Geometry-Aware Noisy Correspondence Mitigation for Cross-Modal Text-Based Person Retrieval

Xinpan Yuan[1]*, Shaomin Xie[1]*, Liujie Hua[2]†, Chengyuan Zhang[3]‡, Guihu Zhao[4], Lin Yuanbo Wu[5]§

[1]School of Computer Science and Artificial Intelligence, Hunan University of Technology
[2]Department of Criminal Science and Technology, Hunan Police Academy
[3]College of Computer Science and Electronic Engineering, Hunan University
[4]National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University
[5]School of Engineering, University of Warwick

Abstract

Text-Based Person Retrieval (TBPR) aims to accurately retrieve target individuals from large-scale image databases using only textual descriptions. Existing methods typically assume a ground-truth correspondence between text and images (i.e., strongly correlated). However, in real-world scenarios, this assumption may not be able to hold for the cross-modal matching due to weak or even corrupted correlations between textual descriptions and visual content, referred to as noisy correspondence (NC). Such NC largely disrupts the correspondence learning between visual and semantic modalities. Though prior works have improved single-modal robustness against noisy labels, systematic modeling of both cross-modal and intra-modal geometric structures in TBPR remains limited attention. In this paper, we propose Geometric Structure Consistency Alignment (GSCA) to TBPR, which leverages cross-modal cosine similarity and intra-modal nearest-neighbor affinity to learn visual-semantic consistency under noisy correspondence. To mitigate the structural corruption caused by noisy pairs, we introduce the Structure Refinement and Mining (SRAM) module. By partitioning training data into clean, ambiguous, and noisy subsets, SRAM enables the model to strategically refine the cross-modal correspondence by mining reliable pairs, thus enhancing the reliability of positive or negative samples discrimination and preserving structural consistency across modalities. Extensive experiments demonstrate that our method achieves state-of-the-art performance across three public datasets. On CUHK-PEDES, it boosts Rank-1 by 1.42% in noise-free conditions, sustaining a robust 74.25% Rank-1 under a 50% noise ratio.

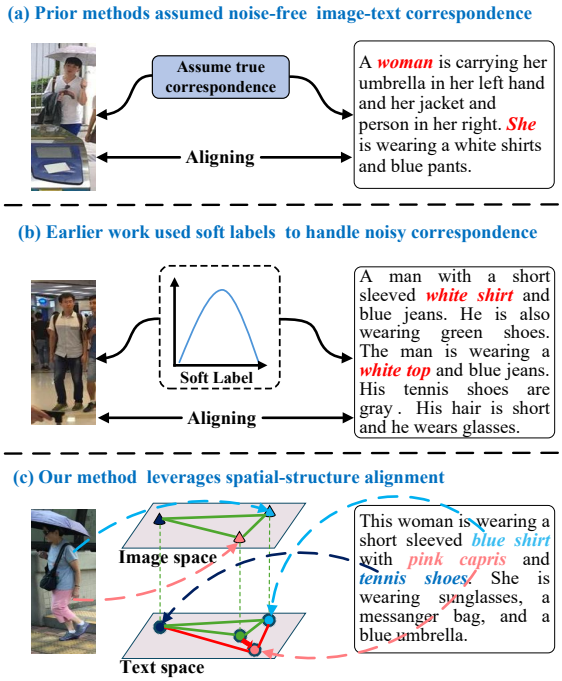


Figure 1: (a) Existing methods assume clean text–image pairs. (b) Noisy-correspondence methods use soft labels for mismatches. (c) Our GSCA exploits cross- and intra-modal geometry to separate clean and noisy samples.

Introduction

Text-Based Person Retrieval (TBPR) (Liu et al. 2025b; Jiang and Ye 2023) aims to retrieve the target individual from a large-scale image gallery based on a given natural language description. It has garnered increasing attention from both academia and industry due to its practical value in surveillance scenarios such as suspect tracking and missing person retrieval (Sun et al. 2025; Liu et al. 2025b; Yu, Wen, and Zheng 2025; Park et al. 2024; Liu et al. 2024). However, TBPR inherently faces the modality gap and significant appearance variations between visual and textual modalities.

Early attempts can be broadly categorized into global and local matching strategies for learning cross-modal similarity. Global methods (Qin, Pu, and Wu 2023; Shu et al. 2022;

*These authors contributed equally.
†The First Corresponding author: liujiehuahnp@163.com
‡The Second Corresponding author: cyzhangcse@hnu.edu.cn
§The Third Corresponding author: yuanbo.lin@warwick.ac.uk

Wu et al. 2021; Zhang et al. 2025) use modality-specific encoders with contrastive learning to enforce holistic visual–semantic alignment, while local approaches (Shao et al. 2022) explicitly align body parts with textual entities to enhance fine-grained discrimination. Recent works (Han et al. 2021; Jiang and Ye 2023; Yan et al. 2023b) further introduce pretrained models such as BERT (Devlin et al. 2019), ViT (Dosovitskiy et al. 2020), and CLIP (Radford et al. 2021) to exploit richer visual and semantic knowledge for either global alignment or fine-grained correspondence discovery. However, as illustrated in Figure 1(a), these methods typically and implicitly assume that all training image–text pairs are strictly and perfectly aligned, and thus inevitably overlook noisy correspondence in real-world applications.

Learning with noisy correspondence in cross-modal retrieval offers a promising way to estimate reliable soft correspondence labels for image-text pairs. As a result, the model is enhanced to judge true matches. For example, as shown in Figure 1(b), NCR (Huang et al. 2021) initiates this line of work by introducing a co-teaching mechanism that filters out high-loss pairs as noisy samples. Subsequent methods (Wang et al. 2025; Zha et al. 2025; Yan et al. 2024; Yang et al. 2023) further improve robustness by leveraging meta-learning and clean subset optimization. However, these approaches still essentially follow the paradigm of uni-modal sample selection (Li, Socher, and Hoi 2020), which struggles to capture deep and complex dependencies in multi-modal structures and often tends to overfit noisy pairs.

Unimodal label noise is usually restricted to a single-modality feature space under an assumed homogeneous distribution (Li, Socher, and Hoi 2020), and can often be alleviated by label smoothing (Lukasik et al. 2020) or noise-robust losses (Yan et al. 2023a). In contrast, multimodal noise arises when textual descriptions are partially relevant or entirely irrelevant to the image, breaking the cross-modal geometric structure (Liu et al. 2025a). As shown in Figure 1(c), noisy correspondences (red edges) incorrectly shrink distances between mismatched pairs, distorting the similarity graph in the shared space and potentially collapsing intra-modal geometry when asymmetrically distributed samples are forced to align. To tackle this, we propose Geometric Structure Consistency Alignment (GSCA) for robust correspondence alignment. At the cross-modal level, GSCA flags low-similarity image–text pairs as potential noisy samples; at the intra-modal level, it models geometry via similarity distributions between a query and other samples in the same modality. The resulting structure-aware soft labels effectively guide alignment, forming a geometry-consistent positive feedback loop that substantially improves the model's overall robustness to noisy correspondence.

