Dual-Granularity Cross-Modal Identity Association for Weakly-Supervised Text-to-Person Image Matching

Yafei Zhang
Kunming University of Science and Technology
Kunming, China
zyfeimail@163.com

Yongle Shang
Kunming University of Science and Technology
Kunming, China
sylmail99@163.com

Huafeng Li*
hfchina99@163.com
Kunming University of Science and Technology
Kunming, China

Abstract

Weakly supervised text-to-person image matching, as a crucial approach to reducing models' reliance on large-scale manually labeled samples, holds significant research value. However, existing methods struggle to predict complex one-to-many identity relationships, severely limiting performance improvements. To address this challenge, we propose a local-and-global dual-granularity identity association mechanism. Specifically, at the local level, we explicitly establish cross-modal identity relationships within a batch, reinforcing identity constraints across different modalities and enabling the model to better capture subtle differences and correlations. At the global level, we construct a dynamic cross-modal identity association network with the visual modality as the anchor and introduce a confidence-based dynamic adjustment mechanism, effectively enhancing the model's ability to identify weakly associated samples while improving overall sensitivity. Additionally, we propose an information-asymmetric sample pair construction method combined with consistency learning to tackle hard sample mining and enhance model robustness. Experimental results demonstrate that the proposed method substantially boosts cross-modal matching accuracy, providing an efficient and practical solution for text-to-person image matching.

CCS Concepts

• **Computing methodologies** → **Computer vision problems**.

Keywords

Text-to-person image matching, Dual-granularity identity matching, Weakly supervised learning, Cross-modal identity association

ACM Reference Format:
Yafei Zhang, Yongle Shang, and Huafeng Li. 2025. Dual-Granularity Cross-Modal Identity Association for Weakly-Supervised Text-to-Person Image Matching. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

*Corresponding author.

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/XXXXXX.XXXXXX

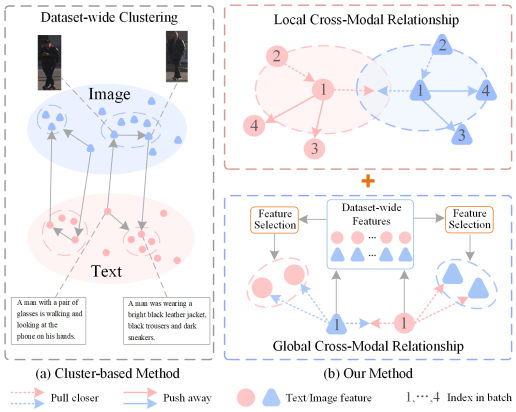

Figure 1: Comparison between the proposed method and existing approaches. While existing methods primarily rely on dataset-wide clustering to establish cross-modal identity associations, our method introduces a dual-granularity mechanism that integrates local and global collaboration, leading to more accurate and robust associations.

1 Introduction

Text-to-person image matching [18] (also known as Text-based person search or Text-to-image person re-identification) is emerging as a key technology in applications such as intelligent security and smart cities. It aims to perform cross-modal matching between natural language descriptions of pedestrians and large-scale image databases, enabling accurate retrieval of target individuals without the need for visual queries. This technology holds strong potential in public security scenarios, including criminal investigations, missing person searches, and intelligent video surveillance. Mainstream supervised methods for text-to-person image matching rely on large-scale, high-quality labeled image-text datasets [6, 18, 25, 28, 30, 42, 49] and employ complex deep neural networks for representation alignment [11, 16, 23, 29, 36, 38, 40]. However, the high cost of annotation presents a major challenge, significantly limiting their scalability in real-world applications.

To address this issue, GTR [2] and GAAP [19] leverage large image-text models to automatically generate textual descriptions for person images, thereby avoiding manual labeling of image-text pairs. However, the generated descriptions often contain noise, which weakens the model's generalization performance and hinders further improvements in cross-modal matching. Weakly supervised text-to-person image matching assumes that each person image is paired with a textual description, while identity relationships between different images remain unknown. This setting not only eliminates the need for manual labeling of identity relationships but

also reduces the noise introduced by unsupervised text generation. Nevertheless, the lack of complete cross-modal identity supervision significantly limits the model's performance. To this end, methods such as CMMT [45] and CPCL [47] apply cross-modal clustering [8] (as shown in Figure 1(a)) to generate pseudo associations and train models in a supervised manner. Unfortunately, these methods struggle with hard samples, as clustering fails to effectively separate visually or semantically similar instances.

To address the aforementioned issues, this paper proposes a dual-granularity identity relationship modeling mechanism that operates at both local and global levels, as illustrated in Figure 1(b). This mechanism enhances the model's ability to distinguish hard samples by emphasizing those with uncertain identity relationships during training, thereby improving overall robustness. Specifically, the model first establishes identity correspondences among intra-modality samples and integrates them with cross-modal associations within each training batch. This facilitates the explicit construction of bidirectional cross-modal relationships— from text to image and image to text. Such modeling reinforces identity-level constraints across modalities, enhances the representational capacity of the base model, and provides a strong foundation for subsequent global relationship reasoning.

At the global relationship level, this paper addresses the heterogeneity of cross-modal data by using the visual modality as a relational anchor. Cross-modal identity associations are constructed by dynamically linking image samples with the entire training set. Leveraging the visual modality's strength in fine-grained representation, image samples are treated as the "feature centers" of cross-modal relationships, helping to mitigate the ambiguity introduced by the semantic abstractness of textual descriptions. This enables the model to establish more reliable global identity associations. To further improve the model's sensitivity to weak associations, a dynamic confidence adjustment mechanism is introduced, which nonlinearly adjusts confidence scores based on real-time feedback from inter-modal similarity responses. As a result, samples with weaker associations receive increased attention during feature extraction, encouraging the model to refine decision boundaries in ambiguous regions. In addition, to compensate for the absence of labeled hard samples, we propose an information-asymmetric sample pair construction and consistency learning strategy. An information perturbation mechanism is applied to generate cross-modal sample pairs with semantic discrepancies, and a cross-modal semantic consistency regularization is employed to guide the model in learning robust representations under asymmetric conditions. This strategy not only addresses the challenge of mining hard cross-modal samples but also significantly enhances the model's robustness in complex matching scenarios. Overall, the contributions of this paper are threefold:

- We integrate identity relationships at both local and global levels by constructing local relationships within a single batch and establishing global cross-modal identity associations using the visual modality as an anchor. This mechanism significantly enhances the model's ability to distinguish hard samples and improves its overall robustness.
- We design a dynamic mechanism that nonlinearly adjusts relationship confidence based on real-time feedback from inter-modality similarity responses. This approach increases model's sensitivity to weakly associated samples and enables it to make finer distinctions near ambiguous decision boundaries.
- We propose an innovative strategy that addresses the absence of clearly labeled hard samples by generating cross-modal sample pairs with semantic discrepancies through an information perturbation strategy. Combined with cross-modality semantic consistency regularization, this method enables the model to learn robust feature representations under information-asymmetric conditions, significantly improving its robustness in complex matching scenarios.

