引用格式：林云博, 丁金, 陆庚有, 等. 文图跨模态知识蒸馏的行人检索算法 [J]. 电子科技大学学报, 2026, 55(4): 518-529.
Lin Yunbo, Ding Jin, Lu Gengyou, et al. Research on person retrieval algorithm based on text-to-image cross-modal knowledge distillation[J]. Journal of University of Electronic Science and Technology of China, 2026, 55(4): 518-529. (in Chinese)

文图跨模态知识蒸馏的行人检索算法

林云博[1]，丁 金[2]，陆庚有[1]，周 雪[1*]，邹见效[1]

（1. 电子科技大学（深圳）高等研究院，深圳 518110；2. 电子科技大学自动化工程学院，成都 611731）

摘要：当前文图跨模态行人检索算法过于依赖视觉语言预训练大模型来提升精度，导致模型参数量庞大、算力要求高等问题，难以满足边缘部署等实际应用需求。基于此，该文从视觉语言预训练大模型的轻量化角度入手，引入多模态联合蒸馏与互补式监督策略，基于华为昇腾平台，提出了一种基于三阶段渐进式知识蒸馏的文图跨模态行人检索算法。该方法区别于传统两阶段仅对各个模态进行独立蒸馏，而是通过模态内–跨模态的蒸馏路径，依次在图像、文本模态内实现特征对齐，最终在共享隐空间中进行跨模态语义关联的协同蒸馏。实验表明，学生模型参数量仅为教师模型的 14.77%，在 CUHK-PEDES、ICFG-PEDES 和 RSTPReid 数据集上与现有轻量化方法相比，其 mAP 指标均得到有效提升，同时在边缘设备上处理单个文本–图像对的耗时约 23 ms，实现了实时推理性能。该研究证实了基于文图跨模态知识蒸馏的行人检索算法的有效性，为人员排查、智能安防等场景国产化落地提供了有效路径。

关键词：文图跨模态；行人检索；知识蒸馏；边缘设备

中图分类号：TP391 **文献标志码：**A DOI: 10.12178/1001-0548.2025233

Research on person retrieval algorithm based on text-to-image cross-modal knowledge distillation

LIN Yunbo[1], DING Jin[2], LU Gengyou[1], ZHOU Xue[1*], and ZOU Jianxiao[1]

(1. Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518110, China;
2. School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract: Current text-to-image cross-modal person retrieval algorithms overly rely on vision-language pre-trained large models to improve accuracy. This leads to issues such as large model parameter sizes and high computational requirements, making it difficult to meet the practical application requirements such as edge deployment. In light of this, this paper focuses on the lightweight aspect of vision-language pre-trained large models. It introduces a multi-modal joint distillation and complementary supervision strategy. Based on Huawei Ascend platform, the paper proposes a text-image cross-modal pedestrian retrieval algorithm based on a three-stage progressive knowledge distillation approach. Unlike the traditional two-stage approach that only performs independent distillation for each modality, this method first conducts distillation within the image and text modalities separately, and finally performs collaborative distillation for cross-modal semantic association in the shared latent space. Experimental results show that the student model has only 14.77% of the parameters of the teacher model. Compared with existing lightweight methods, it achieves effective improvements in mAP on the CUHK-PEDES, ICFG-PEDES, and RSTPReid datasets. It also realizes real-time inference performance with an inference time of approximately 23 ms for processing a single text-image pair on edge devices. This study confirms the effectiveness of the three-stage progressive knowledge distillation method and provides an effective pathway for the domestic implementation in scenarios such as personnel screening and intelligent security.

Key words: text-to-image cross-modal; person retrieval; knowledge distillation; edge devices

随着视频监控技术的普及，全球范围内时刻都在产生海量的监控数据，面对如此庞大的视频数据，在实际应用场景中难以依赖人工进行逐帧分析。行人重识别（person re-identification, ReID）技术根据给定的查询图像在非重叠摄像头网络下检索特定行人目标，已广泛应用于智能视频监控领域。

收稿日期：2025−11−20
基金项目：国家自然科学基金（62372082）；中央高校基本科研业务费（ZYGX2024Z017）；深圳市自然科学基金（JCYJ20240813114206010）
作者简介：林云博，主要从事计算机视觉方面的研究。
*通信作者 E-mail: zhouxue@uestc.edu.cn

然而，在缺乏查询图像的场景下，传统 ReID 技术难以发挥作用。得益于多模态技术的快速发展，文图跨模态行人检索为上述挑战提供了新的解决方案[1]。文图跨模态行人检索旨在利用文本描述，从行人图像库中检索出与该描述语义匹配的目标图像。文图跨模态行人检索方法通过融合自然语言处理和计算机视觉技术，实现了文本与图像之间的跨模态信息对齐，有效地弥补了传统 ReID 技术只依赖图像单模态进行检索的不足，在智能安防、人员排查和智能相册等领域中展现出巨大的社会价值和经济潜力。

在早期的文图跨模态行人检索研究方法中，研究者们主要通过引入特定的网络模块或复杂预处理流程来提升模型性能。大量基于卷积神经网络[2-3]和循环神经网络[4]构建双流编码器的方法涌现，通过分别用独立的编码器对图像和文本进行特征提取，结合复杂的跨模态对齐机制，进而提升模型的检索性能。文献 [5-7] 选择基于相同网络结构提取跨模态特征的方法，进一步缓解文本和图像模态差异带来的挑战。然而，上述策略不仅对检索精度的提升相对有限，而且增加了模型复杂度与计算成本，为模型的边端部署等实际应用带来了一定的挑战。近年来，CLIP 等视觉语言预训练大模型[8-11]得到了广泛研究。这些模型通过在大规模图像−文本对数据集进行预训练，学习到模态对齐的特征表示，在多个下游任务中展现出卓越性能。基于视觉语言大模型的文图跨模态行人检索方法[12-14]应运而生，显著提升了模型的跨模态行人检索能力。虽然基于视觉语言预训练模型设计的文图跨模态行人检索方法在学术界取得了显著成果，但是其平均 140 M 以上参数量的参数规模和高计算成本难以满足实际应用的需求[15]。因此，针对这些具有庞大参数规模的视觉语言大模型，如何设计模型轻量化方法引起了广泛关注。

为了实现视觉语言大模型的轻量化，研究者们通过优化模型结构和参数提出了一系列方法，在保持模型性能的同时，降低计算成本和存储需求，如模型量化、模型剪枝、低秩分解、知识蒸馏（knowledge distillation, KD）[16-22]等。相较于其他神经网络轻量化方法，知识蒸馏方法采用教师−学生模型框架，考虑学生模型的整体网络结构，而非局部参数或结构，利用教师模型提取的知识作为监督信号，提供更丰富的训练信息，具有更好的拓展性。知识蒸馏旨在利用具有高性能的教师模型指导轻量级的学生模型学习，以进一步提升学生模型的判别能力。随着知识蒸馏技术的不断发展，其在多个研究领域展现出强大的应用潜力，如计算机视觉、自然语言处理、语音识别、推荐系统等[23]。针对多模态知识蒸馏中跨模态特征对齐与语义融合的复杂性挑战，研究者们提出了多种视觉语言预训练模型的轻量化方法[17-22]。然而，现有的两阶段知识蒸馏方法对于各个模态进行独立蒸馏，忽略了同时蒸馏图像和文本编码器的跨模型性能学习。同时，通用文图检索的轻量化方法无法满足跨模态行人检索任务的需求，难以继承教师模型对于文本图像细粒度的对齐能力。

本文针对轻量化文图跨模态行人检索任务，提出了一种三阶段渐进式知识蒸馏框架。该框架通过模态内−跨模态的蒸馏路径，依次在图像、文本模态内实现特征对齐，最终在共享隐空间中进行跨模态语义关联的协同蒸馏。具体来说，以双流 CLIP 模型为骨干，构建教师−学生网络架构，并设计逐层递进的蒸馏范式：第一阶段专注于图像模态内的表征一致性，确保学生模型在视觉底层特征空间与教师模型对齐；第二阶段贯彻文本模态内的语义保真性，使学生模型精准捕捉语言细微差异；第三阶段引入跨模态联合蒸馏，旨在引导学生模型在共享的隐空间中，模拟教师模型所建立的深层图文语义关联。最终，通过任务特定损失与跨模态蒸馏损失的互补式监督策略，将大型教师模型中蕴含的复杂跨模态表示，高效地迁移至轻量化学生网络。

