank-1	Rank-5	Rank-10	mAP	mINP
0%	SSAN	43.50	67.80	77.15		
	IVT	46.70	70.00	78.80		
	CFine	50.55	72.50	81.60		
	IRRA	60.20	81.30	88.20	47.17	25.28
	RDE	65.35	83.95	89.90	50.88	28.08
	本文方法	**69.85**	**86.65**	**92.85**	**54.40**	**31.02**
20%	SSAN	35.10	60.00	71.45	28.90	12.08
	IVT	43.65	66.50	75.70	37.22	20.47
	IRRA	58.75	81.90	88.25	46.38	24.78
	CLIP-C	54.45	77.80	86.70	42.58	21.38
	DECL	61.75	80.70	86.90	47.70	26.07
	RDE	64.45	83.50	90.00	49.78	27.43
	本文方法	**68.40**	**86.15**	**91.75**	**52.55**	**28.81**
50%	SSAN	19.40	39.25	50.95	15.95	6.13
	IVT	39.70	63.80	73.95	34.35	18.56
	IRRA	56.65	78.40	86.55	42.41	21.05
	CLIP-C	53.45	76.80	85.50	41.43	21.17
	DECL	56.75	80.55	87.65	44.53	23.61
	RDE	62.85	83.20	89.15	47.67	23.97
	本文方法	**66.00**	**85.40**	**91.00**	**50.73**	**27.32**
80%	SSAN	0.65	3.25	5.95	1.30	0.70
	IVT	15.15	30.00	40.50	14.98	7.79
	IRRA	29.65	46.65	54.50	23.77	11.32
	CLIP-C	47.80	72.70	81.75	37.50	18.09
	DECL	48.15	72.20	80.75	37.31	18.83
	RDE	53.40	76.70	85.55	39.71	18.28
	本文方法	**61.95**	**82.75**	**88.90**	**47.33**	**24.26**

注：黑体数值为最优性能.

在文本-图像对存在错误关联的情况下，传统方法难以有效建模图像与文本之间的真实语义对应关系，容易受到错误监督信息的干扰，导致模型泛化能力下降. 相比之下，本文提出的鲁棒噪声关联学习方法在各个噪声水平下均保持了优越的性能和较强的鲁棒性. 在无噪声条件下，该方法达到 76.87% 的 Rank-1 准确率和 68.58% 的 mAP，优于现有主流方法；在噪声比例为 50% 的情况下，仍可实现 74.50% 的 Rank-1 准确率和 66.00% 的 mAP，性能下降幅度显著小于对比方法. 该结果表明，所提出的语义感知噪声识别机制与互补学习范式能够有效缓解噪声关联带来的负面影响，显著提升模型在复杂、低质量监督环境下的跨模态对齐能力和检索性能.

其次，与现有噪声鲁棒学习方法相比，本文方法在所有评估指标上均展现出更优的性能，体现了其在应对高比例噪声关联场景中的显著优势. 以 RSTPReid 数据集为例，在 80% 噪声比例下，本文方法的 Rank-1 准确率和 mAP 精度分别达到 61.95% 和 47.33%，显著优于当前最优方法 RDE，其对应指标仅为 53.40% 和 39.71%，Rank-1 准确率和 mAP 精度分别提升了 8.55 个百分点和 7.62 个百分点，验证了所提方法的优越性.

3.3 实验分析

1)消融实验. 为了验证所设计的语义感知噪声识别效用函数(SNI)和噪声鲁棒的互补学习范式(NCL)的有效性, 我们将每个组件分别加入到基线模型中, 并对它们的性能进行比较分析. 结果如表 4 所示, 其中基线模型使用标准对比损失函数进行训练. 研究结果表明, 这 2 个因素对于优化结果都至关重要. 具体而言, 加入 SNI 可使 Rank-1 准确率和 mAP 值分别提高 6.93 个百分点和 5.01 个百分点. 此外, 整合 NCL 可使 Rank-1 准确率和 mAP 值分别显著提高 9.00 个百分点和 7.31 个百分点. 值得注意的是, 同时部署 SNI 和 NCL 可产生最令人信服的结果, 这突显了这 2 个精心设计的模块的互补性.