2 Related Work

Supervised Methods. In supervised text-to-person image matching, the identity correspondence between texts and person images is known. How to leverage these annotations to train models has been a research hotspot. In early studies, CNNs, such as VGG [27] and ResNet [9], were commonly used to extract global representations from images, while LSTM [10] or BERT [5] was employed to extract global text representations, as seen in methods like MCCL [34], DCPL [44], and DCIE [48]. The key to these methods lies in designing effective loss functions to enhance global cross-modal alignment. However, relying solely on global information makes it difficult to capture subtle visual cues of individuals, thereby limiting retrieval accuracy.

To address the challenges, explicit local alignment strategies have been proposed based on patch-to-word matching concepts [3]. Methods like MIA [21] and TIPCB [4] adopt multi-stage or hierarchical frameworks to align horizontally segmented image regions with noun phrases, while PMA [12] and ViTAA [35] incorporate prior knowledge such as pose estimation and attribute segmentation to guide finer-grained partitioning. However, granularity mismatch between detailed visual inputs and coarse textual descriptions remains an issue, motivating strategies such as granularity unification and redundancy suppression [15, 26, 32, 41, 49]. In contrast, implicit alignment methods like IRRA [11] and SSAN [6] utilize attention mechanisms or cross-modal interaction to learn local correspondences without explicit region division, offering greater flexibility. Recently, vision-language pre-trained models such as CLIP [24] and ALBEF [17] have gained significant attention, inspiring methods like LAIP [36], Rasa [1], WoRA [29], Cfine [39], PLOT [23], and LSPM [16], which leverage large-scale image-text pretraining to improve cross-modal representation and alignment via fine-tuning or adaptation. Despite their success, these models rely heavily on large labeled datasets, making it essential to develop methods that reduce labeling costs without significantly compromising performance.

Unsupervised and Weakly Supervised Methods. To reduce reliance on large-scale labeled samples, researchers have explored novel methods for unsupervised pedestrian text-image matching. Instead of assuming the availability of pre-existing textual descriptions of pedestrian appearance, these methods generate text descriptions directly from pedestrian images. The generated text-image pairs are then used for supervised training. Since no labeled training set is required, these approaches are classified as unsupervised text-to-person image matching methods. Among them, the quality of the generated text plays a crucial role in model performance. To

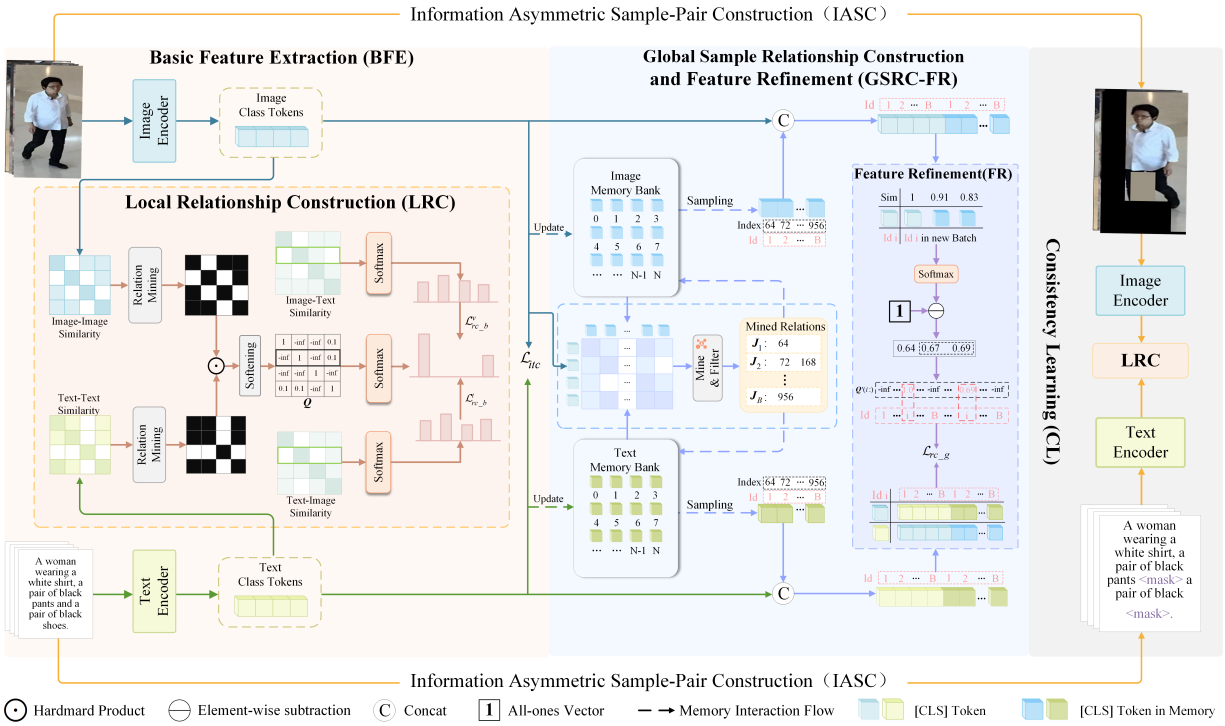

Figure 2: Overall architecture of the proposed method. The input image-text pairs are first processed by the BFE module to extract basic representations and construct local cross-modal identity associations within each batch by integrating information from both modalities. The model then uses the visual modality as an anchor to explore broader cross-modal identity relationships from a global perspective. These global associations are leveraged to further optimize the model, improving its ability to distinguish between different identities. Additionally, information-asymmetric sample pairs are generated through specific transformations on the original inputs, enabling the model to better handle hard samples.

address this issue, GAAP [19] employs Attribute-guided Pseudo Caption Generation and incorporates an Attribute-guided Multi-strategy Cross-modal Alignment module to enhance cross-modal alignment. GTR [2] generates fine-grained pseudo descriptions using a Vision-Language Model with instructional prompts and adopts a confidence-based training strategy during retrieval to mitigate the impact of noise in the generated text. Although these methods are effective, the noise in generated texts still limits the model's generalization ability on real datasets.

In weakly supervised setting, it is typically assumed that each person image has only one corresponding text description, while other correspondences remain unknown. This setup alleviates the challenge of cross-camera labeling for person images. However, relying solely on a single correspondence for training limits model performance. Therefore, in weakly supervised settings, constructing cross-modal relationships between samples and guiding model training accordingly is crucial. The earliest CMMT framework [45] establishes cross-modal correspondences through clustering and improves accuracy via mutual training and relationship optimization. CPCL [47] leverages CLIP's strong cross-modal matching capabilities while using clustering to construct sample correspondences. However, these methods generally rely on clustering algorithms, such as DBSCAN [8], which makes it difficult to obtain precise cross-modal relationships, limiting model performance. Unlike these approaches, this paper introduces a mechanism that integrates intra-batch and global relationship construction to enhance the accuracy of cross-modal associations. By emphasizing samples with weak associations during training, it improves the model's ability to distinguish difficult cases, ensuring robustness.