1　相关工作

1.1　文图跨模态行人检索方法

文图跨模态行人检索任务的目的是根据给定的文本描述，从图像数据库中检索出与输入文本描述最匹配的行人图像。文图跨模态行人检索方法主要可以分为早期基于卷积神经网络与循环神经网络的方法和基于预训练大模型的方法。

早期基于卷积神经网络和循环神经网络构建双流编码器架构的方法中，文献 [1] 提出一个基于门控制注意力机制的循环神经网络框架，采用双流编码器分别对图像和文本进行编码，其中，图像编码器使用 VGG[2] 网络提取视觉特征，通过卷积神经网络捕获行人的全局外观信息；文本编码器使用长短期记忆网络（long short-term memory, LSTM）[4]对自然语言描述进行编码，建模文本序列的时序依赖关系，动态计算每个单词与视觉特征单元之间的关联权重，为后续研究奠定了基础。文献 [24] 采用残差网络（residual network, ResNet）[3] 作为图像

编码器，提出短语级别的局部特征对齐机制，并结合全局特征对齐优化模型，在基准数据集上取得显著的性能提升。文献 [25] 使用 MobileNet[26] 和 Bi-LSTM[27] 分别作为图像和文本编码器，提出跨模态投影匹配损失函数和跨模态投影分类损失函数，直接在全局特征空间进行投影匹配训练。文献 [5] 在文图跨模态行人检索中使用卷积神经网络，同时提取文本和图像特征，提出基于无监督训练的样例损失函数，将匹配的文本–图像对视为一个类别进行优化。这些方法主要通过引入特定网络模块或复杂预处理流程来提升模型性能，然而，在模态间交互、特征细粒度对齐等方面仍然存在较大的提升空间。得益于 Transformer 架构[28] 强大的跨模态特征表示能力，涌现出一系列视觉语言预训练大模型，这些模型在海量图文对上完成了预训练，具备卓越的跨模态对齐能力与泛化性能，并且能够迁移到多种下游任务。文献 [13] 基于 CLIP 模型[8]，设计了 IRRA 模块，通过文本掩码学习机制增强模型对文本–图像隐式关系的建模能力，同时改进了对比学习损失函数，提升了模型对难分负样本的判别能力，其微调后的双编码器模型在多个基准数据集上取得了优越性能。文献 [14] 基于 ALBEF 模型[10] 引入关系感知和敏感度感知学习分支，有效地提升了模型对噪声干扰下弱正样本对的判别能力。这些大规模预训练模型普遍采用复杂的图像和文本编码器架构，导致模型参数量庞大，如文献 [13-14] 的模型参数量均达到 140 M 以上。

1.2 知识蒸馏方法

在模型轻量化研究领域，知识蒸馏作为一种成熟且高效的手段，已被广泛认可和应用。该技术通过将大型复杂模型（教师模型）的知识迁移到小型简单模型（学生模型）中，实现了以强带弱的模型优化策略。文献 [16] 提出知识蒸馏的概念，利用具有高性能的教师模型指导轻量级的学生模型学习，以进一步提升学生模型的判别能力。根据知识迁移粒度的差异，现有知识蒸馏方法可分为 3 类，即基于输出特征[16]、基于中间特征[29] 和基于关系特征[30] 的知识蒸馏。基于输出特征的知识蒸馏不仅让学生模型学习教师模型的软标签，还直接模仿教师模型在输出层产生的特征表示，从而更全面地继承教师模型的决策能力。基于中间特征的知识蒸馏通过引入教师模型中间层的特征激活作为知识载体，有效缓解深度模型蒸馏中仅依赖末层输出导致的表征能力退化问题。基于关系特征的知识蒸馏则突破传统特征对齐范式，通过建模教师模型内部层间关联矩阵或样本间相似性图谱，实现高阶关系知识的传递。由于基于输出特征的知识蒸馏范式仅需利用教师模型末层软标签即可实现知识迁移，无需调用网络内部，兼具实现简单、计算开销低与即插即用等优势，在实际场景中得到了广泛应用。

近年来，知识蒸馏方法在训练策略和知识迁移方式上不断创新，呈现出多种不同的学习方式。其中，文献 [31] 提出在线蒸馏方法，通过学生模型间的相互知识迁移，实现模型性能的协同提升。文献 [32] 提出自蒸馏方法，利用高层特征图对底层输出进行监督，无需额外教师模型，实现了模型内部的知识迁移。文献 [33] 针对训练初期学生模型难以有效学习教师知识的问题，提出渐进式蒸馏方法。该方法通过同步训练教师和学生模型，并阶段性地引入教师监督信息，实现更平滑的知识迁移过程。现有的通用文图跨模态检索场景下的知识蒸馏算法仅适用于检索图像特征差异显著的目标。Cross-model KD 方法是在文献 [16] 提出的经典知识蒸馏方法基础上，结合对比学习策略，将信息噪声对比损失（information noise contrastive estimation loss, InfoNCE Loss）[34] 作为任务损失，通过构建正样本对与负样本对，强化教师模型与学生模型在特征空间中的对齐，以提升学生模型的泛化能力和表征质量。MoTIS[19] 方法在知识蒸馏领域提出了两阶段知识蒸馏框架，分别对图像编码器和文本编码器进行独立蒸馏，以解决传统多模态联合知识蒸馏性能不足的问题，同时还引入难负样本挖掘策略，提升了模型对困难样本的区分能力。ConaCLIP[20] 方法则是在文献 [19] 的基础上，构建教师模型和学生模型之间全连接的知识交互图，在教师–学生或学生–学生之间进行跨模态学习，深入研究不同监督策略对模型性能的影响。CLIP-KD[21] 方法通过多种蒸馏范式（包括关系、特征、梯度和对比学习范式）来指导小型 CLIP 学生模型从大型预训练教师模型中学习知识，最大化了教师和学生模型之间的特征相似性，通过简单的特征模仿损失和交互式对比学习等方式，实现了高效的跨模态知识迁移。TinyCLIP[18] 方法中的 xsmall 模型保留 CLIP 的双塔结构，通过“减层、减通道”实现极致压缩，凭借与教师模型同构的 ViT/Transformer 骨架在视觉语言预训练模型的知识蒸馏中被广泛使用。虽然这些研究方法在跨模态任务中利用知识蒸馏实现了模型轻量化，但是文图跨模态行人检索任务对于细粒度特征对齐的要求更高，行人图像与文本描述之间的语义关联更为复杂，直接应用上述方法难以获得较好的性能。

2　基于三阶段渐进式知识蒸馏的文图跨模态行人检索算法

2.1　知识蒸馏框架

本文提出的三阶段渐进式知识蒸馏框架，其流程首先利用知识蒸馏技术对学生模型进行初始化，以缓解权重初始化对多模态模型收敛性的影响，并对教师模型进行任务微调，最后采用三阶段蒸馏策略，按照图像编码器蒸馏、文本编码器蒸馏、联合蒸馏的顺序，逐步将教师模型的跨模态对齐能力迁移至轻量化学生模型。

图 1 采用 CUHK-PEDES[1] 数据集的行人数据作为示例，展示了本文三阶段渐进式知识蒸馏框架。图中，f_v^{S} 和 f_t^{S} 分别表示学生模型的图像编码器和文本编码器，$f_v^{\mathrm{T}'}$ 和 $f_t^{\mathrm{T}'}$ 分别表示经过任务微调的教师模型的图像编码器和文本编码器。如图 1 所示，第一阶段仅对图像编码器进行模态内蒸馏，第二阶段仅对文本编码器进行模态内蒸馏。对于两阶段知识蒸馏而言，还会在每个阶段利用对应的图像/文本索引，用于计算 InfoNCE 对比损失，以强化单模态表征。然而，两阶段知识蒸馏方法未引入文图跨模态联合优化，导致学生模型无法继承教师在跨模态特征对齐层面的知识。相比之下，本文增加了第三阶段——多模态联合蒸馏。该阶段同时开放图像与文本分支，通过最小化任务损失与蒸馏损失，实现细粒度特征对齐的知识迁移。因此，本文方法在第一阶段和第二阶段仅执行“教师→学生”的特征蒸馏，不计算 InfoNCE 对比损失，从而无需构建图像/文本索引。此外，由于学生模型的容量远小于教师模型，若同时接受图像模态、文本模态和跨模态三类监督，梯度来源会高度复杂并互相竞争。跨模态特征对齐属于高层次知识迁移，其有效学习依赖于图像与文本编码器在单模态内性能达到饱和之后。基于上述原因和动机，本文采用了“先模态内、再跨模态”的三阶段渐进式策略，本质上是为了解耦不同损失的梯度冲突、减少早期训练中的噪声干扰，并确保跨模态监督在合适的时间点介入。

