Computer Science and Application 计算机科学与应用, 2026, 16(2), 348-357
Published Online February 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.162064

Hans汉斯

基于身份与属性感知对齐的文本行人重识别

詹光辉

合肥工业大学计算机与信息学院，安徽 合肥

收稿日期：2026年1月5日；录用日期：2026年2月5日；发布日期：2026年2月13日

摘　要

文本行人重识别(Text-based Person Re-Identification, TextReID)旨在根据自然语言描述在大规模行人图像库中检索对应个体，在智能安防等场景中具有重要应用价值。近年来，大规模的图文预训练模型(如：Contrastive Language-Image Pre-Training, CLIP)被广泛迁移到该任务中，然而CLIP默认采用的实例级对比学习目标与行人重识别数据的真实分布存在结构性错配，同时文本侧对行人属性语义建模不足，导致检索性能不稳定、泛化能力受限。针对上述问题，本文提出一种结构感知CLIP适配框架，从目标空间与语义空间两个层面对CLIP进行结构性修正。具体地，提出身份感知对齐损失(Identity-Aware Alignment, IAA)，将实例级图文对齐提升为身份级分布对齐，避免同一身份样本被误判为负样本；同时提出属性感知掩码建模(Attribute-Aware Masked Modeling, AAMM)，在文本编码阶段重点建模与行人属性相关的词汇语义，增强判别性文本表征。在多个公开数据集上的实验结果表明，我们的方法在Rank-1、mAP等指标上均显著优于现有方法，验证了所提框架的有效性与泛化能力。

关键词

文本行人重识别，CLIP，结构感知对齐，身份感知损失，属性感知掩码建模

Text-Based Person Re-Identification via Identity and Attribute-Aware Alignment

Guanghui Zhan

School of Computer Science and Information Engineering, Hefei University of Technology (HFUT), Hefei Anhui

Received: January 5, 2026; accepted: February 5, 2026; published: February 13, 2026

Abstract

Text-based Person Re-Identification (TextReID) aims to retrieve corresponding pedestrian images

文章引用: 詹光辉. 基于身份与属性感知对齐的文本行人重识别[J]. 计算机科学与应用, 2026, 16(2): 348-357.
DOI: 10.12677/csa.2026.162064

from large-scale galleries based on natural language descriptions and plays an important role in intelligent surveillance. Recently, CLIP has been widely adopted for this task; however, its instance-level contrastive objective mismatches the true data distribution where multiple images and texts correspond to the same identity. Meanwhile, insufficient modeling of attribute-related semantics in textual descriptions further limits its discriminative capability. To address these issues, we propose a structure-aware CLIP adaptation framework that corrects the structural mismatches from both objective and semantic perspectives. Specifically, we introduce an Identity-Aware Alignment (IAA) loss to align identity-level distributions instead of individual instances, preventing same-identity samples from being mistakenly treated as negatives. Moreover, an Attribute-Aware Masked Modeling (AAMM) module is designed to emphasize attribute-related tokens during text encoding, thereby enhancing discriminative textual representations. Extensive experiments on multiple public benchmarks demonstrate that the proposed method significantly outperforms existing approaches in terms of Rank-1 and mAP, validating its effectiveness and generalization ability.

Keywords

Text-Based Person Re-Identification, CLIP, Structure-Aware Alignment, Identity-Aware Loss, Attribute-Aware Masked Modeling

1. 引言

文本检索行人(Text-based Person Re-Identification, TextReID)是多模态检索领域的一个重要研究分支，其核心目标是根据自然语言描述，从大规模非特定摄像头的图像数据库中识别并检索出目标行人[1]。相比于传统的基于图像的行人重识别，文本描述在实际场景中更具易得性，且能够涵盖图像中难以表达的抽象属性信息。随着深度学习和视觉–语言预训练(Vision-Language Pre-training, VLP)技术的飞速发展，以 CLIP [2]为代表的大规模跨模态对比学习模型被广泛应用于该任务中，极大地提升了图像与文本在统一嵌入空间中的对齐效果。然而，直接将通用领域的预训练模型迁移至细粒度的文本行人重识别场景时，依然面临着结构性错配导致的性能瓶颈。