2)超参数敏感性. 本文方法涉及 3 个关键超参数: β、γ和η, 它们的影响如图 3 所示, 所有实验均在噪声为 50% 的 CUHK-PEDES 数据集上进行. 参数γ平衡了所提出的 2 种损失. 结果表明, 本文方法在各种γ值范围内始终优于其他方法, 尤其是在 [0.1, 10.0] 范围内, 在γ = 1.0 时达到最佳性能. 同时, 动量系数β对于确保时间集成过程中的稳定更新至关重要. 当β> 0.5 时, 模型表现出稳定的性能, 而较低的β值可能会延迟更新并导致对噪声的过拟合. 最后, η控制着设计的 SNI 准则中的衰减速度. 包含模态内相似性校正的模型优于η=0 的模型, 证实了所提出的 SNI 准则的有效性. 当η=1.5 时, 本文方法性能最佳.

Table 4 Analysis Kesults of Ablation Studies on CUHK-PEDES Dataset with 50% Noise

表 4 CUHK-PEDES 数据集上的含 50% 噪声的消融实验分析结果 %

方法	Rank-1	Rank-5	Rank-10	mAP	mINP
Baseline	63.10	80.06	88.14	56.82	39.40
Baseline + SNI	70.03	85.71	90.92	61.83	44.87
Baseline + NCL	72.10	88.61	93.05	64.13	47.59
Baseline + SNI + NCL	74.50	89.77	93.93	66.00	49.10

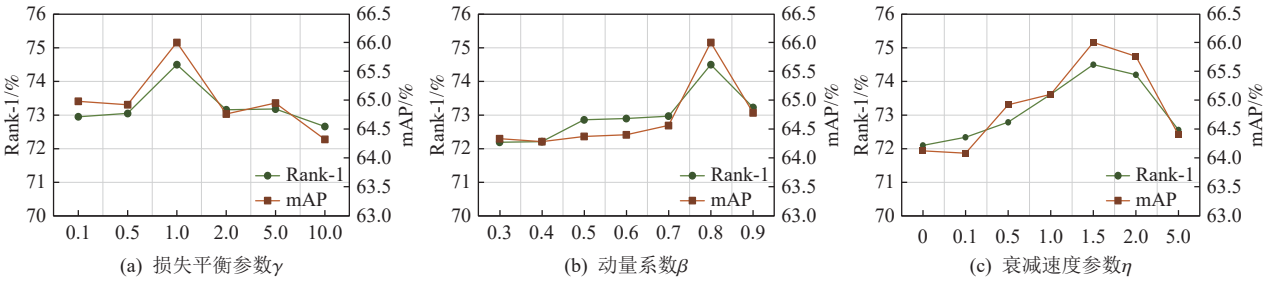


Fig. 3 Analysis of different hyper-parameter combinations

图 3 不同超参数组合的分析

3)训练过程的可视化. 为了进一步说明本文方法在处理噪声关联关系方面的有效性, 在图 4 中展示了整个训练过程中的噪声关联估计准确率曲线和 Rank-1 准确率曲线. 实验结果显示, 在 60 个训练周期内, 噪声关联估计准确率与 Rank-1 检索准确率呈现显著的协同优化趋势: 初始阶段(epoch 0~15)噪声估计准确率从 0.62 快速提升至 0.82, Rank-1 准确率同步由 63.2% 跃升至 69.5%, 该现象表明本文所设计的语义感知噪声识别效用函数在训练初期即可建立有效的噪声检测基准; 优化阶段(epoch 16~40)这 2 条曲线保持稳定增长, 噪声估计准确率以 0.03/epoch 的梯度提升至 0.93, Rank-1 准确率以 0.12/epoch 的速率上升至 73.8%, 验证了动态置信度阈值调整机制的有效性; 最终收敛阶段(epoch 41~60)噪声估计准确率稳定在 0.95±0.01 的高位区间, Rank-1 准确率同步收敛于 74.5% 的稳定状态, 其标准差降至 0.38%, 证实了系统在强噪声条件下的鲁棒性. 随着训练的进行, 噪声关联估计准确率和 Rank-1 准确率同时提高, 表明它们之间存在正相关. 此外, 高达 0.95 的估计准确率突出了该方法的有效性.