3 Proposed Method

3.1 Problem Definition and Method Overview

For the training set $\mathcal{D} = (I_i, T_i)_{i=1}^{N}$, where T_i represents the i-th textual description, I_i is the corresponding pedestrian image, and N is the total number of image-text pairs. In the weakly supervised setting, a text-image pair (I_i, T_j) shares the same pedestrian identity if $i = j$. When $i \neq j$, the identity correspondence between image I_i and text T_j is unknown. The key challenge in this setting lies in mining potential identity relationships among cross-modal samples based on known image-text associations. The overall architecture of the proposed method is illustrated in Figure 2. It consists of three key modules working collaboratively: Basic Feature Extraction (BFE), Global Sample Relationship Construction and Feature Refinement (GSRC-FR), and Information-Asymmetric Sample Pair Construction and Consistency Learning (IASC-CL). The BFE module integrates a text encoder and an image encoder, while reinforcing identity-level association constraints between the two modalities by explicitly constructing cross-modal relationships within a batch. This module establishes a solid foundation for GSRC-FR by jointly optimizing

intra-modal feature representations and inter-modal identity alignment.

The GSRC-FR module automatically identifies sample pairs with clear identity matches based on global cross-modal correspondence and refines the features accordingly. The IASC-CL module addresses the absence of cross-modal hard samples by introducing an information perturbation strategy. Specifically, it applies operations such as feature masking and local region blocking to either the image or text modality to construct sample pairs with differing information across modalities. A contrastive learning paradigm is then employed to ensure the model maintains semantic consistency despite the introduced asymmetry. This mechanism effectively circumvents the difficulty of mining cross-modal hard samples and significantly enhances the model's robustness in complex matching scenarios. Through the cascaded collaboration of these three modules, our method provides a discriminative and robust solution for cross-modal representation learning while fully leveraging the supervision from simple samples.

3.2 Basic Feature Extraction

Feature Encoder. For an input image I_i, we employ the CLIP [24] pre-trained ViT-B/16 encoder to extract its global feature representation, following a mechanism similar to that of ViT [7]. The resulting global image feature is denoted as $\boldsymbol{f}_i^v \in \mathbb{R}^{1\times d}$. Likewise, for a given text description T_i, we use the CLIP pre-trained Transformer [31] encoder to extract its global feature, denoted as $\boldsymbol{f}_i^t \in \mathbb{R}^{1\times d}$. To enforce cross-modal consistency between $\boldsymbol{f}_i^v$ and $\boldsymbol{f}_i^t$, we adopt a contrastive learning objective [22] that incorporates both a text-to-image loss $\mathcal{L}_{t2v}$ and an image-to-text loss $\mathcal{L}_{v2t}$. The text-to-image loss is formulated as:

$$\mathcal{L}_{t2v} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\text{sim}(\boldsymbol{f}_i^t,\boldsymbol{f}_i^v)/\tau)}{\sum_{k=1}^{B}\exp(\text{sim}(\boldsymbol{f}_i^t,\boldsymbol{f}_k^v)/\tau)} \tag{1}$$

where τ is a temperature coefficient that controls the sharpness of the probability distribution, B denotes the batch size, and $\text{sim}(\boldsymbol{a},\boldsymbol{b})$ denotes the cosine similarity between vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. Similarly, the image-to-text loss $\mathcal{L}_{v2t}$ is defined in the same way. The total contrastive loss is given by:

$$\mathcal{L}_{itc} = \mathcal{L}_{t2v} + \mathcal{L}_{v2t} \tag{2}$$

Local Relationship Construction. In a weakly supervised setting, knowing only the correspondence between a single text and its paired image is insufficient for training a high-performance and robust model. To address this issue, an effective approach is to leverage the relationship between a given text and its corresponding image to mine potential identity correspondences across modalities. To this end, this paper proposes an intra-batch cross-modal relationship (i.e. local relationship) construction mechanism. Specifically, for a batch of B paired samples, we compute the intra-modal self-similarity matrices $\boldsymbol{M}^v$ and $\boldsymbol{M}^t$. The element at position (i,j) in $\boldsymbol{M}^v$ is defined as $\boldsymbol{M}^v(i,j) = \text{sim}(\boldsymbol{f}_i^v, \boldsymbol{f}_j^v)$. Similarly, the elements of $\boldsymbol{M}^t$ are given by $\boldsymbol{M}^t(i,j) = \text{sim}(\boldsymbol{f}_i^t, \boldsymbol{f}_j^t)$, where $i, j \in \{1, \cdots, B\}$.

Through the similarity matrices $\boldsymbol{M}^v$ and $\boldsymbol{M}^t$, we can initially infer identity relationships among samples within the same modality. However, due to the influence of identity-irrelevant factors, $\boldsymbol{M}^v$ and $\boldsymbol{M}^t$ may not accurately reflect identity relationships. To address this, we refine $\boldsymbol{M}^v$ and $\boldsymbol{M}^t$ as follows:

$$\tilde{\boldsymbol{M}}^k(i,j) = \begin{cases} 1 & \text{if } \boldsymbol{M}^k(i,j) > \text{th} \\ 0 & \text{if } \boldsymbol{M}^k(i,j) \le \text{th} \end{cases} \tag{3}$$

where $k \in \{v,t\}$ and th is a threshold, set to 0.7 in this paper. Based on $\tilde{\boldsymbol{M}}^k$, we obtain the cross-modal identity correspondence matrix:

$$\boldsymbol{M}^{v,t} = \tilde{\boldsymbol{M}}^v \odot \tilde{\boldsymbol{M}}^t \tag{4}$$

where $\odot$ denotes the Hadamard product. $\boldsymbol{M}^{v,t}$ integrates both intra-modal and cross-modal correlations among samples within a batch. When $\boldsymbol{M}^{v,t}(i,j) = 1$, it indicates a consistent identity correspondence between the i-th image and the j-th text; otherwise, no such correspondence exists. By establishing cross-modal correspondences, the proposed method not only fully utilizes the known correspondence between the i-th image and the i-th text but also incorporates intra-modal relationships, thereby enhancing the reliability of cross-modal relationship prediction.