现有的 TextReID 方法主要沿用 CLIP 的实例级图文对比学习目标，其基本假设是每一图像文本对在特征空间中具有唯一的对应关系。但在真实的重识别数据分布中，同一行人身份(Identity)往往对应多张视角各异的图像以及多段描述侧重不同的文本。这种实例级对齐目标忽略了行人标签的先验信息，强制将属于同一身份的不同样本对视为负样本进行推开。这种处理方式不仅破坏了特征空间的类内紧致性，导致身份表征在空间分布上产生“撕裂”，也限制了模型在处理跨摄像头视角变换时的泛化能力。此外，通用模型在语义建模阶段通常采用统一的掩码策略，未能充分考虑行人重识别任务中关键属性词汇(如衣着颜色、随身物品等)的判别权重，使得核心语义线索容易被背景噪声稀释。

针对上述挑战，本文提出一种基于身份与属性感知对齐的文本检索行人框架，旨在从目标空间与语义建模两个层面实现对预训练模型的结构性修正。在目标空间层面，本文提出身份感知对齐(Identity-Aware Alignment, IAA)模块，利用身份标签引导模型从实例级对齐提升至身份级分布对齐，有效地确保了同一身份样本在嵌入空间中的内聚性。在语义建模层面，本文设计了属性感知掩码建模(Attribute-Aware Masked Modeling, AAMM)模块，通过显著性引导机制强化模型对行人属性相关词汇的感知能力，增强了文本表征的判别力。本文的研究工作不仅为处理跨模态细粒度对齐提供了新的思路，也通过大量的实验

验证了所提方法在复杂检索场景下的有效性。

本文的主要贡献可以概括为以下三个方面：

1) 提出了身份感知对齐(IAA)模块。通过引入行人身份标签作为先验引导，将模型从传统的实例级图文对齐提升至身份级分布对齐，有效地确保了同一身份样本在联合嵌入空间中的紧致性，缓解了跨模态匹配中的冲突问题。

2) 设计了属性感知掩码建模(AAMM)模块。该模块通过显著性引导机制，在文本编码过程中重点挖掘并强化与行人关键属性相关的词汇语义，显著增强了文本特征在细粒度检索任务中的判别能力，使模型能够更精准地捕捉身份相关的关键线索。

3) 在多个公开的文本检索行人基准数据集上进行了详尽的实验分析与消融研究。实验结果表明，本文提出的方法在 Rank-1 与 mAP 等核心评价指标上均优于现有的先进算法。

2. 相关工作

2.1. 基于文本的行人检索

基于文本的行人检索(Text-based Person Re-Identification)最早由 Li 等人[1]提出，旨在通过自然语言描述在大规模候选库中检索特定行人目标。早期的研究工作[1] [3] [4]主要聚焦于跨模态特征的共用嵌入空间构建，利用 VGG [5]和 LSTM [6]分别提取图文特征，并配合交叉熵损失或三元组损失进行端到端对齐。为了挖掘行人表征中的细粒度线索，后续研究者引入了人体关键点检测、语义分割以及属性识别等外部辅助任务，通过显式地建立局部图像区域与文本实体之间的对应关系，显著提升了模型对复杂场景的适应能力。然而，此类局部对齐方法往往依赖于繁重的计算开销，且在处理低质量图像或描述模糊时表现出较差的鲁棒性。

随着大规模视觉–语言预训练模型的兴起，利用 CLIP [2]的强泛化能力来实现高效的跨模态对齐成为当前的主流研究方向。例如，Han 等人[7]利用动量对比学习框架迁移 CLIP 的先验知识，Yan 等人[8]则通过挖掘细粒度信息进一步增强了 CLIP 的表征能力。然而，现有工作大多忽略了 TextReID 数据的特殊分布属性。在通用图文检索任务中，通常假设每个图文对构成唯一的实例级映射，但在行人重识别场景下，同一行人身份(Identity)往往对应多视角图像与多段文本描述。直接应用实例级对比学习目标会忽略身份标签的关联性，导致模型难以学习到稳定的身份级特征表征。

2.2. 跨模态预训练模型

视觉–语言预训练模型旨在通过海量图文对数据学习具有强迁移能力的特征表示。根据交互机制的不同，现有的 VLP 模型主要分为单流架构与双流架构。单流模型通过跨模态 Transformer [9]实现深度语义交互，虽然在理解任务中表现卓越，但其高昂的在线计算成本限制了其在实时检索场景中的应用。以 CLIP [2]和 ALIGN [10]为代表的双流模型则通过独立的特征编码器实现了离线特征计算，凭借极高的检索效率成为大规模行人重识别任务的首选架构。