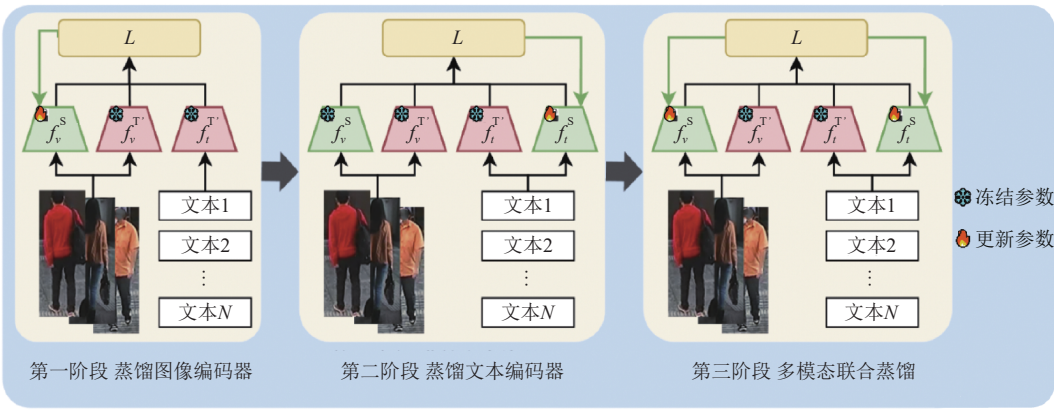


图 1　三阶段渐进式知识蒸馏框架

总体而言，本文的知识蒸馏框架结合文图跨模态行人检索任务的特性，在第一阶段和第二阶段依然采用分阶段独立蒸馏策略，分别对学生模型的图像编码器和文本编码器进行优化，但本文算法中不存在小批量训练中难负样本缺失的问题，因此无需对难负样本进行挖掘；在第三阶段中，引入了文本图像联合知识蒸馏，以充分挖掘教师模型在多模态特征对齐方面的知识，更好地适应文图跨模态行人检索任务的需求。

2.2　多模态联合蒸馏阶段

本文在前两阶段知识蒸馏框架的基础上，新增了文本图像多模态联合蒸馏的第三阶段。这一改进旨在扩展学生模型的学习空间，使其能够更好地捕捉跨模态语义关联。具体而言，第一阶段和第二阶段分别专注于图像和文本单模态的知识迁移，确保学生模型在单一模态中充分吸收教师模型的知识。在第三阶段，则通过设计多模态联合蒸馏损失函数，同时对学生模型的文本编码器和图像编码器进行联合优化，使模型在单模态性能达到饱和后，能够进一步学习跨模态特征对齐能力。这种渐进式的学习策略不仅符合认知规律，也有效提升了模型在文图跨模态行人检索任务中的表现。

第三阶段的损失函数设计如图 2 所示，采用 CUHK-PEDES[1] 数据集的行人数据作为示例，其主要由两个核心部分组成：一是基于学生模型和教师模型交叉模态间的任务损失，二是基于学生模型输出特征与教师模型输出特征间构成的蒸馏损失。其中，任务损失函数确保模型保持基本的检索性能，

蒸馏损失函数则着重于迁移教师模型的判别能力，两者的有机结合提升了模型的整体性能。

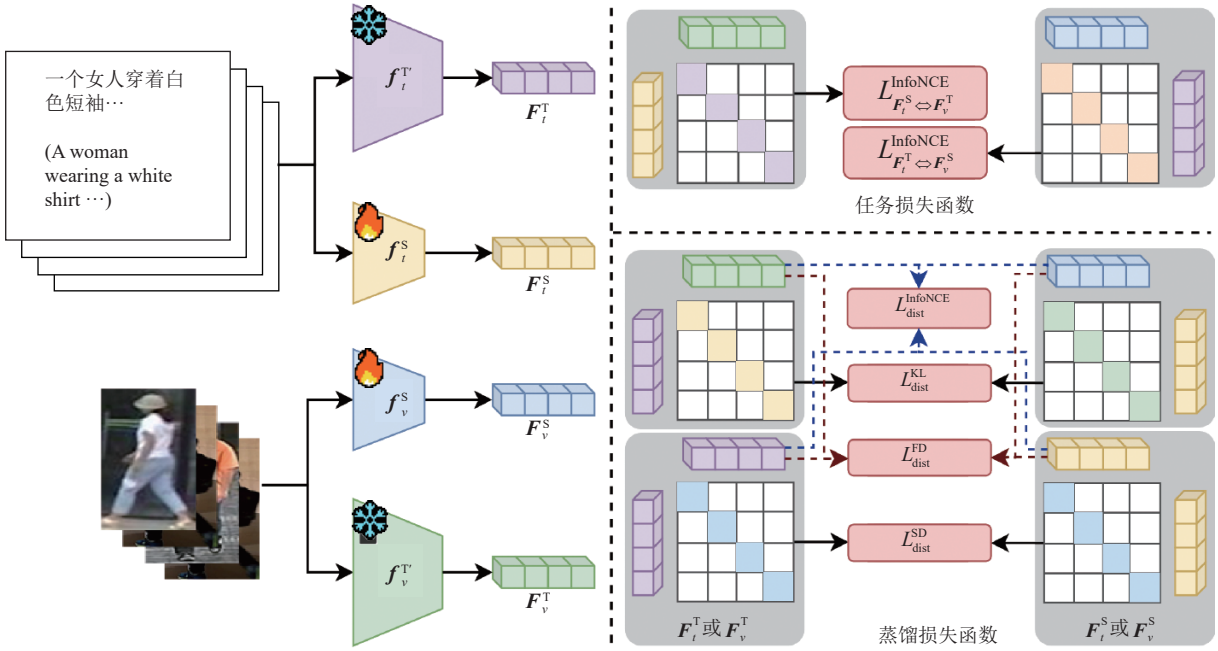


图 2 第三阶段损失函数设计框架示意图

本节主要介绍第三阶段的蒸馏损失。在后续公式中，定义f表示编码器函数，$\boldsymbol{F}$表示经过编码器处理后进行 L2 范数归一化的输出特征。定义输入批量大小为N，d为特征维度大小，文本编码器定义为f_t，图像编码器定义为f_v，教师模型上标为T，学生模型上标为S。通过结合学生模型与教师模型的不同编码器结构，并基于实验效果，本文构建了对比蒸馏损失函数、KL 散度蒸馏损失函数、逐特征距离蒸馏（feature distance, FD）损失函数以及逐相似度距离蒸馏（similarity distance, SD）损失函数作为多模态联合蒸馏阶段的蒸馏损失。这些蒸馏损失函数分别用于在单模态间拟合教师模型的预测概率分布、跨模态匹配间拟合教师模型的预测概率分布、缩小学生模型与教师模型在特征提取上的差距，以及提升学生模型在单模态内对细粒度特征的提取能力。

具体而言，在对比蒸馏损失函数计算中，定义$\boldsymbol{F}^a,\boldsymbol{F}^b\in\boldsymbol{R}^{N\times d}$（$a,b\in\{\mathrm{S},\mathrm{T}\}$）为两个编码器经过 L2 范数归一化后的输出特征。特征$\boldsymbol{F}^a$到特征$\boldsymbol{F}^b$的对比损失计算，表达式为：

$$L^{\mathrm{InfoNCE}}_{\boldsymbol{F}^a\to\boldsymbol{F}^b}=-\frac{1}{N}\sum_{i=1}^{N}\log\left(p_{i,i}\left(\boldsymbol{F}^a,\boldsymbol{F}^b\right)\right) \tag{1}$$

$$p_{i,j}\left(\boldsymbol{F}^a,\boldsymbol{F}^b\right)=\frac{\exp\left(\boldsymbol{F}^a_i(\boldsymbol{F}^b_j)^{\mathrm{T}}/\tau\right)}{\sum_{k=1}^{N}\exp\left(\boldsymbol{F}^a_i(\boldsymbol{F}^b_k)^{\mathrm{T}}/\tau\right)} \tag{2}$$