Furthermore, to address the sample pairs structure collapse caused by noisy correspondence, we introduce the Structure Refinement and Mining (SRAM) module, which jointly models the spatial structures among different sample pairs. Specifically, SRAM first estimates the matching confidence of each image-text pair by leveraging the memory effect of deep neural networks (Jiang et al. 2023), then categorizes them into clean, ambiguous, and noisy subsets. Based on this partitioning, SRAM refines the structural alignment of positive (clean) pairs and mines potential consistency from negative (noisy) ones. In the reconstructed structure space, SRAM introduces a structure-aware contrastive loss that unifies the structural relationships of clean, ambiguous, and noisy sample pairs within a joint optimization framework. The main contributions of this work are as follows:

- We propose Geometric Structure Consistency Alignment (GSCA), which leverages structural differences between and within modalities to effectively differentiate clean and noisy samples, thereby rectifying these structural disparities to enhance alignment robustness.
- We introduce the Structure Refinement and Mining (SRAM) module, which refines the consistency among positive sample pairs and leverages the latent consistency of negative samples to adjust the spatial structures between sample pairs, thereby mitigating the structural collapse induced by noisy correspondence.
- We conducted experiments on the CUHK-PEDES, ICFG-PEDES ,and RSTPReid datasets, achieving R@1 improvements of 1.42%, 0.58% and 1.22%. Even with 50% noisy correspondence, our method maintains robust R@1 of 74.25%, 66.47% and 64.28%.

Related Work

Text-Based Person Retrieval

Text-Based Person Retrieval (TBPR) (Li et al. 2017) aims to retrieve person images with natural language queries. Existing methods (Chen, Xu, and Luo 2018; Li et al. 2017; Ji et al. 2023; Zhong et al. 2024) broadly fall into global and local matching: global approaches learn a shared cross-modal embedding space but often miss fine-grained interactions, while local approaches explicitly align body regions with textual entities (Wu et al. 2018a,b) for precise correspondence at higher computational cost. Recent work, inspired by vision–language pretraining, leverages large pretrained models to capture alignment knowledge. However, these methods typically assume perfectly aligned training pairs, an assumption that fails under ubiquitous noisy correspondence. This paper addresses this inevitable and challenging noisy correspondence problem in TBPR.

Learning with Noisy Correspondence

Noisy Correspondence (NC) in multi-modal or multi-view learning has recently gained increasing attention. Unlike general noisy-label learning (Hu et al. 2021; Li, Socher, and Hoi 2020; Liu et al. 2022; Huang et al. 2023; Yao et al. 2023; Huang et al. 2024; Dang et al. 2025), NC focuses on negative pairs that are incorrectly annotated as positives. Most NC methods address this at the loss level by redesigning contrastive objectives or re-weighting schemes, such as adapting DivideMix to separate clean and noisy pairs (Han et al. 2023), using Beta-mixture models for soft correspondence labels (Yang et al. 2023), or employing meta-learning–based similarity correction networks and robust contrastive losses (Chuang et al. 2022; Hong et al. 2024) to improve robustness under different noise ratios.

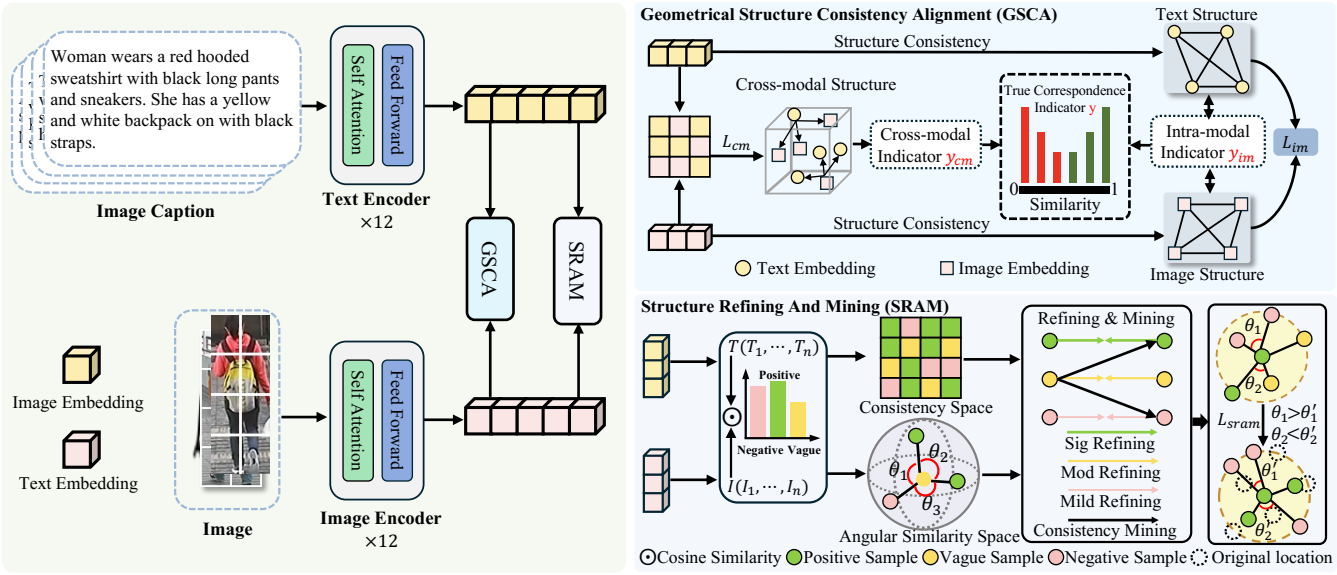


Figure 2: Overview of the proposed GSCA and SRAM. GSCA optimizes cross-modal and intra-modal objectives to maintain geometric structure consistency, distinguishing noisy samples and estimating true correspondence indicators y for robust learning. SRAM partitions data into clean, ambiguous, and noisy subsets, refining positive pair consistency to mitigate NC and mining latent negative pair consistency to preserve structural integrity.

Proposed Method

In this section, we present two plug-and-play modules, GSCA and SRAM, whose overall architecture is shown Figure 2, and detail their designs in the following subsections.