尽管双流 VLP 模型在宏观语义对齐方面表现出色，但其在 TextReID 这类对判别性要求极高的任务中仍表现出一定的局限性。一方面，通用的实例级对齐目标难以刻画行人身份的层次化特征；另一方面，通用的掩码语言建模[11]策略未能针对行人属性词汇进行偏置。在行人描述中，诸如服装颜色、配饰类型等判别性语义词汇在身份识别中起着决定性作用，而随机掩码机制会导致这些核心线索被背景语义所稀释。目前，虽然已有部分工作尝试对 CLIP 进行微调，但如何针对行人任务特有的“属性敏感性”进行结构化语义增强，仍是一个尚未被充分解决的开放性课题。

3. 方法

3.1. 框架概述

本文提出一种基于身份与属性感知对齐的文本检索行人方法，其总体框架如图 1 所示。该框架以 CLIP 双编码器为基础，分别对图像与文本进行特征编码，并在保持 CLIP 原有预训练能力的同时，引入两个结构修正模块：属性感知掩码建模模块(Attribute-Aware Masked Modeling, AAMM)与身份感知对齐模块(Identity-Aware Alignment, IAA)。AAMM 模块从语义空间角度强化文本属性语义建模，IAA 模块从目标空间角度修正实例级对比学习的结构性错配，两者协同提升跨模态身份级检索性能。

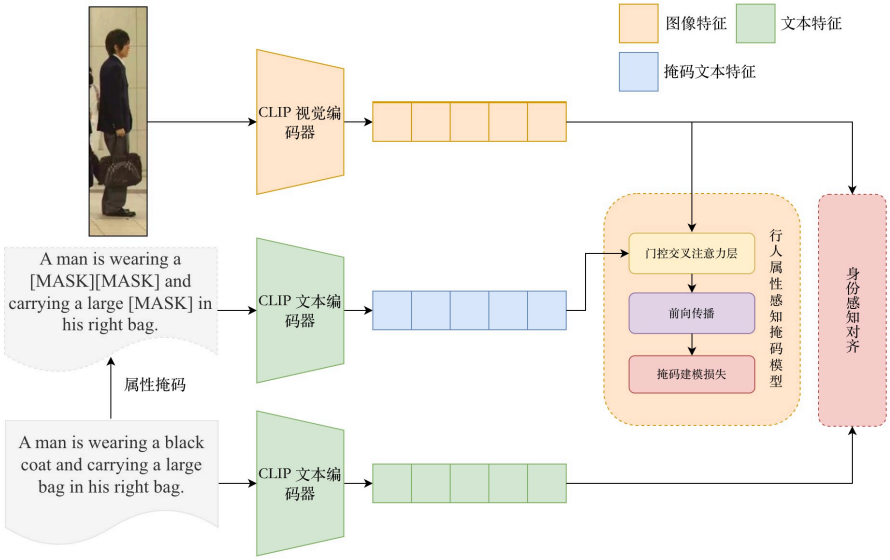

Figure 1. Overall of our proposed framework
图 1. 总体框架图

3.2. 属性感知掩码

3.2.1. 动机

现有掩码语言模型采用均匀随机掩码策略，忽略了行人属性词汇在身份判别中的重要性。为增强文本表征的判别能力，本文提出属性感知掩码建模模块，在掩码阶段对与行人属性相关的词汇赋予更高的掩码概率，从而引导模型重点学习关键属性语义。

3.2.2. 属性感知掩码策略

首先构建行人属性词汇集 A，包含服饰颜色、服装类型、携带物等关键属性词汇。在掩码过程中，对于输入文本中的 token w_i，其被掩码的概率定义为：

$$P_{mask}(w_i)=\begin{cases} p_a, & w_i \in A \\ p_o, & w_i \notin A \end{cases} \quad p_a > p_o \tag{1}$$

3.2.3. 门控交叉注意力

为在掩码属性词的语义恢复过程中引入视觉引导信息，同时避免视觉噪声对原始文本结构的过度干扰，本文引入了一种门控交叉注意力层，用于实现跨模态语义的自适应融合。模型结构如图 2 所示。设掩码文本特征为 $Q \in \mathbb{R}^{N\times L\times D}$，图像特征为 $K,V \in \mathbb{R}^{N\times L_v\times D}$。