式中，τ为温度系数，用于控制概率分布的平滑程度；p为匹配输出概率。进一步地，特征$\boldsymbol{F}^a$与特征$\boldsymbol{F}^b$之间的对比损失函数表达式为：

$$L^{\mathrm{InfoNCE}}_{\boldsymbol{F}^a\Leftrightarrow\boldsymbol{F}^b}=\frac{L^{\mathrm{InfoNCE}}_{\boldsymbol{F}^a\to\boldsymbol{F}^b}+L^{\mathrm{InfoNCE}}_{\boldsymbol{F}^b\to\boldsymbol{F}^a}}{2} \tag{3}$$

学生模型与教师模型在单模态间的对比蒸馏损失函数的表达式为：

$$L^{\mathrm{InfoNCE}}_{\mathrm{dist}}=L^{\mathrm{InfoNCE}}_{\boldsymbol{F}^{\mathrm{S}}_v\to\boldsymbol{F}^{\mathrm{T}}_v}+L^{\mathrm{InfoNCE}}_{\boldsymbol{F}^{\mathrm{S}}_t\to\boldsymbol{F}^{\mathrm{T}}_t} \tag{4}$$

在 KL 散度蒸馏损失函数中，定义特征$\tilde{\boldsymbol{F}}^a$和特征$\tilde{\boldsymbol{F}}^b$的匹配输出概率分布为目标概率分布，特征$\boldsymbol{F}^a$与特征$\boldsymbol{F}^b$的匹配输出概率分布为预测概率分布，其维度均为$\boldsymbol{R}^{N\times d}$。预测概率分布向目标概率分布拟合的 KL 散度计算公式为：

$$L^{\mathrm{KL}}_{\boldsymbol{F}^a\to\boldsymbol{F}^b\tilde{\boldsymbol{F}}^a\to\tilde{\boldsymbol{F}}^b}=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}p_{i,j}\left(\boldsymbol{F}^a,\boldsymbol{F}^b\right)\log\frac{p_{i,j}\left(\boldsymbol{F}^a,\boldsymbol{F}^b\right)}{p_{i,j}\left(\tilde{\boldsymbol{F}}^a,\tilde{\boldsymbol{F}}^b\right)} \tag{5}$$

则学生模型的预测概率分布向教师模型的目标概率分布拟合的 KL 散度蒸馏损失函数计算公式为：

$$L^{\mathrm{KL}}_{\mathrm{dist}}=\frac{L^{\mathrm{KL}}_{\boldsymbol{F}^{\mathrm{S}}_t\to\boldsymbol{F}^{\mathrm{S}}_v\boldsymbol{F}^{\mathrm{T}}_t\to\boldsymbol{F}^{\mathrm{T}}_v}+L^{\mathrm{KL}}_{\boldsymbol{F}^{\mathrm{S}}_v\to\boldsymbol{F}^{\mathrm{S}}_t\boldsymbol{F}^{\mathrm{T}}_v\to\boldsymbol{F}^{\mathrm{T}}_t}}{2} \tag{6}$$

在逐特征距离蒸馏损失函数的计算中，定义逐特征距离损失函数如式 (7) 所示，可以得到学生模

型和教师模型在文本和图像双模态间的逐特征距离蒸馏损失函数，如式 (8) 所示：

$$L_{\boldsymbol{F}^a \Leftrightarrow \boldsymbol{F}^b}^{\mathrm{FD}} = \frac{1}{2Nd}\sum_{i=1}^{N}\sum_{j=1}^{d}\left(F_{i,j}^{a} - F_{i,j}^{b}\right)^2 \tag{7}$$

$$L_{\mathrm{dist}}^{\mathrm{FD}} = L_{\boldsymbol{F}_t^{\mathrm{S}} \Leftrightarrow \boldsymbol{F}_t^{\mathrm{T}}}^{\mathrm{FD}} + L_{\boldsymbol{F}_v^{\mathrm{S}} \Leftrightarrow \boldsymbol{F}_v^{\mathrm{T}}}^{\mathrm{FD}} \tag{8}$$

在逐相似度距离蒸馏损失函数的计算中，定义逐相似度距离损失函数如式 (9) 所示，可以得到学生模型和教师模型间的逐相似度距离蒸馏损失函数，如式 (10) 所示：

$$L_{\boldsymbol{F}^a \to \boldsymbol{F}^b \Leftrightarrow \tilde{\boldsymbol{F}}^a \to \tilde{\boldsymbol{F}}^b}^{\mathrm{SD}} = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\boldsymbol{F}_i^a(\boldsymbol{F}_j^b)^{\mathrm{T}} - \tilde{\boldsymbol{F}}_i^a(\tilde{\boldsymbol{F}}_j^b)^{\mathrm{T}}\right)^2 \tag{9}$$

$$L_{\mathrm{dist}}^{\mathrm{SD}} = \frac{L_{\boldsymbol{F}_t^{\mathrm{S}} \to \boldsymbol{F}_t^{\mathrm{S}} \Leftrightarrow \tilde{\boldsymbol{F}}_t^{\mathrm{T}} \to \tilde{\boldsymbol{F}}_t^{\mathrm{T}}}^{\mathrm{SD}} + L_{\boldsymbol{F}_v^{\mathrm{S}} \to \boldsymbol{F}_v^{\mathrm{S}} \Leftrightarrow \tilde{\boldsymbol{F}}_v^{\mathrm{T}} \to \tilde{\boldsymbol{F}}_v^{\mathrm{T}}}^{\mathrm{SD}}}{2} \tag{10}$$

2.3　互补式监督策略

为了挖掘文本图像联合蒸馏在提升学生模型细粒度特征对齐方面的潜力，以及实现对学生模型的多层次监督，本文提出了一种互补式监督策略。除了蒸馏损失，本文的损失函数设计还综合考虑到了任务监督的重要性，实现了任务损失与蒸馏损失函数的协同设计。

第一阶段和第二阶段的损失函数设计在 MoTIS 方法的基础上，针对文图跨模态行人检索任务进行了适应性调整。由于该任务主要关注行人类别区分，且公开数据集规模相对较小，因此在训练过程中引入图像增强操作至关重要。基于此，本文在第一阶段和第二阶段的损失函数中移除了难例挖掘分支，其余部分保持与 MoTIS 方法一致。

具体而言，在第一阶段，任务损失定义为学生模型图像编码器与教师模型文本编码器之间的对比学习损失表达式为：

$$L_{\mathrm{task}} = L_{\boldsymbol{F}_t^{\mathrm{T}} \Leftrightarrow \boldsymbol{F}_v^{\mathrm{S}}}^{\mathrm{InfoNCE}} \tag{11}$$

蒸馏损失则定义为学生模型图像编码器输出和教师模型图像编码器输出的概率分布之间的对比损失，表达式为：

$$L_{\mathrm{KD}} = L_{\boldsymbol{F}_v^{\mathrm{S}} \to \boldsymbol{F}_v^{\mathrm{T}}}^{\mathrm{InfoNCE}} \tag{12}$$

在第二阶段，任务损失扩展为学生模型文本编码器与教师模型图像编码器之间的对比学习损失，以及学生模型内部文本与图像编码器之间的对比学习损失，表达式为：

$$L_{\mathrm{task}} = L_{\boldsymbol{F}_t^{\mathrm{S}} \Leftrightarrow \boldsymbol{F}_v^{\mathrm{T}}}^{\mathrm{InfoNCE}} + L_{\boldsymbol{F}_t^{\mathrm{S}} \Leftrightarrow \boldsymbol{F}_v^{\mathrm{S}}}^{\mathrm{InfoNCE}} \tag{13}$$

蒸馏损失除了学生模型的文本编码器输出到教师模型的文本编码器输出的对比损失外，还添加了学生模型文本和图像匹配输出概率分布向教师模型文本和图像匹配输出概率分布拟合的 KL 散度损失，表达式为：

$$L_{\mathrm{KD}} = L_{\boldsymbol{F}_t^{\mathrm{S}} \to \boldsymbol{F}_t^{\mathrm{T}}}^{\mathrm{InfoNCE}} + L_{\mathrm{dist}}^{\mathrm{KL}} \tag{14}$$