Geometric Structure Consistency Alignment

Geometric Structure Consistency Alignment aims to maintain structural consistency both across and within modalities, while leveraging the perceived structural differences between and within modalities to identify noisy samples. Previous research (Li et al. 2024) has shown that clean and noisy samples exhibit distinct structural characteristics: clean pairs generally achieve higher cross-modal similarity scores, while noisy pairs tend to distort intra-modal structures, leading to lower similarity. These insights support the effectiveness of structure-aware modeling in distinguishing noisy pairs and enhancing robust alignment.

Cross-Modal Geometric Structure. From the cross-modal perspective, the geometric structure is defined by the similarity between representations from different modalities. For a given query image I_i, its cross-modal geometric structure can be formulated as $\mathcal{Q}_i^{cm} = \{\langle I_i, T_j\rangle\}_{j=1}^{N}$, where T_j denotes the j-th text description within the same batch. To preserve such structure, GSCA minimizes the expected risk of the cross-modal objective, defined as:

$$\mathcal{H}_{\mathcal{L}_{cm}}(f,g) = \min_{(I,T,y)\sim\mathcal{B}} \left[\mathcal{L}_{cm}\left(\langle I, T\rangle, y\right)\right], \qquad (1)$$

where $\mathcal{L}_{cm}$ is the cross-modal loss function, which can be instantiated as a contrastive loss or triplet loss used in traditional retrieval models. The goal is to align cross-modal representations based on the correspondence label y, thereby maximizing the similarity of $\langle I_i, T_i\rangle$.

Intra-Modal Geometric Structure. Matched text-image pairs should exhibit consistent intra-modal structures, whereas unmatched pairs display disordered structures. The intra-modal structure of the i-th sample is given by $\mathcal{Q}_{im}^{i} = \{\langle I_i, I_j\rangle, \langle T_i, T_j\rangle\}_{j=1}^{N}$, where $\langle I_i, I_j\rangle$ and $\langle T_i, T_j\rangle$ represent image-image and text-text similarities, respectively. To maintain such structural consistency, GSCA introduces the intra-modal objective as:

$$\mathcal{H}_{\mathcal{L}_{im}}(f,g) = \min_{(I,T,y)\sim\mathcal{B}} \left[\mathcal{L}_{im}\left(\langle I, I\rangle, \langle T, T\rangle, y\right)\right], \qquad (2)$$

where $\mathcal{L}_{im}$ represents the intra-modal loss function, designed to ensure that matched samples maintain consistent structure within the same modality. Neglecting structural consistency during optimization leads to a decline in the model's text-to-image retrieval performance (Goel et al. 2022; Jiang et al. 2023), highlighting the dependence of cross-modal alignment on the stability of intra-modal structure. Consequently, incorporating intra-modal structural consistency can further significantly enhance the effectiveness of cross-modal alignment.

Noise Discrimination and Purification. In the early training phase, deep neural networks (DNNs) preferentially learn from clean samples before adapting to noisy ones. Leveraging this behavior, GSCA constructs well-established structural representations to effectively distinguish clean samples from noisy ones.

Cross-Modal Discrimination. Based on the well-formed cross-modal structures learned in early training, the representations of clean data pairs are expected to be more tightly aligned than those of noisy pairs. GSCA exploits this structural discrepancy by introducing a function to estimate the

bidirectional cross-modal correspondence indicator:

$$y_i^{cm} = \frac{1}{2}\frac{\exp(\langle I_j, T_j\rangle/\mathcal{T}_1)}{\sum_{j=1}^{N}\exp(\langle I_i, T_j\rangle/\mathcal{T}_1)} + \frac{1}{2}\frac{\exp(\langle I_i, T_i\rangle/\mathcal{T}_1)}{\sum_{i=1}^{N}\exp(\langle I_j, T_i\rangle/\mathcal{T}_1)}, \tag{3}$$

where $\mathcal{T}_1$ denotes the temperature coefficient, which is set to 0.03 in all experiments. Taking the first term as an example, it evaluates the ratio of the similarity between the current pair (I_i, T_i) to the total similarity between I_i and all textual samples. For clean pairs, the similarity of (I_i, T_i) is expected to dominate, resulting in an indicator value close to 1. In contrast, noisy pairs will exhibit lower similarity, and thus the indicator value approaches 0.

Intra-Modal Discrimination. For matched image-text pairs, the intra-modal structures are expected to be similar and mutually aligned, whereas unmatched pairs exhibit distinct patterns that reflect positional divergence. Specifically, we compute the cosine similarity between intra-modal structures to obtain a consistency score, denoted as: $S_i^{im} = cos\left(\langle I_i, I_j\rangle, \langle T_i, T_j\rangle\right)_{j=1}^{N}$. Previous studies (Li et al. 2024) suggested that clean samples consistently achieve higher cosine similarity scores than noisy ones, resulting in a bimodal score distribution across the entire dataset. This distribution can be modeled by a two-component Gaussian Mixture Model (GMM) (Li, Socher, and Hoi 2020), described as:

$$p(S^{im}) = \sum_{k=1}^{K}\alpha_k\phi(S^{im} \mid k), y_i^{im} = \frac{\alpha_{k_l}\phi(S_i^{im} \mid k_l)}{\sum_{k=1}^{N}\alpha_k\phi(S_i^{im} \mid k)}, \tag{4}$$

where α_k denotes the mixture coefficient for the k-th component, and $\phi(S^{im} \mid k)$ is the probability density function of that component. The second equation estimates the intra-modal consistency indicator y_i^{im}, which reflects the probability that the observed sample belongs to the cleaner component ($k = k_l$). This value approaches 1 for clean samples and 0 for noisy samples, effectively distinguishing samples with noisy correspondences.

We have estimated the cross-modal and intra-modal correspondence indicators y^{cm} and y^{im} by exploiting structural differences. To effectively leverage their complementary reliability and accurately identify noisy samples, we define the final label as the minimum:

$$y_i = min(y_i^{cm}, y_i^{im}). \tag{5}$$

Noise Purification. To address the noisy correspondence problem, we optimize both cross-modal and intra-modal objectives. Since the estimated correspondence labels are soft values within the range $[0, 1]$, reflecting the degree of ground-truth alignment, they can be seamlessly incorporated into the loss function for each sample.