(1) 跨模态注意力建模

首先采用标准多头交叉注意力建模文本 token 与视觉区域之间的对应关系：

$$X_{\text{attn}} = MHA\left(LN_t(Q), LN_i(K), LN_i(V)\right) \tag{2}$$

该操作为每个文本 token 生成视觉感知的语义增强表示。

(2) 门控残差融合机制

为在引入视觉语义的同时保持原始文本语义稳定性，本文引入门控融合算子：

$$G = \sigma\left(W_g\left[X_{\text{attn}}; Q\right]\right) \tag{3}$$

$$F = G \odot X_{\text{attn}} + (1-G) \odot Q \tag{4}$$

其中 $\sigma(\cdot)$ 为 Sigmoid 激活函数，$\odot$ 表示逐元素乘法。该门控机制使模型能够根据当前 token 的语义置信度，自适应调节视觉语义注入强度，从而避免无关视觉区域对属性语义恢复造成干扰。得到的融合特征 F 进一步输入跨模态编码器进行语义重整，随后通过 MLM 预测头恢复被掩码词汇，并采用交叉熵损失进行优化：

$$L_{AAMM} = -\sum \log P\left(w_i \mid F\right) \tag{5}$$

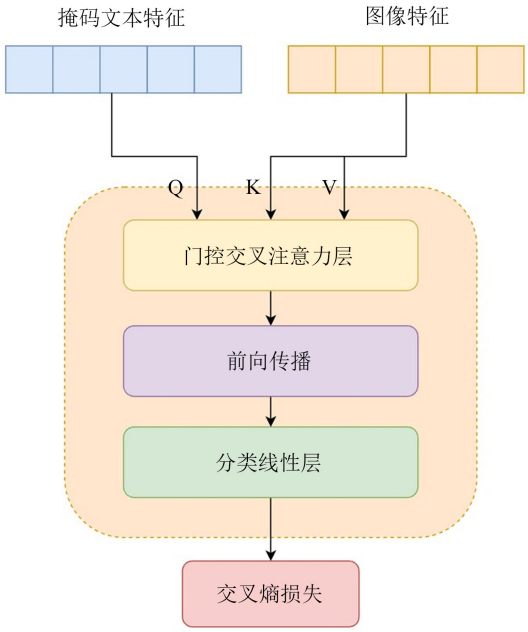

Figure 2. Architecture of the attribute-aware mask module
图 2. 属性感知掩码模块结构图

AAMM 模块仅在训练阶段引入，用于在视觉信息的引导下优化文本编码器的表示能力；在推理阶段，该模块被完全移除，仅保留文本编码器进行特征提取，因此不会引入任何额外的计算开销。尽管 AAMM 模块在推理阶段不参与计算，但其在训练阶段通过跨模态注意力机制直接作用于文本编码器的中间特征，使文本编码器在参数更新过程中显式地感知视觉属性信息。

3.3. 身份感知对齐

3.3.1. 动机

CLIP 默认采用实例级图文对比学习目标，将每一个图像文本对视为相互独立的类别。然而，在

TextReID 任务中，同一行人身份通常对应多张图像及多条文本描述，实例级对齐策略会将同一身份下的其他样本错误地视为负样本，从而导致类内特征分布被撕裂，降低跨模态检索的稳定性与鲁棒性。针对上述问题，本文提出身份感知对齐模块(Identity-Aware Alignment, IAA)，旨在将传统的实例级对齐提升为身份级分布对齐。与 Triplet Loss 和 Circle Loss 等主要在实例层面约束样本距离关系的损失函数不同，IAA 损失从身份分布的角度出发，显式建模同一身份下多文本–多图像之间的相似度分布一致性。此外，相比于基于硬间隔约束的损失形式，KL 散度能够提供更加平滑且稳定的优化目标，这一特性在微调大规模预训练模型(如 CLIP)时尤为重要，有助于在增强身份判别能力的同时保持原有语义空间结构的稳定性。因此，IAA 损失并非用于替代 CLIP 原有的实例级对比学习目标，而是作为一种身份感知的正则项，与实例级对齐目标形成互补，共同促进跨模态特征的有效对齐。

3.3.2. 身份感知目标分布

设一个 batch 中包含 B 个图像文本对，其全局特征分别为 $\{i_1,\cdots,i_B\}$ 与 $\{t_1,\cdots,t_B\}$，对应身份标签为 $\{pid_1,\cdots,pid_B\}$。