在第三阶段，本文针对多个任务损失与蒸馏损失之间可能存在竞争的问题，在任务损失中仅使用学生模型和教师模型交叉模态的对比学习，以缓解学生模型内部的文本和图像编码器在任务学习上的竞争，以及学生模型和教师模型之间交叉模态编码器在任务学习上的竞争，表达式为：

$$L_{\mathrm{task}} = L_{\mathrm{cross}}^{\mathrm{InfoNCE}} = L_{\boldsymbol{F}_t^{\mathrm{T}} \Leftrightarrow \boldsymbol{F}_v^{\mathrm{S}}}^{\mathrm{InfoNCE}} + L_{\boldsymbol{F}_t^{\mathrm{S}} \Leftrightarrow \boldsymbol{F}_v^{\mathrm{T}}}^{\mathrm{InfoNCE}} \tag{15}$$

蒸馏损失则综合了第一阶段和第二阶段的蒸馏损失，并引入了逐特征距离蒸馏损失和逐相似度距离蒸馏损失，分别用于缩小学生模型和教师模型提取特征的差距和学生模型向教师模型在单模态内对细粒度特征提取能力的学习，表达式为：

$$L_{\mathrm{KD}} = L_{\mathrm{dist}}^{\mathrm{InfoNCE}} + L_{\mathrm{dist}}^{\mathrm{KL}} + L_{\mathrm{dist}}^{\mathrm{FD}} + \mathrm{L}_{\mathrm{dist}}^{\mathrm{SD}} \tag{16}$$

最终，各阶段的综合损失通过加权求和的方式计算，表达式为：

$$L = \alpha L_{\mathrm{KD}} + (1-\alpha) L_{\mathrm{task}} \tag{17}$$

式中，权重超参数α采用阶段式衰减策略，其原因是在训练初期赋予蒸馏损失更大权重以快速吸收教师模型知识，在训练后期则增加任务损失权重以更好地适应特定任务需求。这种动态权重调整策略能有效平衡知识迁移和任务适应之间的关系，确保模型的整体性能。

3　实验

3.1　数据集与评价指标

本文采用文图跨模态行人检索领域的 3 个主流数据集：CUHK-PEDES[1]、ICFG-PEDES[35] 和 RSTPReid[36]。

CUHK-PEDES 数据集包含 40 206 张行人图像，覆盖 13 003 个不同的行人身份，每张图像都配有 2 条自然语言描述，描述内容涵盖了行人的外观、动作和姿态等细节。数据集划分为训练集、验证集和测试集。其中，训练集包含 11 003 个身份的 34 054 张图像和 68 108 条文本描述；验证集和测试集分别

包含 1 000 个身份，图像数量分别为 3 078 和 3 074，文本描述数量分别为 6 156 和 6 148。

ICFG-PEDES 数据集包含 4 102 个行人身份的 54 522 张图像，每张图像配有唯一的自然语言描述。其中，所有图像均从 MSMT17 数据库收集；每个文本描述平均有 37.2 个字，包含更细粒度的信息描述。数据集分为训练集和测试集。其中，训练集包含 3 102 个身份的 34 674 个图文对，测试集包含 1 000 个身份的 19 848 个图文对。

RSTPReid 数据集包含来自 4 101 个身份的 20 505 张行人图像，这些图像由 15 个摄像头在不同的时间、地点、天气、光照和视角条件下拍摄。每个身份对应 5 张图像，每张图像均配有两条自然语言描述，且每条描述不少于 23 个单词。数据集分为训练集、验证集和测试集，分别包含 3 701、200 和 200 个身份。

本文在对算法进行评测时，采用文图跨模态行人检索领域常用 Rank-K 和平均精度均值（mean average precision, mAP）两项指标。其中 Rank-K 指标中主要选用 Rank-1、Rank-5 和 Rank-10 进行评价。Rank-K 评价指标用于衡量检索系统在前 K 个检索结果中的准确性和有效性。其表达式为：

$$\text{Rank-}K(q)=\begin{cases}1 & k_q \leqslant K\\ 0 & k_q > K\end{cases} \tag{18}$$

$$\text{Rank-}K=\frac{1}{|Q|}\sum_{q\in Q}\text{Rank-}K(q) \tag{19}$$

式中，k_q为查询图像q的第k个匹配结果；$|Q|$代表查询集合的大小，最终的 Rank-K 指标为所有查询的 Rank-K 值的平均值。

mAP 评价指标通过对所有查询的平均精度进行统计，能够有效评估模型在多样化查询场景下的全局性能表现。其表达式为：

$$\text{mAP}=\frac{1}{|Q|}\sum_{q=1}^{Q}\text{AP}_q \tag{20}$$

式中，AP_q是单个查询q的平均精度（average precision, AP）。

3.2 实验设置

针对文图跨模态行人检索任务对文本与图像间细粒度特征对齐的要求，以及 Transformer 在全局与局部信息捕获方面和在多模态信息映射方面的能力。本文在教师模型与学生模型的选取中均采用了基于 Transformer 架构的模型。具体而言，本文选用 IRRA[13] 方法的主干网络作为教师模型。在学生模型的选取中，考虑到文图跨模态行人检索任务的知识蒸馏研究尚处于初步探索阶段，且模型初始化权重对该任务的性能具有决定性影响，本研究借鉴了 TinyCLIP[18] 方法的 xsmall 模型架构。两个模型的网络结构参数如表 1 所示。与教师模型相比，学生模型在参数量上实现了显著压缩，总参数量仅为教师模型的 14.77%。其中，文本编码器参数量仅为教师模型的 22.22%，图像编码器参数量更是降至教师模型的 9.30%。

表 1 教师模型和学生模型的网络结构参数

模型	特征维度	图像编码器					文本编码器				
		图像输入大小	层数	宽度	头数	参数量/M	token输入长度	层数	宽度	头数	参数量/M
教师模型	512	384×128	12	768	12	63	77	12	512	8	86
学生模型	512	384×128	10	256	4	14	77	3	256	4	8

在知识蒸馏的训练过程中，文本输入在序列前后分别添加“<|startoftext|>”和“<|endoftext|>”标记符，以明确界定文本边界，最终统一截断或填充至 77 个 token 的长度。图像输入首先缩放至 384×128 大小，随后进行概率为 50% 的水平翻转，接着执行 10 个像素的边界填充并进行 384×128 的随机裁剪，最后应用区域比例为 0.02～0.40 的随机擦除数据增强。在训练过程中，每个批量的数据同时通过学生模型和固定权重的教师模型进行处理。

在训练参数设置中，本文选用 Adam[37] 优化器进行梯度计算和反向传播，平滑常数β_1=0.9、β_2=0.999，初始学习率为 1×10^{-5}，权重衰减系数为 4×10^{-5}。每个训练阶段持续 60 个 epoch，批量大小固定为 64。学习率调度策略采用线性预热与余弦衰减相结合的方式，前 5 个 epoch 进行从 1×10^{-6} 开始的线性预热，随后进行余弦衰减。任务损失的温度系数设置为τ=0.02，蒸馏损失的温度系数设置为τ=0.5。损失函数的加权系数α采用阶段式衰减策略，其变化曲线如图 3 所示。为了加速模型收敛并提升模型训练效果，学生模型采用 TinyCLIP[18] 方法中的 xsmall 模型进行参数初始化。

本文所有模型训练研究均在华为昇腾 Atlas800 服务器平台上进行。实验平台的软硬件配置如表 2 所示。

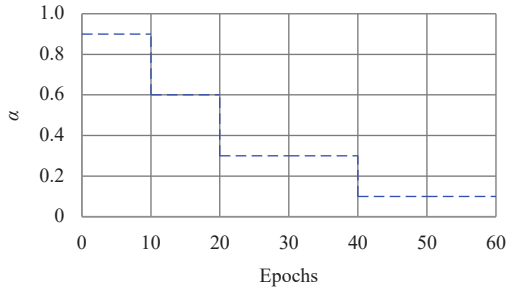


图 3　损失函数加权系数 α 的阶段式衰减曲线图

表 2　模型训练实验平台软硬件信息

类型	名称	型号/版本号
服务器	Atlas800	9 000
硬件	CPU	华为鲲鹏920×4
	NPU	华为昇腾910×8
软件	操作系统	Ubuntu 20.04.6 LTS
	内核	5.4.0-26-generic
	驱动	23.00 商业版
	固件	7.0.0.230 商业版
	CANN	7.0.0.alpha003
	Python	3.8.2
	PyTorch	1.11.0