To mitigate the significant negative impact of mismatched sample pairs on model training, we propose a relationship softening mechanism to reduce their adverse effects:

$$\boldsymbol{Q} = \boldsymbol{I} + \lambda(\boldsymbol{M}^{v,t} - \boldsymbol{I}) \tag{5}$$

where $\boldsymbol{I}$ is the identity matrix and $0 < \lambda < 1$ serves as a balancing factor that controls the influence of mismatched correspondences during training. To further emphasize strong identity correspondences while suppressing weak ones, we refine $\boldsymbol{Q}$ as:

$$\boldsymbol{Q}(i,j) = \begin{cases} -\infty & \text{if } \boldsymbol{Q}(i,j) = 0 \\ \boldsymbol{Q}(i,j) & \text{otherwise} \end{cases} \tag{6}$$

Based on the updated $\boldsymbol{Q}$, we define the target similarity probability $q_{i,j}$ as:

$$q_{i,j} = \frac{\exp(\boldsymbol{Q}(i,j))}{\sum_{k=1}^{B}\exp(\boldsymbol{Q}(i,k))} \tag{7}$$

The cross-modal similarity probability between the i-th image and the j-th text within a batch is given by:

$$p_{i,j} = \frac{\exp\left(\text{sim}\left(\boldsymbol{f}_i^v,\boldsymbol{f}_j^t\right)/\tau\right)}{\sum_{k=1}^{B}\exp\left(\text{sim}\left(\boldsymbol{f}_i^v,\boldsymbol{f}_k^t\right)/\tau\right)} \tag{8}$$

Guided by $q_{i,j}$, we optimize the representations $\boldsymbol{f}_i^v$ and $\boldsymbol{f}_i^t$ by aligning the similarity distribution $p_{i,j}$ with the target distribution $q_{i,j}$, thereby enhancing cross-modal matching capability. To achieve this, we introduce a similarity distribution matching (SDM) loss [11, 44] to optimize the model parameters:

$$\mathcal{L}_{rc\_b}^v = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B} p_{i,j}\log\left(\frac{p_{i,j}}{q_{i,j}+\varepsilon}\right) \tag{9}$$

where ε is a small constant added for numerical stability. Similarly, the within-batch relationship construction loss for text-to-image is defined as $\mathcal{L}_{rc\_b}^t$. The total local relationship construction loss is given by:

$$\mathcal{L}_{rc\_b} = \mathcal{L}_{rc\_b}^v + \mathcal{L}_{rc\_b}^t \tag{10}$$

3.3 Global Relationship Construction and Feature Refinement

Due to the limited number of samples within a single batch, there are relatively few cross-modal sample pairs with the same identity. To address this issue, we propose a GSRC-FR method. This

approach constructs sample pairs with explicit cross-modal correspondences by leveraging the precise relationships between the current sample and same-modality samples from the entire training set. Since the sample relationships are built upon the entire dataset, we refer to this as a global sample construction method. To construct global sample relationships, we use the person image from the current batch as a visual anchor to measure its relevance with all samples in the dataset. For efficient implementation, we maintain Memory Banks for both modalities to store the representations of all training samples. Let $\mathcal{B}^v = \{\hat{\boldsymbol{f}}_i^v\}_{i=1}^N$ and $\mathcal{B}^t = \{\hat{\boldsymbol{f}}_i^t\}_{i=1}^N$ denote the Memory Banks for the image and text modalities, respectively. These Memory Banks are initialized using the representations $\boldsymbol{f}_i^v$ and $\boldsymbol{f}_i^t$ extracted by the encoders and are updated according to Eq. (11). Taking the image Memory Bank as an example, the update rule is defined as:

$$\hat{\boldsymbol{f}}_{i,n}^v = \alpha \boldsymbol{f}_i^v + (1-\alpha)\hat{\boldsymbol{f}}_{i,n-1}^v \tag{11}$$

where $\boldsymbol{f}_i^v$ is the newly extracted representation, $\hat{\boldsymbol{f}}_{i,n-1}^v$ is the stored representation from the previous epoch, $\hat{\boldsymbol{f}}_{i,n}^v$ is the updated representation after the n-th epoch, and $0 < \alpha < 1$ is a smoothing factor that mitigates abrupt changes in the stored features.

Assume $\boldsymbol{M}_m^v(i,j) = \text{sim}(\boldsymbol{f}_i^v, \hat{\boldsymbol{f}}_{j,n-1}^v)$, where $\boldsymbol{M}_m^v(i,j)$ denotes the similarity between the feature representation of the i-th pedestrian image in the current batch and that of the j-th image sample in the Memory Bank. Compared to textual descriptions, pedestrian images contain richer appearance information, which plays a more critical role in determining pedestrian identity. Therefore, we use the relationships among pedestrian images as a bridge to establish cross-modal associations in global relation modeling. For each row of $\boldsymbol{M}_m^v$, we select the indices corresponding to the top k highest similarity scores. However, these top-k samples may still include incorrectly matched instances. To mitigate this issue, we apply Eq. (12) to filter out false matches from the top-k candidates. The final set of selected indices is defined as:

$$\boldsymbol{J}_i = \{j \mid \boldsymbol{M}_m^v(i,j) > \text{th}\}, \quad i \in \{1,\ldots,B\} \tag{12}$$

where th is a threshold value, which is set to 0.7. The pedestrian identities of the image samples indexed by $\boldsymbol{J}_i$ are regarded as being the same as that of the i-th image. We propagate this identity information to the text associated with the i-th image, thereby associating a single text with multiple pedestrian images. Meanwhile, we transfer the pedestrian identities carried by the texts corresponding to the pedestrian images in $\boldsymbol{J}_i$ back to the i-th image, thus linking a pedestrian image with multiple pedestrian texts.

To effectively utilize cross-modal identity correspondences, we compute a similarity matrix $\boldsymbol{S}^{t,v}$ between the texts and pedestrian images within a batch, where $\boldsymbol{S}^{t,v}(i,j) = \text{sim}(\boldsymbol{f}_i^t, \boldsymbol{f}_j^v)$ denotes the similarity between the i-th text and the j-th pedestrian image. In addition, we calculate the similarity between each text in the batch and the selected pedestrian images from the Memory Bank as $\boldsymbol{S}^{t,m}(i,j) = \text{sim}(\boldsymbol{f}_i^t, \hat{\boldsymbol{f}}_{j,n-1}^v)$, where j indexes samples in the sets $\{\boldsymbol{J}_1,\ldots,\boldsymbol{J}_B\}$. By concatenating $\boldsymbol{S}^{t,v}$ and $\boldsymbol{S}^{t,m}$ along the column dimension, we obtain $\boldsymbol{S}' = \text{concat}(\boldsymbol{S}^{t,v}, \boldsymbol{S}^{t,m}) \in \mathbb{R}^{B\times B_1}$, where B_1 denotes the total number of pedestrian images in the current batch and the selected ones from the Memory Bank. Based on $\{\boldsymbol{J}_1,\ldots,\boldsymbol{J}_B\}$ and $\boldsymbol{S}'$, we construct a set $\boldsymbol{J}'$, which contains the column indices in $\boldsymbol{S}'$ corresponding to pedestrian images that share identity relationships with each text. Using $\boldsymbol{J}'$, we define the target relationship matrix $\boldsymbol{Q}' \in \mathbb{R}^{B\times B_1}$ as follows:

$$\boldsymbol{Q}'(i,j) = \begin{cases} 1 & \text{if } j \in \boldsymbol{J}_i' \\ -\infty & \text{otherwise} \end{cases} \tag{13}$$