图文相似度矩阵定义为：

$$S_{ij}=\frac{i_i^{\top}t_j}{\tau} \tag{6}$$

对于第 i 个图像样本，构造其身份感知目标分布：

$$Q_{ij}=\begin{cases}1, & i=j\\ \alpha, & pid_i=pid_j, i\neq j\\ 0, & pid_i\neq pid_j\end{cases} \tag{7}$$

并进行归一化：

$$\tilde{Q}_i=\frac{Q_i}{\sum_j Q_{ij}} \tag{8}$$

3.3.3. 自适应权重 α

为缓解 batch 内身份样本数量不均衡对分布建模的影响，本文引入自适应权重：

$$\alpha=\frac{1}{\overline{k}} \tag{9}$$

其中 $\overline{k}$ 表示 batch 中每个身份的平均样本数。

3.3.4. 损失函数

预测分布定义为：

$$P_{ij}=\mathrm{Softmax}\left(S_{ij}\right) \tag{10}$$

采用双向 KL 散度作为优化目标：

$$L_{IAA}=\frac{1}{B}\sum_i KL\left(\tilde{Q}_i\,\|\,P_i\right)+KL\left(\tilde{Q}_i\,\|\,P_i^{\top}\right) \tag{11}$$

最终训练目标为：

$$L_{total}=L_{IAA}+\lambda L_{MLM} \tag{12}$$

其中 λ 为权重系数。

4. 实验

4.1. 数据集与实验设置

我们使用 CUHK-PEDES [1]和 ICFG-PEDES [17]数据集来评估模型。CUHK-PEDES 是首个专门用于行人文本检索的数据集，包含 13,003 个身份，共 40,206 张图像和 80,412 条文本描述。按照官方划分，训练集包含 11,003 个身份，对应 34,054 张图像和 68,108 条文本描述；验证集和测试集分别包含 3078 和 3074 张图像，以及 6158 和 6156 条文本描述，每个子集均包含 1000 个身份。ICFG-PEDES 包含 4102 个身份，共 54,522 张图像，每张图像对应一条文本描述。数据集按照官方划分，训练集包含 3102 个身份对应的 34,674 对图文，测试集包含剩余 1000 个身份对应的 19,848 对图文。为了评估模型性能，我们采用标准的 Recall-K (Rank-K)指标。Rank-1 表示每条查询文本在测试集中第 1 位返回正确匹配图像的比例。Rank-5 和 Rank-10 分别表示正确匹配图像出现在前 5 或前 10 位的比例。此外，我们还计算平均精度均值(mean Average Precision, mAP)，即所有查询的平均 AP，用于衡量整体检索效果。

4.2. 实验细节

模型训练使用 Adam 优化器，初始学习率设为 1e−5，batch size 为 16，训练 50 个 epoch，实验在单卡 NVIDIA RTX 3070 GPU 上完成。

4.3. 实验结果

表 1 展示和表 2 分别展示了我们的模型在 CUHK-PEDES 和 ICFG-PEDES 数据集上的实验结果。结果表明，我们所提出的属性感知跨模态特征增强框架取得了优异的结果，在数据集 CUHK-PEDES 上的 Rank-1 准确率达到了 70.97%的，比 CLIP 直接微调提升了 2.78 个百分点，在数据集 ICFG-PEDES 上的 Rank-1 准确率达到了 59.97%的，比 CLIP 微调的方法提升了 3.23 个百分点。这验证了我们提出的方法在捕捉行人细粒度属性和增强跨模态对齐能力上的有效性。

Table 1. Experimental results on the CUHK-PEDES dataset
表 1. 在 CUHK-PEDES 数据集上的实验结果

Method	Ref	Rank-1	Rank-5	Rank-10	mAP
CMPM/C [12]	ECCV18	49.37	-	79.27	-
TIMAM [13]	ICCV19	54.51	77.56	79.27	-
ViTAA [14]	ECCV20	54.92	75.18	82.90	51.60
NAFS [15]	arXiv21	59.36	79.13	86.00	54.07
DSSL [16]	MM21	59.98	80.41	87.56	-
SSAN [17]	arXiv21	61.37	80.15	86.73	-
ISANet [18]	arXiv22	63.92	82.15	87.69	-
LBUL [19]	LBUL	64.04	82.66	87.22	-
SAF [20]	ICASSP22	64.13	82.62	88.40	-
TIPCB [21]	Neuro22	64.26	83.19	89.10	-
IVT [22]	ECCVW22	65.59	83.11	89.21	-
CFine [8]	arXiv22	69.57	85.93	91.15	-
CSKT [23]	ICASSP24	69.70	86.92	91.80	62.74
PGA [24]	ICASSP25	69.44	87.03	**92.88**	62.83
Baseline (CLIP)	-	68.19	86.47	91.47	61.12
Ours	-	**70.97**	**88.17**	92.54	**63.10**