We use a contrastive loss for the cross-modal objective:

$$\mathcal{L}_{cm} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\frac{\exp\left(\langle I_i, T_i\rangle/\mathcal{T}_1\right)}{\sum_{j=1}^{N}\exp\left(\langle I_i, T_j\rangle/\mathcal{T}_1\right)}, \tag{6}$$

where y_i is applied directly to each sample before computing the loss. For the intra-modal objective, in addition to applying sample-level purification to the loss itself, it is crucial to avoid distortion in the geometry computation caused by noisy correspondences. Specifically, distances between the query and noisy samples should be excluded. The purified intra-modal loss is defined as:

$$\mathcal{L}_{im} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\sum_{z=1}^{N} y_z\langle I_i, I_z\rangle y_z\langle T_i, T_z\rangle/\mathcal{T}_1}{\sum_{j=1}^{N}\exp\left(\sum_{z=1}^{N} y_z\langle I_i, I_z\rangle y_z\langle T_j, T_z\rangle/\mathcal{T}_1\right)}, \tag{7}$$

where the y_z terms filter out noisy samples from similarity aggregation. Similarly, when calculating intra-modal cosine similarity scores for discrimination, noisy correspondences can interfere with the detection.

Overall Objective. The total loss for the GSCA module is formulated as a weighted sum of the two purified losses:

$$\mathcal{L}_{gsca} = \mathcal{L}_{cm} + \alpha \cdot \mathcal{L}_{im}, \tag{8}$$

where α is a hyperparameter to balance the optimization of both objectives. We set α to 0.4 in experiments.

Structure Refining and Mining

Given N image-text pairs $\{(I_i, T_i, y_i)\}_{i=1}^{N}$, where $y_i \in \{0, 1\}$ indicates whether the image and text refer to the same identity, we consider a challenging scenario where some negative pairs ($y_i = 0$) are incorrectly labeled as positive ($y_i = 1$), resulting in noisy supervision. To mitigate this issue, SRAM first identifies potential noisy pairs by measuring cosine similarity $s(I_i, T_i)$ between features extracted from image and text encoders. To enhance representation discriminability, we adopt the InfoNCE loss:

$$\mathcal{L}_{info} = -\log\frac{\exp(s(I_i, T_j)/\mathcal{T}_2)}{\sum_{j=1}^{N}\exp(s(I_i, T_j)/\mathcal{T}_2)} - \log\frac{\exp(s(I_j, T_i)/\mathcal{T}_2)}{\sum_{j=1}^{N}\exp(s(I_j, T_i)/\mathcal{T}_2)}, \tag{9}$$

where $\mathcal{T}_2 = 0.07$ is a fixed temperature parameter in all experiments. We then design a cooperative data partition strategy that assigns each sample pair to one of three subsets: clean ($\mathcal{D}_p$), noisy ($\mathcal{D}_n$), and ambiguous ($\mathcal{D}_v$), based on the bidirectional similarities p_i^{i2t} and p_i^{t2i}:

$$(I_i, T_i) \in \begin{cases} \mathcal{D}_p & if\ p_i^{i2t} > \gamma\ and\ p_i^{t2i} > \gamma, \\ \mathcal{D}_n & if\ p_i^{i2t} < \gamma\ and\ p_i^{t2i} < \gamma, \\ \mathcal{D}_v & otherwise, \end{cases} \tag{10}$$

where $p_i^{i2t} = s(I_i, T_i)$ and $p_i^{t2i} = s(T_i, I_i)$. The confidence threshold is set to $\gamma = 0.5$ in our experiments.

Consistency Refining. For clean samples $\mathcal{D}_p$, the original labels are likely reliable. We slightly adjust their correspondence strength using p_i^{i2t} and p_i^{t2i}:

$$y_{p_i} = [p_i^{t2i} y_i + (1 - y_i)p_i^{i2t}]\exp(-\theta^2/2\sigma^2), \tag{11}$$

where $\exp\left(-\theta^2/2\sigma^2\right)$ serves as an angle correction factor, σ is a hyperparameter controlling the sensitivity of correction, set to 0.1 in this study. Detailed experimental results are provided in the appendix. The angle is computed as

Method	Ref	Image Enc.	Text Enc.	R@1	R@5	R@10	mAP	mINP
ViTAA (Wang et al. 2020)	ECCV20	RN50	LSTM	54.92	75.18	82.90	51.60	-
DSSL (Zhu et al. 2021)	MM21	RN50	BERT	59.98	80.41	87.56	-	-
LapsCore (Wu et al. 2021)	ICCV21	RN50	BERT	63.40	-	87.80	-	-
LBUL (Wang et al. 2022b)	MM22	RN50	BERT	64.04	82.66	87.22	-	-
Han et al. (Han et al. 2021)	BMVC21	CLIP-RN101	CLIP-Xformer	64.08	81.73	88.19	60.08	-
SAF (Li, Cao, and Zhang 2022)	ICASSP22	ViT-Base	BERT	64.13	82.62	88.40	-	-
TIPCB (Chen et al. 2022)	Neuro22	RN50	BERT	64.26	83.19	89.10	-	-
CAIBC (Wang et al. 2022a)	MM22	RN50	BERT	64.43	82.87	88.37	-	-
AXM-Net (Farooq et al. 2022)	MM22	RN50	BERT	64.44	80.52	86.77	58.73	-
LGUR (Shao et al. 2022)	MM22	DeiT-Small	BERT	65.25	83.12	89.00	-	-
IVT (Shu et al. 2022)	ECCV22	ViT-Base	BERT	65.59	83.11	89.21	-	-
CFine (Yan et al. 2023b)	TIP22	CLIP-ViT	BERT	69.57	85.93	91.15	-	-
IRRA (Jiang and Ye 2023)	CVPR23	CLIP-ViT	CLIP-Xformer	73.38	89.93	93.71	66.13	50.24
LAIP (Wu et al. 2024)	ICME24	CLIP-ViT	CLIP-Xformer	<u>76.72</u>	90.42	93.60	66.05	-
PLOT (Park et al. 2024)	ECCV24	CLIP-ViT	CLIP-Xformer	75.28	90.42	94.12	-	-
RDE (Qin et al. 2024)	CVPR24	CLIP-ViT	CLIP-Xformer	75.94	90.14	94.12	67.56	<u>51.44</u>
IRLT (Liu et al. 2024)	AAAI24	CLIP-ViT	CLIP-Xformer	74.46	90.19	94.01	-	-
WoRA (Sun et al. 2025)	WWW25	CLIP-ViT	CLIP-Xformer	76.38	89.72	93.49	67.22	-
DM-Adapter (Liu et al. 2025b)	AAAI25	CLIP-ViT	CLIP-Transformer	72.17	88.74	92.85	64.33	-
VFE-TPS (Shen et al. 2025)	KBS25	CLIP-ViT	CLIP-Transformer	72.47	88.24	93.24	64.26	-
IRRA+GSCA+SRAM(Ours)	-	CLIP-ViT	CLIP-Xformer	75.67	<u>91.03</u>	<u>94.28</u>	<u>67.63</u>	51.35
RDE+GSCA+SRAM(Ours)	-	CLIP-ViT	CLIP-Xformer	**77.36**	**91.62**	**94.47**	**68.37**	**51.69**

Table 1: Comparison of different methods on CUHK-PEDES. Bold and underline denote the best and second-best results.