However, directly optimizing the model using $\boldsymbol{Q}'$ may cause it to focus predominantly on high-confidence sample pairs, while ignoring pairs with weaker but potentially informative correspondences. This could hinder the learning of more robust and generalizable representations. To address this issue, we propose an adaptive weight adjustment mechanism to regulate the influence of different sample pairs on the model. Specifically, for $j \in \boldsymbol{J}_i'$ and $j \neq i$, we update $\boldsymbol{Q}'(i,j)$ as follows:

$$\boldsymbol{Q}'(i,j) = 1 - \frac{\exp(\text{sim}(\boldsymbol{f}_i^v, \boldsymbol{f}_j^v))}{\sum_{k\in\boldsymbol{J}_i'} \exp(\text{sim}(\boldsymbol{f}_i^v, \boldsymbol{f}_k^v))}. \tag{14}$$

To strengthen the associations between strongly correlated identity samples while suppressing the impact of weakly related pairs, we refine $\boldsymbol{Q}'$ using the weighting strategy defined above. With the updated $\boldsymbol{Q}'$, we compute the target similarity probability $q_{i,j}'$ as:

$$q_{i,j}' = \frac{\exp(\boldsymbol{Q}'(i,j))}{\sum_{k\in\boldsymbol{J}_i'} \exp(\boldsymbol{Q}'(i,k))}. \tag{15}$$

Within the same batch, the predicted matching probability between the i-th text and the j-th image is given by:

$$p_{i,j}' = \frac{\exp(\boldsymbol{S}'(i,j)/\tau)}{\sum_{k=1}^{B_1} \exp(\boldsymbol{S}'(i,k)/\tau)}. \tag{16}$$

Guided by $q_{i,j}'$, we refine the original within-batch representations $\boldsymbol{f}_i^v$ and $\boldsymbol{f}_i^t$ to incorporate richer identity-related information, thereby enhancing their cross-modal matching capability. To achieve this, we employ the SDM loss to optimize the model parameters:

$$\mathcal{L}_{rc\_g}^t = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B_1} p_{i,j}' \log\left(\frac{p_{i,j}'}{q_{i,j}' + \epsilon}\right). \tag{17}$$

Similarly, we define the global relation construction and representation refinement loss from image to text as $\mathcal{L}_{rc\_g}^v$. The final cross-modal global relation consistency loss is formulated as:

$$\mathcal{L}_{rc\_g} = \mathcal{L}_{rc\_g}^v + \mathcal{L}_{rc\_g}^t. \tag{18}$$

3.4 Information Asymmetric Sample-Pair Construction and Consistency Learning

Relying solely on the relationships within simple cross-modal sample pairs (i.e., between text and images) constructed in GSRC-FR is insufficient to address the complex and diverse cross-modal matching scenarios encountered in real-world applications. Directly mining cross-modal sample pairs from the training set for model training cannot ensure the correctness of the correspondences. To overcome this challenge, we propose the IASC-CL method to improve the model's cross-modal matching capability for hard samples. Specifically, given an image-text pair (I_i, T_i), we define an image augmentation operator as:

$$\mathcal{A}^v(I_i) = \text{Augment}(I_i) \tag{19}$$

In Eq. (19), the operator applies a series of variationsto the image, including horizontal flipping, border padding, random cropping, and random erasing. These augmentations simulate appearance variations caused by occlusion, viewpoint shifts, and other real-world factors. We further define a text augmentation operator as:

$$\mathcal{A}^t(T_i) = \text{Mask}(T_i) \tag{20}$$

Here, Mask represents a masking operation that replaces parts of T_i with [MASK], simulating information loss due to incomplete descriptions or varying levels of granularity in the text.

The image-text pair constructed through Eqs. (19) and (20), namely $(\mathcal{A}^v(I_i), \mathcal{A}^t(T_i))$, exhibits weaker explicit semantic correlation compared to the original pair (I_i, T_i), while still preserving underlying identity consistency. We therefore refer to these as information-asymmetric sample pairs, which are designed to simulate the complex and diverse cross-modal matching scenarios encountered in real-world applications. By exposing the model to such information-asymmetric pairs during training, it learns from more varied and challenging relationships, thereby enhancing its cross-modal matching capability in difficult cases. Similar to the within-batch cross-modal relationship construction loss, the consistency constraint for information-asymmetric sample pairs is formulated as:

$$\mathcal{L}_{rc\_h} = \mathcal{L}^v_{rc\_h} + \mathcal{L}^t_{rc\_h} \tag{21}$$

The overall training loss is defined as:

$$\mathcal{L} = \mathcal{L}_{itc} + \mathcal{L}_{rc\_b} + \mathcal{L}_{rc\_g} + \mathcal{L}_{rc\_h} \tag{22}$$

This formulation enables the model to not only handle straightforward matching cases, but also effectively tackle more complex scenarios involving information asymmetry, thereby improving its overall performance and discriminative capability.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. To evaluate the performance of our model, we use three widely used text-to-image person retrieval datasets: CUHK-PEDES [18], ICFG-PEDES [6], and RSTPReid [49].

CUHK-PEDES. The dataset comprises 40,206 person images and 80,412 text descriptions corresponding to 13,003 identities. Each image is paired with at least two text descriptions, with an average length of at least 23 words. The dataset is divided into training, validation, and test sets. The training set contains 11,003 identities (34,054 images, 68,108 text descriptions), the validation set includes 1,000 identities (3,078 images, 6,158 text descriptions), and the test set consists of 1,000 identities (3,074 images, 6,156 text descriptions).

ICFG-PEDES. The dataset consists of 54,522 images corresponding to 4,102 identities, with each image associated with a single text description. Compared to CUHK-PEDES, the text descriptions in ICFG-PEDES are more detailed, averaging 37 words in length. The dataset is split into a training set (34,674 image-text pairs, 3,102 identities) and a test set (19,848 image-text pairs, 1,000 identities).

RSTPReid. The dataset is specifically designed for real-world applications and includes 20,505 images corresponding to 4,101 identities. Each identity has five images captured from different cameras, each paired with two text descriptions of at least 23 words. The training, validation, and test sets contain 3,701, 200, and 200 identities, respectively, corresponding to 18,505, 1,000, and 1,000 images, and 37,010, 2,000, and 2,000 text descriptions, respectively.

Evaluation Metrics. To evaluate model performance, we use the Rank metric of the Cumulative Match Characteristic (CMC) [33] for retrieval assessment. Additionally, we employ mean Average Precision (mAP) [46] and mean Inverse Negative Penalty (mINP) [43] as supplementary retrieval metrics. Higher values for these metrics indicate better retrieval performance.

4.2 Implementation Details

The proposed method is implemented using the PyTorch framework, and all experiments are conducted on a single RTX 4090 GPU. The image encoder employs the CLIP-pretrained ViT-B/16 model, while the text encoder utilizes the CLIP-pretrained text Transformer. All input images are resized to 384×128 pixels, and the maximum text token sequence length is set to 77. The batch size is set to 64, and training is conducted for 40 epochs using the Adam [13] optimizer. The initial learning rate is set to 1×10^{-5}, following a cosine learning rate decay strategy. A 5-epoch warm-up [20] period is applied at the beginning of training, during which the learning rate linearly increases from 1×10^{-6} to 1×10^{-5}. The temperature parameter τ is fixed at 0.02.