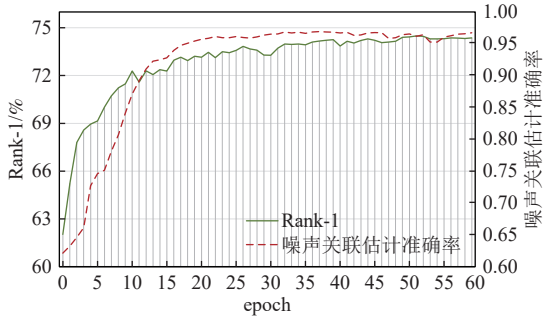
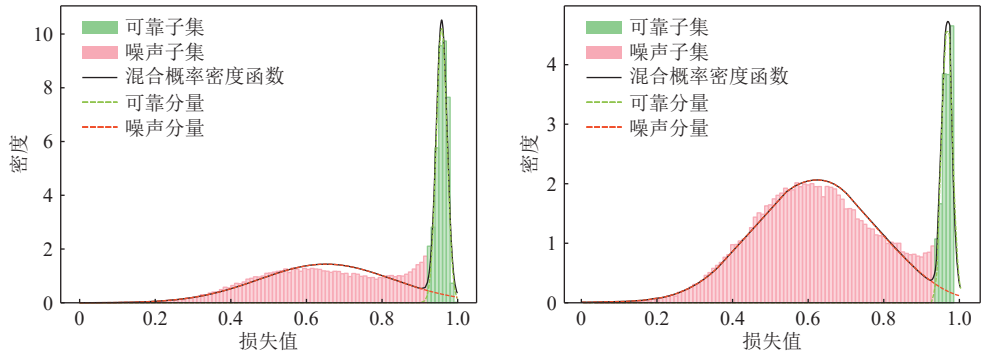


Fig. 4 Visualization of the noisy correspondence estimation accuracy and Rank-1 accuracy in the training process

图 4 噪声关联估计准确率和 Rank-1 准确率的可视化

4)分布的可视化. 为说明高斯混合模型在拟合可靠子集和噪声子集分布的有效性, 本节在图 5 中展示了在 50% 和 80% 噪声比例下损失分布的可视化



(a) 噪声比例为50% (b) 噪声比例为80%

Fig. 5 Visualization of the loss distribution on CUHK-PEDES dataset

图 5 在 CUHK-PEDES 数据集上的损失分布可视化

结果. 实验结果表明, 2 类子集损失呈现显著双峰特性, 且噪声样本的分布峰始终右偏于可靠样本, 这与理论预期的标注噪声导致样本损失升高的特性高度吻合. 此外, 在 2 种噪声场景下高斯混合模型均能精准捕捉分布特征, 且当噪声率从 50% 提升至 80% 时, 噪声子集的占比也随之增加, 验证了所提出的噪声识别方法的有效性.

5)模型运行效率分析. 为了说明本文方法的可部署性, 本节通过表 5 展示了在 50% 噪声比例下, 本文方法与最流行的文本-图像行人检索基线方法 IRRA 及最优噪声关联学习方法 RDE 之间的详细对比. 所有实验均在相同数据集(CUHK-PEDES)、相同硬件环境(NVIDIA Tesla V100)、相同批量大小(64)等条件下进行, 比较指标包括模型参数量、训练过程最大内存占用、每次迭代平均训练时间和 Rank-1 检索性能等. 从表 5 可以看出, 尽管 IRRA 方法在训练时间上表现最佳, 但其 Rank-1 检索性能较低且内存占用较高, 这限制了其在噪声数据环境下的鲁棒性和大规模应用潜力. 而 RDE 方法通过引入噪声鲁棒损失, 显著提升了检索性能, 但也带来了模型参数量的增加. 相比之下, 本文方法在保持较小模型参数量和适中内存占用的同时, 达到了 74.50% 的 Rank-1 检索性能, 充分证明了其在复杂噪声数据场景下的优势与应用前景.