Table 2. Experimental results on the ICFG-PEDES dataset
表 2. 在 ICFG-PEDES 数据集上的实验结果

Method	Rank-1	Rank-5	Rank-10	mAP
CMPM/C [12]	43.51	65.44	74.26	-
ViTAA [14]	50.98	68.79	75.78	-
SSAN [17]	54.23	72.63	79.53	-
IVT [22]	56.04	73.60	80.22	-
ISANet [18]	57.73	75.42	81.72	-
CFine [8]	60.83	76.55	82.42	-
CSKT [23]	58.90	77.31	83.56	33.87
PGA [24]	58.10	76.95	84.06	32.58
Baseline (CLIP)	56.74	75.72	82.26	31.84
Ours	**59.97**	**78.06**	**84.33**	**34.54**

4.4. 消融实验

为了分析各模块对整体性能的贡献，我们在 CUHK-PEDES 数据集上进行了消融实验。实验结果见表 3，结果表明加入属性感知掩码建模模块后，Rank-1 从 68.17%提升到了 70.28%，说明属性感知掩码对文本编码器属性语义学习至关重要。加上身份感知对齐模块后，Rank-1 提升到了 68.81%，表明身份感知对齐能够显著提升图像–文本跨模态对齐性能。同时加入两者后，Rank-1 提升到了 70.97%，验证了两者组合的互补性和整体框架设计的有效性。

Table 3. Ablation experiment on CUHK-PEDES dataset
表 3. 在数据集 CUHK-PEDES 上的消融实验

Method	Rank-1	Rank-2	Rank-10	mAP
CLIP	68.17	86.71	91.58	61.52
+AAMM	70.28	87.93	92.52	62.76
+IAA	68.81	87.02	91.76	62.15
+IAA + AAMM	70.97	88.17	92.54	63.10

5. 总结

本文针对 CLIP 在文本行人重识别任务中存在的目标函数与语义建模双重结构性错配问题，提出一种基于身份与属性感知对齐的文本检索行人方法。该方法从目标空间与语义空间两个层面对 CLIP 进行结构性修正，设计了身份感知对齐损失与属性感知掩码建模模块，实现了身份级分布对齐与判别性属性语义建模的协同优化。在多个公开数据集上的实验结果表明，我们所提方法在 Rank-1 与 mAP 等评价指标上均优于现有主流方法，验证了该框架在跨摄像头检索稳定性与泛化能力方面的优势。该研究为跨模态预训练模型在细粒度检索任务中的适配提供了一种有效的结构性修正思路。

参考文献

[1] Li, S., Xiao, T., Li, H., Zhou, B., Yue, D. and Wang, X. (2017) Person Search with Natural Language Description. 2017

IEEE Conference on Computer Vision and Pattern Recognition (*CVPR*), Honolulu, 21-26 July 2017, 1970-1979. https://doi.org/10.1109/cvpr.2017.551

[2] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S. and Sutskever, I. (2021) Learning Transferable Visual Models from Natural Language Supervision. *International Conference on Machine Learning*, Vienna, 18-24 July 2021, 8748-8763.

[3] Li, S., Xiao, T., Li, H., Yang, W. and Wang, X. (2017) Identity-Aware Textual-Visual Matching with Latent Co-attention. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 1890-1899. https://doi.org/10.1109/iccv.2017.209

[4] Chen, T., Xu, C. and Luo, J. (2018) Improving Text-Based Person Search by Spatial Matching and Adaptive Threshold. 2018 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Lake Tahoe, 12-15 March 2018, 1879-1887. https://doi.org/10.1109/wacv.2018.00208

[5] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. https://arxiv.org/pdf/1409.1556

[6] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[7] Han, X., He, S., Zhang, L. and Xiang, T. (2021) Text-Based Person Search with Limited Data. https://doi.org/10.5244/c.35.10

[8] Yan, S., Dong, N., Zhang, L. and Tang, J. (2023) Clip-Driven Fine-Grained Text-Image Person Re-Identification. *IEEE Transactions on Image Processing*, **32**, 6032-6046. https://doi.org/10.1109/tip.2023.3327924

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., *et al*. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, **30**, 5998-6008.