$\theta = arccos(s(I_i, T_i))$. The objective of angle optimization is to significantly reduce the angles of positive pairs, enlarge those of negative pairs, and maintain moderate angles for ambiguous pairs, thereby enhancing the model's discriminative ability in image-text matching tasks. For ambiguous pairs $\mathcal{D}_v$, we average the confidences to soften the label:

$$y_{v_i} = \frac{y_i(p_i^{t2i} + p_i^{i2t})\exp(-\theta^2/2\sigma^2)}{2 \cdot (1 + |p_i^{i2t} - p_i^{t2i}|)}. \tag{12}$$

For noisy pairs $\mathcal{D}_n$, we adopt the average of the predictions:

$$y_{n_i} = \frac{(p_i^{i2t} + p_i^{t2i})\exp(-\theta^2/2\sigma^2)}{2} y_i. \tag{13}$$

Consistency Mining. We further mine latent consistency from noisy samples. For a given pair (I_i, T_j), we compute directional matching strengths using corrected labels:

$$\omega_{i,j}^{i2t} = \frac{(1 - y_{n_i})s(I_i, T_j)}{\sum_{k \neq i}^{N} s(I_i, T_k)}. \tag{14}$$

Similarly, the computation of $\omega_{i,j}^{t2i}$ aligns with $\omega_{i,j}^{i2t}$. Instead of using all noisy pairs, we filter training pairs based on a learned, adaptive, data-driven threshold:

$$\beta = \frac{1}{N}\left(y_{p_i} N_c + y_{v_i} N_v + y_{n_i} N_n\right). \tag{15}$$

where N_c, N_v and N_n denote the number of pairs in the clean, vague and noisy split. Pairs with $\omega_{i,j}^{i2t} < \beta$ are discarded, and similarly for $\hat{\omega}_{i,j}^{t2i}$:

$$\hat{\omega}_{i,j}^{i2t} = \begin{cases} 0 & if\ \omega_{i,j}^{i2t} < \beta, \\ \omega_{i,j}^{i2t} & otherwise. \end{cases} \tag{16}$$

Noise	Method	R@1	R@5	R@10	mAP	mINP
0%	IRRA	73.38	89.93	93.71	66.13	50.24
	IRRA+GSCA+SRAM	75.67	<u>91.03</u>	<u>94.28</u>	<u>67.63</u>	51.35
	RDE	<u>75.94</u>	90.14	94.12	67.56	<u>51.44</u>
	RDE+GSC+SRAM	**77.36**	**91.62**	**94.47**	**68.37**	**51.69**
20%	IRRA	69.44	87.09	92.20	62.16	45.70
	IRRA+GSCA+SRAM	72.56	88.92	92.97	64.80	48.61
	RDE	<u>74.53</u>	<u>89.23</u>	<u>93.55</u>	<u>66.13</u>	<u>49.63</u>
	RDE+GSC+SRAM	**76.37**	**90.72**	**93.88**	**68.03**	**50.40**
50%	IRRA	62.41	82.23	88.40	55.52	38.48
	IRRA+GSCA+SRAM	70.46	<u>88.21</u>	<u>91.83</u>	63.51	<u>48.10</u>
	RDE	<u>71.25</u>	87.39	91.76	<u>63.59</u>	47.50
	RDE+GSCA+SRAM	**74.25**	**89.07**	**93.36**	**67.57**	**49.36**

Table 2: Performance comparison of various methods under different noise ratios on the CUHK-PEDES dataset, where noise ratio indicates the proportion of noisy samples.

Noise	Method	R@1	R@5	R@10	mAP	mINP
0%	IRRA	63.46	80.25	85.82	38.06	7.93
	IRRA+GSCA+SRAM	65.52	81.43	85.12	39.30	<u>8.06</u>
	RDE	<u>67.68</u>	<u>82.47</u>	<u>87.36</u>	<u>40.06</u>	7.87
	RDE+GSCA+SRAM	**68.26**	**82.76**	**88.12**	**40.67**	**8.30**
20%	IRRA	60.58	78.14	84.20	35.92	6.91
	IRRA+GSCA+SRAM	64.20	80.18	84.43	37.58	7.50
	RDE	<u>66.51</u>	<u>81.70</u>	<u>86.71</u>	<u>39.09</u>	<u>7.56</u>
	RDE+GSCA+SRAM	**66.92**	**81.90**	**87.30**	**39.40**	**7.74**
50%	IRRA	52.53	71.99	79.41	29.05	4.43
	IRRA+GSCA+SRAM	62.80	74.76	79.13	25.47	3.47
	RDE	<u>63.76</u>	<u>79.53</u>	<u>84.91</u>	<u>37.38</u>	<u>6.80</u>
	RDE+GSCA+SRAM	**66.47**	**81.29**	**86.30**	**39.20**	**7.31**

Table 3: Performance comparison under different noise ratios on the ICFG-PEDES dataset.

Noise	Method	R@1	R@5	R@10	mAP	mINP
0%	IRRA	60.20	81.30	88.20	47.17	25.28
	IRRA+GSCA+SRAM	62.74	83.50	90.33	48.52	25.43
	RDE	65.35	83.95	89.90	50.88	28.08
	RDE+GSCA+SRAM	**66.57**	**86.20**	**90.82**	**51.69**	**28.47**
20%	IRRA	54.00	77.15	85.55	43.20	22.53
	IRRA+GSCA+SRAM	61.35	83.18	89.23	48.10	24.75
	RDE	63.85	83.85	89.45	50.27	27.75
	RDE+GSCA+SRAM	**64.68**	**84.25**	**90.30**	**50.29**	**27.88**
50%	IRRA	56.65	78.40	86.55	42.41	21.05
	IRRA+GSCA+SRAM	59.72	81.20	87.80	47.00	24.37
	RDE	62.85	83.20	89.15	47.67	23.97
	RDE+GSCA+SRAM	**64.28**	**83.26**	**89.64**	**48.68**	**24.40**

Table 4: Performance comparison under different noise ratios on the RSTPReid dataset.