Table 5 Running Efficiency of Model on CUHK-PEDES Dataset with 50% Noise

表 5 CUHK-PEDES 数据集上的含 50% 噪声的模型运行效率

方法	模型参数	内存占用/GB	每批量训练时间/s	Rank-1/%
IRRA	191×10^6	12.79	0.398	62.41
RDE	153×10^6	8.72	0.647	71.33
本文方法	150×10^6	11.60	0.696	74.50

4 结 论

本文针对现有文本-图像行人检索任务受限于跨模态数据中噪声关联问题, 提出了一种语义感知噪声关联学习框架. 该框架通过设计语义感知噪声识别效用函数有效识别可靠样本对和噪声样本对. 同时, 采用互补学习范式, 对可靠样本对采取正向学习策略, 借助对比损失强化语义对齐能力. 对于噪声样本对采用反向学习, 鼓励模型在特征空间中将不匹配的样本对进行区分, 防止噪声样本主导梯度方向. 本文提出的方法在保证模型有效学习判别能力的同时, 大幅降低了噪声关联带来的负面影响. 在 3 个广泛使用的基准数据集上的大量实验均取得了优越的性能, 验证了所提出方法的有效性.

作者贡献声明: 胡冰玉提出了算法思路、实验方案以及撰写论文; 徐艺心辅助完成实验, 负责撰写论文; 余珊负责撰写和润色论文; 赵巨峰指导实验完成, 提出论文修改意见; 杨宇翔提出指导意见并修改论文.

参 考 文 献

[1] Li Shuang, Xiao Tong, Li Hongsheng, et al. Person search with natural language description[C]//Proc of the IEEE Conf on computer vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1970−1979

[2] Jiang Ding, Ye Mang. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 2787−2797

[3] Ding Zefeng, Ding Changxing, Shao Zhiyin, et al. Semantically self-aligned network for text-to-image part-aware person re-identification[J]. arXiv preprint, arXiv: 2107.12666, 2021

[4] Shu Xiujun, Wen Wei, Wu Haoqian, et al. See finer, see more: Implicit modality alignment for text-based person retrieval[C]//Proc of European Conf on Computer Vision. Berlin: Springer, 2022: 624−641

[5] Yan Shuanglin, Dong Neng, Zhang Liyan, et al. Clip-driven fine-grained text-image person re-identification[J]. IEEE Transactions on Image Processing, 2023, 32: 6032−6046

[6] Qin Yang, Chen Yingke, Peng Dezhong, et al. Noisy-correspondence learning for text-to-image person re-identification[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 27197−27206

[7] Qin Yang, Chen Yingke, Peng Dezhong, et al. Learning with noisy correspondence for cross-modal matching[C]//Proc of Neural Information Processing Systems. New York: Curran Associates, 2021: 29406−29419

[8] Yang Shuo, Xu Zhaopan, Wang Kai, et al. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 19883−19892

[9] Zhao Zihua, Chen Mengxi, Dai Tianjie, et al. Mitigating noisy correspondence by geometrical structure consistency learning[C]// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 27381−27390

[10] Arpit D, Jastrzebski S, Ballas N, et al. A closer look at memorization in deep networks[C]//Proc of the 34th Int Conf on Machine Learning. New York: PMLR, 2017: 233−242

[11] Zheng Zhedong, Zheng Liang, Garrett M, et al. Dual-path convolutional image-text embeddings with instance loss[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2020, 16(2): 1−23

[12] Jing Ya, Si Chenyang, Wang Junbo, et al. Pose-guided multi-granularity attention network for text-based person search[C]// Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 11189–11196

[13] Zhang Ying, Lu Huchuan. Deep cross-modal projection learning for image-text matching[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 686–701

[14] Zheng Kecheng, Liu Wu, Liu Jiawei, et al. Hierarchical gumbel attention network for text-based person search[C]//Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 3441–3449

[15] Zhang Lu, Cao Feng, Liang Xinyan, et al. Cross-modal retrieval with correlation feature propagation[J]. Journal of Computer Research and Development, 2022, 59(9): 1993−2002 (in Chinese)
(张璐, 曹峰, 梁新彦, 等. 基于关联特征传播的跨模态检索[J]. 计算机研究与发展, 2022, 59(9): 1993−2002)

[16] Yan Shuanglin, Tang Hao, Zhang Liyan, et al. Image-specific information suppression and implicit local alignment for text-based person search[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 35(12): 17973−17986