[10] Jia, C., Yang, Y., Xia, Y., Chen, Y.T., *et al*. (2021) Scaling up Visual and Vision-Language Representation Learning with Noisy Text Supervision. *International Conference on Machine Learning*, Vienna, 18-24 July 2021, 4904-4916.

[11] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, **1**, 4171-4186.

[12] Zhang, Y. and Lu, H. (2018) Deep Cross-Modal Projection Learning for Image-Text Matching. In: *Lecture Notes in Computer Science*, Springer, 707-723. https://doi.org/10.1007/978-3-030-01246-5_42

[13] Sarafianos, N., Xu, X. and Kakadiaris, I. (2019) Adversarial Representation Learning for Text-to-Image Matching. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 5814-5824. https://doi.org/10.1109/iccv.2019.00591

[14] Wang, Z., Fang, Z., Wang, J. and Yang, Y. (2020) Vitaa: Visual-Textual Attributes Alignment in Person Search by Natural Language. In: *Lecture Notes in Computer Science*, Springer, 402-420. https://doi.org/10.1007/978-3-030-58610-2_24

[15] Gao, C., Cai, G., Jiang, X., Zheng, F., *et al*. (2021) Contextual Non-Local Alignment over Full-Scale Representation for Text-Based Person Search. https://arxiv.org/pdf/2101.03036

[16] Zhu, A., Wang, Z., Li, Y., Wan, X., Jin, J., Wang, T., *et al*. (2021) DSSL: Deep Surroundings-Person Separation Learning for Text-Based Person Retrieval. *Proceedings of the* 29*th ACM International Conference on Multimedia*, Chengdu, 20-24 October 2021, 209-217. https://doi.org/10.1145/3474085.3475369

[17] Ding, Z., Ding, C., Shao, Z. and Tao, D. (2021) Semantically Self-Aligned Network for Text-to-Image Part-Aware Person Re-Identification. https://arxiv.org/pdf/2107.12666

[18] Yan, S., Tang, H., Zhang, L. and Tang, J. (2024) Image-Specific Information Suppression and Implicit Local Alignment for Text-Based Person Search. *IEEE Transactions on Neural Networks and Learning Systems*, **35**, 17973-17986. https://doi.org/10.1109/tnnls.2023.3310118

[19] Wang, Z., Zhu, A., Xue, J., Wan, X., Liu, C., Wang, T., *et al*. (2022) Look before You Leap: Improving Text-Based Person Retrieval by Learning a Consistent Cross-Modal Common Manifold. *Proceedings of the* 30*th ACM International Conference on Multimedia*, Lisboa, 10-14 October 2022, 1984-1992. https://doi.org/10.1145/3503161.3548166

[20] Li, S., Cao, M. and Zhang, M. (2022) Learning Semantic-Aligned Feature Representation for Text-Based Person Search. 2022 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Singapore, 23-27 May 2022, 2724-2728. https://doi.org/10.1109/icassp43922.2022.9746846

[21] Chen, Y., Zhang, G., Lu, Y., Wang, Z. and Zheng, Y. (2022) TIPCB: A Simple but Effective Part-Based Convolutional Baseline for Text-Based Person Search. *Neurocomputing*, **494**, 171-181. https://doi.org/10.1016/j.neucom.2022.04.081

[22] Shu, X., Wen, W., Wu, H., Chen, K., Song, Y., Qiao, R., *et al*. (2022) See Finer, See More: Implicit Modality Alignment for Text-Based Person Retrieval. In: *Lecture Notes in Computer Science*, Springer, 624-641.

https://doi.org/10.1007/978-3-031-25072-9_42

[23] Liu, Y., Li, Y., Liu, Z., Yang, W., Wang, Y. and Liao, Q. (2024) Clip-Based Synergistic Knowledge Transfer for Text-Based Person Retrieval. 2024 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Seoul, 14-19 April 2024, 7935-7939. https://doi.org/10.1109/icassp48485.2024.10445963

[24] Huang, Y., Zhang, C., Li, Z., Wang, Z. and Wei, C. (2025) Prototypical Graph Alignment for Text-Based Person Search. 2025 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Hyderabad, 6-11 April 2025, 1-5. https://doi.org/10.1109/icassp49660.2025.10890808