Noise	Method	R@1	R@5	R@10	mAP	mINP
0%	Baseline(RDE)	75.94	90.14	94.12	67.56	51.44
	GSCA	76.63	90.37	94.10	67.71	51.43
	SRAM	76.47	90.25	94.33	67.50	51.17
	GSCA+SRAM	**77.36**	**91.62**	**94.47**	**68.37**	**51.69**
20%	Baseline(RDE)	74.53	89.23	93.55	66.13	49.63
	GSCA	75.60	90.14	93.85	67.47	50.10
	SRAM	76.10	90.20	93.50	67.50	50.24
	GSCA+SRAM	**76.37**	**90.72**	**93.88**	**68.03**	**50.40**
50%	Baseline(RDE)	71.25	87.39	91.76	63.59	47.50
	GSCA	73.46	88.54	92.70	66.52	49.03
	SRAM	73.62	88.65	92.96	66.69	49.19
	GSCA+SRAM	**74.25**	**89.07**	**93.36**	**67.57**	**49.36**

Table 5: Ablation study of GSCA and SRAM on the RDE baseline under different noise levels.

At low noise levels, most sample pairs are clean, and a higher rigid filtering threshold inhibits consistency learning from negative pairs; conversely, a lower threshold, indicating fewer clean samples, requires mining potential consistency from a large set of negative pairs.

Refining and Mining Loss. Finally, we define the joint objective combining refined clean sample pairs and mined negative sample pairs:

$$\mathcal{L}_{sram} = \frac{1}{N}\sum_{i=1}^{N} y_i \mathcal{L}_{info} + \frac{1}{2}\left(\mathcal{L}_n^{i2t} + \mathcal{L}_n^{t2i}\right), \tag{17}$$

where

$$\mathcal{L}_n^{i2t} = \sum_{j\neq i} \hat{\omega}_{i,j}^{t2i} \log \frac{\exp(s(I_i, T_k)/\mathcal{T}_2)}{\sum_{k=1}^{N}\exp(s(I_i, T_k)/\mathcal{T}_2)}. \tag{18}$$

$\mathcal{L}_n^{t2i}$ is defined similarly for the reverse direction. Together, they ensure robust training against noisy correspondences.

Experiment

Evaluation Protocol

We use Rank-k (k = 1, 5, 10) as our primary evaluation metric, which measures the probability that at least one correct person image appears in the top-k results for a text query. For a more comprehensive evaluation, we also report mean Average Precision (mAP) and mean Inverse Negative Penalty (mINP), where higher values of Rank-k, mAP, and mINP indicate better retrieval performance.

Implementation Details

We employ the pretrained CLIP (Radford et al. 2021) as our modality-specific encoder. During training, we apply extensive data augmentation to enhance sample diversity. Specifically, for images we perform random horizontal flipping, random cropping with padding, and random erasing; for textual descriptions we adopt random token masking, substitution, and deletion. The input images are resized to 384×128 pixels, and the maximum token length of input captions is set to 77. The model is optimized with Adam for 60 epochs using a cosine learning-rate scheduler, where the initial learning rate for the original CLIP parameters is 1×10^{-5}. We meticulously conduct all our experiments on a single RTX 3090 24GB GPU.

Comparison with State-of-the-Art Methods

To validate GSCA and SRAM, we conducted experiments on the CUHK-PEDES, ICFG-PEDES, and RSTPReid datasets, using IRRA and RDE as baseline methods.

Results on CUHK-PEDES. As shown in Table 1, the proposed GSCA and SRAM modules significantly enhance model performance under clean settings. Specifically, on the CUHK-PEDES dataset, our modules enhance the R@1 score by 2.29% for the IRRA baseline and by 1.42% for the RDE baseline, demonstrating the effectiveness of the proposed modules in optimizing cross-modal retrieval performance. Furthermore, Table 2 highlights the robustness of our method under noisy conditions. In particular, under a 50% synthetic noise scenario, GSCA and SRAM boost R@1 by 8.05% for IRRA and by 3% for RDE, illustrating their ability to effectively distinguish clean from noisy samples.

Results on ICFG-PEDES. The results in Table 3 demonstrate that both GSCA and SRAM modules consistently enhance cross-modal retrieval performance on the ICFG-PEDES dataset. Under clean conditions, the R@1 score rises steadily by 2.06%, effectively confirming our modules' generalizability. Under noisy conditions with 50% synthetic noise, the proposed modules yield a notable 10.27% R@1 increase for IRRA and 2.71% for RDE, and significantly enhance retrieval robustness under noisy conditions.

Results on RSTPReid. As shown in Table 4, our modules deliver significant performance gains under clean conditions on the RSTPReid dataset. Specifically, GSCA and SRAM enhance the R@1 score by 4.22% for IRRA and 1.87% for RDE. Under the 50% noise condition, our method boosts R@1 by 5.42% for IRRA and 2.28% for RDE, demonstrating robust performance across varying noise conditions.

Ablation Study

To validate the effectiveness and synergy of GSCA and SRAM, we performed ablation studies on the CUHK-PEDES dataset using RDE as the baseline, testing noise ratios of 0%, 20%, and 50%. Results from Table 5 show that each module can independently enhance performance, and their combination yields the best outcomes. At 0% noise,

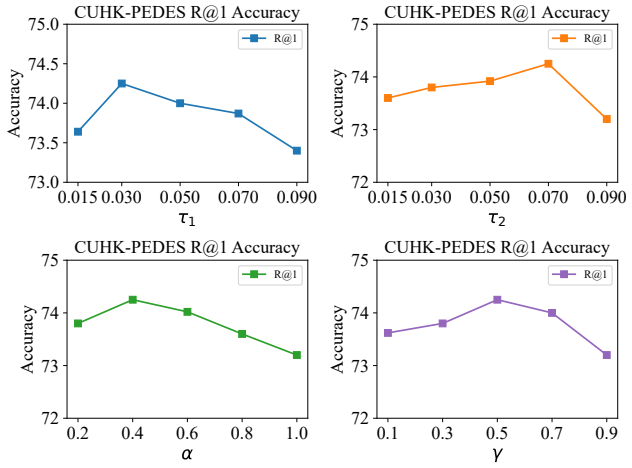


Figure 3: Evaluation of GSCA ($\mathcal{T}_1$, α) and SRAM ($\mathcal{T}_2$, γ) hyper-parameters on retrieval performance under 50% noisy correspondence on CUHK-PEDES.