Table 1: Performance comparison of methods (%) on the CUHK-PEDES dataset. The best performance in weakly supervised methods is highlighted in bold.

Settings	Methods	R1	R5	R10	mAP	mINP
Supervised	MANet [41]	65.64	83.01	88.78	–	–
	LCR^{2}S [38]	67.36	84.19	89.62	59.24	–
	UniPT [25]	68.50	84.67	90.38	–	–
	Cfine [39]	69.57	85.93	91.15	–	–
	IA-PTIM [14]	70.61	86.55	91.65	–	–
	IRRA [11]	73.38	83.93	93.71	66.13	50.24
	LSPM [16]	74.38	89.51	93.42	67.74	53.09
	PP-ReID[40]	74.89	89.90	94.17	67.12	–
	PLOT [23]	75.28	90.42	94.12	–	–
	APTM [42]	76.53	90.04	94.15	66.91	–
	LAIP [36]	76.72	90.42	93.60	66.05	–
	TriMatch [37]	76.84	89.90	94.17	67.12	–
	PFM-EKFP [15]	77.24	93.71	96.98	73.47	–
	LUP-MLLM [30]	78.13	91.19	94.50	68.75	–
Unsupervised	GTR [2]	47.53	68.23	75.91	42.91	–
	GAAP [19]	47.64	67.79	76.08	41.28	–
Weakly Supervised	CMMT [45]	57.10	78.14	85.23	–	–
	CPCL [47]	70.03	87.28	91.78	63.19	47.54
	Ours	**73.06**	**89.21**	**93.44**	**64.88**	**48.15**

4.3 Comparison with State-of-the-Art Methods

To validate the effectiveness of our method, we conduct comparisons with state-of-the-art (SOTA) methods in the supervised, weakly supervised, and unsupervised settings.

Comparison with Supervised Methods. As shown in Tables 1-3, our weakly supervised method achieves competitive performance without relying on costly manual identity annotations. Although it still lags behind the most advanced supervised methods, it has already matched or even outperformed some recently published supervised approaches. This result demonstrates that our method effectively balances performance and annotation cost, making it more practical for real-world applications with limited annotation resources or budget constraints.

Comparison with Unsupervised Methods. In the unsupervised setting, we omit manually annotated text and rely solely on

Table 2: Performance comparison of methods (%) on the ICFG-PEDES dataset. The best performance in weakly supervised methods is highlighted in bold.

Labeled	Method	R1	R5	R10	mAP	mINP
Supervised	LCR^{2}S[38]	57.93	76.08	82.40	38.21	–
	MANet[41]	59.44	76.80	85.75	–	–
	IA-PTIM[14]	59.82	77.05	83.02	–	–
	UniPT[25]	60.09	76.19	82.46	–	–
	Cfine[39]	60.83	76.55	82.42	–	–
	IRRA[11]	63.46	80.25	85.82	38.06	7.93
	LAIP[36]	63.52	79.28	84.57	37.02	–
	LSPM[16]	64.40	79.96	85.41	42.60	11.65
	PP-ReID[40]	65.12	81.57	86.97	42.93	–
	PLOT[23]	65.76	81.39	86.73	–	–
	TriMatch[37]	67.71	85.37	88.02	–	–
	APTM[42]	68.51	82.99	87.56	41.22	–
	PFM-EKFP[15]	69.29	89.10	94.06	47.15	–
	LUP-MLLM[30]	69.37	83.55	88.18	42.42	–
Unsupervised	GAAP[19]	27.12	44.91	53.56	11.43	–
	GTR[2]	28.25	45.21	53.51	13.82	–
Weakly Supervised	CPCL[47]	62.60	79.07	84.46	36.16	**6.31**
	Ours	**63.71**	**79.90**	**85.38**	**36.87**	6.20

image information, leveraging a large visual-language model (VLM) to generate text descriptions for training. During testing, we still use standard test sets with manually annotated text. GTR [2] and GAAP [19] are representative works in this field, aiming to optimize pseudo-text quality, but their performance remains limited. Our weakly supervised method outperforms the best unsupervised approaches by 25.42%, 35.46%, and 15.70% in Rank-1 accuracy on the CUHK-PEDES, ICFG-PEDES, and RSTPReid datasets, respectively. This significant advantage highlights the limitations of current unsupervised methods and their gap from practical application goals. It further supports our view that weakly supervised methods offer a better trade-off between performance and cost-effectiveness at this stage.

Table 3: Performance Comparison of Methods (%) on the RSTPReid dataset. The best performance in weakly supervised methods is highlighted in bold.

Labeled	Method	Performance Metrics				
		R1	R5	R10	mAP	mINP
Supervised	Cfine[39]	50.55	72.50	81.60	–	–
	UniPT[25]	51.85	74.85	82.85	–	–
	LCR^{2}S [38]	54.95	76.65	84.70	40.92	–
	IRRA[11]	60.20	81.30	88.20	47.17	25.28
	PLOT[23]	61.80	82.85	89.45	–	–
	PP-ReID[40]	61.87	83.63	89.70	47.82	–
	LAIP[36]	62.00	83.15	88.50	45.27	–
	IA-PTIM[14]	65.50	86.50	91.00	–	–
	TriMatch[37]	67.40	85.85	90.95	–	–
	APTM[42]	67.50	85.70	91.45	52.56	–
	LUP-MLLM[30]	69.95	87.35	92.30	54.17	–
Unsupervised	GAAP[19]	44.45	65.15	75.30	31.21	–
	GTR[2]	45.60	70.35	79.95	33.30	–
Weakly Supervised	CPCL[47]	58.35	81.05	87.65	45.81	23.87
	Ours	**61.30**	**82.00**	**88.10**	**47.20**	**25.47**

Comparison with Weakly Supervised Methods. CMMT [45] is a pioneering work in text-based person re-identification under weakly supervised conditions and reports results only on the CUHK-PEDES dataset. Therefore, we first validate our method on this benchmark. As shown in Table 1, our method significantly outperforms existing weakly supervised methods, including CMMT [45] and CPCL [47], achieving a Rank-1 accuracy of 73.06% and an mAP of 64.88%. Compared to CPCL, our method improves Rank-1 accuracy and mAP by 3.03% and 1.69%, respectively.

Next, we evaluate our method on the ICFG-PEDES dataset, which contains more fine-grained text descriptions and imposes higher demands on text understanding and cross-modal matching. As shown in Table 2, our method maintains strong performance, achieving a Rank-1 accuracy of 63.71% and an mAP of 36.87%, demonstrating its capability to handle complex and challenging text descriptions. Finally, on the diverse and realistic RSTPReid dataset, our method achieves a Rank-1 accuracy of 61.30% and an mAP of 47.20%. Compared to CPCL, it improves Rank-1 accuracy and mAP by 2.95 and 1.39 percentage points, respectively.