3.3　消融实验

为系统验证本文提出方法的有效性，本文设计了 4 组消融实验，分别从温度系数优化、损失权重调节策略、任务损失设计以及多模态蒸馏损失组合维度展开分析，深入探讨各关键参数与模块对模型性能的影响机制。

3.3.1　蒸馏损失温度系数

基于文图跨模态行人检索任务对细粒度特征判别能力的特殊要求，本文首先对知识蒸馏过程中的温度系数进行系统优化。根据文献 [38] 的理论分析，温度系数 τ 的降低会增强损失函数对困难负样本的梯度惩罚强度，这对于需要更细粒度匹配的文图跨模态行人检索任务具有重要价值。为此，本文在 Cross-model KD 知识蒸馏方法框架下，系统探究了 $\tau \in \{0.5, 1, 4, 8, 12, 20\}$ 对模型性能的影响。在 CUHK-PEDES 数据集上的实验结果如表 3 所示。

实验结果表明，当 τ=0.5 时，模型在 mAP 指标上取得最佳表现（46.06%），同时在 R5 和 R10 指标上保持次优水平。值得注意的是，虽然 τ=4 时 R1 指标达到峰值 49.81%，但其 mAP 值略低于 τ=0.5 的设置。这种现象表明较低的温度系数有利于提升模型在细粒度匹配中的整体判别能力，而过高温度系数（如 $\tau \geqslant 8$）会导致模型过度关注全局特征而忽视局部细节。基于综合性能评估，后续实验均采用 τ=0.5 作为蒸馏损失的温度参数。

表 3　蒸馏损失温度系数消融实验结果

温度参数 τ	R1/%	R5/%	R10/%	mAP/%
0.5	49.76	73.78	82.83	**46.06**
1.0	49.77	73.57	82.78	46.03
4.0	**49.81**	73.67	82.75	46.05
8.0	49.72	**74.04**	82.70	45.89
12.0	49.68	73.65	82.83	46.02
20.0	49.69	73.69	**82.85**	46.02

3.3.2　损失函数加权系数

本文提出的三阶段渐进式知识蒸馏框架各阶段损失函数计算由任务损失和蒸馏损失加权平均计算所得。为了优化任务损失与蒸馏损失的协同作用机制，本文提出并比较了 3 种损失函数加权系数 α 的调节策略：固定权重（α=0.9）、余弦衰减和阶梯式衰减（如图 3 所示）。其中，余弦衰减策略在训练前期（0～10 epoch）固定 τ=0.9，训练后期（50～60 epoch）固定 τ=0.1，中间阶段通过余弦函数进行衰减调节。实验结果如表 4 所示。

实验结果表明，阶梯式衰减策略在各项指标上均取得最优表现，在 CUHK-PEDES、ICFG-PEDES 数据集上 mAP 值分别达到 46.06%、19.70%，较固定权重策略分别提升 0.98%、1.09%，但由于 RSTPReid 数据集规模较小，跨模态匹配关系相对简单，训练性能基本在 11 epoch 左右已达到最优，3 种策略性能差距不大。该策略在训练初期（0～10 epoch）保持较高蒸馏权重（α=0.9），使学生模型快速吸收教师模型的判别知识；在训练中期（10～40 epoch）逐步降低至 α=0.3，增强模型自主优化能力；最终阶段（40～60 epoch）进一步降至 α=0.1，充分释放学生模型的表征潜力。实验结果证明了这种动态平衡机制有效缓解了知识迁移与模型自适应之间的张力。

表 4　损失函数加权系数调节策略消融实验结果　%

α 调节策略	CUHK-PEDES				ICFG-PEDES				RSTPReid			
	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP
0.9（固定）	48.44	73.29	82.28	45.08	35.70	58.76	68.36	18.61	34.55	**64.35**	74.65	**28.54**
余弦衰减	48.73	73.60	82.41	45.46	35.96	58.96	68.60	18.79	**34.90**	61.70	73.80	28.20
阶梯式衰减	**49.76**	**73.78**	**82.83**	**46.06**	**37.40**	**60.22**	**69.92**	**19.70**	34.70	62.50	**74.85**	28.34

3.3.3 任务损失

该消融实验旨在验证多模态联合知识蒸馏中，在学生模型自身任务对比学习和学生模型与教师模型交叉模态对比学习之间是否存在着竞争关系。在该消融实验中，本文在复现 MoTIS 的模型基础上，添加了第三阶段的多模态联合知识蒸馏。该阶段的蒸馏损失函数仅使用$L_{\text{dist}}^{\text{InfoNCE}}$，以验证其与任务损失$L_{\boldsymbol{F}_t^S \Leftrightarrow \boldsymbol{F}_v^S}^{\text{InfoNCE}}$和$L_{\text{cross}}^{\text{InfoNCE}}$之间的互补设计。实验结果如表 5 所示，其中⓪代表本文复现的 MoTIS 方法。

表 5 任务损失消融实验结果

序号	L_{task}		L_{KD}	CUHK-PEDES			
	$L_{\boldsymbol{F}_t^S \Leftrightarrow \boldsymbol{F}_v^S}^{\text{InfoNCE}}$	$L_{\text{cross}}^{\text{InfoNCE}}$	$L_{\text{dist}}^{\text{InfoNCE}}$	R1/%	R5/%	R10/%	mAP/%
⓪	×	×	×	55.49	78.43	86.18	50.95
①	√	×	√	58.48	**80.62**	**87.39**	53.51
②	×	√	√	**58.82**	80.52	87.35	**53.55**
③	√	√	√	57.77	79.61	87.33	52.77

实验结果表明，②与①的结果相当，但从系统整体性能指标 mAP 来看，②表现更优。这体现了第三阶段知识蒸馏中采用的学生模型与教师模型交叉模态对比学习的有效性。然而，在③中，同时使用学生模型自身任务对比学习和学生模型与教师模型交叉模态对比学习反而导致性能略有下降，表明这两种任务损失在知识蒸馏过程中存在竞争关系。具体而言，学生模型自身任务对比学习鼓励学生模型在自身表示空间中建立判别边界，而学生模型与教师模型交叉模态对比学习则要求学生模型对齐教师模型已形成的跨模态语义结构。在一些情况下，前者推动学生向自身最优学习，而后者约束其贴合教师的对齐结构，两者施加方向不一致的梯度约束，表现为检索性能下降。因此，采用单一任务损失与蒸馏损失的互补学习策略能够更有效地提升学生模型性能。

3.3.4 蒸馏损失

为进一步验证第三阶段多模态联合知识蒸馏对学生模型整体性能的影响，以任务损失消融实验的②为基础（记为L_{base}），研究联合文本图像多模态的蒸馏损失$L_{\text{dist}}^{\text{KL}}$、$L_{\text{dist}}^{\text{FD}}$和$L_{\text{dist}}^{\text{SD}}$对学生模型性能的影响。实验结果如表 6 所示。

表 6 蒸馏损失消融实验结果

序号	$L_{\text{dist}}^{\text{KL}}$	$L_{\text{dist}}^{\text{FD}}$	$L_{\text{dist}}^{\text{SD}}$	CUHK-PEDES			
				R1/%	R5/%	R10/%	mAP/%
⓪	×	×	×	58.82	80.52	87.35	53.55
①	√	×	×	59.31	80.90	87.75	53.94
②	×	√	×	58.82	80.51	87.46	53.56
③	×	×	√	59.15	80.60	87.62	53.77
④	√	√	×	59.32	80.88	87.67	53.94
⑤	×	√	√	59.06	80.51	87.62	53.76
⑥	√	×	√	59.44	81.03	87.78	54.07
⑦	√	√	√	**59.52**	**81.08**	**87.80**	**54.12**

实验结果表明，①的提升最稳定，相对基线在 mAP 上带来明显增益，②的各性能指标几乎与基线持平，说明逐特征距离蒸馏损失的单独贡献较弱，③有一定提升，但幅度低于①，最终⑦达到全局最优，体现出本文蒸馏损失间的互补性。该结果表明在L_{base}基础上添加多模态蒸馏损失函数能够进一步提升学生模型性能，可以归因于联合多模态知识蒸馏扩大了学生模型向教师模型学习的空间，从而增强了模型对特定任务的理解能力。最终，本文在第三阶段采用的损失函数设计与⑦一致。