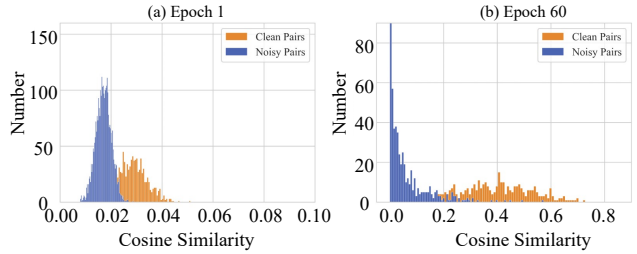
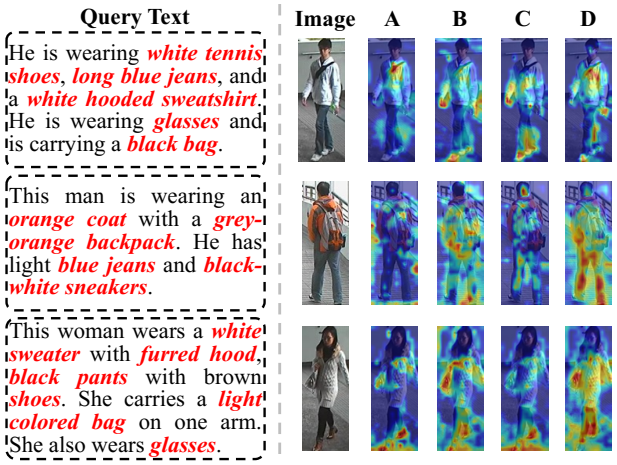


Figure 4: Evidence distribution of clean and noisy sample pairs at the start and end of training under 20% synthetic noise on the CUHK-PEDES dataset.

GSCA achieves 76.63% R@1, SRAM 76.47%, and together they reach 77.36%. Under 50% noise, GSCA scores 73.46%, SRAM 73.62%, and their synergy reaches 74.25%. GSCA models geometric structures cross-modal and intra-modal to accurately distinguish clean from noisy samples, effectively resolving noisy correspondence issues. SRAM refines consistency in positive pairs and extracts latent consistency from negative pairs, significantly alleviating structural collapse. Demonstrated that our proposed method effectively enhances model robustness across various noise ratios.

Parametric Analysis and Visualization

To evaluate key hyper-parameters, we conducted an ablation study on CUHK-PEDES for GSCA cross-modal temperature τ_1, intra-modal weight α, SRAM temperature τ_2 and margin γ. As shown in Figure 3, the optimal settings are $\tau_1 = 0.03$, $\alpha = 0.4$ for GSCA, and $\tau_2 = 0.07$, $\gamma = 0.5$ for SRAM. A moderate $\tau_1 = 0.03$ enhances cross-modal discrimination without feature space collapse, and $\alpha = 0.4$ balances intra-modal and cross-modal objectives, avoiding retrieval degradation. For SRAM, $\tau_2 = 0.07$ ensures robustness to noisy samples, and $\gamma = 0.5$ optimizes samples pair classification boundaries. As shown in Figure 4,



A:IRRA **B**:IRRA+GSCA+SRAM **C**:RDE **D**:RDE+GSCA+SRAM

Figure 5: Grad-CAM visualization of IRRA and RDE baseline models before and after incorporating our method.

our proposed method effectively distinguishes clean samples from noisy ones under a 20% noise condition. Grad-CAM effectively visualizes word-level localization in image-text pairs. Figure 5 clearly and convincingly demonstrates that our method outperforms IRRA and RDE, achieving notably more accurate entity localization.

Conclusion

To address the challenge of noisy correspondence in text-based person Retrieval (TBPR), this paper proposes to integrate geometric structure consistency with structure refinement and mining. By introducing two plug-and-play modules GSCA and SRAM, we effectively mitigate both cross-modal and intra-modal structure inconsistency caused by noisy correspondence in TBPR. GSCA leverages structural differences across and within modalities to accurately distinguish clean and noisy samples, and SRAM refines the structural consistency of positive pairs and mines potential consistency among negative pairs to alleviate representation collapse. Experimental results on three widely used benchmarks demonstrate the superior performance of our method, which maintains strong robustness under noisy conditions.

Acknowledgments

This research was funded by the National Natural Science Foundation of Hunan Province (Grant no. 2025JJ70028, 2025JJ81178, 2024JJ9550) and Scientific Research Project of Education Department of Hunan Province (Grant no. 24A0401). Health Research Project of Hunan Provincial Health Commission (grant no.20254688) National Natural Science Foundation of China 62472161; Hunan Provincial Natural Science Foundation of China 2023JJ30169. Postgraduate Scientific Research Innovation Project of Hunan Province, China[LXBZZ2024316]

References

Chen, T.; Xu, C.; and Luo, J. 2018. Improving text-based person search by spatial matching and adaptive threshold. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1879–1887. IEEE.

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.

Chuang, C.-Y.; Hjelm, R. D.; Wang, X.; Vineet, V.; Joshi, N.; Torralba, A.; Jegelka, S.; and Song, Y. 2022. Robust contrastive learning against noisy views. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16670–16681.

Dang, Z.; Luo, M.; Wang, J.; Jia, C.; Han, H.; Wan, H.; Dai, G.; Chang, X.; and Wang, J. 2025. Disentangled noisy correspondence learning. *IEEE Transactions on Image Processing*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

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

Farooq, A.; Awais, M.; Kittler, J.; and Khalid, S. S. 2022. Axm-net: Implicit cross-modal feature alignment for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 4477–4485.

Goel, S.; Bansal, H.; Bhatia, S.; Rossi, R.; Vinay, V.; and Grover, A. 2022. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35: 6704–6719.

Han, H.; Miao, K.; Zheng, Q.; and Luo, M. 2023. Noisy correspondence learning with meta similarity correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7517–7526.

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

Hong, F.; Yao, J.; Lyu, Y.; Zhou, Z.; Tsang, I.; Zhang, Y.; and Wang, Y. 2024. On harmonizing implicit subpopulations. In *The Twelfth International Conference on Learning Representations*, volume 1.

Hu, P.; Peng, X.; Zhu, H.; Zhen, L.; and Lin, J. 2021. Learning cross-modal retrieval with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5403–5413.

Huang, X.; Liu, Z.; Liu, S.-Y.; and Cheng, K.-T. 2023. Efficient and robust quantization-aware training via adaptive coreset selection. *arXiv preprint arXiv:2306.07215*.

Huang, Z.; Hu, P.; Niu, G.; Xiao, X.; Lv, J.; and Peng, X. 2024. Learning with noisy correspondence. *International Journal of Computer Vision*, 132(9): 3656–3677.

Huang, Z.; Niu, G.; Liu, X.; Ding, W.; Xiao, X.; Wu, H.; and Peng, X. 2021. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34: 29406–29419.

Ji, Z.; Hu, J.; Liu, D.; Wu, L. Y.; and Zhao, Y. 2023. Asymmetric Cross-Scale Alignment for Text-Based Person Search. *IEEE Trans on Multimedia*, 25: 7699 – 7709.

Jiang, D.; and Ye, M. 2023. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2787–2797.

Jiang, Q.; Chen, C.; Zhao, H.; Chen, L.; Ping, Q.; Tran, S. D.; Xu, Y.; Zeng, B.; and Chilimbi, T. 2023. Understanding and constructing latent modality structures in multi-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7661–7671.