4.4 Ablation Study

In our method, we adopt the CLIP-ViT-B/16 model [24] and fine-tune it using the loss function defined in Eq. (2) as the baseline for our approach. Our method primarily consists of three modules: BFE, GSRC-FR, and IASC-CL, with local relationship construction (LRC) being the core component of BFE. To validate the effectiveness of LRC, GSRC-FR, and IASC-CL, we conduct a series of ablation studies, and the results are presented in Table 4. The experimental results show that when LRC is added to the baseline model, the Rank-1 accuracy on the CUHK-PEDES, ICFG-PEDES, and RSTPReid datasets improves by 4.35%, 4.21%, and 1.70%, respectively. Similarly, when GSRC-FR is added to the baseline model, the Rank-1 accuracy increases by 2.87%, 4.27%, and 1.30%, respectively. These results strongly demonstrate the effectiveness of LRC and GSRC-FR.

Table 4: Ablation study on CUHK-PEDES, ICFG-PEDES, and RSTPReid datasets. 'B' denotes the baseline method. Reported performance is Rank-1 accuracy.

Settings	CUHK-PEDES	ICFG-PEDES	RSTPReid
B	64.69	54.24	54.60
B+LRC	69.04	58.45	56.30
B+GSRC-FR	67.56	58.51	55.90
B+LRC+GSRC-FR	70.20	60.30	57.25
B+LRC+GSRC-FR+IASC-CL	73.06	63.71	61.30

Furthermore, when GSRC-FR is added to the B+LRC model, the resulting model, B+LRC+GSRC-FR, achieves Rank-1 accuracy improvements of 5.51%, 6.06%, and 2.65% on the three datasets compared to the baseline model, and improvements of 1.16%, 1.85%, and 0.95% compared to the B+LRC model. This indicates that GSRC-FR effectively complements the limitations of LRC and further enhances the model's performance. To verify the effectiveness of IASC-CL, we add it to the B+LRC+GSRC-FR model, resulting in the final model, B+LRC+GSRC-FR+IASC-CL. As shown in Table 4, compared to the B+LRC+GSRC-FR model, B+LRC+GSRC-FR+IASC-CL achieves Rank-1 accuracy improvements of 2.86%, 3.41%, and 4.05% on the three datasets, respectively. Compared to the baseline model, the final model B+LRC+GSRC-FR+IASC-CL achieves Rank-1 accuracy improvements of 8.37%, 9.47%, and 6.70% on the CUHK-PEDES, ICFG-PEDES, and RSTPReid datasets, respectively. These results fully demonstrate the effectiveness of each module and the advantages of their combined use.

Table 5: Ablation study about BFE on CUHK-PEDES, ICFG-PEDES and RSTPReid datasets. 'B' denotes the baseline method. Reported performance is Rank-1 accuracy.

Settings	CUHK-PEDES	ICFG-PEDES	RSTPReid
B	64.69	54.24	54.60
B+LRCI	67.56	51.32	53.51
B+LRCT	67.61	56.47	55.45
B+LRC	69.04	58.45	56.30

4.5 Further Discussion

Further Analysis of BFE. In LRC, we integrate information from both image and text modalities based on Eq. (4) to determine cross-modal identity relationships within a batch. To verify the effectiveness of this integration approach, we conduct comparative experiments by constructing models that use only image information (LRCI) or only text information (LRCT) and evaluate their performance. As shown in Table 5, when relying solely on single-modal relationships to establish cross-modal identity correspondences within a batch, although the model outperforms the baseline, its performance decreases to varying extents across different datasets compared to the B+LRC method, which utilizes both modalities. This is mainly because relying solely on single-modal relationships for cross-modal identity matching often results in inaccuracies and incompleteness, thereby limiting further performance improvements.

Table 6: Ablation study about GSRC-FR on CUHK-PEDES, ICFG-PEDES and RSTPReid datasets. Reported performance is Rank-1 accuracy.

Settings	CUHK-PEDES	ICFG-PEDES	RSTPReid
BFE	69.04	58.45	56.30
BFE+GSRC-T	67.82	57.42	56.80
BFE+GSRC-IT	68.73	58.22	56.90
BFE+GSRC-FR	70.20	60.30	57.25

Further Analysis of GSRC-FR. Global relation construction serves as the core component of the GSRC-FR module. To evaluate its effectiveness, we compare its performance under three settings: a model using only BFE, a model constructing global identity relations using only text information (BFE+GSRC-T), and a model utilizing both image and text information (BFE+GSRC-IT). As shown in Table 6, on the CUHK-PEDES and ICFG-PEDES datasets, the BFE+GSRC-T model performs worse than BFE, mainly because text information is less informative than pedestrian images, making it ineffective in constructing identity relations and thus hindering performance improvement. Compared to our proposed method, the BFE+GSRC-IT model fails to improve performance and instead suffers a performance drop. This is mainly because, while combining both image and text modalities helps identify more accurate identity relations, it also excludes training samples with uncertain relations, thereby reducing the model's generalization ability.

To further assess the effectiveness of GSRC-FR, we evaluate the accuracy of identity relations constructed by this module, using the ratio of correctly identified identities to the total number of mined identities as the metric. Figure 4 presents the experimental results. As the number of epoch increases, the accuracy of cross-modal

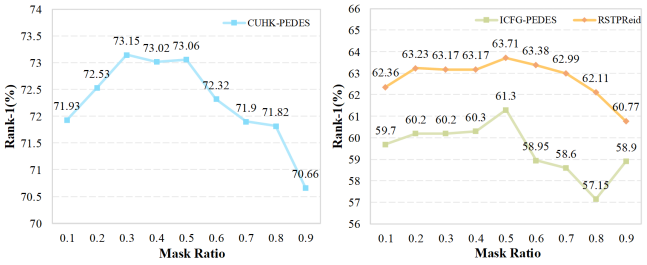

Figure 3: Analysis of the impact of masking ratio on model performance. (a) Effect of different masking ratios on model performance on the CUHK-PEDES dataset. (b) Effect of different masking ratios on model performance on the ICFG-PEDES and RSTPReid datasets.

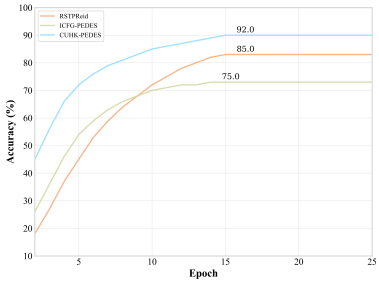

Figure 4: Accuracy of cross-modal relation constructed by GSCS-FR.

identity relations predicted by BFE+GSRC-FR on the CUHK-PEDES dataset reaches 92.0%. Even on the ICFG-PEDES dataset, where performance is relatively lower, the accuracy still reaches 75.0%, which further demonstrates the effectiveness and reliability of the GSRC-FR module.

Further Analysis of IASC-CL. In IASC-CL, applying different masking ratios to text affects performance variably. To find the optimal ratio, we experimented with several values. As shown in Figure 3, on CUHK-PEDES, performance peaks when the ratio is between 0.3–0.5, while on ICFG-PEDES and RSTPReid, the optimal range is 0.4–0.5. Based on this, we set the final masking ratio to 0.5.