3.4 和现有前沿蒸馏方法对比

为验证本文算法的先进性，本文对 InfoNCE 对比损失函数微调学生模型的方法、Cross-model KD[16] 方法、本文复现的两阶段知识蒸馏方法 MoTIS[19]、ConaCLIP[20] 方法和 CLIP-KD[21] 方法进行了对比实验，实验结果如表 7 所示。

表 7 跨模态蒸馏方法性能对比

%

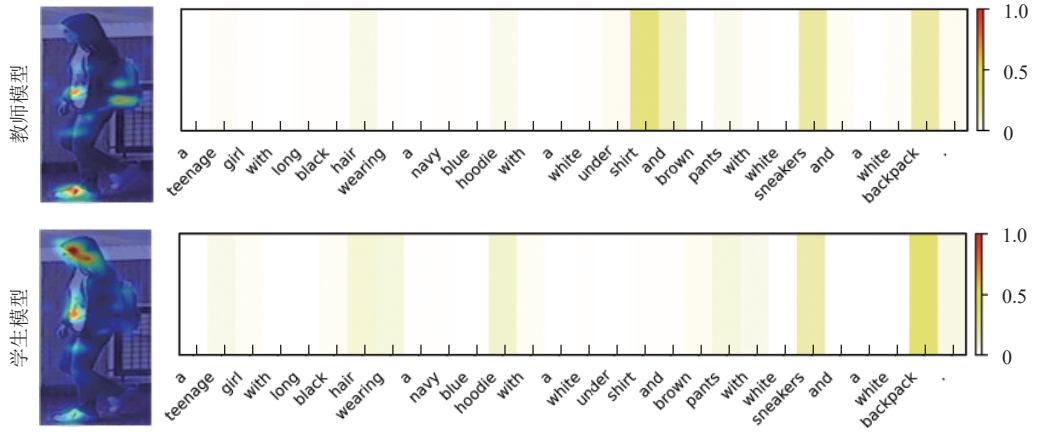
方法	CUHK-PEDES				ICFG-PEDES				RSTPReid			
	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP
InfoNCE-based[8]	50.20	73.78	82.31	46.31	36.58	59.11	69.13	19.12	34.75	62.65	74.80	27.97
Cross-model KD[16]	49.76	73.78	82.83	46.06	37.40	60.22	69.92	19.70	34.70	62.50	74.85	28.34
MoTIS[19]	55.49	78.43	86.18	50.95	42.15	64.74	73.28	22.63	39.30	64.20	75.95	31.49
ConaCLIP[20]	55.98	78.48	86.26	51.36	42.60	65.15	74.01	23.14	39.20	63.85	76.25	31.13
CLIP-KD[21]	53.79	77.76	85.58	50.33	41.58	64.16	72.93	22.50	39.95	67.00	78.60	32.07
本文	**59.52**	**81.08**	**87.80**	**54.12**	**47.63**	**69.06**	**76.72**	**25.77**	**42.00**	**68.75**	**79.05**	**33.52**
教师模型	73.70	89.56	93.58	66.24	63.80	80.43	85.78	38.22	59.55	82.10	88.40	47.07

实验结果表明，本文方法在 3 个数据集上分别保留了教师模型 mAP 指标的 81.70%、67.43%、71.21%，相较于改进基线 MoTIS 方法，mAP 指标分别提升了 3.17%、3.14% 和 2.03%。在 R1 指标上，本文方法较传统的 InfoNCE-based 对比学习微调方法获得了大幅度的性能提升，在 3 个数据集上

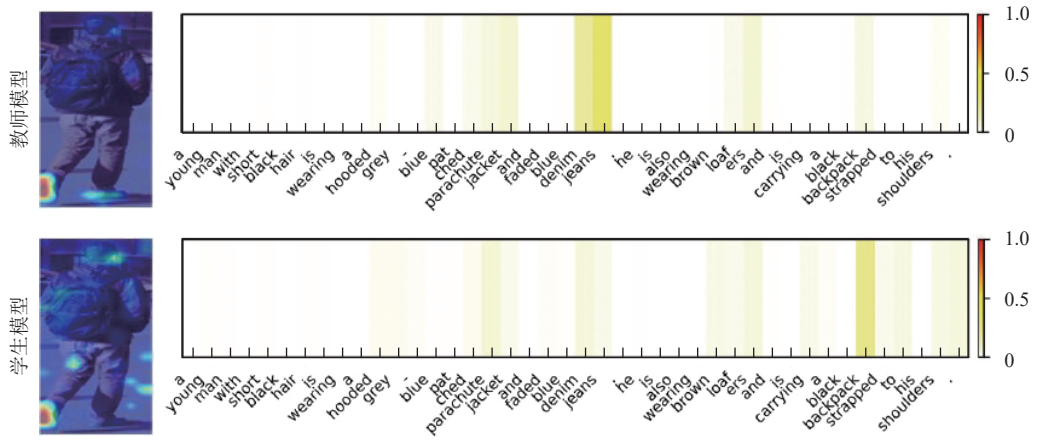
分别提升了 9.32%、11.05%、7.25%。充分验证了本文提出的三阶段渐进式知识蒸馏框架在模型压缩与性能保持方面的有效性。

3.5　可视化分析

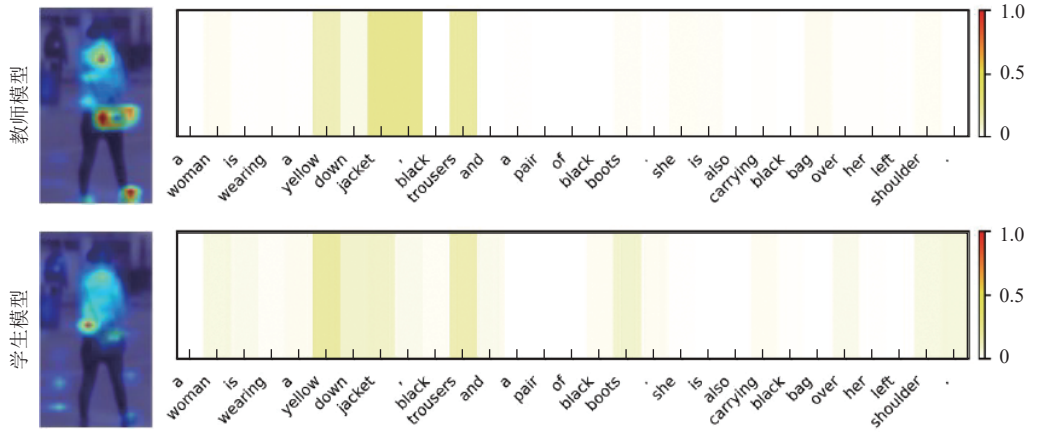
为了深入评估本文提出的三阶段渐进式知识蒸馏框架的跨模态对齐能力，本文从热力图可视化展开系统性分析。本文使用 Grad-CAM[39] 方法对学生模型和教师模型双编码器在最后一个注意力层中对图像和文本信息关注度进行热力图可视化，可视化结果如图 4 所示。



a. CUHK-PEDES测试样例热力图可视化



b. ICFG-PEDES测试样例热力图可视化



c. RSTPReid测试样例热力图可视化

图 4　测试样例热力图可视化

在图像编码器热力图可视化中，学生模型在最后一个注意力层中在 3 个数据集的测试样例中均体现出与教师模型具有部分相同的局部关注信息，如图 4a 中行人的“衬衫”“长裤”和“鞋子”部位，图 4b 中行人“鞋子”和“头部”部位，图 4c 中行人的“外套”“挎包”和“鞋子”部位。在文

本编码器热力图可视化中，学生模型在最后一层也体现出与教师模型具有部分相同的局部信息关注能力，如图 4a 的“backpack”“sneakers”和“hoodie”等局部词语，图 4b 的“ers”和“backpack”等局部词语，图 4c 的“yellow”“jacket”和“trousers”等局部词语。此外，本文发现学生模型在文本编码器最后一层对偏短的文本描述（如图 4a 和图 4c 文本描述）能够具有和教师模型更相近的局部词语关注信息，而在偏长文本（如图 4b 文本描述）学生模型关注的信息与教师模型相比会更分散。