Li, H.; Gu, J.; Song, J.; Zhang, A.; and Gao, L. 2024. One-step Noisy Label Mitigation. *arXiv preprint arXiv:2410.01944*.

Li, J.; Socher, R.; and Hoi, S. C. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*.

Li, S.; Cao, M.; and Zhang, M. 2022. Learning semantic-aligned feature representation for text-based person search. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2724–2728. IEEE.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1970–1979.

Liu, D.; Wu, L. Y.; Li, B.; Zhao, Y.; Ge, Z.; and Zhang, J. 2025a. T-Person-GAN: Text-to-Person Image Generation with Identity-Consistency and Manifold Mix-Up. *Expert Systems with Applications*.

Liu, Y.; Liu, Z.; Lan, X.; Yang, W.; Li, Y.; and Liao, Q. 2025b. Dm-adapter: Domain-aware mixture-of-adapters for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5703–5711.

Liu, Y.; Qin, G.; Chen, H.; Cheng, Z.; and Yang, X. 2024. Causality-inspired invariant representation learning for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14052–14060.

Liu, Y.; Wu, J.; Qu, L.; Gan, T.; Yin, J.; and Nie, L. 2022. Self-supervised correlation learning for cross-modal retrieval. *IEEE Transactions on Multimedia*, 25: 2851–2863.

Lukasik, M.; Bhojanapalli, S.; Menon, A.; and Kumar, S. 2020. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, 6448–6458. PMLR.

Park, J.; Kim, D.; Jeong, B.; and Kwak, S. 2024. Plot: Text-based person search with part slot attention for corresponding part discovery. In *European Conference on Computer Vision*, 474–490. Springer.

Qin, Y.; Chen, Y.; Peng, D.; Peng, X.; Zhou, J. T.; and Hu, P. 2024. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27197–27206.

Qin, Y.; Pu, N.; and Wu, H. 2023. EDMC: efficient multi-view clustering via cluster and instance space learning. *IEEE Transactions on Multimedia*, 26: 5273–5283.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Shao, Z.; Zhang, X.; Fang, M.; Lin, Z.; Wang, J.; and Ding, C. 2022. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th acm international conference on multimedia*, 5566–5574.

Shen, W.; Fang, M.; Wang, Y.; Xiao, J.; Li, D.; Chen, H.; Xu, L.; and Zhang, W. 2025. Enhancing visual representation for text-based person searching. *Knowledge-Based Systems*, 309: 112893.

Shu, X.; Wen, W.; Wu, H.; Chen, K.; Song, Y.; Qiao, R.; Ren, B.; and Wang, X. 2022. See finer, see more: Implicit modality alignment for text-based person retrieval. In *European Conference on Computer Vision*, 624–641. Springer.

Sun, J.; Fei, H.; Ding, G.; and Zheng, Z. 2025. From data deluge to data curation: A filtering-wora paradigm for efficient text-based person search. In *Proceedings of the ACM on Web Conference 2025*, 2341–2351.

Wang, Y.; Wu, Y.; Dai, Z.; Tian, C.; Long, J.; and Chen, J. 2025. Noisy Correspondence Rectification via Asymmetric Similarity Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 21384–21392.

Wang, Z.; Fang, Z.; Wang, J.; and Yang, Y. 2020. Vitaa: Visual-textual attributes alignment in person search by natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 402–420. Springer.

Wang, Z.; Zhu, A.; Xue, J.; Wan, X.; Liu, C.; Wang, T.; and Li, Y. 2022a. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, 5314–5322.

Wang, Z.; Zhu, A.; Xue, J.; Wan, X.; Liu, C.; Wang, T.; and Li, Y. 2022b. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In *Proceedings of the 30th ACM international conference on multimedia*, 1984–1992.

Wu, L.; Wang, Y.; Gao, J.; and Li, X. 2018a. Deep Adaptive Feature Embedding with Local Sample Distributions for Person Re-identification. *Pattern Recognition*, 73: 275–288.

Wu, L.; Wang, Y.; Li, X.; and Gao, J. 2018b. What-and-Where to Match: Deep Spatially Multiplicative Integration Networks for Person Re-identification. *Pattern Recognition*, 76: 727–738.

Wu, Y.; Wang, H.; Wu, M.; Cao, M.; and Zhang, M. 2024. LAIP: learning local alignment from image-phrase modeling for text-based person search. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–10. IEEE.

Wu, Y.; Yan, Z.; Han, X.; Li, G.; Zou, C.; and Cui, S. 2021. Lapscore: language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1624–1633.

Yan, J.; Luo, L.; Deng, C.; and Huang, H. 2023a. Adaptive hierarchical similarity metric learning with noisy labels. *IEEE Transactions on Image Processing*, 32: 1245–1256.

Yan, M.; Huang, H.; Liu, Y.; Zhao, J.; Gao, X.; Xu, C.; Guan, Z.; and Zhao, W. 2024. Truthsr: trustworthy sequential recommender systems via user-generated multimodal content. In *International Conference on Database Systems for Advanced Applications*, 180–195. Springer.

Yan, S.; Dong, N.; Zhang, L.; and Tang, J. 2023b. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 32: 6032–6046.

Yang, S.; Xu, Z.; Wang, K.; You, Y.; Yao, H.; Liu, T.; and Xu, M. 2023. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19883–19892.

Yao, J.; Han, B.; Zhou, Z.; Zhang, Y.; and Tsang, I. W. 2023. Latent class-conditional noise model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9964–9980.

Yu, H.; Wen, J.; and Zheng, Z. 2025. CAMeL: Cross-modality Adaptive Meta-Learning for Text-based Person Retrieval. *IEEE Transactions on Information Forensics and Security*.

Zha, Q.; Liu, X.; Peng, S.-J.; Cheung, Y.-m.; Xu, X.; and Wang, N. 2025. ReCon: Enhancing True Correspondence Discrimination through Relation Consistency for Robust Noisy Correspondence Learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29680–29689.

Zhang, X.; Liu, K.; Wang, X.; Zhou, Z.; and Chen, H. 2025. RMGNet: The Progressive Relationship-Mining Graph Neural Network for Text-to-image Person Re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhong, Y.; Hu, J.; Huang, Y.; Zhang, Y.; and Ji, R. 2024. ERQ: Error Reduction for Post-Training Quantization of Vision Transformers. In *International Conference on Machine Learning (ICML)*.

Zhu, A.; Wang, Z.; Li, Y.; Wan, X.; Jin, J.; Wang, T.; Hu, F.; and Hua, G. 2021. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM international conference on multimedia*, 209–217.