[17] Lu Xiaokai, Feng Jun, Han Yongqiang, et al. GraphMLP-Mixer: A Graph-MLP architecture for efficient multi-behavior sequential recommendation method[J]. Journal of Computer Research and Development, 2024, 61(8): 1917−1929 (in Chinese)
(卢晓凯, 封军, 韩永强, 等. GraphMLP-Mixer: 基于图-多层感知机架构的高效多行为序列推荐方法[J]. 计算机研究与发展, 2024, 61(8): 1917−1929)

[18] Bao Cuizhu, Ding Kai, Dong Jianfeng, et al. Research progress of video question answering technologies[J]. Journal of Computer Research and Development, 2024, 61(3): 639−673 (in Chinese)
(包翠竹, 丁凯, 董建锋, 等. 视频问答技术研究进展[J]. 计算机研究与发展, 2024, 61(3): 639−673)

[19] Liu Mingyang, Wang Ruomei, Zhou Fan, et al. Video question answering scheme base on multimodal knowledge active learning[J]. Journal of Computer Research and Development, 2024, 61(4): 889−902 (in Chinese)
(刘明阳, 王若梅, 周凡, 等. 基于多模态知识主动学习的视频问答方案[J]. 计算机研究与发展, 2024, 61(4): 889−902)

[20] Han Haochen, Miao Kaiyao, Zheng Qinghua, et al. Noisy correspondence learning with meta similarity correction[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 7517−7526

[21] Yang Mouxing, Huang Zhenyu, Hu Peng, et al. Learning with twin noisy labels for visible-infrared person re-identification[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 14308−14317

[22] Oord A, Li Yazhe, Vinyals O. Representation learning with contrastive predictive coding[J]. arXiv preprint, arXiv: 1807.03748, 2018

[23] Permuter H, Francos J, Jermyn I. A study of Gaussian mixture models of color and texture features for image classification and segmentation[J]. Pattern Recognition, 2006, 39(4): 695−706

[24] Liu Sheng, Niles-Weed J, Razavian N, et al. Early-learning regularization prevents memorization of noisy labels[J]. Advances in Neural Information Processing Systems, 2020, 33: 20331−20342

[25] Qin Yang, Peng Dezhong, Peng Xi, et al. Deep evidential learning with noisy correspondence for cross-modal retrieval[C]//Proc of the 30th ACM Int Conf on Multimedia. New York: ACM, 2022: 4948–4956

[26] Hu Peng, Huang Zhenyu, Peng Dezhong, et al. Cross-modal retrieval with partially mismatched pairs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 9595−9610

[27] Ding Zefeng, Ding Changxing, Shao Zhiyin, et al. Semantically self-aligned network for text-to-image part-aware person reidentification[J]. arXiv preprint, arXiv: 2107.12666, 2021

[28] Zhu Aichun, Wang Zijie, Li Yifeng, et al. DSSL: Deep surroundings-person separation learning for text-based person retrieval[C]//Proc of the 29th ACM Int Conf on Multimedia. New York: ACM, 2021: 209–217

[29] Radford A, Kim J, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proc of the 38th Int Conf on Machine Learning. New York: PMLR, 2021: 8748−8763



Hu Bingyu, born in 1997. PhD. Member of CCF. Her main research interests include pattern recognition and computer vision.

胡冰玉, 1997 年生. 博士. CCF 会员. 主要研究方向为模式识别、计算机视觉.



Xu Yixin, born in 2002. Master candidate. Her main research interests include pattern recognition and computer vision.

徐艺心, 2002 年生. 硕士研究生. 主要研究方向为模式识别、计算机视觉.



Yu Shan, born in 2001. Master candidate. Her main research interests include pattern recognition and computer vision.

余　珊, 2001 年生. 硕士研究生. 主要研究方向为模式识别、计算机视觉.



Zhao Jufeng, born in 1985. PhD, associate professor, master supervisor. His main research interests include computational imagining and image processing.

赵巨峰, 1985 年生. 博士, 副教授, 硕士生导师. 主要研究方向为计算成像、图像处理.



Yang Yuxiang, born in 1987. PhD, professor, PhD supervisor. His main research interests include computer vision and artificial intelligence.

杨宇翔, 1987 年生. 博士, 教授, 博士生导师. 主要研究方向为计算机视觉、人工智能.