综合 3 个数据集的测试样例的热力图可视化结果而言，基于本文提出的三阶段渐进式知识蒸馏框架训练得到的学生模型，在双编码器的最后一个注意力层中能够有效学习到与教师模型部分相似的局部信息关注能力。由于学生模型的参数量显著少于教师模型，其在最后一层的信息关注能力与教师模型相比存在一定差异。在图像和文本编码中，学生模型对局部信息的关注范围更为宽泛。

3.6 推理耗时测试

在 Atlas200 上对教师模型和学生模型进行单样例推理耗时测试，并分别统计了文本编码器和图像编码器 OM 离线模型推理耗时，如表 8 所示。

实验结果表明，本文轻量化的学生模型在边缘设备上的推理性能达到实时处理水平，处理单个文本−图像对的推理时间约为 23 ms。其中，文本和图像单样例推理时间分别约为 2 ms 和 20 ms，与教师模型相比推理耗时分别缩短了 5.78 和 3.47 倍。

表 8 边缘设备推理耗时对比

ms

类别	文本编码器耗时	图像编码器耗时
教师模型	14.22	71.06
学生模型	**2.46**	**20.45**

4 结束语

本文提出了一种基于三阶段渐进式知识蒸馏的文图跨模态行人检索算法，通过引入多模态联合蒸馏阶段，使模型在单模态性能达到饱和后，能够进一步学习跨模态特征对齐能力。同时提出互补式监督策略，通过任务损失函数和蒸馏损失函数的有机结合，强化了学生模型与教师模型在跨模态学习中的关联。该框架通过分阶段的蒸馏过程，逐步提升学生模型细粒度特征对齐的能力，加强了跨模态语义理解方面的能力，实现了文图跨模态行人检索模型的轻量化。与其他方法相比，本文方法能够在有效降低模型参数量和计算复杂度的同时，保留较高的检索精度。在公共数据集上，本文进行了消融实验，与现有前沿方法的对比实验和可视化实验，行人检索任务达到了实时处理水平，验证了本文方法的可行性和优越性。在未来的工作中，本文将探索更加精细的蒸馏策略，如在网络中间层引入多模态联合蒸馏，进一步提升学生模型在细粒度特征对齐方面的能力。此外，针对复杂场景下的实时性与鲁棒性，以及模型在不同跨模态任务中的泛化能力，值得开展进一步研究。

参 考 文 献

[1] Li Shuang, Xiao Tong, Li Hongsheng, et al. Person search with natural language description[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 5187-5196.

[2] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc. , 2012: 1097-1105.

[3] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.

[4] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[5] 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 (TOMM), 2020, 16(2): 51.

[6] Ma Tinghuai, Yang Mingming, Rong Huan, et al. Dual-path CNN with max gated block for text-based person re-identification[J]. Image and Vision Computing, 2021, 111: 104168.

[7] Farooq A, Awais M, Kittler J, et al. AXM-Net: Implicit cross-modal feature alignment for person re-identification [C]//Proceedings of the 36th AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2022: 4477-4485.

[8] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C]//Proceedings of the 38th International Conference on Machine Learning. [S.l.]: PMLR, 2021: 8748-8763.

[9] Jia Chao, Yang Yinfei, Xia Ye, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]//Proceedings of the 38th International Conference on Machine Learning. [S.l.]: PMLR, 2021: 4904-4916.

[10] Li Junnan, Selvaraju R R, Gotmare A D, et al. Align before fuse: Vision and language representation learning with momentum distillation[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. [S.l.]: Curran Associates Inc. , 2021: 742.

[11] Li Junnan, Li Dongxu, Xiong Caiming, et al. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]// Proceedings of the 39th International Conference on Machine Learning. Baltimore: PMLR, 2022: 12888-12900.

[12] Yang Shuyu, Zhou Yinan, Zheng Zhedong, et al. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark[C]//Proceedings

of the 31st ACM International Conference on Multimedia. Ottawa: ACM, 2023: 4492-4501.
[13] Jiang Ding, Ye Mang. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval [C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 2787-2797.
[14] Bai Yang, Cao Min, Gao Daming, et al. RaSa: Relation and sensitivity aware representation learning for text-based person search[C]//Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2023: 62.
[15] 石瑞鑫, 智敏, 殷雁君. 多模态行人重识别研究综述[J]. 计算机应用研究, 2025, 42(7): 1921-1929.
Shi Ruixin, Zhi Min, Yin Yanjun. Review of multimodal pedestrian re-identification[J]. Application Research of Computers, 2025, 42(7): 1921-1929. (in Chinese)
[16] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[EB/OL]. (2015-03-09) [2025-10-11]. https://arxiv.org/abs/1503.02531.
[17] Andonian A, Chen Shixing, Hamid R. Robust cross-modal representation learning with progressive self-distillation [C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 16409-16420.
[18] Wu Kan, Peng Houwen, Zhou Zhenghong, et al. TinyCLIP: CLIP distillation via affinity mimicking and weight inheritance[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 21913-21923.
[19] Ren Siyu, Zhu K. Leaner and faster: Two-stage model compression for lightweight text-image retrieval[C]// Proceedings of 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle: ACL, 2022: 4085-4090.
[20] Wang Jiapeng, Wang Chengyu, Wang Xiaodan, et al. CocaCLIP: Exploring distillation of fully-connected knowledge interaction graph for lightweight text-image retrieval[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto: ACL, 2023: 71-80.
[21] Yang Chuanguang, An Zhulin, Huang Libo, et al. CLIP-KD: An empirical study of clip model distillation[C]// Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 15952-15962.
[22] Yang Kaicheng, Gu Tiancheng, An Xiang, et al. CLIP-CID: Efficient clip distillation via cluster-instance discrimination[C]//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia: AAAI, 2025: 21974-21982.
[23] 邵仁荣, 刘宇昂, 张伟, 等. 深度学习中知识蒸馏研究综述[J]. 计算机学报, 2022, 45(8): 1638-1673.
Shao Renrong, Liu Yuang, Zhang Wei, et al. A survey of knowledge distillation in deep learning[J]. Chinese Journal of Computers, 2022, 45(8): 1638-1673. (in Chinese)
[24] Chen Dapeng, Li Hongsheng, Liu Xihui, et al. Improving deep visual representation for person re-identification by global and local image-language association[C]// Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 56-73.
[25] Zhang Ying, Lu Huchuan. Deep cross-modal projection learning for image-text matching[C]//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 707-723.
[26] Howard A G, Zhu Menglong, Chen Bo, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17) [2025-10-11]. https://arxiv.org/abs/1704.04861.
[27] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5/6): 602-610.
[28] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc. , 2017: 6000-6010.
[29] Romero A, Ballas N, Kahou S E, et al. FitNets: Hints for thin deep nets[C]//Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015: 1-13.
[30] Yim J, Joo D, Bae J, et al. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 7130-7138.
[31] Zhang Ying, Xiang Tao, Hospedales T M, et al. Deep mutual learning[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4320-4328.
[32] Zhang Linfeng, Song Jiebo, Gao Anni, et al. Be your own teacher: Improve the performance of convolutional neural networks via self distillation[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 3712-3721.
[33] Yang Chenglin, Xie Lingxi, Su Chi, et al. Snapshot distillation: Teacher-student optimization in one generation [C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 2854-2863.
[34] van den Oord A, Li Yazhe, Vinyals O. Representation learning with contrastive predictive coding[EB/OL]. (2018-07-10) [2025-10-11]. https://arxiv.org/abs/1807.03748.
[35] Ding Zefeng, Ding Changxing, Shao Zhiyin, et al. Semantically self-aligned network for text-to-image part-aware person re-identification[EB/OL]. (2021-07-27) [2025-10-11]. https://arxiv.org/abs/2107.12666.
[36] Zhu Aichun, Wang Zijie, Li Yifeng, et al. DSSL: Deep surroundings-person separation learning for text-based person retrieval[C]//Proceedings of the 29th ACM International Conference on Multimedia. [S.l.]: ACM, 2021: 209-217.
[37] Kingma D P, Ba J. Adam: A method for stochastic optimization[C]//Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015.
[38] Wang Feng, Liu Huaping. Understanding the behaviour of contrastive loss[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 2495-2504.
[39] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 618-626.

责任编辑　刘飞阳