5 Conclusion

This paper proposes a local-and-global dual-granularity identity association mechanism that effectively enhances text-to-person image matching under weakly supervised settings. By explicitly establishing cross-modal associations within each batch, identity constraints across modalities are reinforced, supporting global relationship reasoning. The dynamic association strategy, guided by visual information, alleviates ambiguity caused by abstract textual semantics, improving robustness. Additionally, the cross-modal confidence-based adaptive adjustment mechanism boosts sensitivity to weak associations, while the information-asymmetric sample pair construction effectively avoids the challenge of hard sample mining. Experimental results confirm the proposed method's superiority and robustness on challenging benchmarks, promoting progress in intelligent security and smart city applications. Future work will explore multi-modal dynamic weight adaptation and cross-scenario generalization to further enhance practical deployment.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62276120 and Grant 61966021, the Yunnan Fundamental Research Projects under Grant 202401AS070106, Grant 202301AV070004 and Grant 202501AS070123.

References

[1] Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. 2023. RaSa: Relation and Sensitivity Aware Representation Learning for Text-based Person Search. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. 555–563.

[2] Yang Bai, Jingyao Wang, Min Cao, Chen Chen, Ziqiang Cao, Liqiang Nie, and Min Zhang. 2023. Text-based person search without parallel image-text data. In *Proceedings of the 31st ACM International Conference on Multimedia*. 757–767.

[3] Tianlang Chen, Chenliang Xu, and Jiebo Luo. 2018. Improving text-based person search by spatial matching and adaptive threshold. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1879–1887.

[4] Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, and Yuhui Zheng. 2022. TIPCB: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing* 494 (2022), 171–181.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

[6] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. 2021. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666* (2021).

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR* (2021).

[8] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 226–231.

[9] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Su. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[11] Ding Jiang and Mang Ye. 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.

[12] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. 2020. Pose-guided multi-granularity attention network for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11189–11196.

[13] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

[14] Fan Li, Hang Zhou, Huafeng Li, Yafei Zhang, and Zhengtao Yu. 2025. Person text-image matching via text-feature interpretability embedding and external attack node implantation. *IEEE Transactions on Emerging Topics in Computational Intelligence* 9, 2 (2025), 1202–1215.

[15] Huafeng Li, Shedan Yang, Yafei Zhang, Dapeng Tao, and Zhengtao Yu. 2025. Progressive Feature Mining and External Knowledge-Assisted Text-Pedestrian Image Retrieval. *IEEE Transactions on Multimedia* 27 (2025), 1973–1987.

[16] Jiayi Li, Min Jiang, Jun Kong, Xuefeng Tao, and Xi Luo. 2024. Learning semantic polymorphic mapping for text-based person retrieval. *IEEE Transactions on Multimedia* 26 (2024), 10678–10691.

[17] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*. 9694–9705.

[18] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person Search with Natural Language Description. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 5187–5196.

[19] Zongyi Li, Jianbo Li, Yuxuan Shi, Hefei Ling, Jiazhong Chen, Runsheng Wang, and Shijuan Huang. 2024. Cross-modal generation and alignment via attribute-guided prompt for unsupervised text-based person retrieval. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. 1047–1055.

[20] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 1487–1495.

[21] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. 2020. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing* 29 (2020), 5542–5556.

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[23] Jicheol Park, Dongwon Kim, Boseung Jeong, and Suha Kwak. 2024. PLOT: Text-based person search with part slot attention for corresponding part discovery. In *Computer Vision – ECCV 2024: 18th European Conference*. Springer, 474–490.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 8748–8763.

[25] Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang. 2023. Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11174–11184.

[26] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. 2022. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*. 5566–5574.

[27] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

[28] A Venkata Subramanyam, Vibhu Dubey, Niranjan Sundararajan, and Brejesh Lall. 2023. Dense captioning for Text-Image ReID. In *Proceedings of the Fourteenth Indian Conference on Computer Vision, Graphics and Image Processing*. 1–8.

[29] Jintao Sun, Hao Fei, Zhedong Zheng, and Gangyi Ding. 2024. From Data Deluge to Data Curation: A Filtering-WoRA Paradigm for Efficient Text-based Person Search. *arXiv preprint arXiv:2404.10292* (2024).

[30] Wentan Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. 2024. Harnessing the power of mllms for transferable text-to-image person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17127–17137.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010.

[32] Chengji Wang, Zhiming Luo, Yaojin Lin, and Shaozi Li. 2021. Text-based person search via multi-granularity embedding learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. 1068–1074.

[33] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu. 2007. Shape and appearance context modeling. In *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 1–8.

[34] Yuyu Wang, Chunjuan Bo, Dong Wang, Shuang Wang, Yunwei Qi, and Huchuan Lu. 2019. Language person search with mutually connected classification loss. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2057–2061.

[35] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. 2020. Vitaa: Visual-textual attributes alignment in person search by natural language. In *Computer vision–ECCV 2020: 16th European Conference*. Springer, 402–420.

[36] Yu Wu, Haiguang Wang, Mengxia Wu, Min Cao, and Min Zhang. 2024. LAIP: learning local alignment from image-phrase modeling for text-based person search. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–10.

[37] Shuanglin Yan, Neng Dong, Shuang Li, and Huafeng Li. 2025. TriMatch: Triple Matching for Text-to-Image Person Re-Identification. *IEEE Signal Processing Letters* 32 (2025), 806–810.

[38] Shuanglin Yan, Neng Dong, Jun Liu, Liyan Zhang, and Jinhui Tang. 2023. Learning comprehensive representations with richer self for text-to-image person re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6202–6211.

[39] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. 2023. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing* 32 (2023), 6032–6046.

[40] Shuanglin Yan, Jun Liu, Neng Dong, Liyan Zhang, and Jinhui Tang. 2024. Prototypical Prompting for Text-to-image Person Re-identification. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2331–2340.

[41] Shuanglin Yan, Hao Tang, Liyan Zhang, and Jinhui Tang. 2024. Image-specific information suppression and implicit local alignment for text-based person search. *IEEE Transactions on Neural Networks and Learning Systems* 35, 12 (2024), 17973–17986.

[42] Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. 2023. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*. 4492–4501.

[43] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 6 (2021), 2872–2893.

[44] Ying Zhang and Huchuan Lu. 2018. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 686–701.

[45] Shizhen Zhao, Changxin Gao, Yuanjie Shao, Wei-Shi Zheng, and Nong Sang. 2021. Weakly supervised text-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11395–11404.

[46] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*. 1116–1124.

[47] Yanwei Zheng, Xinpeng Zhao, Chuanlin Lan, Xiaowei Zhang, Bowen Huang, Jibin Yang, and Dongxiao Yu. 2024. CPCL: Cross-Modal Prototypical Contrastive Learning for Weakly Supervised Text-based Person Re-Identification. *arXiv preprint arXiv:2401.10011* (2024).

[48] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 2 (2020), 1–23.

[49] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. 2021. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*. 209–217.